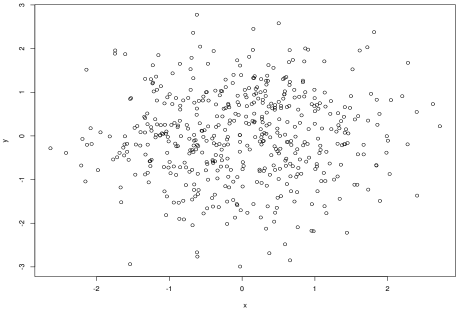
A Short Course in Linear and Logistic Regression

Suman Guha
Assistant Professor
Department of Statistics
Presidency University, Kolkata
July 26, 2020

- Observations taken on two features - say height (x) and weight (y) of individuals.

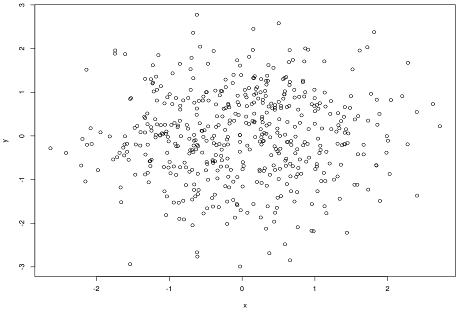
- Situations when (x) and (y) show no interrelationships - no point doing regression.



(a)

Figure: Artificially simulated dataset showing no dependence between x and y.

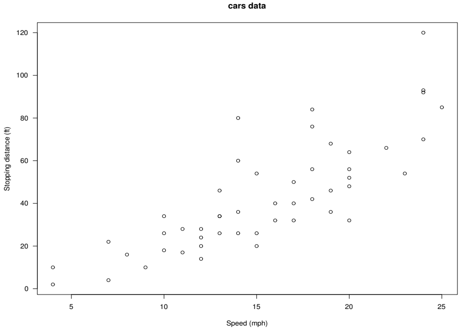
- Observations taken on two features - say height (x) and weight (y) of individuals.
- Situations when (x) and (y) show no interrelationships - no point doing regression.



(a)

Figure: Artificially simulated dataset showing no dependence between x and y.

- Fortunately, most of the time x and y turns out to be dependent!



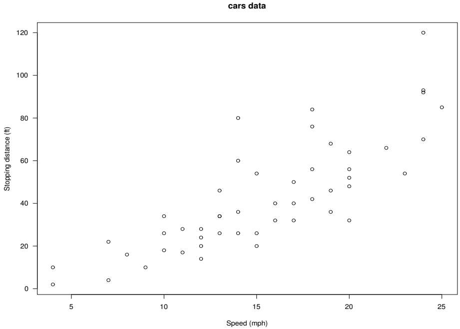
(a)

Figure: (x) car speed in miles per hour vs (y) stopping distance in feet.

- Want an approximate formula ($y \approx f(x)$) of stopping distance (y) in terms of car speed (x) - regression problem.

- Why?

- Fortunately, most of the time x and y turns out to be dependent!



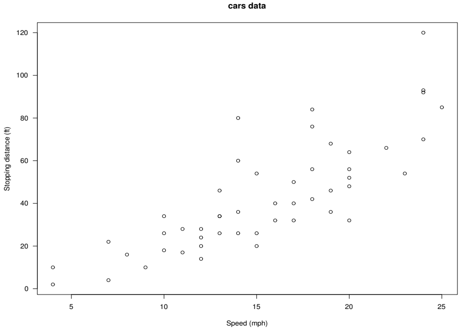
(a)

Figure: (x) car speed in miles per hour vs (y) stopping distance in feet.

- Want an approximate formula ($y \approx f(x)$) of stopping distance (y) in terms of car speed (x) - regression problem.

- Why?

- Fortunately, most of the time x and y turns out to be dependent!



(a)

Figure: (x) car speed in miles per hour vs (y) stopping distance in feet.

- Want an approximate formula ($y \approx f(x)$) of stopping distance (y) in terms of car speed (x) - regression problem.
- Why?

- To understand the nature of dependence between (x) and (y).
- Sometimes (y) may be costly/difficult to measure (total annual income) but (x) may be measured easily (total annual expenditure) - can use the formula to predict y^* using x^*.
- What type of formula? - $f(x) = ax^3 + b\sqrt{x} + c$?
- No, we want a formula of form $f(x) = a + bx$ - equation of a straight line.
-

  Reason? (i) Mathematically simple.

  (ii) Most of the time linear regression perform quite well!
- How to get the value of a, b? - a line that pass through the most middle - obtained by minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$.
- Closed form solution available - $f(x) = (\bar{y} - \frac{Cov(x,y)}{Var(x)}\bar{x}) + \frac{Cov(x,y)}{Var(x)}x = \bar{y} + \frac{Cov(x,y)}{Var(x)}(x - \bar{x})$.

- To understand the nature of dependence between (x) and (y).
- Sometimes (y) may be costly/difficult to measure (total annual income) but (x) may be measured easily (total annual expenditure) - can use the formula to predict y^* using x^*.
- What type of formula? - $f(x) = ax^3 + b\sqrt{x} + c$?
- No, we want a formula of form $f(x) = a + bx$ - equation of a straight line.
-

  Reason?(i) Mathematically simple.

  (ii) Most of the time linear regression perform quite well!
- How to get the value of a, b? - a line that pass through the most middle - obtained by minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$.
- Closed form solution available - $f(x) = (\bar{y} - \frac{Cov(x,y)}{Var(x)}\bar{x}) + \frac{Cov(x,y)}{Var(x)}x = \bar{y} + \frac{Cov(x,y)}{Var(x)}(x - \bar{x})$.

- To understand the nature of dependence between (x) and (y).
- Sometimes (y) may be costly/difficult to measure (total annual income) but (x) may be measured easily (total annual expenditure) - can use the formula to predict y^* using x^*.
- What type of formula? - $f(x) = ax^3 + b\sqrt{x} + c$?
- No, we want a formula of form $f(x) = a + bx$ - equation of a straight line.
-

Reason? (i) Mathematically simple.

(ii) Most of the time linear regression perform quite well!

- How to get the value of a, b? - a line that pass through the most middle - obtained by minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$.
- Closed form solution available - $f(x) = (\bar{y} - \frac{Cov(x,y)}{Var(x)}\bar{x}) + \frac{Cov(x,y)}{Var(x)}x = \bar{y} + \frac{Cov(x,y)}{Var(x)}(x - \bar{x})$.

- To understand the nature of dependence between (x) and (y).
- Sometimes (y) may be costly/difficult to measure (total annual income) but (x) may be measured easily (total annual expenditure) - can use the formula to predict y^* using x^*.
- What type of formula? - $f(x) = ax^3 + b\sqrt{x} + c$?
- No, we want a formula of form $f(x) = a + bx$ - equation of a straight line.
- Reason?(i) Mathematically simple.

  (ii) Most of the time linear regression perform quite well!
- How to get the value of a, b? - a line that pass through the most middle - obtained by minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$.
- Closed form solution available - $f(x) = (\bar{y} - \frac{Cov(x,y)}{Var(x)}\bar{x}) + \frac{Cov(x,y)}{Var(x)}x = \bar{y} + \frac{Cov(x,y)}{Var(x)}(x - \bar{x})$.

- To understand the nature of dependence between (x) and (y).
- Sometimes (y) may be costly/difficult to measure (total annual income) but (x) may be measured easily (total annual expenditure) - can use the formula to predict y^* using x^*.
- What type of formula? - $f(x) = ax^3 + b\sqrt{x} + c$?
- No, we want a formula of form $f(x) = a + bx$ - equation of a straight line.
- Reason?(i) Mathematically simple.
  (ii) Most of the time linear regression perform quite well!
- How to get the value of a, b? - a line that pass through the most middle - obtained by minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$.
- Closed form solution available - $f(x) = (\bar{y} - \frac{Cov(x,y)}{Var(x)}\bar{x}) + \frac{Cov(x,y)}{Var(x)}x = \bar{y} + \frac{Cov(x,y)}{Var(x)}(x - \bar{x})$.

- To understand the nature of dependence between (x) and (y).
- Sometimes (y) may be costly/difficult to measure (total annual income) but (x) may be measured easily (total annual expenditure) - can use the formula to predict y^* using x^*.
- What type of formula? - $f(x) = ax^3 + b\sqrt{x} + c$?
- No, we want a formula of form $f(x) = a + bx$ - equation of a straight line.
-

Reason?(i) Mathematically simple.

(ii) Most of the time linear regression perform quite well!

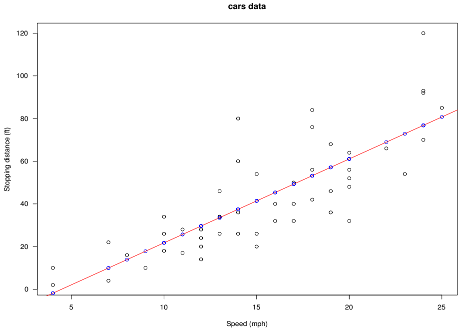
- How to get the value of a, b? - a line that pass through the most middle - obtained by minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$.
- Closed form solution available - $f(x) = (\bar{y} - \frac{Cov(x,y)}{Var(x)}\bar{x}) + \frac{Cov(x,y)}{Var(x)}x = \bar{y} + \frac{Cov(x,y)}{Var(x)}(x - \bar{x})$.

- To understand the nature of dependence between (x) and (y).
- Sometimes (y) may be costly/difficult to measure (total annual income) but (x) may be measured easily (total annual expenditure) - can use the formula to predict y^* using x^*.
- What type of formula? - $f(x) = ax^3 + b\sqrt{x} + c$?
- No, we want a formula of form $f(x) = a + bx$ - equation of a straight line.
-

  Reason?(i) Mathematically simple.

  (ii) Most of the time linear regression perform quite well!
- How to get the value of a, b? - a line that pass through the most middle - obtained by minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$.
- Closed form solution available - $f(x) = (\bar{y} - \frac{Cov(x,y)}{Var(x)}\bar{x}) + \frac{Cov(x,y)}{Var(x)}x = \bar{y} + \frac{Cov(x,y)}{Var(x)}(x - \bar{x})$.

- x_i, y_i - given data. $Y_i = f(x_i)$ is fitted values and $e_i = y_i - Y_i$ - residuals.

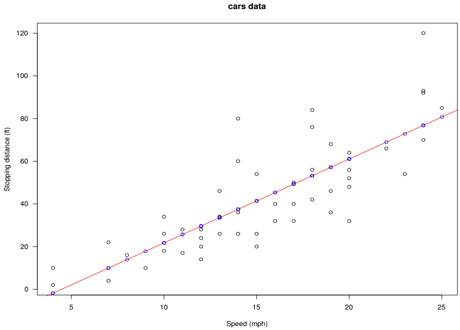


(a)

Figure: Scatter plot with the regression line, fitted values and residuals.

- Minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ wrt a, b - Principle of least squares (LS) - LS regression line.
- LS regression line is highly vulnerable to outlying observation.

- x_i, y_i - given data. $Y_i = f(x_i)$ is fitted values and $e_i = y_i - Y_i$ - residuals.



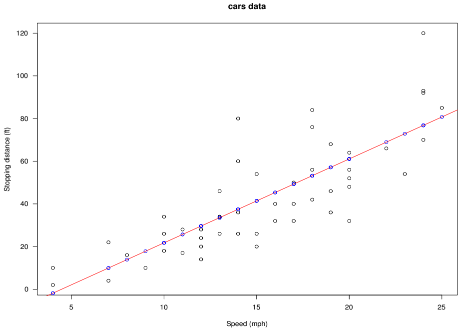
(a)

Figure: Scatter plot with the regression line, fitted values and residuals.

- Minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ wrt a, b - Principle of least squares (LS) - LS regression line.

- LS regression line is highly vulnerable to outlying observation.

- x_i, y_i - given data. $Y_i = f(x_i)$ is fitted values and $e_i = y_i - Y_i$ - residuals.

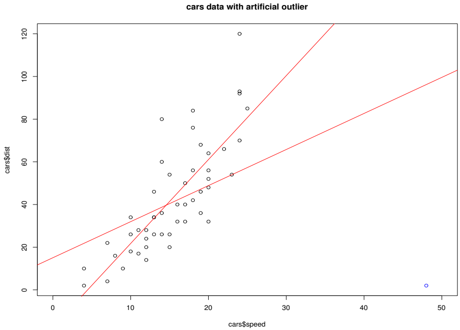


(a)

Figure: Scatter plot with the regression line, fitted values and residuals.

- Minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ wrt a, b - Principle of least squares (LS) - LS regression line.

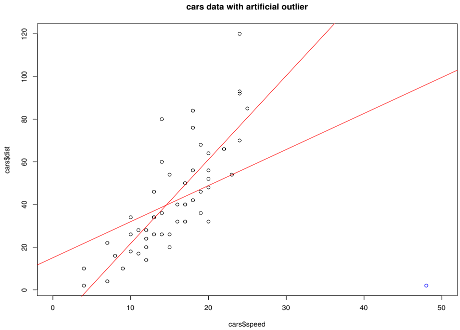
- LS regression line is highly vulnerable to outlying observation.



(a)

Figure: Effect of a single outlier on LS regression line.

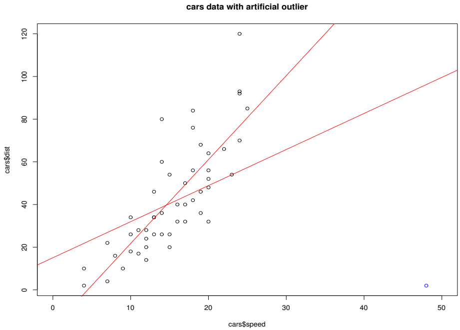
- Two possibilities : (i) detect and drop the outlier (ii) apply an outliers resistant regression.
- Minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ wrt a, b equivalent minimizing $\frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$ (mean of $(y_i - a - bx_i)^2$) wrt a, b.
- Why not minimize Median of $(y_i - a - bx_i)^2$ wrt a, b? - least median square (LMS) regression.



(a)

Figure: Effect of a single outlier on LS regression line.

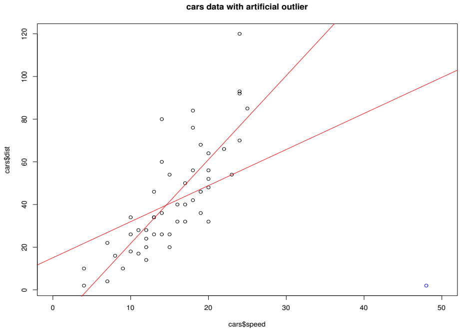
- Two possibilities : (i) detect and drop the outlier (ii) apply an outliers resistant regression.
- Minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ wrt a, b equivalent minimizing $\frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$ (mean of $(y_i - a - bx_i)^2$) wrt a, b.
- Why not minimize Median of $(y_i - a - bx_i)^2$ wrt a, b? - least median square (LMS) regression.



(a)

Figure: Effect of a single outlier on LS regression line.

- Two possibilities : (i) detect and drop the outlier (ii) apply an outliers resistant regression.
- Minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ wrt a, b equivalent minimizing $\frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$ (mean of $(y_i - a - bx_i)^2$) wrt a, b.
- Why not minimize Median of $(y_i - a - bx_i)^2$ wrt a, b? - least median square (LMS) regression.

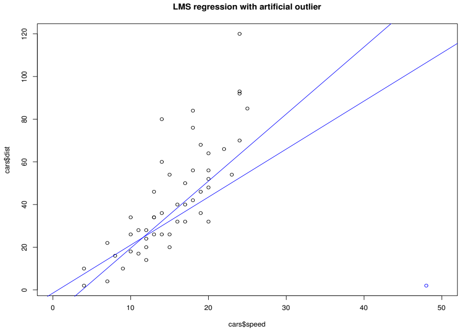


(a)

Figure: Effect of a single outlier on LS regression line.

- Two possibilities : (i) detect and drop the outlier (ii) apply an outliers resistant regression.
- Minimizing $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ wrt a, b equivalent minimizing $\frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$ (mean of $(y_i - a - bx_i)^2$) wrt a, b.
- Why not minimize Median of $(y_i - a - bx_i)^2$ wrt a, b? - least median square (LMS) regression.

- LMS regression line is less affected by outliers - outliers resistant.

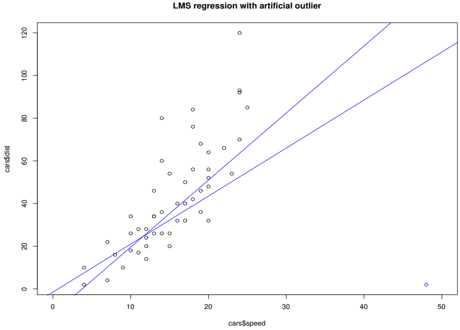


(a)

Figure: Effect of outlier on LMS regression line.

- So far only descriptive statistics.
- Want to understand reliability/accuracy of this regression lines - require specifying suitable statistical model for the data.

- LMS regression line is less affected by outliers - outliers resistant.



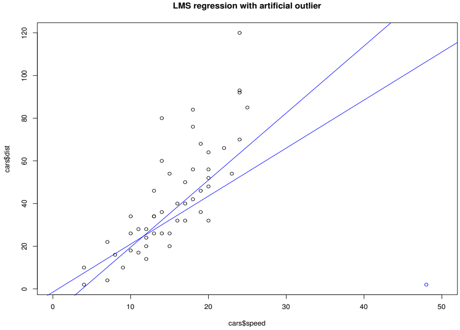
(a)

Figure: Effect of outlier on LMS regression line.

- So far only descriptive statistics.

- Want to understand reliability/accuracy of this regression lines - require specifying suitable statistical model for the data.

- LMS regression line is less affected by outliers - outliers resistant.



(a)

Figure: Effect of outlier on LMS regression line.

- So far only descriptive statistics.
- Want to understand reliability/accuracy of this regression lines - require specifying suitable statistical model for the data.

- **Simple linear regression model :**

$$[Y_1 = y_1, \cdots, Y_n = y_n | X_1 = x_1, \cdots, X_n = x_n] \sim \left(\frac{1}{\sqrt{2\pi}\sigma_\epsilon}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - a - bx_i)^2}{\sigma_\epsilon^2}}$$

- Model parameters - a, b, σ_ϵ.
- The model looks unfamiliar?
- The model is nothing but a family of MVN distributions indexed by unknown parameters a, b, σ_ϵ.
- More familiar specification - $\mathbf{Y} = \mathbf{X}\beta + \epsilon;\ \epsilon \sim MVN(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$.
- $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}$ and $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ are unobserved random errors.

$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$.

- **Simple linear regression model :**

$$[Y_1 = y_1, \cdots, Y_n = y_n | X_1 = x_1, \cdots, X_n = x_n] \sim \left(\frac{1}{\sqrt{2\pi}\sigma_\epsilon}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - a - bx_i)^2}{\sigma_\epsilon^2}}$$

- Model parameters - a, b, σ_ϵ.
- The model looks unfamiliar?
- The model is nothing but a family of MVN distributions indexed by unknown parameters a, b, σ_ϵ.
- More familiar specification - $\mathbf{Y} = \mathbf{X}\beta + \epsilon;\ \epsilon \sim MVN(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$.
- $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}$ and $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ are unobserved random errors.

  $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$.

- **Simple linear regression model :**

$$[Y_1 = y_1, \cdots, Y_n = y_n | X_1 = x_1, \cdots, X_n = x_n] \sim \left(\frac{1}{\sqrt{2\pi}\sigma_\epsilon}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - a - bx_i)^2}{\sigma_\epsilon^2}}$$

- Model parameters - a, b, σ_ϵ.
- The model looks unfamiliar?
- The model is nothing but a family of MVN distributions indexed by unknown parameters a, b, σ_ϵ.
- More familiar specification - $\mathbf{Y} = \mathbf{X}\beta + \epsilon;\ \epsilon \sim MVN(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$.
- $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}$ and $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ are unobserved random errors.

$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$.

- **Simple linear regression model :**

$$[Y_1 = y_1, \cdots, Y_n = y_n | X_1 = x_1, \cdots, X_n = x_n] \sim \left(\frac{1}{\sqrt{2\pi}\sigma_\epsilon}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - a - bx_i)^2}{\sigma_\epsilon^2}}$$

- Model parameters - a, b, σ_ϵ.
- The model looks unfamiliar?
- The model is nothing but a family of MVN distributions indexed by unknown parameters a, b, σ_ϵ.
- More familiar specification - $\mathbf{Y} = \mathbf{X}\beta + \epsilon;\ \epsilon \sim MVN(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$.
- $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}$ and $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ are unobserved random errors. $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$.

- **Simple linear regression model :**

$$[Y_1 = y_1, \cdots, Y_n = y_n | X_1 = x_1, \cdots, X_n = x_n] \sim \left(\frac{1}{\sqrt{2\pi}\sigma_\epsilon}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - a - bx_i)^2}{\sigma_\epsilon^2}}$$

- Model parameters - a, b, σ_ϵ.
- The model looks unfamiliar?
- The model is nothing but a family of MVN distributions indexed by unknown parameters a, b, σ_ϵ.
- More familiar specification - $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon};\ \boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$.
- $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}$ and $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ are unobserved random errors.

$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$.

- **Simple linear regression model :**

$$[Y_1 = y_1, \cdots, Y_n = y_n | X_1 = x_1, \cdots, X_n = x_n] \sim \left(\frac{1}{\sqrt{2\pi}\sigma_\epsilon}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - a - bx_i)^2}{\sigma_\epsilon^2}}$$

- Model parameters - a, b, σ_ϵ.
- The model looks unfamiliar?
- The model is nothing but a family of MVN distributions indexed by unknown parameters a, b, σ_ϵ.
- More familiar specification - $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon};\ \boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$.
- $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}$ and $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ are unobserved random errors. $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating β and σ^2_ϵ.
- mle of β is given by $\hat{\beta} = \mathbf{Q_X y} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ - same as LS regression values.
- mle of σ^2_ϵ is given by $\hat{\sigma^2_\epsilon} = \frac{\sum_{i=1}^n e_i^2}{n}$ - biased.
- An unbiased estimator $\tilde{\sigma^2_\epsilon} = \frac{\sum_{i=1}^n e_i^2}{n-2}$.
- Only concentrate on $\hat{\beta}$ from now on.
- How good/reliable are these estimates? - calculate standard errors.
- $Var(\hat{\beta}) = Var(\mathbf{Q_X y}) = \mathbf{Q_X} Var(\mathbf{y})\mathbf{Q'_X} = \mathbf{Q_X}\sigma^2_\epsilon \mathbf{I}_n \mathbf{Q'_X} = \sigma^2_\epsilon \mathbf{Q_X Q'_X} = \sigma^2_\epsilon(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1} = \sigma^2_\epsilon(\mathbf{X'X})^{-1}$.
- Estimate of $Var(\hat{\beta})$ is $\tilde{\sigma^2_\epsilon}(\mathbf{X'X})^{-1}$ (we use the unbiased estimator $\tilde{\sigma^2_\epsilon}$ not mle $\hat{\sigma^2_\epsilon}$.
- Its diagonal entries - estimate of standard error $\widehat{se(\hat{\beta}_0)}$ and $\widehat{se(\hat{\beta}_1)}$.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating $\boldsymbol{\beta}$ and σ_ϵ^2.
- mle of β is given by $\hat{\beta} = \mathbf{Q_X y} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ - same as LS regression values.
- mle of σ_ϵ^2 is given by $\hat{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n}$ - biased.
- An unbiased estimator $\tilde{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$.
- Only concentrate on $\hat{\beta}$ from now on.
- How good/reliable are these estimates? - calculate standard errors.
- $Var(\hat{\beta}) = Var(\mathbf{Q_X y}) = \mathbf{Q_X}Var(\mathbf{y})\mathbf{Q_X'} = \mathbf{Q_X}\sigma_\epsilon^2\mathbf{I}_n\mathbf{Q_X'} = \sigma_\epsilon^2\mathbf{Q_X Q_X'} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}$.
- Estimate of $Var(\hat{\beta})$ is $\tilde{\sigma_\epsilon^2}(\mathbf{X'X})^{-1}$ (we use the unbiased estimator $\tilde{\sigma_\epsilon^2}$ not mle $\hat{\sigma_\epsilon^2}$.
- Its diagonal entries - estimate of standard error $\widehat{se(\hat{\beta}_0)}$ and $\widehat{se(\hat{\beta}_1)}$.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating $\boldsymbol{\beta}$ and σ_ϵ^2.
- mle of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = \mathbf{Q_X y} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ - same as LS regression values.
- mle of σ_ϵ^2 is given by $\hat{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n}$ - biased.
- An unbiased estimator $\tilde{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$.
- Only concentrate on $\hat{\beta}$ from now on.
- How good/reliable are these estimates? - calculate standard errors.
- $Var(\hat{\beta}) = Var(\mathbf{Q_X y}) = \mathbf{Q_X} Var(\mathbf{y})\mathbf{Q_X'} = \mathbf{Q_X}\sigma_\epsilon^2 \mathbf{I}_n \mathbf{Q_X'} = \sigma_\epsilon^2 \mathbf{Q_X Q_X'} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}$.
- Estimate of $Var(\hat{\beta})$ is $\tilde{\sigma_\epsilon^2}(\mathbf{X'X})^{-1}$ (we use the unbiased estimator $\tilde{\sigma_\epsilon^2}$ not mle $\hat{\sigma_\epsilon^2}$.
- Its diagonal entries - estimate of standard error $\widehat{se(\hat{\beta}_0)}$ and $\widehat{se(\hat{\beta}_1)}$.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating β and σ_ϵ^2.
- mle of β is given by $\hat{\beta} = \mathbf{Q_X y} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ - same as LS regression values.
- mle of σ_ϵ^2 is given by $\hat{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n}$ - biased.
- An unbiased estimator $\tilde{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$.
- Only concentrate on $\hat{\beta}$ from now on.
- How good/reliable are these estimates? - calculate standard errors.
- $Var(\hat{\beta}) = Var(\mathbf{Q_X y}) = \mathbf{Q_X}\, Var(\mathbf{y})\mathbf{Q_X'} = \mathbf{Q_X}\sigma_\epsilon^2 \mathbf{I}_n \mathbf{Q_X'} = \sigma_\epsilon^2 \mathbf{Q_X Q_X'} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}$.
- Estimate of $Var(\hat{\beta})$ is $\tilde{\sigma_\epsilon^2}(\mathbf{X'X})^{-1}$ (we use the unbiased estimator $\tilde{\sigma_\epsilon^2}$ not mle $\hat{\sigma_\epsilon^2}$.
- Its diagonal entries - estimate of standard error $\widehat{se(\hat{\beta}_0)}$ and $\widehat{se(\hat{\beta}_1)}$.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating β and σ_ϵ^2.
- mle of β is given by $\hat{\beta} = \mathbf{Q_X y} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ - same as LS regression values.
- mle of σ_ϵ^2 is given by $\hat{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n}$ - biased.
- An unbiased estimator $\tilde{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$.
- Only concentrate on $\hat{\beta}$ from now on.
- How good/reliable are these estimates? - calculate standard errors.
- $Var(\hat{\beta}) = Var(\mathbf{Q_X y}) = \mathbf{Q_X}\,Var(\mathbf{y})\mathbf{Q_X'} = \mathbf{Q_X}\sigma_\epsilon^2\mathbf{I}_n\mathbf{Q_X'} = \sigma_\epsilon^2\mathbf{Q_X Q_X'} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}$.
- Estimate of $Var(\hat{\beta})$ is $\tilde{\sigma_\epsilon^2}(\mathbf{X'X})^{-1}$ (we use the unbiased estimator $\tilde{\sigma_\epsilon^2}$ not mle $\hat{\sigma_\epsilon^2}$.
- Its diagonal entries - estimate of standard error $\widehat{se(\hat{\beta}_0)}$ and $\widehat{se(\hat{\beta}_1)}$.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating $\boldsymbol{\beta}$ and σ_ϵ^2.
- mle of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = \mathbf{Q_X y} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ - same as LS regression values.
- mle of σ_ϵ^2 is given by $\hat{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n}$ - biased.
- An unbiased estimator $\tilde{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$.
- Only concentrate on $\hat{\boldsymbol{\beta}}$ from now on.
- How good/reliable are these estimates? - calculate standard errors.
- $Var(\hat{\boldsymbol{\beta}}) = Var(\mathbf{Q_X y}) = \mathbf{Q_X} Var(\mathbf{y})\mathbf{Q_X'} = \mathbf{Q_X}\sigma_\epsilon^2\mathbf{I}_n\mathbf{Q_X'} = \sigma_\epsilon^2\mathbf{Q_X Q_X'} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}$.
- Estimate of $Var(\hat{\boldsymbol{\beta}})$ is $\tilde{\sigma_\epsilon^2}(\mathbf{X'X})^{-1}$ (we use the unbiased estimator $\tilde{\sigma_\epsilon^2}$ not mle $\hat{\sigma_\epsilon^2}$.
- Its diagonal entries - estimate of standard error $\widehat{se(\hat{\beta}_0)}$ and $\widehat{se(\hat{\beta}_1)}$.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating β and σ_ϵ^2.
- mle of β is given by $\hat{\beta} = \mathbf{Q_X y} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ - same as LS regression values.
- mle of σ_ϵ^2 is given by $\hat{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n}$ - biased.
- An unbiased estimator $\tilde{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$.
- Only concentrate on $\hat{\beta}$ from now on.
- How good/reliable are these estimates? - calculate standard errors.
- $Var(\hat{\beta}) = Var(\mathbf{Q_X y}) = \mathbf{Q_X}\, Var(\mathbf{y})\mathbf{Q_X'} = \mathbf{Q_X}\sigma_\epsilon^2 \mathbf{I}_n \mathbf{Q_X'} = \sigma_\epsilon^2 \mathbf{Q_X Q_X'} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}$.
- Estimate of $Var(\hat{\beta})$ is $\tilde{\sigma_\epsilon^2}(\mathbf{X'X})^{-1}$ (we use the unbiased estimator $\tilde{\sigma_\epsilon^2}$ not mle $\hat{\sigma_\epsilon^2}$.
- Its diagonal entries - estimate of standard error $\widehat{se(\hat{\beta}_0)}$ and $\widehat{se(\hat{\beta}_1)}$.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating $\boldsymbol{\beta}$ and σ_ϵ^2.
- mle of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = \mathbf{Q_X y} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ - same as LS regression values.
- mle of σ_ϵ^2 is given by $\hat{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n}$ - biased.
- An unbiased estimator $\tilde{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$.
- Only concentrate on $\hat{\boldsymbol{\beta}}$ from now on.
- How good/reliable are these estimates? - calculate standard errors.
- $Var(\hat{\boldsymbol{\beta}}) = Var(\mathbf{Q_X y}) = \mathbf{Q_X}\, Var(\mathbf{y})\mathbf{Q_X'} = \mathbf{Q_X}\sigma_\epsilon^2 \mathbf{I}_n \mathbf{Q_X'} = \sigma_\epsilon^2 \mathbf{Q_X Q_X'} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}$.
- Estimate of $Var(\hat{\boldsymbol{\beta}})$ is $\tilde{\sigma_\epsilon^2}(\mathbf{X'X})^{-1}$ (we use the unbiased estimator $\tilde{\sigma_\epsilon^2}$ not mle $\hat{\sigma_\epsilon^2}$.
- Its diagonal entries - estimate of standard error $\widehat{se(\hat{\beta}_0)}$ and $\widehat{se(\hat{\beta}_1)}$.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating $\boldsymbol{\beta}$ and σ_ϵ^2.
- mle of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = \mathbf{Q_X y} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}$ - same as LS regression values.
- mle of σ_ϵ^2 is given by $\hat{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n}$ - biased.
- An unbiased estimator $\tilde{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$.
- Only concentrate on $\hat{\boldsymbol{\beta}}$ from now on.
- How good/reliable are these estimates? - calculate standard errors.
- $Var(\hat{\boldsymbol{\beta}}) = Var(\mathbf{Q_X y}) = \mathbf{Q_X}\, Var(\mathbf{y})\mathbf{Q_X'} = \mathbf{Q_X}\sigma_\epsilon^2 \mathbf{I}_n \mathbf{Q_X'} = \sigma_\epsilon^2 \mathbf{Q_X Q_X'} = \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} = \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$.
- Estimate of $Var(\hat{\boldsymbol{\beta}})$ is $\tilde{\sigma_\epsilon^2}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$ (we use the unbiased estimator $\tilde{\sigma_\epsilon^2}$ not mle $\hat{\sigma_\epsilon^2}$.
- Its diagonal entries - estimate of standard error $\widehat{se(\hat{\beta}_0)}$ and $\widehat{se(\hat{\beta}_1)}$.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating $\boldsymbol{\beta}$ and σ_ϵ^2.
- mle of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = \mathbf{Q_X y} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ - same as LS regression values.
- mle of σ_ϵ^2 is given by $\hat{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n}$ - biased.
- An unbiased estimator $\tilde{\sigma_\epsilon^2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$.
- Only concentrate on $\hat{\boldsymbol{\beta}}$ from now on.
- How good/reliable are these estimates? - calculate standard errors.
- $Var(\hat{\boldsymbol{\beta}}) = Var(\mathbf{Q_X y}) = \mathbf{Q_X}Var(\mathbf{y})\mathbf{Q_X'} = \mathbf{Q_X}\sigma_\epsilon^2\mathbf{I}_n\mathbf{Q_X'} = \sigma_\epsilon^2\mathbf{Q_X Q_X'} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1} = \sigma_\epsilon^2(\mathbf{X'X})^{-1}$.
- Estimate of $Var(\hat{\boldsymbol{\beta}})$ is $\tilde{\sigma_\epsilon^2}(\mathbf{X'X})^{-1}$ (we use the unbiased estimator $\tilde{\sigma_\epsilon^2}$ not mle $\hat{\sigma_\epsilon^2}$.
- Its diagonal entries - estimate of standard error $\widehat{se(\hat{\beta}_0)}$ and $\widehat{se(\hat{\beta}_1)}$.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $T = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Practitioners prefer p-value - $P(T > |T_{observed}|)$ where $T \sim t_{n-2}$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Joint test of significance $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$
- Test statistic $F = \frac{\hat{\boldsymbol{\beta}}'(X'X)\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Null distribution of test statistic $\sim F_{2,n-2}$ - Cutoff is obtained using $F_{2,n-2}$-distribution table.
- p-value - $P(F > F_{observed})$ where $F \sim F_{2,n-2}$.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $T = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Practitioners prefer p-value - $P(T > |T_{observed}|)$ where $T \sim t_{n-2}$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Joint test of significance $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$
- Test statistic $F = \frac{\hat{\boldsymbol{\beta}}'(X'X)\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Null distribution of test statistic $\sim F_{2,n-2}$ - Cutoff is obtained using $F_{2,n-2}$-distribution table.
- p-value - $P(F > F_{observed})$ where $F \sim F_{2,n-2}$.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $T = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Practitioners prefer p-value - $P(T > |T_{observed}|)$ where $T \sim t_{n-2}$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Joint test of significance $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$
- Test statistic $F = \frac{\hat{\boldsymbol{\beta}}'(X'X)\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Null distribution of test statistic $\sim F_{2,n-2}$ - Cutoff is obtained using $F_{2,n-2}$-distribution table.
- p-value - $P(F > F_{observed})$ where $F \sim F_{2,n-2}$.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $T = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Practitioners prefer p-value - $P(T > |T_{observed}|)$ where $T \sim t_{n-2}$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Joint test of significance $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$
- Test statistic $F = \frac{\hat{\boldsymbol{\beta}}'(X'X)\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Null distribution of test statistic $\sim F_{2,n-2}$ - Cutoff is obtained using $F_{2,n-2}$-distribution table.
- p-value - $P(F > F_{observed})$ where $F \sim F_{2,n-2}$.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $T = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Practitioners prefer p-value - $P(T > |T_{observed}|)$ where $T \sim t_{n-2}$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Joint test of significance $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$
- Test statistic $F = \frac{\hat{\boldsymbol{\beta}}'(X'X)\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Null distribution of test statistic $\sim F_{2,n-2}$ - Cutoff is obtained using $F_{2,n-2}$-distribution table.
- p-value - $P(F > F_{observed})$ where $F \sim F_{2,n-2}$.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $T = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Practitioners prefer p-value - $P(T > |T_{observed}|)$ where $T \sim t_{n-2}$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Joint test of significance $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$
- Test statistic $F = \frac{\hat{\boldsymbol{\beta}}'(X'X)\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Null distribution of test statistic $\sim F_{2,n-2}$ - Cutoff is obtained using $F_{2,n-2}$-distribution table.
- p-value - $P(F > F_{observed})$ where $F \sim F_{2,n-2}$.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $T = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Practitioners prefer p-value - $P(T > |T_{observed}|)$ where $T \sim t_{n-2}$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Joint test of significance $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$
- Test statistic $F = \frac{\hat{\boldsymbol{\beta}}'(X'X)\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Null distribution of test statistic $\sim F_{2,n-2}$ - Cutoff is obtained using $F_{2,n-2}$-distribution table.
- p-value - $P(F > F_{observed})$ where $F \sim F_{2,n-2}$.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $T = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Practitioners prefer p-value - $P(T > |T_{observed}|)$ where $T \sim t_{n-2}$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Joint test of significance $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$
- Test statistic $F = \frac{\hat{\boldsymbol{\beta}}'(X'X)\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Null distribution of test statistic $\sim F_{2,n-2}$ - Cutoff is obtained using $F_{2,n-2}$-distribution table.
- p-value - $P(F > F_{observed})$ where $F \sim F_{2,n-2}$.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $T = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Practitioners prefer p-value - $P(T > |T_{observed}|)$ where $T \sim t_{n-2}$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Joint test of significance $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$
- Test statistic $F = \frac{\hat{\boldsymbol{\beta}}'(X'X)\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Null distribution of test statistic $\sim F_{2,n-2}$ - Cutoff is obtained using $F_{2,n-2}$-distribution table.
- p-value - $P(F > F_{observed})$ where $F \sim F_{2,n-2}$.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $T = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Practitioners prefer p-value - $P(T > |T_{observed}|)$ where $T \sim t_{n-2}$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Joint test of significance $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$
- Test statistic $F = \frac{\hat{\boldsymbol{\beta}}'(X'X)\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Null distribution of test statistic $\sim F_{2,n-2}$ - Cutoff is obtained using $F_{2,n-2}$-distribution table.
- p-value - $P(F > F_{observed})$ where $F \sim F_{2,n-2}$.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $T = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Practitioners prefer p-value - $P(T > |T_{observed}|)$ where $T \sim t_{n-2}$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Joint test of significance $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$
- Test statistic $F = \frac{\hat{\boldsymbol{\beta}}'(X'X)\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Null distribution of test statistic $\sim F_{2,n-2}$ - Cutoff is obtained using $F_{2,n-2}$-distribution table.
- p-value - $P(F > F_{observed})$ where $F \sim F_{2,n-2}$.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $T = \frac{\hat{\beta}_0}{\widehat{se(\hat{\beta}_0)}}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Practitioners prefer p-value - $P(T > |T_{observed}|)$ where $T \sim t_{n-2}$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $T = \frac{\hat{\beta}_1}{\widehat{se(\hat{\beta}_1)}}$.
- Null distribution of test statistic $\sim t_{n-2}$ - Cutoff is obtained using t_{n-2}-distribution table.
- Joint test of significance $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$
- Test statistic $F = \frac{\hat{\boldsymbol{\beta}}'(X'X)\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Null distribution of test statistic $\sim F_{2,n-2}$ - Cutoff is obtained using $F_{2,n-2}$-distribution table.
- p-value - $P(F > F_{observed})$ where $F \sim F_{2,n-2}$.

- Confidence interval for β_0 can be obtained by inverting the test statistic $\frac{\hat{\beta}_0}{\widehat{se(\hat{\beta}_0)}}$.
- Confidence interval for β_1 can be obtained by inverting the test statistic $\frac{\hat{\beta}_1}{\widehat{se(\hat{\beta}_1)}}$.
- Confidence interval $\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}, \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}\right]$.
- $t_{n-2,\frac{\alpha}{2}}$ upper $\frac{\alpha}{2}$-cutoff point.
- Ellipsoidal joint confidence set for β is obtained by inverting the test statistic $\frac{\hat{\beta}'(\mathbf{X}'\mathbf{X})\hat{\beta}}{2\tilde{\sigma}_\epsilon^2}$.
- Confidence ellipsoid - $P(\beta : (\beta - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta - \hat{\beta}) \leq 2\tilde{\sigma}_\epsilon^2 F_{2,n-2,\alpha}) = 1 - \alpha$.
- $F_{2,n-2,\alpha}$ upper α-cutoff point.
- All of the above findings are useless if model fit is poor - need to check if the model is appropriate for the data.
- Model diagnostic checking.

- Confidence interval for β_0 can be obtained by inverting the test statistic $\frac{\hat{\beta}_0}{\widehat{se(\hat{\beta}_0)}}$.
- Confidence interval for β_1 can be obtained by inverting the test statistic $\frac{\hat{\beta}_1}{\widehat{se(\hat{\beta}_1)}}$.
- Confidence interval $\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}, \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}\right]$.
- $t_{n-2,\frac{\alpha}{2}}$ upper $\frac{\alpha}{2}$-cutoff point.
- Ellipsoidal joint confidence set for β is obtained by inverting the test statistic $\frac{\hat{\beta}'(\mathbf{X}'\mathbf{X})\hat{\beta}}{2\tilde{\sigma}_\epsilon^2}$.
- Confidence ellipsoid - $P(\beta : (\beta - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta - \hat{\beta}) \leq 2\tilde{\sigma}_\epsilon^2 F_{2,n-2,\alpha}) = 1 - \alpha$.
- $F_{2,n-2,\alpha}$ upper α-cutoff point.
- All of the above findings are useless if model fit is poor - need to check if the model is appropriate for the data.
- Model diagnostic checking.

- Confidence interval for β_0 can be obtained by inverting the test statistic $\frac{\hat{\beta}_0}{\widehat{se(\hat{\beta}_0)}}$.
- Confidence interval for β_1 can be obtained by inverting the test statistic $\frac{\hat{\beta}_1}{\widehat{se(\hat{\beta}_1)}}$.
- Confidence interval $\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}, \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}\right]$.
- $t_{n-2,\frac{\alpha}{2}}$ upper $\frac{\alpha}{2}$-cutoff point.
- Ellipsoidal joint confidence set for β is obtained by inverting the test statistic $\frac{\hat{\beta}'(\mathbf{X}'\mathbf{X})\hat{\beta}}{2\tilde{\sigma}_\epsilon^2}$.
- Confidence ellipsoid - $P(\beta : (\beta - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta - \hat{\beta}) \leq 2\tilde{\sigma}_\epsilon^2 F_{2,n-2,\alpha}) = 1 - \alpha$.
- $F_{2,n-2,\alpha}$ upper α-cutoff point.
- All of the above findings are useless if model fit is poor - need to check if the model is appropriate for the data.
- Model diagnostic checking.

- Confidence interval for β_0 can be obtained by inverting the test statistic $\frac{\hat{\beta}_0}{\widehat{se(\hat{\beta}_0)}}$.
- Confidence interval for β_1 can be obtained by inverting the test statistic $\frac{\hat{\beta}_1}{\widehat{se(\hat{\beta}_1)}}$.
- Confidence interval $\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}, \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}\right]$.
- $t_{n-2,\frac{\alpha}{2}}$ upper $\frac{\alpha}{2}$-cutoff point.
- Ellipsoidal joint confidence set for β is obtained by inverting the test statistic $\frac{\hat{\beta}'(\mathbf{X}'\mathbf{X})\hat{\beta}}{2\tilde{\sigma}_\epsilon^2}$.
- Confidence ellipsoid - $P(\beta : (\beta - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta - \hat{\beta}) \leq 2\tilde{\sigma}_\epsilon^2 F_{2,n-2,\alpha}) = 1 - \alpha$.
- $F_{2,n-2,\alpha}$ upper α-cutoff point.
- All of the above findings are useless if model fit is poor - need to check if the model is appropriate for the data.
- Model diagnostic checking.

- Confidence interval for β_0 can be obtained by inverting the test statistic $\frac{\hat{\beta}_0}{\widehat{se(\hat{\beta}_0)}}$.
- Confidence interval for β_1 can be obtained by inverting the test statistic $\frac{\hat{\beta}_1}{\widehat{se(\hat{\beta}_1)}}$.
- Confidence interval $\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}, \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}\right]$.
- $t_{n-2,\frac{\alpha}{2}}$ upper $\frac{\alpha}{2}$-cutoff point.
- Ellipsoidal joint confidence set for $\boldsymbol{\beta}$ is obtained by inverting the test statistic $\frac{\hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Confidence ellipsoid - $P(\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq 2\tilde{\sigma}_\epsilon^2 F_{2,n-2,\alpha}) = 1 - \alpha$.
- $F_{2,n-2,\alpha}$ upper α-cutoff point.
- All of the above findings are useless if model fit is poor - need to check if the model is appropriate for the data.
- Model diagnostic checking.

- Confidence interval for β_0 can be obtained by inverting the test statistic $\frac{\hat{\beta}_0}{\widehat{se(\hat{\beta}_0)}}$.
- Confidence interval for β_1 can be obtained by inverting the test statistic $\frac{\hat{\beta}_1}{\widehat{se(\hat{\beta}_1)}}$.
- Confidence interval $\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}, \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}\right]$.
- $t_{n-2,\frac{\alpha}{2}}$ upper $\frac{\alpha}{2}$-cutoff point.
- Ellipsoidal joint confidence set for $\boldsymbol{\beta}$ is obtained by inverting the test statistic $\frac{\hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Confidence ellipsoid - $P(\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq 2\tilde{\sigma}_\epsilon^2 F_{2,n-2,\alpha}) = 1 - \alpha$.
- $F_{2,n-2,\alpha}$ upper α-cutoff point.
- All of the above findings are useless if model fit is poor - need to check if the model is appropriate for the data.
- Model diagnostic checking.

- Confidence interval for β_0 can be obtained by inverting the test statistic $\frac{\hat{\beta}_0}{\widehat{se(\hat{\beta}_0)}}$.
- Confidence interval for β_1 can be obtained by inverting the test statistic $\frac{\hat{\beta}_1}{\widehat{se(\hat{\beta}_1)}}$.
- Confidence interval $\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}, \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}\right]$.
- $t_{n-2,\frac{\alpha}{2}}$ upper $\frac{\alpha}{2}$-cutoff point.
- Ellipsoidal joint confidence set for $\boldsymbol{\beta}$ is obtained by inverting the test statistic $\frac{\hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Confidence ellipsoid - $P(\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq 2\tilde{\sigma}_\epsilon^2 F_{2,n-2,\alpha}) = 1 - \alpha$.
- $F_{2,n-2,\alpha}$ upper α-cutoff point.
- All of the above findings are useless if model fit is poor - need to check if the model is appropriate for the data.
- Model diagnostic checking.

- Confidence interval for β_0 can be obtained by inverting the test statistic $\frac{\hat{\beta}_0}{\widehat{se(\hat{\beta}_0)}}$.
- Confidence interval for β_1 can be obtained by inverting the test statistic $\frac{\hat{\beta}_1}{\widehat{se(\hat{\beta}_1)}}$.
- Confidence interval $\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}} \widehat{se(\hat{\beta}_0)}, \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}} \widehat{se(\hat{\beta}_0)}\right]$.
- $t_{n-2,\frac{\alpha}{2}}$ upper $\frac{\alpha}{2}$-cutoff point.
- Ellipsoidal joint confidence set for $\boldsymbol{\beta}$ is obtained by inverting the test statistic $\frac{\hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Confidence ellipsoid - $P(\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq 2\tilde{\sigma}_\epsilon^2 F_{2,n-2,\alpha}) = 1 - \alpha$.
- $F_{2,n-2,\alpha}$ upper α-cutoff point.
- All of the above findings are useless if model fit is poor - need to check if the model is appropriate for the data.
- Model diagnostic checking.

- Confidence interval for β_0 can be obtained by inverting the test statistic $\frac{\hat{\beta}_0}{\widehat{se(\hat{\beta}_0)}}$.
- Confidence interval for β_1 can be obtained by inverting the test statistic $\frac{\hat{\beta}_1}{\widehat{se(\hat{\beta}_1)}}$.
- Confidence interval $\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}, \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}\right]$.
- $t_{n-2,\frac{\alpha}{2}}$ upper $\frac{\alpha}{2}$-cutoff point.
- Ellipsoidal joint confidence set for $\boldsymbol{\beta}$ is obtained by inverting the test statistic $\frac{\hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}}}{2\tilde{\sigma}_\epsilon^2}$.
- Confidence ellipsoid - $P(\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq 2\tilde{\sigma}_\epsilon^2 F_{2,n-2,\alpha}) = 1 - \alpha$.
- $F_{2,n-2,\alpha}$ upper α-cutoff point.
- All of the above findings are useless if model fit is poor - need to check if the model is appropriate for the data.
- Model diagnostic checking.

- **Appropriateness of Gauss-Markov assumption :**

  (i) Linearity: The relationship between **X** and the mean of **Y** is linear ($E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta$).

  (ii) Homoscedasticity: The variance of residual is the same for $x_1, x_2, \cdots, x_n$.

  (iii) Uncorrelatedness: Observations are uncorrelated of each other.

- Normality: For any fixed value x_i, $[Y_i|X_i = x_i]$ is normally distributed.

- Normality + (iii) Uncorrelatedness: Observations are uncorrelated of each other $\Rightarrow$ Observations are independent of each other.

- **Check for potentially bad points which may lead to poor model fit :**

  (i) Outliers: An outlier is defined as an observation that has a large residual. In other words, the observed value for the point is very different from that predicted by the regression model.

  (ii) Leverage points: A leverage point is defined as an observation that has a value of x_i that is far away from the mean of $x_1, x_2, \cdots, x_n$.

  (iii) Influential observations: An influential observation is defined as an observation that changes the slope of the line. Thus, influential points have a large influence on the fit of the model.

- **Appropriateness of Gauss-Markov assumption :**

(i) Linearity: The relationship between **X** and the mean of **Y** is linear ($E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta$).

(ii) Homoscedasticity: The variance of residual is the same for $x_1, x_2, \cdots, x_n$.

(iii) Uncorrelatedness: Observations are uncorrelated of each other.

- Normality: For any fixed value x_i, $[Y_i|X_i = x_i]$ is normally distributed.

- Normality + (iii) Uncorrelatedness: Observations are uncorrelated of each other $\Rightarrow$ Observations are independent of each other.

- **Check for potentially bad points which may lead to poor model fit :**

(i) Outliers: An outlier is defined as an observation that has a large residual. In other words, the observed value for the point is very different from that predicted by the regression model.

(ii) Leverage points: A leverage point is defined as an observation that has a value of x_i that is far away from the mean of $x_1, x_2, \cdots, x_n$.

(iii) Influential observations: An influential observation is defined as an observation that changes the slope of the line. Thus, influential points have a large influence on the fit of the model.

- **Appropriateness of Gauss-Markov assumption :**

  (i) Linearity: The relationship between $\mathbf{X}$ and the mean of $\mathbf{Y}$ is linear ($E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta$).

  (ii) Homoscedasticity: The variance of residual is the same for $x_1, x_2, \cdots, x_n$.

  (iii) Uncorrelatedness: Observations are uncorrelated of each other.

- Normality: For any fixed value x_i, $[Y_i|X_i = x_i]$ is normally distributed.

- Normality + (iii) Uncorrelatedness: Observations are uncorrelated of each other $\Rightarrow$ Observations are independent of each other.

- **Check for potentially bad points which may lead to poor model fit :**

  (i) Outliers: An outlier is defined as an observation that has a large residual. In other words, the observed value for the point is very different from that predicted by the regression model.

  (ii) Leverage points: A leverage point is defined as an observation that has a value of x_i that is far away from the mean of $x_1, x_2, \cdots, x_n$.

  (iii) Influential observations: An influential observation is defined as an observation that changes the slope of the line. Thus, influential points have a large influence on the fit of the model.

- **Appropriateness of Gauss-Markov assumption :**

  (i) Linearity: The relationship between **X** and the mean of **Y** is linear $(E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta)$.

  (ii) Homoscedasticity: The variance of residual is the same for $x_1, x_2, \cdots, x_n$.

  (iii) Uncorrelatedness: Observations are uncorrelated of each other.

- Normality: For any fixed value x_i, $[Y_i|X_i = x_i]$ is normally distributed.

- Normality + (iii) Uncorrelatedness: Observations are uncorrelated of each other $\Rightarrow$ Observations are independent of each other.

- **Check for potentially bad points which may lead to poor model fit :**

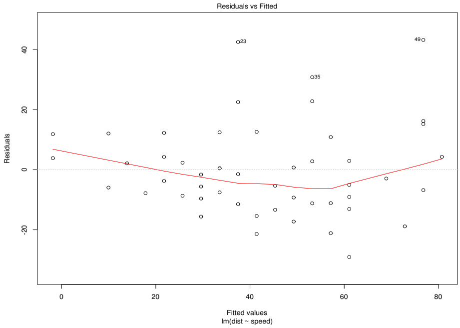
  (i) Outliers: An outlier is defined as an observation that has a large residual. In other words, the observed value for the point is very different from that predicted by the regression model.

  (ii) Leverage points: A leverage point is defined as an observation that has a value of x_i that is far away from the mean of $x_1, x_2, \cdots, x_n$.

  (iii) Influential observations: An influential observation is defined as an observation that changes the slope of the line. Thus, influential points have a large influence on the fit of the model.

- Linearity - Check the fitted value Y_i vs residual e_i plot for any pattern - randomly and closely distributed around $x - axis$ indicates linearity.

- Homoscedasticity - Check the fitted value Y_i vs residual e_i plot to see if the spread is changing as we move along $x - axis$ - changing means heteroscedastic.

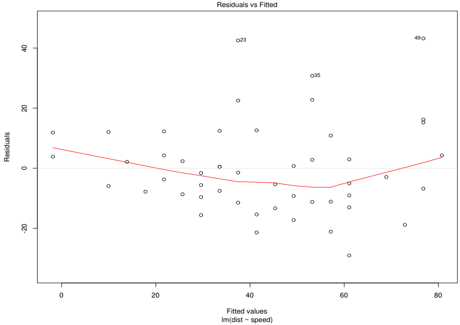


(a)

Figure: Clear indication of nonlinearity and heteroscedasticity.

- Linearity - Check the fitted value Y_i vs residual e_i plot for any pattern - randomly and closely distributed around $x - axis$ indicates linearity.

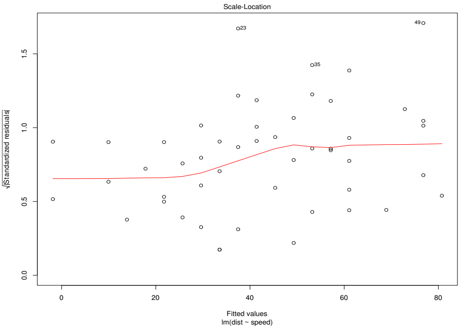
- Homoscedasticity - Check the fitted value Y_i vs residual e_i plot to see if the spread is changing as we move along $x - axis$ - changing means heteroscedastic.



(a)

Figure: Clear indication of nonlinearity and heteroscedasticity.

- Homoscedasticity - Check the fitted value Y_i vs square root of absolute standardised residual $\sqrt{|\frac{e_i}{\tilde{\sigma_\epsilon}\sqrt{1-h_{ii}}}|}$ plot to see if the spread is changing as we move along $x-axis$ - changing means heteroscedastic.

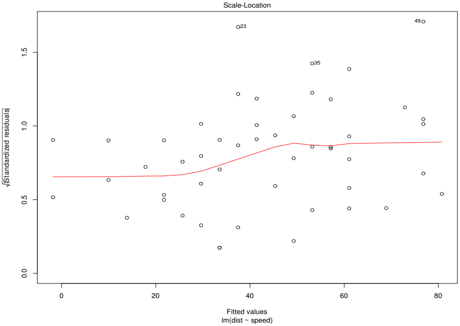


(a)

Figure: Clear indication of heteroscedasticity.

- This plot is more appropriate for homoscedasticity checking as $Var(e_i)$ are different not same as $Var(\epsilon_i)$.

- Homoscedasticity - Check the fitted value Y_i vs square root of absolute standardised residual $\sqrt{|\frac{e_i}{\tilde{\sigma_\epsilon}\sqrt{1-h_{ii}}}|}$ plot to see if the spread is changing as we move along $x - axis$ - changing means heteroscedastic.



(a)

Figure: Clear indication of heteroscedasticity.

- This plot is more appropriate for homoscedasticity checking as $Var(e_i)$ are different not same as $Var(\epsilon_i)$.

- $Var(e_i) = \sigma_\epsilon^2(1 - h_{ii})$ - so, $\widehat{Var(e_i)} = \tilde{\sigma}_\epsilon^2(1 - h_{ii})$.
- So, standardised residual $\frac{e_i}{\sqrt{\widehat{Var(e_i)}}} = \frac{e_i}{\tilde{\sigma}_\epsilon\sqrt{1-h_{ii}}}$.
- h_{ii} is the *i*th leverage value - the *i*th diagonal entry of the matrix $\mathbf{XQ_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P_X}$.
- $\mathbf{P_X}$ (some refers it as hat-matrix $\mathbf{H}$) is an orthogonal projection matrix - idempotent and symmetric - also, $\hat{\mathbf{y}} = \mathbf{P_X y}$.
- One can use Breusch-Pagan Test for checking homoscedasticity - asymptotically χ^2 distributed.
- Uncorrelatedness: Plot the sample autocorrelation function of the residuals.

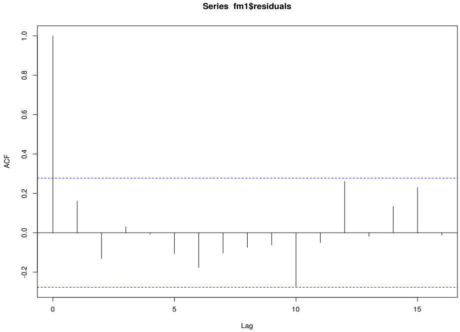
- $Var(e_i) = \sigma_\epsilon^2(1 - h_{ii})$ - so, $\widehat{Var(e_i)} = \tilde{\sigma}_\epsilon^2(1 - h_{ii})$.
- So, standardised residual $\frac{e_i}{\sqrt{\widehat{Var(e_i)}}} = \frac{e_i}{\tilde{\sigma}_\epsilon \sqrt{1-h_{ii}}}$.
- h_{ii} is the ith leverage value - the ith diagonal entry of the matrix $\mathbf{XQ_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P_X}$.
- $\mathbf{P_X}$ (some refers it as hat-matrix $\mathbf{H}$) is an orthogonal projection matrix - idempotent and symmetric - also, $\hat{\mathbf{y}} = \mathbf{P_X}\mathbf{y}$.
- One can use Breusch-Pagan Test for checking homoscedasticity - asymptotically χ^2 distributed.
- Uncorrelatedness: Plot the sample autocorrelation function of the residuals.

- $Var(e_i) = \sigma_\epsilon^2(1 - h_{ii})$ - so, $\widehat{Var(e_i)} = \tilde{\sigma_\epsilon^2}(1 - h_{ii})$.
- So, standardised residual $\frac{e_i}{\sqrt{\widehat{Var(e_i)}}} = \frac{e_i}{\tilde{\sigma_\epsilon}\sqrt{1-h_{ii}}}$.
- h_{ii} is the ith leverage value - the ith diagonal entry of the matrix $\mathbf{XQ_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P_X}$.
- $\mathbf{P_X}$ (some refers it as hat-matrix $\mathbf{H}$) is an orthogonal projection matrix - idempotent and symmetric - also, $\hat{\mathbf{y}} = \mathbf{P_X}\mathbf{y}$.
- One can use Breusch-Pagan Test for checking homoscedasticity - asymptotically χ^2 distributed.
- Uncorrelatedness: Plot the sample autocorrelation function of the residuals.

- $Var(e_i) = \sigma_\epsilon^2(1 - h_{ii})$ - so, $\widehat{Var(e_i)} = \tilde{\sigma}_\epsilon^2(1 - h_{ii})$.
- So, standardised residual $\frac{e_i}{\sqrt{\widehat{Var(e_i)}}} = \frac{e_i}{\tilde{\sigma}_\epsilon\sqrt{1-h_{ii}}}$.
- h_{ii} is the ith leverage value - the ith diagonal entry of the matrix $\mathbf{XQ_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P_X}$.
- $\mathbf{P_X}$ (some refers it as hat-matrix $\mathbf{H}$) is an orthogonal projection matrix - idempotent and symmetric - also, $\hat{\mathbf{y}} = \mathbf{P_X}\mathbf{y}$.
- One can use Breusch-Pagan Test for checking homoscedasticity - asymptotically χ^2 distributed.
- Uncorrelatedness: Plot the sample autocorrelation function of the residuals.

- $Var(e_i) = \sigma_\epsilon^2(1 - h_{ii})$ - so, $\widehat{Var(e_i)} = \tilde{\sigma}_\epsilon^2(1 - h_{ii})$.
- So, standardised residual $\frac{e_i}{\sqrt{\widehat{Var(e_i)}}} = \frac{e_i}{\tilde{\sigma_\epsilon}\sqrt{1-h_{ii}}}$.
- h_{ii} is the ith leverage value - the ith diagonal entry of the matrix $\mathbf{XQ_X} = \mathbf{X(X'X)^{-1}X'} = \mathbf{P_X}$.
- $\mathbf{P_X}$ (some refers it as hat-matrix $\mathbf{H}$) is an orthogonal projection matrix - idempotent and symmetric - also, $\hat{\mathbf{y}} = \mathbf{P_X y}$.
- One can use Breusch-Pagan Test for checking homoscedasticity - asymptotically χ^2 distributed.
- Uncorrelatedness: Plot the sample autocorrelation function of the residuals.

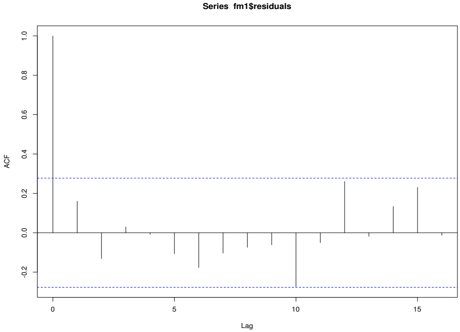
- $Var(e_i) = \sigma_\epsilon^2(1 - h_{ii})$ - so, $\widehat{Var(e_i)} = \tilde{\sigma_\epsilon^2}(1 - h_{ii})$.
- So, standardised residual $\frac{e_i}{\sqrt{\widehat{Var(e_i)}}} = \frac{e_i}{\tilde{\sigma_\epsilon}\sqrt{1-h_{ii}}}$.
- h_{ii} is the ith leverage value - the ith diagonal entry of the matrix $\mathbf{XQ_X} = \mathbf{X(X'X)^{-1}X'} = \mathbf{P_X}$.
- $\mathbf{P_X}$ (some refers it as hat-matrix $\mathbf{H}$) is an orthogonal projection matrix - idempotent and symmetric - also, $\hat{\mathbf{y}} = \mathbf{P_X y}$.
- One can use Breusch-Pagan Test for checking homoscedasticity - asymptotically χ^2 distributed.
- Uncorrelatedness: Plot the sample autocorrelation function of the residuals.



(a)

Figure: Indication of uncorrelatedness.

- Also can perform Durbin-Watson test and Box-Pierce test for checking whether there is any autocorrelation.

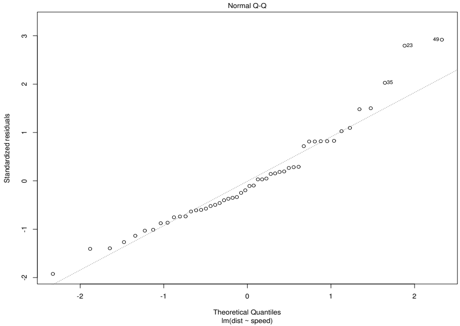


(a)

Figure: Indication of uncorrelatedness.

- Also can perform Durbin-Watson test and Box-Pierce test for checking whether there is any autocorrelation.

- Normality: Q-Q plot of standardised/studentized residuals.

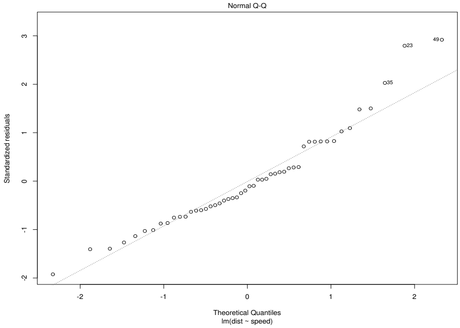


(a)

Figure: Indication of non-normality.

- Also can perform Shapiro-Wilks test and Kolmogorov-Smirnov test for checking departure from normality.

- Normality: Q-Q plot of standardised/studentized residuals.

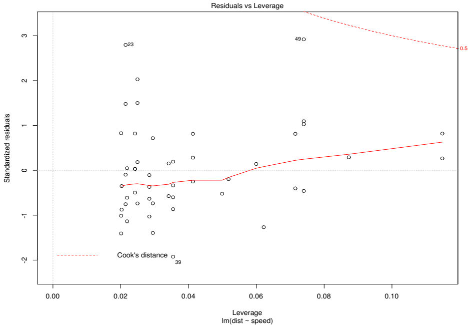


(a)

Figure: Indication of non-normality.

- Also can perform Shapiro-Wilks test and Kolmogorov-Smirnov test for checking departure from normality.

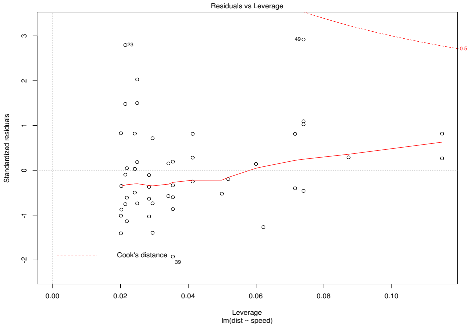
- Outliers : Check the fitted value Y_i vs residual e_i plot for large values - potential outliers.
- Leverage points : Check for points with high leverage values h_{ii}.
- Recall that $0 \leq h_{ii} \leq 1$.
- Influential observations: Can be detected by looking into standardised residuals vs leverage plot.



(a)

Figure: A few influential observations.

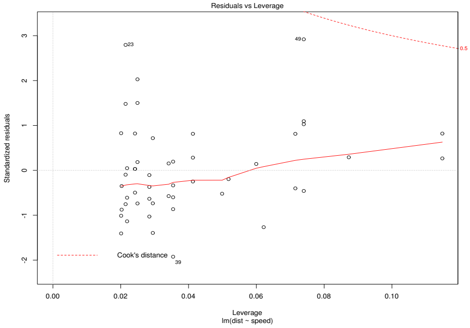
- Outliers : Check the fitted value Y_i vs residual e_i plot for large values - potential outliers.
- Leverage points : Check for points with high leverage values h_{ii}.
- Recall that $0 \leq h_{ii} \leq 1$.
- Influential observations: Can be detected by looking into standardised residuals vs leverage plot.



(a)

Figure: A few influential observations.

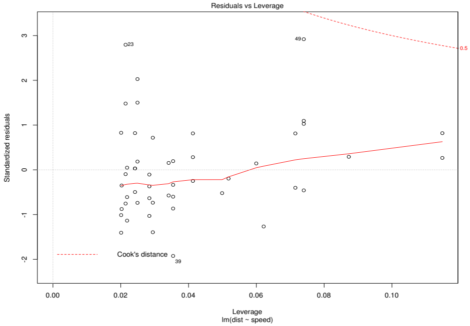
- Outliers : Check the fitted value Y_i vs residual e_i plot for large values - potential outliers.
- Leverage points : Check for points with high leverage values h_{ii}.
- Recall that $0 \leq h_{ii} \leq 1$.
- Influential observations: Can be detected by looking into standardised residuals vs leverage plot.



(a)

Figure: A few influential observations.

- Outliers : Check the fitted value Y_i vs residual e_i plot for large values - potential outliers.
- Leverage points : Check for points with high leverage values h_{ii}.
- Recall that $0 \leq h_{ii} \leq 1$.
- Influential observations: Can be detected by looking into standardised residuals vs leverage plot.



(a)

Figure: A few influential observations.

- Also, some more numerical diagnostic measures are there for detection of potentially influential observations.

- Cook's distance :
$D_i = \frac{1}{2}\left(\frac{e_i}{\tilde{\sigma_\epsilon}\sqrt{1-h_{ii}}}\right)^2 \frac{h_{ii}}{1-h_{ii}} = \frac{1}{2}\,(\text{standardized residual})^2\,\frac{h_{ii}}{1-h_{ii}}$.

- So, Cook's D is a function of studentized residual and leverage value - can be plotted as a nonlinear contours in the residuals vs leverage plot.

- High leverage values (close to 1) means Cook's distance very large - highly influential observation.

- DFFIT : $DFFIT_i =$ difference in fit as we drop the ith observation.

- Relationship between D_i and $DFFIT_i$: $D_i = \frac{1}{2}\frac{\hat{\sigma}^2_{\epsilon(i)}}{\hat{\sigma}^2_{\epsilon}} DFFIT_i^2$.

- If the model diagnostic checking turns out satisfactory then we check for how good the model fits the data.

- Also, some more numerical diagnostic measures are there for detection of potentially influential observations.

- Cook's distance :
$$D_i = \frac{1}{2}\left(\frac{e_i}{\tilde{\sigma}_\epsilon\sqrt{1-h_{ii}}}\right)^2 \frac{h_{ii}}{1-h_{ii}} = \frac{1}{2}\,(\text{standardized residual})^2\,\frac{h_{ii}}{1-h_{ii}}.$$

- So, Cook's D is a function of studentized residual and leverage value - can be plotted as a nonlinear contours in the residuals vs leverage plot.

- High leverage values (close to 1) means Cook's distance very large - highly influential observation.

- DFFIT : $DFFIT_i =$ difference in fit as we drop the ith observation.

- Relationship between D_i and $DFFIT_i$: $D_i = \frac{1}{2}\frac{\hat{\sigma}^2_{\epsilon(i)}}{\hat{\sigma}^2_\epsilon} DFFIT_i^2$.

- If the model diagnostic checking turns out satisfactory then we check for how good the model fits the data.

- Also, some more numerical diagnostic measures are there for detection of potentially influential observations.

- Cook's distance :
$D_i = \frac{1}{2}\left(\frac{e_i}{\tilde{\sigma_\epsilon}\sqrt{1-h_{ii}}}\right)^2 \frac{h_{ii}}{1-h_{ii}} = \frac{1}{2}\,(\text{standardized residual})^2\,\frac{h_{ii}}{1-h_{ii}}.$

- So, Cook's D is a function of studentized residual and leverage value - can be plotted as a nonlinear contours in the residuals vs leverage plot.

- High leverage values (close to 1) means Cook's distance very large - highly influential observation.

- DFFIT : $DFFIT_i =$ difference in fit as we drop the ith observation.

- Relationship between D_i and $DFFIT_i$: $D_i = \frac{1}{2}\frac{\hat{\sigma}^2_{\epsilon(i)}}{\hat{\sigma}^2_\epsilon} DFFIT_i^2$.

- If the model diagnostic checking turns out satisfactory then we check for how good the model fits the data.

- Also, some more numerical diagnostic measures are there for detection of potentially influential observations.

- Cook's distance :
$$D_i = \frac{1}{2}\left(\frac{e_i}{\tilde{\sigma}_\epsilon\sqrt{1-h_{ii}}}\right)^2 \frac{h_{ii}}{1-h_{ii}} = \frac{1}{2}\,(\text{standardized residual})^2\,\frac{h_{ii}}{1-h_{ii}}.$$

- So, Cook's D is a function of studentized residual and leverage value - can be plotted as a nonlinear contours in the residuals vs leverage plot.

- High leverage values (close to 1) means Cook's distance very large - highly influential observation.

- DFFIT : $DFFIT_i =$ difference in fit as we drop the ith observation.

- Relationship between D_i and $DFFIT_i$: $D_i = \frac{1}{2}\frac{\hat{\sigma}^2_{\epsilon(i)}}{\hat{\sigma}^2_\epsilon} DFFIT_i^2$.

- If the model diagnostic checking turns out satisfactory then we check for how good the model fits the data.

- Also, some more numerical diagnostic measures are there for detection of potentially influential observations.
- Cook's distance :
$$D_i = \frac{1}{2}\left(\frac{e_i}{\tilde{\sigma}_\epsilon\sqrt{1-h_{ii}}}\right)^2 \frac{h_{ii}}{1-h_{ii}} = \frac{1}{2}\,(\text{standardized residual})^2\,\frac{h_{ii}}{1-h_{ii}}.$$
- So, Cook's D is a function of studentized residual and leverage value - can be plotted as a nonlinear contours in the residuals vs leverage plot.
- High leverage values (close to 1) means Cook's distance very large - highly influential observation.
- DFFIT : $DFFIT_i =$ difference in fit as we drop the ith observation.
- Relationship between D_i and $DFFIT_i$: $D_i = \frac{1}{2}\frac{\hat{\sigma}^2_{\epsilon(i)}}{\hat{\sigma}^2_\epsilon} DFFIT_i^2$.
- If the model diagnostic checking turns out satisfactory then we check for how good the model fits the data.

- Also, some more numerical diagnostic measures are there for detection of potentially influential observations.

- Cook's distance :
$D_i = \frac{1}{2}\left(\frac{e_i}{\tilde{\sigma_\epsilon}\sqrt{1-h_{ii}}}\right)^2 \frac{h_{ii}}{1-h_{ii}} = \frac{1}{2}\,(\text{standardized residual})^2\,\frac{h_{ii}}{1-h_{ii}}$.

- So, Cook's D is a function of studentized residual and leverage value - can be plotted as a nonlinear contours in the residuals vs leverage plot.

- High leverage values (close to 1) means Cook's distance very large - highly influential observation.

- DFFIT : $DFFIT_i =$ difference in fit as we drop the ith observation.

- Relationship between D_i and $DFFIT_i$: $D_i = \frac{1}{2}\frac{\hat{\sigma}^2_{\epsilon(i)}}{\hat{\sigma}^2_\epsilon} DFFIT_i^2$.

- If the model diagnostic checking turns out satisfactory then we check for how good the model fits the data.

- Also, some more numerical diagnostic measures are there for detection of potentially influential observations.

- Cook's distance :
  $D_i = \frac{1}{2}\left(\frac{e_i}{\tilde{\sigma_\epsilon}\sqrt{1-h_{ii}}}\right)^2 \frac{h_{ii}}{1-h_{ii}} = \frac{1}{2}\,(\text{standardized residual})^2\,\frac{h_{ii}}{1-h_{ii}}.$

- So, Cook's D is a function of studentized residual and leverage value - can be plotted as a nonlinear contours in the residuals vs leverage plot.

- High leverage values (close to 1) means Cook's distance very large - highly influential observation.

- DFFIT : $DFFIT_i =$ difference in fit as we drop the ith observation.

- Relationship between D_i and $DFFIT_i$: $D_i = \frac{1}{2}\frac{\hat{\sigma}^2_{\epsilon(i)}}{\hat{\sigma}^2_\epsilon} DFFIT_i^2$.

- If the model diagnostic checking turns out satisfactory then we check for how good the model fits the data.

- There are several such goodness of fit measure.
- These measures are useful in selection of a single best model among several competing models.
- R-squared - $R^2 = \frac{Var(Y)}{Var(y)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2};\ 0 \leq R^2 \leq 1.$
- Problem of R^2 - tend to select overfitting models.
- Adjusted R-squared - $R^2_{adj} = 1 - \frac{(n-1)(1-R^2)}{(n-2)}$ - higher the better - can be negative!
- AIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + 2(2+1)$ - lower the better.
- BIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + \ln(n)(2+1)$ - lower the better.
- BIC penalizes complex models more severely - better to use BIC than AIC.

- There are several such goodness of fit measure.
- These measures are useful in selection of a single best model among several competing models.
- R-squared - $R^2 = \frac{Var(Y)}{Var(y)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$; $0 \leq R^2 \leq 1$.
- Problem of R^2 - tend to select overfitting models.
- Adjusted R-squared - $R^2_{adj} = 1 - \frac{(n-1)(1-R^2)}{(n-2)}$ - higher the better - can be negative!
- AIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + 2(2+1)$ - lower the better.
- BIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + \ln(n)(2+1)$ - lower the better.
- BIC penalizes complex models more severely - better to use BIC than AIC.

- There are several such goodness of fit measure.
- These measures are useful in selection of a single best model among several competing models.
- R-squared - $R^2 = \frac{Var(Y)}{Var(y)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$; $0 \leq R^2 \leq 1$.
- Problem of R^2 - tend to select overfitting models.
- Adjusted R-squared - $R^2_{adj} = 1 - \frac{(n-1)(1-R^2)}{(n-2)}$ - higher the better - can be negative!
- AIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + 2(2+1)$ - lower the better.
- BIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + \ln(n)(2+1)$ - lower the better.
- BIC penalizes complex models more severely - better to use BIC than AIC.

- There are several such goodness of fit measure.
- These measures are useful in selection of a single best model among several competing models.
- R-squared - $R^2 = \frac{Var(Y)}{Var(y)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2};\ 0 \leq R^2 \leq 1.$
- Problem of R^2 - tend to select overfitting models.
- Adjusted R-squared - $R^2_{adj} = 1 - \frac{(n-1)(1-R^2)}{(n-2)}$ - higher the better - can be negative!
- AIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + 2(2+1)$ - lower the better.
- BIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + \ln(n)(2+1)$ - lower the better.
- BIC penalizes complex models more severely - better to use BIC than AIC.

- There are several such goodness of fit measure.
- These measures are useful in selection of a single best model among several competing models.
- R-squared - $R^2 = \frac{Var(Y)}{Var(y)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2};\ 0 \leq R^2 \leq 1.$
- Problem of R^2 - tend to select overfitting models.
- Adjusted R-squared - $R^2_{adj} = 1 - \frac{(n-1)(1-R^2)}{(n-2)}$ - higher the better - can be negative!
- AIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + 2(2+1)$ - lower the better.
- BIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + \ln(n)(2+1)$ - lower the better.
- BIC penalizes complex models more severely - better to use BIC than AIC.

- There are several such goodness of fit measure.
- These measures are useful in selection of a single best model among several competing models.
- R-squared - $R^2 = \frac{Var(Y)}{Var(y)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$; $0 \leq R^2 \leq 1$.
- Problem of R^2 - tend to select overfitting models.
- Adjusted R-squared - $R^2_{adj} = 1 - \frac{(n-1)(1-R^2)}{(n-2)}$ - higher the better - can be negative!
- AIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + 2(2+1)$ - lower the better.
- BIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + \ln(n)(2+1)$ - lower the better.
- BIC penalizes complex models more severely - better to use BIC than AIC.

- There are several such goodness of fit measure.
- These measures are useful in selection of a single best model among several competing models.
- R-squared - $R^2 = \frac{Var(Y)}{Var(y)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$; $0 \leq R^2 \leq 1$.
- Problem of R^2 - tend to select overfitting models.
- Adjusted R-squared - $R^2_{adj} = 1 - \frac{(n-1)(1-R^2)}{(n-2)}$ - higher the better - can be negative!
- AIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + 2(2+1)$ - lower the better.
- BIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + \ln(n)(2+1)$ - lower the better.
- BIC penalizes complex models more severely - better to use BIC than AIC.

- There are several such goodness of fit measure.
- These measures are useful in selection of a single best model among several competing models.
- R-squared - $R^2 = \frac{Var(Y)}{Var(y)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$; $0 \leq R^2 \leq 1$.
- Problem of R^2 - tend to select overfitting models.
- Adjusted R-squared - $R^2_{adj} = 1 - \frac{(n-1)(1-R^2)}{(n-2)}$ - higher the better - can be negative!
- AIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + 2(2+1)$ - lower the better.
- BIC - $-2\ln(L(\hat{\beta}_{mle}, \hat{\sigma}^2_{\epsilon mle}|\mathbf{y}, \mathbf{X})) + \ln(n)(2+1)$ - lower the better.
- BIC penalizes complex models more severely - better to use BIC than AIC.

- **Multiple linear regression model :**
  More familiar specification - $\mathbf{Y} = \mathbf{X}\beta + \epsilon;\ \epsilon \sim MVN(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$.

- $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots x_{p1} \\ 1 & x_2 & \cdots x_{p2} \\ \vdots & & \\ 1 & x_n & \cdots x_{pn} \end{pmatrix}$ and $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ are unobserved random errors. $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$.

- All the previous developments are applicable.

- **Polynomial regression model :** $[Y_i | X_i = x_i] \overset{ind}{\sim} N(a + bx_i + cx_i^2, \sigma_\epsilon^2)$ is a special case.

- **Multiple linear regression model :**
  More familiar specification - $\mathbf{Y} = \mathbf{X}\beta + \epsilon;\ \epsilon \sim MVN(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$.

- $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots x_{p1} \\ 1 & x_2 & \cdots x_{p2} \\ \vdots & & \\ 1 & x_n & \cdots x_{pn} \end{pmatrix}$ and $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ are unobserved random errors. $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \beta_p \end{pmatrix}$.

- All the previous developments are applicable.

- **Polynomial regression model :** $[Y_i | X_i = x_i] \overset{ind}{\sim} N(a + bx_i + cx_i^2, \sigma_\epsilon^2)$ is a special case.

- **Multiple linear regression model :**
  More familiar specification - $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon};\ \boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$.

- $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots x_{p1} \\ 1 & x_2 & \cdots x_{p2} \\ \vdots & & \\ 1 & x_n & \cdots x_{pn} \end{pmatrix}$ and $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ are unobserved random errors. $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$.

- All the previous developments are applicable.

- **Polynomial regression model :** $[Y_i | X_i = x_i] \overset{ind}{\sim} N(a + bx_i + cx_i^2, \sigma_\epsilon^2)$ is a special case.

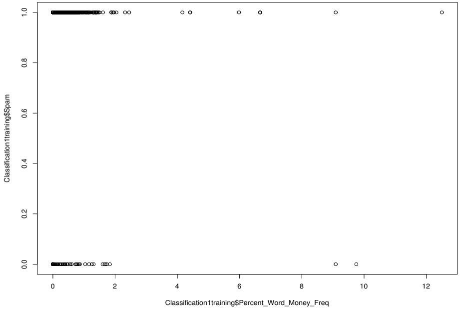
- **Multiple linear regression model :**
More familiar specification - $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon};\ \boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$.

- $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots x_{p1} \\ 1 & x_2 & \cdots x_{p2} \\ \vdots & & \\ 1 & x_n & \cdots x_{pn} \end{pmatrix}$ and $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ are unobserved random errors. $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$.

- All the previous developments are applicable.

- **Polynomial regression model :** $[Y_i | X_i = x_i] \overset{ind}{\sim} N(a + bx_i + cx_i^2, \sigma_\epsilon^2)$ is a special case.

- Observations taken on two features - covariate is continuous say dosage of a drug (x) and response (y) is binary subject is alive/dead (we code it as $0/1$).
- Scatter plot of (x) and (y) does not give much insight!

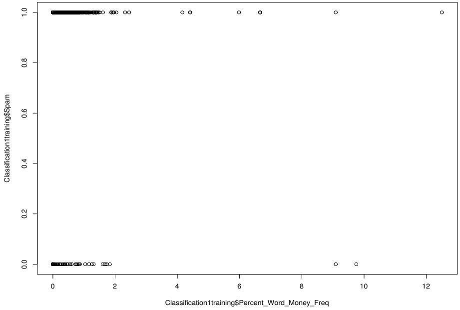


(a)

Figure: Scatter plot of x and $y(0/1)$ - not useful.

- Not much of descriptive statistics can be done.
- Still - need some motivation!

- Observations taken on two features - covariate is continuous say dosage of a drug (x) and response (y) is binary subject is alive/dead (we code it as $0/1$).
- Scatter plot of (x) and (y) does not give much insight!

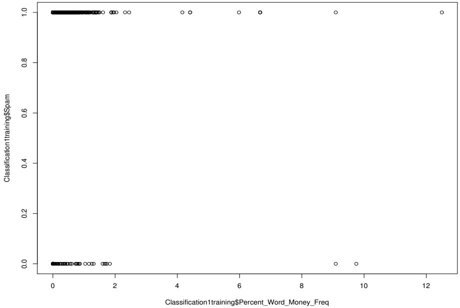


(a)

Figure: Scatter plot of x and $y(0/1)$ - not useful.

- Not much of descriptive statistics can be done.
- Still - need some motivation!

- Observations taken on two features - covariate is continuous say dosage of a drug (x) and response (y) is binary subject is alive/dead (we code it as $0/1$).
- Scatter plot of (x) and (y) does not give much insight!

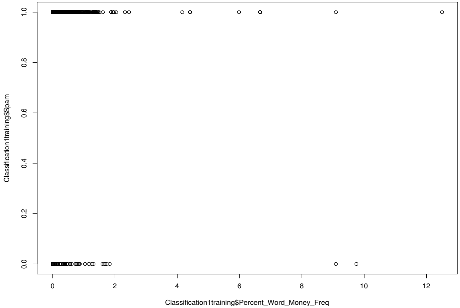


(a)

Figure: Scatter plot of x and $y(0/1)$ - not useful.

- Not much of descriptive statistics can be done.
- Still - need some motivation!

- Observations taken on two features - covariate is continuous say dosage of a drug (x) and response (y) is binary subject is alive/dead (we code it as $0/1$).
- Scatter plot of (x) and (y) does not give much insight!



(a)

Figure: Scatter plot of x and $y(0/1)$ - not useful.

- Not much of descriptive statistics can be done.
- Still - need some motivation!

- In simple linear regression model we have assumption $E(Y|X=x) = a + bx$.
- Now for the logistic regression model we have assumption $E(Y|X=x) = 1 \times P(Y=1|X=x) + 0 \times P(Y=0|X=x) = P(Y=1|X=x) = a + bx$?? - meaningless
- $0 \leq P(Y=1|X=x) \leq 1$ but $-\infty < a + bx < +\infty$ for $b \neq 0$.
- However, $P(Y=1|X=x) = \frac{e^{a+bx}}{1+e^{a+bx}}$ - absolutely meaningful.
- $\frac{e^{a+bx}}{1+e^{a+bx}}$ - logistic distribution - so the name logistic regression.
- $logit(P(Y=1|X=x)) = log(\text{ODDS for Y=1}) = log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = a + bx$ - so the name logit regression.
- If not coded using dummy variables - $P(Y = \text{``}dead\text{''}|X=x) = \frac{e^{a+bx}}{1+e^{a+bx}}$.
- Reason?(i) Very Simple Form.
  (ii) Lots of Similarity with Linear Regression Model.
  (iii) Logistic Regression Model/Logit Regression Model is Highly Successful!

- In simple linear regression model we have assumption $E(Y|X=x) = a + bx$.
- Now for the logistic regression model we have assumption $E(Y|X=x) = 1 \times P(Y=1|X=x) + 0 \times P(Y=0|X=x) = P(Y=1|X=x) = a + bx$?? - meaningless
- $0 \leq P(Y=1|X=x) \leq 1$ but $-\infty < a + bx < +\infty$ for $b \neq 0$.
- However, $P(Y=1|X=x) = \frac{e^{a+bx}}{1+e^{a+bx}}$ - absolutely meaningful.
- $\frac{e^{a+bx}}{1+e^{a+bx}}$ - logistic distribution - so the name logistic regression.
- $logit(P(Y=1|X=x)) = log(\text{ODDS for Y=1}) = log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = a + bx$ - so the name logit regression.
- If not coded using dummy variables - $P(Y = \text{"dead"}|X=x) = \frac{e^{a+bx}}{1+e^{a+bx}}$.
- Reason?(i) Very Simple Form.
  (ii) Lots of Similarity with Linear Regression Model.
  (iii) Logistic Regression Model/Logit Regression Model is Highly Successful!

- In simple linear regression model we have assumption $E(Y|X = x) = a + bx$.
- Now for the logistic regression model we have assumption $E(Y|X = x) = 1 \times P(Y = 1|X = x) + 0 \times P(Y = 0|X = x) = P(Y = 1|X = x) = a + bx$?? - meaningless
- $0 \leq P(Y = 1|X = x) \leq 1$ but $-\infty < a + bx < +\infty$ for $b \neq 0$.
- However, $P(Y = 1|X = x) = \frac{e^{a+bx}}{1+e^{a+bx}}$ - absolutely meaningful.
- $\frac{e^{a+bx}}{1+e^{a+bx}}$ - logistic distribution - so the name logistic regression.
- $logit(P(Y = 1|X = x)) = log(\text{ODDS for Y=1}) = log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = a + bx$ - so the name logit regression.
- If not coded using dummy variables - $P(Y = \text{“}dead\text{''}|X = x) = \frac{e^{a+bx}}{1+e^{a+bx}}$.
-

Reason? (i) Very Simple Form.
(ii) Lots of Similarity with Linear Regression Model.
(iii) Logistic Regression Model/Logit Regression Model is Highly Successful!

- In simple linear regression model we have assumption $E(Y|X=x) = a + bx$.
- Now for the logistic regression model we have assumption $E(Y|X=x) = 1 \times P(Y=1|X=x) + 0 \times P(Y=0|X=x) = P(Y=1|X=x) = a + bx$?? - meaningless
- $0 \leq P(Y=1|X=x) \leq 1$ but $-\infty < a + bx < +\infty$ for $b \neq 0$.
- However, $P(Y=1|X=x) = \frac{e^{a+bx}}{1+e^{a+bx}}$ - absolutely meaningful.
- $\frac{e^{a+bx}}{1+e^{a+bx}}$ - logistic distribution - so the name logistic regression.
- $logit(P(Y=1|X=x)) = log(\text{ODDS for Y=1}) = log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = a + bx$ - so the name logit regression.
- If not coded using dummy variables - $P(Y = \text{“}dead\text{''}|X=x) = \frac{e^{a+bx}}{1+e^{a+bx}}$.
-

Reason?(i) Very Simple Form.
(ii) Lots of Similarity with Linear Regression Model.
(iii) Logistic Regression Model/Logit Regression Model is Highly Successful!

- In simple linear regression model we have assumption $E(Y|X=x) = a + bx$.
- Now for the logistic regression model we have assumption $E(Y|X=x) = 1 \times P(Y=1|X=x) + 0 \times P(Y=0|X=x) = P(Y=1|X=x) = a + bx$?? - meaningless
- $0 \leq P(Y=1|X=x) \leq 1$ but $-\infty < a + bx < +\infty$ for $b \neq 0$.
- However, $P(Y=1|X=x) = \frac{e^{a+bx}}{1+e^{a+bx}}$ - absolutely meaningful.
- $\frac{e^{a+bx}}{1+e^{a+bx}}$ - logistic distribution - so the name logistic regression.
- $logit(P(Y=1|X=x)) = log(\text{ODDS for Y=1}) = log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = a + bx$ - so the name logit regression.
- If not coded using dummy variables - $P(Y = \text{“}dead\text{''}|X=x) = \frac{e^{a+bx}}{1+e^{a+bx}}$.
-

Reason? (i) Very Simple Form.
(ii) Lots of Similarity with Linear Regression Model.
(iii) Logistic Regression Model/Logit Regression Model is Highly Successful!

- In simple linear regression model we have assumption $E(Y|X=x) = a+bx$.
- Now for the logistic regression model we have assumption $E(Y|X=x) = 1 \times P(Y=1|X=x) + 0 \times P(Y=0|X=x) = P(Y=1|X=x) = a+bx$?? - meaningless
- $0 \leq P(Y=1|X=x) \leq 1$ but $-\infty < a+bx < +\infty$ for $b \neq 0$.
- However, $P(Y=1|X=x) = \frac{e^{a+bx}}{1+e^{a+bx}}$ - absolutely meaningful.
- $\frac{e^{a+bx}}{1+e^{a+bx}}$ - logistic distribution - so the name logistic regression.
- $logit(P(Y=1|X=x)) = log(\text{ODDS for Y=1}) = log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = a+bx$ - so the name logit regression.
- If not coded using dummy variables - $P(Y = \text{“}dead\text{''}|X=x) = \frac{e^{a+bx}}{1+e^{a+bx}}$.
-

Reason?(i) Very Simple Form.
(ii) Lots of Similarity with Linear Regression Model.
(iii) Logistic Regression Model/Logit Regression Model is Highly Successful!

- In simple linear regression model we have assumption $E(Y|X=x) = a + bx$.
- Now for the logistic regression model we have assumption $E(Y|X=x) = 1 \times P(Y=1|X=x) + 0 \times P(Y=0|X=x) = P(Y=1|X=x) = a + bx$?? - meaningless
- $0 \leq P(Y=1|X=x) \leq 1$ but $-\infty < a + bx < +\infty$ for $b \neq 0$.
- However, $P(Y=1|X=x) = \frac{e^{a+bx}}{1+e^{a+bx}}$ - absolutely meaningful.
- $\frac{e^{a+bx}}{1+e^{a+bx}}$ - logistic distribution - so the name logistic regression.
- $logit(P(Y=1|X=x)) = log(\text{ODDS for Y=1}) = log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = a + bx$ - so the name logit regression.
- If not coded using dummy variables - $P(Y = \text{“}dead'' | X = x) = \frac{e^{a+bx}}{1+e^{a+bx}}$.

Reason?(i) Very Simple Form.
(ii) Lots of Similarity with Linear Regression Model.
(iii) Logistic Regression Model/Logit Regression Model is Highly Successful!

- In simple linear regression model we have assumption $E(Y|X=x)=a+bx$.
- Now for the logistic regression model we have assumption $E(Y|X=x)= 1\times P(Y=1|X=x)+0\times P(Y=0|X=x)=P(Y=1|X=x)=a+bx$?? - meaningless
- $0\leq P(Y=1|X=x)\leq 1$ but $-\infty < a+bx < +\infty$ for $b\neq 0$.
- However, $P(Y=1|X=x)=\frac{e^{a+bx}}{1+e^{a+bx}}$ - absolutely meaningful.
- $\frac{e^{a+bx}}{1+e^{a+bx}}$ - logistic distribution - so the name logistic regression.
- $logit(P(Y=1|X=x)) = log(\text{ODDS for Y=1}) = log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = a+bx$ - so the name logit regression.
- If not coded using dummy variables - $P(Y=\text{``}dead''|X=x)=\frac{e^{a+bx}}{1+e^{a+bx}}$.
- Reason?(i) Very Simple Form.
  (ii) Lots of Similarity with Linear Regression Model.
  (iii) Logistic Regression Model/Logit Regression Model is Highly Successful!

- Logistic regression model used in
  (a) spam detection based on certain words and characters.
  (b) malignant tumor detection based on certain cell profiles.
  (c) loan defaulters detection based on personal/socio-economic and demographic profiles.

- Difference with linear regression - no closed form solution available.

- **Simple logistic regression model :**

$$[Y_1 = y_1, \cdots, Y_n = y_n | X_1 = x_1, \cdots, X_n = x_n] \sim$$
$$\prod_{i=1}^{n} [P(Y = 1|X = x_i)]^{y_i} [1 - P(Y = 1|X = x_i)]^{1-y_i} =$$
$$\prod_{i=1}^{n} [\frac{e^{a+bx_i}}{1 + e^{a+bx_i}}]^{y_i} [\frac{1}{1 + e^{a+bx_i}}]^{1-y_i}$$

- Model parameters - a, b.

- The model is nothing but a family of product of Bernoulli distributions indexed by unknown parameters a, b.

- More familiar specification - $[Y_i | X_i = x_i] \overset{ind}{\sim} Ber(\frac{e^{a+bx}}{1+e^{a+bx}})$.

- Logistic regression model used in
  (a) spam detection based on certain words and characters.
  (b) malignant tumor detection based on certain cell profiles.
  (c) loan defaulters detection based on personal/socio-economic and demographic profiles.
- Difference with linear regression - no closed form solution available.
- **Simple logistic regression model :**

$$[Y_1 = y_1, \cdots, Y_n = y_n | X_1 = x_1, \cdots, X_n = x_n] \sim$$
$$\prod_{i=1}^{n} [P(Y = 1 | X = x_i)]^{y_i} [1 - P(Y = 1 | X = x_i)]^{1-y_i} =$$
$$\prod_{i=1}^{n} [\frac{e^{a+bx_i}}{1 + e^{a+bx_i}}]^{y_i} [\frac{1}{1 + e^{a+bx_i}}]^{1-y_i}$$

- Model parameters - a, b.
- The model is nothing but a family of product of Bernoulli distributions indexed by unknown parameters a, b.
- More familiar specification - $[Y_i | X_i = x_i] \overset{ind}{\sim} Ber(\frac{e^{a+bx}}{1+e^{a+bx}})$.

- Logistic regression model used in
  (a) spam detection based on certain words and characters.
  (b) malignant tumor detection based on certain cell profiles.
  (c) loan defaulters detection based on personal/socio-economic and demographic profiles.

- Difference with linear regression - no closed form solution available.

- **Simple logistic regression model :**

$$[Y_1 = y_1, \cdots, Y_n = y_n | X_1 = x_1, \cdots, X_n = x_n] \sim$$
$$\prod_{i=1}^{n} [P(Y = 1|X = x_i)]^{y_i} [1 - P(Y = 1|X = x_i)]^{1-y_i} =$$
$$\prod_{i=1}^{n} [\frac{e^{a+bx_i}}{1 + e^{a+bx_i}}]^{y_i} [\frac{1}{1 + e^{a+bx_i}}]^{1-y_i}$$

- Model parameters - a, b.

- The model is nothing but a family of product of Bernoulli distributions indexed by unknown parameters a, b.

- More familiar specification - $[Y_i | X_i = x_i] \overset{ind}{\sim} Ber(\frac{e^{a+bx}}{1+e^{a+bx}})$.

- Logistic regression model used in
  (a) spam detection based on certain words and characters.
  (b) malignant tumor detection based on certain cell profiles.
  (c) loan defaulters detection based on personal/socio-economic and demographic profiles.

- Difference with linear regression - no closed form solution available.

- **Simple logistic regression model :**

$$[Y_1 = y_1, \cdots, Y_n = y_n | X_1 = x_1, \cdots, X_n = x_n] \sim$$
$$\prod_{i=1}^{n} [P(Y = 1|X = x_i)]^{y_i} [1 - P(Y = 1|X = x_i)]^{1-y_i} =$$
$$\prod_{i=1}^{n} [\frac{e^{a+bx_i}}{1 + e^{a+bx_i}}]^{y_i} [\frac{1}{1 + e^{a+bx_i}}]^{1-y_i}$$

- Model parameters - a, b.

- The model is nothing but a family of product of Bernoulli distributions indexed by unknown parameters a, b.

- More familiar specification - $[Y_i | X_i = x_i] \overset{ind}{\sim} Ber(\frac{e^{a+bx}}{1+e^{a+bx}})$.

- Logistic regression model used in
  (a) spam detection based on certain words and characters.
  (b) malignant tumor detection based on certain cell profiles.
  (c) loan defaulters detection based on personal/socio-economic and demographic profiles.

- Difference with linear regression - no closed form solution available.

- **Simple logistic regression model :**

$$[Y_1 = y_1, \cdots, Y_n = y_n | X_1 = x_1, \cdots, X_n = x_n] \sim$$
$$\prod_{i=1}^{n} [P(Y=1|X=x_i)]^{y_i} [1 - P(Y=1|X=x_i)]^{1-y_i} =$$
$$\prod_{i=1}^{n} [\frac{e^{a+bx_i}}{1+e^{a+bx_i}}]^{y_i} [\frac{1}{1+e^{a+bx_i}}]^{1-y_i}$$

- Model parameters - a, b.

- The model is nothing but a family of product of Bernoulli distributions indexed by unknown parameters a, b.

- More familiar specification - $[Y_i | X_i = x_i] \overset{ind}{\sim} Ber(\frac{e^{a+bx}}{1+e^{a+bx}})$.

- Logistic regression model used in
  (a) spam detection based on certain words and characters.
  (b) malignant tumor detection based on certain cell profiles.
  (c) loan defaulters detection based on personal/socio-economic and demographic profiles.

- Difference with linear regression - no closed form solution available.

- **Simple logistic regression model :**

$$[Y_1 = y_1, \cdots, Y_n = y_n | X_1 = x_1, \cdots, X_n = x_n] \sim$$
$$\prod_{i=1}^{n} [P(Y=1|X=x_i)]^{y_i} [1 - P(Y=1|X=x_i)]^{1-y_i} =$$
$$\prod_{i=1}^{n} [\frac{e^{a+bx_i}}{1+e^{a+bx_i}}]^{y_i} [\frac{1}{1+e^{a+bx_i}}]^{1-y_i}$$

- Model parameters - a, b.

- The model is nothing but a family of product of Bernoulli distributions indexed by unknown parameters a, b.

- More familiar specification - $[Y_i|X_i = x_i] \overset{ind}{\sim} Ber(\frac{e^{a+bx}}{1+e^{a+bx}})$.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating the parameter vector $\beta = (a, b)'$.
- mle of β is denoted by $\hat{\beta}$ - unlike linear regression no closed form expression.
- mle is calculated using numerical algorithm - Fisher's scoring algorithm.
- Often the algorithm may not converge - multicollinearity, sparseness and complete separation.
- multicollinearity : when covariate/predictor variables are linearly highly correlated.
- sparseness : for some combinations of covariate variables we do not get any data.
- complete separation : beyond some combination threshold value only $Y = 1$ or only $Y = 0$ responses are obtained.
- For the simple logistic regression model instead of Fisher's scoring one often use Newton-Raphson method.
- For the simple logistic regression model Newton-Raphson method become a iteratively reweighted least squares (IRLS) algorithm.
- IRLS form is highly useful since calculation of least squares is relatively easy.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating the parameter vector $\beta = (a, b)'$.
- mle of β is denoted by $\hat{\beta}$ - unlike linear regression no closed form expression.
- mle is calculated using numerical algorithm - Fisher's scoring algorithm.
- Often the algorithm may not converge - multicollinearity, sparseness and complete separation.
- multicollinearity : when covariate/predictor variables are linearly highly correlated.
- sparseness : for some combinations of covariate variables we do not get any data.
- complete separation : beyond some combination threshold value only $Y = 1$ or only $Y = 0$ responses are obtained.
- For the simple logistic regression model instead of Fisher's scoring one often use Newton-Raphson method.
- For the simple logistic regression model Newton-Raphson method become a iteratively reweighted least squares (IRLS) algorithm.
- IRLS form is highly useful since calculation of least squares is relatively easy.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating the parameter vector $\boldsymbol{\beta} = (a, b)'$.
- mle of $\boldsymbol{\beta}$ is denoted by $\hat{\boldsymbol{\beta}}$ - unlike linear regression no closed form expression.
- mle is calculated using numerical algorithm - Fisher's scoring algorithm.
- Often the algorithm may not converge - multicollinearity, sparseness and complete separation.
- multicollinearity : when covariate/predictor variables are linearly highly correlated.
- sparseness : for some combinations of covariate variables we do not get any data.
- complete separation : beyond some combination threshold value only $Y = 1$ or only $Y = 0$ responses are obtained.
- For the simple logistic regression model instead of Fisher's scoring one often use Newton-Raphson method.
- For the simple logistic regression model Newton-Raphson method become a iteratively reweighted least squares (IRLS) algorithm.
- IRLS form is highly useful since calculation of least squares is relatively easy.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating the parameter vector $\beta = (a, b)'$.
- mle of β is denoted by $\hat{\beta}$ - unlike linear regression no closed form expression.
- mle is calculated using numerical algorithm - Fisher's scoring algorithm.
- Often the algorithm may not converge - multicollinearity, sparseness and complete separation.
- multicollinearity : when covariate/predictor variables are linearly highly correlated.
- sparseness : for some combinations of covariate variables we do not get any data.
- complete separation : beyond some combination threshold value only $Y = 1$ or only $Y = 0$ responses are obtained.
- For the simple logistic regression model instead of Fisher's scoring one often use Newton-Raphson method.
- For the simple logistic regression model Newton-Raphson method become a iteratively reweighted least squares (IRLS) algorithm.
- IRLS form is highly useful since calculation of least squares is relatively easy.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating the parameter vector $\boldsymbol{\beta} = (a, b)'$.
- mle of $\boldsymbol{\beta}$ is denoted by $\hat{\boldsymbol{\beta}}$ - unlike linear regression no closed form expression.
- mle is calculated using numerical algorithm - Fisher's scoring algorithm.
- Often the algorithm may not converge - multicollinearity, sparseness and complete separation.
- multicollinearity : when covariate/predictor variables are linearly highly correlated.
- sparseness : for some combinations of covariate variables we do not get any data.
- complete separation : beyond some combination threshold value only $Y = 1$ or only $Y = 0$ responses are obtained.
- For the simple logistic regression model instead of Fisher's scoring one often use Newton-Raphson method.
- For the simple logistic regression model Newton-Raphson method become a iteratively reweighted least squares (IRLS) algorithm.
- IRLS form is highly useful since calculation of least squares is relatively easy.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating the parameter vector $\beta = (a, b)'$.
- mle of β is denoted by $\hat{\beta}$ - unlike linear regression no closed form expression.
- mle is calculated using numerical algorithm - Fisher's scoring algorithm.
- Often the algorithm may not converge - multicollinearity, sparseness and complete separation.
- multicollinearity : when covariate/predictor variables are linearly highly correlated.
- sparseness : for some combinations of covariate variables we do not get any data.
- complete separation : beyond some combination threshold value only $Y = 1$ or only $Y = 0$ responses are obtained.
- For the simple logistic regression model instead of Fisher's scoring one often use Newton-Raphson method.
- For the simple logistic regression model Newton-Raphson method become a iteratively reweighted least squares (IRLS) algorithm.
- IRLS form is highly useful since calculation of least squares is relatively easy.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating the parameter vector $\boldsymbol{\beta} = (a, b)'$.
- mle of $\boldsymbol{\beta}$ is denoted by $\hat{\boldsymbol{\beta}}$ - unlike linear regression no closed form expression.
- mle is calculated using numerical algorithm - Fisher's scoring algorithm.
- Often the algorithm may not converge - multicollinearity, sparseness and complete separation.
- multicollinearity : when covariate/predictor variables are linearly highly correlated.
- sparseness : for some combinations of covariate variables we do not get any data.
- complete separation : beyond some combination threshold value only $Y = 1$ or only $Y = 0$ responses are obtained.
- For the simple logistic regression model instead of Fisher's scoring one often use Newton-Raphson method.
- For the simple logistic regression model Newton-Raphson method become a iteratively reweighted least squares (IRLS) algorithm.
- IRLS form is highly useful since calculation of least squares is relatively easy.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating the parameter vector $\boldsymbol{\beta} = (a, b)'$.
- mle of $\boldsymbol{\beta}$ is denoted by $\hat{\boldsymbol{\beta}}$ - unlike linear regression no closed form expression.
- mle is calculated using numerical algorithm - Fisher's scoring algorithm.
- Often the algorithm may not converge - multicollinearity, sparseness and complete separation.
- multicollinearity : when covariate/predictor variables are linearly highly correlated.
- sparseness : for some combinations of covariate variables we do not get any data.
- complete separation : beyond some combination threshold value only $Y = 1$ or only $Y = 0$ responses are obtained.
- For the simple logistic regression model instead of Fisher's scoring one often use Newton-Raphson method.
- For the simple logistic regression model Newton-Raphson method become a iteratively reweighted least squares (IRLS) algorithm.
- IRLS form is highly useful since calculation of least squares is relatively easy.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating the parameter vector $\boldsymbol{\beta} = (a, b)'$.
- mle of $\boldsymbol{\beta}$ is denoted by $\hat{\boldsymbol{\beta}}$ - unlike linear regression no closed form expression.
- mle is calculated using numerical algorithm - Fisher's scoring algorithm.
- Often the algorithm may not converge - multicollinearity, sparseness and complete separation.
- multicollinearity : when covariate/predictor variables are linearly highly correlated.
- sparseness : for some combinations of covariate variables we do not get any data.
- complete separation : beyond some combination threshold value only $Y = 1$ or only $Y = 0$ responses are obtained.
- For the simple logistic regression model instead of Fisher's scoring one often use Newton-Raphson method.
- For the simple logistic regression model Newton-Raphson method become a iteratively reweighted least squares (IRLS) algorithm.
- IRLS form is highly useful since calculation of least squares is relatively easy.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating the parameter vector $\boldsymbol{\beta} = (a, b)'$.
- mle of $\boldsymbol{\beta}$ is denoted by $\hat{\boldsymbol{\beta}}$ - unlike linear regression no closed form expression.
- mle is calculated using numerical algorithm - Fisher's scoring algorithm.
- Often the algorithm may not converge - multicollinearity, sparseness and complete separation.
- multicollinearity : when covariate/predictor variables are linearly highly correlated.
- sparseness : for some combinations of covariate variables we do not get any data.
- complete separation : beyond some combination threshold value only $Y = 1$ or only $Y = 0$ responses are obtained.
- For the simple logistic regression model instead of Fisher's scoring one often use Newton-Raphson method.
- For the simple logistic regression model Newton-Raphson method become a iteratively reweighted least squares (IRLS) algorithm.
- IRLS form is highly useful since calculation of least squares is relatively easy.

- The model is fitted using maximum likelihood method.
- Inferential goal - estimating the parameter vector $\boldsymbol{\beta} = (a, b)'$.
- mle of $\boldsymbol{\beta}$ is denoted by $\hat{\boldsymbol{\beta}}$ - unlike linear regression no closed form expression.
- mle is calculated using numerical algorithm - Fisher's scoring algorithm.
- Often the algorithm may not converge - multicollinearity, sparseness and complete separation.
- multicollinearity : when covariate/predictor variables are linearly highly correlated.
- sparseness : for some combinations of covariate variables we do not get any data.
- complete separation : beyond some combination threshold value only $Y = 1$ or only $Y = 0$ responses are obtained.
- For the simple logistic regression model instead of Fisher's scoring one often use Newton-Raphson method.
- For the simple logistic regression model Newton-Raphson method become a iteratively reweighted least squares (IRLS) algorithm.
- IRLS form is highly useful since calculation of least squares is relatively easy.

- Another inferential goal - testing for β.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $Z = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Finite sample null distribution is not available - asymptotic null distribution (assuming no. of data n large) of test statistic $\sim N(0,1)$ - Cutoff is obtained using standard normal table.
- Practitioners prefer p-value - $P(Z > |Z_{observed}|)$ where $Z \sim N(0,1)$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $Z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- asymptotic null distribution of test statistic $\sim N(0,1)$ - Cutoff is obtained using $N(0,1)$-distribution table.
- Asymptotically approximate confidence intervals can be obtained for the parameters β_0 and β_1 inverting the Ztest statistics.
- Goodness of fit measures.
- Want something like R^2.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $Z = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Finite sample null distribution is not available - asymptotic null distribution (assuming no. of data n large) of test statistic $\sim N(0, 1)$ - Cutoff is obtained using standard normal table.
- Practitioners prefer p-value - $P(Z > |Z_{observed}|)$ where $Z \sim N(0, 1)$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $Z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- asymptotic null distribution of test statistic $\sim N(0, 1)$ - Cutoff is obtained using $N(0, 1)$-distribution table.
- Asymptotically approximate confidence intervals can be obtained for the parameters β_0 and β_1 inverting the Ztest statistics.
- Goodness of fit measures.
- Want something like R^2.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $Z = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Finite sample null distribution is not available - asymptotic null distribution (assuming no. of data n large) of test statistic $\sim N(0, 1)$ - Cutoff is obtained using standard normal table.
- Practitioners prefer p-value - $P(Z > |Z_{observed}|)$ where $Z \sim N(0, 1)$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $Z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- asymptotic null distribution of test statistic $\sim N(0, 1)$ - Cutoff is obtained using $N(0, 1)$-distribution table.
- Asymptotically approximate confidence intervals can be obtained for the parameters β_0 and β_1 inverting the Ztest statistics.
- Goodness of fit measures.
- Want something like R^2.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $Z = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Finite sample null distribution is not available - asymptotic null distribution (assuming no. of data n large) of test statistic $\sim N(0,1)$ - Cutoff is obtained using standard normal table.
- Practitioners prefer p-value - $P(Z > |Z_{observed}|)$ where $Z \sim N(0,1)$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $Z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- asymptotic null distribution of test statistic $\sim N(0,1)$ - Cutoff is obtained using $N(0,1)$-distribution table.
- Asymptotically approximate confidence intervals can be obtained for the parameters β_0 and β_1 inverting the Ztest statistics.
- Goodness of fit measures.
- Want something like R^2.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $Z = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Finite sample null distribution is not available - asymptotic null distribution (assuming no. of data n large) of test statistic $\sim N(0,1)$ - Cutoff is obtained using standard normal table.
- Practitioners prefer p-value - $P(Z > |Z_{observed}|)$ where $Z \sim N(0,1)$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $Z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- asymptotic null distribution of test statistic $\sim N(0,1)$ - Cutoff is obtained using $N(0,1)$-distribution table.
- Asymptotically approximate confidence intervals can be obtained for the parameters β_0 and β_1 inverting the Ztest statistics.
- Goodness of fit measures.
- Want something like R^2.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $Z = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Finite sample null distribution is not available - asymptotic null distribution (assuming no. of data n large) of test statistic $\sim N(0,1)$ - Cutoff is obtained using standard normal table.
- Practitioners prefer p-value - $P(Z > |Z_{observed}|)$ where $Z \sim N(0,1)$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $Z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- asymptotic null distribution of test statistic $\sim N(0,1)$ - Cutoff is obtained using $N(0,1)$-distribution table.
- Asymptotically approximate confidence intervals can be obtained for the parameters β_0 and β_1 inverting the Ztest statistics.
- Goodness of fit measures.
- Want something like R^2.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $Z = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Finite sample null distribution is not available - asymptotic null distribution (assuming no. of data n large) of test statistic $\sim N(0, 1)$ - Cutoff is obtained using standard normal table.
- Practitioners prefer p-value - $P(Z > |Z_{observed}|)$ where $Z \sim N(0, 1)$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $Z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- asymptotic null distribution of test statistic $\sim N(0, 1)$ - Cutoff is obtained using $N(0, 1)$-distribution table.
- Asymptotically approximate confidence intervals can be obtained for the parameters β_0 and β_1 inverting the Ztest statistics.
- Goodness of fit measures.
- Want something like R^2.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $Z = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Finite sample null distribution is not available - asymptotic null distribution (assuming no. of data n large) of test statistic $\sim N(0, 1)$ - Cutoff is obtained using standard normal table.
- Practitioners prefer p-value - $P(Z > |Z_{observed}|)$ where $Z \sim N(0, 1)$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $Z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- asymptotic null distribution of test statistic $\sim N(0, 1)$ - Cutoff is obtained using $N(0, 1)$-distribution table.
- Asymptotically approximate confidence intervals can be obtained for the parameters β_0 and β_1 inverting the Ztest statistics.
- Goodness of fit measures.
- Want something like R^2.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $Z = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Finite sample null distribution is not available - asymptotic null distribution (assuming no. of data n large) of test statistic $\sim N(0,1)$ - Cutoff is obtained using standard normal table.
- Practitioners prefer p-value - $P(Z > |Z_{observed}|)$ where $Z \sim N(0,1)$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $Z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- asymptotic null distribution of test statistic $\sim N(0,1)$ - Cutoff is obtained using $N(0,1)$-distribution table.
- Asymptotically approximate confidence intervals can be obtained for the parameters β_0 and β_1 inverting the Ztest statistics.
- Goodness of fit measures.
- Want something like R^2.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $Z = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$.
- Finite sample null distribution is not available - asymptotic null distribution (assuming no. of data n large) of test statistic $\sim N(0,1)$ - Cutoff is obtained using standard normal table.
- Practitioners prefer p-value - $P(Z > |Z_{observed}|)$ where $Z \sim N(0,1)$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $Z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.
- asymptotic null distribution of test statistic $\sim N(0,1)$ - Cutoff is obtained using $N(0,1)$-distribution table.
- Asymptotically approximate confidence intervals can be obtained for the parameters β_0 and β_1 inverting the Ztest statistics.
- Goodness of fit measures.
- Want something like R^2.

- Another inferential goal - testing for $\boldsymbol{\beta}$.
- Individual test of significance $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ (test of intercept).
- Test statistic $Z = \frac{\hat{\beta}_0}{\widehat{se(\hat{\beta}_0)}}$.
- Finite sample null distribution is not available - asymptotic null distribution (assuming no. of data n large) of test statistic $\sim N(0,1)$ - Cutoff is obtained using standard normal table.
- Practitioners prefer p-value - $P(Z > |Z_{observed}|)$ where $Z \sim N(0,1)$.
- Individual test of significance $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (test of slope).
- Test statistic $Z = \frac{\hat{\beta}_1}{\widehat{se(\hat{\beta}_1)}}$.
- asymptotic null distribution of test statistic $\sim N(0,1)$ - Cutoff is obtained using $N(0,1)$-distribution table.
- Asymptotically approximate confidence intervals can be obtained for the parameters β_0 and β_1 inverting the Ztest statistics.
- Goodness of fit measures.
- Want something like R^2.

- Deviance measure : $D_{fitted} = -2\ln(L(\hat{\boldsymbol{\beta}}_{mle}|\mathbf{y},\mathbf{X}))$.
- Null model means only intercept term - no regressors.
- $D_{null} = -2\ln(L(\hat{\beta}_{0mle}|\mathbf{y},\mathbf{X}))$.
- $D_{null} - D_{fitted} \geq 0$.
- If $D_{null} - D_{fitted}$ very large then we can reject the hypothesis of no regression.
- H_0 : all coefficients except β_0 is 0 vs H_1 : not H_0 (test of regression is needed or not/no regressors).
- This test is analogue of F-test in linear regression models.
- Can construct a pseudo-R squared based on Deviance : $R_L^2 = \frac{D_{null} - D_{fitted}}{D_{null}}$
- R_L^2 - larger value indicates good fit.

- Deviance measure : $D_{fitted} = -2\ln(L(\hat{\boldsymbol{\beta}}_{mle}|\mathbf{y}, \mathbf{X}))$.
- Null model means only intercept term - no regressors.
- $D_{null} = -2\ln(L(\hat{\beta}_{0mle}|\mathbf{y}, \mathbf{X}))$.
- $D_{null} - D_{fitted} \geq 0$.
- If $D_{null} - D_{fitted}$ very large then we can reject the hypothesis of no regression.
- H_0 : all coefficients except β_0 is 0 vs H_1 : not H_0 (test of regression is needed or not/no regressors).
- This test is analogue of F-test in linear regression models.
- Can construct a pseudo-R squared based on Deviance : $R_L^2 = \frac{D_{null} - D_{fitted}}{D_{null}}$
- R_L^2 - larger value indicates good fit.

- Deviance measure : $D_{fitted} = -2\ln(L(\hat{\boldsymbol{\beta}}_{mle}|\mathbf{y}, \mathbf{X}))$.
- Null model means only intercept term - no regressors.
- $D_{null} = -2\ln(L(\hat{\beta}_{0mle}|\mathbf{y}, \mathbf{X}))$.
- $D_{null} - D_{fitted} \geq 0$.
- If $D_{null} - D_{fitted}$ very large then we can reject the hypothesis of no regression.
- H_0 : all coefficients except β_0 is 0 vs H_1 : not H_0 (test of regression is needed or not/no regressors).
- This test is analogue of F-test in linear regression models.
- Can construct a pseudo-R squared based on Deviance : $R_L^2 = \frac{D_{null} - D_{fitted}}{D_{null}}$
- R_L^2 - larger value indicates good fit.

- Deviance measure : $D_{fitted} = -2\ln(L(\hat{\boldsymbol{\beta}}_{mle}|\mathbf{y},\mathbf{X}))$.
- Null model means only intercept term - no regressors.
- $D_{null} = -2\ln(L(\hat{\beta}_{0mle}|\mathbf{y},\mathbf{X}))$.
- $D_{null} - D_{fitted} \geq 0$.
- If $D_{null} - D_{fitted}$ very large then we can reject the hypothesis of no regression.
- H_0 : all coefficients except β_0 is 0 vs H_1 : not H_0 (test of regression is needed or not/no regressors).
- This test is analogue of F-test in linear regression models.
- Can construct a pseudo-R squared based on Deviance : $R_L^2 = \frac{D_{null} - D_{fitted}}{D_{null}}$
- R_L^2 - larger value indicates good fit.

- Deviance measure : $D_{fitted} = -2\ln(L(\hat{\boldsymbol{\beta}}_{mle}|\mathbf{y}, \mathbf{X}))$.
- Null model means only intercept term - no regressors.
- $D_{null} = -2\ln(L(\hat{\beta}_{0mle}|\mathbf{y}, \mathbf{X}))$.
- $D_{null} - D_{fitted} \geq 0$.
- If $D_{null} - D_{fitted}$ very large then we can reject the hypothesis of no regression.
- H_0 : all coefficients except β_0 is 0 vs H_1 : not H_0 (test of regression is needed or not/no regressors).
- This test is analogue of F-test in linear regression models.
- Can construct a pseudo-R squared based on Deviance : $R_L^2 = \frac{D_{null} - D_{fitted}}{D_{null}}$
- R_L^2 - larger value indicates good fit.

- Deviance measure : $D_{fitted} = -2\ln(L(\hat{\boldsymbol{\beta}}_{mle}|\mathbf{y},\mathbf{X}))$.
- Null model means only intercept term - no regressors.
- $D_{null} = -2\ln(L(\hat{\beta}_{0mle}|\mathbf{y},\mathbf{X}))$.
- $D_{null} - D_{fitted} \geq 0$.
- If $D_{null} - D_{fitted}$ very large then we can reject the hypothesis of no regression.
- H_0 : all coefficients except β_0 is 0 vs H_1 : not H_0 (test of regression is needed or not/no regressors).
- This test is analogue of F-test in linear regression models.
- Can construct a pseudo-R squared based on Deviance : $R_L^2 = \frac{D_{null} - D_{fitted}}{D_{null}}$
- R_L^2 - larger value indicates good fit.

- Deviance measure : $D_{fitted} = -2\ln(L(\hat{\boldsymbol{\beta}}_{mle}|\mathbf{y}, \mathbf{X}))$.
- Null model means only intercept term - no regressors.
- $D_{null} = -2\ln(L(\hat{\beta}_{0mle}|\mathbf{y}, \mathbf{X}))$.
- $D_{null} - D_{fitted} \geq 0$.
- If $D_{null} - D_{fitted}$ very large then we can reject the hypothesis of no regression.
- H_0 : all coefficients except β_0 is 0 vs H_1 : not H_0 (test of regression is needed or not/no regressors).
- This test is analogue of F-test in linear regression models.
- Can construct a pseudo-R squared based on Deviance : $R_L^2 = \frac{D_{null} - D_{fitted}}{D_{null}}$
- R_L^2 - larger value indicates good fit.

- Deviance measure : $D_{fitted} = -2\ln(L(\hat{\boldsymbol{\beta}}_{mle}|\mathbf{y}, \mathbf{X}))$.
- Null model means only intercept term - no regressors.
- $D_{null} = -2\ln(L(\hat{\beta}_{0mle}|\mathbf{y}, \mathbf{X}))$.
- $D_{null} - D_{fitted} \geq 0$.
- If $D_{null} - D_{fitted}$ very large then we can reject the hypothesis of no regression.
- H_0 : all coefficients except β_0 is 0 vs H_1 : not H_0 (test of regression is needed or not/no regressors).
- This test is analogue of F-test in linear regression models.
- Can construct a pseudo-R squared based on Deviance : $R_L^2 = \frac{D_{null} - D_{fitted}}{D_{null}}$
- R_L^2 - larger value indicates good fit.

- Deviance measure : $D_{fitted} = -2\ln(L(\hat{\boldsymbol{\beta}}_{mle}|\mathbf{y},\mathbf{X}))$.
- Null model means only intercept term - no regressors.
- $D_{null} = -2\ln(L(\hat{\beta}_{0mle}|\mathbf{y},\mathbf{X}))$.
- $D_{null} - D_{fitted} \geq 0$.
- If $D_{null} - D_{fitted}$ very large then we can reject the hypothesis of no regression.
- H_0 : all coefficients except β_0 is 0 vs H_1 : not H_0 (test of regression is needed or not/no regressors).
- This test is analogue of F-test in linear regression models.
- Can construct a pseudo-R squared based on Deviance : $R_L^2 = \frac{D_{null} - D_{fitted}}{D_{null}}$
- R_L^2 - larger value indicates good fit.

- **Multiple logistic regression** :

-

$$[Y_1 = y_1, \cdots, Y_n = y_n | \mathbf{X} = \mathbf{x}] \sim \prod_{i=1}^{n} [\frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{y_i} [\frac{1}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{1-y_i}$$

- Model parameters - $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$.
- Everything is same as simple logistic regression.
- One additional issue - multicollinearity or aliasing.
- multicollinearity : some of the regressors/predictors are linearly highly correlated.
- multicollinearity makes some estimates very unreliable!

- **Multiple logistic regression** :
-

$$[Y_1 = y_1, \cdots, Y_n = y_n | \mathbf{X} = \mathbf{x}] \sim$$

$$\prod_{i=1}^{n} [\frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}{1 + e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{y_i} [\frac{1}{1 + e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{1-y_i}$$

- Model parameters - $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$.
- Everything is same as simple logistic regression.
- One additional issue - multicollinearity or aliasing.
- multicollinearity : some of the regressors/predictors are linearly highly correlated.
- multicollinearity makes some estimates very unreliable!

- **Multiple logistic regression** :
- $$[Y_1 = y_1, \cdots, Y_n = y_n | \mathbf{X} = \mathbf{x}] \sim$$
$$\prod_{i=1}^{n} [\frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}{1 + e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{y_i} [\frac{1}{1 + e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{1-y_i}$$
- Model parameters - $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$.
- Everything is same as simple logistic regression.
- One additional issue - multicollinearity or aliasing.
- multicollinearity : some of the regressors/predictors are linearly highly correlated.
- multicollinearity makes some estimates very unreliable!

- **Multiple logistic regression** :
- $$[Y_1 = y_1, \cdots, Y_n = y_n | \mathbf{X} = \mathbf{x}] \sim$$ $$\prod_{i=1}^{n} [\frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}}}]^{y_i} [\frac{1}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}}}]^{1 - y_i}$$
- Model parameters - $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$.
- Everything is same as simple logistic regression.
- One additional issue - multicollinearity or aliasing.
- multicollinearity : some of the regressors/predictors are linearly highly correlated.
- multicollinearity makes some estimates very unreliable!

- **Multiple logistic regression** :
-

$$[Y_1 = y_1, \cdots, Y_n = y_n | \mathbf{X} = \mathbf{x}] \sim$$
$$\prod_{i=1}^{n} [\frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{y_i} [\frac{1}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{1-y_i}$$

- Model parameters - $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$.
- Everything is same as simple logistic regression.
- One additional issue - multicollinearity or aliasing.
- multicollinearity : some of the regressors/predictors are linearly highly correlated.
- multicollinearity makes some estimates very unreliable!

- **Multiple logistic regression** :
-

$$[Y_1 = y_1, \cdots, Y_n = y_n | \mathbf{X} = \mathbf{x}] \sim$$
$$\prod_{i=1}^{n} [\frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{y_i} [\frac{1}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{1-y_i}$$

- Model parameters - $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$.
- Everything is same as simple logistic regression.
- One additional issue - multicollinearity or aliasing.
- multicollinearity : some of the regressors/predictors are linearly highly correlated.
- multicollinearity makes some estimates very unreliable!

- **Multiple logistic regression** :
- $$[Y_1 = y_1, \cdots, Y_n = y_n | \mathbf{X} = \mathbf{x}] \sim \prod_{i=1}^{n} [\frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{y_i} [\frac{1}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{1-y_i}$$
- Model parameters - $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$.
- Everything is same as simple logistic regression.
- One additional issue - multicollinearity or aliasing.
- multicollinearity : some of the regressors/predictors are linearly highly correlated.
- multicollinearity makes some estimates very unreliable!

- **Multiple logistic regression** :

-

$$[Y_1 = y_1, \cdots, Y_n = y_n | \mathbf{X} = \mathbf{x}] \sim \prod_{i=1}^{n} [\frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}{1 + e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{y_i} [\frac{1}{1 + e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}]^{1-y_i}$$

- Model parameters - $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$.
- Everything is same as simple logistic regression.
- One additional issue - multicollinearity or aliasing.
- multicollinearity : some of the regressors/predictors are linearly highly correlated.
- multicollinearity makes some estimates very unreliable!

- Calculate variance inflation factors VIF_j for each of the p regressors.
- Perform a multiple linear regression of the j th covariate on the remaining $(p-1)$ covariates - calculate the R_j^2 (R-squared).
- $VIF_j = \frac{1}{1-R_j^2}$
- High VIF means highly correlated covariate - $VIF_j > 5$ is high (thumb rule).
- Unlike linear regression there are different notions of residuals - Deviance residual, Pearson residual and Anscombe residual.
- Similar diagnostic plots based on them can be devised like linear regression problems.

- Calculate variance inflation factors VIF_j for each of the p regressors.
- Perform a multiple linear regression of the j th covariate on the remaining $(p-1)$ covariates - calculate the R_j^2 (R-squared).
- $VIF_j = \frac{1}{1-R_j^2}$
- High VIF means highly correlated covariate - $VIF_j > 5$ is high (thumb rule).
- Unlike linear regression there are different notions of residuals - Deviance residual, Pearson residual and Anscombe residual.
- Similar diagnostic plots based on them can be devised like linear regression problems.

- Calculate variance inflation factors VIF_j for each of the p regressors.
- Perform a multiple linear regression of the j th covariate on the remaining $(p-1)$ covariates - calculate the R_j^2 (R-squared).
- $VIF_j = \frac{1}{1-R_j^2}$
- High VIF means highly correlated covariate - $VIF_j > 5$ is high (thumb rule).
- Unlike linear regression there are different notions of residuals - Deviance residual, Pearson residual and Anscombe residual.
- Similar diagnostic plots based on them can be devised like linear regression problems.

- Calculate variance inflation factors VIF_j for each of the p regressors.
- Perform a multiple linear regression of the j th covariate on the remaining $(p-1)$ covariates - calculate the R_j^2 (R-squared).
- $VIF_j = \frac{1}{1-R_j^2}$
- High VIF means highly correlated covariate - $VIF_j > 5$ is high (thumb rule).
- Unlike linear regression there are different notions of residuals - Deviance residual, Pearson residual and Anscombe residual.
- Similar diagnostic plots based on them can be devised like linear regression problems.

- Calculate variance inflation factors VIF_j for each of the p regressors.
- Perform a multiple linear regression of the j th covariate on the remaining $(p-1)$ covariates - calculate the R_j^2 (R-squared).
- $VIF_j = \frac{1}{1-R_j^2}$
- High VIF means highly correlated covariate - $VIF_j > 5$ is high (thumb rule).
- Unlike linear regression there are different notions of residuals - Deviance residual, Pearson residual and Anscombe residual.
- Similar diagnostic plots based on them can be devised like linear regression problems.

- Calculate variance inflation factors VIF_j for each of the p regressors.
- Perform a multiple linear regression of the j th covariate on the remaining $(p-1)$ covariates - calculate the R_j^2 (R-squared).
- $VIF_j = \frac{1}{1-R_j^2}$
- High VIF means highly correlated covariate - $VIF_j > 5$ is high (thumb rule).
- Unlike linear regression there are different notions of residuals - Deviance residual, Pearson residual and Anscombe residual.
- Similar diagnostic plots based on them can be devised like linear regression problems.