Statistics for Engineers Lecture 9
Linear Regression

Chong Ma

Department of Statistics
University of South Carolina
chongm@email.sc.edu

April 17, 2017

Outline

1. Introduction to regression
2. Simple linear regression model
3. Least squares estimation
4. Model assumptions and sampling distribution
5. Estimating the error variance
6. Statistical inference for β_0 and β_1
7. Confidence and Prediction Intervals

Introduction to regression

A problem arising in engineering, economics, medicine, and other areas is that of investigating the relationship between two or more variables. In such settings, the goal is to model a random variable Y (often continuous) as a function of one or more independent variables, say, $x_1, x_2, \ldots, x_k$. Mathematically, we can express this model as

$$Y = g(x_1, x_2, \ldots, x_k) + \varepsilon$$

where $g : \mathbb{R}^k \to \mathbb{R}$ is a function (whose form may or may not be specified). This is called a **regression model**. In this course, we will consider models of the form

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}_{g(x_1, x_2, \ldots, x_k)} + \varepsilon$$

That is, g is a linear function of $\beta_0, \beta_1, \ldots, \beta_k$. We call this a **linear regression model**.

Terminology:

- The **response variable** Y is random (but we do get to observe its value).
- The **independent variable** $x_1, x_2, \ldots, x_k$ are fixed (and observed).
- The **response parameters** $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are unknown. They are to be estimated based on the observed data.
- The **error term** ε is random (not observed). The presence of the **random error** ε conveys the fact that the relationship between the dependent variable Y and the independent variables $x_1, x_2, \ldots, x_k$ through g is not deterministic. Instead, the term ε "absorbs" all variation in Y that is not explained by $g(x_1, x_2, \ldots, x_k)$.

Remark: The term "linear" does not refer to the shape of the regression function g. It refers to how the regression parameters $\beta_0, \beta_1, \ldots, \beta_k$ enter the g function.

Outline

1. Introduction to regression
2. Simple linear regression model
3. Least squares estimation
4. Model assumptions and sampling distribution
5. Estimating the error variance
6. Statistical inference for β_0 and β_1
7. Confidence and Prediction Intervals

Simple linear regression model

A **simple linear regression model** includes only one independent variable x and is of the form

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

The population regression function $g(x) = \beta_0 + \beta_1 x$ is a straight line with intercept β_0 and slope β_1. These parameters describe the population of individuals for which this model is assumed. Note if $E(\varepsilon) = 0$, then

$$E(Y) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x + E(\varepsilon) = \beta_0 + \beta_1 x$$

Therefore, the interpretations for β_0 and β_1 are as follows.

- β_0 quantifies the population mean of Y when $x = 0$.
- β_1 quantifies the population-level change in $E(Y)$ brought about by one-unit change in x.

Simple linear regression model

Example As part of a waste removal project, a new compression machine for processing sewage sludge is being studied. Engineers are interested in the following variables:

$$Y = \text{ moisture control of compressed pellets (measured as a percent)}$$
$$x = \text{ machine filtration rate (kg-DS/m/hr)}$$

Engineers collect observations of (x, Y) from a random sample of $n = 20$ sewage specimens; the data are given below.

Obs	x	Y	Obs	x	Y
1	125.3	77.9	11	159.5	79.9
2	98.2	76.8	12	145.8	79.0
⋮	⋮	⋮	⋮	⋮	⋮
9	161.2	80.1	19	159.6	79.0
10	178.9	80.2	20	110.7	78.6

Simple linear regression model

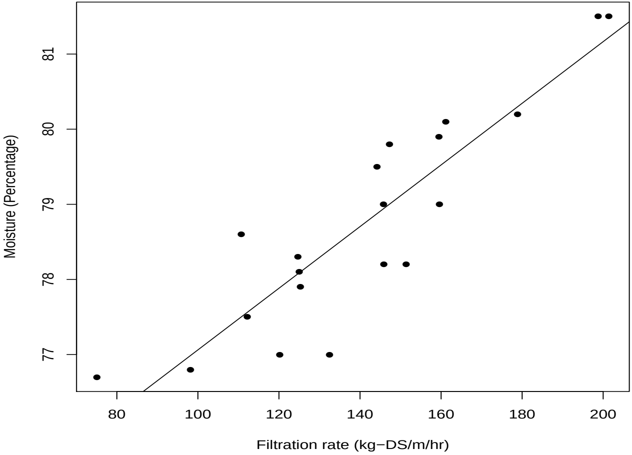

Figure 1. Scatterplot of pellet moisture Y (measured as a percentage) as a function of machine filtration rate x (measured in kg-DS/m/hr).

Simple linear regression model

Figure 1 displays the sample data in a **scatterplot**. This sample information suggests the variables Y and x are **linearly related**, although there is a large amount of variation that is unexplained.

- This unexplained variability could arise from other independent variables (e.g., applied temperature, pressure, sludge mass, etc.) that also influence the moisture percentage Y but are not present in the model.
- It could also arise from measurement error or just random variation in the sludge compression process.

Inference: What does the sample information suggest about the population? Do we have evidence that Y and x are linearly related in the population?

Outline

1. Introduction to regression
2. Simple linear regression model
3. Least squares estimation
4. Model assumptions and sampling distribution
5. Estimating the error variance
6. Statistical inference for β_0 and β_1
7. Confidence and Prediction Intervals

Least sqaures estimation

Fitting a regression model refers to estimating the population regression parameters in the model with the observed sample information(data). In the simple linear regression context, suppose we have a random sample of observations $(x_i, Y_i), i = 1, 2, \ldots, n$ and postulate the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i, i = 1, 2, \ldots, n$$

Our goal is to estimate β_0 and β_1. The most common method of estimating the population parameters β_0 and β_1 is the **method of least squares**. The **least squares method** is to find the optimal values of β_0 and β_1 such that minimizes

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i))^2$$

Least sqaures estimation

Denote the least squares estimators by b_0 and b_1, respectively, that is, the values of β_0 and β_1 that minimizes $Q(\beta_0, \beta_1)$. A two-variable calculus minimization argument can be used to find minimizers of $Q(\beta_0, \beta_1)$. Taking partial derivatives of $Q(\beta_0, \beta_1)$, we obtain

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i) \overset{\text{set}}{=} 0$$

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)x_i \overset{\text{set}}{=} 0$$

Solving for β_0 and β_1 gives the **least squares estimators**

$$b_0 = \bar{Y} - b_1\bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}}$$

The estimated model is written as $\hat{Y} = b_0 + b_1 x$.

Least sqaures estimation

We use R to calculate the equation of the least squares regression line for the sewage data.
The least squares estimates for the sewage data are

$$b_0 = 72.959, b_1 = 0.041$$

Therefore, the estimated model is

$$\hat{Y} = 72.959 + 0.041x$$

or, in other words,

$$\widehat{\text{moisture}} = 72.959 + 0.041\text{Filtrationrate}$$

Least sqaures estimation

Remarks: The estimated model is also called the **prediction equation**, because we can now predict the value of Y (moisture percentage) for a given value of x (filtration rate). For example, when the filtration rate is $x = 150$ (kg-DS/m/hr), we would predict the moisture percentage to be

$$\hat{Y}(150) = 72.959 + 0.041(150) \approx 79.11$$

Of course, the prediction comes directly from the sample of observations used to fit the regression model. Therefore, we will eventually want to quantify the **uncertainty** in this prediction, i.e., how variable is this prediction?

Outline

1. Introduction to regression
2. Simple linear regression model
3. Least squares estimation
4. Model assumptions and sampling distribution
5. Estimating the error variance
6. Statistical inference for β_0 and β_1
7. Confidence and Prediction Intervals

Model assumptions and sampling distribution

Interest: We investigate the properties of the least squares estimators b_0 and b_1 as estimators of the population-level regression parameters β_0 and β_1 in the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \ldots, n$$

Assumption: $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$. **Results:** Under the above assumption, we can derive the following results for the simple linear model.

- **Result 1:** $Y \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$ In other words, the response variable Y is normally distributed with mean $\beta_0 + \beta_1 x$ and variance σ^2.
- **Result 2:** The least squares estimators b_0 and b_1 are unbiased estimators of β_0 and β_1, respectively, that is

$$E(b_0) = \beta_0, E(b_1) = \beta_1$$

Model assumptions and sampling distribution

- **Result 3:** The least squares estimators b_0 and b_1 have normal sampling distributions; specially,

$$b_0 \sim \mathcal{N}(\beta_0, c_{00}\sigma^2) \text{ and } b_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2)$$

where

$$c_{00} = \frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \text{ and } c_{11} = \frac{1}{SS_{xx}}$$

These distributions are needed to construct confidence intervals and perform hypothesis tests for β_0 and β_1.

Outline

1. Introduction to regression
2. Simple linear regression model
3. Least squares estimation
4. Model assumptions and sampling distribution
5. Estimating the error variance
6. Statistical inference for β_0 and β_1
7. Confidence and Prediction Intervals

Estimating the error variance

In the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, we now turn our attention to estimating σ^2, the **error variance**. **Recall:** As we did in estimating β_0 and β_1 (the population level regression parameters), we will use the observed data $(x_i, Y_i), i = 1, 2, \ldots, n$ to estimate the error variance σ^2. The error variance is also a population level parameter and quantifies how variable the population is for a given model.
Terminology: Define the ith **fitted value** by

$$\hat{Y}_i = b_0 + b_1 x_i$$

where b_0 and b_1 are the least squares estimators.

Estimating the error variance

Each observation has its own fitted value. Defie the ith **residual** by

$$e_i = Y_i - \hat{Y}_i$$

In the simple linear regression model, we have the following fact

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = 0$$

Note $b_0 = \bar{Y} - b_1 \bar{x}$, then

$$\begin{aligned} \sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 x_i)) \\ &= \sum_{i=1}^{n} Y_i - n(b_0 + b_1 \bar{x}) = n\bar{Y} - n\bar{Y} \quad (\bar{Y} = b_0 + b_1 \bar{x}) \\ &= 0 \end{aligned}$$

Estimating the error variance

Define the **residual sum of squares** by

$$\mathrm{SS}_{res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

In the simple linear regression model, the **residual mean squares**

$$\mathrm{MS}_{res} = \frac{\mathrm{SS}_{res}}{n-2}$$

is an unbiased estimator of σ^2, that is,

$$E(\mathrm{MS}_{res}) = \sigma^2$$

The quantity

$$\hat{\sigma} = \sqrt{\mathrm{MS}_{res}} = \sqrt{\frac{\mathrm{SS}_{res}}{n-2}}$$

estimates σ and is called the **residual standard error**.

Estimating the error variance

```
> summary(fit)
Call:
lm(formula = moisture ~ filtration.rate)
Residuals:
     Min       1Q   Median       3Q      Max
-1.39552 -0.27694  0.03548  0.42913  1.09901

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     72.958547   0.697528 104.596  < 2e-16 ***
filtration.rate  0.041034   0.004837   8.484 1.05e-07 ***
---
Residual standard error: 0.6653 on 18 degrees of freedom
Multiple R-squared:  0.7999,Adjusted R-squared:  0.7888
F-statistic: 71.97 on 1 and 18 DF,  p-value: 1.052e-07
```

Outline

1. Introduction to regression
2. Simple linear regression model
3. Least squares estimation
4. Model assumptions and sampling distribution
5. Estimating the error variance
6. Statistical inference for β_0 and β_1
7. Confidence and Prediction Intervals

Statistical inference for β_0 and β_1

In the simple linear regression

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

we are dealing with the question, **"What does the sample information from an estimated regression model suggest about the population?"** Put another way, we pursue **statistical inference** for the population level regression parameters β_0 and β_1. In practice,

- Inference for the slope parameter β_1 is of primary interest because of its connection to the independent variable x in the model. For example, if $\beta_1 = 0$, then Y and x are not linearly related in the population.
- Statistical inference for β_0 is less meaningful, unless one is explicitly interested in the mean of Y when $x = 0$. We will not pursue this.

Statistical inference for β_0 and β_1

Under the regression model assumptions, the following sampling distribution arises:

$$t = \frac{b_1 - \beta_1}{\sqrt{\frac{\mathrm{MS}_{res}}{\mathrm{SS}_{xx}}}} \sim t(n-2)$$

- **Confidence Interval:** the $100(1-\alpha)$ percent confidence interval

$$[b_1 \pm t_{n-2,\alpha/2}\sqrt{\frac{\mathrm{MS}_{res}}{\mathrm{SS}_{xx}}}]$$

- **Hypothesis test:** $\mathrm{H}_0 : \beta_1 = 0$ v.s. $\mathrm{H}_1 : \beta_1 \neq 0$

$$\text{p-value} = P(|T| > |t|) = 2P(T > |t|)$$

If p-value$< \alpha$, we reject H_0; otherwise, do not reject H_0

Statistical inference for β_0 and β_1

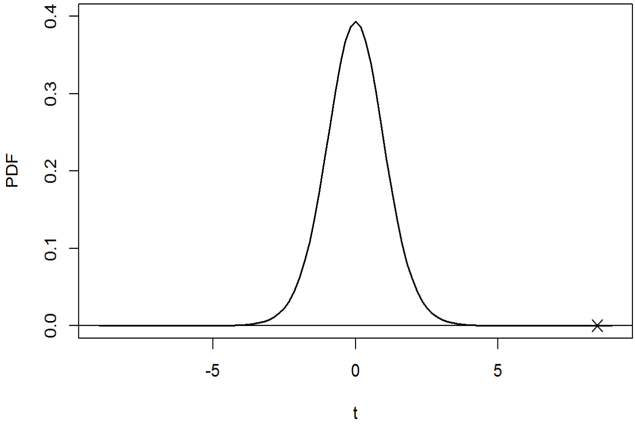

Figure 2. Sewage data: t_{18} pdf, which is the sampling distribution of t when $\mathrm{H}_0 : \beta_1 = 0$ is true. The "$\times$" represents $t = 8.484$.

Statistical inference for β_0 and β_1

- **Confidence interval**

```
> confint(fit,level=0.95)
                       2.5 %      97.5 %
(Intercept)      71.49309400 74.42399995
filtration.rate   0.03087207  0.05119547
```

  Interpretation: We are 95% confident that the population parameter β_1 is between 0.0309 and 0.0511. Further, it means **for every one unit increase in the machine filtration rate x, we are 95% confident that the population mean absorption $E(Y)$ will increase between 0.0309 and 0.0511 percent.**

- **Hypothesis test:** use **summary** function in R to perform the hypothesis test. Since p-value $< 2 \times 10^{-16}$, reject H_0. We have sufficient evidence to conclude that β_1 is not equal 0.

Outline

1. Introduction to regression
2. Simple linear regression model
3. Least squares estimation
4. Model assumptions and sampling distribution
5. Estimating the error variance
6. Statistical inference for β_0 and β_1
7. Confidence and Prediction Intervals

Confidence and prediction intervals

We are often interested in learning about the response Y at a certain setting of the independent variable, say $x = x_0$. For the sewage data, for example, suppose we are interested in the moisture percentage Y when the filtration rate is $x = 150$ kg-DS/m/hr. Two potential goals arise:

- Be interested in **estimating the population mean** of Y when $x = x_0$, that is $E(Y|x_0) = \beta_0 + \beta_1 x_0$.
- Be interested in **predicting a new response Y** at $x = x_0$, that is $Y^*(x_0) = \beta_0 + \beta_1 x_0 + \varepsilon$.

Goals: We would like to create $100(1-\alpha)$ percent intervals for the population mean $E(Y|x_0)$ and for the new response $Y^*(x_0)$. The former is called a **confidence interval** and the latter is called a **prediction interval**.
Point Estimator: the same for $E(Y|x_0)$ and $Y^*(x_0)$, which is denoted by

$$\hat{Y}(x_0) = b_0 + b_1 x_0$$

Confidence and prediction intervals

Confidence Interval: A $100(1-\alpha)$ percent confidence interval for the population mean $E(Y|x_0)$ is given by

$$\hat{Y}(x_0) \pm t_{n-2,\alpha/2}\sqrt{\mathrm{MS}_{res}\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\mathrm{SS}_{xx}}\right]}$$

Prediction Interval: A $100(1-\alpha)$ percent confidence interval for the population mean $Y^*(x_0)$ is given by

$$\hat{Y}(x_0) \pm t_{n-2,\alpha/2}\sqrt{\mathrm{MS}_{res}\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\mathrm{SS}_{xx}}\right]}$$

Confidence and prediction intervals

- **Comparison:** The two intervals have the same form and are nearly identical.
  - The extra "1" in the prediction interval's standard error arises from the additional uncertainty associated with ε.
  - The prediction interval is always wider than the according confidence interval, provided x_0 and α are fixed.
- **Interval length:** The length of both intervals depends on the value of x_0.
  - The standard error in either interval will be smallest when $x_0 = \bar{x}$ and will get larger the further x_0 is from $\bar{x}$ in either direction.
  - This makes intuitive sense, namely, we would expect to have the most "confidence" in the fitted model near the "center" of the observed data.
- **Warning:** Sometimes estimate $E(Y|x_0)$/predict $\bar{Y}^*(x_0)$ for values of x_0 outside the range of x values in the study. This is called **extrapolation** and can be very dangerous.

Confidence and prediction intervals

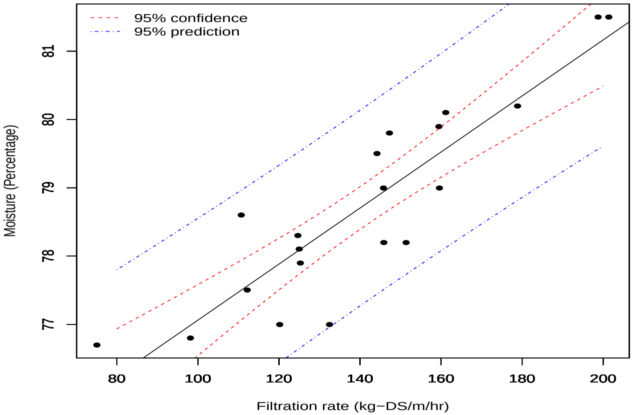

Figure 3. Scatterplot of pellet moisture Y (measured as a percentage) as a function of machine filtration rate x (measured in kg-DS/m/hr). The least squares regression line is added. 95% confidence/prediction bands are added.

Confidence and prediction intervals

- A 95% confidence interval for $E(Y|x_0 = 150)$ is $(78.79, 79.44)$. When the filtration rate is $x_0 = 150$ kg-DS/m/hr, we are 95% confident that **the population mean moisture percentage** is between 78.79 and 79.44 percent.
- A 95% prediction interval for $Y^*(x_0 = 150)$ is $(77.68, 80.55)$. When the filtration rate is $x_0 = 150$ kg-DS/m/hr, we are 95% confident that **the moisture percentage for a single specimen** is between 78.79 and 79.44 percent.