Chapter 11: SIMPLE LINEAR REGRESSION (SLR) AND CORRELATION

Part 3: Hypothesis tests for β_0 and β_1 Coefficient of Determination, R^2

Sections 11-4 & 11-7.2

- For SLR, a common hypothesis test is the test for a linear relationship between X and Y.

$H_0 : \beta_1 = 0$ (no linear relationship)
$H_1 : \beta_1 \neq 0$

- Under the assumption $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, we have

$$\hat{\beta}_0 \sim N\left(\beta_0,\ \ \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1,\ \ \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

- Test of interest

  $H_0 : \beta_1 = 0$ (no linear relationship)
  $H_1 : \beta_1 \neq 0$

- Since we will be estimating σ^2, we will use a t-statistic:

$$T_0 = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

Under H_0 true, $T_0 \sim t_{n-2}$.

From our observed test statistic t_0, we can compute a p-value and make decison on the hypothesis test.

Example: The chloride concentration data (revisited)

Testing for a linear relationship between chloride concentration (Y) and % of watershed in roadways (X)

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Estimates:

$$\hat{\beta}_1 = 20.567$$

$$se(\hat{\beta}_1) = \sqrt{\frac{13.8092}{3.0106}} = 2.1417$$

Test statistic:

$$t_0 = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{20.567}{2.1417} = 9.603$$

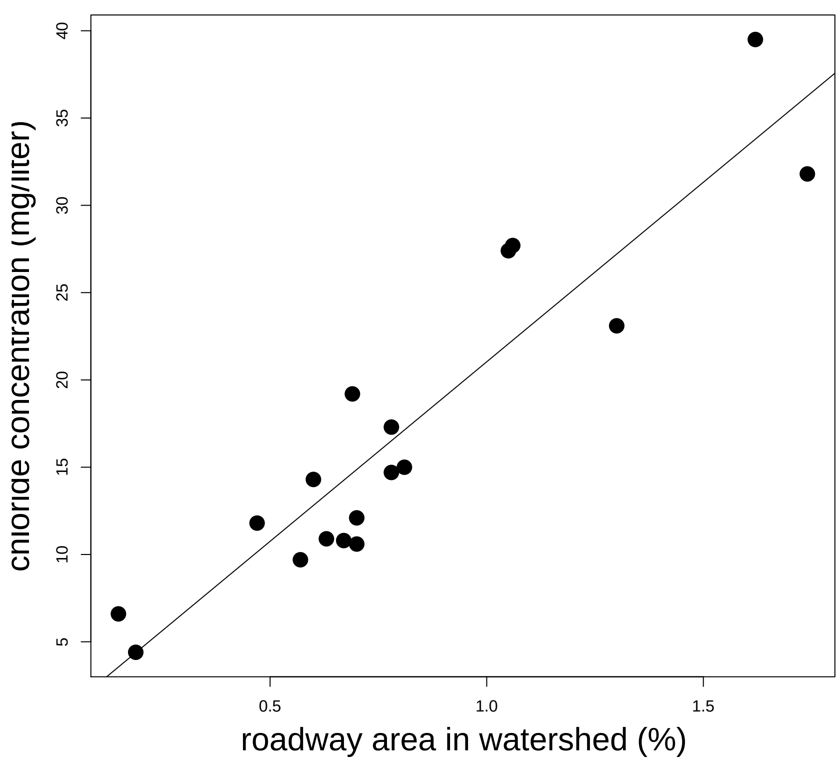
Under H_0 true, $T_0 \sim t_{16}$

P-value:

$$2 \times P(T_0 > 9.603) = 4.81 \times 10^{-8}$$
$$\{\text{very small}\}$$

Reject H_0.

There IS statistically significant evidence that the slope is not 0, so there is evidence of a linear relationship between chloride concentration and % of watershed in roadways.



- Similarly, we can run a hypothesis test that the intercept equals 0...

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

The test statistic:

$$T_0 = \frac{\hat{\beta}_0 - 0}{se(\hat{\beta}_0)} = \frac{\hat{\beta}_0}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)}}$$

Under H_0 true, $T_0 \sim t_{n-2}$.

- **Example**: The chloride concentration data (revisited)

Testing if the intercept is zero.

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

Estimates:

$$\hat{\beta}_0 = 0.4705$$

$$se(\hat{\beta}_0) = \sqrt{13.8092\left(\frac{1}{18} + \frac{0.8061^2}{3.0106}\right)} = 1.9358$$

Test statistic:

$$t_0 = \frac{\hat{\beta}_0 - 0}{se(\hat{\beta}_0)} = \frac{0.4705}{1.9358} = 0.2431$$

Under H_0 true, $T_0 \sim t_{16}$

P-value:
$2 \times P(T_0 > 0.2431) = 0.8110$

Fail to reject H_0. We do not have evidence to suggest the intercept is anything other than zero. (So, a watershed with no roadways essentially has a chloride concentration of 0 mg/liter.)

MINITAB OUTPUT:

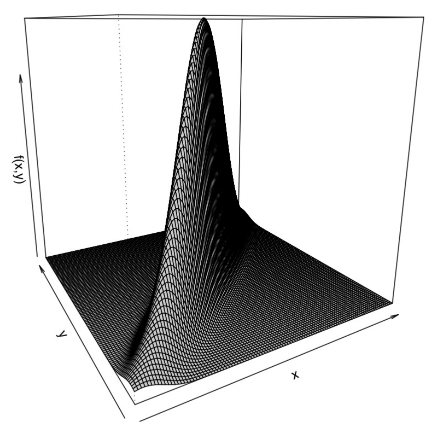
```
Regression Analysis: y versus x


The regression equation is
y = 0.47 + 20.6 x


Predictor     Coef  SE Coef     T      P
Constant     0.470    1.936  0.24  0.811
x           20.567    2.142  9.60  0.000


S = 3.71607
```

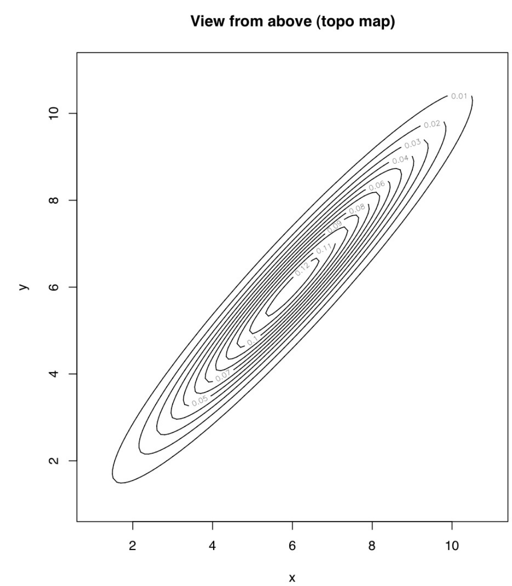
Correlation

Section 11-8

- Earlier we discussed the correlation coefficient between Y and X, denoted as ρ, where
$$\rho = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

- For example, in the bivariate normal:





$\rho = 0.95$

- ρ is a parameter of interest to be estimated from the data.

- The *sample correlation coefficient* r (denoted R in our book) measures the **strength of a linear relationship** in the observed data.

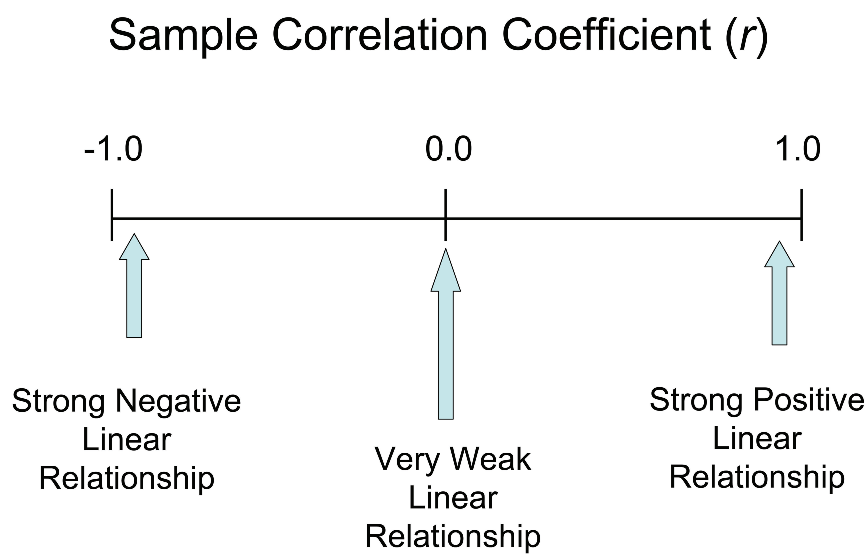
- r has a number of different formulas...

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$= \frac{S_X}{S_Y} \cdot \hat{\beta}_1$$

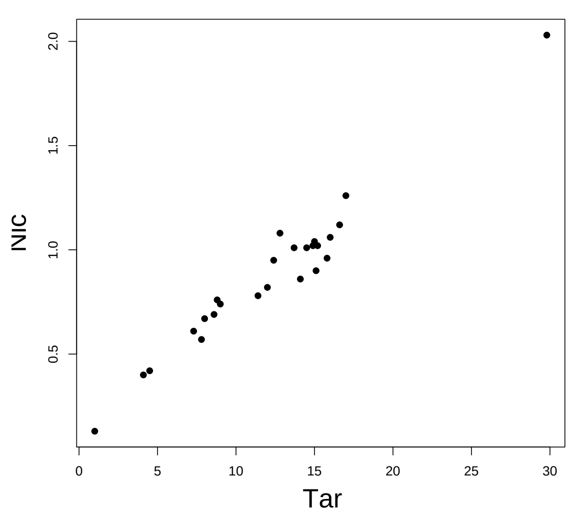
- The sample correlation coefficient r estimates the population correlation coefficient ρ

- Possible values for r:



Correlation Example: Cigarette data

```
> correlation(Tar,Nic)          0.9766076
```



With r near $+1$, this shows a very strong positive linear association.

- r...
  - is a unitless measure, and $-1 \leq r \leq 1$
  - near -1 or +1 shows a strong linear relationship
  - near 0 suggests no relationship
  - a positive r is associated with an estimated positive slope
  - a negative r is associated with an estimated negative slope
  - r is NOT used to measure strength of a curved line
  - In simple linear regression, r^2 is the Coefficient of Determination R^2 discussed next.

Simple Linear Regression

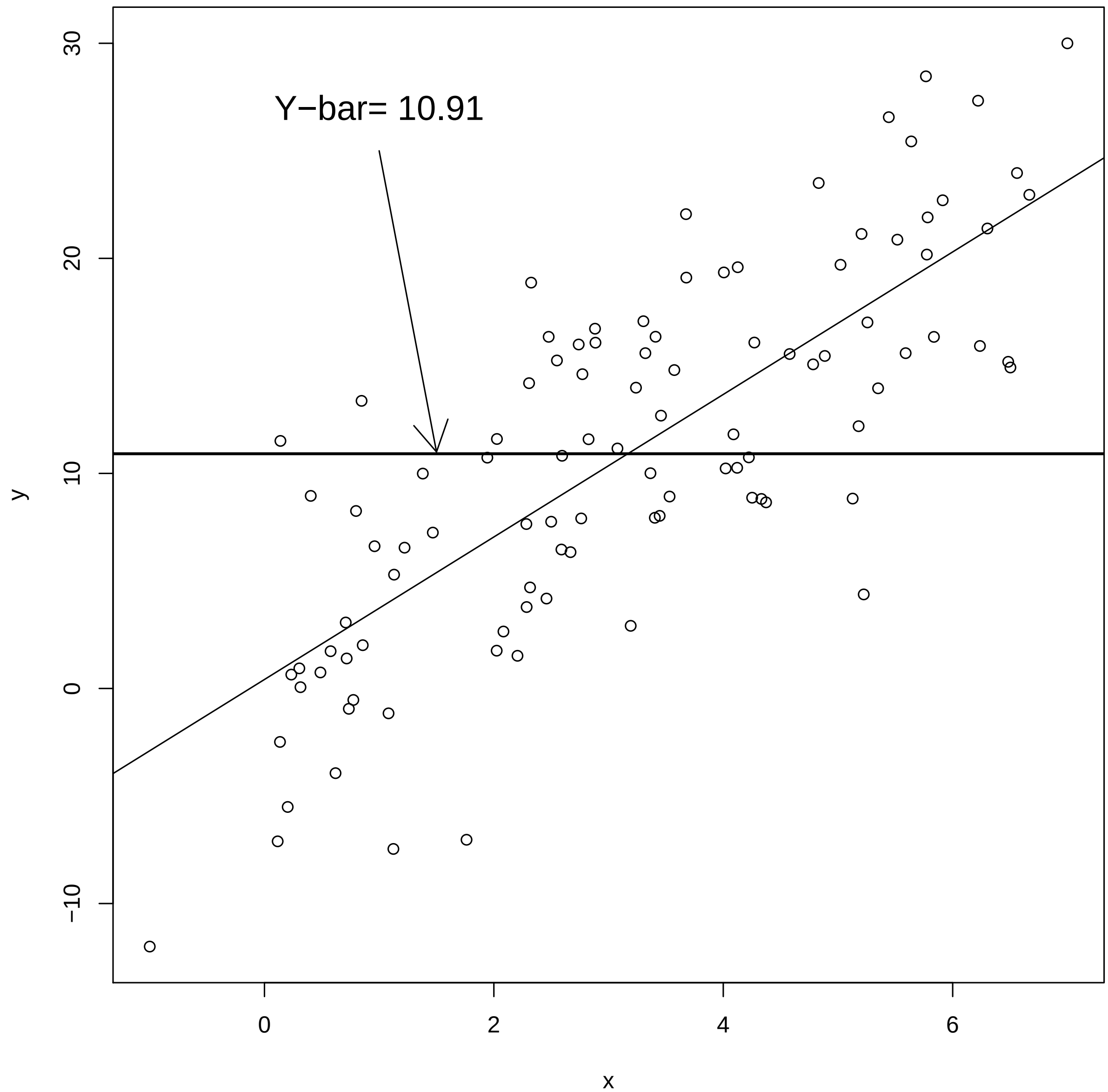
Total corrected sum of squares (SS_T)
Secton 11-4.2

- We use the total corrected sum of squares of Y, or SS_T , to quantify the **total variability in the response**.

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

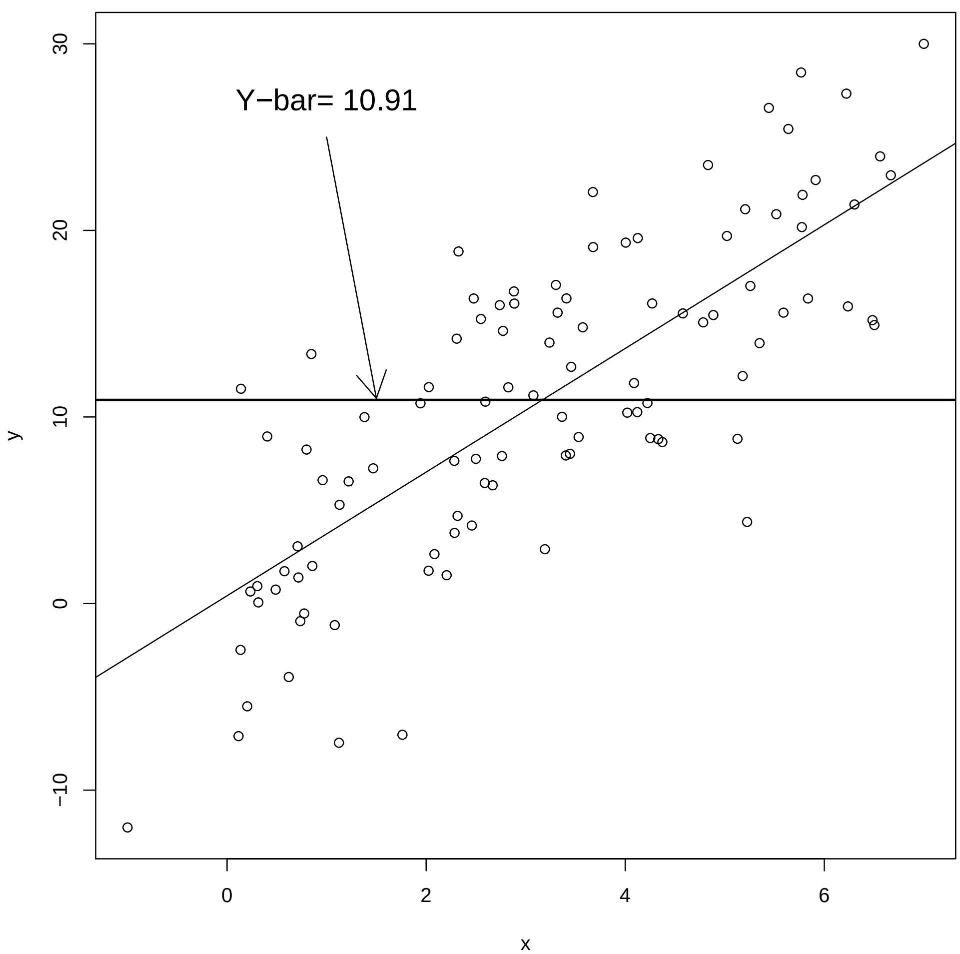
- Total sum of squares quantifies the overall squared distance of the Y-values from the overall mean of the responses $\bar{Y}$

  We can look at this graphically...



- For regression, we can 'decompose' the distance of an observation y_i from the overall mean $\bar{y}$ and write:

$$y_i - \bar{y} \;=\; \underbrace{(y_i - \hat{y}_i)}_{\substack{\text{distance from}\\ \text{observation to}\\ \text{fitted line}}} \;+\; \underbrace{(\hat{y}_i - \bar{y})}_{\substack{\text{distance from}\\ \text{fitted line to}\\ \text{overall mean}}}$$



- Which leads to the equation:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

or

$$SS_T = SS_E + SS_R$$

where SS_R is the *regression sum of squares*

- Total variability has been decomposed into "explained" variability (SS_R) and "unexplained" variability (SS_E)

- In general, when the proportion of total variability that is explained is high, we have a good fitting model

- The proportion of total variability that is explained by the model is called the **Coefficient of Determination** (denoted R^2):

  - $R^2 = \frac{SS_R}{SS_T}$
  - $R^2 = 1 - \frac{SS_E}{SS_T}$
  - $0 \leq R^2 \leq 1$
  - R^2 near 1 suggests a good fit to the data
  - if $R^2 = 1$, ALL points fall *exactly* on the line
  - Different disciplines have different views on what is a *high* R^2, in other words what is a good model...

* social scientists may get excited about an R^2 near 0.30

* a researcher with a designed experiment may want to see an R^2 near 0.80 or higher

NOTE: Coefficient of Determination is discussed in section 11-7.2

Example: The chloride concentration data (revisited)

MINITAB OUTPUT:

```
Regression Analysis: y versus x

The regression equation is
y = 0.47 + 20.6 x

Predictor     Coef  SE Coef     T      P
Constant     0.470    1.936  0.24  0.811
x           20.567    2.142  9.60  0.000

S = 3.71607       R-Sq = 85.22%
```

Coefficient of Determination: $R^2 = \frac{SS_R}{SS_T} = 0.8522$

R^2 interpretation:

85.22% of the total variability in chloride concentration is explained by the model (or by the percentage of roadway area in watershed, since this is the only predictor in the model).