Chapter 11: Simple linear regression
(Ott & Longnecker Sections: 11.1-11.5)

Duzhe Wang

Part 2

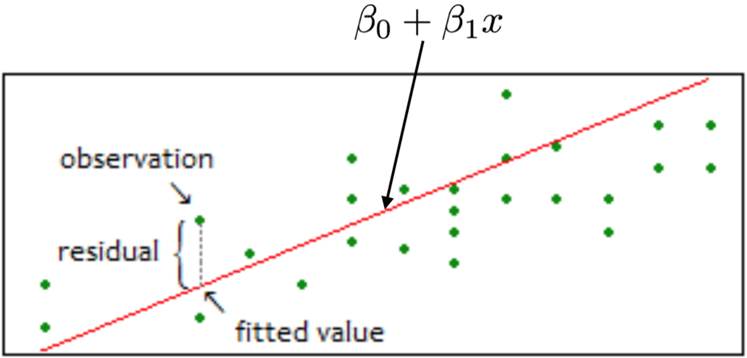
https://dzwang91.github.io/stat371/

UNIVERSITY OF WISCONSIN–MADISON

Review of last lecture

- Denote the height of son i by y_i, the height of father i by x_i, and the random error by ϵ_i, so that the model becomes:
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$
- Our goal is to estimate the values of β_0 and β_1 from data.

Estimation of β_0 and β_1



- We want the vertical distance from the line to the points to be small.

Estimation of β_0 and β_1

- Suppose $\hat{\beta}_0$ and $\hat{\beta}_1$ are our estimates, then the estimated (fitted) value y for any given x is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

- The difference between this and the actual y_i (observed) is $y_i - \hat{y}_i$, we call this the residual.
- For given choices of $\hat{\beta}_0$ and $\hat{\beta}_1$, we can measure how well the line fits by measuring the residual sum of squares (RSS) produced by these choices:

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2.$$

Estimation of β_0 and β_1

- By some calculations,

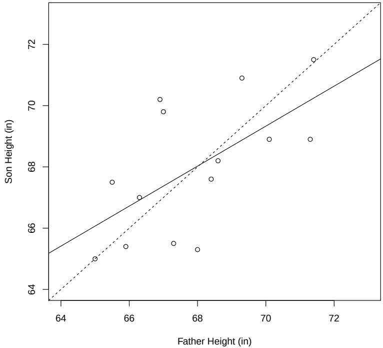
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

- These choices of $\hat{\beta}_0$ and $\hat{\beta}_1$ produce what is called the least squares line.
- The residual sum of squares for the least squares line has a special name: the sum of squared errors or SSE. SSE is the smallest possible residual sum of squares in the universe of all possible lines.

Example

- For the father and son data, these values work out to: $\hat{\beta}_1 = 0.65$ and $\hat{\beta}_0 = 23.64$.
- To assess the quality of the fit, we can add the line (solid) to the scatterplot. We also add a line with slope 1 and intercept zero (dashed) for comparison purposes:



Hypothesis testing in regression model

- In which situation does knowing x tell us anything about y?

Hypothesis testing in regression model

- In which situation does knowing x tell us anything about y? It does if the slope is different than zero.

Hypothesis testing in regression model

- In which situation does knowing x tell us anything about y? It does if the slope is different than zero.
- So, we are often interested in testing:

$$H_0 : \beta_1 = 0 \text{ vs. } H_A : \beta_1 \neq 0.$$

Hypothesis testing in regression model

- In which situation does knowing x tell us anything about y? It does if the slope is different than zero.
- So, we are often interested in testing:

$$H_0 : \beta_1 = 0 \text{ vs. } H_A : \beta_1 \neq 0.$$

- If
  1. The model is correct. (A straight line makes sense for the data.)
  2. The observations are independent.
  3. The variance around the true regression line is constant for all values of x.
  4. The random error around the true line is normal.

Hypothesis testing in regression model

- In which situation does knowing x tell us anything about y? It does if the slope is different than zero.
- So, we are often interested in testing:

$$H_0 : \beta_1 = 0 \text{ vs. } H_A : \beta_1 \neq 0.$$

- If
  1. The model is correct. (A straight line makes sense for the data.)
  2. The observations are independent.
  3. The variance around the true regression line is constant for all values of x.
  4. The random error around the true line is normal.
- Assumptions 2-4 are equivalent to $\epsilon_i \sim iid\ N(0, \sigma^2)$.

Hypothesis testing in regression model continued

- t-test statistic:

$$t = \frac{\hat{\beta}_1}{\widehat{SE(\hat{\beta}_1)}}.$$

where

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

and

$$\hat{\sigma}^2 = \frac{SSE}{n-2}.$$

- If H_0 is true, then t has a T-distribution on $n - 2$ degrees of freedom.

Hypothesis testing in regression model continued

- t-test statistic:

$$t = \frac{\hat{\beta}_1}{\widehat{SE(\hat{\beta}_1)}}.$$

where

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

and

$$\hat{\sigma}^2 = \frac{SSE}{n-2}.$$

- If H_0 is true, then t has a T-distribution on $n-2$ degrees of freedom.
- For the father and son data, $\hat{\sigma} = 1.78$, so $\widehat{SE(\hat{\beta}_1)} = 0.24$, and $t = 2.70$. Comparing this to a t_{12}, the p-value is 0.0193. So we would reject at the 5% level, and conclude that father's height is related to son's height.

Prediction

- Prediction: use our fitted line to guess at the y value that would result from a new x value.

Prediction

- Prediction: use our fitted line to guess at the y value that would result from a new x value.
- Fitted values:

$$(\text{Fitted value for } x = x^*) = \hat{y}|x^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

Prediction

- Prediction: use our fitted line to guess at the y value that would result from a new x value.
- Fitted values:

$$\text{(Fitted value for } x = x^*) = \hat{y}|x^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

- $\mathbb{E}\left(\hat{y}|x^*\right) = \beta_0 + \beta_1 x^*$, so $\hat{y}$ is unbiased.

Prediction

- Prediction: use our fitted line to guess at the y value that would result from a new x value.
- Fitted values:

$$(\text{Fitted value for } x = x^*) = \hat{y}|x^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

- $\mathbb{E}\left(\hat{y}|x^*\right) = \beta_0 + \beta_1 x^*$, so $\hat{y}$ is unbiased.
-

$$SE(\hat{y}|x^*) = \sigma\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Plug in $\hat{\sigma}$ for σ to obtain the estimated SE.

Prediction

- Prediction: use our fitted line to guess at the y value that would result from a new x value.
- Fitted values:

$$(\text{Fitted value for } x = x^*) = \hat{y}|x^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

- $\mathbb{E}(\hat{y}|x^*) = \beta_0 + \beta_1 x^*$, so $\hat{y}$ is unbiased.
- $$SE(\hat{y}|x^*) = \sigma\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

  Plug in $\hat{\sigma}$ for σ to obtain the estimated SE.
- Example: Suppose we want to predict the average son's height when the father is $x^* = 70$ inches tall. Our estimate and standard error would be:

$$\hat{y}|(x^* = 70) = 23.64 + 0.65 * 70 = 69.14$$

$$\text{estimated } SE(\hat{y}|(x^* = 70)) = 1.78\sqrt{1/14 + \tfrac{(70-67.93)^2}{54.25}} = 0.69.$$

R^2

- If we define $SSTot = \sum_{i=1}^{n}(y_i - \bar{y})^2$, we can create a quantity called R^2:

$$R^2 = \frac{SSTot - SSE}{SSTot}.$$

- This can be interpreted as the fraction of the total sum of squares that is explained by the regression line. Often we say that it is the fraction of the total variability explained by the regression line.

R^2

- If we define $SSTot = \sum_{i=1}^{n}(y_i - \bar{y})^2$, we can create a quantity called R^2:

$$R^2 = \frac{SSTot - SSE}{SSTot}.$$

- This can be interpreted as the fraction of the total sum of squares that is explained by the regression line. Often we say that it is the fraction of the total variability explained by the regression line.
- For the father and son data, $R^2 = 0.38$. So we can say that about 38% of the variability in sons' heights can be explained by fathers' heights.