STAT 2593

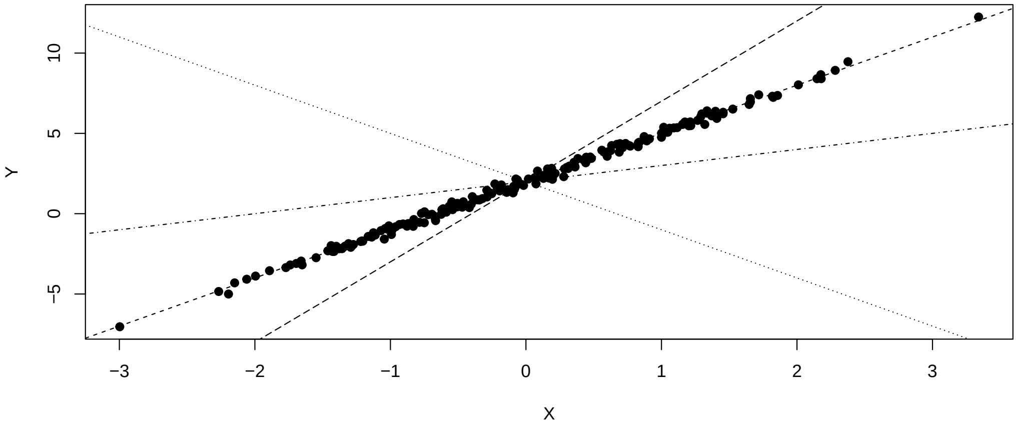
Lecture 039 - Estimating Model Parameters

Dylan Spicker

Estimating Model Parameters

Learning Objectives

1. Understand how the best linear regression line is solved for.
2. Interpret the regression coefficients and understand the limitations.
3. Understand the variance breakdown, what is explained by the model, and what is not.
4. Understand the coefficient of determination and its limitations.



−5
0
5
10
Y
−3
−2
−1
0
1
2
3
X

Estimating Parameters for the Simple Linear Regression

- We want the *best* possible line, as defined by β_0 and β_1.

Estimating Parameters for the Simple Linear Regression

- We want the *best* possible line, as defined by β_0 and β_1.
- We will use estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$, which gives $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

Estimating Parameters for the Simple Linear Regression

- We want the *best* possible line, as defined by β_0 and β_1.
- We will use estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$, which gives $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$.
  - The mistakes we make we call **residuals**, and denote this $e_i = y_i - \widehat{y}_i$.

Estimating Parameters for the Simple Linear Regression

- We want the *best* possible line, as defined by β_0 and β_1.
- We will use estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$, which gives $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$.
  - The mistakes we make we call **residuals**, and denote this $e_i = y_i - \widehat{y}_i$.
  - The idea is to minimize the **squared residuals**, given by $\sum_{i=1}^{n} e_i^2$.

Estimating Parameters for the Simple Linear Regression

- We want the *best* possible line, as defined by β_0 and β_1.
- We will use estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$, which gives $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$.
  - The mistakes we make we call **residuals**, and denote this $e_i = y_i - \widehat{y}_i$.
  - The idea is to minimize the **squared residuals**, given by $\sum_{i=1}^n e_i^2$.
- Doing this results in

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

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

Regression Parameter Interpretation

- Recall that β_0 is the intercept of the regression line.

Regression Parameter Interpretation

- Recall that β_0 is the intercept of the regression line.
  - $\widehat{\beta}_0$ is the *best guess* we have at the intercept, given the data.

Regression Parameter Interpretation

- Recall that β_0 is the intercept of the regression line.
  - $\widehat{\beta}_0$ is the *best guess* we have at the intercept, given the data.
  - If $X = 0$, this is the value that we would except Y to take on.

Regression Parameter Interpretation

- Recall that β_0 is the intercept of the regression line.
  - $\widehat{\beta}_0$ is the *best guess* we have at the intercept, given the data.
  - If $X = 0$, this is the value that we would except Y to take on.
  - Note, $X = 0$ does not always have a substantive meaning.

Regression Parameter Interpretation

- Recall that β_0 is the intercept of the regression line.
  - $\widehat{\beta}_0$ is the *best guess* we have at the intercept, given the data.
  - If $X = 0$, this is the value that we would except Y to take on.
  - Note, $X = 0$ does not always have a substantive meaning.
- Recall that β_1 is the slope of the regression line.

Regression Parameter Interpretation

- Recall that β_0 is the intercept of the regression line.
  - $\widehat{\beta}_0$ is the *best guess* we have at the intercept, given the data.
  - If $X = 0$, this is the value that we would except Y to take on.
  - Note, $X = 0$ does not always have a substantive meaning.
- Recall that β_1 is the slope of the regression line.
  - $\widehat{\beta}_1$ is the *best guess* we have at the slope, given the data.

Regression Parameter Interpretation

- Recall that β_0 is the intercept of the regression line.
  - $\widehat{\beta}_0$ is the *best guess* we have at the intercept, given the data.
  - If $X = 0$, this is the value that we would except Y to take on.
  - Note, $X = 0$ does not always have a substantive meaning.
- Recall that β_1 is the slope of the regression line.
  - $\widehat{\beta}_1$ is the *best guess* we have at the slope, given the data.
  - For a 1 unit increase in X, we expect that Y will change by $\widehat{\beta}_1$.

Regression Parameter Interpretation

- Recall that β_0 is the intercept of the regression line.
  - $\widehat{\beta}_0$ is the *best guess* we have at the intercept, given the data.
  - If $X = 0$, this is the value that we would except Y to take on.
  - Note, $X = 0$ does not always have a substantive meaning.
- Recall that β_1 is the slope of the regression line.
  - $\widehat{\beta}_1$ is the *best guess* we have at the slope, given the data.
  - For a 1 unit increase in X, we expect that Y will change by $\widehat{\beta}_1$.
  - Be careful for *extrapolation* and for *bad model fits*.

Error Sums of Squares

- We call the sum of squared residuals the **error sum of squares**, $\text{SSE} = \sum_{i=1}^{n} e_i^2$.

Error Sums of Squares

- We call the sum of squared residuals the **error sum of squares**, $\text{SSE} = \sum_{i=1}^{n} e_i^2$.
    - This is the variation left unexplained by the model.

Error Sums of Squares

- We call the sum of squared residuals the **error sum of squares**, $\text{SSE} = \sum_{i=1}^{n} e_i^2$.
  - This is the variation left unexplained by the model.
- We can estimate the variance, σ^2, as $\hat{\sigma}^2 = \frac{SSE}{n-2}$.

Error Sums of Squares

- We call the sum of squared residuals the **error sum of squares**, SSE $= \sum_{i=1}^{n} e_i^2$.
  - This is the variation left unexplained by the model.
- We can estimate the variance, σ^2, as $\hat{\sigma}^2 = \frac{SSE}{n-2}$.
- We can also consider the **total variation** in Y, which is given by

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = S_y^2.$$

Error Sums of Squares

- We call the sum of squared residuals the **error sum of squares**, $\text{SSE} = \sum_{i=1}^{n} e_i^2$.
  - This is the variation left unexplained by the model.
- We can estimate the variance, σ^2, as $\hat{\sigma}^2 = \frac{SSE}{n-2}$.
- We can also consider the **total variation** in Y, which is given by

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = S_y^2.$$

  - We will always have that $\text{SSE} \leq \text{SST}$.

Error Sums of Squares

- We call the sum of squared residuals the **error sum of squares**, $\mathrm{SSE} = \sum_{i=1}^{n} e_i^2$.
  - This is the variation left unexplained by the model.
- We can estimate the variance, σ^2, as $\hat{\sigma}^2 = \frac{SSE}{n-2}$.
- We can also consider the **total variation** in Y, which is given by

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = S_y^2.$$

  - We will always have that $\mathrm{SSE} \leq \mathrm{SST}$.
  - The difference, $\mathrm{SST} - \mathrm{SSE}$, is the variation **explained by the model**.

Error Sums of Squares

- We call the sum of squared residuals the **error sum of squares**, $\text{SSE} = \sum_{i=1}^{n} e_i^2$.
  - This is the variation left unexplained by the model.
- We can estimate the variance, σ^2, as $\hat{\sigma}^2 = \frac{SSE}{n-2}$.
- We can also consider the **total variation** in Y, which is given by

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = S_y^2.$$

  - We will always have that $\text{SSE} \leq \text{SST}$.
  - The difference, $\text{SST} - \text{SSE}$, is the variation **explained by the model**.
  - This is called the **regression sums of squares**, denoted SSR.

The Coefficient of Determination

- If we consider the ratio of the variance that is explained by the model we write

$$r^2 = \frac{\text{SSR}}{\text{SST}}.$$

The Coefficient of Determination

- If we consider the ratio of the variance that is explained by the model we write

$$r^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}.$$

    - This is the **r-squared** value, or the **coefficient of determination**.

The Coefficient of Determination

- If we consider the ratio of the variance that is explained by the model we write

$$r^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}.$$

  - This is the **r-squared** value, or the **coefficient of determination**.
  - It gives the proportion of variance in Y which is captured by the model.

The Coefficient of Determination

- If we consider the ratio of the variance that is explained by the model we write

$$r^2 = \frac{\text{SSR}}{\text{SST}}.$$

  - This is the **r-squared** value, or the **coefficient of determination**.
  - It gives the proportion of variance in Y which is captured by the model.
  - It is typically used to indicate the strength of the relationship, with values closer to 1 being preferable.

The Coefficient of Determination

- If we consider the ratio of the variance that is explained by the model we write
$$r^2 = \frac{\text{SSR}}{\text{SST}}.$$
  - This is the **r-squared** value, or the **coefficient of determination**.
  - It gives the proportion of variance in Y which is captured by the model.
  - It is typically used to indicate the strength of the relationship, with values closer to 1 being preferable.
- Note: The coefficient of determination has received a lot of criticism. It is probably best to steer largely clear of it!

Summary

- Linear regression estimates are determined through the **least squares procedure**.
- There is a closed form expression for both the slope and the intercept.
- The intercept gives the value we expect to observe at $X = 0$ and the slope captures the expected change in outcome for a unit change in X.
- The total variance can be decomposed into the error sum of squares and the regression sum of squares.
- The proportion of variance which is explained by the model is called the coefficient of determination.