Least Squares Estimation

DASC 512

Least Squares Estimation

Goal: Be able to use the method of least squares to fit a hypothesized model to fit sample data

We'll be assuming a first-order linear model as the deterministic component for now.

$$y = \beta_0 + \beta_1 x + \epsilon$$

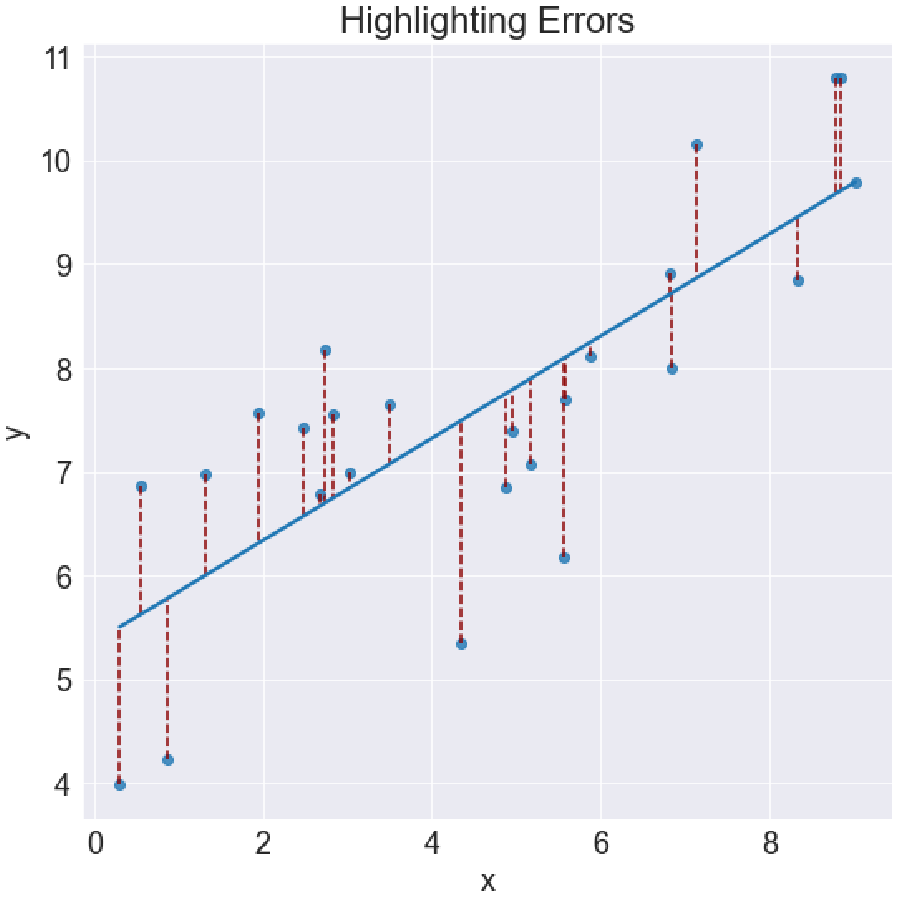
The method of least squares is one method to find values of β_0 and β_1. It minimizes the squared errors in prediction – hence the name!

What is Least Squares?

Least Squares Estimation finds the coefficient values that minimize the sum of squared errors

Thinking back on ANOVA, minimizing SS Error also maximizes F

Why not just minimize sum of error?



Parameter Estimation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Least squares estimates for the slope and intercept are

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Ordinary Least Squares Model

The Ordinary Least Squares (OLS) model is the model produced by the least squares estimates, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

It has the following properties:

- The sum of all errors equals zero $\sum_{i=1}^{n} y_i - \hat{y} = 0 \;\Rightarrow\; \mu_\epsilon = 0$
- The Sum of Squared Errors (SSE) is smaller than for any other straight-line model (i.e., the σ_ϵ^2 is minimized)

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

The fitted least squares line is only valid on the range of data that was used to create it. Outside of that range, we are extrapolating.

Interpreting the Betas

β_0 is the y-intercept

- May not have practical meaning if $x = 0$ is outside range of data
- If it does, it is the expected value of y when $x = 0$

β_1 is the slope

- Remember back to high school: “rise over run”
- Represents the expected change in y for every one-unit increase in x
- This relationship often will not hold outside range of sample data

Example: Rocket Propellant

Pretend we are all rocket scientists!

We have gathered a sample of rocket propellants, noted their age in weeks, and measured their sheer strength in pounds per square inch (psi)

What is the relationship between age and sheer strength?

Example: Rocket Propellant

Pretend we are all rocket scientists!

We have gathered a sample of rocket propellants, noted their age in weeks, and measured their sheer strength in pounds per square inch (psi)

What is the relationship between age and sheer strength?

We now have our hypothesized, parameterized model!

$$y = 2628 - 37.15x + \epsilon$$

Example: Rocket Propellant

$$y = 2628 - 37.15x + \epsilon$$

Breaking this model down, we expect:

- Brand new propellant to have sheer strength of 2628 psi
  - Note: Lowest observed value is 2 weeks, so this may be extrapolation.
- Propellant will lose 37.15 psi of sheer strength per week
  - Note: Up to 25 weeks. Beyond that, the linear model may not hold.

And what about that epsilon?

Because we now have the model

$$y = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon = 2628 - 37.15x + \epsilon$$

we can also calculate the observed errors

$$\epsilon_i = y_i - \hat{\beta}_1 x_i - \hat{\beta}_0 = y_i + 37.15x_i - 2628$$

These are called the residuals, and characterizing them is the focus of our next lesson on checking model assumptions

Equations Recap

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

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

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Next time…

Checking Assumptions of an OLS Model