STAT 302: Linear Regression

Sheridan Grant

University of Washington Statistics Department

slgstats@uw.edu

April 15, 2020

Linear Regression: Motivation

Many universities use the SAT to determine who to admit. Is this a good idea?

Linear Regression: Motivation

Many universities use the SAT to determine who to admit. Is this a good idea? Reasons Against:

- Advantages students with certain cultural backgrounds, wealthier students who can afford multiple attempts, coaching
- Can study to do better on the test without actually getting much smarter (easily "gamed")
- Annoying AF; a hassle; better things to do at 8am Saturday (sleep)

Linear Regression: Motivation

Many universities use the SAT to determine who to admit. Is this a good idea? Reasons Against:

- Advantages students with certain cultural backgrounds, wealthier students who can afford multiple attempts, coaching
- Can study to do better on the test without actually getting much smarter (easily "gamed")
- Annoying AF; a hassle; better things to do at 8am Saturday (sleep)

Reasons For:

- *Might* tell us who is best prepared for rigorous study

Linear Regression: Motivation

Many universities use the SAT to determine who to admit. Is this a good idea? Reasons Against:

- Advantages students with certain cultural backgrounds, wealthier students who can afford multiple attempts, coaching
- Can study to do better on the test without actually getting much smarter (easily "gamed")
- Annoying AF; a hassle; better things to do at 8am Saturday (sleep)

Reasons For:

- *Might* tell us who is best prepared for rigorous study

How do we investigate? See if there is a relationship between SAT and first-year GPA.

Linear Regression: Motivation

Your instructor taking the SAT, 2010, colorized

Theory of Linear Regression

Suppose

$$Y_i = \beta_0 + \beta X_i + \epsilon_i$$

for $i \in \{1, \ldots, n\}$

- "Bivariate Data"—observe n realizations of *random variables* X, Y

Theory of Linear Regression

Suppose

$$Y_i = \beta_0 + \beta X_i + \epsilon_i$$

for $i \in \{1, \ldots, n\}$

- "Bivariate Data"—observe n realizations of *random variables* X, Y
- Assume "independent observations." Many ways to violate this, including:
  - Same student took the test twice, is in data set twice
  - Students cheated off one another

Theory of Linear Regression

Suppose

$$Y_i = \beta_0 + \beta X_i + \epsilon_i$$

for $i \in \{1, \ldots, n\}$

- "Bivariate Data"—observe n realizations of *random variables* X, Y
- Assume "independent observations." Many ways to violate this, including:
  - Same student took the test twice, is in data set twice
  - Students cheated off one another
- Assume $\epsilon_i \perp\!\!\!\perp X_i$ (independent errors). Many ways to violate this, including:
  - Higher-SAT students all took harder classes than lower-SAT students (positive correlation between X and ϵ)
  - Higher-SAT students lied about GPAs to look good, lower-SAT students didn't (negative correlation between X and ϵ)

Least Squares

Then: we can estimate β_0, β, understand how good our estimates are, interpret them.

Least Squares

Then: we can estimate β_0, β, understand how good our estimates are, interpret them.
Given *estimates* $\hat{\beta}_0, \hat{\beta}$, we have

- *predicted outcomes*: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta} X_i$
- *residuals*: $Y_i - \hat{Y}_i = \hat{\epsilon}_i$

Least Squares

Then: we can estimate β_0, β, understand how good our estimates are, interpret them.

Given *estimates* $\hat{\beta}_0, \hat{\beta}$, we have

- *predicted outcomes*: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta} X_i$
- *residuals*: $Y_i - \hat{Y}_i = \hat{\epsilon}_i$

Least squares seeks to minimize the squared residuals, solving the problem

$$\begin{aligned} &\min_{\beta_0, \beta} \sum_{i=1}^{n} \hat{\epsilon}_i^2 \\ = &\min_{\beta_0, \beta} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta X_i)]^2 \end{aligned}$$

Least Squares

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta X_i)]^2 \\
&= -\sum_{i=1}^{n} [Y_i - (\beta_0 + \beta X_i)]
\end{aligned}
$$

Least Squares

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta X_i)]^2 \\
&= -\sum_{i=1}^{n} [Y_i - (\beta_0 + \beta X_i)]
\end{aligned}
$$

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \beta} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta X_i)]^2 \\
&= -\sum_{i=1}^{n} X_i [Y_i - (\beta_0 + \beta X_i)]
\end{aligned}
$$

Least Squares

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{\beta}X_i]$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0)X_i}{\sum_{i=1}^{n} X_i^2}$$

These are two equations in two variables, so we can solve for

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

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}\bar{X}$$

Univariate Linear Model Interpretation

An increase in X of 1 unit is **associated** with an increase of $\hat{\beta}$ units in Y

- "An increase of 100 points on the SAT is associated with an improvement in GPA of 0.2 points"
- Why "associated" instead of "causes" or "leads to?" If you cheat on the SAT and improve your score by 100 points, will your college GPA improve by 0.2 points?
- **NOT A CAUSAL RELATIONSHIP**
- Longer but clearer interpretation: "Person A, whose SAT score is 100 points higher than person B, is expected to have a 0.2-point higher GPA in college."

Linear Models with Multiple Covariates

Previously, asssumed X was 1-dimensional (SAT score). Can we do a better job predicting college GPA (the "outcome," or "response") with higher-dimensional X ("covariates," or "predictors")?

- SAT Math
- SAT Verbal (could be more or less informative than math)
- HS GPA (less easily gamed than SAT)

Linear Models with Multiple Covariates

Let (M_i, V_i, G_i) be the *i*th student's SAT Math, SAT Verbal, HS GPA.

$$Y_i = \beta_0 + \beta_M M_i + \beta_V V_i + \beta_G G_i + \epsilon_i$$

- Interpretation of β_M: "an increase of 100 points on the SAT Math, *holding SAT Verbal and HS GPA fixed*, is associated with a 0.1 point increase in College GPA."
- Because SAT Math is related to SAT Verbal and HS GPA, an increase of 100 points on SAT Math is also associated with increases in SAT Verbal and HS GPA, which are *also* associated with changes in College GPA via β_V and β_G