Chapter-3

Katlyn H. Degamo

2023-03-06

General least squares for linear equations

- Consider again the parent and child height data from Galton.
- Let's try fitting the best line. Let Y_i be the i^{th} child's height and X_i be the i^{th} parental heights.
- Consider finding the best line of the form

$$ChildHeight = \beta_0 + ParentHeight\beta_1$$

- Let's try using least squares by minimizing the following equation over β_0 and β_1:

$$\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 X_i)]^2$$

- Minimizing this equation will minimize the sum of the squared distances between the fitted line at the parents' heights ($\beta_1 X_i$) and the observed child heights (Y_i). The result actually has a closed form.
- Specifically, the least squares of the line:

$$Y = \beta_0 + \beta_1 X$$

through the data pairs (X_i, Y_i) with Y_i as the outcome obtains the line $Y = \hat{\beta}_0 + \hat{\beta}_1 X$ where:

$$\hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)} and \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- At this point, a couple of notes are in order. +First, the slope, $\hat{\beta}_1$, has the units of Y/X. +Secondly, the intercept, $\hat{\beta}_0$, has the units of Y. The line passes through the point $(\bar{X}, \bar{Y})$. If you center your Xs and Ys first, then the line will pass through the origin. +Moreover, the slope is the same one you would get if you centered the data,$(X_i - \bar{X}, Y_i - \bar{Y})$, and either fit a linear regression or regression through the origin.
- To elaborate, regression through the origin, assuming that $\hat{\beta}_o = 0$, yields the following solution to the least squares criteria:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

- This is exactly the correlation times the ratio in the standard deviations if the both the Xs and Ys have been centered first. It is interesting to think about what happens when you reverse the role of X and Y.Specifically, the slope of the regression line with X as the outcome and Y as the predictor is $Cor(Y, X)Sd(X)/Sd(Y)$. If you normalized the data, $[\frac{X_i - \bar{X}}{Sd(X)}, \frac{Y_i - \bar{Y}}{Sd(Y)}]$, the slope is simply the correlation, $Cor(Y, X)$, regardless of which variable is treated as the outcome.
- Click me

Revisiting Galton's data

- Finding Galton's data using linear regression.
- We can see that the result of lm is identical to hard coding the fit ourselves.
- Now let's show that regression through the origin yields an equivalent slope if you center the data first.
- Now let's show that normalizing variables results in the slope being the correlation.
- Click me

Showing the OLS result

- Linear regression slope estimate is

$$\hat{\beta}_i = Cor(Y, X)\frac{Sd(y)}{Sd(x)}$$

- Let do regression through origin. I want to fit the best line $Y = X\beta$ through my data.My data is $Y_1, ..., Y_n$ and $X_1, ..., X_n$.
- Minimize the criteria $\sum_{i=1}^{n}(Y_i - X_i\beta)$. Let $\hat{\beta}$ be the solution.
- Solution:

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - X_i\beta) &= \sum_{i=1}^{n}(Y_i - X_i\hat{\beta} + X_i\hat{\beta} - X_i\beta)^2 \\
&= \sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2 - 2\sum_{i=1}^{n}(Y_i - X_i\hat{\beta})(X_i\hat{\beta} - X_i\beta) + \sum_{i=1}^{n}(X_i\hat{\beta} - X_i\beta)^2 \\
&\geq \sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2 - 2\sum_{i=1}^{n}(Y_i - X_i\hat{\beta})(X_i\hat{\beta} - X_i\beta) \\
&\geq \sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2 \\
&\sum_{i=1}^{n}(Y_i - X_i\hat{\beta})X_i = o \\
&\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}
\end{aligned}$$

.

- Example: $X_i, ..., X_n = 1$

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

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

- Click me
- Book