Simple Linear Regression

Date: 2017-11-28

What it is

In its simplest form, simple linear regression is used to predict the value of a numerical **outcome**[1] (y) from a numerical **predictor**[2] (x). The predicted value $\hat{y}$ will have the smallest squared error $(y_i - \bar{y})^2$ of any linear predictor based on only x.

[1] Also called the "response" or "dependent variable"
[2] Also called the "explanatory variable" or "independent variable"

When to use it

Linear regression is used to estimate the functional form of a linear relationship between two variables[3]. In other words, it estimates the slope and intercept of the line that best represents the trend between the variables. As with most statistical methods, we make some simplifying assumptions when using simple linear regression which should be checked before the regression model is put forth as a potential solution[4]. These assumptions are:

[3] As opposed to correlation, which estimates the *strength* of the relationship

[4] See *Regression Diagnostics*

1. There is a linear trend in the data
2. Errors/residuals are normally distributed
3. Errors/residuals have a constand variance across the range of x and y.

How to use it

To run a linear regression in R, use the `lm()` function[5]. The regression variables should be included in a formula of the form `y ~ x`, along with the data set. The regression model is then saved for future use. For example, if we want to use height to predict weight, we would use

[5] `lm` stands for "**l**inear **m**odel."

```
myreg = lm(weight ~ height, data = d)
```

Most information you would want from the regression model can be obtained through the `summary()` function. This output will include intercept and slope calculations with hypothesis tests for each, the residual standard deviation, R^2, and an F test for model fit:

```
summary(myreg)
```

```
##
## Call:
## lm(formula = weight ~ height, data = d)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -107.27  -28.46   -6.90   18.31  324.95
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -180.845     31.426  -5.755 1.52e-08 ***
## height         5.351      0.468  11.435  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.51 on 498 degrees of freedom
## Multiple R-squared:  0.208,  Adjusted R-squared:  0.2064
## F-statistic: 130.7 on 1 and 498 DF,  p-value: < 2.2e-16
```

The output above tells us the estimated regression model is

$$weight = -180.845 + 5.351(height)$$

and the residuals have a standard deviation of 42.51.

Why it works