Simple Linear Regression

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 values $\hat{y}$ will have the smallest squared error $(y - \hat{y})^2$ of any linear predictor based on only x.

When to use it

Simple linear regression is used to estimate the functional form of a linear relationship between two variables[3]. 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:

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

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 physical health to predict mental health, we would use

```
myreg = lm(menthealth ~ physhealth, data=mydat)
```

Most information you would want from the regression model can be obtained through the `coef()` and `summary()` functions. The `coef` function simply gives the estimated intercept and slope while the `summary` function 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[6]:

```
coef(myreg)
```

```
## (Intercept)  physhealth
##   2.5324618   0.2803984
```

[1] Also called the “response” or “dependent variable”

[2] Also called the “explanatory variable” or “independent variable”

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

[4] See *Regression Diagnostics*

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

[6] The `summary(myreg)` example is omitted here because the output takes up about half a page

The output above tells us the estimated regression model is

$$menthealth = 2.53 + 0.28(physhealth)$$

So, it someone has had 10 poor physical health days in the past 30 days, our best guess for the number of poor mental health days they have had is $2.53 + 0.28(10) = 2.58 + 2.80 = 5.38$.

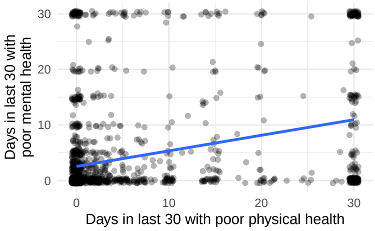

Figure 1: Scatterplot with regression line included.

Why it works

Simple linear regression chooses the slope and intercept estimates to make the best possible predictions of y, the outcome. This is done by considering the squared error $(y - \hat{y})^2$ for each observation in the data set and choosing the slope and intercept values which give the smallest total squared error when added across all observations[7].

If we use the notation $\hat{y} = b_0 + b_1 x$, the formula for the slope of the line is

$$b_1 = r\frac{s_y}{s_x}$$

and the formula for the intercept is

$$b_0 = \bar{y} - b_1\bar{x}$$

We see that correlation plays a role in the slope: the stronger the correlation, the steeper the slope. If there is no correlation, then the slope is 0. In addition, if x is one unit larger, then $\hat{y}$ will be b_1 units larger (or smaller if b_1 is negative).

Usually we are most interested in the slope. The intercept helps balance the equation so the line passes through the point $(\bar{x}, \bar{y})$. We cna see this by re-arranging the formula for the intercept:

$$\begin{aligned} b_0 &= \bar{y} - b_1\bar{x} \\ b_0 + b_1\bar{x} &= \bar{y} \end{aligned}$$

So, the slope gives you the average difference in y you can expect for a one-unit difference in x, and the intercept allows the equation to go through the means of x and y. These estimates work in tandem to give us the best predictions we can make using a simple straight line.

[7] This is called the **sum of squared error** and is the motivation for the other name linear regression is known by: **least squares regression**