Chapter 5.3 - Simple Linear Models

"Statisticians, like artists, have the bad habit of falling in love with their models." - George Box

The "perfect" model doesn't really exist. Models are crude representations of reality, like model airplanes, they're never going to achieve true accuracy. But if we wanted to get as close to perfect as possible we would want a model that's flexible, easy, intuitive, and works more often than not.

That model is the linear model:

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad , \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{I}\sigma^2)$$

This form of the linear model, the *general* linear model, hits all of those pre-requisites we discussed. Data can easily be manipulated to work with it, it can handle any number of parameters so long as they're less than the total amount of samples ($n > p$), and it readily produces highly interpretable results. Most graduate statistics programs will spend *at least* two semesters covering the theories and applications of the linear model. For our purposes we'll be discussing a reduced form, the *simple* linear model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \quad \epsilon_i \sim N(0, \sigma^2)$$

Assumptions

The shift from least squares to the linear model seems minor, but we've made a mammoth change; a *distributional assumption* on the residuals.

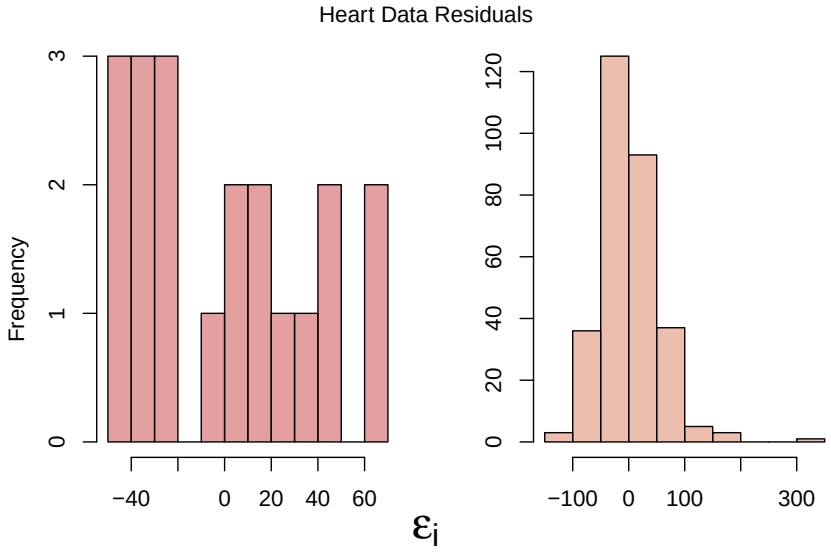
$$\epsilon_i \sim N(0, \sigma^2)$$

The same assumptions we had to meet with least squares apply to the linear model:

- The expected value of the residuals is 0
- We assume homoscedasticity (constant variance)
- There exists linearity in the process we're modeling

We've now added in the assumption of **normality** in the residuals. This is an incredibly useful assumption to work with because we're able to capture (and model) the "random noise" in the data generating process.

It should be noted that the *data* isn't what we're assuming to be normal, but rather the *residuals*. One does not inherently beget the other, but sample size increases tend to result in that normal distribution (thanks to our dear friend the CLT):



A useful feature of this assumption is that σ^2 component. Where we can estimate the variance of the residuals, we can in-turn estimate the variance of regression parameters.

This concept will become more useful as we dive deeper into Chapter 6, but for now we can see how the standard `lm()` function in R (which stands for linear model, if we hadn't figured that out) will give us the "standard errors" (analogous to standard deviation) of regression coefficients.

```r
m=lm(Y~X) # this model uses the 'full' sample size of the heart study
summary(m)[4] # coefficients with standard errors, t-values, and p-values
```

```
## $coefficients
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 179.967471 17.7115839 10.161004 4.771714e-21
## X             1.219441  0.3213438  3.794819 1.786286e-04
```

The method we use to obtain these values is different from least squares, but in the next section we'll show that the coefficient estimates of a linear model will *match* those of least squares.

Maximum Likelihood Estimation

> This portion is fairly advanced relative to the expectation of the textbook. It's been included to help students better understand the linear model but you shouldn't fret if this goes over your head.

When we discussed the normal distribution we introduced its *probability distribution function* (PDF) and threw it aside. We need to revisit this function in order to handle a model that *assumes normality*. (We'll be using $\exp(x)$ to denote e^x)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

If we assume n observations, $x_1, x_2, ..., x_n$, are *independent and identically distributed* (normally) we can consider them to each have a probability of observation that is *modeled* with the normal PDF. From there we can *combine* their individual PDFs into a *joint probability distribution*. We can use this joint PDF to construct a *likelihood* function, describing the likelihood that these observations occur given the parameters we've have:

$$L(\mu, \sigma^2 | x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

Interestingly, the regression parameters of the linear model are estimating the *mean* in this likelihood function. A simple numerical proof can help explain this concept. Imagine an "intercept-only" model:

$$y_i = \beta_0 + \epsilon_i \quad , \quad \epsilon_i \sim N(0, \sigma^2)$$

When we fit this intercept-only model to the total heart study data, using serum cholesterol as the response, we'll find that the mean of y will match the value of β_0.

```r
# intercept only models can be fit in R by setting the only predictor as '1'
cat("Intercept =", coef(lm(Y~1)),"\n")
```

```
## Intercept = 246.264
```

```r
cat("Mean of Y =", mean(Y))
```

```
## Mean of Y = 246.264
```

This leads us to another method of denoting the simple linear model:

$$y_i \sim N(\beta_0 + \beta_1 x_i \ , \ \sigma^2)$$

We can also use a separate statement to define μ:

$$y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \beta_0 + \beta_1 x_i$$

We're constructing this likelihood function specifically for the *response* (y_i) using the *predictors* (x_i) and the regression parameters($\beta_0, \beta_1, \sigma^2$) so we'll need to do some rearranging in addition to substituting the regression equation for μ:

$$L(y_i|x_i, \beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\right)$$

This likelihood function now informs us on the probability of observing y_i **given** (remember your conditional probabilities) the values we have for x_i, β_0, β_1, and σ^2. So higher values imply a higher likelihood of those observations of y_i matching the values of the data and parameters. Naturally we'd like to *maximize* this likelihood, which is a shockingly simple algorithm!

1. Calculate the first partial derivative of the likelihood function (w.r.t. the parameter of interest)
2. Set the equation equal to 0 and solve for the parameter of interest

We'll find that the likelihood function is a hassle to work with during this process, but we can take the *natural logarithm* (we'll denote this as log) of the likelihood function and work with it the same way (this is referred to as the "log-likelihood"):

$$\ell(y_i|x_i, \beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i-1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

Solving the following operations:

$$\frac{\partial \ell(y_i|x_i, \beta_0, \beta_1, \sigma^2)}{\partial \beta_0} = 0$$

$$\frac{\partial \ell(y_i|x_i, \beta_0, \beta_1, \sigma^2)}{\partial \beta_1} = 0$$

$$\frac{\partial \ell(y_i|x_i, \beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} = 0$$

Will produce the following results:

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

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

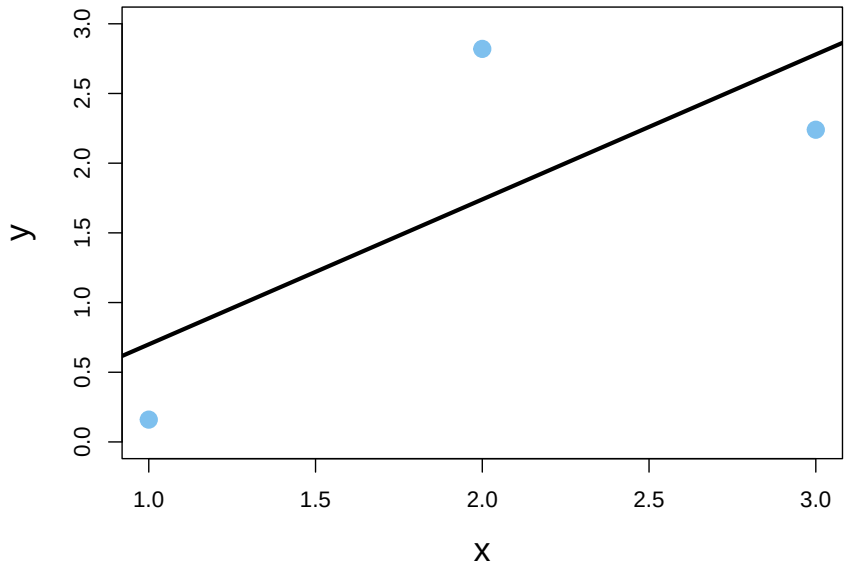
$\hat{\beta}_0$ and $\hat{\beta}_1$ match the least squares estimators because of the *assumption of independence* in the residuals. This means that whatever solution we find for least squares estimators will *also* maximize the likelihood function for the analogous linear model. This is why the `lm()` function can be considered a least squares regression model *if we only consider the coefficients.*

The defining difference, and reason we use the linear model, is that we can quantify the variance in our estimates from a linear model. Without a distributional assumption on the residuals we can't do this.

> As a final note, the distribution of the residuals **doesn't need to be normal**. If we change the distributional assumption from normal to some other named (or non-named) distribution then we've constructed a **generalized** linear model (GLM). GLMs are an extremely useful tool for applied sciences but are fairly complex and not worth discussion in this chapter.

The Coefficient of Determination: R^2

When we fit a least squares regression line to our trivial data set in chapter 5.2, it was pretty obvious that it wasn't *passing through* any of the points:



This implies (as we should be aware) that the model is inaccurate. Quantifying the level a model is *accurate* is an obsession of many scientists and statisticians alike. The linear model (and least squares) provide a simple measurement of this "accuracy" or "goodness of fit", R^2.

If we were to take the differences between the observed and predicted values of the response, we would have calculated the *residuals*. If we square these values we're given the *squared difference* between the observations and predictions:

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	0.16	0.7	-0.54	0.2916
2	2.82	1.74	1.08	1.1664
3	2.24	2.78	-0.54	0.2916

Just like with variance, this squaring step gives us a *magnitude* without having to worry about *sign*. If we sum up these squared differences we have the *total magnitude* of the squared difference between our observations and our predictions:

$$0.2916 + 1.1664 + 0.2916 = 1.7496$$

This is a pretty useful measurement since we can consider it to be a quantification of how *incorrect* our model is from reality. But we're left with two problems.

- Whatever unit this magnitude is in holds a lot of the context as to what we've actually measured.
- We don't know how "significant" the magnitude is even with the context of units.

Think back to concept behind z-scores: A difference in weight of 2 lbs. may be significant when we're talking about one group of people while it could be completely irrelevant for another.

We need to develop some way of contextualizing this squared difference within the data. We may have already come up with a good candidate for achieving this— the difference between the observations and their average. We'll want to square these values to ensure everythings in the same dimensions:

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$	$y - \bar{y}$	$(y - \bar{y})^2$
1	0.16	0.7	-0.54	0.2916	-1.58	2.4964
2	2.82	1.74	1.08	1.1664	1.08	1.1664
3	2.24	2.78	-0.54	0.2916	0.50	0.2500

$$2.4964 + 1.1664 + 0.2500 = 3.9128$$

What should we do with it now? Well, it'd be disappointing if our instinct at this point wasn't to take a ratio between these two sums:

$$\frac{1.7496}{3.9128} = 0.4471$$

What this ratio represents is how **incorrect** the model is. Since it's a ratio it's bounded between 0 and 1; if we take the *complement* of this ratio we should have a measurement of how **correct** the model is:

$$1 - \frac{1.7496}{3.9128} = 0.5529$$

This value, 0.5529, is the R^2 ("R squared") of the model, also known as the "coefficient of determination". It represents the proportion of the variance that's explained by the independent variable relative to the total variance in the data. Another way of thinking of it is as a measurement of the "goodness of fit" of the regression line. The general formula for R^2 is:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Where RSS is the "residual sum of squares":

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

And TSS is the "total sum of squares":

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

But there's an alternate formula for R^2 (several if we really dive into it):

$$R^2 = r \times r$$

Where r is the *correlation coefficient*. It's another showcase of the general laziness of statisticians that we would call the squaring of r "R squared".

$\bar{x}$	$\bar{y}$	s_x	s_y
2	1.74	1	1.399

$$r = \frac{1}{n-1}\sum_i \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$\frac{1}{2}\left(\left(\left(\frac{1-2}{1}\right)\left(\frac{0.16-1.74}{1.399}\right)\right)+\left(\left(\frac{2-2}{1}\right)\left(\frac{2.82-1.74}{1.399}\right)\right)+\left(\left(\frac{3-2}{1}\right)\left(\frac{2.24-1.74}{1.399}\right)\right)\right)$$

$$\frac{1}{2}(1.129378 + 0 + 0.3573981) = \frac{1}{2}(1.486776) = 0.743388$$

$$0.743388^2 = 0.553$$

While R^2 is a good metric to use when checking the "quality" of a simple linear regression model it quickly falls apart when we introduce more than one parameter. We should also be weary of using R^2 as a substitute for *scientific* inference. Just because the value of R^2 is high doesn't mean the model is inherently representing the process correctly. Likewise, a low value of R^2 is not a black-and-white indicator of a "bad" model.

We have to use our scientific expertise *in tandem* with statistical models to make them worthwhile. Otherwise we may as well throw up our hands and say that scientific inquiry is completely meaningless.