Chapter 3 Summary: Linear Regression

Linear regression (yes, the one from middle or high school) may seem boring and not worth your time, but it is! It is widely used and many fancier models are generalizations or extensions of linear regression. So seriously, this is important.

3.1 Simple Linear Regression

Simple Linear Regression assumes that the response variable, Y has a linear relationship to a single predictor variable X.

$$Y \approx \beta_0 + \beta_1 X \tag{3.1}$$

> Note
>
> This is the same relationship as $y = mx + b$ you likely remember from school.
> β_0 represents the intercept, b while β_1 represents the slope, m.
> In the future, we will use this same notation when Y and X are vectors, usually writing them in bold $\mathbf{Y} \approx \beta_0 + \beta_1 \mathbf{X}$

β_0 and β_1 are referred to as the **coefficients** or **parameters**.
We will use our training data to estimate these parameters and then indicate and reprsent our **model** as

$$\hat{Y} \approx \hat{\beta}_0 + \hat{\beta}_1 X \tag{3.2}$$

> Note
>
> We call this "hat notation." In statistics, a hat over something indicates it's either an estimator or an estimated value.

3.1.1 Estimating the Coefficients

Since we do not know the true relationship between X and Y, we use the training data to estimate β_0 and β_1

> Note
>
> Just like you did in middle/high school when calculating the line of best fit

We will use **least squares** as the criteria to determine what values to use for our parameters, β_0 and β_1.

The ith **residual** is the difference between the ith response variable and our prediction for that variable

$$e_i = y_i - \hat{y}_i$$

The **residual sum of squares** is defined

$$\begin{aligned}\text{RSS} &= e_1^2 + e_2^2 + \cdots + e_n^2 \\ &= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2\end{aligned}$$

The parameters can be calculated as:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x}x)^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x}\end{aligned} \tag{3.4}$$

where $\bar{y}$ and $\bar{x}$ are the sample mean, defined below

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{and} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

3.1.2 Assessing the Accuracy of the Coefficient Estimates

Standard Error

Standard error tells us the average amount that an estimate differs from the actual value.

$$\text{Var}(\hat{\mu}) = \text{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n} \tag{3.7}$$

$$\text{SE}(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \tag{3.8}$$

$$\text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{also 3.8}$$

where $\sigma^2 = \text{Var}(\epsilon)$ While we generally don't know σ^2, it can be estimated from the data.
The estimate of σ is called **residual standard error**

$$\sigma = \text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}}$$

Confidence Intervals

Standard errors can be used to compute **confidence intervals**. The 95% confidence interval for β_0 and β_1 are approximately:

$$\begin{aligned}&\hat{\beta}_0 \pm 2 \cdot \text{SE}(\hat{\beta}_0) \\ &\hat{\beta}_1 \pm 2 \cdot \text{SE}(\hat{\beta}_1)\end{aligned} \tag{3.11}$$

This means there is approximately a 95% chance that the interval

$$\left[\beta_0 - 2 \cdot \text{SE}(\hat{\beta}_0), \beta_0 + 2 \cdot \text{SE}(\hat{\beta}_0)\right] \tag{almost 3.10}$$

contains the true value for β_0. **NOTE: This is an incorrect interpretation of a confidence interval , but it is what the book writes. Please see the warning below**

> Note
>
> 1.96 is closer to the correct value than 2. This value comes from the Z-score value for a 97.5% quantile of a t-distribution. You are likely to see these in any statistics class.

Warning

A 95% confidence interval does **NOT** mean we are 95% confident that the true parameters lies within that range.
What it really means is that we sampled the data 100 times, each time calculating the parameter and confidence interval 95% of those confidence intervals would contain the true value of the parameter.
It's a small distinction, but I wanted to make it, even if the book did not.
That said, it's generally okay to say we're 95% confidence the true value is in the interval, even if it's incorrect. In pure statistics, the distinction is generally considered

3.2 Multiple Linear Regression

3.2.1 Estimating the Regression Coefficients

3.2.2 Some Important Questions

3.3 Other Considerations in the Regression Model

3.4 The Marketing Plan

3.5 Comparison of Linear Regression with K-Nearest Neighbors