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.

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{3.2}$$

> **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

$$\begin{aligned}
\text{RSS} &= e_1^2 + e_2^2 + \cdots + e_n^2 \\
&= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 \\
&= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2
\end{aligned}$$

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.

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

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

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

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

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

Confidence Intervals

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

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

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

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

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.

Interesting Tidbit!

Did you know that the t-test was developed by in order to make better beer? William Sealy Gosset, while the head experimental brewer at Guinness, developed the t-test as a way to study the quality of the barley used in brewing Guinness.

Warning

A 95% confidence interval does **NOT** mean we there is a 95% chance that the true parameter lies within the range.
What it really means is that if 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.
We generally say that we are 95% confident that the true parameter lies in the range, not that the probability is 0.95.

Hypothesis Testing

Standard errors can also be used to perform hypothesis testing. The most common of which is the *null hypothesis*. We will test to see if the data provides evidence to reject the null hypothesis (that two variables/phenomena/results have no relationship) in favor of the alternative hypothesis (that there is a relationship)

$$\underbrace{H_0}_{\text{Null hypothesis}} : \text{There is no relationship between } X \text{ and } Y \implies \beta_1 = 0 \tag{3.12}$$

$$\underbrace{H_1}_{\text{Alternative hypothesis}} : \text{There is a relationship between } X \text{ and } Y \implies \beta_1 \neq 0 \tag{3.13}$$

We compute the **t-statistic**:

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} \tag{3.14}$$

If $\beta_1 = 0$, this will have a t-distribution with $n - 2$ degrees of freedom. The probability of observing a value greater than or equal to $|t|$ is called a **p-value** . If the p-value is small, there it is unlikely that the relationship between predictor and response is due to chance. We **reject** the null hypothesis (and accept the alternative hypothesis, if the p-value is "small enough")

DEAN, ADD MORE

3.1.3 Assessing the Accuracy of the Model

Residual Standard Error

Suppose we knew the exact true model, $Y \approx \beta_0 + \beta_1 X$ (from equation 3.1), recall that there is the irreducible error, ϵ, associated with every term. Residual Standard Error (RSE) will attempt to estimate the standard deviation of that irreducible error.

It does this by seeing how well your model fits the data! Convenient, let's calculate the standard deviation of our residuals!

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.15}$$

RSE tells us about the spread of the observed values from the predicted ones, how far off the model's predictions are, on average.. A lower RSE means the predicted values are closer to the observed ones.

RSE is in whatever units the Y variable is in, so it can be hard to understand what it really means. For that, let's look at...

R^2 Statistic

R^2 also measures the accuracy of the model, but does it as a proportion so the values are always between 0 and 1.

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{3.17}$$

where TSS, the **total sum of squares** is defined

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

- TSS is the total variance in the response, Y. Think of it as the amount of variability in the response (not dependent on the model)
- RSS measures the amount of variability that is left unexplained after performing the regression
- TSS - RSS then measures the amount of variability in the response that is explained by performing the regression.
- R^2 measures the proportion of variability in Y that can be explained using X

An R^2 value of 1 means the model perfectly explains all variability. An R^2 of 0 means the model explains none of it. It does just as well as predicting the mean.

Should I add correlation?

3.2 Multiple Linear Regression

What happens if we have more than one dependent variable? If we have p input variables, a multiple linear regression model takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{3.19}$$

X_j is the jth predictor, β_j is the corresponding coefficient. It can be interpreted as the average effect on Y of a one unit increase in X_j, given that no other predictor variables change.

3.2.1 Estimating the Regression Coefficients

Our multiple linear regression model will be

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p. \tag{3.21}$$

Like in simple linear regression, we will choose our βs in order to minimize the residual sum of squares.

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

Note

While not covered by ISLP, in order to make the book more approachable (and not dependent on linear algebra), you will often see multiple linear regression represented in matrix form. If we let

$$\boldsymbol{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{bmatrix} \qquad \text{and} \qquad \mathbf{X}_i = \begin{bmatrix} 1 & x_{i1} & x_{i2} & \cdots & x_{in} \end{bmatrix}$$

then

$$\hat{y}_i = \mathbf{X}_i \boldsymbol{\beta}$$

Note that you will likely see this represented as multiple rows in $\mathbf{X}$ and $\mathbf{Y}$ representing different observations. I've omitted that to make it easier to understand.

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