Linear Regression Analysis

Scott Shaffer

Fall 2019

Contents

1 Introduction

1.1 Statistical Models

This paper is on a certain sort of statistical models. In any model we begin with a **response variable** Y and a **predictor variable** X. (or several predictors $X_1, X_2, X_3, ..., X_m$). A function of X gives us an estimate of Y (hence why X and Y are called the predictor and response respectively). Now this is where statistical models diverge from their counterpart, deterministic models. In a deterministic model we would propose that $Y = f(X)$. However, we call $f(X)$ an estimate of the response because it is truly a mere estimate. There is still some difference between $f(X)$ and Y, which we call the **error** ε. Therefore to make a statistical model we propose that the response is the sum of the estimate and the error i.e.

$$Y = f(X) + \varepsilon. \tag{1.1}$$

1.2 Linear Statistical Models

In the context of linear regression, linear refers to a property of the response estimate $f(X)$. Suppose that the response estimate $f(X)$ is a function of $p-1$ predictors $X_1, X_2, ..., X_{p-1}$ and has p parameters labeled $\beta_0, \beta_1, ..., \beta_{p-1}$. The model is linear if $f(X)$ is a linear combination of the parameters $\beta_0, \beta_1, ..., \beta_p$. Hence a linear regression model has the form

$$Y = \beta_0 + \beta_1 f_1(X_1) + \beta_2 f_2(X_2) + ... + \beta_{m-1} f_{m-1}(X_{m-1}) + \varepsilon. \tag{1.2}$$

A misconception that many have about linear regression models is that they assume linear means Y is linear in the predictors, but we can take any arbitrary function of the predictor. The following **is** a linear regression model:

$$Y = \beta_0 + \beta_1 e^{X_1} + \beta_2 \cos X_2 + \beta_3 \Gamma(X_3) + \varepsilon.$$

Whereas the following **is not** a linear regression model:

$$Y = \ln\left(\frac{\beta_1 + \beta_2 X_1}{X_2}\right) + \varepsilon.$$

What's convenient for us is that we can reformulate many nonlinear models into linear ones. With the variables above, consider defining the variables

$$\begin{aligned} Y' &= e^Y, \\ X_1' &= \frac{1}{X_2}, \\ X_2' &= \frac{X_1}{X_2}. \end{aligned}$$

Then we can write a new model:

$$Y' = \beta_1 X_1' + \beta_2 X_2' + \varepsilon,$$

which is linear.

1.3 Error

As the variable in our model that makes the model statistical, and not deterministic, ε is no doubt important to statistical modeling. But we should take a second to discuss what exactly ε is. This question requires us to go outside of statistics to understand. I'll provide a couple of examples to elucidate the concept. The three following examples illustrate different interpretations of error. However, no matter the source of error the mathematics that we use to estimate the error remains the same. That said this section is meant to give the reader a better conceptual understanding of error, rather than better mathematical tools.

Error is an issue with our measuring tools: suppose for instance we have an old analog style scale (where a plate is displaced by mass placed on top of it). We measure that the plate is 20in high with no weight on it. Then using the tension in the spring we work out Newton's force laws and find that for every pound placed on the plate, the plate should lower 2 in. Our model would have the form $Y = 20 - 2X$ where Y is the weight of the object on the plate and X is the weight of the object. However when we place an object that weighs 5lb's on the plate, we get our a ruler and measure that the height of the plate is 10.1in, which is 0.1 more inches than we predicted. Surely we have confidence that the force of gravity isn't working differently than we predict. The more reasonable conclusion is that we simply didn't read the ruler correctly. We could rewrite our model as $Y = 20 - 2X + \varepsilon$ where ε is a measure of how many inches I add to the actual value of Y because my measurement is imperfect. In this example error is a good term for ε because ε is truly the error we make in our measurement.

Error is the affect of unseen variables: Let X be the height of a father and let Y be the height of the father's son. Intuitively we expect that the taller the father then the taller the son. In fact we expect $Y = X$. Now if you measured a father and his son and found their heights differed by 6in, you wouldn't believe it's because you measured son's height incorrectly, so you wouldn't believe there's a deficiency in the measuring tool as in the last example. What sets the weight on the scale and the father and son examples apart is that in the first we hypothesize that plate height is determined by object weight. We know that the height of the son isn't determined by the height of the father, the correlation coefficient between the two isn't one. In this example error seem to mean more or less "error in our model" rather than "error in our estimate" where the estimate, recall, is that function of the predictor $f(X)$.

Error is the affect of a truly random world: In the last example we could have added terms to our model in order to account for other variables e.g. height of the mother, diet, etc. However, we'd still be left with the fact that height is largely determined by genes being copied from the parents to the child. When a genes get copied it requires processes which occur on atomic scales. On these scales the movement of particles is described by quantum mechanics. The model we use to describe the motion of a particle on a quantum scale is itself a statistical one, but the error in this model is due to the fact that nature itself seems to be probabilistic. Hence in this case error isn't an inadequacy of our measurement or our estimation, but rather an inherent property of how things work.

Error is a random variable: This is perhaps the best way to describe what the error is. In this way error is no different than the flip of a coin. In a coin flip we say the coin has a $1/2$ chance of landing on each side. In reality a given coin flip is going to be determined by the mechanics of the force applied to it and the air it travels through, yet by using probability we can be pretty confident that 1,000 heads in a row won't happen. Likewise we won't know the source of our error, and we won't even know it's value, but the existence of the random variable ε in our model has practical applications in the same way the hypothetical $1/2$ chance a coin lands on heads has practical applications. [2]

2 Simple Linear Regression Models

Recall we said a statistical model has the form $Y = f(X) + \varepsilon$, with our response Y, our predictor X, and our error ε. We can have more than one predictor, but in simple linear regression we'll stick to one. The estimate $f(X)$ in a simple linear regression model has two parameters β_0 and β_1. The function itself if $f(X) = \beta_0 + \beta_1 X$.

Therefore the simple regression model is

$$Y = \beta_0 + \beta_1 X + \varepsilon. \tag{2.1}$$

This is the relationship we propose between the response and predictor, but of course we'll be taking more than one sample of Y and X. We'll suppose there are m samples of Y and X, and we'll index a given sample with i. Hence for a given trial

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \tag{2.2}$$

There are several important assumptions we'll make about Y and ε. Many of the later theorems in this paper will rely on these properties [1]:

1. **Homoscedasticity** The variance of the error does not change:

$$\sigma^2[\varepsilon_i] = \sigma_\varepsilon^2 \tag{2.3}$$

2. **Expected Error is 0** We always expect the error to be 0:

$$\mathrm{E}[\varepsilon] = 0 \tag{2.4}$$

3. **Normality** The error is normally distributed:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{2.5}$$

4. **Predictor has no Error** Unlike the response Y, the predictor X is not a random variable:

5. **Responses are Uncorrelated** Lastly, we assume that $\forall i, j \leq m$ Y_i and Y_j are uncorrelated:

$$\sigma^2[Y_i + Y_j] = \sigma^2[Y_i] + \sigma^2[Y_j] \tag{2.6}$$

The following properties are implications of those previous.

6. **Expected Response is Estimate** This is an implication of assumptions 2 and 4:

$$\mathrm{E}[Y_i] = \mathrm{E}[\beta_0 + \beta_1 X_i + \varepsilon_i] = \mathrm{E}[\beta_0 + \beta_1 X_i] + \mathrm{E}[\varepsilon_i] = \beta_0 + \beta_1 X_i. \quad (2.7)$$

Recall our estimate of the response $f(X) = \beta_0 + \beta_1 X$. This means

$$\mathrm{E}[Y_i] = f(X_i). \quad (2.8)$$

7. **Distribution of the Response** This is an implication of assumptions 1, 3, and 4. The distribution of Y maintains the shape of ε:

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 X, \sigma_\varepsilon^2) \quad (2.9)$$

Given that the variance in Y and ε are the same we can use σ_ε^2 and σ_Y^2 interchangeably i.e.

$$\sigma_\varepsilon^2 = \sigma_Y^2 \quad (2.10)$$

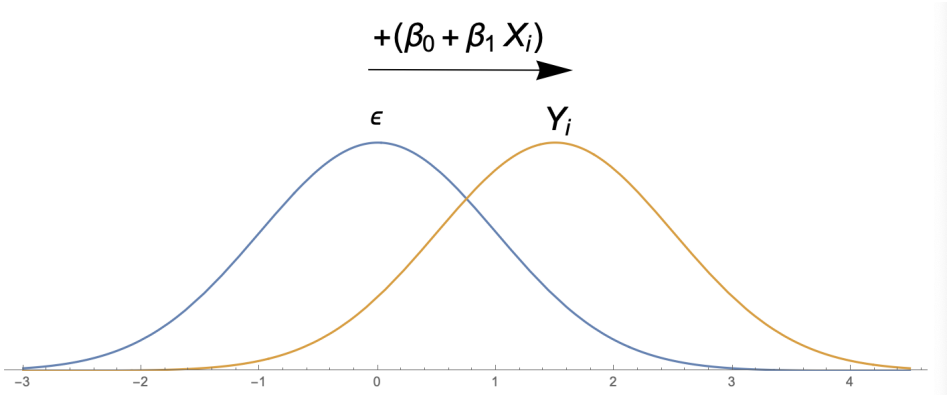

Figure 1: Y_i is modeled as the error plus a constant

All of these assumptions are fairly common in the realm of linear regression models, but its the assumption that ε is normally distributed that really does a lot of heavy lifting. This might already seem pretty clear since, recall from property 7, it gives us a distribution for Y. However this assumption takes us further. The condition that ε is normal will tell us how to find confidence intervals for our estimated regression coefficients and will give us prediction intervals for Y.

Now suppose we're given two variables X and Y. There are some hypothetical true regression coefficients, β_0 and β_1, of the simple linear regression

model. We use the notation $\hat{\beta}$ to talk about the estimation of a regression coefficient β. Likewise $\hat{\sigma}_\varepsilon^2$ is an estimate for σ_ε^2. By property 7 of the regression model, β_0, β_1, and σ_ε^2 completely parameterize the distribution of the response Y. Hence it will be our job to find the best values we can for $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_\varepsilon^2$.

2.1 Least Squares Estimators

To understand least squares we first introduce the definition of a **residual** e. Residuals are a way of estimating the error in the response. A residual is the difference between the response Y and the estimated response $\hat{Y}$.

$$e = Y - \hat{Y} = Y - (\hat{\beta}_0 + \hat{\beta}_1 X) \tag{2.11}$$

Residual can be easily confused with error, for good reason. The equations for the two are almost identical. We wrote $Y = \beta_0 + \beta_1 X + \varepsilon$ and wrote $Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + e$. The confusion arises in part because the Y in the former and latter equations aren't exactly the same. In the former equation Y is still written as a random variable (written in terms of another random variable, ε). In the latter equation Y is an observed quantity (sampled from the distribution $\mathcal{N}(0, \sigma_Y^2)$). That's why e is a known quantity but ε is random. Conceptually speaking the two are very different but related quantities. e is simply the output of a formula, the difference of the observed and estimated response, whereas ε is an unknown random variable that determines the variance in the response. Because ε is the unknown variable that impact the variation of the response, and e measures the variation in the response, we can use the values of e to estimate ε.

We now move on to discuss the quantity for which the least squares estimators are named, the **residual sum of squares**, RSS. As the name implies, the RSS is the sum over the squares of all the residuals, or in other words

$$\text{RSS} = \sum_{i=1}^{m} (Y - \hat{Y})^2 = \sum_{i=1}^{m} e_i^2. \tag{2.12}$$

The least squares regression model is defined to be that model which minimizes the RSS, thus the name. We now proceed to elaborate on several properties of the least squares model, which we'll also call the OLS model (ordinary least squares):

1. The OLS model minimizes the residual sum of squares.

2. Furthermore, the sum of all residuals is zero. This will be shown in the derivation of the regression coefficients.

$$\sum_{i=1}^{m} e_i = 0 \tag{2.13}$$

3. As a consequence of Equation 2.13, the sum of observed values equals the sum of fitted values

$$\sum_{i=1}^{m} Y_i = \sum_{i=1}^{m} \hat{Y}_i \tag{2.14}$$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{m} Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) = \sum_{i=1}^{m} Y_i - \hat{Y}_i = 0 \Rightarrow \sum_{i=1}^{m} Y_i = \sum_{i=1}^{m} \hat{Y}_i$$

4. The regression coefficients are as follows, see Appendix A for a derivation

$$\hat{\beta}_0 = \frac{1}{n}(\sum Y_i - \hat{\beta}_1 \sum X_i) = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{2.15}$$

$$\hat{\beta}_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \tag{2.16}$$

One last thing the reader should note about the OLS regression coefficients is that even when we are using different linear regression models we can still use the OLS method to find the regression coefficients.

2.2 Best Linear Unbiased Estimators

What follows is a proof of the Gauss-Markov theorem. I've taken $\hat{\beta}_0$ and $\hat{\beta}_1$ to be the parameter estimators given by the OLS method. In proving the Gauss Markov Theorem we have to briefly discuss BLUEs. We haven't yet encountered any of the three parts of the term (best, linear and unbiased).

1. **Linear** Linear in this context is different than the linear we discussed in the Introduction. In fact it's essentially the opposite. Linear Regression is linear because the response is a linear combination of the parameters [1]. An estimate of a regression coefficient is linear if it is a linear combination of the observed response. Written formally

$$\hat{\beta} = \sum k_i Y_i$$

2. **Unbiased** An estimator of the regression coefficient is unbiased if the expected value of the estimator is the regression coefficient itself:

$$\mathrm{E}[\hat{\beta}] = \beta$$

3. **Best** An estimator is the best of its class if it has minimum variance amongst all estimators of that class. So for all B in the class of linear unbiased estimators

$$\sigma^2[\hat{\beta}] \leq \sigma^2[B]$$

Now we'll move on to the theorem.

Theorem 2.1. Gauss-Markov Theorem Given each error term is uncorrelated, the error terms are homoscedastic, and $\mathrm{E}[\varepsilon] = 0$, the OLS coefficient estimators are the best linear unbiased estimators (BLUE's).

Proof. For a proof that the OLS estimators are linear and unbiased refer to Appendix A. In what follows we'll prove that the OLS estimators have minimum variance amongst all other linear, unbiased estimators i.e. they are the "best" of their class.
Choose B_1 to be a linear, unbiased estimator of β_1; $\hat{\beta}_1$ will continue to be the OLS estimate of β_1. Furthermore we'll refer to the variance of the error and response variable, σ_ε^2, given in Equation 2.9.
Since B_1 is linear we have that

$$B_1 = \sum c_i Y_i$$

We want to show that if B_1 is the best estimator, then $B_1 = \hat{\beta}_1$; or equivalently, the constants c_i are equal to the least squares constants k_i in Equation A.7.

$$\begin{aligned}
\mathrm{E}[B_1] &= E[\sum c_i Y_i] \\
&= \sum c_i E[Y_i] \\
&= \sum c_i(\beta_0 + \beta_1 X_i) \\
&= \beta_0 \sum c_i + \beta_1 \sum c_i X_i
\end{aligned}$$

Now using B_1 is unbiased i.e. $\mathrm{E}[B_1] = \beta_1$

$$\sum c_i = 0, \quad \sum c_i X_i = 1. \tag{2.17}$$

Now we know

$$\begin{aligned}\sigma^2[B_1] &= \sigma^2[\sum c_i Y_i] \\ &= \sum \sigma^2[c_i Y_i] \\ &= \sum c_i^2 \sigma_\varepsilon^2 \\ &= \sigma_\varepsilon^2 \sum c_i^2\end{aligned}$$

So to minimize $\sigma^2[B_1]$ we need to find expressions for c_i which minimize $\sum c_i^2$ while also meeting the conditions in Equation 2.17. We follow the method given in [1, pg.44]:
Now define $c_i = k_i + d_i$

$$\begin{aligned}\sigma^2[B_1] &= \sigma^2 \sum (k_i + d_i)^2 \\ &= \sigma^2(\sum k_i^2 + \sum d_i^2 + \sum 2k_i d_i)\end{aligned} \tag{2.18}$$

Furthermore we can prove $\sum 2k_i d_i = 0$, we'll have to use Equation A.15

$$\begin{aligned}\sum k_i d_i &= \sum k_i d_i \\ &= \sum k_i (c_i - k_i) \\ &= \sum c_i k_i - \sum k_i^2 \\ &= \sum \Big[\frac{c_i(X_i - \bar{X})}{\sum (X_i - X)^2}\Big] - \frac{1}{\sum (X_i - \bar{X})^2} \\ &= \frac{\sum c_i X_i}{\sum (X_i - \bar{X})^2} - \frac{\bar{X} \sum c_i}{\sum (X_i - \bar{X})^2} - \frac{1}{\sum (X_i - \bar{X})^2} \\ &= \frac{1}{\sum (X_i - \bar{X})^2} - 0 - \frac{1}{\sum (X_i - \bar{X})^2} \\ &= 0\end{aligned} \tag{2.19}$$

Now we have

$$\begin{aligned}\sigma^2[B_1] &= \sigma_\varepsilon^2(\sum k_i^2 + \sum d_i^2 + \sum 2k_i d_i) \\ &= \sigma_\varepsilon^2 \sum k_i^2 + \sigma_\varepsilon^2 \sum d_i^2 + \sigma_\varepsilon^2 2 \sum k_i d_i \\ &= \sigma^2(\hat{\beta}_1) + \sigma_\varepsilon^2 \sum d_i^2\end{aligned} \tag{2.20}$$

Because B_1 is an arbitrary unbiased, linear estimator, we find that for all estimators in this class, the variance must be at least as large as the least squares variance with the addition of of the second term in Equation 2.20.

We can minimize this expression if all the $d_i = 0$. If all $d_i = 0$, B_1 has the minimum variance of all linear unbiased estimators. Furthermore, in this case for all i $c_i = k_i + d_i = k_i$, meaning that best linear unbiased estimator $B_1 = \hat{\beta}_1$.
Therefore $\hat{\beta}_1$ is the best linear unbiased estimator of β_1. ∎

When we introduced the simple linear regression model, we made assumptions about the model which make the Gauss-Markov theorem valid. Hence through the remainder of the paper we'll use OLS method to find the regression coefficient estimates.

2.3 Inferential Statistics in Simple Linear Regression

As we saw previously, the distribution of Y, for all X, is determined the regression coefficients β_0 and β_1 and by the variance in the error σ_ε^2. The Gauss-Markov theorem has given us regression coefficient estimators we can use, but we still need to a way to estimate the variance in our error.

As I stated before the residuals estimate the error in our response variable. That said, the sample variance of the residuals is an estimator of the variance of the error. Applying the sample variance to the residuals gives us the **mean-squared error**, or MSE (this term is a flippant use of the term error, as the MSE is really the mean sqaured residual) [3].

$$\text{MSE} = \frac{1}{m}\sum(Y_i - \hat{Y}_i)^2 \tag{2.21}$$

We'll use the MSE to estimate σ_ε^2, and in fact this estimator is unbiased [1,pg. 34]. Having an estimator for the variance, we can estimate the variance of our regression coefficients. We'll derive the variance of the slop coefficient β_1. This follows the method in [1, pg. 43], and requires us to use the linear form of β_1 found in Appendix A.

$$\begin{aligned}
\sigma^2[\hat{\beta}_1] &= \sigma^2[\sum k_i Y_i] \\
&= \sum k_i^2 \sigma^2[Y_i] \\
&= \sum k_i^2 \sigma_Y^2 \\
&= \sigma_Y^2 \sum k_i^2 \\
&= \frac{\sigma_Y^2}{\sum(X_i - \overline{X})^2}
\end{aligned}$$

Now we wish to find an estimator of the variance in β_1, $\hat{\sigma}^2[\hat{\beta}_1]$. We can produce an unbiased estimator of $\hat{\sigma}^2[\hat{\beta}_1]$ using the MSE quantity.

$$\hat{\sigma}^2[\hat{\beta}_1] = \frac{\text{MSE}}{\sum(X_i - \overline{X})^2} \tag{2.22}$$

.

The fact that $\hat{\sigma}^2[\hat{\beta}_1]$ is unbiased follows immediately

$$\begin{aligned}
\text{E}[\hat{\sigma}^2[\hat{\beta}_1]] &= \text{E}\left[\frac{\text{MSE}}{\sum(X_i - \overline{X})^2}\right] \\
&= \frac{\text{E[MSE]}}{\sum(X_i - \overline{X})^2} \\
&= \frac{\sigma_Y^2}{\sum(X_i - \overline{X})^2}.
\end{aligned}$$

The estimated variance of β_1 is also called the **standard error** of β_1. We'll use the standard error in order to make a confidence interval around β_1. To do that we'll make use of the following theorem,

Theorem 2.1. Given $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ then the following quantity is distributed as t with $m-2$ degrees of freedom,

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}^2[\hat{\beta}_1]}$$

Theorem 2.1 is actually a particular case of a more general theorem we'll see later, so the proof will be given there. Given a p-value α the $1-\alpha$ confidence interval is

$$\hat{\beta}_1 \pm t(1-\alpha/2; m-2)\hat{\sigma}^2[\hat{\beta}_1]$$

Conversely to find how big a confidence interval around $\hat{\beta}_1$ would have to be in order to contain a hypothetical value of β_1 we use

$$P(\hat{\beta}_1 - t(1-\alpha/2; m-2)\hat{\sigma}^2[\hat{\beta}_1] \leq \beta_1 \leq \hat{\beta}_1 + t(1-\alpha/2; m-2)\hat{\sigma}^2[\hat{\beta}_1]) = 1-\alpha$$

2.4 Applying a Simple Linear Regression Model

Knowing how the least squares coefficients are derived, and being able to make inferences about a linear regression model, we'll apply our theory to an example. In Figure 3.3 I've graphed real data on the arctic sea ice extent, the area of the arctic sea covered in ice. In particular I've taken the arctic

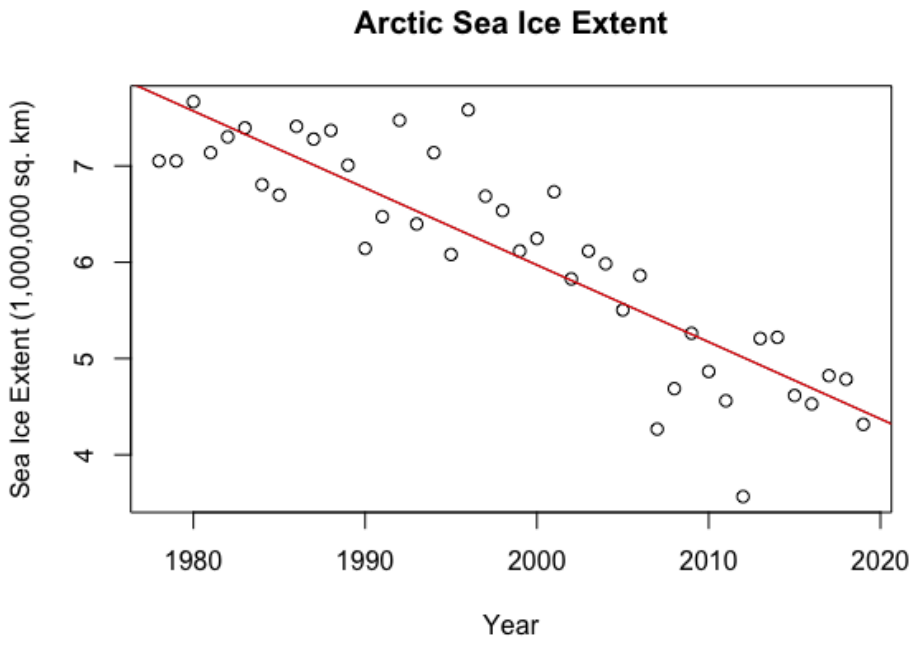

Figure 2: Applying the OLS simple linear regression to a set of data

Coefficient	Estimate	Std. Error	t-score	p-value
Intercept	166.0	13.6	12.24	4.23e-15
Slope	-0.0800	0.0068	-11.79	1.37e-14

Table 1: Simple Linear Regression R Statistics

sea ice extent measured in September, and plotted the measured value from every year over the last 40 years. [6]

The statistics in the model summary table were all generated using R Studio. The Intercept and Slope Estimates are, as you'd expect, the estimated regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively. This linear regression model used the OLS method to generate the regression coefficients. The standard error reported in the table is the standard error we discuss at the beginning of the Inferential Statistics in Simple Regression section. The given t-score and p-value test the hypothesis that the regression coefficients are zero. Of course it's no surprise that the intercept is nonzero because we do see and have seen ice in the arctic sea. More importantly The p-value of

the slope coefficient makes us very confident that the level of arctic sea ice is not remaining constant.

There is one last important quantity to report in this regression model: the **coefficient of determination**, denoted R^2 and often called "R-squared". This is an estimation of the correlation between Y and X. Hence R^2 can be between 0 and 1, with values close to 0 suggesting a poor correlation between X and Y, and a value close to 1 suggesting a strong correlation between Y and X. We calculate R-squared from the formula

$$R^2 = 1 - \frac{\text{RSS}}{m * \text{MSE}}. \tag{2.23}$$

For the arctic sea ice example $R^2 = 0.7709$.

The R^2 from the simple linear regression model is ok, but perhaps we can find a model that lowers the value sum of squared residuals. This will take us into our next section, where we'll discuss polynomial regression.

3 Other Regression Models

3.1 Polynomial Regression

In a simple linear regression model, we have one predictor and one outcome variable. In polynomial regression we again have one predictor and one outcome variable, x and Y respectively; however, in this model we take Y to be a linear combination of powers of x:

$$Y_i = \beta_0 X_i + \beta_1 X_i + \beta_2 X_i^2 + ... + \beta_{p-1} X_i^{p-1} + \varepsilon. \tag{3.1}$$

The first important property of polynomial regression to know is that the fit always gets better as we increase the order of the polynomial we are using. In fact if we have sampled the response variable n times, a polynomial of order $n-1$ can be constructed that passes through every point that we've sampled.

Secondly, we still have our OLS method of calculating the regression coefficients.

3.2 Inferential Statistics in Polynomial Regression

One of the means we have to compare models is the nested F-test. This is useful when we have one model which is a special case of the other. What this means is that the full model takes the form

$$Y = \beta_0 + \beta_1 f_1(X) + \beta_2 f_2(X) + \varepsilon \tag{3.2}$$

For some arbitrary functions f_1 and f_2 of X. The nested model takes the form

$$Y = \beta_0 + \beta_1 f_1(X) \tag{3.3}$$

Model 3.3 is the special case of 3.2 where $\beta_2 = 0$. However, it's important to note that comparing these two models is not as simple as hypothesis testing whether or not β_2 is 0. This is because the values of the other coefficients β_0 and β_1 might be (and probably will be) different when applying the full than when applying the nested model.

Let Model 1 indicate the nested model and Model 2 indicate the more complex model. The F ratio is the ratio of the percent change in residual sum of squares to the percent change in degrees of freedom [4]:

$$F = \frac{\frac{\mathrm{RSS}_1 - \mathrm{RSS}_2}{\mathrm{RSS}_2}}{\frac{(m-p_1)-(m-p_2)}{m-p_2}}. \tag{3.4}$$

Model 2 will always have a smaller residual sum of squares than Model 1, so $\mathrm{RSS}_2 < \mathrm{RSS}_1$. Furthermore, Model 2 will always have fewer degrees of freedom than Model 1, so $m - p_2 < m - p_1$. Therefore the F ratio is always positive. If F is less than one, than the weighted change in degrees of freedom is greater than the weighted change in RSS. If the number of degrees of freedom is changing a lot, yet the residual sum of squares is not, then the more complex model isn't very helpful. So if F is much less than 1, we'd choose the nested model. Conversely if F is much greater than 1 the residuals are getting smaller yet the degrees of freedom isn't changing much. This is evidence to add more parameters to the model. So if F is much greater than 1, we'd choose the more complex model.

3.3 Applying a polynomial Regression

We'll now apply 2nd order and 3rd order polynomial regressions to the set of arctic sea ice data we encountered in simple linear regression. I've shown the result of such a regression in Figures 3 and 4. The R summary statistics are given in Tables 2 and 3.

Furthermore we'll do a nested F-test to compare the simple linear model to the 2nd order polynomial model. I found that the F ratio between the simple linear model and the 2nd order polynomial is 4.82. This implies that the sum of squared errors changes much more than the degrees of freedom between the two models. This makes sense considering visually, the second order looks like a much better fit than the first order.

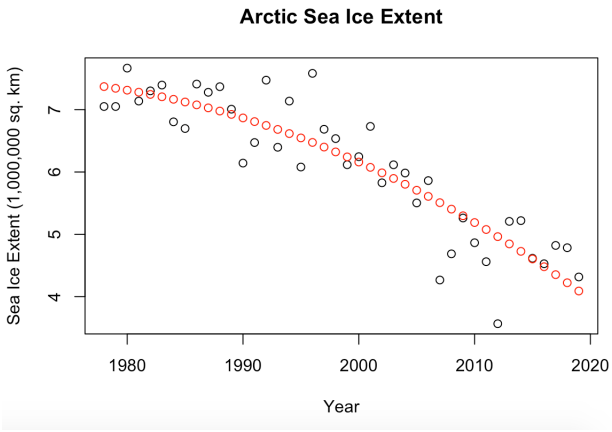

Figure 3: 2nd order polynomial regression applied to arctic sea ice levels

Coefficient	Estimate	Std. Error	t-score	p-value
β_0	-5.08e+03	2.391e+03	-2.126	0.0399
β_1	5.274e+00	2.393e+00	2.162	0.0368
β_2	-1.314e-03	5.987e-04	-2.195	0.0341

Table 2: 2nd Order Polynomial Regression R Statistics

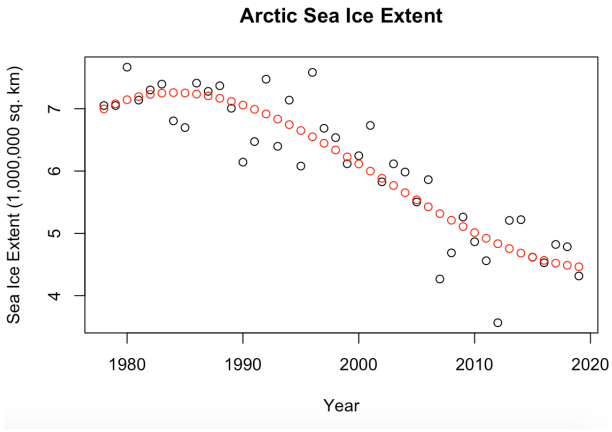

Figure 4: 3rd order polynomial regression applied to arctic sea ice levels

Coefficient	Estimate	Std. Error	t-score	p-value
β_0	-9.376e+05	4.299e+05	-2.181	0.0355
β_1	1.405e+03	6.454e+02	2.177	0.0358
β_2	-7.018e-01	3.229e-01	-2.173	0.0361
β_3	1.168e-04	5.386e-05	2.169	0.0364

Table 3: 3rd Order Polynomial Regression R Statistics

3.4 Multivariate Regression

In simple regression and polynomial regression we had one predictor and one response, now we'll move on to having multiple predictor variables. A multiple regression model with p parameters variables is written

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1}. \tag{3.5}$$

With many predictors in our regression model, it will be convenient to use a notation that captures everything in the model more succinctly. We'll do that using vectors and matrices. We'll have: a **response vector** $\boldsymbol{Y}$, which holds the value of the response at every trial; a **design matrix** $\boldsymbol{X}$ which holds the value of each predictor at each trial, rows will separate trials and columns will separate predictors; a **parameter vector** $\boldsymbol{\beta}$ that holds the coefficients; and an **error vector** with the error ε at each trial.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.6}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ 1 & X_{31} & X_{32} & \cdots & X_{3,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{m1} & X_{m2} & \cdots & X_{mp-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon \\ \varepsilon \\ \varepsilon \\ \vdots \\ \varepsilon \end{pmatrix} \tag{3.7}$$

Remark 3.4.1. Indexing predictor variables A potential point of confusion that arises because of how we define the design matrix is how we index our predictor variables. The fact that *rows* indicate trial number and *columns* indicate variable number, means if we were to write out the i^{th} trial we'd write it as

$$Y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

So the first subscript, i, indexes the trial number, but the second subscript indexes the predictor variable. This can be problematic because when we simply write out the predictor variable X_i, the subscript i indexes the predictor variable. So the convention is now that the first subscript indexes trial number and the second subscript indexes the parameter.

We can still use the method of least squares to find the regression coefficients in a multivariate linear regression model. The vector which contains our regression coefficient estimators is given by

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \tag{3.8}$$

Below we'll give a brief derivation of the OLS estimators:

$$\begin{aligned} \boldsymbol{e}^\top \boldsymbol{e} &= \begin{bmatrix} e_1 & e_2 & \dots & \dots & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{bmatrix} \\ &= e_1^2 + e_2^2 + ...e_n^2 \\ &= \sum e_i^2 \end{aligned} \tag{3.9}$$

Equation 3.9 tells us that we need to minimize the quantity $\boldsymbol{e}^\top \boldsymbol{e}$. [2] We'll use the same calculus approach that we applied in Appendix A. First we will need the identity that:

$$\begin{aligned} \boldsymbol{e}^\top \boldsymbol{e} &= (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &= \boldsymbol{Y}^\top \boldsymbol{Y} - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{Y}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{Y}^\top \boldsymbol{Y} - 2\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{Y} + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}} \end{aligned} \tag{3.10}$$

In Equation 3.10 we took advantage of the fact that $\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{Y} = (\boldsymbol{Y}^\top \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^\top)^\top = \boldsymbol{Y}^\top \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^\top$ The last identity is true because the quantity is a scalar.

$$\begin{aligned} &\frac{\partial \boldsymbol{e}^\top \boldsymbol{e}}{\partial \widehat{\boldsymbol{\beta}}} = -2\boldsymbol{X}^\top \boldsymbol{Y} + 2\boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}} = 0 \\ &\Rightarrow \boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{Y} \\ &\Rightarrow \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \end{aligned} \tag{3.11}$$

Which is the result we wanted. The derivation of the oLS estimators demonstrates that matrix notation really does make our lives easier. What took us three pages in Appendix A took only half a page here.

Remark 3.4.2. Full Rank: Its possible that the columns of our design matrix are linearly *dependent*, which is another way to say $\boldsymbol{X}$ does not have full rank. This causes a big issues. If $\boldsymbol{X}$ does not have full rank then $(\boldsymbol{X}^\top \boldsymbol{X})$ does not have full rank. Hence by the invertible matrix theorem $(\boldsymbol{X}^\top \boldsymbol{X})$ is not invertible. This is an issue because, as we just saw, we need to invert $(\boldsymbol{X}^\top \boldsymbol{X})$ to find the OLS regression coefficients. [2]

3.5 Inferential Statistics in Multiple Regression

The two theorems that I'll be proving below state that certain variables are t-distributed. Recall that the t-distribution is characterized by a standard normal distribution and a chi-squared distribution. Both proofs rely on producing a standard normal and chi-squared distribution in order to arrive at the result.

Lemma 3.1. If $x \sim \mathcal{N}(0, I)$ and A is symmetric and idempotent, then $x^T A x$ is distributed χ^2_ν where v is the rank of A. [2]

Theorem 3.2. Let $\hat{\boldsymbol{\beta}}$ be the OLS parameter vector. Given that $\varepsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}\right)$, then $\forall k \leq m$, $\frac{\hat{\beta}_k - \beta_k}{s(\hat{\beta}_k)} \sim t(m-p)$.

Proof. We begin with the definition of the OLS parameter vector from Equation 3.8:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \\
&= \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
&= \boldsymbol{\beta} + \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{\varepsilon}.
\end{aligned}$$

Therefore $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon}$. Hence by our assumption about the distribution of $\boldsymbol{\varepsilon}$

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right).$$

This implies that for a given trial k

$$\hat{\beta}_k - \beta \sim \mathcal{N}\left(0, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{kk}\right),$$

where $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{kk}$ is the k^{th} diagonal element of the matrix $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. Therefore we get the standard normal variable z_k:

$$z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{kk}}} \sim \mathcal{N}(0,1)$$

Now we introduce the residual maker matrix $\boldsymbol{M}$, a matrix defined such that $\boldsymbol{MY} = \boldsymbol{e}$, and the hat matrix $\boldsymbol{H} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top$. Hence

$$\boldsymbol{M} = \boldsymbol{I}_m - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top = \boldsymbol{I}_m - \boldsymbol{H}$$

.

We wish to show that $\boldsymbol{M}$ is both idempotent and symmetric. Beginning with idempotent:

$$\begin{aligned}
\boldsymbol{MM} &= \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top\right)\left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top\right) \\
&= \boldsymbol{I}^2 - 2\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top + \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X}\left(X^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \\
&= \boldsymbol{I} - 2\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top + \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \\
&= \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \\
&= \boldsymbol{M}
\end{aligned}$$

To show that $\boldsymbol{M}$ is symmetric we first show $\boldsymbol{H}$ is symmetric then use the sum of symmetric matrices is symmetric[5].

$$\begin{aligned}
\boldsymbol{H}^\top &= (\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top)^\top \\
&= (\boldsymbol{X}^\top)^\top ((\boldsymbol{X}^\top \boldsymbol{X})^{-1})^\top \boldsymbol{X}^\top \\
&= \boldsymbol{X}((\boldsymbol{X}^\top \boldsymbol{X})^\top)^{-1} \boldsymbol{X}^\top \\
&= \boldsymbol{X}(\boldsymbol{X}^\top (\boldsymbol{X}^\top)^\top)^{-1} \boldsymbol{X}^\top \\
&= \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \\
&= \boldsymbol{H}
\end{aligned}$$

Since the hat matrix $\boldsymbol{H}$ is symmetric and $\boldsymbol{M} = \boldsymbol{I}_m - \boldsymbol{H}$, we know $\boldsymbol{M}$ is symmetric. Now we know that the residual maker matrix is symmetric and idempotent. Therefore by our lemma the following quantity is distributed

as chi squared with $m - p$ degrees of freedom:

$$\left(\frac{\boldsymbol{\varepsilon}}{\sqrt{\sigma_\varepsilon^2}}\right)^\top \boldsymbol{M} \left(\frac{\boldsymbol{\varepsilon}}{\sqrt{\sigma_\varepsilon^2}}\right)$$

Hence we produce the following t-distribution:

$$t_k = \frac{z_k}{\sqrt{\left(\frac{\boldsymbol{\varepsilon}}{\sqrt{\sigma_\varepsilon^2}}\right)^\top \boldsymbol{M} \left(\frac{\boldsymbol{\varepsilon}}{\sqrt{\sigma_\varepsilon^2}}\right) / (n-p)}}$$

$$\begin{aligned} t_k &= \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma_\varepsilon^2 S_{kk}}}}{\sqrt{\frac{(m-p)\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}/(m-p)}} \\ &= \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{S_{kk}}}}{\sqrt{\hat{\sigma}_\varepsilon^2}} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}_\varepsilon^2 S_{kk}}} \\ &= \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}^2\left(\hat{\beta}_k\right)} \end{aligned}$$

Which is the result we wanted ∎

Theorem 3.2 gives us a way to write out confidence intervals of our regression coefficients.

$$\hat{\beta}_1 - [t\ \hat{\sigma}_\varepsilon^2(\hat{\beta}_1)] \leq \beta_1 \leq \hat{\beta}_1 + [t\ \hat{\sigma}_\varepsilon^2(\hat{\beta}_1)] \tag{3.12}$$

The next theorem will be important when predicting the outcome of future samples of our response variable Y

Appendix A Simle Linear OLS coefficients

In this appendix we'll derive the least squares regression coefficients. With this done we'll then use algebra to show that the least squares coefficients are linear, and then we'll prove that they are also unbiased estimators of the true regression coefficients β_0 and β_1.

A.1 Deriving OLS Coefficients

Remember that with the least squares model we're trying to minimize the sum below

$$\text{RSS} = \sum_{i=1}^{N} e_i^2 \tag{A.1}$$

To do this we'll take the calculus approach of finding a critical point.
For $\hat{\beta}_0$.

$$\begin{aligned}
\frac{\partial \text{RSS}}{\partial \beta_0} &= -2\sum_{i=1}^{N}(Y_i - \beta_0 - \beta_i X_i) = 0 \\
&\Rightarrow \sum_{i=1}^{N}(Y_i - \beta_0 - \beta_i X_i) = 0 \\
&\Rightarrow \sum_{i=1}^{N} Y_i - \sum_{i=1}^{N}\beta_0 - \sum_{i=1}^{N} X_i = 0 \\
&\Rightarrow \ N\,\beta_0 = \sum_{i=1}^{N} Y_i - \beta_1 \sum_{i=1}^{N} X_i \\
&\Rightarrow \beta_0 = \frac{1}{m}(\sum_{i=1}^{N} Y_i - \beta_1 \sum_{i=1}^{N} X_i) = \overline{Y} - \beta_1 \overline{X}
\end{aligned}$$

Thus

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Identities A.2 and A.3 are needed for the derivation of $\hat{\beta}_1$.

$$\begin{aligned}
\sum_i \left(X_i - \overline{X}\right)^2 &= \sum_i X_i^2 - \sum_i X_i \overline{X} - \sum_i \overline{X} X_i + \sum_i \overline{X}^2 \\
&= \sum_i X_i^2 - 2\overline{X} \sum_i X_i + m\overline{X}^2 \\
&= \sum_i X_i^2 - 2\overline{X} * m\overline{X} + m\overline{X}^2 \\
&= \sum_i X_i^2 - m\overline{X}^2 \\
&= \sum_i X_i^2 - \overline{X} \sum_i X_i
\end{aligned} \tag{A.2}$$

$$\begin{aligned}
\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) &= \sum_i X_i * Y_i - \sum_i X_i * \overline{Y} - \sum_i \overline{X} * Y_i + \sum_i \overline{XY} \\
&= \sum_i X_i * Y_i - m\overline{XY} - m\overline{XY} + m\overline{XY} \\
&= \sum_i X_i * Y_i - m\overline{XY} \\
&= \sum_i X_i * Y_i - \overline{Y} \sum_i X_i
\end{aligned} \tag{A.3}$$

Now we'll derive $\hat{\beta}_1$ by the same calculus approach we used with $\hat{\beta}_0$.

$$
\begin{aligned}
&\frac{\partial \text{RSS}}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_i X_i) = 0 \\
\Rightarrow \quad &\sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i) = 0 \\
\Rightarrow \quad &\sum_{i=1}^{n} X_i(Y_i - (\overline{Y} - \beta_1\overline{X}) - \beta_1 X_i) = 0 \\
\Rightarrow \quad &\sum_{i=1}^{n} X_i Y_i - X_i\overline{Y} + X_i\beta_1\overline{X} - \beta_1 X_i^2 = 0 \\
\Rightarrow \quad &\sum_{i=1}^{n} X_i Y_i - \overline{Y}\sum_{i=1}^{n} X_i + \beta_1\overline{X}\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \beta_1 X_i^2 = 0 \\
\Rightarrow \quad &\beta_1\overline{X}\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \beta_1 X_i^2 = \overline{Y}\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i Y_i \\
\Rightarrow \quad &\hat{\beta}_1 = \frac{\overline{Y}\sum X_i - \sum X_i Y_i}{\overline{X}\sum X_i - \sum X_i^2} \\
\Rightarrow \quad &\beta_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}
\end{aligned}
\tag{A.4}
$$

A.2 In Linear Form

Now we'll show that $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear estimators. That's to say that each estimator is a linear combination of the observed Y_i; or equivalently, for all i there exist constants k_i, j_i such that

$$
\begin{aligned}
\hat{\beta}_0 &= \sum j_i Y_i \\
\hat{\beta}_1 &= \sum k_i Y_i
\end{aligned}
\tag{A.5}
$$

In Derivation A.6 we'll need to use the identity that $\sum X_i - \overline{X} = 0$. Now for $\hat{\beta}_1$

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2} \\
&= \frac{\sum(X_i - \overline{X})Y_i - \sum(X_i - \overline{X})\overline{Y}}{\sum(X_i - \overline{X})^2} \\
&= \frac{\sum(X_i - \overline{X})Y_i - \overline{Y}\sum(X_i - \overline{X})}{\sum(X_i - \overline{X})^2} \\
&= \frac{\sum(X_i - \overline{X})Y_i}{\sum(X_i - \overline{X})^2} \\
&= \sum \frac{(X_i - \overline{X})Y_i}{\sum(X_i - \overline{X})^2} \\
&= \sum k_i Y_i
\end{aligned} \tag{A.6}$$

Hence

$$k_i = \frac{(X_i - \overline{X})}{\sum(X_i - \overline{X})^2} \tag{A.7}$$

Now for $\hat{\beta}_0$

$$\begin{aligned}
\hat{\beta}_0 &= \frac{1}{m}(\sum Y_i - \hat{\beta}_1 \sum X_i) \\
&= \frac{1}{m}(\sum Y_i - \sum(k_i Y_i) \sum X_i) \\
&= \frac{1}{m}(\sum Y_i - \sum(k_i Y_i \sum X_i)) \\
&= \frac{1}{m}(\sum Y_i(1 - k_i \sum X_i)) \\
&= \sum(\frac{1}{m} - k_i \overline{X})Y_i \\
&= \sum j_i Y_i
\end{aligned} \tag{A.8}$$

Hence

$$j_i = \frac{1}{m} - k_i \overline{X} = \frac{1}{m} - \frac{(X_i - \overline{X})\overline{X}}{\sum(X_i - X)^2} \tag{A.9}$$

A.3 Unbiased

We wish to show the least squares regression coefficients are unbiased, that is

$$\mathrm{E}[\hat{\beta}_0] = \beta_0, \quad \mathrm{E}[\hat{\beta}_1] = \beta_1 \tag{A.10}$$

When handling $\hat{\beta}_1$, we'll need to use that $\sum(X_i - \overline{X}) = 0$ for the OLS estimator. Now for $\hat{\beta}_1$:

$$\begin{aligned}
\mathrm{E}[\hat{\beta}_1] =& E\Big[\frac{\sum(X_i - \overline{X})Y_i}{\sum(X_i - \overline{X})X_i}\Big] \\
=& \frac{\sum[(X_i - \overline{X})\mathrm{E}[Y_i]]}{\sum(X_i - \overline{X})X_i} \\
=& \frac{\sum[(X_i - \overline{X})(\beta_0 + \beta_1 X_i)]}{\sum(X_i - \overline{X})X_i} \\
=& \frac{\sum[\beta_0(X_i - \overline{X}) + \beta_1 X_i(X_i - \overline{X})]}{\sum(X_i - \overline{X})X_i} \\
=& \frac{\sum[\beta_0(X_i - \overline{X})] + \sum[\beta_1 X_i(X_i - \overline{X})]}{\sum(X_i - \overline{X})X_i} \\
=& \frac{\sum \beta_0(X_i - \overline{X})}{\sum(X_i - \overline{X})X_i} + \frac{\sum \beta_1 X_i(X_i - \overline{X})}{\sum(X_i - \overline{X})X_i} \\
=& \beta_0 \frac{\sum(X_i - \overline{X})}{\sum(X_i - \overline{X})X_i} + \beta_1 \frac{\sum X_i(X_i - \overline{X})}{\sum(X_i - \overline{X})X_i} \\
=& 0 + \beta_1 \\
=& \beta_1
\end{aligned} \tag{A.11}$$

When handling $\hat{\beta}_0$ we'll first need the identity in Equation A.12

$$\begin{aligned}
\mathrm{E}[\overline{Y}] &= \mathrm{E}[\frac{1}{m}\sum Y_i] \\
&= \frac{1}{m}\mathrm{E}[\sum Y_i] \\
&= \frac{1}{m}\sum \mathrm{E}[Y_i] \\
&= \frac{1}{m}\sum[\beta_0 + \beta_1 X_i] \\
&= \frac{1}{m}\sum \beta_0 + \frac{1}{m}\sum \beta_1 X_i \\
&= \frac{1}{m} n\beta_0 + \beta_1 \frac{\sum X_i}{n} \\
&= \beta_0 + \beta_1 \overline{X}
\end{aligned} \tag{A.12}$$

Now for $\hat{\beta}_0$

$$\begin{aligned} \mathrm{E}[\hat{\beta}_0] &= \mathrm{E}[\overline{Y} - \hat{\beta}_1\overline{X}] = \mathrm{E}[\overline{Y}] - \mathrm{E}[\hat{\beta}_1\overline{X}] \\ &= \beta_0 + \beta_1\overline{X} - \mathrm{E}[\hat{\beta}_1]\mathrm{E}[\overline{X}] = \beta_0 + \beta_1\overline{X} - \beta_1\overline{X} \\ &= \beta_0 \end{aligned} \tag{A.13}$$

Having verified Equation A.10 we conclude that the OLS estimator is unbiased.

A.4 Linear Coefficients Squares

Equations A.14 and A.15 are both needed to show that $\hat{\beta}_0$ is linear.

$$\begin{aligned} \sum k_i &= \sum \frac{X_i - \overline{X}}{\sum(X_i - \overline{X})^2} \\ &= \frac{1}{\sum(X_i - \overline{X})^2} \sum(X_i - \overline{X}) = \frac{0}{\sum(X_i - \overline{X})^2} = 0 \end{aligned} \tag{A.14}$$

$$\begin{aligned} \sum k_i^2 &= \Big[\frac{X_i - \overline{X}}{\sum(X_i - \overline{X}}\Big]^2 \\ &= \frac{1}{[\sum(X_i - \overline{X})^2]^2} \sum(X_i - \overline{X})^2 = \frac{1}{\sum(X_i - \overline{X})^2} \end{aligned} \tag{A.15}$$

References

[1] Kutner, Nachtsheim, Neter. Applied Linear Regression Models, 4th ed. McGraw-Hill Inc., 2004

[2] Seber. Linear Regression Analysis. John Wiley & Sons, Inc., 1977.

[3] B. W. Lindgren. Statistical Theory, 2nd ed. Collier Macmillian Limited., 1968

[4] Koopmans. Intriductio to Contemporary Statistical Methods, 2nd ed. Duxbury Press, 1987

[5] Curtis. Linear Algebra: an Introductory Approach. Spring-Verlag, 1986

[6] Journal of Statistics Education, Volume 21, Number 1 (2013), http://jse.amstat.org/v21n1/witt.pdf