Linear Regression

Andrew Andreas

August 15, 2024

1 Introduction

Please report any errors to andrew.andreas0@gmail.com

2 Simple Linear Regression

The simple linear regression model is a linear model of a single explanatory variable x_i and is given by

$$Y = \alpha + \beta x + W$$

meaning that the dependent variable Y depends on the observed value $X = x$ through three unknown parameters, α, β and σ^2, which is the variance of $W \sim \mathcal{N}(0, \sigma^2)$

Y is just a linear combination of constants (α and β) and a normal random variable, W, and hence conditionally also follows a normal distribution

$$Y|X \sim \mathcal{N}(\alpha + \beta x, \sigma^2)$$

2.1 Model Assumptions

Constructing the simple linear regression model this way gives rise to some key assumptions we have made about the data

- **Homoskedasticity**
  - Which means that we have constant error regardless of the predictor value because σ^2 is a constant
- **Conditional Independence**
  - We assume that $W_i, i = 1, ..., n$ are independent and thus $Y_i|X = x_i$ are conditionally independent

These assumptions do not always hold in practice and there are various ways we can deal with this

2.2 Estimating the Coefficients

There are two common approaches to estimating the parameters of the linear regression model - the *Least Squares* criterion (of which there are multiple) and via *Maximum Likelihood Estimation*. Under the assumptions of normality of the error term W, we arrive at the same estimate for β, which we denote $\widehat{\beta}$ and for α, denoted $\widehat{\alpha}$ under OLS and MLE.

2.2.1 Ordinary Least Squares

Under the least squares criterion we look to minimize the sum of squared differences between the actual data points y_i and their estimate $\widehat{y_i}$

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{\alpha} + \widehat{\beta} x_i)^2 \end{aligned}$$

We can visualise this below as trying to find the estimates of the slope and intercept of the fitted line that minimizes the sum of squared differences between the line and the actual data points.

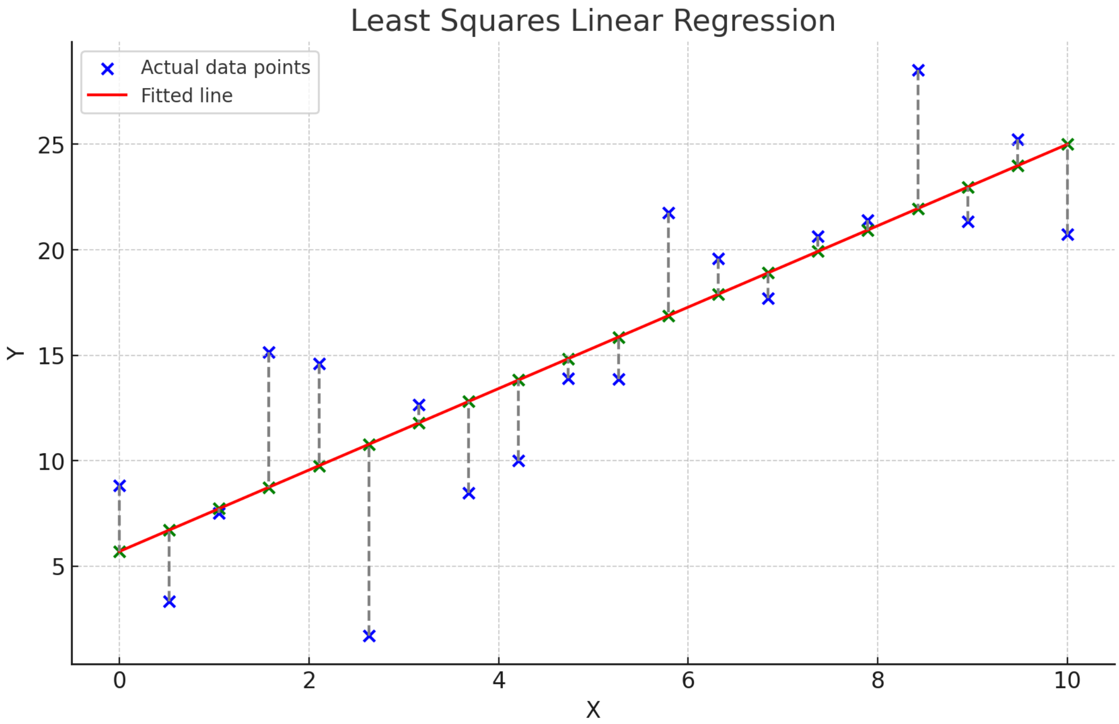

Figure 1: Plot of the least squares regression line

Finding the values of $\widehat{\alpha}$ and $\widehat{\beta}$ which minimize RSS can be done through calculus and yields the following estimators

$$\begin{aligned}\widehat{\alpha} =& \overline{y} - \widehat{\beta}\overline{x} \\ \widehat{\beta} =& \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}\end{aligned}$$

As mentioned previously the maximum likelihood estimates of β and α are the same as under the least squares criterion due to the assumption that the error terms are normally distributed. As such, we won't derive these here but this can be done quite easily by working with the log-likelihood and some calculus.

The estimate of the population variance σ^2 is given by S^2 which is the sample variance. It benefits from being an unbiased estimator of the true population variance, σ^2.

$$S^2 = \frac{1}{n-p}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

where p represents the number of predictors (or degrees of freedom). In this case $p = 2$ since we have predictors $\widehat{\alpha}$ and $\widehat{\beta}$.

We have all that is required to produce the *Fitted Values*, which is just a term to describe $\widehat{y_i} = \widehat{\alpha} + \widehat{\beta}x_i$ and can alternatively be expressed as

$$\begin{aligned}\widehat{y_i} =& \widehat{\alpha} + \widehat{\beta}x_i \\ =& \overline{y} - \widehat{\beta}\overline{x} + \widehat{\beta}x_i \\ =& \overline{y} - \widehat{\beta}(x_i - \overline{x})\end{aligned}$$

The fitted values are what define our regression line as seen in Figure 1

2.3 Sampling Distributions

To get a sense of uncertainty in the classical sense we must consider the sampling distributions of our estimators. To do so we exploit the fact that i) we can write the estimators $\widehat{\alpha}$ and $\widehat{\beta}$ as linear combinations of the random variable Y_j and ii) that linear combinations of normal random variables are also normally distributed (since Y_j is conditionally normal).

Let $Z = \sum_{j=1}^{n} k_j Y_j$, then the mean is given by

$$\begin{aligned}\mathbb{E}(Z) =&\mathbb{E}\big[\sum_{j=1}^{n} k_j Y_j\big]\\ =&\mathbb{E}\big[\sum_{j=1}^{n} k_j(\alpha+\beta x_j)\big]\\ =&\mathbb{E}\big[\sum_{j=1}^{n} k_j\alpha\big]+\mathbb{E}\big[\sum_{j=1}^{n} k_j\beta x_j)\big]\\ =&\alpha\sum_{j=1}^{n} k_j+\beta\sum_{j=1}^{n} k_j x_j\end{aligned}$$

and the variance is given by

$$\begin{aligned}\mathbb{V}(Z) =&\mathbb{V}\big[\sum_{j=1}^{n} k_j Y_j\big]\\ =&\sum_{j=1}^{n} k_j\mathbb{V}(Y_j)\\ =&\sigma^2\sum_{j=1}^{n} k_j^2\end{aligned}$$

Which tells us that if we can describe our estimators as linear combinations of Y then we only need find three terms $\sum_{j=1}^{n} k_j^2$, $\sum_{j=1}^{n} k_j$ and $\sum_{j=1}^{n} k_j x_j$

As we have stated before, the linear regression model assumes the error terms are uncorrelated which implies conditional independence between $Y_j, j = 1, ..., n$. The importance of this plays out in the variance where we just have sums of individual variances σ^2. If however, we have correlated errors, then the true variance will contain covariance terms $\mathrm{Cov}(Y_j, Y_i), j \neq i$ which aren't accounted for in our model. This means our model is underestimating the variance in our data and cna lead to two key problems:

- Confidence intervals that are narrower than they should be
- p-values which are lower than they should be which may lead to incorrectly conclude a dependent variable is statistically significant

It will be useful to consider the covariance between the estimators, particularly when we look at the sampling distribution of the residuals. Since we can express these estimators as linear combinations of the target variable then

$$\begin{aligned}\mathrm{Cov}(\widehat{\alpha},\widehat{\beta}) =&\mathrm{Cov}\big(\sum_{j=1}^{n} k_j Y_j,\sum_{j=1}^{n} n_j Y_j\big)\\ =&\sum_{j=1}^{n} k_j n_j\mathbb{V}(Y_j)\\ =&\sum_{j=1}^{n} k_j n_j\sigma^2\end{aligned}$$

2.3.1 Sampling Distribution for Slope Term

For $\widehat{\beta}$,

$$k_j = \frac{(x_j - \overline{x})}{\sum_{j=1}^{n}(x_j - \overline{x})^2}$$

and so the sampling distribution is

$$\widehat{\beta} \sim \mathcal{N}\Big(\beta, \frac{\sigma^2}{\sum_{j=1}^n (x_j - \overline{x})^2}\Big)$$

2.3.2 Sampling Distribution for the Intercept

For $\widehat{\alpha}$,

$$l_j = \frac{1}{n} - \overline{x} k_j$$

which yields a sampling distribution of

$$\widehat{\alpha} \sim \mathcal{N}\Big(\alpha, \sigma^2 (\frac{1}{n} + \frac{\overline{x}^2}{\sum_{j=1}^n (x_j - \overline{x})^2})\Big)$$

2.3.3 Sampling Distribution for the Fitted Values

The ith fitted value is given by $\widehat{y_i} = \widehat{\alpha} + \widehat{\beta} x_i$ and due to the linearity of the expectation operator

$$\begin{aligned}\mathbb{E}(\widehat{Y_i}) =& \mathbb{E}(\widehat{\alpha} + \widehat{\beta} x_i) \\ =& \alpha + \beta x_i\end{aligned}$$

The variance of the ith fitted value is

$$\begin{aligned}\mathbb{V}(\widehat{Y_i}) =& \mathbb{V}(\widehat{\alpha} + \widehat{\beta} x_i) \\ =& \mathbb{V}(\widehat{\alpha}) + 2x_i \mathrm{Cov}(\widehat{\alpha}, \widehat{\beta}) + x_i^2 \mathbb{V}(\widehat{\beta}) \\ =& \sigma^2 \Big(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{j=1}^n (x_j - \overline{x})^2}\Big) - 2x_i \frac{\sigma^2 \overline{x}}{\sum_{j=1}^n (x_j - \overline{x})^2} + x_i^2 \frac{\sigma^2}{\sum_{j=1}^n (x_j - \overline{x})^2} \\ =& \sigma^2 \Big(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{s_{xx}}\Big)\end{aligned}$$

In the third line of the variance derivation we computed the covariance using the linear combination values we defined earlier and also introduced the standard notation s_{xx} for the numerator

$$\mathrm{Cov}(\widehat{\alpha}, \widehat{\beta}) = \sigma^2 \sum_{j=1}^n l_j k_j \tag{1}$$

Therefore the sampling distribution for the fitted values is

$$\widehat{Y_i} | X_i \sim \mathcal{N}\Big(\alpha + \beta x_i, \sigma^2 \Big(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{s_{xx}}\Big)\Big)$$

2.3.4 Sampling Distribution for the residuals

$$\begin{aligned}r_i =& Y_i - \widehat{Y_i} \\ =& Y_i - \overline{Y} + \widehat{\beta}(x_i - \overline{x})\end{aligned}$$

Let $p_i = Y_i - \overline{Y}$ where we can write this as a linear combination of the random variable Y_j as before with

$$c_j = \delta_{ij} - \frac{1}{n}$$

so that

$$\begin{aligned}Z =& \sum_{i=1}^n (\delta_{ij} - \frac{1}{n}) Y_j \\ =& \sum_{i=1}^n \delta_{ij} Y_j - \frac{1}{n} \sum_{i=1}^n Y_j \\ =& Y_i - \overline{Y}\end{aligned}$$

This has a normal distribution which follows

$$Y_i - \widehat{Y} \sim \mathcal{N}\Big(\beta(x_i - \overline{x}), \sigma^2(1 - \frac{1}{n})\Big)$$

The other term in the residual is just $\widehat{\beta}(x_i - \overline{x})$ which is just a constant $(x_i - \overline{x})$ multiplied by a normal random variable in $\widehat{\beta}$ and thus is a linear transformation of a normal random variable and is itself normal. It has the following distribution

$$\widehat{\beta}(x_i - \overline{x}) \sim \mathcal{N}\Big(\beta(x_i - \overline{x}), \frac{\sigma^2(x_i - \overline{x})^2}{s_{xx}}\Big)$$

The distribution of the residuals then also follows a normal distribution since it is a sum of normally distributed random variables with $\mathbb{E}(r_i) = 0$. The covariance term can be computed by again exploiting the same trick we did for the sampling distribution of the fitted values by substituting for p_i and $\widehat{\beta}(x_i - \overline{x})$ their linear combinations

$$\begin{aligned}
\text{Cov}(p_i, \widehat{\beta}(x_i - \overline{x})) =&(x_i - \overline{x})^2\text{Cov}\Big((x_i - \overline{x})\sum_{j=1}^{n} c_j Y_j, \sum_{j=1}^{n} k_j Y_j\Big) \\
=&(x_i - \overline{x})\text{Cov}\Big(\sum_{j=1}^{n} c_j Y_j, \sum_{i=1}^{n} k_j Y_j\Big) \\
=&\sigma^2(x_i - \overline{x})\sum_{j=1}^{n} c_j k_j \\
=&\sigma^2(x_i - \overline{x})\sum_{j=1}^{n}\big(\delta_{ij} - \frac{1}{n}\big)\frac{(x_j - \overline{x})}{s_{xx}} \\
=&\frac{\sigma^2(x_i - \overline{x})^2}{s_{xx}}
\end{aligned}$$

So the variance term is then given by

$$\begin{aligned}
\mathbb{V}(r_i) =&\mathbb{V}(p_i) - 2\text{Cov}(\widehat{\alpha}, \widehat{\beta}) + \mathbb{V}(\widehat{\beta}) \\
=&\sigma^2\Big(1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{s_{xx}}\Big)
\end{aligned}$$

So the distribution of a single residual r_i is given by

$$r_i \sim \mathcal{N}\Big(0, \sigma^2\Big(1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{s_{xx}}\Big)\Big)$$

meaning that the expected value of the residuals is zero, which is what we would hope. We can also see this by noticing the expected fitted value is an unbiased estimator of the *conditional mean* of $Y|X$, though not of the observed values of $Y|X$ since this includes the error term W.

It is also the case that the residuals are correlated

$$\text{Cov}(r_i, r_j) = \sigma^2\Big(- \frac{1}{n} - \frac{(x_i - \overline{x})(x_j - \overline{x})}{s_{xx}}\Big)\Big)$$

Despite the fact that $\mathbb{V}(r_i) \neq \sigma^2$ and that the residuals are correlated and depend on x_i, it can be shown that as $n \to \infty, s_{xx} \to \infty$ and $1/n \to 0$, so $\mathbb{V}(r_i) \to \sigma^2$

2.3.5 Sampling Distribution for the Sample Variance

The OLS estimator for the variance, σ^2, in a simple linear regression model is

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

which is distributed as

$$S^2 \sim \frac{\sigma^2}{n-2}\chi^2(n-2)$$

and is also independent of the estimators of the coefficients.

Note: Since we rarely know the model variance σ^2, we use S^2 as an estimator for it and in practice if we are concerned with inference for any of the quantities in the sampling distributions we will replace σ^2 with S^2

2.4 Inference

Inference for parameters or values of interest in the (classical) linear regression model involves the following two-step methodology:

- Construction of an appropriate test statistic
- Construction of confidence intervals to assess the uncertainty in our estimates

Our desired outcome is to produce a test statistic, T, which follows the t-distribution and then proceed to construct a $100\%(1-\alpha)$ confidence interval. By definition, a random variable $Y = Z/\sqrt{V/v}$ is said to be t-distributed with v degrees of freedom, where Z is a standard normal random variable and V follows the χ^2 distribution with v degrees of freedom

$$Y = \frac{\sqrt{v}Z}{\sqrt{V}}, \quad \text{where } Z \sim \mathcal{N}(0,1) \text{ and } V \sim \chi^2(v)$$

Since all of the variables of interest in the sampling distributions are normally distributed we transform them into standard normal random variables by standardising them

$$Z = \frac{\widehat{\mu}-\mu}{\sigma}, \ Z \sim \mathcal{N}(0,1)$$

Then we can define another random variable, U such that

$$U = \frac{(n-2)S^2}{\sigma^2}, \ U \sim \chi^2(n-2)$$

So that

$$\begin{aligned} T =& \frac{Z}{\sqrt{U/(n-2)}} \\ =& \frac{\sqrt{n-2}(\widehat{\mu}-\mu)}{\sigma} \times \frac{\sigma}{\sqrt{n-2}S} \\ =& \frac{\widehat{\mu}-\mu}{S} \end{aligned}$$

Therefore by definition we have that $T \sim t(n-2)$

A quick note on notation

$$t_\alpha = F^{-1}(\alpha) = x_\alpha$$

this represents the α quantile (or inverse distribution function) for the $t(n-2)$ distribution.

Having created our test statistic T, we can construct a confidence interval of size α to quantify our uncertainty in our regression model. By exploiting the symmetric nature of the t-distribution about 0 (specifically that the $t_\alpha = -t_{1-\alpha}$), we have that

$$P\big(\widehat{\mu} - St_{1-\alpha/2} < \mu < \widehat{\mu} + St_{1-\alpha/2}\big)$$

Note that S in the above generic formula for the confidence interval of size α represents the standard error of the estimator and neglects any of the additional terms in standard deviation. Specifically, for the regression coefficients the confidence intervals are given by

$$\left(\widehat{\alpha} - t_{1-p/2}s\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}}, \widehat{\alpha} + t_{1-p/2}s\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}}\right)$$

and

$$\left(\widehat{\beta} - t_{1-p/2}\frac{s}{\sqrt{S_{xx}}}, \widehat{\beta} + t_{1-p/2}\frac{s}{\sqrt{S_{xx}}}\right)$$

I have another set of notes on the Hypothesis Testing and Confidence Intervals which discuss some more of the details here.

2.4.1 Predictive Inference

Another desirable task besides estimating parameters is to make predictions of future values, which is termed predictive inference and is an example of out-of-sample prediction. Since the model is trained on data $\{(x_1, y_1), ..., (x_n, y_n)\}$ and we are trying to make a prediction of the value Y_0, we are making predictions of a data point that our model has not seen. The simple linear regression model is of the form

$$Y_0 = \alpha + \beta x_0 + W_0$$

α and β are unknown parameters but we have found estimates for these in the form of $\widehat{\alpha}$ and $\widehat{\beta}$ so that

$$\widehat{Y_0} = \widehat{\alpha} + \widehat{\beta} x_0$$

While we shall see that the estimated value for Y_0 is $\widehat{\alpha} + \widehat{\beta} x_0$, the process of inference differs. In its definition, Y_0 has an error term W_0 which provides additional uncertainty since it is independent of $Y_1, ..., Y_n$ and this should be accounted for.

If we define a new random variable $P = \widehat{\alpha} + \widehat{\beta} x_0 - Y_0$, which is itself normally distributed since both $\widehat{Y_0}$ and Y_0 are, it follows the normal distribution (note we have replaced σ^2 with the sample variance S^2

$$P \sim \mathcal{N}\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)\right)$$

We can then construct a test statistic in the same way which follows the $t(n-2)$ distribution

$$Z = \frac{P - 0}{\sigma\sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)}} \sim \mathcal{N}(0, 1)$$

and

$$U = \frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2)$$

So we can create a random variable which follows the t-distribution, as we did previously, defined as

$$\begin{aligned} T =& \frac{Z}{\sqrt{U}/\sqrt{n-2}} \\ =& \frac{P}{\sigma k} \times \frac{\sigma\sqrt{n-2}}{\sqrt{n-2}S}, \text{ where } k = \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right) \\ =& \frac{P}{kS} \\ =& \frac{\widehat{\alpha} + \widehat{\beta} x_0 - Y_0}{kS} \end{aligned}$$

Then we can create a prediction interval of size p which takes the form

$$\left(\widehat{\alpha} + \widehat{\beta} x_0 - t_{1-p/2} s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}, \widehat{\alpha} + \widehat{\beta} x_0 + t_{1-p/2} s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}\right)$$

Note that this interval is wider than the interval for the fitted values since we have the additional uncertainty of W_0.

3 Multiple Linear Regression

The linear regression model is an additive linear model for a set of input-output pairs $(y_i, \mathbf{x}_i)$ where $y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^p$.

$$\begin{aligned} y_i =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \\ y_i =& \sum_{j=1}^{p} \beta_j X_{i,j} + \epsilon_i \end{aligned}$$

For the n observations in the dataset we can rewrite this in matrix form for simplicity

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a matrix of the n observations, with each observation $\mathbf{x}_i$ containing $(p-1)$ independent variables with a 1 in the first position to account for the β_0 term. $\boldsymbol{\beta}$ is the p-dimensional vector of the regression coefficients and $\mathbf{W}$ is a n-dimensional vector of the independent error terms, ϵ_i.

Since we do not know the true population regression coefficients we will need to estimate them from the observed data. These estimates are denoted $\widehat{\boldsymbol{\beta}}$ to distinguish it as an estimator. Therefore our regression model is given by

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

3.1 Ordinary Least Squares

Then we can define the error in the predictions as $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$ and then taking the squared error

$$\begin{aligned} \mathbf{e}^T\mathbf{e} =& \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^T \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \\ =& \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} \end{aligned}$$

If we want to find an expression for $\widehat{\boldsymbol{\beta}}$, which are the values of $\boldsymbol{\beta}$ that minimize the squared error, we can take the derivative of this expression and equate this to zero

$$\begin{aligned} \mathbf{0} =& -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} \\ \mathbf{X}^T\mathbf{y} =& \mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\beta}} =& \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\ =& \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{w}) \\ =& \boldsymbol{\beta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{w} \end{aligned}$$

provided that the inverse matrix $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ exists.

3.2 Maximum Likelihood

Under the assumption that the irreducible error terms are independent and follow a normal distribution, the OLS solution is the same as the MLE solution to the linear regression model. We begin with the same model formulation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}, \quad \mathbf{w} \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{\Sigma})$$

Then we can see that conditionally, $\mathbf{y}$ follows a multivariate normal distribution as well

$$\mathbf{y}|\mathbf{X} \sim \boldsymbol{\mathcal{N}}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

with the joint density function given by

$$\mathbf{f}(\mathbf{y}|\mathbf{X}) = (2\pi)^{-n/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\Big(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\Big)$$

Using the assumption that the data are independent and identically distributed means we can rewrite the multivariate normal and its joint density function as a product of the univariate normal distributions

$$
\begin{aligned}
\mathbf{y}|\mathbf{X} \sim & \boldsymbol{\mathcal{N}}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \\
\mathbf{f}(\mathbf{y}|\mathbf{X}) = & \prod_{i=1}^{n} f(y_i|\mathbf{x}_i)
\end{aligned}
$$

Therefore the log-likelihood function is

$$
\begin{aligned}
\boldsymbol{\mathcal{L}}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}) = & \sum_{i=1}^{n} f(y_i|\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2) \\
= & -\frac{n}{2}\ln(2\pi) - n\ln\left(\sigma\right) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})
\end{aligned}
$$

3.2.1 Estimating β

We seek to maximise the log-likelihood function with respect to $\boldsymbol{\beta}$, then this is equivalent to minimizing

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Then taking derivatives with respect to $\boldsymbol{\beta}$, which follows the maximization process we did previously yields the same solution

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

3.2.2 Estimating σ^2

Then maximizing with respect to σ^2 yields

$$
\begin{aligned}
\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}) = & \frac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\
= & \frac{\mathbf{e}^T\mathbf{e}}{n}
\end{aligned}
$$

though note this is the MLE estimate which is a biased estimate. Instead we more often use the unbiased estimate akin to the simple linear regression case where

$$
\begin{aligned}
S^2 = & \frac{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}{n-p} \\
= & \frac{\mathbf{e}^T\mathbf{e}}{n-p}
\end{aligned}
$$

More detail on this is given below when we consider the sampling distribution of σ^2

3.3 Sampling Distributions for the Parameters

3.3.1 Sampling Distribution for the Regression Coefficients

Recall that we can write $\widehat{\boldsymbol{\beta}}$ as

$$\boldsymbol{\beta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{w}$$

The sampling distribution of $\widehat{\boldsymbol{\beta}}$ is therefore a normal distribution

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

3.3.2 Sampling Distribution of the Residuals

Understanding the distribution of the residuals is useful when it comes to estimating the variance. The residuals have been defined above as $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$ with mean

$$\begin{aligned}\mathbb{E}(\mathbf{e}) =& \mathbb{E}(\mathbf{y} - \widehat{\mathbf{y}})\\ =& \mathbb{E}(\mathbf{y}) - \mathbb{E}(\widehat{\mathbf{y}})\\ =& 0\end{aligned}$$

We then define $\mathbf{H} = \mathbf{X}(\mathbf{X^T X})^{\mathbf{-1}}\mathbf{X^T}$ which is a projection matrix such that $\mathbf{HH} = \mathbf{H}$.

$$\begin{aligned}\mathbf{e} =& \mathbf{y} - \widehat{\mathbf{y}}\\ =& \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\\ =& \mathbf{y} - \mathbf{X}\big(\mathbf{X}^T\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{y}\\ =& \mathbf{y} - \mathbf{H}\mathbf{y}\\ =& (\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}$$

Then the residuals have variance given by

$$\begin{aligned}\mathbb{V}(\mathbf{e}) =& \mathbb{V}\big((\mathbf{I} - \mathbf{H})\mathbf{y}\big)\\ =& (\mathbf{I} - \mathbf{H})^T\mathbb{V}(\mathbf{y})(\mathbf{I} - \mathbf{H})\\ =& \sigma^2(\mathbf{I} - \mathbf{H})\end{aligned}$$

Then we can see that the residuals are normally distributed

$$\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H}))$$

3.3.3 Sampling Distribution for Variance

We mentioned earlier that the MLE estimate for σ^2 is biased and that it is usually preferred to use the unbiased estimator.

Construction of an unbiased estimator means that we wish to find $\mathbb{E}(S^2) = \sigma^2$. We can do this by considering the **Residual Sum of Squares**, $R = \mathbf{e}^T\mathbf{e} = \sum_{i=1}^n e_i^2$

$$\begin{aligned}\mathbb{E}\big(S^2\big) =& \mathbb{E}\bigg(\frac{\mathbf{e}^T\mathbf{e}}{n-p}\bigg)\\ =& \frac{1}{n-p}\mathbb{E}\big(\mathbf{e}^T\mathbf{e}\big)\\ =& \frac{1}{n-p}\mathbb{E}\big(\sum_{i=1}^n e_i^2\big)\\ =& \frac{1}{n-p}\bigg(\sum_{i=1}^n \mathbb{V}\big(e_i\big) + \sum_{i=1}^n \mathbb{E}\big(e_i\big)^2\bigg)\\ =& \frac{1}{n-p}\bigg(\sum_{i=1}^n \mathbb{V}\big(e_i\big)\bigg)\\ =& \frac{\sigma^2}{n-p}\sum_{i=1}^n(\mathbf{I} - \mathbf{H})_{ii}\\ =& \frac{\sigma^2}{n-p}\big(\sum_{i=1}^n(\mathbf{I})_{ii} - \sum_{i=1}^n(\mathbf{H})_{ii}\big)\\ =& \frac{\sigma^2}{n-p}\big(n-p\big)\\ =& \sigma^2\end{aligned}$$

where on the third to last line we use $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^n \mathbf{A}_{ii}$ and $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$

It can be shown that

$$S^2 \sim \frac{\sigma^2}{n-p}\chi^2(n-p)$$

a proof of which is provided in the appendix

3.4 Inference in Multiple Regression

In the case of multiple regression we can gauge uncertainty about our estimates of $\hat{\beta}_i$ individually or we can consider uncertainty with respect to linear combinations of arbitrary $\hat{\beta}_i, i = 1, ..., n$

3.4.1 Uncertainty for Individual Regression Coefficients

Similar to the case for simple univariate regression where we created a test statistic $T \sim t(n-2)$ which we then used to compute confidence intervals. We can follow a similar approach for uncertainty quantification of individual regression coefficients.

Since $\hat{\beta} \sim \boldsymbol{\mathcal{N}}\Big(\boldsymbol{\beta}, \sigma^2 \big(\mathbf{X}^T\mathbf{X}\big)^{-1}\Big)$, a multivariate normal distribution, then we know $\hat{\beta}_i \sim \mathcal{N}\Big(\beta_i, \sigma^2\xi\Big)$ where $\xi = \big(\mathbf{X}^T\mathbf{X}\big)^{-1}_{ii}$ so that a new random variable

$$Y = \frac{\hat{\beta}_i - \beta_i}{S\sqrt{\xi}} \sim t_{n-p-1}$$

The proof of which is given below. Let $Z = \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{\xi}} \sim \mathcal{N}(0,1)$ and $E = \frac{S^2(n-p)}{\sigma^2} \sim \chi^2(n-p)$

$$\begin{aligned} Y =& \frac{1}{\sigma\sqrt{\xi}}\frac{\beta_i - \beta_i}{1} \times \frac{1}{S\sqrt{\xi}}\frac{\sigma\sqrt{\xi}}{1} \\ =& Z \times \frac{\sigma}{S}\frac{\sqrt{n-p}}{\sqrt{n-p}} \\ =& \frac{Z\sqrt{n-p}}{E} \end{aligned}$$

where by definition this is a t-distributed random variable with $n-p$ degrees of freedom which completes the proof

3.4.2 Uncertainty for Multiple Regression Coefficients

If we seek to get uncertainty estimates for multiple regression coefficients we can use a linear combination using the matrix $\mathbf{K} \in \mathbb{R}^{k \times p+1}, k \leq p+1$ which when applied to $\boldsymbol{\beta}$ yields its image $\mathbf{K}\boldsymbol{\beta}$.

If for example we wanted to test $H_0 : \beta_i = 0, i = 2, ..., p+1$ then we would construct $\mathbf{K}$ such that

$$\mathbf{K}\boldsymbol{\beta} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p+1} \end{pmatrix} = \begin{pmatrix} \beta_2 \\ \beta_3 \\ \vdots \\ \beta_{p+1} \end{pmatrix}$$

As we do not know $\boldsymbol{\beta}$ we can use $\widehat{\boldsymbol{\beta}}$ as an estimator such that $\mathbf{K}\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\Big(\mathbf{K}\widehat{\boldsymbol{\beta}}, \sigma^2\mathbf{K}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{K}^\mathbf{T}\Big)$.

Then we define F as a random variable which follows the F-distribution with p and $n-p-1$ degrees of freedom and can be used to for the construction of confidence intervals and hypothesis testing

$$F := \frac{||\mathbf{K}\widehat{\boldsymbol{\beta}} - \mathbf{K}\boldsymbol{\beta}||^2_{\mathbf{K}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{K}^\mathbf{T}}}{kS^2} \sim F_{(p,n-p)}$$

before we proceed we state a few results which we will need before we can complete the proof

Definition. F-Distribution The F-distribution with υ_1 and υ_2 degrees of freedom is the distribution of

$$F := \frac{S_1/\upsilon_1}{S_2/\upsilon_2}$$

where S_1 and S_2 are independent random variables that follow the $\chi^2(\upsilon_1)$ and $\chi^2(\upsilon_2)$ distribution, respectively.

Given an invertible matrix $\mathbf{Q}$ then we define

$$||\kappa||^2_{\mathbf{Q}} = \kappa^T \mathbf{Q}^{-1} \kappa$$

where we are weighting the result by the components of $\mathbf{Q}$

Then if we consider the numerator of F

$$\begin{aligned}
||\mathbf{K}\widehat{\boldsymbol{\beta}} - \mathbf{K}\boldsymbol{\beta}||^2_{\mathbf{K}(\mathbf{X^TX})^{-1}\mathbf{K^T}} =& ||\mathbf{K}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2_{\mathbf{K}(\mathbf{X^TX})^{-1}\mathbf{K^T}} \\
=& ||\mathbf{K}(\mathbf{X^TX})^{-1}\mathbf{X}^T\mathbf{w}||^2_{\mathbf{K}(\mathbf{X^TX})^{-1}\mathbf{K^T}} \\
=& \mathbf{w}^T\mathbf{X}(\mathbf{X^TX})^{-1}\mathbf{K}^T\mathbf{K}(\mathbf{X^TX})^{-1}\mathbf{K^T}\mathbf{K}(\mathbf{X^TX})^{-1}\mathbf{X}^T\mathbf{w} \\
=& \mathbf{w}^T\mathbf{R}\mathbf{w}
\end{aligned}$$

As we know, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathbf{R}$ is idempotent (as was $\mathbf{H}$) and symmetric, meaning that we can use *Cochran's theorem* much the same way we did for $\mathbf{H}$ in the appendix to conclude that

$$\frac{\mathbf{w}^T\mathbf{R}\mathbf{w}}{\sigma^2} \sim \chi^2(k)$$

where the degrees freedom are given by the rank of $\mathbf{Q} = \mathbf{K}(\mathbf{X^TX})^{-1}\mathbf{K^T}$ which can be shown to be $\min\{k, p\} = k$, which is the number of independent variables we are testing where $k \leq p$

So the numerator follows the $\chi^2(k)$ so then if we set

$$\frac{1}{\sigma^2}\frac{||\mathbf{K}\widehat{\boldsymbol{\beta}} - \mathbf{K}\boldsymbol{\beta}||^2_{\mathbf{K}(\mathbf{X^TX})^{-1}\mathbf{K^T}}}{k} \times \frac{\sigma^2}{1}\frac{n-p}{S^2(n-p)}$$

where

- $S_1 = \frac{1}{\sigma^2}||\mathbf{K}\widehat{\boldsymbol{\beta}} - \mathbf{K}\boldsymbol{\beta}||^2_{\mathbf{K}(\mathbf{X^TX})^{-1}\mathbf{K^T}} \sim \chi^2(k)$
- $\upsilon_1 = k$
- $S_2 = \frac{S^2(n-p)}{\sigma^2} \sim \chi^2(n-p)$
- $\upsilon_2 = n - p$

So this is exactly the definition of the F-distribution with k and $n-p$ degrees of freedom, where $k = p - 1$ in the scenario where we choose all coefficients but β_1

4 Appendix

4.1 Distribution of S^2

The residuals can be written as

$$\begin{aligned}
\mathbf{e}^T\mathbf{e} =& (\mathbf{y} - \widehat{\mathbf{y}})^T(\mathbf{y} - \widehat{\mathbf{y}}) \\
=& \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y} \\
=& \mathbf{w}^T(\mathbf{I} - \mathbf{H})\mathbf{w}
\end{aligned}$$

Since $\mathbf{H}$ is symmetric then $\mathbf{I} - \mathbf{H}$ is also symmetric, meaning that we can diagonalize this matrix. Furthermore, since $\mathbf{H}$ is symmetric, its eigenvectors form an orthonormal basis in $\mathbb{R}^n$, represented by the matrices $\mathbf{U}$ and $\mathbf{D}$, where $\mathbf{D}$ contains the eigenvalues of $\mathbf{H}$ and $\mathbf{U}$ the corresponding eigenvectors

$$\mathbf{D} = \mathbf{U}\mathbf{H}\mathbf{U^T}$$

which implies that

$$\mathbf{H} = \mathbf{U^T}\mathbf{D}\mathbf{U}$$

In addition, if we consider the eigenvalues of $\mathbf{H}$

$$\mathbf{H}\mathbf{x} = \lambda\mathbf{x}$$

and the eigenvalues of $\mathbf{H^2}$

$$\begin{aligned}\mathbf{H}(\mathbf{Hx}) =& \lambda\mathbf{Hx}\\ =& \lambda^2\mathbf{x}\end{aligned}$$

and since $\mathbf{H^2} = \mathbf{H}$, we must have $\lambda = \lambda^2$, which can only happen when $\lambda = \{0, 1\}$, so that the eigenvalues of $\mathbf{H}$ are either 0 or 1.

Furthermore, we know that since $(\mathbf{X^T X})^{-1} \in \mathbb{R}^p$ is invertible, so it must have rank p, i.e., p linearly independent columns. It turns out that the rank of $\mathbf{H}$ is also p. The importance of this is that the number of non-zero diagonal elements of $\mathbf{D}$ is exactly p.

Then we can rewrite the RSS as

$$\begin{aligned}\mathbf{e}^T\mathbf{e} =& \mathbf{w}^T(\mathbf{I} - \mathbf{U^T D U})\mathbf{w}\\ =& \mathbf{w}^T\mathbf{w} - \mathbf{w}^T(\mathbf{U^T D U})\mathbf{w}\\ =& \mathbf{w}^T\mathbf{w} - (\mathbf{Uw})^\mathbf{T}\mathbf{D}(\mathbf{Uw})\end{aligned}$$

We know $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \mathbf{w} \in \mathbb{R}^n$.

Then if we consider $\boldsymbol{\eta} = \mathbf{Uw} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ since $\mathbf{U^T U} = \mathbf{I}$ since $\mathbf{U}^T = \mathbf{U}^{-1}$ by definition of being orthonormal.

The second term is a quadratic form

$$\begin{aligned}\boldsymbol{\eta}^T\mathbf{D}\boldsymbol{\eta} =& \sum_{i=1}^n\sum_{j=1}^n \eta_i D_{ij}\eta_j\\ =& \sum_{i=1}^n \eta_i D_{ii}\eta_i\\ =& \sum_{i=1}^n \eta_i^2 D_{ii}\\ =& \begin{cases}\eta_i^2, & \text{if } D_{ii} = 1\\ 0, & \text{otherwise}\end{cases}\end{aligned}$$

This, along with $\mathbf{w}^T\mathbf{w}$, is the sum of squared *(nearly)* standard normal random variables, which can be made into standard normal random variables by dividing $\mathbf{e}^T\mathbf{e}$ by σ^2.

$\frac{\mathbf{w^T w}}{\sigma^2}$ is a sum of n squared standard normal random variables and thus follows the $\chi^2(n)$ distribution. $\frac{\boldsymbol{\eta}^T\mathbf{D}\boldsymbol{\eta}}{\sigma^2} = \sum_{i=1}^n \frac{\eta_i^2 D_{ii}}{\sigma^2}$ is the sum of p squared standard normal random variables and generally since the sum of two standard normal random variables is not also standard normal, we have the sum of $n - p$ squared standard normal random variables which therefore follows the $\chi^2(n - p)$ distribution.

We restate the result for clarity

$$\frac{\mathbf{e^T e}}{\sigma^2} \sim \chi^2(n - p)$$

and therefore

$$S^2 = \frac{\mathbf{e^T e}}{(n - p)} \sim \frac{\sigma^2}{n - p}\chi^2(n - p)$$

4.2 Independence of Hw and (I − H)w

Taking the inner product of $\mathbf{Hw}$ and $(\mathbf{I} - \mathbf{H})\mathbf{w}$ and exploiting the associativity of matrix multiplication shows that these vectors are orthogonal

$$\begin{aligned}(\mathbf{Hw})^T(\mathbf{I} - \mathbf{H})\mathbf{w} =& \mathbf{w^T}(\mathbf{H}(\mathbf{I} - \mathbf{H}))\mathbf{w}\\ =& \mathbf{w^T}(\mathbf{0})\mathbf{w}\\ =& \mathbf{0}\end{aligned}$$

A result that follows from this is that $\widehat{\boldsymbol{\beta}}$ and S^2 are independent since we can write these terms as functions of $\mathbf{Hw}$ and $(\mathbf{I}-\mathbf{H})\mathbf{w}$.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} =& \widehat{\boldsymbol{\beta}} + (\mathbf{X^T X})^{-1}\mathbf{X^T w} \\
=& \widehat{\boldsymbol{\beta}} + (\mathbf{X^T X})^{-1}\mathbf{X^T X}(\mathbf{X^T X})^{-1}\mathbf{X^T w} \\
=& \widehat{\boldsymbol{\beta}} + (\mathbf{X^T X})^{-1}\mathbf{X^T H w}
\end{aligned}$$

and

$$\begin{aligned}
S^2 =& \frac{\mathbf{e^T e}}{(n-p)} \\
=& \frac{1}{n-p}\left(\mathbf{y}-\mathbf{y}\right)^T\left(\mathbf{y}-\mathbf{y}\right) \\
=& \frac{1}{n-p}\left(\mathbf{y}-\mathbf{Hy}\right)^T\left(\mathbf{y}-\mathbf{Hy}\right) \\
=& \frac{1}{n-p}\mathbf{y^T}(\mathbf{I}-\mathbf{H})^{\mathbf{T}}(\mathbf{I}-\mathbf{H})\mathbf{y} \\
=& \frac{1}{n-p}\mathbf{y^T}(\mathbf{I}-\mathbf{H})\mathbf{w}
\end{aligned}$$

which completes the proof.