SHEET - An Introduction to Statistical Learning
Chapter 3 - Linear Regression

4 December 2023

1 Linear Regression

1.1 Courses' Demonstrations

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X_i. Then $e_i = y_i - \hat{y}_i$ represents the ith residual - this is the difference between the ith observed response value and the ith response value that is predicted by our linear model. We define the residual sum of squares (RSS) as

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

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are

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

where $\bar{y}$ is the sample mean, defined as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Demonstration : $\hat{\beta}_0$ and $\hat{\beta}_1$

We search RSS as

$$f(\hat{\beta}_0, \hat{\beta}_1) = \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

We get the minimum where

$$\frac{\partial f}{\partial \hat{\beta}_k} = 0 \text{ with k} \in \{0, 1\}$$

We have

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

Then

$$\begin{cases} \dfrac{\partial f}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \dfrac{\partial f}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \end{cases}$$

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \sum_{i=1}^{n} x_i y_i - n\hat{\beta}_0 \bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \end{cases}$$

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \sum_{i=1}^{n} x_i y_i - n(\bar{y} - \hat{\beta}_1 \bar{x})\bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \end{cases}$$

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \sum_{i=1}^{n} (x_i y_i) - n\bar{y}\bar{x} - \hat{\beta}_1 (\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = 0 \end{cases}$$

We have

$$\begin{cases} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i y_i - \bar{y}\bar{x}) \\ \sum_{i=1}^{n} (x_i^2 - \bar{x}^2) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \end{cases}$$

Finally

$$\begin{cases} \hat{\beta}_{\mathbf{0}} = \bar{\mathbf{y}} - \hat{\beta}_{\mathbf{1}} \bar{\mathbf{x}} \\ \hat{\beta}_{\mathbf{1}} = \dfrac{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - \bar{\mathbf{y}})(\mathbf{x_i} - \bar{\mathbf{x}})}{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{x_i} - \bar{\mathbf{x}})^{\mathbf{2}}} \end{cases}$$

By developping we can have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(y_i)(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Demonstration : $\hat{\beta}_1$

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

In the same way

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

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_1 x_i + \beta_0 + \varepsilon_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\boldsymbol{\hat{\beta}_1} &= \boldsymbol{\beta_1} + \mathbf{\frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}
\end{aligned}$$

We assume that the True relationship between X and Y takes the form

$$Y = f(X) + \varepsilon$$

for some unknown function f, where ε is a mean-zero random error term. If f is to be approximated by a linear function then we can write the relationship as

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3.5}$$

How accurate is the sample mean $\hat{\mu}$ as an estimate of μ ? In general, we answer this question by computing the standard error of $\hat{\mu}$, written as the standard error SE($\hat{\mu}$). A reasonable estimate is $\hat{\mu} = \bar{y}$. We have the well-known formula :

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

where σ is the standard deviation of each of the realizations y_i of Y.

Demonstration : $Var(\hat{\mu})$

We have

$$\begin{aligned} \hat{\mu} &= \bar{y} \\ Var(X+Y) &= Var(X) + Var(Y) \\ Var(aX) &= a^2 Var(X) \textbf{ with a = cste} \\ Var(y_i) &= \sigma^2 \end{aligned}$$

Then

$$\begin{aligned} Var(\hat{\mu}) = Var(\bar{y}) &= Var(\frac{1}{n}\sum_{i=1}^{n} y_i) \\ &= \frac{1}{n^2} Var(\sum_{i=1}^{n} y_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(y_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\ \mathbf{Var}(\hat{\mu}) &= \frac{\boldsymbol{\sigma^2}}{\mathbf{n}} \qquad = \end{aligned}$$

In a similar vein, we can wonder how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values β_0 and β_1 . To compute the standard errors associated with β_0 and β_1 , we use the following formulas :

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

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

where $\sigma^2 = Var(\varepsilon)$

Preliminaries :

We have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} w_i y_i \quad \text{with} \quad w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sum_{i=1}^{n} w_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n(\bar{x} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0$$

$$\begin{aligned}\sum_{i=1}^{n} w_i x_i &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x} + \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\sum_{i=1}^{n}((x_i - \bar{x})\bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\bar{x}\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1 + 0 = 1\end{aligned}$$

$$\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$$

Démonstration : $E[\hat{\beta}_1]$

$$\begin{aligned}\mathbf{E}[\hat{\beta}_\mathbf{1}] &= E\left[\sum_{i=1}^{n} w_i y_i\right] = E\left[\sum_{i=1}^{n} w_i(\beta_1 x_i + \beta_0 + \varepsilon_i)\right] \\ &= E[\beta_1 \sum_{i=1}^{n} w_i x_i] + E[\beta_0 \sum_{i=1}^{n} w_i] + \sum_{i=1}^{n} E[w_i \varepsilon_i] \\ &= E[\beta_1 * 1] + E[\beta_0 * 0] + \sum_{i=1}^{n} E[\varepsilon_i] E[w_i] \\ \mathbf{E}[\hat{\beta}_\mathbf{1}] &= \beta_\mathbf{1}\end{aligned}$$

Démonstration : $E[\hat{\beta}_0]$

$$\mathbf{E}[\hat{\beta}_\mathbf{0}] = E[\bar{y} - \hat{\beta}_1\bar{x}] = E[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)] = \frac{1}{n}\sum_{i=1}^{n}E[y_i - \hat{\beta}_1 x_i]$$
$$= \frac{1}{n}\sum_{i=1}^{n}E[(\beta_1 x_i + \beta_0 + \varepsilon_i) - \hat{\beta}_1 x_i] = \frac{1}{n}\sum_{i=1}^{n}(E[(\beta_1]E[x_i] + E[\beta_0] + E[\varepsilon_i]) - E[\hat{\beta}_1]E[x_i])$$
$$= \frac{1}{n}\sum_{i=1}^{n}(\beta_1 x_i + \beta_0 + 0 - \beta_1 x_i) = \frac{1}{n}\sum_{i=1}^{n}\beta_0 = \beta_\mathbf{0}$$

Démonstration : $Var(\hat{\beta}_1)$

We have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} w_i y_i \quad \text{with} \quad w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$Var(\hat{\beta}_1) = Var(\sum_{i=1}^{n} w_i yi) = \sum_{i=1}^{n} Var(w_i y_i) = \sum_{i=1}^{n} {w_i}^2 Var(y_i) = \sigma^2 \sum_{i=1}^{n} {w_i}^2$$
$$\mathbf{Var}(\hat{\beta}_\mathbf{1}) = \frac{\sigma^\mathbf{2}}{\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{x_i} - \bar{\mathbf{x}})^\mathbf{2}}$$

Démonstration : $Var(\hat{\beta}_0)$

We have

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$
$$Cov(X, cste) = E[(X - E[X])(cste - E[cste])] = 0$$
$$Var(\bar{y}) = \frac{\sigma^2}{n}$$

Then

$$Var(\hat{\beta}_0) = Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x}Cov(\bar{y}, \hat{\beta}_1)$$
$$= \frac{\sigma^2}{n} + \bar{x}^2 Var(\hat{\beta}_1) \text{ because } \bar{y} = cste$$
$$= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\mathbf{Var}(\hat{\beta}_\mathbf{0}) = [\frac{\mathbf{1}}{\mathbf{n}} + \frac{\bar{\mathbf{x}}^\mathbf{2}}{\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{x_i} - \bar{\mathbf{x}})^\mathbf{2}}]\sigma^\mathbf{2}$$

The estimate of σ is known as the residual standard error, and is given by the formula residual standard error

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

Preliminaries :

$$\beta_1 - \hat{\beta}_1 = -\frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\begin{aligned}
\sum_{\mathbf{i=1}}^{\mathbf{n}} \hat{\varepsilon}_{\mathbf{i}} &= \sum_{i=1}^{n}(y_i - \hat{y}_i) = n\bar{y} - \sum_{i=1}^{n}\hat{y}_i \\
&= n(\hat{\beta}_1\bar{x} + \hat{\beta}_0) - \sum_{i=1}^{n}(\hat{\beta}_1 x_i + \hat{\beta}_0) = 0 \\
\frac{\partial f}{\partial \hat{\beta}_1} &= -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\
\sum_{i=1}^{n}\hat{\varepsilon}_i x_i &= \sum_{i=1}^{n}(y_i - \hat{y}_i)x_i = \sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i - \hat{\beta}_0)x_i = 0 \\
\sum_{i=1}^{n}\hat{\varepsilon}_i\hat{y}_i &= \sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{\beta}_1 x_i + \hat{\beta}_0) = \hat{\beta}_1\sum_{i=1}^{n}\hat{\varepsilon}_i x_i + \hat{\beta}_0\sum_{i=1}^{n}\varepsilon_i = 0
\end{aligned}$$

Demonstration : $\sum_{i=1}^{n} \varepsilon_i^2$
We have

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + 2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + 2\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{y}_i-\bar{y}) + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\
\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 &= \sum_{i=1}^{n}((\hat{\beta}_1 x_i+\hat{\beta}_0)-(\hat{\beta}_1\bar{x}+\hat{\beta}_0))^2 \\
&= \hat{\beta}_1{}^2\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
E[X^2] &= Var(X) + (E[X])^2
\end{aligned}$$

Then

$$\begin{aligned}
\sum_{i=1}^{n}\hat{\varepsilon}_i{}^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2 - \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\
&= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1{}^2\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
E[\sum_{i=1}^{n}\hat{\varepsilon}_i{}^2] &= \sum_{i=1}^{n}E[y_i^2] - nE[\bar{y}^2] - E[\hat{\beta}_1{}^2]\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \sum_{i=1}^{n}(Var(y_i)+E[y_i]^2)) - n(Var(\bar{y})+E[\bar{y}]^2) - (Var(\hat{\beta}_1)+E[\hat{\beta}_1]^2)\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \sum_{i=1}^{n}(\sigma^2+(\beta_1 x_i+\beta_0)^2) - n(\frac{\sigma^2}{n}+(\beta_1\bar{x}+\beta_0)^2) - (\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}+\beta_1^2)\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= n\sigma^2 - \sigma^2 - \sigma^2 + \sum_{i=1}^{n}(\beta_1 x_i+\beta_0)^2 - n(\beta_1\bar{x}+\beta_0)^2 - \beta_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \sigma^2(n-2)
\end{aligned}$$

$$\mathbf{RSE^2} = \frac{\mathbf{RSS}}{\mathbf{n-2}}$$

Recall that in the simple linear regression setting, in order to determine whether there is a relationship between the response and the predictor we can simply check whether $\beta_1 = 0$. In the multiple regression setting with p predictors, we need to ask whether all of the regression coefficients are zero, i.e. whether $\beta_1 = \beta_2 = \quad = \beta_p = 0$. As in the simple linear regression setting, we use a hypothesis test to answer this question. We test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \quad = \beta_p = 0$$

versus the alternative

$$Ha : \text{ at least one } \beta_j \text{ is non-zero}$$

This hypothesis test is performed by computing the F-statistic,

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \tag{3.23}$$

where, as with simple linear regression, $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and $RSS = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$ If the linear model assumptions are correct, one can show that

$$E[RSS/(n-p-1)] = \sigma^2$$

and that, provided H0 is true,

$$E[(TSS - RSS)/p] = \sigma^2.$$

Preliminaries :

$$\begin{aligned}
Y &= X\beta + \varepsilon \\
\hat{Y} &= X\hat{\beta} \\
MSE &= (Y - \hat{Y})^T(Y - \hat{Y}) = (Y - X\hat{\beta})^T(Y - X\hat{\beta}) \\
\frac{\partial MSE}{\partial \hat{\beta}} &= \frac{\partial (Y - X\hat{\beta})^T(Y - X\hat{\beta})}{\partial \hat{\beta}} = X^T(Y - X\hat{\beta}) = 0 \\
\Rightarrow \hat{\beta} &= (X^TX)^{-1}X^TY \\
\Rightarrow \hat{Y} &= X\hat{\beta} = X(X^TX)^{-1}X^TY \\
(I_n - X(X^TX)^{-1}X^T)Y &= (I_n - X(X^TX)^{-1}X^T)(X\beta + \varepsilon) \\
&= (I_n - X(X^TX)^{-1}X^T)(X\beta) + (I_n - X(X^TX)^{-1}X^T)\varepsilon) \\
&= (I_n - X(X^TX)^{-1}X^T)\varepsilon) \\
E[Tr(X)] &= Tr(E[X]) \\
Tr(AB) &= Tr(BA)
\end{aligned}$$

Demonstration : If linear assumption is True $E[RSS/(n-p-1)] = \sigma^2$

$$\begin{aligned}
RSS &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = Tr((Y-\hat{Y})(Y-\hat{Y})^T) \\
&= Tr((Y - X(X^TX)^{-1}X^TY)(Y - X(X^TX)^{-1}X^TY)^T) \\
&= Tr(((I_n - X(X^TX)^{-1}X^T)\varepsilon)((I_n - X(X^TX)^{-1}X^T)\varepsilon)^T) \\
&= Tr((I_n - X(X^TX)^{-1}X^T)\varepsilon\varepsilon^T((I_n - X(X^TX)^{-1}X^T)^T) \\
\textbf{We have} & \\
Tr(\varepsilon\varepsilon^T) &= n\sigma^2 \\
Tr(\varepsilon\varepsilon^T(X(X^TX)^{-1}X)^T) &= Tr((X(X^TX)^{-1}X^T)\varepsilon\varepsilon^T) \\
Tr((X(X^TX)^{-1}X^T)\varepsilon\varepsilon^T(X(X^TX)^{-1}X^T)^T) &= Tr((X(X^TX)^{-1}X^T)^T(X(X^TX)^{-1}X^T)\varepsilon\varepsilon^T) \\
&= Tr(X(X^TX)^{-1}X^T\varepsilon\varepsilon^T) \\
\textbf{Then} & \\
E[RSS] &= E[n\sigma^2 - Tr(X(X^TX)^{-1}X^T\varepsilon\varepsilon^T)] \\
&= n\sigma^2 - Tr(E[X(X^TX)^{-1}X^T\varepsilon\varepsilon^T]) \\
&= n\sigma^2 - Tr(X(X^TX)^{-1}X^TE[\varepsilon\varepsilon^T]) \\
&= n\sigma^2 - Tr(X(X^TX)^{-1}X^TE[(\varepsilon - E[\varepsilon])(\varepsilon^T - E[\varepsilon^T)]]) \\
&= n\sigma^2 - Tr(X(X^TX)^{-1}X^TVar(\epsilon) \\
&= n\sigma^2 - \sigma^2 Tr(X^TX(X^TX)^{-1}) \\
&= n\sigma^2 - \sigma^2 Tr(I_{p+1}) \\
&= n\sigma^2 - \sigma^2(p+1) \\
\mathbf{E[RSS]} &= \boldsymbol{\sigma^2(n-p-1)}
\end{aligned}$$

Demonstration : if H_0 True then $E[(TSS - RSS)/p] = \sigma^2$

$$\begin{aligned}
E[TSS] &= E[\sum_{i=1}^{n}(y_i - \bar{y})^2] = E[\sum_{i=1}^{n}(y_i^2 - 2y_i\bar{y} + \bar{y}^2)] \\
&= E[\sum_{i=1}^{n}(y_i^2) - n\bar{y}^2] = \sum_{i=1}^{n} E[y_i^2] - nE[\bar{y}^2] \\
E[X^2] &= Var(X) + E[X]^2 \\
E[TSS] &= \sum_{i=1}^{n}(Var(y_i) + E[y_i]^2) - n(Var(\bar{y}) + E[\bar{y}]^2) \\
&= \sum_{i=1}^{n}(\sigma^2 + E[\beta_0 + \varepsilon]^2) - n(\frac{\sigma^2}{n} + E[\beta_0]^2) \\
&= n(\sigma^2 + nE[\beta_0]^2) - n\frac{\sigma^2}{n} - nE[\beta_0]^2 \\
\mathbf{E[TSS]} &= \boldsymbol{\sigma^2(n-1)} \\
E[TSS - RSS] &= E[TSS] - E[RSS] = \sigma^2(n-1) - \sigma^2(n-p-1) \\
\mathbf{E[TSS - RSS]} &= \boldsymbol{\sigma^2 p}
\end{aligned}$$

Hence, when there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1. On the other hand, if H_a is true, then $E(TSS - RSS)/p > \sigma^2$, so we expect F to be greater than 1.

Demonstration : if H_a True then $E(TSS - RSS)/p > \sigma^2$

If H_a is True then

$$
\begin{aligned}
E[RSS] &= \sigma^2(n-p-1) \\
E[TSS] &= E[\sum_{i=1}^{n}(y_i - \bar{y})^2] = E[\sum_{i=1}^{n}(\sum_{j=1}^{n}(x_{ij} - \bar{x})\beta_j + \varepsilon_i)^2] \\
&= \sum_{i=1}^{n} E[\sum_{j=1}^{n}(x_{ij} - \bar{x})\beta_j + \varepsilon_i)^2] \\
&= \sum_{i=1}^{n}(E[(\sum_{j=1}^{n}(x_{ij} - \bar{x})\beta_j)^2] + 2E[(\sum_{j=1}^{n}(x_{ij} - \bar{x})\beta_j)\varepsilon_i] + E[\varepsilon_i^2]) \\
&= \sum_{i=1}^{n} E[(\sum_{j=1}^{n}(x_{ij} - \bar{x})\beta_j)^2] + \sigma^2(n-1) \\
E[TSS] &\geq \sigma^2(n-1) \\
E[TSS - RSS] &\geq \sigma^2(n-1) - E[RSS] = \sigma^2(n-1) - \sigma^2(n-p-1) \\
E[TSS - RSS] &\geq \sigma^2 p \\
\mathbf{E[TSS - RSS]/p} &\geq \boldsymbol{\sigma^2} \\
F &= \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \\
E[RSS/(n-p-1)] &= \sigma^2 \\
F &= \frac{(TSS - RSS)/p}{\sigma^2} \\
F &\geq \frac{\sigma^2}{\sigma^2} \\
\mathbf{F} &\geq \mathbf{1}
\end{aligned}
$$

Demonstration : if H_0 True then $F = 1$

We have

$$
\begin{aligned}
E[TSS] &= \sigma^2(n-1) \\
RSS &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = Tr((Y-\hat{Y})(Y-\hat{Y})^T) \\
&= Tr((Y - X(X^TX)^{-1}X^TY)(Y - X(X^TX)^{-1}X^TY)^T) \\
&= Tr((X\beta + \varepsilon - X(X^TX)^{-1}X^T(X\beta+\varepsilon)(X\beta + \varepsilon - X(X^TX)^{-1}X^T(X\beta+\varepsilon))^T) \\
&= Tr((\varepsilon - X(X^TX)^{-1}X^T(\varepsilon)(\varepsilon - X(X^TX)^{-1}X^T(\varepsilon))^T) \\
&= Tr((\varepsilon - X(X^TX)^{-1}X^T(\varepsilon)(\varepsilon - X(X^TX)^{-1}X^T(\varepsilon))^T) \\
&= Tr(((I_n - X(X^TX)^{-1}X^T)\varepsilon)((I_n - X(X^TX)^{-1}X^T)\varepsilon)^T) \\
\mathbf{E[RSS]} &= \boldsymbol{\sigma^2(n-p-1)} \\
\mathbf{E[TSS-RSS]} &= \boldsymbol{\sigma^2 p} \\
F &= \frac{(TSS-RSS)/p}{RSS/(n-p-1)} = \frac{\sigma^2}{\sigma^2} \\
\mathbf{F} &= \mathbf{1}
\end{aligned}
$$

In order to quantify an observation's leverage, we compute the leverage statistic. A large value of this statistic indicates an observation with high leverage. For a simple linear regression

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n} (x_i' - \bar{x})^2} \tag{3.37}$$

It is clear from this equation that h_i increases with the distance of x_i from $\bar{x}$. There is a simple extension of h_i to the case of multiple predictors, though we do not provide the formula here. The leverage statistic h_i is always between $\frac{1}{n}$ and 1, and the average leverage for all the observations is always equal to $\frac{p+1}{n}$. So if a given observation has a leverage statistic that greatly exceeds $\frac{p+1}{n}$, then we may suspect that the corresponding point has high leverage.

Demonstration : The leverage statistic h_i is always between $\frac{1}{n}$ and 1

We have

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n} (x_i' - \bar{x})^2}$$

if $\mathbf{x_i} \to \bar{\mathbf{x}}$

$$h_i \to \frac{1}{n} + \frac{(\bar{x} - \bar{x})^2}{\sum_{i'=1}^{n} (x_i' - \bar{x})^2} \to \frac{1}{n}$$

if $\mathbf{x_i} \to -\infty$ or $\mathbf{x_i} \to +\infty$

$$h_i \to \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(x_i - \bar{x})^2} \to \frac{1}{n} + 1$$

$\mathbf{h_i}$is convex on $\bar{\mathbf{x}}$

$$\frac{1}{n} \leq h_i \leq \frac{1}{n} + 1$$

Moreover

$$H = X(X^T X)^{-1} X^T$$

$$H^2 = H$$

$$h_{ii} = h_{ii}^2 + \sum_{i \neq j}^{n} h_{ij}^2$$

$$h_{ii} \geq h_{ii}^2$$

$$1 \geq h_{ii} \geq 0$$

Demonstration : the average leverage for all the observations is always equal to $\frac{p+1}{n}$

We have

$$
\begin{aligned}
y &= X\beta + \varepsilon \text{ with } dim(\beta) = p+1 \\
H &= X(X^TX)^{-1}X^T \\
H^2 &= H \\
\sum_{i=1}^{n} h_{ii} &= Tr(H) = Tr(X(X^TX)^{-1}X^T) \\
&= Tr(X^TX(X^TX)^{-1}) = Tr(I_{p+1}) = p+1 \\
\frac{1}{n}\sum_{i=1}^{n} h_{ii} &= \frac{p+1}{n}
\end{aligned}
$$

We call this situation multicollinearity. Instead of inspecting the correlation matrix, a better way to assess multicollinearity is to compute the variance inflation factor (VIF). The VIF is the ratio of the variance $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ it fit on its own. The smallest possible value for VIF is 1, which indicates the complete absence of collinearity. Typically in practice there is a small amount of collinearity among the predictors. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. The VIF for each variable can be computed using the formula.

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

Demonstration :

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

Wehave

$$
\begin{aligned}
S(\hat{\beta}) &= ||Y - X\hat{\beta}|| \\
&= (Y - X\hat{\beta})(Y - X\hat{\beta})^T \\
\hat{\beta} &= (X^TX)^{-1}X^TY \\
&= (X^TX)^{-1}X^T(X\beta + \epsilon) \\
&= \beta + (X^TX)^{-1}X^T\epsilon \\
Var(\hat{\beta}) &= E[\hat{\beta}^2] - E[\hat{\beta}]^2 \\
&= E[(\beta + (X^TX)^{-1}X^T\epsilon)(\beta + (X^TX)^{-1}X^T\epsilon)^T] - E[\hat{\beta}]^2 \\
&= E[(X^TX)^{-1}X^T\epsilon)(X^TX)^{-1}X^T\epsilon)^T] \\
&= E[(X^TX)^{-1}X^T\epsilon\epsilon^T((X^TX)^{-1}X^T)^T] \\
&= E[(X^TX)^{-1}X^T\sigma^2((X^TX)^{-1}X^T)^T] \\
&= \sigma^2E[(X^TX)^{-1}(X^TX)(((X^TX)^{-1})^T] \\
&= \sigma^2E[(X^TX)^{-1}] \\
&= \sigma^2(X^TX)^{-1} \\
\mathbf{Var}(\beta_{\mathbf{jj}}) &= \sigma^{\mathbf{2}}[(\mathbf{X^TX})^{-\mathbf{1}}]_{\mathbf{jj}}
\end{aligned}
$$