STAT 331 - Applied Linear Models

Cameron Roopnarine

Last updated: November 5, 2020

1 Introduction to Regression Models

Lecture 1 | 2020-09-08

DEFINITION 1.1: Response variable

A **response (dependent) variable** is the primary variable of interest, denoted by a capital roman letter Y.

DEFINITION 1.2: Explanatory Variable

An **explanatory (independent, predictor) variable** are variables that impact the response, denoted by x_i for $i = 1, \dots, p$.

DEFINITION 1.3: Regression Model

A **regression model** deals with modelling the functional relationship between a response variable and one or more explanatory variables.

EXAMPLE 1.4: Alligators in Florida

Let Y be the length in metres of an alligator and $x_1 := \{0, 1\}$ (male or female). The mass in an alligators stomach consists of fish (x_2), invertebrates (x_3), reptiles (x_4), birds (x_5), and other ($x_6, \dots, x_p$). We imagine we can explain Y in terms of $(x_1, \dots, x_p)$ using some function such that $Y = f(x_1, \dots, x_p)$.

In this course, we will be looking at linear models.

DEFINITION 1.5: Linear model

A general **linear model** is defined as $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$ where Y is the response variable, $(x_1, \dots, x_p)$ are the p explanatory variables, $(\beta_0, \beta_1, \dots, \beta_p)$ are the model parameters, and ε is the random error. We assume that $(x_1, \dots, x_p)$ are fixed constants, β_0 is the intercept of Y, $(\beta_1, \dots, \beta_p)$ all quantify effect on x_j on Y, and $\varepsilon \sim N(0, \sigma^2)$.

REMARK 1.6

In general, the model will not perfectly explain the data.

> "All models are wrong, but some are useful."

$Y \sim N\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2\right)$ since $\mathbb{E}[Y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ and $\mathbb{V}(Y) = \mathbb{V}(\varepsilon) = \sigma^2$.

2 Simple Linear Regression

Lecture 2 | 2020-09-09

DEFINITION 2.1: Simple linear regression

A **simple linear regression** is a linear model that uses only one explanatory variable; that is, $Y = \beta_0 + \beta_1 x + \varepsilon$. The **data** in a simple linear regression consists of pairs (x_i, y_i) where $i = 1, \dots, n$.

REMARK 2.2

Before fitting any model, we might want to make a scatterplot to visualize if there is a linear relationship between x and y, or calculate the *correlation*.

DEFINITION 2.3: Correlation

The **correlation** of random variables X and Y is $\rho_{XY} = \dfrac{\mathsf{Cov}(X,Y)}{\mathsf{Sd}(X)\mathsf{Sd}(Y)}$.

DEFINITION 2.4: Sample correlation

The **sample correlation** of all pairs (x_i, y_i) is

$$\begin{aligned} r &= \frac{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})}} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \end{aligned}$$

REMARK 2.5

The sample correlation measures the strength and direction of the linear relationship between X and Y. Note that $-1 \leqslant r \leqslant 1$. If $|r| \approx 1$, then there is a strong linear relationship, and if $|r| \approx 0$ then there is a lack of linear relationship. Also, if $r > 0$, then there is a positive relationship, and if $r < 0$ then there is a negative relationship. It does not tell us how to predict Y from X. To do so, we need to estimate β_0 and β_1.

DEFINITION 2.6: Simple linear regression model

For data (x_i, y_i) for $i = 1, \dots, n$, the **simple linear regression model** is $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with the assumption that $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Therefore, $Y_i \sim N(\mu_i = \beta_0 + \beta_1 x_i, \sigma^2)$.

DEFINITION 2.7: Method of least squares

The method of estimating β_0 and β_1 by minimizing $S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$ is referred to as the **method of least squares**.

REMARK 2.8

The least squares is equivalent to maximum likelihood estimate when $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

THEOREM 2.9: Least Square Estimates (LSEs) for SLR

Minimizing $S(\beta_0, \beta_1)$, gives the least square estimates

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \textit{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Proof of: 2.9

$\dfrac{\partial S}{\partial \beta_0} = 2\displaystyle\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)](-1)$ and $\dfrac{\partial S}{\partial \beta_1} = 2\displaystyle\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)](-x_i)$.
Now,

$$\frac{dS}{d\beta_0} := 0 \iff \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \iff \beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\begin{aligned} \frac{dS}{d\beta_1} := 0 &\overset{\text{plug } \beta_0}{\iff} \sum_{i=1}^{n}[y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i]\, x_i = 0 \\ &\iff \sum_{i=1}^{n} x_i(y_i - \bar{y}) - \beta_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}) = 0 \\ &\iff \beta_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \end{aligned}$$

REMARK 2.10

We use a hat on the β's to show that they are estimates.

DEFINITION 2.11: Fitted value, Residual

The expression $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is called the **fitted value** that corresponds to the ith observation with x_i as the explanatory variable. The difference between y_i and $\hat{\mu}_i$, and $e_i = y_i - \hat{\mu}_i$ is referred to as the **residual**. It is the vertical distance between the observation y_i and the estimated line $\hat{\mu}_i$ evaluated at x_i.

Lecture 3 | 2020-09-14

For $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, the equation of fitted line is given by $y = \hat{\beta}_0 + \hat{\beta}_1 x$. Our interpretation of the parameters is as follows.

- $\hat{\beta}_0$ is the estimate of the expected response when $x = 0$ (but not always meaningful if outside range of x_i's in data)
- $\hat{\beta}_1$ is the estimate of expected change in response for unit increase in x
- σ^2 is the "variability around the line" where $\sigma^2 = \mathbb{V}(\varepsilon_i) = \mathbb{V}(Y_i)$

Q: How should we estimate σ^2?

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i) \quad \text{and} \quad e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Our intuition tells us to use variability in the residuals to estimate σ^2, so we use

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$$

where the first term looks like sample variance of e_i's. The second equality follows since $\bar{e} = \bar{y} - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}) = 0$ by definition of our β_0 estimate.

DEFINITION 2.12: Residual sum of squares

$\text{SS(Res)} = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 = \sum_{i=1}^n e_i^2$, is known as the **residual (error) sum of squares**.

REMARK 2.13

The $n-2$ will be looked at in more detail later, but for now it suffices to say that the degrees of freedom is $n-2$ or equivalently, $n-$ number of parameters estimated. It allows $\hat{\sigma}^2$ to be an unbiased estimator for the true value of σ^2; that is, $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$ whenever $\hat{\sigma}^2$ is viewed as a random variable.

THEOREM 2.14: Linear Combination of Independent Normal Random Variables

If $Y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \dots, n$ independently, then

$$\sum_{i=1}^n a_i Y_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right)$$

Proof of: 2.14

The proof is completed in STAT 330 with moment generating functions.

Viewing $\hat{\beta}_1$ as a random variable:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i - \bar{Y}\overbrace{\sum_{i=1}^n (x_i - \bar{x})}^{0}}{\sum_{i=1}^n (x_i - \bar{x})x_i - \bar{x}\underbrace{\sum_{i=1}^n (x_i - \bar{x})}_{0}} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})x_i} = \sum_{i=1}^n a_i Y_i$$

where $a_i = \dfrac{x_i - \bar{x}}{\sum_{i=1}^n x_i (x_i - \bar{x})}$. Therefore,

$$\mathbb{E}[\hat{\beta}_1] = \sum_{i=1}^n a_i \mathbb{E}[Y_i] = \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^n x_i (x_i - \bar{x})} = \frac{\beta_0 \overbrace{\sum_{i=1}^n (x_i - \bar{x})}^{0} + \beta_1 \sum_{i=1}^n x_i (x_i - \bar{x})}{\sum_{i=1}^n x_i (x_i - \bar{x})} = \beta_1$$

Now, we calculate the variance of $\hat{\beta}_1$:

$$\mathbb{V}(\hat{\beta}_1) = \sum_{i=1}^n a_i^2 \mathbb{V}(Y_i) = \sigma^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n x_i (x_i - \bar{x})\right]^2} = \sigma^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{S_{xx}}$$

Using our calculations from $\hat{\beta}_1$, and viewing $\hat{\beta}_0$ as a random variable:

$$\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\bar{Y}] - \bar{x}\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\frac{\sum_{i=1}^n Y_i}{n}\right] - \bar{x}\beta_1 = \frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_i)}{n} - \beta_1 \bar{x} = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0$$

Now, we calculate the variance of $\hat{\beta}_0$:

$$\mathbb{V}(\hat{\beta}_1) = \mathbb{V}(\bar{Y} - \beta_1\bar{x}) = \mathbb{V}(\bar{Y}) + (-\bar{x}^2)\mathbb{V}(\beta_1) = \mathbb{V}\left(\frac{\sum_{i=1}^n Y_i}{n}\right) + \bar{x}^2\left(\frac{\sigma^2}{S_{xx}}\right) = \frac{n\sigma^2}{n^2} + \frac{\sigma^2 x^2}{S_{xx}}$$

Also, since $\hat{\beta}_1$ and $\hat{\beta}_0$ are linear combination of Normal random variables, they follow a Normal distribution. Therefore, we get the following theorem.

THEOREM 2.15: Distribution of LSEs

The distribution of the least square estimates are given by

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \quad \textit{and} \quad \hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

Since $\mathbb{E}[\hat{\beta}_1] = \beta_1$, we say $\hat{\beta}_1$ is an unbiased estimator of β_1. This implies that when the experiment is repeated a large number of times, the average of the estimates $\hat{\beta}_1$; that is, $\mathbb{E}[\hat{\beta}_1]$ coincides with the true value of β_1. A similar argument can be made for β_0.

Then, $\dfrac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$, but σ is unknown, so need to use $\hat{\sigma}$ to get $\dfrac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t(n-2)$.

DEFINITION 2.16: Standard deviation and standard error of $\hat{\beta}_1$

The **standard deviation** of $\hat{\beta}_1$ is defined as $\mathsf{Sd}(\hat{\beta}_1) = \sigma/\sqrt{S_{xx}}$. The **estimated** standard deviation of $\hat{\beta}_1$ is also referred to as the **standard error** of the estimate $\hat{\beta}_1$, and we write $\mathsf{Se}(\hat{\beta}_1) = \hat{\sigma}/\sqrt{S_{xx}}$.

DEFINITION 2.17: Student t distribution

Suppose $Z \sim N(0,1)$ and $U \sim \chi^2(\nu)$, with Z and U independent. Then, $T = Z/\sqrt{U/\nu}$ has a **Student t distribution** with ν degrees of freedom.

THEOREM 2.18

For a simple linear regression model,

$$\frac{\hat{\sigma}^2(n-2)}{\sigma^2} = \frac{SS(Res)}{\sigma^2} \sim \chi^2(n-2)$$

Proof of: 2.18

Too hard for sure.

Using the theorem stated, we justify the fact that replacing σ with $\hat{\sigma}$ gives us a $t(n-2)$ distribution.

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{\dfrac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\dfrac{\hat{\sigma}^2(n-2)}{\sigma^2}\left(\dfrac{1}{n-2}\right)}} = \frac{Z}{\sqrt{U/\nu}} = T \sim t(n-2)$$

where $\dfrac{\hat{\sigma}^2(n-2)}{\sigma^2} = U$, $\nu = n-2$, and $Z = \dfrac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}}$. A $(1-\alpha)$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm c\,\mathsf{Se}(\hat{\beta}_1)$$

where c is the $1-\dfrac{\alpha}{2}$ quantile of $t(n-2)$; that is, $P(|T| \leqslant c) = 1-\alpha$ or $P(T \leqslant c) = 1-\dfrac{\alpha}{2}$ where $T \sim t(n-2)$.

Hypothesis test: H_0: $\beta = 0$ versus H_A: $\beta_1 \neq 0$. If H_0 is true, then $\hat{\beta}_1/\mathsf{Se}(\hat{\beta}_1) \sim t(n-2)$, so calculate the **t statistic** $t = \hat{\beta}_1/\mathsf{Se}(\hat{\beta}_1)$, and reject H_0 at level α if $|t| > c$ where c is $1-\dfrac{\alpha}{2}$ quantile of $t(n-2)$. Therefore, p-value $= P(|T| \geqslant |t|) = 2P(T \geqslant |t|)$.

Lecture 4 | 2020-09-16

Suppose we want to predict the response y for a new value of x, say $x = x_0$. Then, SLR model says $Y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$ where Y_0 is a random variable for response when $x = x_0$; that is, $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. The fitted model predicts the *value* of y to be $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Also, $\mathbb{E}[\hat{Y}_0] = \mathbb{E}[\hat{\beta}_0] + x_0\mathbb{E}[\hat{\beta}_1] = \beta_0 + \beta_1 x_0 = \mathbb{E}[Y_0]$, since $\hat{\beta}_i$ for $i = 0, 1$ are unbiased. Therefore, we can say that $\hat{Y}_0$ is an unbiased estimate of the random variable for the mean of Y_0. For the variance of $\hat{Y}_0$ we write

$$\begin{aligned}
\hat{Y}_0 &= \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{Y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_0 \\
&= \bar{Y} + \hat{\beta}_1(x_0 - \bar{x}) \\
&= \sum_{i=1}^n \left[\frac{Y_i}{n} + (x_0-\bar{x})\left(\frac{(x_i-\bar{x})(Y_i-\bar{Y})}{S_{xx}}\right)\right] \\
&= \sum_{i=1}^n \left[\frac{Y_i}{n} + (x_0-\bar{x})\left(\frac{(x_i-\bar{x})Y_i}{S_{xx}}\right)\right] \\
&= \sum_{i=1}^n \left[\frac{1}{n} + \frac{(x_0-\bar{x})(x_i-\bar{x})}{S_{xx}}\right] Y_i \\
&= \sum_{i=1}^n a_i Y_i
\end{aligned}$$

where $a_i = \dfrac{1}{n} + \dfrac{(x_0-\bar{x})(x_i-\bar{x})}{S_{xx}}$. Therefore,

$$\begin{aligned}
\mathbb{V}(Y_0) &= \sum_{i=1}^n \left[\frac{1}{n} + \frac{(x_0-\bar{x})(x_i-\bar{x})}{S_{xx}}\right]^2 \\
&= \sum_{i=1}^n \left[\frac{1}{n^2} + \frac{2(x_0-\bar{x})(x_i-\bar{x})}{nS_{xx}} + \frac{(x_0-\bar{x})^2(x_i-\bar{x})^2}{(S_{xx})^2}\right] \\
&= \sum_{i=1}^n \left[\frac{1}{n^2}\right] + \frac{2(x_0-\bar{x})}{nS_{xx}}\sum_{i=1}^n (x_i-\bar{x}) + \frac{(x_0-\bar{x})^2}{(S_{xx})^2}\sum_{i=1}^n (x_i-\bar{x})^2 \\
&= \frac{1}{n} + \frac{2(x_0-\bar{x})}{S_{xx}}(0) + \frac{(x_0-\bar{x})^2}{(S_{xx})^2}(S_{xx}) \\
&= \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}
\end{aligned}$$

We proved the following theorem.

THEOREM 2.19: Distribution of Prediction

The distribution of the prediction random variable is given by

$$\hat{Y}_0 \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$$

DEFINITION 2.20: Prediction error

The random variable for **prediction error** is defined as $Y_0 - \hat{Y}_0$ where Y_0 and $\hat{Y}_0$ are independent and $\hat{Y}_0$ is a function of $Y_1, \dots, Y_n$.

$$\mathbb{E}[Y_0 - \hat{Y}_0] = \mathbb{E}[Y_0] - \mathbb{E}[\hat{Y}_0] = 0$$

$$\mathbb{V}(Y_0 - \hat{Y}_0) = \mathbb{V}(Y_0) + (-1)^2\mathbb{V}(\hat{Y}_0) = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

We proved the following theorem.

THEOREM 2.21: Distribution of Prediction Error

The distribution of the prediction error is given by

$$Y_0 - \hat{Y}_0 \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$$

Since σ is unknown, we use $\hat{\sigma}$ and get the following:

$$\frac{Y_0 - \hat{Y}_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

Intuition for prediction error composed of 2 terms:

- $\mathbb{V}(Y_0)$: random error of new observation
- $\mathbb{V}(\hat{Y}_0)$ (predictor): estimating β_0 and β_1

Those are 2 sources of uncertainty.

REMARK 2.22

Be careful that the prediction may not make sense if x_0 is outside the range of the x_i's in the data.

A $(1-\alpha)$ prediction interval for the mean response $y_0 = \beta_0 + \beta_1 x_0$ at x_0 is

$$\hat{y}_0 \pm c\,\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

where c is the $1 - \frac{\alpha}{2}$ quantile of $t(n-2)$.

EXAMPLE 2.23: Orange production 2018 in FL

We are given the following information.

- x: acres
- y: # boxes of oranges (thousands)
- (x_i, y_i) recorded for each of 25 FL counties
- $r = 0.964$
- $\bar{x} = 16133$
- $\bar{y} = 1798$
- $S_{xx} = 1.245 \times 10^{10}$
- $S_{xy} = 1.453 \times 10^{9}$

Now, $\hat{\beta}_1 = S_{xy}/S_{xx} = 0.1167$ has a positive slope, therefore x and y are positively correlated. The expected number of boxes produced is estimated to be about 117 higher per an additional acre.
Computing $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = -85.3$, we see that it is not meaningful to interpret, since it is the expected production if there were 0 acres (outside the range of x_i) as no county has $x = 0$.
Now suppose $\mathsf{SS}(\text{Res}) = 1.31 \times 10^7$ the residuals are the differences between y_i and the fitted regression line.

- $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^n e_i^2}{n-2} = \dfrac{1.31 \times 10^7}{25-2} = 5.7 \times 10^5$
- $\mathsf{Se}(\hat{\beta}_1) = \dfrac{\hat{\sigma}}{\sqrt{S_{xx}}} = 0.00676$
- To test H_0: $\beta_1 = 0$, calculate $t = (\hat{\beta}_1 - 0)/\mathsf{Se}(\hat{\beta}_1) = 0.1167/0.00676 \approx 17.3$, then elect the 0.975 quantile (for demonstration purposes) of $t(23)$ which is 2.07.
- Note that 17.3 is very unlikely to see in $t(23)$.

Since $17.3 \gg 2.07$, we reject H_0 at $\alpha = 0.05$ level, and conclude there's a significant linear relationship between acres and oranges produced.
The 95% confidence interval for β_1 is given by $0.1167 \pm 2.07(0.00676)$, which does not contain 0.

$$p\text{-value} = P(|t_{23}| \geqslant 17.3) = 2P(t_{23} \geqslant 17.3) \approx 1.2 \times 10^{-14}$$

Predict the # of boxes in thousands produced if we had 10000 acres to grow oranges.

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 = -85.3 + (0.1167)(10000) \approx 1082$$

The 95% prediction interval is given by

$$1082 \pm 2.07\sqrt{5.69 \times 10^5}\sqrt{1 + \frac{1}{25} + \frac{(6133)^2}{1.245 \times 10^{10}}} = [-512.0407, 2675.595]$$

REMARK 2.24

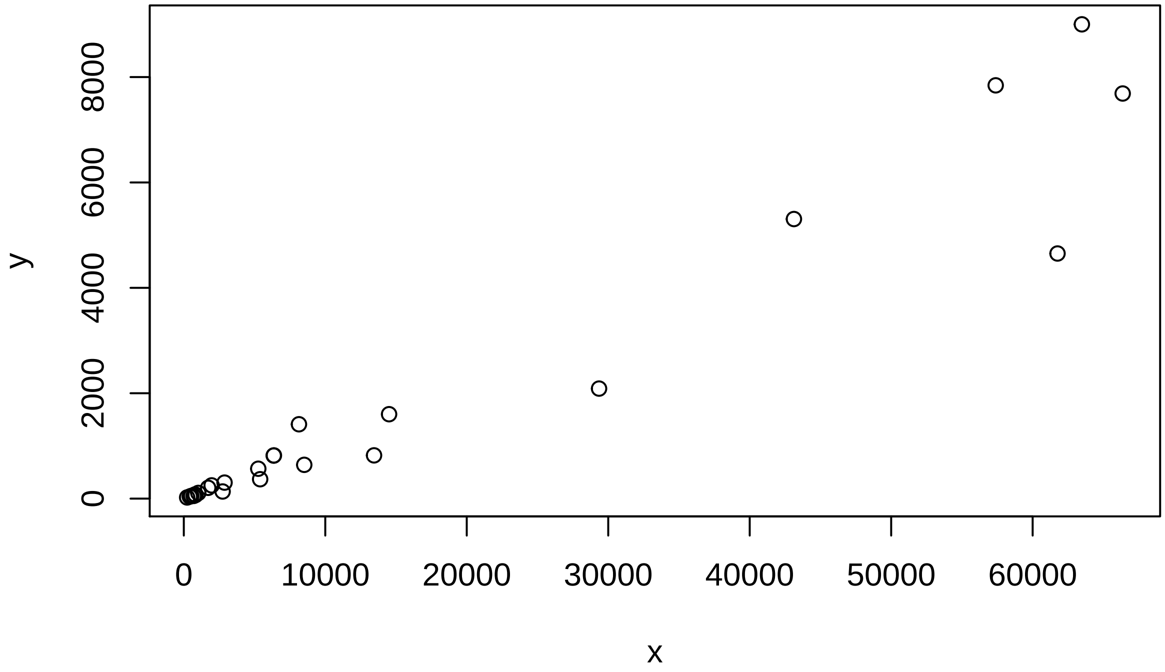
We are **not** trying to establish causation.

The example done in R is included in the next page.

```r
# Read data from florange.csv and input it into the dat vector.
dat <- read.csv("florange.csv")
# Done to make the predict function work well.
x <- dat$acres
y <- dat$boxes
# Output the first 6 rows in dat.
head(dat)
```

```
##      county boxes acres
## 1   Brevard    51   696
## 2 Charlotte   821 13447
## 3   Collier  2088 29351
## 4    DeSoto  7688 66365
## 5    Glades   368  5396
## 6    Hardee  5306 43126
```

```r
# Draw a scatterplot with x-axis as `acres` and y-axis as `boxes`.
plot(x,y)
```



```r
# Compute some common variables with common functions.
r <- cor(x,y)
xbar <- mean(x)
ybar <- mean(y)
cat("r:", r, "xbar:", xbar, "ybar:", ybar)
```

```
## r: 0.9635098 xbar: 16132.64 ybar: 1797.56
```

Therefore, $r = 0.9635098$, $\bar{x} = 16132.64$, and $\bar{y} = 1797.56$.

```r
# Compute some common variables manually.
Sxx <- sum( (x - xbar)^2 )
Sxy <- sum( (x - xbar) * (y - ybar) )
cat("Sxx: ", Sxx, "Sxy: ", Sxy)
```

```
## Sxx:  12450023404 Sxy:  1453128337
```

Therefore, $S_{xx} = 12450023404 = 1.245 \times 10^{10}$ and $Sxy = 1453128337 = 1.453 \times 10^{9}$.

```
# R's lm function fits linear models
lm.1 <- lm(y~x)
summary(lm.1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2470.81    -6.17    71.72   106.46  1677.32
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -85.391989 186.178031  -0.459    0.651
## x             0.116717   0.006761  17.263 1.16e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 754.4 on 23 degrees of freedom
## Multiple R-squared:  0.9284, Adjusted R-squared:  0.9252
## F-statistic:   298 on 1 and 23 DF,  p-value: 1.164e-14
```

From the summary, we can see that $\hat{\beta}_0 = -85.391989$, $\hat{\beta}_1 = 0.116717$, $\text{Se}(\hat{\beta}_1) = 0.006761$, $t = 17.263$, p-value $= 1.64 \times 10^{-14}$, and $\hat{\sigma} = 754.4$.

```
# Sum Squared Fitted Values
sum(lm.1$fitted.values^2)
```

```
## [1] 250385207
```

```
# Sum Squared Residuals
sum(lm.1$residuals^2)
```

```
## [1] 13089860
```

Therefore, $SS(\text{Res}) = \sum_{i=1}^{n} e_i^2 = 13089860 = 1.31 \times 10^{7}$.

```
# Manual calculation of sigma^2 estimate
sum(lm.1$residuals^2) / 23
```

```
## [1] 569124.3
```

Therefore, $\hat{\sigma}^2 = 69124.3 = 5.7 \times 10^{5}$.

```
# Manual calculation of sigma estimate
sqrt(sum(lm.1$residuals^2) / 23)
```

```
## [1] 754.4033
```

Therefore, $\hat{\sigma} = 754.4$.

```
# t distribution values
qt(0.975,23)
```

```
## [1] 2.068658
```

Therefore, $c = 2.07$.

```
# 95% confidence interval
confint(lm.1)
```

```
##                   2.5 %      97.5 %
## (Intercept) -470.5305905 299.7466119
## x              0.1027305   0.1307034
```

```
# 95% prediction interval with predicted boxes if we had 10000 acres
predict(lm.1, data.frame(x=10000), interval="prediction")
```

```
##        fit       lwr      upr
## 1 1081.777 -512.0407 2675.595
```

Q: Is σ the same for all values of y?

A: It appears to not in the sense that the variance appears to be higher with respect to higher acres. Sigma will be smaller when there's less acres. Later, this will be testing equal variance or homoscedastic assumption. Later, when we talk about variable transformations we can consider taking the logarithm.

Q: Are the error terms plausibly independent? In other words, does knowing one e_i (residual) help predict e_j (another residual) for a different county?

A: There's diagnostics for checking this. However, intuitively there could be some common factors at play when two counties are geographically close.

3 Multiple Linear Regression

Lecture 5 | 2020-09-21

DEFINITION 3.1: Multiple linear regression

A **multiple linear regression** (MLR) model is defined as

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

which links a response variable y to several independent explanatory variables $x_1, x_2, \dots, x_p$.

EXAMPLE 3.2: Rocket MLR

- x_1: nozzle area (large or small, 0 or 1)
- x_2: mixture in propellant, ratio oxidized fuel
- Y: thrust

Want to develop linear relationship between response y and x_1, x_2; that is, we want to develop a linear relationship between thrust and both nozzle area and mixture in propellant.

In a MLR, there are n observations, where each consists of p response variables (y_i), and p explanatory variables $(x_{i1}, x_{i2}, \dots, x_{ip})$. Then,

$$Y_i \sim N(\underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}_{\mathbb{E}[Y_i]=\mu_i}, \sigma^2)$$

or $Y_i = \mu_i + \varepsilon_i$ where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. We can write in vector/matrix form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \cdots + \beta_p x_{2p} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

Which we can more commonly write as $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1(p-1)} & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2(p-1)} & x_{2p} \\ \vdots & & & \ddots & & \vdots \\ 1 & x_{(n-1)1} & x_{(n-1)2} & \cdots & x_{(n-1)(p-1)} & x_{(n-1)p} \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n(p-1)} & x_{np} \end{bmatrix}_{n\times(p+1)} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

DEFINITION 3.3: Random vector

We call $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)^\top$ a **random vector**.

DEFINITION 3.4: Mean vector

The **mean vector** of $\boldsymbol{Y}$ is defined as $\mathbb{E}[\boldsymbol{Y}] = (\mathbb{E}[Y_1], \mathbb{E}[Y_2], \dots, \mathbb{E}[Y_n])^\top$.

DEFINITION 3.5: Covariance matrix

The **covariance matrix** (or **variance-covariance matrix**) of $\boldsymbol{Y}$ is defined as

$$\mathbb{V}(\boldsymbol{Y}) = \begin{bmatrix} \mathbb{V}(Y_1) & \mathsf{Cov}(Y_1, Y_2) & \cdots & \mathsf{Cov}(Y_1, Y_{n-1}) & \mathsf{Cov}(Y_1, Y_n) \\ \mathsf{Cov}(Y_2, Y_1) & \mathbb{V}(Y_2) & \cdots & \mathsf{Cov}(Y_2, Y_{n-1}) & \mathsf{Cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathsf{Cov}(Y_{n-1}, Y_1) & \mathsf{Cov}(Y_{n-1}, Y_2) & \cdots & \mathbb{V}(Y_{n-1}) & \mathsf{Cov}(Y_{n-1}, Y_n) \\ \mathsf{Cov}(Y_n, Y_1) & \mathsf{Cov}(Y_n, Y_2) & \cdots & \mathsf{Cov}(Y_n, Y_{n-1}) & \mathbb{V}(Y_n) \end{bmatrix}_{n \times n}$$

PROPOSITION 3.6: Properties of Covariance Matrix

Let $\boldsymbol{Y}$ be a random vector and $\boldsymbol{a} \in \mathbb{R}^n$, then the covariance matrix has the following properties.

(1) Symmetric since $\mathsf{Cov}(Y_i, Y_j) = \mathsf{Cov}(Y_j, Y_i)$; that is $\mathbb{V}(\boldsymbol{Y})^\top = \mathbb{V}(\boldsymbol{Y})$.

(2) Positive semi-definite since $\boldsymbol{a}^\top \mathbb{V}(\boldsymbol{Y})\boldsymbol{a} \geqslant 0$ for all $\boldsymbol{a} \in \mathbb{R}^n$.

(3) $\mathbb{V}(\boldsymbol{Y}) = \mathbb{E}\left[(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])^\top\right]$

Proof of: 3.6

Trivial.

PROPOSITION 3.7: Properties of Random Vector

Let $\boldsymbol{a}$ be a $1 \times n$ matrix (row vector) of constants and A be an $n \times n$ matrix of constants, then the random vector has the following properties.

(1) $\mathbb{E}[\boldsymbol{a}\boldsymbol{Y}] = \boldsymbol{a}\boldsymbol{Y}$

(2) $\mathbb{E}[A\boldsymbol{Y}] = A\mathbb{E}[\boldsymbol{Y}]$

(3) $\mathbb{V}(\boldsymbol{a}\boldsymbol{Y}) = \boldsymbol{a}\mathbb{V}(\boldsymbol{Y})\boldsymbol{a}^\top$

(4) $\mathbb{V}(A\boldsymbol{Y}) = A\mathbb{V}(\boldsymbol{Y})A^\top$

Proof of: 3.7

We prove prove property (4) only.

$$\begin{aligned} \mathbb{V}(A\boldsymbol{Y}) &= \mathbb{E}[(A\boldsymbol{Y} - \mathbb{E}[A\boldsymbol{Y}])(A\boldsymbol{Y} - \mathbb{E}[A\boldsymbol{Y}])^\top] \\ &= \mathbb{E}[(A\boldsymbol{Y} - A\mathbb{E}[\boldsymbol{Y}])(A\boldsymbol{Y} - A\mathbb{E}[\boldsymbol{Y}])^\top] \\ &= \mathbb{E}[A(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])(A(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]))^\top] \\ &= \mathbb{E}[A(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])^\top A^\top] \\ &= A\mathbb{E}[(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])^\top]A^\top \\ &= A\mathbb{V}(\boldsymbol{Y})A^\top \end{aligned}$$

EXAMPLE 3.8: Calculations with MLR Varaibles

Let $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^\top$. Suppose $\mathbb{E}[\boldsymbol{Y}] = (3, 1, 2)^\top$. Let $\mathbb{V}(Y) = \begin{bmatrix} 4 & 1/2 & -2 \\ 1/2 & 1 & 0 \\ -2 & 0 & 3 \end{bmatrix}$ and $\boldsymbol{a} = (1, -1, 2)$ and $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$. Note that $\boldsymbol{a}$ is a 1×3 row vector. Compute the following.

(i) $\mathbb{E}[\boldsymbol{a}\boldsymbol{Y}]$

(ii) $\mathbb{V}(\boldsymbol{a}\boldsymbol{Y})$

(iii) $\mathbb{E}[A\boldsymbol{Y}]$

(iv) $\mathbb{V}(A\boldsymbol{Y})$

Solution. We do the first two and leave the rest as an exercise.

(i) $\mathbb{E}[\boldsymbol{a}\boldsymbol{Y}] = \boldsymbol{a}\mathbb{E}[\boldsymbol{Y}] = \begin{bmatrix} 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} = 1(3) - 1(1) + 2(2) = 6.$

(ii)

$$\begin{aligned}
\mathbb{V}(\boldsymbol{a}\boldsymbol{Y}) &= \boldsymbol{a}\mathbb{V}(\boldsymbol{Y})\boldsymbol{a}^\top \\
&= \begin{bmatrix} 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 4 & 1/2 & -2 \\ 1/2 & 1 & 0 \\ -2 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} \\
&= \begin{bmatrix} 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 4(1) + (1/2)(-1) - 2(2) \\ (1/2)(1) + 1(-1) + 0(2) \\ -2(1) + 0(-1) + 3(2) \end{bmatrix} \\
&= \begin{bmatrix} 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} -1/2 \\ -1/2 \\ 4 \end{bmatrix} \\
&= 1(-1/2) - 1(-1/2) + 2(4) \\
&= 8
\end{aligned}$$

DEFINITION 3.9: Multivariate normal distribution

Let $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$ be a random vector. We say that $Y \sim \text{MVN}(\mu, \sigma)$; that is, Y follows a **multivariate normal distribution** (MVN) when

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right\}$$

where $\boldsymbol{\mu}$ is defined as the **mean vector**, and Σ is defined as the **covariance matrix**. Note that Σ^{-1} is the inverse of the covariance matrix and $|\Sigma|$ is the determinant of Σ.

THEOREM 3.10: Properties of Multivariate Normal Distribution

Let $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top \sim MVN(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{a}$ be a $1 \times n$ row vector of constants and A be an $n \times n$ matrix of constants.

(1) Linear transformations of MVN is MVN, so

$$\boldsymbol{a}\boldsymbol{Y} \sim MVN(\boldsymbol{a}\boldsymbol{\mu}, \boldsymbol{a}\Sigma\boldsymbol{a}^\top)$$

$$A\boldsymbol{Y} \sim MVN(A\boldsymbol{\mu}, A\Sigma A^\top)$$

(2) Marginal distribution of Y_i is Normal,

$$Y_i \sim N(\mu_i, \Sigma_{ii})$$

In fact, any subset of Y_i's is MVN

(3) Conditional MVN is MVN, e.g. $Y_1 \mid Y_2, \dots, Y_n$

(4) Another property:

$$\text{Cov}(Y_i, Y_j) = 0 \iff Y_i, Y_j \textit{ independent}$$

that is, Y_i and Y_j are uncorrelated.

$$\Sigma_{ij} = 0$$

Lecture 6 | 2020-09-23

Recall that last lecture, for a MLR, we have $\boldsymbol{Y} = X\boldsymbol{B} + \boldsymbol{\varepsilon}$ with the assumption that $\varepsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Therefore, for a random vector $\boldsymbol{\varepsilon}$, we have

$$\boldsymbol{\varepsilon} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 & 0 \\ 0 & \sigma^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 & 0 \\ 0 & 0 & \cdots & 0 & \sigma^2 \end{bmatrix} \right) = (\boldsymbol{0}_{n\times 1}, \sigma^2 I_{n\times n})$$

since $\text{Cov}(\varepsilon_1, \varepsilon_2) = 0$ due to independence.

Thus, $\boldsymbol{Y} \sim \text{MVN}(X\boldsymbol{B}, \sigma^2 I)$.

DEFINITION 3.11: Least squares for MLR

We define the **least squares for a multiple linear regression model** as

$$S(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^{n} (y_i - \underbrace{(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}_{\mathbb{E}[Y_i]=\mu_i})^2$$

THEOREM 3.12: Least Square Estimates (LSEs) for MLR

Minimizing $S(\beta_0, \beta_1, \dots, \beta_p)$, gives the least squares estimate $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}$.

Proof of: 3.12

The first partial is $\frac{\partial S}{\partial \beta_0} = \sum\limits_{i=1}^{n} 2(y_i - \mu_u)(-1)$, and all other partials for $j = 1, \dots, p$ are

$$\frac{\partial S}{\partial \beta_j} = \sum_{i=1}^{n} 2(y_i - \mu_i)(-x_{ij})$$

Set $\dfrac{\partial S}{\partial \beta_0} = 0$ and $\dfrac{\partial S}{\partial \beta_j} = 0$ for $j = 1, \dots, p$ to get

$$\begin{cases} \displaystyle\sum_{i=1}^{n} (y_i - \mu_i) = 0 \iff \boldsymbol{1}_{n\times n}^\top (\boldsymbol{y} - \boldsymbol{\mu}) = 0 \\ \displaystyle\sum_{i=1}^{n} (y_i - \mu_i) x_{ij} = 0 \iff \boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{\mu}) = 0 \quad j = 1, \dots, p \end{cases}$$

since we recall that

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \boldsymbol{1}_{n\times 1} & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{p-1} & \boldsymbol{x}_p \end{bmatrix}$$

Therefore,

$$X^\top (\boldsymbol{y} - X\boldsymbol{B}) = 0 \iff X^\top \boldsymbol{y} - X^\top X\boldsymbol{B} = 0 \iff X^\top X \boldsymbol{B} = X^\top \boldsymbol{y} \iff \boldsymbol{B} = (X^\top X)^{-1} X^\top \boldsymbol{y}$$

assuming $X^\top X$ is invertible (full rank of $p+1$ or linearly independent columns). So, the LS solution for $\boldsymbol{B}$ is given by $\hat{\boldsymbol{B}} = (X^\top X)^{-1} X^\top \boldsymbol{y}$.

DEFINITION 3.13: Residuals for MLR

The **residuals** for a multiple linear regression model is defined as

$$e_i = y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots \hat{\beta}_p x_{ip})}_{\text{fitted value } \mu_i}$$

or equivalently, $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{B}}$ and $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}$.

The estimate σ^2 based on e_i's is

$$\hat{\sigma}^2 = \frac{\text{SS(Res)}}{n-(p+1)} = \frac{\sum_{i=1}^n e_i^2}{n-p-1} = \frac{\boldsymbol{e}^\top \boldsymbol{e}}{n-p-1}$$

since d.f. is $n - (\text{no. estimated parameters})$. When viewed as a random variable,

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$$

Inference for $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_p)^\top = (X^\top X)^{-1} X^\top \boldsymbol{Y}$.

Note that $\hat{\boldsymbol{\beta}}$ is a matrix of constants and $\boldsymbol{Y}$ is a random vector, and $\boldsymbol{Y} \sim \text{MVN}(X\boldsymbol{\beta}, \sigma^2 I)$, so

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(X^\top X)^{-1} X^\top \boldsymbol{Y}] \\ &= (X^\top X)^{-1} X^\top \mathbb{E}[\boldsymbol{Y}] \\ &= (X^\top X)^{-1} (X^\top X)\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

That is, $\mathbb{E}[\hat{\beta}_0], \dots, \mathbb{E}[\hat{\beta}_p] = \beta_p$ all unbiased.

$$\begin{aligned} \mathbb{V}((X^\top X)^{-1} X^\top \boldsymbol{Y}) &= (X^\top X)^{-1} X^\top \mathbb{V}(\boldsymbol{Y}) \left[(X^\top X)^{-1} X^\top\right]^\top \\ &= (X^\top X)^{-1} X^\top \sigma^2 I (X^\top)^\top \left[(X^\top X)^{-1}\right]^\top && X^\top X \text{ symmetric} \\ &= \sigma^2 (X^\top X)^{-1} (X^\top X)(X^\top X)^{-1} \end{aligned}$$

Since $\hat{\boldsymbol{\beta}}$ is a linear transformation of $\boldsymbol{Y}$ we have $\hat{\boldsymbol{\beta}} \sim \text{MVN}(\boldsymbol{\beta}, \sigma^2 \underbrace{(X^\top X)^{-1}}_{V})$. We proved the following theorem.

THEOREM 3.14: Distribution of $\hat{\beta}_j$

The distribution of a given $\hat{\beta}_j$ is

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 V_{jj})$$

from marginal property of MVN.

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{V_{jj}}} \sim N(0,1)$$

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{V_{jj}}} \sim t(n-p-1)$$

DEFINITION 3.15: Standard error for $\hat{\beta}_j$

We define the **standard error** of $\hat{\beta}_j$ as

$$\text{Se}(\hat{\beta}_j) = \hat{\sigma}\sqrt{V_{jj}}$$

So, a $(1-\alpha)$ confidence interval for β_j is

$$\hat{\beta}_j \pm c\mathsf{Se}(\hat{\beta}_j)$$

where c is $(1-(\alpha/2))$ quantile of $t(n-p-1)$.

To test $H_0\colon \beta_j = 0$ vs $H_A\colon \beta_j \neq 0$, calculate t-statistic $t = \dfrac{\hat{\beta}_j}{\mathsf{Se}(\hat{\beta}_j)}$ reject at level α if $|t| > c$ and p-value is $2P(T \geqslant |t|)$ where $T \sim t(n-p-1)$.

Interpretation of $\hat{\boldsymbol{\beta}}$: fitted linear regression model says $\widehat{\mathbb{E}[Y]}$ (estimate of the expected response) is $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$.

- $\hat{\beta}_0$ is the estimate of expected response when all explanatory variables are equal to 0.
- $\hat{\beta}_j$ is the estimated change in expected response for a unit increase in x_j, when holding all other explanatory variables constant, e.g.

$$\hat{\beta}_0 + \hat{\beta}_1(x_1+1) + \cdots + \hat{\beta}_p x_p - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p) = \hat{\beta}_1$$

REMARK 3.16

When it's written V_{jj}, that means the $j+1^{\text{th}}$ column and $j+1^{\text{th}}$ row since we start from index 0 for these matrices. Some unfortunate events may have happened on the quiz to me due to this.

EXAMPLE 3.17: Rocket MLR

Let $n=12$, $\hat{\boldsymbol{\beta}} = (473.6, 16.7, -1.09)^\top = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)^\top$.

- x_1: nozzle area $(1 = L, 0 = S)$
- x_2: propellant ratio
- Y: thrust

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{12} e_i^2}{12-1-2}} = \sqrt{\frac{\boldsymbol{e}^\top \boldsymbol{e}}{9}} = 2.655$$

Interpretation of $\hat{\boldsymbol{\beta}}$:

- $\hat{\beta}_1$ estimated change in expected thrust is 16.7 when changing small to large nozzle while holding other variables (propellant ratio) constant.
- $\hat{\beta}_2$ estimated thrust to decrease by 1.09 on average for a unit increase in propellant ratio while holding other variables (nozzle area) constant.

Given $\mathsf{Se}(\hat{\beta}_2) = 0.94$, we compute the t-statistic for $H_0\colon \beta_2 = 0$ vs $H_A\colon \beta_2 \neq 0$ which is $t = -1.09/0.94 = -1.16$.

$$p\text{-value} = 2P(T \geqslant 1.16) = 0.275 \text{ from R where } T \sim t(9)$$

Do not reject H_0 (e.g. $\alpha = 0.05$), therefore propellent ratio does not significantly influence thrust.

Lecture 7 | 2020-09-28

Recall that $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim \text{MVN}(X\boldsymbol{\beta}, \sigma^2 I)$, and

- Estimates: $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{Y}$
- Fitted values: $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$
- Residuals: $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}$
- Constants: $X = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_p \end{bmatrix}_{n\times(p+1)}$

- Values of responses: $\boldsymbol{y} = (y_1, y_2, \dots, y_n)^\top \in \mathbb{R}^n$

Author's Note: Geometric interpretation of data is omitted in these notes because I'm simply too lazy.

The span of X is $\text{Span}(X) = \{b_0\mathbf{1} + b_1\boldsymbol{x}_1 + \dots + b_p\boldsymbol{x}_p : b_0, \dots, b_p \in \mathbb{R}\} \subset \mathbb{R}^n$ which is all linear combinations of columns of X which is a subspace of $\mathbb{R}^n$, and by assumption we know $\text{rank}(X) = p + 1$.

We can say $\text{Span}(X)$ represents all possible vector values $X\boldsymbol{b}$ where $\boldsymbol{b} = (b_0, b_1, \dots, b_p)^\top$.

Generally, $\boldsymbol{y} \notin \text{Span}(X)$, so since the linear model is an approximation, ε variability not explained by model.

Intuitively, it makes sense to choose an estimate $\hat{\boldsymbol{\beta}}$ so that $X\hat{\boldsymbol{\beta}}$ is as close to $\boldsymbol{y}$ as possible. Therefore, $\boldsymbol{e}$ must be orthogonal to $\text{Span}(X) \iff \boldsymbol{e}$ is orthogonal to all columns of X.

$$\begin{aligned}
\mathbf{1}^\top \cdot (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) &= 0 \\
\boldsymbol{x}_1^\top \cdot (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) &= 0 \\
&\vdots \\
\boldsymbol{x}_p^\top \cdot (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) &= 0
\end{aligned}$$

which is the same as LS estimates. We also know $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$ and $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}$.

DEFINITION 3.18: Hat matrix

The **hat matrix** is defined as $H = X(X^\top X)^{-1}X^\top$.

PROPOSITION 3.19: Properties of Hat Matrix

Let H be a hat matrix, then H has the following properties.
(1) H is symmetric; that is, $H = H^\top$.
(2) H is idempotent; that is, $H^2 = HH = H$.
(3) $I - H$ is symmetric idempotent; that is, $(I - H)^2 = (I - H)(I - H) = I - H$.

Proof of: 3.19

We prove all three because it's easy.
(1) $H^\top = [X(X^\top X)^{-1}X^\top]^\top = X(X^\top X)^{-1}X^\top = H$.
(2) $HH = X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top = H$.
(3) $(I-H)(I-H) = I(I-H) - H(I-H) = II - IH - HI + HH = I - 2H + HH = I - 2H + H = I - H$.

Let's view $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{e}$ as random vectors

$$\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}} = X(X^\top X)^{-1}X^\top\boldsymbol{Y} = H\boldsymbol{Y}$$

$$\begin{gathered}
\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{\mu}} = I\boldsymbol{Y} - H\boldsymbol{Y} = (I - H)\boldsymbol{Y} \\
\mathbb{E}[\hat{\boldsymbol{\mu}}] = \mathbb{E}[H\boldsymbol{Y}] = H\mathbb{E}[\boldsymbol{Y}] = X(X^\top X)^{-1}X^\top \underbrace{X\boldsymbol{\beta}}_{\mathbb{E}[\boldsymbol{Y}]} = X\boldsymbol{\beta} \\
\mathbb{V}(\hat{\boldsymbol{\mu}}) = \mathbb{V}(H\boldsymbol{Y}) = H\mathbb{V}(\boldsymbol{Y})H^\top = H\sigma^2 IH^\top = \sigma^2(HH^\top) = \sigma^2 H \\
\mathbb{E}[\boldsymbol{e}] = \mathbb{E}[(I - H)\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{Y}] - \mathbb{E}[H\boldsymbol{Y}] = X\boldsymbol{\beta} - X\boldsymbol{\beta} = 0 \\
\mathbb{V}(\boldsymbol{e}) = (I - H)\mathbb{V}(\boldsymbol{Y})(I - H)^\top = \sigma^2(I - H)(I - H)^\top = \sigma^2(I - H)
\end{gathered}$$

So since $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{e}$ are linear transformations of $\boldsymbol{Y}$ we have proved the following theorem.

THEOREM 3.20: Distribution of $\hat{\boldsymbol{\mu}}$ and e

$\hat{\boldsymbol{\mu}}$ and $\hat{e}$ have the following distribution.

$$\hat{\boldsymbol{\mu}} \sim MVN(X\boldsymbol{\beta}, \sigma^2 H)$$

$$\hat{\boldsymbol{e}} \sim MVN(0, \sigma^2(I-H))$$

Suppose we want to predict response for $\boldsymbol{x}_0$ where the first 1 represents the intercept in the row vector.

$$\boldsymbol{x}_0 = \begin{bmatrix} 1 & x_{01} & x_{02} & \cdots & x_{0p} \end{bmatrix}_{1\times(p+1)}$$

Let Y_0 random variable representing the response associated with $\boldsymbol{x}_0$. The MLR says

$$Y_0 \sim N(\beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}, \sigma^2)$$

So we predict the value

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p} = \boldsymbol{x}_0\hat{\boldsymbol{\beta}}$$

which represents the estimated mean response given $x_{01}, x_{02}, \ldots, x_{0p}$. Corresponding distribution has

$$\mathbb{E}[\hat{Y}_0] = \boldsymbol{x}_0\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{x}_0\boldsymbol{\beta} = \mathbb{E}[Y_0]$$

$$\mathbb{V}(\hat{Y}_0) = \boldsymbol{x}_0\mathbb{V}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_0^\top = \boldsymbol{x}_0\sigma^2(X^\top X)^{-1}\boldsymbol{x}_0^\top$$

We have proved the following theorem.

THEOREM 3.21: Distribution of Predictor

The distribution of $\hat{Y}_0$ which is a function of $Y_1, \ldots, Y_n$ is

$$\hat{Y}_0 \sim N(\boldsymbol{x}_0\boldsymbol{\beta}, \sigma^2\boldsymbol{x}_0(X^\top X)^{-1}\boldsymbol{x}_0^\top)$$

$$\frac{\hat{Y}_0 - \boldsymbol{x}_0\boldsymbol{\beta}}{\sigma\sqrt{\boldsymbol{x}_0(X^\top X)^{-1}\boldsymbol{x}_0^\top}} \sim N(0,1)$$

$$\frac{\hat{Y}_0 - \boldsymbol{x}_0\boldsymbol{\beta}}{\hat{\sigma}\sqrt{\boldsymbol{x}_0(X^\top X)^{-1}\boldsymbol{x}_0^\top}} \sim t(n-(p+1)) = t(n-p-1)$$

A $(1-\alpha)$ confidence interval for the mean response $y_0 = \boldsymbol{x}_0\hat{\boldsymbol{\beta}}$ given $\boldsymbol{x}_0$ is

$$\hat{y}_0 \pm c\hat{\sigma}\sqrt{\boldsymbol{x}_0(X^\top X)^{-1}\boldsymbol{x}_0^\top}$$

where c is the $1-\alpha/2$ quantile of $t(n-p-1)$.

Prediction error: $Y_0 - \hat{Y}_0$ which are independent since Y_0 is a random variable with variance σ^2 and $\hat{Y}_0$ is a function of $Y_1, \ldots, Y_n$. Therefore,

$$\mathbb{E}[Y_0 - \hat{Y}_0] = \boldsymbol{x}_0\boldsymbol{\beta} - \boldsymbol{x}_0\boldsymbol{\beta} = 0$$

$$\mathbb{V}(Y_0 - \hat{Y}_0) = \mathbb{V}(Y_0) + (-1)^2\mathbb{V}(\hat{Y}_0) = \sigma^2 + \sigma^2(\boldsymbol{x}_0(X^\top X)^{-1}\boldsymbol{x}_0^\top)$$

We have proved the following theorem.

THEOREM 3.22: Distribution of Prediction Error

The distribution of the prediction error is

$$Y_0 - \widehat{Y}_0 \sim N(0, \sigma^2(1 + \boldsymbol{x}_0(X^\top X)^{-1}\boldsymbol{x}_0^\top))$$

A $(1-\alpha)$ prediction interval for the mean response $y_0 = \boldsymbol{x}_0\hat{\boldsymbol{\beta}}$ given $\boldsymbol{x}_0$ is

$$\hat{y}_0 \pm c\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0(X^\top X)^{-1}\boldsymbol{x}_0^\top}$$

where c is the $1-\alpha/2$ quantile of $t(n-p-1)$.

REMARK 3.23

Our intuition tells us that the prediction interval is wider than the confidence interval for mean. In other words, estimating an average is “easier” than an individual response.

Lecture 8 | 2020-09-30

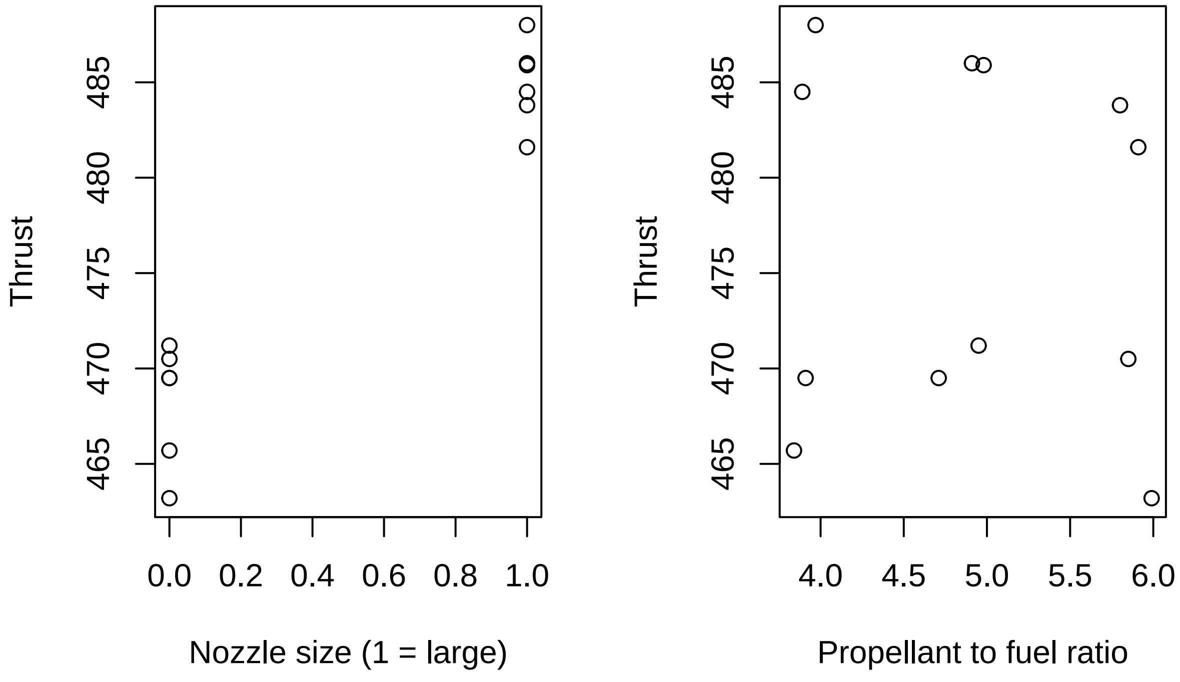
The example done in R is included in the next page.

```r
## NASA rocket data example

## From: R.S. Jankovsky, T.D. Smith, A.J. Pavli (1999). "High-Area-Ratio Rocket
## Nozzle at High Combustion Chamber Pressure-Experimental and Analytical
## Validation".

# setwd(...) first if your CSV file is somewhere else
rocket <- read.csv(file="rocket.csv")
# output all data in rocket vector
rocket
```

```
##    thrust nozzle propratio
## 1   488.0      1      3.97
## 2   481.6      1      5.91
## 3   485.9      1      4.98
## 4   486.0      1      4.91
## 5   484.5      1      3.89
## 6   483.8      1      5.80
## 7   463.2      0      5.99
## 8   471.2      0      4.95
## 9   469.5      0      3.91
## 10  470.5      0      5.85
## 11  469.5      0      4.71
## 12  465.7      0      3.84
```

Y (thrust) is the response variable, and there are two explanatory variables x_1, x_2 (nozzle, propratio) where nozzle is coded as 1 if it's large.

```r
# Scatter plots where mfrow is used to put multiple
# plots on one image
par(mfrow = c(1,2))
plot(rocket$nozzle, rocket$thrust, ylab="Thrust", xlab="Nozzle size (1 = large)")
plot(rocket$propratio, rocket$thrust, ylab="Thrust", xlab="Propellant to fuel ratio")
```



Left is

nozzle size vs thrust. Right is propellant relationship vs thrust.

```
# Fit MLR using lm
m1 <- lm(thrust ~ nozzle + propratio, data = rocket)
summary(m1)
```

```
##
## Call:
## lm(formula = thrust ~ nozzle + propratio, data = rocket)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8459 -1.7555  0.5934  1.2906  3.3008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 473.6039     4.7158 100.430 4.88e-15 ***
## nozzle       16.7383     1.5329  10.919 1.71e-06 ***
## propratio    -1.0948     0.9414  -1.163    0.275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.655 on 9 degrees of freedom
## Multiple R-squared:  0.9303, Adjusted R-squared:  0.9148
## F-statistic: 60.05 on 2 and 9 DF,  p-value: 6.238e-06
```

On the left it's Y (response variable) and on the right it's x_1, x_2 (explanatory variables). From summary, we get the estimate vector $\hat{\beta} = (473.6039, 16.7383, -1.0948)^\top$.

```
# Manual beta estimates where rep is used to make the columns of 1s
X <- cbind(rep(1, 12), rocket$nozzle, rocket$propratio) # X matrix
y <- matrix(rocket$thrust, ncol = 1) # response vector
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
beta_hat
```

```
##            [,1]
## [1,] 473.603924
## [2,]  16.738319
## [3,]  -1.094822
```

`solve` is used for the inverse. `%*%` is used for matrix-matrix multiplication, and `t(X)` is used for transposing X.

```
# Manual sigma estimate
mu_hat <- X %*% beta_hat # fitted values
e <- y - mu_hat # residuals
sigma_hat <- sqrt((t(e) %*% e) / 9) # Note n-p-1 = 12-2-1 = 9
sigma_hat
```

```
##        [,1]
## [1,] 2.6545
```

```
sigma_hat <- sqrt( sum(e^2) / 9) # equivalent
sigma_hat
```

```
## [1] 2.6545
```

- $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$

- $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}$
- $\hat{\sigma} = \sqrt{\left(\sum_{i=1}^{n} e_i^2\right)/9} = 2.6545$, or
- $\hat{\sigma} = \sqrt{(\boldsymbol{e}^\top \boldsymbol{e})/9} = 2.6545$

```
# Covariance matrix of beta_hat
vcov(m1)
```

```
##             (Intercept)      nozzle   propratio
## (Intercept)   22.238325 -1.02316688 -4.32080608
## nozzle        -1.023167  2.34987593 -0.03102117
## propratio     -4.320806 -0.03102117  0.88631920
```

```
sqrt(diag(vcov(m1))) # SEs of individual betas
```

```
## (Intercept)      nozzle   propratio
##   4.7157528   1.5329305   0.9414453
```

```
# Manual
se_beta <- sigma_hat * sqrt(diag(solve(t(X) %*% X)))
se_beta
```

```
## [1] 4.7157528 1.5329305 0.9414453
```

- $Se(\hat{\boldsymbol{\beta}}) = \hat{\sigma}\sqrt{(X^\top X)^{-1}} = (4.71, 1.53, 0.94)^\top$

```
# Estimate the mean response for units with small nozzle and propellant ratio 5.5
# include a 95% CI
predict(object = m1, newdata = data.frame(nozzle = 0, propratio = 5.5),
        interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 467.5824 464.7929 470.3719
```

Therefore, $\hat{y}_0 = 467.58$. The 95% confidence interval for the mean response given $\boldsymbol{x}_0$ is $[464.7929, 470.3719]$.

```
# Manual calculation
x0 <- matrix(c(1, 0, 5.5), nrow = 1)
y0_hat <- x0 %*% beta_hat
y0_hat
```

```
##          [,1]
## [1,] 467.5824
```

```
# mu0 is also known as \hat{Y}_0
se_mu0 <- sigma_hat * sqrt(x0 %*% solve(t(X) %*% X) %*% t(x0))
se_mu0
```

```
##          [,1]
## [1,] 1.233132
```

```
crit_val <- qt(0.975,9)
ci_lo <- y0_hat - crit_val*se_mu0
ci_hi <- y0_hat + crit_val*se_mu0
c(y0_hat, ci_lo, ci_hi)
```

```
## [1] 467.5824 464.7929 470.3719
```

- $\boldsymbol{x}_0 = \begin{bmatrix} 1 & 0 & 5.5 \end{bmatrix}$

- $\hat{y}_0 = \boldsymbol{x}_0\hat{\boldsymbol{\beta}} = 467.5824$
- $Se(\hat{Y}_0) = \hat{\sigma}\sqrt{\boldsymbol{x}_0(X^\top X)^{-1}\boldsymbol{x}_0^\top} = 1.233132$

Therefore, $\hat{y}_0 = 467.58$. The 95% confidence interval for the mean response given $\boldsymbol{x}_0$ is $[464.7929, 470.3719]$.

```
# Predict the value of the response for a unit with small nozzle and propellant ratio 5.5
# include a 95% PI
predict(object = m1, newdata = data.frame(nozzle = 0, propratio = 5.5),
        interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 467.5824 460.9612 474.2036
```

Therefore, $y_0 = 467.5824$. The 95% prediction interval for the response (y_0) given $\boldsymbol{x}_0$ is $[460.9612474.2036]$.

```
# Manual calculation for an individual
x0 <- matrix(c(1, 0, 5.5), nrow = 1)
y0_hat <- x0 %*% beta_hat
se_y0 <- sigma_hat * sqrt(1+ x0 %*% solve(t(X) %*% X) %*% t(x0))
se_y0
```

```
##          [,1]
## [1,] 2.926941
```

```
crit_val <- qt(0.975,9)
pi_lo <- y0_hat - crit_val*se_y0
pi_hi <- y0_hat + crit_val*se_y0
c(y0_hat, pi_lo, pi_hi)
```

```
## [1] 467.5824 460.9612 474.2036
```

- $Se(Y_0 - \hat{Y}_0) = \hat{\sigma}\sqrt{1 + \boldsymbol{x}_0(X^\top X)^{-1}\boldsymbol{x}_0^\top} = 2.926941$

Handling categorical variables: when there are explanatory variables with values that fall into one of several categories.

- e.g. nozzle large/small, if just binary, code as 1 and 0
- ordered small, medium, large or not red, blue green

Approach: can convert to indicator variables or treat as numerical if it makes sense to do so.

Example: CQI (2018)

Extract a few variables:

	Acidity	Method
1	8.7	Washed-wet
2	8.3	Washed-wet
3	8.2	Natural-dry
4	8.4	Semi-washed/pulped

Flavour (response)

How to set up X? For example,

$$x_{i2} = \begin{cases} 0 & \text{dry} \\ 1 & \text{semi} \\ 2 & \text{wet} \end{cases}$$

Not generally appropriate unless we think a response is linear according to this scheme.

More flexible approach: indicator/dummy variables

$$x_{i2} = \begin{cases} 1 & \text{semi} \\ 0 & \text{otherwise} \end{cases}, \quad x_{i3} = \begin{cases} 1 & \text{wet} \\ 0 & \text{otherwise} \end{cases}$$

Therefore,

$$X = \begin{bmatrix} 1 & 8.7 & 0 & 1 \\ 1 & 8.3 & 0 & 1 \\ 1 & 8.2 & 0 & 0 \\ 1 & 8.4 & 1 & 0 \end{bmatrix}$$

Why not $x_{i4} = \begin{cases} 1 & \text{dry} \\ 0 & \text{otherwise} \end{cases}$? If we did that, we would have

$$X = \begin{bmatrix} 1 & 8.7 & 0 & 1 & 0 \\ 1 & 8.3 & 0 & 1 & 0 \\ 1 & 8.2 & 0 & 0 & 1 \\ 1 & 8.4 & 1 & 0 & 0 \end{bmatrix}$$

This has linearly dependent columns since $\boldsymbol{x}_4 = \boldsymbol{1} - \boldsymbol{x}_2 - \boldsymbol{x}_3$. There is no new information and X would not have full rank.

Model: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$.

Interpretation:

- mean flavour if acidity $= x_{01}$ and method dry is $\beta_0 + \beta_1 x_{01}$.
- mean flavour if acidity $= x_{01}$ and method wet is $\beta_0 + \beta_1 x_{01} + \beta_3$.
- mean flavour if acidity $= x_{01}$ and method semi is $\beta_0 + \beta_1 x_{01} + \beta_2$.

- β_2 is the difference between semi and dry in expected response (holding acidity constant)
- β_3 is the difference between wet and dry in expected response (holding acidity constant)
- $\beta_2 - \beta_3$ is the difference between semi and wet (holding other variables constant)

$\hat{\boldsymbol{\beta}} \sim \text{MVN}(\boldsymbol{\beta}, \sigma^2 V)$ where $V = (X^\top X)^{-1}$.

- We know $\hat{\beta}_j \sim N(\beta_j, \sigma^2 V_{jj})$ with $\mathsf{Se}(\hat{\beta}_j) = \hat{\sigma}\sqrt{V_{jj}}$ where $j = 0, \ldots, p$.
- What about $\beta_2 - \beta_3$?

$$\mathbb{V}(\hat{\beta}_2 - \hat{\beta}_3) = \mathbb{V}(\hat{\beta}_2) - \mathbb{V}(\hat{\beta}_3) - 2\mathsf{Cov}(\hat{\beta}_2, \hat{\beta}_3) = \sigma^2 V_{22} + \sigma^2 V_{33} - 2\sigma^2 V_{23}$$

Therefore,

$$\mathsf{Se}(\hat{\beta}_2 - \hat{\beta}_3) = \hat{\sigma}\sqrt{V_{22} + V_{33} - 2V_{23}}$$

Now, we can construct a CI for $\beta_2 - \beta_3$.

In general, for an explanatory variable with k categories. We need $k-1$ indicator variables.

Lecture 9 | 2020-10-05

Analysis of variance (ANOVA): how well does our regression model fit our response variable?

Variability in response can be measured by "total sum of squares:"

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

as seen in HW1, it's closely related to sample variance of $y_1, \ldots, y_n$, which is $\text{SS(Total)}/(n-1)$.

ANOVA decomposes $\text{SS(Total)} = \text{SS(Reg)} + \text{SS(Res)}$ where SS(Reg) is the regression sum of squares and SS(Res) is the residual sum of squares.

The regression sum of squares is variation explained by the model and the residual sum of squares is the variation not explained by the regression model.

Using the fact that

$$y_i - \bar{y} = y_i - \hat{\mu}_i + \hat{\mu}_i - \bar{y}$$

When regression fits data well, the observations y_i tend to be much closer to $\hat{\mu}_i$. Note that $\bar{y}$ is line a regression line with $\beta_1 = 0$.

Mathematically,

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\text{SS(Total)}} = \underbrace{\sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2}_{\text{SS(Reg)}} + \underbrace{\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2}_{\text{SS(Res)}}$$

since we showed that $\sum_{i=1}^{n} (\hat{\mu}_i - \bar{y}) \underbrace{(y_i - \hat{\mu}_i)}_{e_i} = 0$ in HW1 for SLR. It's also true for MLR since

$$\sum_{i=1}^{n} (\hat{\mu}_i - \bar{y}) e_i = \sum_{i=1}^{n} (e_i \hat{\mu}_i) - \bar{y} \sum_{i=1}^{n} e_i = \hat{\boldsymbol{\mu}}^\top \boldsymbol{e} - \bar{y} \mathbf{1}^\top \boldsymbol{e} = 0$$

Recall: $\mathbf{1}^\top \boldsymbol{e} = 0$ is one of LS equations, and $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$ is in $\text{Span}(X)$, so $\boldsymbol{e}$ is orthogonal to $\text{Span}(X)$, so $\hat{\boldsymbol{\mu}}^\top \boldsymbol{e} = 0$.

F is used to test the overall significance of regression (later).

Table 1: ANOVA Table

Source	d.f.	SS	Mean Square	F
Regression	p	SS(Reg)	SS(Reg)$/p$	MS(Reg)/MS(Res)
Residual	$n-p-1$	SS(Res)	SS(Res)$/(n-p-1) = \hat{\sigma}^2$	
Total	$n-1$	SS(Total)		

We call the **coefficient of determination** R^2 = SS(Reg)/SS(Total) = $1-$ SS(Res)/SS(Total). clearly, $0 \leqslant R^2 \leqslant 1$. It is the proportion of variation (in our response variable) that is explained by the regression model. Larger R^2 means the fitted values are closer to the observations y_i, which means the residuals are small; that is, smaller SS(Res). Note that (HW1) in SLR, R^2 is equivalent to the square of the sample correlation between x and y based on $(x_1, y_1), \dots, (x_n, y_n)$.

Table 2: Rocket ANOVA Table

Source	d.f.	SS	Mean Square	F
Regression	2	846.2	423.1	60
Residual	9	63.42	7.05	
Total	11	909.62		

Response thrust $R^2 = 846.2/909.62 \approx 0.93$. R^2 interpretation: regression model with nozzle size and propellant ratio explains 93% of variation in thrust (response).

Lecture 10 | 2020-10-07

Hypothesis testing based on F distribution

So far we've tested H_0: $\beta_j = 0$ vs H_A: $\beta_j \neq 0$ involving individual parameters, using t distribution.

Now consider hypothesis test of the form H_0: $A\boldsymbol{\beta} = \mathbf{0}$ where A is a matrix of constraints specifying linear combinations of parameters.

EXAMPLE 3.24: Coffee Continued

The full model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

- Y_i is the flavour
- x_{i1} is acidity
- x_{i2} is 1 if semi, and 0 otherwise.
- x_{i3} is 1 if wet, and 0 otherwise.

Example 1.

- H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ versus
- H_A: at least one of $\beta_1, \beta_2, \beta_3$ not 0.
- If H_0 is true, the model reduces to $Y_i = \beta_0 + \varepsilon_i$.
- This tests overall significance of regression (whether any of predictors impact response)
- $A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. Note that row i considers the constraint of $\beta_i = 0$ for $i = 1, 2, 3$ in this example.

Example 2.

- H_0: $\beta_2 = \beta_3 = 0$

- If H_0 is true, $Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$
- Q: Is reduced model with only acidity plausible?
- $A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. Note that $A\boldsymbol{\beta} = \mathbf{0}_{1\times 2}$

Example 3.

- H_0: $\beta_2 - \beta_3 = 0$
- H_A: $\beta_2 \neq \beta_3$
- $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2(x_{i2} + x_{i3}) + \varepsilon_i$ where $(x_{i2} + x_{i3})$ is 1 if semi/wet and 0 if dry.
- Do the wet and semi methods have the same impact on the response (holding acidity constant)?
- $A = \begin{bmatrix} 0 & 0 & 1 & -1 \end{bmatrix}$

In general, with ℓ constraints. A is an $\ell \times (p+1)$ matrix with rank ℓ. Recall that

$$\text{Span}(X) = \{\beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_p \boldsymbol{x}_p\}$$

Let

$$\text{Span}(X)_A = \{\beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_p \boldsymbol{x}_p : A\boldsymbol{\beta} = 0\}$$

which is a subspace of Span(X) since any vector in Span$(X)_A$ is also in Span(X). We call Span$(X)_A$ the Span(X) with constraint A on $\boldsymbol{\beta}$.

Let $\hat{\boldsymbol{\mu}}_A$ denote the fitted values from fitting the reduced model. The residual if we hit the model with $A\boldsymbol{\beta} = \mathbf{0}$ is $\boldsymbol{e}_A = \boldsymbol{y} - \hat{\boldsymbol{\mu}}_A$.

If H_0: $A\boldsymbol{\beta} = \mathbf{0}$ is true, then $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}_A$ should be close; that is, the model makes similar predictions whether we set $A\boldsymbol{\beta} = \mathbf{0}$ or not when fitting the model.

So to assess whether H_0 is plausible, look at $\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|$ where $\|\cdot\|$ is Euclidean or L_2 norm. That is,

$$\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\| = \sqrt{(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A)^\top (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A)}$$

If it's "large" or "small" (close to 0) where large gives evidence against H_0 and small gives evidence for H_0.

By Pythagoras,

$$\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}_A\|^2 = \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2 + \|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2 \quad \text{or} \quad \|\boldsymbol{e}_A\|^2 = \|\boldsymbol{e}\|^2 + \|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2$$

or equivalently $\boldsymbol{e}_A^\top \boldsymbol{e}_A = \boldsymbol{e}^\top \boldsymbol{e} + \|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2$ where $\boldsymbol{e}_A^\top \boldsymbol{e}_A$ is the sum of squares residual in the reduced model and $\boldsymbol{e}^\top \boldsymbol{e}$ is the sum of squares residual in the full model.

We define $\boldsymbol{e}_A^\top \boldsymbol{e}_A = \text{SS(Res)}_A$ and $\boldsymbol{e}^\top \boldsymbol{e} = \text{SS(Res)}$.

Thus, $\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2 = \text{SS(Res)}_A - \text{SS(Res)} \geqslant 0$ additional sum of squares explained by full model vs reduced one with constraints A.

Practical implications:

- SS(Res) cannot decrease when constraints applied.
- Equivalently, full model always has small (or equal) SS(Res) for a fixed SS(Tot) and thus higher R^2 compared to a reduced model.

Define test statistic:

$$F = \frac{(\text{SS(Res)}_A - \text{SS(Res)})/\ell}{\text{SS(Res)}/(n-p-1)} = \frac{(\text{SS(Res)}_A - \text{SS(Res)})/\ell}{\hat{\sigma}^2}$$

DEFINITION 3.25: F distribtion

If $U \sim \chi^2(a)$ and $V \sim \chi^2(b)$ are independent. We say F follows an F **distribution** if

$$F = \frac{U/a}{V/b}$$

and wite $F \sim F(a, b)$.

Here, we have these facts when H_0 is true

$$V = \frac{\hat{\sigma}^2(n-p-1)}{\sigma^2} \sim \chi^2(n-p-1)$$

$$U = \frac{\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2}{\sigma^2} \sim \chi^2(\ell)$$

where U and V are independent. Therefore,

$$F = \frac{\dfrac{\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2}{\sigma^2}\left(\dfrac{1}{\ell}\right)}{\dfrac{\hat{\sigma}^2(n-p-1)}{\sigma^2}\left(\dfrac{1}{n-p-1}\right)} \sim F(\ell, n-p-1)$$

when H_0 is true. Reject H_0: $A\boldsymbol{\beta} = \mathbf{0}$ at level α if F is greater than $(1-\alpha)$ quantile of $F(\ell, n-p-1)$ and p-value is $P(Y \geqslant F)$ where $Y \sim F(\ell, n-p-1)$.

Relation to T distribution: Say $Y \sim t(a)$

$$Y = \frac{Z}{\sqrt{U/a}}$$

where $Z \sim N(0,1)$ and $U \sim \chi^2(a)$ are independent. Squaring everything,

$$Y^2 = \frac{Z^2}{U/a}$$

and we know $Z^2 \sim \chi^2(1)$. Therefore, $Y^2 \sim F(1, a)$ (we divide by 1 in the numerator).

Thus, if our hypothesis test has one constraint, then F test is equal to t test of same hypothesis; for example, H_0: $\beta_1 = 0$ versus H_A: $\beta_1 \neq 0$.

Lecture 11 | 2020-10-19

Recall the general linear hypothesis: H_0: $A\boldsymbol{\beta} = \mathbf{0}$ vs H_A: $A\boldsymbol{\beta} \neq \mathbf{0}$ where A gives ℓ constraints.

$$F \text{ statistic} = \frac{(\text{SS(Res)}_A - \text{SS(Res)})/\ell}{\text{SS(Res)}/(n-p-1)} = \frac{(\text{SS(Res)}_A - \text{SS(Res)})/\ell}{\hat{\sigma}^2}$$

compare to $F(\ell, n-p-1)$.

Special case: overall test of significance

“are any predictors related to response?”

- H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
- H_A: $\beta_j \neq 0$ for at least one j

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & \ddots & & & \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \\ \beta_p \end{bmatrix}$$

If H_0 is true: $Y_i = \beta_0 + \varepsilon_i$ where $Y_i \sim N(\beta_0, \sigma^2)$.

Fit reduced model; that is, in this case estimate β_0 using least squares, minimize $\sum_{i=1}^n (y_i - \beta_0)^2$, which can be shown $\hat{\beta}_0 = \bar{y}$. So,

$$\text{SS(Res)}_A = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = \text{SS(Total)}$$

Then,

$$F = \frac{(\text{SS(Total)} - \text{SS(Res)})/p}{\text{SS(Res)}/(n-p-1)} = \frac{\text{SS(Reg)}/p}{\text{SS(Res)}/(n-p-1)} = \frac{\text{MS(Reg)}}{\text{MS(Res)}} \leftarrow F \text{ statistic on ANOVA table}$$

Lecture 12 | 2020-10-21

Multicollinearity: occurs when some explanatory variables have a **strong linear** relationship amongst themselves. For example, this might occur exactly

$$\boldsymbol{x}_3 = \alpha_0 \mathbf{1} + \alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2$$

in which case the columns of X would be **linearly dependent** and $X^\top X$ does not have an inverse. Practically, there is no new info including $\boldsymbol{x}_3$ when $\boldsymbol{x}_1, \boldsymbol{x}_2$ are in the model. **Approximately**,

$$\boldsymbol{x}_3 \approx \alpha_0 \mathbf{1} + \alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2$$

in which case the columns of X are close to being linearly dependent which cases $\mathbb{V}(\hat{\beta}_j)$ to be **inflated**, in turn leads to inaccurate confidence intervals and conclusions of hypothesis tests for the regression parameters, in practice. $\text{Se}(\hat{\beta}_j)$ when fitting models can change drastically when adding/removing variables from the model.

EXAMPLE 3.26: Hockey (NHL)

In the NHL we have Goals + Assists = Points. Suppose we want to predict a forward's salary. Define

- x_1 = Goals
- x_2 = Assists
- x_3 = Points

However, $x_3 = x_1 + x_2$ so we have exact multicollinearity.

EXAMPLE 3.27: Burmese Pythons in Florida (2017)

- y = fat content
- x_1 = mass
- x_2 = overall length
- x_3 = snout-to-vent length

It turns out that x_2 and x_3 are highly correlated. Including all variables in regression lead to inflated $\text{Se}(\hat{\beta}_2)$ and $\text{Se}(\hat{\beta}_3)$.

3.1 Detection of Multicollinearity

If two predictors are related

- Scatterplot matrix [all possible pairs of scatterplots b/w $y, x_1, x_2, \dots, x_p$]
- Correlation matrix (all pairwise correlations)

DEFINITION 3.28: Variance inflation factor

For multicollinearity between more than two predictors, we can define the **variance inflation factor** (VIF).

$$\text{VIF}_j = \frac{\mathbb{V}(\hat{\beta}_j)}{\mathbb{V}(\hat{\beta}_j^*)}$$

for $j = 1, \dots, p$, where $\hat{\beta}_j$ is the estimate of β_j with all predictors in the model, and $\hat{\beta}_j^*$ estimate of β_j based on regression with x_j only.

THEOREM 3.29

$$\text{VIF}_j \geqslant 1$$

Fit MLR of x_j in terms of other predictors; that is,

$$x_{ij} = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_{j-1} x_{i(j-1)} + \alpha_{j+1} x_{i(j+1)} + \cdots + \alpha_p x_{ip} + \varepsilon_{ij}$$

and compute R^2 for this model, call it R_j^2.

Intuition: if R_j^2 is close to 1, x_j is strongly related linearly to other predictors. It can be shown that

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

Values of VIF larger than 10 are taken as solid evidence of multicollinearity; that is, $R_j^2 > 0.9$.

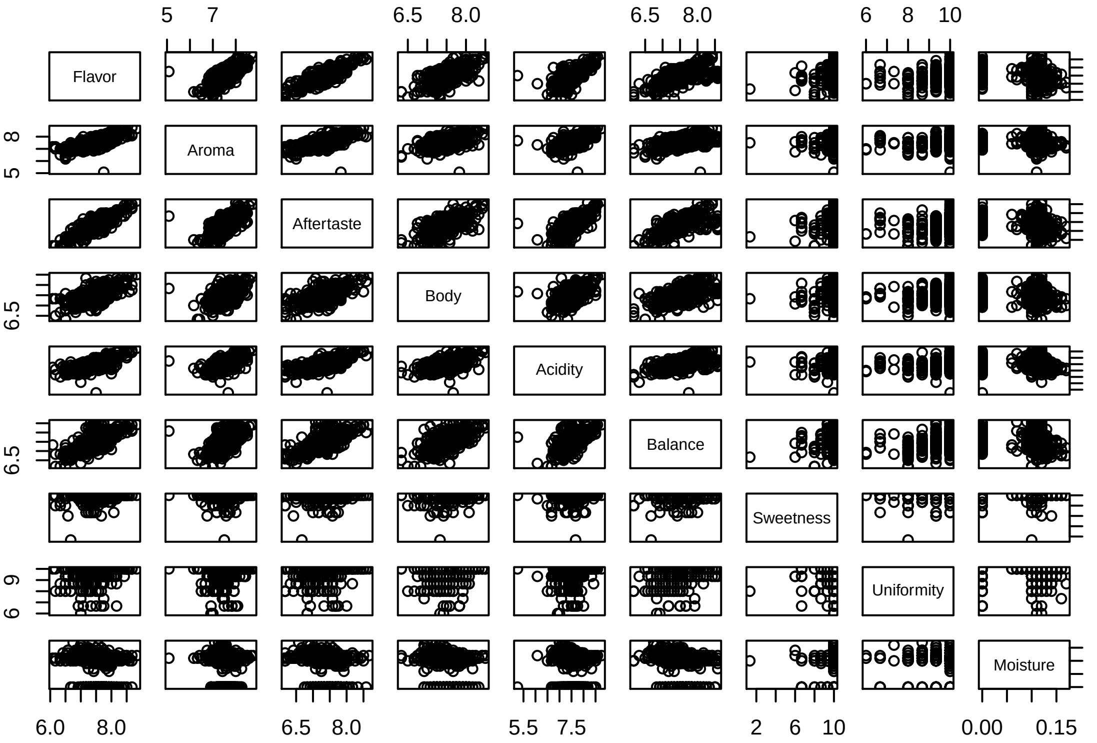
Procedure:

- remove predictors with largest VIF, if it exceeds 10
- repeat until no more multicollinearity

```
## Coffee example  (Coffee Quality Institute, 2018) continued
coffee <- read.csv("coffee_arabica.csv")

# cor(coffee) # doesn't work as there's a categorical variable
cor(coffee[,-1]) # e.g., remove first column
```

```
##                  Aroma     Flavor Aftertaste        Body     Acidity    Balance
## Aroma       1.00000000  0.7339782  0.6892744  0.56699932  0.60115765  0.6156508
## Flavor      0.73397820  1.0000000  0.8582783  0.67694834  0.73845546  0.7324530
## Aftertaste  0.68927440  0.8582783  1.0000000  0.67407704  0.69408861  0.7657979
## Body        0.56699932  0.6769483  0.6740770  1.00000000  0.60795391  0.6924568
## Acidity     0.60115765  0.7384555  0.6940886  0.60795391  1.00000000  0.6417994
## Balance     0.61565084  0.7324530  0.7657979  0.69245676  0.64179938  1.0000000
## Sweetness   0.06955938  0.1345364  0.1185760  0.03977892  0.06906093  0.1016718
## Uniformity  0.14785498  0.2132347  0.2143116  0.07195778  0.14876428  0.2180726
## Moisture   -0.11567549 -0.1327342 -0.1745366 -0.21009097 -0.10391684 -0.2161964
##             Sweetness Uniformity    Moisture
## Aroma      0.06955938 0.14785498 -0.11567549
## Flavor     0.13453644 0.21323472 -0.13273418
## Aftertaste 0.11857600 0.21431157 -0.17453658
## Body       0.03977892 0.07195778 -0.21009097
## Acidity    0.06906093 0.14876428 -0.10391684
## Balance    0.10167183 0.21807265 -0.21619640
## Sweetness  1.00000000 0.34756414  0.08049300
## Uniformity 0.34756414 1.00000000  0.02105693
## Moisture   0.08049300 0.02105693  1.00000000
```

```
# pairs without response: pairs(coffee[,-1])
# pairs with response, this is what we want
pairs(~ Flavor + Aroma + Aftertaste + Body +
        Acidity + Balance + Sweetness + Uniformity + Moisture, data=coffee)
```



```r
# Code our own indicators, so that we can more easily interpret VIFs
# 1 = wet, 0 otherwise
coffee$wet <- ifelse(coffee$Processing.Method == 'Washed / Wet', 1, 0)
# 1 = semi/dry, 0 otherwise
coffee$semi <- ifelse(coffee$Processing.Method == 'Semi-washed / Semi-pulped',
                      1, 0)
```

Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i(10)} + \varepsilon_i$$

where

- $y =$ flavour
- $x_1 = 1$ if wet, 0 otherwise
- $x_2 = 1$ if semi, 0 otherwise
- $x_3 =$ Aroma
- $x_4 =$ Aftertaste
- $x_5 =$ Body
- $x_6 =$ Acidity
- $x_7 =$ Balance
- $x_8 =$ Sweetness
- $x_9 =$ Uniformity

- x_{10} = Moisture

```
# Full MLR with our own coded indicators
mfull <- lm(Flavor~ wet + semi + Aroma + Aftertaste +
      Body + Acidity + Balance + Sweetness + Uniformity + Moisture, dat=coffee)
summary(mfull)
```

```
## 
## Call:
## lm(formula = Flavor ~ wet + semi + Aroma + Aftertaste + Body + 
##     Acidity + Balance + Sweetness + Uniformity + Moisture, data = coffee)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.68587 -0.08465  0.00079  0.08910  0.63633 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.728757   0.168516  -4.325 1.67e-05 ***
## wet         -0.033061   0.011024  -2.999  0.00277 ** 
## semi        -0.001396   0.022021  -0.063  0.94947    
## Aroma        0.220302   0.020447  10.774  < 2e-16 ***
## Aftertaste   0.468759   0.023912  19.603  < 2e-16 ***
## Body         0.096140   0.024334   3.951 8.28e-05 ***
## Acidity      0.216751   0.021194  10.227  < 2e-16 ***
## Balance      0.046806   0.022558   2.075  0.03823 *  
## Sweetness    0.025507   0.010150   2.513  0.01211 *  
## Uniformity   0.016297   0.009803   1.663  0.09669 .  
## Moisture     0.169012   0.102480   1.649  0.09938 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.148 on 1108 degrees of freedom
## Multiple R-squared:  0.8091, Adjusted R-squared:  0.8073 
## F-statistic: 469.5 on 10 and 1108 DF,  p-value: < 2.2e-16
```

```
# Full MLR alternative, using factor command
mfull_alternative <- lm(Flavor~ factor(Processing.Method) + Aroma + Aftertaste +
      Body + Acidity + Balance + Sweetness + Uniformity + Moisture, dat=coffee)
```

Suppose we want to check the VIF for $j = 1$; that is, x_1. Now, we fit:

$$x_{i1} = \alpha_0 + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4} + \alpha_5 x_{i5} + \alpha_6 x_{i6} + \alpha_7 x_{i7} + \alpha_8 x_{i8} + \alpha_9 x_{i9} + \alpha_{10} x_{i(10)} + \varepsilon_i$$

```
wet_reg <- lm(wet ~ semi + Aroma + Aftertaste + Body + Acidity + Balance +
                Sweetness + Uniformity + Moisture,dat=coffee)
summary(wet_reg)
```

```
## 
## Call:
## lm(formula = wet ~ semi + Aroma + Aftertaste + Body + Acidity + 
##     Balance + Sweetness + Uniformity + Moisture, data = coffee)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.0015 -0.0283  0.1770  0.2522  0.7704 
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.81748    0.45838   1.783 0.074794 .
## semi        -0.75675    0.05551 -13.632  < 2e-16 ***
## Aroma        0.09690    0.05562   1.742 0.081774 .
## Aftertaste  -0.13169    0.06502  -2.026 0.043054 *
## Body        -0.21885    0.06596  -3.318 0.000936 ***
## Acidity      0.18696    0.05746   3.254 0.001173 **
## Balance     -0.10804    0.06136  -1.761 0.078563 .
## Sweetness    0.08373    0.02753   3.041 0.002413 **
## Uniformity   0.03547    0.02668   1.329 0.184053
## Moisture     0.59486    0.27858   2.135 0.032956 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4031 on 1109 degrees of freedom
## Multiple R-squared:  0.1911, Adjusted R-squared:  0.1845
## F-statistic: 29.11 on 9 and 1109 DF,  p-value: < 2.2e-16
```

```
r2_wet <- summary(wet_reg)$r.squared
r2_wet
```

```
## [1] 0.191077
```

R_j: In our case, $R_1 = 0.191077$.

```
VIF_wet <- 1 / (1 - r2_wet)
VIF_wet
```

```
## [1] 1.236212
```

VIF_j: $\text{VIF}_1 = 1.236212$. Interpretation: in a regression with all the variables compared to a regression with just this one, the estimated variance has increased by a factor of 1.24, which is not a very large inflation. The variable wet is not very linearly correlated or dependent on the other predictors that we have in the model.

```
Aroma_reg <- lm(Aroma ~ wet + semi + Aftertaste +
      Body + Acidity + Balance + Sweetness + Uniformity + Moisture, dat=coffee)
r2_Aroma <- summary(Aroma_reg)$r.squared
r2_Aroma
```

```
## [1] 0.5204716
```

```
VIF_Aroma <- 1 / (1 - r2_Aroma)
VIF_Aroma
```

```
## [1] 2.085382
```

$R_3 = 0.5204716$, $\text{VIF}_3 = 2.085382$.

```
Aftertaste_reg <- lm(Aftertaste ~ wet + semi + Aroma +
       Body + Acidity + Balance + Sweetness + Uniformity + Moisture, dat=coffee)
r2_Aftertaste <- summary(Aftertaste_reg)$r.squared
r2_Aftertaste
```

```
## [1] 0.7101012
```

```
VIF_Aftertaste <- 1 / (1 - r2_Aftertaste)
VIF_Aftertaste
```

```
## [1] 3.449479
```

```
library(car)
```

```
## Loading required package: carData
```

```
vif(mfull) # VIF function in the "car" library
```

```
##        wet       semi      Aroma Aftertaste       Body    Acidity    Balance
##   1.236212   1.178004   2.085382   3.449479   2.317728   2.232210   3.002813
##  Sweetness Uniformity   Moisture
##   1.159602   1.209901   1.086101
```

No serious signs of inflation, all VIFs are less than 10.

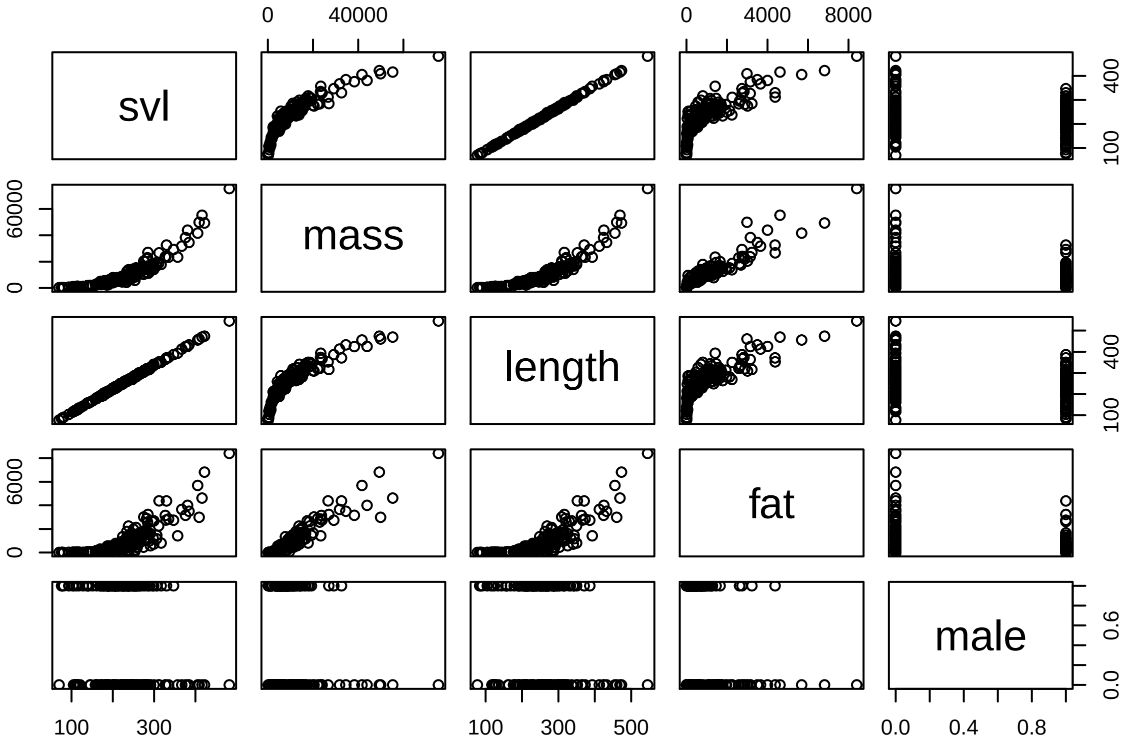
```
## Python in FL everglades example (2017)
## Sex, length, total mass, fat mass, and specimen condition data for
## 248 Burmese pythons (Python bivittatus) collected in the Florida Everglades

python <- read.csv("FLpython.csv")
head(python)
```

```
##   sex  svl mass length    fat
## 1   F 70.0  186   77.5  6.000
## 2   M 76.0  310   83.8 11.000
## 3   M 77.0  260   86.1  6.000
## 4   M 78.0  262   87.1  8.000
## 5   M 81.0  306   91.1  4.000
## 6   M 93.5  605  104.6 18.959
```

```
python$male <- ifelse(python$sex == 'M', 1, 0) # 1 = M, 0 =F

cor(python[,-1])
```

```
##               svl       mass     length        fat       male
## svl     1.0000000  0.8843022  0.9994935  0.8098652 -0.1602418
## mass    0.8843022  1.0000000  0.8858256  0.9419114 -0.2190993
## length  0.9994935  0.8858256  1.0000000  0.8114658 -0.1593512
## fat     0.8098652  0.9419114  0.8114658  1.0000000 -0.2933111
## male   -0.1602418 -0.2190993 -0.1593512 -0.2933111  1.0000000
```

```
pairs(python[,-1])
```



```r
mpf <- lm(fat ~ male + svl + mass + length, data = python)
summary(mpf)
```

```
##
## Call:
## lm(formula = fat ~ male + svl + mass + length, data = python)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2445.77  -137.41    -5.29   110.00  1527.27
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.021e+02  1.331e+02   1.518    0.130
## male        -1.971e+02  4.732e+01  -4.165 4.32e-05 ***
## svl         -3.370e+00  1.125e+01  -0.300    0.765
## mass         1.178e-01  5.302e-03  22.210  < 2e-16 ***
## length       1.594e+00  1.010e+01   0.158    0.875
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 360.9 on 243 degrees of freedom
## Multiple R-squared:  0.897,  Adjusted R-squared:  0.8953
## F-statistic:   529 on 4 and 243 DF,  p-value: < 2.2e-16
```

```
vif(mpf)
```

```
##        male          svl         mass      length
##    1.058699  994.546545    4.813078 1007.484200
```

```
mpf_l <- lm(length ~ male + svl + mass, data=python)
1/(1-summary(mpf_l)$r.squared)
```

```
## [1] 1007.484
```

Misleading conclusion: svl and length are both irrelevant (this is not the case). Also, the standard errors are very large.

```
# remove "length" based on VIF
mpf2 <- lm(fat ~ male + mass + svl, data = python)
summary(mpf2)
```

```
##
## Call:
## lm(formula = fat ~ male + mass + svl, data = python)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2444.44  -137.38    -6.66   109.22  1530.81
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  204.09840  132.30121   1.543   0.1242
## male        -196.71705   47.16396  -4.171 4.22e-05 ***
## mass           0.11788    0.00524  22.495  < 2e-16 ***
## svl           -1.59841    0.76433  -2.091   0.0375 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 360.2 on 244 degrees of freedom
## Multiple R-squared:  0.897,  Adjusted R-squared:  0.8957
## F-statistic: 708.2 on 3 and 244 DF,  p-value: < 2.2e-16
```

```
vif(mpf2)
```

```
##     male     mass      svl
## 1.056139 4.720065 4.611903
```

Svl now has a significant *t*-statistic.

Lecture 13 | 2020-10-26

Model selection: Given p explanatory variables, find the subset $k \leqslant p$ of explanatory variables ("reduced model") that gives us the "best" model:

- goodness of fit
- interpretability
- predictive performance

Some related concepts:

1. F tests compare between 2 specific models where test adequacy of a "reduced" model (subset, "nested") relative to full model.

   Quiz 4: $\beta_1 = \beta_2$ in part d-f

2. Multicollinearity: can affect interpretability of $\hat{\beta}_j$ usual interpretation "holding other variables constant" doesn't really work when x_j is strongly correlated with other predictors.
3. R^2 is the proportion of variability in the response explained by the regression model. It always increases when adding variables.
4. $\hat{\sigma}^2$ is estimated residual variable, used for prediction, want $\hat{\sigma}^2$ small to give good predictive performance

Two key ingredients:

- Metric (or criterion) for comparing different models with potentially different number of predictors
- selection/search strategy (which models should we fit and test?)

Examples of metrics for model selection:

DEFINITION 3.30: Adjusted R^2

$$R^2_{\text{adj}} = 1 - \frac{\text{SS(Res)}/(n-k-1)}{\text{SS(Total)}/(n-1)} = 1 - \frac{\hat{\sigma}^2}{s^2}$$

for model with k predictors.

Compared to

$$R^2 = 1 - \frac{\text{SS(Res)}}{\text{SS(Total)}}$$

- $\text{SS(Res)}/(n-k-1)$ estimated $\hat{\sigma}^2$ for model with k predictors
- $\text{SS(Total)}/(n-1)$ is the sample variance of responses y_i.

$$\begin{aligned}
R^2_{\text{adj}} = 1 - \frac{n-1}{n-k-1}(1-R^2) &= 1 - \left(1 + \frac{k}{n-k-1}\right)(1-R^2) \\
&= 1 - \left[1(1-R^2) - \left(\frac{k}{n-k-1}\right)(1-R^2)\right] \\
&= 1 - \left[1 - R^2 - \left(\frac{k}{n-k-1}\right)(1-R^2)\right] \\
&= 1 - 1 + R^2 - \frac{k}{n-k-1}(1-R^2) \\
&= R^2 - (1-R^2)\frac{k}{n-k-1}
\end{aligned}$$

Intuition: R^2_{adj} accounts for number variables in model, *penalizes* inclusion of unimportant predictors; that is, SS(Res) has little decrease when adding such variables. Meanwhile, R^2 always increases with more predictors, but R^2_{adj} can decrease if SS(Res) change is small.

While R^2_{adj} loses its usual interpretation of R^2, but can be used as a measure of "goodness of fit" and model selection criterion (e.g. pick subset of predictors that gives the highest R^2_{adj}).

EXAMPLE 3.31

Given

- $n = 25$
- SS(Total) $= 20$
- $p = 6$

Suppose we're considering on a subset of $k = 4$ predictors, and find:

	red	full
	$k = 4$	$p = 6$
SS(Total)	20	20
SS(Res)	10	9.8
R^2	$10/20 = 0.5$	$9.8/20 = 0.49$
R^2_{adj}	$1 - \frac{10/(25-4-1)}{20/(25-1)} = 0.4$	$1 - \frac{9.8/(25-6-1)}{20/(25-1)} \approx 0.347$
$\hat{\sigma}^2$	$10/(25-4-1) = 0.5$	$9.8/(25-6-1) \approx 0.544$

- $n - k - 1$ d.f. Res in reduced
- $n - p - 1$ d.f. Res in full

Remarks:

- $R^2_{\text{adj}} < R^2$, but as $n \to \infty$, $R^2_{\text{adj}} \to R^2$.
- model with higher R^2_{adj} has lower $\hat{\sigma}^2$, thus is a reasonable metric for model selection

Akaike Information Criterion (AIC)

Let n be sample size, q is number of parameters [in MLR k predictors + 1 (intercept) + 1 (σ^2)]

$$\text{AIC} = 2q - 2\ln[L(\hat{\theta})]$$

where $L(\hat{\theta})$ is the likelihood function evaluated at $\hat{\theta}$ (parameter estimates). Note that LS estimates of $\boldsymbol{\beta}$ are equivalent to MLE under the usual normal assumptions on $\boldsymbol{\varepsilon}$. Also, $2q$ is the penalty for including more predictors. With more parameters, $L(\hat{\theta})$ increases, offset by penalty $2q$.

If we want to measure just the difference, we can do

$$\text{AIC} = n\ln\left[\frac{\text{SS(Res)}}{n}\right] + 2(p+1)$$

Therefore, model with lower AIC is preferred; that is, differences in AIC matter not the value itself.

Bayesian Information Criterion (BIC)

Similar to AIC, but more strongly penalizes inclusion of more variables.

$$\text{BIC} = q\ln(n) - 2\ln[L(\hat{\theta})]$$

where $q\ln(n)$ depends on sample size.

If we want to measure just the difference, we can do

$$\text{BIC} = n\ln\left[\frac{\text{SS(Res)}}{n}\right] + (p+1)\ln(n)$$

Recap:

- R^2, AIC, BIC are all based on comparing the fitted models. In other words, they look at the explanatory power of the model.
- They all have penalties to try to prevent "overfitting." That is, having too many variables might end up modelling spurious relationships that are actually noise.

Mean Square Prediction Error (MSPE)

Consider predictive performance of model on *new* data; that is, data *not* used in fitting of models. "Is model generalizable to new data?" Overfitted models tend to have high prediction error.

For example, via cross-validation schemes. We've given 4 examples of metrics/criteria for comparing models. Imagine we have p predictors:

- $\binom{p}{1}$ 1 predictors
- $\binom{p}{2}$ 2 predictors
- $\vdots$
- $\binom{p}{p}$ p predictors

$$\sum_{j=0}^{p} \binom{p}{j} = 2^p$$

Occam's Razor: "The simplest explanation is usually the best one."—William Ockham

Lecture 14 | 2020-10-28

Model Selection

- Criteria: R^2_{adj}, AIC, BIC, MSPE, etc. explicitly penalizes unnecessarily complex models.
- Search strategies (use with chosen criterion)

(i) Brute force: fit all possible regressions. With p predictors, we have $\sum_{j=0}^{p} \binom{p}{j} = 2^p$ possible models to fit.

  - Finds optimal model that may be computationally intensive (or infeasible) if p is large.

  Idea: Find a "good" (useful) model in reasonable computational time (not necessarily optimal). Many strategies focus on adding/removing variables one at a time.

(ii) Forward selection: add one variable at a time to model.

  - Start with model that only has intercept β_0.
  - Fit p simple linear regression models

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_j + \boldsymbol{\varepsilon} \quad j = 1, \dots, p$$

  - Pick the best of p models (with 1 predictor) according to chosen criterion, and add that variable x_j to model.
  - Fit $(p-1)$ models containing x_j and one other variable.
    - If none of $(p-1)$ models improves criterion, stop.
    - Pick the best of $(p-1)$ models according to criterion, so now we have 2 variables in the model.

  Continue adding 1 variable at a time in this way until we can no more variables improve the criterion. The final model is one with the best criterion after we *stop*; that is, no further improvement is possible.

Note: Much faster than brute force as the maximum number of models to fit is:

$$p + (p-1) + \cdots + 2 + 1 = \sum_{i=1}^{p} i = \frac{p(p+1)}{2}$$

which is $\mathcal{O}(p^2)$ compared to $\mathcal{O}(2^p)$ for all possible regressions.

(iii) Backward direction: remove one variable at a time to model.

- Start with model that has p predictors.
- Fit p models that result from removing one variable from the regression; that is, each one has $(p-1)$ variables.
- Pick the best of p models according to criterion.
  - Eliminate that variable x_j from model.
  - Fit $(p-1)$ models that remove x_j and one other variable from model.
  - Pick best of $(p-1)$ models (2 variables removed).

Continue removing 1 variable at a time in this way until we can no more variables improve the criterion. Same computational complexity as forward selection.

(iv) Forward-backwards (allows individual variables to be both added/removed)

- Start as in forward selection
- If we have k variables in model:
  - Backwards: fit k models with $(k-1)$ variables. If any of these improve criterion, remove the variable.
  - Forwards: fit $(p-k)$ models with $(k+1)$ variables. If any of these improve criterion, add that variable.

- These are the basic "stepwise" selection models to get a "good" (useful) model.
- Many other have sophisticated procedures available. For example, stochastic search, lasso.
- We've assumed that $n > p$ because otherwise $(X^\top X)$ is not invertible. More specialized methods needed if number of predictors is larger than sample size.

```
## Coffee example  (Coffee Quality Institute, 2018) continued
coffee <- read.csv("coffee_arabica.csv")

mfull <- lm(Flavor~ factor(Processing.Method) + Aroma + Aftertaste +
    Body + Acidity + Balance + Sweetness + Uniformity + Moisture, dat=coffee)
summary(mfull)$adj.r.squared
```

```
## [1] 0.8073297
```

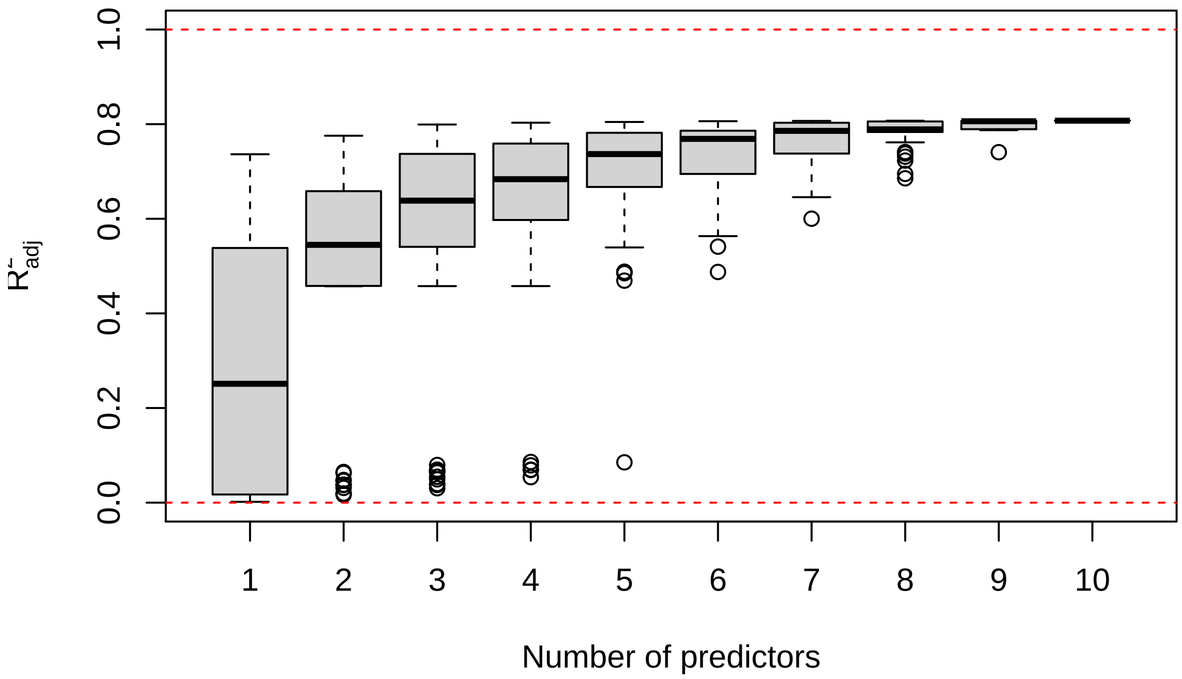
```
AIC(mfull)
```

```
## [1] -1087.524
```

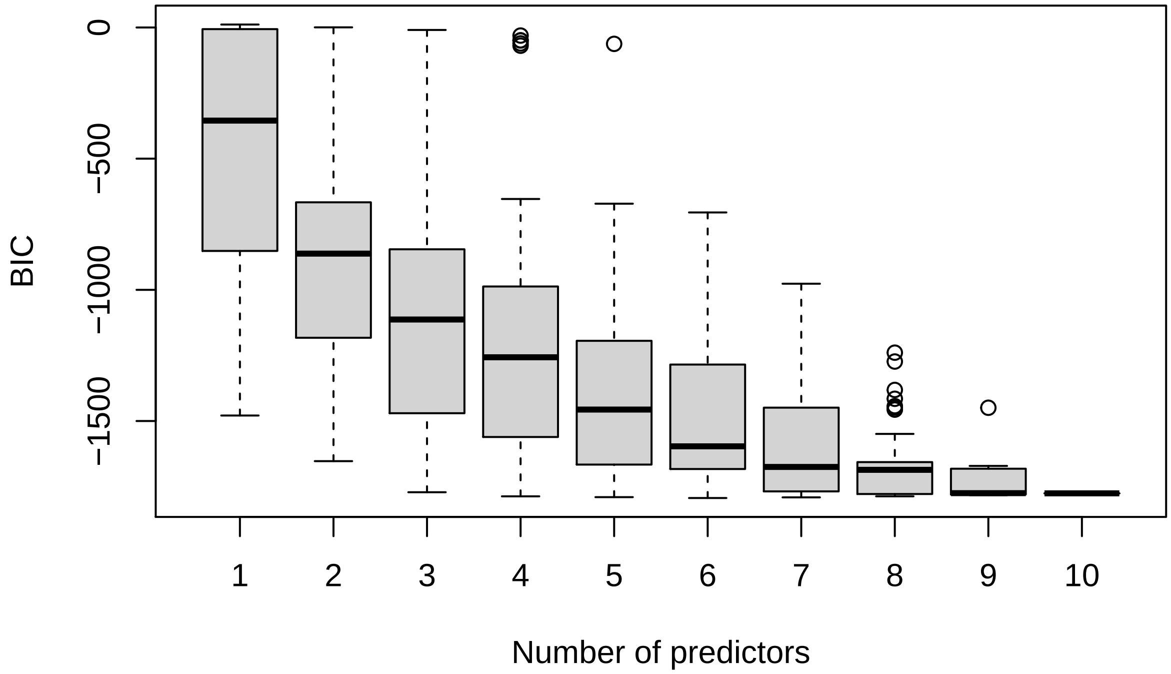
```
BIC(mfull)
```

```
## [1] -1027.282
```

```
library(leaps)
all_regs <- regsubsets(Flavor ~ ., data = coffee, nvmax = 10, nbest = 2^10,
                       really.big = TRUE)
all_regs_summ <- summary(all_regs)
# all_regs_summ$which
# all_regs_summ$adjr2
# all_regs_summ$bic

# Organize results according to number of variables in model
p <- 10
k <- c(rep(1, choose(p,1)),
           rep(2, choose(p,2)),
           rep(3, choose(p,3)),
           rep(4, choose(p,4)),
           rep(5, choose(p,5)),
           rep(6, choose(p,6)),
           rep(7, choose(p,7)),
           rep(8, choose(p,8)),
           rep(9, choose(p,9)),
           rep(10, choose(p,10)))
boxplot(all_regs_summ$adjr2 ~ k, xlab = "Number of predictors", ylab =
          expression(R[adj]^2), ylim = c(0,1))
abline(h = c(0,1), lty = 2, col = "red")
```



```r
boxplot(all_regs_summ$bic ~ k, xlab = "Number of predictors", ylab = "BIC")
```



```
max(all_regs_summ$adjr2)
```

```
## [1] 0.8075027
```

```
bestR2adj <- which.max(all_regs_summ$adjr2)
min(all_regs_summ$bic)
```

```
## [1] -1793.389
```

```
bestBIC <- which.min(all_regs_summ$bic)

# Find out which predictors in those models
all_regs_summ$which[bestR2adj,]
```

```
##                                 (Intercept) 
##                                        TRUE 
## Processing.MethodSemi-washed / Semi-pulped 
##                                       FALSE 
##              Processing.MethodWashed / Wet 
##                                        TRUE 
##                                      Aroma 
##                                        TRUE 
##                                 Aftertaste 
##                                        TRUE 
##                                       Body 
##                                        TRUE 
##                                    Acidity 
##                                        TRUE 
##                                    Balance
##                                       TRUE
##                                  Sweetness
##                                       TRUE
##                                 Uniformity
##                                       TRUE
##                                   Moisture
##                                       TRUE
```

```
all_regs_summ$which[bestBIC,]
```

```
##                                (Intercept)
##                                       TRUE
## Processing.MethodSemi-washed / Semi-pulped
##                                      FALSE
##              Processing.MethodWashed / Wet
##                                       TRUE
##                                      Aroma
##                                       TRUE
##                                 Aftertaste
##                                       TRUE
##                                       Body
##                                       TRUE
##                                    Acidity
##                                       TRUE
##                                    Balance
##                                      FALSE
##                                  Sweetness
##                                       TRUE
##                                 Uniformity
##                                      FALSE
##                                   Moisture
##                                      FALSE
```

```
coffee$wet <- ifelse(coffee$Processing.Method == 'Washed / Wet', 1,
                     0) # 1 = wet, 0 otherwise
coffee$semi <- ifelse(coffee$Processing.Method == 'Semi-washed / Semi-pulped',
                      1, 0) # 1 = semi/dry, 0 otherwise
coffee$Processing.Method <- NULL

m_bestr2adj <- lm(Flavor~ wet + Aroma + Aftertaste +
            Body + Acidity + Balance + Sweetness + Uniformity + Moisture,
            dat=coffee)
summary(m_bestr2adj)
```

```
##
## Call:
## lm(formula = Flavor ~ wet + Aroma + Aftertaste + Body + Acidity +
##     Balance + Sweetness + Uniformity + Moisture, data = coffee)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68587 -0.08469  0.00080  0.08923  0.63660
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.728709   0.168439  -4.326 1.65e-05 ***
## wet         -0.032797   0.010197  -3.216  0.00134 ** 
## Aroma        0.220278   0.020434  10.780  < 2e-16 ***
## Aftertaste   0.468749   0.023901  19.612  < 2e-16 ***
## Body         0.096194   0.024308   3.957 8.06e-05 ***
## Acidity      0.216754   0.021185  10.232  < 2e-16 ***
## Balance      0.046793   0.022547   2.075  0.03819 *  
## Sweetness    0.025480   0.010136   2.514  0.01209 *  
## Uniformity   0.016291   0.009798   1.663  0.09665 .  
## Moisture     0.168439   0.102033   1.651  0.09906 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1479 on 1109 degrees of freedom
## Multiple R-squared:  0.8091, Adjusted R-squared:  0.8075 
## F-statistic: 522.1 on 9 and 1109 DF,  p-value: < 2.2e-16
```

```r
AIC(m_bestr2adj)
```

```
## [1] -1089.52
```

```r
BIC(m_bestr2adj)
```

```
## [1] -1034.298
```

```r
m_bestBIC <- lm(Flavor~ wet + Aroma + Aftertaste +
                  Body + Acidity + Sweetness , dat=coffee)
summary(m_bestBIC)
```

```
## 
## Call:
## lm(formula = Flavor ~ wet + Aroma + Aftertaste + Body + Acidity + 
##     Sweetness, data = coffee)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.65627 -0.08781  0.00032  0.08529  0.63010 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.609003   0.159910  -3.808 0.000148 ***
## wet         -0.032852   0.010198  -3.221 0.001313 ** 
## Aroma        0.225969   0.020378  11.089  < 2e-16 ***
## Aftertaste   0.490988   0.021938  22.381  < 2e-16 ***
## Body         0.103438   0.022926   4.512 7.11e-06 ***
## Acidity      0.225638   0.020994  10.748  < 2e-16 ***
## Sweetness    0.033445   0.009582   3.491 0.000501 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1484 on 1112 degrees of freedom
## Multiple R-squared:  0.8073, Adjusted R-squared:  0.8063 
## F-statistic: 776.4 on 6 and 1112 DF,  p-value: < 2.2e-16
```

```
AIC(m_bestBIC)
```

```
## [1] -1085.26
```

```
BIC(m_bestBIC)
```

```
## [1] -1045.098
```

```
# Let's also try stepwise methods
library(MASS)

# Full model and empty model with just intercept
full <- lm(Flavor ~ ., data = coffee)
empty <- lm(Flavor ~ 1, data = coffee)

# default stepAIC uses AIC criterion
stepAIC(object = empty, scope = list(upper = full, lower = empty), direction
        = "forward")
```

```
## Start:  AIC=-2432.31
## Flavor ~ 1
##
##               Df Sum of Sq     RSS     AIC
## + Aftertaste  1    93.607  33.465 -3923.3
## + Acidity     1    69.294  57.778 -3312.3
## + Aroma       1    68.457  58.615 -3296.1
## + Balance     1    68.173  58.899 -3290.7
## + Body        1    58.232  68.840 -3116.2
## + Uniformity  1     5.778 121.294 -2482.4
## + wet         1     2.313 124.759 -2450.9
## + Sweetness   1     2.300 124.772 -2450.8
## + Moisture    1     2.239 124.833 -2450.2
## + semi        1     0.331 126.741 -2433.2
## <none>                    127.072 -2432.3
##
## Step:  AIC=-3923.33
## Flavor ~ Aftertaste
##
##               Df Sum of Sq    RSS     AIC
## + Acidity     1    4.9955 28.470 -4102.2
## + Aroma       1    4.9082 28.557 -4098.8
## + Body        1    2.2551 31.210 -3999.4
## + Balance     1    1.7369 31.729 -3981.0
## + Sweetness   1    0.1384 33.327 -3926.0
## + Uniformity  1    0.1143 33.351 -3925.2
## + wet         1    0.0871 33.378 -3924.2
## <none>                    33.465 -3923.3
## + Moisture    1    0.0382 33.427 -3922.6
## + semi        1    0.0179 33.448 -3921.9
##
## Step:  AIC=-4102.23
## Flavor ~ Aftertaste + Acidity
##
##               Df Sum of Sq    RSS     AIC
## + Aroma       1   3.02166 25.448 -4225.8
## + Body        1   0.89556 27.575 -4136.0
## + Balance     1   0.65424 27.816 -4126.2
## + wet         1   0.22561 28.244 -4109.1
## + Sweetness   1   0.17094 28.299 -4107.0
## + Uniformity  1   0.11428 28.356 -4104.7
## <none>                   28.470 -4102.2
## + semi        1   0.04453 28.425 -4102.0
## + Moisture    1   0.01991 28.450 -4101.0
##
## Step:  AIC=-4225.78
## Flavor ~ Aftertaste + Acidity + Aroma
##
##              Df Sum of Sq    RSS     AIC
## + Body        1   0.50898 24.939 -4246.4
## + Balance     1   0.32565 25.123 -4238.2
## + wet         1   0.26887 25.180 -4235.7
## + Sweetness   1   0.19006 25.258 -4232.2
## + Uniformity  1   0.11405 25.334 -4228.8
## <none>                   25.448 -4225.8
## + semi        1   0.04166 25.407 -4225.6
## + Moisture    1   0.01953 25.429 -4224.6
##
## Step:  AIC=-4246.39
## Flavor ~ Aftertaste + Acidity + Aroma + Body
##
##              Df Sum of Sq    RSS     AIC
## + Sweetness   1  0.223266 24.716 -4254.5
## + wet         1  0.183460 24.756 -4252.6
## + Uniformity  1  0.172293 24.767 -4252.1
## + Balance     1  0.132479 24.807 -4250.3
## + Moisture    1  0.058517 24.881 -4247.0
## <none>                   24.939 -4246.4
## + semi        1  0.040446 24.899 -4246.2
##
## Step:  AIC=-4254.45
## Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness
##
##              Df Sum of Sq    RSS     AIC
## + wet         1  0.228511 24.488 -4262.8
## + Balance     1  0.118556 24.598 -4257.8
## + Uniformity  1  0.075546 24.641 -4255.9
## <none>                   24.716 -4254.5
## + Moisture    1  0.038620 24.677 -4254.2
## + semi        1  0.037654 24.678 -4254.2
##
## Step:  AIC=-4262.84
## Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness + wet
##
##              Df Sum of Sq    RSS     AIC
## + Balance     1  0.099946 24.388 -4265.4
## + Uniformity  1  0.084340 24.403 -4264.7
## + Moisture    1  0.046681 24.441 -4263.0
## <none>                   24.488 -4262.8
## + semi        1  0.000241 24.487 -4260.9
```

```
##
## Step:  AIC=-4265.42
## Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness + wet +
##     Balance
##
##              Df Sum of Sq    RSS     AIC
## + Uniformity  1  0.063931 24.324 -4266.4
## + Moisture    1  0.063069 24.325 -4266.3
## <none>                    24.388 -4265.4
## + semi        1  0.000236 24.387 -4263.4
##
## Step:  AIC=-4266.36
## Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness + wet +
##     Balance + Uniformity
##
##            Df Sum of Sq    RSS     AIC
## + Moisture  1  0.059626 24.264 -4267.1
## <none>                  24.324 -4266.4
## + semi      1  0.000150 24.324 -4264.4
##
## Step:  AIC=-4267.1
## Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness + wet +
##     Balance + Uniformity + Moisture
##
##        Df  Sum of Sq    RSS     AIC
## <none>               24.264 -4267.1
## + semi  1 8.7985e-05 24.264 -4265.1
```

```
##
## Call:
## lm(formula = Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness +
##     wet + Balance + Uniformity + Moisture, data = coffee)
##
## Coefficients:
## (Intercept)   Aftertaste      Acidity        Aroma         Body    Sweetness
##    -0.72871      0.46875      0.21675      0.22028      0.09619      0.02548
##         wet      Balance   Uniformity     Moisture
##    -0.03280      0.04679      0.01629      0.16844
```

```r
# Let's get stepAIC to use BIC by specifying the penalty k = log(n)
# Forward
stepAIC(object = empty, scope = list(upper = full, lower = empty), direction
        = "forward", k = log(nrow(coffee)))
```

```
## Start:  AIC=-2427.29
## Flavor ~ 1
##
##              Df Sum of Sq     RSS     AIC
## + Aftertaste  1    93.607  33.465 -3913.3
## + Acidity     1    69.294  57.778 -3302.2
## + Aroma       1    68.457  58.615 -3286.1
## + Balance     1    68.173  58.899 -3280.7
## + Body        1    58.232  68.840 -3106.2
## + Uniformity  1     5.778 121.294 -2472.3
## + wet         1     2.313 124.759 -2440.8
## + Sweetness   1     2.300 124.772 -2440.7
## + Moisture    1     2.239 124.833 -2440.2
## <none>                    127.072 -2427.3
## + semi        1     0.331 126.741 -2423.2
##
## Step:  AIC=-3913.29
## Flavor ~ Aftertaste
##
##              Df Sum of Sq    RSS     AIC
## + Acidity     1    4.9955 28.470 -4087.2
## + Aroma       1    4.9082 28.557 -4083.7
## + Body        1    2.2551 31.210 -3984.3
## + Balance     1    1.7369 31.729 -3965.9
## <none>                    33.465 -3913.3
## + Sweetness   1    0.1384 33.327 -3910.9
## + Uniformity  1    0.1143 33.351 -3910.1
## + wet         1    0.0871 33.378 -3909.2
## + Moisture    1    0.0382 33.427 -3907.5
## + semi        1    0.0179 33.448 -3906.9
##
## Step:  AIC=-4087.17
## Flavor ~ Aftertaste + Acidity
##
##              Df Sum of Sq    RSS     AIC
## + Aroma       1   3.02166 25.448 -4205.7
## + Body        1   0.89556 27.575 -4115.9
## + Balance     1   0.65424 27.816 -4106.2
## + wet         1   0.22561 28.244 -4089.0
## <none>                    28.470 -4087.2
## + Sweetness   1   0.17094 28.299 -4086.9
## + Uniformity  1   0.11428 28.356 -4084.6
## + semi        1   0.04453 28.425 -4081.9
## + Moisture    1   0.01991 28.450 -4080.9
##
## Step:  AIC=-4205.7
## Flavor ~ Aftertaste + Acidity + Aroma
##
##              Df Sum of Sq    RSS     AIC
## + Body        1   0.50898 24.939 -4221.3
## + Balance     1   0.32565 25.123 -4213.1
## + wet         1   0.26887 25.180 -4210.6
## + Sweetness   1   0.19006 25.258 -4207.1
## <none>                    25.448 -4205.7
## + Uniformity  1   0.11405 25.334 -4203.7
## + semi        1   0.04166 25.407 -4200.5
## + Moisture    1   0.01953 25.429 -4199.5
##
## Step:  AIC=-4221.29
## Flavor ~ Aftertaste + Acidity + Aroma + Body
##
##              Df Sum of Sq    RSS     AIC
## + Sweetness   1  0.223266 24.716 -4224.3
## + wet         1  0.183460 24.756 -4222.5
## + Uniformity  1  0.172293 24.767 -4222.0
## <none>                     24.939 -4221.3
## + Balance     1  0.132479 24.807 -4220.2
## + Moisture    1  0.058517 24.881 -4216.9
## + semi        1  0.040446 24.899 -4216.1
##
## Step:  AIC=-4224.33
## Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness
##
##              Df Sum of Sq    RSS     AIC
## + wet         1  0.228511 24.488 -4227.7
## <none>                     24.716 -4224.3
## + Balance     1  0.118556 24.598 -4222.7
## + Uniformity  1  0.075546 24.641 -4220.7
## + Moisture    1  0.038620 24.677 -4219.1
## + semi        1  0.037654 24.678 -4219.0
##
## Step:  AIC=-4227.7
## Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness + wet
##
##              Df Sum of Sq    RSS     AIC
## <none>                     24.488 -4227.7
## + Balance     1  0.099946 24.388 -4225.3
## + Uniformity  1  0.084340 24.403 -4224.5
## + Moisture    1  0.046681 24.441 -4222.8
## + semi        1  0.000241 24.487 -4220.7
```

```
##
## Call:
## lm(formula = Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness +
##     wet, data = coffee)
##
## Coefficients:
## (Intercept)   Aftertaste      Acidity        Aroma         Body    Sweetness
##    -0.60900      0.49099      0.22564      0.22597      0.10344      0.03345
##         wet
##    -0.03285
```

```
m_f <- stepAIC(object = empty, scope = list(upper = full, lower = empty),
               direction = "forward", trace = 0, k = log(nrow(coffee)))
summary(m_f)
```

```
##
## Call:
## lm(formula = Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness +
##     wet, data = coffee)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65627 -0.08781  0.00032  0.08529  0.63010
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.609003   0.159910  -3.808 0.000148 ***
## Aftertaste   0.490988   0.021938  22.381  < 2e-16 ***
## Acidity      0.225638   0.020994  10.748  < 2e-16 ***
## Aroma        0.225969   0.020378  11.089  < 2e-16 ***
## Body         0.103438   0.022926   4.512 7.11e-06 ***
## Sweetness    0.033445   0.009582   3.491 0.000501 ***
## wet         -0.032852   0.010198  -3.221 0.001313 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1484 on 1112 degrees of freedom
## Multiple R-squared:  0.8073, Adjusted R-squared:  0.8063
## F-statistic: 776.4 on 6 and 1112 DF,  p-value: < 2.2e-16
```

```r
# Backward
stepAIC(object = full, scope = list(upper = full, lower = empty),
        direction = "backward", k = log(nrow(coffee)))
```

```
## Start:  AIC=-4209.89
## Flavor ~ Aroma + Aftertaste + Body + Acidity + Balance + Sweetness +
##     Uniformity + Moisture + wet + semi
##
##              Df Sum of Sq    RSS     AIC
## - semi        1    0.0001 24.264 -4216.9
## - Moisture    1    0.0596 24.324 -4214.2
## - Uniformity  1    0.0605 24.325 -4214.1
## - Balance     1    0.0943 24.358 -4212.6
## - Sweetness   1    0.1383 24.402 -4210.5
## <none>                    24.264 -4209.9
## - wet         1    0.1970 24.461 -4207.9
## - Body        1    0.3418 24.606 -4201.3
## - Acidity     1    2.2904 26.554 -4116.0
## - Aroma       1    2.5422 26.806 -4105.4
## - Aftertaste  1    8.4155 32.679 -3883.7
##
## Step:  AIC=-4216.9
## Flavor ~ Aroma + Aftertaste + Body + Acidity + Balance + Sweetness +
##     Uniformity + Moisture + wet
##
##              Df Sum of Sq    RSS     AIC
## - Moisture    1    0.0596 24.324 -4221.2
## - Uniformity  1    0.0605 24.325 -4221.1
## - Balance     1    0.0942 24.358 -4219.6
## - Sweetness   1    0.1383 24.402 -4217.6
## <none>                    24.264 -4216.9
## - wet         1    0.2263 24.490 -4213.5
## - Body        1    0.3426 24.607 -4208.2
## - Acidity     1    2.2905 26.555 -4123.0
## - Aroma       1    2.5426 26.807 -4112.4
## - Aftertaste  1    8.4155 32.680 -3890.7
##
## Step:  AIC=-4221.18
## Flavor ~ Aroma + Aftertaste + Body + Acidity + Balance + Sweetness +
##     Uniformity + wet
##
##              Df Sum of Sq    RSS     AIC
## - Uniformity  1    0.0639 24.388 -4225.3
## - Balance     1    0.0795 24.403 -4224.5
## <none>                    24.324 -4221.2
## - Sweetness   1    0.1553 24.479 -4221.1
## - wet         1    0.2189 24.543 -4218.2
## - Body        1    0.3205 24.644 -4213.5
## - Acidity     1    2.3500 26.674 -4125.0
## - Aroma       1    2.5685 26.892 -4115.9
## - Aftertaste  1    8.3791 32.703 -3897.0
##
## Step:  AIC=-4225.26
## Flavor ~ Aroma + Aftertaste + Body + Acidity + Balance + Sweetness +
##     wet
##
##               Df Sum of Sq    RSS     AIC
## - Balance     1    0.0999 24.488 -4227.7
## <none>                    24.388 -4225.3
## - wet         1    0.2099 24.598 -4222.7
## - Sweetness   1    0.2522 24.640 -4220.8
## - Body        1    0.2905 24.678 -4219.0
## - Acidity     1    2.3543 26.742 -4129.2
## - Aroma       1    2.5711 26.959 -4120.1
## - Aftertaste  1    8.5578 32.945 -3895.7
##
## Step:  AIC=-4227.7
## Flavor ~ Aroma + Aftertaste + Body + Acidity + Sweetness + wet
##
##               Df Sum of Sq    RSS     AIC
## <none>                    24.488 -4227.7
## - wet         1    0.2285 24.716 -4224.3
## - Sweetness   1    0.2683 24.756 -4222.5
## - Body        1    0.4483 24.936 -4214.4
## - Acidity     1    2.5437 27.031 -4124.1
## - Aroma       1    2.7077 27.195 -4117.4
## - Aftertaste  1   11.0308 35.518 -3818.6
```

```
##
## Call:
## lm(formula = Flavor ~ Aroma + Aftertaste + Body + Acidity + Sweetness +
##     wet, data = coffee)
##
## Coefficients:
## (Intercept)        Aroma   Aftertaste         Body      Acidity    Sweetness
##    -0.60900      0.22597      0.49099      0.10344      0.22564      0.03345
##         wet
##    -0.03285
```

```r
m_b <- stepAIC(object = full, scope = list(upper = full, lower = empty),
               direction = "backward", trace = 0, k = log(nrow(coffee)))
summary(m_b)
```

```
##
## Call:
## lm(formula = Flavor ~ Aroma + Aftertaste + Body + Acidity + Sweetness +
##     wet, data = coffee)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.65627 -0.08781  0.00032  0.08529  0.63010
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.609003   0.159910  -3.808 0.000148 ***
## Aroma        0.225969   0.020378  11.089  < 2e-16 ***
## Aftertaste   0.490988   0.021938  22.381  < 2e-16 ***
## Body         0.103438   0.022926   4.512 7.11e-06 ***
## Acidity      0.225638   0.020994  10.748  < 2e-16 ***
## Sweetness    0.033445   0.009582   3.491 0.000501 ***
## wet         -0.032852   0.010198  -3.221 0.001313 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1484 on 1112 degrees of freedom
## Multiple R-squared:  0.8073, Adjusted R-squared:  0.8063
## F-statistic: 776.4 on 6 and 1112 DF,  p-value: < 2.2e-16
```

```
# Forward-backward
stepAIC(object = empty, scope = list(upper = full, lower = empty),
        direction = "both", k = log(nrow(coffee)))
```

```
## Start:  AIC=-2427.29
## Flavor ~ 1
##
##              Df Sum of Sq     RSS     AIC
## + Aftertaste  1    93.607  33.465 -3913.3
## + Acidity     1    69.294  57.778 -3302.2
## + Aroma       1    68.457  58.615 -3286.1
## + Balance     1    68.173  58.899 -3280.7
## + Body        1    58.232  68.840 -3106.2
## + Uniformity  1     5.778 121.294 -2472.3
## + wet         1     2.313 124.759 -2440.8
## + Sweetness   1     2.300 124.772 -2440.7
## + Moisture    1     2.239 124.833 -2440.2
## <none>                    127.072 -2427.3
## + semi        1     0.331 126.741 -2423.2
##
## Step:  AIC=-3913.29
## Flavor ~ Aftertaste
##
##              Df Sum of Sq     RSS     AIC
## + Acidity     1     4.995  28.470 -4087.2
## + Aroma       1     4.908  28.557 -4083.7
## + Body        1     2.255  31.210 -3984.3
## + Balance     1     1.737  31.729 -3965.9
## <none>                     33.465 -3913.3
## + Sweetness   1     0.138  33.327 -3910.9
## + Uniformity  1     0.114  33.351 -3910.1
## + wet         1     0.087  33.378 -3909.2
## + Moisture    1     0.038  33.427 -3907.5
## + semi        1     0.018  33.448 -3906.9
## - Aftertaste  1    93.607 127.072 -2427.3
##
```

```
## Step:  AIC=-4087.17
## Flavor ~ Aftertaste + Acidity
##
##              Df Sum of Sq    RSS     AIC
## + Aroma       1    3.0217 25.448 -4205.7
## + Body        1    0.8956 27.574 -4115.9
## + Balance     1    0.6542 27.816 -4106.2
## + wet         1    0.2256 28.244 -4089.0
## <none>                    28.470 -4087.2
## + Sweetness   1    0.1709 28.299 -4086.9
## + Uniformity  1    0.1143 28.356 -4084.6
## + semi        1    0.0445 28.425 -4081.9
## + Moisture    1    0.0199 28.450 -4080.9
## - Acidity     1    4.9955 33.465 -3913.3
## - Aftertaste  1   29.3075 57.778 -3302.2
##
## Step:  AIC=-4205.7
## Flavor ~ Aftertaste + Acidity + Aroma
##
##              Df Sum of Sq    RSS     AIC
## + Body        1    0.5090 24.939 -4221.3
## + Balance     1    0.3257 25.123 -4213.1
## + wet         1    0.2689 25.179 -4210.6
## + Sweetness   1    0.1901 25.258 -4207.1
## <none>                    25.448 -4205.7
## + Uniformity  1    0.1141 25.334 -4203.7
## + semi        1    0.0417 25.407 -4200.5
## + Moisture    1    0.0195 25.429 -4199.5
## - Aroma       1    3.0217 28.470 -4087.2
## - Acidity     1    3.1089 28.557 -4083.7
## - Aftertaste  1   15.5890 41.037 -3678.0
##
## Step:  AIC=-4221.29
## Flavor ~ Aftertaste + Acidity + Aroma + Body
##
##              Df Sum of Sq    RSS     AIC
## + Sweetness   1    0.2233 24.716 -4224.3
## + wet         1    0.1835 24.756 -4222.5
## + Uniformity  1    0.1723 24.767 -4222.0
## <none>                    24.939 -4221.3
## + Balance     1    0.1325 24.807 -4220.2
## + Moisture    1    0.0585 24.881 -4216.9
## + semi        1    0.0404 24.899 -4216.1
## - Body        1    0.5090 25.448 -4205.7
## - Acidity     1    2.4119 27.351 -4125.0
## - Aroma       1    2.6351 27.574 -4115.9
## - Aftertaste  1   11.9225 36.862 -3791.1
##
## Step:  AIC=-4224.33
## Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness
##
##              Df Sum of Sq    RSS     AIC
## + wet         1    0.2285 24.488 -4227.7
## <none>                    24.716 -4224.3
## + Balance     1    0.1186 24.598 -4222.7
## - Sweetness   1    0.2233 24.939 -4221.3
## + Uniformity  1    0.0755 24.641 -4220.7
## + Moisture    1    0.0386 24.677 -4219.1
## + semi        1    0.0377 24.678 -4219.0
## - Body        1    0.5422 25.258 -4207.1
## - Acidity     1    2.4163 27.132 -4127.0
## - Aroma       1    2.6440 27.360 -4117.6
## - Aftertaste  1   11.4557 36.172 -3805.2
##
## Step:  AIC=-4227.7
## Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness + wet
##
##              Df Sum of Sq    RSS     AIC
## <none>                    24.488 -4227.7
## + Balance     1    0.0999 24.388 -4225.3
## + Uniformity  1    0.0843 24.403 -4224.5
## - wet         1    0.2285 24.716 -4224.3
## + Moisture    1    0.0467 24.441 -4222.8
## - Sweetness   1    0.2683 24.756 -4222.5
## + semi        1    0.0002 24.487 -4220.7
## - Body        1    0.4483 24.936 -4214.4
## - Acidity     1    2.5437 27.031 -4124.1
## - Aroma       1    2.7077 27.195 -4117.4
## - Aftertaste  1   11.0308 35.518 -3818.6
```

```
##
## Call:
## lm(formula = Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness +
##     wet, data = coffee)
##
## Coefficients:
## (Intercept)   Aftertaste      Acidity        Aroma         Body    Sweetness
##    -0.60900      0.49099      0.22564      0.22597      0.10344      0.03345
##         wet
##    -0.03285
```

```
m_h <- stepAIC(object = empty, scope = list(upper = full, lower = empty),
               direction = "both", trace = 0, k = log(nrow(coffee)))
summary(m_h)
```

```
##
## Call:
## lm(formula = Flavor ~ Aftertaste + Acidity + Aroma + Body + Sweetness +
##     wet, data = coffee)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65627 -0.08781  0.00032  0.08529  0.63010
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.609003   0.159910  -3.808 0.000148 ***
## Aftertaste   0.490988   0.021938  22.381  < 2e-16 ***
## Acidity      0.225638   0.020994  10.748  < 2e-16 ***
## Aroma        0.225969   0.020378  11.089  < 2e-16 ***
## Body         0.103438   0.022926   4.512 7.11e-06 ***
## Sweetness    0.033445   0.009582   3.491 0.000501 ***
## wet         -0.032852   0.010198  -3.221 0.001313 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1484 on 1112 degrees of freedom
## Multiple R-squared:  0.8073, Adjusted R-squared:  0.8063
## F-statistic: 776.4 on 6 and 1112 DF,  p-value: < 2.2e-16
```

```
# 10 variables is still a fairly small problem:  in this example
# all 3 approaches identify the same BIC-based model as the exhaustive search.
```

Checking model assumptions

Recall: $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \text{MVN}(0, \sigma^2 I_n)$. Practically, this means

$$\varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

- independence among all error terms
- normally distributed
- since $\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ is implied by $\mathbb{E}[\varepsilon_i] = 0$ for any $x_{i1}, \ldots, x_{ip}$, the linear model is appropriate; that is, it correctly explains response on average.
- constant error variance σ^2

We could assess these assumptions via ε_i's but we can't observe ε_i directly. Rather, we do have an approximation via residuals e_i from the fitted model.

Recall, $\boldsymbol{e} \sim \text{MVN}(0, (I - H)\sigma^2)$. So $\boldsymbol{e}$ and $\boldsymbol{\varepsilon}$ are related:

$$\boldsymbol{e} = \boldsymbol{Y} - X\hat{\boldsymbol{\beta}} = (I - H)\boldsymbol{Y} = (I - H)(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (X\boldsymbol{\beta} - HX\boldsymbol{\beta}) + (I - H)\boldsymbol{\varepsilon} = (I - H)\boldsymbol{\varepsilon}$$

Note: we can't "solve" $\boldsymbol{\varepsilon}$ since $(I - H)$ is not invertible since it is not full rank: recall H and $(I - H)$ are idempotent, so

$$\text{tr}(H) = p + 1 = \text{tr}(X)$$

$$\text{tr}(I - H) = n - (p + 1) < n$$

Similarly, this does not imply $\boldsymbol{Y} = \boldsymbol{\varepsilon}$. So, $e_i = \varepsilon_i - \sum_{j=1}^n h_{ij}\varepsilon_j$ which means e_i is a good approximation to ε when entries of h_{ij} of H are small (which is "usually" the case, especially when n is large).

$$e_i \sim \mathcal{N}(0, \sigma^2(1 - h_{ii})) \iff \frac{e_i - 0}{\sigma\sqrt{1 - h_{ii}}} \sim \mathcal{N}(0, 1)$$

If we plug in $\hat{\sigma}$, that defines the studentized residuals.

$$d_i \equiv \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

Common practice is to use $\boldsymbol{e}$ for the following residual plots/diagnostics, (using d_i is also possible) to check model assumptions:

Plot $\boldsymbol{e}$ versus $\hat{\boldsymbol{\mu}}$ which was shown in A2 were mutually independent since they are multivariate normal with covariance 0.

Typical "good" scatterplot will have a random scatter around $y = 0$ (no visible patterns).

Problematic scatterplot is when variance of e_i is not constant (cone) increases with fitted values.

In general, plot of $\boldsymbol{e}$ and $\hat{\boldsymbol{\mu}}$ can show deviations from independence, constant variance if those assumptions are violated.

Plot $\boldsymbol{e}$ versus $\boldsymbol{x}_j$ for each $j = 1, \ldots, p$ in model when not many predictors. This can help detect non-linearity between $\boldsymbol{x}_j$ and $\boldsymbol{y}$, not as practical when p is large.

Typical "good" scatterplot will have a random scatter around $y = 0$ (no visible patterns).

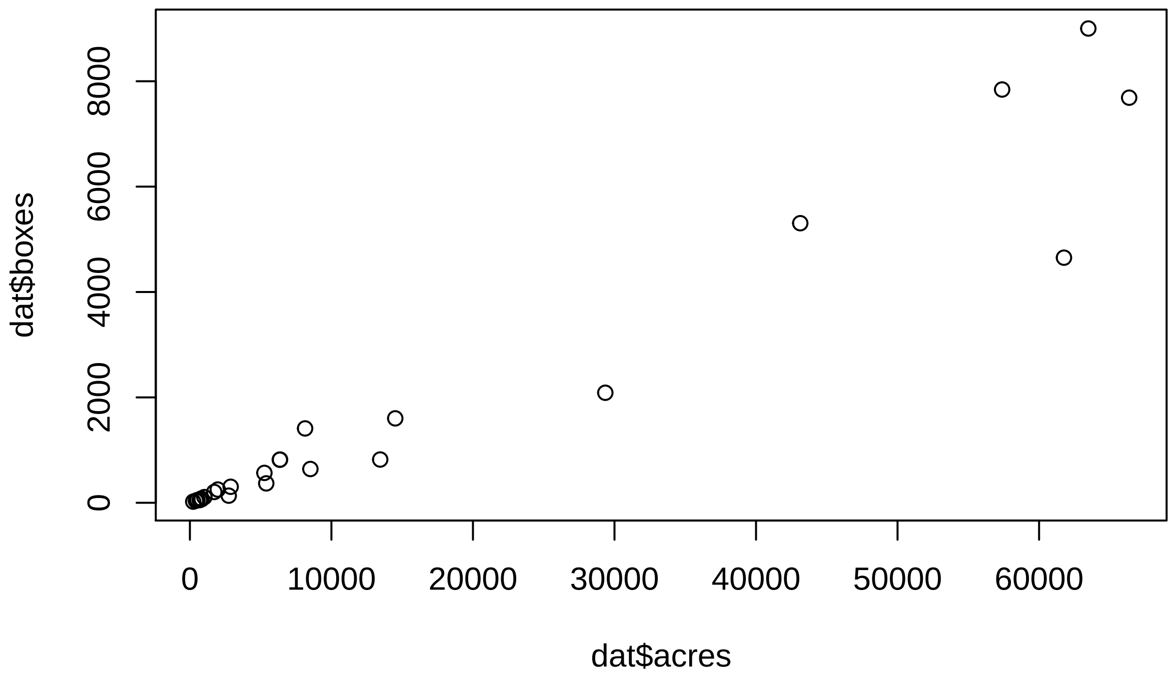
If the observation numbers were collected in some order (time, space, etc.), also plot e_i versus indices i to check for any patterns (again, look for random scatter)

Histogram of e: is it bell shaped and symmetric? to assess Normal assumption, but the histogram can't easily detect overly fat/thing tails.

QQ plot of e more formally assess Normality: scatterplot of ordered quantiles from 2 distributions. In our case: empirical quantiles from from residuals (data) versus theoretical quantiles from assumed Normal distribution. If quantiles roughly match, the points will roughly fall on the 45° line through origin.

If these sets of residuals plots show violations, then that could affect the validity of CIs, HTs, etc.

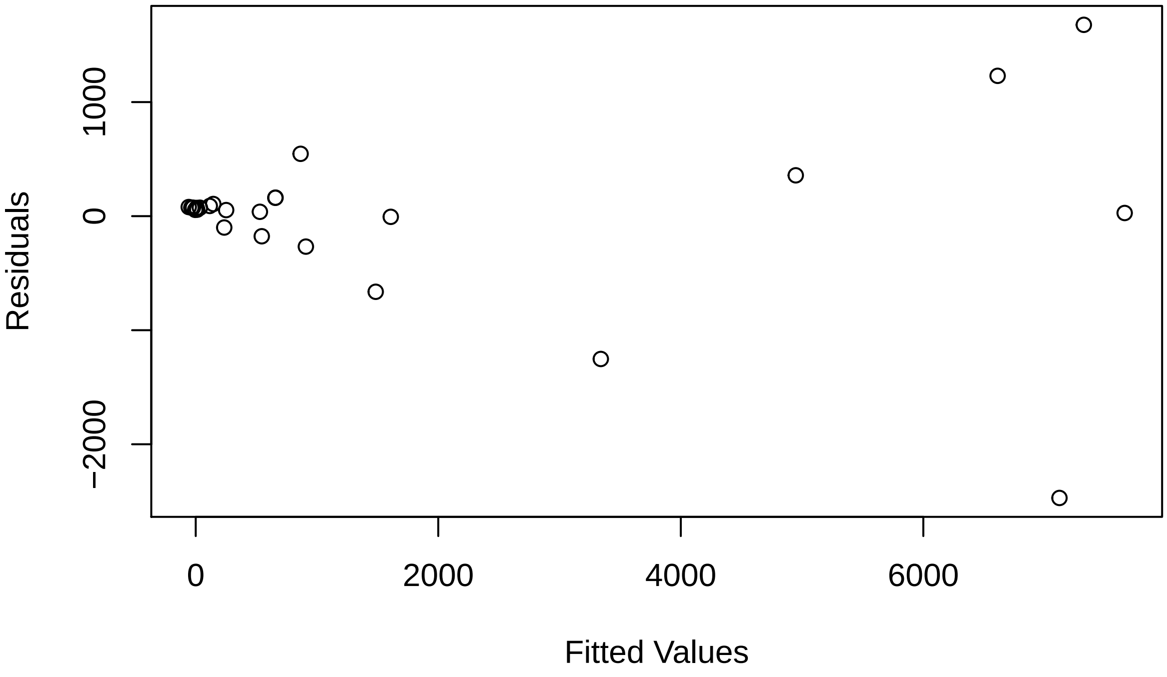
```r
### Residual plots/diagnostics demo

## Florida oranges revisited
dat <- read.csv("florange.csv")
plot(dat$acres,dat$boxes)
```



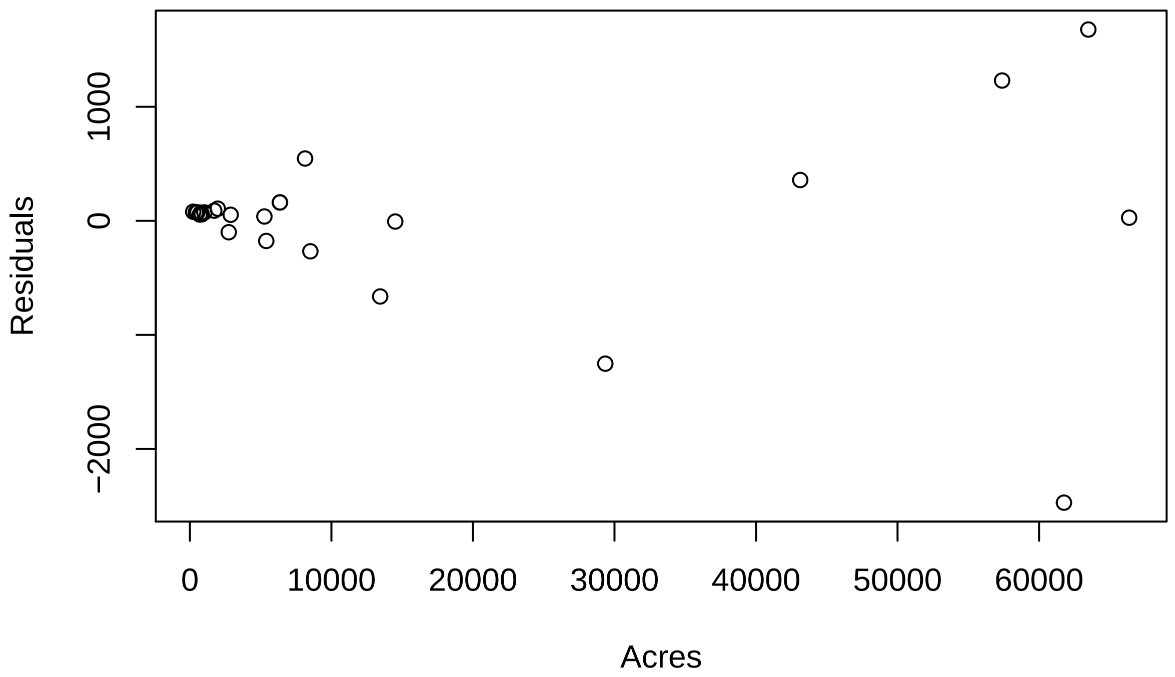
```r
lm.1 <- lm(dat$boxes~dat$acres)
summary(lm.1)
```

```
##
## Call:
## lm(formula = dat$boxes ~ dat$acres)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2470.81    -6.17    71.72   106.46  1677.32
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -85.391989 186.178031  -0.459    0.651
## dat$acres     0.116717   0.006761  17.263 1.16e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 754.4 on 23 degrees of freedom
## Multiple R-squared:  0.9284, Adjusted R-squared:  0.9252
## F-statistic:   298 on 1 and 23 DF,  p-value: 1.164e-14
```

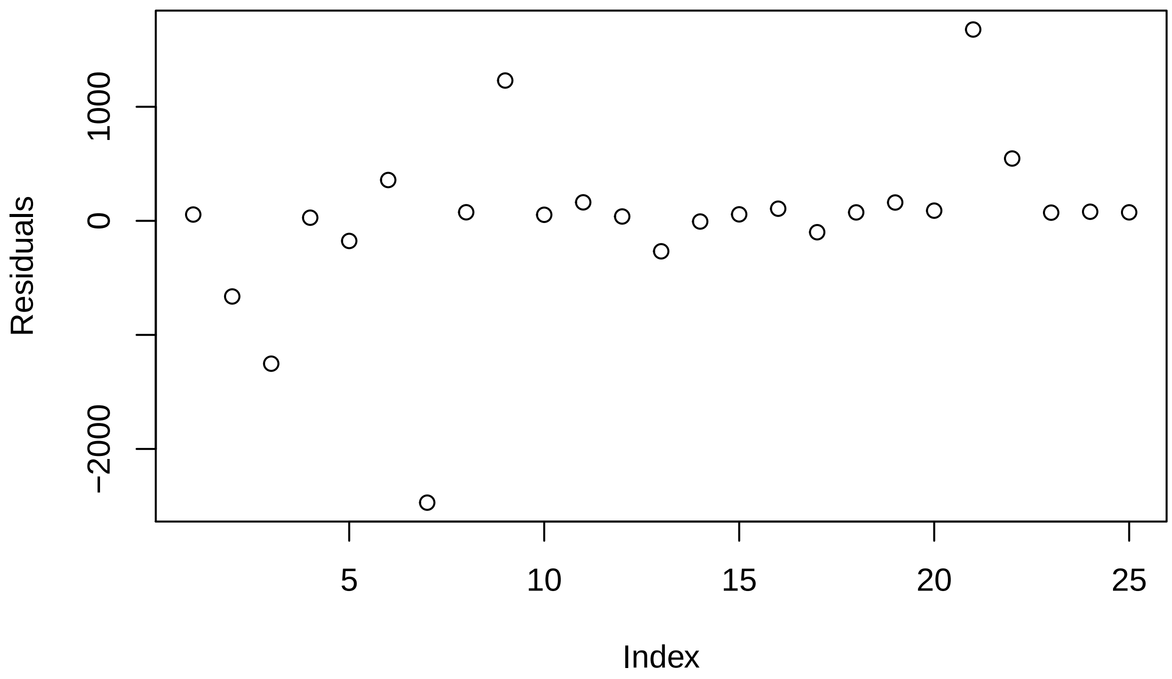
```
# Residual plot: vs fitted values
plot(lm.1$fitted.values, lm.1$residuals, xlab = "Fitted Values", ylab = "Residuals")
```



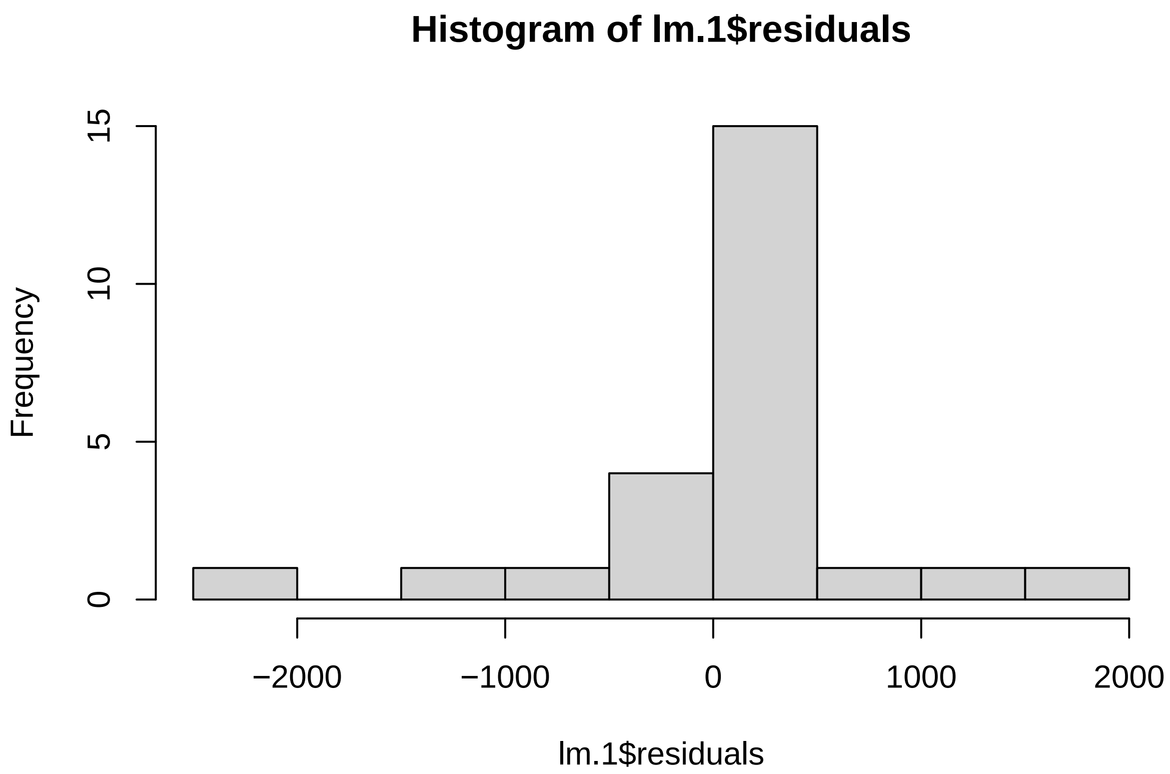
```
# Residual plot: vs predictor (just one in this case)
plot(dat$acres, lm.1$residuals, xlab = "Acres", ylab = "Residuals")
```



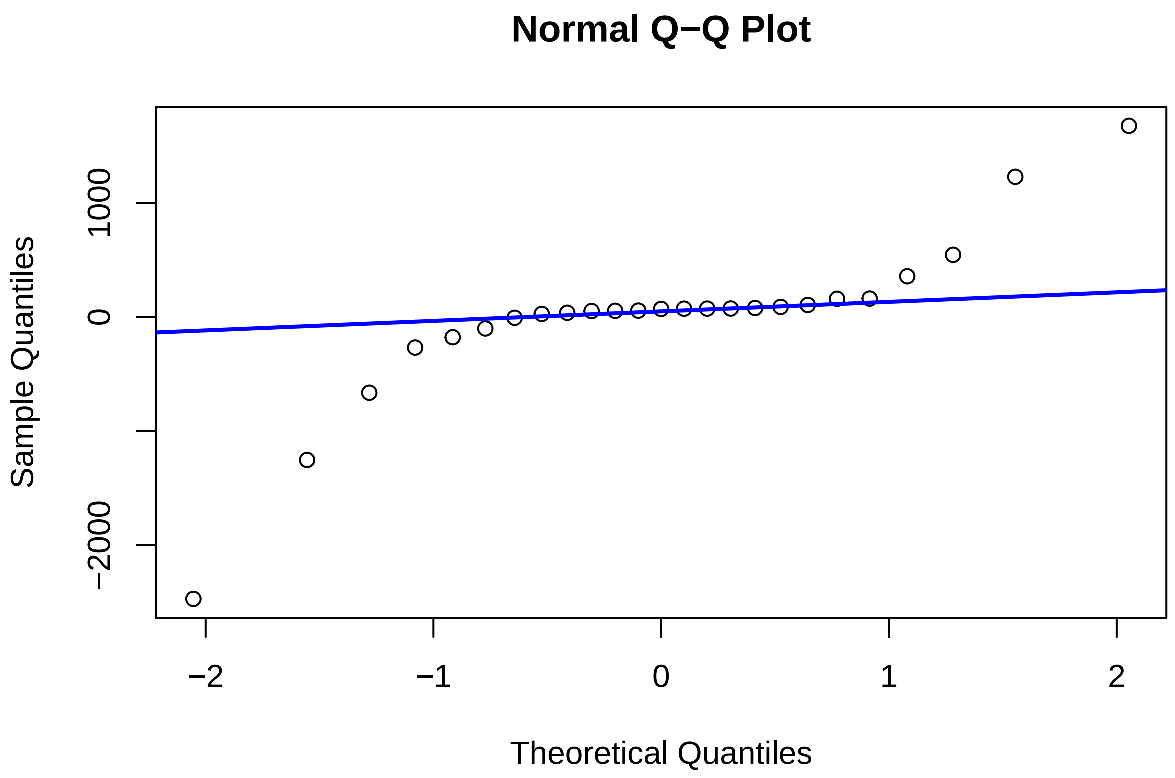
```
# Residual plot: vs i (just to demo plot; no time/space ordering here)
plot(1:nrow(dat), lm.1$residuals, xlab = "Index", ylab = "Residuals")
```



```
# Histogram of residuals
hist(lm.1$residuals)
```



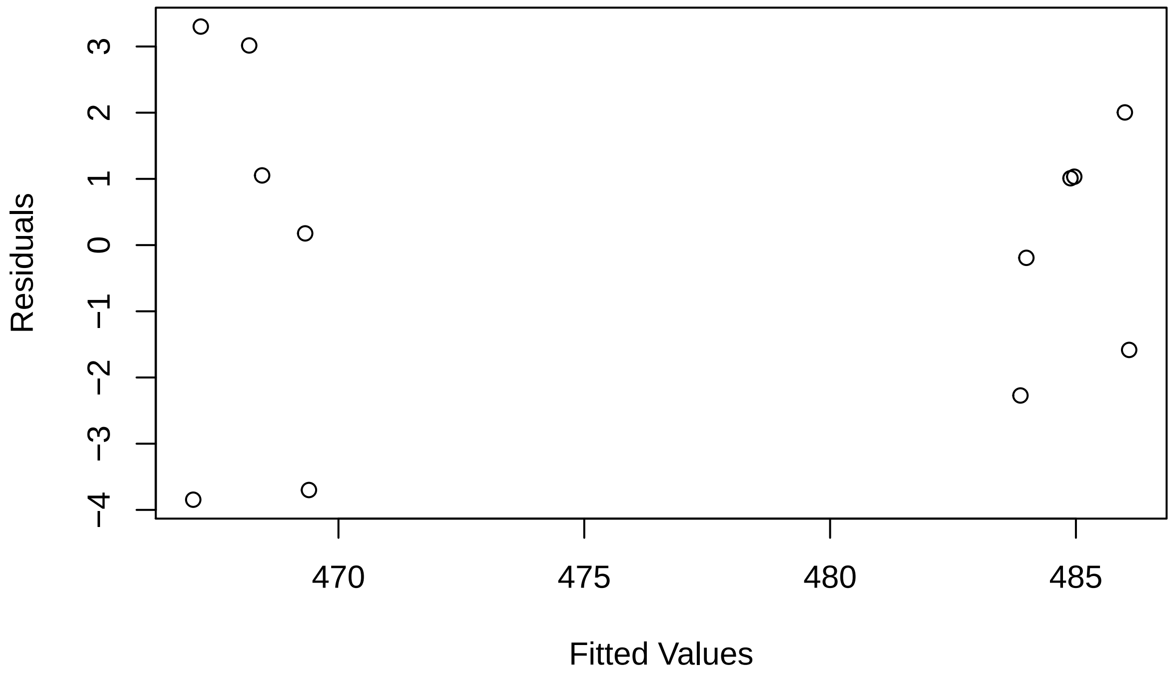
```
# QQ plot of residuals
qqnorm(lm.1$residuals)
qqline(lm.1$residuals, col="blue", lwd = 2)
```



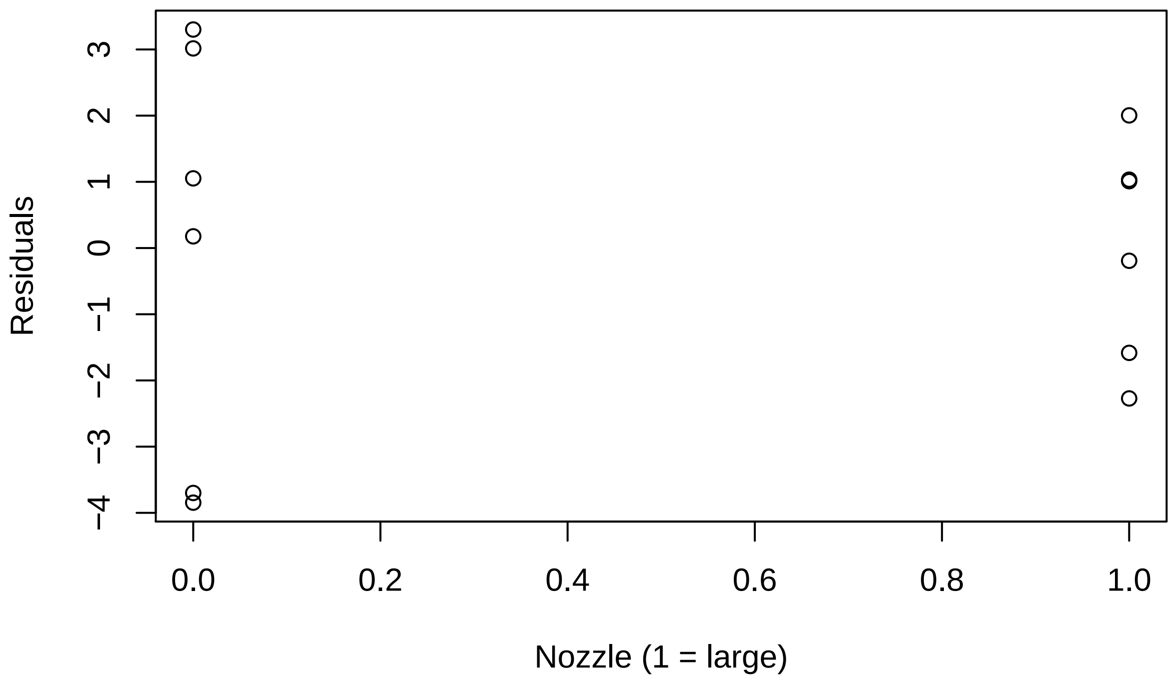
```
## Rocket data revisited
rocket <- read.csv(file="rocket.csv")
mr <- lm(thrust ~ nozzle + propratio, data = rocket)
summary(mr)
```

```
##
## Call:
## lm(formula = thrust ~ nozzle + propratio, data = rocket)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8459 -1.7555  0.5934  1.2906  3.3008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 473.6039     4.7158 100.430 4.88e-15 ***
## nozzle       16.7383     1.5329  10.919 1.71e-06 ***
## propratio    -1.0948     0.9414  -1.163    0.275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.655 on 9 degrees of freedom
## Multiple R-squared:  0.9303, Adjusted R-squared:  0.9148
## F-statistic: 60.05 on 2 and 9 DF,  p-value: 6.238e-06
```

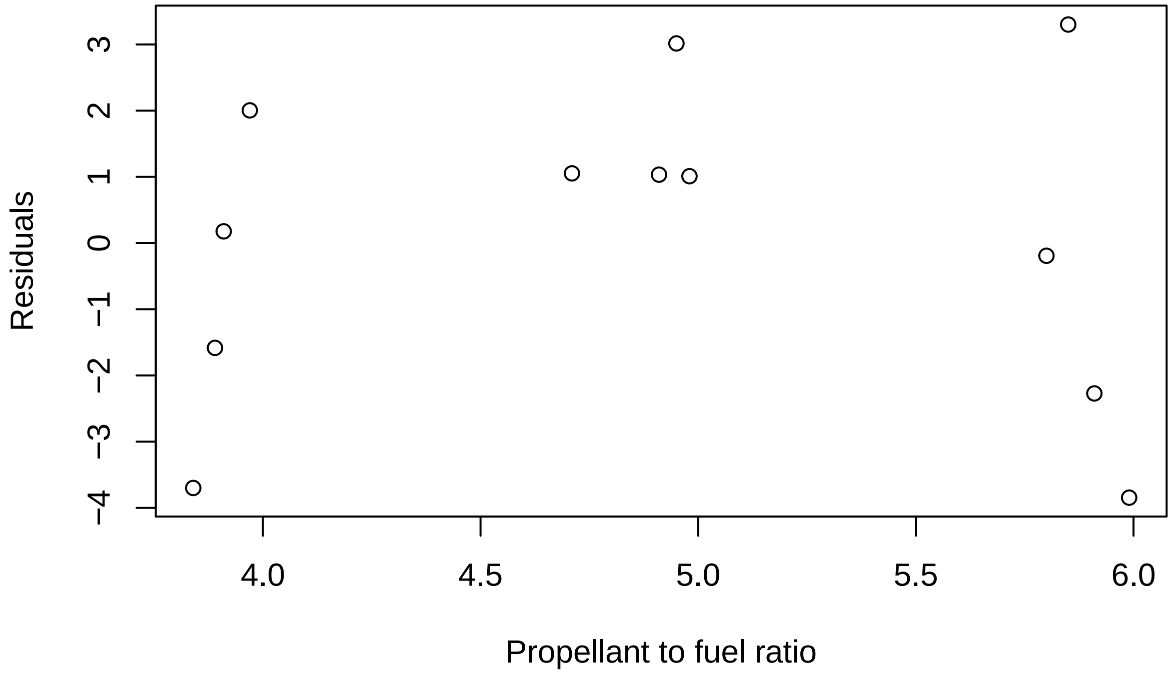
```
# Residual plot: vs fitted values
plot(mr$fitted.values, mr$residuals, xlab = "Fitted Values",
     ylab = "Residuals")
```



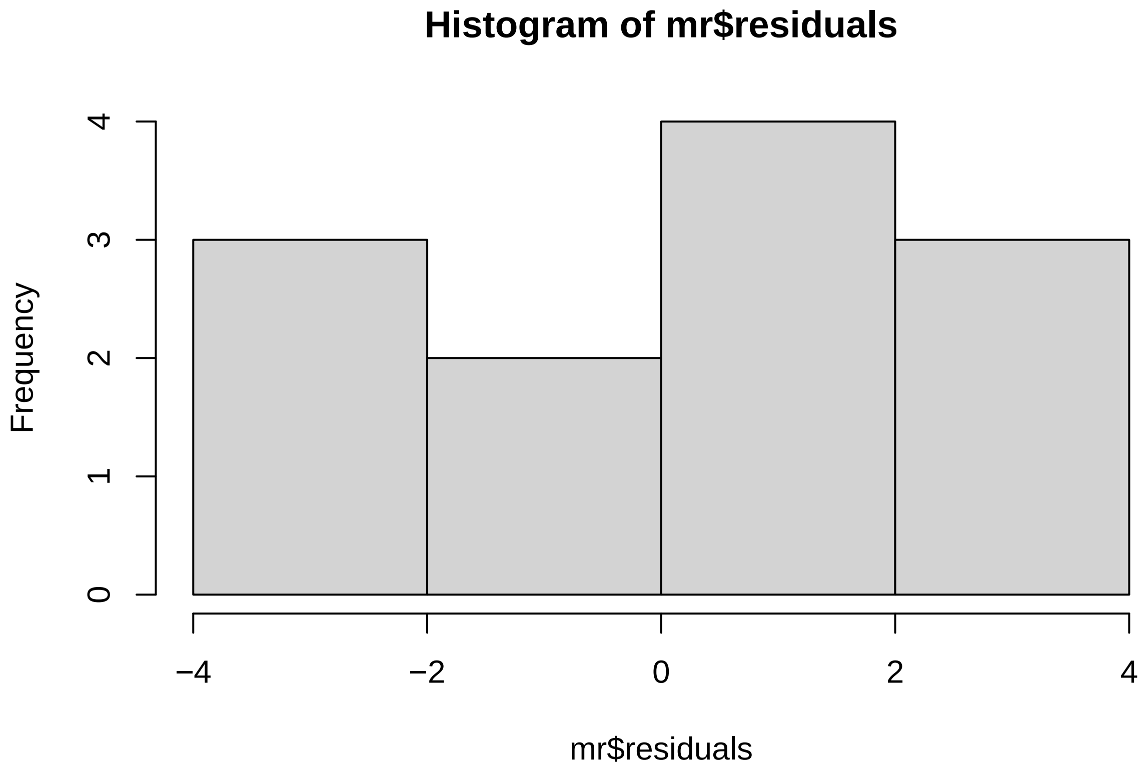
```
# Residual plot: vs predictors
plot(rocket$nozzle, mr$residuals, xlab = "Nozzle (1 = large)",
     ylab = "Residuals")
```



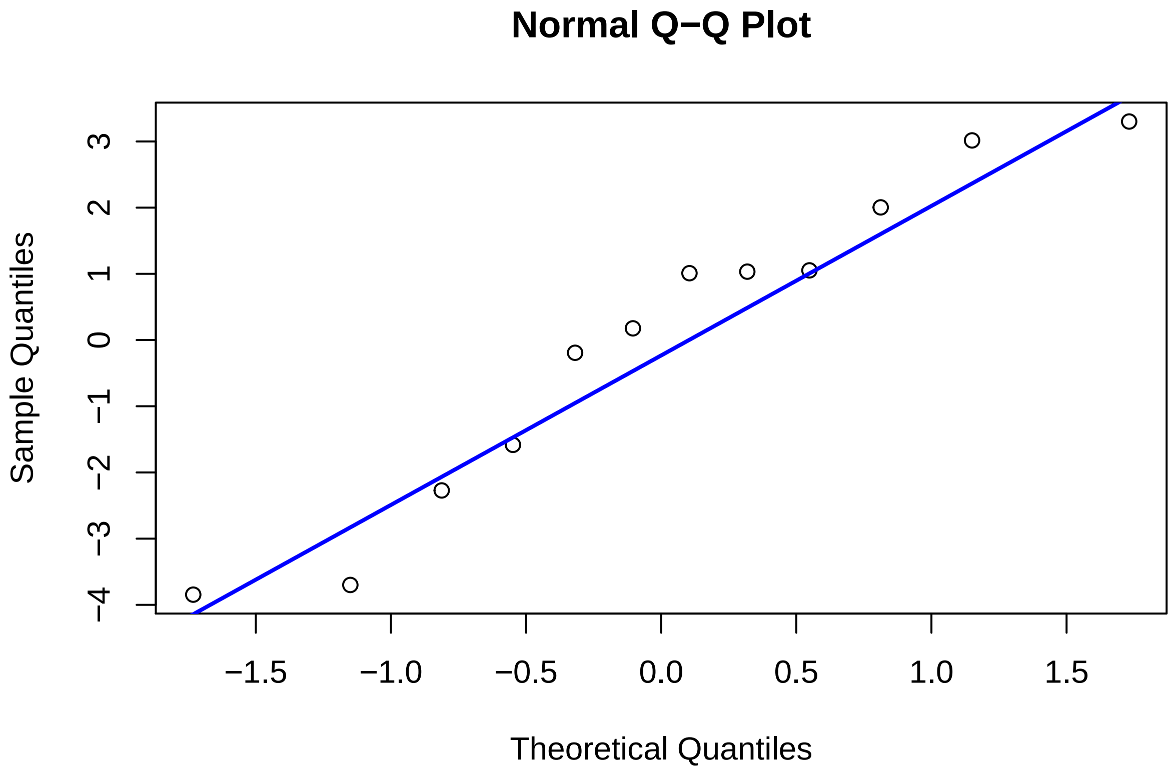
```r
plot(rocket$propratio, mr$residuals, xlab = "Propellant to fuel ratio",
     ylab = "Residuals")
```



```
# Histogram of residuals
hist(mr$residuals)
```



```
# QQ plot of residuals
qqnorm(mr$residuals)
qqline(mr$residuals, col="blue", lwd = 2)
```


Normal Q−Q Plot
Sample Quantiles
−4
−3
−2
−1
0
1
2
3
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles