STAT 331 - Applied Linear Models

Cameron Roopnarine

Last updated: October 10, 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 $\mathsf{E}[Y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ and $\mathsf{Var}(Y) = \mathsf{Var}(\varepsilon) = \sigma^2$.

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

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

Proof. $\dfrac{\partial S}{\partial \beta_0} = 2\displaystyle\sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)](-1)$ 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 $\sigma^2 = \mathsf{Var}(\varepsilon_i) = \mathsf{Var}(Y_i)$

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

SSE $=$ 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, $\mathsf{E}[\hat{\sigma}^2] = \sigma^2$ 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. 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,

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

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

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

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

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

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

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

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 $\mathsf{E}[\hat{\beta}_1] = \beta_1$, 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, $\mathsf{E}[\hat{\beta}_1]$ 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,

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

Proof. Too hard probably. □

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 $1 - \frac{\alpha}{2}$ quantile of $t(n-2)$; that is, $P(|T| \leqslant c) = 1 - \alpha$ or $P(T \leqslant c) = 1 - \frac{\alpha}{2}$ 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 $1 - \frac{\alpha}{2}$ 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, $\mathsf{E}[\hat{Y}_0] = \mathsf{E}[\hat{\beta}_0] + x_0 \mathsf{E}[\hat{\beta}_1] = \beta_0 + \beta_1 x_0 = \mathsf{E}[Y_0]$, 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 $a_i = \frac{1}{n} + \frac{(x_0 - \bar{x})(x_i - \bar{x})}{S_{xx}}$. Therefore,

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

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 $Y_1, \ldots, Y_n$.

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

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

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:

- $\mathsf{Var}(Y_0)$: random error of new observation
- $\mathsf{Var}(\hat{Y}_0)$ (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 y_0:

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

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

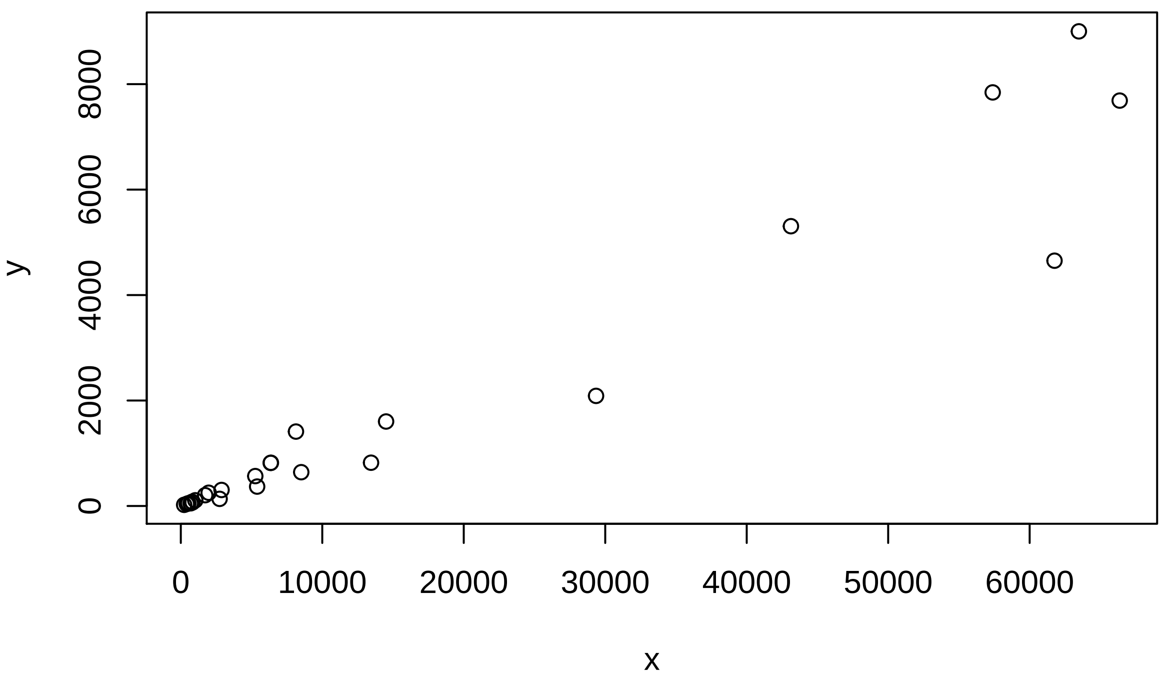
We are **not** trying to establish causation.

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
# 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 $\hat{\beta}_0 = -85.391989$, $\hat{\beta}_1 = 0.116717$, $\mathrm{Se}(\hat{\beta}_1) = 0.006761$, $t = 17.263$, p-value $= 1.64 \times 10^{-14}$, and $\hat{\sigma} = 754.4$.

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

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

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

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

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

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

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

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

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

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

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

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

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

Therefore, $c = 2.07$.

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

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

```r
# 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 $x_1, x_2, \ldots, x_p$.

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 $(x_{i1}, x_{i2}, \ldots, x_{ip})$. Then,

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

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

DEFINITION 3.4: Mean vector

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

DEFINITION 3.5: Covariance matrix

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

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

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* $\mathsf{Var}(\boldsymbol{Y})^\top = \mathsf{Var}(\boldsymbol{Y})$.

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

(3) $\mathsf{Var}(\boldsymbol{Y}) = \mathsf{E}\left[(\boldsymbol{Y} - \mathsf{E}[\boldsymbol{Y}])\,(\boldsymbol{Y} - \mathsf{E}[\boldsymbol{Y}])^\top\right]$

Proof. Trivial. □

PROPOSITION 3.7: Properties of Random Vector

et $\boldsymbol{a}$ 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) $\mathsf{E}[\boldsymbol{a}\boldsymbol{Y}] = \boldsymbol{a}\boldsymbol{Y}$

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

(3) $\mathsf{Var}(\boldsymbol{a}\boldsymbol{Y}) = \boldsymbol{a}\mathsf{Var}(\boldsymbol{Y})\boldsymbol{a}^\top$

(4) $\mathsf{Var}(A\boldsymbol{Y}) = A\mathsf{Var}(\boldsymbol{Y})A^\top$

Proof. We prove prove property (4) only.

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

□

EXAMPLE 3.8: Calculations with MLR Varaibles

Let $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^\top$. Suppose $\mathsf{E}[\boldsymbol{Y}] = (3, 1, 2)^\top$. Let $\mathsf{Var}(Y) = \begin{bmatrix} 4 & 1/2 & -2 \\ 1/2 & 1 & 0 \\ -2 & 0 & 3 \end{bmatrix}$ 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) $\mathsf{E}[\boldsymbol{a}\boldsymbol{Y}]$

(ii) $\mathsf{Var}(\boldsymbol{a}\boldsymbol{Y})$

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

(iv) $\mathsf{Var}(A\boldsymbol{Y})$

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

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

(ii)

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

DEFINITION 3.9: Multivariate normal distribution

Let $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$ be a random vector. We say that $Y \sim \mathrm{MVN}(\mu, \sigma)$; 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:

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

that is, Y_i and Y_j are uncorrelated.

$$
\Sigma_{ij} = 0
$$

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

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

since $\mathsf{Cov}(\varepsilon_1, \varepsilon_2) = 0$ 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

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

THEOREM 3.12: Least Square Estimates (LSEs) for MLR

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

Proof. The first partial is $\frac{\partial S}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \mu_u)(-1)$, and all other partials for $j = 1, \ldots, p$ 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 $j = 1, \ldots, p$ to get

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

since we recall that

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

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

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

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

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

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

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

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

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

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: $\beta_j = 0$ vs H_A: $\beta_j \neq 0$, 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 $\widehat{\mathsf{E}[Y]}$ (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.

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

EXAMPLE 3.16: 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: $\beta_2 = 0$ vs H_A: $\beta_2 \neq 0$ 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: $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top \in \mathbb{R}^n$

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

The span of X is $\text{Span}(X) = \{b_0\mathbf{1} + b_1\boldsymbol{x}_1 + \cdots + b_p\boldsymbol{x}_p : b_0, \ldots, b_p \in \mathbb{R}\} \subset \mathbb{R}^n$ 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 $\boldsymbol{b} = (b_0, b_1, \ldots, b_p)^\top$.

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 3.17: Hat matrix

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

PROPOSITION 3.18: 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. 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

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

$$\begin{aligned}
\boldsymbol{e} &= \boldsymbol{Y} - \hat{\boldsymbol{\mu}} = I\boldsymbol{Y} - H\boldsymbol{Y} = (I - H)\boldsymbol{Y} \\
\mathsf{E}[\hat{\boldsymbol{\mu}}] &= \mathsf{E}[H\boldsymbol{Y}] = H\mathsf{E}[\boldsymbol{Y}] = X(X^\top X)^{-1}X^\top \underbrace{X\boldsymbol{\beta}}_{\mathsf{E}[\boldsymbol{Y}]} = X\boldsymbol{\beta} \\
\mathsf{Var}(\hat{\boldsymbol{\mu}}) &= \mathsf{Var}(H\boldsymbol{Y}) = H\mathsf{Var}(\boldsymbol{Y})H^\top = H\sigma^2 I H^\top = \sigma^2(HH^\top) = \sigma^2 H \\
\mathsf{E}[\boldsymbol{e}] &= \mathsf{E}[(I-H)\boldsymbol{Y}] = \mathsf{E}[\boldsymbol{Y}] - \mathsf{E}[H\boldsymbol{Y}] = X\boldsymbol{\beta} - X\boldsymbol{\beta} = 0 \\
\mathsf{Var}(\boldsymbol{e}) &= (I-H)\mathsf{Var}(\boldsymbol{Y})(I-H)^\top = \sigma^2(I-H)(I-H)^\top = \sigma^2(I-H)
\end{aligned}$$

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

THEOREM 3.19: Distribution of $\hat{\mu}$ and e

$\hat{\boldsymbol{\mu}}$ and $\boldsymbol{e}$ 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

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

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

We have proved the following theorem.

THEOREM 3.20: 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 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,

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

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

We have proved the following theorem.

THEOREM 3.21: 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 y_0 is given by

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

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

REMARK 3.22

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

Lecture 8 | 2020-09-30

LEARN: `rocket`

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 $\boldsymbol{x}_4 = \mathbf{1} - \boldsymbol{x}_2 - \boldsymbol{x}_3$. 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 $\hat{\boldsymbol{\beta}}_j \sim N(\beta_j, \sigma^2 V_{jj})$ with $\mathsf{Se}(\hat{\sigma}\sqrt{V_{jj}})$
- What about $\beta_2 - \beta_3$?

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

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:"

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

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

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

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

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 span(X), so $\boldsymbol{e}$ is orthogonal to span(X), so $\hat{\boldsymbol{\mu}}^\top \boldsymbol{e} = 0$.

Table 1: ANOVA Table

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

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

We call the **coefficient of determination** $R^2 = SS(\text{Reg})/SS(\text{Total}) = 1 - SS(\text{Res})/SS(\text{Total})$. 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 $SS(\text{Res})$. 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).