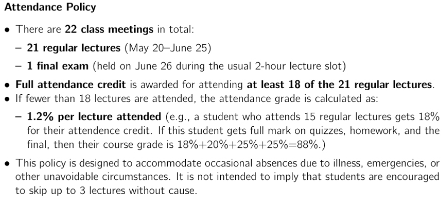
Rest of Chapter 14

MA 116

June 2025

Attendance policy

OLD

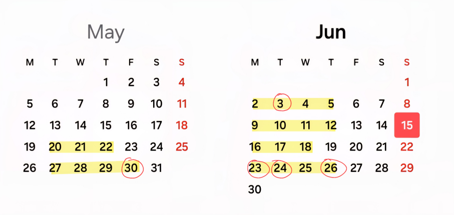


Attendance Policy

- There are **22 class meetings** in total:
  - **21 regular lectures** (May 20–June 25)
  - **1 final exam** (held on June 26 during the usual 2-hour lecture slot)
- **Full attendance credit** is awarded for attending **at least 18 of the 21 regular lectures**.
- If fewer than 18 lectures are attended, the attendance grade is calculated as:
  - **1.2% per lecture attended** (e.g., a student who attends 15 regular lectures gets 18% for their attendence credit. If this student gets full mark on quizzes, homework, and the final, then their course grade is 18%+20%+25%+25%=88%.)
- This policy is designed to accommodate occasional absences due to illness, emergencies, or other unavoidable circumstances. It is not intended to imply that students are encouraged to skip up to 3 lectures without cause.

UPDATED

You are allowed to miss up to 3 regular lectures and still get full attendance credits. May 30, Jun 3, 23, 24, 26 are not considered regular lectures.



May

M	T	W	T	F	S	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Jun

M	T	W	T	F	S	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

#Regular lectures=17
If fewer than 14 lectures are attended, you get 1.6% $\times$ n attendance credits where n=number of regular lectures attended.

Quiz 4 and Final Exam

Quiz 4: two questions

Question 1: Given a sample of small size, calculate $\overline{x}$, $\overline{y}$, s_x, s_y, r. Find the lease squares regression line of this sample. Draw a residual plot. Deduct from the residual plot that a linear relation between x and y is reasonable. Conduct a hypothesis test to test whether there is a (positively/negatively associated) linear relation between the response variable and the explanatory variable.

Question 2: Short answer question about estimators in 14.1 and 14.2

Final Exam: June 26 3-5pm (two hours)

Covers Quiz 1-4 content and multiple regression + qualitative (dummy) variable.

Reminder

You have to score at least 30% on the final exam to pass this course

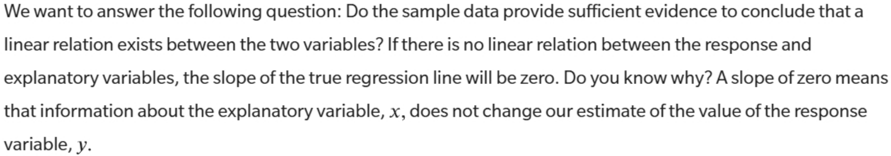
A calculator that's able to compute square roots is required for Quiz 4 and the final exam.

Review: 14.2

Confidence interval for a mean response vs. prediction interval for an individual response. [diagram]

- What are the point estimator for both interval estimators?
- What is the quantity that we want to estimate, if we are using a (confidence interval for a mean response/prediction interval for an individual response)?
- Fix a value of x, which interval has a larger length?

Example. Review 14.1: Hypothesis test about β_1



We want to answer the following question: Do the sample data provide sufficient evidence to conclude that a linear relation exists between the two variables? If there is no linear relation between the response and explanatory variables, the slope of the true regression line will be zero. Do you know why? A slope of zero means that information about the explanatory variable, x, does not change our estimate of the value of the response variable, y.

1. the null hypothesis $H_0 : \beta_1 = 0$ means there is no linear relation between the explanatory and response variables.
2. the alternative hypothesis $H_1 : \beta_1 \neq 0$ means there is linear relation between the explanatory and response variables.
3. the alternative hypothesis $H_1 : \beta_1 > 0$ means there is linear relation between the explanatory and response variables that is positively associated.
4. the alternative hypothesis $H_1 : \beta_1 > 0$ means there is linear relation between the explanatory and response variables that is negatively associated.

Example of hypothesis test about β_1

Suppose we want to test if temperature (explanary variable) is in linear relation to garlic leaves growing speed (response variable).

x_i	y_i
50	1
60	3
70	3
80	3

The table shows a sample we get.

1. Step I. Calculate $\overline{x}$, $\overline{y}$, s_x and s_y.
2. Step II. Calculate r.

Example of hypothesis test about β_1

Suppose we want to test if temperature (explanary variable) is in linear relation to garlic leaves growing speed (response variable), to a significance level $\alpha = 0.05$. Before we determine H_0 and H_1, we first calculate all the values that might be used in our hypothesis test.

x_i	y_i
50	1
60	3
70	3
80	3

The table shows a sample we get.

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775$.

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775.$
3. Calculate
$$b_1 = r\frac{s_y}{s_x}$$
and
$$b_0 = \overline{y} - b_1\overline{x}.$$

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775.$
3. $$b_1 = r\frac{s_y}{s_x} = 0.06$$

   and

   $$b_0 = \overline{y} - b_1\overline{x} = -1.4.$$

   Therefore, the least squares regression line we obtained from this sample is

   $$\hat{y} = 0.06x - 1.4.$$

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775$.
3. Step III. $\hat{y} = 0.06x - 1.4$.
4. Step IV. Calculate the residues $y_i - \hat{y}_i$ and draw a residual plot for this least squares regression line.

x_i	y_i	$\hat{y}_i = 0.06x_i - 1.4$	residual
50	1		
60	3		
70	3		
80	3		

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775$.
3. Step III. $\hat{y} = 0.06x - 1.4$.
4. Step IV. Calculate the residues and draw a residual plot for this least squares regression line. Since the sample data points seem to scatter around the zero line in the residual plot without a pattern, we say that a linear relationship between x and y is reasonable.

x_i	y_i	$\hat{y}_i = 0.06x_i - 1.4$	residual
50	1	1.6	-0.6
60	3	2.2	0.8
70	3	2.8	0.2
80	3	3.4	-0.4

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775$.
3. Step III. $\hat{y} = 0.06x - 1.4$.
4. Step IV.

x_i	y_i	$\hat{y}_i = 0.06x_i - 1.4$	residual
50	1	1.6	-0.6
60	3	2.2	0.8
70	3	2.8	0.2
80	3	3.4	-0.4

5. Step V. Calculate

$$s_e = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_i \text{residual}^2}{n-2}}$$

and

$$s_{b_1} = \frac{s_e}{(\sqrt{n-1})s_x}$$

.

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775.$
3. Step III. $\hat{y} = 0.06x - 1.4$.
4. Step IV.

x_i	y_i	$\hat{y}_i = 0.06x_i - 1.4$	residual
50	1	1.6	-0.6
60	3	2.2	0.8
70	3	2.8	0.2
80	3	3.4	-0.4

5. Step V.

$$s_e = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_i \text{residual}^2}{n-2}} = \sqrt{\frac{1.2}{2}} = 0.77,$$

$$s_{b_1} = \frac{s_e}{(\sqrt{n-1})s_x} = 0.034.$$

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775.$
3. Step III. $\hat{y} = 0.06x - 1.4$.
4. Step IV.

x_i	y_i	$\hat{y}_i = 0.06x_i - 1.4$	residual
50	1	1.6	-0.6
60	3	2.2	0.8
70	3	2.8	0.2
80	3	3.4	-0.4

5. Step V. $s_e = \sqrt{\dfrac{\sum_i (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\dfrac{\sum_i \text{residual}^2}{n-2}} = \sqrt{\dfrac{1.2}{2}} = 0.77,$

   $s_{b_1} = \dfrac{s_e}{(\sqrt{n-1})s_x} = 0.034.$
6. Step VI. Hypothesis test. H_0, $H_1 : \beta_1 \neq 0$, test type. Critical value(s). Calculate test statistic. Conclusion.

Suppose we want to test if temperature (explanary variable) is in linear relation to garlic leaf growing speed (response variable) at $\alpha = 0.05$.

$H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$. Two-tailed test.
Assume that H_0 is true. Since we have verified that a linear relation between x and y is reasonable by looking at the residual plot, we further assume that ϵ_i is normally distributed so that the new variable

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

follows Student's t-distribution with $df = n - 2 = 2$.
My critical values are $\pm t_{0.025} = \pm 4.303$ at $df = 2$.
My test statistic is $t_0 = \dfrac{b_1 - \beta_1}{s_{b_1}} = \dfrac{0.06 - 0}{0.034} = 1.76$, which does not fall in the critical region.
Conclusion: There is not sufficient evidence to conclude that there is a linear relation between temperature and garlic leaf growing speed.

14.2 Example 1

A real estate analyst wants to predict the **selling price of houses** (denoted by y, response variable) based on their **size** (in square feet) (denoted by x, explanatory variable). In particular, the real estate analyst wants to estimate the average selling price of houses that are about 1500 sq ft. The real estate analyst has a sample that consisting of data points of the form (x_i, y_i), where x_i varies discretely between 300 and 3500.

1. Suppose this real estate analyst wants to obtain a point estimator of the average selling price of houses that are about 1500 sq ft from the sample they have. Describe the steps of calculation.
2. To obtain an interval estimator of the average selling price of houses that are about 1500 sq ft, the real estate analyst would calculate a confidence interval for a mean response. If, instead of a **confidence interval for a mean response**, the real estate analyst calculates a **prediction interval for an individual response**, what quantity is being estimated?

Multiple linear regression: quantitative data

Multiple linear regression equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

where y is called the dependent variable, $x_1, \ldots, x_k$ are the independent variables. (Not to be confused with data points x_i!)

Like the simple linear regression case, $\beta_1, \ldots, \beta_k$ are population parameters. From a sample, one may use technologies or more advanced math tools (linear algebra, not covered in this course) to obtain a **least squares prediction equation**

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k.$$

(I.e. from a sample, the technology/math tool used calculates b_i for $0 \leq i \leq k$ with the intention of finding a linear relation that minimizes sum of square residuals.)

Example

A collector of antique grandfather clocks sold at auction believes that the price received for the clocks depends linearly on both the age of the clocks and the number of bidders at the auction. Thus, he hypothesizes a model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ where y=Auction proce (dollars), x_1=Age of clock (years), x_2=Number of bidders.

A sample of 32 auction prices of grandfather clocks, along with their age and the number of bidders, is collected.

index	Age x_1	Number of bidders x_2	Auction Price y
1	127	13	1235
2	115	12	1080
3	127	7	845
$\vdots$	$\vdots$	$\vdots$	$\vdots$
32	18	7	1262

A statistician uses technology to find out that the **least squares prediction equation** of this sample is

$$\hat{y} = -1339 + 12.74x_1 + 85.95x_2.$$

This is equivalent to say that this statistician finds that $b_0 = -1339$, $b_1 = 12.74$, and $b_2 = 85.95$ for this sample. b_0 is a point estimator for β_0, b_1 is a point estimator for β_1, b_2 is a point estimator for β_2.

Interpret β_i

For a multiple linear regression equation $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$, β_1 represents the slope of the line relating y to x_1 for fixed $x_2, ..., x_k$. That is, β_1 measures the change in μ_y for every one unit increase in x_1 when the other independent variables in the model are held fixed. A similar statement can be made about β_i, $i \geq 1$.

In this example, we interpret $b_1 = 12.74$ by the following statement.

We estimate the mean auction price μ_y of an antique clock to increase \$12.74 for every 1-year increase in age x_1 when the number of bidders x_2 is held fixed.

We can say this because b_1 is a point estimator of β_1.

Example.

Suppose we are given the least squares prediction equation of this sample, $\hat{y} = -1339 + 12.74x_1 + 85.95x_2$.

1. Predict the auction price of a grandfather clock if it is of an age 150 years, and there are 10 bidders at the auction.
2. A collector owns a grandfather clock of age 150 years. Suppose this collector wants to make at least \$1500 from the auction of this clock. Predict the smallest number of bidders for the collector to achieve this amount. (Round up to the next integer!)

Hypothesis test about β_i, $i > 0$

Consider the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon.$$

If the random error ϵ obeys the following four assumptions, then we may make inference about β_i, $i > 0$. [these assumptions will not be tested; take it for granted]

1. Random errors are independent
2. It has a mean=0
3. It has a variance equal to σ^2, the variance of y
4. It is a normal variable

If the random error ϵ obeys these four assumptions, the new variable

$$t = \frac{b_i - \beta_i}{s_{b_i}}$$

follows Student's t-distribution with $df = n - (k + 1)$.

Example

In the setting of our previous example, test the hypothesis that the mean auction price of a clock increases as the number of bidders increases when the age is held constant, at $\alpha = 0.05$. **Given that $\mathbf{s_{b_2} = 8.729}$.**

$H_0 : \beta_2 = 0$, $H_1 : \beta_2 > 0$. Right-tailed test.
Assume that H_0 is true. The new variable

$$t = \frac{b_2 - \beta_2}{s_{b_2}} = \frac{b_2}{8.729}$$

follows the Student's t-distribution with
$df = n - (k+1) = 32 - (2+1) = 29$.
The critical value is $t_{0.05}$ with $df = 29$, which is 1.699.
The test statistic is $t_0 = \dfrac{85.95}{8.729} = 9.85$, which falls in the critical region.
Conclusion: There is sufficient evidence to conclude that the mean auction price of a clock increases as the number of bidders increases when the age is held constant.

Qualitative (Dummy) Variable Models

A researcher wants to model how age associates to whether an American voted or not in the last federal election. In particular, this researcher wants to investigate if the average age of Americans who did not vote is larger than the average age of Americans who voted.

- In this set-up, age is a quantitative data, but voted or not is a qualitative/categorical data.
- Recall–we may use inference about two population means, independent samples, to test if the average age of Americans who did not vote is larger than the average age of Americans who voted. I.e., if $\mu_1 > \mu_2$.
- What if the researcher wants to investigate whether the average age differs across voters of different political parties? In this setting, *different political parties* is a categorical data with more than two categories, so we can no longer use inference about two population means.

Let's first introduce **another set-up** for hypothesis test that tests if the average age of Americans who did not vote is larger than the average age of Americans who voted. We attempt to roughly let "voted or not" be the explanatory variable, and age be the response variable.

In this setting, **categorical independent variable with two levels** is a variable u that either takes a value "voted" or "did not vote".

Consider a dummy variable x which assumes 0 if the random individual voted (i.e. $u=$"voted"), and 1 if not (i.e. $u=$"did not vote"). Then x is a quantitative variable. **u="voted" is called the base level.**

Now, a data point like $(1, 32)$ represents an American who did not vote in the last federal election and is 32 years old. A data point like $(0, 64)$ represents an American who voted in the last federal election and is 64 years old.

Consider a line model $y = \beta_0 + \beta_1 x$ such that β_0 is defined to be the mean for base level, in this example, the average age of Americans who voted, and β_1 is defined to be the mean for level assigned "1" minus mean for base level.

A hypothesis test may be conducted with $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 > 0$.

Only concepts and H_0, H_1 of this kind of hypothesis test are testable. The process of conducting this hypothesis test will not be tested.

More generally, we can consider a categorical random variable u with k levels. Let μ_i, $1 \leq i \leq k$ represents the mean value of the response variable y for level i. Choose level 1 to be the base level. Define dummy variables x_i for $1 \leq i \leq k-1$ such that x_i assumes value 1 if u is at level $i+1$, and x_i assumes value 0 otherwise. This gives us $k-1$ dummy variables $x_1, \ldots, x_{k-1}$.

Consider a linear model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1}$$

such that $\beta_0 = \mu_1$, $\beta_i = \mu_{i+1} - \mu_1$ for $1 \leq i \leq k-1$. Such a relation gives $\mu_j = \beta_0 + \beta_{j-1}$ for $2 \leq j \leq k$.

If we have a sample in which data points are of the form (u, y), we may change to dummy variables to get a sample in which data points are of the form $(x_1, ..., x_{k-1}, y)$, for example $(1, 0, ..., 0, 36)$ or $(0, 1, 0, ..., 3)$.

Note that in such a data point, **at most one** dummy variable x_i can assume 1.

Now we have a quantitative sample data set in which data points are k-tuples. We may consider the lease squares prediction equation (least squares regression line) of this sample, i.e., the line

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_{k-1} x_{k-1}$$

that minimizes the sum of square residuals.

The technologies or math tools that are able to find the lease squares prediction equation in the previous section can be directly applied to this sample. Why?

Theorem

Given a quantitative sample data set in which data points are of the form $(x_1, \ldots, x_{k-1}, y)$ where x_i are dummy variables, the line that minimizes sum of square residuals $\hat{y} = b_0 + b_1 x_1 + \cdots + b_{k-1} x_{k-1}$ (i.e. the least squares regression line) **is** given by:

- b_0 is the sample mean at the base level ($b_0 = \overline{y}_1$)
- b_1 is the sample mean at level 2 minus the sample mean at the base level ($b_1 = \overline{y}_2 - \overline{y}_1$)
- $\cdots$
- b_{k-1} is the sample mean at level k minus the sample mean at the base level ($b_{k-1} = \overline{y}_k - \overline{y}_1$)

This theorem is a mathematical consequence of the lease squares regression line of dummy variables that can only assume 0 or 1. Therefore, we have two equivalent definitions of the least square regression line $\hat{y} = b_0 + b_1 x_1 + \cdots + b_{k-1} x_{k-1}$ of dummy variables.

RMK.

- We are choosing level 1 to be the base level
- For any $i > 0$, b_i is a sample statistic that estimates $\beta_i = \mu_{i+1} - \mu_1$.

Test of hypothesis: H_0 and H_1

We may want to test whether the mean value of y is the same across all levels of the categorical variable u, i.e. $\mu_1 = \mu_2 = \cdots = \mu_k$? This is equivalent to testing

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$$

$$H_1 : \text{at least one of } \beta_i \text{ differs from } 0$$

Example

USGA wants to compare the mean driving distances of four different golf ball brands (A, B, C, and D). A robot golfer hits a sample of 3 balls from each brand to get the following sample.

index	A	B	C	D
1	251.2	263.2	269.7	251.6
2	245.1	262.9	263.2	248.6
3	248.0	265.0	277.5	249.4

Then our categorical variable u can assume A, B, C, D. We choose level A to be the base level and define dummy variables x_1, x_2, x_3 by the following rules.

$$x_1 = \begin{cases} 1, & \text{if } u = B \\ 0, & \text{if not} \end{cases} \qquad x_2 = \begin{cases} 1, & \text{if } u = C \\ 0, & \text{if not} \end{cases} \qquad x_3 = \begin{cases} 1, & \text{if } u = D \\ 0, & \text{if not} \end{cases}$$

We propose a hypothetical model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

in which β_0 represents the mean driving distance for golf balls of Brand A (i.e. $\beta_0 = \mu_A$); $\beta_1 = \mu_B - \mu_A$, $\beta_1 = \mu_C - \mu_A$, $\beta_3 = \mu_D - \mu_A$.

We may again use technologies or more advanced math tools to find b_0, b_1, b_2, b_3 from our sample using least squares regression. This process is called **fitting sample data points to a (proposed, hypothetical) model**.

However, by our theorem, we can find the least squares regression line $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$ of this sample by hand.

index	A	B	C	D
1	251.2	263.2	269.7	251.6
2	245.1	262.9	263.2	248.6
3	248.0	265.0	277.5	249.4

Sample point examples

(A,251.2)→ (0,0,0,251.2)
(C,263.2) → (0,1,0,263.2)

We calculate $\overline{y}_A = 248.1$, $\overline{y}_B = 263.7$, $\overline{y}_C = 270.13$, $\overline{y}_D = 249.87$.
Then $b_0 = \overline{y}_A = 248.1$.
$b_1 = \overline{y}_B - \overline{y}_A = 15.6$, $b_2 = \overline{y}_C - \overline{y}_A = 22.03$, $b_3 = \overline{y}_D - \overline{y}_A = 1.77$.

The least squares regression line is

$$\hat{y} = 248.1 + 15.6x_1 + 22.03x_2 + 1.77x_3.$$

Question

Suppose USGA wants to compare the mean driving distances of four different golf ball brands (A, B, C, and D). A robot golfer hits a sample of **30** balls from each brand to get a sample of size $30 \times 4 = 120$. A statistician then finds a line $\hat{y} = 249 - 4x_1 + 12x_2 + 2x_3$ that minimizes sum of square residuals for this sample.

1. What is the sample mean of driving distances of the 30 Brand A balls?
2. What is the sample mean of driving distances of the 30 Brand C balls?

Reference

The following examples are from *Statistics*, by McClave and Sincich, 13th edition, page 658, 700.

A collector of antique grandfather clocks sold at auction believes that the price received for the clocks depends linearly on both the age of the clocks and the number of bidders at the auction. Thus, he hypothesizes a model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ where y=Auction price (dollars), x_1=Age of clock (years), x_2=Number of bidders.

Example

USGA wants to compare the mean driving distances of four different golf ball brands (A, B, C, and D). A robot golfer hits a sample of 10 balls from each brand to get the following sample.