# Unit 30: Inference for Regression

## PREREQUISITES

Students should be familiar with the topic of least-squares regression lines, which was covered in Unit 11, Fitting Lines to Data. Students need some background in significance tests, confidence intervals, and the *t*-distributions. Coverage of the material in Unit 26, Small Sample Inference for One Mean, is a prerequisite for this unit. In addition, students must have some familiarity with material on normal distributions covered in Units 8 and 9, Normal Calculations and Checking Assumptions of Normality, respectively.

## ADDITIONAL TOPIC COVERAGE

Additional coverage of inference for simple linear regression can be found in *The Basic Practice of Statistics*, Chapter 24, Inference for Regression. To extend the topic, see Chapter 28, Multiple Regression (in Optional Companion Chapters).

## ACTIVITY DESCRIPTION

In the activity, students use clues left by a thief – his/her step length and forearm length – to estimate the height of the thief. But first students must build models, one to estimate height from forearm length and the other to estimate height from step length. Students now have two competing models to predict the thief's height. Their choice will depend on each model's standard error of the estimate, $s_e$ .

## MATERIALS

Rulers and/meter stick (optional, if you plan to collect your own data).

In this activity, data on height, step length, and forearm length from 9$^{th}$ and 10$^{th}$ grade students are provided. However, your class can collect data of their own and either add it to the data

provided in the activity or substitute it for the data provided. If your students are older – college students or 11th- or 12th-grade students – your data might have a different pattern than the data contained in the activity (especially for the male students).

Particularly if you add data, you could extend this project by creating separate models for each gender. The footprints appeared to be from male sneakers (but sometimes females wear male sneakers). So using a model developed from the male student data could have a smaller $s_e$ than a model using all students (both male and female).

# THE VIDEO SOLUTIONS

1. Peregrine falcons were not able to hatch their eggs due to eggshell thinning. Scientists believed the cause of eggshell thinning was due to DDT or its derivative DDE.

2. The log-concentration of DDE was the explanatory variable and eggshell thickness was the response variable.

3. The pattern of the data showed a negative, linear relationship.

4. A line fit to data from the entire population.

5. Because they depend on the sample data and can vary from sample to sample.

6. $H_o$ : Amount DDE and eggshell thickness have no linear relationship (or $\beta = 0$).

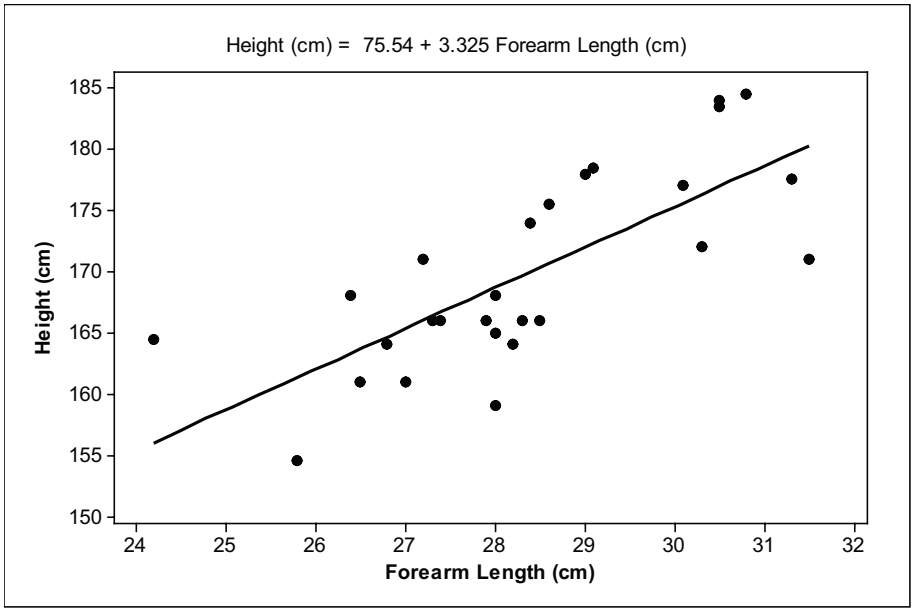
$H_a$ : Amount DDE and eggshell thickness have a negative linear relationship (or $\beta < 0$).

7. The null hypothesis was rejected.

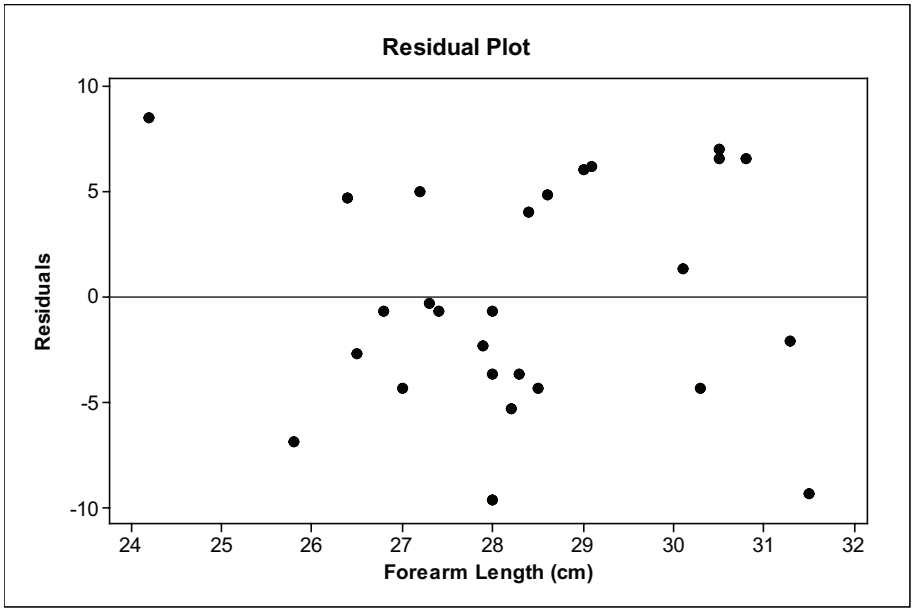
8. Yes, their populations have increased since DDT was banned in the U.S. and Western European countries.

# UNIT ACTIVITY SOLUTIONS

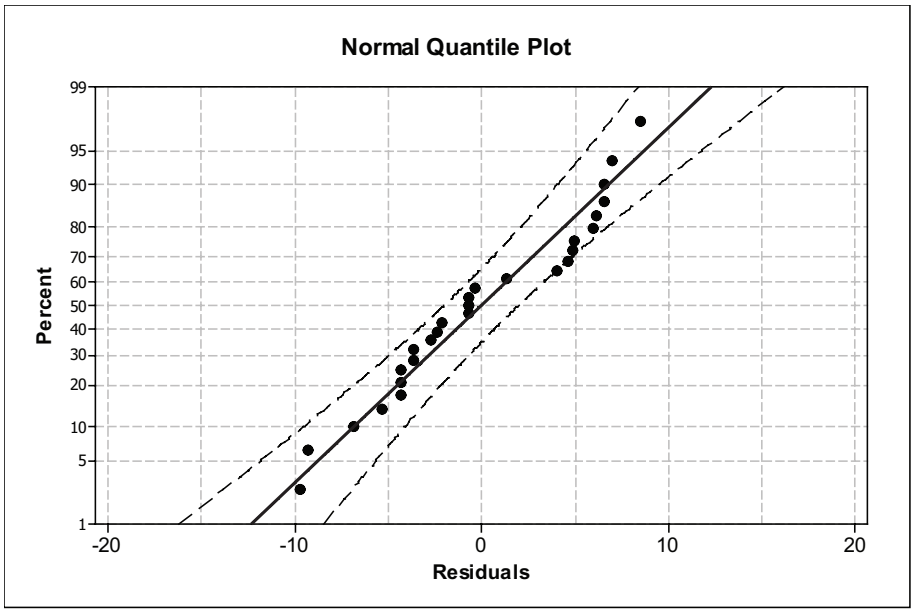
1. a. Equation of least-squares line: Height = 75.53 + 3.325 Forearm Length.




b. The dots in the residual plot below appear randomly scattered with no strong patterns. The vertical spread of the dots appears to stay the same as *x* increases. So, Conditions 1 and 4 appear to be reasonably satisfied.




The normal quantile plot of the residuals shown below is fairly linear. The dots stay pretty much within the curved bands that Minitab adds to the plot. So, there does not appear to be a strong departure from normality. There is no evidence of an extreme outlier.




Finally, these data were a random sample of 9th- and 10th-grade students. So, the heights for fixed values of the explanatory variables are independent of each other.

In conclusion, all four conditions for inference are reasonably met.

c. $s_e \approx 5.386$ cm

2. a. First, we need to compute the standard error of the slope, $s_b = \dfrac{s_e}{\sqrt{\sum (x - \bar{x})^2}}$.

We know the numerator but need to determine the denominator.

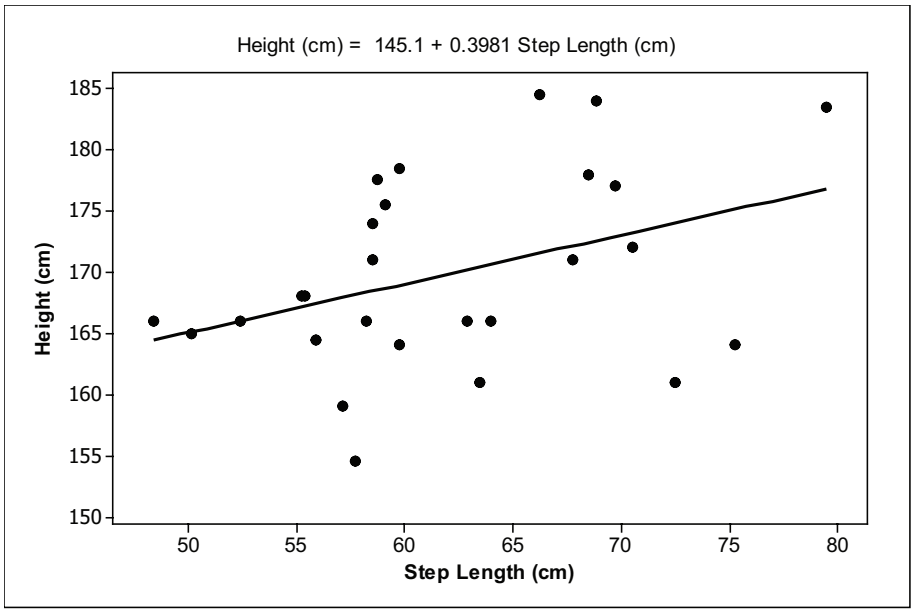
We used Excel: $s_b = \dfrac{5.386}{\sqrt{80.9067}} \approx 0.599$.

$t = \dfrac{3.325 - 0}{0.599} \approx 5.55$; $df$ = 25; $p$ = 0.000.

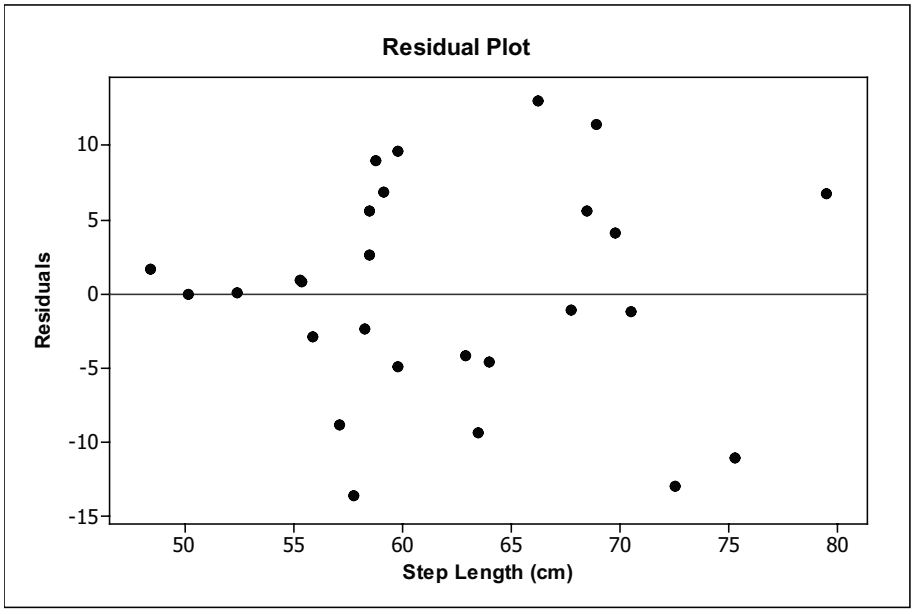
Reject the null hypothesis. Conclude $\beta > 0$; in other words, there is a positive linear relationship between height and forearm length.

b. We use a *t*-table to find *t** = 2.060. Now, we are ready to calculate a 95% confidence interval for $\beta$: $3.325 \pm (2.060)(0.599) \approx 3.325 \pm 1.234$, or from 2.091 to 4.559. Therefore, for each 1 centimeter increase in forearm length, we would expect an increase in height of between 2.09 cm and 4.56 cm.

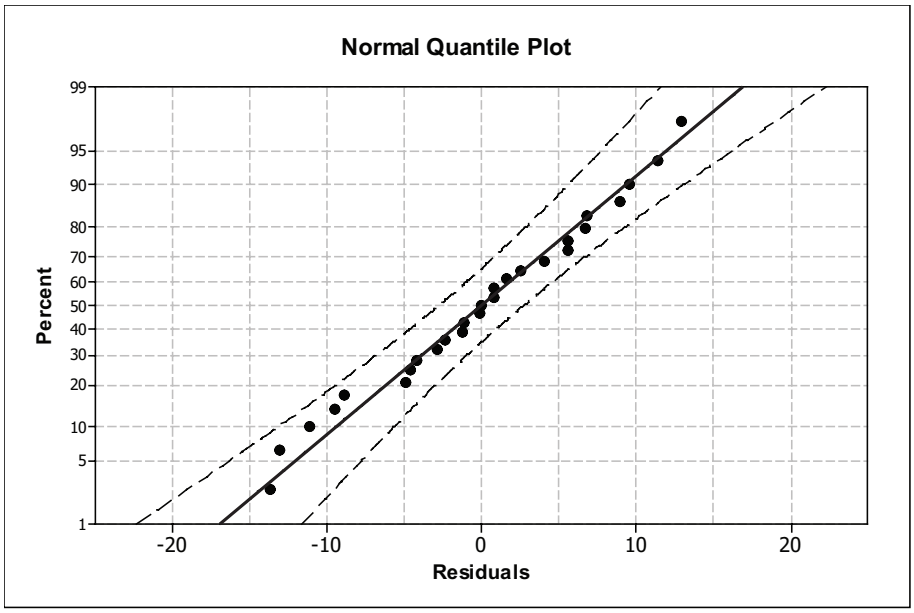
3.a. Equation of least-squares line: Height = 145.1 + 0.3981 Step Length.




b. The dots in the residual plot below appear randomly scattered with no strong patterns. The vertical spread of the dots appears to stay the same as *x* increases with one exception. The vertical spread of the first three dots is very small – indicating less variability in height for people with the smallest step lengths. However, the vertical spread in the rest of the plot looks good. So, Conditions 1 and 4 appear to be reasonably satisfied.




The normal quantile plot of the residuals shown below is fairly linear. It is a reasonable assumption that the residuals follow an approximately normal distribution.




Finally, these data were a random sample of 9th- and 10th-grade students. So, the heights for fixed values of the explanatory variables are independent of each other.

In conclusion, all four conditions for inference are reasonably met.

c. $s_e \approx 7.424$ cm

4. a. First, we need to compute the standard error of the slope, $s_b = \dfrac{s_e}{\sqrt{\sum (x - \bar{x})^2}}$.

We know the numerator but need to determine the denominator.

We used Excel to calculate $s_b$: $s_b = \dfrac{7.424}{\sqrt{1527.64}} \approx 0.1899$

$t = \dfrac{0.3981 - 0}{0.1899} \approx 2.096$; $df$ = 25; $p$ = 0.02318. Reject the null hypothesis.

Conclude that there is a positive linear relationship between height and step length; in other words, $\beta > 0$.

b. We use a *t*-table to find *t** = 2.060. Now, we are ready to calculate the 95% confidence interval for $\beta$: $0.3981 \pm (2.060)(0.1899) \approx 0.398 \pm 0.391$, or from 0.007 to 0.789. Therefore, for each 1 centimeter increase in step length, we would expect an increase in height of between 0.007 cm and 0.789 cm.

5. a. The model based on the explanatory variable forearm length will produce more precise estimates. The standard error of the estimate for the forearm length model is $s_e \approx 5.386$ compared to $s_e \approx 7.424$ for the model based on step length.

b. Students may decide to answer this question in several different ways.

Sample answer: Because the least-squares line based on forearm length is associated with a smaller value for $s_e$, we decided to use that linear model for our prediction. For the point estimate we used the forearm length of 26.5 cm, midway between 26 cm and 27 cm and for the smallest and largest estimate, we used the forearm lengths of 26 cm and 27 cm. These three point estimates allowed us to complete the sentence as follows:

We predict that the thief is 163.7 cm tall. But the thief might be as short as 162.0 cm or as tall as 165.3. (From about 5'3¾'' to 5'5''.)

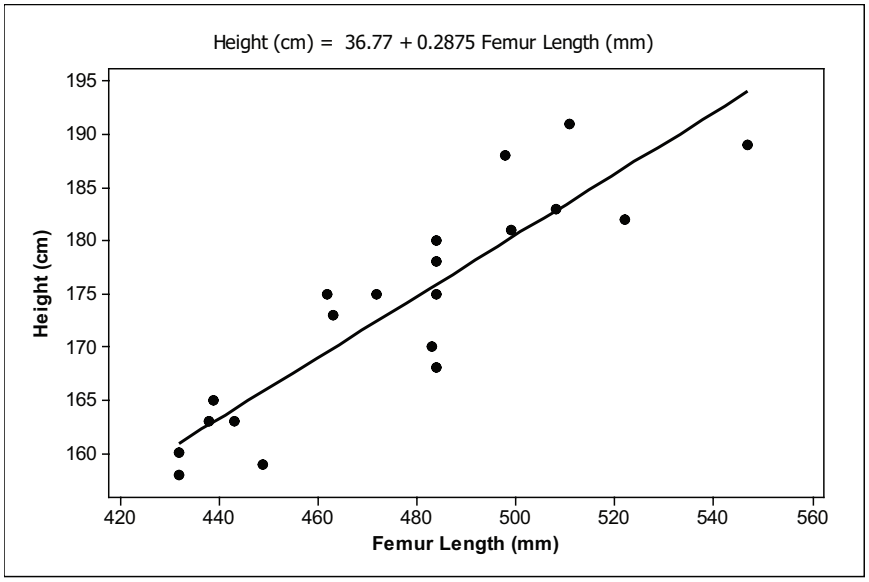
However, in our statement above, we did not use the fact that $s_e \approx 5.386$ (we only used it to select the model). If we want to give more conservative bounds, we would subtract $s_e$ from our lower bound and add it to our upper bound to give from between 156.7 cm to 170.7 cm. (From 5'1⅝'' to 5'6''.) However, if we want to give very conservative bounds, we could subtract $2s_e$ from our lower bound and add it to our upper bound, giving 151.9 cm to 176.1 cm. However, that bound is probably too wide to be of much use in finding the thief. (From around 5' to 5'9''.)

(Another possible answer might be to give 163.7 cm for the point estimate and 163.7 cm ±5.386 cm or 163.7 cm ±10.772 cm for the bounds.)

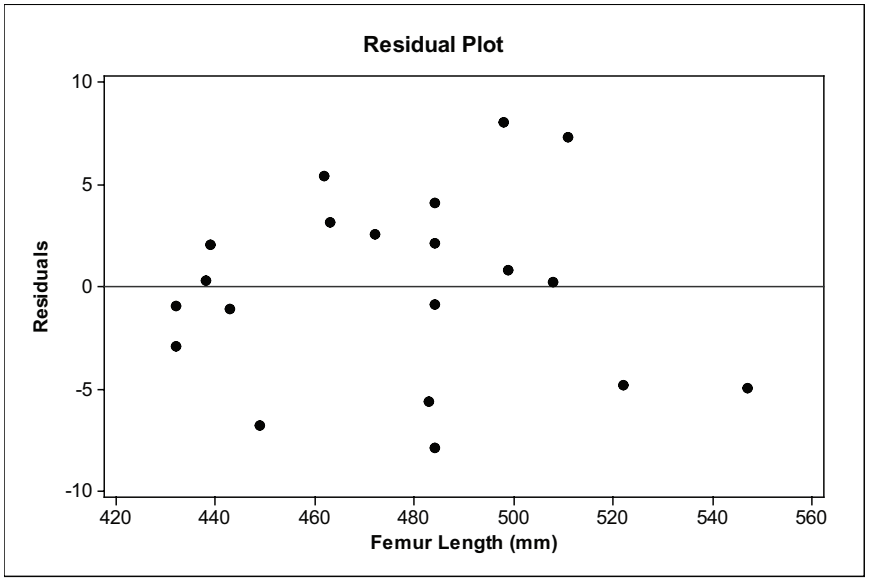
# EXERCISE SOLUTIONS

1. a. The pattern appears to be linear and the association to be positive. (See solution to 1(b) for the scatterplot.)

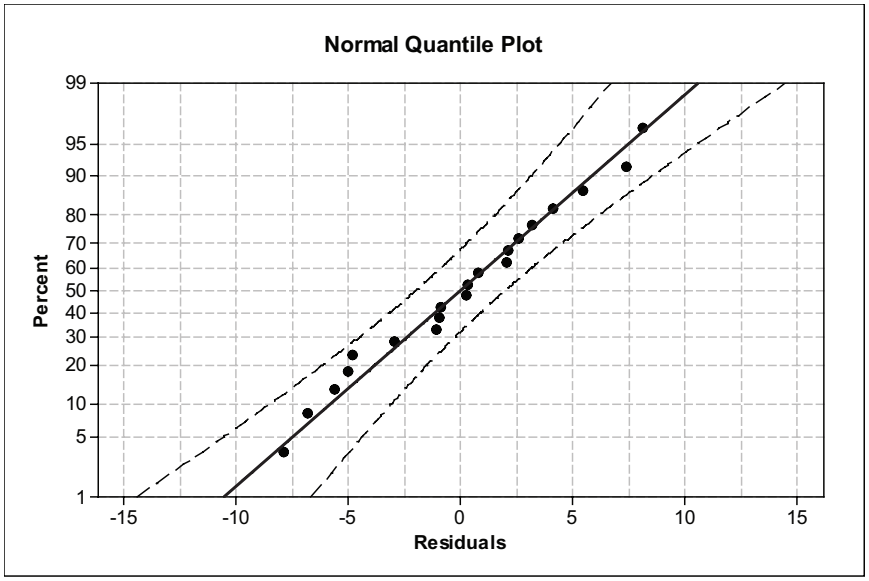
b. Equation of the least-squares line: $y = 36.77 + 0.2875x$




c. Sample answer: The dots in the residual plot appear to be randomly scattered with a good split of points above and below the horizontal axis. There is some evidence that the vertical spread of the residuals is smaller for small femur lengths than for larger femur lengths. However, that difference seems relatively small and could be due to the small sample size. So, conditions 1 and 4 seem to be reasonably satisfied.



A normal quantile plot of the residuals is fairly linear. So, it is reasonable to assume the residuals follow a normal distribution. Condition 2 appears to be satisfied.




Finally, these data are a random sample. So, the heights are independent of each other. We conclude that Conditions 1 – 4 are reasonably satisfied.

2. a. The residuals are listed below:

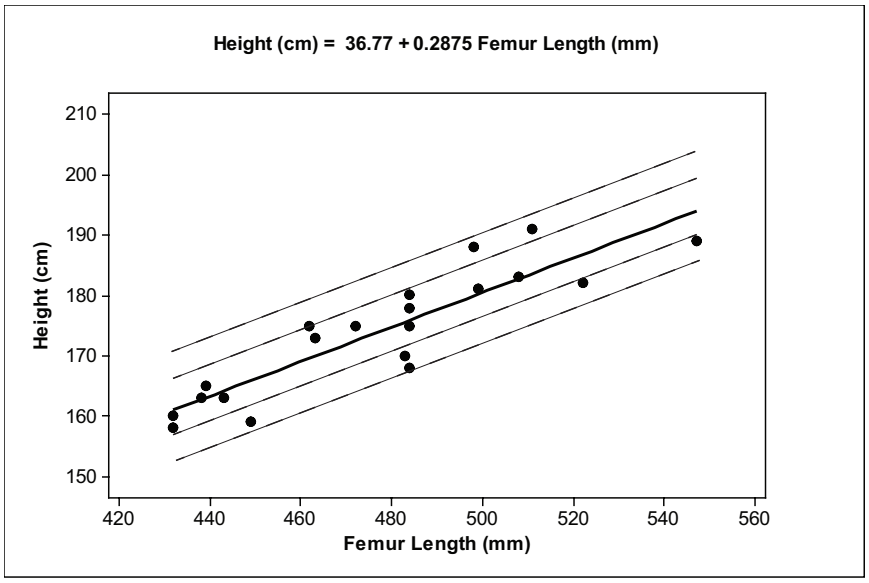
| | | | | |
|---|---|---|---|---|
| -2.95031 | 8.07699 | 3.13827 | -1.11242 | 7.33994 |
| -5.00880 | 2.10150 | -4.82217 | 0.32490 | 5.42574 |
| -6.83721 | 0.78953 | -7.89850 | 2.55109 | -0.89850 |
| -0.95031 | 2.03744 | -5.61103 | 4.10150 | 0.20234 |

To find the SSE, we square the residuals and then find the sum:

$$\text{SSE} = (-2.9531)^2 + \ldots + (0.20234)^2 \approx 391.679$$

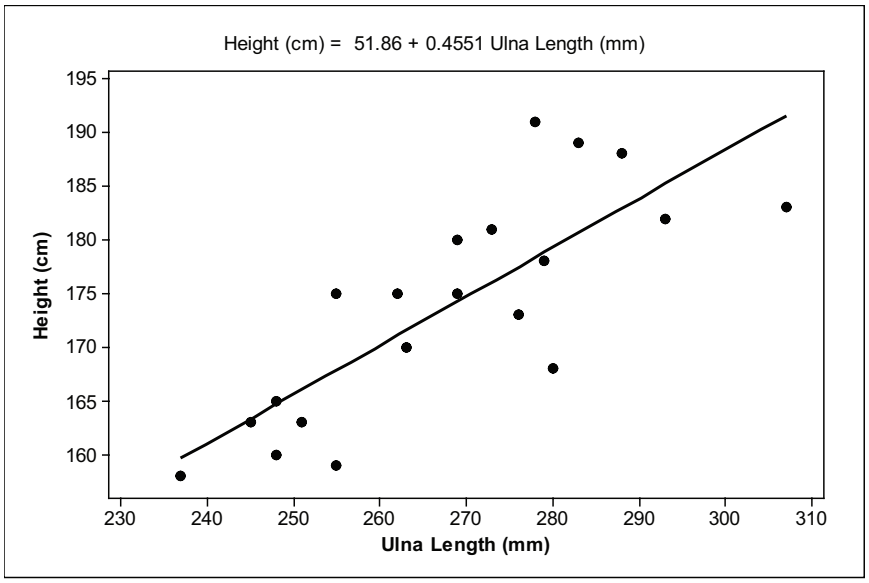
$$s_e = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{391.679}{20-2}} \approx 4.665 \text{ cm}$$

b. $y = 36.77 + 0.2875x \pm 4.665$; $y = 36.77 + 0.2875x \pm 9.330$




c. In this case, all 20 data points are trapped between the outermost error bands.

3. a. Equation of least-squares line: $y = 51.86 + 0.4551x$




b. The residuals are given below:

| | | | | |
|---|---|---|---|---|
| -1.7150 | 5.0755 | -4.4635 | -0.3557 | 12.6264 |
| 8.3509 | -0.8287 | -3.2000 | -3.0863 | 3.9078 |
| -8.9066 | 4.9018 | -11.2838 | 7.0934 | 0.7222 |
| -4.7210 | 0.2790 | -1.5473 | 5.7222 | -8.5712 |

SSE = 726.122

$$s_e = \sqrt{\frac{726.112}{20-2}} \approx 6.351 \text{ cm}$$

c. Height prediction from ulna length: $y = 51.86 + 0.4551(287) \approx 182.47$ cm.
Height prediction from femur length: $y = 36.77 + 0.2875(520) \approx 186.27$ cm.

The prediction based on femur length is likely to be more reliable. The $s_e$ for the model based on femur length is 4.665 cm compared to 6.351 cm for the model based on ulna length.

4. a. $t = 8.63$; $df = 18$; $p \approx 0$; Reject the null hypothesis and conclude that there is a positive linear relationship between height and femur length.

b. First, we need to calculate the standard error of the slope, $s_b$: $s_b = \frac{s_e}{\sqrt{\sum (x-\bar{x})^2}}$

We use Excel to calculate $\sum (x-\bar{x})^2 = \sum (x-476.70)^2 \approx 19598.2$

$$s_b = \frac{4.665}{\sqrt{19598.2}} \approx 0.0333$$

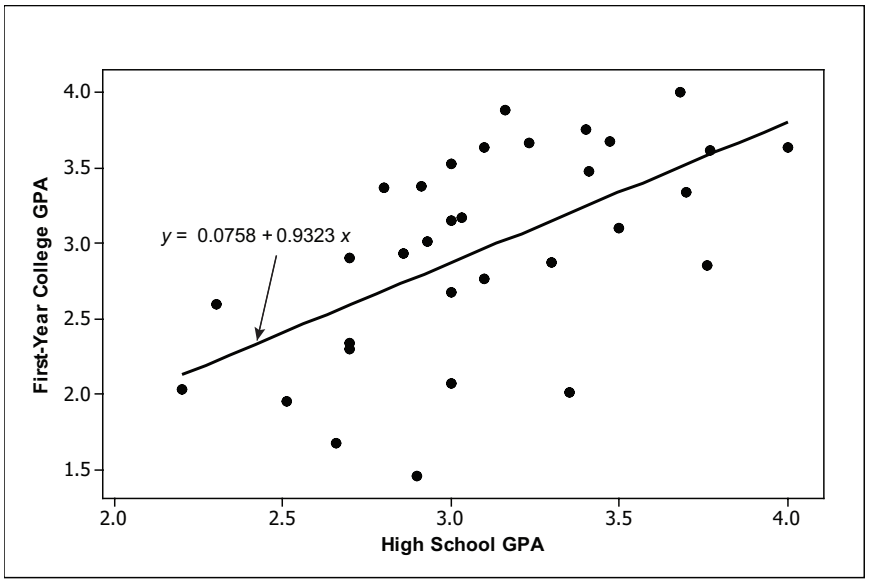
Next, we use a *t*-table to find the *t*-critical value for a 95% confidence interval: $t^* = 2.101$.

The 95% confidence interval for $\beta$: $0.2875 \pm (2.101)(0.0333) \approx 0.2875 \pm 0.0700$,
or from 0.2175 to 0.3575.

# REVIEW QUESTIONS SOLUTIONS

1. From Figure 30.14, we note that the dots appear randomly scattered, roughly half above the horizontal axis and roughly half below. This indicates that Condition 1, linearity, is satisfied. In addition, the vertical spread of the dots on this plot appears roughly the same as x increases. Therefore, Condition 4, equal standard deviations, is satisfied. A plot of the residuals is somewhat linear – at least the dots stay within the curved bands provided by Minitab. So, it is reasonable to assume the residuals follow at least an approximate normal distribution. Finally, the eggs were from a random sample, so the egg thicknesses were independent of each other. Conditions 1 – 4 are reasonably satisfied and therefore, the results of the inference shown in the video are trustworthy.

2. a. The form appears linear. The relationship is positive.

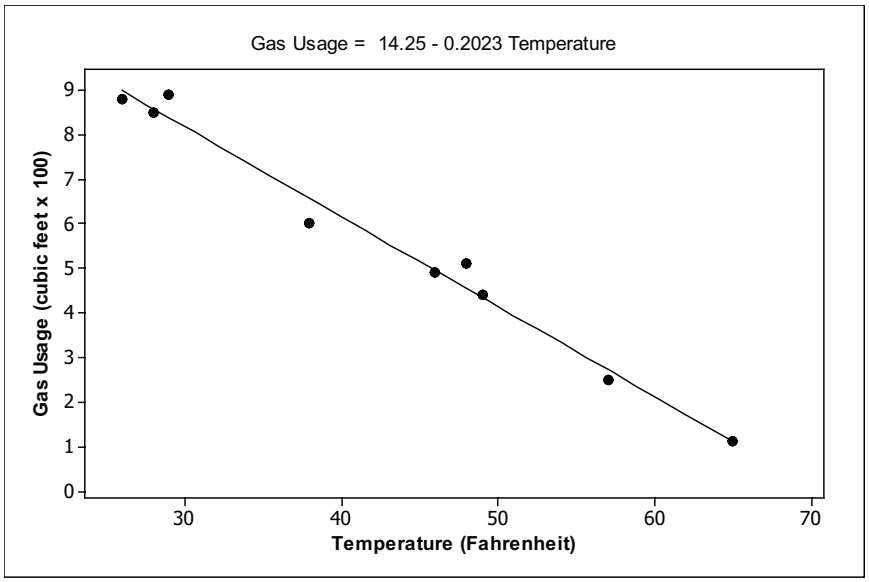



b. See (a) for scatterplot and line: $y = 0.0758 + 0.9323x$

c. $t = \dfrac{0.9323 - 0}{0.2370} \approx 3.93$; $df = 32\text{-}2 = 30$

d. $p \approx 0.0002$; reject the null hypothesis and conclude that there is a positive linear relationship between students' high school GPAs and their first-year college GPAs.

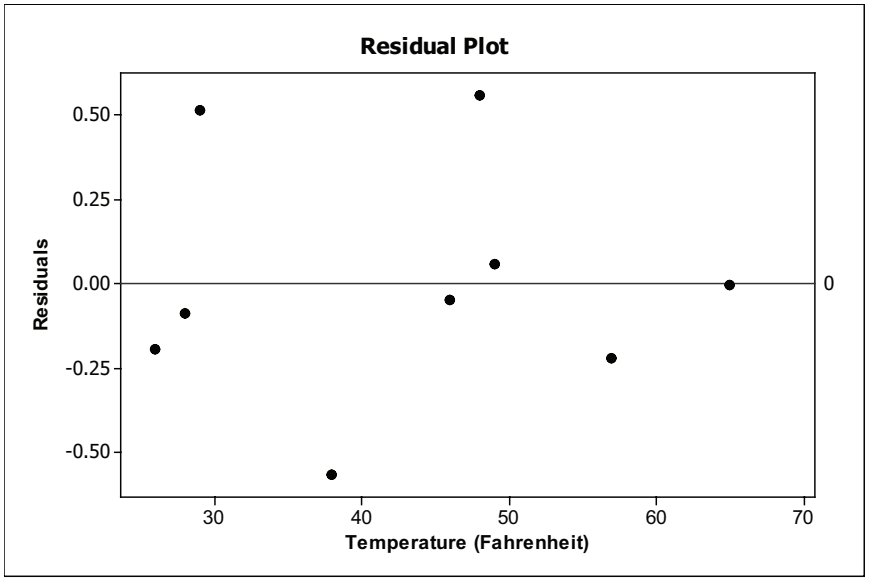
3. a. The relationship appears to be a negative linear relationship.



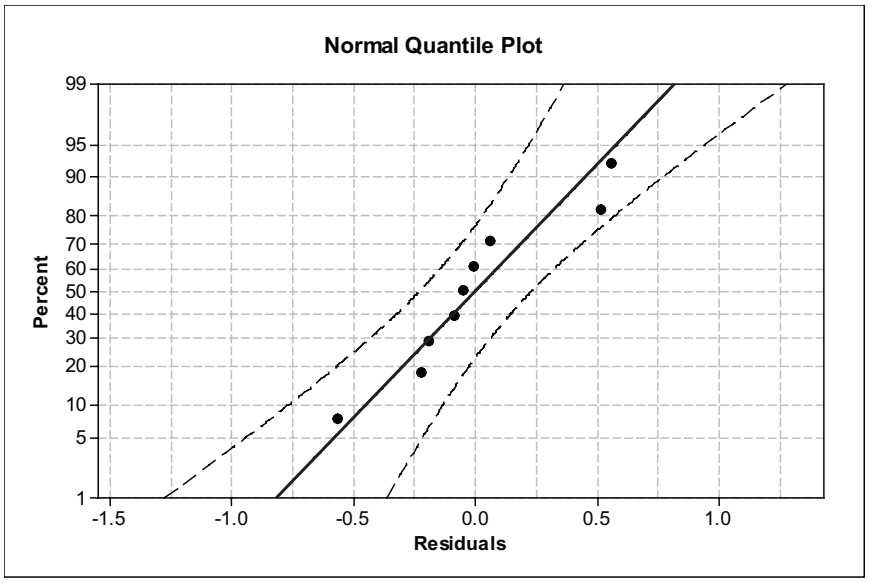

b. See scatterplot in solution to 2(a).

The equation of the least-squares line is $y = 14.25 – 0.2023x$.

c. The dots in the residual plot below appear to be randomly scattered with some above and below the horizontal axis. So, the line is adequate to describe the pattern in the data and Condition 1 is satisfied. In addition, the vertical spread of the dots stays roughly the same as temperature increases. So, Condition 4 is satisfied.




The pattern of the normal probability plot below is roughly linear. Given that the dots stay within the curved bands provided by Minitab, we can say that there are no strong departures from normality in the residuals. So, Condition 4 is reasonably satisfied.




Observed gas usage for different months should depend only on temperature and be independent of each other.

So, given the conditions are satisfied, and we can proceed with inference.

d. $s_e = \sqrt{\frac{\text{SSE}}{n-2}}$

We begin our calculations with the residuals, which were calculated using technology:

| | | | | |
|---|---|---|---|---|
| 0.555956 | -0.048548 | -0.566567 | 0.513162 | -0.193595 |
| -0.089090 | 0.058209 | -0.223773 | -0.005754 | |

$$\text{SSE} = (0.555956)^2 + \ldots + (-0.005754)^2 = 0.994689$$

$$s_e = \sqrt{\frac{0.994689}{9-2}} \approx 0.37696$$

$$s_b = \frac{s_e}{\sqrt{\sum (x-\bar{x})^2}};\ \bar{x} = 42.89;$$

the deviations of the temperatures, x, from their mean is given below:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 5.11 | 3.11 | -4.89 | -13.89 | -16.89 | -14.89 | 6.11 | 14.11 | 22.11 |

$$\sum (x-\bar{x})^2 = (5.11)^2 + \ldots + (22.11)^2 = 1484.89$$

$$s_b = \frac{0.37696}{\sqrt{1484.89}} \approx 0.00978$$

e. Sample answer: We would expect to need more gas as the temperature gets colder. So, it makes sense to test $H_a : \beta < 0$; in other words, the alternative is for a negative linear relationship.

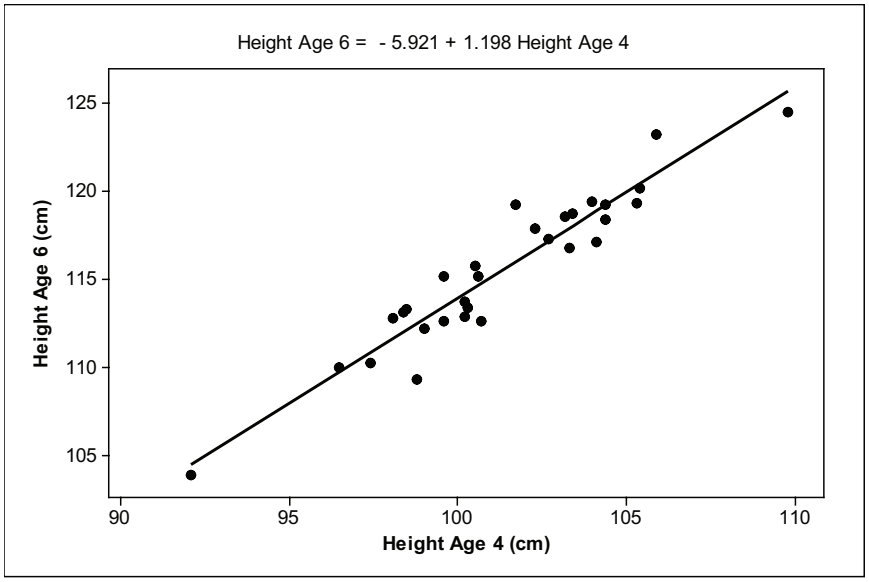
$t = \frac{-0.2023 - 0}{0.00978} \approx -20.68$; $df$ = 7; p ≈ 0.

Reject the null hypothesis and conclude that there is a negative linear relationship between temperature and gas usage.

f. We use the *t*-table to find *t**, the *t*-critical value from a *t*-distribution with *df* = 7. We find *t** = 2.365. Now we have everything that we need to construct a 95% confidence interval for $\beta$: -0.2023 ± (2.365)(0.00978) ≈ -0.202 ± 0.023, or from -0.225 to -0.179.

Interpretation: For each 1°F increase in temperature, the average daily gas usage decreases by between 0.179 cubic feet × 100 and 0.225 cubic feet × 100.

4. a. The equation of the least-squares line is *y* = -5921 + 1.198*x*. The scatterplot and a graph of the least-squares line appear below.




b. First, we use a *t*-table, *df* = 28, to determine *t** = 2.048

95% confidence interval for $\beta$: 1.198 ± (2.048)(0.07437) ≈ 1.198 ± 0.152, or from 1.046 to 1.350.

Interpretation in context: For each 1 cm increase in height at age 4, we expect an increase of between 1.046 cm and 1.350 cm in height at age 6.