Chapter 23 – Inferences for Regression

Section 23.1

1. Graduate earnings.

The linear regression model that predicts earnings from SAT score is $\widehat{Earn} = 14,468.1 + 27.262(SAT)$. This model predicts that graduates earn, on average, $27.26 per year more for each additional point in SAT score.

2. Shoot to score 2016.

The linear regression model that predicts the number of goals from the number of shots is $\widehat{Goals} = 1.13495 + 0.099267(Shots)$. This model predicts that the number of goals made increases, on average, by 0.0993 for each additional shot taken.

Section 23.2

3. Earnings II.

The residual plot has no structure, and there are not any striking outliers. The histogram of residuals is symmetric and bell-shaped. All conditions are satisfied to proceed with the regression analysis.

4. Shoot to score II.

The scatterplot shows a linear relationship between *Shots* and *Goals*. The residual plot looks scattered, but there may be some increase in variation of Goals as Shots increases. We will proceed with caution. The normal probability plot looks straight. All conditions seem satisfied to go ahead with the regression analysis.

Section 23.3

5. Earnings, part III.

The error standard deviation is s = $5603. This is the standard deviation of the residuals, which helps us to understand the amount of spread about the regression model.

6. Shoot to score, hat trick.

The error standard deviation is s = 2.983. This number indicates the amount of variation in the data points about the linear regression model. The standard deviation of the residuals is about 2.98 goals.

7. Earnings, part IV.

The standard error for the slope tells us how much the slope of the regression equation would vary from sample to sample. If we took many samples, and found a regression equation for each, we would expect most of these slopes to be within a couple of standard deviations of the true slope of the association between graduation rate and acceptance rate. The P-value is essentially zero, which confirms that the slope is statistically significantly different than zero.

8. Shoot to score, another one.

The standard error for the slope tells us how much the slope of the regression equation would vary from sample to sample. If we took many samples, and found a regression equation for each, we would expect most of these slopes to be within a couple of standard deviations of the true slope of the association between number of shots and number of goals. This number is very small, which means we should be able to make precise predictions. The P-value is essentially zero, which confirms that the slope is statistically significantly different than zero.

Section 23.4

9. Earnings, part V.

Since the P-value is so low, we reject the null hypothesis. We can conclude that there is evidence of a linear relationship between *Earn* and *SAT*. In other words, the slope of the relationship is not zero. It seems that the those who score higher on their SAT tend to earn more.

10. Shoot to score, number five.

Since the P-value is low, we reject the null hypothesis. The coach can conclude there is evidence of a linear relationship between shooting and scoring. In other words, as a player takes more *Shots*, he should expect to score more *Goals*, on average.

11. Earnings VI.

$b_1 \pm t^*_{n-2} \times SE(b_1) = 27.2642 \pm (t^*_{704}) \times 1.545 \approx (24.23, 30.30)$; We are 95% confident that the true slope relating *Earn* and *SAT* is between 24.23 and 30.30 dollars per year per SAT point. In other words, for each additional point of SAT score, the model predicts an increase in earnings of between 24.23 and 30.30 dollars per year.

12. Shoot to score, overtime.

$b_1 \pm t^*_{n-2} \times SE(b_1) = 0.099267 \pm (t^*_{63}) \times 0.0125 \approx (0.074, 0.124)$; With 95% confidence, the coach can conclude that, the true slope of the relationship between *Goals* and *Shots* is between 0.074 and 0.124 goals per shot on goal. In other words, on average, players who take an additional *Shot* can expect to score between 0.074 to 0.124 *Goals* more.

Section 23.5

13. Earnings and need.

a) $\widehat{Earn} = 23{,}974.2 + 23.1880(SAT) - 8500.75(\%need)$

b) The model predicts that, on average, *Earn* is expected to increase by approximately $23.18 per year for each additional SAT point, after allowing for the effects of *%need.* This differs from the previous interpretation by taking the variable, *%need* into account.

14. Shoot to score, time on ice.

a) $\widehat{Goals} = -3.90580 + 0.070019(Shots) + 0.458088(TOI\,/\,G)$

b) The model predicts that, on average, *Goals* is expected to increase by approximately 0.07 goals per game for each additional shot taken, after allowing for the effects of *TimeOnIce/Game.* This differs from the previous interpretation by taking the variable, *TOI/G* into account.

c) $b_1 \pm t^*_{n-(k+1)} \times SE(b_1) = 0.458088 \pm (t^*_{62}) \times 0.2037 \approx (0.051, 0.865)$

We are 95% confident that coefficient of *TimeOnIce/Game* is between 0.051 and 0.865.

15. Earnings and more.

a) *ACT* and *SAT* are highly correlated with each other. After all, they are very similar measures. Thus *SAT*, after allowing for the effect of *ACT*, is not really a measure of test performance but rather a measure of how students who take the *SAT* may differ from those who take the *ACT* at the colleges in question.

b) $b_1 \pm t^*_{n-(k+1)} \times SE(b_1) = 10.1117 \pm (t^*_{683}) \times 4.336 \approx (1.56, 18.66)$

We are 95% confident that the coefficient of *SAT* in the linear relationship predicting *Earn* from *SAT, %need,* and *ACT* is between 1.56 and 18.66. This interval contains much smaller values as plausible values.

c) We are less confident that this coefficient is different than zero. The collinearity with *ACT* has inflated the variance of the coefficient of *SAT*, leading to a smaller t-ratio and larger P-value.

16. Shoot to score, double overtime.

a) *Shoot%* is collinear with *TimeOnIce/Game*. Perhaps players with high shooting percentages get to spend more time on the ice. The result is that, after accounting for the effects of *Shoot%*, there is little left for *TOI/G* to account for.

16. (continued)

b) $b_1 \pm t^*_{n-(k+1)} \times SE(b_1) = -0.040008 \pm (t^*_{61}) \times 0.0656 \approx (-0.171,\ 0.091)$

We are 95% confident that the coefficient of *TOI/G* in the linear relationship predicting *Goals* from *Shots, TOI/G,* and *Shoot%* is between -0.171 and 0.091. This interval now contains zero.

c) Because *Shoot%* and *TOI/G* are collinear, including *Shoot%* in the model leaves little for *TOI/G* to account for. The collinearity inflates the variance of the coefficient of *TOI/G*. The corresponding P-value is now not significant.

Section 23.6

17. Earnings, predictions.

A prediction interval for an individual SAT score will be too wide to be of much use in predicting future earnings.

18. Shoot to score, predictions.

It is difficult to make accurate individual predictions from a regression. The prediction interval is likely to be too wide to be useful for the coach.

Chapter Exercises.

19. Earnings, planning.

No, regression models describe the data as they are. They cannot predict what would happen if a change were made. We cannot conclude that earning a higher SAT score will lead to higher earnings; that would suppose a causal relationship that clearly isn't true.

20. Shoot to score, shootout.

No, regression models describe the data. They do not tell what would happen if values were changed. We cannot conclude from the regression that a player would score more simply by taking more shots; that would suppose a causal relationship that doesn't consider other variables. (For example, simply spraying shots near the goal is likely to lose the puck and be counter-productive.)

21. Tracking hurricanes 2016.

a) The equation of the line of best fit for these data points is $\widehat{Error\_24th} = 133.024 - 2.05999(Year)$, where *Year* is measured in years since 1970. According to the linear model, the error made in predicting a hurricane's path was about 133 nautical miles, on average, in 1970. It has been declining at rate of about 2.06 nautical miles per year.

b) H_0: There has been no change in prediction accuracy. $(\beta_1 = 0)$

H_A: There has been a change in prediction accuracy. $(\beta_1 \neq 0)$

c) Assuming the conditions have been met, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with 46 – 2 = 44 degrees of freedom. We will use a regression slope *t*-test.

The value of $t = -11.9$. The P-value $\leq$ 0.0001 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence that the prediction accuracies have in fact been changing during the time period.

d) 76.2% of the variation in the prediction accuracy is accounted for by the linear model based on year.

22. Drug use 2013.

a) The equation of the line of best fit for these data points is $\widehat{\%Cocaine} = 0.297 + 0.124(\%Cannabis)$.

According to the linear model, the average percentage of people in these countries who use cocaine increases by about 0.124 percentage points for each additional 1% of people who use cannabis.

b) H_0: There is no linear relationship between marijuana use and use of other drugs. $(\beta_1 = 0)$

H_A: There is a linear relationship between marijuana use and use of other drugs. $(\beta_1 \neq 0)$

c) Assuming the conditions have been met, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (32 – 2) = 30 degrees of freedom. We will use a regression slope *t*-test.

The value of $t = 3.23$. The P-value of 0.003 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence that the percentage of people who use cocaine is related to the percentage of people who use cannabis. Countries with a high percentage of people using cannabis tend to have a high percentage of people using cocaine.

d) 25.8% of the variation in the percentage of people using cocaine can be accounted for by the percentage of people using cannabis.

e) The use of cocaine is associated with cannabis use, but there is no proof of a cause-and-effect relationship between the two variables. There may be lurking variables present.

23. Sea ice.

a) $\widehat{Extent} = 73.7928 - 4.42138(MeanGlobalTemp)$; According to the model, we would expect 73.79 square kilometers of sea ice in the northern Artic if the mean global temperature were 0 degrees Celsius. The model predicts a decrease of 4.42 square kilometers of sea ice in the northern Arctic for each additional degree in mean global temperature.

b) **Straight enough condition:** The scatterplot shows a moderate, linear relationship with no striking outliers, but the residuals plot shows a possible bend.
Does the plot thicken? condition: The residuals plot shows slightly greater variability on the left than on the right.
Nearly Normal condition: The Normal probability plot is reasonably straight.

We should proceed with caution because the conditions are almost satisfied.

c) The standard deviation of the residuals is 0.68295.

d) The standard error of the slope of the regression line is 0.5806.

e) The standard error is the estimated standard deviation of the sampling distribution of the slope coefficient. Over many random samples from this population (or with a bootstrap), we'd expect to see slopes of the samples varying by this much.

f) No. There is evidence of an association between the mean global temperature and the extent of sea ice in the northern Artic, but we cannot establish causation from this study.

24. Saratoga house prices.

a) $\widehat{Price} = -3.12 + 94.5(Size)$; The model suggests that the prices of Saratoga homes increase by about \$94.5 for each additional square foot.

b) A negative intercept makes no sense, but the P-value of 0.51 indicates that we can't discern a difference between our estimated value and zero. It makes sense that a house of zero square feet should cost \$0.

c) Amounts by which house prices differ from predictions made by this model vary, with a standard deviation of about \$54,000 per thousand square feet.

d) The standard error of the slope is \$2.395 per square foot.

24. (continued)

e) If we constructed other models based on different samples of homes, we'd expect the slopes of the regression lines to vary, with a standard deviation of about $2.395 per square foot.

25. More sea ice.

Since conditions have been satisfied in Exercise 23, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (37 – 2) = 35 degrees of freedom.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = -4.42138 \pm (t^*_{35}) \times 0.5806 \approx (-5.60, -3.24)$$

We are 95% confident that the extent of sea ice in the northern artic is decreasing at a rate of between 3.24 and 5.60 square kilometers for each additional degree Celsius in mean global temperature.

26. Second home.

a) **Straight enough condition:** The scatterplot is straight enough, and the residuals plot looks unpatterned.
Randomization condition: The houses were selected at random.
Does the plot thicken? condition: The residuals plot shows no obvious trends in the spread.
Nearly Normal condition, Outlier condition: The histogram of residuals is unimodal and symmetric, and shows no outliers.

b) Since conditions have been satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (1064 – 2) = 1062 degrees of freedom.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = 94.4539 \pm (t^*_{1062}) \times 2.393 \approx (89.8, 99.2)$$

We are 95% confident that Saratoga housing costs increase at a rate of between $89.8 and $99.2 per square foot.

27. Hot dogs.

a) H_0: There's no association between calories and sodium content of all-beef hot dogs. $(\beta_1 = 0)$

H_A: There is an association between calories and sodium content. $(\beta_1 \neq 0)$

b) Assuming the conditions have been met, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (13 – 2) = 11 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is $\widehat{Sodium} = 90.9783 + 2.29959(Calories)$.

The value of $t = 4.10$. The P-value of 0.0018 means that the association we see in the data is very unlikely to occur by chance alone. We reject the null hypothesis, and conclude that there is evidence of a linear association between the number of calories in all-beef hotdogs and their sodium content. Because of the positive slope, there is evidence that hot dogs with more calories generally have higher sodium contents.

28. Cholesterol.

a) H_0: There is no linear relationship between age and cholesterol. $(\beta_1 = 0)$

H_A: Cholesterol levels change with age. $(\beta_1 \neq 0)$

b) Assuming the conditions have been met, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (1406 – 2) = 1404 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is $\widehat{Cholesterol} = 194.232 + 0.772(Age)$.

The value of $t = 3$. The P-value of 0.0056 means that the association we see in the data is very unlikely to occur by chance alone. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between age and cholesterol. Because of the positive slope, there is evidence that cholesterol levels tend to increase with age.

29. Second frank.

a) Among all-beef hot dogs with the same number of calories, the sodium content varies, with a standard deviation of about 60 mg.

b) The standard error of the slope of the regression line is 0.5607 milligrams of sodium per calorie.

c) If we tested many other samples of all-beef hot dogs, the slopes of the resulting regression lines would be expected to vary, with a standard deviation of about 0.56 mg of sodium per calorie.

30. More cholesterol.

a) Among adults of the same age, cholesterol levels vary, with a standard deviation of about 46 points.

b) The standard error of the slope of the regression line is 0.2574 cholesterol points per year of age.

c) If we tested many other samples of adults, the slopes of the resulting regression lines would be expected to vary with a standard deviation of 0.26 cholesterol points per year of age.

31. Last dog.

$b_1 \pm t^*_{n-2} \times SE(b_1) = 2.29959 \pm (2.201) \times 0.5607 \approx (1.07, 3.53)$; We are 95% confident that for every additional calorie, all-beef hot dogs have, on average, between 1.07 and 3.53 mg more sodium.

32. Cholesterol, finis.

$b_1 \pm t^*_{n-2} \times SE(b_1) = 0.771639 \pm (t^*_{1404}) \times 0.2574 \approx (0.27, 1.28)$; We are 95% confident that, on average, adult cholesterol levels increase by between 0.27 and 1.28 points per year of age.

33. Marriage age 2017.

a) H₀: The difference in age between men and women at first marriage has not been changing. $(\beta_1 = 0)$

H$_A$: The difference in age between men and women at first marriage has been changing. $(\beta_1 < 0)$

b) **Straight enough condition:** The scatterplot is straight enough.
Independence assumption: We are examining a relationship over time, so there is reason to be cautious, but the residuals plot shows no evidence of dependence.
Does the plot thicken? condition: The residuals plot shows no obvious trends in the spread.
Nearly Normal condition, Outlier condition: The histogram is reasonably unimodal and symmetric, and shows no obvious skewness or outliers. The normal probability plot is somewhat curved, so we should be cautious when making claims about the linear association.

c) Since conditions have been satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with 78 – 6 = 76 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is $\widehat{(Men - Women)} = 29.6175 - 0.013782(Year)$.

The value of $t = -13.3$. The P-value of less than 0.0001 (even though this is the value for a two-tailed test, it is still very small) means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a negative linear relationship between difference in age at first marriage and year. The difference in marriage age between men and women appears to be decreasing over time.

34. Used Civics 2017.

a) A linear model is probably appropriate, since the plot seems reasonably straight.

b) H₀: There is no linear relationship between price and mileage of Honda Civics. $(\beta_1 = 0)$

H$_A$: There is a linear relationship between price and mileage of Honda Civics. $(\beta_1 \neq 0)$

34. (continued)

c) The equation of the regression line is $\widehat{Price} = 17{,}164.3 - 84.1570(miles(000))$.

According to the model, the average asking price for a used Honda Civic decreases by about $84.16 for each additional thousand miles in mileage.

Assuming the conditions for inference have been satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (55 – 2) = 53 degrees of freedom.

$$t = \frac{b_1 - \beta_1}{SE(b_1)} = \frac{-84.1570 - 0}{5.907} \approx -14.25;\ \text{The P-value is less than 0.0001.}$$

d) With such a low P-value, we reject the null hypothesis and conclude that there is evidence of a linear relationship between *Price* and *miles*(000). It appears that used Civics with more miles on the odometer have lower asking prices and, specifically, that a Civic has an asking price about $84 lower per 1,000 miles driven.

e) The residual plot "thickens" at the right. Greater variation at larger values suggests that a re-expression might help, perhaps to log(*Price*) or $\sqrt{Price}$.

35. Marriage age 2015, again.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = -0.015821 \pm t^*_{35} \times 0.0010 \approx (-0.018,\ -0.014)$$

We are 95% confident that the mean difference in age between men and women at first marriage decreases by between 0.014 and 0.018 years in age for each year that passes.

36. Used Civics 2017, again.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = -84.157 \pm t^*_{53} \times 5.907 \approx (-96.0,\ -72.3)$$

We are 95% confident that the advertised price of a used Honda Civic is expected to decrease by an average of between $72.30 and $96.00 for each additional thousand miles on the odometer.

37. Streams.

a) H$_0$: There is no linear relationship between BCI and pH. $(\beta_1 = 0)$

H$_A$: There is a linear relationship between BCI and pH. $(\beta_1 \neq 0)$

b) Assuming the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with 163 – 2 = 161 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is $\widehat{BCI} = 2733.37 - 197.694(pH)$.

$$t = \frac{b_1 - \beta_1}{SE(b_1)} = \frac{-197.694 - 0}{25.57} \approx -7.73;\ \text{The P-value (two-sided!) is less than 0.0001.}$$

c) Since the P-value is so low, we reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between BCI and pH. Streams with higher pH tend to have lower BCI.

38. Civics again.

a) After accounting for the *Age* of a used Civic, its price is, on average, $54 less per thousand miles driven. Age is (understandably) correlated with mileage, so *miles*(000) has had some of its effects on price already accounted for.

b) The slope of this plot is –53.6; because a partial regression plot will have the same slope as the corresponding coefficient in the multiple regression model.

c) The car has low mileage and a high price. It makes the standard error of the coefficient (which has a term about the variance of *x* in its denominator) smaller. That increases the *t*-ratio and reduces the P-value.

39. Streams again.

a) H_0: After accounting for Alkali content, *BCI* and *pH* are not (linearly) related. $(\beta_1 = 0)$

H_A: After accounting for Alkali content, *BCI* and *pH* are associated. $(\beta_1 \neq 0)$

b) With a very large P-value of 0.47, we fail to reject H0. We failed to find evidence that *pH* is related to *BCI* after allowing for the effects of *Alkali*.

c) *pH* is likely to be correlated with *Alkali*. After allowing for the effect of *Alkali*, there is no remaining effect of *pH*. The collinearity has inflated the standard error of the coefficient of *pH*, reducing its *t*-ratio.

d) The slope of this plot is –1.45740; because a partial regression plot will have the same slope as the corresponding coefficient in the multiple regression model.

e) The point on the far left of the plot is influential, so the least squares slope of the partial regression plot is larger (closer to zero) than it would be without it. The coefficient of *Alkali* larger than it would be if the point were not included.

40. Fuel economy.

a) H_0: There is no linear relationship between the weight of a car and its mileage. $(\beta_1 = 0)$

H_A: There is a linear relationship between the weight of a car and its mileage. $(\beta_1 \neq 0)$

b) **Straight enough condition:** The scatterplot is straight enough to try a linear model.
Independence assumption: The residuals plot is scattered.
Does the plot thicken? condition: The residuals plot indicates some possible “thickening” as the predicted values increases, but it’s probably not enough to worry about.
Nearly Normal condition, Outlier condition: The histogram of residuals is unimodal and symmetric, with one possible outlier. With the large sample size, it is okay to proceed.

Since conditions have been satisfied, the sampling distribution of the regression slope can be modeled by a Student’s *t*-model with 50 – 2 = 48 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is $\widehat{MPG} = 48.7393 - 8.2136(Weight)$, where *Weight* is measured in thousands of pounds.

c) The value of $t = -12.2$. The P-value of less than 0.0001 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between weight of a car and its mileage. Cars that weigh more tend to have lower gas mileage.

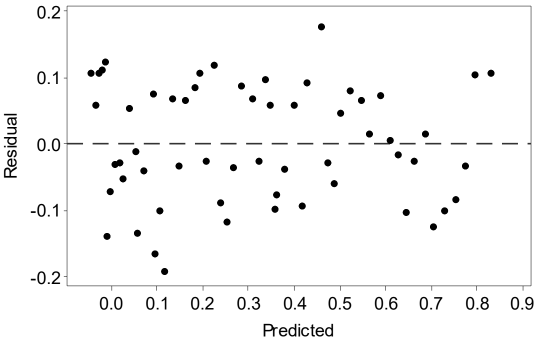
41. Climate change 2016.

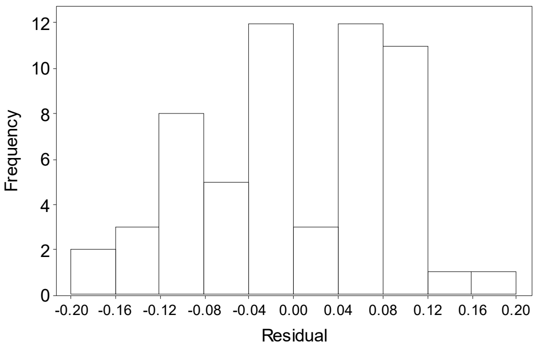
a) $\widehat{Temp} = -3.17933 + 0.00992(CO_2)$

b) HH_0: There is no linear relationship between CO_2 level and global temperature. $(\beta_1 = 0)$

H_A: There is a linear relationship between CO_2 level and global temperature. $(\beta_1 \neq 0)$

Straight enough condition: The scatterplot is straight enough to try a linear model.
Independence assumption: The residuals plot is scattered.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition, Outlier condition: The histogram of residuals is reasonably unimodal and symmetric.





Since conditions have been satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (58 – 2) = 56 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is $\widehat{Temp} = -3.17933 + 0.00992(CO_2)$.

The value of $t = 22.1$. The P-value of less than 0.0001 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between CO_2 level and global temperature. Higher CO_2 levels are associated with higher global temperature anomaly.

c) They may be useful. The standard deviation of the residuals is small and the plot is straight. $R^2 = 89.7\%$, so 89.7% of the variability in global temperature anomaly is accounted for by the linear relationship.

d) No, the model does not prove that increasing CO_2 levels are causing global warming. The model is consistent with the claim that increasing CO_2 is causing global climate change, but it does not by itself prove that this is the mechanism. Other scientific studies showing how CO_2 can trap heat are necessary for that.

42. Fuel economy, part II.

a) Since conditions have been satisfied previously, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (50 – 2) = 48 degrees of freedom. We will use a regression slope *t*-interval, with 95% confidence.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = -8.2136 \pm t^*_{48} \times 0.674 \approx (-9.57, -6.86)$$

b) We are 95% confident that the mean mileage of cars decreases by between 6.86 and 9.57 miles per gallon for each additional 1000 pounds of weight.

43. Climate change, part II.

a) Since conditions have been satisfied previously, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (58 – 2) = 56 degrees of freedom. We will use a regression slope *t*-interval, with 90% confidence.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = 0.00992 \pm t^*_{58} \times 0.0004 \approx (0.0092,\ 0.011)$$

b) We are 90% confident that the mean global temperature anomaly increases by between 0.0092 and 0.011 degrees per ppm of CO_2.

44. Fuel economy, part III.

a) The regression equation predicts that cars that weigh 2500 pounds will have a mean fuel efficiency of $48.7393 - 8.2136(2.5) = 28.2053$ miles per gallon.

$$\hat{y}_\nu \pm t^*_{n-2}\sqrt{SE^2(b_1)\cdot(x_\nu - \bar{x})^2 + \frac{s_e^2}{n}} = 28.2053 \pm t^*_{48}\sqrt{0.674^2\cdot(2.5-2.8878)^2 + \frac{2.413^2}{50}} \approx (27.34,\ 29.07)$$

We are 95% confident that cars weighing 2500 pounds will have mean fuel efficiency between 27.34 and 29.07 miles per gallon.

b) The regression equation predicts that cars that weigh 3450 pounds will have a mean fuel efficiency of $48.7393 - 8.2136(3.45) = 20.40238$ miles per gallon.

$$\hat{y}_\nu \pm t^*_{n-2}\sqrt{SE^2(b_1)\cdot(x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2} = 20.40238 \pm t^*_{48}\sqrt{0.6738^2\cdot(3.45-2.8878)^2 + \frac{2.413^2}{50} + 2.413^2}$$
$$\approx (15.44,\ 25.36)$$

We are 95%confident that a car weighing 3450 pounds will have fuel efficiency between 15.44 and 25.36 miles per gallon.

45. Climate change, again.

a) The regression equation predicts that a CO_2 level of 450 ppm will have a mean global temperature anomaly of $-3.17933 + 0.00992(450) = 1.28467$ degrees Celsius.

$$\hat{y}_\nu \pm t^*_{n-2}\sqrt{SE^2(b_1)\cdot(x_\nu - \bar{x})^2 + \frac{s_e^2}{n}} = 1.28467 \pm t^*_{56}\sqrt{0.0004^2\cdot(450-352.566)^2 + \frac{0.0885^2}{58}} \approx (1.19,\ 1.37)$$

We are 90% confident that years in which CO_2 levels are 450 ppm will have a mean global temperature anomaly of between 1.19 and 1.37 degrees Celsius.

(This interval was calculated from the original data set, using technology. Minor differences due to level of precision in various inputs are fine. These differences are unlikely to change our conclusions about the association.)

b) The regression equation predicts that a CO_2 level of 450 ppm will have a mean global temperature anomaly of $-3.17933 + 0.00992(450) = 1.28467$ degrees Celsius.

$$\hat{y}_\nu \pm t^*_{n-2}\sqrt{SE^2(b_1)\cdot(x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2} = 1.28467 \pm t^*_{56}\sqrt{0.0004^2\cdot(450-352.566)^2 + \frac{0.0885^2}{58} + 0.0885^2}$$
$$\approx (1.08,\ 1.48)$$

45. (continued)

We are 90%confident that a year in which CO_2 levels are 450 ppm will have a global temperature anomaly of between 1.08 and 1.48 degrees Celsius.

(This interval was calculated from the original data set, using technology. Minor differences due to level of precision in various inputs are fine. These differences are unlikely to change our conclusions about the association.)

c) Yes, 1.3 degrees Celsius is a plausible value, since it is within the interval.

46. Cereals.

a) H$_0$: There is no linear relationship between the number of calories and the sodium content of cereals. $(\beta_1 = 0)$

H$_A$: There is a linear relationship between the number of calories and the sodium content of cereals. $(\beta_1 \neq 0)$

Since these data were judged acceptable for inference, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with 77 – 2 = 75 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is $\widehat{Sodium} = 21.4143 + 1.29357(Calories)$.

$t = \dfrac{b_1 - \beta_1}{SE(b_1)} = \dfrac{1.29357 - 0}{0.4738} \approx 2.73;$ The value of t = 2.73. The P-value of 0.0079 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between the number of calories and sodium content of cereals. Cereals with higher numbers of calories tend to have higher sodium contents.

b) Only 9% of the variability in sodium content can be explained by the number of calories. The residual standard deviation is 80.49 mg, which is pretty large when you consider that the range of sodium content is only 320 mg. Although there is strong evidence of a linear association, it is too weak to be of much use. Predictions would tend to be very imprecise.

47. Brain size.

a) H$_0$: There is no linear relationship between brain size and IQ. $(\beta_1 = 0)$

H$_A$: There is a linear relationship between brain size and IQ. $(\beta_1 \neq 0)$

Since these data were judged acceptable for inference, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (21 – 2) = 19 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is $\widehat{IQ\_Verbal} = 24.1835 + 0.0988(Size)$.

$t = \dfrac{b_1 - \beta_1}{SE(b_1)} = \dfrac{0.0988 - 0}{0.0884} \approx 1.12;$ The P-value of 0.2775 means that the association we see in the data is likely to occur by chance. We fail to reject the null hypothesis, and conclude that there is no evidence of a linear relationship between brain size and verbal IQ score.

b) Since $R^2 = 6.5\%,$ only 6.5% of the variability in verbal IQ can be accounted for by brain size. This association is very weak. There are three students with large brains who scored high on the IQ test. Without them, there appears to be no association at all.

48. Cereals, part 2.

Straight enough condition: The scatterplot is not straight.
Independence assumption: The residuals plot shows a curved pattern.
Does the plot thicken? condition: The spread of the residuals is not consistent. The residuals plot "thickens" as the predicted values increase.
Nearly Normal condition, Outlier condition: The histogram of residuals is skewed to the right, with an outlier.

These data are not appropriate for inference.

49. City climate.

Straight enough condition: The scatterplot is not straight.
Independence assumption: The residuals plot shows a curved pattern.
Does the plot thicken? condition: The spread of the residuals is not consistent. The residuals plot shows decreasing variability as the predicted values increase.
Nearly Normal condition, Outlier condition: The histogram of residuals is skewed to the right, with an outlier.

These data are not appropriate for inference.

50. See ice?

a) A P-value of 0.64 provides no evidence that sea ice *Extent* is related to *Mean global Temp* after allowing for *Year*.

b) No. Global temperature is probably strongly correlated with the year because temperatures have been rising consistently in recent years. That collinearity would inflate the standard error of the coefficient and thereby reduce the *t*-ratio.

c) The strength of the collinearity of *Mean global Temp* and *Year since 1970* is overwhelming the effect of *Mean global Temp* alone.

51. Ozone and population.

a) H_0: There is no linear relationship between population and ozone level. $(\beta_1 = 0)$

H_A: There is a linear relationship between population and ozone level. $(\beta_1 \neq 0)$

Assuming the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with 16 – 2 = 14 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is $\widehat{Ozone} = 18.892 + 6.650(Population)$, where ozone level is measured in parts per million and population is measured in millions.

$t = \dfrac{b_1 - \beta_1}{SE(b_1)} = \dfrac{6.650 - 0}{1.910} \approx 3.48;$ The P-value of 0.0037 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a positive linear relationship between ozone level and population. Cities with larger populations tend to have higher ozone levels.

b) City population is a good predictor of ozone level. Population explains 84% of the variability in ozone level and *s* is just over 5 parts per million.

52. Sales and profits.

a) H₀: There is no linear relationship between sales and profit. $\left(\beta_1 = 0\right)$

H_A: There is a linear relationship between sales and profit. $\left(\beta_1 \neq 0\right)$

Assuming the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with 79 – 2 = 77 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is $\widehat{Profits} = -176.644 + 0.092498(Sales)$, with both profits and sales measured in millions of dollars.

$t = \dfrac{b_1 - \beta_1}{SE(b_1)} = \dfrac{0.092498 - 0}{0.0075} \approx 12.33$; The P-value of essentially 0 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between sales and profits. Companies with higher sales tend to have higher profits.

b) A company's sales won't be of much use in predicting profits. $R^2 = 66.2\%$, so only 66.2% of the variability in profits can be accounted for by sales. The standard deviation of residuals, *s*, is nearly half a billion dollars, but the mean profit for these companies is less than half of that.

53. Ozone, again

a) $b_1 \pm t^*_{n-2} \times SE\left(b_1\right) = 6.65 \pm (1.761) \times 1.910 \approx (3.29, 10.01)$

We are 90% confident that each additional million people will increase mean ozone levels by between 3.29 and 10.01 parts per million.

b) The regression equation predicts that cities with a population of 600,000 people will have ozone levels of 18.892 + 6.650(0.6) = 22.882 parts per million.

$$\hat{y}_\nu \pm t^*_{n-2}\sqrt{SE^2(b_1)\cdot(x_\nu - \bar{x})^2 + \frac{s_e^2}{n}} = 22.882 \pm (1.761)\sqrt{1.91^2 \cdot (0.6 - 1.7)^2 + \frac{5.454^2}{16}} \approx (18.47, 27.29)$$

We are 90% confident that the mean ozone level for cities with populations of 600,000 will be between 18.47 and 27.29 parts per million.

54. More sales and profits.

a) There are 77 degrees of freedom, so use $t^*_{75} = 1.992$ as a conservative estimate from the table.

$b_1 \pm t^*_{n-2} \times SE\left(b_1\right) = 0.092498 \pm (1.992) \times 0.0075 \approx (0.078, 0.107)$

We are 95% confident that each additional million dollars in sales will increase mean profits by between $78,000 and $107,000.

b) The regression equation predicts that corporations with sales of $23,000 million dollars will have profits of $-176.644 + 0.092498(23{,}000) = 1950.603$ million dollars.

$$\hat{y}_\nu \pm t^*_{n-2}\sqrt{SE^2(b_1)\cdot(x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$$

$$= 1950.603 \pm (1.992)\sqrt{0.0075^2 \cdot (23{,}000 - 4178.29)^2 + \frac{466.2^2}{79} + 466.2^2} \approx (974.98, 2926.63)$$

We are 95% confident that Eli Lilly's profits will be between 974,980,000 and $2,926,630,000. This interval is too wide to be of any use.

55. Tablet computers 2014.

a) Since there are 34 – 2 = 32 degrees of freedom, there were 34 tablet computers tested.

b) **Straight enough condition:** The scatterplot is roughly straight, but scattered.
Independence assumption: The residuals plot shows no pattern.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition: The Normal probability plot of residuals is reasonably straight.

c) H$_0$: There is no linear relationship between maximum brightness and battery life. $(\beta_1 = 0)$

H$_A$: There is a positive linear relationship between maximum brightness and battery life. $(\beta_1 > 0)$

Since the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with 34 – 2 = 32 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is $\widehat{Hours} = 5.38719 + 0.00904(\mathit{ScreenBrightness})$, where battery life is measured in hours and screen brightness is measured in cd/m^2.

The value of $t \approx 2.02$. The P-value of 0.0522 means that the association we see in the data is not unlikely to occur by chance. We fail to reject the null hypothesis, and conclude that there is little evidence of a positive linear relationship between battery life and screen brightness.

d) Since $R^2 = 11.3\%$, only 11.3% of the variability in battery life can be accounted for by screen brightness. The residual standard deviation is 2.128 hours. That's pretty large, considering the range of battery life is only about 9 hours. Even if we concluded that there was some evidence of a linear association, it is too weak to be of much use. Predictions would tend to be very imprecise.

e) The equation of the line of best fit for these data points is $\widehat{Hours} = 5.38719 + 0.00904(\mathit{ScreenBrightness})$, battery life measure in hours and screen brightness measured in cd/m^2.

f) There are 32 degrees of freedom, so use $t_{32}^* = 1.694$.

$$b_1 \pm t_{n-2}^* \times SE(b_1) = 0.00904 \pm (1.694) \times 0.0045 \approx (0.00142,\ 0.0167)$$

g) We are 90% confident that the mean battery life increases by between 0.00142 and 0.0167 hours for each additional cd/m^2 of screen brightness.

56. Crawling.

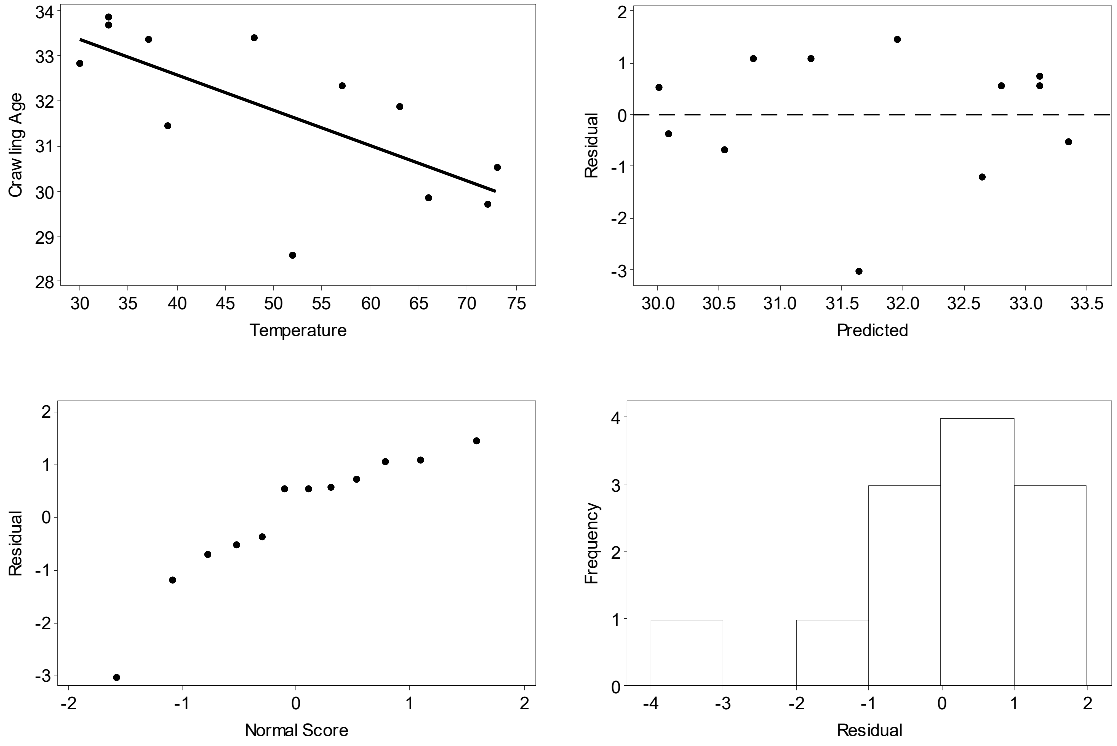
a) If the data had been plotted for individual babies, the association would appear to be weaker, since individuals are more variable than averages.

b) H$_0$: There is no linear relationship between 6-month temperature and crawling age. $(\beta_1 = 0)$

H$_A$: There is a linear relationship. $(\beta_1 \neq 0)$

Straight enough condition: The scatterplot is straight enough to try linear regression.
Independence assumption: The residuals plot shows no pattern, but there may be an outlier. If the month of May were just one data point, it would be removed. However, since it represents the average crawling age of several babies, there is no justification for its removal.
Does the plot thicken? condition: The spread of the residuals is consistent
Nearly Normal condition: The Normal probability plot of residuals isn't very straight, largely because of the data point for May. The histogram of residuals also shows this outlier.

56. (continued)



Since we had difficulty with the conditions for inference, we will proceed cautiously. These data may not be appropriate for inference. The sampling distribution of the regression slope can be modeled by a Student's *t*-model with 12 – 2 = 10 degrees of freedom. We will use a regression slope *t*-test.

```
Dependent variable is:      Age
No Selector
R squared = 49.0%     R squared (adjusted) = 43.9%
s =  1.319  with  12 - 2 = 10  degrees of freedom

Source        Sum of Squares      df    Mean Square     F-ratio
Regression    16.6933              1      16.6933         9.59
Residual      17.4028             10       1.74028

Variable      Coefficient     s.e. of Coeff     t-ratio        prob
Constant      35.6781         1.318             27.1         ≤ 0.0001
Temp          -0.077739       0.0251             -3.10          0.0113
```

The equation of the line of best fit for these data points is $\widehat{Age} = 35.6781 - 0.077739(Temp)$, with average crawling age measured in weeks and average temperature in °F.

The value of $t \approx -3.10$. The P-value of 0.0113 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between average temperature and average crawling age. Babies who reach six months of age in warmer temperatures tend to crawl at earlier ages than babies who reach six months of age in colder temperatures.

c) $b_1 \pm t^*_{n-2} \times SE(b_1) = -0.077739 \pm (2.228) \times 0.0251 \approx (-0.134, -0.022)$

We are 95% that the average crawling age decreases by between 0.022 weeks and 1.34 weeks when the average temperature increases by 10°F.

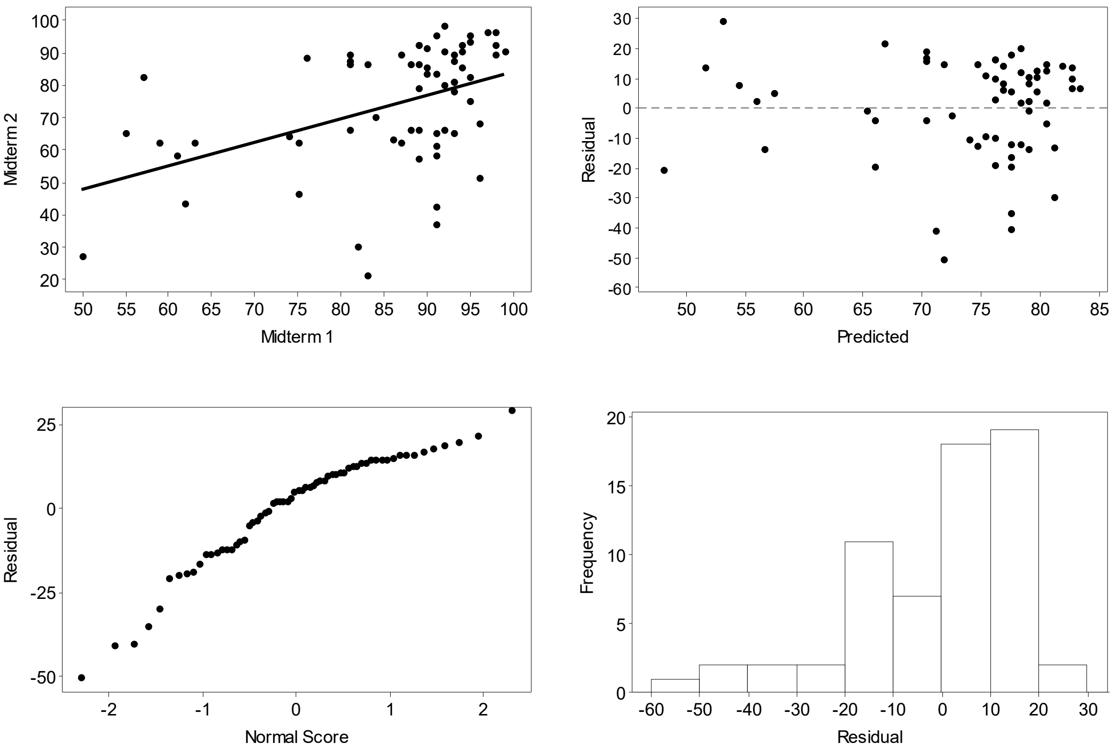
57. Midterms.

a) The regression output is shown below.

```
Dependent variable is:      Midterm 2
No Selector
R squared = 19.9%    R squared (adjusted) = 18.6%
s =  16.78  with  64 - 2 = 62  degrees of freedom

Source          Sum of Squares      df    Mean Square     F-ratio
Regression      4337.14              1        4337.14        15.4
Residual        17459.5             62        281.604

Variable        Coefficient     s.e. of Coeff     t-ratio      prob
Constant        12.0054         15.96             0.752        0.4546
Midterm 1       0.720990          0.1837          3.92         0.0002
```

The model is $\widehat{Midterm2} = 12.005 + 0.721(Midterm1)$.

b) **Straight enough condition:** The scatterplot shows a weak, positive relationship between Midterm 2 score and Midterm 1 score. There are several outliers, but removing them only makes the relationship slightly stronger. The relationship is straight enough to try linear regression.
Independence assumption: The residuals plot shows no pattern.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition, Outlier condition: The histogram of the residuals is unimodal, slightly skewed with several possible outliers. The Normal probability plot shows some slight curvature.



Since we had some difficulty with the conditions for inference, we should be cautious in making conclusions from these data. The small P-value of 0.0002 for the slope would indicate that the slope is statistically distinguishable from zero, but the R^2 value of 0.199 suggests that the relationship is weak. Midterm 1 isn't a useful predictor of Midterm 2.

57. (continued)

c) The student's reasoning is not valid. The R^2 value is only 0.199 and the value of *s* is 16.8 points. Although correlation between Midterm 1 and Midterm 2 may be statistically significant, it isn't of much practical use in predicting Midterm 2 scores. It's too weak.

58. Midterms?

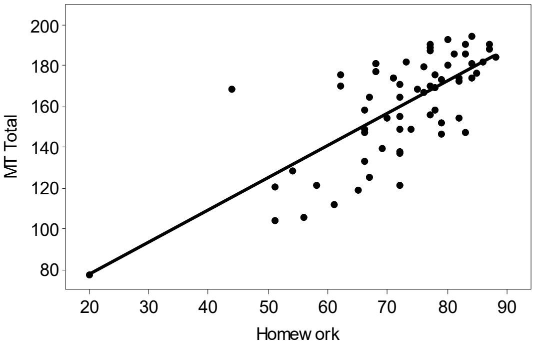
a) The regression output is shown below.

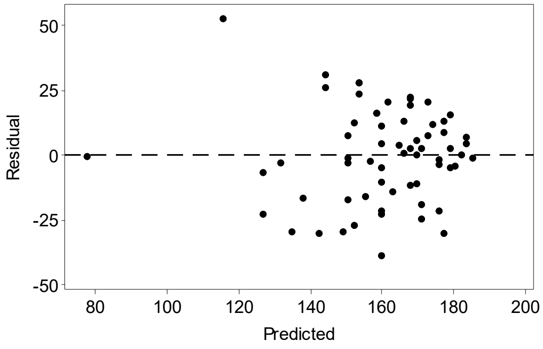
```
Dependent variable is:        M1+M2
No Selector
R squared = 50.7%    R squared (adjusted) = 49.9%
s =  18.30  with  64 - 2 = 62  degrees of freedom

Source         Sum of Squares      df     Mean Square       F-ratio
Regression     21398.1              1         21398.1          63.9
Residual       20773.0             62         335.048

Variable       Coefficient     s.e. of Coeff     t-ratio        prob
Constant       46.0619         14.46             3.19          0.0023
Homework       1.58006           0.1977          7.99        ≤ 0.0001
```

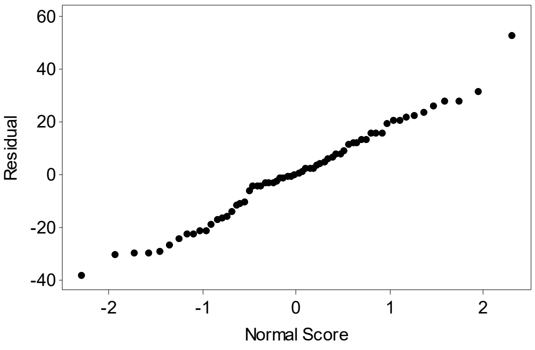
The model is $\widehat{MTtotal} = 46.062 + 1.580(Homework)$.

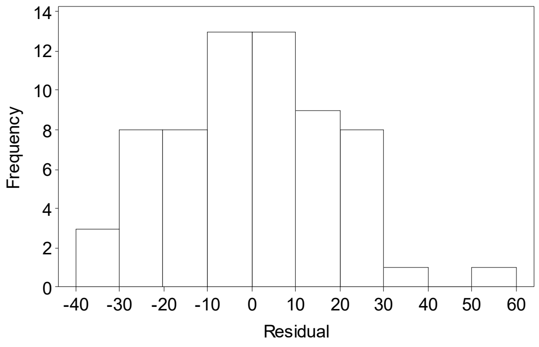
b) **Straight enough condition:** The scatterplot shows a moderate, positive relationship between Midterm total and homework. There are two outliers, but removing them does not significantly change the model. The relationship is straight enough to try linear regression.
Independence assumption: The residuals plot shows no pattern.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition: The histogram of the residuals is unimodal and symmetric, and the Normal probability plot is reasonably straight.

Since the conditions are met, linear regression is appropriate. The small P-value for the slope would indicate that the slope is statistically distinguishable from zero.









58. (continued)

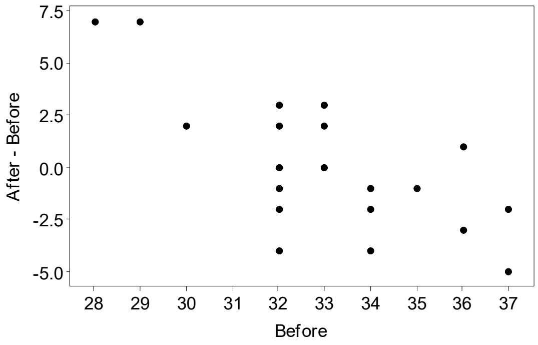
c) The R^2 value of 0.507 suggests that the overall relationship is fairly strong. However, this does not mean that midterm total is accurately predicted from homework scores. The error standard deviation of 18.30 indicates that a prediction in midterm total could easily be off by 20 to 30 points. If this is significant number of points for deciding grades, then homework score alone will not suffice.

59. **Strike two.**

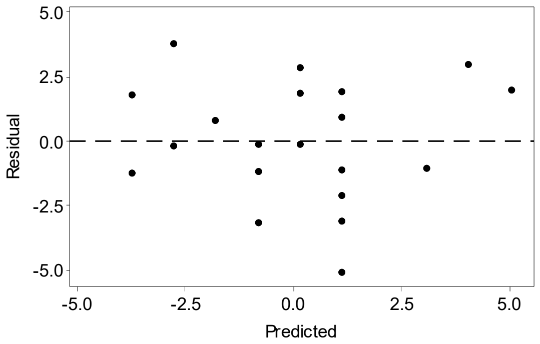
H$_0$: Effectiveness is independent of the player's initial ability. $(\beta_1 = 0)$

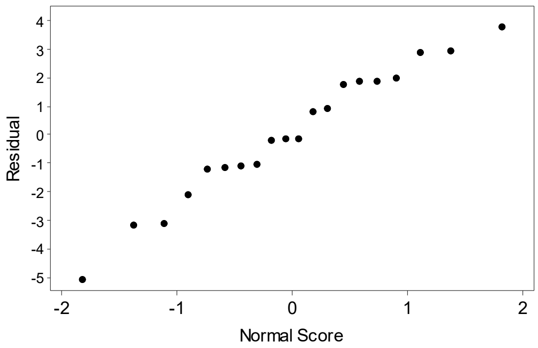
H$_A$: Effectiveness of the video depends on the player's initial ability. $(\beta_1 \neq 0)$

Straight enough condition: The scatterplot is straight enough.



Independence assumption: The residuals plot shows no pattern.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition: The Normal probability plot is straight.





Since the conditions for inference are inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with 20 – 2 = 18 degrees of freedom. We will use a regression slope *t*-test.

The equation of the line of best fit for these data points is $\widehat{(After - Before)} = 32.316 - 0.9748(Before),$ where we are counting the number of strikes thrown before and after the training program.

The value of $t \approx -4.34$. Since the P-value is 0.004, reject the null hypothesis, and conclude that there is evidence of a linear relationship between the player's initial ability and the effectiveness of the program. The negative slope indicates that the method is more effective for those whose initial performance was poorest and less effective for those whose initial performance was better. This appears to be a case of regression to the mean. Those who were above average initially tended to do worse after training. Those who were below average initially tended to improve.

60. All the efficiency money can buy 2011.

a) We'd like to know if there is a linear association between price and fuel efficiency in cars. We have data on 2011 model year cars, with information on highway MPG and retail price.
H$_0$: There is no linear relationship between MPG and retail price. $(\beta_1 = 0)$
H$_A$: Highway MPG and retail price are linearly associated. $(\beta_1 \neq 0)$

b) The scatterplot fails the Straight enough condition. There is no evidence of any association whatsoever between highway MPG and retail price. We can't really perform a test.

c) Since the conditions are not satisfied, we cannot continue this analysis.

61. Education and mortality.

a) **Straight enough condition:** The scatterplot is straight enough.
Independence assumption: The residuals plot shows no pattern. If these cities are representative of other cities, we can generalize our results.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition, Outlier condition: The histogram of the residuals is unimodal and symmetric with no outliers.

b) H$_0$: There is no linear relationship between education and mortality. $(\beta_1 = 0)$
H$_A$: There is a linear relationship between education and mortality. $(\beta_1 \neq 0)$

Since the conditions for inference are inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with 58 – 2 = 56 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is

$\widehat{Mortality} = 1493.26 - 49.9202(Education).$

$t = \dfrac{b_1 - \beta_1}{SE(b_1)} = \dfrac{-49.9202 - 0}{8.000} \approx -6.24$; The P-value of essentially 0 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between the level of education in a city and its mortality rate. Cities with lower education levels tend to have higher mortality rates.

c) We cannot conclude that getting more education is likely to prolong your life. Association does not imply causation. There may be lurking variables involved.

d) For 95% confidence, $t_{56}^* \approx 2.00327.$

$$b_1 \pm t_{n-2}^* \times SE\left(b_1\right) = -49.9202 \pm (2.003) \times 8.000 \approx (-65.95, -33.89)$$

e) We are 95% confident that the mean number of deaths per 100,000 people decreases by between 33.89 and 65.95 deaths for an increase of one year in average education level.

f) The regression equation predicts that cities with an adult population with an average of 12 years of school will have a mortality rate of $1493.26 - 49.9202(12) = 894.2176$ deaths per 100,000. The average education level was 11.0328 years.

$$\hat{y}_\nu \pm t_{n-2}^* \sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}} = 894.2176 \pm (2.003)\sqrt{8.00^2 \cdot (12 - 11.0328)^2 + \frac{47.92^2}{58}}$$
$$\approx (874.239, 914.196)$$

We are 95% confident that the mean mortality rate for cities with an average of 12 years of schooling is between 874.239 and 914.196 deaths per 100,000 residents.

62. Property assessments.

a) **Straight enough condition:** The scatterplot is straight enough.
Independence assumption: The residuals plot shows no pattern. If these cities are representative of other cities, we can generalize our results.
Does the plot thicken? condition: The spread of the residuals is consistent
Nearly Normal condition: The Normal probability plot is fairly straight.

b) H_0: There is no linear relationship between size and assessed valuation. $\left(\beta_1 = 0\right)$

H_A: Larger houses have higher assessed values. $\left(\beta_1 > 0\right)$

Since the conditions for inference are inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (18 – 2) = 16 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is
$\widehat{Assess} = 37{,}108.8 + 11.8987(SqFt)$.

$t = \frac{b_1 - \beta_1}{SE(b_1)} = \frac{11.89987 - 0}{4.290} \approx 2.77$; The P-value of 0.0068 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between the size of a home and its assessed value. Larger homes tend to have higher assessed values.

c) $R^2 = 32.5\%$; This model accounts for 32.5% of the variability in assessments.

d) For 90% confidence, $t_{16}^* \approx 1.746$.

$$b_1 \pm t_{n-2}^* \times SE\left(b_1\right) = 11.8987 \pm (1.746) \times 4.290 \approx (4.41, 19.39)$$

e) We are 90% confident that the mean assessed value increases by between \$441 and \$1939 for each additional 100 square feet in size.

f) The regression equation predicts that houses measuring 2100 square feet will have an assessed value of $37108.8 + 11.8987(2100) = \$62{,}096.07$. The average size of the houses sampled is 2003.39 square feet.

$$\hat{y}_\nu \pm t_{n-2}^* \sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2} = 62096.07 \pm (2.120)\sqrt{4.290^2 \cdot (2100 - 2003.39)^2 + \frac{4682^2}{18} + 4682^2}$$
$$\approx (51860,\ 72332)$$

We are 95% confident that the assessed value of a home measuring 2100 square feet will have an assessed value between \$51,860 and \$72,332. There is no evidence that this home has an assessment that is too high. The assessed value of \$70,200 falls within the prediction interval.

The homeowner might counter with an argument based on the mean assessed value of all homes such as this one.

$$\hat{y}_\nu \pm t_{n-2}^* \sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}} = 62096.07 \pm (2.120)\sqrt{4.290^2 \cdot (2100 - 2003.39)^2 + \frac{4682^2}{18}}$$
$$\approx (\$59{,}597,\ \$64{,}595)$$

The homeowner might ask the city assessor to explain why his home is assessed at \$70,200, if a typical 2100-square-foot home is assessed at between \$59,597 and \$64,595.

63. Embrace or protect?

a) $\widehat{\text{Logit}(Embrace)} = 0.5796 - 0.0149age$

b) Yes, the P-value is < 0.01, which means there is strong evidence to suggest the association.

c) The coefficient on age is negative, so an older person is less likely to respond "Embrace."

64. Cost of higher education.

a) $\widehat{\text{Logit}(Type)} = -13.1461 + 0.08455Top10\% + 0.000259\$ / Student$

b) Yes, the percent of students in the top 10% is statistically significant, since the P-value of 0.033 is less than $\alpha = 0.05$.

c) Yes, the amount of money spent per student is statistically significant, since the P-value of 0.003 is less than $\alpha = 0.05$.