Chapter 27 – Inferences for Regression

1. **Hurricane predictions.**

   a) The equation of the line of best fit for these data points is $\widehat{Error} = 453.22 - 8.37(Year)$, where *Year* is measured in years since 1970. According to the linear model, the error made in predicting a hurricane's path was about 453 nautical miles, on average, in 1970. It has been declining at rate of about 8.37 nautical miles per year.

   b) H$_0$: There has been no change prediction accuracy. $(\beta_1 = 0)$

   H$_A$: There has been a change 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 (34 – 2) = 32 degrees of freedom. We will use a regression slope *t*-test.

   The value of t = -6.92. 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) 58.8% of the variation in the prediction accuracy is accounted for by the linear model based on year.

2. **Drug use.**

   a) The equation of the line of best fit for these data points is $\%\widehat{OtherDrugs} = -3.068 + 0.615(\%Marijuana)$. According to the linear model, the percentage of ninth graders in these countries who use other drugs increases by about 0.615% for each additional 1% of ninth graders who use marijuana.

   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 (11 – 2) = 9 degrees of freedom. We will use a regression slope *t*-test.

   The value of t = 7.85. The *P*-value of 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 percentage of ninth graders who use other drugs is related to the percentage of ninth graders who use marijuana. Countries with a high percentage of ninth graders using marijuana tend to have a high percentage of ninth graders using other drugs.

   d) 87.3% of the variation in the percentage of ninth graders using other drugs can be accounted for by the percentage of ninth graders using marijuana.

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

3. Movie budgets.

a) $\widehat{Budget} = -31.387 + 0.714(RunTime)$**.** The model suggests that each additional minuter of run time for a movie costs about $714,000.

b) A negative intercept makes no sense, but the *P*-value of 0.07 indicates that we can't discern a difference between our estimated value and zero. The statement that a movie of zero length should cost $0 makes sense.

c) Amounts by which movie costs differ from predictions made by this model vary, with a standard deviation of about $33 million.

d) The standard error of the slope is 0.1541 million dollars per minute.

e) If we constructed other models based on different samples of movies, we'd expect the slopes of the regression lines to vary, with a standard deviation of about $154,000 per minute.

4. House prices.

a) $\widehat{Price} = -0.321 + 94.5(Size)$**.** 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 0.50 indicates that we can't discern a difference between our estimated value and zero. The statement that a house of zero square feet should cost $0 makes sense.

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 2.393 dollars per square foot.

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 $2.39 per square foot.

5. Movie budgets, the sequel.

a) **Straight enough condition:** The scatterplot is straight enough, and the residuals plot looks unpatterned.
Independence assumption: 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 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 (120 – 2) = 118 degrees of freedom.

$$b_1 \pm t^*_{n-2} \times SE\left(b_1\right) = 0.714 \pm (t^*_{118}) \times 0.1541 \approx (0.41, 1.02)$$

We are 95% confident that the cost of making longer movies increases at a rate of between 0.41 and 1.02 million dollars per minute.

6. 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.

7. 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 of all-beef hot dogs. $(\beta_1 \neq 0)$

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.

8. Cholesterol 2007.

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

H_A: Cholesterol levels tend to increase with age. $(\beta_1 > 0)$

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 0.0028 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.

9. 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.

10. 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.

11. Last dog.

$$b_1 \pm t^*_{n-2} \times SE\left(b_1\right) = 2.29959 \pm (2.201) \times 0.5607 \approx (1.03, 3.57)$$

We are 95% confident that for every additional calorie, all-beef hot dogs have, on average, between 1.03 and 3.57 mg more sodium.

12. Cholesterol, finis.

$$b_1 \pm t^*_{n-2} \times SE\left(b_1\right) = 0.771639 \pm (t^*_{1404}) \times 0.2574 \approx (0.27, 1.28)$$

We are 95% confident that, on average, adult cholesterol levels increase by between 0.27 and 1.28 points per year of age.

13. Marriage age 2003.

a) H_0: The difference in age between men and women at first marriage has not been decreasing since 1975. $\left(\beta_1 = 0\right)$

H_A: The difference in age between men and women at first marriage has been decreasing since 1975. $\left(\beta_1 < 0\right)$

b) Straight enough condition: The scatterplot is not provided, but the residuals plot looks unpatterned. The scatterplot is likely to be 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 not particularly unimodal and symmetric, but shows no obvious skewness or outliers.

c) Since conditions have been satisfied, the sampling distribution of the regression slope can be modeled by a Student's t-model with (28 – 2) = 26 degrees of freedom. We will use a regression slope t-test. The equation of the line of best fit for these data points is:

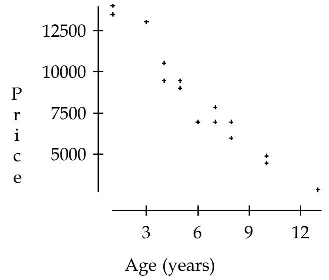
$$(\widehat{Men - Women}) = 61.8 - 0.030(Year)$$

The value of $t = -7.04$. 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.

14. Used cars 2007.

a) A scatterplot of the used cars data is at the right.

b) A linear model is probably appropriate. The plot appears to be linear.



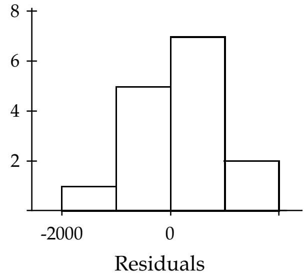
c)

```
Dependent variable is:   Price
No Selector
R squared = 94.4%     R squared (adjusted) = 94.0%
s =  816.2  with  15 - 2 = 13  degrees of freedom

Source         Sum of Squares   df   Mean Square   F-ratio
Regression     146917777         1   146917777      221
Residual       8660659          13      666205

Variable       Coefficient   s.e. of Coeff   t-ratio     prob
Constant       14285.9       448.7             31.8     ≤ 0.0001
Age (years)    -959.046        64.58          -14.9     ≤ 0.0001
```

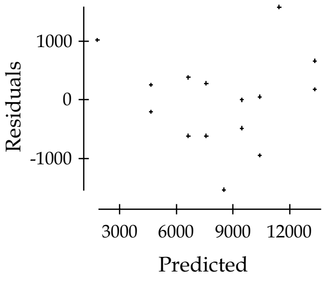
The equation of the regression line is:

$\widehat{Price} = 14286 - 959(Age)$.

According to the model, the average asking price for a used Toyota Corolla decreases by about $959 dollars for each additional year in age. Let's take a closer look.



d) **Straight enough condition:** The scatterplot is straight enough to try a linear model.
Independence assumption: Prices of Toyota Corollas of different ages might be related, but the residuals plot looks fairly scattered. (The fact that there are several prices for some years draws our eyes to some patterns that may not exist.)
Does the plot thicken? condition: The residuals plot shows no obvious patterns in the spread.
Nearly Normal condition, Outlier condition: The histogram is reasonably unimodal and symmetric, and shows no obvious skewness or outliers.



Since conditions have been satisfied, the sampling distribution of the regression slope can be modeled by a Student's t-model with (17 – 2) = 15 degrees of freedom.

15. Marriage age 2003, again.

$$b_1 \pm t^*_{n-2} \times SE\left(b_1\right) = -0.02997 \pm (2.056) \times 0.0043 \approx (-0.039, -0.021)$$

We are 95% confident that the mean difference in age between men and women at first marriage decreases by between 0.021 and 0.039 years in age for each year that passes.

16. Used cars 2007, again.

$$b_1 \pm t^*_{n-2} \times SE\left(b_1\right) = -959 \pm (2.160) \times 64.58 \approx (-1099, -819.5)$$

We are 95% confident that the advertised price of a used Toyota Corolla is decreasing by an average of between $819.50 and $1099 for each additional year in age.

17. Fuel economy.

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

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

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: $\widehat{MPG} = 48.7393 - 8.21362(Weight)$, where *Weight* is measured in thousands of pounds.

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.

18. SAT scores.

a) H_0: There is no linear relationship between SAT Verbal and Math scores. $\left(\beta_1 = 0\right)$

H_A: There is a linear relationship between SAT Verbal and Math scores. $\left(\beta_1 \neq 0\right)$

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 spread of the residuals is consistent.
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 (162 – 2) = 160 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is: $\widehat{Math} = 209.554 + 0.675075(Verbal)$.

The value of $t = 11.9$. 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 SAT Verbal and Math scores. Students with higher SAT-Verbal scores tend to have higher SAT-Math scores.

19. Fuel economy, part II.

a) Since conditions have been satisfied in Exercise 7, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (50 – 2) = 48 degrees of freedom. (Use $t^*_{45} = 2.014$ from the table.) We will use a regression slope *t*-interval, with 95% confidence.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = -8.21362 \pm (2.014) \times 0.6738 \approx (-9.57, -6.86)$$

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.

20. SATs, part II.

a) Since conditions have been satisfied in Exercise 8, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (162 – 2) = 160 degrees of freedom. (Use $t^*_{140} = 1.656$ from the table.) We will use a regression slope *t*-interval, with 90% confidence.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = 0.675075 \pm (1.656) \times 0.0568 \approx (0.581, 0.769)$$

b) We are 90% confident that the mean Math SAT scores increase by between 0.581 and 0.769 point for each additional point scored on the Verbal test.

21. *Fuel economy, part III.

a) The regression equation predicts that cars that weigh 2500 pounds will have a mean fuel efficiency of $48.7393 - 8.21362(2.5) = 28.20525$ miles per gallon.

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

$$= 28.20525 \pm (2.014)\sqrt{0.6738^2 \cdot (2.5 - 2.8878)^2 + \frac{2.413^2}{50}} \approx (27.34, 29.07)$$

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 $48.7393 - 8.21362(3.45) = 20.402311$ miles per gallon.

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

$$= 20.402311 \pm (2.014)\sqrt{0.6738^2 \cdot (3.45 - 2.8878)^2 + \frac{2.413^2}{50} + 2.413^2} \approx (15.44, 25.37)$$

We are 95%confident that a car weighing 3450 pounds will have fuel efficiency between 15.44 and 25.37 miles per gallon.

22. SATs again.

a) The regression equation predicts that students with an SAT-Verbal score of 500 will have a mean SAT-Math score of $209.554 + 0.675075(500) = 547.0915$.

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

$$= 547.0915 \pm (1.656)\sqrt{0.0568^2 \cdot (500 - 596.292)^2 + \frac{71.75^2}{162}} \approx (534.09,\ 560.10)$$

We are 90% confident that students with scores of 500 on the SAT-Verbal will have a mean SAT-Math score between 534.09 and 560.10.

b) The regression equation predicts that students with an SAT-Verbal score of 710 will have a mean SAT-Math score of $209.554 + 0.675075(710) = 688.85725$.

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

$$= 688.85725 \pm (1.656)\sqrt{0.0568^2 \cdot (710 - 596.296)^2 + \frac{71.75^2}{162} + 71.75^2} \approx (569.19,\ 808.52)$$

We are 90%confident that a student scoring 710 on the SAT-Verbal would have an SAT-Math score of between 569.19 and 808.52. Since we are talking about individual scores, and not means, it is reasonable to restrict ourselves to possible scores, so we are 90% confident that the class president scored between 570 and 800 on the SAT-Math test.

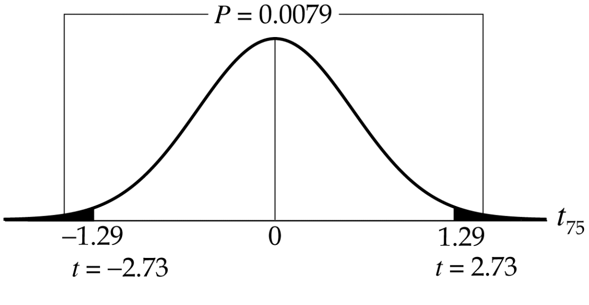
23. Cereal.

a) H0: There is no linear relationship between the number of calories and the sodium content of cereals. $(\beta_1 = 0)$

HA: 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: $\widehat{Sodium} = 21.4143 + 1.29357(Calories)$.

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.

24. 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. $(\beta_1 \neq 0)$

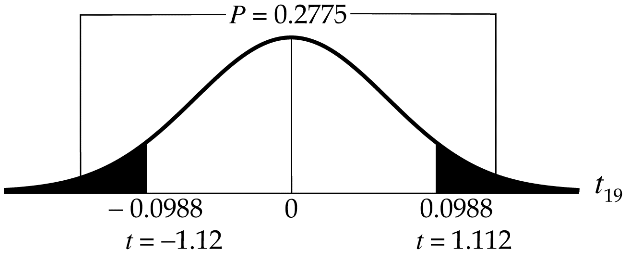
Since these data were judged acceptable for inference, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (21 – 2) = 19 degrees of freedom. (There are 21 dots on the scatterplot. I counted!) We will use a regression slope *t*-test. The equation of the line of best fit for these data points is: $IQ\_\hat{V}erbal = 24.1835 + 0.098842(Size)$.

The value of $t \approx 1.12$. 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.

$$t = \frac{b_1 - \beta_1}{SE(b_1)}$$

$$t = \frac{0.098842 - 0}{0.0884}$$

$$t \approx 1.12$$



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.

25. Another bowl.

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.

26. Winter.

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.

27. Acid rain.

a) H0: There is no linear relationship between BCI and pH. $(\beta_1 = 0)$

HA: 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: $\widehat{BCI} = 2733.37 - 197.694(pH)$.

c) The value of $t \approx -7.73$. The *P*-value (two-sided!) 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 BCI and pH. Streams with higher pH tend to have lower BCI.

$$t = \frac{b_1 - \beta_1}{SE(b_1)}$$

$$t = \frac{-197.694 - 0}{25.57}$$

$$t \approx -7.73$$

28. El Niño.

a) The regression equation is $\widehat{Temp} = 15.3066 + 0.004(CO_2)$, with CO_2 concentration measured in parts per million from the top of Mauna Loa in Hawaii, and temperature in degrees Celsius.

b) H0: There is no linear relationship between temperature and CO_2 concentration. $(\beta_1 = 0)$
HA: There is a linear relationship between temperature and CO_2 concentration. $(\beta_1 \neq 0)$

Since the scatterplots and residuals plots showed that the data were appropriate for inference, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (37 – 2) = 35 degrees of freedom. We will use a regression slope *t*-test.

$$t = \frac{b_1 - \beta_1}{SE(b_1)}$$

$$t = \frac{0.004 - 0}{0.0009}$$

$$t \approx 4.44$$

The value of $t \approx 4.44$. The *P*-value (two-sided!) of about 0.00008 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 concentration and temperature. Years with higher CO_2 concentration tend to be warmer, on average.

c) Since $R^2 = 33.4\%$, only 33.4% of the variability in temperature can be accounted for by the CO_2 concentration. Although there is strong evidence of a linear association, it is weak. Predictions would tend to be very imprecise.

29. Ozone.

a) H0: There is no linear relationship between population and ozone level. $(\beta_1 = 0)$
HA: There is a positive linear relationship between population and ozone level. $(\beta_1 > 0)$

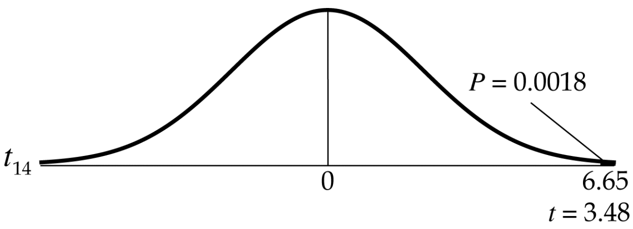
Assuming the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (16 – 2) = 14 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is: $\widehat{Ozone}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

The value of $t \approx 3.48$. The *P*-value of 0.0018 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.

$$t = \frac{b_1 - \beta_1}{SE(b_1)}$$

$$t = \frac{6.650 - 0}{1.910}$$

$$t \approx 3.48$$



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.

30. Sales and profits.

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

Assuming the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (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.

The value of $t \approx 12.33$. 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.

$$t = \frac{b_1 - \beta_1}{SE(b_1)}$$

$$t = \frac{0.092498 - 0}{0.0075}$$

$$t \approx 12.33$$

b) A company's sales may be of some help in predicting profits. $R^2 = 66.2\%$, so 66.2% of the variability in profits can be accounted for by sales. Although *s* is nearly half a billion dollars, the mean profit for these companies is over 4 billion dollars.

31. Ozone, again

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

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.

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

$$= 22.882 \pm (1.761)\sqrt{1.91^2 \cdot (0.6 - 1.7)^2 + \frac{5.454^2}{16}} \approx (18.47, 27.29)$$

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.

32. More sales and profits.

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

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

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 $9,000 million dollars will have profits of $-176.644 + 0.092498(9000) = 655.838$ million dollars.

$$\hat{y}_v \pm t_{n-2}^{*}\sqrt{SE^2(b_1)\cdot(x_v - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$$

$$= 655.838 \pm (1.992)\sqrt{0.0075^2 \cdot (9000 - 4178.29)^2 + \frac{466.2^2}{79} + 466.2^2} \approx (-281.46,\ 1593.14)$$

We are 95% confident that the Eli Lilly's profits will be between –$281,460,000 and $1,593,140,000. This interval is too wide to be of any use.

(If you use $t_{77}^{*} = 1.991297123$, your interval will be $(-281.1,\ 1592.8)$)

33. Start the car!

a) Since the number of degrees of freedom is 33 – 2 = 31, there were 33 batteries tested.

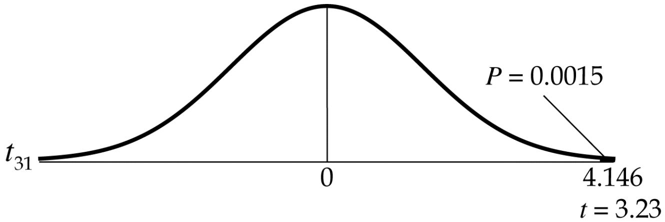
b) **Straight enough condition:** The scatterplot is roughly straight, but very scattered.
Independence assumption: The residuals plot shows no pattern.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition: The Normal probability plot of residuals is reasonably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

c) H_0: There is no linear relationship between cost and power. $(\beta_1 = 0)$

H_A: There is a positive linear relationship between cost and power. $(\beta_1 > 0)$

Since the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (33 – 2) = 31 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is: $\widehat{Power} = 384.594 + 4.14649(Cost)$, with power measured in cold cranking amps, and cost measured in dollars.

The value of $t \approx 3.23$. The *P*-value of 0.0015 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 cost and power. Batteries that cost more tend to have more power.



d) Since $R^2 = 25.2\%$, only 25.2% of the variability in power can be accounted for by cost. The residual standard deviation is 116 amps. That's pretty large, considering the range battery power is only about 400 amps. 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.

e) The equation of the line of best fit for these data points is: $\widehat{Power} = 384.594 + 4.14649(Cost)$, with power measured in cold cranking amps, and cost measured in dollars.

f) There are 31 degrees of freedom, so use $t_{30}^* = 1.697$ as a conservative estimate from the table.

$$b_1 \pm t_{n-2}^* \times SE(b_1) = 4.14649 \pm (1.697) \times 1.282 \approx (1.97,\ 6.32)$$

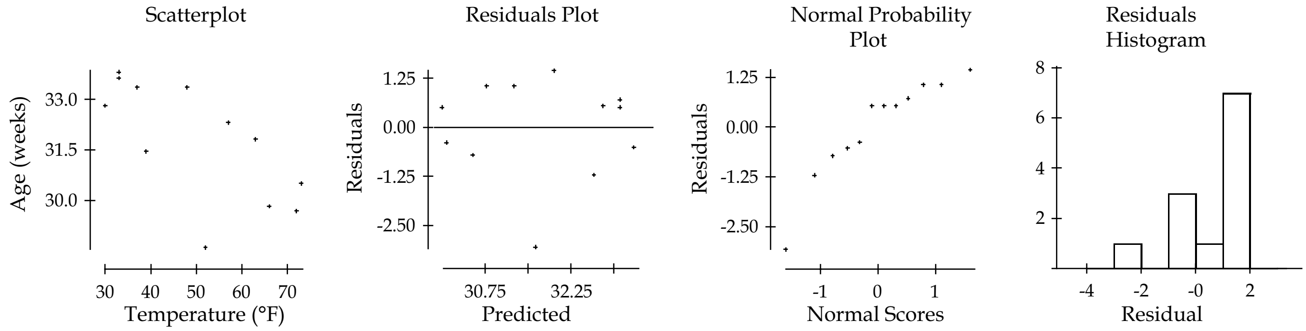
g) We are 95% confident that the mean power increases by between 1.97 and 6.32 cold cranking amps for each additional dollar in cost.

34. 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) H0: There is no linear relationship between 6-month temperature and crawling age. $(\beta_1 = 0)$

HA: There is a linear relationship between 6-month temperature and crawling age. $(\beta_1 \neq 0)$

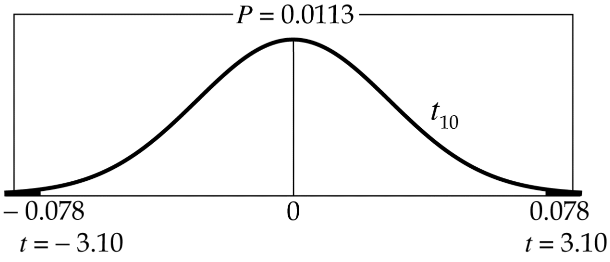


Straight enough condition: The scatterplot is straight enough to try linear regression.
Independence assumption: The residuals plot shows no pattern, but there may be an outlier. If the month of May were just one data point, it would be removed. However, since it represents the average crawling age of 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.

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: $A\hat{g}e = 35.6781 - 0.077739(Temp)$, 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.

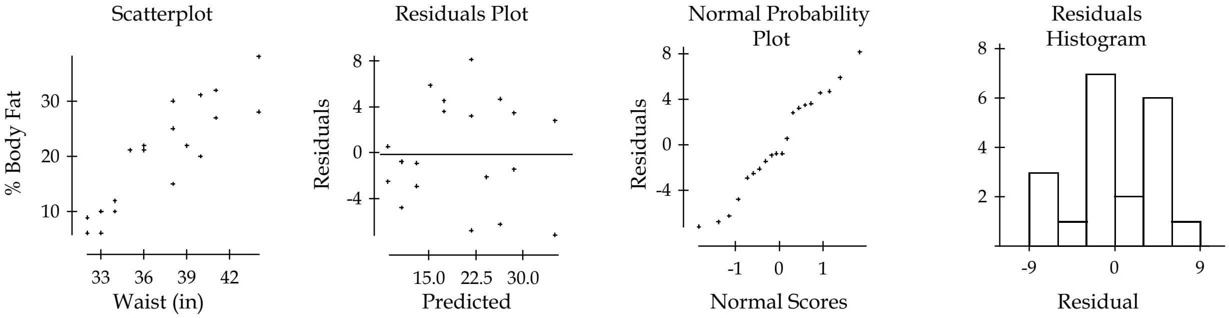
c) $b_1 \pm t^*_{n-2} \times SE(b_1) = -0.077739 \pm (2.228) \times 0.0251 \approx (-0.134, -0.022)$

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.

35. Body fat.

a) H0: There is no linear relationship between waist size and percent body fat. $(\beta_1 = 0)$

HA: There is a linear relationship between waist size and percent body fat. $(\beta_1 \neq 0)$



Straight enough condition: The scatterplot 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 Normal probability plot of residuals is straight, and the histogram of the residuals is unimodal and symmetric with no outliers.

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.

```
Dependent variable is:  Body Fat %
No Selector
R squared = 78.7%     R squared (adjusted) = 77.5%
s =  4.540  with  20 - 2 = 18  degrees of freedom

Source        Sum of Squares   df   Mean Square   F-ratio
Regression    1366.79           1     1366.79      66.3
Residual      370.960          18     20.6089

Variable      Coefficient   s.e. of Coeff   t-ratio     prob
Constant      -62.5573      10.16           -6.16       ≤ 0.0001
Waist (in)    2.22152        0.2728          8.14       ≤ 0.0001
```

The equation of the line of best fit for these data points is: $\%\widehat{BodyFat} = -62.5573 + 2.22152(Waist)$.

The value of $t \approx 8.14$. 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 waist size and percent body fat. People with larger waists tend to have a higher percentage of body fat.

b) The regression equation predicts that people with 40-inch waists will have $-62.5573 + 2.22152(40) = 26.3035\%$ body fat. The average waist size of the people sampled was approximately 37.05 inches.

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

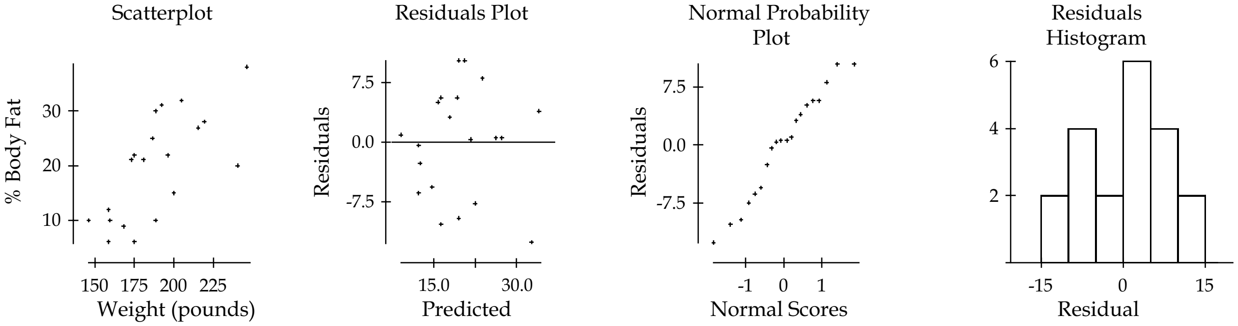
$$= 26.3035 \pm (2.101)\sqrt{0.2728^2 \cdot (40 - 37.05)^2 + \frac{4.54^2}{20}}$$

$$\approx (23.58,\ 29.03)$$

We are 95% confident that the mean percent body fat for people with 40-inch waists is between 23.58% and 29.03%.

36. Body fat, again.

a)



Straight enough condition: The scatterplot 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 Normal probability plot of residuals is straight, and the histogram of the residuals is unimodal and symmetric with no outliers.

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-interval.

```
Dependent variable is:  Body Fat %
No Selector
R squared = 48.5%     R squared (adjusted) = 45.7%
s =  7.049  with  20 - 2 = 18  degrees of freedom

Source        Sum of Squares   df   Mean Square   F-ratio
Regression    843.325           1     843.325      17.0
Residual      894.425          18      49.6903

Variable      Coefficient   s.e. of Coeff   t-ratio    prob
Constant      -27.3763      11.55           -2.37      0.0291
Weight (lb)   0.249874       0.0607          4.12      0.0006
```

The equation of the line of best fit for these data points is: $\%\widehat{BodyFat} = -27.3763 + 0.249874(Weight)$.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = 0.249874 \pm (1.734) \times 0.0607 \approx (0.145, 0.355)$$

b) We are 90% confident that the mean percent body fat increases between 1.45% and 3.55% for an additional 10 pounds in weight.

c) The regression equation predicts that a person weighing 165 pounds would have $-27.3763 + 0.249874(165) = 13.85291\%$ body fat. The average weight of the people sampled was 188.6 pounds.

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

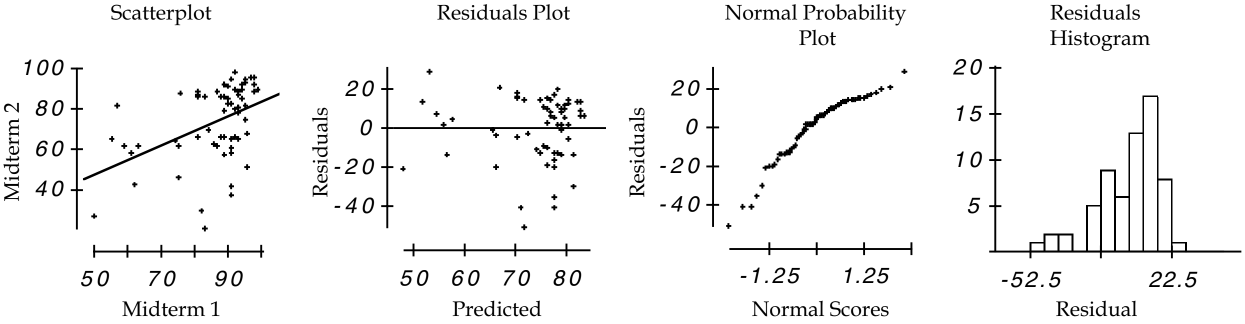
$$= 13.85291 \pm (2.101)\sqrt{0.0607^2 \cdot (165 - 188.6)^2 + \frac{7.049^2}{20} + 7.049^2} \approx (-1.61, 29.32)$$

We are 95% confident that a person weighing 165 pounds would have between 0% (–1.61%) and 29.32% body fat.

37. Grades.

a) The regression output is to the right. The model is: $\widehat{Midterm}2 = 12.005 + 0.721(Midterm1)$

```
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
```



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.

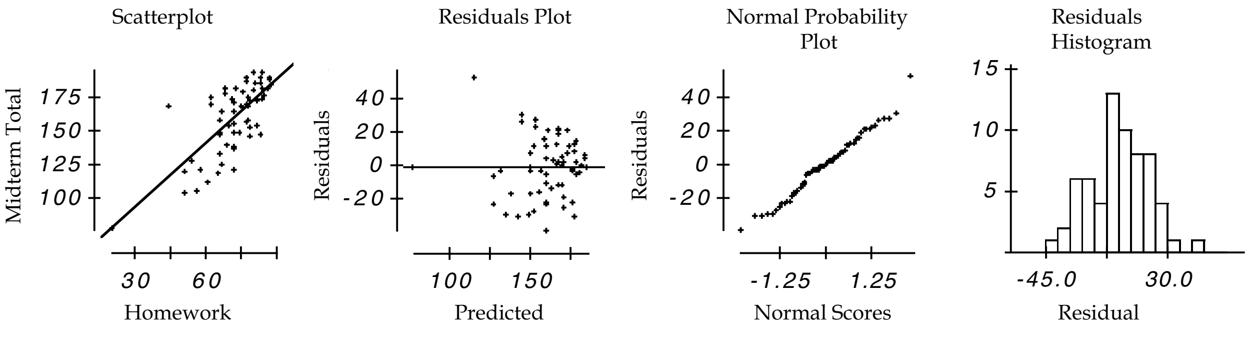
c) The student's reasoning is not valid. The R^2 value is only 0.199 and the value of *s* is 16.8 points. Although correlation between Midterm 1 and Midterm 2 may be statistically significant, it isn't of much practically use in predicting Midterm 2 scores. It's too weak.

38. Grades?

a) The regression output is to the right.
The model is:
$\widehat{MTtotal} = 46.062 + 1.580(Homework)$

```
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
```

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..


Scatterplot

Residuals Plot

Normal Probability Plot

Residuals Histogram

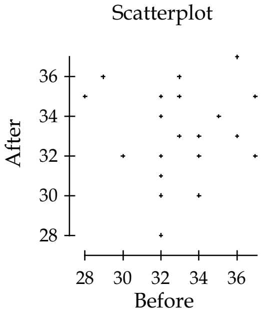
Since the conditions are met, linear regression is appropriate. The small *P*-value for the slope would indicate that the slope is statistically distinguishable from zero.

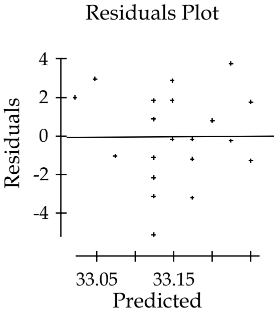
c) The R^2 value of 0.507 suggests that the overall relationship is fairly strong. However, this does not mean that midterm total is accurately predicted from homework scores. The error standard deviation of 18.30 indicates that a prediction in midterm total could easily be off by 20 to 30 points. If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

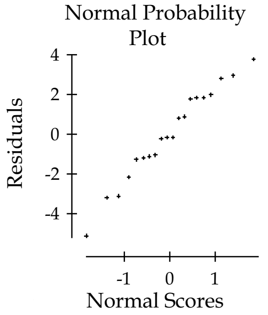
39. Strike two.

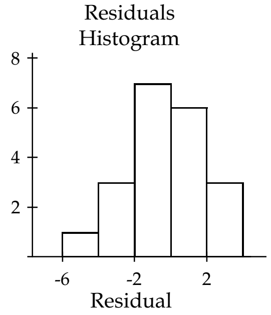
H_0: The effectiveness of the video is independent of the player's initial ability. $(\beta_1 = 0)$

H_A: The effectiveness of the video depends on the player's initial ability. $(\beta_1 \neq 0)$









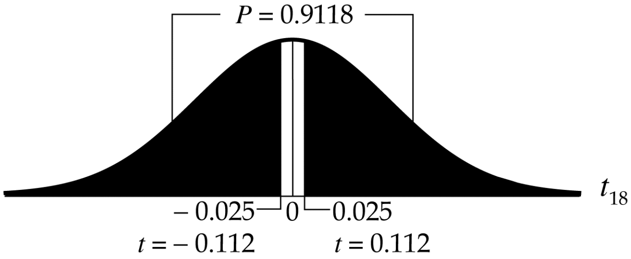
Straight enough condition: The scatterplot is straight enough to try linear regression, although it looks very scattered, and there doesn't appear to be any association.
Independence assumption: The residuals plot shows no pattern.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition, Outlier condition: The Normal probability plot of residuals is very straight, and the histogram of the residuals is unimodal and symmetric with no outliers.

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.

```
Dependent variable is:  After
No Selector
R squared = 0.1%     R squared (adjusted) = -5.5%
s =  2.386  with  20 - 2 = 18  degrees of freedom

Source        Sum of Squares   df   Mean Square   F-ratio
Regression    0.071912          1     0.071912      0.013
Residual      102.478          18     5.69323

Variable    Coefficient    s.e. of Coeff    t-ratio    prob
Constant    32.3161        7.439            4.34       0.0004
Before      0.025232       0.2245           0.112      0.9118
```

The equation of the line of best fit for these data points is: $\widehat{After} = 32.3161 + 0.025232(Before)$, where we are counting the number of strikes thrown before and after the training program.

The value of $t \approx 0.112$. The *P*-value of 0.9118 means that the association we see in the data is quite likely to occur by chance. We fail to reject the null hypothesis, and conclude that there is no evidence of a linear relationship between the number of strikes thrown before the training program and the number of strikes thrown after the program. The effectiveness of the program does not appear to depend on the initial ability of the player.



40. All the efficiency money can buy.

a) We'd like to know if there is a linear association between price and fuel efficiency in cars. We have data on 2004 model year cars, with information on highway MPG and retail price.

H$_0$: There is no linear relationship between highway 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. It shows a bend and it has an outlier. There is also some spreading from right to left, which violates the "Does the plot thicken?" condtion.

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

41. Education and mortality.

a) **Straight enough condition:** The scatterplot is straight enough to try linear regression.
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: $\widehat{Mortality} = 1493.26 - 49.9202(Education)$.

The value of $t \approx -6.24$. 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.

$$t = \frac{b_1 - \beta_1}{SE(b_1)}$$

$$t = \frac{-49.9202 - 0}{8.000}$$

$$t \approx -6.24$$

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, $t^*_{56} \approx 2.00327$.

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

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.

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

$$= 894.2176 \pm (2.003)\sqrt{8.00^2 \cdot (12 - 11.0328)^2 + \frac{47.92^2}{58}} \approx (874.239, 914.196)$$

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.

42. Property assessments.

a) **Straight enough condition:** The scatterplot is straight enough to try linear regression.
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. $(\beta_1 = 0)$

H$_A$: Larger houses have higher assessed values. $(\beta_1 > 0)$

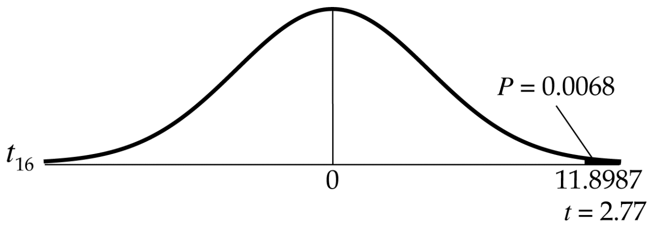
Since the conditions for inference are inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (18 – 2) = 16 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is: $Ass\hat{e}\$\$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

The value of $t \approx 2.77$.
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.

$$t = \frac{b_1 - \beta_1}{SE(b_1)}$$

$$t = \frac{11.8987 - 0}{4.290}$$

$$t \approx 2.77$$



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

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

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

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.

$$\hat{y}_v \pm t_{n-2}^* \sqrt{SE^2(b_1) \cdot (x_v - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$$

$$= 62096.07 \pm (2.120)\sqrt{4.290^2 \cdot (2100 - 2003.39)^2 + \frac{4682^2}{18} + 4682^2} \approx (51860, 72332)$$

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.

$$\hat{y}_v \pm t_{n-2}^* \sqrt{SE^2(b_1) \cdot (x_v - \bar{x})^2 + \frac{s_e^2}{n}}$$

$$= 62096.07 \pm (2.120)\sqrt{4.290^2 \cdot (2100 - 2003.39)^2 + \frac{4682^2}{18}} \approx (\$59{,}597, \$64{,}595)$$

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.