10.3: Inference for Regression and Correlation

How do you really say you have a correlation? Can you test to see if there really is a correlation? Of course, the answer is yes. The hypothesis test for correlation is as follows:

Hypothesis Test for Correlation:

1. State the random variables in words.
   x = independent variable
   y = dependent variable
2. State the null and alternative hypotheses and the level of significance
   $H_o : \rho = 0$ (There is no correlation)
   $H_A : \rho \neq 0$ (There is a correlation)
   or
   $H_A : \rho < 0$ (There is a negative correlation)
   or
   $H_A : \rho > 0$ (There is a postive correlation)
   Also, state your α level here.
3. State and check the assumptions for the hypothesis test
   The assumptions for the hypothesis test are the same assumptions for regression and correlation.
4. Find the test statistic and p-value
   $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$
   with degrees of freedom = df = n - 2
   p-value: Using the TI-83/84: tcdf(lower limit, upper limit, *df*)

   > **Note**
   >
   > If $H_A : \rho < 0$, then lower limit is -1E99 and upper limit is your test statistic. If $H_A : \rho > 0$, then lower limit is your test statistic and the upper limit is 1E99. If $H_A : \rho \neq 0$, then find the p-value for $H_A : \rho < 0$, and multiply by 2.

   Using R: pt(*t*, *df*)

   > **Note**
   >
   > If $H_A : \rho < 0$, then use pt(*t*, *df*), If $H_A : \rho > 0$, then use $1 - \text{pt}(t, df)$. If $H_A : \rho \neq 0$, then find the p-value for $H_A : \rho < 0$, and multiply by 2.

5. Conclusion
   This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value < α, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_o.
6. Interpretation
   This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

> **Note**
>
> The TI-83/84 calculator results give you the test statistic and the p-value. In R, the command for getting the test statistic and p-value is cor.test(independent variable, dependent variable, alternative = "less" or "greater"). Use less for $H_A : \rho < 0$, use greater for $H_A : \rho > 0$, and leave off this command for $H_A : \rho \neq 0$.

Example 10.3.1 Testing the claim of a linear correlation

Is there a positive correlation between beer's alcohol content and calories? To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer," 2011), and the data is in Example $10.3.1$. Test at the 5% level.

Solution

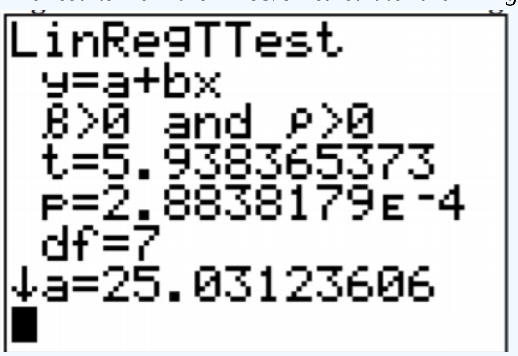
1. State the random variables in words.
   x = alcohol content in the beer
   y = calories in 12 ounce beer
2. State the null and alternative hypotheses and the level of significance.
   Since you are asked if there is a positive correlation, $\rho > 0$.
   $H_o : \rho = 0$
   $H_A : \rho > 0$
   $\alpha = 0.05$
3. State and check the assumptions for the hypothesis test.
   The assumptions for the hypothesis test were already checked in *Example* $10.3.2$.
4. Find the test statistic and p-value.
   The results from the TI-83/84 calculator are in *Figure* $10.3.1$.
   

   ***Figure* $10.3.1$** *: Results for Linear Regression Test on TI-83/84*
   Test statistic: $t \approx$ 5.938 and p-value: $p \approx 2.884 \times 10^{-4}$
   The results from R are
   cor.test(alcohol, calories, alternative = "greater")
   Pearson's product-moment correlation
   data: alcohol and calories
   t = 5.9384, df = 7, p-value = 0.0002884
   alternative hypothesis: true correlation is greater than 0
   95 percent confidence interval:
   0.7046161 1.0000000
   sample estimates:
   cor
   0.9134414
   Test statistic: $t \approx$ 5.9384 and p-value: $p \approx 0.0002884$
5. Conclusion
   Reject H_o since the p-value is less than 0.05.
6. Interpretation
   There is enough evidence to show that there is a positive correlation between alcohol content and number of calories in a 12-ounce bottle of beer.

Prediction Interval

Using the regression equation you can predict the number of calories from the alcohol content. However, you only find one value. The problem is that beers vary a bit in calories even if they have the same alcohol content. It would be nice to have a range instead of a single value. The range is called a prediction interval. To find this, you need to figure out how much error is in the estimate from the regression equation. This is known as the **standard error of the estimate**.

Definition

Standard Error of the Estimate
This is the sum of squares of the residuals

$$s_e = \sqrt{\frac{\sum (y-\hat{y})^2}{n-2}}$$

This formula is hard to work with, so there is an easier formula. You can also find the value from technology, such as the calculator.

$$s_e = \sqrt{\frac{SS_y - b^* SS_{xy}}{n-2}}$$

Example 10.3.2 finding the standard error of the estimate

Find the standard error of the estimate for the beer data. To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer," 2011), and the data are in Example 10.3.1.

Solution

x = alcohol content in the beer

y = calories in 12 ounce beer

Using the TI-83/84, the results are in *Figure* 10.3.2.

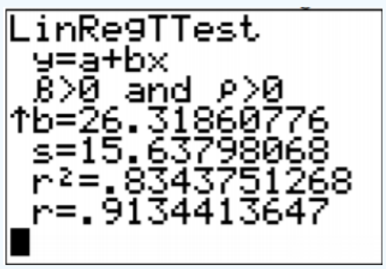


Figure 10.3.2: Results for Linear Regression Test on TI-83/84

The *s* in the results is the standard error of the estimate. So $s_e \approx 15.64$.

To find the standard error of the estimate in R, the commands are
lm.out = lm(dependent variable ~ independent variable) – this defines the linear model with a name so you can use it later. Then
summary(lm.out) – this will produce most of the information you need for a regression and correlation analysis. In fact, the only thing R doesn't produce with this command is the correlation coefficient. Otherwise, you can use the command to find the regression equation, coefficient of determination, test statistic, p-value for a two-tailed test, and standard error of the estimate.

The results from R are
lm.out=lm(calories~alcohol)
summary(lm.out)

Call:
lm(formula = calories ~ alcohol)

Residuals:

Min	1Q	Median	3Q	Max
-30.253	-1.624	2.744	9.271	14.271

Coefficients:

	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	25.031	24.999	1.001	0.350038
alcohol	26.319	4.432	5.938	0.000577

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.64 on 7 degrees of freedom
Multiple R-squared: 0.8344, Adjusted R-squared: 0.8107
F-statistic: 35.26 on 1 and 7 DF, p-value: 0.0005768

From this output, you can find the y-intercept is 25.031, the slope is 26.319, the test statistic is t = 5.938, the p-value for the two-tailed test is 0.000577. If you want the p-value for a one-tailed test, divide this number by 2. The standard error of the estimate is the residual standard error and is 15.64. There is some information in this output that you do not need.

If you want to know how to calculate the standard error of the estimate from the formula, refer to Example $10.3.3$.

Example $10.3.3$ finding the standard error of the estimate from the formula

Find the standard error of the estimate for the beer data using the formula. To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer," 2011), and the data are in Example $10.3.1$.

Solution

x = alcohol content in the beer

y = calories in 12 ounce beer

From Example $10.3.3$:

$$\begin{aligned} SS_y &= \sum(y-\bar{y})^2 = 10335.56 \\ SS_{xy} &= \sum(x-\bar{x})(y-\bar{y}) = 327.6666 \\ n &= 9 \\ b &= 26.3 \end{aligned}$$

The standard error of the estimate is

$$\begin{aligned} s_e &= \sqrt{\frac{SS_y - b^* SS_{xy}}{n-2}} \\ &= \sqrt{\frac{10335.56 - 26.3(327.6666)}{9-2}} \\ &= 15.67 \end{aligned}$$

Prediction Interval for an Individual *y*

Given the fixed value x_0, the prediction interval for an individual *y* is

$\hat{y} - E < y < \hat{y} + E$

where

$\hat{y} = a + bx$

$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_x}}$

$df = n - 2$

Note

To find $SS_x = \sum(x - \overline{x})^2$ remember, the standard derivation formula from chapter 3 $s_x = \sqrt{\frac{\sum(x - \overline{x})^2}{n-1}}$

So, $s_x = \sqrt{\frac{SS_x}{n-1}}$

Now solve for SS_x

$SS_x = s_x^2(n-1)$

You can get the standard deviation from technology.
R will produce the prediction interval for you. The commands are (Note you probably already did the lm.out command. You do not need to do it again.)
lm.out = lm(dependent variable ~ independent variable) – calculates the linear model
predict(lm.out, newdata=list(independent variable = value), interval="prediction", level=C) – will compute a prediction interval for the independent variable set to a particular value (put that value in place of the word value), at a particular C level (given as a decimal)

Example 10.3.4 find the prediction interval

Find a 95% prediction interval for the number of calories when the alcohol content is 6.5% using a random sample taken of beer's alcohol content and calories ("Calories in beer," 2011). The data are in Example 10.3.1.

Solution

x = alcohol content in the beer

y = calories in 12 ounce beer

Computing the prediction interval using the TI-83/84 calculator:
From Example 10.3.2

$$\begin{aligned}\hat{y} &= 25.0 + 26.3x \\ x_o &= 6.50 \\ \hat{y} &= 25.0 + 26.3(6.50) = 196 \text{ calories}\end{aligned}$$

From Example #10.3.2

$$s_e \approx 15.64$$

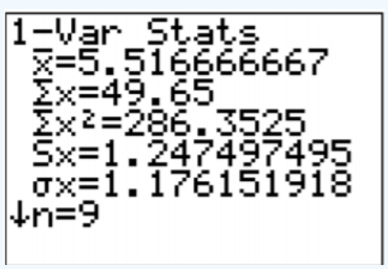


Figure 10.3.3: Results of 1-Var Stats on TI-83/84

$$\begin{aligned}\overline{x} &= 5.517 \\ s_x &= 1.247497495 \\ n &= 9\end{aligned}$$

Now you can find

$$\begin{aligned}SS_x &= s_x^2(n-1) \\ &= (1.247497495)^2(9-1) \\ &= 12.45 \\ df &= n - 2 = 9 - 2 = 7\end{aligned}$$

Now look in table A.2. Go down the first column to 7, then over to the column headed with 95%.

$$t_c = 2.365$$

$$\begin{aligned} E &= t_c s_e \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_x}} \\ &= 2.365(15.64)\sqrt{1 + \frac{1}{9} + \frac{(6.50 - 5.517)^2}{12.45}} \\ &= 40.3 \end{aligned}$$

Prediction interval is

$$\begin{aligned} &\hat{y} - E < y < \hat{y} + E \\ &196 - 40.3 < y < 196 + 40.3 \\ &155.7 < y < 236.3 \end{aligned}$$

Computing the prediction interval using R:

predict(lm.out, newdata=list(alcohol=6.5), interval = "prediction", level=0.95)

	fit	lwr	upr
1	196.1022	155.7847	236.4196

fit = $\hat{y}$ when x = 6.5%. lwr = lower limit of prediction interval. upr = upper limit of prediction interval. So the prediction interval is $155.8 < y < 236.4$.

Statistical interpretation: There is a 95% chance that the interval $155.8 < y < 236.4$ contains the true value for the calories when the alcohol content is 6.5%.

Real world interpretation: If a beer has an alcohol content of 6.50% then it has between 156 and 236 calories.

Example 10.3.5 Doing a correlation and regression analysis using the ti-83/84

Example $10.3.1$ contains randomly selected high temperatures at various cities on a single day and the elevation of the city.

Table 10.3.1: Temperatures and Elevation of Cities on a Given Day

Elevation (in feet)	7000	4000	6000	3000	7000	4500	5000
Temperature (°F)	50	60	48	70	55	55	60

a. State the random variables.
b. Find a regression equation for elevation and high temperature on a given day.
c. Find the residuals and create a residual plot.
d. Use the regression equation to estimate the high temperature on that day at an elevation of 5500 ft.
e. Use the regression equation to estimate the high temperature on that day at an elevation of 8000 ft.
f. Between the answers to parts d and e, which estimate is probably more accurate and why?
g. Find the correlation coefficient and coefficient of determination and interpret both.
h. Is there enough evidence to show a negative correlation between elevation and high temperature? Test at the 5% level.
i. Find the standard error of the estimate.
j. Using a 95% prediction interval, find a range for high temperature for an elevation of 6500 feet.

Solution

a. x = elevation

y = high temperature

b.

a. A random sample was taken as stated in the problem.

b. The distribution for each high temperature value is normally distributed for every value of elevation.

i. Look at the scatter plot of high temperature versus elevation.

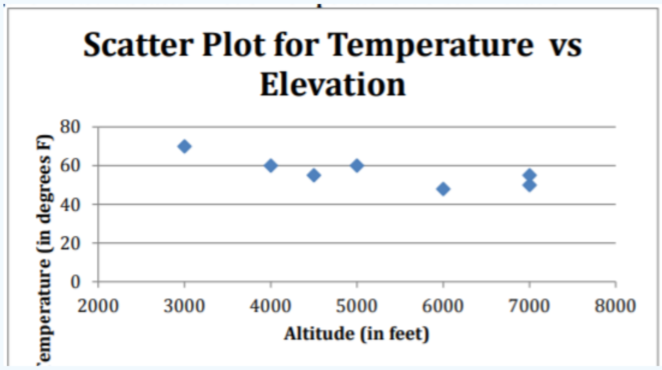


Figure $\mathbf{10.3.4}$: *Scatter Plot of Temperature Versus Elevation*

The scatter plot looks fairly linear.

ii. There are no points that appear to be outliers.

iii. The residual plot for temperature versus elevation appears to be fairly random. (See *Figure* 10.3.7.) It appears that the high temperature is normally distributed.

All calculations computed using the TI-83/84 calculator.

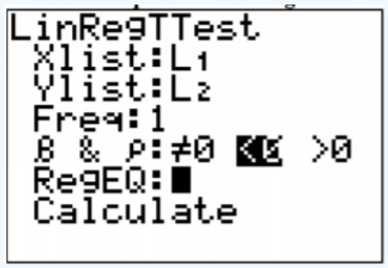


Figure 10.3.5: Setup for Linear Regression on TI-83/84 Calculator

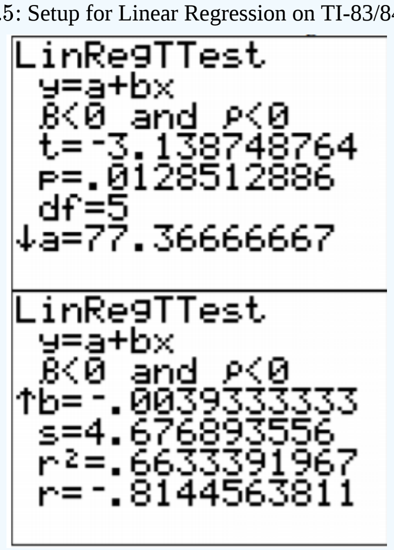


Figure 10.3.6: Results for Linear Regression on TI-83/84 Calculator

$$\hat{y} = 77.4 - 0.0039x$$

c.

Table 10.3.2: Residuals for Elevation vs. Temperature Data

x	y	$\hat{y}$	$y - \hat{y}$
7000	50	50.1	-0.1

x	y	$\hat{y}$	$y-\hat{y}$
4000	60	61.8	-1.8
6000	48	54.0	-6.0
3000	70	65.7	4.3
7000	55	50.1	4.9
4500	55	59.85	-4.85
5000	60	57.9	2.1

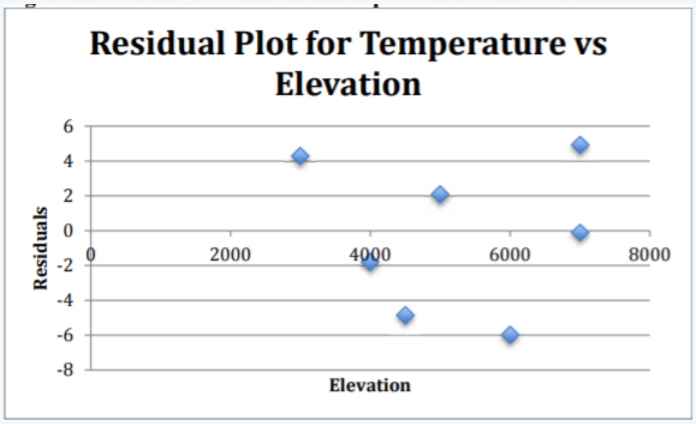


Figure 10.3.7: Residual Plot for Temperature vs. Election

The residuals appear to be fairly random.

d.

$x_o = 5500$
$\hat{y} = 77.4 - 0.0039(5500) = 55.95^{\circ}F$

e.

$x_o = 8000$
$\hat{y} = 77.4 - 0.0039(8000) = 46.2^{\circ}F$

f. Part d is more accurate, since it is interpolation and part e is extrapolation.

g. From *Figure* $10.3.6$, the correlation coefficient is $r \approx -0.814$, which is moderate to strong negative correlation.

From *Figure* $10.3.6$, the coefficient of determination is $r^2 \approx 0.663$, which means that 66.3% of the variability in high temperature is explained by the linear model. The other 33.7% is explained by other variables such as local weather conditions.

h.

1. State the random variables in words.
   x = elevation
   y = high temperature
2. State the null and alternative hypotheses and the level of significance
   $H_o : \rho = 0$
   $H_A : \rho < 0$
   $\alpha = 0.05$
3. State and check the assumptions for the hypothesis test The assumptions for the hypothesis test were already checked part b.
4. Find the test statistic and p-value
   From *Figure* $10.3.6$,
   Test statistic:

$t \approx -3.139$
p-value:
$p \approx 0.0129$

5. Conclusion
   Reject H_o since the p-value is less than 0.05.
6. Interpretation
   There is enough evidence to show that there is a negative correlation between elevation and high temperatures.

i. From *Figure* $10.3.6$,

$s_e \approx 4.677$

j. $\hat{y} = 77.4 - 0.0039(6500) \approx 52.1^{\circ} F$

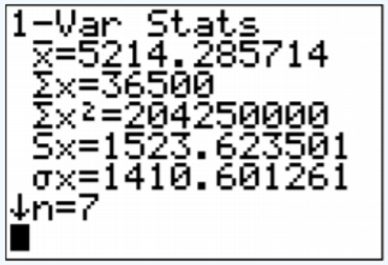


Figure 10.3.8: Results of 1-Var Stats on TI-83/84

$\bar{x} = 5214.29$
$s_x = 1523.624$
$n = 7$

Now you can find

$$\begin{aligned} SS_x &= s_x^2(n-1) \\ &= (1523.623501)^2(7-1) \\ &= 13928571.43 \\ df &= n-2 = 7-2 = 5 \end{aligned}$$

Now look in table A.2. Go down the first column to 5, then over to the column headed with 95%.

$t_c = 2.571$

So

$$\begin{aligned} E &= t_c s_e \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_x}} \\ &= 2.571(4.677)\sqrt{1 + \frac{1}{7} + \frac{(6500 - 5214.29)^2}{13928571.43}} \\ &= 13.5 \end{aligned}$$

Prediction interval is

$$\begin{aligned} &\hat{y} - E < y < \hat{y} + E \\ &52.1 - 13.5 < y < 52.1 + 13.5 \\ &38.6 < y < 65.6 \end{aligned}$$

Statistical interpretation: There is a 95% chance that the interval $38.6 < y < 65.6$ contains the true value for the temperature at an elevation of 6500 feet.

Real world interpretation: A city of 6500 feet will have a high temperature between 38.6°F and 65.6°F. Though this interval is fairly wide, at least the interval tells you that the temperature isn't that warm.

Example 10.3.6 doing a correlation and regression analysis using r

Example 10.3.1 contains randomly selected high temperatures at various cities on a single day and the elevation of the city.

a. State the random variables.
b. Find a regression equation for elevation and high temperature on a given day.
c. Find the residuals and create a residual plot.
d. Use the regression equation to estimate the high temperature on that day at an elevation of 5500 ft.
e. Use the regression equation to estimate the high temperature on that day at an elevation of 8000 ft.
f. Between the answers to parts d and e, which estimate is probably more accurate and why?
g. Find the correlation coefficient and coefficient of determination and interpret both.
h. Is there enough evidence to show a negative correlation between elevation and high temperature? Test at the 5% level.
i. Find the standard error of the estimate.
j. Using a 95% prediction interval, find a range for high temperature for an elevation of 6500 feet.

Solution

a. x = elevation

y = high temperature

b.

a. A random sample was taken as stated in the problem.
b. The distribution for each high temperature value is normally distributed for every value of elevation.
    i. Look at the scatter plot of high temperature versus elevation.
    R command: plot(elevation, temperature, main="Scatter Plot for Temperature vs Elevation", xlab="Elevation (feet)", ylab="Temperature (degrees F)", ylim=c(0,80))

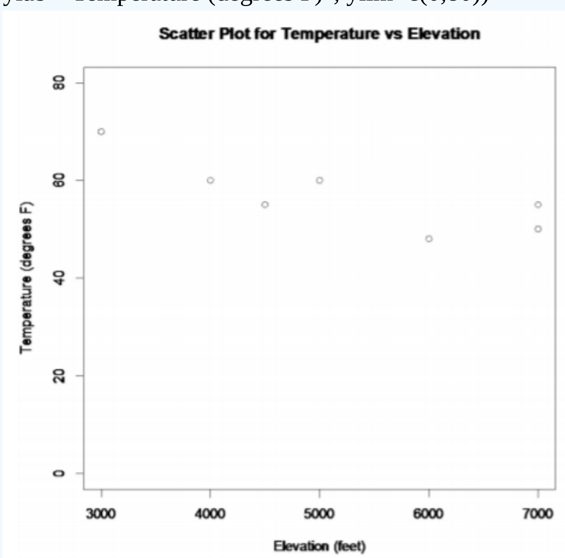


***Figure* 10.3.9** *: Scatter Plot of Temperature Versus Elevation*

The scatter plot looks fairly linear.

    ii. The residual plot for temperature versus elevation appears to be fairly random. (See *Figure 10.3.10*.)
    It appears that the high temperature is normally distributed.

Using R:
Commands:
lm.out=lm(temperature ~ elevation)
summary(lm.out)

Output:
Call:
lm(formula = temperature ~ elevation)

Residuals:

1	2	3	4	5	6	7
0.1667	-1.6333	-5.7667	4.4333	5.1667	-4.6667	2.3000

Coefficients:

	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	77.36667	6.769182	11.429	$8.98e-05***$
elevation	-0.003933	0.001253	-3.139	$0.0257*$

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 4.677 on 5 degrees of freedom
Multiple R-squared: 0.6633, Adjusted R-squared: 0.596
F-statistic: 9.852 on 1 and 5 DF, p-value: 0.0257

From the output you can see the slope = -0.0039 and the y-intercept = 77.4. So the regression equation is:

$\hat{y} = 77.4 - 0.0039x$

c. R command: (notice these are also in the summary(lm.out) output, but if you have too many data points, then R only gives a numerical summary of the residuals.)

residuals(lm.out)

1	2	3	4	5	6	7
0.1666667	-1.63333333	-5.766667	4.43333333	5.1666667	-4.66666667	2.3000000

So for the first x of 7000, the residual is approximately 0.1667. This means if you find the $\hat{y}$ for when x is 7000 and then subtract this answer from the y value of 50 that was measured, you would obtain 0.1667. Similar process is computed for the other residual values.

To plot the residuals, the R command is

plot(elevation, residuals(lm.out), main="Residual Plot for Temperautre vs Elevation", xlab="Elevation (feet)", ylab="Residuals")
abline(0,0)

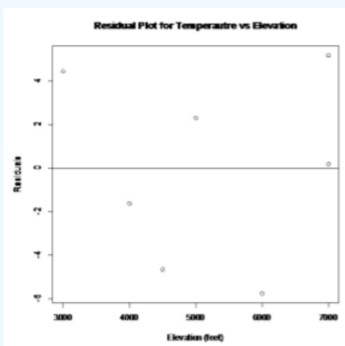


Figure 10.3.10: Residual Plot for Temperature vs. Elevation

The residuals appear to be fairly random.

d. $x_o = 5500$
$\hat{y} = 77.4 - 0.0039(5500) = 55.95^{\circ}F$

e. $x_o = 8000$
$\hat{y} = 77.4 - 0.0039(8000) = 46.2^{\circ}F$

f. Part d is more accurate, since it is interpolation and part e is extrapolation.

g. The R command for the correlation coefficient is
cor(elevation, temperature)
[1] -0.8144564

So, $r \approx -0.814$, which is moderate to strong negative correlation.

From summary(lm.out), the coefficient of determination is the Multiple R-squared.

So $r^2 \approx 0.663$, which means that 66.3% of the variability in high temperature is explained by the linear model. The other 33.7% is explained by other variables such as local weather conditions.

h.

1. State the random variables in words.
   x = elevation
   y = high temperature
2. . State the null and alternative hypotheses and the level of significance
   $H_o : \rho = 0$
   $H_A : \rho < 0$
   $\alpha = 0.05$
3. State and check the assumptions for the hypothesis test.
   The assumptions for the hypothesis test were already checked part b.
4. Find the test statistic and p-value
   The R command is cor.test(elevation, temperature, alternative = "less")
   Pearson's product-moment correlation
   data: elevation and temperature
   t = -3.1387, df = 5, p-value = 0.01285
   alternative hypothesis: true correlation is less than 0
   95 percent confidence interval:
   -1.0000000 -0.3074247
   sample estimates:
   cor
   -0.8144564
   Test statistic: $t \approx -3.1387$ and p-value: $p \approx 0.01285$
5. Conclusion
   Reject H_o since the p-value is less than 0.05.
6. Interpretation
   There is enough evidence to show that there is a negative correlation between elevation and high temperatures.

i. From summary(lm.out), Residual standard error: 4.677.

So, $s_e \approx 4.677$

j. R command is predict(lm.out, newdata=list(elevation = 6500), interval = "prediction", level=0.95)

	fit	lwr	upr
1	51.8	38.29672	65.30328

So when x = 6500 feet, $\hat{y} = 51.8^{\circ} F$ and $38.29672 < y < 65.30328$.

Statistical interpretation: There is a 95% chance that the interval $38.3 < y < 65.3$ contains the true value for the temperature at an elevation of 6500 feet.

Real world interpretation: A city of 6500 feet will have a high temperature between 38.3°F and 65.3°F. Though this interval is fairly wide, at least the interval tells you that the temperature isn't that warm.

Homework

Exercise 10.3.1

For each problem, state the random variables. The data sets in this section are in the homework for section 10.1 and were also used in section 10.2. If you removed any data points as outliers in the other sections, remove them in this sections homework too.

1. When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of their metacarpal bone one (in cm) were collected and are in Example $10.3.5$ ("Prediction of height," 2013).
   a. Test at the 1% level for a positive correlation between length of metacarpal bone one and height of a person.
   b. Find the standard error of the estimate.
   c. Compute a 99% prediction interval for height of a person with a metacarpal length of 44 cm.
2. Example $10.3.6$ contains the value of the house and the amount of rental income in a year that the house brings in ("Capital and rental," 2013).
   a. Test at the 5% level for a positive correlation between house value and rental amount.
   b. Find the standard error of the estimate.
   c. Compute a 95% prediction interval for the rental income on a house worth $230,000.
3. The World Bank collects information on the life expectancy of a person in each country ("Life expectancy at," 2013) and the fertility rate per woman in the country ("Fertility rate," 2013). The data for 24 randomly selected countries for the year 2011 are in Example $10.3.7$.
   a. Test at the 1% level for a negative correlation between fertility rate and life expectancy.
   b. Find the standard error of the estimate.
   c. Compute a 99% prediction interval for the life expectancy for a country that has a fertility rate of 2.7.
4. The World Bank collected data on the percentage of GDP that a country spends on health expenditures ("Health expenditure," 2013) and also the percentage of women receiving prenatal care ("Pregnant woman receiving," 2013). The data for the countries where this information is available for the year 2011 are in Example $10.3.8$.
   a. Test at the 5% level for a correlation between percentage spent on health expenditure and the percentage of women receiving prenatal care.
   b. Find the standard error of the estimate.
   c. Compute a 95% prediction interval for the percentage of woman receiving prenatal care for a country that spends 5.0 % of GDP on health expenditure.
5. The height and weight of baseball players are in Example $10.3.9$ ("MLB heightsweights," 2013).
   a. Test at the 5% level for a positive correlation between height and weight of baseball players.
   b. Find the standard error of the estimate.
   c. Compute a 95% prediction interval for the weight of a baseball player that is 75 inches tall.
6. Different species have different body weights and brain weights are in Example $10.3.10$ ("Brain2bodyweight," 2013).
   a. Test at the 1% level for a positive correlation between body weights and brain weights.
   b. Find the standard error of the estimate.
   c. Compute a 99% prediction interval for the brain weight for a species that has a body weight of 62 kg.
7. A random sample of beef hotdogs was taken and the amount of sodium (in mg) and calories were measured. ("Data hotdogs," 2013) The data are in Example $10.3.11$.
   a. Test at the 5% level for a correlation between amount of calories and amount of sodium.
   b. Find the standard error of the estimate.
   c. Compute a 95% prediction interval for the amount of sodium a beef hotdog has if it is 170 calories.
8. Per capita income in 1960 dollars for European countries and the percent of the labor force that works in agriculture in 1960 are in Example $10.3.12$ ("OECD economic development," 2013).
   a. Test at the 5% level for a negative correlation between percent of labor force in agriculture and per capita income.
   b. Find the standard error of the estimate.
   c. Compute a 90% prediction interval for the per capita income in a country that has 21 percent of labor in agriculture.
9. Cigarette smoking and cancer have been linked. The number of deaths per one hundred thousand from bladder cancer and the number of cigarettes sold per capita in 1960 are in Example $10.3.13$ ("Smoking and cancer," 2013).
   a. Test at the 1% level for a positive correlation between cigarette smoking and deaths of bladder cancer.
   b. Find the standard error of the estimate.
   c. Compute a 99% prediction interval for the number of deaths from bladder cancer when the cigarette sales were 20 per capita.

10. The weight of a car can influence the mileage that the car can obtain. A random sample of cars weights and mileage was collected and are in Example $10.3.14$("Passenger car mileage," 2013).
    a. Test at the 5% level for a negative correlation between the weight of cars and mileage.
    b. Find the standard error of the estimate.
    c. Compute a 95% prediction interval for the mileage on a car that weighs 3800 pounds.

Answer

For hypothesis test just the conclusion is given. See solutions for entire answer.

1. a. Reject Ho, b. $s_e \approx 4.559$, c. $151.3161\text{cm} < y < 187.3859\text{cm}$

3. a. Reject Ho, b. $s_e \approx 3.204$, c. $62.945 \text{ years } < y < 81.391\text{years}$

5. a. Reject Ho, b. $s_e \approx 15.33$, c. $176.02 \text{ inches } < y < 240.92\text{inches}$

7. a. Reject Ho, b. $s_e \approx 48.58$, c. $348.46\text{mg} < y < 559.38\text{mg}$

9. a. Reject Ho, b. $s_e \approx 0.6838$, c. $1.613 \text{ hundred thousand } < y < 5.432 \text{ hundred thousand}$

Data Source:

Brain2bodyweight. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index...ain2BodyWeight

Calories in beer, beer alcohol, beer carbohydrates. (2011, October 25). Retrieved from www.beer100.com/beercalories.htm

Capital and rental values of Auckland properties. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/rentcap.html

Data hotdogs. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index...D_Data_HotDogs

Fertility rate. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SP.DYN.TFRT.IN

Health expenditure. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS

Life expectancy at birth. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SP.DYN.LE00.IN

MLB heightsweights. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index...HeightsWeights

OECD economic development. (2013, December 04). Retrieved from lib.stat.cmu.edu/DASL/Datafiles/oecdat.html

Passenger car mileage. (2013, December 04). Retrieved from lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html

Prediction of height from metacarpal bone length. (2013, September 26). Retrieved from http://www.statsci.org/data/general/stature.html

Pregnant woman receiving prenatal care. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SH.STA.ANVC.ZS

Smoking and cancer. (2013, December 04). Retrieved from lib.stat.cmu.edu/DASL/Datafil...cancerdat.html