Probability & Statistics Workbook Solutions

Regression

krista king
MATH

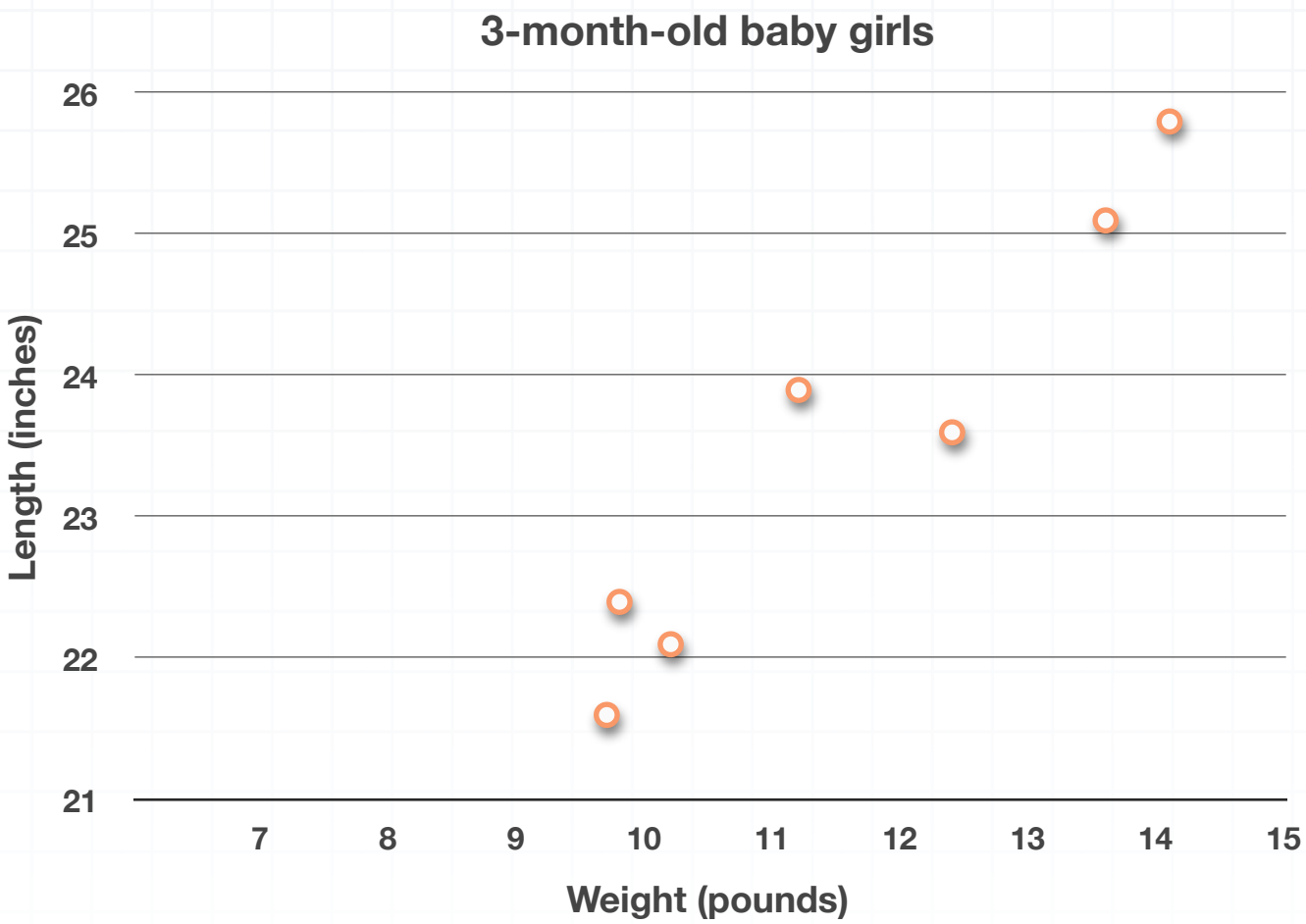
SCATTERPLOTS AND REGRESSION

■ 1. The table gives weight in pounds and length in inches for 3-month-old baby girls. Graph the points from the table in a scatterplot and describe the trend.

Weight (lbs)	Length (in)
9.7	21.6
10.2	22.1
12.4	23.6
13.6	25.1
9.8	22.4
11.2	23.9
14.1	25.8

Solution:

Sketch the scatterplot.



The points rise from left to right and are fairly linear. We can say that there is a strong positive linear correlation between the points. There do not appear to be any outliers in the data.

2. The following values have been computed for a data set of 14 points. Calculate the line of best fit.

$$\sum x = 86$$

$$\sum y = 89.7$$

$$\sum xy = 680.46$$

$$\sum x^2 = 654.56$$

Solution:

We're told that there are 14 items in the data set, so $n = 14$.

To find the line of best fit, we need its slope and y-intercept. The slope is given by

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$b = \frac{14(680.46) - (86)(89.7)}{14(654.56) - (86)^2}$$

$$b = \frac{9{,}526.44 - 7{,}714.2}{9{,}163.84 - 7{,}396}$$

$$b = \frac{1{,}812.24}{1{,}767.84}$$

$$b \approx 1.0251$$

The y-intercept is given by

$$a = \frac{\sum y - b\sum x}{n}$$

$$a = \frac{89.7 - 1.0251(86)}{14}$$

$$a = \frac{89.7 - 88.1599}{14}$$

$$a = \frac{1.5401}{14}$$

$$a \approx 0.1100$$

So the line of best fit is

$$\hat{y} = bx + a$$

$$\hat{y} = 1.0251x + 0.1100$$

■ 3. For the data set given in the table, calculate each of the following values:

$$n,\ \sum x,\ \sum y,\ \sum xy,\ \sum x^2,\ \left(\sum x\right)^2$$

Month	1	2	3	4	5	6	7	8	9	10	11	12
Temperature	73	73	75	75	77	79	79	81	81	81	77	75

Solution:

Expand the original table to calculate the values.

Month, x	Temperature, y	xy	x^2
1	73	$1(73)=73$	$1^2=1$
2	73	$2(73)=146$	$2^2=4$
3	75	$3(75)=225$	$3^2=9$
4	75	$4(75)=300$	$4^2=16$
5	77	$5(77)=385$	$5^2=25$
6	79	$6(79)=474$	$6^2=36$
7	79	$7(79)=553$	$7^2=49$
8	81	$8(81)=648$	$8^2=64$
9	81	$9(81)=729$	$9^2=81$
10	81	$10(81)=810$	$10^2=100$
11	77	$11(77)=847$	$11^2=121$
12	75	$12(75)=900$	$12^2=144$

Summing the first column gives $\sum x = 78$

Summing the second column gives $\sum y = 926$

Summing the third column gives $\sum xy = 6{,}090$

Summing the fourth column gives $\sum x^2 = 650$

Squaring the sum from the first column gives $\left(\sum x\right)^2 = 78^2 = 6{,}084$

4. Use the Average Global Sea Surface Temperatures data shown in the table to create a line of best fit for the data. Consider 1910 as year 10. Use the equation to predict the average global sea surface temperature in the year 2050.

Year	Temperature, F
1910	-1.11277
1920	-0.71965
1930	-0.58358
1940	-0.17977
1950	-0.55318
1960	-0.30358
1970	-0.30863
1980	0.077197
1990	0.274842
2000	0.232502
2010	0.612718

Solution:

Start by expanding the table.

Year	Temperature, F	xy	x^2
10	-1.11277	-11.1277	100
20	-0.71965	-14.393	400
30	-0.58358	-17.5074	900
40	-0.17977	-7.1908	1,600
50	-0.55318	-27.659	2,500
60	-0.30358	-18.2148	3,600
70	-0.30863	-21.6041	4,900
80	0.077197	6.17576	6,400
90	0.274842	24.73578	8,100
100	0.232502	23.2502	10,000
110	0.612718	67.39898	12,100

Summing the first column gives $\sum x = 660$

Summing the second column gives $\sum y = -2.5639$

Summing the third column gives $\sum xy = 3.86392$

Summing the fourth column gives $\sum x^2 = 50{,}600$

Squaring the sum from the first column gives

$$\left(\sum x\right)^2 = 660^2 = 435{,}600$$

To find the regression line for the data, we need the slope and y-intercept of the line. The slope is

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$b = \frac{11(3.86392) - (660)(-2.5639)}{11(50{,}600) - 435{,}600}$$

$$b = \frac{42.50312 + 1{,}692.174}{556{,}600 - 435{,}600}$$

$$b = \frac{1{,}734.67712}{121{,}600}$$

$$b \approx 0.0143$$

The y-intercept is

$$a = \frac{\sum y - b\sum x}{n}$$

$$a = \frac{-2.5639 - 0.0143(660)}{11}$$

$$a = \frac{-2.5639 - 9.415188}{11}$$

$$a = \frac{-11.979088}{11}$$

$$a \approx -1.0890$$

Then the equation of the trend line is

$$\hat{y} = bx + a$$

$$\hat{y} = 0.0143x - 1.0890$$

To predict average global sea surface temperature in 2050, we'll need to plug 150 into this equation.

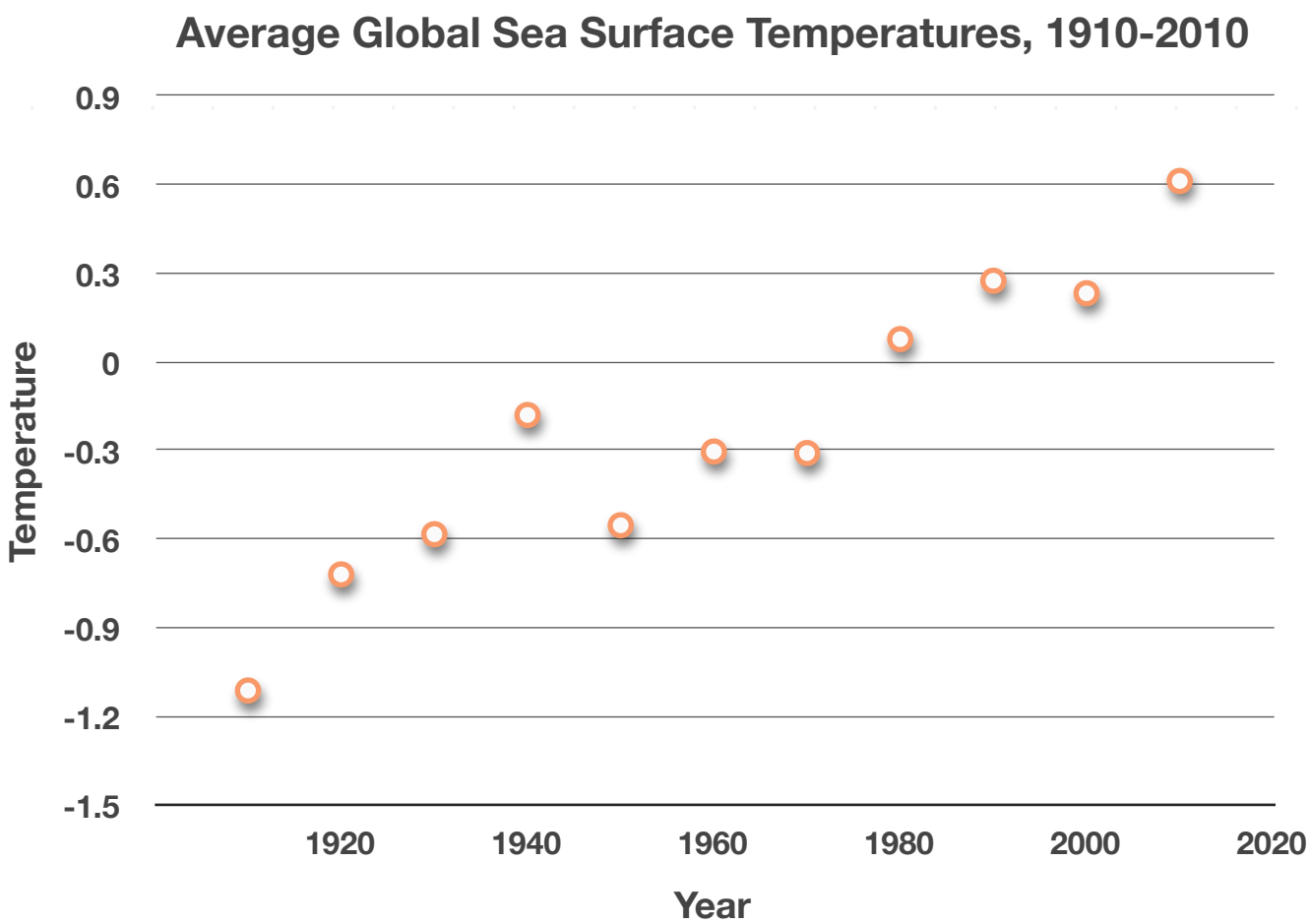
$$\hat{y} = 0.0143(150) - 1.0890$$

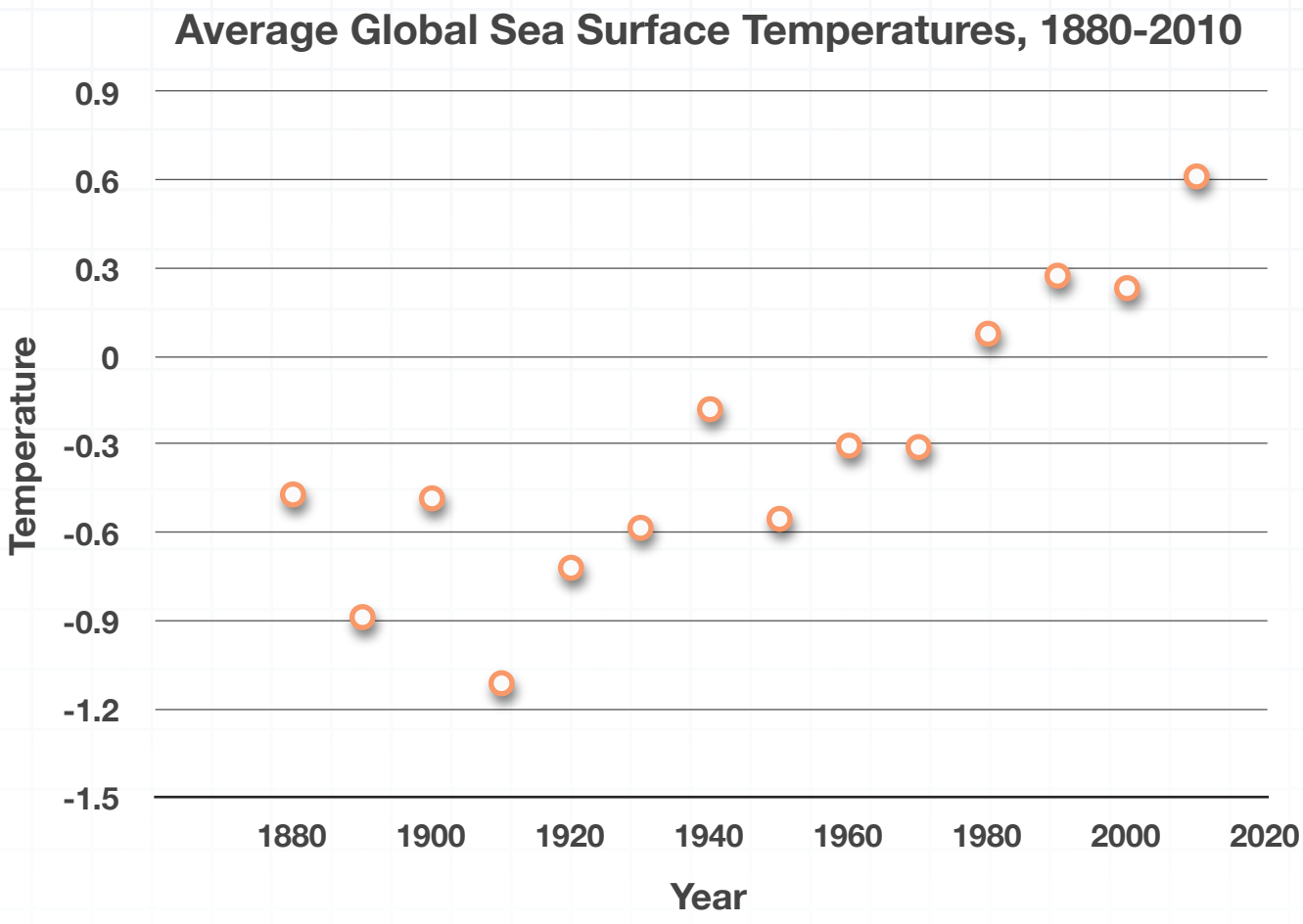
$$\hat{y} = 2.145 - 1.0890$$

$$\hat{y} = 1.056$$

So the predicted sea surface temperature in 2050 is about $1.056°$ F.

■ 5. Compare the scatterplots. The second graph includes extra data starting in 1880. How does this compare to the plot that only shows 1910 to 2010? Explain trends in the data, and how the regression line changes by adding in these extra points. Which trend line would be best for predicting the temperature in 2050?





Solution:

Adding in these extra three points make the graph from 1880 to 2010 appear more scattered and not as linear as the graph that only includes the points from 1910 to 2010.

Both data sets have a positive correlation because the general trend of the scatterplot is to increase as we move from left to right, but we might consider graphs that are exponential in shape instead of linear. If we use a line of best fit for the data from 1880 to 2010, it might not be as accurate as a line predicting only the points from 1910 to 2010.

In other words, the best fit line for 1880 to 2010 would have a weaker correlation than the line for 1910 to 2010, because the additional points to the left of the graph are more spread out.

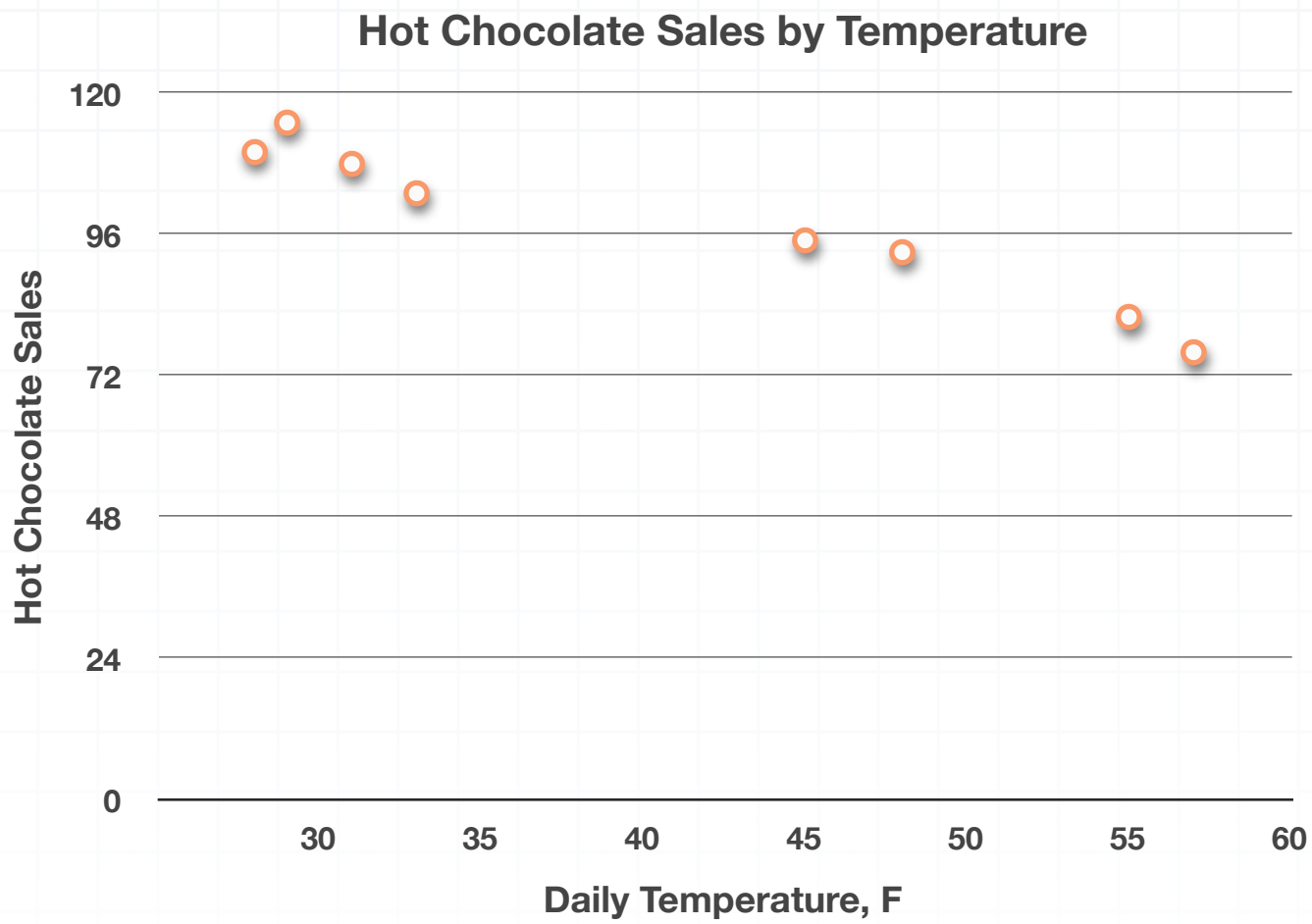
But even though cutting off the points makes the line of best fit have a stronger correlation, it would be good to include them in the data so that our line of best fit is not misleading.

■ 6. A small coffee shop wants to know how hot chocolate sales are affected by daily temperature. Find the rate of change of hot chocolate sales, with respect to temperature.

Daily Temperature, F	Hot Chocolate Sales
28	110
29	115
31	108
33	103
45	95
48	93
55	82
57	76

Solution:

Create a scatterplot.



From the plot we can see there's a relatively strong, negative linear relationship with no outliers. The rate of change is the slope, so we need to look at the slope of the line of best fit for the data set. The formula for the slope of the best-fit line is

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

Extend the table to find the values we need for the formula.

	Daily Temperature, F	Hot Chocolate Sales	xy	x^2
	28	110	3,080	784
	29	115	3,335	841
	31	108	3,348	961
	33	103	3,399	1,089
	45	95	4,275	2,025
	48	93	4,464	2,304
	55	82	4,510	3,025
	57	76	4,332	3,249
Sum:	**326**	**782**	**30,743**	**14,278**

Plug these values into the slope formula.

$$b = \frac{8(30{,}743) - (326)(782)}{8(14{,}278) - (326)^2}$$

$$b = \frac{245{,}944 - 254{,}932}{114{,}224 - 106{,}276}$$

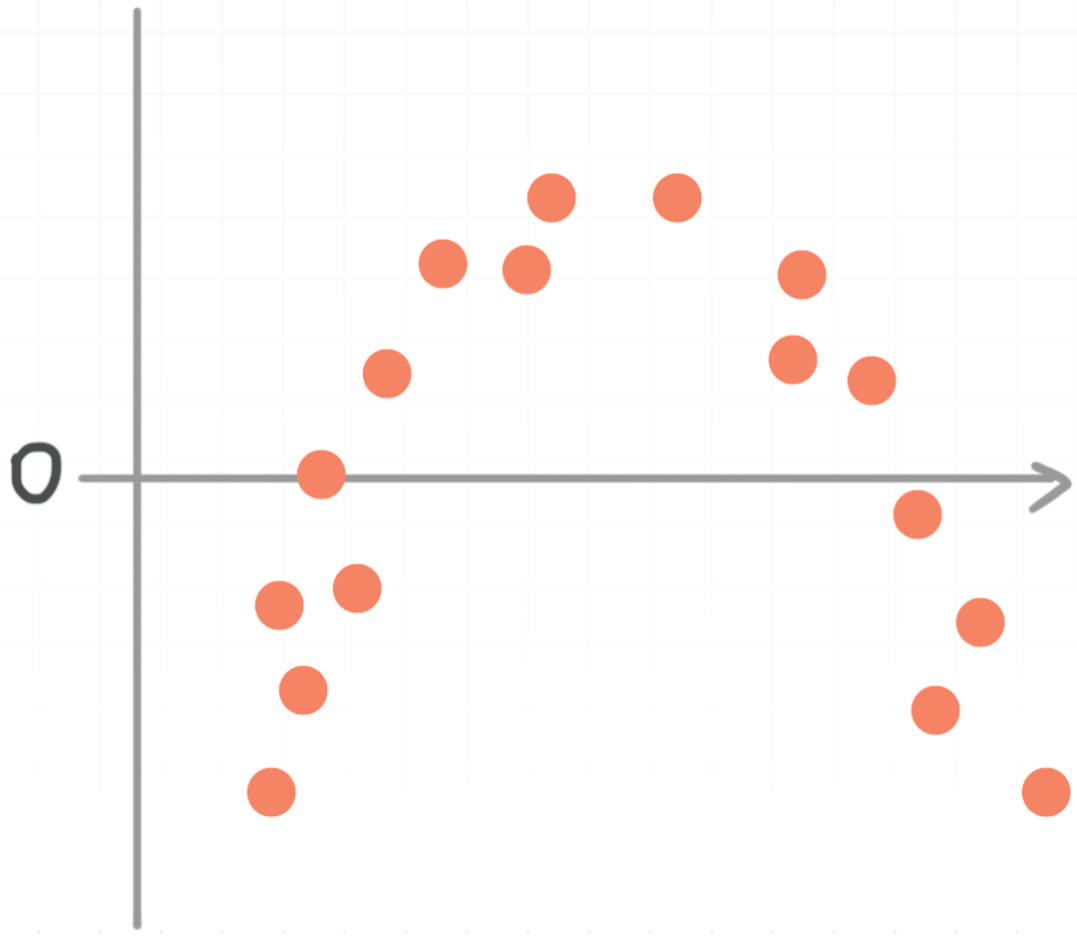
$$b = \frac{-8{,}988}{7{,}946}$$

$$b \approx -1.1311$$

The units of the slope are “hot chocolate sales per degree Fahrenheit.” So the shop can expect hot chocolate sales to decrease by 1.1311 cups for every one degree increase in temperature.

CORRELATION COEFFICIENT AND THE RESIDUAL

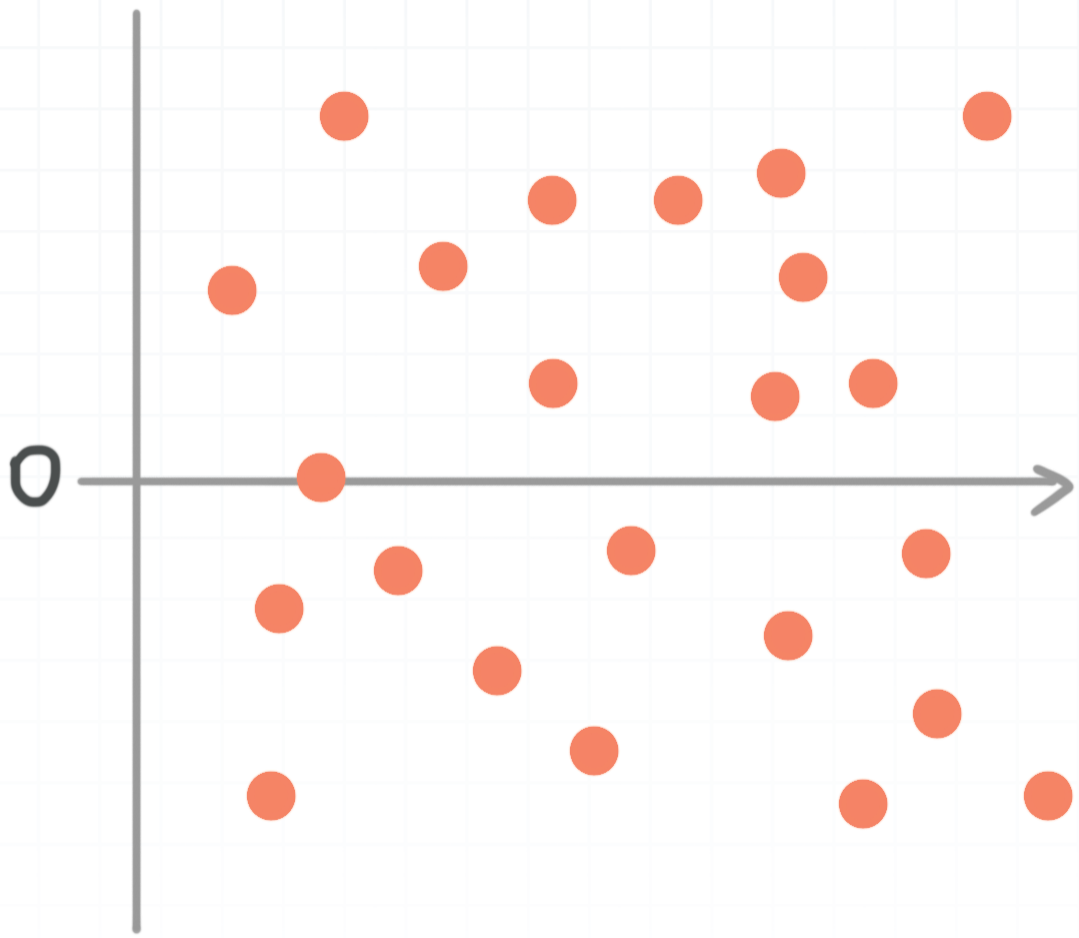
1. What does the shape of this residual plot tell us about the line of best fit that was created for the data?



Solution:

The shape of this graph tells us that the linear model is probably not the best choice for our data set, and that we should consider another type of regression curve, probably one that's quadratic.

■ 2. What does the shape of this residual plot tell us about the line of best fit that was created for the data?



Solution:

The points in this residual plot are evenly spaced around the line $y = 0$. It has about the same number of points on the left and right, and about the same number of points above and below 0. It doesn't appear to have outliers or interesting features. So the line of best fit for the data is probably a good one and can be useful for making predictions.

■ 3. Calculate and interpret the correlation coefficient for the data set.

x	y
54	0.162
57	0.127
62	0.864
77	0.895
81	0.943
93	1.206

Solution:

To find the correlation coefficient, we use the formula

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

We need to start by finding the means and standard deviations for both x and y. The means are

$$\bar{x} = \frac{54 + 57 + 62 + 77 + 81 + 93}{6}$$

$$\bar{x} \approx 70.6667$$

and

$$\bar{y} = \frac{0.162 + 0.127 + 0.864 + 0.895 + 0.943 + 1.206}{6}$$

$$\bar{y} \approx 0.6995$$

and the standard deviations are

$$s_x = \sqrt{\frac{\sum_{i=1}^{6}(x_i - \bar{x})^2}{n-1}}$$

$$s_x \approx \sqrt{\frac{277.7790 + 186.7790 + 75.1117 + 40.1107 + 106.7770 + 498.7760}{5}}$$

$$s_x \approx 15.3970$$

and

$$s_y = \sqrt{\frac{\sum_{i=1}^{6}(y_i - \bar{y})^2}{n-1}}$$

$$s_y \approx \sqrt{\frac{0.2889 + 0.3278 + 0.0271 + 0.0382 + 0.0593 + 0.2565}{5}}$$

$$s_y \approx 0.4467$$

Plug these values into the correlation coefficient formula.

$$r = \frac{1}{6-1}\left[\left(\frac{54 - 70.6667}{15.3970}\right)\left(\frac{0.162 - 0.6995}{0.4467}\right) + \left(\frac{57 - 70.6667}{15.3970}\right)\left(\frac{0.127 - 0.6995}{0.4467}\right)\right.$$

$$+\left(\frac{62 - 70.6667}{15.3970}\right)\left(\frac{0.864 - 0.6995}{0.4467}\right) + \left(\frac{77 - 70.6667}{15.3970}\right)\left(\frac{0.895 - 0.6995}{0.4467}\right)$$

$$\left.+\left(\frac{81 - 70.6667}{15.3970}\right)\left(\frac{0.943 - 0.6995}{0.4467}\right) + \left(\frac{93 - 70.6667}{15.3970}\right)\left(\frac{1.206 - 0.6995}{0.4467}\right)\right]$$

$$r = \frac{1}{5}\Big[(-1.0825)(-1.2033) + (-0.8876)(-1.2816) + (-0.5629)(0.3683)$$

$$+(0.4113)(0.4189) + (0.6711)(0.5217) + (1.4505)(1.1339)\Big]$$

$$r = \frac{1}{5}(1.3026 + 1.1375 - 0.2073 + 0.1723 + 0.3501 + 1.6447)$$

$$r = \frac{1}{5}(4.3999)$$

$$r \approx 0.88$$

The positive correlation coefficient tells us that the regression line has a positive slope. The fact that the positive value is closer to 1 than it is to 0 tells us the data is strongly correlated, or that it most likely has a strong linear relationship. If we looked at a scatterplot of the data and sketched in the regression line, we'd see that this was true.

■ 4. Calculate the residuals, draw the residual plot, and interpret the results. Compare the results to the r-value in the previous problem. The equation of the line of best fit for the data is

$$\hat{y} = 0.0257x - 1.1142$$

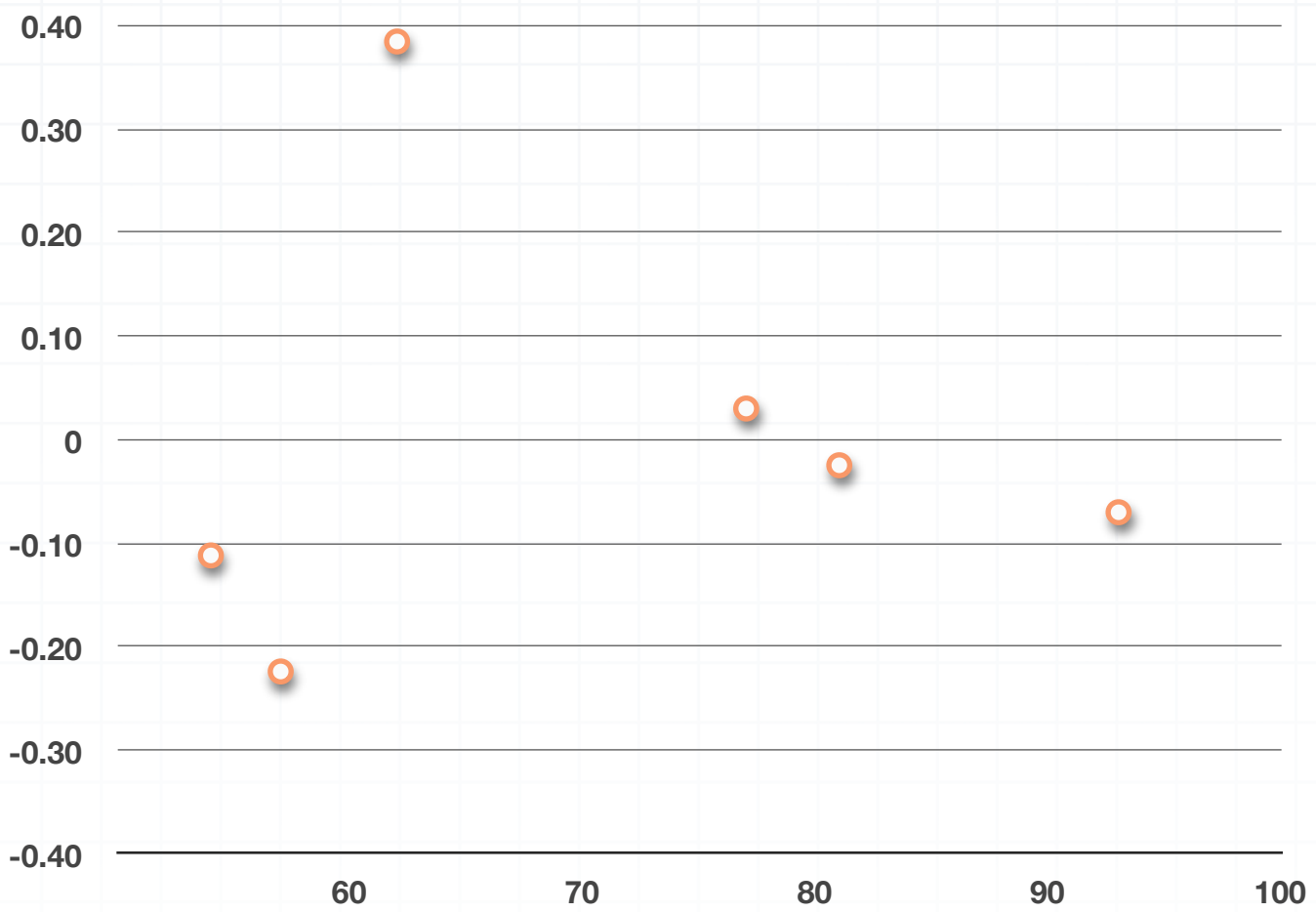
x	y
54	0.162
57	0.127
62	0.864
77	0.895
81	0.943
93	1.206

Solution:

Create a table to find the residual of each value.

x	Actual y	Predicted y	Residual
54	0.162	0.2736	-0.1116
57	0.127	0.3507	-0.2237
62	0.864	0.4792	0.3848
77	0.895	0.8647	0.0303
81	0.943	0.9675	-0.0245
93	1.206	1.2759	-0.0699

A plot of the residuals is



From the residual plot, it looks like the data had an outlier at $x = 62$. We already have a somewhat strong positive linear correlation from the correlation coefficient from the previous problem of $r \approx 0.88$, so it's likely the relationship would be even stronger without the outlier.

■ 5. The table shows average global sea surface temperature by year. Calculate and interpret the correlation coefficient for the data set. Leave the years as they are.

Year	Temperature, F
1880	-0.47001
1890	-0.88758
1900	-0.48331
1910	-1.11277
1920	-0.71965
1930	-0.58358
1940	-0.17977
1950	-0.55318
1960	-0.30358
1970	-0.30863
1980	0.077197
1990	0.274842
2000	0.232502
2010	0.612718

Solution:

Since this is a larger set of data, it can be nice to use a program like Excel to expand the table.

	Year	Temperature, F	$(x_i-x)^2$	$(y_i-y)^2$
	1880	-0.47001	4,225	0.02414
	1890	-0.88758	3,025	0.32827
	1900	-0.48331	2,025	0.02845
	1910	-1.11277	1,225	0.63703
	1920	-0.71965	625	0.16404
	1930	-0.58358	225	0.07233
	1940	-0.17977	25	0.01819
	1950	-0.55318	25	0.05691
	1960	-0.30358	225	0.00012
	1970	-0.30863	625	0.00004
	1980	0.077197	1,225	0.15353
	1990	0.274842	2,025	0.34748
	2000	0.232502	3,025	0.29935
	2010	0.612718	4,225	0.85997
Sum:	**27,230**	**-4.40480**	**22,750**	**2.98986**
Mean:	**0**	**-0.31463**		

The standard deviation for x and y are

$$s_x = \sqrt{\frac{\sum_{i=1}^{14}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{22{,}750}{13}} \approx 41.8330$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{14}(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{2.98986}{13}} \approx 0.4796$$

Now that we have the means and standard deviations, we can find the correlation coefficient. If we expand the table, then we'll be able to pull just the one sum out of the table to plug into the correlation coefficient formula.

	Year	Temp, F	(x_i-x)	$(x_i-x)/s_x$	(y_i-y)	$(y_i-y)/s_y$	$((x_i-x)/s_x)((y_i-y)/s_y)$
	1880	-0.47001	-65	-1.55380	-0.15538	0.00000	0.00000
	1890	-0.88758	-55	-1.31475	-0.57295	-1.19464	1.57066
	1900	-0.48331	-45	-1.07571	-0.16868	-0.35171	0.37834
	1910	-1.11277	-35	-0.83666	-0.79814	-1.66418	1.39235
	1920	-0.71965	-25	-0.59761	-0.40502	-0.84450	0.50468
	1930	-0.58358	-15	-0.35857	-0.26895	-0.56078	0.20108
	1940	-0.17977	-5	-0.11952	0.13486	0.28119	-0.03361
	1950	-0.55318	5	0.11952	-0.23855	-0.49739	-0.05945
	1960	-0.30358	15	0.35857	0.01105	0.02304	0.00826
	1970	-0.30863	25	0.59761	0.00600	0.01251	0.00748
	1980	0.077197	35	0.83666	0.39183	0.81699	0.68354
	1990	0.274842	45	1.07571	0.58947	1.22909	1.32214
	2000	0.232502	55	1.31475	0.54713	1.14081	1.49988
	2010	0.612718	65	1.55380	0.92735	1.93359	3.00440
Sum:							**10.47975**

The correlation coefficient is then

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

$$r = \frac{1}{14-1}(10.98314)$$

$$r = \frac{1}{13}(10.98314)$$

$$r \approx 0.8449$$

There's a strong positive linear relationship between the year and the temperature of the ocean's surface.

■ 6. Calculate the residuals and create the residual plot for the data in the table. Compare this with the r-value we calculated in the last question and interpret the results. Use the equation for the regression line $\hat{y} = 0.0143x - 28.332$.

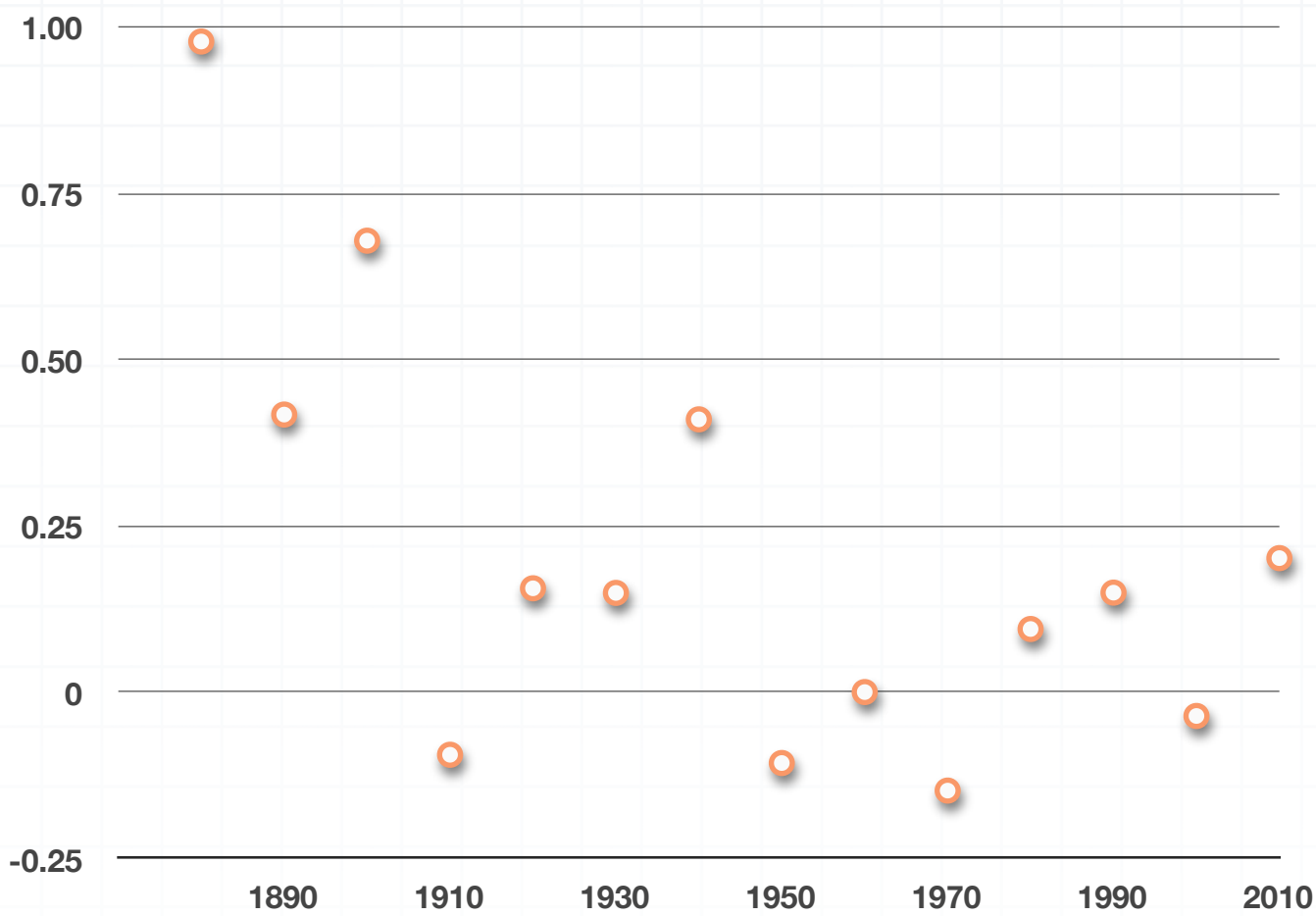
Year	Temperature, F
1880	-0.47001
1890	-0.88758
1900	-0.48331
1910	-1.11277
1920	-0.71965
1930	-0.58358
1940	-0.17977
1950	-0.55318
1960	-0.30358
1970	-0.30863
1980	0.077197
1990	0.274842
2000	0.232502
2010	0.612718

Solution:

Use the equation of the regression line to find predicted values of temperature, and add those values to the table.

Year	Actual y	Predicted y	Residual
1880	-0.47001	-1.448	0.97799
1890	-0.88758	-1.305	0.41742
1900	-0.48331	-1.162	0.67869
1910	-1.11277	-1.019	-0.09377
1920	-0.71965	-0.876	0.15635
1930	-0.58358	-0.733	0.14942
1940	-0.17977	-0.59	0.41023
1950	-0.55318	-0.447	-0.10618
1960	-0.30358	-0.304	0.00042
1970	-0.30863	-0.161	-0.14763
1980	0.077197	-0.018	0.095197
1990	0.274842	0.125	0.149842
2000	0.232502	0.268	-0.035498
2010	0.612718	0.411	0.201718

Make a plot of the residuals.



The residual plot is a good example of why finding the correlation coefficient is not enough.

This plot makes it look like using an exponential regression would be a better fit for the data. The residuals are not above and below the line $y = 0$ in a random pattern. Which means that even though the correlation coefficient of $r \approx 0.8449$ says there is a strong positive linear relationship between year and temperature, it probably won't do as good of a job as we think at making predictions for the future because another type of model would be better.

COEFFICIENT OF DETERMINATION AND RMSE

■ 1. Linda read an article about the predictions of high school students and their GPA. The article studied three factors, the number of volunteer organizations each student participated in, the number of hours spent on homework, and the student's individual scores on standardized tests.

The article concluded that the number of hours spent on homework are the best predictor of GPA, because they found $24\,\%$ of the variance in GPA to be from hours spent on homework, $15\,\%$ from the number of volunteer organizations, and $11.5\,\%$ from individual scores on standardized tests.

What is the coefficient of determination for the line-of-best-fit that has y -values of high school GPA and x-values of hours spent on homework? Is the line of best fit a good predictor of the data? Why or why not?

Solution:

Remember the percent of the variation in y that can be explained by the x -values is the coefficient of determination or the r^2 value.

In this context, the percent of the variance in GPA due to hours spent on homework is $24\,\%$. So, we're talking about a least squares line where $r^2 = 0.24$. This is a very weak positive relationship, so the line of best fit is probably not a good predictor of the connection between hours spent on homework and GPA.

■ 2. For the data in the table, calculate the sum of the squared residuals based on the mean of the y-values.

x	y
1	3.1
2	3.4
3	3.7
4	3.9
5	4.1

Solution:

First calculate $\bar{y}$.

$$\bar{y} = \frac{3.1 + 3.4 + 3.7 + 3.9 + 4.1}{5}$$

$$\bar{y} = \frac{18.2}{5}$$

$$\bar{y} = 3.64$$

The formula for a residual is

$$\text{residual} = \text{actual} - \text{predicted}$$

In this case, the predicted value is the mean of the y-values, $\bar{y} = 3.64$. Let's expand the table and calculate the residuals.

x	y	e
1	3.1	-0.54
2	3.4	-0.24
3	3.7	0.06
4	3.9	0.26
5	4.1	0.46

Now we just need to find the squares of these residuals and add them together.

	x	y	e	e^2
	1	3.1	-0.54	0.2916
	2	3.4	-0.24	0.0576
	3	3.7	0.06	0.0036
	4	3.9	0.26	0.0676
	5	4.1	0.46	0.2116
Sum:				**0.632**

So the sum of the squared residuals is about 0.632.

■ 3. Use the same data as the previous question to calculate the sum of the squared residuals based on the least squares regression line, $\hat{y} = 0.25x + 2.89$.

Solution:

The formula for a residual is

$$\text{residual} = \text{actual} - \text{predicted}$$

In this case the predicted value is based on the regression line, $\hat{y} = 0.25x + 2.89$. Let's expand the table and calculate the residuals.

x	Actual y	Predicted y	e
1	3.1	3.14	-0.04
2	3.4	3.39	0.01
3	3.7	3.64	0.06
4	3.9	3.89	0.01
5	4.1	4.14	-0.04

Now we just need to find the squares of these residuals and add them together.

	x	Actual y	Predicted y	e	e^2
	1	3.1	3.14	-0.04	0.0016
	2	3.4	3.39	0.01	0.0001
	3	3.7	3.64	0.06	0.0036
	4	3.9	3.89	0.01	0.0001
	5	4.1	4.14	-0.04	0.0016
Sum:					**0.007**

So the sum of the squared residuals is 0.007.

■ 4. Based on the previous two questions, in which we found the sum of the squared residuals based on the mean of the y-values and then the line of best fit, what percentage of error did we eliminate by using the least squares line? What is the term for this error?

Solution:

The sum of the squared residuals for the mean of the y-values was

$$\sum \text{residuals}^2 = 0.632$$

The sum of the squared residuals for the line of best fit was

$$\sum \text{residuals}^2 = 0.007$$

This means using the line of best fit reduces the error by

$$0.632 - 0.007$$

$$0.625$$

This is

$$\frac{0.625}{0.632} = 0.9889 = 98.89\,\%$$

of a reduction in error by using the least squares regression line. This is another way to calculate the coefficient of determination, so $r^2 = 0.9889$.

5. What is the RMSE of the data set and what does it mean?

x	y
1	3.1
2	3.4
3	3.7
4	3.9
5	4.1

Solution:

To find RMSE, we'll use the formula

$$\text{RMSE} = \sqrt{\frac{\sum \text{residuals}^2}{n}}$$

We already calculated the residual sum of squares.

	x	Actual y	Predicted y	e	e^2
	1	3.1	3.14	-0.04	0.0016
	2	3.4	3.39	0.01	0.0001
	3	3.7	3.64	0.06	0.0036
	4	3.9	3.89	0.01	0.0001
	5	4.1	4.14	-0.04	0.0016
Sum:					**0.007**

The sum of the squared residuals was 0.007, so RMSE will be

$$\text{RMSE} = \sqrt{\frac{0.007}{4}} \approx 0.0418$$

RMSE is the standard deviation of the residuals, which means that

- $68\,\%$ of the data points will be within ± 0.0418 of the regression line,
- $95\,\%$ of the data points will be within $\pm 2(0.0418)$ of the regression line, and
- $99.7\,\%$ of the data points will be within $\pm 3(0.0418)$ of the regression line.

Since the RMSE we found is a small standard deviation, the data points are going to be more tightly clustered around the line-of-best-fit and the correlation in the data will be stronger.

■ 6. Calculate the RMSE for the data set, given that the least squares line is $\hat{y} = 0.0028x + 1.2208$.

x	y
5	1.25
10	1.29
12	1.17
15	1.24
17	1.32

Solution:

To find RMSE, we'll use the formula

$$\text{RMSE} = \sqrt{\frac{\sum \text{residuals}^2}{n}}$$

We can calculate the residual sum of squares.

	x	Actual y	Predicted y	e	e^2
	5	1.25	1.2348	0.0152	0.00023104
	10	1.29	1.2488	0.0412	0.00169744
	12	1.17	1.2544	-0.0844	0.00712336
	15	1.24	1.2628	-0.0228	0.00051984
	17	1.32	1.2684	0.0516	0.00266256
Sum:					**0.01223424**

The sum of the squared residuals was 0.01223424, so RMSE will be

$$\text{RMSE} = \sqrt{\frac{0.01223424}{4}} \approx 0.0553$$

CHI-SQUARE TESTS

1. We want to know whether a person's geographic region of the United State affects their preference of cell phone brand. We randomly sample people across the country and ask them about their brand preference. What can we conclude using a chi-square test at $95\,\%$ confidence?

	iPhone	Android	Other	Totals
Northeast	72	33	8	113
Southeast	48	26	7	81
Midwest	107	50	10	167
Northwest	59	33	10	102
Southwest	61	27	9	97
Totals	347	169	44	560

Solution:

Start by computing expected values.

	iPhone	Android	Other	Totals
Northeast	72 (70.02)	33 (34.10)	8 (8.88)	113
Southeast	48 (50.19)	26 (24.44)	7 (6.36)	81
Midwest	107 (103.48)	50 (50.40)	10 (13.12)	167
Northwest	59 (63.20)	33 (30.78)	10 (8.01)	102
Southwest	61 (60.11)	27 (29.27)	9 (7.62)	97
Totals	347	169	44	560

Now we'll check our sampling conditions. The problem told us that we took a random sample, and all of our expected values are at least 5, so we've met the random sampling and large counts conditions. And even though we're sampling without replacement, 560 is far less than 10% of the US population, so we've met the independence condition as well.

We'll state the null and alternative hypotheses.

H_0: Cell phone brand preference isn't affected by geographic region.

H_a: Cell phone brand preference is affected by geographic region.

Calculate χ^2.

$$\chi^2 = \frac{(72-70.02)^2}{70.02} + \frac{(33-34.10)^2}{34.10} + \frac{(8-8.88)^2}{8.88}$$

$$+\frac{(48-50.19)^2}{50.19} + \frac{(26-24.44)^2}{24.44} + \frac{(7-6.36)^2}{6.36}$$

$$+\frac{(107-103.48)^2}{103.48} + \frac{(50-50.40)^2}{50.40} + \frac{(10-13.12)^2}{13.12}$$

$$+\frac{(59-63.20)^2}{63.20}+\frac{(33-30.78)^2}{30.78}+\frac{(10-8.01)^2}{8.01}$$

$$+\frac{(61-60.11)^2}{60.11}+\frac{(27-29.27)^2}{29.27}+\frac{(9-7.62)^2}{7.62}$$

$$\chi^2 \approx 0.0560 + 0.0355 + 0.0872$$

$$+0.0956 + 0.0996 + 0.0644$$

$$+0.1197 + 0.0032 + 0.7420$$

$$+0.2791 + 0.1601 + 0.4944$$

$$+0.0132 + 0.1760 + 0.2499$$

$$\chi^2 \approx 2.6759$$

The degrees of freedom are

$$\text{df} = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

$$\text{df} = (5-1)(3-1)$$

$$\text{df} = (4)(2)$$

$$\text{df} = 8$$

With df $= 8$ and $\chi^2 \approx 2.6759$, the χ^2-table gives

	Upper-tail probability p											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67

We're off the chart on the left, which means we will definitely not exceed the alpha level $\alpha = 0.05$. Therefore, we'll fail to reject the null hypothesis, and conclude that geographic region of the country does not affect cell phone brand preference.

■ 2. A beverage company wants to know if gender affects which of their products people prefer. They take a random sample of fewer than $10\,\%$ of their customers, and ask them in a blind taste test which beverage they prefer. What can the company conclude using a chi-square test at $\alpha = 0.1$?

	Beverage			
	A	B	C	Totals
Men	35	34	31	100
Women	31	33	36	100
Totals	66	67	67	200

Solution:

Start by computing expected values.

	Beverage			
	A	B	C	Totals
Men	35 (33.0)	34 (33.5)	31 (33.5)	100
Women	31 (33.0)	33 (33.5)	36 (33.5)	100
Totals	66	67	67	200

Now we'll check our sampling conditions. The problem told us that we took a random sample and that we sampled less than $10\,\%$ of the population, so we've met the random sampling and independence conditions. And all of our expected values are at least 5, so we've met the large counts condition as well.

We'll state the null and alternative hypotheses.

H_0: Gender does not affect beverage preference.

H_a: Gender affects beverage preference.

Calculate χ^2.

$$\chi^2 = \frac{(35-33)^2}{33} + \frac{(34-33.5)^2}{33.5} + \frac{(31-33.5)^2}{33.5}$$

$$+\frac{(31-33)^2}{33} + \frac{(33-33.5)^2}{33.5} + \frac{(36-33.5)^2}{33.5}$$

$$\chi^2 = \frac{4}{33} + \frac{0.25}{33.5} + \frac{6.25}{33.5} + \frac{4}{33} + \frac{0.25}{33.5} + \frac{6.25}{33.5}$$

$$\chi^2 \approx 0.1212 + 0.0075 + 0.1866 + 0.1212 + 0.0075 + 0.1866$$

$$\chi^2 = 0.6306$$

The degrees of freedom are

$$\text{df} = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

$$\text{df} = (2 - 1)(3 - 1)$$

$$\text{df} = (1)(2)$$

$$\text{df} = 2$$

With $\text{df} = 2$ and $\chi^2 = 0.6306$, the χ^2-table gives

	Upper-tail probability p											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.81	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73

We're off the chart on the left, which means we will definitely not exceed the alpha level $\alpha = 0.05$. Therefore, we'll fail to reject the null hypothesis, and conclude that gender does not affect beverage preference.

■ 3. A coffee company wants to know whether or not drink and pastry choice are related among their customers. The company randomly sampled fewer than $10\,\%$ of their customers, and recorded their drink and pastry orders. What can the restaurant conclude using a chi-square test at $99\,\%$ confidence?

	Bagel	Muffin	Totals
Coffee	38	34	72
Tea	25	29	54
Totals	63	63	126

Solution:

Start by computing expected values.

	Bagel	Muffin	Totals
Coffee	38 (36)	34 (36)	72
Tea	25 (27)	29 (27)	54
Totals	63	63	126

Now we'll check our sampling conditions. The problem told us that we took a random sample and that we sampled less than $10\,\%$ of the population, so we've met the random sampling and independence conditions. And all of our expected values are at least 5, so we've met the large counts condition as well.

We'll state the null and alternative hypotheses.

H_0: Pastry preference isn't affected by beverage preference.

H_a: Pastry preference is affected by beverage preference.

Calculate χ^2.

$$\chi^2 = \frac{(38-36)^2}{36} + \frac{(34-36)^2}{36} + \frac{(25-27)^2}{27} + \frac{(29-27)^2}{27}$$

$$\chi^2 = \frac{4}{36} + \frac{4}{36} + \frac{4}{27} + \frac{4}{27}$$

$$\chi^2 = \frac{2}{9} + \frac{8}{27}$$

$$\chi^2 \approx 0.52$$

The degrees of freedom are

$$\text{df} = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

$$\text{df} = (2-1)(2-1)$$

$$\text{df} = (1)(1)$$

$$\text{df} = 1$$

With df $= 1$ and $\chi^2 = 0.52$, the χ^2-table gives

	Upper-tail probability p											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.81	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73

We're off the chart on the left, which means we will definitely not exceed the alpha level $\alpha = 0.01$. Therefore, the coffee company will fail to reject the null hypothesis, and conclude that beverage preference does not affect pastry preference.

4. A school district wants to know whether or not GPA is affected by elective preference. They randomly sampled fewer than $10\,\%$ of their students, and recorded their elective preference an GPA. What can the school district conclude using a chi-square test at $\alpha = 0.1$?

	GPA range				
	<2	2	3	4+	Totals
Music	12	26	31	34	103
Theater	21	22	23	21	87
Art	36	29	29	32	126
Totals	69	77	83	87	316

Solution:

Start by computing expected values.

	GPA range				
	<2	2	3	4+	Totals
Music	12 (22.49)	26 (25.10)	31 (27.05)	34 (28.36)	103
Theater	21 (19.00)	22 (21.20)	23 (22.85)	21 (23.95)	87
Art	36 (27.51)	29 (30.70)	29 (33.09)	32 (34.69)	126
Totals	69	77	83	87	316

Now we'll check our sampling conditions. The problem told us that we took a random sample and that we sampled less than $10\,\%$ of the population, so we've met the random sampling and independence conditions. And all of our expected values are at least 5, so we've met the large counts condition as well.

We'll state the null and alternative hypotheses.

H_0: Elective choice doesn't affect GPA.

H_a: Elective choice affects GPA.

Calculate χ^2.

$$\chi^2 = \frac{(12-22.49)^2}{22.49} + \frac{(26-25.10)^2}{25.10} + \frac{(31-27.05)^2}{27.05} + \frac{(34-28.36)^2}{28.36}$$

$$+\frac{(21-19.00)^2}{19.00} + \frac{(22-21.20)^2}{21.20} + \frac{(23-22.85)^2}{22.85} + \frac{(21-23.95)^2}{23.95}$$

$$+\frac{(36-27.51)^2}{27.51} + \frac{(29-30.70)^2}{30.70} + \frac{(29-33.09)^2}{33.09} + \frac{(32-34.69)^2}{34.69}$$

$$\chi^2 \approx 4.89 + 0.03 + 0.58 + 1.12 + 0.21 + 0.03$$

$$+0.00 + 0.36 + 2.62 + 0.09 + 0.51 + 0.21$$

$$\chi^2 = 10.65$$

The degrees of freedom are

$$\text{df} = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

$$\text{df} = (3-1)(4-1)$$

$$\text{df} = (2)(3)$$

$$\text{df} = 6$$

With df $= 6$ and $\chi^2 = 10.65$, the χ^2-table gives

	Upper-tail probability p											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.52	22.11
6	7.84	8.56	9.45	**10.64**	**12.59**	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02

The χ^2 value just clears $\alpha = 0.1$, which means that the school district can reject the null hypothesis and conclude that elective choice affects GPA. If they had set a higher confidence level of $95\,\%$ (with $\alpha = 0.05$), they would not have been able to reject the null.

■ 5. An airline wants to know if people travel constantly throughout the year, or if travel is more concentrated at specific times. They recorded flights taken each quarter, and recorded them in a table (in hundreds of thousands). What can the airline conclude using a chi-square test at $95\,\%$ confidence?

Quarter	Jan-Mar	Apr-Jun	Jul-Sep	Oct-Dec	Total
Flights	3.97	4.58	4.73	5.14	18.42

Solution:

With 18.42 (or $18{,}420{,}000$) total flights, the expected number of flights in each quarter would be $18.42/4 = 4.605$.

Quarter	Jan-Mar	Apr-Jun	Jul-Sep	Oct-Dec	Total
Flights	3.97	4.58	4.73	5.14	18.42
Expected	4.605	4.605	4.605	4.605	18.42

We'll state the null and alternative hypotheses.

H_0: Number of flights taken is not affected by quarter.

H_a: Number of flights taken is affected by quarter.

Calculate χ^2.

$$\chi^2 = \frac{(3.97 - 4.605)^2}{4.605} + \frac{(4.58 - 4.605)^2}{4.605} + \frac{(4.73 - 4.605)^2}{4.605} + \frac{(5.14 - 4.605)^2}{4.605}$$

$$\chi^2 \approx 0.0876 + 0.0001 + 0.0034 + 0.0622$$

$$\chi^2 = 0.1533$$

The degrees of freedom are $n - 1 = 4 - 1 = 3$. With df $= 3$ and $\chi^2 = 0.1533$, the χ^2-table gives

	Upper-tail probability p											
df	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.0025**	**0.001**	**0.0005**
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	**4.11**	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00

We're off the chart on the left, which means we will definitely not exceed the alpha level $\alpha = 0.05$. Therefore, the airline will fail to reject the null hypothesis, and conclude that number of flights taken is not affected by quarter.

■ 6. A sandwich company wants to know how their sales are affected by time of day. They recorded sandwiches sold during each part of the day. What can the sandwich company conclude using a chi-square test at $\alpha = 0.1$?

Time of day	Midday	Afternoon	Evening	Total
Sales	213	208	221	642

Solution:

With 642 total sandwiches sold, the expected number of sandwiches in each period would be $642/3 = 214$.

Time of day	Midday	Afternoon	Evening	Total
Sales	213	208	221	642
Expected	214	214	214	642

We'll state the null and alternative hypotheses.

H_0: Number of sandwiches sold is not affected by time of day.

H_a: Number of sandwiches sold is affected by time of day.

Calculate χ^2.

$$\chi^2 = \frac{(213-214)^2}{214} + \frac{(208-214)^2}{214} + \frac{(221-214)^2}{214}$$

$$\chi^2 = \frac{1}{214} + \frac{36}{214} + \frac{49}{214}$$

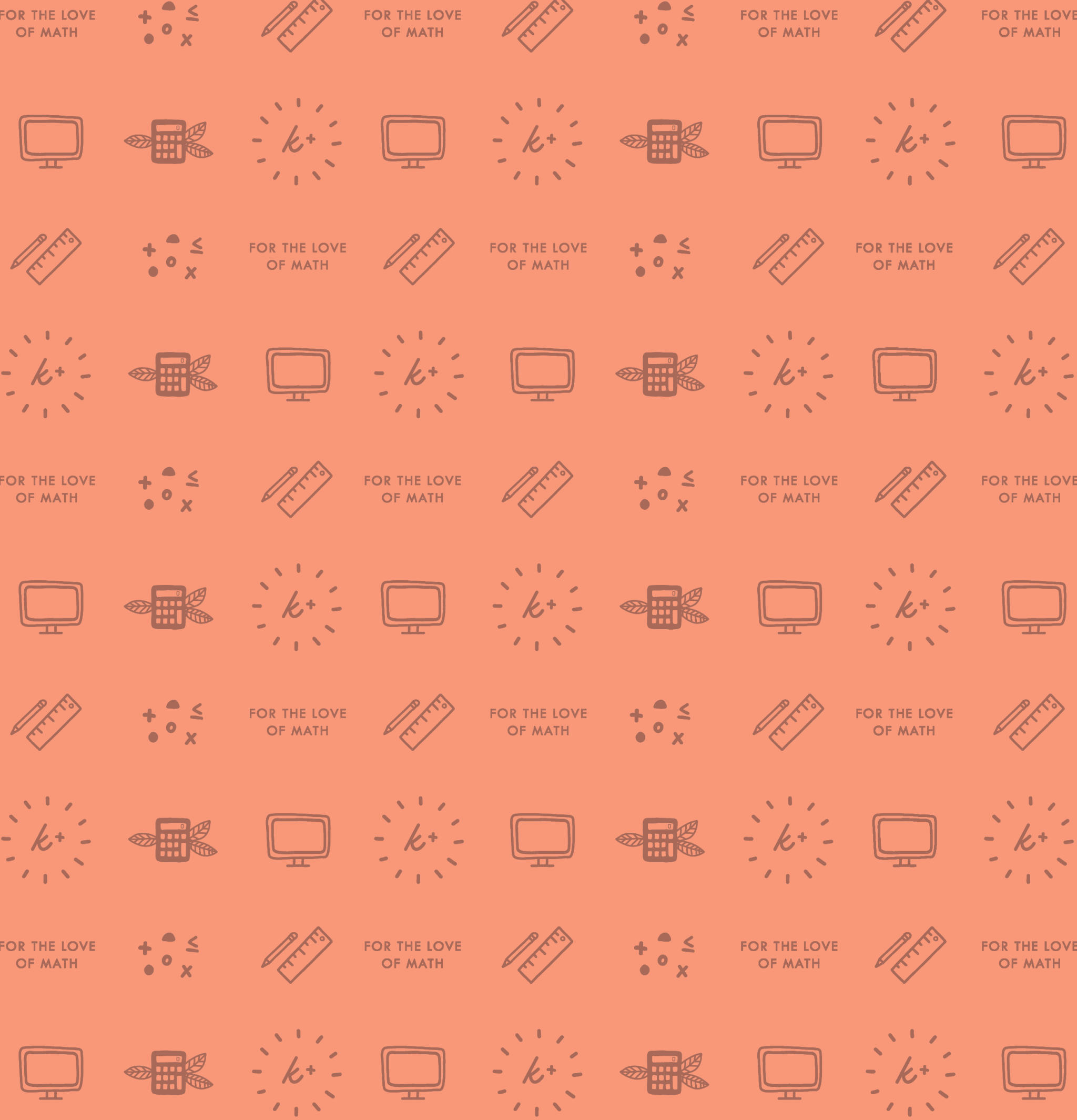
$$\chi^2 = \frac{86}{214}$$

$$\chi^2 \approx 0.4019$$

The degrees of freedom are $n - 1 = 3 - 1 = 2$. With df $= 2$ and $\chi^2 = 0.4019$, the χ^2-table gives

	Upper-tail probability p											
df	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.0025**	**0.001**	**0.0005**
1	1.32	1.64	2.07	2.71	3.81	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	**2.77**	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73

We're off the chart on the left, which means we will definitely not exceed the alpha level $\alpha = 0.1$. Therefore, the sandwich company will fail to reject the null hypothesis, and conclude that number of sandwiches sold is not affected by time of day.


FOR THE LOVE
OF MATH