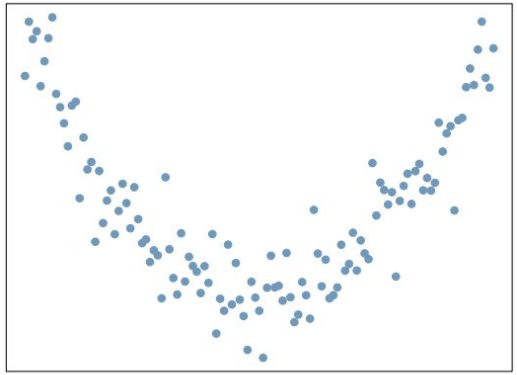
Lecture 9 practice

Stats 7 Summer Session II 2022

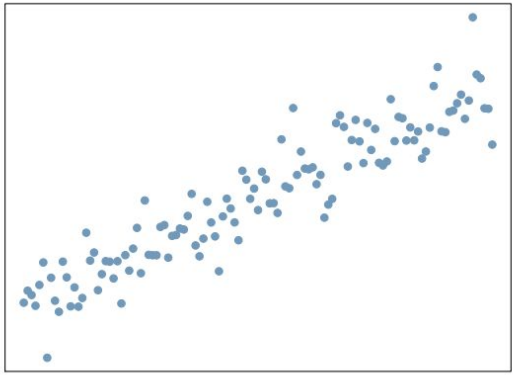
Identify relationships

For each plot, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.



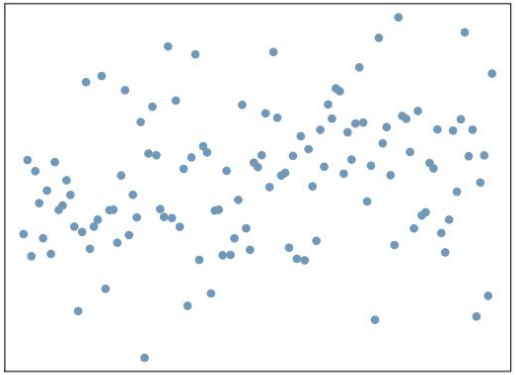
(a)

Strong relationship, but a straight line would not fit the data.



(b)

Strong relationship, and a linear fit would be reasonable.

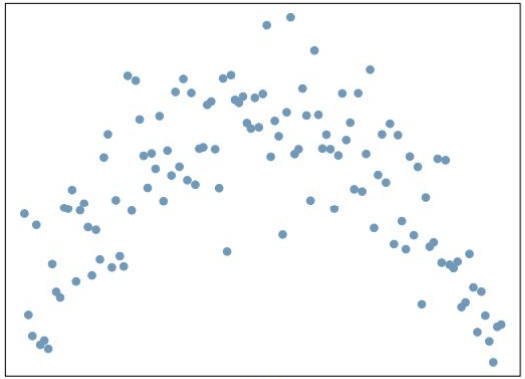


(c)

Weak relationship, and trying a linear fit would be reasonable.

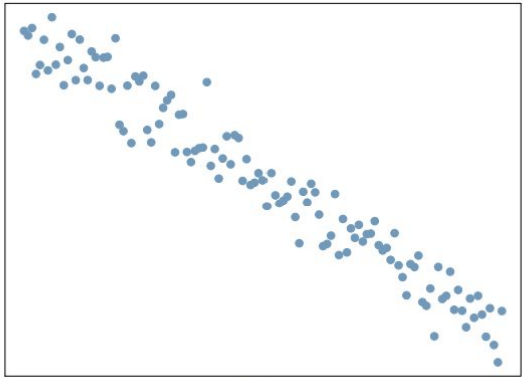
Identify relationships

For each plot, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.



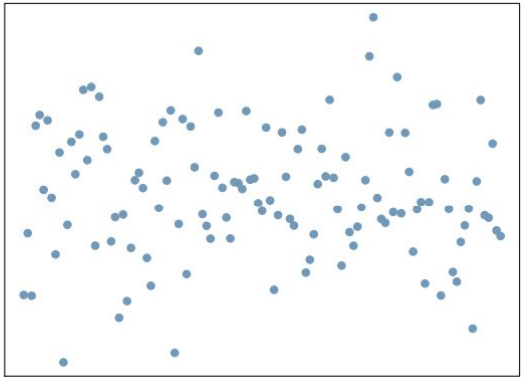
(d)

Moderate relationship, but a straight line would not fit the data.



(e)

Strong relationship, and a linear fit would be reasonable

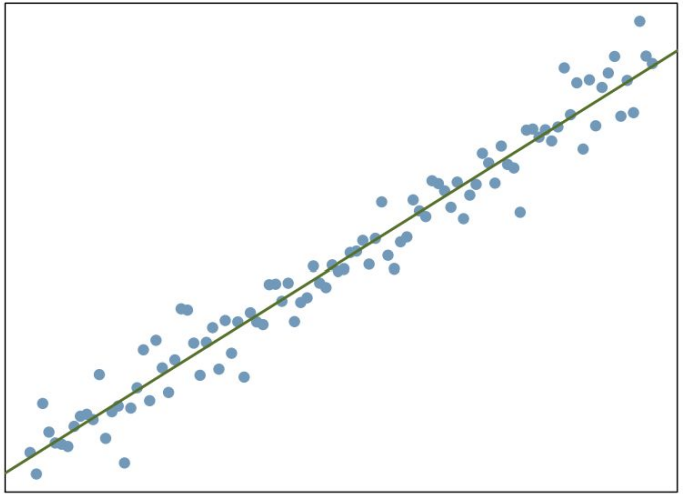


(f)

Weak relationship, and trying a linear fit would be reasonable.

Visualize the residuals

The scatterplots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe what those plots would look like.

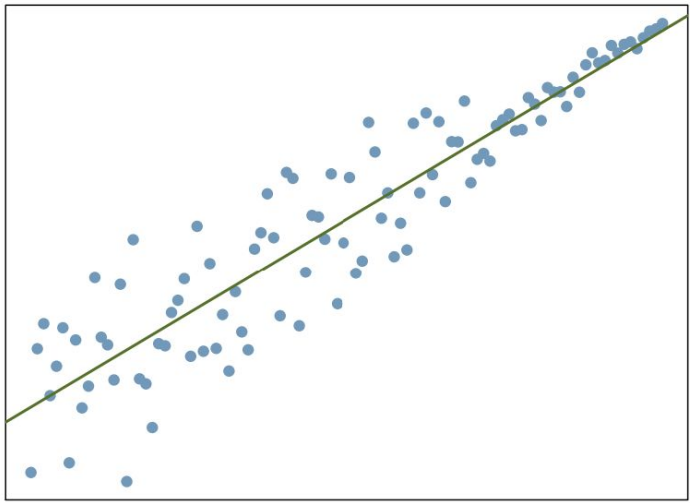


(a)

- The residual plot will show randomly distributed residuals around 0.
- The variance is also approximately constant.

Visualize the residuals

The scatterplots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe what those plots would look like.

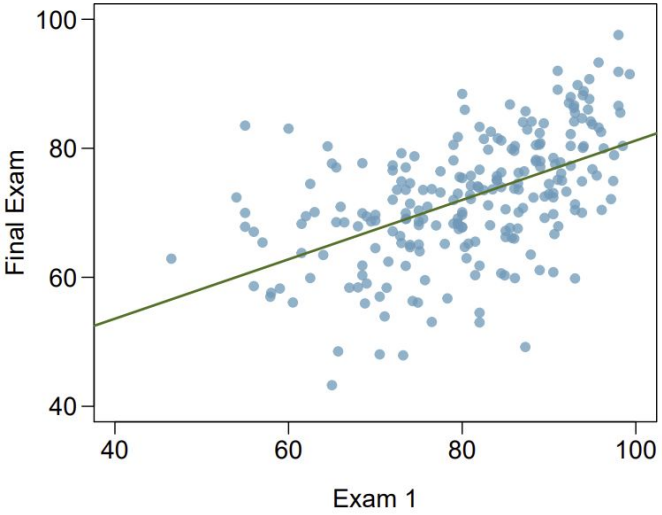


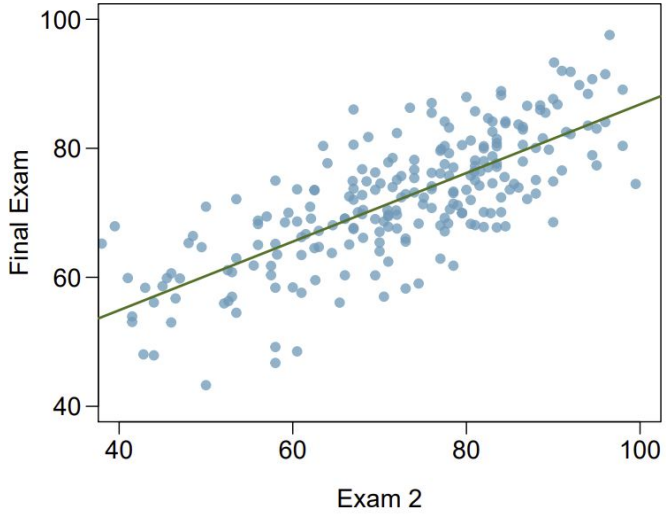
(b)

- The residuals will show a fan shape, with higher variability for smaller x.
- There will also be many points on the right above the line.
- There is trouble with the model being fit here.

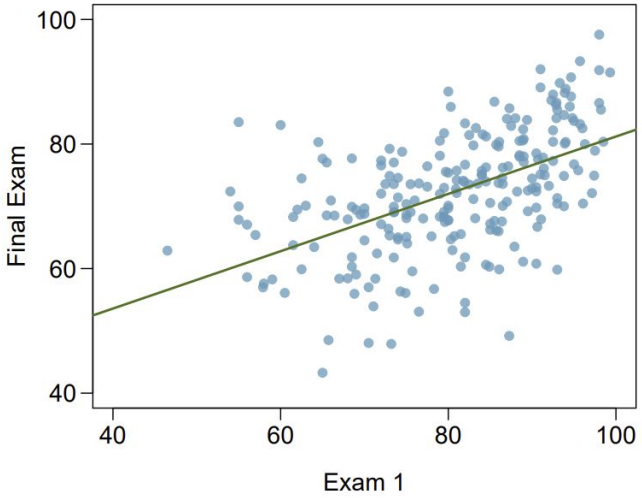
Exams and grades

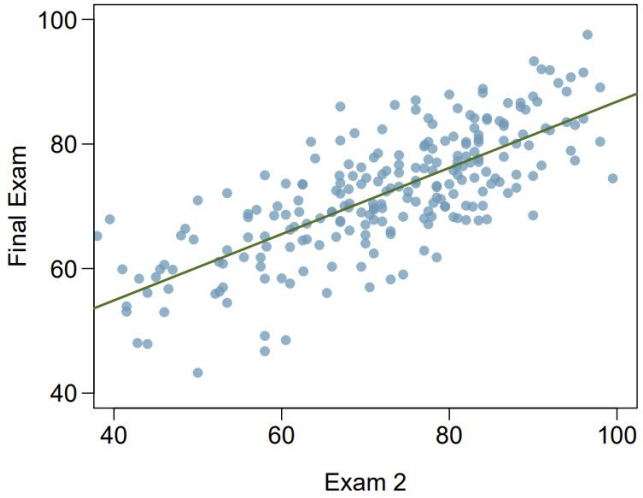
The two scatterplots below show the relationship between final and mid-semester exam grades recorded during several years for a Statistics course at a university.





Exams and grades

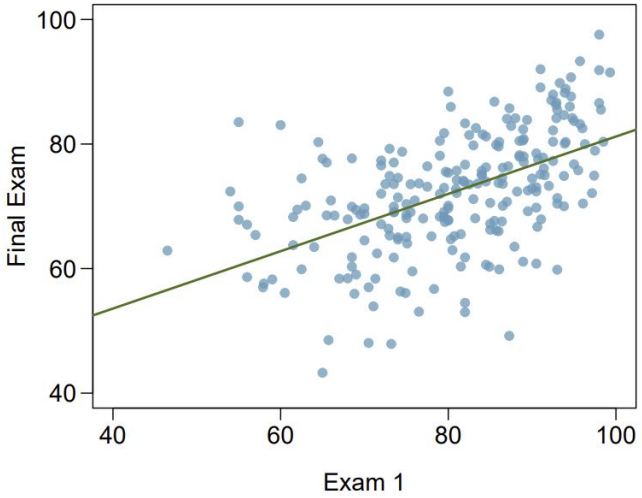


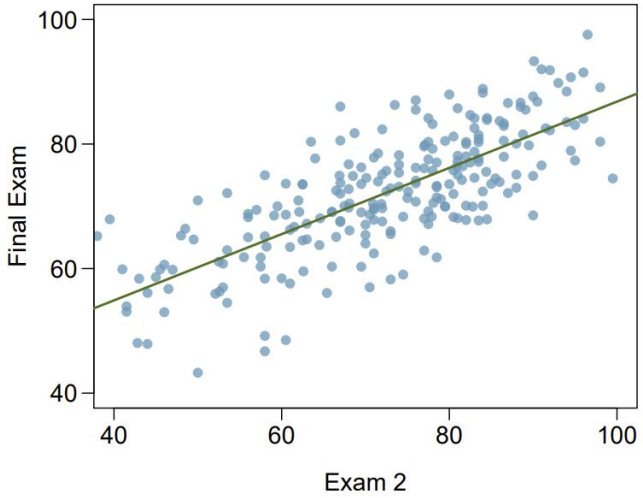


(a) Based on these graphs, which of the two exams has the strongest correlation with the final exam grade? Explain.

Exam 2 since there is less of a scatter in the plot of final exam grade versus exam 2. Notice that the relationship between Exam 1 and the Final Exam appears to be slightly nonlinear.

Exams and grades





(b) Can you think of a reason why the correlation between the exam you chose in part (a) and the final exam is higher?

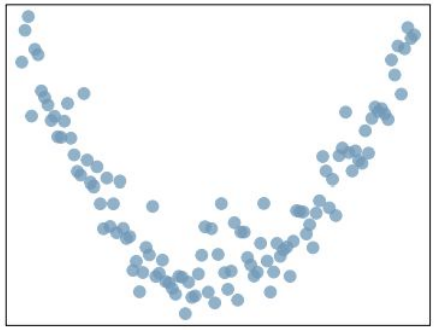
Exam 2 and the final are relatively close to each other chronologically, or Exam 2 may be cumulative so has greater similarities in material to the final exam.

Match the correlation

Match each correlation to the corresponding scatterplot.

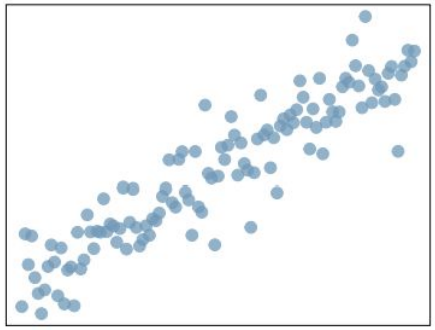
(a) R = -0.7 (b) R = 0.45 (c) R = 0.06 (d) R = 0.92

R = 0.06



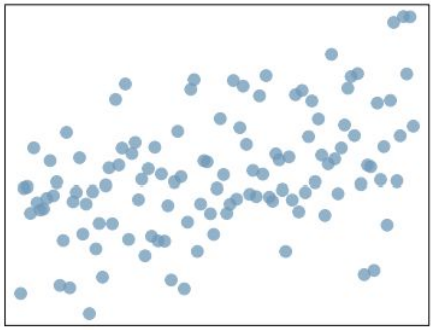
(1)

R = 0.92



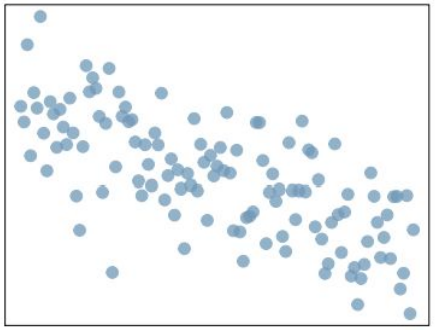
(2)

R = 0.45



(3)

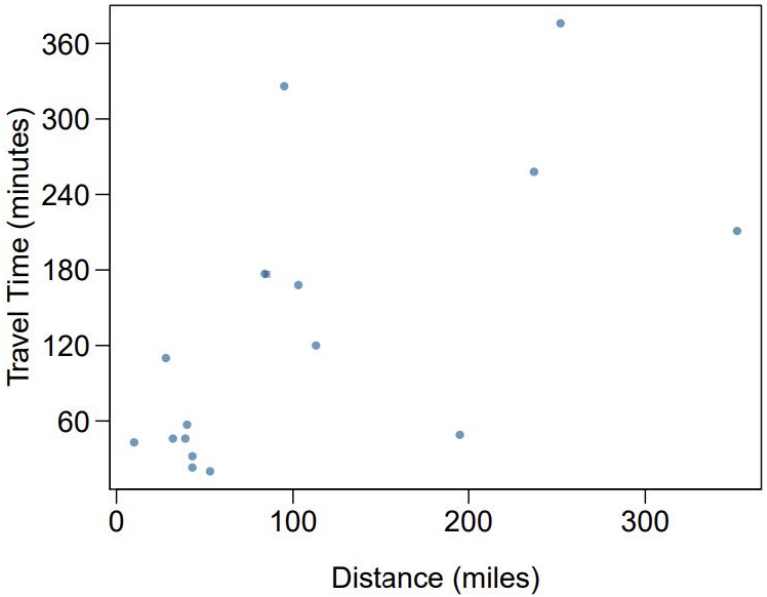
R = -0.7



(4)

The Coast Starlight

The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes).

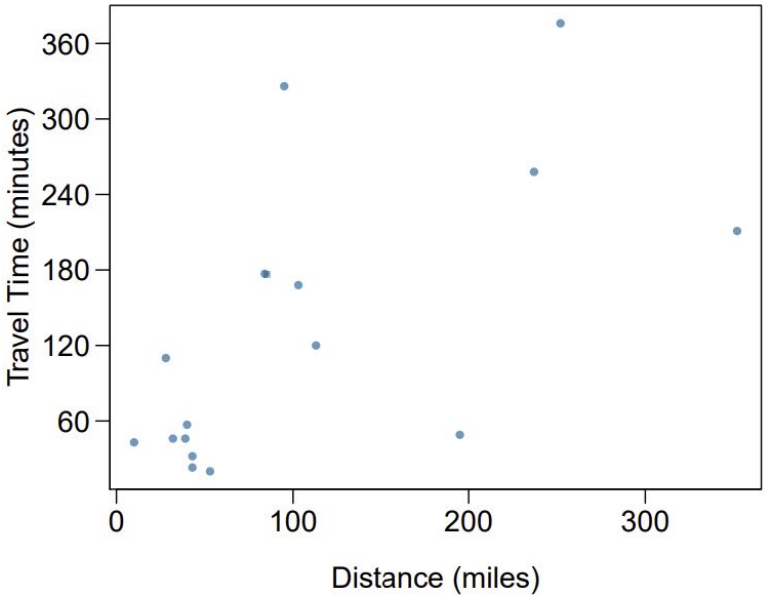


(a) Describe the relationship between distance and travel time.

There is a somewhat weak, positive, possibly linear relationship between the distance traveled and travel time. There is clustering near the lower left corner that we should take special note of.

The Coast Starlight

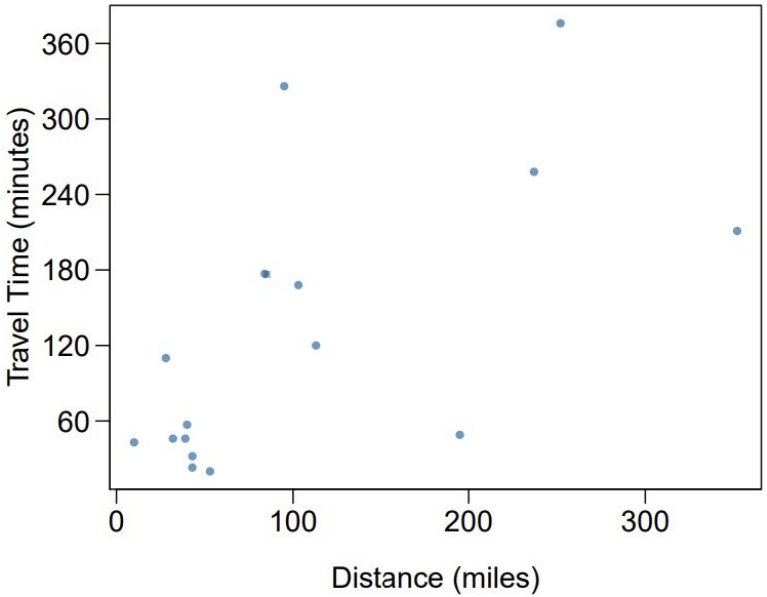
(b) How would the relationship change if travel time was instead measured in hours, and distance was instead measured in kilometers?



Changing the units will not change the form, direction or strength of the relationship between the two variables. If longer distances measured in miles are associated with longer travel time measured in minutes, longer distances measured in kilometers will be associated with longer travel time measured in hours.

The Coast Starlight

(c) Correlation between travel time (in miles) and distance (in minutes) is r = 0.636. What is the correlation between travel time (in kilometers) and distance (in hours)?



Changing units doesn't affect correlation: r = 0.636.

Correlation

What would be the correlation between the ages of husbands and wives if men always married woman who were

(a) 3 years younger than themselves?

(b) 2 years older than themselves?

(c) half as old as themselves?

In each part, we can write the husband ages as a linear function of the wife ages.

(a) $ageH = ageW + 3$. (b) $ageH = ageW - 2$. (c) $ageH = 2 \times ageW$.

Since the slopes are positive and these are perfect linear relationships, the correlation will be exactly 1 in all three parts.

Units of regression

Consider a regression predicting weight (kg) from height (cm) for a sample of adult males. What are the units of the correlation coefficient, the intercept, and the slope?

Correlation: no units.

Intercept: kg.

Slope: kg/cm.

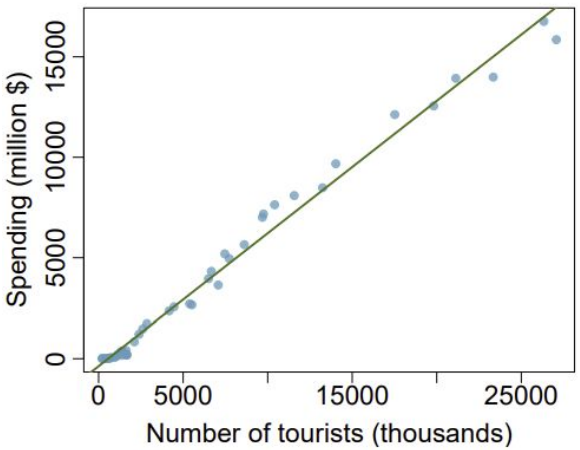
Over-under estimation

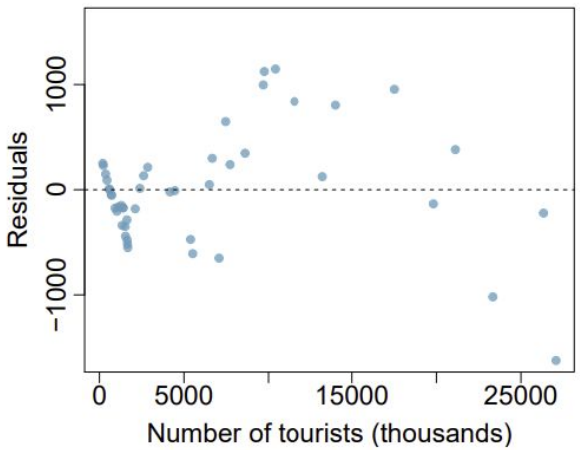
Suppose we fit a regression line to predict the shelf life of an apple based on its weight. For a particular apple, we predict the shelf life to be 4.6 days. The apple's residual is -0.6 days. Did we over or under estimate the shelf-life of the apple? Explain your reasoning.

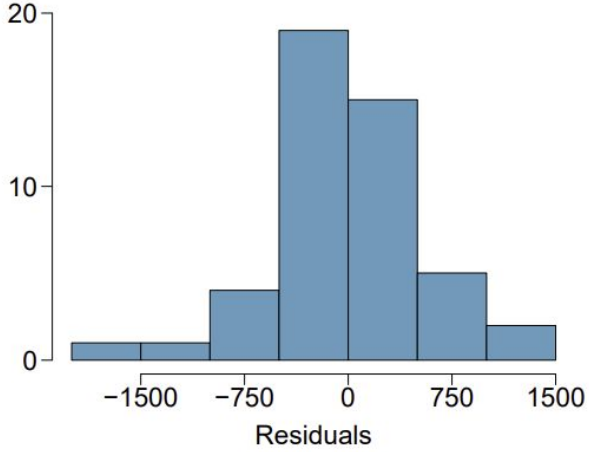
Over-estimate. Since the residual is calculated as *observed − predicted*, a negative residual means that the predicted value is higher than the observed value.

Tourism spending

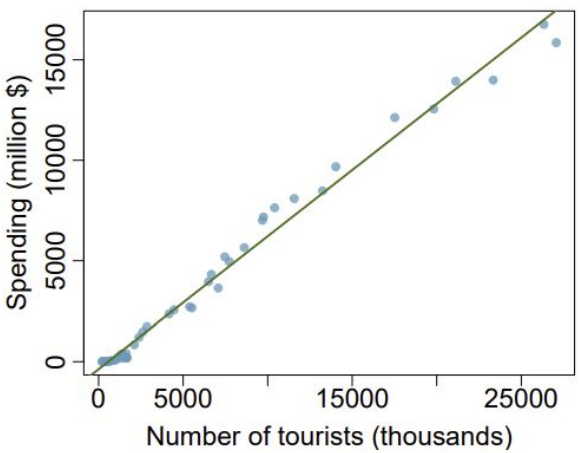
The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year. Three plots are provided: scatterplot showing the relationship between these two variables along with the least squares fit, residuals plot, and histogram of residuals.

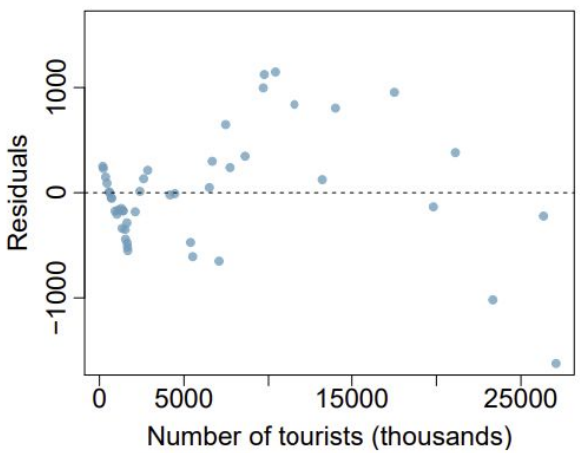


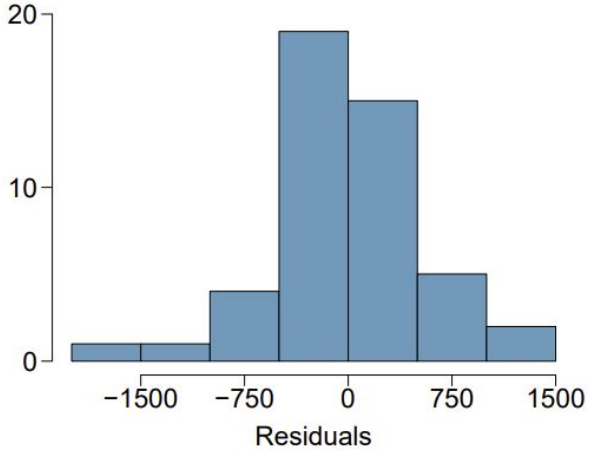




Tourism spending



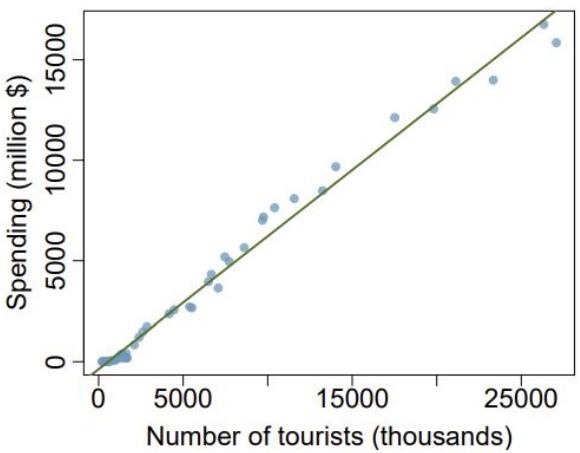


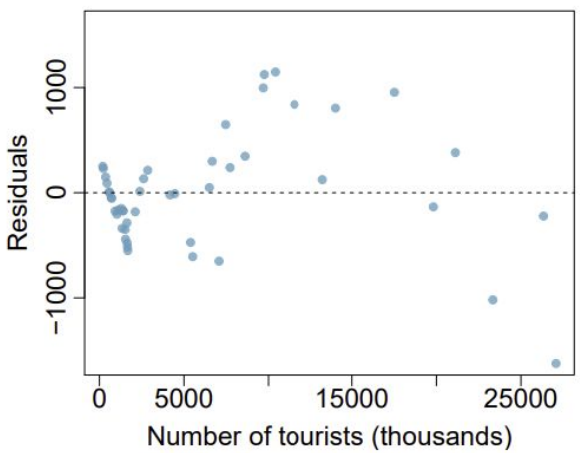


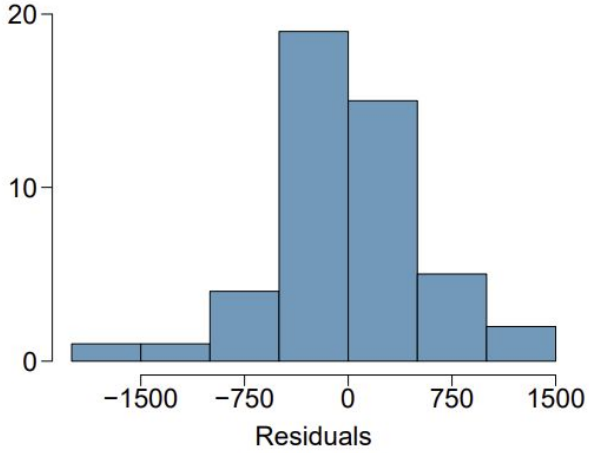
(a) Describe the relationship between number of tourists and spending.

There is a positive, very strong, linear association between the number of tourists and spending.

Tourism spending





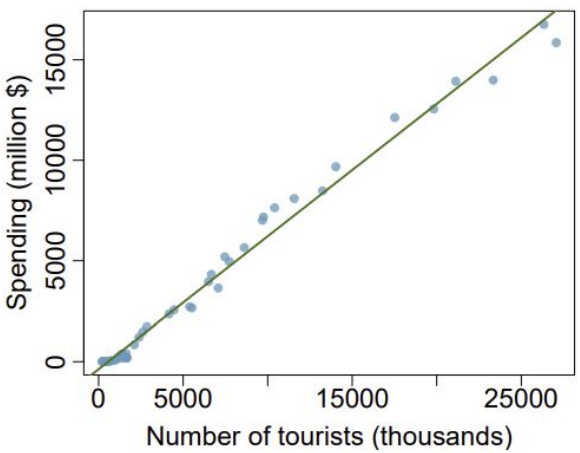


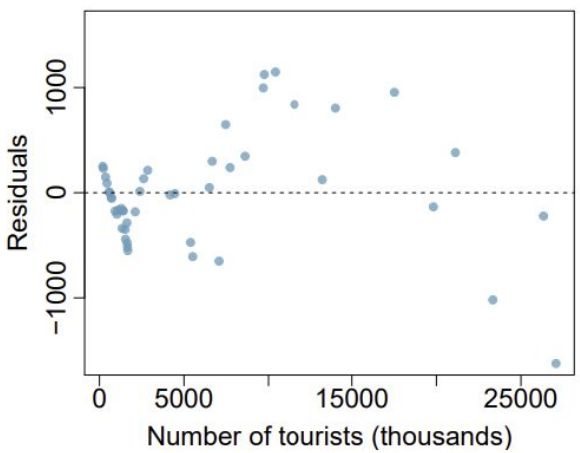
(b) What are the explanatory and response variables?

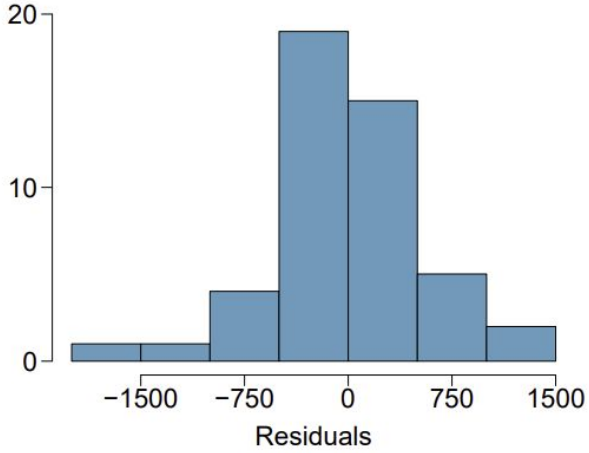
Explanatory: number of tourists (in thousands).

Response: spending (in millions of US dollars).

Tourism spending



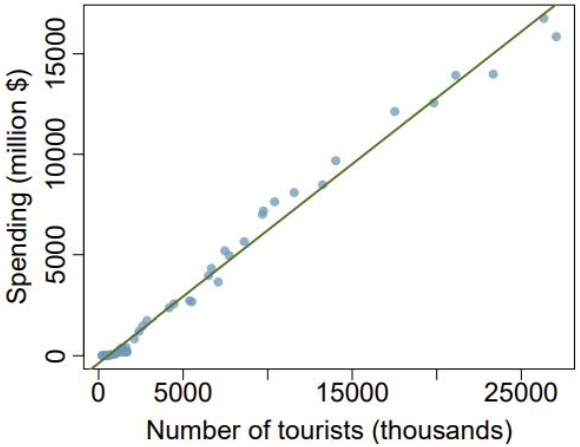


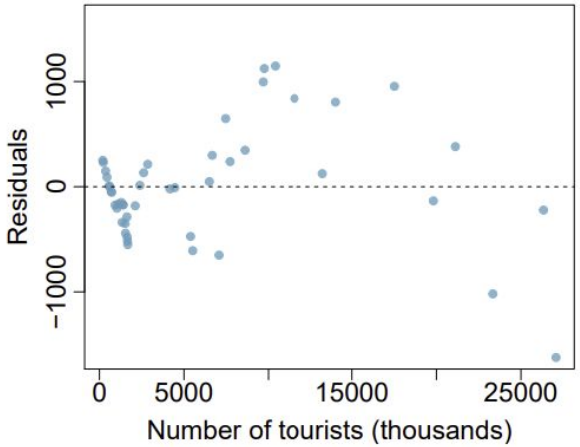


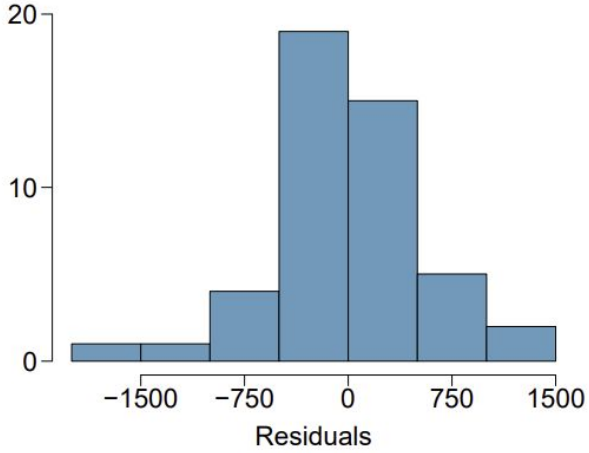
(c) Why might we want to fit a regression line to these data?

We can predict spending for a given number of tourists using a regression line. This may be useful information for determining how much the country may want to spend in advertising abroad, or to forecast expected revenues from tourism.

Tourism spending







(d) Do the data meet the conditions required for fitting a least squares line? In addition to the scatterplot, use the residual plot and histogram to answer this question.

Even though the relationship appears linear in the scatterplot, the residual plot actually shows a nonlinear relationship. This is not a contradiction: residual plots can show divergences from linearity that can be difficult to see in a scatterplot. A simple linear model is inadequate for modeling these data.

The Coast Starlight (continued)

The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 108 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.

(a) Write the equation of the regression line for predicting travel time.

First calculate the slope: $b_1 = R \times s_y / s_x = 0.636 \times 113 / 99 = 0.726$.
Next, make use of the fact that the regression line passes through the point $(\bar{x}, \bar{y})$: $\bar{y} = b_0 + b_1 \times \bar{x}$
Plug in the known values and solve for b_0: 51.

$$\widehat{\text{travel time}} = 51 + 0.726 \times \text{distance}$$

The Coast Starlight (continued)

The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 108 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.

(b) Interpret the slope and the intercept in this context.

$$\widehat{\text{travel time}} = 51 + 0.726 \times \text{distance}$$

b_1: For each additional mile in distance, the model predicts an additional 0.726 minutes in travel time.

b_0: When the distance traveled is 0 miles, the travel time is expected to be 51 minutes. It does not make sense to have a travel distance of 0 miles in this context.

The Coast Starlight (continued)

The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 108 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.

(c) Calculate R^2 of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret R^2 in the context of the application.

$R^2 = 0.636^2 = 0.40$.

About 40% of the variability in travel time is accounted for by the model, i.e. explained by the distance traveled.

The Coast Starlight (continued)

$$\widehat{\text{travel time}} = 51 + 0.726 \times \text{distance}$$

(d) The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.

Predicted travel time = 51 + 0.726 x distance = 51 + 0.726 x 103 = 126 mins

(Note: we should be cautious in our predictions with this model since we have not yet evaluated whether it is a well-fit model.)

The Coast Starlight (continued)

$$\widehat{\text{travel time}} = 51 + 0.726 \times \text{distance}$$

(e) It actually takes the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value.

Residual = observed - predicted

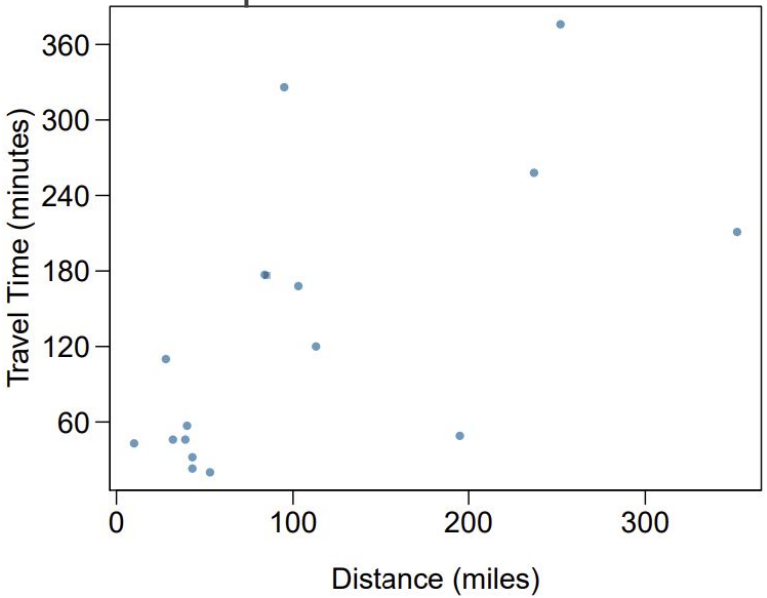
$$e_i = y_i - \hat{y}_i = 168 - 126 = 42 \text{ minutes}$$

A positive residual means that the model underestimates the travel time.

The Coast Starlight (continued)

$$\widehat{\text{travel time}} = 51 + 0.726 \times \text{distance}$$

(f) Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?



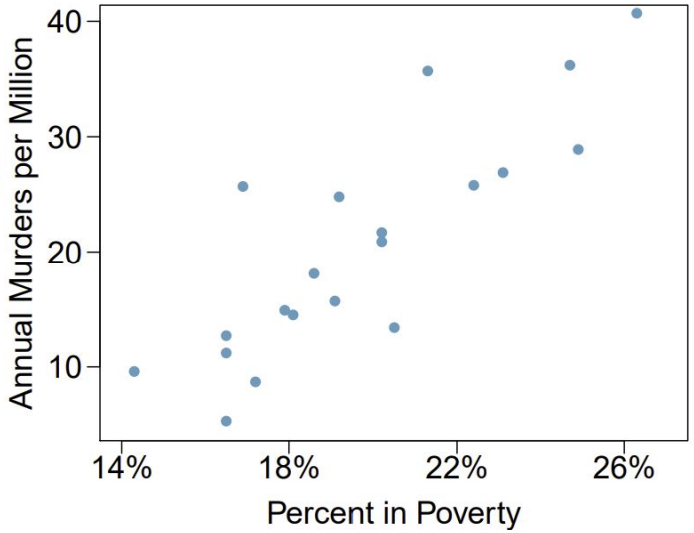
No, this calculation would require extrapolation because the maximum distance in our data is about 300 miles.

Murders and poverty

The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000

$s = 5.512$ $R^2 = 70.52\%$ $R^2_{adj} = 68.89\%$



(a) Write out the linear model.

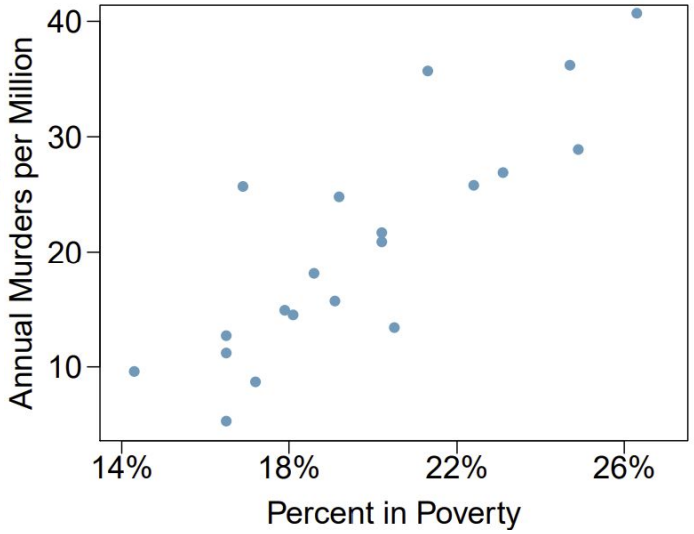
$$\widehat{murder} = -29.901 + 2.559 \times poverty\%$$

Murders and poverty

The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000

$s = 5.512$ $R^2 = 70.52\%$ $R^2_{adj} = 68.89\%$



(b) Interpret the intercept.

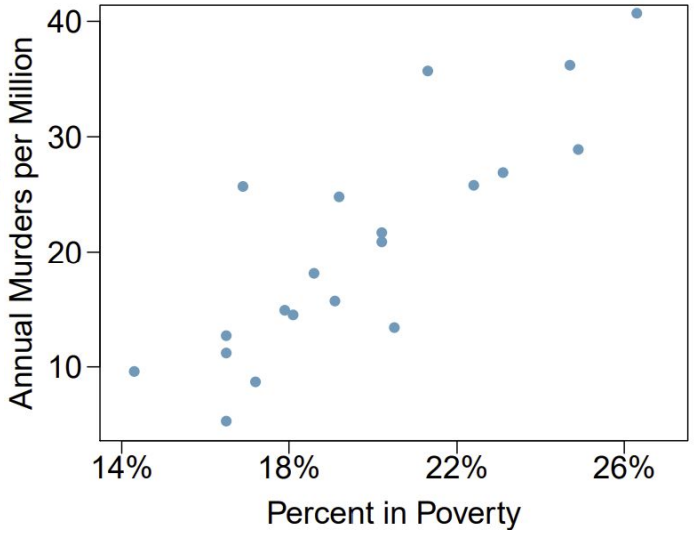
Expected murder rate in metropolitan areas with no poverty is -29. 901 per million. This is obviously not a meaningful value, it just serves to adjust the height of the regression line.

Murders and poverty

The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000

$s = 5.512$ $R^2 = 70.52\%$ $R^2_{adj} = 68.89\%$



(c) Interpret the slope.

For each additional percentage increase in poverty, we expect murders per million to be higher on average by 2.559.

Murders and poverty

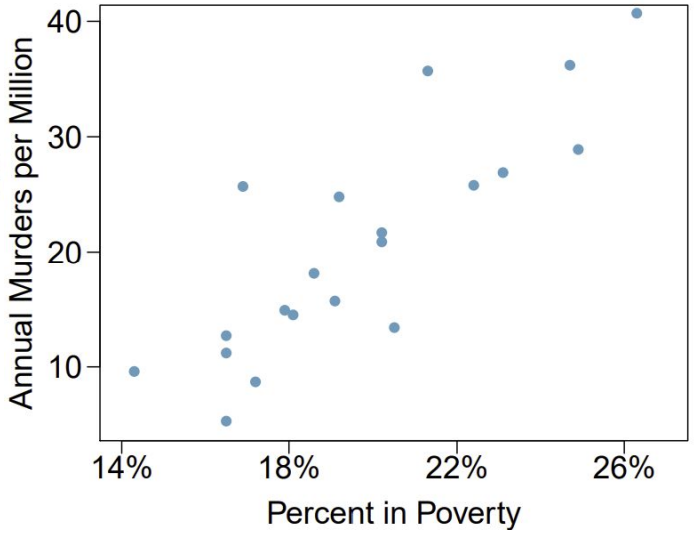
The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000

$s = 5.512$ $R^2 = 70.52\%$ $R^2_{adj} = 68.89\%$

(d) Interpret R^2 .

Poverty level explains 70.52% of the variability in murder rates in metropolitan areas.

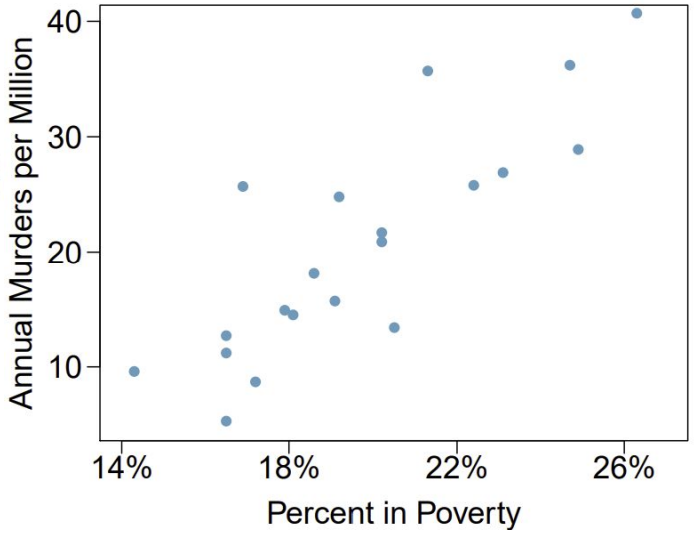


Murders and poverty

The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000

$s = 5.512$ $R^2 = 70.52\%$ $R^2_{adj} = 68.89\%$



(e) Calculate the correlation coefficient.

We want to calculate R and we know R^2

$\sqrt{(0.7052)} = 0.83898$

The data are strongly linearly correlated.

Credits

Examples adapted from OpenIntro Statistics (4th edition) by David Diez, Mine Cetinkaya-Rundel, and Christopher D Barr https://www.openintro.org/book/os/ protected under the Creative Commons License