# PRACTICE PROBLEMS: Cautions about *r2*

Although the *r2* value is a useful summary measure of the strength of the linear association between x and y, it really shouldn't be used in isolation. And certainly, its meaning should not be over-interpreted. These practice problems are intended to illustrate these points.

**Question 1: A large *r2* value does not imply that the estimated regression line fits the data well.**

The American Automobile Association has published data (Defensive Driving: Managing Time and Space, 1991) that looks at the relationship between the average stopping distance ( *y = distance*, in feet) and the speed of a car (*x = speed*, in miles per hour). The data set [carstopping.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/carstopping.txt)contains 63 such data points.

1. Use Minitab to create a fitted line plot of the data. (See Minitab Help Section - [Creating a fitted line plot](https://onlinecourses.science.psu.edu/stat501/node/116)). Does a line do a good job of describing the trend in the data?
2. Interpret the *r2* value. Does car speed explain a large portion of the variability in the average stopping distance? That is, is the *r2* value large?
3. Summarize how the title of this section is appropriate.

**Question 2: One data point can greatly affect the *r2* value**

The [mccoo.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/mccoo.txt)data set contains data on the running back Eric McCoo's rushing yards (*mccoo*) for each game of the 1998 Penn State football season. It also contains Penn State's final score (*score*).

1. Use Minitab to create a fitted line plot. (See Minitab Help Section - [Creating a fitted line plot](https://onlinecourses.science.psu.edu/stat501/node/116)). Interpret the *r2* value, and note its size.
2. Remove the one data point in which McCoo ran 206 yards. Then, create another fitted line plot on the reduced data set. Interpret the *r2* value. Upon removing the one data point, what happened to the*r2* value?
3. When a correlation coefficient is reported in research journals, there often is not an accompanying scatter plot. Summarize why reported correlation values should be accompanied with either the scatter plot of the data or a description of the scatter plot.

**Question 3: Association is not causation!**

Association between the predictor *x* and response *y* should not be interpreted as implying that *x* causes the changes in *y*. There are many possible reasons for why there is an association between *x* and *y*, including:

1. The predictor *x* does indeed cause the changes in the response *y*.
2. The causal relation may instead be reversed. That is, the response *y* may cause the changes in the predictor *x*.
3. The predictor *x* is a contributing but not sole cause of changes in the response variable *y*.
4. There may be a "lurking variable" that is the real cause of changes in *y* but also is associated with *x*, thus giving rise to the observed relationship between *x* and *y*.
5. The association may be purely coincidental.

It is not an easy task to definitively conclude the causal relationships in #1- #3. It generally requires designed experiments and sound scientific justification. #5 is related to Type I errors in the regression setting. The exercises in this section and the next are intended to illustrate #4, that is, examples of lurking variables.

**Drug law expenditures and drug-induced deaths**

"Time" is often a lurking variable. If two things (e.g. road deaths and chocolate consumption) just happen to be increasing over time for totally unrelated reasons, a scatter plot will suggest there is a relationship, regardless of it existing only because of the lurking variable "time." The data set [drugdea.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/drugdea.txt) contains data on drug law expenditures and drug-induced deaths (Duncan, 1994). The data set gives figures from 1981 to 1991 on the U.S. Drug Enforcement Agency budget (*budget*) and the numbers of drug-induced deaths in the United States (*deaths*).

1. Create a fitted line plot treating *deaths* as the response *y* and *budget* as the predictor *x*. Do you think the budget caused the deaths?
2. Create a fitted line plot treating *budget* as the response *y* and *deaths* as the predictor *x*. Do you think the deaths caused the budget?
3. Create a fitted line plot treating *budget* as the response *y* and *year* as the predictor *x*.
4. Create a fitted line plot treating *deaths* as the response *y* and *year* as the predictor *x*.
5. What is going on here? Summarize the relationships between *budget*, *deaths*, and *year* and explain why it might appear that as drug-law expenditures increase, so do drug-induced deaths.

**Question 4: Infant death rates and breast feeding**

The data set [infant.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/infant.txt) contains data on infant death rates (*death*) in 14 countries (1989 figures, deaths per 1000 of population). It also contains data on the percentage of mothers in those countries who are still breast feeding (*feeding*) at six months, as well as the percentage of the population who have access to safe drinking water (*water*).

1. Create a fitted line plot treating *death* as the response *y* and *feeding* as the predictor *x*. Based on what you see, what causal relationship might you be tempted to conclude?
2. Create a fitted line plot treating *feeding* as the response *y* and *water* as the predictor *x*. What relationship does the plot suggest?
3. What is going on here? Summarize the relationships between *death*, *feeding*, and *water* and explain why it might appear that as the percentage of mothers breast feeding at six months increases, so does the infant death rate.

**Question 5: Does a statistically significant*P*-value for H0 : β1 = 0 imply that β1 is meaningfully different from 0?**

Recall that just because we get a small *P*-value and therefore a "statistically significant result" when testing *H*0 : *β*1 = 0, it does not imply that *β*1 will be meaningfully different from 0. This exercise is designed to illustrate this point. The data set [practical.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/practical.txt) contains 1000 (*x*, *y*) data points.

1. Create a fitted line plot and perform a standard regression analysis on the data set. (See Minitab Help Sections [Performing a basic regression analysis](https://onlinecourses.science.psu.edu/stat501/node/130) and [Creating a fitted line plot](https://onlinecourses.science.psu.edu/stat501/node/116)).
2. Interpret the *r2* value. Does there appear to be a strong linear relation between *x* and *y*?
3. Use the Minitab output to conduct the test *H*0 : *β*1 = 0. (We'll cover this formally in Lesson 2, but for the purposes of this exercise reject *H*0 if the *P*-value for *β*1 is less than 0.05.) What is your conclusion about the relationship between *x*and *y*?
4. Use the Minitab output to calculate a 95% confidence interval for *β*1. (Again, we'll cover this formally in Lesson 2, but for the purposes of this exercise use the formula *b*1 ± 2 × *se* (*b*1). Since the sample is so large, we can just use a *t*-value of 2 in this confidence interval formula.) Interpret your interval. Suppose that if the slope *β*1 is 1 or more, then the researcher's would deem it to be meaningfully different from 0. Does the interval suggest, with 95% confidence, that *β*1 is meaningfully different from 0?
5. Summarize the apparent contradiction you've found. What do you think is causing the contradiction? And, based on your findings, what would you suggest you should always do, whenever possible, when analyzing data?

**Question 6: A large r-squared value does not necessarily imply useful predictions**

The [oldfaithful.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/oldfaithful.txt) data set contains data on 21 consecutive eruptions of Old Faithful geyser in Yellowstone National Park. It is believed that one can predict the time until the next eruption (*next*), given the length of time of the last eruption (*duration*).

1. Use Minitab to quantify the degree of linear association between *next* and *duration*. That is, determine and interpret the *r2* value.
2. Use Minitab to obtain a 95% prediction interval for the time until the next eruption if the last eruption lasted 3 minutes. (See Minitab Help Section - [Performing a multiple regression analysis - with options](https://onlinecourses.science.psu.edu/stat501/node/244)). Interpret your prediction interval. (We'll cover Prediction Intervals formally in Lesson 3, so just use your intuitive notion of what a Prediction Interval might mean for this exercise.)
3. Suppose you are a "ratrace tourist" who knows that you can only spend up to one hour waiting for the next eruption to occur. Is the prediction interval too wide to be helpful to you?
4. Is the title of this section appropriate?

# Brief solutions to Practice Problems

**Question 1**

* 1. – The plot shows a strong positive association between the variables that curves upwards slightly
  2. – of the sample variation in can be explained by the variation . This is a relatively large value.
  3. – The value of is relatively high but the estimated regression line misses the curvature in the data

**Question 2**

2.1 – The plot shows a slight positive association between the variables with of the sample variation in explained by the variation in McCoo.

2.2 - decreases to just 7.9% on removal of the one data point with yards.

**Question 3**

* 1. The plot shows a moderate positive association between the variables but with more variation on the right side of the plot.
  2. This plot also shows a moderate positive association between the variables but with more variation on the right side of the plot.
  3. This plot shows a strong positive association between the variables.
  4. This plot shows a moderate positive association that is very similar to the *deaths vs budget* plot.
  5. Year appears to be a lurking variable here and the variables *deaths* and *budget* most likely have little to do with one another.

**Question 4**

* 1. The plot shows moderate positive association between the variables, possibly suggesting that as *feeding* increases, so too does *death*.
  2. The plot shows a moderate negative association between the variables, possibly suggesting that as *water* increases, so too does *feeding*.
  3. Higher values of *water* tend to be associated with lower values of both *feeding* and *death*, so low values of *feeding* and *death* tend to occur together. Similarly, lower values of *water* tend to be associated with higher values of both *feeding* and *death*, so high values of *feeding* and *death* tend to occur together. *Water* is a lurking variable here and is likely the real driver behind infant death rates.

**Question 5**

* 1. The fitted regression equation is
  2. of the sample variation in can be explained by the variation in There appears to be a moderate linear association between the variables.
  3. The -value is 0.000 suggesting a significant linear association between and
  4. The interval is 0.09980 2(0.0058) or (0.0882, 0.1114). Since this interval excludes the researchers’ threshold of 1, is not meaningfully different from 0.
  5. The large sample size results in a sample slope that is significantly different from 0, but not meaningfully different from 0. The scatterplot, which should always accompany a somple linear regression analysis., illustrates.

**Question 6**

6.1 of the sample variation in next can be explained by the variation in duration. There is a relatively high degree of linear association between the variable

6.2 The 95% prediction interval is (47.2377, 73.5289), which means we’re 95% confident that the time until the next eruption, if the eruption lasted 3 minutes, will be between 47.2 and 73.5 minutes.

6.3 If we can only wait 60 minutes, this interval is too wide to be helpful to us since it extends beyond 60 minutes.