CHAPTER SEVEN: LINEAR CORRELATION

Review March 28 ~ Project April 20

A line of best fit is a line that best fits the points in a data set. It is used to model real-world phenomena and make predictions. However, it is important to keep in mind that just because two variables are correlated does not necessarily mean that one affects the other, nor that one can precisely predict the other.

7-A The Line of Best Fit

Monday • 3/20

model • linear regression • line of best fit • interpolation • extrapolation • residual

1. Find the equation of the line of best fit for a data set.
2. Use a line of best fit equation to predict a value of the dependent variable.
3. Calculate a residual for a point on a regression line.

7-B Statistically Significant Correlations

Thursday • 3/23

data • sample • correlation coefficient • population • p value • statistically significant • data snooping

1. Calculate r for a data set, and interpret the value.
2. Determine whether or not the correlation found in a sample is statistically significant.

7-C Causal Relationships

Tuesday • 3/28

affect • effect • causal relationship • confounding variable

1. Distinguish between *affect* and *effect* meaning *influence*.
2. Discuss possible reasons for correlation in a sample.
3. Critically evaluate a causal claim from a sample correlation.

7-A The Line of Best Fit

A Mathematical MODEL is an equation representing a real-world phenomenon. It is used to make predictions about data values.

In some cases, especially in physical sciences like physics, models are developed based on theoretical calculations. In other cases, especially in behavioral sciences like psychology, models are developed from sample data. This is done by REGRESSION, which is a process to find the line or curve through a set of data points that comes closest to all of the data points. Many different types of regression can be done, the most common of which is LINEAR Regression. The line found for a data set by linear regression is called the LINE OF BEST FIT.

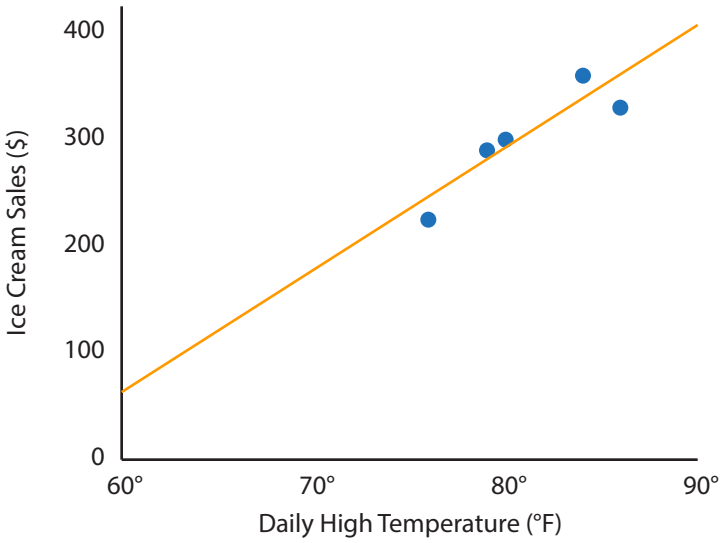
For a line of best fit, the variable $\hat{y}$ (y-hat) is often used in place of y to show that it represents predicted values rather than actual data values.

❶ Find the equation of the line of best fit for a data set.

1. Push [STAT] and choose `EDIT`.
2. Clear any previous data by moving the cursor onto `L1` and pushing [CLEAR] and [ENTER].
3. Type each x value, followed by [ENTER] each time.
4. Repeat steps 2 and 3 in `L2` and with the corresponding y values.
5. Push [STAT] and choose `TESTS`.
6. Choose `LinRegTTest...`.
7. Choose `Calculate`.
8. The equation of the regression line is $\hat{y} = bx + a$. Write this equation using the a and b shown on the calculator.

❶ daily high temperature (°F): 76 79 86 84 80
ice cream Annika sells ($): 224 289 330 360 299

8. `b=11.453125` `a=-627.303125` $\hat{y} = 11.5x - 627$



A line of best fit can be used to predict y-values.

INTERPOLATION is predicting a y value for an x value within the range of x values in the sample.

EXTRAPOLATION is predicting a y value for an x value outside of the range of x values in the sample. Extrapolation often leads to poor or even impossible predictions, because the trend found in the data may not continue outside of the range studied.

❷ Use a line of best fit equation to predict a value of the dependent variable.

1. Plug the specified value of x into the line of best fit equation and calculate $\hat{y}$.
2. If the specfied value of x is not between the lowest and highest values of x in the data set, be aware that the predicted value may be off by a huge amount.

❷ Predict Annika's ice cream sales for days with the following high temperatures.

a) 78°
1. $\hat{y} = 11.5(78) - 627 = \$270$
2. interpolation
(78 is between 76 and 86.)

b) 86°
$\hat{y} = 11.5(86) - 627 = \$362$
interpolation
(86 is the highest x value in the sample.)

c) 50°
$\hat{y} = 11.5(50) - 627 = \$\text{-}52$ (not realistic)
extrapolation
(50 is lower than all the x's in the sample.)

A RESIDUAL is the difference between an actual data value and the value predicted by a regression equation. A line of best fit is defined as the line with the smallest sum of squared residuals.

❸ Calculate a residual for a point on a regression line.

1. Calculate the predicted value $\hat{y}$ (see ❷).
2. Subtract the predicted value $\hat{y}$ from the actual value y.

❸ Calculate Annika's residual for 86°.

1. $\hat{y} = \$362$ (see ❷b)
2. The residual is $330 - 362 = -32$.

7-B Statistically Significant Correlations

DATA is a plural word that essentially means *observed values*.

A SAMPLE is the group from which data are collected.

How closely two variables are correlated in a sample is measured by the Linear CORRELATION COEFFICIENT r. r ranges from -1, representing a perfect negative correlation, to 1, representing a perfect positive correlation. A correlation coefficient of 0 would represent no correlation.

A positive correlation means higher values of x tend to correspond with higher values of y, and lower values of x tend to correspond with lower values of y.

A negative correlation means higher values of x tend to correspond with lower values of y, and lower values of x tend to correspoind with higher values of y.

❶ Calculate r for a data set, and interpret the value.

1. Do steps 1-7 in ❶ in 7-A.
8. If $r > 0$, the correlation is positive. If $r < 0$, the correlation is negative.
9. If r is close to 0, the correlation is weak, meaning the two variables are not closely related.

❶ Calculate the correlation coefficient for Dallis's data, and explain what it means.

1. $r = .90$
8. The correlation is positive, meaning higher temperatures tend to correspond with higher ice cream sales, and vice versa.
9. r is almost as far from 0 as possible, meaning the relationship between temperature and ice cream sales is very strong.

A POPULATION is the entire group intended to be represented by the sample. The population does not necessarily consist only of things that currently exist. For example, the population of coin flips includes all coin flips that have ever happened and that ever will happen.

r is calculated from sample data to estimate the correlation coefficient for the whole population. Due to the possibility of coincidence in the sample, sample data can never *prove* anything about a population. In particular, if r is close to zero or the sample size is small, any correlation found in the sample could easily be a coincidence. On the other hand, the farther r is from zero and the larger the sample size is, the more confident we can be that a correlation exists in the population overall; nevertheless, at no point is this possibility "proven."

The *P* VALUE for a sample correlation is the probability that another random sample of the same size would coincidently show as strong of a correlation in the predicted direction, given the two variables are not actually correlated. If a predicted correlation has a very small p value, the correlation in the sample was not likely to occur coincidentally, meaning the two variables are likely to be correlated beyond just the sample. Specifically, if $p < .05$, the data are considered STATISTICALLY SIGNIFICANT, meaning the sample data are strong enough that conclusions can be drawn about the relationship between the variables for the whole population, not just for the sample (although it is still possible that these conclusions are incorrect). If $p > .05$, the data are not statistically significant, and no predictions or conclusions about the population should be made.

P values only have meaning for relationships that were predicted before the data were known. Calculating p values without establishing specific hypotheses beforehand is DATA SNOOPING, and frequently leads to invalid conclusions.

P values for sample correlations are found with a calculator. The sample size needed to reach statistical significance (that is, for p to be below .05) depends on how strong the correlation is in the sample, and is approximately $n = \frac{3}{r^2}$.

❷ Determine whether or not the correlation found in a sample is statistically significant.

1. Prior to knowing the sample data, state whether the correlation is expected to be positive or negative, and justify the prediction.
2. Do steps 1-6 in ❶ in 7-A.
7. Highlight `>0` if you predicted a positive correlation, or highlight `<0` if you predicted a negative correlation. If you separately justified both possible directions (which is not common), highlight `≠0`.
8. Choose `Calculate`.
9. Identify the p value. If $p < .05$, the data are statistically significant, meaning conclusions can be drawn and predictions can be made.

❷ Are Dallis's data statistically significant?

1. We predict a positive correlation, because people like ice cream more when it is hot.
7. `>0`
9. $p = .019 < .05$, so her data are statistically signficant. It is appropriate for Dallas to conclude that there is a positive correlation between temperature and ice cream sales and for her to use her line of best fit to make predictions about how much she will sell based on temperature.

7-C Causal Relationships

Thursday • 3/26

The words AFFECT and EFFECT both can mean influence. In this case, *affect* is a verb and *effect* is a noun.

❶ Distinguish between *affect* and *effect* meaning *influence*.

1. If it is preceded by an adjective or article (*the*, *an*, *some*, *three*, *significant*, etc.), it is a noun: effect.
2. If it has a subject, it is a verb: affect. The subject will usually be an independent variable or a confounding variable (for example, *caffeine*, *study habits*, or *socioeconomic status*).
3. Only *affected* (not *effected*) means *influenced*, and only *affecting* (not *effecting*) means *influencing*.

❶ Alcohol __ffects fine motor control. Nausea, amnesia, and unconsciousness are also possible __ffects. Smaller people are particularly __ffected, although anyone can experience the __ffects of alcohol.

a) has the subject *alcohol*: affects
b) preceded by the adjective *possible*: effects
c) means *influenced*: affected
d) preceded by the article *the*: effects

A CAUSAL Relationship is one in which a correlation is due to the effect the independent variable has on the dependent variable.

A CONFOUNDING Variable is a variable that creates a possible explanation for the relationship between the independent variable and the dependent variable that is not causal, that is, that does not involve the independent variable having any effect, directly or indirectly, on the dependent variable. Due to confounding variables, correlation does not imply causation.

❷ Discuss possible reasons for correlation in a sample.

1. The correlation could be just due to a coincidence in the sample, and no actual relationship exists. This is likely if $p > .05$.
2. The relationship could involve causation: The independent variable affects the dependent variable (as originally assumed).
3. The relationship could involve reverse causation: The dependent variable affects the independent variable.
4. The relationship could be due to one or more confounding variables: Outside variables affect the independent variable and the dependent variable simultaneously.

❷ Identify possible reasons for the correlation between exercise and health.

1. It is not likely that this is just a coincidence, since many samples have shown a significant correlation.
2. This relationship is likely to be causal because exercise causes good health, such as by burning calories, boosting endorphins and good cholesterol, and aiding distribution of oxygen and nutrients throught the body.
3. This relationship is likely to involve reverse-causation, because good health causes people to be more able and motivated to exercise.
4. This relationship may be partially due to confounding variables. For example, people who live in poor inner-city areas breathe in more pollutants and have less access to and money for healthful foods and health care, and these same people tend to have less access to and cultural precident for athletics and fitness programs and routines.

A low p value does not automatically mean that two variables are correlated in a meaningful way. There are four things to conisder when evaluating a sample correlation.

1. Is the sample representative of the population?
2. How strong is the correlation? That is, how far is r from 0?
3. Is the correlation unlikely to be a coincidence? This is, was justification given beforehand and $p < .05$?
4. Are there established reasons the relationship may be causal? Or are there confounding variables that could be the actual reason the two variables studied are correlated?

❸ Critically evaluate a causal claim from a sample correlation.

1. Consider the sampling method. If every member of the population has an equal chance of being included in the sample, then the sample is appropriate, but if the sample is collected by convenience or otherwise does not represent the population overall, then it is important to consider how the sample used limits the extent to which r represents the population.
2. Consider how close r is to 1 or -1. If r is close to zero, the correlation may not have much practical significance even if it is a legitimate correlation in the population.
3. If $p > .05$, or if there was no justification for the correlation given prior to data collection, the correlation is likely to be a coincidence.
4. If there are confounding variables that could justify part or all of the correlation between the variables, this makes it impossible to determine how much of an effect, if any, the independent variable actually has on the dependent variable.

❸ From 1999 to 2010, the number of lawyers in Georgia has been positively correlated with the annual number of Americans who die by becoming tangled in their bed sheets: $r = .961$, $p = .0000003$.

1. The years 1999 to 2010 are not random, but this is not a major concern since there is probably nothing different about these years than others with respect to Georgia lawyers or death by bed sheets.
2. r is very close to 1. The relationship is very strong.
3. $p < .05$, but because no prediction was made beforehand, this p value is meaningless. (And clearly, a direct connection between these two variables would be nonsensical.)
4. Any amount of this correlation that is not coincidental is likely due to the confounding variable of population size: As population increases over time, there are more lawyers and there are also more accidental deaths.