Homework for Chapter 4: Describing Relationships

1. What is a conditional distribution?

A conditional distribution is the distribution of one variable given the value of another variable.

1. The following figure (using fictional data) describes the relationship between Income level and rating on a scale testing for signs of Depression.  
   Chart, scatter chart

   Description automatically generated
   1. How does the conditional mean of Depression change as Income increases?

The conditional mean of Depression decreases as Income increases.

* 1. Does the graph indicate that lower income causes depression? Why or why not?

No. Though from the graph, higher income levels are associated with lower ratings on the scale testing for signs of Depression, we don’t know if there is any other factor that affects both of the variables, namely if the negative relationship between income levels and depression can be entirely explained by an unknown variable.

1. The below fictional table depicts data collected from 3000 university students on their classification (Freshman, Sophomore, Junior, Senior) and whether or not they receive financial aid. The table shows a cross tabulation of classification and receipt of financial aid.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Financial Aid | Freshman | Sophomore | Junior | Senior |
| Yes | 508 | 349 | 425 | 288 |
| No | 371 | 337 | 384 | 338 |

* 1. Calculate the probability of receiving financial aid given that a student is a Senior. 288/(288+338) = 0.4600639
  2. Calculate the probability that a student is a Senior given that they receive financial aid. 288/(508+349+425+288) = 0.1834395
  3. Calculate the probability of receiving financial aid given that a student is a Freshman. 508/(508+371) = 0.5779295

1. Describe two advantages and one disadvantage of using line-fitting methods as opposed to calculating local means.

Unlike calculating local means, line-fitting methods can give us the conditional mean of Y for any value of X we can think of, even if we don’t have data for that specific value. Also, since the line is estimated using all the data, rather than just local data, the results are more precise. However, fitting a line requires us to pick a kind of shape of the relationship ahead in order for the line-fitting procedure to pick the best version of the shape we give it to fit a line. If the shape we pick is all wrong to start with, our estimate of the conditional mean will be all wrong.

1. Consider the line of best fit: .
   1. What is the conditional predicted mean of when ? = -17
   2. What is the conditional predicted mean of when ? 4 - 3.5\*(-2) = 11
2. Which of the following terms describes a measurement of how much two variables vary with each other? c
   1. Variance
   2. Conditional mean
   3. Covariance
   4. Local mean
3. What is the difference between covariance and correlation?

Covariance measures how much two variables vary with each other, whereas correlation describes that how the change in one variable results in the change in another variable. Covariance indicates the direction of the linear relationship between variables. Correlation measures both the strength and direction of the linear relationship between two variables. The correlation is achieved by taking the covariance between X and Y and dividing by both the standard deviation of X and the standard deviation of Y. The scale of the variables can affect their covariance, but does not affect their correlation. Covariance has a definite unit based on the units the original variables were in, but the correlation is a number without units. While the value of the covariance can be any real number, the values of the correlation can only range from -1 to 1.

1. Figure A and Figure B below depict the (fictional-data) relationship between scores on a math exam and an intelligence measure from data collected from a fictional sample of 100 students.   
   Chart, scatter chart

   Description automatically generated
   1. What type of shape is fitted in Figure A? A straight line (linear regression).
   2. What kind of shape is fitted in Figure B? A curve/parabola (quadratic regression).
   3. Which shape is a better fit for the data, and how can you tell?

The second shape, the parabola, is a better fit for the data because the pattern in the data points resembles a parabola rather than a line. It shows that the intelligence score is not increasing at a steady rate, but first decreases as the math exam score increases and then increases with the math exam score. Also, the total deviation between the actual values and the predictions from figure B is smaller than that from figure A: the parabola has a lower sum of squared residuals.

* 1. For Figure A, describe the residuals for different ranges of math exam scores. Are the observed data below or above the line/curve? Are the residuals evenly scattered around the line/curve?

When the math exam score is below -1, almost all of the observed data are above the line, making the residuals for this range of math exam scores positive (the residuals are not evenly scattered around the line). When the math exam score is around -1, the observed data are either below or above the line and the residuals are evenly scattered around the line. When the math exam score is between -1 and 1, almost all of the observed data are below the line, making the residuals for this range of math exam scores negative (the residuals are not evenly scattered around the line). When the math exam score is around 1, the observed data are either below or above the line and the residuals are evenly scattered around the line. When the math exam score is above 1, all the observed data are above the line, making the residuals for this range of math exam scores positive (the residuals are not evenly scattered around the line).

1. The below table contains fictional data on 5 employees from a company, repotting on how well they get along with their coworkers (GetAlong) and their level of job satisfaction (Satisfaction). The Prediction variable is the predicted satisfaction level, or the conditional mean of satisfaction, for each value of GetAlong derived after fitting a line of best fit using ordinary least squares estimation.

|  |  |  |  |
| --- | --- | --- | --- |
| GetAlong | Satisfaction | Prediction | Residual |
| 4.7 | 5.07 | 4.72 | 0.35 |
| 4.21 | 4.05 | 4.28 | -0.23 |
| 5.42 | 5.33 | 5.38 | -0.05 |
| 4.14 | 4.02 | 4.22 | -0.20 |
| 3.3 | 3.59 | 3.45 | 0.14 |

* 1. Fill in the “residual” column.
  2. Describe how ordinary least squares uses residuals when fitting a line.

Ordinary Least Squares (OLS) picks the line that gives the lowest sum of squared residuals. OLS takes the value of the difference between each of the observations’ actual values and the conditional means assigned by the line, squares it, then adds up all the predictions across all the data. Then OLS picks the values of β0 and the β1 in the line Y = β0+ β1X that make that sum of squared residuals as small as possible by taking the covariance of X and Y and divides it by the variance of X to determine the slope β1, and then picking an intercept β0 for the line that makes the mean of the residuals 0.

1. We’ll be thinking here about the process of controlling for a variable. Consider the example: What is the relationship between first generation status and graduation rate in a population of college students?
   1. Give an example of a variable that might explain why first-generation status and graduation rate are correlated other than one causing the other.

There might be a negative correlation between first generation status and graduation rate in a population of college students: The graduation rate is lower for college students with first-generation status than those who are not the first generation. But family income might be an extraneous variable that explains why first-generation status and graduation rate are negatively correlated. Lower family income may more likely lead to the first-generation status of a college student in that for students with lower family income, it is less likely that their parents have a bachelor's degree or above which makes the students more likely to be the first generation college students. In the meantime, lower family income may decrease students’ graduation rate in that students may not be able to afford the four years’ tuitions or to finish the four years’ study while working at the same time to pay for their study and dependents. In this case, it is actually the family income that creates the negative association between first generation status and graduation rate.

* 1. Describe in five steps how you would subtract out the part of the first-generation/graduation-rate relationship that is explained by the variable you listed in part a.

If we want to remove the part of the first-generation/graduation-rate relationship that is explained by a student’s family income, we can first get the mean of the first-generation status conditional on family income, and then take the residual, getting only the variation in first-generation status that has nothing to do with family income. Then we would take the mean of graduation rate conditional on family income, and take the residual, getting only the variation in graduation rate that has nothing to do with family income. Finally, we can thus get the mean of the graduation-rate residual conditional on the first generation-status residual.

* 1. How would you interpret the first-generation/graduation-rate relationship you get after performing the steps in part b?

If the graduation rate mean doesn’t change much conditional on different values of first-generation status, then the entire relationship is explained by family income. If the graduation rate mean does still change much conditional on different values of first-generation status, maybe there is a relationship between first generation status and graduation rate, and we can do further research.