STAT 109 Lecture Notes

9 Linear Models

In this chapter, we will explore the relationship between **two paired quantitative variables**. Nothing in Chapter 9 makes sense out of this context.

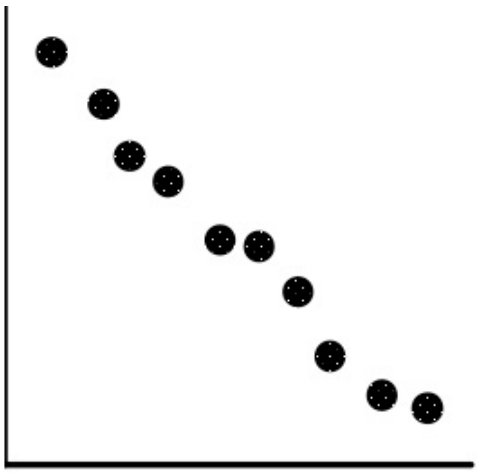
9.1 Correlation

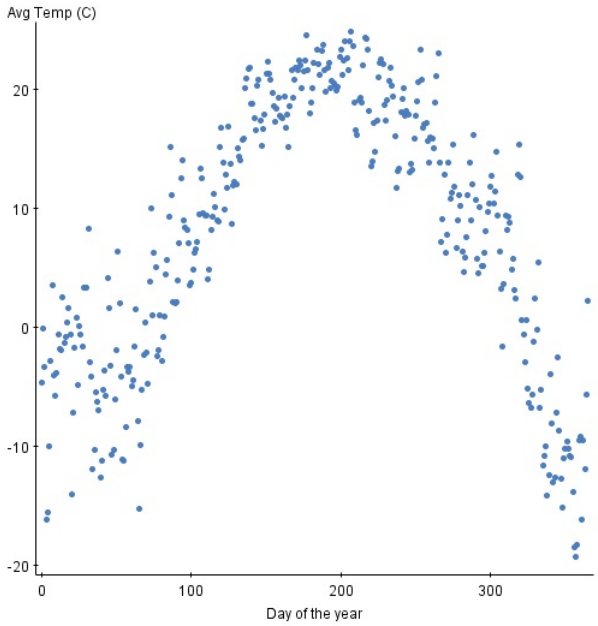
Recall from Section 2.2 that a scatterplot is a visual representation comparing two quantitative variables. The two variables must make a *paired data set*, meaning that each value of one variable corresponds to a value of the other, and both of these values "come from" the same individual subject. The variable plotted on the x (horizontal) axis is usually called the independent, or **explanatory**, variable. The variable plotted on the y (vertical) axis is usually called the dependent, or **response**, variable.

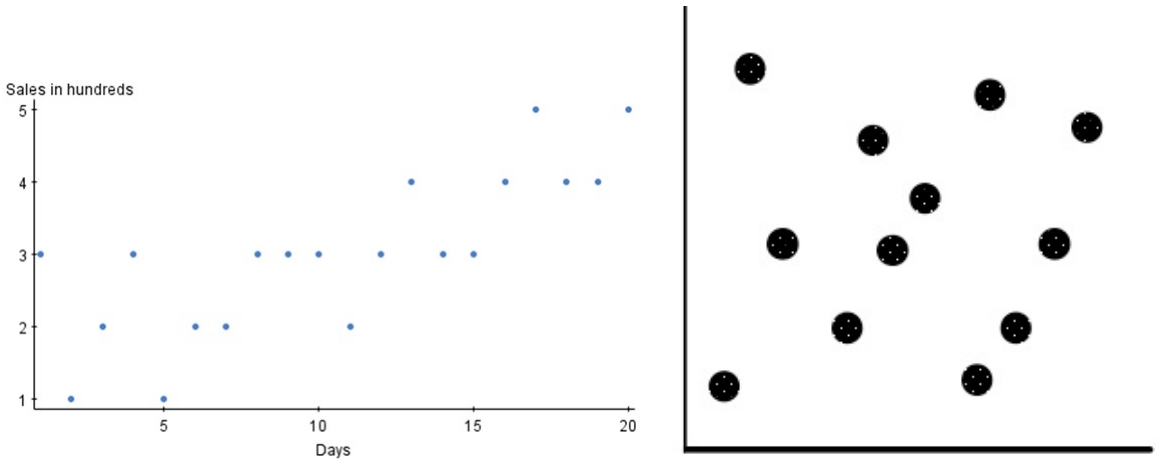
Remember describing the shape of a distribution? Our major points of interest were modes, symmetry, and skewness. We will have certain qualities of interest for scatterplots as well that describe the association of the variables.

- **Direction:** If the dots in the scatterplot tend upward as you move from left to right, the association is positive. If the dots in the scatterplot tend downward as you move from left to right, the association is negative.
- **Form:** If the dots appear as a cloud or swarm of points stretched out in a generally consistent, straight form, we say the association is linear. Otherwise, we say it is nonlinear.

Example 9.1: Describe the association of each scatterplot shown below.







Definition 1: Correlation is a numerical value which measures the strength of a linear association of two quantitative variables.

- Correlation is always between -1 and 1.
- If correlation is negative, the association is negative. If correlation is positive, the association is positive.
- If correlation is close to 1 or -1, the association is **strong and linear**. This means the dots in the scatterplot will be almost in a straight line.
- If correlation is close to 0, then the association is **weak or nonlinear**.

This numerical value is also called the correlation coefficient. A sample correlation coefficient is denoted by r, and a population correlation coefficient is denoted by ρ.

There are several (equivalent) ways of writing the formula for the sample correlation coefficient r of the quantitative variables x and y. Here are a few:

$$r = \frac{\sum(z_x z_y)}{n-1} = \frac{\sum[(x-\overline{x})(y-\overline{y})]}{(n-1)(s_x)(s_y)} = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{n\sum(x^2) - (\sum x)^2}\sqrt{n\sum(y^2) - (\sum y)^2}}$$

In the above equation, (x, y) must be a pair of data, and each sum is over all data pairs. There are n data pairs. The mean and standard deviation of x are $\overline{x}$ and s_x, respectively, and the mean and standard deviation of y are $\overline{y}$ and s_y, respectively. The z-scores of x and y are z_x and z_y, respectively.

The first formula is useful if you've already computed the z-scores of your x and y values. The middle formula is useful if you've already computed the mean and standard deviation of x and y. The last formula is the most complicated, but it doesn't require anything but the x and y values themselves.

Example 9.2: Calculate the correlation of the data set below by hand.

x	y
6	8
8	2
11	5
16	3
17	-1
29	-3

Example 9.3: Below is a table displaying data from our first day of class survey. In Chapter 2, we created a scatterplot for this data. Now compute the correlation.

Height (in)	Hours sleep per night	Height (in)	Hours sleep per night
75	7	66	7
75	6	66	7
67	8	66	6
64	7	71	8
62	7	74	7.5
74	9	69	7
71	6	64	6
67	8	73	6
73	8	73	6
66	8	63	11
68	7	60	7
73	7	68	7
67	6	63	9

Example 9.4: The table below shows the number (in thousands) of licensed drivers in different age groups by gender and number of fatal crashes in each age group by gender.

Age Range	Number of Male Drivers	Number of Fatal Crashes by Male Drivers	Number of Female Drivers	Number of Fatal Crashes by Female Drivers
< 16	12	227	12	77
16-20	6424	5180	6139	2113
21-24	6941	5016	6816	1531
25-34	18068	8595	17664	2780
35-44	20406	7990	20063	2742
45-54	19898	7118	19984	2285
55-64	14340	4527	14441	1514
65-74	8194	2274	8400	938
> 75	4803	2022	5375	980

1. Use Excel to produce a scatter diagram that shows both male and female data on the same plot. Describe the association of each data set.

2. Using Excel, compute the correlation coefficient between number of licensed drivers and number of fatal crashes for each gender.

3. Which gender has a stronger linear relationship between number of licensed drivers and number of fatal crashes?

4. Does this data justify auto insurance companies' choice to charge different rates based on gender?

Interpreting Correlation

- When r is positive, the association is positive. When r is negative, the association is negative.
- Correlation is always between -1 and 1. (If you get a value outside of this range, check for errors!)
- When the correlation is close to 1 or -1, the association is strong and linear. When the correlation is close to zero it is weak and linear OR it is nonlinear. Correlation measures linearity. So, two variables can have a strong association but low correlation if their association is not linear.
- Correlation is symmetric with respect to x and y. That means the correlation of the scatterplot (x, y) is the same as the correlation of (y, x). (If you flip the x and y values in the previous examples and re-compute, you will get the same r!)
- Correlation has no units. It does not maintain the units of either variable, and it is not a percentage, nor is it measured in "correlations."
- Correlation is highly sensitive to outliers. It can be useful to report correlation including the outliers, then again excluding the outliers.
- **Strong correlation does not imply causation.** For example, suppose we record the heights and weights of all the children in an elementary school, and find that the correlation is 0.91. This does not mean that being tall *causes* a child to be heavy, or that being short *causes* a child to be light. There can be a relationship between the two variables that is not causal. We may be able to explain changes in BOTH of the variables by some third lurking variable (like, in this example, age).
- **Correlation is only a useful statistic when (1) both variables are quantitative, (2) the association appears to be linear, and (3) there don't appear to be any outliers in the data.**

Comment: You are not responsible for the information on pages 476-479 in the book. This requires parts of Chapters 6 and 7 that we skipped.

9.2 Linear Regression

Definition 2: When a scatterplot appears to be linear, we would like to apply a linear model to the data. The least squares line, aka line of best fit, aka least squares regression line, is the line that most closely approximates the data with the equation of a line.

The linear model will predict the value of the response variable y given values of the explanatory variable x. Predicted values will be denoted by $\widehat{y}$. **This model is only appropriate when we are comparing two quantitative variables with a linear association.**

The equation of a line in the Cartesian plane is given by $y = mx + b$, where m is the slope and b is the point where the line crosses the y-axis. We will use this same type of equation.

The slope of the line that fits the data best will depend on the correlation (how tightly the data appears to be linear), and the variation of the actual data points from the line (the deviations - or standard deviation - of the data).

In general, the slope depends on the changes in y and the changes in x:

$$m = \frac{\Delta y}{\Delta x}$$

The regression line slope is given by the below formula:

$$m = r\frac{s_y}{s_x}$$

The units for the slope is: "y-units per x-units." For example, if we consider the last example in the previous section, the slope is given in "fatal crashes per driver." Think about it: r has no units, s_y has the units of the y variable, and s_x has the units of the x variable. All together, it's $\frac{y-\text{unit}}{x-\text{unit}}$, or, in words, "$y$-unit per x-unit."

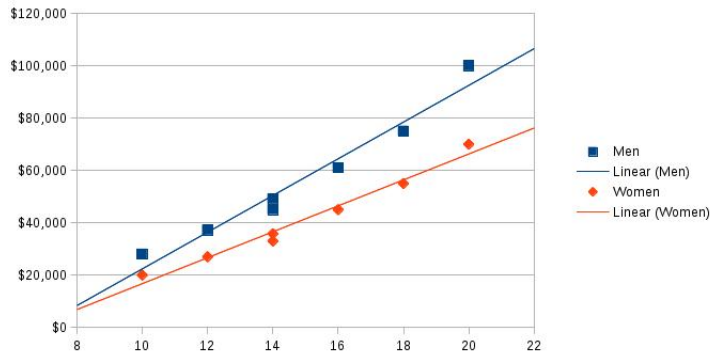
Note: Your book gives the formula below for the slope of the regression line. I will not use this formula in class, but you may use it if you like.

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

The regression line always goes through the point $(\overline{x}, \overline{y})$, so we can use algebra to find the y-intercept of the line:

$$b = \overline{y} - m\overline{x}$$

Example 9.5: Below is the scatterplot of years of education with median salary in 2007. The data is separated by gender.



Men: $\overline{x} = 13.41$, $\overline{y} = 47516$, $s_x = 3.25$, $s_y = 16935$, and $r = 0.7865$

Women: $\overline{x} = 13.38$, $\overline{y} = 34139$, $s_x = 3.01$, $s_y = 11011$, and $r = 0.7937$

1. Which variable is the explanatory variable? Which is the response variable?

2. Does the data appear to be linear for the men? for the women?

3. Compute the equation of the regression line for the men and for the women.

4. Suppose an individual has 17 years of education. Use the linear model to predict that person's salary if he is a man. Now use the linear model to predict that person's salary if she is a woman.

One last question may be in your mind: why is it called a *regression* line? That word makes no sense! It comes from the fact that when applying this linear model, $\widehat{y}$ tends to be closer to $\overline{y}$ than x is to $\overline{x}$. This is called regression to the mean, hence the term "regression line."

Residuals

Definition 3: The difference between an observed value (y) and the value predicted by the model ($\widehat{y}$) is called the residual. Residuals are denoted by the letter e, so

$$e = y - \widehat{y}$$

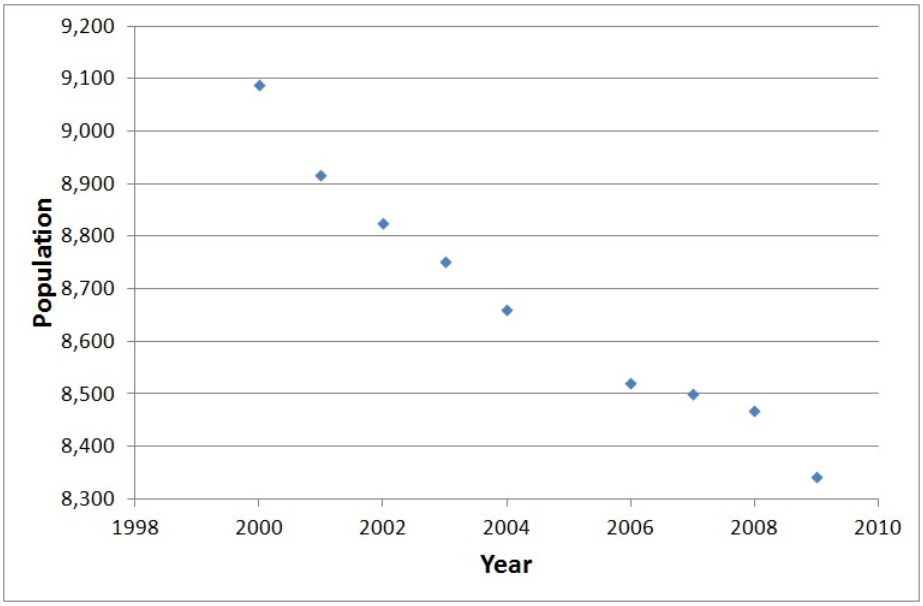
Note that when e is positive, the actual value is larger than the predicted value. When e is negative, the actual value is less than the predicted value.

If the linear model is going to approximate the data well (run as close to the actual data values as possible), then the residuals will need to be as small as possible. As we did when we computed standard deviation, we only want to compare positive values, so we will square the residuals. We want the total of the residuals to be as small as possible, so we add up all the square residuals. The line with the smallest sum of the square residuals is closest to the greatest amount of data points. This line turns out to be the same regression line we calculated earlier, and that's why we also call it the least squares line or the best fit line. It's the line that "best fits" the data, resulting from the "least squares" of the residuals.

Residuals are one way of measuring the error in our model. For statisticians, "error" does not mean we've made a mistake or that we are wrong, it simply means that we are approximating values, not using exact values. Error is the measurement of how close our approximation is.

Example 9.6: Below is the scatterplot for the population of my hometown from 2000-2004 and 2006-2009. Answer the questions below.

$$\overline{x} = 2004.44, \overline{y} = 8674, s_x = 3.206, s_y = 214.192, r = -0.983$$



- Is linear regression appropriate for this data set?

- Find the equation of the regression line.

- The population in 2007 was 8,499. What is the residual for that particular x value?

- What is the predicted population for the year 2005?

A few notes about predicted values and linear regression:

1. You should not extrapolate outside the range of the data set. In the previous example, what is the expected 2010 population? What about the expected 1950 population? The census found the actual 2010 population to be 7,637 and the 1950 population to be 13,293. Is this an acceptable model outside the years 2000-2009?
2. The inverse of the regression line is NOT the regression line when you switch the variables to plot (y, x) instead of (x, y). You will need to completely recompute m and b if you decide you'd like to change the roles of your response and explanatory variables.

Example 9.7: Below is a table of data showing final course grade and grade on Exam 2 for a sample of 26 students in my previous statistics courses. We wish to use the Exam 2 grade to predict the final course grade.

Exam 2 Grade	Final course grade
106	102.42
96	95.29
93	94.72
92	102.02
87.5	87.96
85	87.80
82	94.01
80	57.58
79	84.28
74	79.98
72	72.19
70.5	74.74
70	92.43
68	60.17
66	74.30
61	66.51
56	69.08
54	61.67
52	78.63
52	39.55
46	54.53
44	71.43
36	56.20
30	24.18
0	13.43
0	35.02

1. Create a scatterplot for the data. Does the linear model appear to be appropriate based on the scatterplot? Are there any outliers?

2. What is the correlation between Exam 2 grade and Final Course grade? Does the linear model appear to be appropriate based on the correlation coefficient?

3. Which variable is the explanatory variable for the linear model? response variable?

4. What is the equation of the regression line?

5. If you earned a 50 on Exam 2, what should you expect your final course grade to be? If you earned a 75? If you earned a 95? If you earned a 0?

6. On your own: What does the model predict YOUR final course grade to be, based on your actual grade on Exam 2? (At the end of the semester, you can compute the residual using your actual final grade in the course.)

7. If you found any outliers in part 1, delete them and recompute the correlation and the regression line. Are the results very different?