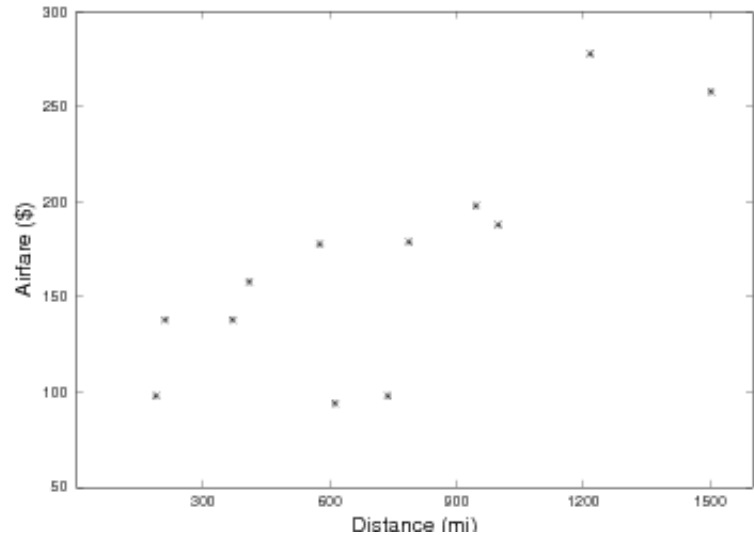


MATH 143: Introduction to Probability and Statistics

Worksheet 4 for Wed., Sept. 23: Least Squares Regression

Consider the data (from 1995) concerning distances and airfares for flights originating in Baltimore, MD given in the table, with the given scatterplot.

Destination	Dist.	Fare
Atlanta	576	178
Boston	370	138
Chicago	612	94
Dallas/Fort Worth	1216	278
Detroit	409	158
Denver	1502	258
Miami	946	198
New Orleans	998	188
New York	189	98
Orlando	787	179
Pittsburgh	210	138
St. Louis	737	98



A natural goal is to try to use the distance of a destination to predict the airfare for flying there, and the simplest model for this prediction is to assume that a straight line summarizes the relationship between distance and airfare.

1. Place a straightedge over the scatterplot above so that the edge forms a line which roughly summarizes the relationship between distance and airfare. Draw this line on the scatterplot.
2. Roughly what airfare does your line predict for a destination which is 300 miles away?
3. Roughly what airfare does your line predict for a destination which is 1500 miles away?

The equation of a line can be represented as $y = a + bx$, where y denotes the variable being predicted (which is plotted on the vertical axis), x denotes the variable being used for the prediction (which is plotted on the horizontal axis), a is the value of the y -intercept of the line, and b is the value of the slope of the line. In this case x represents *distance* and y *airfare*.

4. Use your answers to 2 and 3 above to find the *slope* of your line, remembering that
slope $b = \frac{\text{rise}}{\text{run}} = \frac{\text{change in } y}{\text{change in } x}$.
5. Use your answers to 4 and 2 above to determine the *intercept* of your line. (Note the vertical axis starts at 50.)
6. Put your answers to 5 and 6 together to produce the equation of your line. It is good form to replace the generic x and y symbols in the equation with the actual variable names, in this case *distance* and *airfare* respectively.

Naturally, we would prefer a better way of choosing the line describing the relationship, over simply drawing one that “seems about right”. Since there are infinitely many lines to select from, we need some criterion for choosing the “best” one. The most commonly used criterion is *least squares*, which says to choose the line that minimizes the sum of squared vertical distances from the points to the line, as determined by the sample data. The most convenient expressions for calculating the coefficients a and b of our *least squares line* relate them to the means and standard deviations of the two variables along with the correlation coefficient between them:

$$b = r \frac{s_y}{s_x}, \quad \text{and} \quad a = \bar{y} - b\bar{x}.$$

7. Calculate the means \bar{x} , \bar{y} for the data in the table.
8. The correlation coefficient and standard deviations for this data are $r = 0.7949855$, $s_x = 402.68582$ and $s_y = 59.454273$. Use these statistics to calculate the coefficients a and b of the least squares line.
9. Write the equation of the least squares line.
10. Using this expression for the least squares line, find what airfares are predicted for destinations that are 300 and 1500 miles away, respectively.
11. Go back to the scatterplot and add the two points from 10. Connect them with a line, and compare this new line with the one you drew before.
12. What airfare would the regression line predict for a flight to San Francisco, which is 2842 miles from Baltimore? Would you take this prediction as seriously as one, say, for a destination 900 miles away?
13. Fill in the predicted airfares for destinations 900, 901, 902 and 903 miles from Baltimore in the table below.

Distance	900	901	902	903
Predicted airfare				

What pattern do you notice? By how many dollars is each prediction higher than the preceding one? Give a brief interpretation of the slope coefficient b for our regression line.

14. By how much does the regression line predict airfare to rise for each additional 50 miles that a destination is farther away?

A common theme in statistical modeling is to think of each data point as being comprised of two parts: the part that is explained by the model (called the *fit*) and the “leftover” part (called the *residual* that is either the result of chance variation or of other variables not included in the model. In the context of least squares regression, the *fitted* value for an observation is simply the y -value that the regression line would predict for the x -value of that observation. The *residual* is the difference between the actual y -value and the fitted one (residual = actual - fitted). So the residual measures the vertical distance from the observed y -value to the regression line.

15. If you look back at the original listing of distances and airfares, you find that Atlanta is 576 miles from Baltimore. What airfare would the regression line have predicted for Atlanta (i.e., what is its *fitted value*)?
16. The actual airfare to Atlanta at that time was \$178. Determine the *residual* for Atlanta.

17. Fill in the missing values in the table. Which city has the largest (in absolute value) residual? What were its distance and airfare? By how much did the regression line err in predicting its airfare? Was it an overestimate or underestimate? In general, what can be said about those predicted values which are *overestimated*? How do you identify these when looking at the scatterplot with regression line overlaid?

Destination	Dist.	Fare	Fitted	Residual
Atlanta	576	178		
Boston	370	138	126.70	
Chicago	612	94		-61.10
Dallas/Fort Worth	1216	278	226.00	52.00
Detroit	409	158	131.27	26.73
Denver	1502	258	259.57	-1.56
Miami	946	198	194.30	3.70
New Orleans	998	188	200.41	-12.41
New York	189	98	105.45	-7.45
Orlando	787	179	175.64	3.36
Pittsburgh	210	138	107.92	30.08
St. Louis	737	98	169.77	-71.77

The standard deviation of the residuals (the collection of numbers in the column on the far right above) is 36.067. There is a numerical measure of how much of the variability in the data is explained by the model and how much is “left over”.

18. Find the ratio of this standard deviation of the residuals to the standard deviation of the airfares, and then square this value.
19. Add to the number above the square of the correlation (i.e., r^2). What is the result?
20. What proportion of the variability in airfares is “explained” by the regression line with distance?