Boston University
Department of Mathematics and Statistics

MA 214

Applied Statistics

Lab Session 5: Regression

In this lab we will use the JMP software to explore linear regression and gain practice in fitting a least squares regression line to data and making correct interpretations about the fitted results.

Preparation

We will be using the same data set as the first few weeks of the semester, so if necessary please again download the JMP data set called Televisions.jmp from the Blackboard site under the Course Documents link. Recall that in this file, for each of the forty largest countries in the world (according to 1990 population figures), data are given for the country's life expectancy at birth, number of people per television set, and number of people per physician. SOURCE: *The World Almanac and Book of Facts 1993* (1993), New York: Pharos Books.

Variable Descriptions:

Columns

1	Country
2	Life expectancy
3	People per television
4	People per physician
5	Female life expectancy
6	Male life expectancy

Missing values are denoted with *.

Questions

Previously we have taken a look at the relationship between the life expectancy of a country and its television density; that is, the approximate number of people per television set in a country. Now let us further investigate this relationship using the concept of linear regression.

1. First, open the Televisions.jmp dataset and go to the Analyze -> Fit Y by X command option. Since we will try to predict the life expectancy of a country using the number of people per television as the covariate, place the LifeExp column in the “Y, Response” role, and the “People/Television” column in the “X, Factor” role, and then click the “OK” button. JMP will then show you a scatterplot of the life expectancy values by the people/television values. Based on this scatterplot, do you think there is a linear relationship between the two variables? If so, does this relationship seem positive or negative; strong or weak, etc.?

   Based on the scatterplot, it appears that there is a negative relationship between the two variables. As the number of people per television increases, the life expectancy tends to decrease. However, the relationship does not seem to be very strong, as the points are quite scattered and do not follow a clear linear trend. There are also some countries with missing data for the number of people per television, which are not included in the scatterplot.

2. Now let us fit a regression line to the data. From the red triangle on the new scatterplot that you constructed in part **(a)**, select the “Fit Line” option. JMP will then overlay the fitted line on the scatterplot and provide you with a summary of the fit as well as an Analysis of Variance Table. Use the line and the summaries to answer the following questions.

   i. Write the fitted regression line as it appears in the JMP output. Label β_1 as the slope of the regression line and β_0 as the intercept. Interpret each of these values within the context of the data.

      The fitted regression line is:

      LifeExp = 69.648132 - 0.0362642*People/Television

      This equation represents the best fit line through the data points in the scatterplot, according to the least squares criterion.

      β_0 = 69.648132: This is the intercept of the regression line. It represents the estimated life expectancy when the number of people per television is zero. In other words, it's the estimated life expectancy in a hypothetical situation where every person has a television. However, this interpretation may not be meaningful or realistic in the real-world context, as it's unlikely for every person in a country to have a television.

      β1 = -0.0362642: This is the slope of the regression line. It

represents the estimated change in life expectancy for each additional person per television. Specifically, for each increase of one person per television, the life expectancy decreases by approximately 0.036 years, or about 13.2 days. This suggests that, on average, countries where televisions are more widely available (fewer people per television) tend to have higher life expectancies. However, this is a statistical association and does not necessarily imply a causal relationship. There may be other factors at play that are associated with both television availability and life expectancy.

ii. Scroll down to the "Parameter Estimates" box in the JMP regression output. Here you will find the estimates for the coefficients (the intercept and the slope), their standard errors, and their t test statistics as well as the corresponding p-values for the two-sided tests of hypothesis that each estimated coefficient is equal to zero. Record the p-value for testing whether or not the slope is equal to zero. Based on this p-value, would you conclude that the number of people per television set is a useful predictor in determining the average life expectancy of a country?

H0: $\beta_1 = 0$
H1: $\beta_1 \neq 0$

p-value: less than 0.001

Conclusion: In this case, the p-value is less than 0.001, which is much smaller than 0.05. Therefore, we would reject the null hypothesis and conclude that the number of people per television set is a statistically significant predictor of the average life expectancy of a country.

However, while the number of people per television set is statistically significant, the practical significance or the size of the effect might be small. The slope of -0.036264 suggests that for each increase of one person per television, the life expectancy decreases by approximately 0.036 years, or about 13.2 days. This is a relatively small effect. Furthermore, correlation does not imply causation, and there may be other factors at play that are associated with both television availability and life expectancy.

iii. There are a few hidden menus in JMP that allow you to access additional summaries about your data. We will now use one of

them to construct a 95% confidence interval for each of the estimated parameters in the regression line. Again scroll down to the "Parameter Estimates" box in the JMP regression output and right click somewhere inside the box. From the "Columns" option on the menu that appears, select "Lower 95%" and then repeat this process and select "Upper 95%". You will now see two additional columns of information in the "Parameter Estimates" box. Record the confidence intervals for β_0 and β_1 in the space below.

The 95% confidence intervals for the estimated parameters in the regression line are as follows:

For the intercept (\(β_0\)): The confidence interval is (67.415083, 71.881181). This means that we are 95% confident that the true value of the intercept is between 67.415083 and 71.881181. In the context of the data, this is the estimated range of the average life expectancy when the number of people per television is zero.

For the slope (\(β_1\)): The confidence interval is (-0.052361, -0.020168). This means that we are 95% confident that the true value of the slope is between -0.052361 and -0.020168. In the context of the data, this is the estimated range of the change in life expectancy for each additional person per television.

iv. What is the value of the coefficient of determination? What is your interpretation?

The value of the coefficient of determination is 0.36705, which means that approximately 36.7% of the variation in life expectancy can be explained by the number of people per television. This leaves about 63.3% of the variation in life expectancy that is not explained by this model, which could be due to other factors not included in the model, random variation, or the fact that the relationship is not perfectly linear.

3. It is possible that a linear relationship stronger or weaker than the one you described in parts **(a and b)** between the two variables is being hidden by the presence of “unusual” observations, called outliers, in the data. Which points seem particularly unusual that you may wish to consider removing from the data before continuing with the linear regression techniques?

   I might remove the point, with the data #21, Myanmar (Burma), LifeExp = 54.5, and People/Television = 592, and the data #02, Bangladesh, LifeExp = 53.5, and People/Television = 315, since can be considered as the observations that lie an abnormal distance from other values in a random sample from a population. In a sense, they are points that do not follow the general trend of the rest of the data.

4. Suppose you wanted to exclude the three points that are furthest from the main cluster of the dataset from the study. If your hover your mouse on these points in the scatterplot, the row numbers of the data points will be displayed. Close the scatterplot that you made in part **(a)** and now return to the Televisions data table. Scroll through the table and search for the three entries with the largest numbers of people per television. On each one, right click on the row number and select the “Exclude/Unexclude” option. When an entry is excluded, a small red circle with a line through it will appear next to its number of the left-hand side of the data table. Once you are finished, repeat the steps in part **(a)** to generate a new scatterplot by fitting LifeExp by People/Television in the “Fit Y by X” menu, and comment on the effect of excluding these three points.

   If we exclude these three countries and generate a new scatterplot, we would expect the overall pattern of the data to be clearer, as the remaining points would be more closely clustered together. The relationship between life expectancy and the number of people per television might appear stronger, as the outliers that were potentially distorting the pattern have been removed.

   a. Now we will repeat the steps in part **(b)** using the modified data set.

      i. Write the fitted regression line as it appears in the JMP output. Label β_1 as the slope of the regression line and β_0 as the intercept. Interpret each of these values within the context of the data.

         The fitted regression line is:

         (LifeExp = 71.482553 - 0.1528474*People/Television)

This equation represents the best fit line through the data points in the scatterplot, according to the least squares criterion, after excluding the three countries with the highest number of people per television.

β_0 = 71.482553: This is the intercept of the regression line. It represents the estimated life expectancy when the number of people per television is zero. In other words, it's the estimated life expectancy in a hypothetical situation where every person has a television. However, this interpretation may not be meaningful or realistic in the real world context, as it's unlikely for every person in a country to have a television.

β_1 = -0.1528474: This is the slope of the regression line. It represents the estimated change in life expectancy for each additional person per television. Specifically, for each increase of one person per television, the life expectancy decreases by approximately 0.153 years, or about 56 days. This suggests that, on average, countries where televisions are more widely available (fewer people per television) tend to have higher life expectancies. However, this is a statistical association and does not necessarily imply a causal relationship. There may be other factors at play that are associated with both television availability and life expectancy.

It's worth noting that the slope is more negative in this model compared to the original model that included all countries. This suggests that the relationship between life expectancy and the number of people per television is stronger when the three countries with the highest number of people per television are excluded.

ii. Record the p-value for testing whether or not the slope is equal to zero. Based on this p-value, would you conclude that the number of people per television set is a useful predictor in determining the average life expectancy of a country?

H0: $\beta_1 = 0$
H1: $\beta_1 \neq 0$

p-value: 0.0008

Conclusion: The null hypothesis (H0) is that the slope β_1 is equal

to zero, which would mean that the number of people per television set has no effect on the average life expectancy of a country. The alternative hypothesis (H1) is that the slope β_1 is not equal to zero, which would mean that the number of people per television set does have an effect on the average life expectancy. A p-value less than 0.05 is typically considered statistically significant, meaning that we would reject the null hypothesis in favor of the alternative hypothesis. In this case, the p-value is 0.0008, which is much smaller than 0.05. Therefore, we would reject the null hypothesis and conclude that the number of people per television set is a statistically significant predictor of the average life expectancy of a country.

However, while the number of people per television set is statistically significant, the practical significance or the size of the effect might be small. The slope of -0.152847 suggests that for each increase of one person per television, the life expectancy decreases by approximately 0.153 years, or about 56 days. This is a relatively small effect. Furthermore, correlation does not imply causation, and there may be other factors at play that are associated with both television availability and life expectancy.

iii. Record the confidence intervals for β_0 and β_1 in the space below.

The 95% confidence intervals for the estimated parameters in the regression line are as follows:

For the intercept β_0: The confidence interval is (69.078583, 73.886522). This means that we are 95% confident that the true value of the intercept is between 69.078583 and 73.886522. In the context of the data, this is the estimated range of the average life expectancy when the number of people per television is zero.

For the slope β_1: The confidence interval is (-0.236913, -0.068781). This means that we are 95% confident that the true value of the slope is between -0.236913 and -0.068781. In the context of the data, this is the estimated range of the change in life expectancy for each additional person per television.

iv. Record the new value of the coefficient of determination. Does the new value suggest that the modified model is a better fit to the data? If not, what might be the reason for it?

In this case, an R^2 of 0.293112 means that approximately 29.3% of the variation in life expectancy can be explained by the number of people per television. This is less than the R^2 value of 0.36705 from the original model that included all countries.

This decrease in R^2 suggests that the modified model (after excluding the three countries with the highest number of people per television) is not a better fit to the data than the original model. The lower R^2 value indicates that the modified model explains less of the variation in life expectancy.

The reason for this could be that the three countries that were excluded were actually contributing to the explanation of the variation in life expectancy. Even though these countries were outliers in terms of the number of people per television, they might have been important in terms of their life expectancy values. By excluding these countries, we might have removed some of the information about the relationship between life expectancy and the number of people per television, resulting in a lower R^2 value.

5. Let us now turn our attention to the prediction of average life expectancies for a country with a new number of people per television. For instance, suppose we wanted to determine the average life expectancy for a country with 10 people per television.

   a. To start answering this question, close your regression results window from question **1)**, return to the Televisions data set in JMP (with excluded rows), and scroll down to the very bottom. On the first empty row, double click on the People/Television cell and enter the value of 10. Notice that the rest of the variables are marked as missing values. Construct the scatterplot for the new data table, and fit a regression line by following the steps from **1) (b)**. Confirm that the regression line has not changed with the addition of your new data point. This is because JMP does not take missing data into account when fitting a regression line.

   b. Now we will actually use our new entry in the Televisions data table to arrive at our desired prediction. From the main JMP toolbar, go to the “Analyze

-> Fit Model" command and then place "LifeExp" in the "Y" role and add "People/Television" to the "Construct Model Effects" section. Press the "Run" button to have JMP run the regression as usual but to now display much more information about our assumed linear model. Of particular interest to us now are the predictions for each data point. From the red triangle at the very top left option, "Response LifeExp", go to "Save Columns" and select "Mean Confidence Interval". Repeat this process to also select "Indiv Confidence Interval" from the "Save Columns" option. Now when you look at the Televisions data table you will see four new columns. Two of these (Lower 95% Mean LifeExp, Upper 95% Mean LifeExp) will give the bounds of a 95% confidence interval for the average life expectancy of a country and the other two (Lower 95% Indiv LifeExp, Upper 95% LifeExp) will give the bounds of a 95% prediction interval for the life expectancy of a country.

i. Scroll down to the bottom of the updated data table and record the 95% confidence interval and 95% prediction interval for the new data point with 10 people per television:

   95% Confidence Interval (for the mean life expectancy)

   (67.903882192, 72.004275273)

   95% Prediction Interval (for the life expectancy)

   (58.039045903, 81.869111562)

ii. Which of the intervals above is wider, and why?

   The 95% prediction interval is wider than the 95% confidence interval.

   The reason the prediction interval is wider than the confidence interval is because the prediction interval accounts for more sources of uncertainty.

   The confidence interval estimates the average life expectancy for all countries with 10 people per television. It accounts for the uncertainty in estimating the mean response but assumes that the true line (the relationship between life expectancy and people per

television) is known exactly.

On the other hand, the prediction interval estimates the life expectancy for a single country with 10 people per television. It accounts not only for the uncertainty in estimating the mean response, but also for the additional variability in individual responses around the true line. This additional variability makes the prediction interval wider than the confidence interval.

In other words, it's more uncertain to predict a single observation than to estimate the mean of all observations, hence the prediction interval is wider.

iii. What is the value of X (i.e. People/Television) near which the 95% confidence interval and 95% prediction interval are closest?

Here is a summary of the process we followed to answer the question:

Data Preparation: We started with a dataset that included life expectancy and the number of people per television for various countries. We added a new data point with a "People/Television" value of 10, and ran a regression analysis to predict life expectancy based on the number of people per television.

Calculate Confidence and Prediction Intervals: In the regression output, we saved the 95% confidence intervals and 95% prediction intervals for each data point to new columns in the data table. These intervals represent the range of values within which we can be 95% confident that the true mean life expectancy (for the confidence interval) or a single life expectancy value (for the prediction interval) will fall.

Calculate Interval Widths: We created two new columns in the data table to calculate the width of each interval (upper limit - lower limit). This gave us a measure of the uncertainty associated with each prediction.

Calculate Interval Width Difference: We created another new column to calculate the absolute difference between the widths of the confidence interval and prediction interval for each data point. This gave us a measure of how close the two intervals were for

each value of "People/Television".

Find Minimum Interval Width Difference: We sorted the data table by the "Interval Width Difference" column in ascending order. The first row in the sorted table represented the value of "People/Television" where the difference between the widths of the confidence interval and prediction interval was smallest.

Based on this analysis, we found that the value of "People/Television" near which the 95% confidence interval and 95% prediction interval are closest is approximately 13.33, which corresponds to the country Kenya in our dataset. This suggests that for countries with around 13 people per television, the uncertainty in predicting the mean life expectancy and a single life expectancy value is relatively small.