Boston University
Department of Mathematics and Statistics

MA 214

Applied Statistics

Lab Session 6: Regression II

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

Preparation

Once again we will be using the JMP data set called Televisions.jmp which is available on 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

1) Let us continue exploring the relationship between the life expectancy of a country and the number of people per television set by fitting a number of models to these two variables.

(a) First construct a scatterplot for the life expectancy of a country versus its number of people per television set in the usual way, by going to the “Analyze” -> “Fit Y by X” command, placing “LifeExp” in the “Y” role, adding “People/Television” to the “X, Factor” role, and finally pressing the “OK” button. Based on this scatterplot and the criteria discussed in class for determining the presence of extreme covariate values (in our case, People/Television), would you remove any data points before continuing with the regression analysis? If so, then go to the JMP data table and exclude those points, and then record those points below with an explanation for why you chose to remove them.

From the scatterplot, we can observe a few points that could be considered as outliers or extreme covariate values. These are the points with a high number of people per television set (greater than 300). These points are significantly different from the rest of the data and could potentially skew the regression analysis.

The points that could be considered for removal are:

Bangladesh: LifeExp = 53.5, People/Television = 315
Ethiopia: LifeExp = 51.5, People/Television = 503
Myanmar (Burma): LifeExp = 54.5, People/Television = 592

These points are chosen for removal because they have a significantly higher number of people per television set compared to other countries, which makes them extreme values in the context of this dataset. Removing these points could help in obtaining a more accurate regression model that better represents the general trend in the data.

(b) Next, using your potentially revised data set, fit a linear regression model to predict the life expectancy of a country using its number of people per television set in a slightly different way than usual, by going to the “Analyze” -> “Fit Model” command, placing “LifeExp” in the “Y” role, adding “People/Television” to the “Construct Model Effects” box, changing the “Emphasis” to “Minimal Report”, and finally pressing the “Run” button.

(i) From the red triangle at the top of the JMP output for the “Fit Model” procedure, go to “Row Diagnostics” twice in order to select both “Plot Residuals by Predicted” and “Plot Residuals by Row”.

(ii) Based on the plot of the residuals by the predicted values and the criteria discussed in class, do you think that there is evidence to suggest a nonlinear relationship between life expectancy and people per television? Please explain your answer.

So we have two plots, in which the first plot shows residuals by predicted values, and the second plot shows residuals by row.

From the plot of residuals by predicted values, we can see that the residuals are not randomly scattered around the horizontal axis (residual = 0), and there seems to be a pattern or trend in the residuals. This suggests that the relationship between life expectancy and people per television might not be linear.

In a good linear regression model, we would expect the residuals to be randomly scattered around the horizontal axis with no obvious pattern or trend. This is because the residuals represent the error between the observed and predicted values, and in a good model, this error should be random and not systematic.

Therefore, based on the plot of residuals by predicted values, there is evidence to suggest a nonlinear relationship between life expectancy and people per television.

(iii) Based on the plot of the residuals by row (order of data collection) and the criteria discussed in class, do you think that there is evidence to suggest that the data points are not independent of one another? Please explain your answer.

Essentially, in a good regression model, we would expect the residuals to be randomly scattered around the horizontal axis (residual = 0) with no obvious pattern or trend. This is because the residuals represent the error between the observed and predicted values, and in a good model, this error should be random and not systematic.

In the plot of residuals by row for this data, we do not see any clear patterns or trends. The residuals seem to be randomly scattered around the horizontal axis. This suggests that the data points are likely independent of one another.

(iv) To verify the normality assumption, from the red triangle next to the response, choose save column > residual. The residuals will be saved in a new column of the data frame. From the Analyze menu, choose distribution platform, and select residual column as the Y variable. To get the normal quantile plot, use the red triangle next to the Residual and choose Normal Quantile Plot. Does the plot suggest that the normality assumption is satisfied? Please explain your answer.

In this case, the points in the Q-Q plot do not perfectly follow the straight line, especially at the tails. However, they are not too far off, and the deviations could be due to random chance. Therefore, based on this plot, we could say that the normality assumption is approximately satisfied.

2) Now let us explore the effect of transforming the life expectancy data, the people per television set data, or both, on the fit of a linear model.

(a) Based on the guidelines we've discussed in class about transforming data, let's use the "Fit Model" command to try a variety of different transformations and look for the best one. As before, go to "Analyze" -> "Fit Model", place "LifeExp" in the "Y" Role, add "People/Television" to the "Construct Model Effects" box, and change the "Emphasis" to "Minimal Report". Now, after highlighting "People/Television" inside the "Construct Model Effects" box, or "LifeExp" in the "Y" Role, or both, go to the red triangle near the bottom of the "Fit Model" window next to the word "Transform" and apply the appropriate transformation to the data in order to fill in the table below. Remember to further use the "Plot Residual by Predicted" and "Plot Residual by Row" commands to check the adequacy of each model. You may enter your own ideas for transformations in the empty rows at the end of the table.

Transformation on X	Transformation on Y	R^2	Adequate?	Reason(s)
None	**None**	**0.293**	**Yes**	The residuals are randomly scattered around the zero line, suggesting that the assumption of homoscedasticity is met.

$\sqrt{X}$	**None**	0.405	Yes	The residuals are randomly scattered around the zero line, suggesting that the assumption of homoscedasticity is met. The residuals by row show no clear pattern, suggesting that the data points are independent.
None	$\sqrt{Y}$	0.223	Yes	The residuals are randomly scattered around the zero line, suggesting that the assumption of homoscedasticity is met. The residuals by row show no clear pattern, suggesting that the data points are independent.
$\log X$	$\log Y$	0.573	Yes	The residuals are randomly scattered around the zero line, suggesting that the assumption of homoscedasticity is met. The residuals by row show no clear pattern, suggesting that the data points are independent.

$\log X$	**None**	0.585	Yes	The residuals are randomly scattered around the zero line, suggesting that the assumption of homoscedasticity is met. The residuals by row show no clear pattern, suggesting that the data points are independent.
Any other transformations you wish to try? List the transformation(s) and your results here. X^2	None	0.074	No	The residuals are not randomly scattered around the zero line, suggesting that the assumption of homoscedasticity may not be met. There is a clear pattern such as a funnel shape, which would indicate heteroscedasticity.
Any other transformation? 1/X	None	0.548	No	The residuals are not randomly scattered around the zero line, suggesting that the assumption of homoscedasticity may not be met. There is a clear pattern such as a funnel shape, which would indicate heteroscedasticity.

(b) After exploring all of the transformations, decide on one model to use in the next part of the analysis. Indicate which model you have selected in the space below, and also express the model in the original unit of the response variable Y.

Based on the R-squared values and the adequacy of the models, the transformation with the highest R-squared value and that meets the assumptions of homoscedasticity and independence of residuals is the model with a log transformation on X (People/Television) and no transformation on Y (LifeExp). This model has an R-squared value of 0.585, which is the highest among all the models that meet the assumptions.

So, the selected model is:

Log(People/Television) -> LifeExp

Expressing the model in the original unit of the response variable Y, we get:

LifeExp = $\beta 0 + \beta 1$ * log(People/Television)

where $\beta 0$ is the intercept and $\beta 1$ is the coefficient for log(People/Television). This model suggests that the life expectancy of a country is linearly related to the logarithm of the number of people per television set in that country.

3) Now that you have settled on a model to use in the regression analysis, use the JMP software to build a 95% prediction interval for the life expectancy of a country that has 6 people per television. Recall that to do this you must first add a new row to the Televisions data set and assign a value of 6 to the “People/Television” column. Then, after using the “Analyze” -> “Fit Model” command to fit your chosen model, you must go to the red triangle at the top left of the output window and select “Indiv Confidence Interval” from the “Save Columns” option under the red triangle. This will place the 95% prediction interval lower bound and upper bound for each data point as values of two new columns in the Televisions data table.

Record the 95% prediction interval derived from your model for the life expectancy of a country that has 6 people per television in the space below:

95% prediction interval: (61.62, 78.94). This means that we can be 95% confident that the true life expectancy for a country with 6 people per television will fall within this interval.

4) (Optional) What kind of fits do you think you would have seen for the transformed data had you not excluded any data points? If you have time, go back and explore the adequacy of each model for the different combinations of transformed data where no data points are excluded. Can you find a model of transformed data that actually performs better when **no** data points are excluded? What does this tell you about removing extreme observations from the data set?

Removing extreme observations from a dataset can sometimes improve the fit of a model, but it's not always the case. Extreme observations, or outliers, can have a significant impact on the fit of a model, especially if the model is sensitive to these values, like linear regression models. However, these extreme observations can sometimes provide valuable information about the variability in the data, and removing them might lead to an oversimplified model that doesn't capture the full complexity of the data.

If we were to include all data points, including the extreme ones, in the transformations and model fitting, we might see different results. The fit of the models might be worse due to the influence of the extreme values, or it might be better if the extreme values are actually representative of the underlying relationship in the data.

For example, if the extreme values are not just random outliers but are indicative of a nonlinear relationship or a relationship with higher variability at certain levels of the predictor variable, then a model that includes these points and accounts for this complexity might actually have a better fit.

It’s also important to consider the context and the implications of removing or including extreme values. If the extreme values are due to measurement errors or other factors that are not representative of the population we're interested in, then it might be justified to remove them. On the other hand, if the extreme values are true observations from the population, then they should be included in the analysis.