

**BA240 Team Project**  
**Leslie Lum**  
**Fall 2009**

**Phase 1 (10 points) – Execute a team contract, identify a project and create a project plan. Due Oct 6.**

Grading Rubric (2 points each)

	2 Complete Missing	1 Partial	0
Method of communication defined			
Team roles and tasks defined on sample project plan. You may revise tasks as needed but are required to track time on task.			
Decision making and conflict resolution defined.			
Hypothesis and structure of analysis defined.			
Data set selected and qualifies. There will be one y variable and several x variables. There must be at least 50 data.			

**Number all your responses in your submission.**

**Phase 2 (30 points)**

**Data Selection and Analysis INDIVIDUAL ASSIGNMENT– Due Oct 29.**

Choose one independent variables for each team member to regress against the dependent variable. Save this analysis for your final paper. If your data is extremely highly correlated ( $r > .9$ ), it is probably measuring the same thing so look for data that is less correlated.

4 points each except outliers and central tendency	Complete	Partial	Missing
Plot histograms of your x and y data -Reasonable intervals -Labels correct -No gaps			
Show the five number summary. Assess normality using empirical rule and box plot.			

Assess normality using normality plot			
Outliers identified on graph			
Mean, median, mode identified on graph			
Scatterplot -Labels correct -x and y correct			
Correlation and interpretation			
Least squares line and analysis			

Each team member is to choose one of the independent data sets to analyze along with the dependent data set. Review the Excel video on histograms and linear regression before you do your own. It will take three or four viewings of the videos to really learn how to do it. Each variable should contain more than 50 data. Number all your answers in your submission. Assist your team members in making sure that their submission is correct.

1. Explain how the data was collected. Cite websites. List any limitations of the data.
2. Calculate the correlation and draw conclusions about the relationship between the independent and dependent variables. If correlations are extremely high (over 90%) look for other data. High correlations signify that the data might measure the same thing.
3. Plot histograms using reasonable intervals for each set.
4. Analyze whether the distributions are normal by using the empirical rule comparing plus and minus one, two and three standard deviations. Show this on the graph.
5. Complete a normality plot using Excel/Data Analysis/Regression.
6. Identify any outliers in each distribution.
7. Calculate the mean, median, and mode. Explain what these say in terms of the distribution.
8. Create a scatter plot of your independent variable against the dependent variable.
9. Draw the least squares line and discuss what the linear model says about the relationship between the dependent and independent variables.
10. Compare all team member results and select the data you want to include in the final analysis. You need a minimum of three independent variables. There is no maximum.

### Phase 3 (45 points)

Modeling – Draft Due **Dec 3**. Final Due **Dec. 7**

	Complete	Partial	Missing
1. Multiple Regression Modeling using Excel -used all possible Combinations (6)			
2. Outlier removal and model Using Excel show how you selected your outliers -First outlier removal -Second outlier removal (6)			
3. Final Modeling Express the model in mathematical and verbal terms. Table with selected output <ul style="list-style-type: none"> <li>• ANOVA table (F test)</li> <li>• Intercept significance</li> <li>• Slope significance</li> <li>• Adjusted R square (6)</li> </ul>			
4. Prediction -Prediction on 3 x variables with analysis against actual -Prediction intervals -Confidence intervals -Interpretation (6)			
5. Assumption Check -Residual analysis for mean of 0. -Variance constancy -Variance independence -Normality plot against straight line (6)			
6. Discussion of Final Model <ul style="list-style-type: none"> <li>• Did this model align with your original hypothesis of the relationships?</li> <li>• How did outlier removal affect your model?</li> <li>• What is good about the model? Did it advance your knowledge of the relationships?</li> <li>• What are the drawbacks of the model? Does it have good predictive ability?</li> </ul> (5)			
7. Conclusion Summary of findings. What other studies should be done to advance knowledge? What improvements			

should be made for future study? (5)			
8. Appendix All raw output Original data (or first page and last page) (5)			

## Team Project Checklist

- ✓ **Title Page** (Free Design, however, **MUST** includes the followings)
  - Title
  - Date
  - Your Name
  - Course Name
  - Instructor's Name
  
- ✓ **Table of Contents**
  
- ✓ **Report (5 – 10 pages)**
  - I. **Introduction**
    - Definition of the area of study
    - Identify the problem and purpose of study
  
  - II. **Data**
    - How data was collected, write down the source if possible.
    - Data Introduction (Identify Variables, i.e. independent variables, dependent variables)
    - Adequacy of the data (Reliability of the source and sufficiency of the data)
    - Summary Statistics (including central tendency, variability, five-number summary, 95% confidence intervals) with complete interpretation for each variable.
  
  - III. **Modeling**
    - Techniques used in analysis (i.e. Regression Analysis / Least Square Line)
    - Multiple Regression Modeling:
      - i. Run all possible combination of models: for each model, specify both dependent and independent variables.
      - ii. Comparison Table for model significance
      - iii. Comparison Table for parameter estimates and parameter significance.
      - iv. Comparison Table for model goodness of fit. (this is essentially the adjusted r-square)
      - v. Select your Best Model, and state why you believe this is the best model. Do steps vi to ix to the best model.
      - vi. Excel outputs and interpretations for model significance
      - vii. Excel outputs and interpretations for parameter estimates and parameter significance.
      - viii. Excel output and interpretation for model goodness of fit.
      - ix. Identify normal and extreme outliers, show Excel output.

- Best Model Improvement:
  - i. Remove the outliers, and re-run your best model.
  - ii. Excel outputs and interpretations for model significance
  - iii. Excel outputs and interpretations for parameter estimates and parameter significance.
  - iv. Excel output and interpretation for model goodness of fit, Compare it with the Original Model, and comment on the difference.
  - v. Identify normal and extreme outliers, show Excel output.
- Final modeling:
  - i. Remove the outliers, and re-run the model one more time.
  - ii. Write down the model, specify both dependent and independent variables.
  - iii. Write down the regression equation and interpret it.
  - iv. Excel outputs and interpretations for model significance
  - v. Excel outputs and interpretations for parameter estimates and parameter significance
  - vi. Excel output and interpretation for model goodness of fit. Compare it with the Original Model, and Model after 1<sup>st</sup> outlier removal, and comment on the difference.
  - vii. Randomly assign some values (at least 3 sets) for independent variables, using the regression equation to predict the dependent variable.

#### **IV. Discussion for Final Model**

- Identify if there is any normal and/or extreme outliers from the final model, show Excel output, and comment/discuss it.
- Assumption check for final model overall quality (Residual Analysis), show Excel outputs.
- Does your model appear to be a VALID one? Why or why not?
- Recommendation of the model (good things about the model)
- Drawbacks of the model (bad things about the model)

#### **V. Conclusion**

- Future Study (any possible improvement)
- Final paragraph (summary)

#### ✓ **Appendix (No Limit)**

- Include ALL raw data and/or transformed data
- ALL Raw outputs

## Tips and Notes

**1)** When you are doing INDIVIDUAL variable summary statistics (in Part II Data), **DO NOT** delete "outliers", you need to keep every single value (unless it's missing). For Scatter Plot, you also need to use the original data.

**2)** Scatter Plot is between **TWO** variables ONLY (Specifically, Y Vs. X). So, you need to use your dependent variable Y against **EACH** independent variable X. That is, how many independent variables do you have, how many scatter plots you need to produce. Remember to add trend line and show the equation and R-square on the scatter plot. You can simply calculate the correlation using R-square, remember to check the SIGN.

**3)** Parameters are the y-intercept and slopes, so you need to find out these estimates ( $b_0$ ,  $b_1$ ,  $b_2$  ...) from the output, and then form your regression equation (the least square line). When you have multiple independent variables, you need to be careful on the interpretations: that is, predict the slopes for independent variables INDIVIDUALLY, which means ONE at a time. When you predict one of the slopes, remember to FIX the others.

**4)** Compare your models (original model, model after first outlier removal, and model after second outlier removal) in Discussion (Part IV). Which model is better? Keep in mind: Adjusted R-square is NOT the ONLY factor to make decision. You also need to check whether or not the model and/or parameters are significant (look at the p-values).