Homework 3
STAT 334, Prof. Walker, Spring 2019
Due at the beginning of class on Thursday, April 25

1. Real Estate (Textbook Appendix C.7 p. 682) File: **Real Estate2.jmp**

Price Home price (in thousands of dollars)
Sqft Finished area of home (in square feet)

- Make a scatterplot and regression table of **Price** vs. **Sqft**. As part of your regression output, make the necessary residual plots.

(a) List any regression assumptions that are violated. (You can use your answers from HW2 here.)
(b) Based on your answer in part (a), what variable should you transform first?
(c) Find transformations of Y and/or X that fix the problems you mentioned in parts (a)-(c). (Use only common power transformations in which the power is evenly divisible by 0.5.) Write down the formulas for the transformations that you used for the variable(s). (For example, write $\sqrt{Y}$ and not $p = 0.5$.) Use the residual plots and tests to make your decision about which transformation to use. If you cannot fix violations of all the assumptions, choose transformations that fix as many as you can.

- Print the regression table and residual plots for your final transformed model and turn them in with your answers.

(d) Write down the estimated regression function (after the transformations).
(e) Is there a linear association between the transformed y and x using $\alpha = .05$. Do the test and state your conclusion in the context of the data. If so, interpret the meaning of the slope of the transformed regression with 95% confidence. (If your transformation includes a log, write your interpretation in terms of the original, untransformed scales of the variables.)
(f) With 95% confidence, estimate the average value of all homes with 2000 square feet of finished space using your transformed regression. Write the limits of the interval in the original units (dollars). Write a sentence interpreting the results of the interval.
(g) With 95% confidence, estimate the value of one home with 2000 square feet of finished space using your transformed regression. Write the limits of the interval in the original units (dollars). Write a sentence interpreting the results of the interval.

2. Solution Concentration (See Problem 3.15, p. 150) File: **Solution.jmp**

Concentration Concentration of the solution (molar or "moles per liter")
Hours Hours that have passed since the solution was prepared

(a) List any regression assumptions that are violated. (You can use your answer from HW2 here.)
(b) Based on your answer in part (a), what variable should you transform first?
(c) Find transformations of Y and/or X that fix the problems you mentioned in parts (a)-(d). (Use only common power transformations in which the power is evenly divisible by 0.5.) Write down the formulas for the transformations that you used for the variable(s). (For example, write $\sqrt{Y}$ and not $p = 0.5$.) Use the residual plots and tests to make your decision about which transformation to use. If you cannot fix violations of all the assumptions, choose transformations that fix as many as you can.

- Print the regression table and residual plots for your final transformed model and turn them in with your answers.

(d) Write down the estimated regression function (after the transformations).
(e) Is there a linear association between the transformed y and x using $\alpha = 0.05$. Do the test and state your conclusion in the context of the data. If so, interpret the meaning of the slope of the transformed regression with 95% confidence. (If your transformation includes a log, write your interpretation in terms of the original, untransformed scales of the variables.)
(f) With 95% confidence, estimate the average concentration of all samples after 8 hours using your transformed regression. Write the limits of the interval in the original units. Write a sentence interpreting the results of the interval.
(g) With 95% confidence, estimate the concentration of one sample after 8 hours using your transformed regression. Write the limits of the interval in the original units. Write a sentence interpreting the results of the interval.

Homework 3
STAT 334, Prof. Walker, Spring 2019
Due at the beginning of class on Thursday, April 25

3. Sales Growth (See Problem 3.17, p. 150-151) File: **Sales Growth.jmp**

Sales Yearly sales (in thousands of units)
Year Coded year (0 = 10 years ago, …, 9 = 1 year ago)

(a) Using residuals plots and/or tests, check each regression assumption. List each assumption, then tell the plot and/or test you used, whether the assumption is OK or violated, and describe why you made that decision. Do any tests at $\alpha = 0.1$ for more power.

(b) Based on your answer in part (a), what variable should you transform first?

(c) Find transformations of Y and/or X that fix the problems you mentioned in parts (a)-(d). (Use only common power transformations in which the power is evenly divisible by 0.5.) Write down the formulas for the transformations that you used for the variable(s). (For example, write $\sqrt{Y}$ and not $p = 0.5$.) Use the residual plots and tests to make your decision about which transformation to use. If you cannot fix violations of all the assumptions, choose transformations that fix as many as you can.

- Print the regression table and residual plots for your final transformed model and turn them in with your answers.

(d) Write down the estimated regression function (after the transformations).

(e) Is there a linear association between the transformed y and x using $\alpha = 0.05$. Do the test and state your conclusion in the context of the data. If so, interpret the meaning of the slope of the transformed regression with 95% confidence. (If your transformation includes a log, write your interpretation in terms of the original, untransformed scales of the variables.)

(f) With 95% confidence, estimate the sales next year using your transformed regression. Write the limits of the interval in the original units. Write a sentence interpreting the results of the interval.

(g) Explain what problem there might be in using the interval in part (f)?

Doing regressions in JMP

- If you need to make confidence or prediction intervals for x_v, enter the x_v value in the last row of the predictor variable column before running the regression.
- Optional: If you want to do tests or confidence intervals with an α other than 0.05 (95% confidence), click the red triangle next to "Model Specification", select **Set Alpha Level**, and enter the new α level in the blank.

1. Run the **Analyze>Fit Model** command.
2. Select the response variable from the variable list at the left and click the **Y** button.
3. Select the explanatory variable(s) from the variable list at the left and click the **Add** button. The explanatory variable name(s) should appear in the blank under **Construct Model Effects**.

After this, the Model Dialog will show "Personality: Standard Least Squares" and "Emphasis: Effect Leverage".

4. Click the **Run** button.

A JMP report window for Fit Model will appear. The different parts of this window are described in more detail on p. 47 (Multiple Linear Regression) of the file *JMP Documentation.pdf* in the "Software Help" section of our PolyLearn course. I will describe the highlights below.

The **Parameter Estimates** section of the report window contains the estimated coefficients (sample intercept and sample slope), their standard errors, test statistics, and p-values. The asterisk (*) beside a p-value just means that it is less than α, which is 0.05 unless you changed it in the Fit Model dialog above.

To add confidence intervals for the slope and intercept to the parameter estimates table, click on the red triangle at the top of the window (beside the word "Response") and choose **Regression Reports>Show All Confidence Intervals**. This shows only the confidence intervals for the intercept and slope. It does not show confidence intervals for the mean response or prediction intervals. For those, see below.

To compute confidence and prediction intervals for x_v, click the red triangle and select **Save Columns>Mean Confidence Interval Formula** or **Save Columns>Indiv Confidence Interval Formula**. The intervals will appear in the spreadsheet. Only use the intervals in the last row of the spreadsheet next to the x_v value that you entered. (See the first bullet point above.)

Plot of Residuals vs. Predicted

A plot of residuals vs. predicted values appears in the report window under the heading **Residual by Predicted Plot**. While this is not quite as good as plotting studentized residuals vs. predicted values, this plot will still do the job—although sometimes it is poorly scaled making it hard to read.

If you want a plot of studentized residuals vs. predicted values, you first need to store the studentized residuals and predicted values in the data table. Click the red triangle at the top of the regression report window and select **Save Columns>Studentized Residuals** and **Save Columns>Predicted Values**. Select the **Analyze Fit Y-by-X** command, choose the studentized residuals as **Y** and the predicted values as **X**. Click OK. If you want to add a horizontal line at residual = 0 on the plot, click the red triangle above the graph, and choose **Fit Mean**.

Index Plot of Residuals

To check the independence assumption for time series data, you will need an **index plot of the residuals**. To plot the ordinary residuals vs. row number, click the red triangle and choose **Row Diagnostics>Plot Residuals by Row**. To make an index plot of the studentized residuals, click the red triangle at the top of the regression report window and select **Save Columns>Studentized Residuals**. Then, choose the **Analyze>Modeling>Time Series** command and enter the residuals in the box labeled **Y, Time Series**. Click **OK**. The plot at the top of the output is the index plot of residuals.

Normal Probability Plot of Residuals

Click the red triangle at the top of the regression report window and select **Save Columns>Studentized Residuals**. Choose the **Analyze>Distribution** command, select the residuals that you just saved as the Y column, and click OK. When the report window appears, click on the lower of the two red triangles at the top of the window (the one next to the "Residual" variable name. Select **Normal Quantile Plot** from the list of options. A normal probability plot of residuals should appear to the right of the histogram and boxplot in the report window.

Testing Non-normality
First, save the residuals into a column, then use the **Analyze>Distribution** command. Select the residuals column as **Y**, then click **OK**. From the red triangle menu next to the variable name in the output, select **Continuous Fit>Normal**. Then, from the red triangle menu next to the Fitted Nornal output, select **Goodness of Fit**. The Shapiro-Wilk test statistic and p-value will be displayed. (JMP does not do the Anderson-Darling Normalilty test.)

Testing Unequal Variance
If the *x*-values in the data set are not replicated, do this step first.
Go to the JMP spreadsheet, select the column containing the predictor variable, and choose the **Cols>Utilities>Make Binning Formula** command. Under **Bin Shape**, type in a number for **width** that gives you the number of bins that you want with a reasonable sample size in each. (You can look at the display of alternating bands to see how many observations will be in each bin.) The **offset** value sets where the bins start, but you usually won't need to adjust the offset. When you are done, click **Make Formula Columns** and a new categorical variable containing the bin definitions will appear in the spreadsheet.

Save the residuals into a column, then choose the **Analyze>Fit Y-by-X** command. Enter the residuals as **Y, Response** and enter the categorical predictor variable (or categorical bin variable, see above) as **X, Factor**. Click **OK**. From the red triangle menu, select **Unequal Variances**. Use the Brown-Forsythe test rather than Levene's test to avoid any problems with outliers.

Testing for Autocorrelation
In the main regression output window, click the red triangle menu and select **Row Diagnostics>Durbin-Watson Test**. From the red triangle next to the Durbin-Watson output, select **Significance P-Value** to add a p-value for the Durbin-Watson test. **Warning:** The p-value only works for testing for positive autocorrelation. If you are testing for any autocorrelation or only negative autocorrelation, use the Durbin-Watson table to determine the test result.

Transforming Variables
There are two ways to transform variables in JMP.
1. Have JMP transform the data for you by selecting transformation in the Fit Model window.
2. Create transformed variables in the spreadsheet, then use the Fit Model command normally.

Usually, you will want to have JMP do the transformations for you.

Letting JMP Transform the Data
1. Enter the variables as usual in the **Y** or **Construct Model Effects** blanks of the Fit Model window.
2. Click the name of the variable to transform in the **Y** or **Construct Model Effects** list.
3. Click the red triangle next to **Transform** at the bottom of the Fit Model window and select the transformation that you want. (Note: "Log" in JMP is "ln", the natural log. If you want the base-10 log, you'll need to create the column manually.)

- You can transform both *y* and *x* using this method.
- JMP will reverse transform all predictions and prediction intervals back onto the original scale. It will not reverse transform residuals, slope, or intercept (or their confidence intervals).

Creating the Transformed Variables
1. Choose the **Cols>New Column** command (or click into an empty column of the spreadsheet).
2. Type a name for the new column. (Make sure that Data Type is Numeric and Modeling Type is Continuous.)
3. Click on the Column Properties button and choose Formula.
4. A formula window will appear. Choose variables to transform from the list on the left and select transforms from the list on the right. (Log and square root appear under the Transcendental heading.) Create the formula in the window, then click OK.

Choosing a Good Transformation
A good transformation fixes the violations of the regression assumptions. You must verify that the transformation fixes the violations by rechecking the assumptions for the transformed regression. JMP can suggest the first transformation to try by using the **Factor Profiling>Box Cox Y Transformation** from red triangle menu on the output. This shows the power of the Y transformation that minimizes the regression SSE. This doesn't guarantee that it will fix the violations, but it is a good transformation to start with. JMP does not suggest X transformations.