Homework 2
STAT 334, Spring 2019
Due at the beginning of class on Thursday, April 18

1. Problem 2.10, p. 91

2. **GPA** (See Problem 1.19, p. 35 for a description of the data.) File: **GPA.jmp**
- Make a scatterplot and regression table of **GPA** vs. **ACT**. As part of your regression output, make the necessary residual plots.

(a) In the context of the data, write a sentence interpreting the results of the 95% confidence interval for the mean response when $x_v = 20$.
(b) In the context of the data, write a sentence interpreting the results of the 95% prediction interval a new observation when $x_v = 20$.
(c) Use residual plots to check the regression assumptions. For each assumption, tell what you looked at to judge the assumption, whether you think the assumption is violated or not, and explain how you reached your decision.
(d) Test normality at $\alpha = .05$. What is your conclusion?
(e) Test at $\alpha = .05$ to check for non-constant variance. What is your conclusion?
(f) Perform an F-test for lack of fit with $\alpha = .05$. Write out the hypotheses of the test, state the p-value, and write a sentence describing the conclusion of the test.

3. **Plastic** (See Problem 1.22, p. 36 for a description of the data.) File: **Plastic.jmp**
- Make a scatterplot and regression table of **Hardness** vs. **Time**. As part of your regression output, make the necessary residual plots.

(a) In the context of the data, write a sentence interpreting the results of the 95% confidence interval for the mean response when $x_v = 36$.
(b) In the context of the data, write a sentence interpreting the results of the 95% prediction interval a new observation when $x_v = 36$.
(c) Use residual plots to check the regression assumptions. For each assumption, tell what you looked at to judge the assumption, whether you think the assumption is violated or not, and explain how you reached your decision.
(d) Test normality at $\alpha = .05$. What is your conclusion?
(e) Test at $\alpha = .05$ to check for non-constant variance. What is your conclusion?
(f) Perform an F-test for lack of fit with $\alpha = .05$. Write out the hypotheses of the test, state the p-value, and write a sentence describing the conclusion of the test.

4. **Sales** (See Problem 12.17, p. 505 for a description of the data.) File: **Sales.jmp**
- Make a scatterplot and regression table of **Company** sales vs. **Industry** sales. As part of your regression output, make the necessary residual plots.

(a) Use residual plots to check the regression assumptions. For each assumption, tell what you looked at to judge the assumption, whether you think the assumption is violated or not, and explain how you reached your decision.
(b) Test normality at $\alpha = .05$. What is your conclusion?
(c) Split the data into bins and test at $\alpha = .05$ to check for non-constant variance. What is your conclusion?
(d) Perform the Durbin-Watson test for positive autocorrelation at $\alpha = .05$. Write out the hypotheses of the test, state the critical values or p-value, and write a sentence describing the conclusion of the test.

(continued on the next page)

Homework 2
STAT 334, Spring 2019
Due at the beginning of class on Thursday, April 18

5. Real Estate (See Appendix C.7 p. 682 for a description of the data.) 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) Use residual plots to check the regression assumptions. For each assumption, tell what you looked at to judge the assumption, whether you think the assumption is violated or not, and explain how you reached your decision.

(b) Test normality at $\alpha = .05$. What is your conclusion?

(c) Split the data into bins and test at $\alpha = .05$ to check for non-constant variance. What is your conclusion?

(d) Explain why a lack of fit test on these data might not be reliable.

6. 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

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

(a) Use residual plots to check the regression assumptions. For each assumption, tell what you looked at to judge the assumption, whether you think the assumption is violated or not, and explain how you reached your decision.

(b) Test normality at $\alpha = .05$. What is your conclusion?

(c) Test at $\alpha = .05$ to check for non-constant variance. What is your conclusion?

(d) Perform an F-test for lack of fit with $\alpha = .05$. Write out the hypotheses of the test, state the p-value, and write a sentence describing the conclusion of the test.

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 (or an interval with other than 95% confidence), click the red triangle next to "Model Specification" and select **Set Alpha Level**. Enter the new α level in the blank, then click **OK**.

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 should appear in the blank under **Construct Model Effects**. (For multiple regression, just continue to select more variables and click **Add** until all of the explanatory variables appear in the blank under **Construct Model Effects**.)

After the variables are selected, the Personality blank in the window will say "Standard Least Squares" and the Emphasis blank will say "Effect Leverage". Do not worry about these. Do not change them.

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 in the file *JMP Multiple Regression.pdf* in the "JMP Help" section of our Blackboard 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 Limit Formula** or **Save Columns>Indiv Confidence Limit Formula**. (You can hold down the **Shift** key while issuing these commands to choose a different α level for the interval.) 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.) If you change the x_v value the formula columns containing the interval limits will update.

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 (deleted) 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 (deleted) 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>Specialized 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 Correlated Errors (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 want to two-sided test for autocorrelation or negative autocorrelation only, use *R* or adjust the *JMP* p-value as described in the class notes.

Testing for Lack of Fit
The Lack of Fit Test appears in JMP's default regression output. If it does not appear, click the red triangle in the regression output window and choose **Regression Reports>Lack of Fit**. If the Lack of Fit command is grayed out, it means that there are not enough unique values in the explanatory variable to conduct a lack of fit test.