**Stat 401 A – Lab 9**

**Goals:** In this lab, we will learn how to exclude observations from analyses or graphs and do more with regression. Specifically, we will see how to:

 test hypotheses or construct confidence intervals for regression parameters,

 construct a confidence interval for the line (mean Y at given X)

 construct a prediction interval for observations (single Y at given X)

 do the ANOVA lack of fit test

We will use the ad sales data used last week for the first three items and an expanded adsales data file for the last item. You should already have the adsales.txt data file from last week. If not, download it from the datasets page on the class web site.

**Omitting rows**

1. Read the ad sales data. JMP provides two mechanisms to omit data:
	1. exclude: to omit one or more rows from analysis (or other calculations
	2. hide: to omit one or more rows from graphs
2. Select the row (or shift-click or ctrl-click to select multiple rows) that you want to omit
3. Left-click Rows on the main menu, then select Exclude, Hide, or Hide and Exclude
	1. The red circle indicates excluded observations
	2. The block goggles indicate hidden observations
4. If you want to use these observations in a subsequent analysis, either
	1. select them again and click Exclude, Hide, or Hide and Exclude to clear the exclusions
	2. or select Rows / Clear Row States to clear all exclusions
5. We don’t need it, but in the future, you may want to use the extensive facilities for selecting rows based on the data. Look in Rows / Row Selection to see what is possible.

**Doing more with regression**

1. Read the data and use Analyze / Fit Y by X / Fit line to fit a straight-line regression. (Same as last week). You should see a results window with the following:



1. Most of what we want is part of the default output in the window above. To find:

s, the estimate of the error standard deviation: look for Root Mean Square Error

the standard error of a regression slope: look in the Parameter Estimates box at the Std.Error column. The first row is the intercept; the second is the slope, labelled by the variable name (ADEXP\_X)

test of slope = 0: look in Parameter Estimates box at the Prob > |t| column. That has the two-sided p-values. If you wanted the p-value for the test of intercept = 0, look in the intercept row.

1. To get the **95% confidence interval for the slope**, you have to ask for it. This is an optional statistic reported in the Parameter Estimates box at the bottom of the numeric results. In the Parameter Estimates box, right click on the variable name (ADEXP\_X), select Columns from the popup menu and click on Lower 95%. Repeat (right click / Columns) and click on Upper 95%. Those two columns are the 95% confidence interval for each parameter.

As far as I know, JMP will only report 95% intervals for regression parameters. If you need others (e.g., 90% interval or 99% interval), you have to compute it yourself from the estimate and standard error.

1. JMP provides lots of other things, you just have to ask for them.

Since these are extensions to the Linear Fit, they are all found by left clicking the red triangle by Linear Fit (between the plot and the numeric results).

Note: This is NOT the red triangle at the top of the window by Bivariate Fit. It's then one below the plot.

The popup window provides lots of options. It should look like:



To plot confidence intervals for the line (mean): click Confid Curves Fit or Confid Shaded Fit.

Curves adds lines to the plot; Shaded shades the regions between the two confidence lines.

To plot prediction intervals for observations: click Confid Curves Indiv or Confid Shaded Indiv.

Curves and Shaded work just like they do for confidence intervals.

To change the coverage (from the default 95%): click Set α level. 0.05 gives 95% intervals, 0.10 gives 90% intervals. You can change this after plotting the intervals and JMP will change the plot.

To store the numbers for confidence intervals or prediction intevals: click Mean Confidence Limit Formula (for confidence intervals) or Indiv Confidence Limit Formula (for prediction intervals). The intervals are added as new columns in the adsales data table. You can get the numbers for each row Use Set α level to change from 95% to desired coverage.

**Confidence or Prediction intervals for new X values:**

The confidence intervals and/or prediction intervals are computed for the observations in the data set. To get intervals for new values of X, add that X value (or multiple X values) as dummy X values to the data table. Here are the detailed instructions, using the adsales data as the example. Click on the data table window to make it the active window, click on an empty cell in the ADEXP\_X column and enter the desired X value. The intervals are calculated for that X value (so long as Mean Confidence Limit Formula and/or Indiv Confidence Limit Formula has been checked). A dot indicates a missing value (e.g. in the Y column).

To get predicted values for each observation: Click Save Predicted. Predicted values are computed for observations used to fit the regression and any additional X values you provide.

**To get residuals for each observation**: Click Save Residuals. Residuals are only defined for the observations used to fit the regression. If you added new X values, JMP should indicate a dot (missing value) for those residuals.

To get a **residual plot and a QQ plot of residuals**: click Plot Residuals. A plot of Residuals vs Predicted values (what we have called a residual plot) is the top plot. A QQ plot of residuals (to check normality) is the last plot. The other three plots are less useful.

Note: If the plots look weird (e.g. a poor scale for the X axis), it is probably because JMP has gotten confused by the additional dummy X values. close all windows except the JMP Home Window, reload the data, restart Analyze/Fit Y by X, select the Y variable, the X variable, and Fit Line again, and then chose Plot Residuals.

5. You can get essentially the same information and sometimes even more by using the Analyze / Fit Model dialog instead of the Fit Y by X dialog. It’s mostly a matter of finding which menu to look in. Here’s a quick guide:

confidence intervals for parameters: Red triangle by Response (very top left): Regression Reports, then select Show all confidence intervals. The Parameter Estimates box now includes 95% intervals for each regression parameter.

predictions: Red triangle by Response (very top left): Save Columns, then select Predicted values or Prediction Formula. Choosing the first (Interval) gives you values for the X values currently in the data set. Choosing the second (Formula) allows you to add new X values and the data table will automatically update with the predicted value.

predictions for observations not in the data set: Basically, add the desired X values to the data set. Leave the Y value missing (a dot). That’s because we want JMP to make a prediction at a desired X but we don’t want to change the estimated regression line, but. Need an X value for the first and a missing Y value to not change the estimates. If you use the Predicted values version, you need to add the new data points before you fit the regression. If you use the Prediction Formula, you can add new data points after fitting the regression.

confidence intervals for observations: Red triangle by Response (very top left): Save Columns, then select Mean Confidence Interval or Mean Confidence Limit Formula. Same deal as with predictions; the formula version will automatically update when you add new observations.

prediction intervals for observations: Red triangle by Response (very top left): Save Columns, then select Individual Confidence Interval or Individual Confidence Limit Formula. Yes, a bit confusing (perhaps) but what the rest of the world calls a prediction interval is an individual confidence interval for JMP.

6. ANOVA lack of fit test. This is automatically calculated by JMP when the data are appropriate (i.e., when you have replicate observations at the same X value). The adsales2.txt data has repeat observations at some X values.

If you use Analyze / Fit Y by X, look for the Lack of Fit box between the Summary of Fit and the ANOVA table. That’s minimized (vertical light gray triangle). Click that small triangle and the test box will open up. The line labelled Pure Error is the information about the ANOVA (means) model. The Lack of fit line is the difference between the regression fit and the ANOVA fit. The p-value for that test is labelled Prob > F. If it’s small (e.g. < 0.05), you have evidence that the postulated regression model **does not fit** the data. In this case, significance is bad. If you don’t see the words Lack of Fit, it is probably because the data doesn’t have any replicates (same X values, so the test can’t be done).

If you use Analyze / Fit Model, you see the Lack of Fit output by default – it’s between the ANOVA table and the Parameter Estimates box.