Quantitative Methods (GEO 441)
SPSS Lab 5: Regression
Dr. Paul Marr

Please copy the file **S:\GEO\Marr\Quantitative Methods\SPSS Example Data\Bivariate Regression.sav**, **Non-linear Regression.sav**, and **Multiple Regression.sav** to your portable media.

- Start SPSS.

Bivariate Regression

1. Open **Bivariate Regression.sav**.
2. Graphs > Legacy Dialogs > Scatter/Dot
    a. Choose *Simple Scatter* then click **Define**.
    b. Move **Median Income($)** to the *Y-axis* field and **College Graduates** to the *X-axis* field.
    c. Click **Ok**.
        i. Note that the scatterplot shows a weak relation between the number of college graduates and median income.
        ii. These data need to be normalized based on total population.
3. Transform > Computer Variable…
    a. Enter **PctGrad** in the *Target Variable* field.
    b. Enter **(COLLGRADS/POP1990)*100** into the *Numeric Expression* field.
    c. Click **Ok**.
4. Graphs > Legacy Dialogs > Scatter/Dot
    a. Choose *Simple Scatter* then click **Define**.
    b. Move **Median Income($)** to the *Y-axis* field and **PctGrad** to the *X-axis* field.
    c. Click **Ok**.
        i. Note that the scatterplot shows a stronger relation between the number of college graduates and median income.
        ii. Also note that one observation is well away from the main grouping.
5. Analyze > Regression > Linear…
    a. Move **Median Income ($)** to the *Dependent* field.
    b. Move **PctGrad** to the *Independent(s)* field.
    c. Click on the **Statistics…** button. Check the *Durbin-Watson* and *Casewise diagnostics* boxes.
    d. Click **Ok**.
        i. What is the R^2? What is the standard error of the estimate?
        ii. Is the model significant?
        iii. What is the model equation?
    e. Examine the *Casewise Diagnostics* table.
        i. Note that the standard residual value for case 101 is -4.497 stdev from the mean.
6. In the *Output Navigation Pane*, click on the scatterplot graph. Double click the scatterplot in the *Output Window* to activate the *Graph Editor*.
    a. Click on the *Case Label* button and then click on the obviously dissimilar case.
        i. What is its case number?
    b. Go to the *Data Editor* window and write down the FIPS number for that case.
7. Click on the *Recall* button and select **Linear Regression**.
    a. Move **FIPS** to the *Selection Variable* field.

b. Click on the **Rule…** button.
    i. Change the logical operator to 'Not equal to'.
    ii. Type in the FIPS code you wrote down into the *Value* field.
    iii. Click **Continue**.
c. Click the **Plots…** button.
    i. Move ***ZPRED** to the *Y-axis* field and ***ZRESID** to the *X-axis* field.
    ii. Check the *Normal probability plot* box.
    iii. Click **Continue**.
d. Click the **Save…** button.
    i. Check the **Unstandardized** *Predicted Values* box and the **Unstandardized** *Residuals* box.
    ii. Click **Continue**.
e. Click **Ok**.
    i. What is the R^2? What is the standard error of the estimate?
    ii. Is the model significant?
    iii. What is the model equation?
    iv. Are the residuals normally distributed?
    v. Note that the predicted and residual values have been saved.

<u>Multiple Regression</u>

1. Open **Multiple Regression.sav**.
2. Analyze > Regression > Linear…
    a. Add **Life Expectancy** to the *Dependent* field.
    b. Add the following variables to the *Independent(s)* field.
        i. Add Percent Out-of-Pocket Health Expenditures
        ii. Per Capita Health Expenditures in USD
        iii. Infant Mortality Rate per 1000
        iv. Births per Woman
    c. Click the **Statistics…** button.
        i. Check the *Collinearity diagnostics* box.
        ii. Check the *Casewise diagnostics* box.
        iii. Click **Continue**.
    d. Click **Ok**.
        i. What is the adjusted R^2? What is the standard error of the estimate?
        ii. Is the model significant? Are all of the independent variable significant?
    e. Examine the *Coefficients* table.
        i. Note the low *Tolerance* values for **Infant Mortality** and **Births per Woman**. The tolerance value is the amount of variation the variable contributes that is NOT correlated with other variables. Therefore, low tolerance values are not good.
    f. Examine the *Collinearity Diagnostics* table.
        i. Note that the highest *Variance Proportion* values for **Infant Mortality** and **Births per Woman** are on the same model dimension.
        ii. This suggests that these two variables are highly correlated.
        iii. Since the t statistic probability for Births per Woman is not significant, and it is correlated with Infant Mortality, it will be removed from the analysis.
3. Click on *Recall* and select **Linear Regression**.
    a. Remove **Births per Woman** from the *Independent(s)* field.

b. Click on **Plots…**
    i. Move ***ZPRED** to the *Y-axis* field and ***ZRESID** to the *X-axis* field.
    ii. Check the *Normal probability plot* box.
    iii. Click **Continue**.
c. Click **Ok**.
d. Examine the *Model Summary Table*.
    i. What is the R^2 and standard error of the estimate?
e. Examine the *ANOVA* table.
    i. Is the model significant?
f. Examine the *Coefficients* table.
    i. Are the independent variables significant?
    ii. Are the *Tolerance* values reasonable?
g. Examine the *Collinearity Diagnostics* table.
    i. Note that the highest *Variance Proportions* for each variable occur on different model dimensions.
h. Examine the *Casewise Diagnostics* table.
    i. Several observations have larger than expected residual values.
i. Examine the *PP-PLOT* and *SCATTERPLOT* graphs.
    i. There do not appear to be any observations that are obviously outliers.
j. What is our final mode equation?

<u>Non-Linear Regression</u>

1. Open **Non-linear Regression.sav**.
2. Graphs > Legacy Dialogs > Scatter/Dot
    a. Choose *Simple Scatter* then click **Define**.
    b. Move **Percent Commuters** to the *Y-axis* field and **Distance Range Midpoint** to the *X-axis* field.
    c. Click **Ok**.
        i. Note that the scatterplot shows a definite relation between the percentage of commuters and the distance traveled to work.
3. Analyze > Regression > Linear…
    a. Move **Percent Commuters** to the *Dependent* field and **Distance Range Midpoint** to the *Independent(s)…* field.
    b. Click **Ok**.
        i. What is the adjusted R^2 for this model?
        ii. Is the model significant?
        iii. What is the model equation?
4. From the *Output Navigation Pane* click on the graph for US Commuting Patterns.
    a. In the *Output Window* double click on the graph to open the *Graph Editor*.
    b. Click the *Add Fit Line* button.
        i. Note that the Add Fit Line function uses the unadjusted R^2 value.
        ii. Do the observations fall above and below the line of best fit equally?
    c. Close the *Graph Editor*.
5. Analyze > Regression > Curve Estimation…
    a. Move **Percent Commuters** to the *Dependent(s)…* field, move **Distance Range Midpoint** to the *Independent, Variable* field.
    b. Check the **Exponential** box under *Models*.

c. Check the *Include constant in equation* and *Plot models* boxes.
d. Click the **Save…** button and check the *Predicted Values* and *Residuals* boxes.
e. Check the *Display ANOVA table* box.
f. Click **Ok** and **Ok** again.
    i. What is the adjusted R^2 for this model?
    ii. Is the model significant?
    iii. What is the model equation?
    iv. To determine the predicted values (on my calculator):

    where *a* = 52.045 and *b* = -0.093 (from the SPSS output)

    $\hat{y} = a \times e^{-b(x)}$ or 52.045 × 0.093 +/− × 3 (-) SHIFT e^x = for x = 3miles

    v. Note that this technique results in predicted values that are in the original units. Log transforming the data is also appropriate but would result in logged units that would need to be transformed back into their original units to be interpreted.
g. Make the *Data Editor* window active.
    i. Note that the predicted and residual values were saved as variables.
    ii. Also note that the predicted percent commuters for a 3 mile commute is 39.42571 which is over-predicted by 4.42571 (obs – exp).
    iii. Therefore, negative values denote over-prediction.