**STAT 501 – Homework 4 (covering Lesson 4) – due date September 20th**

**Instructions**: Use Word to type your answers within this document. Then, submit your answers in the appropriate dropbox in ANGEL by the due date. The point distribution is located next to each question. If there are multiple parts, then the points are divided equally over the subparts.

1. **(3x8 = 24 points)** Say whether the following statements about residual analysis for simple linear regression are true or false and briefly explain your answers.
   1. If the sample mean of the residuals is zero, then this supports the linearity condition for the model.
   2. An ideal residual plot for a valid model will display residuals with a strong positive or negative linear trend.
   3. An ideal residual plot for a valid model will display residuals with similar variation no matter the value on the horizontal axis.
   4. We should only assess the linearity and equal variance conditions after first confirming the normality condition using a normal probability plot.
   5. A residual plot that has a predictor variable excluded from the model on the horizontal axis provides no useful information about the adequacy of the model.
   6. We can accept the validity of our model if at least one of the four LINE conditions is supported.
   7. We should question the validity of our model only if all four LINE conditions seem in doubt.
   8. Residual analysis is entirely objective so that every question has a right or wrong answer.
2. **(2x4 = 8 points)** The dataset “Auto” gives information on 392 different cars including their mpg, weight, horsepower, displacement, year of make, country of origin etc. A car buyer is interested in regressing mpg (Y) on horsepower (X). Given below is the residual plot of the regression.



1. What departure, if any, from the simple linear regression model assumptions can be assessed from this plot?



1. The residuals are now regressed on another predictor weight. Residual plot of this regression is given below. Would you say that inclusion of weight as an additional predictor has improved the residual plot? Briefly justify your answer.
2. **(8 points)** “Auto” data includes cars from three different origins: 1 (USA), 2 (Europe) and 3 (Japan). The variable *acceleration* (time to accelerate) is regressed on the variable *displacement*, separately for the cars from different origins. Following are the residual plots from three separate regression equations.

|  |
| --- |
|  |
|  |
|  |

Compare the residual plots for the cars of three different origins in terms of (i) whether linearity holds (ii) whether the response follows normal distribution and (iii) whether the errors have equal variances.

1. (**6x5 = 30 points**) Consider the “Desired Height” dataset. The data are from females in two introductory statistics classes at Penn State. The variables are *x* = Height (student's height in centimeters) and *y* = DesiredHt (how tall the student said they would like to be, in centimeters, if they could be any height at all). (Note: Students gave responses in inches, but we converted them to centimeters.)
   1. Draw a scatterplot of *y* = DesiredHt versus *x* = Height. Include the graph and write a brief description of the graph's main features in the context of simle linear regression. [In Minitab, the menu sequence is Graph > Scatterplot. To copy a Minitab graph, right click on it and use “Copy Graph” in the resulting menu.]
   2. Determine the estimated sample regression equation for a straight-line relationship between desired height and actual height. [In Minitab, use the menu sequence Stat > Regression > Regression > Fit Regression Model (v17, exclude last part in v16).]
   3. The output that you created for the previous part will include a value of *R*2. What is that value? Write a sentence that interprets the value in the context of this situation.
   4. In the output that you created, find evidence that the observed relationship is statistically significant. Describe the evidence that you found (for instance, the test for . . . had a *p*-value of . . ., which means . . . .)
   5. Create a plot of residuals versus predicted values (fits) for this situation. Include the plot and briefly discuss what the plot indicates about the validity of the model and assumptions about the errors. [In Minitab, click the Graphs button in the Regression dialog and then select “Residuals versus Fits.”]
   6. Create a histogram of the residuals and a normal probability plot of the residuals. Include the plots and briefly discuss what they indicate about the validity of the model and assumptions about the errors. [In Minitab, click the Graphs button in the Regression dialog and then select “Histogram of residuals” and also “Normal probability plot of residuals.”]
2. (**6x5 = 30 points**) Consider the “Compare” dataset. There is a single response variable, *Y*, and four potential predictor variables, *X1*, *X2*, *X3*, and *X4*. Fit four simple linear regression models, with each of the predictor variables, *X1*, *X2*, *X3*, and *X4* in turn. For each model create a residual plot with residuals on the vertical axis and fitted values on the horizontal axis.
   1. Compare the values of *S* (root mean square error), *R2*, and the slope *t*-statistic for each model (the clearest way to present this information is in a table). *Based solely on these values*, order the four models from “best” to “worst” (in terms of predicting or explaining *Y*). If two or more models appear to be essentially equivalent based on these criteria, make a note of this.
   2. What does the residual plot for the model with *X1* as the predictor indicate about the validity of this regression model and assumptions made about the errors?
   3. What does the residual plot for the model with *X2* as the predictor indicate about the validity of this regression model and assumptions made about the errors?
   4. What does the residual plot for the model with *X3* as the predictor indicate about the validity of this regression model and assumptions made about the errors?
   5. What does the residual plot for the model with *X4* as the predictor indicate about the validity of this regression model and assumptions made about the errors?
   6. Given your answers for parts (b) to (e), which of the four models now appears to be “best?” Explain your answer.