**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.

False. The least squares method works in a way where the sample mean of the residuals is zero for the line of best fit. Therefore that fact in itself doesn’t support the linearity condition for the model.

* 1. An ideal residual plot for a valid model will display residuals with a strong positive or negative linear trend.

False. The ideal residual plot wherein the assumption that the relationship is linear is reasonable, has the residuals that "bounce randomly" around the 0 line. Any pattern / trend means that linearity assumption is doubtful.

* 1. An ideal residual plot for a valid model will display residuals with similar variation no matter the value on the horizontal axis.

True. This is a validation of the equal variance assumption of the linear regression model. However it is worth noting that non-constancy of error variance tends to be less serious, leading to less efficient estimates and invalid error variance estimates.

* 1. We should only assess the linearity and equal variance conditions after first confirming the normality condition using a normal probability plot.

False. We know that the confidence and prediction intervals are not very sensitive to departures from normality (CI is less sensitive than PI) while linearity and equal variance conditions are more important.

Data analysis is an art based on science and such hard and fast rules about sequences of tests don’t serve will.

* 1. 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.

False. Residual plots against predictor variables not a part of the model can actually be used to validate whether their addition to the model will be valuable or not. Therefore iteratively adding predictor variables and having their residual plots against fitted values and predictor variables that are not a part of the model is an important analytics exercise.

* 1. We can accept the validity of our model if at least one of the four LINE conditions is supported.

False. All the conditions “L I N E” are important and critical for the validity of the model. Though the fidelity of the model to the various assumptions varies in terms of their impact on the accuracy of the model, the assumptions however are the building blocks of the linear regression model.

* 1. We should question the validity of our model only if all four LINE conditions seem in doubt.

False. Several types of departures from the simple linear regression model can be identified by diagnostic tests of the residuals. Their effect on the accuracy of the model varies:

* Model misspecification due to either nonlinearity or the omission of important predictor variables tends to be serious
* Nonconstancy of error variance tends to be less serious
  1. Residual analysis is entirely objective so that every question has a right or wrong answer.

False. Data analysis is an art (subjective decisions!) based on science (objective tools!)

It requires developing an intuition and comes through gaining perspective from experience.

1. **(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?

The following departures are evident from the plot:

1. First, the pattern seems curved (slight U) that indicates that the wrong type of equation was used. Curvilinear may be more appropriate.
2. This plot of residuals versus fits shows that the residual variance increases as the fitted values increase. This violates the assumption of constant error variance.



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.

Yes. The inclusion of weight predictor has reduced both of the effects highlighted in part a.

1. First, the pattern still seems curved but is markedly less curved than the previous version.
2. This plot of residuals versus weight still shows variation in the variance of the residuals. However even in this case the variation is markedly reduced.
3. **(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.

**Whether linearity holds**: The preferable method of detection of linearity is the examination of residual plots. In case of a well-behaved residual vs. fits plot, the residuals "bounce randomly" around the 0 line. This suggests that the assumption that the relationship is linear is reasonable.

In our case# 3 it seems that the residuals bounce randomly (though with different variance at different levels of fitted value) indicating linearity. We do see a similar pattern in case #2’s residual vs fitted value plot. However we see some nonlinear pattern in the case #1’s residual vs fitted value plot casting a doubt on the assumption of linearity.

**Whether the response follows normal distribution**: For the normality detection we rely on the normal probability plot and the histogram. From the figures, it is clear that case # 1 meets this condition - the normal probability plot of the residuals is approximately linear supporting the condition that the error terms are normally distributed.

Case #2 indicates that the distribution of the residuals is skewed to the right casting a slight doubt on the normality assumption.

Case #3 indicates that the distribution of the residuals is skewed to the left again casting a slight doubt on the normality assumption.

**Whether the errors have equal variances**: The preferable method here again is the examination of residual plots. We see that for case 1 while the residual bounce randomly there is an increase in the variance as the fitted value increases – leading to a fanning effect.

Case # 2 indicates equal variance. There are thinner tails in the plots that may require some further analysis but other than that there is no observable pattern.

Case # 1 indicates equal variance. The graph does show some cyclical pattern, however the variance seems to be close to being constant across the different values of the fit.

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.]



The scatterplot shows the following features

1. The plot indicates that the mean desired height increases as height increases.
2. The variance of desired height at different height level is almost but not completely uniform
   1. 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).]

The model from minitab gives:

**Regression Analysis: DesiredHt versus Height**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 1 3716.4 3716.39 197.72 0.000

Height 1 3716.4 3716.39 197.72 0.000

Error 249 4680.3 18.80

Lack-of-Fit 16 634.5 39.66 2.28 0.004

Pure Error 233 4045.8 17.36

Total 250 8396.7

Model Summary

S R-sq R-sq(adj) R-sq(pred)

4.33549 44.26% 44.04% 43.25%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 77.84 6.45 12.06 0.000

Height 0.5518 0.0392 14.06 0.000 1.00

Regression Equation

DesiredHt = 77.84 + 0.5518 Height

* 1. 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.

The value of R2 is 44.26%

We can say that 44.26% of the variation in the DesiredHt is reduced by taking into account Height. Or, we can say that 44.26% of the variation in DesiredHt is 'explained by' Height.

* 1. 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 . . . .)

The evidence lies in the following section of the output:

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 77.84 6.45 12.06 0.000

Height 0.5518 0.0392 14.06 0.000 1.00

Here we can see that the P-Value for the height co-efficient has value < 0.001

For the testing of the null hypothesis that the population slope β1 equals 0, the t-statistic = 14.06 and the p-value = 0.000 (< 0.001)

Since the p-value < alpha = 0.01 and therefore we reject the null hypothesis that the population slope equals 0. Therefore, the conclusion we draw is that there is a relationship between the variables DesiredHt and Height.

* 1. 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.”]



The residual vs fit plot indicates:

1. There is no obvious pattern in the residuals indicating that the model is valid. The residuals bounce randomly around the 0 line.
2. The variance of the residuals is varying slightly for different values of the fitted values (it is almost but not completely uniform). However it is worth noting that non-constancy of error variance tends to be less serious, leading to less efficient estimates and invalid error variance estimates.
   1. 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.”]





The plots above indicate that the residuals are not perfectly normal. The histogram and the normal probability plot suggests that the residuals are slightly skewed (to the right).

1. (**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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | S | R2 | T-Statistic (slope) | P-Value | Model Rank |
| X1 | 1.06996 | 80.65% | 14.14 | 0.000 | 4 |
| X2 | 1.01855 | 82.46% | 15.02 | 0.000 | 1 |
| X3 | 1.01873 | 82.46% | 15.02 | 0.000 | 2 |
| X4 | 1.01878 | 82.45% | 15.02 | 0.000 | 3 |

Note: While model 2 (with variable X2) has slightly less root mean square error than 3, they are essentially equivalent. Even model 4 is only very slightly worse than model 2 and 3. Therefore model 2, 3 and 4 are essentially equivalent.

* 1. 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?



The residuals indicate the following:

1. It appears that the residuals bounce randomly around the residual = 0 line. There is no obvious pattern in the residuals indicating that the model is valid
2. The variance of the residuals appears to be almost constant across the different fitted values indicating that the equal variance assumption is valid.
   1. 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?



The residuals indicate the following:

1. The plot has a "fanning" effect. That is, the residuals are close to 0 for small x values and are more spread out for large x values. This casts a doubt on the equal variance assumption.
2. The three residual entries between fitted value 2 and 3 indicate some pattern that is different from the remaining points. This may cast a doubt on the validity of the model but requires more analysis since there isn’t any observable curvilinear pattern.
3. The residuals (other than the 3 mentioned above) bounce randomly around the residual = 0 line.
   1. 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?



The residuals indicate the following:

1. There is some non linear pattern in the layout of the residuals. This casts a doubt on the assumption of linearity and therefore the validity of the model.
2. The spread of the residuals in the residuals vs. fits plot varies in some complex fashion. This casts a doubt on the equal variance assumption.
   1. 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?



The residuals indicate the following:

1. There is some non linear pattern in the layout of the residuals. This casts a doubt on the assumption of linearity and therefore the validity of the model.
2. The spread of the residuals in the residuals vs. fits plot varies in some complex fashion. This casts a doubt on the equal variance assumption.
   1. Given your answers for parts (b) to (e), which of the four models now appears to be “best?” Explain your answer.

The best model now appears to be where the predictor is X1. This was ranked the 4th in the initial answer.

The results reinforce the point that data analysis is an art based on science. There are multiple facets and together they give a holistic picture. One can’t blindly rely on any single concept / idiom and make absolute statements.