**STAT 501 – Homework 11(covering lessons 12 and 13) – Due Sunday Nov 22**

**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.

1. **(4+5+5+4+4+5+3+5+4+5+3+3 = 50 points)**

Open the “AgeSaving” dataset. The variables *savings*, *age*, and *income* were originally obtained from 50 individuals with the goal of investigating the dependence of *savings* on *age* and *income*. Data from a further 10 individuals were obtained at a later date (variables *savingsnew*, *agenew*, and *incomenew* contain the original 50 observations plus these 10 new ones). As you work through this problem, complete the table below titled ‘Results of Regression Analyses.” The row corresponding to part (b) has already been completed for you.

**Results of Regression Analyses**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Part | Model | LINE conditions satisfied? | | | Sample equation | S | R2 adj |
| **Linearity** | **Normality** | **Equal variance** |
| b | *savings* vs *age* | N | Y | Y | savings = 1.92 + 1.5247 age | 3.56198 | 90.76% |
| c | *savings* vs *age, income* | N | Y | Y | savings = -11.27 + 0.830 age + 0.952 income | 3.46443 | 91.26% |
| f | *savingsnew* vs *agenew, incomenew* | N | Y | Y | savingsnew = -14.53 + 0.6757 agenew + 1.177 incomenew | 3.50033 | 90.27% |
| h | *savingsnew* vs *agenew, incomenew, agenew2* | Y | Y | Y | savingsnew = 3.49 - 0.441 agenew + 1.179 incomenew + 0.01636 agenew\*agenew | 3.40931 | 90.77% |
| j | *savingsnew* vs *agenewc, incomenew, agenewc2* | Y | Y | Y | savingsnew = 7.22 + 0.6624 agenewc + 1.179 incomenew + 0.01636 agenewc\*agenewc | 3.40931 | 90.77% |

1. Construct a Matrix Plot of *savings*, *age*, and *income*. Briefly describe what this plot tells us about potential problems if we fit an MLR model with both *age* and *income* as predictors.



The problem is high level of multicollinearity between the predictors (age and income).

**Correlation: savings, age, income**

savings age

age 0.954

0.000

income 0.952 0.983

0.000 0.000

1. Fit an SLR model for *Y=savings* and *X=age*. Obtain a Residuals vs Fits plot and a Normal Probability Plot of the residuals. Briefly comment on your findings and confirm the entries in the table above.





* The residuals are normally distributed. P-value for Ryan-Joiner test > 0.1 and so we don’t have enough evidence to doubt normality of the residuals.
* The residuals vs fits plot indicates that the equal variance condition is also met
* The residuals vs fits plot indicates a non linear relationship between Y and X.

1. You should have found some evidence that the linearity assumption is questionable for the model in part (b). In an attempt to improve the model, we’ll next try adding the *income* variable to the model. Fit an MLR model for *Y=savings* and *X1=age* and *X2=income*. Obtain a Residuals vs Fits plot and a Normal Probability Plot of the residuals. Briefly comment on your findings and complete the entries in the part (c) row of the table above.





* The residuals are normally distributed. P-value for Ryan-Joiner test > 0.1 and so we don’t have enough evidence to doubt normality of the residuals.
* The residuals vs fits plot indicates that the equal variance condition is also met
* The residuals vs fits plot indicates a non linear relationship between Y and X.

1. What are the Variance Inflation Factors for *age* and *income* in the model from part (c)? What regression pitfall does this suggest and what can we do to mitigate this type of problem?

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant -11.27 7.21 -1.56 0.125

age 0.830 0.366 2.27 0.028 29.29

income 0.952 0.492 1.93 0.059 29.29

The high values of VIF indicate the signs of multicollinearity amongst the predictors.

In order to mitigate we could opt to remove one of the two predictors from the model. Alternatively, if we have a good scientific reason for needing both of the predictors to remain in the model, we could go out and collect more data.

1. Construct a Matrix Plot of *savingsnew*, *agenew*, and *incomenew*. Briefly compare this plot with the Matrix Plot in part (a).



The plot suggests that a correlation still exists between agenew and incomenew — it is just a weaker correlation now.

**Correlation: savingsnew, agenew, incomenew**

savingsnew agenew

agenew 0.886

0.000

incomenew 0.905 0.773

0.000 0.000

Reduced from 0.983 to 0.773

1. Fit an MLR model for *Y=savingsnew* and *X1=agenew* and *X2=incomenew*. Obtain a Residuals vs Fits plot and a Normal Probability Plot of the residuals. Briefly comment on your findings and complete the entries in the part (f) row of the table above.





* The residuals are normally distributed. P-value for Ryan-Joiner test > 0.1 and so we don’t have enough evidence to doubt normality of the residuals.
* The residuals vs fits plot indicates that the equal variance condition is also met
* The residuals vs fits plot indicates that the linearity relationship between Y and X has improved but there is still some non linear relationship to be addressed.

1. What are the Variance Inflation Factors for *agenew* and *incomenew* in the model from part (f)? Has the regression pitfall from part (d) been mitigated?

The minitab output:

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant -14.53 3.52 -4.13 0.000

agenew 0.6757 0.0934 7.24 0.000 2.48

incomenew 1.177 0.138 8.55 0.000 2.48

The table shows that the VIF values have reduced from 29.29 to 2.48

This indicates that the regression pitfall from part has been mitigated.

1. You should have found some evidence that the linearity assumption remains questionable for the model in part (f). In an attempt to improve the model, we’ll next try adding an *agenew2* variable to the model. Fit an MLR model for *Y=savingsnew* and *X1=agenew*, *X2=incomenew*, and *X3=agenew2*. (An easy way to do this in Minitab v17 is to click the Model button in the Regression Dialog, highlight *agenew* in the top-left Predictors box, then click “Add” to the right of “Terms through order: 2” so that *agenew\*agenew* appears in the “Terms in the model” list.) Obtain a Residuals vs Fits plot and a Normal Probability Plot of the residuals. Briefly comment on your findings and complete the entries in the part (h) row of the table above.





* The residuals are normally distributed. P-value for Ryan-Joiner test > 0.1 and so we don’t have enough evidence to doubt normality of the residuals.
* The residuals vs fits plot indicates that the equal variance condition is also met
* The residuals vs fits plot indicates that the linearity relationship between Y and X has improved but there is still some non linear relationship to be addressed.

1. What are the Variance Inflation Factors for *agenew* and *agenew2* in the model from part (h)? What regression pitfall does this suggest and what can we do to mitigate this type of problem?

Minitab output:

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 3.49 9.55 0.36 0.717

agenew -0.441 0.560 -0.79 0.435 94.15

incomenew 1.179 0.134 8.79 0.000 2.48

agenew\*agenew 0.01636 0.00809 2.02 0.048 92.51

The problem is high level of multicollinearity between the predictors (agenew and agenew\* agenew). We have structural multicollinearity since a mathematical artifact caused by creating new predictors from other predictors i.e. the predictor agenew\* agenew from the predictor agenew

In order to correct this we have to do "centering the predictors." Centering a predictor merely entails subtracting the mean of the predictor values in the data set from each predictor value.

1. Use the Minitab calculator to create a centered *agenew* variable stored in a variable called *agenewc* and defined as “agenew-mean(agenew).” Then fit an MLR model for *Y=savingsnew* and *X1=agenewc*, *X2=incomenew*, and *X3=agenewc2*. Obtain a Residuals vs Fits plot and a Normal Probability Plot of the residuals. Briefly comment on your findings and complete the entries in the part (j) row of the table above.





* The residuals are normally distributed. P-value for Ryan-Joiner test > 0.1 and so we don’t have enough evidence to doubt normality of the residuals.
* The residuals vs fits plot indicates that the equal variance condition is also met
* The residuals vs fits plot indicates a linear relationship between Y and X.

1. What are the Variance Inflation Factors for *agenewc* and *agenewc2* in the model from part (j)? Has the regression pitfall from part (i) been mitigated?

The output:

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 7.22 5.20 1.39 0.170

agenewc 0.6624 0.0912 7.26 0.000 2.50

incomenew 1.179 0.134 8.79 0.000 2.48

agenewc\*agenewc 0.01636 0.00809 2.02 0.048 1.01

The table shows that the VIF values have reduced from 94.15 to 2.5 for agenewc and from 92.51 to 1.01 for agenewc\*agenewc

This indicates that the regression pitfall from part has been mitigated.

1. Use the model from part (h) to predict *savingsnew* for an individual with *agenew* = 30 and *incomenew* = 45 (a point prediction is sufficient, no need for an interval). Then use the model from part (j) to predict *savingsnew* for this same individual (i.e., since mean(*agenew*) = 33.7167, with *agenewc* = –3.7167 and *incomenew* = 45). Explain your findings.

**Model with centering (part j)**

For the case: agenew = 30, agenewc = 30 - 33.7167 = –3.7167, and incomenew = 45

We have:

Regression Equation

savingsnew = 7.22 + 0.6624 agenewc + 1.179 incomenew + 0.01636 agenewc\*agenewc

Variable Setting

agenewc -3.7167

incomenew 45

Fit SE Fit 95% CI 95% PI

58.0455 1.27895 (55.4835, 60.6076) (50.7511, 65.3399)

**Model without centering (part h)**

For the case: agenew = 30 and incomenew = 45

We have:

Regression Equation

savingsnew = 3.49 - 0.441 agenew + 1.179 incomenew + 0.01636 agenew\*agenew

Variable Setting

agenew 30

incomenew 45

Fit SE Fit 95% CI 95% PI

58.0455 1.27895 (55.4835, 60.6076) (50.7511, 65.3400)

We see that we get the same value for the FIT for both the cases.

1. **(14+6+6 = 26 points)**

Use the “AgeSaving” dataset used in the previous problem. Fit the following six models:

* (where *Y=savings*, *X1=age*)
* (where *Y=savings*, *X2=income*)
* (where *Y=savings*, *X1=age,* *X2=income*)
* (where *Y=savingsnew,* *X1=agenew*)
* (where *Y=savingsnew,* *X2=incomenew*)
* (where *Y=savingsnew,* *X1=agenew,* *X2=incomenew*)

1. Use the model results to complete the following table (the first row has been completed for you):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | b1 | se(b1) | b2 | se(b2) |
| *savings* vs *age* | 1.5247 | 0.0694 | XXX | XXX |
| *savings* vs *income* | XXX | XXX | 2.0493 | 0.0948 |
| *savings* vs *age, income* | 0.830 | 0.366 | 0.952 | 0.492 |
| *savingsnew* vs *agenew* | 1.2930 | 0.0887 | XXX | XXX |
| *savingsnew* vs *incomenew* | XXX | XXX | 1.947 | 0.120 |
| *savingsnew* vs *agenew, incomenew* | 0.6757 | 0.0934 | 1.177 | 0.138 |

1. The first three models use data in which the two predictors are sufficiently highly correlated to create data-based multicollinearity (as explored in the previous problem). Briefly describe how the results in the first three rows of the table illustrate how:
   1. The estimated regression coefficient of any one variable depends on which other predictor variables are included in the model;
   2. The precision of the estimated regression coefficients decreases as more predictor variables are added to the model.

The following can be said about the results:

1. When predictor variables are correlated, the estimated regression coefficient of any one variable depends on which other predictor variables are included in the model. Here's the relevant portion of the table:

|  |  |  |
| --- | --- | --- |
| Model | b1 | b2 |
| *savings* vs *age* | 1.5247 | XXX |
| *savings* vs *income* | XXX | 2.0493 |
| *savings* vs *age, income* | 0.830 | 0.952 |

Note that, depending on which predictors we include in the model, we obtain different estimates of the slope parameter for x1 and x2

When x1 = age is the only predictor included in our model, we claim that for every additional year (age), savings increases by 1.5247. On the other hand, if x1 = age and x2 = income are both included in our model, we claim that for every additional year in age, holding income constant, savings increases by only 0.830

When x2 = income is the only predictor included in our model, we claim that for every additional unit (income), savings increases by 2.0493. On the other hand, if x1 = age and x2 = income are both included in our model, we claim that for every additional unit in income, holding age constant, savings increases by only 0.952

1. When predictor variables are correlated, the precision of the estimated regression coefficients decreases as more predictor variables are added to the model. Here's the relevant portion of the table:

|  |  |  |
| --- | --- | --- |
| Model | se(b1) | se(b2) |
| *savings* vs *age* | 0.0694 | XXX |
| *savings* vs *income* | XXX | 0.0948 |
| *savings* vs *age, income* | 0.366 | 0.492 |

The standard error for the estimated slope b1 obtained from the model including only x1 and x2 is 5.3 times greater than the standard error for the estimated slope b1 obtained from the model including x1 alone.

The standard error for the estimated slope b2 obtained from the model including only x1 and x2 is 5.2 times greater than the standard error for the estimated slope b2 obtained from the model including x2 alone.

1. The last three models use additional data such that the correlation between the two predictors has been reduced. The hope is that the data-based multicollinearity has been mitigated. Do the results in the last three rows of the table support the following assertions?
   1. The estimated regression coefficient of any one variable no longer depends on which other predictor variables are included in the model;
   2. The precision of the estimated regression coefficients remains approximately the same as more predictor variables are added to the model.

*[Note: You should find that while the results support assertion (ii) to some extent, they don’t particularly support assertion (i) at all. The take-home message here is that in most observational studies there will be varying degrees of correlation between predictors that we can’t do anything about. This means that it is almost always the case that the estimated regression coefficient of any one variable depends on which other predictor variables are included in the model. The only time this is not the case is in datasets where the correlation between predictors is close to zero (which generally only occurs in designed experiments). The rest of the time we need to be careful to include all relevant predictor variables and interpret regression coefficients correctly (that is with reference to all the other predictor variables in the model).]*

The relevant results are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | b1 | se(b1) | b2 | se(b2) |
| *savingsnew* vs *agenew* | 1.2930 | 0.0887 | XXX | XXX |
| *savingsnew* vs *incomenew* | XXX | XXX | 1.947 | 0.120 |
| *savingsnew* vs *agenew, incomenew* | 0.6757 | 0.0934 | 1.177 | 0.138 |

1. Though the correlation has reduced, the estimated regression coefficient of any one variable still depends on which other predictor variables are included in the model. Here's the relevant portion of the table:

|  |  |  |
| --- | --- | --- |
| Model | b1 | b2 |
| *savingsnew* vs *agenew* | 1.2930 | XXX |
| *savingsnew* vs *incomenew* | XXX | 1.947 |
| *savingsnew* vs *agenew, incomenew* | 0.6757 | 1.177 |

Note that, depending on which predictors we include in the model, we obtain different estimates of the slope parameter for x1 and x2

When x1 = agenew is the only predictor included in our model, we claim that for every additional year (agenew), savingsnew increases by 1.2930. On the other hand, if x1 = agenew and x2 = incomenew are both included in our model, we claim that for every additional year in agenew, holding incomenew constant, savingsnew increases by only 0.6757

When x2 = incomenew is the only predictor included in our model, we claim that for every additional unit (incomenew), savingsnew increases by 1.947. On the other hand, if x1 = agenew and x2 = incomenew are both included in our model, we claim that for every additional unit in incomenew, holding agenew constant, savingsnew increases by only 1.177

Therefore assertion (i) is not true.

1. When predictor variables are less correlated, the precision of the estimated regression coefficients remains stable as more predictor variables are added to the model. Here's the relevant portion of the table:

|  |  |  |
| --- | --- | --- |
| Model | se(b1) | se(b2) |
| *savingsnew* vs *agenew* | 0.0887 | XXX |
| *savingsnew* vs *incomenew* | XXX | 0.120 |
| *savingsnew* vs *agenew, incomenew* | 0.0934 | 0.138 |

The standard error for the estimated slope b1 obtained from the model including only x1 and x2 is 1.05 times (5% only) greater than the standard error for the estimated slope b1 obtained from the model including x1 alone.

The standard error for the estimated slope b2 obtained from the model including only x1 and x2 is 1.15 times (15% only) greater than the standard error for the estimated slope b2 obtained from the model including x2 alone.

1. **(6x4 = 24 points)** Use the “CarStopping” dataset. The dataset gives data about the relationship between the stopping distance of a car and the speed of the car when the brakes are applied. Earlier in the course, we used this data to show that transforming the ***y***-variable might help us model a curve and non-constant variance at the same time. Here we will not transform ***y***, but instead we will model the curve with a quadratic model and use weighted least squares to model the non-constant variance.

The variables are the car’s stopping distance (StopDist), the car’s speed (minus 19) (spdmn19) and the square of the car’s speed minus 19 (spdsqrd). The value “19” is approximately the average speed in the data set. By centering speed, we lessen the correlation between the linear and squared terms in a quadratic model. The overall fit is the same whether we center or not, but we get a bit cleaner look at the linear and quadratic contributions with centering.

1. Graph ***y*** = StopDist versus ***x*** = spdmn19. Comment on the important features of the relationship.



* The plot indicates a non linear (curvilinear) relationship between Y and X.
* The plot indicates that the equal variance condition is not met. The variance is higher as spdmn19 increases

1. Fit a multiple linear regression model with **y** = StopDist and **x**-variables spdmn19 and spdsqrd. Store the Fits (predicted values) and Residuals (use the *Storage* button in the Regression Dialog). Plot the Residuals versus Fits (use the *Graphs* button). Describe the difficulty that is indicated by this residual plot.



* The residuals vs fits plot indicates that the equal variance condition is not met. The variance shows a megaphone effect.
* The residuals vs fits plot alludes to a non linear relationship between Y and X.

1. Refer to the multiple regression results from part (b). Fill in the values for the coefficients and standard errors in the table below.

|  |  |  |
| --- | --- | --- |
| **Coefficient** | **Coefficient Value** | **Standard Error** |
| ***b*0 *(***constant) | 32.91 | 1.76 |
| ***b*1** (linear term) | 2.902 | 0.135 |
| ***b*2** (quadratic term) | 0.0666 | 0.0129 |

1. Follow these steps to determine appropriate weights for a weighted least squares model:

* Use Minitab’s calculator to create a new variable named absres defined by the expression abs(RES1), where RES1 represents theresiduals that you stored in part (b).
* Fit a simple linear regression model with **y** = absres and **x**-variable FITS1, where FITS1 represents thefitted values that you stored in part (b). Store the Fits (use the *Storage* button in the Regression Dialog).



* Use Minitab’s calculator to create a new variable named weights defined by the expression 1 / FITS2^2 , where FITS2 represents the fitted values that you just stored. These will be possible weights for a weightedregression. The idea for this is that the correct weights are 1 / sd2, where we’ve estimated the standard deviation function in the second step.

Now, refit the multiple linear regression model from part (b) using weighted least squares (use the *Options* button in the Regression Dialog and put the new weights variable in the *Weights* box.). For the weighted regression, fill in the values for the coefficientsand standard errors in the table below.

|  |  |  |
| --- | --- | --- |
| **Coefficient** | **Coefficient Value** | **Standard Error** |
| ***b*0 *(***constant) | 33.08 | 1.35 |
| ***b*1** (linear term) | 2.908 | 0.132 |
| ***b*2** (quadratic term) | 0.0650 | 0.0122 |

1. Briefly compare the results in parts (c) and (d).

*[Note: Typically, with weighted least squares, the coefficient values often stay about the same and the changes in standard error may be slight. Overall regression results tend to be somewhat robust in the presence of non-constant variance.]*

The relevant table for comparison is:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficient** | **OLS Coef Value** | **WLS Coef Value** | **OLS Std Error** | **WLS Std Error** |
| ***b*0 *(***constant) | 32.91 | 33.08 | 1.76 | 1.35 |
| ***b*1** (linear term) | 2.902 | 2.908 | 0.135 | 0.132 |
| ***b*2** (quadratic term) | 0.0666 | 0.0650 | 0.0129 | 0.0122 |

From the above table it is clear that the coefficient values stay about the same and there are only slight changes in standard error.

1. For the weighted regression that you did in part (d), graph the studentized residuals versus fits. (Remember that Minitab calls studentized residuals “standardized” residuals.) Briefly discuss whether the plot looks about as it should (a horizontal random band with constant variance). [In Minitab, use the *Graphs* button of the regression dialog and at the top of the next dialog box select *Standardized residuals*.]

*[Note: With weighted least squares, it is crucial that you use studentized residuals to evaluate the aptness of the model, since these take into account the weights that are used to model the changing variance. The usual residuals don't do this and will maintain the same non-constant variance pattern no matter what weights have been used in the analysis.]*



The above plot of the studentized residuals versus the fits when using the weighted least squares method shows that we have corrected for the megaphone shape since the studentized residuals appear to be more randomly scattered about 0: