**STAT 501 – Homework 11 Solutions Fall 2015**

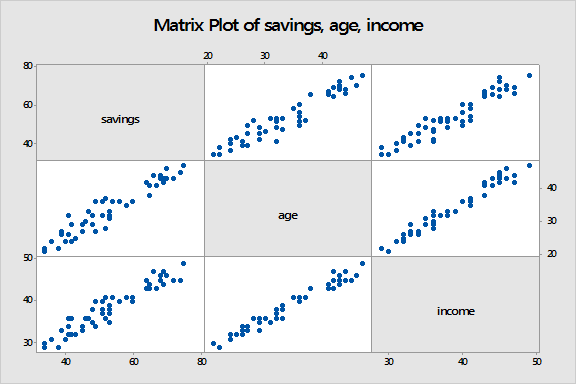
1. **(4+5+5+4+4+5+3+5+4+5+3+3 = 50 points)**

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 Matrix Plot indicates a high correlation between age and income, which could cause data-based multicollinearity problems in an MLR model with both age and income as predictors, including reduced precision of the estimated regression coefficients and increased difficulty interpreting the coefficients.***

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.

Model Summary

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

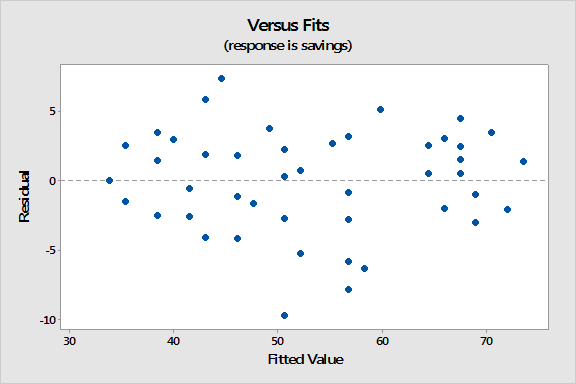
3.56198 90.95% 90.76% 90.34%

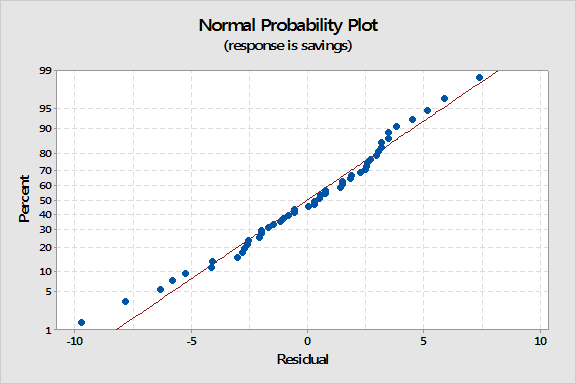
Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 1.92 2.40 0.80 0.426

age 1.5247 0.0694 21.96 0.000 1.00





***The Residuals vs Fits plot exhibits some evidence of nonlinearity since the residuals in the middle tend to be lower than the residuals on the left and the right (i.e., there is a slight “v-shaped” trend). However, the equal variance condition seems fine. The Normal Probability plot supports the normality condition.***

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.

Model Summary

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

3.46443 91.61% 91.26% 90.63%

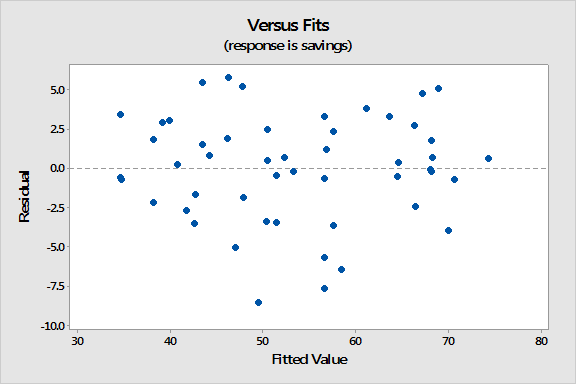
Coefficients

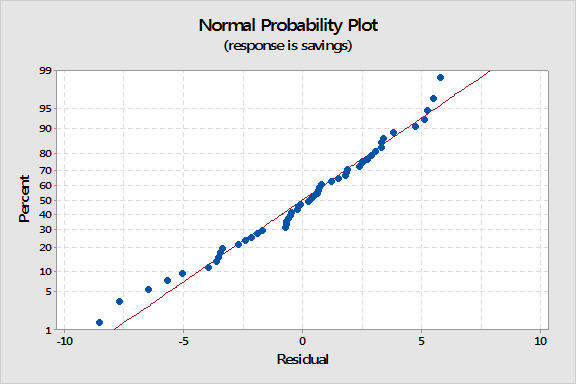
Term Coef SE Coef T-Value P-Value VIF

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



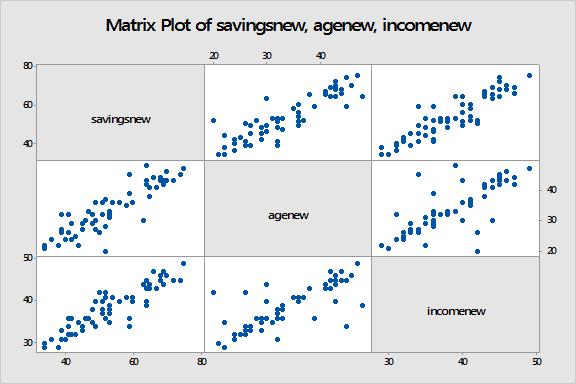


***The Residuals vs Fits plot still exhibits some evidence of v-shaped nonlinearity, while the equal variance condition still seems fine and the Normal Probability plot continues to support the normality condition.***

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?

***The VIFs are 29.29 > 10, which indicates a multicollinarity problem. To mitigate data-based multicollinearity like this we could try removing one or more of the violating predictors from the regression model or try to collect additional data under different experimental or observational conditions.***

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



***The Matrix Plot indicates less correlation between the two predictor variables than the Matrix Plot in part (a), which could mitigate the data-based multicollinarity problem evident in the model in part (c).***

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.

Model Summary

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

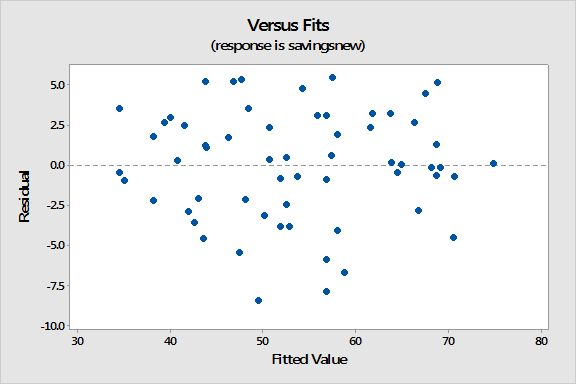
3.50033 90.60% 90.27% 89.64%

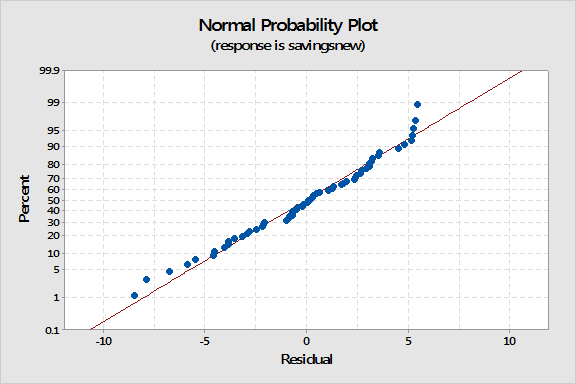
Term Coef SE Coef T-Value P-Value VIF

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 Residuals vs Fits plot still exhibits some evidence of v-shaped nonlinearity, while the equal variance condition still seems fine and the Normal Probability plot continues to support the normality condition.***

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 VIFs are 2.48 < 4, which indicates that the data-based multicollinarity problem 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.

Model Summary

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

3.40931 91.24% 90.77% 90.13%

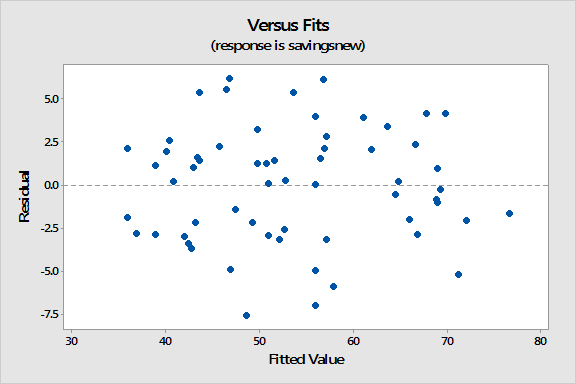
Term Coef SE Coef T-Value P-Value VIF

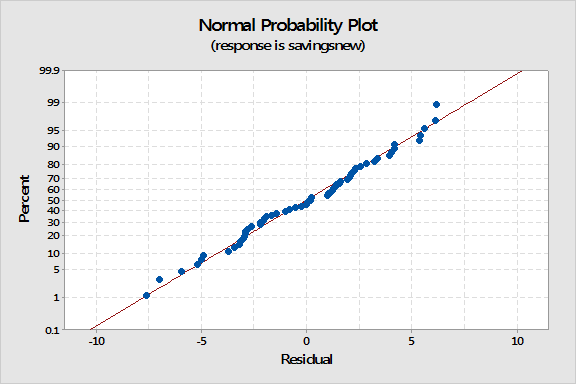
Constant 3.49 9.55 0.36 0.717

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 Residuals vs Fits plot no longer exhibits any evidence of nonlinearity, while the equal variance condition still seems fine and the Normal Probability plot continues to support the normality condition.***

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?

***The VIFs are 94.15 and 92.51 > 10, which indicates a multicollinarity problem. To mitigate structural multicollinearity like this we could try centering the agenew predictor.***

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.

Model Summary

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

3.40931 91.24% 90.77% 90.13%

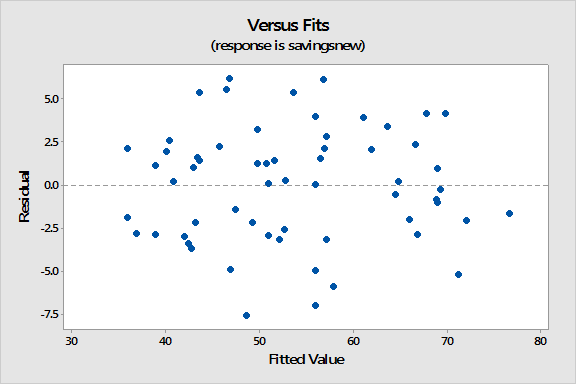
Term Coef SE Coef T-Value P-Value VIF

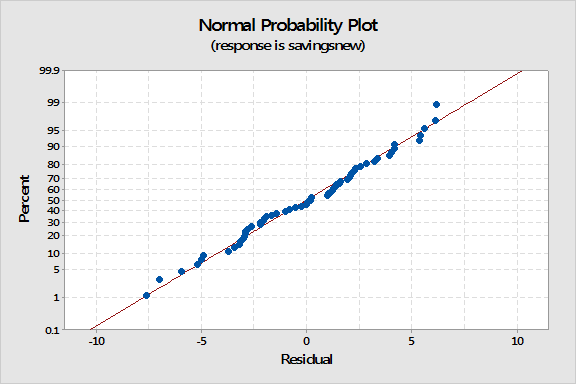
Constant 7.22 5.20 1.39 0.170

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 Residuals vs Fits plot and the Normal Probability plot are identical to the model in part (h), as are the values for S and adjusted R2. This is because the fitted values (and the residuals) are unchanged after centering a predictor.***

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 VIFs are 2.50 and 1.01 < 4, which indicates that the structural multicollinarity problem 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.

Regression Equation

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

Regression Equation

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

***The predictions are the same because the fitted values are unchanged after centering a predictor.***

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:
2. The estimated regression coefficient of any one variable depends on which other predictor variables are included in the model;

***Estimate b1 changes from 1.52 to 0.83 when X2 is added to the model, while estimate b2 changes from 2.05 to 0.95 when X1 is added.***

1. The precision of the estimated regression coefficients decreases as more predictor variables are added to the model.

***The standard error of b1 increases from 0.069 to 0.366 when X2 is added to the model, while the standard error of b2 increases from 0.095 to 0.492 when X1 is added.***

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?
2. The estimated regression coefficient of any one variable no longer depends on which other predictor variables are included in the model;

***Estimate b1 changes from 1.29 to 0.68 when X2 is added to the model, while estimate b2 changes from 1.95 to 1.18 when X1 is added. So, the results in the table do not support this assertion.***

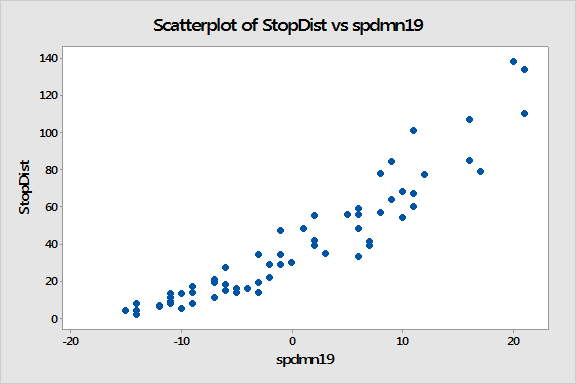
1. The precision of the estimated regression coefficients remains approximately the same as more predictor variables are added to the model.

***The standard error of b1 increases from 0.089 to 0.093 when X2 is added to the model, while the standard error of b2 increases from 0.12 to 0.14 when X1 is added. So, the results in the table do support this assertion.***

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.

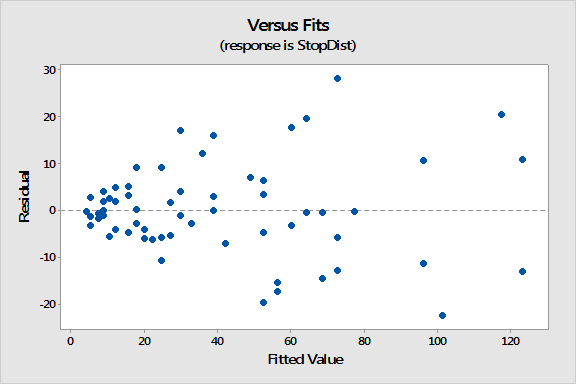
The variables are the car’s stopping distance (StopDist), the car’s speed (minus 19) (spdmn19) and the square of the car’s speed minus 19 (spdsqrd). The value “19” is approximately the average speed in the data set. By centering speed, we lessen the correlation between the linear and squared terms in 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.



***There is a curvilinear pattern with non-constant variance.***

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.



***There is non-constant variance. The variance of the residuals increases as the predicted values increase.***

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(RESI1), where RESI1 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 (predicted values) into a column of your worksheet (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 thevalues 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.

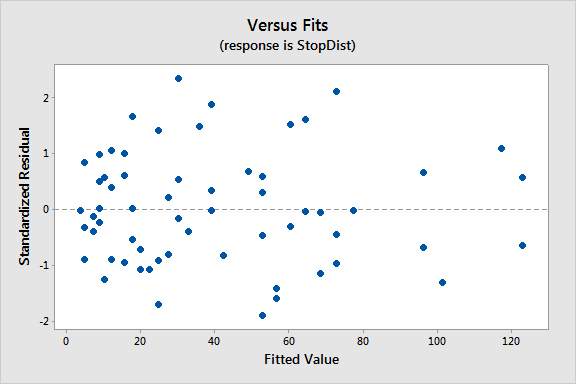
Now, refit the multiple linear regression model from part (b) using weighted least squares (use the *Options* button in the Regression Dialog and put the new weights variable in the *Weights* box.). For the weighted regression, fill in the values for the coefficientsand standard errors in the table below.

|  |  |  |
| --- | --- | --- |
| **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).

***The coefficient estimates are nearly the same in the two regressions. The standard errors are slightly smaller in the weighted regression case.***

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*.]



***The plot of studentized residuals looks more or less like a horizontal random band (as it should). The variance might be slightly smaller for the smallest fitted values but it does not look to be a serious problem*.**