Homework 11

Roly Vicaría
STAT501 Fall 2015

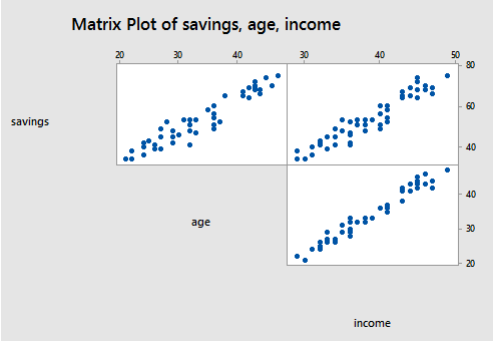
November 23, 2015

0.0.1 Question 1

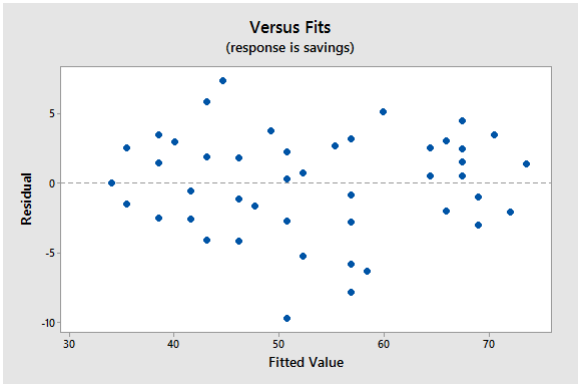
Results of Regression Analyses

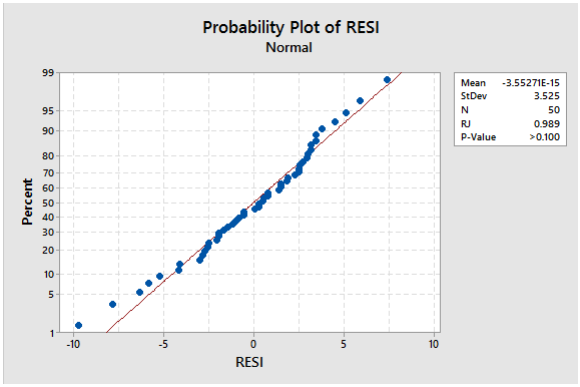
Part	Model	L	N	E	Sample equation	S	R^2 adj
b	savings vs age	N	Y	Y	savings = 1.92 + 1.5247 age	3.56198	99.76%
c	savings vs age, income	N	Y	Y	savings = -11.27 + 0.830 age + 0.952 income	3.46443	91.26%
f	savingsnew vs agenew, incomenew	N	Y	N	savingsnew = -14.53 + 0.6757 agenew + 1.177 incomenew	3.50033	90.27%
h	savingsnew vs agenew, incomenew, $agenew^2$	Y	Y	N	savingsnew = 3.49 - 0.441 agenew + 1.179 incomenew + 0.01636 agenew*agenew	3.40931	90.77%
j	savingsnew vs agenewc, incomenew, $agenewc^2$	Y	Y	N	savingsnew = 7.22 + 0.6624 agenewc + 1.179 incomenew + 0.01636 agenewc*agenewc	3.40931	90.77%

a) The matrix plot shows that savings has a strong correlation with age and a strong correlation with income. It also shows that age has a strong correlation with income. It suggests that an MLR model with both predictors will have larger standard error for the estimated coefficients and marginal (if any) contribution to reducing the error sum of squares from the second predictor.

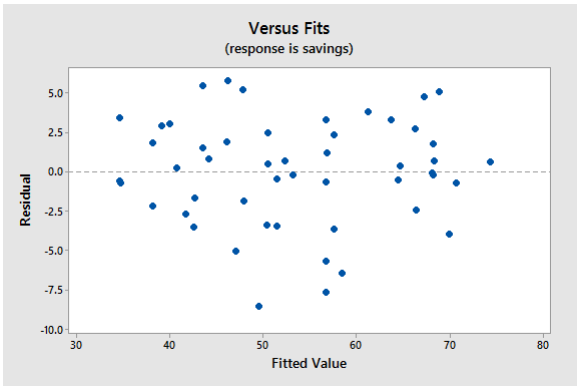


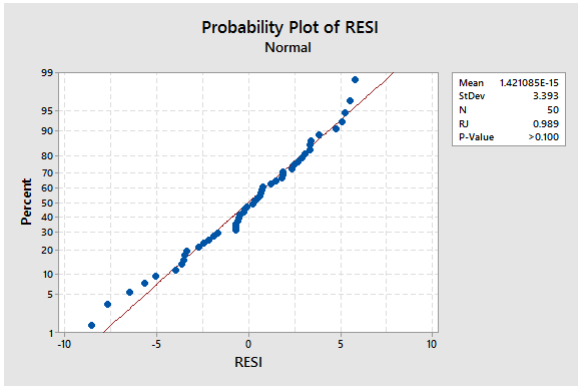
b) The residuals vs fits plot for this model look good. The visual inspection suggests equal variances. The lack of linearity isn't obvious, but there is a subtle up-down-up pattern in the data. The normality test for the residuals confirms that they are normally distributed.





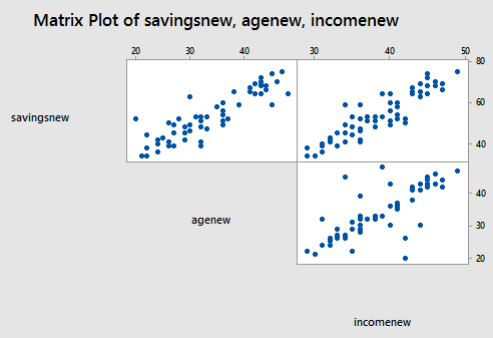
c) The residuals vs fits plot shows that the up-down-up pattern from the previous model is actually accentuated by adding the income predictor to the model. The variances appear equal, however. And the normality test confirms that they are normally distributed.



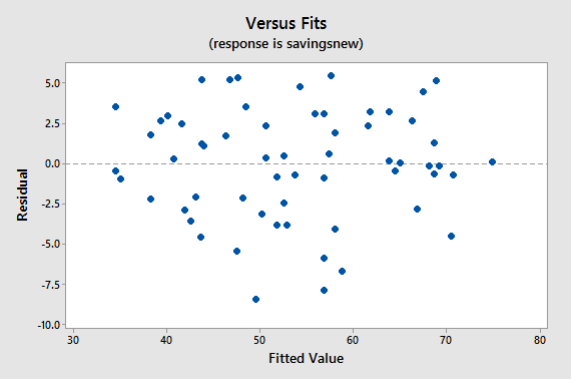


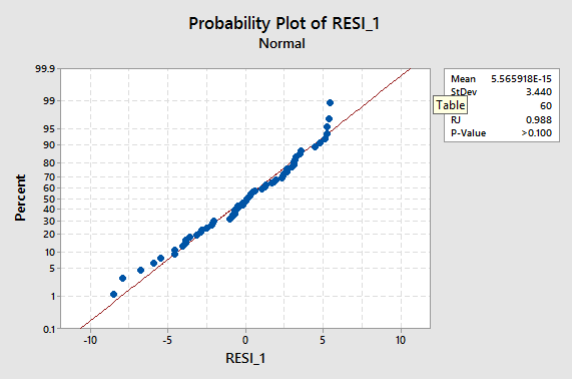
d) From the model in part (c), age and income both had the same VIF value of 29.29. This indicates a high level of correlation between the two predictors. This suggests that the MLR model that includes these two predictors would have less precision of the estimated regression coefficients. The estimated coefficients will be dependent on each other and have different values than from independent SLR models. We can mitigate these problems by removing one of the predictors from the model.

e) Comparing this matrix plot with that from part (a), it looks like agenew and incomenew are highly correlated, but slightly less so than the age and income from part (a).



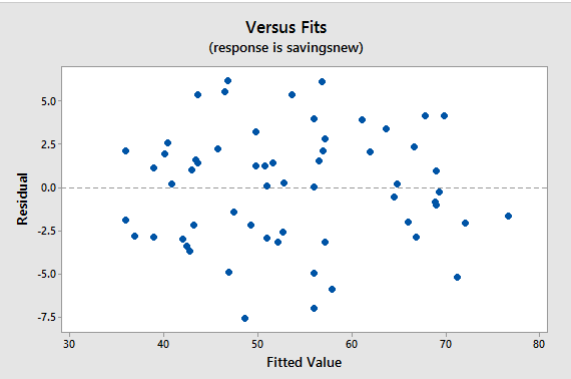
f) The residuals vs fits plot still shows the up-down-up pattern from previous plots. It looks however, that now the variance is no longer equal. It seems less at the ends and greater in the center.

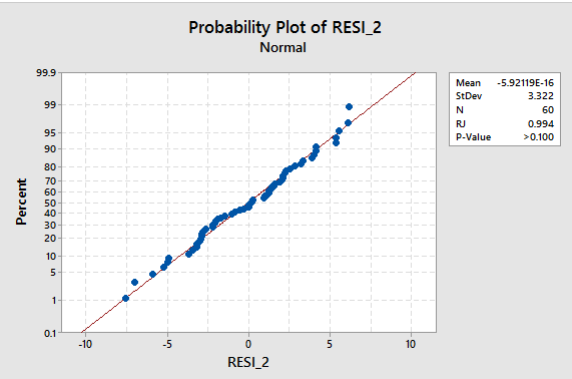




g) From the model in part (f), agenew and incomenew both have the same VIF value of 2.48. This indicates a slight level of correlation between the two predictors, but nothing that would warrant further investigation according to the rule of thumb given in the notes. It appears as though the regression pitfalls have been mitigated.

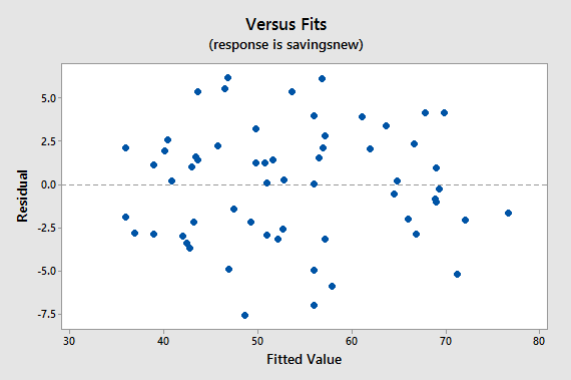
h) I believe that the model generated by adding the $agenew^2$ term has addressed the linearity issues of the previous models, but it still seems to have non-equal variances, larger in the center and smaller at the ends. The probability plot confirms that the errors are normally distributed.

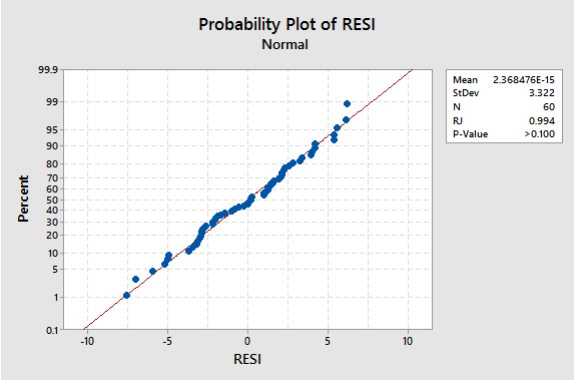




i) The VIFs in the previous part are 94.15 for agenew and 92.51 for $agenew^2$. This indicates a strong correlation between these predictors that could affect the precision of the estimated coefficients, as well as causing the estimated coefficients to be dependent on each other. We can remove one of the predictors to mitigate this.

j) The residuals vs fits plot of this model looks identical to that from part (h). The regression equation has changed slightly since the agenewc column has a different range of values.





k) The VIFs for agenewc and $agenewc^2$ are 2.5 and 1.01, respectively. It would appear that the pitfalls from part (i) have been mitigated.

l) The prediction of savingsnew for an individual with agenew = 30 and incomenew = 45, using the model from part (h), is 58.0455. Using the model from part (j), is 58.0455. Both models arrive at the exact same prediction which makes sense since we saw that multicollinearity doesn't have a huge impact on predictions especially for values close to the means of the predictors.

0.0.2 Question 2

a)

Model	b_1	$se(b_1)$	b_2	$se(b_2)$
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

b) i. We can see that in the first 3 models, the estimated coefficients vary significantly from the SLR models to the MLR model. The age coefficient went from 1.5247 in the SLR to 0.830 in the MLR. Same thing from income: from 2.0493 to 0.952. Both reduced by about half.

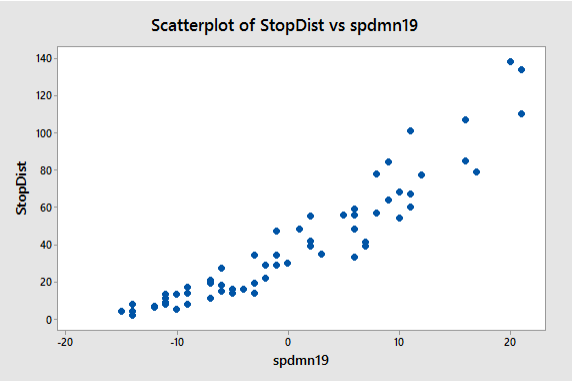
ii. We see that the standard error for both predictors increases by a little over 500% from the SLRs to the MLR.

c) i. The estimated coefficient of agenew shows that it's still dependent on incomenew since it drops by about 50% from one model to the other. The incomenew coefficient shows better invariance across models, but still drops considerably.

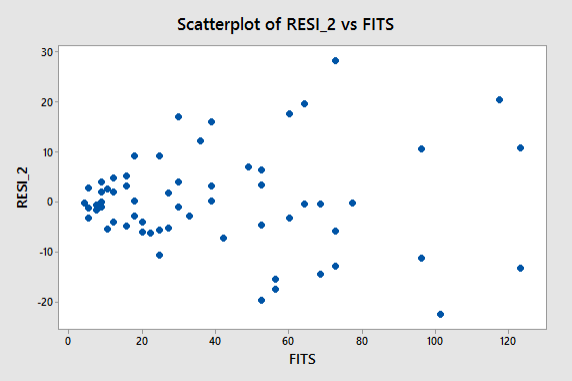
ii. The standard errors of the estimated coefficients do remain approximately the same from the SLRs to the MLR.

0.0.3 Question 3

a) The plot of StopDist vs spdmn19 shows a strong positive correlation between the variables. The relationship appears to be non-linear since it's curving upwards. The variance in StopDist also seems to increase as spdmn19 increases.



b) $StopDist = 32.91 + \ 2.902\ spdmn19 + \ 0.0666\ spdsqrd$



The residual plot indicates that the variance is not constant. It increases for larger fitted values. Also, there seems to be a larger concentration of data points for smaller fitted values and grows more sparse as we move to the right.

c)

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

d)

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

e) The results are almost the same for both models. The coefficient values and standard errors varied only slightly. The largest change was the standard error for the constant term which changed from 1.76 to 1.35.

f) The plot of studentized residuals vs fits looks much better than the previous plot in terms of correcting the error variance. This plot appears much closer to a horizontal band with constant variance. There does still appear to be a denser concentration of data points for smaller fitted values, and gets sparser for larger fitted values.