Laboratory 5
Regression Assumptions & Multicollinearity
Homework 5 Answer Key

Assumption Checking

Random sampling to get X_1 (n=100) and E (n=100) from a normal distribution population, N (0,1). Let $X_2 = X_1 * X_1$; $X_3 = X_1 * X_1 * X_1$; $Y1= X_1 + X_2 + X_3 + E$; $Y2= X_1 + X_2 + X_3 + E^2$. Please run the following SAS program to fit the regression models Y1= X_1 X_2 X_3 and Y2= X_1 X_2 X_3. Check the assumption of multiple regression by examining and plotting the residuals. Report and interpret the results (both testing and graphs) for model Y2= X_1 X_2 X_3

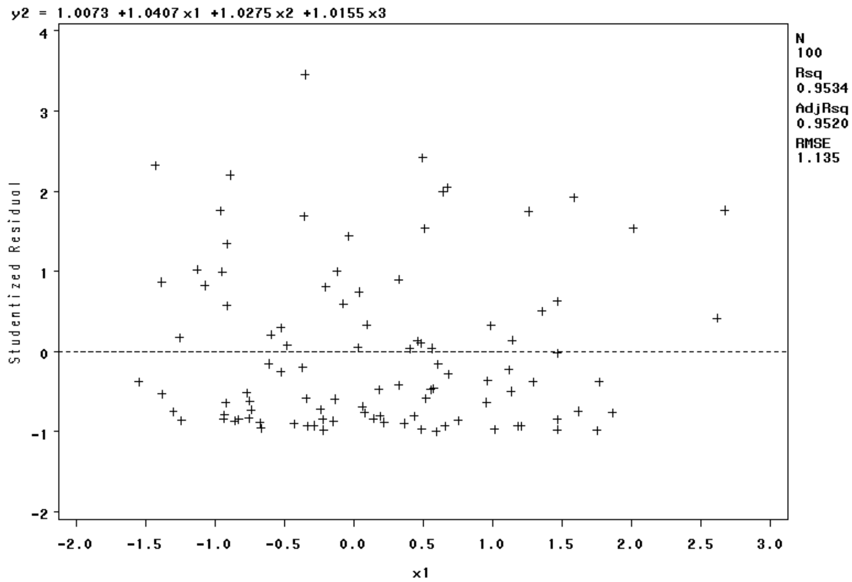
Each random sample will give different results. Please report the results you got. The following is the result of the analysis for model Y2= X_1 X_2 X_3

1. Analysis of Residuals

```
                    The UNIVARIATE Procedure
                       Variable:  residual

                              Moments

N                         100    Sum Weights                100
Mean                        0    Sum Observations             0
Std Deviation      1.11763328    Variance            1.24910415
Skewness           1.16501396    Kurtosis            0.66703599
Uncorrected SS     123.661311    Corrected SS        123.661311
Coeff Variation             .    Std Error Mean      0.11176333

                        Tests for Normality

   Test                  --Statistic---    -----p Value------

   Shapiro-Wilk          W     0.854342    Pr < W     <0.0001
   Kolmogorov-Smirnov    D     0.170954    Pr > D     <0.0100
   Cramer-von Mises      W-Sq   0.83288    Pr > W-Sq  <0.0050
   Anderson-Darling      A-Sq  4.933847    Pr > A-Sq  <0.0050
```

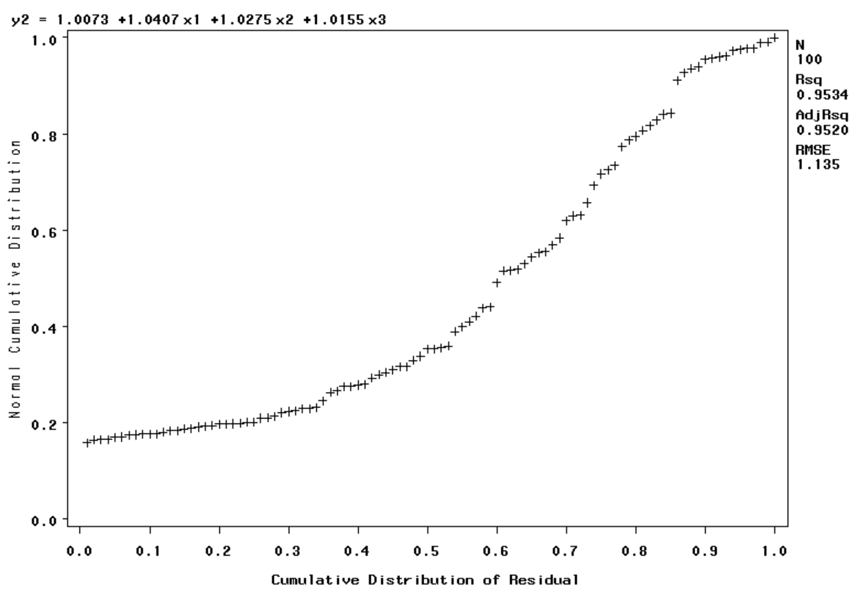
The residual is not normally distributed (P<0.01 for Normality of Residuals)

2. Scatterplot of Residuals vs. X_1



The residuals are not approximately evenly distributed half above the 0-line and half below the 0-line. The residuals are unevenly distributed with more falling below the 0-line than above it. The mean of the residual is not 0 when x1 from -2.0 to 3. This graph suggests that the relationship between y and x1 is not linear.

3. Normal Probability Plot of Residuals



It is not a straight line, suggesting that the residuals are not normally distributed.

Collinearity

Random sampling to get X_1 (n=30) from a normal distribution population, N $(8,2^2)$, and e (n=30) from another normal distribution population, N (0,1). Let $X_2 = 2X_1 + e$; $Y = 3X_1 + 1 + e$. Please run the following SAS program to fit the regression models $Y= X_1$, $Y= X_2$, and $Y= X_1 \ X_2$. Why is X_2 significant in model $Y= X_2$, but not in model $Y= X_1 \ X_2$? Please run this program 10 times. Make a table to record the parameter estimate, SE, and P value of the final model for each `model y=x1 x2/selection=forward`. What did you learn from this computer experiment by comparing the β and SE?

Each random sample will give different results. Please report the results you got. There are the results of one set of analysis from
`model y=x1 x2/selection=forward`

Model	X	β	SE	P
Model 1	X1	2.97601	0.15006	<0.0001
Model 2	X1	3.37690	0.31533	<0.0001
	X2	-0.14771	0.15416	0.3465
Model 3	X1	4.05555	0.37985	<0.0001
	X2	-0.48721	0.18900	<0.0157
Model 4	X1	3.1694	0.08872	<0.0001
Model 5	X1	3.43573	0.30387	<0.0001
	X2	-0.25692	0.15000	0.0982
Model 6	X1	2.50198	0.44491	<0.0001
	X2	0.20357	0.23428	0.3926
Model 7	X1	2.85146	0.07560	<0.0001
Model 8	X1	3.47863	0.37415	<0.0001
	X2	-0.28419	0.17875	0.1235
Model 9	X1	3.34755	0.31164	<0.0001
	X2	-0.18800	0.13318	0.1695
Model 10	X1	3.31532	0.08261	<0.0001

1. When analysis separately, X1 and X2 are significantly associated with Y (using models Y=X1 and Y=X2). But when both X1 and X2 are present in the model Y=X1 X2, X2 is not significant. This is because X1 and X2 are highly correlated themselves. It doesn't mean there is not an association between Y and X2.

2. The regression coefficients are very unreliable. The β for X1 ranged from 2.5 to 4.1. The β for X2 ranged from negative (-0.49) to positive (0.20). The range of SE for X1 is from 0.0756 to 0.4449. The difference of SE is about 6 times.