Regression Analysis Projet

VEE Fall 2014

by ZHAO AI SUN

Introduction

The purpose of this project is to illustrate how we can analyze the linear regression model and the residuals in the model, and how do we test homoscedasticity and heteroscedasticity by three different tests in SAS.

First, I will explain how to quickly check the linear regression model and run normality tests (the Shapiro-Wilk Test and the Kolmogorov-Smirnov test) by SAS output. Part of the project is to discuss how to analyze the residuals by verifying graphics in SAS. Finally, I try to detect homoscedasticity and heteroscedasticity by Spearman test, White test and Breusch- Pagan test.

Consider the data from the website: http://lx2.saas.hku.hk/staff/kaing/tdg/data8/d103.txt
where:

States: Different states of the USA.
MA: number of married people / 10 000 inhabitants
D: Number of divorced people/ 10 000
DR: Number of doctors / 100 000 inhabitants
DN: Number of Dentists / 100 000 inhabitants
HS: Number of officers / 1,000 people
CR: Number of crimes / 100 000 inhabitants
M: Number of people killed / 100 000
PI: Number of prisons / 100,000 residents
RP: % vote for a Republican candidate for the presidential election
VT: % of voting for a presidential candidate among the population of voting age
PH: Percentage (in 1979) people with a phone
INC: Income (dollars) per capita in 1972
PL: Number of persons / 1000 people living below the poverty line

```
*************SAS Code   *************;
data donnee;
input State $ MA   D   DR   DN   HS   CR   M   PI   RP   VT   PH   INC  PL;
cards;
```

State	MA	D	DR	DN	HS	CR	M	PI	RP	VT	PH	INC	PL
ME	109	56	146	45	678	4368	28	61	456	648	54	4430	120
NH	102	59	159	53	703	4680	25	35	577	578	58	5105	79
VT	105	46	211	58	697	4988	22	67	444	583	52	4372	135
MA	78	30	258	71	723	6079	41	56	419	593	58	5660	71
RI	79	39	206	56	617	5933	44	65	372	590	57	5281	87
CT	82	45	242	73	703	5882	47	68	482	612	64	6552	67
NY	81	37	261	74	662	6912	127	123	467	480	54	5736	94
NJ	75	32	184	66	664	6401	69	76	520	551	66	6107	81
PA	80	34	183	55	648	3736	68	68	496	520	62	5273	97
OH	93	55	157	49	677	5431	81	125	515	554	56	5289	94
IN	110	77	126	43	670	4930	89	114	560	577	57	4995	81
IL	97	46	182	54	661	5275	106	94	496	578	66	5881	105
MI	97	48	154	53	686	6676	102	163	490	598	60	5562	91
WI	84	36	151	58	703	4799	29	85	479	677	55	5225	77
MN	91	37	185	62	724	4799	26	49	425	704	57	5436	83
IA	96	39	122	50	723	4747	22	86	513	629	59	5232	79
MO	109	57	158	48	641	5433	111	112	512	589	58	5021	120
ND	92	32	126	47	676	2964	12	28	642	651	63	4891	106
SD	130	39	102	43	689	3243	7	88	605	674	56	4362	131
NE	89	40	145	61	743	4305	44	89	655	568	61	5234	96
KS	105	54	150	46	731	5379	69	106	579	570	61	5580	80
DE	75	53	160	46	695	6777	69	183	472	549	64	5779	82
MD	111	41	257	59	693	6630	95	183	442	502	62	5846	77
VA	113	45	170	49	642	4620	86	161	530	480	53	5250	105
WV	94	53	133	39	533	2552	71	64	452	528	44	4360	151
NC	80	49	150	38	553	4640	106	244	493	439	53	4371	147
SC	182	47	134	36	571	5439	114	238	494	407	50	4061	172
GA	134	65	144	42	587	5604	138	219	409	417	55	4512	180
FL	117	79	188	50	648	8402	145	208	555	496	59	5028	144
KY	96	45	134	42	533	3434	88	99	491	500	48	4255	177
TN	135	68	158	48	549	4498	108	153	487	489	53	4315	158
AL	129	70	124	35	555	4934	132	149	488	490	50	4186	164
MS	112	56	106	32	523	3417	145	132	494	521	47	3677	261
AR	119	93	119	33	562	3811	92	128	481	516	47	4062	185
LA	103	38	149	40	583	5454	157	211	512	537	52	4727	193
OK	154	79	128	42	656	5053	100	151	605	528	57	5095	138
TX	129	69	152	42	645	6143	169	210	553	456	55	5336	152
MT	104	65	127	57	725	5024	40	94	568	652	56	4769	115
ID	148	71	108	55	715	4782	31	87	665	685	54	4502	103
WY	144	78	107	49	753	4986	62	113	626	542	56	6089	87
CO	118	60	199	61	781	7333	69	96	551	568	57	5603	91
NM	131	80	147	41	657	5979	131	106	549	514	46	4384	193
AZ	121	82	187	49	725	8171	103	160	606	452	53	4915	138
UT	122	56	164	64	802	5881	38	64	728	655	53	4274	85
NV	1474	168	138	49	757	8854	200	230	625	413	63	5999	88
WA	120	69	178	68	763	6915	51	106	497	580	56	5762	85
OR	87	70	177	69	755	6687	51	120	483	616	55	5208	89
CA	88	61	226	63	740	7833	145	98	527	495	61	6114	104
AK	123	86	118	56	796	6210	97	143	543	583	36	7141	67
HI	128	55	203	68	730	7482	87	65	425	436	47	5645	79

```
;
   run;

  data donne1;
  set donnee;
  if MA="." then delete;
  run;
  Data analyse;
  set donne1(drop=DN   HS   CR  PI   RP   PH);
  run;
  proc print data=analyse;
  run;
******* Part 1 A linear model without intercept;
******  SAS OUTPUT 1******
proc reg;
model M=MA D DR VT INC PL /noint;
output out=Ia residual=res student=stud;
run;
***** Part 2 Normality tests;
****** SAS OUTPUT 2******
proc univariate data=Ia normal plot;
var res stud;
run;
***** Part3 Graphics - normality verification;
************  SAS OUTPUT 3 **********
proc reg data=analyse;
model M= MA D DR VT INC PL /noint;
plot r.*p.;
plot r.*npp.;
plot r.*M ;
run;
********* Part 4 Detecting Homoscedasticity *******
***** Spearman: Low correlation => Homoscedasticity;
******  SAS OUTPUT 4a ******
proc corr data=analyse spearman ;
var M MA D DR VT INC PL;
run ;
***** Heteroscedasticity test: The White Test & Breusch Pagan Test;
******  SAS OUTPUT 4b ******
PROC REG DATA=analyse ;
model M=MA D DR VT INC PL /noint ACOV;
output out=Ib r=residus;
run;
proc model data=analyse;
parms b1 b2 b3 b4 b5 b6;
M=b1*MA+ b2*D+ b3*DR+ b4*VT+ b5*INC+b6*PL;
fit M / white pagan=(MA D DR VT INC PL);
run;
quit;
********************************************
```

Part 1 A linear model without intercept

```
************* SAS Code **********
proc reg;
model M=MA D DR VT INC PL /noint;
output out=Ia residual=res student=stud;
run;
*********************************
```

SAS OUTPUT 1

```
                                   The SAS System        14:41 Monday, December 6, 2014   2

                                   The REG Procedure
                                     Model: MODEL1
                                 Dependent Variable: M

                     NOTE: No intercept in model. R-Square is redefined.

                                  Analysis of Variance
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	414092	69015	**150.97**	**<.0001**
Error	44	20115	457.16002		
Uncorrected Total	50	434207			

Root MSE	21.38130	**R-Square**	**0.9537**
Dependent Mean	81.78000	Adj R-Sq	0.9474
Coeff Var	26.14490		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
MA	1	**0.02906**	0.02442	1.19	0.2406
D	1	**0.31291**	0.22206	1.41	0.1658
DR	1	**0.06281**	0.09189	**0.68**	**0.4979**
VT	1	**-0.28638**	0.03356	-8.53	<.0001
INC	1	**0.02596**	0.00517	5.02	<.0001
PL	1	**0.64371**	0.07142	9.01	<.0001

We have a linear model without intercept to estimate y = M by using six coefficients:

$$y = \beta_1\ x_1 + \beta_2\ x_2 + \beta_3\ x_3 + \beta_4\ x_4 + \beta_5\ x_5 + \beta_6\ x_6 + \varepsilon$$

The values of the estimated parameters give the equation for the fitted model:

$$M = 0.02906\ \text{MA} + 0.31291\ \text{D} + 0.06281\ \text{DR} - 0.28638\ \text{VT} + 0.02596\ \text{INC} + 0.64371\ \text{PL}$$

This is a multiple linear regression model.

- The coefficient of determination:

  R^2 = R-Square = 0.9537. R^2 provides the proportion of variability in Y explained by the regression as 95.37%. R^2 close to 1 will be associated with a good fit.

- Test the hypothesis:
  F-value = 150.97 is used to test the null hypothesis:
  $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ against $H_1:$ at least one coefficient $\beta_i \neq 0,\ i = 1, \cdots, 6$
  From the output of SAS, associated P-value <0.0001, this causes a rejection of this hypothesis at the 5% level. It indicates that at least one coefficient beta is not zero.

- The t-statistic:
  The value t is used to test the hypothesis on individual parameters.
  For example, a statistic t = 0.68 is to test the null hypothes β_3=0, at the 5% level. The test shows if there is variation in M due to DR. The P value for the hypothesis (β_3=0) is 0.4979. The null hypothesis is accepted. It means that there is no great effect due to the variable DR. DR can be removed. Similarly, we can also remove MA and D.
  The P value for the hypothesis (β_4=0) is <0.0001. This is significant. We can reject the null hypothesis. It means that there is effect due to the variable VT. You can't remove VT.
  The results of the t-statistic show that the variables INC, PL have the same situation as the variable VT.
  Conclusion: if you want to remove an explanatory variable, DR is always the first choice because it has very big P-value.

Part 2 Tests for Normality

```
************* SAS Code*********
proc univariate data=Ia normal plot;
var res stud;
run;
*********************************
```

SAS OUTPUT 2

```
                     The SAS System        14:41 Monday, December 6, 2014   3
                      The UNIVARIATE Procedure
                     Variable:  res  (Residual)

                        Tests for Normality

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

Shapiro-Wilk          W      0.960049    Pr < W      0.0893
Kolmogorov-Smirnov    D      0.110306    Pr > D      0.1305
Cramer-von Mises      W-Sq   0.136736    Pr > W-Sq   0.0362
Anderson-Darling      A-Sq   0.783313    Pr > A-Sq   0.0410


                      The UNIVARIATE Procedure
                     Variable:  res  (Residual)

                       Normal Probability Plot
```

```
 42.5+                                          * ++*
     |                                       ** ++
     |                                       +++
     |                                   ***+
     |                                 ***++
     |                                 ++
     |                              +**
     |                            +***
     |                          ++**
     |                       ++***
     |                   *******
     |                  +*+
     |               ****
     |         ** ***+
     |           ++
     |       *+++
     |      ++
-42.5+   *++
      +----+----+----+----+----+----+----+----+----+----+
          -2        -1         0        +1        +2
```

The UNIVARIATE Procedure

Variable: stud (Studentized Residual)

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	**W**	**0.964624**	**Pr < W**	**0.1388**
Kolmogorov-Smirnov	**D**	**0.114773**	**Pr > D**	**0.0966**
Cramer-von Mises	W-Sq	0.121324	Pr > W-Sq	0.0579
Anderson-Darling	A-Sq	0.680639	Pr > A-Sq	0.0750

The UNIVARIATE Procedure

Variable: stud (Studentized Residual)

Normal Probability Plot

```
  2.1+                                          * ++*
     |                                       ** ++
     |                                        ++
     |                                     *++
     |                                  ***+
     |                                 **++
     |                                *++
  0.7+                               +*
     |                             ++*
     |                           ++**
     |                         ++ *
     |                       ++***
     |                     +****
     |                   ***
 -0.7+                  **
     |                **
     |            ****
     |         ** ++
     |          ++
     |       *++
     |      ++
 -2.1+   *++
      +----+----+----+----+----+----+----+----+----+----+
          -2        -1         0        +1        +2
```

❖ Normality tests:

➢ The Shapiro-Wilk Test: This test also tests the normality of our residuals by comparing them with expected values. The test returns a W statistic, which informs us the normality of the data.
The Shapiro-Wilk Test accepts the normality assumption: The statistic W = 0.960049 for residuals and W = 0.964624 for studentized residuals. For both of them, W is very close to 1, which indicates that it is approximately a normal distribution. P-Value = 0.0893> 0.05 for residuals, P-Value = 0.1388 > 0.05 for studentized residuals.For both of them, P-Value < W, so we accept the null hypothesis of normality.

➢ The Kolmogorov-Smirnov test. This test is often used when the sample size is large.
The Kolmogorov-Smirnov also accepts residuals normality assumption. The statistic D = 0.110306 and P Value = 0.1305> 0.05, so we accept the null hypothesis.

❖ The graphic "Normal Probability Plot" is a line for both residuals and studentized residuals, which shows that the errors are normally distributed.

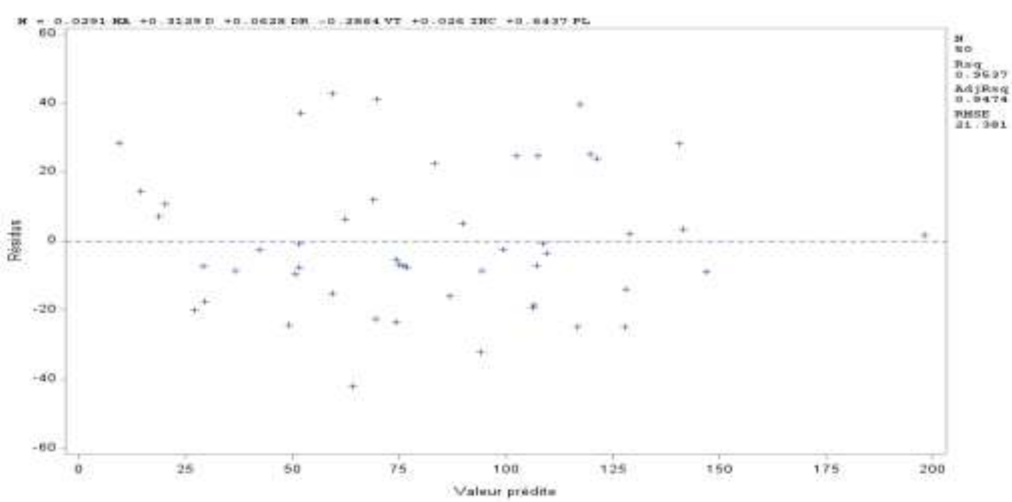
Conclusion: By "Normal Probability Plot" and our various tests, we conclude that the residuals are normally distributed. Indeed, the "Normal Probability Plot" has the appearance of a line and tests confirm the hypothesis.

Part 3 Residual Analysis in Regression by graphics

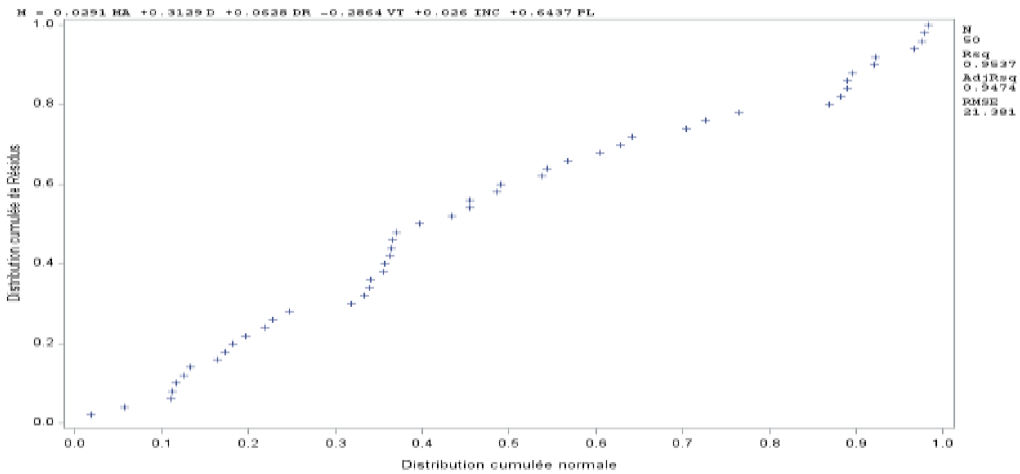
```
*************SAS Code*************
proc reg data=analyse;
model M= MA D DR VT INC PL /noint;
plot r.*p.;
plot r.*npp.;
plot r.*M ;
run;
*********************************
```

SAS OUTPUT 3

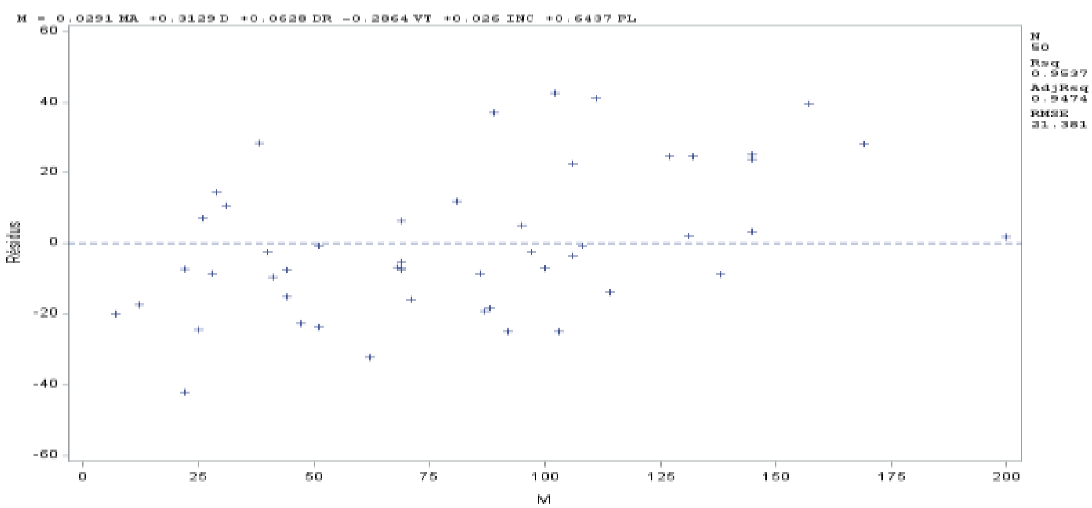
Graphic 1



Graphic 2



Graphic 3



Graphic 1

We have the following equation: Residuals = M - Predictive Value. The graph (residuals* predicted) shows a scatter relatively evenly distributed randomly between -60 and 60 (the interval is not small) and it is also symmetrical around the x-axis. Indeed, as the hypothesis associated with the model are correct, residuals and predicted values are not correlated. Therefore, the trace of the points should not have any particular structure.

Graphic 2

We can verify the distribution of residuals by QQ-Plot - "Normal Probability Plot" . Using *plot r. * Nqq.; run;* in SAS, we can visualize the distribution of residuals in this model. We see that the points lie on or very close to a straight line. Therefore, it is compatible with the normal distribution.

Graphic 3
If we make a graphics of residuals depending on the response Y (M here) to see the quality of the regression, we can also find a linear regression. The residuals randomly distribute between -60 to 60.

Conclusion: we have the same conclusion for these 3 graphics.

Part 4 Detecting Heteroscedasticity

The ordinary least squares (OLS) makes the assumption that the error ε in the regression model had a constant variance σ^2 for all x ,which means variances $var(\varepsilon_i) = \sigma^2$ do not depend on the x-value. This is one of the Gauss-Markov condition, which states that $var(\varepsilon_i) = var(y_i)$ is a constant σ^2. Consequently, each probability distribution for y (response variable) has the same standard deviation regardless of the x-value (predictor). This assumption is homoscedasticity.
If the error terms do not have constant variance, they are said to be heteroscedastic.

Very frequently, we can determine if heteroscedasticity is likely to be present and also determine what corrective measures might be taken.
At first, we will see if their variance (or quantities proportional to them) can be guessed. There are several statistical tests in SAS can help us to test the equality of variance, such as Spearman test, White test, and Breusch- Pagan test, etc.
To check to see if heteroscedasticity is present, another way is through the residuals plots to see whether the variance of error is constant. (See part 3 Residual Analysis in Regression by graphics).

If in fact that variance of error is not constant then it is better to modify model by using weighted least squares (WLS) method to get the estimators rather than using the ordinary least squares (OLS) method. Transformation of variables can also be used to stabilize variances. If the response variable represents a count, then a Poisson distribution can be considered for modelling the response. In a Poisson regression, the unequal variance is expected due to the nature of the count data.

```
**************SAS Code **************
PROC REG DATA=analyse ;
model M=MA D DR VT INC PL /noint ACOV;
output out=Ib r=residus;
run;
****************************************
```

SAS OUTPUT 4

```
The SAS System       21:51 Tuesday, December 7, 2014  12
The REG Procedure
Model: MODEL1
Dependent Variable: M

Consistent Covariance of Estimates
```

Variable	MA	D	DR	VT	INC	PL
MA	0.0003573061	-0.004052164	-0.000173725	0.00010143	0.0000119602	0.0007607911
D	-0.004052164	0.0535420807	0.0076283856	0.0009424341	-0.000578784	-0.011005308

DR	-0.000173725	0.0076283856	0.0090140361	0.0008693459	-0.000413447	-0.001811813
VT	0.00010143	0.0009424341	0.0008693459	0.0009808474	-0.000125022	-0.000867191
INC	0.0000119602	-0.000578784	-0.000413447	-0.000125022	0.0000291511	0.0001498898
PL	0.0007607911	-0.011005308	-0.001811813	-0.000867191	0.0001498898	0.004743984

Testing for Heteroscedasticity by SAS

The regression model is specified as yi = xiβ+εi, where the εi's are identically and independently distributed: E(ε) = 0 and E(ε'ε) =σ^2I. If the εi's are not independent or their variances are not constant, the parameter estimates are unbiased, but the estimate of the covariance matrix is inconsistent. In the case of heteroscedasticity, the ACOV option provides a consistent estimate of the covariance matrix. If the regression data are from a simple random sample, the ACOV option produces the covariance matrix. This matrix is $(X'X)^{-1} (X'\text{diag}(\varepsilon_i^2)\, X)\, (X'X)^{-1}$ where εi= yi - xib

ACOV in the SAS model statement displays the estimated asymptotic covariance matrix of the estimates under the hypothesis of heteroscedasticity.
With the ACOV option, the point estimates of the coefficients are exactly the same as in ordinary OLS, but we will calculate the standard errors based on the asymptotic covariance matrix.
The standard error obtained from the asymptotic covariance matrix is considered to be more robust and can deal with a collection of minor concerns about failure to meet assumptions, such as minor problems about normality, heteroscedasticity, or some observations that exhibit large residuals, leverage or influence. For such minor problems, the standard error based on ACOV may effectively deal with these concerns.

Part 4a Spearman test

The Spearman rank-order correlation coefficient (Spearman's correlation, for short) is a nonparametric measure of the strength and direction of association that exists between two variables measured on at least an ordinal scale. This test determines if two variables are related and specifies the degree of relationship. It is denoted by the symbol r_s (or the Greek letter ρ, pronounced rho).

For a sample of size n, the n raw scores Xi,Yi are converted to ranks xi,yi , and ρ is computed from:

$$\rho = r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

where: di=xi-yi, the difference of the ranks for each pair of variables

n = number of pairs of variables.

Procedure in the use of the Spearman test for homoscedasticity testing:

- The hypotheses in Spearman test are H0 : ρ =0, Homoscedasticity against H1 : Heteroscedasticity.
- Fit the regression to the data on X and Y variables, then obtain the residuals εi.
- Use the absolute values of εi. Enter the ranks for the absolute values of εi and the ranks for the Xi variable, then compute the Spearman correlation coefficient. If the regression model involves more than one X variable, r_s can be computed between εi and each of X variables separately

For the sample greater than 8 or (some people will say 10), the significance can be tested by using t test. How to calculate probability values?

- When n is 10 or more, r_s is approximated by a t distribution with n-2 degrees of freedom. When the null hypothesis is H0 : ρs = 0 the standardized t statistic can be written $t = r_s * \sqrt{\frac{n-2}{1-r_s^2}}$

  If t value is greater than $t_{\alpha/2,n-2}$ value, then heteroscedasticity exists or there's unequality of variance.
- When n is greater than 30, the significance of r_s can be tested by using standard normal Z with the following formula: $z = \frac{r_s - 0}{1/\sqrt{n-1}} = r_s * \sqrt{n-1}$

  $If\ \ z > z_{\alpha/2}\ \ and\ \ z < z_{-\alpha/2}$ then heteroscedasticity exists.

Note that this method should not be used in cases where the data set is truncated; that is, when the Spearman correlation coefficient is desired for the top X records (whether by pre-change rank or post-change rank, or both), the user should use the Pearson correlation coefficient formula.

```
******  SAS Code 4a ************
***** Spearman Test;
proc corr data=analyse spearman ;
var M MA D DR VT INC PL;
run ;
******************************
```

SAS OUTPUT 4a

```
The SAS System       17:42 Tuesday, December 7, 2014   2
The CORR Procedure

7  Variables:    M         MA        D         DR        VT        INC       PL
```

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
M	50	81.78000	45.13214	83.50000	7.00000	200.00000
MA	50	135.50000	194.50012	107.00000	75.00000	1474
D	50	57.78000	22.62768	55.00000	30.00000	168.00000
DR	50	161.86000	40.15227	153.00000	102.00000	261.00000
VT	50	551.40000	75.14096	552.50000	407.00000	704.00000
INC	50	5130	719.77324	5217	3677	7141
PL	50	115.68000	42.19268	100.00000	67.00000	261.00000

Spearman Correlation Coefficients, N = 50
Prob > |r| under H0: Rho=0

	M	MA	D	DR	VT	INC	PL
M	1.00000	0.32536	0.40782	-0.00139	-0.77546	-0.02181	0.49146
		0.0211	0.0033	0.9923	<.0001	0.8805	0.0003
MA	0.32536	1.00000	0.67392	-0.44367	-0.30563	-0.29935	0.35531
	0.0211		<.0001	0.0013	0.0309	0.0347	0.0113
D	0.40782	0.67392	1.00000	-0.32055	-0.27943	-0.12900	0.25001
	0.0033	<.0001		0.0232	0.0494	0.3720	0.0799
DR	-0.00139	-0.44367	-0.32055	1.00000	-0.08376	0.48274	-0.35598

	0.9923	0.0013	0.0232		0.5630	0.0004	0.0112
VT	-0.77546	-0.30563	-0.27943	-0.08376	1.00000	0.06344	-0.44180
	<.0001	0.0309	0.0494	0.5630		0.6616	0.0013
INC	-0.02181	-0.29935	-0.12900	0.48274	0.06344	1.00000	-0.72357
	0.8805	0.0347	0.3720	0.0004	0.6616		<.0001
PL	0.49146	0.35531	0.25001	-0.35598	-0.44180	-0.72357	1.00000
	0.0003	0.0113	0.0799	0.0112	0.0013	<.0001	

- Definition in SAS: Spearman rank-order correlation is a nonparametric measure of association based on the ranks of the data values. The formula in SAS is:

$$\theta = \frac{\sum_i ((R_i - \bar{R})(S_i - \bar{S}))}{\sqrt{\sum_i (R_i - \bar{R})^2 \sum (S_i - \bar{S})^2}}$$

where R_i is the rank of x_i, S_i is the rank of y_i, $\bar{R}$ is the mean of the R_i values, and $\bar{S}$ is the mean of the S_i values.

- For example: For M, the first line is 'Spearman correlation coefficient';
  the second line is 'P-value'.

- The result of SAS OUTPUT shows that Spearman test shows that we can accept the null hypothesis H0: Homoscedasticity.

Part 4b White test and Breusch- Pagan test

The result of the OLS regression (Ordinary Least Square) is presented in the SAS output.
To detect homoscedastic, we used the White test and the Breusch- Pagan test.

```
******************* SAS Code 4b ****************
PROC REG DATA=analyse ;
model M=MA D DR VT INC PL /noint ACOV;
output out=Ib r=residus;
run;
proc model data=analyse;
parms b1 b2 b3 b4 b5 b6;
M=b1*MA+ b2*D+ b3*DR+ b4*VT+ b5*INC+b6*PL;
fit M / white pagan=(MA D DR VT INC PL);
run;
quit;
************************************************
```

SAS OUTPUT 4b

The MODEL Procedure

The Equation to Estimate is

M = F(b1(MA), b2(D), b3(DR), b4(VT), b5(INC), b6(PL))

Heteroscedasticity Test

Equation	Test	Statistic	DF	Pr > ChiSq	Variables
M	White's Test	28.42	27	0.3898	Cross of all vars
	Breusch-Pagan	1.66	6	0.9479	MA,D,DR,VT,INC, PL,1

❖ The White test:

In statistics, the White test is a statistical test that establishes whether the residual variance of a variable in a regression model is constant. This test does not assume that the residuals are normally distributed. The test uses the null hypothesis H0 : no heteroscedasticity against HA : there is heterocedasticity of some form. It is easy to implement. The principle is that we effect an auxiliary regression: regress squared residuals on all variables, their squares and all possible non-redundant cross-products. For example: We take the square and products increasing the variables of the model. Under the assumption of homoscedastic, it is shown that $n * R^2$ follows a Chi-square with df (degree of freedom) = number of regressors. A disadvantage is that in the presence of several estimators, products quickly consume the degrees of freedom.

The result of SAS OUTPUT shows that White test gives a probability of chi-square = 0.3898> 0.05. We cannot reject the null hypothesis of homoscedasticity.

❖ The Breusch- Pagan test:

The Breusch-Pagan tests whether the estimated variance of the residuals from a regression is dependent on the values of the independent variables. Mechanically it is very similar to White's test. Breusch-Pagan tests the null hypothesis that the error variances are all equal ($H_{0:}$ homoskedasticity) versus the alternative that the error variances are a multiplicative function of one or more variables.
For example, the alternative hypothesis states that the error variances increase (or decrease) as the predicted values of Y increase.

The test statistic is 0.5 * ESS of this regression and it follows a chi ^ 2 with df = (k-1) where k is the number of variables used in the regression. This test has the advantage of being independent of an arbitrary choice. This test, however, assumes that the residuals are normally distributed.
The result of SAS OUTPUT shows that the Breusch-Pagan test for this model gives a probability of chi ^ 2 = 0.9479> 0.05. We cannot reject the null hypothesis of homoscedasticity.

Conclusion: We can conclude that all these tests show there is no heteroskedasticity in this model.

Conclusion

When solving the four parts of the project, I applied the theories and knowledge we learned from the the course Regression Analysis.
I build a linear model without intercept in SAS. I want to verify if the errors are normally distributed and if they are homoskedastic errors. I have illustrated all this using a number of realistic results from SAS output.