```
* Dr. Steven S. Vickner;
* MSBA 635 - Data Analytics II;
 Heteroskedasticity example

* print data;

proc print data=tmp1.food (obs=5);
run;
```

The SAS System 16:59 Tuesday, January 15, 2019 1

Obs	food_exp	income
1	115.22	3.69
2	135.98	4.39
3	119.34	4.75
4	114.96	6.03
5	187.05	12.47

```
* display data attributes;

proc contents data=tmp1.food;
run;
```

The SAS System 16:59 Tuesday, January 15, 2019 2

The CONTENTS Procedure

Data Set Name	TMP1.FOOD	Observations	40
Member Type	DATA	Variables	2
Engine	V9	Indexes	0
Created	12/21/2018 08:29:04	Observation Length	16
Last Modified	12/21/2018 08:29:04	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	YES
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information

Data Set Page Size	4096
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	252
Obs in First Data Page	40
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	C:\Users\nxnguy01\Desktop\food.sas7bdat
Release Created	9.0401M3
Host Created	X64_8PRO

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Label

```
1    food_exp    Num    8    household food expenditure per week
2    income      Num    8    weekly household income
```

Sort Information

Sortedby	income
Validated	YES
Character Set	ANSI

```
* sort data;

*sorts from lowest to highest amount of income, the explanatory variable is
organized from low to high

proc sort data=tmp1.food;
by income;
run;

* estimate regression using proc reg;

*model statement model_exp are a function of income
*output line out is foodout.
*r stands for residual
*p=predicted value
*the income is statistically significant <.0001 is less than 0.05 at the 99%
confidence level and 99% confidence level.
*as income increases by 1 unit food increases by 10.20964 units
*positive intercept of 83.41600
*the fit statistics: root MSE, 38.50% of variability is explained by the
model

options nolabel;
proc reg data=tmp1.food;
model food_exp = income;
output out=foodout r=ehat p=yhat;
run;
quit;
```


The REG Procedure
Model: MODEL1
Dependent Variable: food_exp

Number of Observations Read	40
Number of Observations Used	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	190627	190627	23.79	<.0001
Error	38	304505	8013.29410		
Corrected Total	39	495132			

Root MSE	89.51700	R-Square	0.3850
Dependent Mean	283.57350	Adj R-Sq	0.3688
Coeff Var	31.56748		

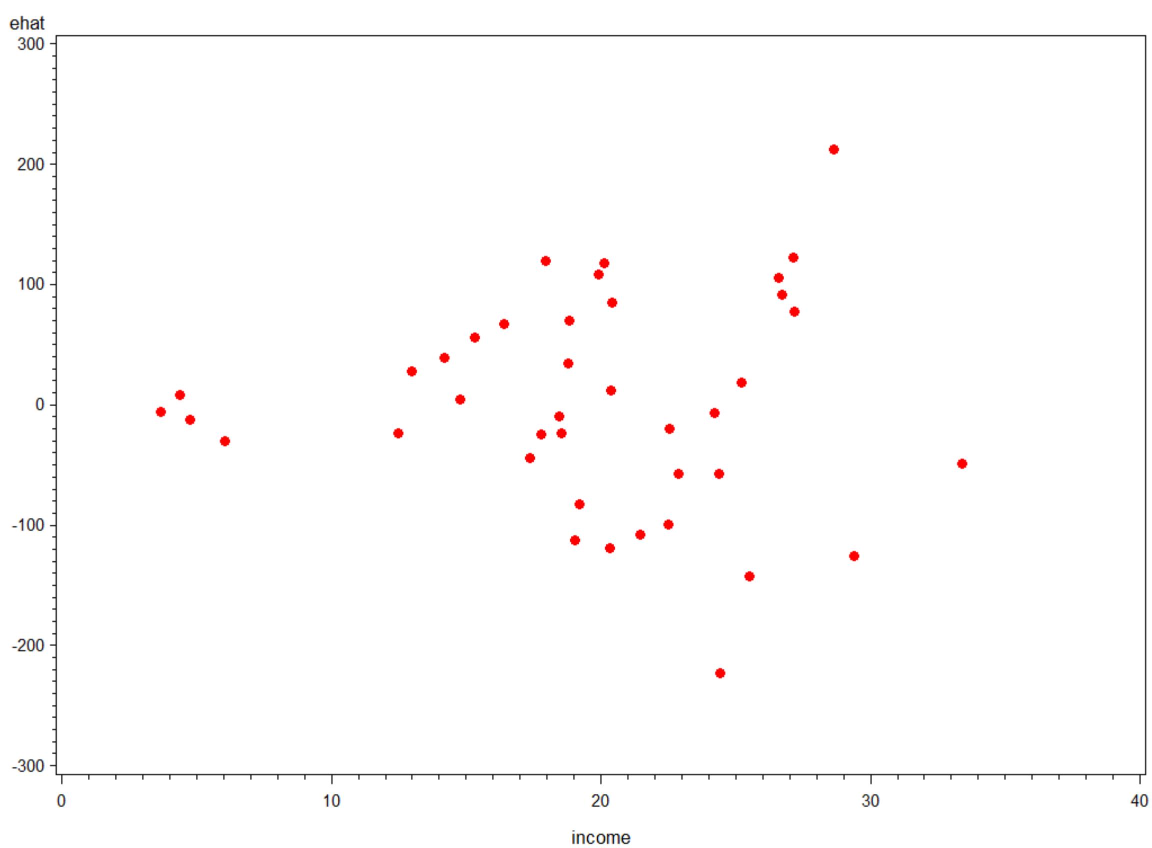
Parameter Estimates

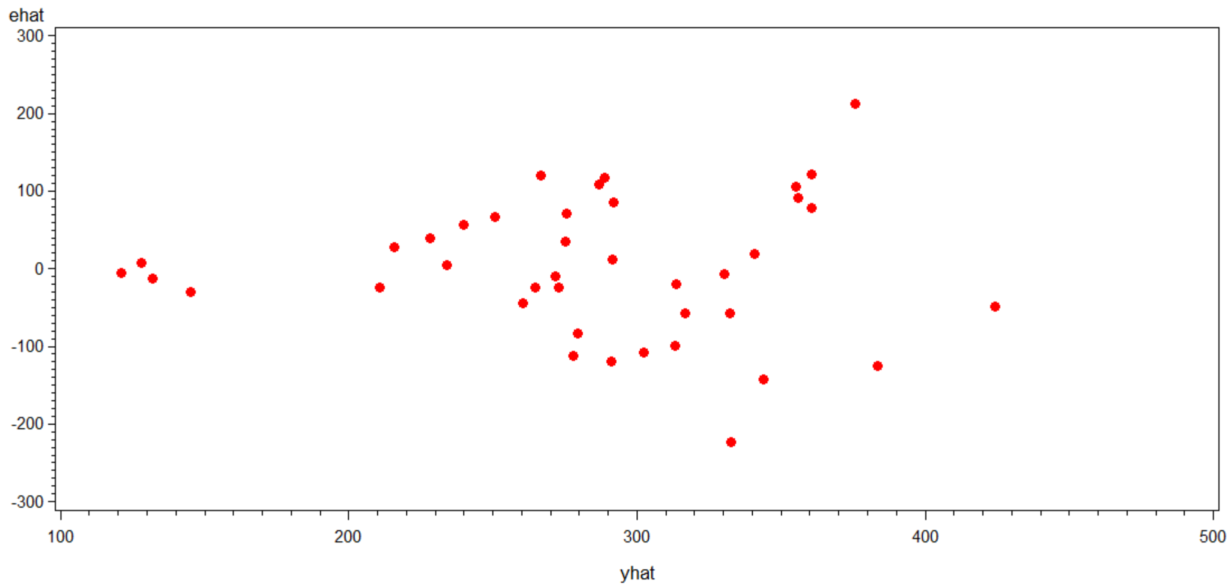
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	83.41600	43.41016	1.92	0.0622
income	1	10.20964	2.09326	4.88	<.0001

```
* produce scatterplot;

*you have syntax here for a scatterplot, if you want to use it just change
the variables
*the first plot, the empirical residuals, those are your errors. They add to
zero because the sum to zero
* Heteroskedasticity is about testing if the variables are constant or not.
*ehat is actual - expected for all 40 observations.
*plot those residuals by expenditures
*if you see cone shape then you have constant violation of error in your
model
*you need some strategy to mitigate it
*theres a bit of a funnel/cone shape in the data. Go from household of less
income to more income there is more diversion from actual - expected
*if you have high income you spend a lot of money on food, if you don't you
spend less on food
*our horizontal axis is our explanatory variable. Lower income has less
dispersion of those empirical residuals (which are actuals - expected)
*you are not predicting well at higher incomes, there is a real wide spread
here

symbol1 value=dot color=red;
proc gplot data=work.foodout;
plot ehat*income=1 / overlay;
plot ehat*yhat=1 / overlay;
run;
quit;
```




ehat
300
200
100
0
-100
-200
-300
100
200
300
400
500
yhat

```
* construct new data sets for Goldfeld-Quandt test;

*set that permanent data set, if the first observations are >20.
*data is ordered by income, first is the smallest 20, second is the largest
20

data gq1;
set tmp1.food;
if _n_ le 20;
run;

data gq2;
set tmp1.food;
if _n_ gt 20;
run;

* estimate regression using proc reg on gq1 data;

options nolabel;
proc reg data=work.gq1;
model food_exp = income;
run;
quit;
```

The SAS System 16:59 Tuesday, January 15, 2019 4

The REG Procedure
Model: MODEL1
Dependent Variable: food_exp

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	75194	75194	21.03	0.0002
Error	18	64346	3574.77175		
Corrected Total	19	139540			

Root MSE	59.78939	R-Square	0.5389
Dependent Mean	240.18300	Adj R-Sq	0.5133
Coeff Var	24.89327		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	72.96174	38.83435	1.88	0.0766
income	1	11.50038	2.50751	4.59	0.0002

```
* estimate regression using proc reg on gq2 data;

*run a regression on the second one. Slope is much larger in value but it has
slipped in overall significance
*this is only significant at the alpha 0.10 level because it is 0.0707
*we see in this particular regression that higher income individuals actual -
expected was spread out
*

options nolabel;
proc reg data=work.gq2;
model food_exp = income;
run;
quit;
```

The SAS System 16:59 Tuesday, January 15, 2019 5

The REG Procedure
Model: MODEL1
Dependent Variable: food_exp

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	47688	47688	3.69	0.0707
Error	18	232595	12922		
Corrected Total	19	280282			

Root MSE	113.67465	R-Square	0.1701
Dependent Mean	326.96400	Adj R-Sq	0.1240
Coeff Var	34.76672		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-24.91465	184.92486	-0.13	0.8943
income	1	14.26400	7.42509	1.92	0.0707

```
* this approach always works;

*to assemble the goldfield quant test,
*20 observations in first observation, 20 observations in second observation
*you want sigma 1 and sigma 2
*in your first output sigma 1=3574.77175, it is the Mean Square Error. Used
in the G-Q test
*in second regression do the same. It is 12922 for the Mean Square Error
here.
*take the 3574.77175/12922. This is the Goldfeld Quant test statistic
*(n2-k) is the numerator degree of freedom
*(n1-k) is the denominator degree of freedom
*this will give you the critical value and give you the p-value
*20-2=18, 18 is the degree of freedom
*k=number of variables

data gqtest;
n1 = 20;
n2 = 20;
k = 2;
sig2_1 = 3574.77175;
sig2_2 = 12922;
gq = sig2_2/sig2_1;
fc = finv(0.95,(n2 - k),(n1 - k));
pval_gq = 1-probf(gq,(n2 - k),(n1 - k));
run;

* print data;

*the goldfield quant=3.61478
*in the f-distribution table the f-critical is row 20 column about 20 =
2.2172, you get close to it
*3.61478>2.21720 so we reject the null hypothesis
*assume mean is 0, reject the null hypothesis that the residuals have
consistantancy
*the variance of the residuals is not constant
*heterosketisity means constant variance

proc print data=work.gqtest;
run;
```

The SAS System 16:59 Tuesday, January 15, 2019 6

Obs	n1	n2	k	sig2_1	sig2_2	gq	fc	pval_gq
1	20	20	2	3574.77	12922	3.61478	2.21720	.004596285

```
* construct variable for Lagrange Multiplier test;

*ehat squared is ehat*ehat
*take each residual of actual - expected and square it

data lmtestdata;
set work.foodout;
ehat2 = ehat*ehat;
run;

* estimate regression using proc reg on lmtest data;

*now dependent variable is ehat2

options nolabel;
proc reg data=work.lmtestdata;
model ehat2 = income;
run;
quit;
```

```
  The SAS System         16:59 Tuesday, January 15, 2019   7

                                    The REG Procedure
                                     Model: MODEL1
                                Dependent Variable: ehat2

                         Number of Observations Read          40
                         Number of Observations Used          40
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	851193272	851193272	8.60	0.0057
Error	38	3759556169	98935689		
Corrected Total	39	4610749441			

Root MSE	9946.64208	R-Square	0.1846
Dependent Mean	7612.62940	Adj R-Sq	0.1632
Coeff Var	130.65974		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-5762.36984	4823.50094	-1.19	0.2396
income	1	682.23258	232.59204	2.93	0.0057

```
* this approach always works;

*this is the lagrange multiplier test
*here you are back to 40 observations
*you do it on the entire dataset
*the r squared=0.1846
*chisq=chi squared test statistic
*chiq critical is 0.95,
*degrees of freedom (s-1) is the parameters -1
*test whether the residuals are constant or not

data lmtest;
s = 2;
n = 40;
r2 = 0.1846;
chisq = n*r2;
csc = cinv(0.95,(s - 1));
pval_chisq = 1-probchi(chisq,(s - 1));
run;

* print data;

*test statistical is greater than the critical value. Reject the null that
the variance is constant. The variance of the error terms is not constant
*

proc print data=work.lmtest;
run;
```

The SAS System 16:59 Tuesday, January 15, 2019 8

Obs	s	n	r2	chisq	csc	pval_chisq
1	2	40	0.1846	7.384	3.84146	.006580665

```
* contruct variables for White test;

*create in this data step the same residual square, ehat*ehat
*also create income squared, income*income

data wtestdata;
set work.foodout;
ehat2 = ehat*ehat;
income2 = income*income;
run;
```

```
* estimate regression using proc reg on wtest data;

*whites test models your dependent variable. Added in another term instead of
having income just linear it is introduced as a quadratic
*not a whole lot of interpretation
*the r-square 0.1889

options nolabel;
proc reg data=work.wtestdata;
model ehat2 = income income2;
run;
quit;
```

The SAS System 16:59 Tuesday, January 15, 2019 9

The REG Procedure
Model: MODEL1
Dependent Variable: ehat2

Number of Observations Read	40
Number of Observations Used	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	870864356	435432178	4.31	0.0208
Error	37	3739885085	101077975		
Corrected Total	39	4610749441			

Root MSE	10054	R-Square	0.1889
Dependent Mean	7612.62940	Adj R-Sq	0.1450
Coeff Var	132.06678		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2908.78281	8100.10876	-0.36	0.7216
income	1	291.74573	915.84618	0.32	0.7519
income2	1	11.16529	25.30953	0.44	0.6617

```
* this approach always works;

*you have 3 parameters, slope on income, slope on income squared

data wtest;
s = 3;
n = 40;
r2 = 0.1889;
chisq = n*r2;
csc = cinv(0.95,(s - 1));
pval_chisq = 1-probchi(chisq,(s - 1));
run;

* print data;

*does same as lagrange multiplier test.
*chsq test stat=7.556, cisq critical =5.99146, reject null hypothesis. It is
non-constant
*0.022868 is less that 0.025 so it is significant


proc print data=work.wtest;
run;
```

The SAS System 16:59 Tuesday, January 15, 2019 10

Obs	s	n	r2	chisq	csc	pval_ chisq
1	3	40	0.1889	7.556	5.99146	0.022868

```
* construct variables for other model specifications to possibly eliminate
heteroskedasticity;

*look at the plots, ehats compared to predicted values, GQ test, lagrange
multiplier test, whites test
*null hypothesis is to reject heteroskedastcity, we have it in our data.
*think of heteroscedasticity in the following 5 ways
*1. Do a natural log transformation to dampen the variance, so log the
dependent and independent variable
*2. Per capita transformation - Divide through by something to control for
the relative size per unit

data newdata;
set work.foodout;
lfood_exp = log(food_exp);
lincome = log(income);
run;
```

```
* estimate regression using proc reg for other model specifications to
possibly eliminate heteroscedasticity;

*set of regressions to run
*the first model is lin lin. Food expeditures linear, income linear.
*there are 4 total models
*1. Dependent variable here is food_exp
*2.lfood_exp is the dependent variable, create the 2 graphs and run your 3
tests to see if there is multicolinearlity
*3. Create 2 graphs and run 3 test
*4. 4th model, do lin lin. Know that this is what we all use. /white
*are parameter estimates the same as the original parameter estimates from
first regression? Yes, what's different is use the columns with the
heteroscedasticity consistent column numbers instead. The income here is
significant.*
*the heteroscedasticity consistent columns are the errors
*always do your graphs and always do your tests
*fitted observations are your regression lines
*heteroscedasticity is introduced through poor sampling design
*you would do this because they are used in segmentation type analysis.
*you'll have heteroscedasticity because you haven't segmented the market well
*to build a better model take high income models out and build a low income
model. Then build a separate model for high incomes


options nolabel;
proc reg data=work.newdata;
model food_exp = lincome;
model lfood_exp = income;
model lfood_exp = lincome;
model food_exp = income / white;
run;
quit;
```


The REG Procedure
Model: MODEL1
Dependent Variable: food_exp

Number of Observations Read	40
Number of Observations Used	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	176520	176520	21.05	<.0001
Error	38	318612	8384.53584		
Corrected Total	39	495132			

Root MSE	91.56711	R-Square	0.3565
Dependent Mean	283.57350	Adj R-Sq	0.3396

Coeff Var	32.29043

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-97.18642	84.23744	-1.15	0.2558
lincome	1	132.16584	28.80461	4.59	<.0001

The REG Procedure
Model: MODEL2
Dependent Variable: lfood_exp

Number of Observations Read	40
Number of Observations Used	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.93046	2.93046	27.27	<.0001
Error	38	4.08304	0.10745		
Corrected Total	39	7.01350			

Root MSE	0.32779	R-Square	0.4178
Dependent Mean	5.56502	Adj R-Sq	0.4025
Coeff Var	5.89024		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.78024	0.15896	30.07	<.0001
income	1	0.04003	0.00767	5.22	<.0001

The REG Procedure
Model: MODEL3
Dependent Variable: lfood_exp

Number of Observations Read	40
Number of Observations Used	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.12262	3.12262	30.50	<.0001
Error	38	3.89088	0.10239		
Corrected Total	39	7.01350			

Root MSE	0.31999	R-Square	0.4452
Dependent Mean	5.56502	Adj R-Sq	0.4306
Coeff Var	5.74997		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.96357	0.29437	13.46	<.0001
lincome	1	0.55588	0.10066	5.52	<.0001

The REG Procedure
Model: MODEL4
Dependent Variable: food_exp

Number of Observations Read	40
Number of Observations Used	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	190627	190627	23.79	<.0001
Error	38	304505	8013.29410		
Corrected Total	39	495132			

Root MSE	89.51700	R-Square	0.3850
Dependent Mean	283.57350	Adj R-Sq	0.3688
Coeff Var	31.56748		

Parameter Estimates

						--Heteroscedasticity Consistent-		
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	1	83.41600	43.41016	1.92	0.0622	26.76835	3.12	0.0035
income	1	10.20964	2.09326	4.88	<.0001	1.76327	5.79	<.0001