```
* MSBA 635 - Data Analytics II;
*this is multicollinearity example
*Multicollinearity is a state of very high intercorrelations or inter-
associations among the dependent variables. It is a type of disturbance in
the data, and if present in the data the statistical inferences made about
the data may not be reliable
*it generally occurs when there are high correlations between two or more
predictor variables. AKA one predictor variable can be used to predict the
other
* print data;
proc print data=tmp1.cars (obs=5);
run;
```

The SAS System 16:53 Tuesday, January 15, 2019 2

Obs	mpg	cyl	eng	wgt
1	18	8	307	3504
2	15	8	350	3693
3	18	8	318	3436
4	16	8	304	3433
5	17	8	302	3449

```
* display data attributes;
proc contents data=tmp1.cars;
run;
```

The SAS System 16:53 Tuesday, January 15, 2019 1

The CONTENTS Procedure

Data Set Name	TMP1.CARS	Observations	392
Member Type	DATA	Variables	4
Engine	V9	Indexes	0
Created	11/14/2010 11:28:36	Observation Length	32
Last Modified	11/14/2010 11:28:36	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information

Data Set Page Size	4096
Number of Data Set Pages	4
First Data Page	1
Max Obs per Page	126
Obs in First Data Page	80
Number of Data Set Repairs	0
Filename	C:\Users\nxnguy01\Desktop\cars.sas7bdat
Release Created	9.0202M3
Host Created	W32_VSPRO

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Label
2	cyl	Num	8	number of cylinders
3	eng	Num	8	engine displacement in cubic inches
1	mpg	Num	8	miles per gallon
4	wgt	Num	8	vehicle weight in pounds

```
* estimate regression using proc reg;
*here the dependent variable is mpg, and the independent variables are: cyl,
eng, and wgt
*this regression will give us one explanatory variable cy1. Theres an inverse
relation between number of cylinders and miles per gallon.
*P-value for cy1 is <.001 and less than 0.05 so it is statistically
significant
*60.47% of variability is explained by the model

options nolabel;
proc reg data=tmp1.cars;
model mpg = cyl;
run;
quit;
```


The REG Procedure
Model: MODEL1
Dependent Variable: mpg

Number of Observations Read	392
Number of Observations Used	392

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	14403	14403	596.56	<.0001
Error	390	9415.91039	24.14336		
Corrected Total	391	23819			

Root MSE	4.91359	R-Square	0.6047
Dependent Mean	23.44592	Adj R-Sq	0.6037
Coeff Var	20.95712		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	42.91551	0.83487	51.40	<.0001
cyl	1	-3.55808	0.14568	-24.42	<.0001

```
* estimate regression using proc reg;
*dependent variable is still mpg. Engine and weight of the vehicle were added
to the cyl independent variable for the equation
*the r-square went up nearly 10 percentage points. (From 0.6047 to 0.6993)
As you introduced other variables cyl is no longer statistically significant
*the second result eng size isn't significant either (0.1253)
*wgt is statistically significant with the p-value <.0001
*things on the right hand side of the equation are highly correlated. This is
multicollinearlity.
*evidence of multicollinearlity, r-square went up, robustness (cyl) went away
and just one independent variable (wgt) is shining through
*cyl and eng are tested simultaneously here. They both equal 0 here. The
results tell us separately they aren't statistically significant, but
together they are because of multicolinearlity.
*all 3 of these things are measuring basically the same thing (cyl, eng, wgt)
do an f-test on them. If you fail to reject the null=0 then you have good
reason to boot them out.
options nolabel;
proc reg data=tmp1.cars;
model mpg = cyl eng wgt;
testcol: test cyl=0, eng=0;
run;
quit;
```


The REG Procedure
Model: MODEL1
Dependent Variable: mpg

Number of Observations Read	392
Number of Observations Used	392

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	16656	5552.14810	300.76	<.0001
Error	388	7162.54916	18.46018		
Corrected Total	391	23819			

Root MSE	4.29653	R-Square	0.6993
Dependent Mean	23.44592	Adj R-Sq	0.6970
Coeff Var	18.32528		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	44.37096	1.48069	29.97	<.0001
cyl	1	-0.26780	0.41307	-0.65	0.5172
eng	1	-0.01267	0.00825	-1.54	0.1253
wgt	1	-0.00571	0.00071392	-8.00	<.0001

*the results tell us that there is **multicolinearlity**
*reject the null hypothesis that those coefficients are equal to 0, you have m**ulticollinearlity**
*do a correlation matrix
*options nolabel will get rid of your labels if you don't want to see them
*the f-test says they are statistically significant. You don't reject the null that they are statistically significant, you reject the null that the values are equal to 0
*with this you go to infinity for the f distribution so your critical value would be **3.00.** The numerator 2 in the top column for DF, once you get past 120 the value will be 3.00 because the denominator is 388
*if looking at t-critical values use 0.975 for a 95% confidence interval. 0.25 in one tail and 0.25 in the other.
*in the F-distribution it say 95th percentile at the top so its for the 0.05.
*

The REG Procedure
Model: MODEL1

Test testcol Results for Dependent Variable mpg

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	79.34228	4.30	0.0142
Denominator	388	18.46018		

```
* obtain correlation matrix;

*you want to see if they are correlated or not
*they are super highly correlated, because they are basically measuring the
same thing
*big heavy cars have a lot of engines with heavy cylinders
*these are saying that the raw Pearson correlation coefficients are
correlated.
*Eng is <.001 under the cyl column,
*wgt is <.0001 under cyl column,
*cyl is <.0001 under eng column,
*wgt is <.0001 under eng column,
*cyl is <.0001 under wgt column
*eng is <.0001 under wgt column
*its Multicollinearity, its saying you are measuring the same thing with all
three of those

options nolabel;
proc corr data=tmp1.cars;
var cyl eng wgt;
run;
quit;
```


The CORR Procedure

3 Variables: cyl eng wgt

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
cyl	392	5.47194	1.70578	2145	3.00000	8.00000
eng	392	194.41199	104.64400	76210	68.00000	455.00000
wgt	392	2978	849.40256	1167213	1613	5140

Pearson Correlation Coefficients, N = 392
Prob > |r| under H0: Rho=0

	cyl	eng	wgt
cyl	1.00000	0.95082 <.0001	0.89753 <.0001
eng	0.95082 <.0001	1.00000	0.93299 <.0001
wgt	0.89753 <.0001	0.93299 <.0001	1.00000

```
* estimate regression using proc reg with vif option;

*vif=variance inflation factors or variance inflation. The VIF detects
multicollinearity in regression analysis. The VIF estimates how much the
variance of a regression coefficient is inflated due to multicollinearity in
the model. VIF=1/(1-Ri2)
*if above the value of 10 you've got trouble.
*if VIF is above 10 you have multicollinearlity, which its above 10 on cy1
(10.51551), and eng (15.78646) so kick those out. Keep wgt (7.7882) because
it is less than 10
* [in the term project make sure to do vif in model specification phase, just
put /vif to do it].
*throw it out of the model if it's over 10 because it is causing the
multicolinearlity. Throw out the cyl and eng since both are over 10
*when doing a cross-sectional regression is when you check the vif
*depended variable cy1 as a function of eng and wgt. Then switch it, eng is
the dependent variable cy1 and weight are the function; and then try it for
weight.

options nolabel;
proc reg data=tmp1.cars;
model mpg = cyl eng wgt / vif;
run;
quit;
```

The SAS System 16:53 Tuesday, January 15, 2019 7

The REG Procedure
Model: MODEL1
Dependent Variable: mpg

Number of Observations Read	392
Number of Observations Used	392

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	16656	5552.14810	300.76	<.0001
Error	388	7162.54916	18.46018		
Corrected Total	391	23819			

Root MSE	4.29653	R-Square	0.6993
Dependent Mean	23.44592	Adj R-Sq	0.6970
Coeff Var	18.32528		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	44.37096	1.48069	29.97	<.0001	0
cyl	1	-0.26780	0.41307	-0.65	0.5172	10.51551

eng	1	-0.01267	0.00825	-1.54	0.1253	15.78646
wgt	1	-0.00571	0.00071392	-8.00	<.0001	7.78872

```
* estimate regression using proc reg with collin option;
*condition index will indicate that the inversion of the matrix is
numerically unstable. Computed by finding the square root of the max
eigenvalue/min eigenvalue.
*the /Collin is another set of metrics to look it. It is the condition index.
10-30 is moderate, above 30 it is severe and means that multicollinearity may
exist. If between 0-10 then you don't have much of a problem. On the chart
below. Advantage of the condition index is it shows which variables are
causing the problem
*The "Numbers" column are the order of the parmaters- intercept, cyl, eng,
wgt
*The diagnostics would indicate to drop 2 of the 3 explanatory variables, as
did the VIF diagnostics. They do not have to recommend dropping the same two
as that doesn't matter since all 3 explanatory variables measures the same
thing
*1.if we find multicolinearlity then the first thing to do is collect more
data. You probability have too small of a data set to identify the right hand
side of your line. More data will introduce more heterogeneity to your data.
*more data helps mitigate this problem by definition, better opportunity to
break the multicolinearlity
*2.you could also re-specify the model and abandon the 2 problematic
variables and keep wgt.
*the condition index corroborated what we found because the value is so high
*3.you can also use factor analysis. It allows you to take 3 variables like
cyl, eng, wgt and collapse all 3 of those measures into one measure and use
it as the explanatory variable. It reduces dimensionality of your data. Its
basically a weighted average of the 3 variables. Use a data reduction
technique here. It says bigger cars with bigger engines that are heavier have
lower mpg. Lighter cars with smaller engines have higher mpg.
*4. Use another technique like neural networks, instead of using regression

options nolabel;
proc reg data=tmp1.cars;
model mpg = cyl eng wgt / collin;
run;
quit;
```


The REG Procedure
Model: MODEL1
Dependent Variable: mpg

Number of Observations Read	392
Number of Observations Used	392

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F

Model	3	16656	5552.14810	300.76	<.0001
Error	388	7162.54916	18.46018		
Corrected Total	391	23819			

Root MSE	4.29653	R-Square	0.6993
Dependent Mean	23.44592	Adj R-Sq	0.6970
Coeff Var	18.32528		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	44.37096	1.48069	29.97	<.0001
cyl	1	-0.26780	0.41307	-0.65	0.5172
eng	1	-0.01267	0.00825	-1.54	0.1253
wgt	1	-0.00571	0.00071392	-8.00	<.0001

Collinearity Diagnostics

		Condition	Proportion of Variation			
Number	Eigenvalue	Index	Intercept	cyl	eng	wgt
1	3.86663	1.00000	0.00134	0.00055885	0.00089840	0.00064110
2	0.12033	5.66860	0.09914	0.00045810	0.05169	0.00000188
3	0.00829	21.60042	0.02754	0.40752	0.00185	0.68206
4	0.00475	28.52804	0.87198	0.59147	0.94556	0.31730