3 Lack of Fit and Multicollinearity in Regression Models

3.1 Lack of Fit in Regression

- When data is collected and a regression model is fit, we want to assess if the fitted regression model adequately models the response variable.
- The ability to assess the quality of the fitted models is possible when replications are taken at several combinations of the predictor variables.
- One simple check for model adequacy is to perform a **lack-of-fit** test. To perform the lack-of-fit test, the error sum of squares, SSE, is partitioned into two components: *pure error* and *lack-of-fit error*:

$$SSE =$$

- The **pure error sum of squares**, SSE_P, is computed using only the responses from replicated observations:

$$SSE_P =$$

where m is the number of unique combinations of predictor variables with multiple observations, and n_i is the number of replicates for the i^{th} combination of variables.

SSE_P can also be written as $SSE_P =$ where s_i^2 is the sample variance of the n_i observations for the i^{th} combination of variables.

$MSE_P = \dfrac{SSE_P}{q-m}$ where $q = n_1 + n_2 + \ldots + n_m$. Note $q - m = \sum_{i=1}^{m}(n_i - 1)$, the total degrees of freedom from pooling across m sets of repeated observations.

- The **lack-of-fit sum of squares**: $SSE_{LOF} = SSE - SSE_P$, and the degrees of freedom are $d.f.(\text{LOF}) = d.f.(\text{Error}) - d.f.(\text{Pure Error}) = N - p - 1 - q + m$.
- The ANOVA table for a model with p terms (not including the intercept):

Source of Variation	d.f.	Sum of Squares	Mean Squares	F-statistic
Regression	p	SS_R	MSR	$F =$
Error	$N-p-1$	SS_E	MSE	
Lack of Fit	$N-p-1-q+m$	SSE_{LOF}	MSE_{LOF}	$F =$
Pure Error	$q-m$	SSE_P	MSE_P	
Total	$N-1$	SS_T		

- A statistically significant lack-of-fit F-statistic implies that the terms in the model do not account for all of the assignable cause variation in the response variable.
- Because the pure error mean square, MSE_P, measures the variation associated with replicate observations, a large F-ratio indicates that the MS_{LOF} is a measure of the variation that exceeds the variation due to pure (replication) error.
- When a large F-ratio occurs, alternative model specifications should be considered (such as a transformation of the response or regressor variables, or inclusion of additional polynomial terms).

3.1.1 Lack of Fit Example

From *Statistical Design and Analysis of Experiments* by Mason, Gunst, and Hess.

- For a particular brand of automobile tire, an experiment was run to examine the relative tire tread wear. Two factors (temperature T and miles M) were believed to affect tread wear (y).
- Temperature T is in degrees Fahrenheit. Miles M is the number of miles the tires were tested on a wet road surface.
- There are $m = 8$ unique combinations of (T, M), and $n_i = 4$ replicates taken for each of the $m = 8$ combinations. The experimental data is given in the following table. We will perform lack-of-fit tests for several regression models.

Temp (T)	Miles (M)	Rep (j)	Pure Error df	Wear y_{ij}	Sum of squares for (T, M) combination
53.2	388	1		2.25847	
53.2	388	2	3	2.19915	$\sum_{j=1}^{4}(y_{1j} - \overline{y}_{1\cdot})^2$
53.2	388	3		2.19068	$= .0405525$
53.2	388	4		1.99153	
53.3	438	1		2.27837	
53.3	438	2	3	2.22698	$\sum_{i=1}^{4}(y_{2j} - \overline{y}_{2\cdot})^2$
53.3	438	3		2.03854	$= .0445465$
53.3	438	4		2.05139	
66.0	58	1		2.68891	
66.0	58	2	3	2.63003	$\sum_{i=1}^{4}(y_{3j} - \overline{y}_{3\cdot})^2$
66.0	58	3		2.59078	$= .0068071$
66.0	58	4		2.58587	
70.3	7	1		2.35556	
70.3	7	2	3	2.41333	$\sum_{i=1}^{4}(y_{4j} - \overline{y}_{4\cdot})^2$
70.3	7	3		2.34222	$= .0052346$
70.3	7	4		2.42667	
76.9	28	1		2.15361	
76.9	28	2	3	2.20181	$\sum_{i=1}^{4}(y_{5j} - \overline{y}_{5\cdot})^2$
76.9	28	3		2.22892	$= .0034212$
76.9	28	4		2.16867	
78.4	25	1		2.09884	
78.4	25	2	3	2.25000	$\sum_{i=1}^{4}(y_{6j} - \overline{y}_{6\cdot})^2$
78.4	25	3		2.04360	$= .0242599$
78.4	25	4		2.08721	
88.1	275	1		2.07979	
88.1	275	2	3	2.23404	$\sum_{i=1}^{4}(y_{7j} - \overline{y}_{7\cdot})^2$
88.1	275	3		2.07713	$= .0209151$
88.1	275	4		2.21011	
89.6	324	1		2.01934	
89.6	324	2	3	2.26796	$\sum_{i=1}^{4}(y_{8j} - \overline{y}_{8\cdot})^2$
89.6	324	3		2.10221	$= .0426335$
89.6	324	4		2.01105	
			$df_P = 24$		$SS_P = .1883703$

Pure error mean square $MS_P = .1883703/24 = .00785$

- Let $x_1 = T$ and $x_2 = M$. Perform a lack-of-fit test for the following models:

$$\begin{array}{lrcl} \text{Model (1)} & y_{ij} & = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_{ij} \\ \text{Model (2)} & y_{ij} & = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon_{ij} \\ \text{Model (3)} & y_{ij} & = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \epsilon_{ij} \end{array}$$

- We know that $SS_P = .1883703$, $df_P = 24$, and $MS_P = .00785$. These values associated with pure error **do not depend on the choice of model.**

- The next step is to fit each model, and calculate $SS_{LOF} = SS_E - SS_P$, $df_{LOF} = df_E - df_P$, and $MS_{LOF} = SS_{LOF}/df_{LOF}$ for each model.

For Model (1):

Source of Variation	d.f.	Sum of Squares	Mean Squares	F-statistic	p-value
Regression	2	.37764	.18882	$F = 7.60$	.0022
Error	29	.72045	.02484		
Lack of Fit	*5*	*.53208*	*.10642*	*F =13.56*	*<.0001*
Pure Error	*24*	*.18837*	*.00785*		
Total	31	1.09808			

There is a significant lack-of-fit for Model 1 (p-value $< .0001$). We should consider adding additional terms to the model (such as an interaction term ($\beta_{12}x_1x_2$) or squared terms ($\beta_{11}x_1^2$ or $\beta_{22}x_2^2$)). If we add these three terms, we now have Model 2.

For Model (2):

Source of Variation	d.f.	Sum of Squares	Mean Squares	F-statistic	p-value
Regression	5	.90271	.18054	$F = 24.03$	$< .0001$
Error	26	.19538	.00751		
Lack of Fit	*2*	*.00701*	*.00350*	*F =0.446*	*.6451*
Pure Error	*24*	*.18837*	*.00785*		
Total	31	1.09808			

There is no significant lack-of-fit for Model 2 (p-value = .6451). We do not need to add any additional terms to the model. It may be possible to remove terms and still not reject the lack-of-fit test. If we remove $\beta_{22}x_2^2$, then we get Model 3.

For Model (3):

Source of Variation	d.f.	Sum of Squares	Mean Squares	F-statistic	p-value
Regression	4	.84060	.21015	$F = 22.04$	$< .0001$
Error	27	.25748	.00954		
Lack of Fit	*3*	*.06911*	*.02304*	*F =2.93*	*.0538*
Pure Error	*24*	*.18837*	*.00785*		
Total	31	1.09808			

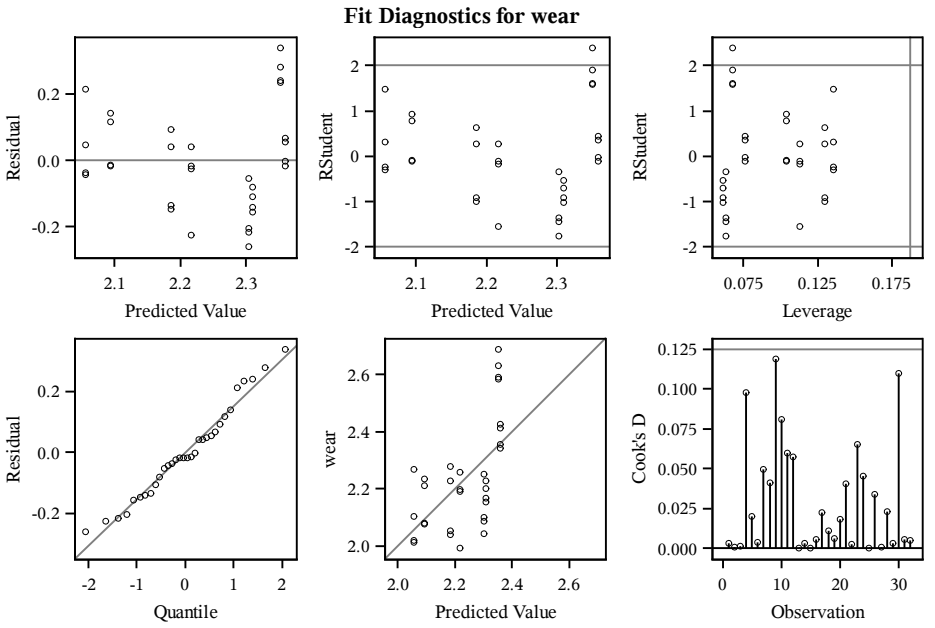
There is marginal evidence (p-value = .0538) for lack-of-fit at $\alpha = .05$ significance level.

3.1.2 Using PROC REG in SAS

PROC REG Output for Model 1: $y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_{ij}$.

LOF Test and Collinearity Check for MODEL 1

The REG Procedure
Model: MODEL1
Dependent Variable: wear



Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	0.37764	0.18882	7.60	0.0022
Error	29	0.72045	0.02484		
Lack of Fit	5	0.53208	0.10642	13.56	<.0001
Pure Error	24	0.18837	0.00785		
Corrected Total	31	1.09808			

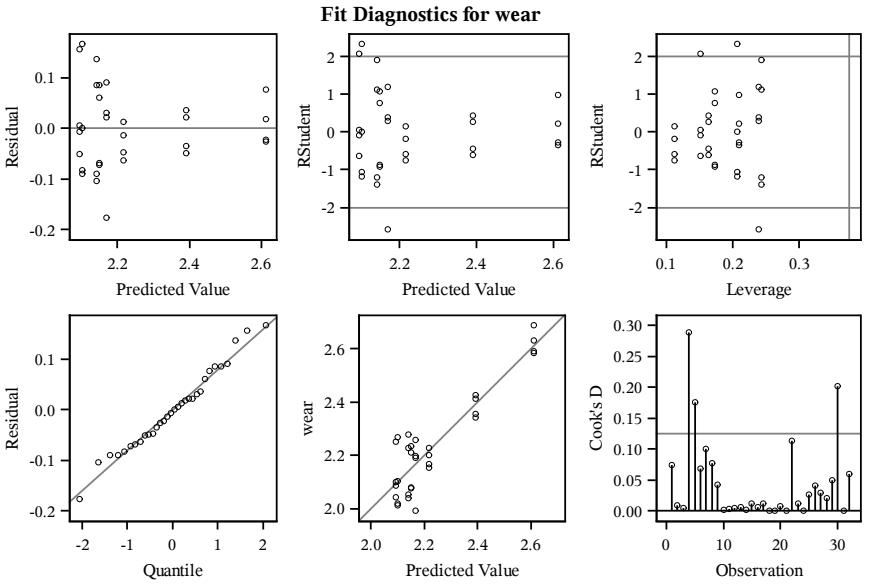
Root MSE	0.15762	**R-Square**	0.3439
Dependent Mean	2.23446	**Adj R-Sq**	0.2987
Coeff Var	7.05390		

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	2.21919	0.02829	78.46	<.0001	0
T	1	-0.10078	0.04056	-2.48	0.0190	1.10054
M	1	-0.13424	0.03713	-3.62	0.0011	1.10054

PROC REG Output for Model 2: $y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon_{ij}$.

LOF Test and Collinearity Check for MODEL 2

The REG Procedure
Model: MODEL1
Dependent Variable: wear



Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	0.90271	0.18054	24.03	<.0001
Error	26	0.19538	0.00751		
Lack of Fit	2	0.00701	0.00350	0.45	0.6451
Pure Error	24	0.18837	0.00785		
Corrected Total	31	1.09808			

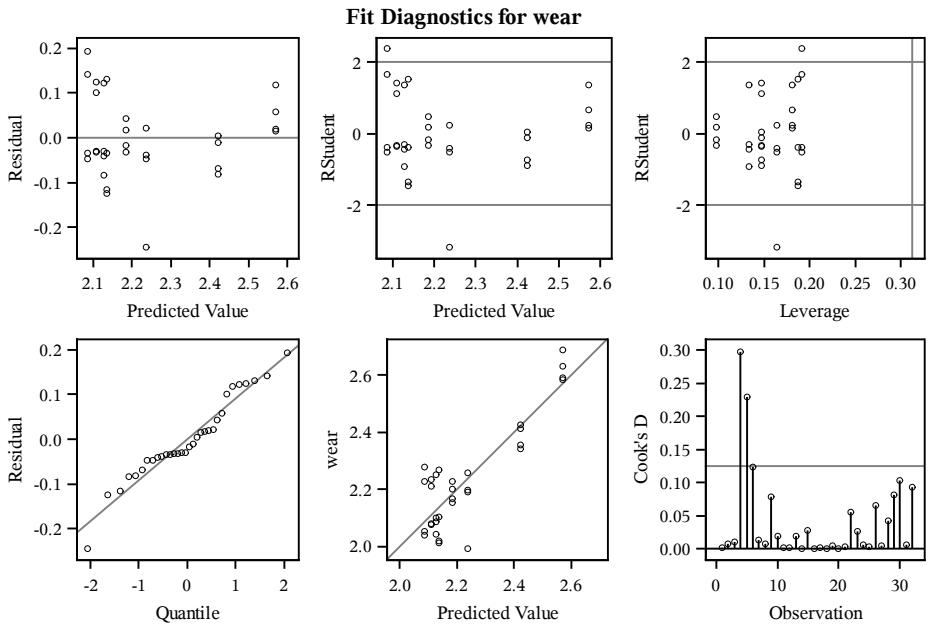
Root MSE	0.08669	**R-Square**	0.8221
Dependent Mean	2.23446	**Adj R-Sq**	0.7879
Coeff Var	3.87952		

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	3.16128	0.28727	11.00	<.0001	0
T	1	-0.18238	0.06095	-2.99	0.0060	8.21415
M	1	0.55280	0.23565	2.35	0.0269	146.52531
TM	1	0.30089	0.07217	4.17	0.0003	5.14839
T2	1	-1.22596	0.42645	-2.87	0.0080	151.00370
M2	1	-0.24177	0.11328	-2.13	0.0424	5.78881

SAS PROG REG Output for Model 3: $y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \epsilon_{ij}$.

LOF Test and Collinearity Check for MODEL 3

The REG Procedure
Model: MODEL1
Dependent Variable: wear



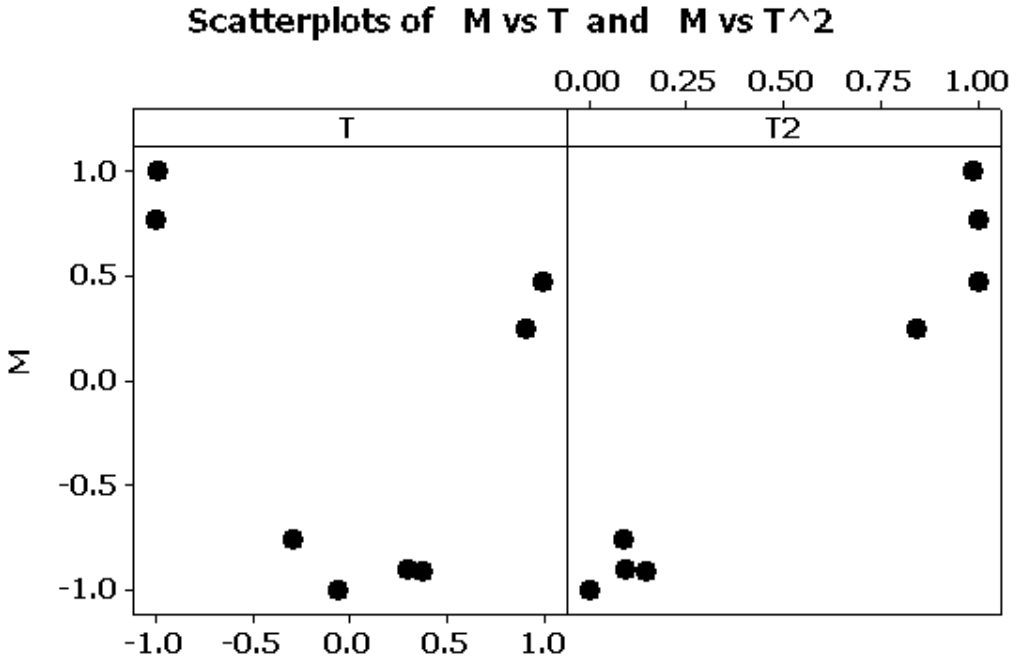
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	0.84060	0.21015	22.04	<.0001
Error	27	0.25748	0.00954		
Lack of Fit	3	0.06911	0.02304	2.94	0.0538
Pure Error	24	0.18837	0.00785		
Corrected Total	31	1.09808			

Root MSE	0.09765	**R-Square**	0.7655
Dependent Mean	2.23446	**Adj R-Sq**	0.7308
Coeff Var	4.37036		

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	2.35173	0.06399	36.75	<.0001	0
T	1	-0.31994	0.04253	-7.52	<.0001	3.15156
M	1	-0.11738	0.03878	-3.03	0.0054	3.12647
TM	1	0.38078	0.07503	5.08	<.0001	4.38483
M2	1	-0.08827	0.11255	-0.78	0.4397	4.50268

Original and Coded Data

Obs	wear	temp	T	miles	M	T2
1	2.25847	53.2	-1.00000	388	0.76798	1.00000
2	2.19915	53.2	-1.00000	388	0.76798	1.00000
3	2.19068	53.2	-1.00000	388	0.76798	1.00000
4	1.99153	53.2	-1.00000	388	0.76798	1.00000
5	2.27837	53.3	-0.99451	438	1.00000	0.98904
6	2.22698	53.3	-0.99451	438	1.00000	0.98904
7	2.03854	53.3	-0.99451	438	1.00000	0.98904
8	2.05139	53.3	-0.99451	438	1.00000	0.98904
9	2.68891	66.0	-0.29670	58	-0.76334	0.08803
10	2.63003	66.0	-0.29670	58	-0.76334	0.08803
11	2.59078	66.0	-0.29670	58	-0.76334	0.08803
12	2.58587	66.0	-0.29670	58	-0.76334	0.08803
13	2.35556	70.3	-0.06044	7	-1.00000	0.00365
14	2.41333	70.3	-0.06044	7	-1.00000	0.00365
15	2.34222	70.3	-0.06044	7	-1.00000	0.00365
16	2.42667	70.3	-0.06044	7	-1.00000	0.00365
17	2.15361	76.9	0.30220	28	-0.90255	0.09132
18	2.20181	76.9	0.30220	28	-0.90255	0.09132
19	2.22892	76.9	0.30220	28	-0.90255	0.09132
20	2.16867	76.9	0.30220	28	-0.90255	0.09132
21	2.09884	78.4	0.38462	25	-0.91647	0.14793
22	2.25000	78.4	0.38462	25	-0.91647	0.14793
23	2.04360	78.4	0.38462	25	-0.91647	0.14793
24	2.08721	78.4	0.38462	25	-0.91647	0.14793
25	2.07979	88.1	0.91758	275	0.24362	0.84196
26	2.23404	88.1	0.91758	275	0.24362	0.84196
27	2.07713	88.1	0.91758	275	0.24362	0.84196
28	2.21011	88.1	0.91758	275	0.24362	0.84196
29	2.01934	89.6	1.00000	324	0.47100	1.00000
30	2.26796	89.6	1.00000	324	0.47100	1.00000
31	2.10221	89.6	1.00000	324	0.47100	1.00000
32	2.01105	89.6	1.00000	324	0.47100	1.00000



PROC REG SAS Code for Models 1, 2, and 3:

```
DM 'LOG; CLEAR; OUT; CLEAR;';

ODS PRINTER PDF file='C:\COURSES\ST578\SAS\LOF_WEAR.pdf';
ODS LISTING;
OPTIONS PS=54 LS=72 NODATE NONUMBER;

*********************************************;
*** LACK OF FIT ANALYSIS FOR TREADWEAR DATA ***;
*********************************************;

DATA in; INPUT wear temp miles @@; LINES;
2.25847 53.2 388   2.19915 53.2 388    2.19068 53.2 388   1.99153 53.2 388
2.27837 53.3 438   2.22698 53.3 438    2.03854 53.3 438   2.05139 53.3 438
2.68891 66.0  58   2.63003 66.0  58    2.59078 66.0  58   2.58587 66.0  58
2.35556 70.3   7   2.41333 70.3   7    2.34222 70.3   7   2.42667 70.3   7
2.15361 76.9  28   2.20181 76.9  28    2.22892 76.9  28   2.16867 76.9  28
2.09884 78.4  25   2.25000 78.4  25    2.04360 78.4  25   2.08721 78.4  25
2.07979 88.1 275   2.23404 88.1 275    2.07713 88.1 275   2.21011 88.1 275
2.01934 89.6 324   2.26796 89.6 324    2.10221 89.6 324   2.01105 89.6 324
 ;
DATA in; SET in;

 *** CODE THE VARIABLE LEVELS WITH MIN = -1 and MAX = +1 ***;
     T = (temp-71.4)/18.2;
     M = (miles-222.5)/215.5;
     TM = T*M;
     T2 = T**2;
     M2 = M**2;

PROC PRINT DATA = in;
     VAR wear temp T miles M T2;
TITLE 'Original and Coded Data';

PROC REG DATA=in;
     MODEL wear = T M / LACKFIT VIF;
TITLE 'LOF Test and Collinearity Check for MODEL 1';

PROC REG DATA=in;
     MODEL wear = T M TM T2 M2 / LACKFIT VIF;
TITLE 'LOF Test and Collinearity Check for MODEL 2';

PROC REG DATA=in;
     MODEL wear = T M TM M2 / LACKFIT VIF;
TITLE 'LOF Test and Collinearity Check for MODEL 3';
RUN;
```

3.2 Multicollinearity in Regression

- In regression problems, strong linear dependencies may exist among the regressor (predictor) variables in the regression model. If strong linear dependencies exist, then **multicollinearity** exists among the regressors.
- Matrix interpretation: Let $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_k}]$ be the design matrix where $\mathbf{x_j}$ is the j^{th} column of $\mathbf{X}$. Multicollinearity occurs when there are strong linear dependencies among the columns of $\mathbf{X}$. That is, there exits a set of constants c_j (not all 0) for which $\sum_{j=1}^{k} c_j \mathbf{x_j} \approx 0$.

- Multicollinearity can seriously affect the estimates of the model parameters, and, therefore, affect the interpretation and applicability of the final model.

- When a set of regressors are conditioned in such a way that the regressors are highly collinear, the least squares procedure can not detect the structure of the multicollinearity in the data. With multicollinearity, the least squares method cannot separate the individual effects of collinear variables.

- There are numerous methods for detecting the presence of multicollinearity. I will discuss several common methods:

  1. Look at sample correlation coefficients r_{ij} for each pair of regressor variables x_i and x_j. If $|r_{ij}|$ is close to 1, then x_i and x_j are strongly collinear. This method, however, will only detect collinearity among pairs of variables. If more than two variables are multicollinear, this method will not always enable the detection of the multicollinearity .

  2. If the F-test for the fit of the regression model is significant, but the F-tests for the individual regression coefficients are not significant, then multicollinearity may be present.

  3. The eigenvalues of the $\mathbf{X'X}$ matrix for design matrix $\mathbf{X}$ provide a measure of multicollinearity. One or more eigenvalues near 0 imply that multicollinearity is present. If λ_{max} and λ_{min} are the largest and smallest eigenvalues of $\mathbf{X'X}$, then the **condition number** $\kappa =$

     Rule of thumb: If $\kappa < 100$, there is little problem with multicollinearity . Also, the number of small eigenvalues is directly related to the number of multicollinearities.

  4. The **variance inflation factor (VIF)** is defined to be:

     $$\text{VIF}(b_j) = \qquad\qquad\qquad\qquad j = 1, 2, \ldots, k$$

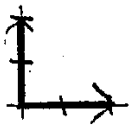
     where R_j^2 is the coefficient of multiple determination resulting from the regression of x_j on the remaining $k-1$ regressors (b_j is the model coefficient associated with x_j). A strong multicollinearity between x_j and the other regressors variables is reflected in R_j^2 close to 1. Thus, the VIF will increase as $R_j^2 \longrightarrow 1$.

  5. We say that the variance of b_j is "inflated" by the quantity $\dfrac{1}{(1-R_j^2)}$. That is, the VIFs represent the inflation that each regression coefficient experiences above the case when the correlation matrix of the regressor variables is an identity matrix.

     Some authors have suggested that if any VIFs exceed 10 ($R_j^2 = .90$), then multicollinearity is a problem. Other authors consider this value is too liberal and suggest that the VIFs should not exceed 4 or 5 ($R_j^2 = .75, .80$).

- Some computing packages provide **tolerances**. The tolerance of $b_j = \dfrac{1}{\text{VIF}(b_j)} =$

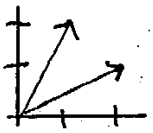
  Thus, we do not want tolerances to be too small. For example if a VIF > 10, then the tolerance is $< .10$.

$A = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$  $|A| = 4$

CONDITION NUMBER $\frac{\lambda_{MAX}}{\lambda_{min}} = \frac{2}{2} = 1$

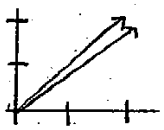
$|A - \lambda I| = \begin{vmatrix} 2-\lambda & 0 \\ 0 & 2-\lambda \end{vmatrix} = (2-\lambda)^2$

EIGENVALUES OF A ARE ROOTS OF $(2-\lambda)^2 = 0 \Rightarrow \boxed{\lambda = 2, 2}$

$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$  $|A| = 4 - 1 = 3$ $\frac{\lambda_{max}}{\lambda_{min}} = \frac{3}{1} = 3$

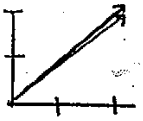
$|A - \lambda I| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = (2-\lambda)^2 - 1 = \lambda^2 - 4\lambda + 3 = (\lambda - 1)(\lambda - 3)$

EIGENVALUES OF A ARE $\boxed{\lambda = 1, 3}$

$A = \begin{bmatrix} 2 & 1.9 \\ 1.9 & 2 \end{bmatrix}$  $|A| = 4 - 3.61 = .39$ $\frac{\lambda_{max}}{\lambda_{min}} = \frac{3.9}{.1} = 39$

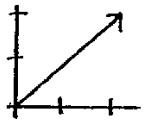
$|A - \lambda I| = \begin{vmatrix} 2-\lambda & 1.9 \\ 1.9 & 2-\lambda \end{vmatrix} = (2-\lambda)^2 - 3.61 = \lambda^2 - 4\lambda + .39 = (\lambda - 3.9)(\lambda - .1)$

EIGENVALUES OF A ARE $\boxed{\lambda = .1, 3.9}$

$A = \begin{bmatrix} 2 & 1.99 \\ 1.99 & 2 \end{bmatrix}$  $|A| = 4 - 3.9601 = .0399$ $\frac{\lambda_{max}}{\lambda_{min}} = \frac{3.99}{.01} = 399$

$|A - \lambda I| = \begin{vmatrix} 2-\lambda & 1.99 \\ 1.99 & 2-\lambda \end{vmatrix} = (2-\lambda)^2 - 3.9601 = \lambda^2 - 4\lambda + .0399 = (\lambda - 3.99)(\lambda - .01)$

EIGENVALUES OF A ARE $\boxed{\lambda = .01, 3.99}$

$A = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$  $|A| = 4 - 4 = 0$ $\frac{\lambda_{max}}{\lambda_{min}} = \frac{4}{0} \to \infty$

$|A - \lambda I| = \begin{vmatrix} 2-\lambda & 2 \\ 2 & 2-\lambda \end{vmatrix} = (2-\lambda)^2 - 4 = \lambda^2 - 4\lambda = \lambda(\lambda - 4)$

EIGENVALUES OF A ARE $\boxed{\lambda = 0, 4}$