STAT 525 FALL 2018

Chapter 10
Diagnostics in Multiple Regression

Professor Min Zhang

Model Adequacy for a Predictor

Partial Coefficient of Determination

Consider model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- Define two variables:
  - residuals of predicting Y as function of $X_2, \cdots, X_{p-1}$

$$e_i(Y|X_2, \cdots, X_{p-1}) = Y_i - \widehat{Y}_i(X_2, \cdots, X_{p-1})$$

  - residuals of predicting X_1 as function of $X_2, \cdots, X_{p-1}$

$$e_i(X_1|X_2, \cdots, X_{p-1}) = X_{i1} - \widehat{X}_{i1}(X_2, \cdots, X_{p-1})$$

- $R^2_{Y1|2:(p-1)}$ equals to R^2 for regressing $e_i(Y|X_2, \cdots, X_{p-1})$ on $e_i(X_1|X_2, \cdots, X_{p-1})$
  - Measures *additional* information in X_1 helping predict Y
  - Equals to squared partial correlation $r^2_{Y1|2:(p-1)}$
- Similarly consider each X_j, $j = 2, \cdots, p-1$

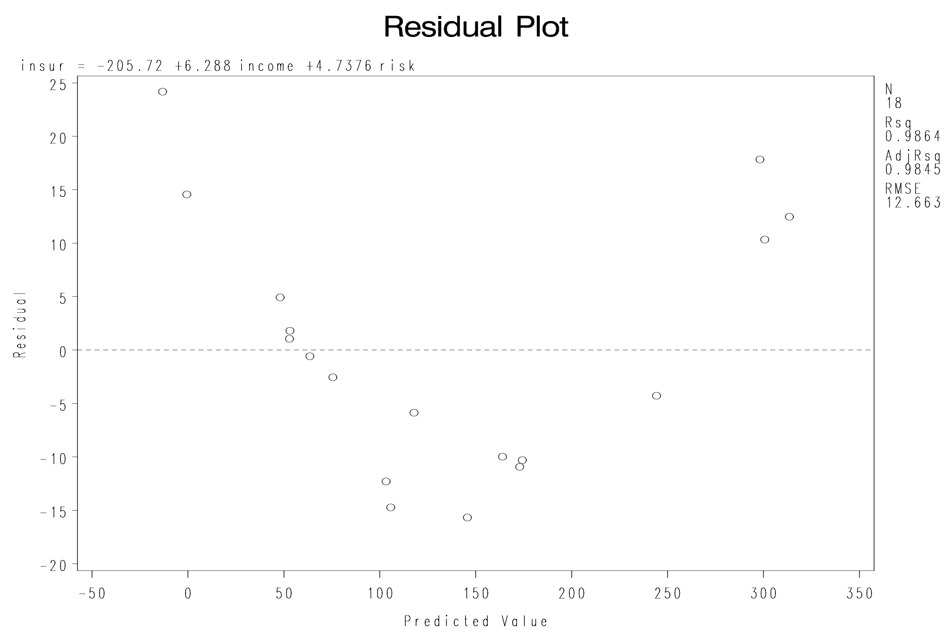
Partial Regression Plots

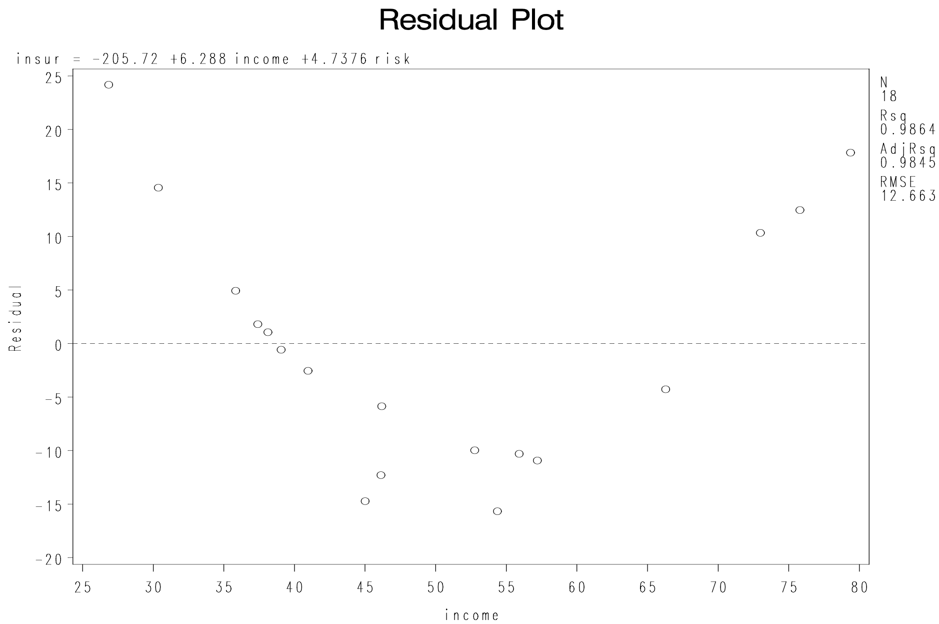
- For X_j, plot $e_i(Y|all\ X_k\ with\ k \neq j)$ vs $e_i(X_j|all\ X_k\ with\ k \neq j)$
- Also called added variable plots or adjusted variable plots
- Shows strength of the marginal relationship given other variables in the model model
  - Provides visual display of relationship
  - Allows check of “adjusted” relationship
- Can detect:
  - Nonlinear relationship
  - Heterogeneous variance
  - Unusual observations
- NOTE: *Partial Residual Plots* are based on fitting the full model only
  - Plot $e_i + \widehat{\beta}_j X_{ij}$ vs. X_{ij}
  - More meaningful way to show the relationship between Y and X_j

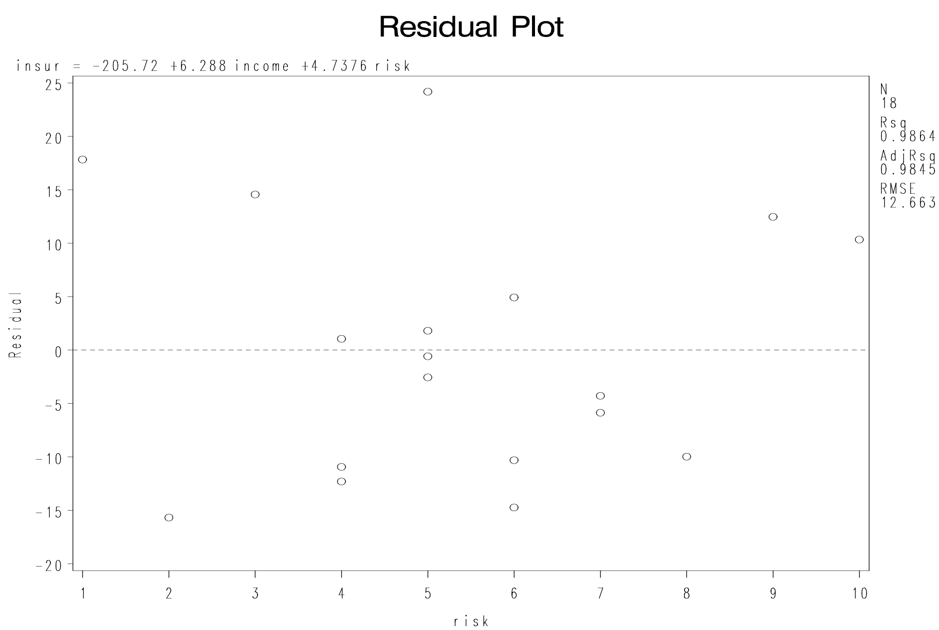
Example on Page 386

- Surveyed 18 managers age 30-39. Interested in relating the amount of life insurance carried to risk aversion and salary.
- Y is amount of life insurance carried
- Two predictor variables
  - X_1 average annual income during past two years
  - X_2 risk aversion score (higher $\rightarrow$ more averse)

```
options nocenter ls=75;
data a1;
    infile 'D:\nobackup\tmp\CH10TA01.TXT';
    input income risk insur;

proc reg data=a1;
    model insur = income risk;
    output out=out1 r=resid p=pred;

proc sort data=out1; by pred; run;
symbol v=circle i=none c=black;
proc gplot data=out1;
    plot resid*pred resid*income resid*risk/vref=0;
run; quit;
```




Residual Plot
insur = -205.72 +6.288 income +4.7376 risk
N
18
Rsq
0.9864
AdjRsq
0.9845
RMSE
12.663
Residual
income
25
20
15
10
5
0
-5
-10
-15
-20
25 30 35 40 45 50 55 60 65 70 75 80


Residual Plot
insur = -205.72 +6.288 income +4.7376 risk
N
18
Rsq
0.9864
AdjRsq
0.9845
RMSE
12.663
Residual
risk
25
20
15
10
5
0
-5
-10
-15
-20
1 2 3 4 5 6 7 8 9 10

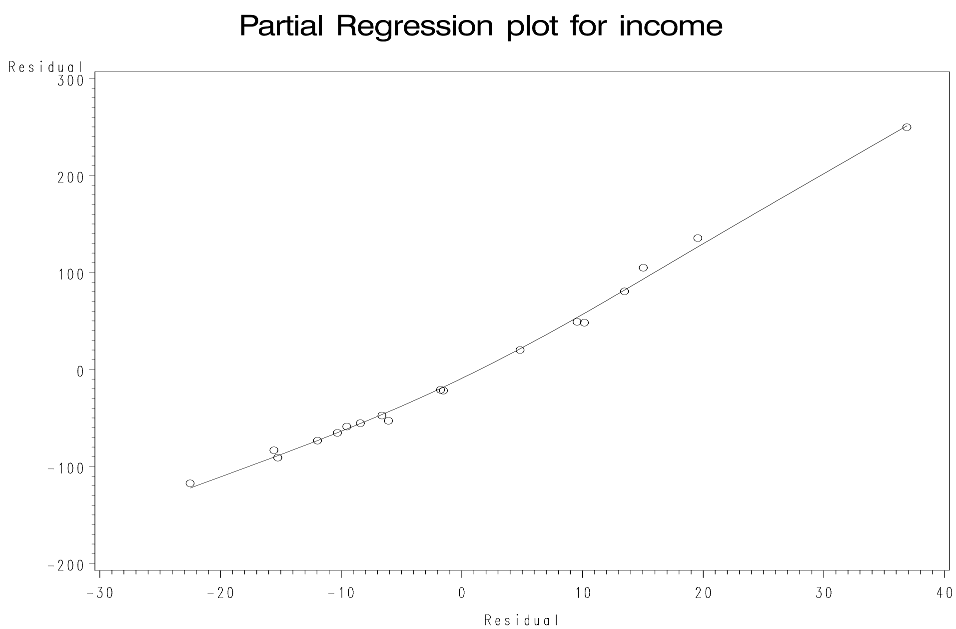
Partial Regression Plots from `PROC REG`

```
proc reg data=a1;
    model insur = income risk/partial;
run; quit;
```

- Provide line printer plots for partial regression plots for each predictor.
- We would rather to generate graphics displays.

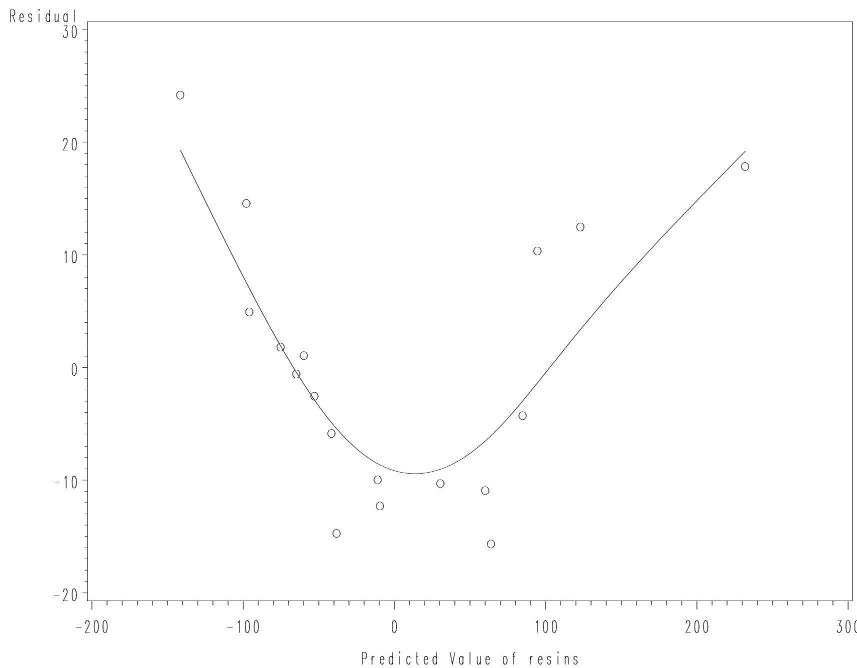
Partial Regression Plot of X_1 (Income)

```
proc reg data=a1;
    model insur income = risk;
    output out=a2 r=resins resinc;
run;

proc sort data=a2; by resinc; run;

symbol v=circle i=sm70 c=black;
proc gplot data=a2;
    plot resins*resinc;
run;
```



e(insurance|risk) vs. e(income|risk)

```
/* --- Is a linear relationship appropriate? ---*/
proc reg data=a2;
    model resins = resinc;
    output out=new1 r=res p=pred;
run;

symbol v=circle i=sm70 c=black;
proc gplot data=new1;
    plot res*pred;
run; quit;
```



residuals of `model resins = resinc`

Output of `model resins = resinc`

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.10593E-14	2.88985	0.00	1.0000
resinc	Residual	1	6.28803	0.19767	31.81	<.0001

Output of `model insur=income risk`

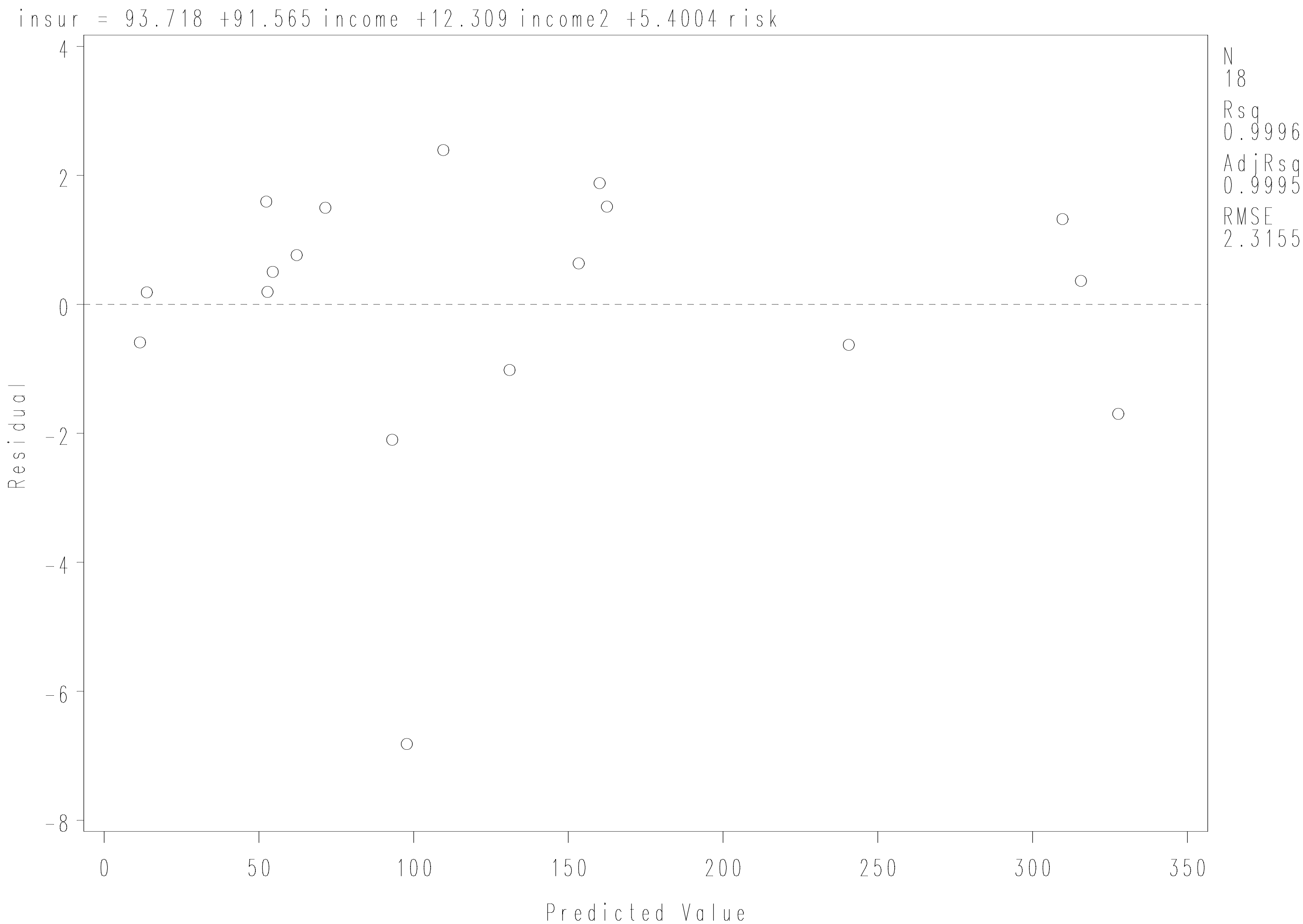
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-205.71866	11.39268	-18.06	<.0001
income	1	6.28803	0.20415	30.80	<.0001
risk	1	4.73760	1.37808	3.44	0.0037

- Note that the parameter estimate for the slope is the same as the parameter estimate for `risk` in the full model

```
data a2; set a1;
    sincome = income;

proc standard data=a2 out=a3 mean=0 std=1;
    var sincome;

data a3; set a3;
    sincome2=sincome*sincome;

/* --- Include quadratic term of income in the model ---*/
proc reg data=a3;
    model insur = sincome sincome2 risk;
    output out=out2 r=resid p=pred;
run;

proc sort data=out2; by pred; run;
symbol v=circle i=none c=black;
proc gplot data=out2;
    plot resid*pred/vref=0;
run; quit;
```

Residual Plot



Output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	176249	58750	10958.0	<.0001
Error	14	75.05895	5.36135		
Corrected Total	17	176324			

Root MSE	2.31546	R-Square	0.9996
Dependent Mean	134.44444	Adj R-Sq	0.9995
Coeff Var	1.72224		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	93.71759	1.63501	57.32	<.0001
sincome	1	91.56523	0.65352	140.11	<.0001
sincome2	1	12.30855	0.59042	20.85	<.0001
risk	1	5.40039	0.25399	21.26	<.0001

Residuals for Diagnostics

- $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$
  - $\mathbf{I} - \mathbf{H}$ symmetric and idempotent
- Expected value $E(\mathbf{e}) = \mathbf{0}$
- Covariance matrix

$$\begin{aligned} \boldsymbol{\sigma}^2(\mathbf{e}) &= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' \\ &= \sigma^2(\mathbf{I} - \mathbf{H}) \end{aligned}$$

  - $\text{Var}(e_i) = \sigma^2 \cdot (1 - h_{ii})$ where $h_{ii} = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i$
  - $Cov(e_i, e_j) = \sigma^2 \cdot (0 - h_{ij}) = -\sigma^2 h_{ij}$
- Estimated variance and covariance
  - $\widehat{Var}(e_i) = MSE \cdot (1 - h_{ii})$
  - $\widehat{Cov}(e_i, e_j) = -MSE \cdot h_{ij}$

Residuals

- Ordinary residual

$$e_i = Y_i - \widehat{Y}_i \quad \rightarrow \quad \mathbf{e} \sim \text{MVN}(\mathbf{0}, (\mathbf{I} - \mathbf{H})\sigma^2)$$

  – residuals do not have the same variance, but depend on $\mathbf{X}_i$

- Semi-studentized residual

$$r_i = \frac{e_i}{\sqrt{\text{MSE}}}$$

  – denominator is not an estimate of SD of e_i

- Studentized residual

$$r_i = \frac{e_i}{\sqrt{\text{MSE}(1 - h_{ii})}}$$

  – denominator is the estimate of SD of e_i

- All these residuals are optimistic in that the corresponding observations were part of the regression fit

Deleted Residual

- Deleted residual

$$d_i = Y_i - \widehat{Y}_{i(i)} = \frac{e_i}{1 - h_{ii}}$$

  - $(\mathbf{X}_i, Y_i)$ was not used to fit the model
  - can calculate d_i in a single model fit

- Standard deviation of deleted residuals

$$\begin{aligned} s^2\{d_i\} &= MSE_{(i)} \cdot (1 + \mathbf{X}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{X}_i) \\ &= \frac{MSE_{(i)}}{1 - h_{ii}} \end{aligned}$$

- Studentized deleted residual

$$\begin{aligned} t_i &= \frac{d_i}{s\{d_i\}} = \frac{e_i}{1 - h_{ii}} \cdot \sqrt{\frac{1 - h_{ii}}{MSE_{(i)}}} \\ &= \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}} \end{aligned}$$

Studentized Deleted Residuals

- Useful for identifying outlying Y observation
  - Test $H_{i0} : E[Y_i] = X_i\beta$ vs $H_{ia} : E[Y_i] \neq X_i\beta$
- If there are no outlying observations,

$$t_i \sim t_{n-1-p}$$

  - can compare t_i to this reference distribution
  - adjust for n tests using Bonferroni
  - an outlier has $|t_i| > t_{1-\alpha/(2n)}(n-1-p)$
  - t_i are not independent

Example on Page 386 (Continued)

```
options nocenter ls=75;

/* INFLUENCE: to report studentized deleted residuals */
proc reg data=a3;
    model insur = sincome sincome2 risk/r influence;
run; quit;
```

Output

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual
1	91.0000	97.8164	0.7181	-6.8164	2.201	-3.097
2	162.0000	160.1201	0.9577	1.8799	2.108	0.892
3	11.0000	11.5901	1.5574	-0.5901	1.713	-0.344
4	240.0000	240.6278	0.8580	-0.6278	2.151	-0.292
5	73.0000	71.5019	0.6656	1.4981	2.218	0.675
6	311.0000	309.6777	1.4363	1.3223	1.816	0.728
7	316.0000	315.6359	2.0100	0.3641	1.150	0.317
8	154.0000	153.3645	0.9829	0.6355	2.096	0.303
9	164.0000	162.4847	0.8211	1.5153	2.165	0.700
10	54.0000	52.4068	0.7346	1.5932	2.196	0.726
11	53.0000	52.8060	0.8340	0.1940	2.160	0.0898
12	326.0000	327.6975	1.4378	-1.6975	1.815	-0.935
13	55.0000	54.4957	0.7142	0.5043	2.203	0.229
14	130.0000	131.0179	1.2720	-1.0179	1.935	-0.526
15	112.0000	109.6080	0.8185	2.3920	2.166	1.104
16	91.0000	93.0992	0.8093	-2.0992	2.169	-0.968
17	14.0000	13.8135	1.2042	0.1865	1.978	0.0943
18	63.0000	62.2363	0.6776	0.7637	2.214	0.345

Output Statistics

Obs	-2-1 0 1 2	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS
1	|******| |	0.255	-5.3155	0.0962	0.0147	-1.7339
2	| |* |	0.041	0.8848	0.1711	1.2842	0.4020
3	| | |	0.025	-0.3333	0.4524	2.3742	-0.3029
4	| | |	0.003	-0.2822	0.1373	1.5215	-0.1126
5	| |* |	0.010	0.6618	0.0826	1.2842	0.1986
6	| |* |	0.083	0.7153	0.3848	1.8735	0.5656
7	| | |	0.077	0.3063	0.7535	5.3027	0.5356
8	| | |	0.005	0.2931	0.1802	1.5981	0.1374
9	| |* |	0.018	0.6866	0.1258	1.3342	0.2604
10	| |* |	0.015	0.7127	0.1006	1.2830	0.2384
11	| | |	0.000	0.0866	0.1297	1.5420	0.0334
12	| *| |	0.137	-0.9308	0.3856	1.6912	-0.7373
13	| | |	0.001	0.2210	0.0951	1.4643	0.0717
14	| *| |	0.030	-0.5120	0.3018	1.7786	-0.3366
15	| |** |	0.044	1.1138	0.1249	1.0675	0.4209
16	| *| |	0.033	-0.9653	0.1222	1.1616	-0.3601
17	| | |	0.001	0.0909	0.2705	1.8390	0.0553
18	| | |	0.003	0.3338	0.0856	1.4216	0.1022

With 18 observations and 3 predictors, the df for the studentized deleted residuals are 13. The P-value associated with Obs #1 is 0.00014. Using Bonferroni, we'd compare this to $0.1/(2*18) = 0.00278$. Conclusion: observation does appear to be unusual.

Identifying Outlying X: Hat Matrix Diagonals

- Diagonals $0 \leq h_{ii} \leq 1$ and sum to p
- Also known as the leverage of ith case
- Is a measure of distance between the X value and the mean of the X values for all n cases $(\overline{X}_1, \overline{X}_2, ..., \overline{X}_{p-1})$
- Since $\hat{\mathbf{Y}} = \mathbf{HY}$

$$\widehat{Y}_i = h_{i1}Y_1 + h_{i2}Y_2 + \ldots + h_{in}Y_n$$

- Thus h_{ii} is a measure of how much Y_i is contributing to the prediction of $\widehat{Y}_i$

Hat Matrix Diagonals

- Residual

$$\begin{aligned} \mathbf{e} &= (\mathbf{I} - \mathbf{H})\mathbf{Y} \\ Var(\mathbf{e}) &= (\mathbf{I} - \mathbf{H})\sigma^2 \\ Var(e_i) &= (1 - h_{ii})\sigma^2 \end{aligned}$$

- Large h_{ii} means small residual variance
  - $\widehat{Y}_i$ will be close to Y_i (i.e., model is forced to fit this observation closely)

- Observations with large h_{ii} considered influential
  - large h_{ii} if it is more than double of the average value, i.e., $h_{ii} > 2p/n$

- Can compute $\mathbf{X}'_{new}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{new}$ to check for hidden extrapolation

Example on Page 386 (Continued)

```
options nocenter ls=75;

proc reg data=a3;
    model insur = sincome sincome2 risk/r influence;
    id sincome risk;
run; quit;
```

Output

Obs	sincome	risk	Hat Diag H	Cov Ratio	DFFITS
1	-0.323145222	6	0.0962	0.0147	-1.7339
2	0.4607431878	4	0.1711	1.2842	0.4020
3	-1.490427964	5	0.4524	2.3742	-0.3029
4	1.0448345518	7	0.1373	1.5215	-0.1126
5	-0.583241377	5	0.0826	1.2842	0.1986
6	1.4759281779	10	0.3848	1.8735	0.5656
7	1.8863221101	1	0.7535	5.3027	0.5356
8	0.175447406	8	0.1802	1.5981	0.1374
9	0.377944412	6	0.1258	1.3342	0.2604
10	-0.765938675	4	0.1006	1.2830	0.2384
11	-0.912636506	6	0.1297	1.5420	0.0334
12	1.6559255166	9	0.3856	1.6912	-0.7373
13	-0.811837997	5	0.0951	1.4643	0.0717
14	0.2789458757	2	0.3018	1.7786	-0.3366
15	-0.24754634	7	0.1249	1.0675	0.4209
16	-0.251146287	4	0.1222	1.1616	-0.3601
17	-1.264531303	3	0.2705	1.8390	0.0553
18	-0.705639567	5	0.0856	1.4216	0.1022

The critical value in this case would be if a diagonal value was greater than $2(4)/18 = 0.44$. It does appear that there are some outlying X observations (Obs #3 and #7). For Obs #7, the largest income and lowest risk. For Obs #3, the smallest income.

Identifying Influential Cases

DFFITS

- Difference between the fitted values $\widehat{Y}_i$ and the predicted values $\widehat{Y}_{i(i)}$
- Measures influence of case i on $\widehat{Y}_i$

$$\text{DFFITS}_i = \frac{\widehat{Y}_i - \widehat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

  – The denominator is the estimated standard deviation of $\widehat{Y}_i$, obtained using the $MSE_{(i)}$

  – t_i is the studentized deleted residual

- Adjusts studentized deleted residual by function of h_{ii}
- Concern if absolute value greater than 1 for small data sets, or greater than $2\sqrt{p/n}$ for large data sets

Output

Obs	sincome	risk	Hat Diag H	Cov Ratio	DFFITS
1	-0.323145222	6	0.0962	0.0147	-1.7339
2	0.4607431878	4	0.1711	1.2842	0.4020
3	-1.490427964	5	0.4524	2.3742	-0.3029
4	1.0448345518	7	0.1373	1.5215	-0.1126
5	-0.583241377	5	0.0826	1.2842	0.1986
6	1.4759281779	10	0.3848	1.8735	0.5656
7	1.8863221101	1	0.7535	5.3027	0.5356
8	0.175447406	8	0.1802	1.5981	0.1374
9	0.377944412	6	0.1258	1.3342	0.2604
10	-0.765938675	4	0.1006	1.2830	0.2384
11	-0.912636506	6	0.1297	1.5420	0.0334
12	1.6559255166	9	0.3856	1.6912	-0.7373
13	-0.811837997	5	0.0951	1.4643	0.0717
14	0.2789458757	2	0.3018	1.7786	-0.3366
15	-0.24754634	7	0.1249	1.0675	0.4209
16	-0.251146287	4	0.1222	1.1616	-0.3601
17	-1.264531303	3	0.2705	1.8390	0.0553
18	-0.705639567	5	0.0856	1.4216	0.1022

This is a small data set, so we'll be concerned about values greater than 1 in scale. In this case, Obs #1 has strong influence. Recall this observation had a very large studentized deleted residual. None of the others are a concern.

Cook's Distance

- Measures influence of a case on all $\widehat{Y}_i$'s
- Standardized version of sum of squared differences between fitted values with and without case i

$$D_i = \frac{\sum_{j=1}^{n} (\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{p \cdot \mathrm{MSE}}$$

  - can be obtained in a single fit
- Compare with $F(p, n-p)$
- Concern if D_i is above the 50%-tile of $F(p, n-p)$

Output

```
                      Output Statistics
                                              Cook's
Obs          sincome risk   -2-1 0 1 2             D   RStudent
  1     -0.323145222    6 |******|      |      0.255    -5.3155
  2     0.4607431878    4 |      |*     |      0.041     0.8848
  3     -1.490427964    5 |      |      |      0.025    -0.3333
  4     1.0448345518    7 |      |      |      0.003    -0.2822
  5     -0.583241377    5 |      |*     |      0.010     0.6618
  6     1.4759281779   10 |      |*     |      0.083     0.7153
  7     1.8863221101    1 |      |      |      0.077     0.3063
  8      0.175447406    8 |      |      |      0.005     0.2931
  9      0.377944412    6 |      |*     |      0.018     0.6866
 10     -0.765938675    4 |      |*     |      0.015     0.7127
 11     -0.912636506    6 |      |      |      0.000     0.0866
 12     1.6559255166    9 |     *|      |      0.137    -0.9308
 13     -0.811837997    5 |      |      |      0.001     0.2210
 14     0.2789458757    2 |     *|      |      0.030    -0.5120
 15      -0.24754634    7 |      |**    |      0.044     1.1138
 16     -0.251146287    4 |     *|      |      0.033    -0.9653
 17     -1.264531303    3 |      |      |      0.001     0.0909
 18     -0.705639567    5 |      |      |      0.003     0.3338
```

With 18 observations and 3 predictors, the df for the F are 4 and 14. The 30, 40, and 50%-tiles are 0.553, 0.707, and 0.881 respectively. None of the observations appear to have an undue amount of influence.

DFBETAS

- Measures influence of case i on each of the regression coefficients
- Standardized version of the difference between regression coefficient computed with and without case i

$$\text{DFBETAS}_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{\text{MSE}_{(i)} c_{kk}}}$$

where c_{kk} is the k-th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$

  - The denominator is the SE of β_k, obtained using $MSE_{(i)}$ to estimate σ^2

- Concern if greater than 1 for small data sets or greater than $2/\sqrt{n}$ for large data sets

Output

Output Statistics

			DFBETAS			
Obs	sincome	risk	Intercept	sincome	sincome2	risk
1	-0.323145222	6	-0.4440	0.0662	0.9168	-0.3686
2	0.4607431878	4	0.3372	0.2513	-0.2579	-0.2064
3	-1.490427964	5	0.0874	0.2513	-0.2312	-0.0525
4	1.0448345518	7	-0.0067	-0.0692	0.0230	-0.0299
5	-0.583241377	5	0.0831	-0.0566	-0.0580	-0.0108
6	1.4759281779	10	-0.3129	0.1183	0.1704	0.3901
7	1.8863221101	1	0.2554	0.2235	0.2233	-0.3381
8	0.175447406	8	-0.0162	0.0245	-0.0712	0.0788
9	0.377944412	6	0.1121	0.1333	-0.1799	0.0084
10	-0.765938675	4	0.1267	-0.0988	-0.0084	-0.0773
11	-0.912636506	6	-0.0064	-0.0244	0.0091	0.0126
12	1.6559255166	9	0.3453	-0.1728	-0.3486	-0.3821
13	-0.811837997	5	0.0137	-0.0427	0.0063	0.0030
14	0.2789458757	2	-0.3279	-0.1746	0.1861	0.2583
15	-0.24754634	7	-0.0046	-0.0195	-0.2036	0.2003
16	-0.251146287	4	-0.2937	-0.0774	0.2177	0.1654
17	-1.264531303	3	0.0101	-0.0383	0.0317	-0.0150
18	-0.705639567	5	0.0310	-0.0471	-0.0097	-0.0003

Nothing looks real troubling here except for Obs #1 and its influence on the quadratic coefficient. Since this had such a large residual, we will remove it and refit the model.

Analysis of Variance

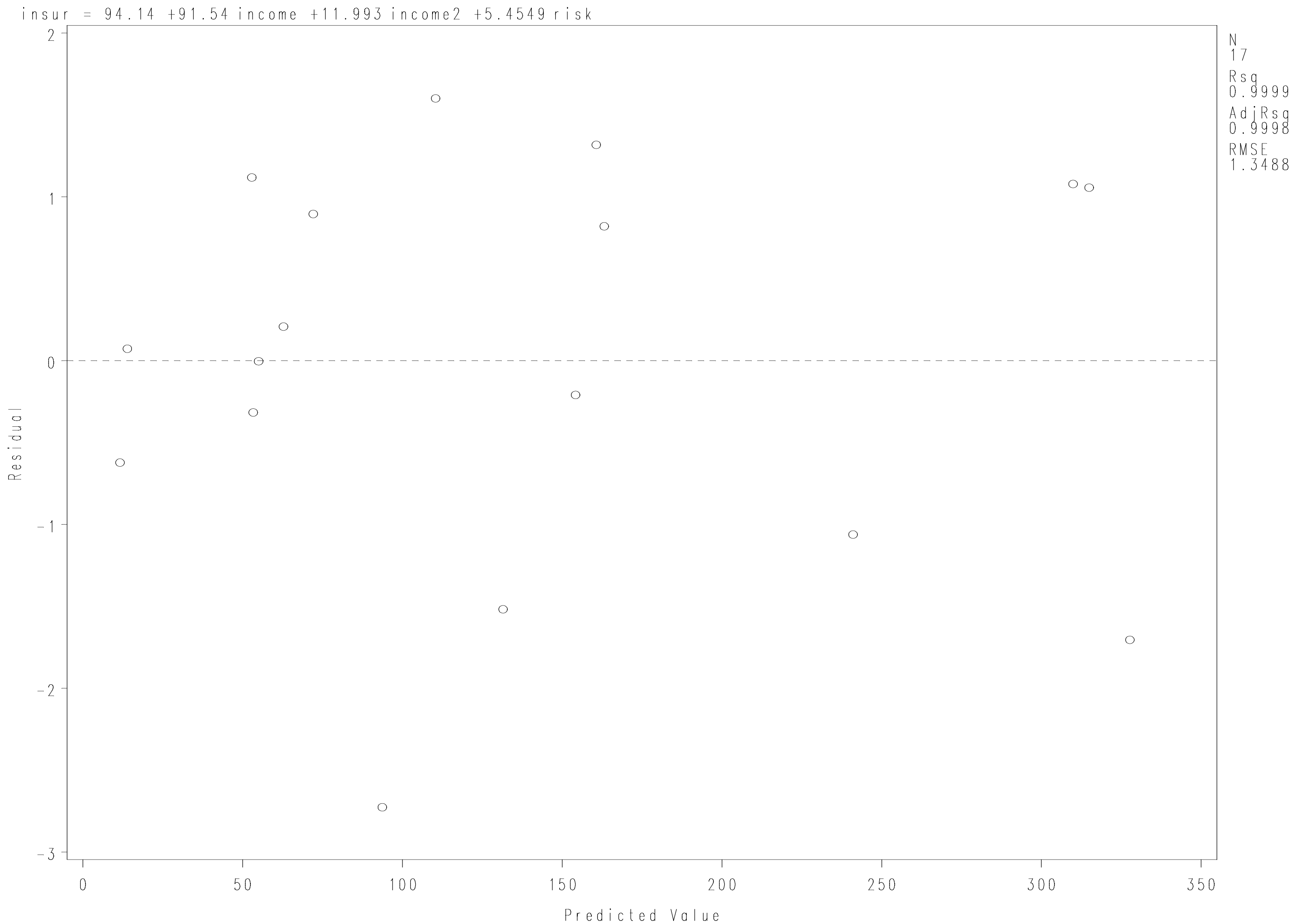
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	174302	58101	31934.2	<.0001
Error	13	23.65205	1.81939		
Corrected Total	16	174326			

Root MSE	1.34885	R-Square	0.9999
Dependent Mean	137.00000	Adj R-Sq	0.9998
Coeff Var	0.98456		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	94.14049	0.95577	98.50	<.0001
sincome	1	91.54004	0.38073	240.43	<.0001
sincome2	1	11.99324	0.34902	34.36	<.0001
risk	1	5.45493	0.14831	36.78	<.0001

Residual Plot

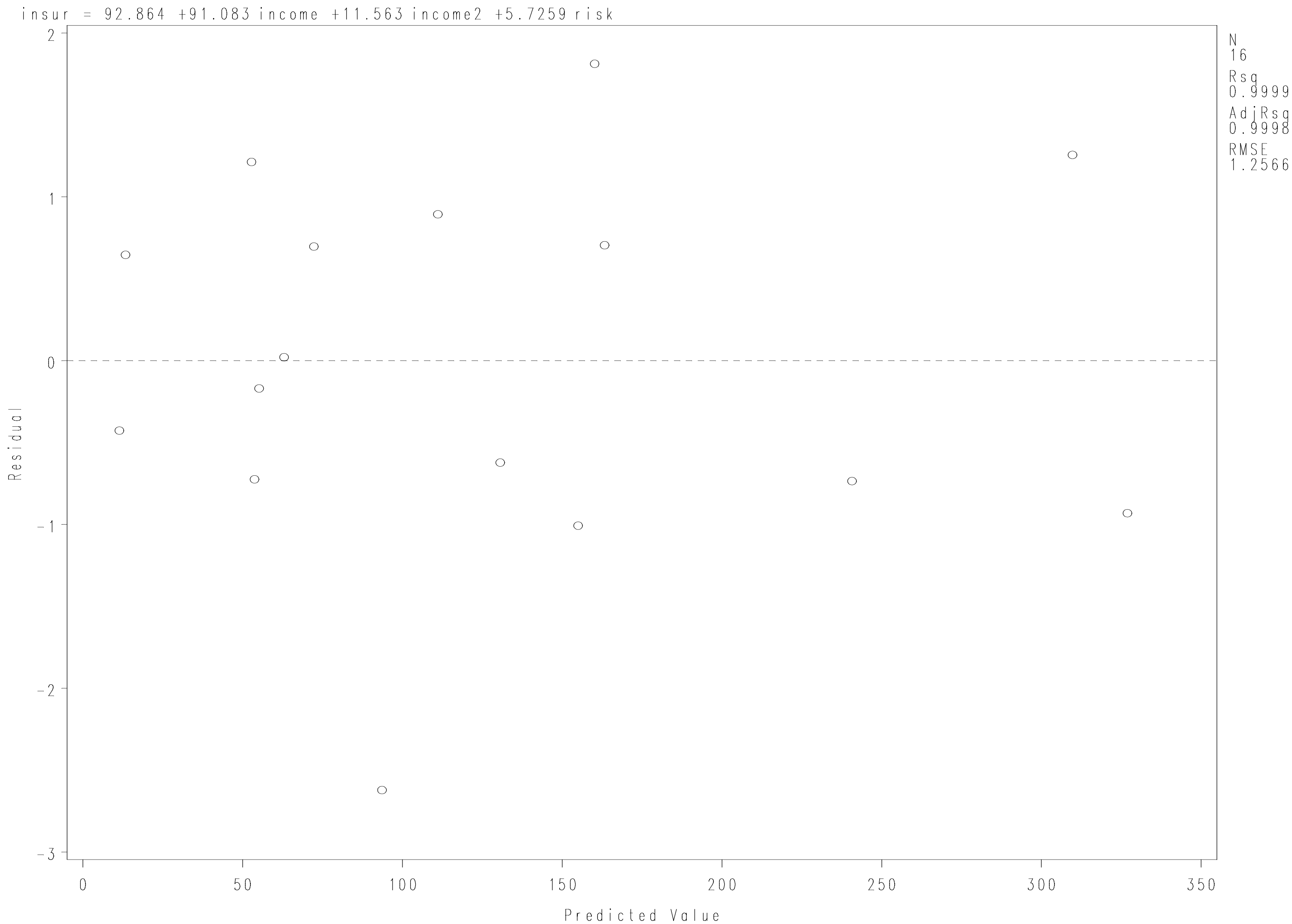


Output

			DFBETAS			
Obs	income	risk	Intercept	income	income2	risk
1	0.4607431878	4	0.4210	0.3079	-0.3285	-0.2467
2	-1.490427964	5	0.1587	0.4590	-0.4154	-0.0960
3	1.0448345518	7	-0.0246	-0.2058	0.0768	-0.0927
4	-0.583241377	5	0.0906	-0.0592	-0.0692	-0.0069
5	1.4759281779	10	-0.4422	0.1685	0.2325	0.5595
6	1.8863221101	1	1.4336	1.2882	1.3223	-1.9612
7	0.175447406	8	0.0074	-0.0138	0.0439	-0.0465
8	0.377944412	6	0.1100	0.1238	-0.1770	0.0124
9	-0.765938675	4	0.1591	-0.1213	-0.0204	-0.0899
10	-0.912636506	6	0.0163	0.0690	-0.0221	-0.0367
11	1.6559255166	9	0.6402	-0.3214	-0.6388	-0.7095
12	-0.811837997	5	-0.0002	0.0006	-0.0000	-0.0001
13	0.2789458757	2	-0.9070	-0.4778	0.5234	0.6995
14	-0.24754634	7	0.0076	-0.0251	-0.2646	0.2479
15	-0.251146287	4	-0.8138	-0.2068	0.6230	0.4303
16	-1.264531303	3	0.0068	-0.0254	0.0205	-0.0098
17	-0.705639567	5	0.0155	-0.0221	-0.0066	0.0007

Now Obs #6 is influential. This was Obs #7 before we discarded the first observation. It would be worth investigating how much the model changes with and without this observation.

Residual Plot



Multicollinearity Diagnostics: VIF

- Use **V**ariance **I**nflation **F**actor (VIF) for quantitative diagnostic of multicollinearity
- VIF_k is the the kth diagonal element of r_{XX}^{-1}

$$\text{VIF}_k = (r_{XX}^{-1})_{kk} = \frac{1}{1 - R_k^2}$$

  – where R_k^2 is the coefficient of multiple determination of X_k regressed versus all other $p - 2$ variables.

- In standardized regression

$$Var(\mathbf{b}^*) \;=\; (\sigma^*)^2 \; \mathbf{r}_{X'X}^{-1}$$

$$Var(\mathbf{b}_k^*) \;=\; (\sigma^*)^2 (r_{XX}^{-1})_{kk} = (\sigma^*)^2 \; \text{VIF}_k$$

  – the larger VIF_k, the larger the variance of the estimated coefficient

- VIF of 10 or more suggests strong multicollinearity
- Also compare mean VIF to 1

$$\overline{\text{VIF}} = \frac{\sum \text{VIF}_k}{p-1}$$

  - $\overline{\text{VIF}}$ considerably larger than one indicates serious multicollinearity problems

- Tolerance(TOL) = 1/VIF
  - SAS gives TOL results for each predictor
  - Trouble if TOL $< .1$

Example on Page 386 (Continued)

```
options nocenter ls=75;

/* TOL: tolerance*/
proc reg data=a3;
    model insur = sincome sincome2 risk/tol;
    id sincome risk;
run; quit;
```

Output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	174302	58101	31934.2	<.0001
Error	13	23.65205	1.81939		
Corrected Total	16	174326			

Root MSE	1.34885	R-Square	0.9999
Dependent Mean	137.00000	Adj R-Sq	0.9998
Coeff Var	0.98456		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance
Intercept	1	94.14049	0.95577	98.50	<.0001	.
sincome	1	91.54004	0.38073	240.43	<.0001	0.74314
sincome2	1	11.99324	0.34902	34.36	<.0001	0.79731
risk	1	5.45493	0.14831	36.78	<.0001	0.92021

Chapter Review

- Partial Regression Plots to Check Model Adequacy for a Predictor Variable
- Identifying Outlying Y Observations
  - Use studentized deleted residuals
- Identifying Outlying X's
- Identifying Influential Cases
  - DFFITS, DFBETAS
  - Cook's Distance
- Multicollinearity Diagnostics Using Variance Inflation Factor