Topic 14 - Regression Diagnostics

STAT 525 - Fall 2013

Outline

- Partial Regression Plots
- Residuals
  - Studentized
  - Studentized Deleted
- Identifying outlying X's
- Identifying influential cases

Example Page 386

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

Partial Regression Plots

- Also called added variable plots or adjusted variable plots
- Recall partial correlation / coefficient of determination
- These provide visual display of that relationship
- One plot for each X_i
- Allows check of "adjusted" relationship

Partial Regression Plots

- Procedure for X_i vs Y
  - Predict Y using other X's
  - Predict X_i using other X's
  - Plot residuals from first regression vs residuals from the second regression
- Shows the **strength** of the marginal relationship in terms of the **full** model
- Can detect:
  - Nonlinear relationship
  - Heterogeneous variance
  - Unusual observations

SAS Commands

```
proc reg data=a1;
  model insur=income risk/r partial influence tol;
  id income risk; plot r.*(p. risk income);

proc reg data=a1;
   model insur risk = income;          ************************
   output out=a2 r=resins resris;      Generates added variable
proc sort data=a2; by resris;          plot for risk. Similar
proc gplot data=a2;                    approach for income
   plot resins*resris;                 ************************
proc reg data=a2; model resins = resris;
   output out=new1 r=res p=pred;
proc gplot data=new1;
   plot res*pred;
run;
```


Output

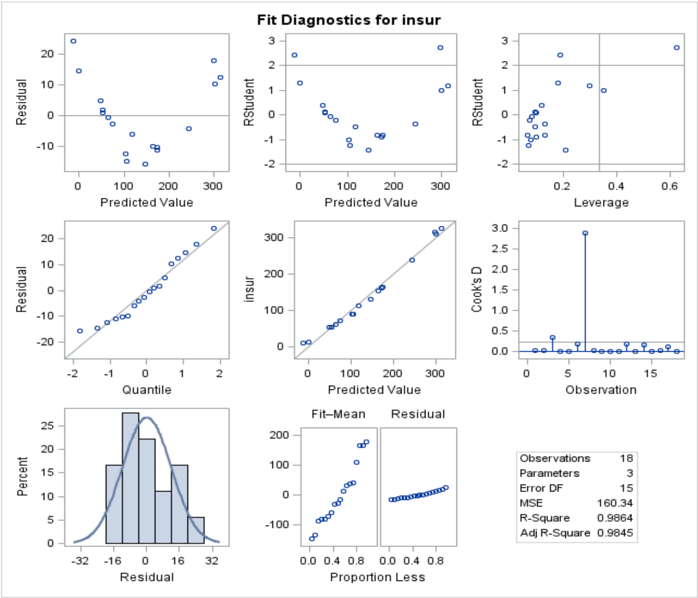
Analysis of Variance

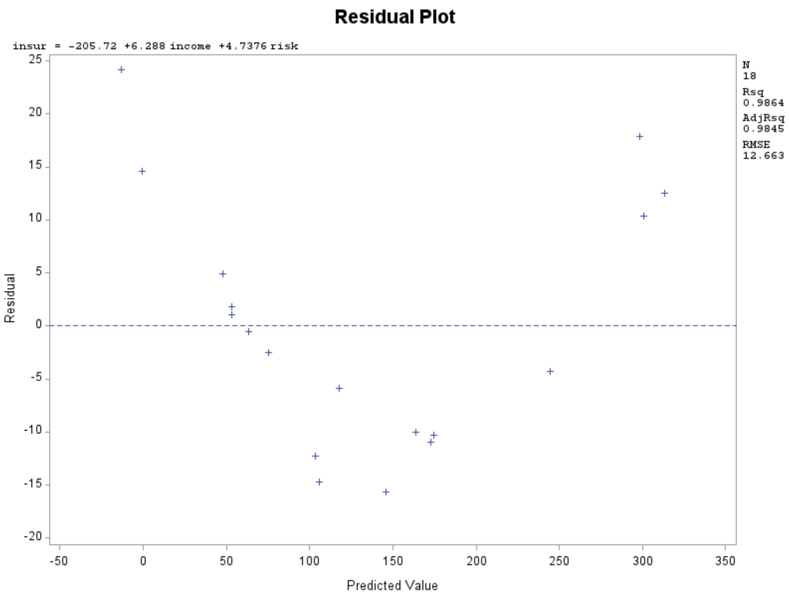
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	173919	86960	542.33	<.0001
Error	15	2405.14763	160.34318		
Corrected Total	17	176324			

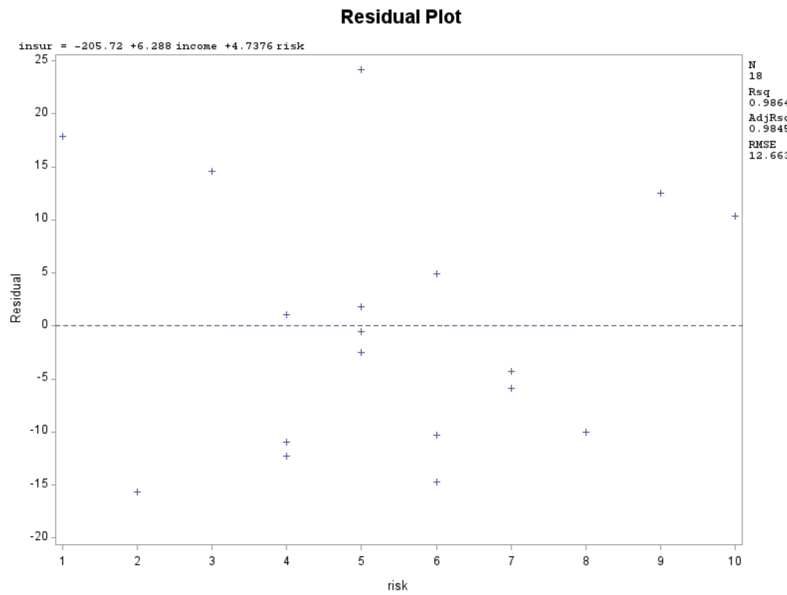
Root MSE	12.66267	R-Square	0.9864
Dependent Mean	134.44444	Adj R-Sq	0.9845
Coeff Var	9.41851		

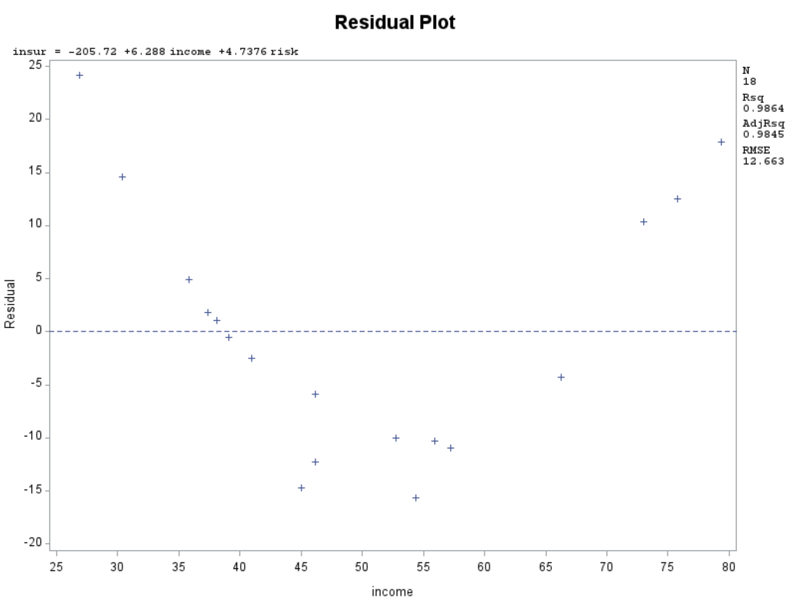
Parameter Estimates

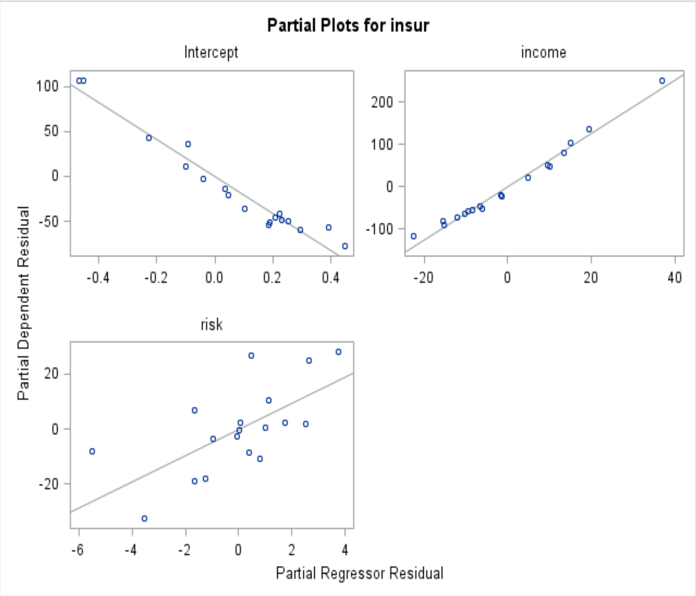
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-205.71866	11.39268	-18.06	<.0001
income	1	6.28803	0.20415	30.80	<.0001
risk	1	4.73760	1.37808	3.44	0.0037




Residual Plot
insur = -205.72 +6.288 income +4.7376 risk
N
18
Rsq
0.9864
AdjRsq
0.9845
RMSE
12.663
Residual
Predicted Value


Residual Plot
insur = -205.72 +6.288 income +4.7376 risk
N
18
Rsq
0.9864
AdjRsq
0.9845
RMSE
12.663
Residual
risk


Residual Plot
insur = -205.72 +6.288 income +4.7376 risk
N
18
Rsq
0.9864
AdjRsq
0.9845
RMSE
12.663
Residual
income


Partial Plots for insur
Intercept
income
risk
Partial Dependent Residual
Partial Regressor Residual

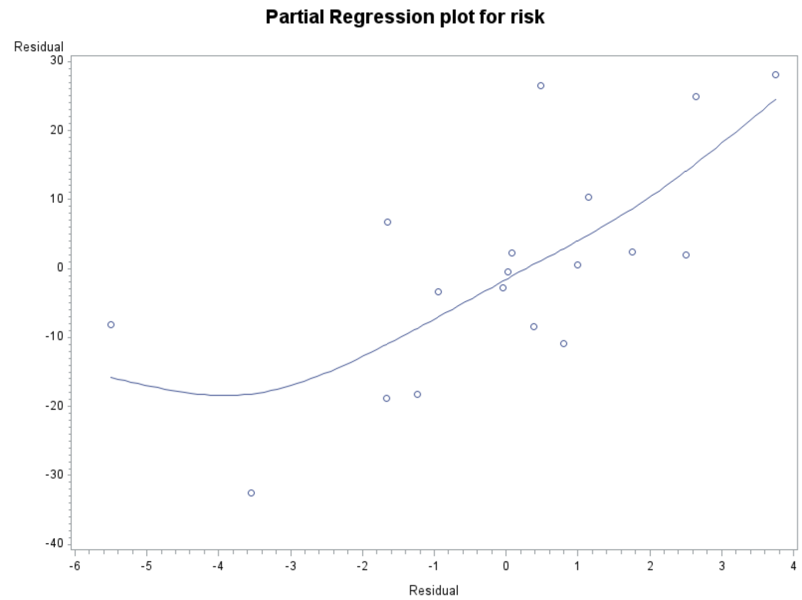
Output

```
                 Analysis of Variance
                         Sum of         Mean
Source             DF     Squares       Square  F Value   Pr > F
Model               1  1895.04339  1895.04339     12.61   0.0027
Error              16  2405.14763   150.32173
Corrected Total    17  4300.19102

Root MSE             12.26058    R-Square     0.4407 **Partial R-Square
Dependent Mean     -1.204E-14    Adj R-Sq     0.4057
Coeff Var         -1.01834E17

                Parameter Standard
Variable   DF    Estimate    Error  t Value  Pr > |t|
Intercept   1 -9.4683E-15  2.88985    -0.00    1.0000
resris      1     4.73760  1.33432     3.55    0.0027
```

**Note that the parameter estimate for the slope is the same as the parameter estimate for RISK in the full model



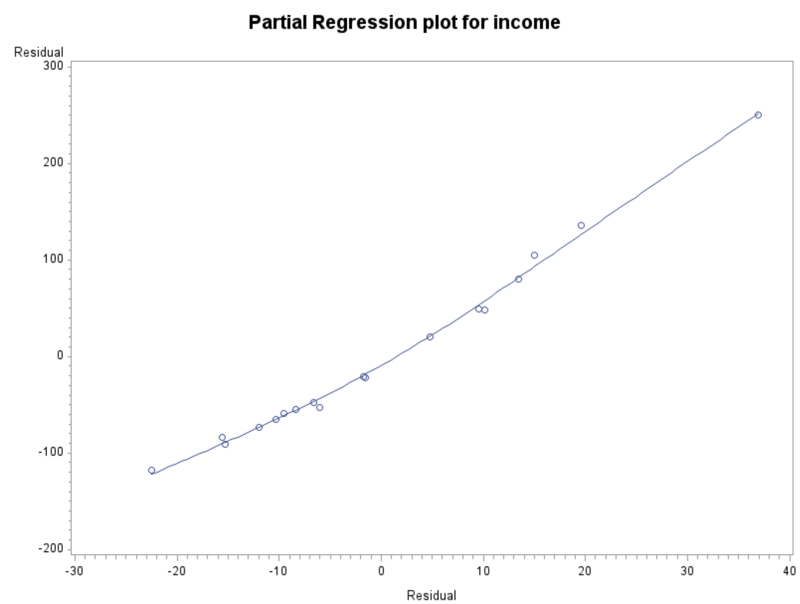
Output

```
                 Analysis of Variance
                         Sum of        Mean
Source             DF     Squares      Square  F Value   Pr > F
Model               1      152119      152119  1011.96   <.0001
Error              16  2405.14763   150.32173
Corrected Total    17      154524

Root MSE             12.26058    R-Square     0.9844  **Partial R-square
Dependent Mean    -6.3159E-15    Adj R-Sq     0.9835
Coeff Var         -1.94121E17

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

**Note that the parameter estimate for the slope is the same as the parameter estimate for INCOME in the full model



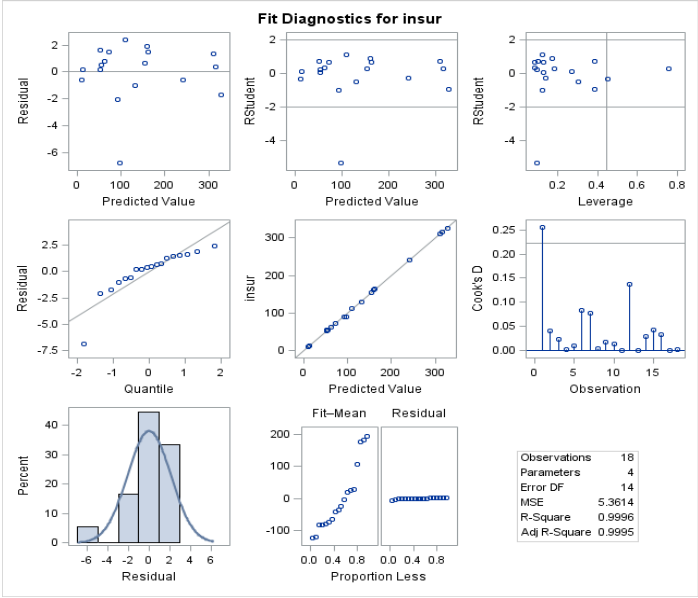
Output - Adding Income^2

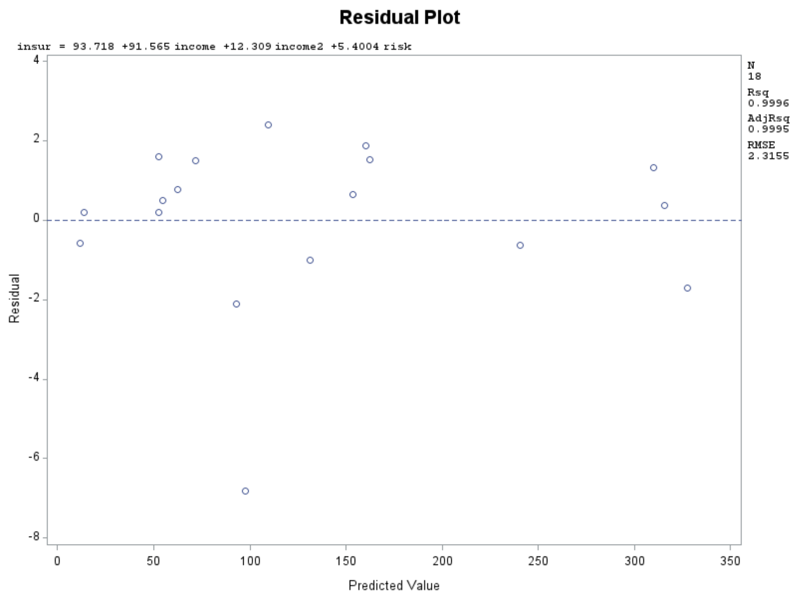
Analysis of Variance

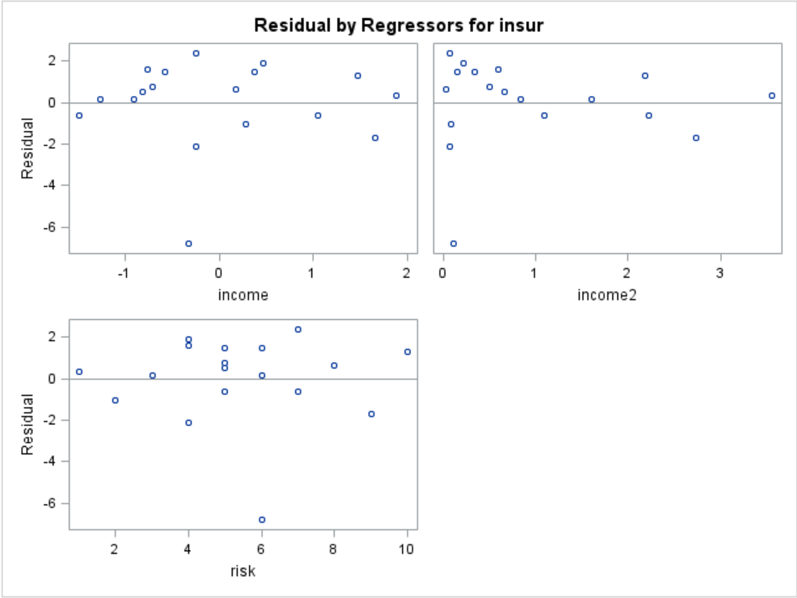
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	176249	58750	10958.0	<.0001
Error	14	75.05895	5.36135		
Corrected Total	17	176324			

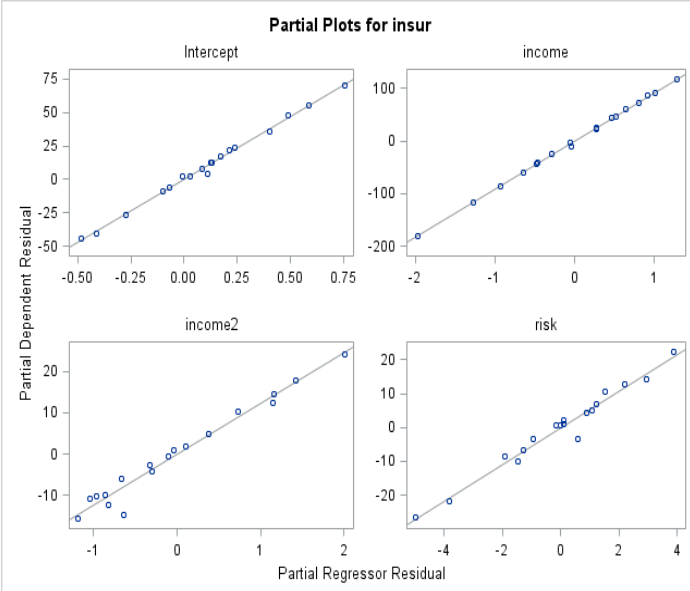
Root MSE	2.31546	R-Square	0.9996
Dependent Mean	134.44444	Adj R-Sq	0.9995
Coeff Var	1.72224		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	93.71759	1.63501	57.32	<.0001
income	1	91.56523	0.65352	140.11	<.0001
income2	1	12.30855	0.59042	20.85	<.0001
risk	1	5.40039	0.25399	21.26	<.0001









Residuals

- Standard residual

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

- Studentized residual

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

- Studentized deleted residual

$$\begin{aligned} t_i &= \frac{Y_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(\text{i})}/(1 - \text{h}_{\text{ii}})}} \sim t(n-p-1) \\ &= e_i \left[\frac{n-p-1}{\text{SSE}(1 - h_{ii}) - e_i^2} \right]^{1/2} \end{aligned}$$

Studentized Deleted Residual

- Can express deleted residual as

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

- Based on following identity (Gauss 1821)

$$\left(\mathbf{X}'_{(\text{i})}\mathbf{X}_{(\text{i})}\right)^{-1} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1} X_i X_i' (\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}}$$

- Relationship between $\text{MSE}_{(i)}$ and MSE

$$(n-p)\text{MSE} = (n-p-1)\text{MSE}_{(i)} + \frac{e_i^2}{1 - h_{ii}}$$

Output

Output Statistics

Obs	income	risk	Residual	Std Error Residual	Student Residual
1	-0.323145222	6	-6.8164	2.201	-3.097
2	0.4607431878	4	1.8799	2.108	0.892
3	-1.490427964	5	-0.5901	1.713	-0.344
4	1.0448345518	7	-0.6278	2.151	-0.292
5	-0.583241377	5	1.4981	2.218	0.675
6	1.4759281779	10	1.3223	1.816	0.728
7	1.8863221101	1	0.3641	1.150	0.317
8	0.175447406	8	0.6355	2.096	0.303
9	0.377944412	6	1.5153	2.165	0.700
10	-0.765938675	4	1.5932	2.196	0.726
11	-0.912636506	6	0.1940	2.160	0.0898
12	1.6559255166	9	-1.6975	1.815	-0.935
13	-0.811837997	5	0.5043	2.203	0.229
14	0.2789458757	2	-1.0179	1.935	-0.526
15	-0.24754634	7	2.3920	2.166	1.104
16	-0.251146287	4	-2.0992	2.169	-0.968
17	-1.264531303	3	0.1865	1.978	0.0943
18	-0.705639567	5	0.7637	2.214	0.345

Output

```
                                  Output Statistics
                                                              Cook's
Obs income             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 studentized deleted residuals are 13. The P-value associated with Obs #1 is 0.00014. Using Bonferroni, we'd compare this to .05/18 = 0.00278. Conclusion: observation does appear to be unusual.

Hat Matrix Diagonals

- Used to identify outlying X observations
- 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 $\hat{\mathbf{Y}} = \mathbf{HY}$

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

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

Hat Matrix Diagonals

- Residual $e_i = (1 - h_{ii})Y_i$
- Var(e_i)=$\sigma^2(1 - h_{ii})$
- If h_{ii} large, small residual variance
- This implies $\hat{Y}_i$ will be close to Y_i (i.e., model forced to fit observation closely)
- Look for large h_{ii} : usually considered large if it is more than double the mean leverage value (p/n)
- When predicting, can compute $\mathbf{X}'_{\mathbf{i(new)}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{\mathbf{i(new)}}$ to see if there is the danger of extrapolating

Output

Obs	income	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.

DFFITS

- Measures influence of case i on $\hat{Y}_i$
- Closely related to h_{ii}

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

- 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	income	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 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 case i on <u>all</u> $\hat{Y}_i$'s
- Standardized version of sum of squared differences between fitted values with and without case i

$$D_i = \frac{\sum (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\text{MSE}}$$

- Compare with $F(p, n-p)$
- Concern if D_i above the 50%-tile
- Can show $D_i = \text{MSE}_{(i)}(\text{DFFITS}_i)^2/(p\text{MSE})$...Thus another cutoff is $4/n$ provided $\text{MSE}_{(i)}/\text{MSE} \approx 1$

Output

```
                         Output Statistics
                                                      Cook's
Obs income           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 <u>each</u> 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 from $(\mathbf{X}'\mathbf{X})^{-1}$

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

Output

Output Statistics

			DFBETAS			
Obs	income	risk	Intercept	income	income2	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.

Output

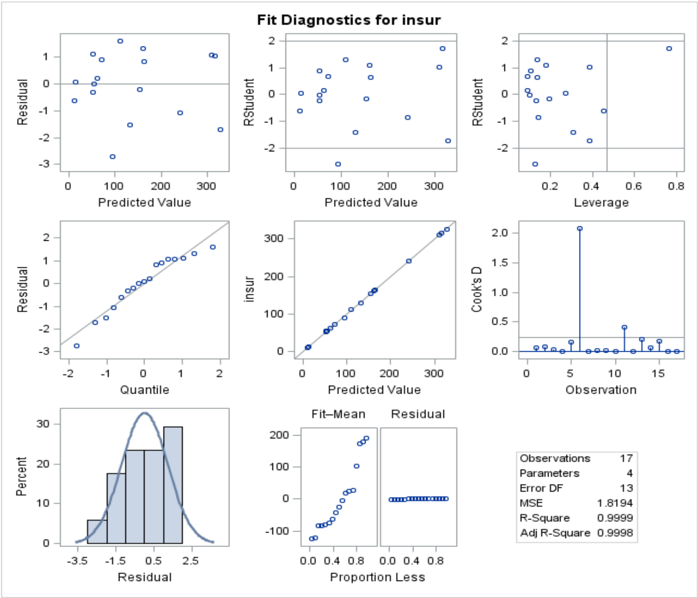
Analysis of Variance

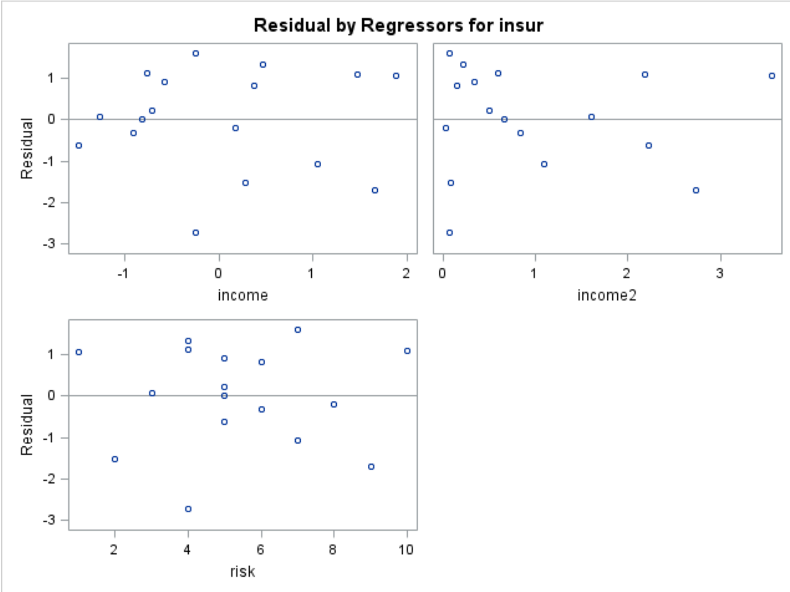
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	174302	58101	31934.2	<.0001
Error	13	23.65205	1.81939		
Corrected Total	16	174326			

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|
Intercept	1	94.14049	0.95577	98.50	<.0001
income	1	91.54004	0.38073	240.43	<.0001
income2	1	11.99324	0.34902	34.36	<.0001
risk	1	5.45493	0.14831	36.78	<.0001





Output

Obs	income	risk	Intercept	income	income2	risk
			DFBETAS			
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.

Variance Inflation Factor

- The VIF is related to the variance of the estimated regression coefficients
- Looks at standardized model using correlation transformation
- Can show $\boldsymbol{\sigma}^2(\mathbf{b}) = (\sigma')^2 r_{XX}^{-1}$
- VIF_k is the the kth diagonal element of r_{XX}^{-1}
- Can show $\mathrm{VIF}_k = (1 - R_k^2)^{-1}$

Variance Inflation Factor

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

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

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

Output

```
                  Analysis of Variance

                       Sum of      Mean
Source           DF   Squares    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 Standard
Variable  DF  Estimate    Error t Value Pr > |t| Tolerance

Intercept  1  94.14049  0.95577   98.50  <.0001         .
income     1  91.54004  0.38073  240.43  <.0001  0.74314
income2    1  11.99324  0.34902   34.36  <.0001  0.79731
risk       1   5.45493  0.14831   36.78  <.0001  0.92021
```


Background Reading

- KNNL Chapter 10
- knnl386.sas
- KNNL Sections 11.1, 11.5, 11.6