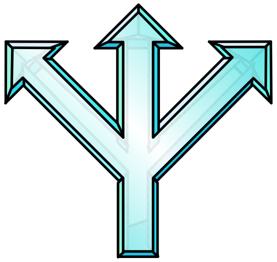
Statistical Techniques II

EXST7015

Multiple Regression (Part 2)



Multiple Regression

- **Much of the remaining aspects of multiple regression pertain to "observation diagnostics" and tests of the assumptions. These were mostly covered with simple linear regression, (SLR) but will be reviewed here.**
- **One difference from SLR is that where previously we used X_i we now use Yhat. For example, residual plots are usually plotted on Yhat.**

Observation Diagnostics

- **Continuing with the Plant available phosphorus (MLR Example 1).**
- **Below is a listing of some observation diagnostics (see handout).**

Output Statistics

<table>
<tr><th>Obs</th><th>Dep Var Y</th><th>Predicted Value</th><th>Std Error Mean Predict</th><th colspan="2">95% CL Mean</th><th colspan="2">95% CL Predict</th><th>Residual</th><th>Std Error Residual</th><th>Student Residual</th><th>-2-1 0 1 2</th></tr>
<tr><td>1</td><td>64.0000</td><td>65.4050</td><td>10.5752</td><td>42.7235</td><td>88.0865</td><td>16.9378</td><td>113.8723</td><td>-1.4050</td><td>16.941</td><td>-0.0829</td><td>| | |</td></tr>
<tr><td>2</td><td>60.0000</td><td>68.7126</td><td>11.0062</td><td>45.1068</td><td>92.3185</td><td>19.8060</td><td>117.6192</td><td>-8.7126</td><td>16.664</td><td>-0.523</td><td>| *| |</td></tr>
<tr><td>3</td><td>71.0000</td><td>53.5624</td><td>11.6962</td><td>28.4766</td><td>78.6481</td><td>3.9245</td><td>103.2002</td><td>17.4376</td><td>16.187</td><td>1.077</td><td>| |** |</td></tr>
<tr><td>4</td><td>61.0000</td><td>67.1854</td><td>8.6364</td><td>48.6623</td><td>85.7085</td><td>20.5193</td><td>113.8515</td><td>-6.1854</td><td>18.007</td><td>-0.344</td><td>| | |</td></tr>
<tr><td>5</td><td>54.0000</td><td>59.5460</td><td>9.4223</td><td>39.3371</td><td>79.7548</td><td>12.1854</td><td>106.9065</td><td>-5.5460</td><td>17.608</td><td>-0.315</td><td>| | |</td></tr>
<tr><td>6</td><td>77.0000</td><td>61.0848</td><td>13.7970</td><td>31.4932</td><td>90.6765</td><td>9.0244</td><td>113.1453</td><td>15.9152</td><td>14.438</td><td>1.102</td><td>| |** |</td></tr>
<tr><td>7</td><td>81.0000</td><td>64.1718</td><td>10.1592</td><td>42.3825</td><td>85.9610</td><td>16.1156</td><td>112.2279</td><td>16.8282</td><td>17.193</td><td>0.979</td><td>| |* |</td></tr>
<tr><td>8</td><td>93.0000</td><td>77.9457</td><td>6.2695</td><td>64.4989</td><td>91.3925</td><td>33.0521</td><td>122.8393</td><td>15.0543</td><td>18.961</td><td>0.794</td><td>| |* |</td></tr>
<tr><td>9</td><td>93.0000</td><td>89.8131</td><td>9.9864</td><td>68.3944</td><td>111.2318</td><td>41.9238</td><td>137.7024</td><td>3.1869</td><td>17.294</td><td>0.184</td><td>| | |</td></tr>
<tr><td>10</td><td>51.0000</td><td>79.3503</td><td>8.3893</td><td>61.3571</td><td>97.3435</td><td>32.8919</td><td>125.8086</td><td>-28.3503</td><td>18.123</td><td>-1.564</td><td>| ***| |</td></tr>
<tr><td>11</td><td>76.0000</td><td>77.9063</td><td>5.1132</td><td>66.9397</td><td>88.8730</td><td>33.6922</td><td>122.1205</td><td>-1.9063</td><td>19.305</td><td>-0.0987</td><td>| | |</td></tr>
<tr><td>12</td><td>96.0000</td><td>99.4135</td><td>7.3614</td><td>83.6249</td><td>115.2021</td><td>53.7637</td><td>145.0632</td><td>-3.4135</td><td>18.564</td><td>-0.184</td><td>| | |</td></tr>
<tr><td>13</td><td>77.0000</td><td>102.3025</td><td>7.1198</td><td>87.0320</td><td>117.5730</td><td>56.8294</td><td>147.7757</td><td>-25.3025</td><td>18.658</td><td>-1.356</td><td>| **| |</td></tr>
<tr><td>14</td><td>93.0000</td><td>90.2967</td><td>8.9528</td><td>71.0949</td><td>109.4985</td><td>43.3570</td><td>137.2363</td><td>2.7033</td><td>17.851</td><td>0.151</td><td>| | |</td></tr>
<tr><td>15</td><td>95.0000</td><td>107.2807</td><td>8.3375</td><td>89.3984</td><td>125.1629</td><td>60.8652</td><td>153.6961</td><td>-12.2807</td><td>18.147</td><td>-0.677</td><td>| *| |</td></tr>
<tr><td>16</td><td>54.0000</td><td>67.0830</td><td>7.3513</td><td>51.3159</td><td>82.8500</td><td>21.4406</td><td>112.7253</td><td>-13.0830</td><td>18.568</td><td>-0.705</td><td>| *| |</td></tr>
<tr><td>17</td><td>168.0000</td><td>119.1961</td><td>11.2752</td><td>95.0132</td><td>143.3789</td><td>70.0083</td><td>168.3838</td><td>48.8039</td><td>16.483</td><td>2.961</td><td>| |*****|</td></tr>
<tr><td>18</td><td>99.0000</td><td>112.7443</td><td>9.8642</td><td>91.5877</td><td>133.9009</td><td>64.9717</td><td>160.5169</td><td>-13.7443</td><td>17.364</td><td>-0.792</td><td>| *| |</td></tr>
</table>

Obs. Diagnostics *(continued)*

- The first columns are the observed value of Y_i and the predicted value, $Yhat_i$.
- These are followed by the "Std Error of the Mean Predict" value. You are not responsible for this value.

Output Statistics

Obs	Dep Var Y	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual	Std Error Residual	Student Residual	-2-1 0 1 2
1	64.0000	65.4050	10.5752	42.7235	88.0865	16.9378	113.8723	-1.4050	16.941	-0.0829	| | |
2	60.0000	68.7126	11.0062	45.1068	92.3185	19.8060	117.6192	-8.7126	16.664	-0.523	| *| |
3	71.0000	53.5624	11.6962	28.4766	78.6481	3.9245	103.2002	17.4376	16.187	1.077	| |** |
4	61.0000	67.1854	8.6364	48.6623	85.7085	20.5193	113.8515	-6.1854	18.007	-0.344	| | |
5	54.0000	59.5460	9.4223	39.3371	79.7548	12.1854	106.9065	-5.5460	17.608	-0.315	| | |
6	77.0000	61.0848	13.7970	31.4932	90.6765	9.0244	113.1453	15.9152	14.438	1.102	| |** |
7	81.0000	64.1718	10.1592	42.3825	85.9610	16.1156	112.2279	16.8282	17.193	0.979	| |* |
8	93.0000	77.9457	6.2695	64.4989	91.3925	33.0521	122.8393	15.0543	18.961	0.794	| |* |
9	93.0000	89.8131	9.9864	68.3944	111.2318	41.9238	137.7024	3.1869	17.294	0.184	| | |
10	51.0000	79.3503	8.3893	61.3571	97.3435	32.8919	125.8086	-28.3503	18.123	-1.564	| ***| |
11	76.0000	77.9063	5.1132	66.9397	88.8730	33.6922	122.1205	-1.9063	19.305	-0.0987	| | |
12	96.0000	99.4135	7.3614	83.6249	115.2021	53.7637	145.0632	-3.4135	18.564	-0.184	| | |
13	77.0000	102.3025	7.1198	87.0320	117.5730	56.8294	147.7757	-25.3025	18.658	-1.356	| **| |
14	93.0000	90.2967	8.9528	71.0949	109.4985	43.3570	137.2363	2.7033	17.851	0.151	| | |
15	95.0000	107.2807	8.3375	89.3984	125.1629	60.8652	153.6961	-12.2807	18.147	-0.677	| *| |
16	54.0000	67.0830	7.3513	51.3159	82.8500	21.4406	112.7253	-13.0830	18.568	-0.705	| *| |
17	168.0000	119.1961	11.2752	95.0132	143.3789	70.0083	168.3838	48.8039	16.483	2.961	| |***** |
18	99.0000	112.7443	9.8642	91.5877	133.9009	64.9717	160.5169	-13.7443	17.364	-0.792	| *| |

Obs. Diagnostics *(continued)*

- **The following columns are important. They give the confidence interval for the regression line and for individual observations. The default is a 95% CI.**
- **These both center on Yhat for each observation, but have a different variance.**
- **The variance for the regression line (SLR) is**

$$S^2_{\hat{Y}_{y|x}} = MSE\left(\frac{1}{n} + \frac{\left(X_i - \bar{X}.\right)^2}{\sum\left(X_i - \bar{X}.\right)^2}\right)$$

Obs. Diagnostics (*continued*)

- **Recall that** $$Y_i = b_0 + b_1 X_i + e_i$$
-
- $$Y_i = \hat{Y} + e_i$$
-
- **The variance for** $\hat{Y}$ **is**

$$S^2_{\hat{Y}_{Y|X}} = MSE\left(\frac{1}{n} + \frac{\left(X_i - \overline{X.}\right)^2}{\sum\left(X_i - \overline{X.}\right)^2}\right)$$

Obs. Diagnostics *(continued)*

- **and the variance for Y_i is one mean square error term larger because it adds e_i with a variance of MSE.**
-
- $$S^2_{Y_{Y|X}} = MSE\left(1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}.\right)^2}{\sum\left(X_i - \overline{X}.\right)^2}\right)$$
-
- **There are comparable equations in matrix algebra for the variance that include all variances and covariances for the regression coefficients.**

Obs. Diagnostics *(continued)*

- The bottom line,
- The columns labeled "Lower 95% mean" and "Upper 95% mean" are confidence intervals on the regression line itself (Yhat). These would be used for most applications where you are asked to give a confidence interval on predicted values.

Obs. Diagnostics *(continued)*

- **The columns labeled "Lower 95% Predict" and "Upper 95% Predict" are confidence intervals on individual observations (Y_i). These would be used for evaluating individual observations and giving confidence intervals for a single new observation.**

Obs. Diagnostics *(continued)*

- The last 3 columns give the residual and the studentized residual with the standard error of that residual.
- The studentized residual is standardized to a mean of zero and variance of 1. It may be used to evaluate outliers.

Output Statistics

Obs	Dep Var Y	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual	Std Error Residual	Student Residual	-2-1 0 1 2
1	64.0000	65.4050	10.5752	42.7235	88.0865	16.9378	113.8723	-1.4050	16.941	-0.0829	| | |
2	60.0000	68.7126	11.0062	45.1068	92.3185	19.8060	117.6192	-8.7126	16.664	-0.523	| *| |
3	71.0000	53.5624	11.6962	28.4766	78.6481	3.9245	103.2002	17.4376	16.187	1.077	| |** |
4	61.0000	67.1854	8.6364	48.6623	85.7085	20.5193	113.8515	-6.1854	18.007	-0.344	| | |
5	54.0000	59.5460	9.4223	39.3371	79.7548	12.1854	106.9065	-5.5460	17.608	-0.315	| | |
6	77.0000	61.0848	13.7970	31.4932	90.6765	9.0244	113.1453	15.9152	14.438	1.102	| |** |
7	81.0000	64.1718	10.1592	42.3825	85.9610	16.1156	112.2279	16.8282	17.193	0.979	| |* |
8	93.0000	77.9457	6.2695	64.4989	91.3925	33.0521	122.8393	15.0543	18.961	0.794	| |* |
9	93.0000	89.8131	9.9864	68.3944	111.2318	41.9238	137.7024	3.1869	17.294	0.184	| | |
10	51.0000	79.3503	8.3893	61.3571	97.3435	32.8919	125.8086	-28.3503	18.123	-1.564	| ***| |
11	76.0000	77.9063	5.1132	66.9397	88.8730	33.6922	122.1205	-1.9063	19.305	-0.0987	| | |
12	96.0000	99.4135	7.3614	83.6249	115.2021	53.7637	145.0632	-3.4135	18.564	-0.184	| | |
13	77.0000	102.3025	7.1198	87.0320	117.5730	56.8294	147.7757	-25.3025	18.658	-1.356	| **| |
14	93.0000	90.2967	8.9528	71.0949	109.4985	43.3570	137.2363	2.7033	17.851	0.151	| | |
15	95.0000	107.2807	8.3375	89.3984	125.1629	60.8652	153.6961	-12.2807	18.147	-0.677	| *| |
16	54.0000	67.0830	7.3513	51.3159	82.8500	21.4406	112.7253	-13.0830	18.568	-0.705	| *| |
17	168.0000	119.1961	11.2752	95.0132	143.3789	70.0083	168.3838	48.8039	16.483	2.961	| |***** |
18	99.0000	112.7443	9.8642	91.5877	133.9009	64.9717	160.5169	-13.7443	17.364	-0.792	| *| |

Obs. Diagnostics (*continued*)

- **The second section of the observation diagnostics contains new material.**

Output Statistics

	Cook's		Hat Diag	Cov		--------DFBETAS--------			
Obs	D	RStudent	H	Ratio	DFFITS	Intercept	X1	X2	X3
1	0.001	-0.0799	0.2804	1.8655	-0.0499	0.0184	0.0389	-0.0292	-0.0126
2	0.030	-0.5088	0.3037	1.7853	-0.3361	-0.1038	0.0829	0.1986	-0.2064
3	0.151	1.0840	0.3430	1.4483	0.7832	0.7596	-0.0292	-0.3246	-0.4587
4	0.007	-0.3324	0.1870	1.5994	-0.1594	-0.0176	0.0855	0.0327	-0.0885
5	0.007	-0.3046	0.2226	1.6817	-0.1630	-0.1550	0.0089	0.0663	0.0863
6	0.277	1.1115	0.4773	1.7897	1.0622	-0.3836	-0.7486	0.9230	-0.2049
7	0.084	0.9772	0.2588	1.3667	0.5774	0.2736	-0.0944	0.2170	-0.5008
8	0.017	0.7829	0.0986	1.2410	0.2589	0.1768	0.0437	-0.1658	0.0267
9	0.003	0.1778	0.2501	1.7762	0.1027	0.0187	0.0226	-0.0717	0.0733
10	0.131	-1.6594	0.1765	0.7574	-0.7682	0.1975	0.2350	-0.6252	0.2841
11	0.000	-0.0952	0.0656	1.4354	-0.0252	-0.0154	-0.0019	0.0074	0.0036
12	0.001	-0.1774	0.1359	1.5416	-0.0703	-0.0116	-0.0505	0.0070	0.0144
13	0.067	-1.4021	0.1271	0.8773	-0.5350	0.0663	-0.3410	-0.0228	-0.0195
14	0.001	0.1460	0.2010	1.6724	0.0732	0.0317	0.0399	0.0006	-0.0478
15	0.024	-0.6631	0.1743	1.4261	-0.3046	0.1548	-0.1245	-0.0635	-0.1224
16	0.019	-0.6913	0.1355	1.3467	-0.2737	-0.0470	0.1597	0.0298	-0.1113
17	1.026	4.6666	0.3188	0.0386	3.1922	-1.8192	1.4167	0.1771	1.9528
18	0.051	-0.7804	0.2440	1.4814	-0.4433	0.0137	-0.3603	0.0259	0.0449

Obs. Diagnostics *(continued)*

- **The graphics on the left shows the standardized residuals as deviations from the mean to the left (negative) and right (positive). The numbers at the top of the column are standard deviation units, two asterisks for each std dev unit.**

Obs. Diagnostics *(continued)*

- **The RSTUDENT column is the column most used to examine for outliers. It contains DELETED standardized residuals, so the observation is excluded for the calculation of the regression line and mean square error.**
- **The observations follow a t distribution with n-p-1 degrees of freedom (p is the number of parameters including the intercept).**

Using Studentized residuals

- **Bonferroni adjustment**
  - **Doing more tests increases your chance of error.**
  - **It is possible to do 20, 100, even 1000 tests and have no Type I errors (at α=0.05), but the chance of an error goes up.**
  - **The rate of increase is not linear, so twice as many tests does not double your chance of error.**

Using Studentized residuals *(continued)*

- **Bonferroni adjustment (continued)**
  - **However, as an approximation Bonferroni noted that the probability of error would be NO GREATER than the sum of the α values of the individual tests.**
    - **For example, do one test at α and have α chance of error**
    - **Do two tests and have no more than 2α chance of error**
    - **Do 10 tests and error rate is $< 10\alpha$**

Using Studentized residuals *(continued)*

- Bonferroni adjustment (continued)
  - Bonferroni suggested a simple fix. If we were to do 2 tests at $\alpha/2$, then the two tests together would have no more than a $2*\alpha/2$ error rate, giving us α overall.
  - If we were to do 10 tests at $\alpha/10$, then the two tests together would have no more than a $10*\alpha/10$ error rate (= α).
  - Two tailed tests are already $\alpha/2$, so we actually want $\alpha/4$ for two tests and $\alpha/20$ for 10 tests.

Using Studentized residuals *(continued)*

- **To make this correction simply choose the t value to reflect the smaller α value. For studentized residuals use $t_{(1-\alpha)/2n,\ n-p}$ d.f.·**
- **For deleted residuals use $t_{(1-\alpha)/2n,n-p-1}$ d.f. where there is an extra "-1" because of the deleted value.**

Using Studentized residuals *(continued)*

- **For our numerical analysis the Bonferroni adjusted critical value would be,**
- **$t_{\alpha/2,\ n\text{-}p\ d.f.}$ = 2.144788596 (unadjusted)**
- **$t_{\alpha/2n,n\text{-}p\text{-}1}$ = 3.621389624**
- **The RSTUDENT value for one observation (#17) exceeds this value and is a probable outlier.**

Obs. Diagnostics *(continued)*

- **The second section of observation diagnostics also has a value called the Hat Diag values (H). These come from the HAT MATRIX we discussed briefly earlier in the matrix section.**

Output Statistics

Obs	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS X1	DFBETAS X2	DFBETAS X3
1	0.001	-0.0799	0.2804	1.8655	-0.0499	0.0184	0.0389	-0.0292	-0.0126
2	0.030	-0.5088	0.3037	1.7853	-0.3361	-0.1038	0.0829	0.1986	-0.2064
3	0.151	1.0840	0.3430	1.4483	0.7832	0.7596	-0.0292	-0.3246	-0.4587
4	0.007	-0.3324	0.1870	1.5994	-0.1594	-0.0176	0.0855	0.0327	-0.0885
5	0.007	-0.3046	0.2226	1.6817	-0.1630	-0.1550	0.0089	0.0663	0.0863
6	0.277	1.1115	0.4773	1.7897	1.0622	-0.3836	-0.7486	0.9230	-0.2049
7	0.084	0.9772	0.2588	1.3667	0.5774	0.2736	-0.0944	0.2170	-0.5008
8	0.017	0.7829	0.0986	1.2410	0.2589	0.1768	0.0437	-0.1658	0.0267
9	0.003	0.1778	0.2501	1.7762	0.1027	0.0187	0.0226	-0.0717	0.0733
10	0.131	-1.6594	0.1765	0.7574	-0.7682	0.1975	0.2350	-0.6252	0.2841
11	0.000	-0.0952	0.0656	1.4354	-0.0252	-0.0154	-0.0019	0.0074	0.0036
12	0.001	-0.1774	0.1359	1.5416	-0.0703	-0.0116	-0.0505	0.0070	0.0144
13	0.067	-1.4021	0.1271	0.8773	-0.5350	0.0663	-0.3410	-0.0228	-0.0195
14	0.001	0.1460	0.2010	1.6724	0.0732	0.0317	0.0399	0.0006	-0.0478
15	0.024	-0.6631	0.1743	1.4261	-0.3046	0.1548	-0.1245	-0.0635	-0.1224
16	0.019	-0.6913	0.1355	1.3467	-0.2737	-0.0470	0.1597	0.0298	-0.1113
17	1.026	4.6666	0.3188	0.0386	3.1922	-1.8192	1.4167	0.1771	1.9528
18	0.051	-0.7804	0.2440	1.4814	-0.4433	0.0137	-0.3603	0.0259	0.0449

Obs. Diagnostics *(continued)*

- **The hat diag values (h_{ii}) are the elements on the main diagonal of the hat matrix.**
- **They are measures of how close a value is to the center of all X_i values, or the center of the "X space".**
- **If a value is right at the center (mean) of all X_i values, then h_{ii} is zero.**
- **As an observation has X values further removed from the center the h_{ii} value increases.**

Obs. Diagnostics *(continued)*

- **These are also called leverage values.**
- **They value has two applications,**
- **First it is used in many calculations relating to variance. We know that as we move away from the mean of X in a simple linear regression, the variance of the regression line increases. The same thing happens for multiple regression and the h_{ii} value is used to make the adjustment in many variance calculations.**

Obs. Diagnostics *(continued)*

- **The second application of h_{ii} is as a measure of how "unusual" an observation is in terms of its X values. If the observation is near the middle of the X space, h_{ii} will be small. If an observation is far from the mean of all of the X's, the h_{ii} value will be large.**
-

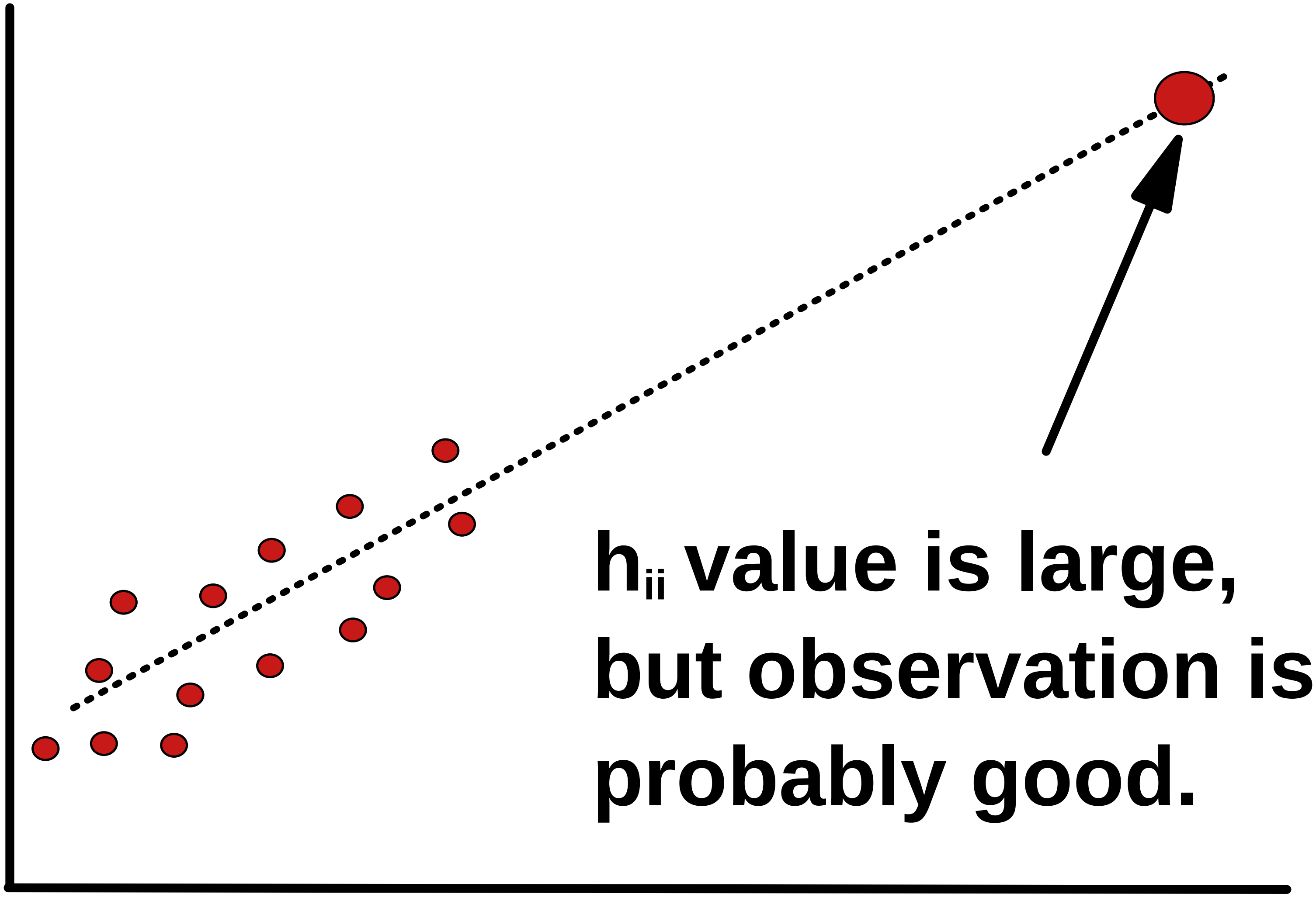
Obs. Diagnostics *(continued)*

- **Characteristics of hat diag values**
  - **h_{ii} values sum to p**
  - **mean h_{ii} = p/n**
  - **Generally, h_{ii} values greater than 0.5 are "large" while those between 0.2 and 0.5 are moderately large**
  - **Also, values over twice the mean (2p/n) are considered large.**

Obs. Diagnostics *(continued)*

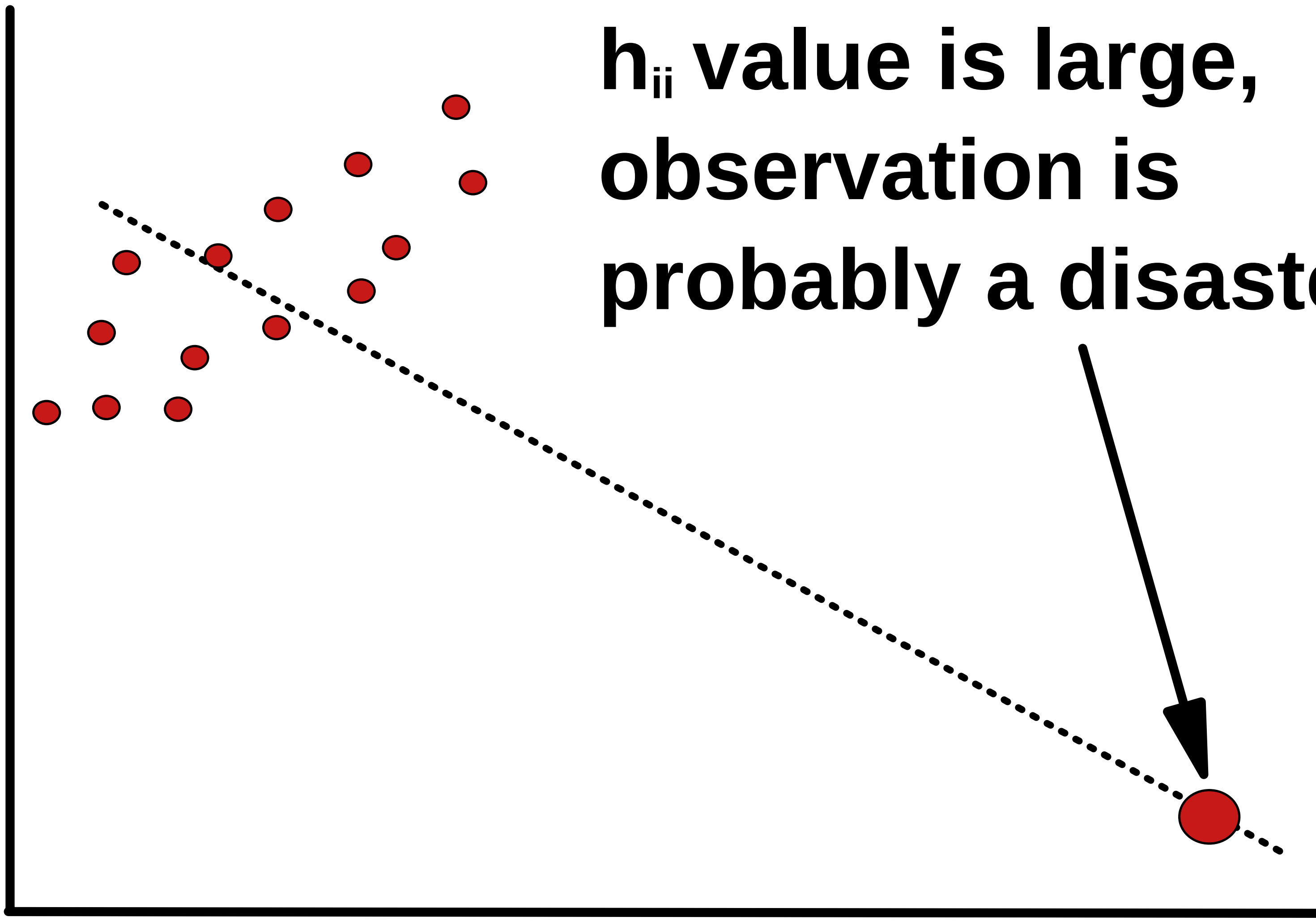
- **This is sometimes used to detect "outliers" in X, but it is important to note that a large hat value does not necessarily mean problems. It may be the most valuable observation in the data set.**
-

Hat diag values



Hat diag values *(continued)*

h_{ii} value is large, observation is probably a disaster.



Obs. Diagnostics *(continued)*

- **There are also 3 influence diagnostics in the second section of observation output. These are**
  - **Cook's D**
  - **DFFITS**
  - **DFBETAS (p columns in all)**
- **All of these are deleted statistics. They compare the performance of the line with the observation included to the performance with the observation excluded.**

Influence diagnostics

- **DFFITS**
  - **Compares the Yhat value for the point included with the Yhat value for the point excluded.**
- **DFBETAS (p columns in all)**
  - **Compares each b_i value for the point included with the b_i value for the point excluded.**
- **Cook's D**
  - **Jointly measures the changes in all b_i values together when obs. is deleted.**

Influence diagnostics *(continued)*

- **DFFITS**
  - **The value is "standardized"**
  - **for small to medium size databases, DFFITS should not exceed 1, while for large databases it should not exceed 2*sqrt(p/n)**
- **DFBetas**
  - **The value is "standardized"**
  - **for small to medium size databases, DFBETAS should not exceed 1, while for large databases it should not exceed 2/sqrt(n)**

Influence diagnostics *(continued)*

- Cook's D
  - The boundary of a simultaneous regional confidence region for all regression coefficients
  - b) this does not follow an F distribution, but it is useful to compare it to the percentiles of the F distribution [F1-a; p, n-p] where a change of < 10th or 20th percentile shows little effect, while the 50th percentile is considered large

Influence diagnostics *(continued)*

- See F tables on the Internet for P(>F)=0.5
  - For our example with n=18 and p=4 the F values of interest are
  - P(>F) = 0.7 is 0.55 (30th percentile), values exceeding this would be moderate
  - P(>F) = 0.5 is 0.88 (50th percentile), values exceeding this have large influence
- These values can be obtained from EXCEL using the FINV function.

Influence diagnostics *(continued)*

- A note on influence diagnostics,
  - Just because an observation is influential does not mean it is an outlier.
  - We saw that large h_{ii} values could be either good or bad. The same is true for influential values.
- However, if a point is both influential AND it is an outlier, then problems are much more serious.

Numerical analysis

- **For our numerical analysis.**
  - **Bonferroni adjusted Rstudent = 3.621**
  - **One observations exceeds this value, observation #17 is a possible outlier.**

Output Statistics

	Cook's		Hat Diag	Cov		--------DFBETAS--------			
Obs	D	RStudent	H	Ratio	DFFITS	Intercept	X1	X2	X3
1	0.001	-0.0799	0.2804	1.8655	-0.0499	0.0184	0.0389	-0.0292	-0.0126
2	0.030	-0.5088	0.3037	1.7853	-0.3361	-0.1038	0.0829	0.1986	-0.2064
3	0.151	1.0840	0.3430	1.4483	0.7832	0.7596	-0.0292	-0.3246	-0.4587
4	0.007	-0.3324	0.1870	1.5994	-0.1594	-0.0176	0.0855	0.0327	-0.0885
5	0.007	-0.3046	0.2226	1.6817	-0.1630	-0.1550	0.0089	0.0663	0.0863
6	0.277	1.1115	0.4773	1.7897	1.0622	-0.3836	-0.7486	0.9230	-0.2049
7	0.084	0.9772	0.2588	1.3667	0.5774	0.2736	-0.0944	0.2170	-0.5008
8	0.017	0.7829	0.0986	1.2410	0.2589	0.1768	0.0437	-0.1658	0.0267
9	0.003	0.1778	0.2501	1.7762	0.1027	0.0187	0.0226	-0.0717	0.0733
10	0.131	-1.6594	0.1765	0.7574	-0.7682	0.1975	0.2350	-0.6252	0.2841
11	0.000	-0.0952	0.0656	1.4354	-0.0252	-0.0154	-0.0019	0.0074	0.0036
12	0.001	-0.1774	0.1359	1.5416	-0.0703	-0.0116	-0.0505	0.0070	0.0144
13	0.067	-1.4021	0.1271	0.8773	-0.5350	0.0663	-0.3410	-0.0228	-0.0195
14	0.001	0.1460	0.2010	1.6724	0.0732	0.0317	0.0399	0.0006	-0.0478
15	0.024	-0.6631	0.1743	1.4261	-0.3046	0.1548	-0.1245	-0.0635	-0.1224
16	0.019	-0.6913	0.1355	1.3467	-0.2737	-0.0470	0.1597	0.0298	-0.1113
17	1.026	4.6666	0.3188	0.0386	3.1922	-1.8192	1.4167	0.1771	1.9528
18	0.051	-0.7804	0.2440	1.4814	-0.4433	0.0137	-0.3603	0.0259	0.0449

Numerical analysis (*continued*)

- For our numerical analysis.
  - Hat diag: p/n = 4/18 = .222, 2p/n = 0.444
  - One observations exceeds this value, but not 0.5 (the other criteria). No strong evidence of an "outlier" of X_i values.

Output Statistics

Obs	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	X1	X2	X3
1	0.001	-0.0799	0.2804	1.8655	-0.0499	0.0184	0.0389	-0.0292	-0.0126
2	0.030	-0.5088	0.3037	1.7853	-0.3361	-0.1038	0.0829	0.1986	-0.2064
3	0.151	1.0840	0.3430	1.4483	0.7832	0.7596	-0.0292	-0.3246	-0.4587
4	0.007	-0.3324	0.1870	1.5994	-0.1594	-0.0176	0.0855	0.0327	-0.0885
5	0.007	-0.3046	0.2226	1.6817	-0.1630	-0.1550	0.0089	0.0663	0.0863
6	0.277	1.1115	0.4773	1.7897	1.0622	-0.3836	-0.7486	0.9230	-0.2049
7	0.084	0.9772	0.2588	1.3667	0.5774	0.2736	-0.0944	0.2170	-0.5008
8	0.017	0.7829	0.0986	1.2410	0.2589	0.1768	0.0437	-0.1658	0.0267
9	0.003	0.1778	0.2501	1.7762	0.1027	0.0187	0.0226	-0.0717	0.0733
10	0.131	-1.6594	0.1765	0.7574	-0.7682	0.1975	0.2350	-0.6252	0.2841
11	0.000	-0.0952	0.0656	1.4354	-0.0252	-0.0154	-0.0019	0.0074	0.0036
12	0.001	-0.1774	0.1359	1.5416	-0.0703	-0.0116	-0.0505	0.0070	0.0144
13	0.067	-1.4021	0.1271	0.8773	-0.5350	0.0663	-0.3410	-0.0228	-0.0195
14	0.001	0.1460	0.2010	1.6724	0.0732	0.0317	0.0399	0.0006	-0.0478
15	0.024	-0.6631	0.1743	1.4261	-0.3046	0.1548	-0.1245	-0.0635	-0.1224
16	0.019	-0.6913	0.1355	1.3467	-0.2737	-0.0470	0.1597	0.0298	-0.1113
17	1.026	4.6666	0.3188	0.0386	3.1922	-1.8192	1.4167	0.1771	1.9528
18	0.051	-0.7804	0.2440	1.4814	-0.4433	0.0137	-0.3603	0.0259	0.0449

Numerical analysis *(continued)*

- **DFFITS: > 1 (small databases), or (>2*sqrt(p/n) = 0.943) for large.**
- **Observation #17 is influential.**

Output Statistics

Obs	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	X1	X2	X3
1	0.001	-0.0799	0.2804	1.8655	-0.0499	0.0184	0.0389	-0.0292	-0.0126
2	0.030	-0.5088	0.3037	1.7853	-0.3361	-0.1038	0.0829	0.1986	-0.2064
3	0.151	1.0840	0.3430	1.4483	0.7832	0.7596	-0.0292	-0.3246	-0.4587
4	0.007	-0.3324	0.1870	1.5994	-0.1594	-0.0176	0.0855	0.0327	-0.0885
5	0.007	-0.3046	0.2226	1.6817	-0.1630	-0.1550	0.0089	0.0663	0.0863
6	0.277	1.1115	0.4773	1.7897	1.0622	-0.3836	-0.7486	0.9230	-0.2049
7	0.084	0.9772	0.2588	1.3667	0.5774	0.2736	-0.0944	0.2170	-0.5008
8	0.017	0.7829	0.0986	1.2410	0.2589	0.1768	0.0437	-0.1658	0.0267
9	0.003	0.1778	0.2501	1.7762	0.1027	0.0187	0.0226	-0.0717	0.0733
10	0.131	-1.6594	0.1765	0.7574	-0.7682	0.1975	0.2350	-0.6252	0.2841
11	0.000	-0.0952	0.0656	1.4354	-0.0252	-0.0154	-0.0019	0.0074	0.0036
12	0.001	-0.1774	0.1359	1.5416	-0.0703	-0.0116	-0.0505	0.0070	0.0144
13	0.067	-1.4021	0.1271	0.8773	-0.5350	0.0663	-0.3410	-0.0228	-0.0195
14	0.001	0.1460	0.2010	1.6724	0.0732	0.0317	0.0399	0.0006	-0.0478
15	0.024	-0.6631	0.1743	1.4261	-0.3046	0.1548	-0.1245	-0.0635	-0.1224
16	0.019	-0.6913	0.1355	1.3467	-0.2737	-0.0470	0.1597	0.0298	-0.1113
17	1.026	4.6666	0.3188	0.0386	3.1922	-1.8192	1.4167	0.1771	1.9528
18	0.051	-0.7804	0.2440	1.4814	-0.4433	0.0137	-0.3603	0.0259	0.0449

Numerical analysis *(continued)*

- **DFBetas: < 1 (small database) , while for large databases it should not exceed 2/sqrt(n) = .471**
- **Observation #17 is influential.**

Output Statistics

Obs	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS X1	DFBETAS X2	DFBETAS X3
1	0.001	-0.0799	0.2804	1.8655	-0.0499	0.0184	0.0389	-0.0292	-0.0126
2	0.030	-0.5088	0.3037	1.7853	-0.3361	-0.1038	0.0829	0.1986	-0.2064
3	0.151	1.0840	0.3430	1.4483	0.7832	0.7596	-0.0292	-0.3246	-0.4587
4	0.007	-0.3324	0.1870	1.5994	-0.1594	-0.0176	0.0855	0.0327	-0.0885
5	0.007	-0.3046	0.2226	1.6817	-0.1630	-0.1550	0.0089	0.0663	0.0863
6	0.277	1.1115	0.4773	1.7897	1.0622	-0.3836	-0.7486	0.9230	-0.2049
7	0.084	0.9772	0.2588	1.3667	0.5774	0.2736	-0.0944	0.2170	-0.5008
8	0.017	0.7829	0.0986	1.2410	0.2589	0.1768	0.0437	-0.1658	0.0267
9	0.003	0.1778	0.2501	1.7762	0.1027	0.0187	0.0226	-0.0717	0.0733
10	0.131	-1.6594	0.1765	0.7574	-0.7682	0.1975	0.2350	-0.6252	0.2841
11	0.000	-0.0952	0.0656	1.4354	-0.0252	-0.0154	-0.0019	0.0074	0.0036
12	0.001	-0.1774	0.1359	1.5416	-0.0703	-0.0116	-0.0505	0.0070	0.0144
13	0.067	-1.4021	0.1271	0.8773	-0.5350	0.0663	-0.3410	-0.0228	-0.0195
14	0.001	0.1460	0.2010	1.6724	0.0732	0.0317	0.0399	0.0006	-0.0478
15	0.024	-0.6631	0.1743	1.4261	-0.3046	0.1548	-0.1245	-0.0635	-0.1224
16	0.019	-0.6913	0.1355	1.3467	-0.2737	-0.0470	0.1597	0.0298	-0.1113
17	1.026	4.6666	0.3188	0.0386	3.1922	-1.8192	1.4167	0.1771	1.9528
18	0.051	-0.7804	0.2440	1.4814	-0.4433	0.0137	-0.3603	0.0259	0.0449

Numerical analysis *(continued)*

- Cook's D: F value is 0.55 at the 30th percentile, and 0.88 at the 50th.
- Again, #17 appears to be influential.

Output Statistics

Obs	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS X1	DFBETAS X2	DFBETAS X3
1	0.001	-0.0799	0.2804	1.8655	-0.0499	0.0184	0.0389	-0.0292	-0.0126
2	0.030	-0.5088	0.3037	1.7853	-0.3361	-0.1038	0.0829	0.1986	-0.2064
3	0.151	1.0840	0.3430	1.4483	0.7832	0.7596	-0.0292	-0.3246	-0.4587
4	0.007	-0.3324	0.1870	1.5994	-0.1594	-0.0176	0.0855	0.0327	-0.0885
5	0.007	-0.3046	0.2226	1.6817	-0.1630	-0.1550	0.0089	0.0663	0.0863
6	0.277	1.1115	0.4773	1.7897	1.0622	-0.3836	-0.7486	0.9230	-0.2049
7	0.084	0.9772	0.2588	1.3667	0.5774	0.2736	-0.0944	0.2170	-0.5008
8	0.017	0.7829	0.0986	1.2410	0.2589	0.1768	0.0437	-0.1658	0.0267
9	0.003	0.1778	0.2501	1.7762	0.1027	0.0187	0.0226	-0.0717	0.0733
10	0.131	-1.6594	0.1765	0.7574	-0.7682	0.1975	0.2350	-0.6252	0.2841
11	0.000	-0.0952	0.0656	1.4354	-0.0252	-0.0154	-0.0019	0.0074	0.0036
12	0.001	-0.1774	0.1359	1.5416	-0.0703	-0.0116	-0.0505	0.0070	0.0144
13	0.067	-1.4021	0.1271	0.8773	-0.5350	0.0663	-0.3410	-0.0228	-0.0195
14	0.001	0.1460	0.2010	1.6724	0.0732	0.0317	0.0399	0.0006	-0.0478
15	0.024	-0.6631	0.1743	1.4261	-0.3046	0.1548	-0.1245	-0.0635	-0.1224
16	0.019	-0.6913	0.1355	1.3467	-0.2737	-0.0470	0.1597	0.0298	-0.1113
17	1.026	4.6666	0.3188	0.0386	3.1922	-1.8192	1.4167	0.1771	1.9528
18	0.051	-0.7804	0.2440	1.4814	-0.4433	0.0137	-0.3603	0.0259	0.0449

Numerical analysis *(continued)*

- **So we have one possible outlier (large Rstudent value for # 17) and all influence diagnostics show this observation to be influential.**
  - **This makes for a particularly serious case.**
  - **An outlier that is not influential is less of a problem. But if influential it must be thoroughly evaluated.**

Numerical analysis *(continued)*

- The "unusual" value in the X space for observation #6 is probably not a problem. It just barely exceeded the limit, and the observation is neither an outlier nor influential.
- Note that although observation #17 did not exceed our Hat value criteria, it is the second largest Hat value.

Partial Residual Plots

- **We know that the Y variable gets adjusted for each X_i variable as it enters a model.**
- **For a regression of Y on X_1, the remaining e_i are the variation left over that X_1 could not explain.**
- **What variation is left for X_2? Would fitting X_2 after X_1 be like a regression the e_i (left over from X_1) on X_2?**

Partial Residual Plots *(continued)*

- Kinda, but not quite.
- One thing that happens in the matrix algebra, that is not apparent elsewhere, is that when X_1 adjusts Y, it also adjusts X_2!
- Strange as it sounds, fitting X_2 after X_1 is like a regression of the residuals from Y, after fitting X_1, on the residuals of X_2 after it is fitted to X_1!

Partial Residual Plots *(continued)*

- **That is,**
- **Y fitted on X_1 gives residuals, say, e_{Yi}**
- **If we were to regress X_2 on X_1 we would have residuals e_{X2i}**
- **Fitting X_2 in a multiple regression after fitting X_1 is like a regression of e_{Yi} on e_{X2i}.**

Partial Residual Plots *(continued)*

- **We usually use residual plots on X or Yhat to check for problems.**
- **However, when residuals are plotted on a variable not included in the analysis they can help determine if that variable may explain the remaining variation.**
- **This is interpreted like a scatter plot.**

Partial Residual Plots *(continued)*

- **Partial residual plots are residual plots fitted to all variables, except the variable of interest, and then plotted on the variable of interest after it is adjusted for all other variables.**
- **They are requested in SAS by the "PARTIAL" option on the model statement.**
- **See Handout**

Partial Resid Plots *(continued)*

- **Partial residual plot for X1.**

```
      -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----
    Y |                                                                                     1                         |
      |                                                                                                               |
   60 +                                                                                                               +
P     |                                                                                                               |
l     |                                                                                                               |
a     |                                                                                                               |
n     |                                                                                                               |
t     |                                                                                                               |
      |                                                                                                               |
a  40 +                                                                                                               +
v     |                                                                                                               |
a     |                                                                                                               |
i     |                                                                                                               |
l     |                                                                                                               |
a     |                                                                                                               |
b     |                                                                                                               |
l  20 +                                                                                    1                          +
e     |                                                        1       1                                              |
      |                                                                                       1              1        |
p     |                                                 1                   1                                         |
h     |                                                                                                               |
o     |                                                                                                               |
s     |                                                                                                               |
p   0 +                                                           1                1                                  +
h     |                                                                                                               |
o     |                                                                                                               |
r     |                                                       1                                                       |
u     |                                                                                  1                            |
s     |                                                                                                               |
      |    1                                      1                                                                   |
  -20 +                                                                                                               +
      |                                1                                                                              |
      |                                                                                                               |
      |            1                     1                                                                            |
      |                                                                                                               |
      |                                                                                                               |
      |                                            1                                                                  |
  -40 +                                                                                                               +
      -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----
          -18   -16   -14   -12   -10    -8    -6    -4    -2     0     2     4     6     8    10    12    14    16
                                                 Inorganic phosphorus   X12
```

Regression in GLM

- PROC GLM and PROC MIXED do regression, but do not have all of the regression diagnostics available that we find in PROC REG.
- However, they do have a few advantages.
  - They facilitate the inclusion of class variable (something we will be interest in later), and
  - They provide tests of both Type I and TYPE II SS (as well as Types III and IV).

Regression in GLM *(continued)*

- The formatting is different, but most of the same information is available.

```
The GLM Procedure
Dependent Variable: Y

                                    Sum of
Source                   DF        Squares   Mean Square   F Value   Pr > F
Model                     3     6806.11145    2268.70382      5.69   0.0092
Error                    14     5583.49966     398.82140
Corrected Total          17    12389.61111

R-Square     Coeff Var      Root MSE        Y Mean
0.549340      24.57069      19.97051      81.27778
```

Regression in GLM *(continued)*

- Tests of both SS 1 and SS 3 are given by default.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	5957.022495	5957.022495	14.94	0.0017
X2	1	18.646037	18.646037	0.05	0.8319
X3	1	830.442921	830.442921	2.08	0.1710

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	4394.149832	4394.149832	11.02	0.0051
X2	1	15.897886	15.897886	0.04	0.8446
X3	1	830.442921	830.442921	2.08	0.1710

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	43.65219779	18.01021075	2.42	0.0295
X1	1.78477968	0.53769551	3.32	0.0051
X2	-0.08339706	0.41770557	-0.20	0.8446
X3	0.16113269	0.11166524	1.44	0.1710

Regression in GLM *(continued)*

- Note that the Type I and Type III are the same as in PROC REG (recall extra SS), but tests are provided. These F test values are calculated by dividing each SS (Sequential or Partial) by the MSE.
- Also note that the t-tests of the parameter estimates are the same as the tests of the Partial SS.
-

Regression in GLM *(continued)*

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	5957.022495	5957.022495	14.94	0.0017
X2	1	18.646037	18.646037	0.05	0.8319
X3	1	830.442921	830.442921	2.08	0.1710

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	4394.149832	4394.149832	11.02	0.0051
X2	1	15.897886	15.897886	0.04	0.8446
X3	1	830.442921	830.442921	2.08	0.1710

Parameter	Estimate	T for H0: Parameter=0	Pr>|T|	Std Error of Estimate
INTERCEPT	43.65219779	2.42	0.0295	18.01021075
X1	1.78477968	3.32	0.0051	0.53769551
X2	-0.08339706	-0.20	0.8446	0.41770557
X3	0.16113269	1.44	0.1710	0.11166524

Multiple Regression First Example

- The first example ends here.
- Summarization of Multiple regression will be done with the second example.
-