MULTIPLE REGRESSION DIAGNOSTICS

Documents prepared for use in course B01.1305,
New York University, Stern School of Business

Cover photo: Hyacinths 2003

This problem concerns the regression of Y on $(X_1, X_2, \ldots, X_k)$ based on n data points. The model we use is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

where the ε_i's are independent statistical noise terms with mean value zero and standard deviation σ. The subscripting scheme is done so that X_{ij} is the value of the j^{th} independent variable (X_j) for data point i.

We wish to distinguish these types of problem points (described here with generic subscripts g, h, and ℓ):

> For point g, the values of the independent variables $(X_{g1}, X_{g2}, \ldots, X_{gk})$ are reasonable when compared to the other data points, but the noise term ε_g is very far from zero.
>
> For point h, the values of the independent variables $(X_{h1}, X_{h2}, \ldots, X_{hk})$ are reasonable when compared to the other data points, but the model fails. That is, Y_h does not have a distribution centered at $\beta_0 + \beta_1 X_{h1} + \beta_2 X_{h2} + \ldots + \beta_k X_{hk}$.
>
> For point ℓ, the values of the independent variables, namely $X_{\ell 1}, X_{\ell 2}, \ldots, X_{\ell k}$, are unusual when compared to the other data points.

As we will see, point g is likely to be designated an outlier, because its corresponding residual e_g will be far from zero.

Point h could create all sorts of problems, but most likely it will resemble points of type g because its failure to fit the model will be reflected in a residual e_h which is far from zero. The least squares calculation will accommodate the other points very well, leaving point h with a residual far from zero.

Point ℓ will be called a ***high leverage*** point. High leverage points generally do not produce unusual residuals, but they have the potential to do great harm to the regression. There are many notions of harm, but here we refer to one notion: the regression coefficients would be very different if this point were omitted.

Unless the sample size n is very small, point g is not likely to create much trouble. This point will be easily picked out from the residual versus fitted plot. It is probably worthy of special note. Removal of this point

will have very little impact on the fitted regression line's coefficients

will increase the value of R^2, perhaps substantially

will reduce the value of s_ε, perhaps substantially

will shorten the prediction intervals for new points, perhaps substantially

Should point g be removed? In terms of the fitted regression, it doesn't matter (which is a vote for not removing the point). In terms of other calculations, the removal of point g improves thing; one now has to balance the improvement in the other statistics with the appearance of data-massaging.

Point h will be operationally hard to distinguish from point g, and it should be treated the same.

Point ℓ is troubling. Generally, we recommend that high leverage points be removed before the regression work starts. By the time that the regression work is completed, some of the predictors may have been removed, and the status of point ℓ may have changed. You will have to make subjective decisions about whether this point should be reincluded. It is hard to give completely general advice, but here are some considerations:

(1) If the sample size is truly large, say $n > 400$, then it's not worth the trouble to remove a small number of high leverage points.

(2) You should be concerned if your data set has a substantial number of high leverage points. Here "substantial number" is subjective, but would certainly cover ten high leverage points when $n = 40$ or twenty high leverage points when $n = 400$.

(3) Binary independent variables which are unbalanced (say 95% of values are "0" and 5% of values are "1") can easily create high leverage situations. Data points which are marked as high leverage because of this kind of situation need not be removed.

(4) Do not get into a cycle of point removal for high leverage issues. Remove points for high leverage *only* at the initial run.

(5) A decision to transform any of the independent variables will require a complete restart of the problem. That is, you'll have to start over in terms of checking for high leverage.

The data in file X:\SOR\B011305\HO\EX1233.MTP call for the regression of SALARY on predictors NumExpl, Margin, and IPCost. This problem appears in Hildebrand and Ott.

(a) Perform the regression of SALARY on the three predictors. Within **Stat** ⇒ **Regression** ⇒ **Regression** ⇒, ask for **Storage** ⇒ **Hi (leverages)** ⇒. Also, be sure to ask for the residual versus fitted plot through **Graphs** ⇒ **Residuals versus fits** ⇒. Report the fitted regression equation.

SOLUTION: Here is the regression output:

Regression Analysis: Salary versus NumExpl, Margin, IPCost

```
The regression equation is
Salary = 25.5 + 0.00389 NumExpl + 0.0957 Margin + 0.216 IPCost

Predictor        Coef     SE Coef          T        P
Constant      25.5378      0.6430      39.72    0.000
NumExpl      0.003894    0.001718       2.27    0.027
Margin        0.09572     0.03653       2.62    0.011
IPCost        0.21635     0.06920       3.13    0.003

S = 0.9999      R-Sq = 38.4%     R-Sq(adj) = 35.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3      39.291      13.097     13.10    0.000
Residual Error    63      62.983       1.000
Total             66     102.274


Unusual Observations
Obs    NumExpl     Salary         Fit      SE Fit    Residual    St Resid
  7         42     27.500      29.804       0.198      -2.304      -2.35R
 12        389     28.900      29.706       0.623      -0.806      -1.03 X
 46        130     25.700      28.426       0.247      -2.726      -2.81R
 56        371     32.400      30.729       0.510       1.671       1.94 X
 66         43     31.300      29.084       0.144       2.216       2.24R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

The residual-versus-fitted plot (not shown) seems to be reasonable.

(b) Use the F statistic and its p-value to indicate whether the overall regression is significant. Use the individual t statistics to decide whether the three predictors are needed.

SOLUTION: As $F = 13.10$ on (3, 63) degrees of freedom reports a p-value of 0.000, we can certainly claim that the overall regression is significant. Also, each of the three predictors has a t statistic with a p-value below 0.05, so we would judge each of the three predictors to be significant also.

(c) Minitab will use an X on two points with this message:

```
X denotes an observation whose X value gives it large influence.
```

Here we should interpret "large influence" as meaning that the points have the potential to seriously alter the regression results. This is a warning, not a claim that the points really have altered the results. The strategic question can be handled by examining the leverage values (which Minitab calls **Hi**), and these will appear in the data window in a column called HI1, which was created through your action in part (a). Find the numeric values for HI1 for the two points which got the X message.

SOLUTION: For point 12, the value of **Hi** is 0.388716, and for point 56 it is 0.259877.

(d) A reasonable standard for the leverage values uses a threshold of concern. A leverage value (**Hi** in Minitab) is potentially troublesome if it exceeds $3\frac{k+1}{n}$, where k is the number of predictors (here 3) and n is the number of data points. Do the points identified in (c) exceed this threshold of concern? Can you see what is potentially troublesome about these points?

SOLUTION: We have $n = 67$, so that $3\frac{k+1}{n} = 3\frac{3+1}{67} \approx 0.1791$. Certainly the leverage values (**Hi**) for both points 12 and 56 easily exceed this threshold. Now, why are these points unusual? It seems that these two have outrageously large values for NumExpl. Point 12 is also extremely unusual in its combination of (NumExpl, Margin).

(e) The cautious approach to regression requires that high leverage points be set aside and that the regression should be repeated without these points. Give the regression on the remaining 65 points.

Here are two artful ways to omit points from a Minitab regression.

Method 1:

Start by copying the dependent variable column to a new column. Use **Manipulate** ⇒ **Copy Columns** ⇒ to make a copy of SALARY in a new column; you might call this new column as SALBACKUP.

In the original SALARY column, type the missing data code * over the values for the points you want to omit (12 and 56).

Repeat the regression, using exactly the same commands as before. The output will also include some useful facts about the two omitted points.

Method 2:

Start by marking the entire data area, including the variable names, as a block. Press Ctrl+C to copy this to the Windows clipboard.

Use **File** ⇒ **New** ⇒ **Worksheet** to create a new worksheet. With the cursor in the name box for C1, press Ctrl+V. This will create a copy of your original data in the new worksheet.

You would like to give a new name to this copy. Click on the **Show Worksheets Folder** icon; this appears as a small square button with an image of three cascading data sheets. The Project Manager window will open up, and the new worksheet will appear as a folder icon (with the name Worksheet 2). Click with the *right* mouse button on this folder, and then select **Rename**. Type an appropriate name, such as Ex1233_NO12_56.

Return to the new worksheet, which should now appear with its new name. In the SALARY column, type the missing data code * over the values for the points you want to omit (12 and 56).

Repeat the regression, using exactly the same commands as before. The output will include some useful facts about the two omitted points.

You might wish to save the original data together with your modified worksheet, and it's convenient to make this a project. Use **File** ⇒ **Save Project As** ⇒. A recommended name would be EX1233.MPJ.

If it should happen that this regression with 65 points gets some X messages, it would be reasonable to ignore them. That is, we do not want to get into a cycle of omitting points.

SOLUTION: Here is the regression omitting these points:

Regression Analysis: Salary versus NumExpl, Margin, IPCost

```
The regression equation is
Salary = 25.8 + 0.00264 NumExpl + 0.0830 Margin + 0.226 IPCost

65 cases used 2 cases contain missing values

Predictor        Coef     SE Coef          T        P
Constant      25.7927      0.6454      39.96    0.000
NumExpl      0.002637    0.002516       1.05    0.299
Margin        0.08302     0.03898       2.13    0.037
IPCost        0.22569     0.07067       3.19    0.002

S = 0.9841      R-Sq = 36.3%     R-Sq(adj) = 33.2%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3      33.666      11.222     11.59    0.000
Residual Error    61      59.075       0.968
Total             64      92.741
```

```
Unusual Observations
Obs    NumExpl      Salary         Fit      SE Fit    Residual    St Resid
  7         42      27.500      29.789       0.215      -2.289       -2.38R
 12        389           *      29.402       0.918           *           * X
 37         28      27.700      29.699       0.199      -1.999       -2.07R
 41        230      29.300      28.923       0.435       0.377        0.43 X
 46        130      25.700      28.358       0.261      -2.658       -2.80R
 56        371           *      30.304       0.727           *           * X
 64        279      29.400      29.854       0.511      -0.454       -0.54 X
 66         43      31.300      29.122       0.149       2.178        2.24R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

You should note that points 12 and 56 still appear in the listing with an X mark. We will choose to ignore other points with X marks (once we have passed the initial regression).

(f) Make a comparison between the two regression results in terms of the F statistic, the t statistics, R^2, and s_ε . Also, does the regression in (e) suggest that any of the three predictors could be removed? This removal will be followed up in (g) and (h).

SOLUTION:

Calculation		Full regression $n = 67$ (a)		Reduced regression $n = 65$ (e)	
F		13.10 on (3, 63) df		11.59 on (3, 61) df	
R^2		38.4%		36.3%	
S (s_ε)		0.9999		0.9841	
b_{NumExpl}	(t_{NumExplo})	0.003894	(2.27)	0.002637	(1.05)
b_{Margin}	(t_{Margin})	0.09572	(2.62)	0.08302	(2.13)
b_{IPCost}	(t_{IPCost})	0.21635	(3.13)	0.22569	(3.19)

This does seem to suggest that variable NumExpl could well be removed.

(g) Remove the variable for which the t statistic in part (f) was inside the interval (-2, 2). Does the printout indicate that anything has changed with regard to points 12 and 56?

SOLUTION: Here is the regression on only (Margin, IPCost):

Regression Analysis: Salary versus Margin, IPCost

```
The regression equation is
Salary = 25.9 + 0.0933 Margin + 0.207 IPCost

65 cases used 2 cases contain missing values

Predictor        Coef     SE Coef          T        P
Constant      25.9400      0.6304      41.15    0.000
Margin        0.09334     0.03774       2.47    0.016
IPCost        0.20719     0.06849       3.03    0.004

S = 0.9849      R-Sq = 35.2%     R-Sq(adj) = 33.1%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      32.602      16.301     16.81    0.000
Residual Error    62      60.138       0.970
Total             64      92.741
```

```
Unusual Observations
Obs     Margin     Salary         Fit      SE Fit    Residual    St Resid
  7       23.4     27.500      29.908       0.183      -2.408       -2.49R
 37       19.8     27.700      29.798       0.175      -2.098       -2.16R
 38        6.7     27.300      27.220       0.415       0.080        0.09 X
 46       15.6     25.700      28.248       0.239      -2.548       -2.67R
 59        9.8     27.000      27.153       0.397      -0.153       -0.17 X
 66       18.3     31.300      29.206       0.125       2.094        2.14R
```

This model certainly fits well. We should also note that points 12 and 56 are no longer marked as **Hi**, or high leverage.

(h) Once the problem is down to two predictors, it appears that points 12 and 56 are no longer troublesome. Restore them to the regression, and compare the findings to that of the regression in (g). The work that you did in step (e) allows you to recover the SALARY for points 12 and 56.

SOLUTION: Here is that regression:

```
The regression equation is
SALARY = 25.9 + 0.102 Margin + 0.196 IPCost

Predictor        Coef       StDev          T        P
Constant      25.9084      0.6416      40.38    0.000
Margin        0.10220     0.03757       2.72    0.008
IPCost        0.19559     0.07077       2.76    0.007

S = 1.032       R-Sq = 33.4%     R-Sq(adj) = 31.3%

Analysis of Variance

Source            DF          SS          MS          F        P
Regression         2      34.157      17.079      16.05    0.000
Residual Error    64      68.117       1.064
Total             66     102.274

Unusual Observations
Obs     Margin   SALbacku          Fit   StDev Fit    Residual    St Resid
  7       23.4     27.500       29.984       0.188      -2.484       -2.45R
 37       19.8     27.700       29.829       0.181      -2.129       -2.10R
 38        6.7     27.300       27.211       0.420       0.089        0.09 X
 46       15.6     25.700       28.307       0.249      -2.607       -2.60R
 56       22.3     32.400       29.645       0.182       2.755        2.71R
 59        9.8     27.000       27.192       0.412      -0.192       -0.20 X
 66       18.3     31.300       29.250       0.129       2.050        2.00R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

Here's a comparison:

The fitted equation in (g) was

```
Salary = 25.9 + 0.0933 Margin + 0.207 IPCost
```

The fitted equation in (h) was

```
SALARY = 25.9 + 0.102 Margin + 0.196 IPCost
```

Other facts:

Calculation	(g) Two predictors $n = 65$	(h) Two predictors $n = 67$
F	41.15	40.38
t for Margin	2.47	2.72
t for IPCost	3.03	2.76
S (s_ε)	0.9849	1.032
R^2	35.2%	33.4%

The results of (g) and (h) are generally similar. One might actually say that (g) is a somewhat better fit to the data, but this is a close call.

You should feel comfortable with the removal of the variable NumEmpl. It's then less critical whether you do or do not include points 12 and 56.

You might observe that all this action has led us to this simple resolution:

> The relationship between SALARY and NumExpl is dominated by two data points, 12 and 56.
> When points 12 and 56 are removed, the relationship between SALARY and NumEmpl disappears.

Suppose that you have a data base involving variables Y, A, B, C, D and that this data base has $n = 74$ points. You wish to consider the regression of Y on (A, B, C, D). Concerns about high leverage values (HI in Minitab) cause you to consider the removal, possibly temporarily, of points 14 and 68. Here are two distinct strategies:

STRATEGY 1: This strategy keeps you within the original worksheet.

Begin by making a copy of the dependent variable. Do **Calc** ⇒ **Calculator** ⇒. In the box next to **Store result in variable:** type the name YCOPY (or any similar suggestive name). In the box **Expression:** simply type Y.

Next in the spreadsheet move to the entry in row 14 under Y and type * , which is Minitab's missing data code. Similarly place * in row 68 under Y.

Now do the regression again of Y on (A, B, C, D). The resulting printout will show a regression with this message:

```
72 cases used 2 cases contain missing values
```

An advantage to this strategy is that regression work will still show leverage (HI) values for the omitted points 14 and 68. This can be useful in deciding later whether you might readmit these points to the regression.

Another advantage is that the point sequencing is maintained; that is, point number 74 is still point number 74. Clean accounting is enormously helpful.

STRATEGY 2: This strategy creates a new worksheet.

In your worksheet, mark the entire relevant data *including the variable names* as a block. Put the cursor in the name block for the first column, and while holding down the shift key, move the cursor to the last row of the final column. Press Ctrl-C to mark this block.

Do **File** ⇒ **New** ⇒ **Minitab worksheet**. This will create a new worksheet. (The Alt-W or **Window** feature will allow you to move among worksheets.) In this new worksheet, place the cursor in the name box for column 1 and do Ctrl-V. This will copy the contents of the old worksheet into the new worksheet. You should use **Window** ⇒ **Manage Worksheets** ⇒ **Description** ⇒ to leave yourself notes about the worksheets.

Perform the appropriate editing in the new worksheet. For instance, you can remove an entire row through **Manipulate** ⇒ **Erase Variables**. Removing rows will alter the sequencing; if you remove two rows then the final row will have number 72. Of course, you may still choose to use the editing style in Strategy 1 (which will preserve the original sequencing).

While this worksheet is active, you should use **Editor** ⇒ **Worksheet description** to leave yourself a reminder as to what you've done. For instance, this is the right place to indicate that you've removed points 14 and 68. You might also consider using **Window** ⇒ **Manage worksheets** to give this new worksheet a descriptive name.

At the conclusion of your work, you should save everything as a *project*. This will keep your two (or more) worksheets together in a single file with extension MPJ.

A USEFUL MULTIPLE REGRESSION ENDING

This document deals a data set on trash hauling information collected over 40 districts. Because these districts differed substantially in size, all variables were logged. The objective was to explain lwaste (logarithm of solid waste generated) in terms of five predictors.

At the initial stage of the work, point 10 was identified as "large influence" or "high leverage" by Minitab.

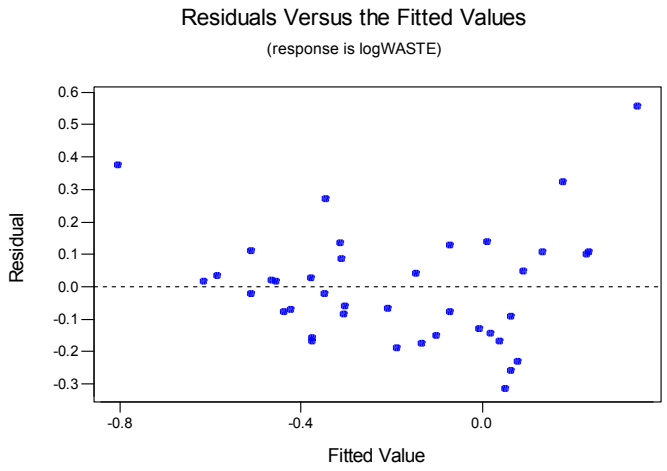
> Minitab uses the cutoff $3\frac{k+1}{n}$ for determining high leverage points. You do not need to actually go through the work of finding the exact leverage value, though it can be interesting. Here the leverage value for point 10 was found to be 0.484069. By comparison, $3\frac{k+1}{n} = 3\frac{5+1}{40} = 0.45$.

Point 10 definitely has the potential to make trouble, so we set it aside.

Here's the regression using 39 points and all five predictors.

```
The regression equation is
logWASTE = - 0.541 - 0.0195 logIND + 0.0603 logMETALS + 0.0407 logTRUCK
           - 0.129 logRETAIL + 0.244 logHOTEL

39 cases used 1 cases contain missing values

Predictor        Coef       StDev          T        P
Constant      -0.5405      0.1407      -3.84    0.001
logIND       -0.01949     0.02343      -0.83    0.411
logMETAL      0.06027     0.02119       2.84    0.008
logTRUCK      0.04070     0.02472       1.65    0.109
logRETAI     -0.12913     0.05849      -2.21    0.034
logHOTEL      0.24390     0.05747       4.24    0.000

S = 0.1920      R-Sq = 70.0%     R-Sq(adj) = 65.4%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         5     2.83611     0.56722     15.38    0.000
Residual Error    33     1.21675     0.03687
Total             38     4.05285

Source       DF      Seq SS
logIND        1     1.67707
logMETAL      1     0.15937
logTRUCK      1     0.11933
logRETAI      1     0.21616
logHOTEL      1     0.66418

Unusual Observations
Obs     logIND   logWASTE         Fit   StDev Fit    Residual    St Resid
  2       7.11     0.9030      0.3440      0.0878      0.5590        3.27R
  5       2.53    -0.4292     -0.8069      0.1066      0.3776        2.36R
 10      -0.69          *     -0.3431      0.1860           *           * X
 15       3.75     0.5020      0.1781      0.1117      0.3239        2.07R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

The residual-versus-fitted plot for this is the following:



Because the visual impression of expanding residuals seems to come from the single point at the upper right, we will *not* interpret this plot as requesting a logarithm transformation of the dependent variable.

We like this regression, except for the fact that some of the *t* statistics are weak. Let's note that $R^2 = 70.0\%$, and $s_\varepsilon = 0.1920$.

We'll show now a final version for this regression (without telling the whole story as to how we got here).

```
The regression equation is
logWASTE = - 0.700 + 0.0554 logMETALS + 0.142 logHOTEL

39 cases used 1 cases contain missing values

Predictor        Coef       StDev          T        P
Constant     -0.70049     0.07410      -9.45    0.000
logMETAL      0.05545     0.01402       3.96    0.000
logHOTEL      0.14211     0.02772       5.13    0.000

S = 0.2011      R-Sq = 64.1%     R-Sq(adj) = 62.1%

Analysis of Variance

Source            DF          SS          MS          F        P
Regression         2      2.5967      1.2983      32.10    0.000
Residual Error    36      1.4562      0.0404
Total             38      4.0529

Source       DF      Seq SS
logMETAL      1      1.5335
logHOTEL      1      1.0632

Unusual Observations
Obs    logMETAL   logWASTE         Fit    StDev Fit    Residual    St Resid
  2        6.58     0.9030      0.3590       0.0755      0.5440        2.92R
 15        1.50     0.5020     -0.0317       0.0633      0.5337        2.80R
 20        4.84    -0.4020     -0.5306       0.1067      0.1286        0.75 X

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

We see that R^2 has dropped, but only to 64.1%. We're happy to tolerate this drop in R^2 to reduce the problem to just two predictors. We see that Minitab has found another large influence point, point 20, but we're going to react only at the beginning of the work to such messages.

The residual versus fitted plot here looks similar to the original.

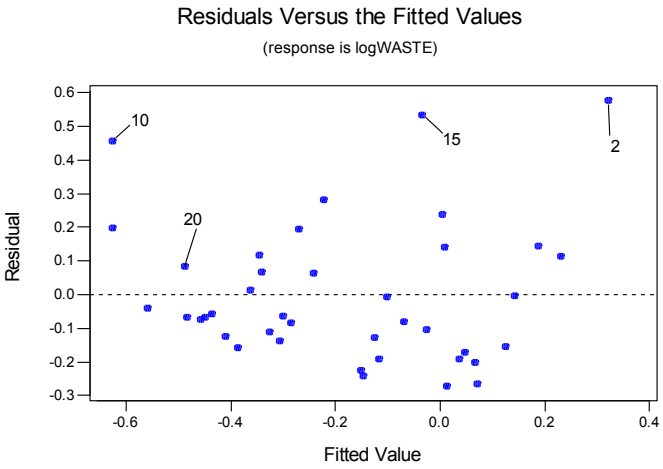
Now that we're down to only two predictors, maybe point 10 is not troublesome any more. Let's restore point 10 and see what happens:

```
The regression equation is
logWASTE = - 0.643 + 0.0508 logMETALS + 0.129 logHOTEL

Predictor        Coef       StDev          T        P
Constant     -0.64349     0.07469      -8.62    0.000
logMETAL      0.05078     0.01476       3.44    0.001
logHOTEL      0.12936     0.02893       4.47    0.000

S = 0.2139      R-Sq = 58.3%     R-Sq(adj) = 56.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      2.3611      1.1805     25.81    0.000
Residual Error    37      1.6921      0.0457
Total             39      4.0532

Source       DF      Seq SS
logMETAL      1      1.4469
logHOTEL      1      0.9142

Unusual Observations
Obs   logMETAL   logWASTE         Fit   StDev Fit    Residual    St Resid
  2       6.58     0.9030      0.3230      0.0787      0.5800        2.92R
 10      -0.69    -0.1672     -0.6262      0.0700      0.4590        2.27R
 15       1.50     0.5020     -0.0343      0.0673      0.5363        2.64R
 20       4.84    -0.4020     -0.4874      0.1118      0.0854        0.47 X

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

Point 10 is no longer a high leverage point. You might note that we've paid a penalty in R^2, a drop from 64.1% to 58.3%, just for putting in this one point. You might look back at the original data. Point 10 is really unusual.

Should we react to the fact that point 20 is now identified as having high leverage? Probably not, as the process of editing out points could go on indefinitely.

Here is the residual versus fitted plot, with the interesting points marked:



There are a number of numerical relationships in computer regression output that may appear as coincidences. In this document, we will not count the basic accounting facts as coincidences. (An example of a basic accounting fact is $SS_{\text{regression}} + SS_{\text{error}} = SS_{\text{total}}$.) Here is a list of some of these. This uses n = number of data points and K = number of independent variables.

THESE FACTS ALWAYS HOLD:

R^2 = R-squared = $\dfrac{SS_{\text{regression}}}{SS_{\text{total}}}$

R^2_{adj} = adjusted R-squared = $1 - \left(\dfrac{s_\varepsilon}{\text{SD}(Y)}\right)^2$

R = Multiple correlation = $\sqrt{R^2}$

s_ε = Standard error of estimate = $\sqrt{MS_{\text{error}}}$

SD(Y) = SD(dependent variable) = $\sqrt{\dfrac{SS_{\text{total}}}{n-1}}$

$F = \dfrac{n-1-K}{K} \times \dfrac{R^2}{1-R^2}$

THESE HOLD FOR SIMPLE REGRESSION (K = 1):

$t^2 = F$ (using t for slope)

P-value for t (for slope) = P-value for F

r = ordinary correlation = $\pm\sqrt{R^2}$ (using + if $b > 0$ and - if $b < 0$)

VIF (variance inflation factor) cannot be given

THESE HOLD FOR THE CASE OF TWO PREDICTORS (K = 2):

The two VIF (variance inflation factor) values are equal

Some software provides tolerance instead of VIF, but tolerance = $\dfrac{1}{\text{VIF}}$.