PLSC 503: "Multivariate Analysis for Political Research"

Multivariate Regression: Residuals, Model Fit, and Outliers
March 16, 2017

Model Fit

There's not much to say about model fit in the multivariate context, at least not that we haven't covered in bivariate regression. Some points:

- R^2, $R^2_{adj.}$, RMSE, and the like all remain useful means of assessing model fit (as far as they go).
  - Generally speaking, the greater the number of variables you have, and/or the smaller the number of observations, the more useful $R^2_{adj.}$ is.
  - Also: Remember all the warnings about when one can and cannot use and compare R^2s.
- We can also use an F-test to assess model fit in the most general sense:
  - R (and Stata , and most other programs) automatically report an F-statistic for the joint null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_K = 0$. Suffice it to say that we can almost always reject this at a high level of confidence.
  - As we said before, we can also use F-tests to assess "nested" hypotheses; e.g., that groups of variables have effects that are (jointly) zero, and so to assess whether or not adding one or more variables significantly improves the fit of any given model.

Outliers, Influence, and So Forth

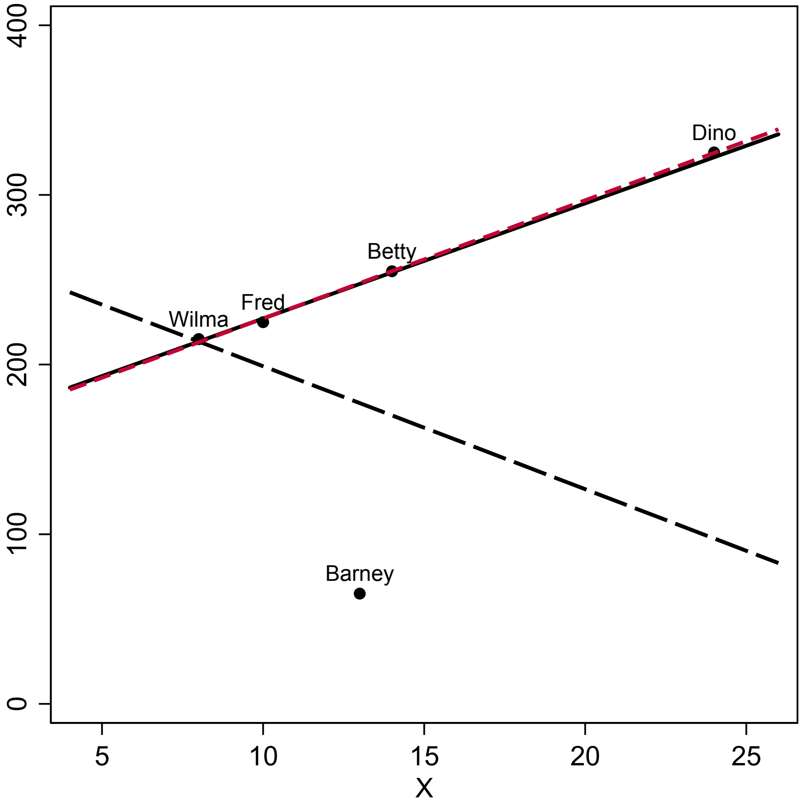
Of more interest today is the issue of *outliers* – observations that are somehow unusual, either in their value of Y_i, of one or more X_is, or some combination thereof. It's important to think about outliers, since individual observations with unusual and/or extreme values can have a disproportionate influence on one's regression results.

Terms

First, a little terminology:

- **Leverage** is the term for the degree of potential influence on the coefficient estimates that a given observation can (but not necessarily does) have.
- **Discrepancy** refers to the extent to which an observation is "unusual," or "different" from the rest of the data.

Figure 1: Discrepancy, Leverage, and Influence



Note: Solid line is the regression fit for Wilma, Fred, and Betty only. Long-dashed line is the regression for Wilma, Fred, Betty, and Barney. Short-dashed line is the regression for Wilma, Fred, Betty and Dino.

- **Influence** is just that: How much effect does a particular observation's value(s) on Y and $\mathbf{X}$ have on the coefficient estimates. Fox writes:

$$\text{Influence} = \text{Leverage} \times \text{Discrepancy}$$

- An **outlier** is an observation that has an unusual value on the dependent variable Y given its particular combination of values on $\mathbf{X}_i$.

Consider Figure 1:

- Here, Dino is an example of an observation with high leverage, but relatively little discrepancy (that is, he's very close to the regression line defined by Wilma, Fred, and Betty). This means that his inclusion in the data has little or no effect on the regression line (that's the short-dashed line); his influence is low because his discrepancy is low, even though his leverage is high.

- Barney, by contrast, has both moderately high leverage (though lower than Dino) and high discrepancy, which means he has substantial influence on the regression results (the long-dashed line).

These terms will become both clearer and more useful as we consider means of locating and assessing the impact of outliers in our data.

Leverage

The most common way of assessing an observation's leverage is through its "hat value," commonly denoted h_i. If we think about the predicted value of Y, we can note that:

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= \mathbf{HY}\end{aligned}$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

This $N \times N$ thing $\mathbf{H}$ is sometimes known as the "hat matrix," and its diagonal elements h_i (sometimes referred to as the "orthogonal projection of the hat matrix," and denoted h_{ii}) are:

$$h_i = \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i' \tag{1}$$

Seen in this light, the hat value essentially a multivariate generalization of the variation in $\mathbf{X}_i$ around its mean "divided by" the total variation in the (observed) $\mathbf{X}$ around its centroids. Note a few things about h_i:

1. It is solely a function of $\mathbf{X}$ – that is, it indicates how "unusual" a particular observation is vis-à-vis others *on the covariates only* (*not* on Y).

2. $h_i \in [1/N, 1]$.

3. $\bar{h} = K/N$.

4. This suggests that observations with $h_i \geq 2(K/N)$ can (and often are) considered candidates for "outlier" or "high-influence" status.

As we'll see, the h_i play a pervasive and critical role in assessing things like leverage, influence, and the like.

Discrepancy (or, Fun With Residuals)

Residuals would seem to be the natural place to go to look for "outlying" observations as well. There are a number of useful things we can do with residuals in the context of assessing influential data.

Standardized Residuals

First, note that the variability in the estimated residuals is:

$$\widehat{\mathrm{Var}(\hat{u}_i)} = \hat{\sigma}^2[1 - \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i'] \tag{2}$$

so that the standard error of the estimated residuals is:

$$\begin{aligned}\widehat{\mathrm{s.e.}(\hat{u}_i)} &= \hat{\sigma}\sqrt{[1 - \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i']} \\ &= \hat{\sigma}\sqrt{1 - h_i} \end{aligned} \tag{3}$$

This means that high-leverage observations tend to have (relatively) small residuals. That makes sense, given that such observations are disproportionately "pulling" the regression plane in their direction, but it also means that just looking at the variation in the residuals itself won't help us much here.

One alternative is to calculate a "standardized" residual:

$$\tilde{u}_i = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1 - h_i}} \tag{4}$$

which in effect "nets out" the variability in $\mathbf{X}_i$ for $\hat{u}_i$. Standardized residuals are effectively more "comparable" than are the usual residuals we look at, at least for purposes of detecting outliers and the like, and so they're used a good bit (as we'll see below).

Studentized Residuals

That's a start, but (as Fox notes) (4) has some unfortunate distributional properties that prevent us from using it. Specifically, because $\hat{\sigma}^2$ (and therefore $\hat{\sigma}$) contains $\hat{u}_i^2$, the numerator and the denominator are not independent of one another. The standard way of dealing with this is to calculate a slightly different version of $\hat{\sigma}^2$ – call it $\hat{\sigma}^2_{-i}$, that is based on regressing Y on $\mathbf{X}$ for the $N - 1$ observations in the data that are *not* observation i.

In point of fact, we don't even have to *estimate* that regression; it can be shown that:

$$\hat{\sigma}^2_{-i} = \frac{\hat{\sigma}^2(N - K)}{N - K - 1} - \frac{\hat{u}_i^2}{(N - K - 1)(1 - h_i)} \tag{5}$$

Once we have this, we can then calculate:

$$\hat{u}'_i = \frac{\hat{u}_i}{\hat{\sigma}_{-i}\sqrt{1-h_i}} \tag{6}$$

These "Studentized" residuals are similar to the standardized ones we talked about above, in that they are on a "common" scale. As with all residuals, they are centered around zero; moreover, $\hat{u}'_i$ are distributed according to a t distribution with $N-K-1$ degrees of freedom. That means that we should expect roughly 95 percent of them to fall within the interval $[-2,2]$; it also means that we can use them to do hypothesis testing in a number of different ways.

Digression: Interestingly, if we define a regression

$$Y_i = \mathbf{X}_i\beta + Z_i\gamma + u_i \tag{7}$$

where

$$\begin{aligned} Z_i &= 1 \text{ for observation "}i\text{"} \\ &= 0 \text{ otherwise} \end{aligned}$$

then the normal t-test value for $H_0 : \gamma = 0$ (that is, $\frac{\hat{\beta}_Z}{\widehat{\text{s.e.}(\hat{\beta}_Z)}}$) turns out to be equal to the Studentized residual $\hat{u}'_i$ (we'll demonstrate this a bit later...).

Influence

As we noted above, the *influence* of a particular observation is a combination of its leverage and its "discrepancy" – in short, how unusual it is combined with "where it's located." Intuitively, influence is just that – the effect of a particular observation on the coefficient estimates. That suggests that a simple, direct estimate of that influence is the difference between the coefficient estimate we get when we include the observation in question in the data and that we get when we exclude it:

$$D_{ki} = \hat{\beta}_k - \hat{\beta}_{k(-i)} \tag{8}$$

where the $(-i)$ notation, as above, indicates the coefficient estimate obtained when we exclude observation i from the estimation. Of course, since the estimates $\hat{\boldsymbol{\beta}}$ are scaled very differently (depending on the scale of their respective $\mathbf{X}$s), it's wise to rescale each of the D_{ki} to account for that:

$$D^*_{ki} = \frac{D_{ki}}{\widehat{\text{s.e.}(\hat{\beta}_{k(-i)})}} \tag{9}$$

The quantities in Eq. (8) are often referred to as "DFBETA$_{ki}$", and those in (9) as "DFBETAS$_{ki}$" (where the "S" at the end of the latter stands for "standardized"). Note that because DFBETAs are signed by the numerator,

- Positive values of DFBETA_{ki} / DFBETAS_{ki} correspond to observations which, by their presence, *decrease* the estimate of $\hat{\beta}_k$, while

- Negative values of DFBETA_{ki} / DFBETAS_{ki} occur for observations which have the effect of *increasing* the estimate of $\hat{\beta}_k$.

Cook's D and DFITS

DFBETAs are useful, in that they can reveal almost immediately (through graphical inspection) when particular observations' values on some X_k are particularly influential. They can also be used to construct summary statistics of each observation's influence on the regression model. One such statistic is *Cook's D*, which is based on the idea that – in theory – one could conduct an F-test on each observation for the general hypothesis that $\beta_k = \hat{\beta}_{k(-i)} \forall k \in K$. The simple formula for Cook's D is:

$$\begin{aligned} D_i &= \frac{\tilde{u}_i^2}{K} \times \frac{h_i}{1-h_i} \\ &= \frac{h_i \hat{u}_i^2}{K\hat{\sigma}^2(1-h_i)^2} \end{aligned} \tag{10}$$

that is, as the squared standardized residual, divided by K, and multiplied by $h_i/(1-h_i)$. As Fox notes, the first part of (10) is a measure of discrepancy, while the second is a measure of leverage; high values on both are necessary to generate high influence.

A similar measure is inelegantly known as "DFITS" (or sometimes "DFFITS"):

$$\text{DFITS}_i = \hat{u}_i' \sqrt{\frac{h_i}{1-h_i}} \tag{11}$$

In most cases, this is very similar to D_i – in fact, it is almost always the case that $D_i \approx \frac{\text{DFITS}_i^2}{K}$.

Variance: The Other Side of Influence

Beyond their influence on $\hat{\boldsymbol{\beta}}$, outliers may also be of importance for what they do for one's standard errors. In particular, observations with exceptionally large or small values of Y or $\mathbf{X}$ that also lie "close" to the regression hyperplane – while not exerting any particularly large influence on $\hat{\boldsymbol{\beta}}$ – may nonetheless work to decrease our estimated standard errors. Consider our example in Figure 1 above: While including Dino in the model doesn't change the estimated slope $\hat{\beta}$ much, it does reduce its estimated standard error appreciably:

```
> # No Barney OR Dino...
> summary(lm(Y~X,data=subset(flintstones,name!="Dino" & name!="Barney")))

Residuals:
     2      4      5
 0.714 -2.143  1.429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  159.286      6.776    23.5    0.027 *
X              6.786      0.619    11.0    0.058 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.67 on 1 degrees of freedom
Multiple R-squared: 0.992,Adjusted R-squared: 0.984
F-statistic:  120 on 1 and 1 DF,  p-value: 0.0579

> # No Barney (Dino included...)
> summary(lm(Y~X,data=subset(flintstones,name!="Barney")))

Residuals:
        2         3         4         5
-8.88e-16  2.63e-01 -2.11e+00  1.84e+00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  157.368      2.465    63.8  0.00025 ***
X              6.974      0.161    43.3  0.00053 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.99 on 2 degrees of freedom
Multiple R-squared: 0.999,Adjusted R-squared: 0.998
F-statistic: 1.87e+03 on 1 and 2 DF,  p-value: 0.000534
```

Substantively, this may be a good thing or a bad thing:

- If the data are really just that – a case, like any other, that lies close to the regression line – then we should be happy, since it suggests that the model does a good job of predicting Y with $\mathbf{X}$ even at relatively "odd" values of $\mathbf{X}$ (and so the reduction in the estimated standard errors are warranted).

- On the other hand, if the "outlier" is in fact really a bad observation, then it may cause us to overstate the precision of our estimates.

Statistically, we'd like to have a way to see whether in fact this is happening or not – that is, whether a particular observation has a large effect on the variance-covariance estimates of our parameters. The standard means of assessing this is the COVRATIO statistic:

$$\text{COVRATIO}_i = \left[(1-h_i)\left(\frac{N-K-1+\hat{u}_i'^2}{N-K}\right)^K\right]^{-1} \tag{12}$$

COVRATIO is a standardized indicator of the extent to which a particular observation influences the precision of the estimated coefficients. The middle-point for COVRATIO is the value 1.0, so that:

- Observations with $\text{COVRATIO}_i > 1$ *increase* the precision of the estimates (that is, drive the estimated standard errors down).

- Those with $\text{COVRATIO}_i < 1$ *decrease* the precision of the estimates, and drive the standard error estimates up.

Either way, large (absolute) values of COVRATIO_i suggest that the observation in question is in some way overly influential on the variance–covariance estimates of the parameters.

An Example

We update an example from a classic work by Dahl (1957) on the Supreme Court as a "countermajoritarian institution." The dependent variable (`nulls`) is the number of federal laws struck down by the Court in each Congress, from the First to the 104th (that is, from 1789 to 1996); this variable has a mean of 1.3, with a minimum of zero and a maximum of 10. The key covariates are three:

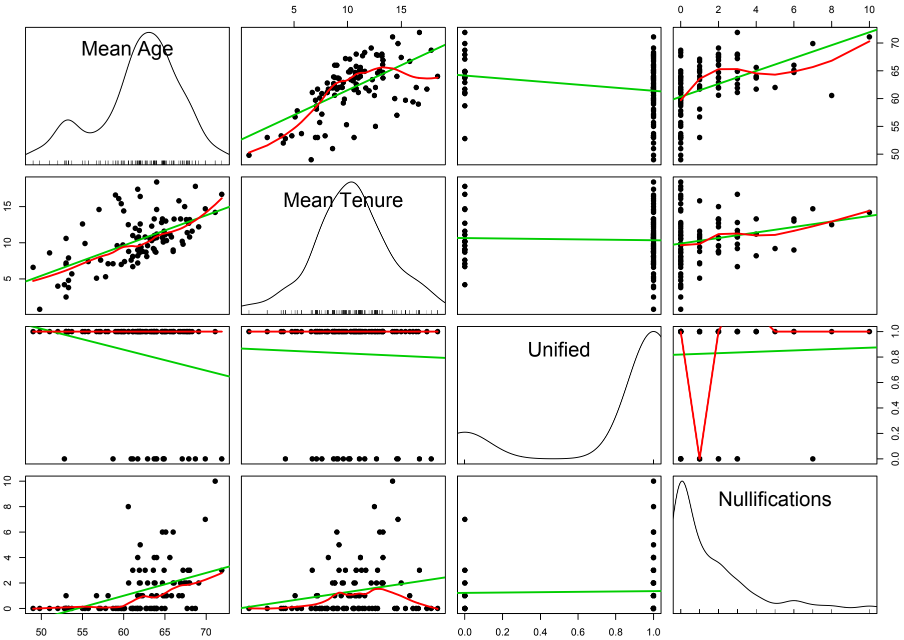
1. `age` – the mean age of the members of the Supreme Court.

2. `tenure` – the mean tenure (also in years) of the members of the Court.

3. `unified` – a dummy variable indicating whether (= 1) or not (= 0) the Congress was "unified" (controlled by the same party) in that period.

Plotting the data, we get:

Figure 2: Data on Judicial Review of Federal Laws, 1789-1996

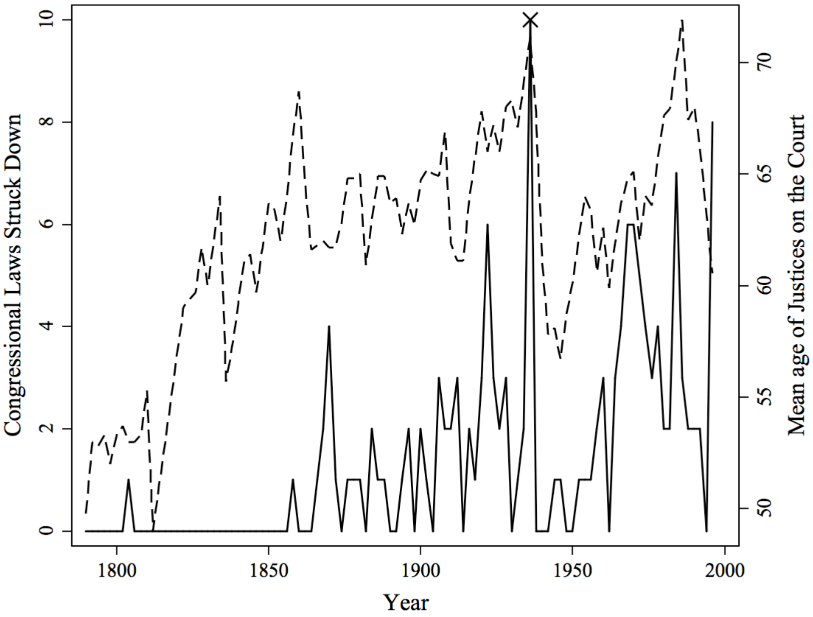


The regression looks like this:

```
> Fit<-lm(nulls~age+tenure+unified)
> summary(Fit)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7857 -1.0773 -0.3634  0.4238  6.9694

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.10340    2.54324  -4.759 6.57e-06 ***
age           0.21886    0.04484   4.881 4.01e-06 ***
tenure       -0.06692    0.06427  -1.041    0.300
unified       0.71760    0.45844   1.565    0.121
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.715 on 100 degrees of freedom
Multiple R-squared: 0.2324,Adjusted R-squared: 0.2093
F-statistic: 10.09 on 3 and 100 DF,  p-value: 7.241e-06
```

The question is whether these results are perhaps being driven by one or more outlying observations. Even if you don't have an encyclopedic knowledge of U.S. political history, a time-series plot of the relevant variables could give one some cause for concern. Note that there is a large "spike" around 1935-36 (the 74th Congress), corresponding to the New Deal / Court-packing crisis (this is denoted on the figure by the ✕ in Figure 3).

Figure 3: Data on Judicial Review of Federal Laws and Mean Supreme Court Age, by Year, 1789-1996



We should thus likely evaluate the extent to which this single observation (or any other) is driving our results. To do this, we can generate some residuals, etc.:

```
> FitResid<-(nulls - predict(Fit)) #residuals
> FitStandard<-rstandard(Fit) # standardized residuals
> FitStudent<-rstudent(Fit) #studentized residuals
> FitCooksD<-cooks.distance(Fit) #Cook's D
> FitDFBeta<-dfbeta(Fit) #DFBeta
> FitDFBetaS<-dfbetas(Fit) #DFBetaS
> FitCOVRATIO<-covratio(Fit) #COVRATIOs
```

With these in hand, we can begin to undertake some diagnostics for influential observations and/or outliers.

Studentized Residuals

Recall what was said earlier about the relationship between Studentized residuals and t-tests. Note here that if we examine the Studentized residual for the 74th Congress:

```
> FitStudent[74]
      74
4.415151
```

...it is equal to the t-statistic for a dummy variable coded 1 for that Congress, and 0 otherwise:

```
> Congress74<-rep(0,length=104)
> Congress74[74]<-1

> summary(lm(nulls˜age+tenure+unified+Congress74))

Residuals:
    Min      1Q  Median      3Q     Max
-2.4955 -0.9316 -0.3135  0.5047  7.0192

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.17290    2.37692  -4.280 4.33e-05 ***
age           0.18820    0.04177   4.505 1.82e-05 ***
tenure       -0.06356    0.05905  -1.076    0.284
unified       0.55159    0.42282   1.305    0.195
Congress74    7.14278    1.61779   4.415 2.58e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.576 on 99 degrees of freedom
Multiple R-squared: 0.3586,Adjusted R-squared: 0.3327
F-statistic: 13.84 on 4 and 99 DF,  p-value: 5.304e-09
```

"Bubble Plots" of Leverages and Residuals

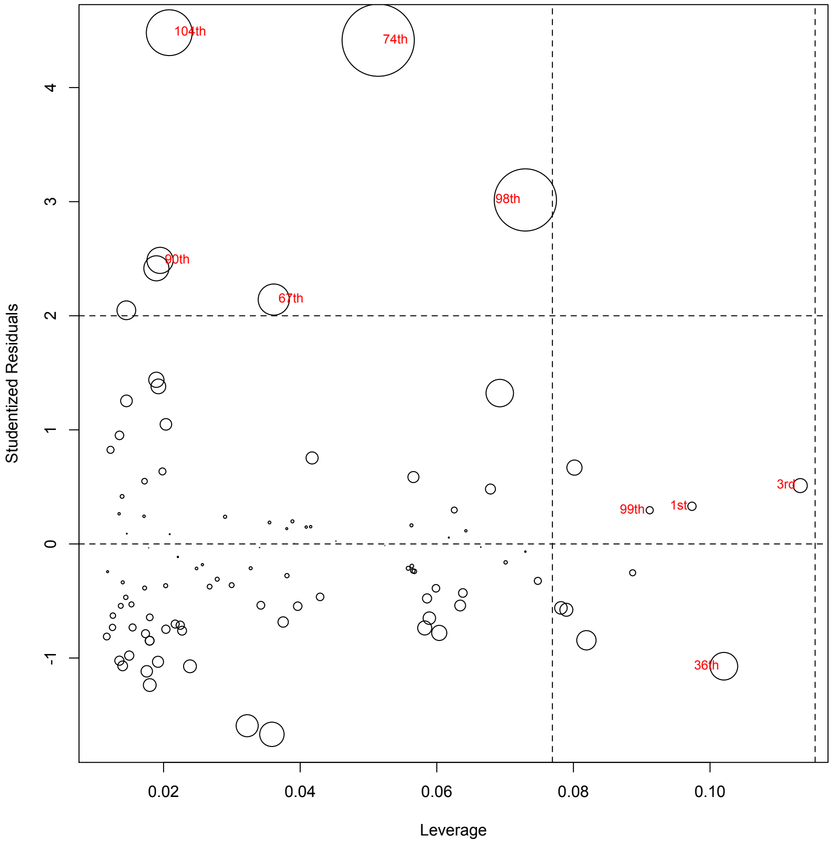
Fox (pp. 250-52) suggests the use of what he terms "bubble plots" in searching for influential observations. Recall that our main summary measure of influence – Cook's D – is a combination of an observation's leverage and its discrepancy, with the former a function of it's "hat" value h_i and the latter its Studentized residual. A "bubble plot" is thus just a scatterplot, with

- Each observation's leverage (that is, h_i) plotted on one axis,
- Its discrepancy (that is, its Studentized residual $\hat{u}'_i$) plotted on the other, and

- Symbols chosen such that the size of the symbol is proportional to Cook's D_i for that observation.

Such a plot looks like this:

Figure 4: "Bubble Plot" of h_i, $\hat{u}'_i$, and Cooks D_i



Note several things here:

- The size of the symbols is proportional to D_i, which in turn is a multiplicative function of the square of the Studentized residuals (the Y axis) and the leverage (the X axis). So, observations that are farther away from $Y = 0$, and/or have higher values of X, will have larger symbols.
- The plot also tells us whether the large influence of an observation is due to high discrepancy, high leverage, or both. So,

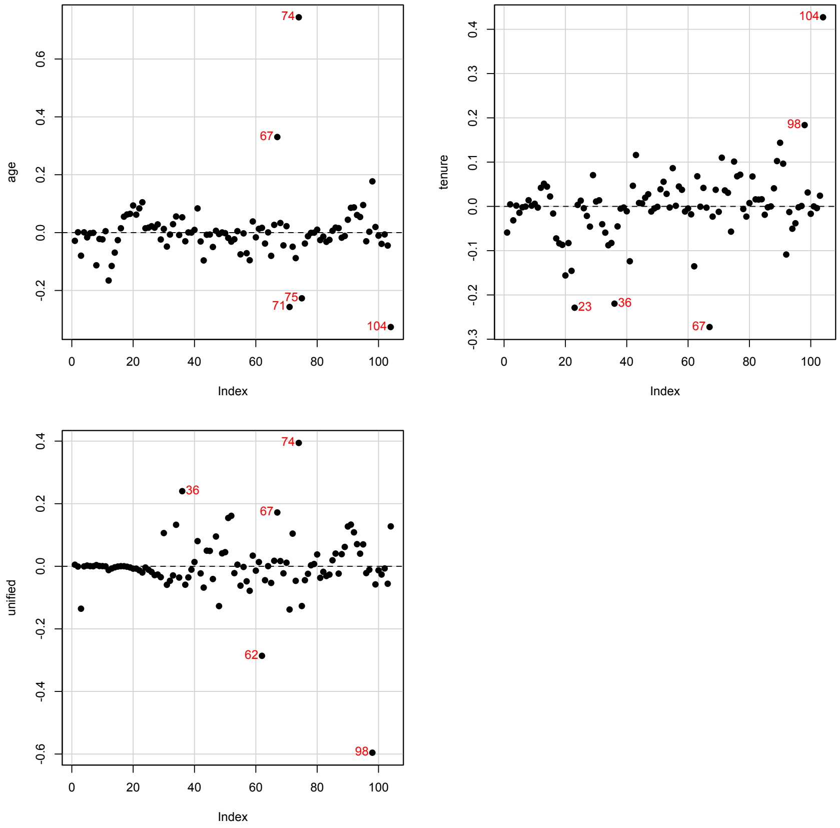
- The 104th Congress has relatively low leverage (that is, its values of $\mathbf{X}_i$ are "close to" overall levels of $\mathbf{X}$), but is very discrepant (in particular, it has a disproportionately large value of Y_i given $\mathbf{X}_i$), while
- The 74th and 98th Congresses demonstrate *both* high discrepancy and high leverage.

This plot thus suggests (doesn't *prove*, just suggests) that the 74th, 98th, and 104th (and possibly a few others) are relatively influential observations.

DFBETAs as Indicators of Influence

We can also plot the DFBETAs to assess whether particular observations are especially influential in the estimation of particular (that is, covariate-specific) parameters. While there are lots of clever ways to do this, perhaps the simplest is just to plot them against the "index" (the number of the observation, which here is the same as the number of the Congress):

Figure 5: Plot of DFBETAs for *Age*, *Tenure*, and *Unified*



Remember: Positive values of DFBETA$_{ki}$ indicate that the observation in question decreases the estimated coefficient $\hat{\beta}_k$, while negative values means that observation increases it. Accordingly, these results suggest that:

- The 74th Congress is particularly influential in the estimate for the `age` variable, and that its presence in the data has the effect of decreasing the estimated influence of `age`.
- The estimate for `tenure` is disproportionately influenced by the 104th Congress, and in the same way: the presence of that observation in the data decreases that variable's estimated effect.
- Finally, for `unified`, we see a large, positive DFBETA for the 74th Congress, and a large, negative one for the 98th.

COVRATIO Plots

Again, while there's no standard way of doing it, a common approach is to plot values of COVRATIO$_i$ against some other relevant indicator (say, an index value). Here, we again use the Congress:

Recall that COVRATIO gives us a summary measure of the extent to which the observation influences the precision of the standard error estimates; values greater than 1.0 mean that the observation decreases our standard error estimates, while those less than 1.0 increase them. This suggests that:

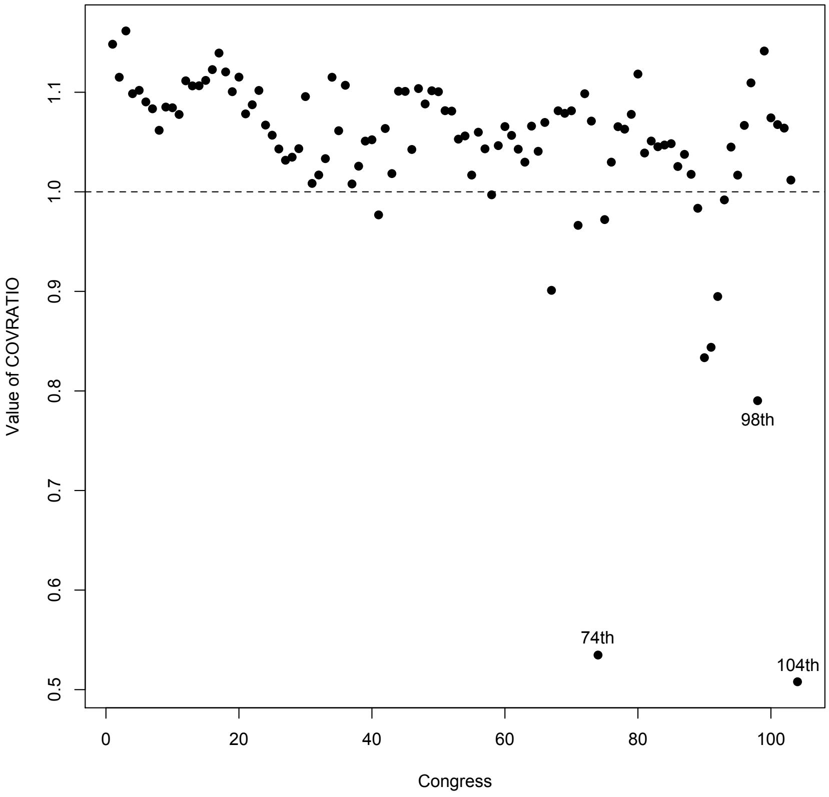
- The 74th, 104th, and (to a lesser extent) 98th Congresses all have disproportionately large influences on the variance-covariance estimates of our $\hat{\boldsymbol{\beta}}$s, and
- All three act to *increase* the size of our standard errors (that is, to decrease the precision of our estimates).

What To Do?

All of these analyses suggest that the observations for the 74th (1935-36), 98th (1983-84), and 104th (1995-96) Congresses are "outliers," that may or may not be disproportionately affecting our results. If we felt that – for whatever reason – these observations were somehow "wrong," and so should be omitted, we can do that:

```
> Outlier<-rep(0,104)
> Outlier[74]<-1
> Outlier[98]<-1
> Outlier[104]<-1
> DahlSmall<-Dahl[which (Outlier==0),]
```

Figure 6: Scatterplot of COVRATIO Values, by Congress



```
> summary(lm(nulls˜age+tenure+unified,data=DahlSmall))

Call:
lm(formula = nulls ˜ age + tenure + unified, data = DahlSmall)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3966 -0.8415 -0.1860  0.5367  4.4694

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.38536    1.99470  -5.206 1.08e-06 ***
age           0.19302    0.03512   5.496 3.13e-07 ***
tenure       -0.10069    0.04974  -2.024   0.0457 *
unified       0.76645    0.36069   2.125   0.0361 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 1.319 on 97 degrees of freedom
Multiple R-squared: 0.2578,Adjusted R-squared: 0.2349
F-statistic: 11.23 on 3 and 97 DF,  p-value: 2.167e-06
```

Interestingly,

- The new results aren't that different from the original ones:
  - The estimate for `age` is a bit smaller, as is its standard error,
  - That for `tenure` is also smaller, but much more precisely estimated,
  - The estimate for `unified` is actually a bit larger, and with a smaller standard error estimate as well.
- The R^2 and $R^2_{adj.}$ are both a little bit larger, and the RMSE is a (fair) bit smaller.
- All of these things suggest that the biggest influence the three outlying observations had was on the precision of the estimates;
- To the extent that they "drove up" the coefficient estimates, they did so only modestly.

Philosophical Issues (Or, What *is* an outlier?)

If/when we figure out that one or more of our observations are outliers, what then?

The answer depends largely (in fact, almost entirely) on *why* the observation(s) in question are "unusual."

Mistakes

Sometimes, data are just wrong: miscoded, mismeasured, mis-entered, or whatever. (In fact, going searching for outliers in otherwise well-behaved relationships is a first-rate way to "clean" data and check for such errors). If that's the case,

- ...and you can fix the error, then by all means do so, and use the fixed data.
- If, for whatever reason, you can't do so, then omit the offending observation. (As an aside: Such an instance is a perfect example of a good time to use any of the various missing-data imputation approaches; this is because – almost by construction – data that are missing in this way are "missing at random," and perhaps even "missing completely at random." We won't go into missing data imputation techniques much in this class, but you should keep it in mind).

Weird Observations

Perhaps there's no mistake: The data for a particular observation are just strange. In that case, the relevant question is, "Why are they so strange?" Two potential answers come to mind:

1. *The data are strange because something unusual/weird/singular happened to that data point.*

   - The next question then becomes, "Is that 'something' important for the theory/theories I'm trying to assess?" If the answer is "yes," then it may be time to reconsider your model specification in light of the "unusual" data.
   - If the answer is "no" (e.g., when the oddity in question is just a random event), then you can (and probably should) drop the offending observation from the analysis.

2. *The data are strange for no apparent reason.*

   - This is the hardest one to deal with, and there are no good answers. If it's possible, doing a little digging into the "history" of that data point is probably in order. Beyond that, however, it is really a judgement call.
   - Relatedly, this is probably an instance where a "rerun-the-model, footnote-the-results" sort of approach is in order, if only to check the robustness of your findings and maintain transparency.

In fact, there are never any hard-and-fast rules for situations like this. As in all such situations, theory – along with a thorough substantive knowledge of what you're studying – is the best guide.