3.3: Influential Observations and Outliers

Dr. Bean - Stat 5100

1 Why Care About Influential Observations/Outliers?

When we specify a model form of

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + \varepsilon$$

we assume that all observations in the data are generated from the same source (i.e. the theoretical line).

If we have observations that are **not** from the same source as the rest, OLS regression will try to **force** the model to fit the data, perhaps compromising the estimated coefficients and or inference.

Two things to watch for (not mutually exclusive):

- **Outliers** - observations with values of Y that are not well-explained by the model.
- **Influential Points** - observations that unduly influence the estimated coefficients b_k or predicted values $\hat{Y}$.

(Individual) Is it possible for a model outlier to not be reflected in a boxplot of Y? Explain why or why not.

Yes. A value of why can be exceptionally far away from the line *given its X-values*, while still being in a reasonable range for Y overall.

2 Ways to detect outliers or influential points

- (Primary) Scatterplots of X_k vs Y
- Other Diagnostics for Influential Observations
  - Hat matrix diagonals
  - DFBETAS
  - DFFITS
  - Cooks Distance
- Other Diagnostics for Outliers
  - Residuals
  - Studentized Residuals
  - Studentized Deleted Residuals

2.1 Hat Matrix Diagonals

Recall the linear algebra representation of the OLS regression model:

$$Y = X\beta + \varepsilon \qquad b = (X'X)^{-1}X'Y$$

$$\hat{Y} = Xb = X(X'X)^{-1}X'Y = HY$$

In other words, the predicted values of Y are simply linear combinations of the actual values of Y where each observation "weight" is determined by the X matrix.
Let $h_{i,l}$ be the element in row i and column l of H
- sometimes called "leverage" (influence of obs. i on its fitted value)

Since $\hat{Y} = HY$, then $\hat{Y}_i = \sum_{l=1}^{n} h_{i,l} Y_l$

(Individual) What would a "larger" diagonal element $h_{i,i}$ mean?

It means that the value of Y_i is more influential in its own prediction ($\hat{Y}_i$). We care about this because if the influence of a particular point is large enough, then the model is likely fitting that particular point at the sacrifice of the rest of the data.

We usually consider a point to be influential if:

- rule of thumb: $h_{i,i} > \frac{2p}{n}$ or $h_{i,i} > \frac{3p}{n}$
- can plot $h_{i,i}$ against observation number, with reference lines at $2p/n$ (SAS default) and/or $3p/n$

Another graphical diagnostic with $h_{i,i}$:

- leverage plots/partial regression/added variable plots); for X_1:
  1. Regress X_1 on $X_2, \ldots, X_{p-1}$ and obtain residuals $e_{X_1|X_2,\ldots,X_{p-1}}$
  2. Regress Y on $X_2, \ldots, X_{p-1}$ and obtain residuals $e_{Y|X_2,\ldots,X_{p-1}}$
  3. Plot $e_{Y|X_2,\ldots,X_{p-1}}$ vs. $e_{X_1|X_2,\ldots,X_{p-1}}$, and add regression line
     - slope will be b_1 from multiple regression model
     - Helps to visualize the marginal effect of adding X_1 in the model after already including all other X variables.
     - Influential points fall significantly farther away from the line than other points.
- (possible) modification here: point-size in leverage plot proportional to corresponding $h_{i,i}$ NOT shown in the SAS output provided in HO 3.3.1.
  - then this is called a proportional leverage plot
  - influential observations will be the points with big "bubbles" that appear to "pull" the regression line in their direction

2.2 DFBETAS

Provide a measure of how **different** ("DF") an estimate of β_k would be if we removed one observation from the data.

$$
\begin{aligned}
b_k &= \text{estimate of } \beta_k \text{ using full data} \\
b_{k(i)} &= \text{estimate of } \beta_k \text{ when observation } i \text{ is ignored} \\
MSE_{(i)} &= \text{Mean SS for error when observation } i \text{ is ignored} \\
C_{kk} &= k^{th} \text{ diagonal element of } (X'X)^{-1} \\
DFBETAS_{k(i)} &= \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} C_{kk}}}
\end{aligned}
$$

Interpreting DFBETAS:

- DFBETAS$_{k(i)}$ positive: obs. i "pulls" b_k up
- DFBETAS$_{k(i)}$ negative: obs. i "pulls" b_k down

How "large" to declare observation i "influential" on b_k?

- *Rough* rule of thumb:

  $|DFBETAS_{k(i)}| > 1$ for $n \leq 30$

  $|DFBETAS_{k(i)}| > 2/\sqrt{n}$ for $n > 30$ (SAS)

- Graphical diagnostics probably better for DFBETAS:
  - Histograms or boxplots for each k
  - Proportional leverage plot with "bubble" size prop. to DFBETAS$_{k(i)}$
  - Plot DFBETAS$_{k(i)}$ against obs. number for each k (Provided by SAS, unlike the others)

2.3 DFFITS

Similar to DFBETAS: how different would $\hat{Y}_i$ be
if observation i were not used to fit the model

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{i,i}}}$$

How large DFFITS to declare obs. i as influential on $\hat{Y}_i$?

- *Rough* rule of thumb:

  $|DFFITS_i| > 1$ for $n \leq 30$

  $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ for $n > 30$ (SAS)

- Good graphical diagnostics for DFFITS:
  - Plot DFFITS vs. Observation Number

- Plot Residuals vs. Predicted Values, with point sizes proportional to corresponding DFFITS_i

(DFBETAS_{ij} vs. DFFITS_i) vs. $h_{i,i}$

- somewhat related, so "conclusions" will quite often agree
- BUT: if two or more points exert "influence" together then the drop-one diagnostics (DFBETAS and DFFITS) may not detect them
  - these are leverage points - need to look at $h_{i,i}$

2.4 Cooks Distance

Kind of an overall measure of effect of obs. i on all of the $\hat{Y}_l$ values:

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{p \cdot \text{MSE}}$$

Diagnostics:

- Numerical:
  - simple: compare D_i with $4/n$ (SAS)
  - more useful: compare D_i with the $F_{p,n-p}$ distribution (See 3.3.1 pg 8 for example of how to do this "by hand")
    * percentile 10-20: little influence
    * percentile 50+: major influence
- Graphical: plot D_i (or percentile from $F_{p,n-p}$) vs. observation number i

2.5 Residuals

$e_i = Y_i - \hat{Y}_i$

Sometimes a large $|e_i|$ indicates an outlier

- not well-explained by fitted model
- but how "large" it needs to be depends on the residuals:
  - Recall $\varepsilon \sim N(0, \sigma^2)$, so $e_i \sim N(0, \sigma^2(1 - h_{ii}))$
    – because $\hat{Y} = HY$ results in $e = Y - HY = (I - H)Y$
  - Could compare e_i with the normal critical values, but need to estimate variance (including σ^2) $\Rightarrow$ normal approx. not appropriate; need Student's t

2.6 Studentized Residuals

$$r_i = \frac{e_i}{\sqrt{MSE \cdot (1 - h_{ii})}} \qquad (MSE = \hat{\sigma}^2)$$

If ε_i iid $N(0, \sigma^2)$, then the r_i follow the t_{n-p} distribution; diagnostics:

- Numerical: compare $|r_i|$ with upper $\alpha/2$ critical value of t_{n-p}
- Graphical: plot $\hat{Y}_i$ vs. r_i, with ref. lines at upper $\alpha/2$ critical value of t_{n-p}

2.7 Studentized Deleted Residuals

If obs. i really is an outlier, then including it in the data will inflate MSE
- So consider dropping it and re-calculating the studentized residual:

$$e_i^* = \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}} \qquad \text{(Text uses } t_i \text{ instead of } e_i^*\text{)}$$

2.8 Other Diagnostics (similar to studentized residuals)

- plot $\hat{Y}_i$ vs. e_i^*
- compare to $|e_i^*|$ to some critical value of t_{n-p} (for each of $i = 1, \ldots, n$)

  BUT: $\alpha =$ probability of type I error (calling obs. i outlier when it's not)

  - actually want α to be probability of *at least one* type I error in all n tests
    – a family-wise error rate
  - many ways to adjust the critical value; here, we'll use Bonferroni correction:

    compare $|e_i^*|$ to upper $\alpha/(2n)$ critical value of t_{n-p}

3 Remedial Measures for Influential Observations or Outliers

1. Look for:
   - typos in data (more common than would like to think)
   - fundamental differences in observations
     - drop obs. if from a different "population"
   - very skewed distributions of predictors
     - remember that in general, there is no assumption regarding the distribution of X's
     - sometimes transforming X will reduce influence of obs. with extreme values
2. Look at potential changes to model:
   - will a transformation "bring in" the observations?
   - should a curvilinear or other predictor be added?
     - look at leverage plot for the possible predictor

    - any trend suggests adding it to model

3. Could obtain estimates differently (instead of OLS, robust regression - more in Module 4):
    - LAD (least absolute deviation) regression
    - IRLS (iteratively reweighted least squares) regression