Stat 5100 Handout #10.b – Influential Observations and Outliers

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

There may be points (individual observations) that are:

- not "well-explained" by the model
  - may be called <u>outliers</u> (usually outliers in Y)

  } not necessarily lone points in boxplot of residuals

- unduly influencing the model fit (the b_k estimates or the $\hat{Y}$ predicted values)
  - may be called <u>influential observations</u> (usually outliers in X's)

Based on only a consideration of the residuals, one is not necessarily a subset of the other
- depends on the nature of influence and the sample size

Use both numerical and <u>graphical</u> diagnostics (to enhance, not replace, scatter plots):

- Main diagnostics for Influential Observations: - have undue influence on some aspect of model fit
  1. Hat matrix diagonals
  2. DFBETAS
  3. DFFITS
  4. Cooks Distance
- Main diagnostics for Outliers: - not well explained by model ↳ gives $\hat{Y}_i$
  ↳ $|Y_i - \hat{Y}_i|$ large
  5. (Residuals)
  6. (Studentized Residuals)
  7. Studentized Deleted Residuals ← H_0: obs. i not an outlier

<u>1. Hat matrix diagonals</u> (leverage)

Recall (from Ch. 5) that H projects Y down to column space of X:

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

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

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

What would a "larger" diagonal element $h_{i,i}$ mean?
– Y_i is more influential in determining $\hat{Y}_i$

How large must $h_{i,i}$ be to declare observation i as "influential"?

- rule of thumb: $h_{i,i} > \boxed{\frac{2p}{n}}$ or $h_{i,i} > \frac{3p}{n}$
- can plot $h_{i,i}$ against observation number, with reference lines at $2p/n$ and $3p/n$
  ↳ SAS ref. line in RStudent vs. Leverage Plot

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

- leverage plots (partial regression plots); for X_1: for each predictor

  1. Regress X_1 on $X_2, \ldots, X_{p-1}$ and obtain residuals $\boxed{e_{X_1|X_2,\ldots,X_{p-1}}}$ represents what's left over (information-wise) in X_1 after accounting for all other predictors
  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
     - useful as "added variable" plot – check for curvilinearity

skip-

- (possible) modification here: point-size in leverage plot proportional to corresponding $h_{i,i}$
  - 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. DFBETAS

"DF" means "different" here

- How different would est. of β_k's be without observation in 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:

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

$$|DFBETAS_{k(i)}| > \boxed{2/\sqrt{n}} \qquad \text{for } n > 30$$ SAS ref. line

- Graphical diagnostics probably better for DFBETAS:
  - Histograms or boxplots for each ~~j~~ k
  - Proportional leverage plot with "bubble" size prop. to DFBETAS$_{k(i)}$
  - SAS → Plot DFBETAS$_{k(i)}$ against obs. number for each k

    i   predictor

3. DFFITS

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

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

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

- *Rough* rule of thumb:

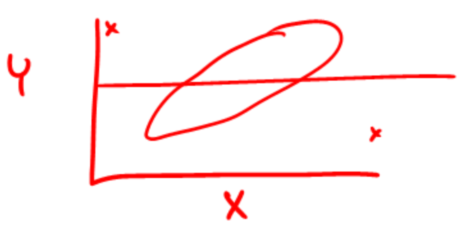
$$|DFFITS_i| > 1 \qquad \text{for } n \leq 30$$

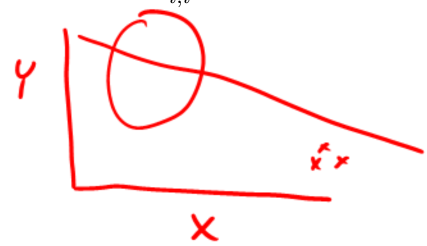
$$|DFFITS_i| > \boxed{2\sqrt{\tfrac{p}{n}}} \qquad \text{for } n > 30$$ → SAS ref line

- Good graphical diagnostics for DFFITS:
  - SAS → Plot DFFITS vs. Observation Number
  - Plot Residuals vs. Predicted Values, with point sizes proportional to corresponding DFFITS$_i$

(DFBETAS$_{ij}$ vs. DFFITS$_i$) vs. $h_{i,i}$ (leverage)

- 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 $h_{i,i}$





4. Cooks Distance

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

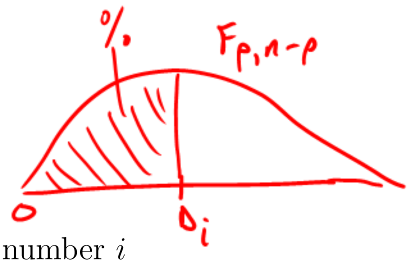
undue influence on overall model fit

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

Diagnostics:

- Numerical:
  - simple: compare D_i with $\boxed{4/n}$ → SAS ref. line
  - more useful: compare D_i with the $F_{p,n-p}$ distribution
    - percentile 10-20: little influence
    - percentile 50+: major influence

check 'by hand'



- Graphical: plot D_i (or percentile from $F_{p,n-p}$) vs. observation number i

SAS →

(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

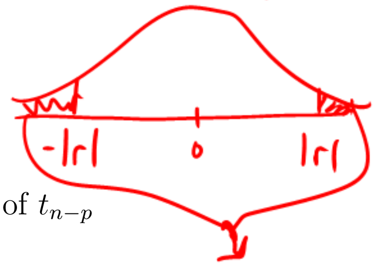
$Var(e_i) = \sigma^2 \cdot (1 - h_{ii})$

$SD(e_i) = \sqrt{Var(e_i)}$

$\hat{\sigma}^2 = MSE$

(6. Studentized Residuals)

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



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}

7. Studentized Deleted Residuals   RStudent (in SAS)

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:

$$\text{test statistic} \rightarrow e_i^* = \frac{e_i}{\sqrt{\underbrace{MSE_{(i)}}_{\text{MSE from model fit w/out obs. } i}(1 - h_{ii})}} \qquad \text{(Text uses } t_i \text{ instead of } e_i^*)$$

Diagnostics similar to Studentized Residuals:

H_0: obs. i not an outlier

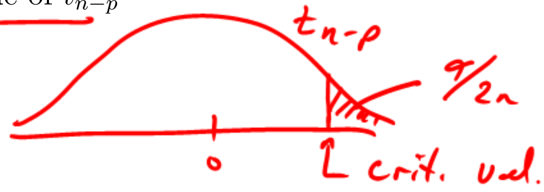
- plot $\hat{Y}_i$ vs. e_i^*
- compare to $|e_i^*|$ to some critical value of t_{n-p} (for each of $i = 1, \ldots, n$)

  BUT: α = 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 $\overbrace{\text{Bonferroni}}^{\text{use } \alpha/(\#\text{ tests}) \text{ instead of } \alpha}$ correction:

    (by hand in SAS)

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



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; see Ch. 11):
   - LAD (least absolute deviation) regression
   - IRLS (iteratively reweighted least squares) regression