Multiple Regression
Part 6: Influence and Outliers

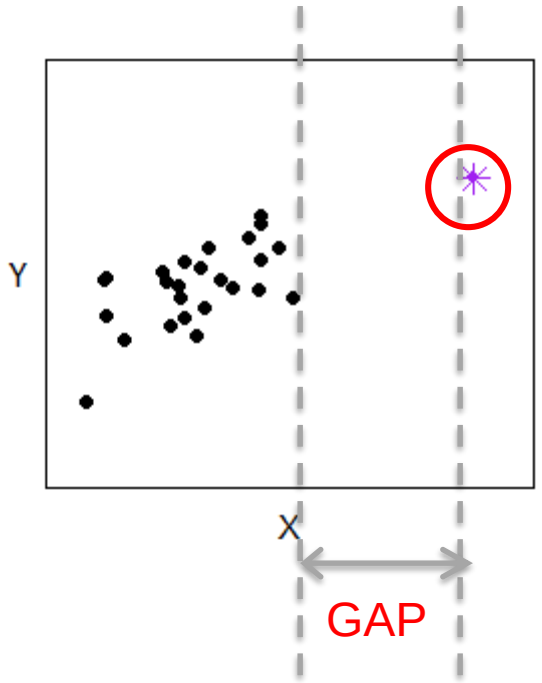
STAT 705: Regression and Analysis of Variance

Extreme Observations

- An ‘observation’ is one row in the data set
  - Includes the measured Y and all the X’s for the subject
  - We call this a ‘data point’
- Ways in which an observation can be extreme
  - The combination of X’s may be unusual
  - The value of Y, given the X’s, may be unusual
  - Both the X’s and the Y may be unusual
- Classify rows according to how they affect the model
  - Leverage, Outlier, Influence

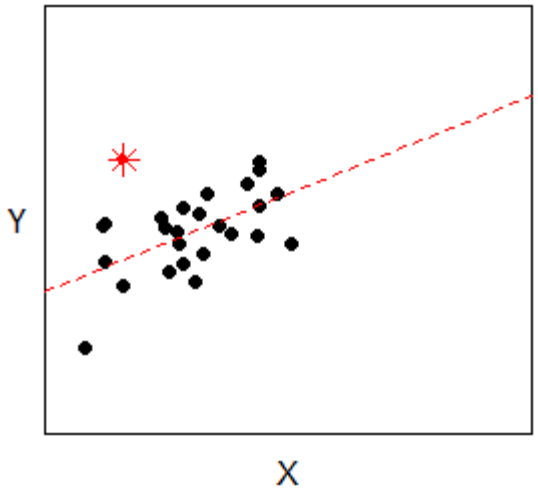
Leverage

- A data point has high leverage if its combination of values for the X's is unusual in relation to all the other rows in the data
- If there is only one X variable, points with high leverage appear separated (to the right or the left) of the other points.
- The Y value is not used in calculating leverage
- The point may or may not seem to follow the least squares line



Outliers

- Are observations that have unusually large or small Y values, in relation to the values of the X's that are recorded for subject
- Measured by the residuals
  - Large positive or large negative
  - Ordinary residuals may have large variance (estimated by MSE)
  - 'Large' residuals are large *relative to the MSE*



Studentized Residuals

- By assumption, $\varepsilon_i \sim N(0,\sigma^2)$
- Error variance (σ^2) is estimated by MSE
- 'Standardize' the residuals
  - Subtract the mean and divide by standard deviation
  - These are not independent, so distribution is not known
- 'Studentize' the residuals
  - Divide the standardized residuals by square root of (1 - leverage)
  - These follow an approximate t distribution

Identifying Outliers

From the theory of normal probability distributions

- Approximately 95% of observations should fall within 2 std. dev. of the mean
  - Observations that fall outside 2 std. dev. are potential outliers
- Approximately 99.7% of observations should fall with 3 std. dev. of the mean
  - Observations that fall outside 3 std. dev. are extreme outliers
- Studentized residuals have mean 0 and std. dev. 1
  - Potential outliers have studentized residual > 2 or < -2
  - Extreme outliers have studentized residual > 3 or < -3

Influence

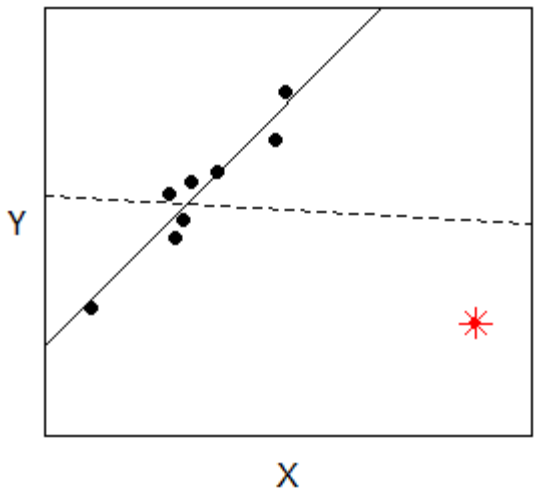
- The influence of a point is a measure of how much the fitted model would change if the point was removed from the data set
- The influence can be measured as change in
  - the fitted values
  - estimates for the individual coefficients
- A point can be influential because
  - it has high leverage
  - it is an outlier
  - both high leverage and outlier

Visualize Influence

The point marked with a red star is influential

If we exclude this point

- Y = 1.17 + 1.27X
- R^2 = 91.3%
- RMSE = 0.92



If we include this point

- Y = 14.36 - 0.07X
- R^2 = 0.7%
- RMSE = 3.55

Identifying Influential Points

- Approach: "Leave one out"
  - Omit a single observation, re-fit the model and evaluate how inference changes
- Criteria for evaluating change in inference:
  - DFFITS - Influence on Single Fitted Values
  - Cook's distance - Influence on all Fitted Values
  - DFBETAS – Influence on Regression Coefficients
- All of these can be generated in SAS

```
proc reg data=fat;
    model bodyfat = triceps midarm /influence r;
run;
```

Influence: SAS Output

This is the body fat data.
This table is generated by the 'influence' option on the model statement.
Each observation (row) in the data has a value for each measure.
(Some columns have been deleted.)

Output Statistics								
Obs	Residual	Cook's D	RStudent	Hat Diag H	DFFITS	DFBETAS		
						Intercept	triceps	midarm
1	-1.8481	0.048	-0.8084	0.1785	-0.3768	-0.0142	0.3087	-0.2152
2	3.4606	0.039	1.4734	0.0538	0.3514	0.0058	-0.0755	0.0837
3	-2.8462	0.478	-1.5271	0.3988	-1.2439	1.0563	0.0525	-1.0572
19	-3.0128	0.036	-1.2703	0.0648	-0.3343	-0.1127	0.1537	-0.0321
20	0.9583	0.003	0.3839	0.0501	0.0881	0.0140	-0.0006	-0.0024

Cook's D

Studentized Residuals

Leverage

DFFITS

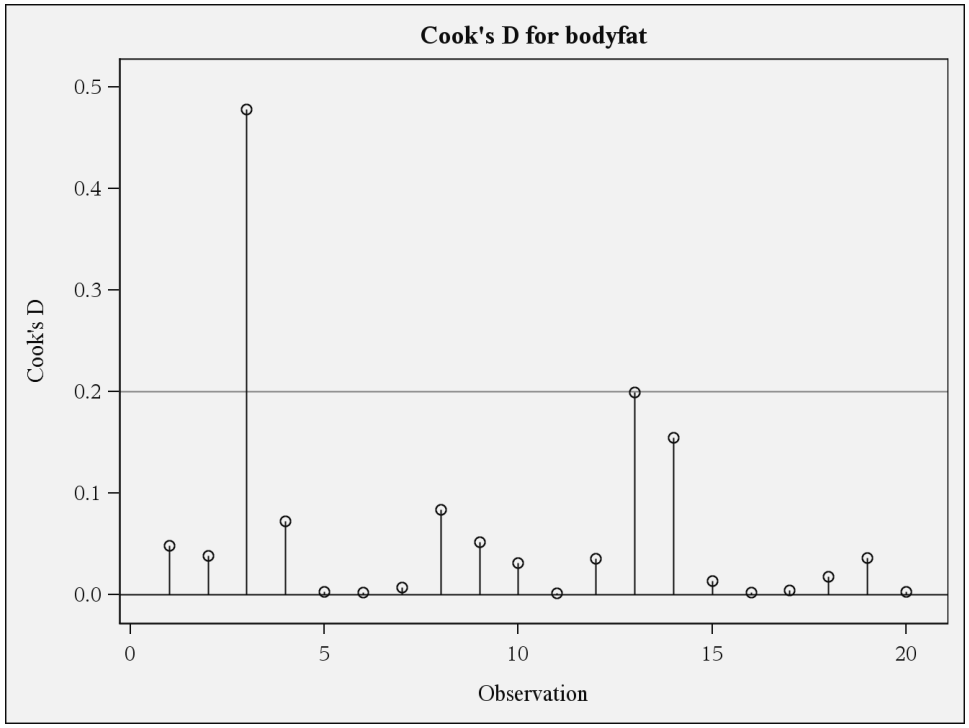
DFBETAS for β_0, β_1, and β_2

Evaluate Each Observation

- Studentized residuals: Values > 2 or < -2
  - Unusual Y value for the observed X's
  - Values > 3 or < -3 are extreme outliers
- Leverage: Values $> \frac{2}{n} \cdot (\#\,\text{parameters})$
  - Unusual combination of X values
- DFBETAS: Values $> \frac{2}{\sqrt{n}}$
  - Observation influences the specific parameter estimate

Cook's D

- Results for Cook's distance are shown in a graph
- Taller points above the horizontal line are more influential
- Before analyzing these data, we should make sure there are no errors in observation 3



What You Should Know

- How to generate influence measures in SAS
- Identify influential observations & outliers

Notes:

- These techniques simply identify unusual points in the data.
- We do NOT automatically remove influential points.
- First make sure there are no errors in the data.
- Then determine if there are any unusual conditions under which the observation was collected (that might explain why this point is different).
- Ultimately, it is the experience and knowledge of the researcher that dictates whether to keep or remove a data point.