Unusual and influential data

Chapter 11

- Unusual data points:
  - What to do with them?
  - Leverage: hat values
  - Outliers: standardized/studentized residuals
  - Influence: Cook's distance
  - Added variable plots can help find clusters of points that are jointly influential

What to do with unusual data?

- Neither ignore them, nor throw them out without thinking
- Check for data entry errors
- Think of reasons why observation may be different
- Change the model
- Fit model with and without the observations to see the effect
- Robust regression

Unusual data points

- Univariate outlier:
  - Unusual value for one of the X's or for Y
- Leverage point: point with unusual combination of independent variables
- Regression outlier:
  - Large residual (in absolute value)
  - The value of Y *conditional* on X is unusual
- Influential point: points with large influence on the regression coefficients
- Influence = Leverage $\times$ 'Outlyingness'
- See examples

Leverage points

5 / 15

Leverage

- Leverage is measured by the so-called "hat values"
- Hat values: $\hat{Y}_j = h_{1j}Y_1 + \cdots + h_{nj}Y_n = \sum_{i=1}^{n} h_{ij}Y_i$
- In matrix notation, h_{ij} are the elements of the hat matrix $H = X(X^TX)^{-1}X^T$. H is called the hat matrix since $\hat{Y} = HY$.
- The weight h_{ij} captures the contribution of Y_i to the fitted value $\hat{Y}_j$
- The number $h_i \equiv h_{ii} = \sum_{j=1}^{n} h_{ij}^2$ summarizes the leverage of Y_i on *all* fitted values
- Note the dependent variable Y is not involved in the computation of the hat values

6 / 15

Leverage

- Range of the hat values: $1/n \leq h_i \leq 1$
- Average of the hat values: $\bar{h} = (k+1)/n$
- Rule of thumb: leverage is large is $h_i > 2(k+1)/n$. Draw a horizontal line at this value
- R-function: `hatvalues()`
- See example

7 / 15

Regression outliers

8 / 15

Residuals

- Residuals: $E_i = Y_i - \hat{Y}_i$. R-function `resid()`.
- Even if statistical errors have constant variance, the residuals do not have constant variance: $V(E_i) = \sigma_\epsilon^2(1 - h_i)$.
- Hence, high leverage points tend to have small residuals, which makes sense because these points can 'pull' the regression line towards them.

9 / 15

Standardized/studentized residuals

- We can compute versions of the residuals with constant variance:
  - Standardized residuals E_i' and studentized residuals E_i^*:

$$E_i' = \frac{E_i}{S_E\sqrt{1-h_i}} \quad \text{and} \quad E_i^* = \frac{E_i}{S_{E(-i)}\sqrt{1-h_i}}.$$

  - Here $S_{E(-i)}$ is an estimate of σ_ϵ when leaving out the ith observation.
  - R-functions `rstandard()` and `rstudent()`.

10 / 15

Testing for outliers

- Look at studentized residuals by eye.
- If the model is correct, then E_i^* has t-distribution with $n-k-2$ degrees of freedom.
- If the model is true, about 5% of observations will have studentized residuals outside of the ranges $[-2, 2]$. It is therefore reasonable to draw horizontal lines at ± 2.
- We can use Bonferroni test to determine if largest studentized residual is an outlier: divide your cut-off for significant p-values (usually 0.05) by n.

11 / 15

Influential points

12 / 15

Influence

- Influence = Leverage $\times$ 'Outlyingness'
- Cook's distance:

$$D_i = \frac{h_i}{1-h_i} \times \frac{E_i'^2}{k+1}$$

- Cook's distance measures the difference in the regression estimates when the ith observation is left out
- Rule of thumb: Cook's distance is large if $D_i > 4/(n-k-1)$
- R-command: `cooks.distance()`

13 / 15

Joint influence

- See example
- Use added variable plots to detect this

14 / 15

Some more useful R-commands

- `indentify()`: to identify points in the plot
- `plot(m)`: gives 4 plots:
  - Residuals against fitted values
  - QQ-plot of standardized residuals
  - Scale-location plot
  - Cook's distance plot
- `influence.measures(m)`: contains various measures of influence.

15 / 15