7. Model diagnostics and unusual and influential data

Unusual and influential data

- Outline:
  - What to do with them?
  - Leverage: hat values
  - Outliers: standardized/studentized residuals
  - Influence: Cook's distance
  - Checking model assumptions

Outliers and the ozone layer

- In 1985, data gathered by the British Antarctic Survey showed that ozone levels for Antarctica had dropped 10% below normal January levels.
- This was surprising, as the Nimbus 7 satellite hadn't recorded such low ozone concentrations.
- After examining the satellite data more closely, it turned out that the satellite had been recording these low concentration levels for 9 years. But they were being treated as outliers by a computer program and discarded!
- The damage to our atmosphere went undetected and untreated for 9 years because outliers were discarded without being examined. So don't just toss out outliers, as they may be the most valuable members of a dataset!

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 (will be discussed later)

4 / 21

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

5 / 21

Leverage points

Leverage

- Leverage point: point with unusual combination of the independent variables
- Leverage is measured by the so-called "hat values"
- These are entries from the hat matrix $P = X(X^TX)^{-1}X^T$; $\hat{Y} = PY$
- $\hat{Y}_j = P_{j1}Y_1 + \cdots + P_{jn}Y_n = \sum_{i=1}^{n} P_{ji}Y_i$
- The weight P_{ji} captures the contribution of Y_i to the fitted value $\hat{Y}_j$
- Since $P^TP = P$, we have $\sum_{j=1}^{n} P_{ji}^2 = P_{ii}$. The value P_{ii} summarizes the contribution of Y_i to *all* fitted values.
- Note the dependent variable Y is not involved in the computation of the hat values

Leverage

- Range of the hat values: $1/n \leq P_{ii} \leq 1$
- Average of the hat values: mean$(P_{11}, \ldots, P_{nn}) = (p+1)/n$, where p is the number of independent variables in the model
- Rough rule of thumb: leverage is large if $P_{ii} > 2(p+1)/n$. Draw a horizontal line at this value
- R-function: `hatvalues()`
- See example

Residuals

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

10 / 21

Standardized/studentized residuals

- We can compute versions of the residuals with constant variance:
  - Standardized residuals $\hat{\epsilon}'_i$ and studentized residuals $\hat{\epsilon}^*_i$:

$$\hat{\epsilon}'_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - P_{ii}}} \quad \text{and} \quad \hat{\epsilon}^*_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(-i)}\sqrt{1 - P_{ii}}}.$$

  - Here $\hat{\sigma}_{(-i)}$ is an estimate of σ when leaving out the ith observation.
  - R-functions `rstandard()` and `rstudent()`.

11 / 21

Testing for outliers

- Look at studentized residuals by eye.
- If the model is correct, then $\hat{\epsilon}_i^*$ has t-distribution with $n-p-2$ 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 significance level α by n.

12 / 21

Influential points

13 / 21

Influence

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

$$D_i = \frac{P_{ii}}{1-P_{ii}} \times \frac{\hat{\epsilon}_i'^2}{p+1}$$

- Cook's distance measures the difference in the regression estimates when the ith observation is left out:
  - $D_i = (\hat{\beta}_{(-i)} - \hat{\beta})^T \mathsf{Var}^{-1}(\hat{\beta})(\hat{\beta}_{(-i)} - \hat{\beta})$
  - $D_i = \sum_{j=1}^n (\hat{Y}_{(-i)j} - \hat{Y}_j)^2 / (p\hat{\sigma}^2)$
- Rough rule of thumb: Cook's distance is large if $D_i > 4/(n-p-1)$
- R-command: `cooks.distance()`
- Beware of jointly influential points (see example)

14 / 21

Some more useful R-commands

- `indentify()`: to identify points in the plot
- `plot(m, which=c(1:5))` gives 5 plots:
  - Tukey-Anscombe plot: Residuals versus fitted values
  - QQ-plot of standardized residuals
  - Scale-location plot: Square root of standardized residuals versus fitted values
  - Cook's distance
  - Standardized residuals versus leverage

Checking model assumptions

Linearity

- Assumption: $E(\epsilon_i) = 0$ for all $i = 1, \ldots, n$
- Plots to use:
  - Tukey-Anscombe plot (residuals versus fitted values; **most important diagnostic plot!**)
  - Plot residuals versus each independent variable
  - Added variable plot for each independent variable
- Possible solutions:
  - Transform variables
  - Adapt model (add variables or quadratic terms)

Added variable plot

- Recall how to make an added variable plot for X_j:
  - Regress Y on all independent variables except for X_j. Obtain the residuals $\hat{\epsilon}_i^{(1)}$, $i = 1, \ldots, n$
  - Regress X_j on all the other independent variables. Obtain the residuals $\hat{\epsilon}_i^{(2)}$, $i = 1, \ldots, n$
  - Plot $\hat{\epsilon}_i^{(1)}$ versus $\hat{\epsilon}_i^{(2)}$, $i = 1, \ldots, n$

18 / 21

Constant variance

- Assumption: $\text{Var}(\epsilon_i) = \sigma^2$ for all $i = 1, \ldots, n$
- Plots to use:
  - Tukey-Anscombe plot
  - Plot residuals versus each independent variable
  - Added variable plot for each independent variable
- Possible solutions:
  - Use weighted least squares when the form of non-constant variance is known
  - Transform the dependent variable

19 / 21

Uncorrelated errors

- Assumption: Cor$(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$
- Plots to use:
  - When time series data: plot residuals versus time, and plot successive residuals
  - Think about data collection process
- Possible solutions:
  - Generalized least squares

20 / 21

Normality

- Assumption: $\epsilon_i \sim N(0, \sigma^2)$ for all $i = 1, \ldots, n$
- Plots to use:
  - QQ plot of residuals
- Consequences of nonnormality:
  - Levels are still valid when sample size is large, but there may be more efficient procedures than least squares
- Possible solutions:
  - Do nothing
  - Transform dependent variable

21 / 21