7. Unusual and influential data

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unusual and influential data</td>
<td>2</td>
</tr>
<tr>
<td>What to do with unusual data?</td>
<td>3</td>
</tr>
<tr>
<td>Unusual data points</td>
<td>4</td>
</tr>
<tr>
<td><strong>Leverage points</strong></td>
<td>5</td>
</tr>
<tr>
<td>Leverage</td>
<td>6</td>
</tr>
<tr>
<td>Leverage</td>
<td>7</td>
</tr>
<tr>
<td><strong>Regression outliers</strong></td>
<td>8</td>
</tr>
<tr>
<td>Residuals</td>
<td>9</td>
</tr>
<tr>
<td>Standardized/studentized residuals</td>
<td>10</td>
</tr>
<tr>
<td>Testing for outliers</td>
<td>11</td>
</tr>
<tr>
<td><strong>Influential points</strong></td>
<td>12</td>
</tr>
<tr>
<td>Influence</td>
<td>13</td>
</tr>
<tr>
<td>Joint influence</td>
<td>14</td>
</tr>
<tr>
<td>Some more useful R-commands</td>
<td>15</td>
</tr>
</tbody>
</table>
Unusual and influential data

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

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)

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

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 $H = X(X^T X)^{-1} X^T$; $\hat{Y} = HY$
- $\hat{Y}_j = h_{j1} Y_1 + \cdots + h_{jn} Y_n = \sum_{i=1}^n h_{ji} Y_i$
- The weight $h_{ji}$ captures the contribution of $Y_i$ to the fitted value $\hat{Y}_j$
- The number $h_{ii} = \sum_{j=1}^{n} h_{ji}^2$ 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: $1/n \leq h_i \leq 1$
- Average of the hat values: $\bar{h} = (p+1)/n$, where $p$ is the number of independent variables in the model
- Rough rule of thumb: leverage is large is $h_i > 2(p+1)/n$. Draw a horizontal line at this value
- R-function: hatvalues()
- See example

Regression outliers

Residuals

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

- We can compute versions of the residuals with constant variance:
  - Standardized residuals \( \hat{\epsilon}'_i \) and studentized residuals \( \hat{\epsilon}^*_i \):
    \[
    \hat{\epsilon}'_i = \frac{\hat{\epsilon}_i}{\hat{\sigma} \sqrt{1 - h_i}} \quad \text{and} \quad \hat{\epsilon}^*_i = \frac{\epsilon_i}{\hat{\sigma}_{(-i)} \sqrt{1 - h_i}}.
    \]
  - Here \( \hat{\sigma}_{(-i)} \) is an estimate of \( \sigma \) when leaving out the \( i \)th observation.
  - R-functions rstandard() and rstudent().

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 \( \pm 2 \).
- 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 \( n \).

Influential points

Influence

- Influence = Leverage \times ‘Outlyingness’
- Cook’s distance:
  \[
  D_i = \frac{h_i}{1 - h_i} \times \frac{\hat{\epsilon}^2_i}{p + 1}
  \]
- Cook’s distance measures the difference in the regression estimates when the \( i \)th observation is left out
- Rough rule of thumb: Cook’s distance is large if \( D_i > 4/(n - p - 1) \)
- R-command: cooks.distance()
Some more useful R-commands

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