

## 7. Model diagnostics and unusual and influential data

Unusual and influential data . . . . .	2
What to do with unusual data? . . . . .	3
Unusual data points . . . . .	4
<b>Leverage points</b> . . . . .	<b>5</b>
Leverage . . . . .	6
Leverage . . . . .	7
<b>Regression outliers</b> . . . . .	<b>8</b>
Residuals . . . . .	9
Standardized/studentized residuals . . . . .	10
Testing for outliers . . . . .	11
<b>Influential points</b> . . . . .	<b>12</b>
Influence . . . . .	13
Some more useful R-commands . . . . .	14
<b>Checking model assumptions</b> . . . . .	<b>15</b>
Linearity . . . . .	16
Added variable plot . . . . .	17
Constant variance . . . . .	18
Uncorrelated errors . . . . .	19
Normality . . . . .	20

## Unusual and influential data

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

2 / 20

## 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)

3 / 20

## 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

4 / 20

**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  $P = X(X^T X)^{-1} X^T$ ;  $\hat{Y} = PY$
- $\hat{Y}_j = P_{j1}Y_1 + \dots + P_{jn}Y_n = \sum_{i=1}^n P_{ji}Y_i$
- The weight  $P_{ji}$  captures the contribution of  $Y_i$  to the fitted value  $\hat{Y}_j$
- Since  $P^T P = P$ , 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

6 / 20

**Leverage**

- Range of the hat values:  $1/n \leq P_{ii} \leq 1$
- Average of the hat values:  $\text{mean}(P_{11}, \dots, P_{nn}) = (p + 1)/n$ , 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

7 / 20

**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.

9 / 20

## Standardized/studentized residuals

- We can compute versions of the residuals with constant variance:

- ◆ Standardized residuals  $\hat{\epsilon}'_i$  and studentized residuals  $\hat{\epsilon}^*_i$ :

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

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

10 / 20

## 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 significance level  $\alpha$  by  $n$ .

11 / 20

## Influential points

12 / 20

### 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  $i$ th observation is left out:
  - ◆  $D_i = (\hat{\beta}_{(-i)} - \hat{\beta})^T \text{Var}^{-1}(\hat{\beta})(\hat{\beta}_{(-i)} - \hat{\beta})$
  - ◆  $D_i = \sum_{j=1}^n (\hat{Y}_{(-i)j} - \hat{Y}_j)^2 / (p\hat{\sigma}^2)$
- 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)

13 / 20

## Some more useful R-commands

- `identify()`: 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

14 / 20

## Checking model assumptions

15 / 20

### Linearity

- Assumption:  $E(\epsilon_i) = 0$  for all  $i = 1, \dots, n$
- 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)

16 / 20

### 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  $\hat{\epsilon}_i^{(1)}$ ,  $i = 1, \dots, n$
  - ◆ Regress  $X_j$  on all the other independent variables. Obtain the residuals  $\hat{\epsilon}_i^{(2)}$ ,  $i = 1, \dots, n$
  - ◆ Plot  $\hat{\epsilon}_i^{(1)}$  versus  $\hat{\epsilon}_i^{(2)}$ ,  $i = 1, \dots, n$

17 / 20

### Constant variance

- Assumption:  $\text{Var}(\epsilon_i) = \sigma^2$  for all  $i = 1, \dots, n$
- 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

18 / 20

### Uncorrelated errors

- Assumption:  $\text{Cor}(\epsilon_i, \epsilon_j) = 0$  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

19 / 20

### Normality

- Assumption:  $\epsilon_i \sim N(0, \sigma^2)$  for all  $i = 1, \dots, n$
- 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

20 / 20