36-617: Applied Linear Models

Regression Basics

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Reading, HW, etc.

- Reading
  - Quiz on Ch 3 today (in class!)
  - For next week Ch 5 (Skip Ch 4 for now)
- HW 02 due tonight 1159 pm
- HW 03 (more on Ch 3 and IDMRAD) out later today

Outline

- The Anscombe plots – why diagnostics matter
- Monday: The standard R diagnostic plots:
  - Residuals,QQ plots, Scale-location, Leverage
  - Recommendations
- Wednesday: Transformations
  - Intuitive / substantive theory-driven
  - Variance stabilization
  - "Automagic": Box-Cox
  - Perspective and recommendations

…and Examples…

The Anscombe plots: why summary() statistics are not enough

- Anscombe data set #1

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x1            0.5001     0.1179   4.241  0.00217 **

Residual standard error: 1.237 on 9 deg of freedom
Multiple R-Squared: 0.6665, Adj R-squared: 0.6295
F-statistic: 17.99 on 1 and 9 DF, p-value: 0.002170
```

- Anscombe data set #3

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0025     1.1245   2.670  0.02562 *
X3            0.4997     0.1179   4.239  0.00218 **

Residual standard error: 1.236 on 9 deg of freedom
Multiple R-Squared: 0.6663, Adj R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF, p-value: 0.002176
```

- Anscombe data set #2

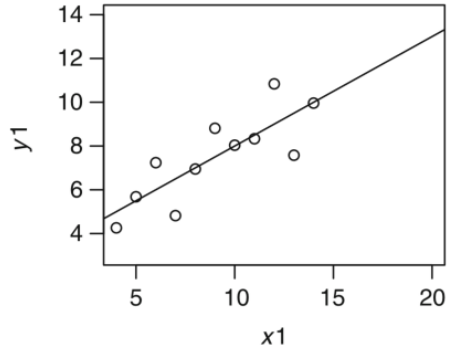
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576 *
x2             0.500      0.118   4.239  0.00218 **

Residual standard error: 1.237 on 9 deg of freedom
Multiple R-Squared: 0.6662, Adj R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF, p-value: 0.002179
```

- Anscombe data set #4

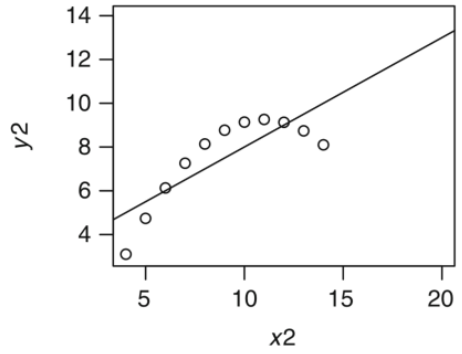
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671  0.02559 *
x4            0.4999     0.1178   4.243  0.00216 **

Residual standard error: 1.236 on 9 deg of freedom
Multiple R-Squared: 0.6667, Adj R-squared: 0.6297
F-statistic: 18 on 1 and 9 DF, p-value: 0.002165
```

The Anscombe plots: why summary() statistics are not enough

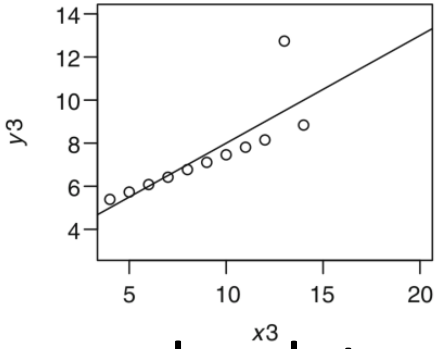
- Anscombe data set #1



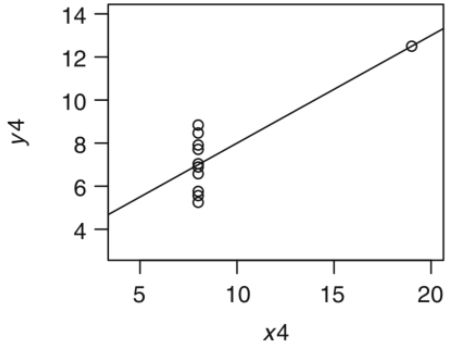
- Anscombe data set #2



- Anscombe data set #3

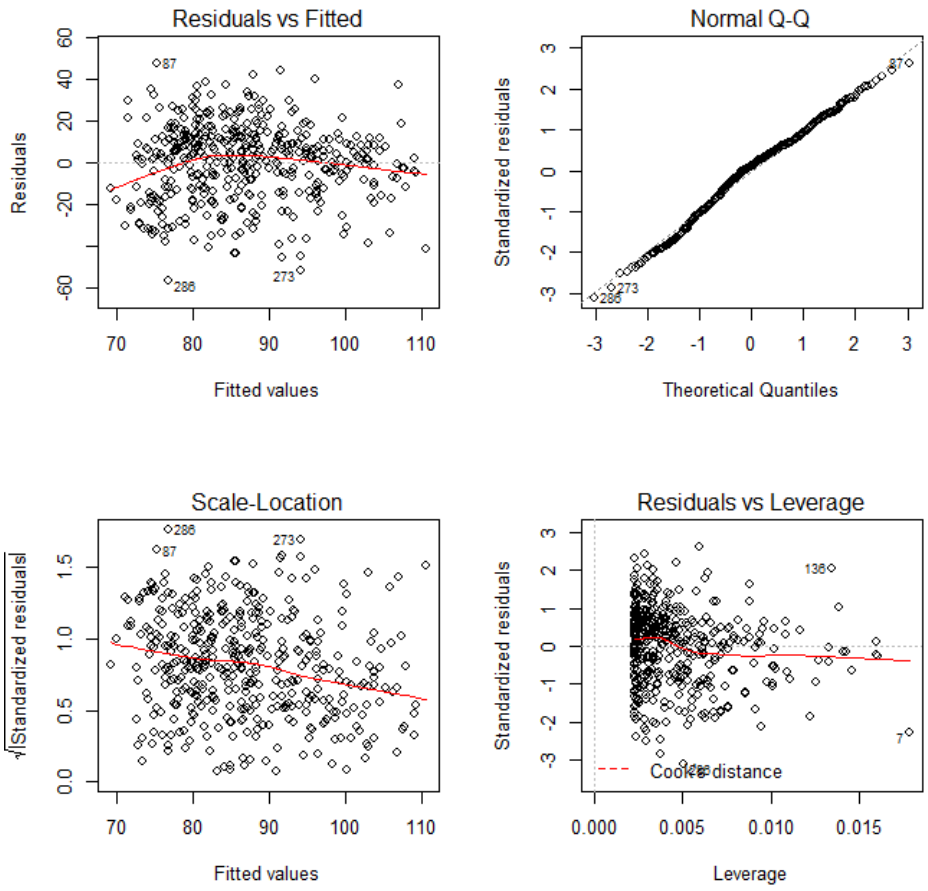


- Anscombe data set #4



The standard R residual plots: casewise diagnostics

```
> kidiq <- read.csv("kidiq.csv")
> fit.lm.1 <- lm(kid.score ~ mom.iq,
+ data=kidiq)
> summary(fit.lm.1)

             Est StdErr t-val Pr(>|t|)
(Intcpt) 25.80    5.92   4.36 1.63e-05 ***
mom.iq    0.61    0.06 10.42  < 2e-16 ***

Residual std err: 18.27 on 432 d.f.
R-sq:  0.201, Adj R-squared:  0.1991
F: 108.6 on 1 and 432 DF,  p < 2.2e-16

> par(mfrow=c(2,2))
> plot(fit.lm.1)
```



Residuals…

- Last time we saw, for the ***fitted values*** $\hat{y}_i$,

$$\begin{aligned} \hat{y}_i &= \sum_{j=1}^{n} h_{ij} y_j, \qquad \text{where } h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX} \\ \text{so Var}\,(\hat{y}_i) &= \sum_{i=1}^{n} h_{ij}^2 \sigma^2 \;=\; h_{ii}\sigma^2 \quad \text{(ex!)} \end{aligned}$$

- Therefore, if we define ***residuals*** $\hat{e}_i = y_i - \hat{y}_i$,

$$\begin{aligned} \hat{e}_i &= y_i - \hat{y}_i \;=\; (1 - h_{ii}) y_i - \sum_{j \neq i} h_{ij} y_j \\ \text{Var}\,(\hat{e}_i) &= \text{Var}\,(y_i - \hat{y}_i) \;=\; \text{Var}\,(y_i) - 2\text{Cov}\,(y_i, \hat{y}_i) + \text{Var}\,(\hat{y}_i) \\ &= \sigma^2 - 2\text{Cov}\,(y_i, \sum_{j=1}^{n} h_{ij} y_j) + h_{ii}\sigma^2 \\ &= \sigma^2 - 2\text{Cov}\,(y_i, h_{ii} y_i) + h_{ii}\sigma^2 \\ &= \sigma^2 - 2h_{ii}\sigma^2 + h_{ii}\sigma^2 \\ &= (1 - h_{ii})\sigma^2 \end{aligned}$$

Standardized Residuals…

- So far…

$$\begin{aligned} \hat{y}_i &= \sum_{j=1}^{n} h_{ij} y_j & \text{Var}(\hat{y}_i) &= h_{ii}\sigma^2 \\ \hat{e}_i &= y_i - \hat{y}_i = (1-h_{ii})y_i - \sum_{j\neq i} h_{ij} y_j & \text{Var}(\hat{e}_i) &= (1-h_{ii})\sigma^2 \end{aligned}$$

and we can calculate

$$\frac{1}{n}\sum_i h_{ii} = \frac{1}{n}\sum_i \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX}\right) = \frac{2}{n}$$

  - So h_{ii} tend to be small, on average (but they still vary)…
  - A common way to look for outliers is to plot the *standardized residuals*

$$r_i = \hat{e}_i / \sqrt{(1-h_{ii})S^2} \quad (S^2 = \tfrac{1}{n-2} RSS \text{ estimates } \sigma^2)$$

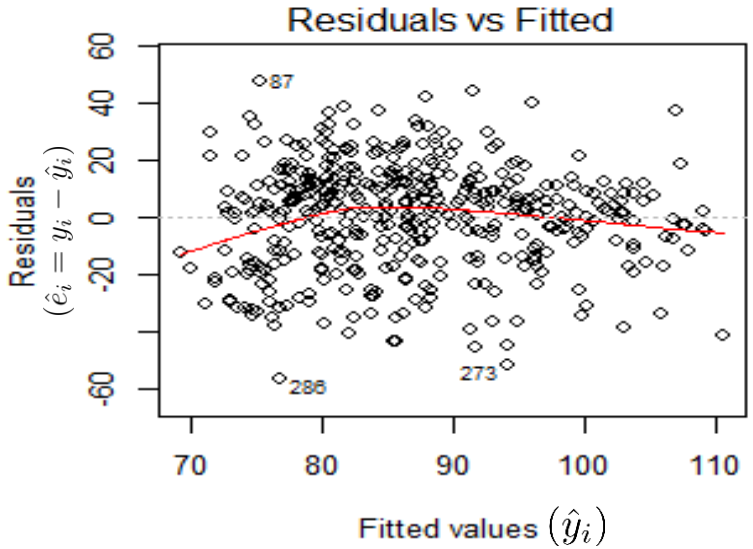
Normality of Residuals…

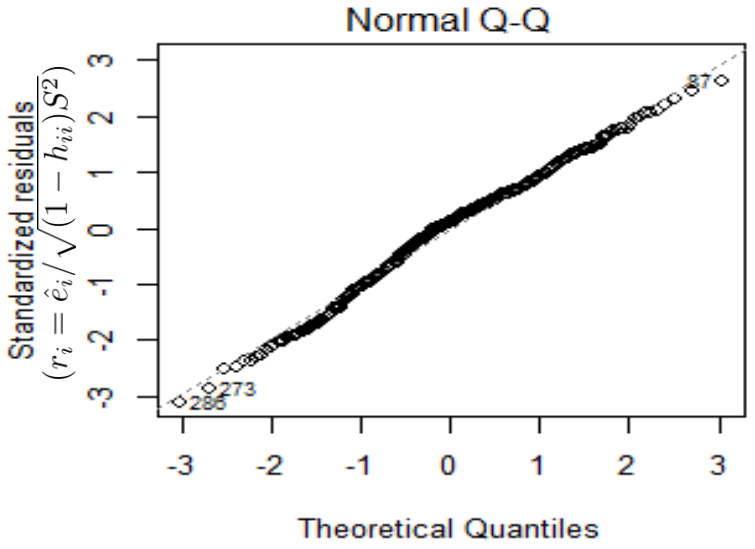
- Look again at

$$\hat{e}_i \;=\; y_i - \hat{y}_i \;=\; (1 - h_{ii})y_i - \sum\nolimits_{j \neq i} h_{ij} y_j$$

  - *When the errors ϵ_i are normal*, the y_i are normal and therefore so are the fitted values $\hat{y}_i$ and residuals $\hat{e}_i$.
  - *When the y_i are not normal,* the CLT will still tend to make $\sum_{j \neq i} h_{ij} y_j$ "look" normal
    - When the sample size is modest, $\sum_{j \neq i} h_{ij} y_j$ will dominate and the $\hat{e}_i$ may "look" normal
    - When the sample size is larger, $(1 - h_{ii})y_i$ will dominate, and the $\hat{e}_i$ will be better at revealing non-normality of y_i and ϵ_i.

The R Residual and QQ plots





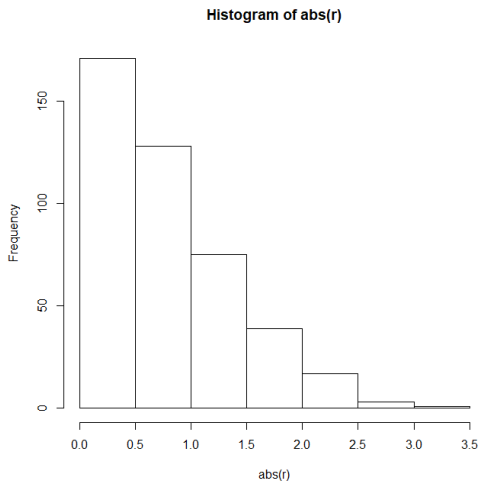
- $E[\hat{e}_i] = E[y_i] - E[\hat{y}_i] = 0$, so should center at zero
  - loess curve helps eye
    - Ignore "edge effects"
- Normal -> Should see no "vertical pattern"…

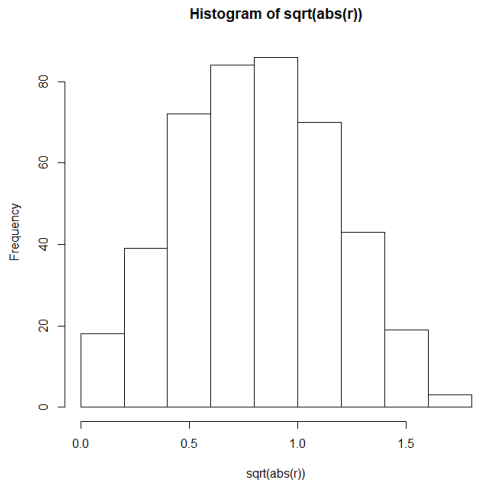
- $(1 - h_{ii})y_i$ dominates, so we are seeing evidence about normality of ϵ_i .
- The three largest r_i are labelled. ($|r_i| > 2$ would be better!)

Checking for (non)constant variance

- Under the hypothesis that $\text{Var}(\epsilon_i) \equiv \sigma^2$,
  - $\text{Var}(\hat{e}_i) = (1 - h_{ii})\sigma^2$ so $\hat{e}_i$ can't be used to test constant variance
  - $r_i = \hat{e}_i / \sqrt{(1 - h_{ii})S^2}$ does have const variance (≈ 1), so can use variation in the size of $|r_i|$ to test constant variance…
- Unfortunately $|r_i|$ are skewed, but we can remove skewing by taking a square root…

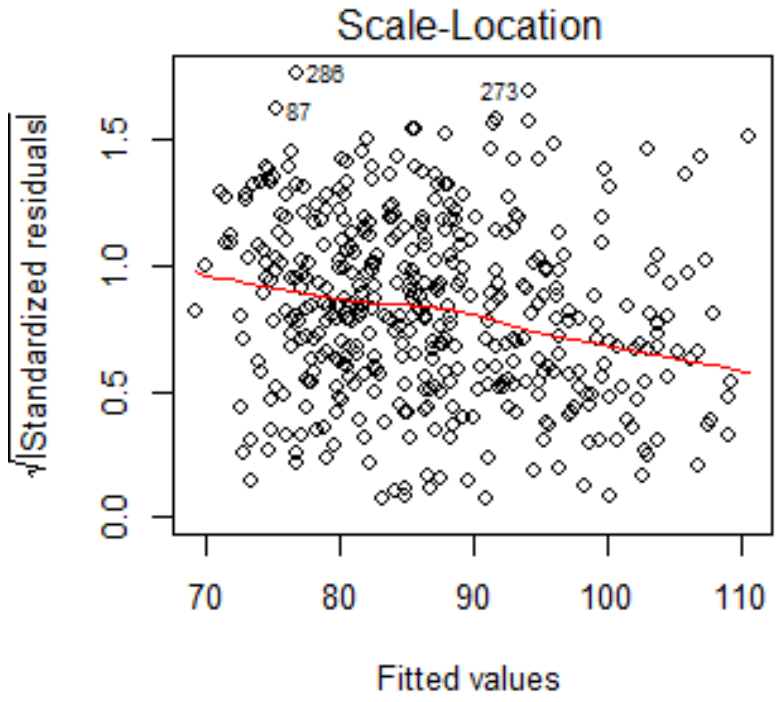
```
r <- rstandard(fit.lm.1)

hist(abs(r))

hist(sqrt(abs(r)))
```





The R Scale-Location Plot

- If $\text{Var}(\epsilon_i) \equiv \sigma^2$ then r_i should have constant variance
  - should see no vertical patterns
  - Loess line helps eye
  - Careful of edge effects
- Designed to catch patterns that depend on x_i (or y_i?)
- Patterns can be caused by
  - Nonconstant variance in ϵ_i
  - Nonlinear relationship between x_i and y_i



- Again, three largest $|r_i|$ get labelled…

Leverage h_{ii} …

- We saw before that

$$\bar{h} = \frac{1}{n}\sum_i h_{ii} = \frac{1}{n}\sum_i \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX}\right) = \frac{2}{n}$$

- We can also calculate

$$\sum_j h_{ij} = \sum_j \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX}\right) = 1$$

- We know

$$\hat{y}_i = \sum_j h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$$

  - h_{ii} measures how much y_i affects $\hat{y}_i$ $\quad (\frac{1}{n} \leq h_{ii} \leq 1)$
  - When $h_{ii} \approx 1$, *SXX* must be dominated by $(x_i - \bar{x})^2$ and therefore all other $(x_j - \bar{x}) \approx 0$, so $h_{ij} \approx \frac{1}{n}$ and the other y_j will have less effect on $\hat{y}_i$

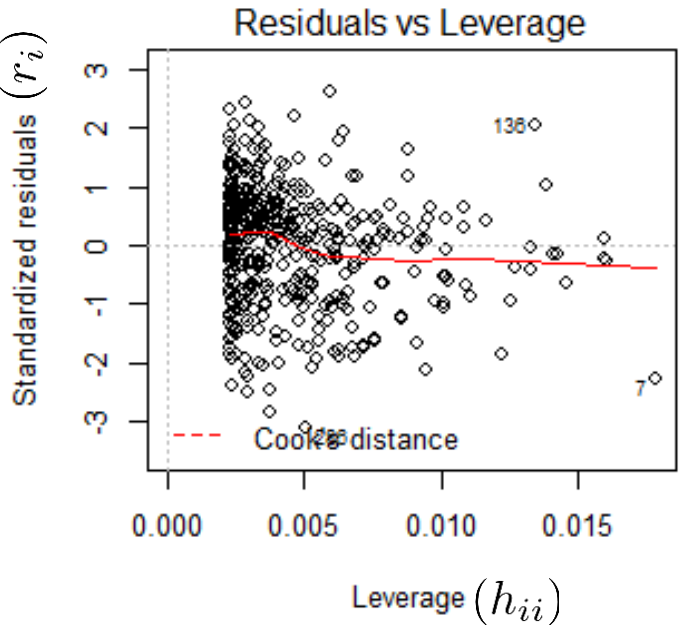
Leverage h_{ii} & Cooks' Distance D_i

- Leverage $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX}$ measures how far x_i is from $\bar{x}$.
- To have an effect, $\hat{e}_i = y_i - \hat{y}_i$ must be large also.
- How much effect can be measured by Cook's D_i:

$$\begin{aligned} D_i &= \frac{\sum_{j=1}^{n} (\hat{y}_{j(i)} - \hat{y}_j)^2}{2S^2} \\ &= \frac{r_i^2}{2} \frac{h_{ii}}{1 - h_{ii}} \quad \text{(not obvious)!} \end{aligned}$$

where $\hat{y}_{j(i)}$ is the fitted value for y_j , omitting the pair (x_i, y_i) from the data set.

The R Leverage Plot



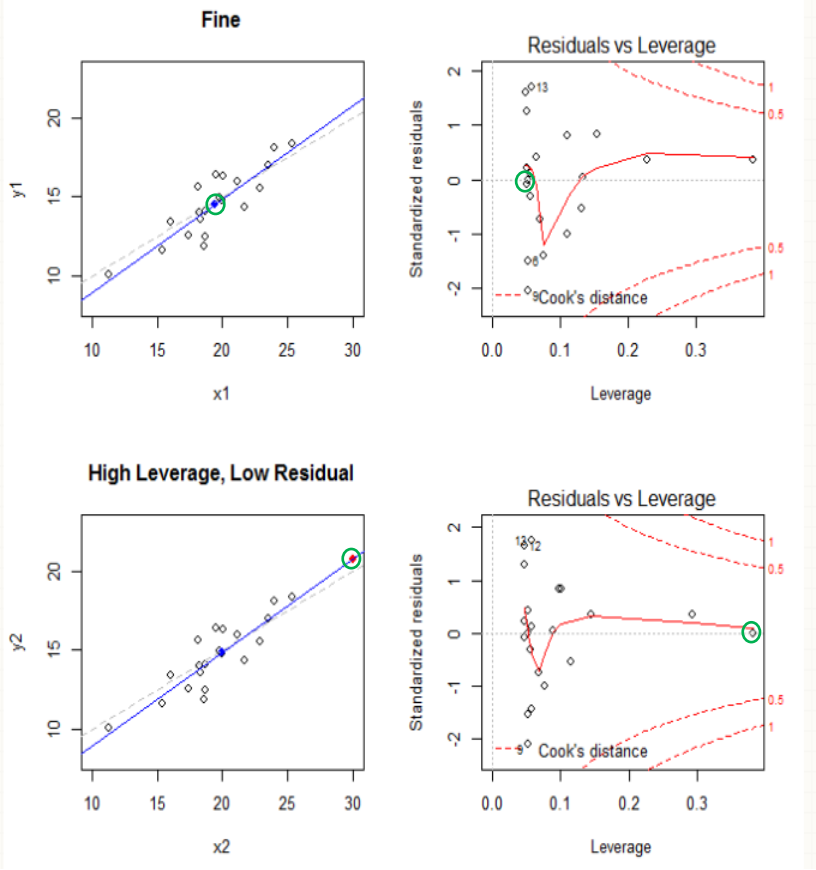
- We need both high leverage h_{ii} and high standardized residual r_i to worry…
  - Rule of thumb:

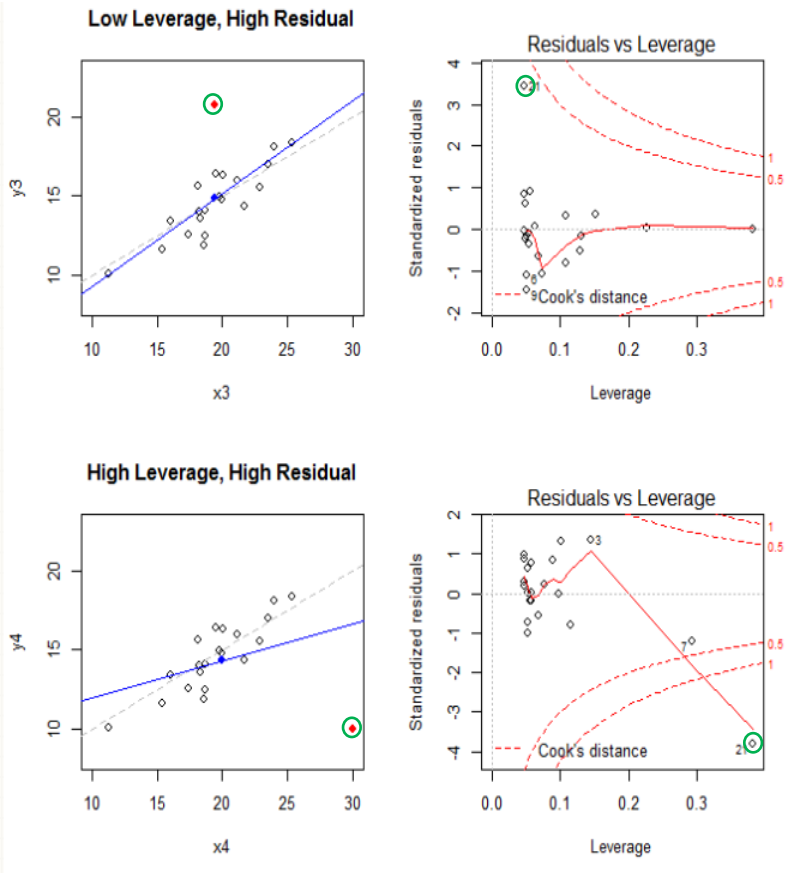
$$h_{ii} > 2 \cdot \bar{h} \ (= 4/n) \quad \text{and}$$
$$|r_i| > 2$$

- Actual effect measured by $D_i = \frac{r_i^2}{2} \frac{h_{ii}}{1 - h_{ii}}$

- D_i is not large enough to show in this example
- Nevertheless, the 3 highest D_i's are labelled.

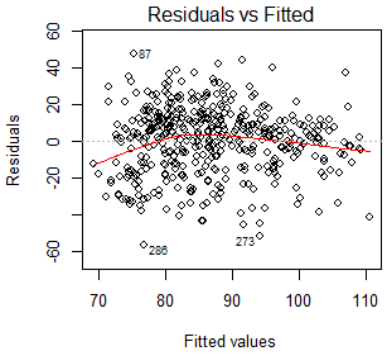
Some Leverage Examples



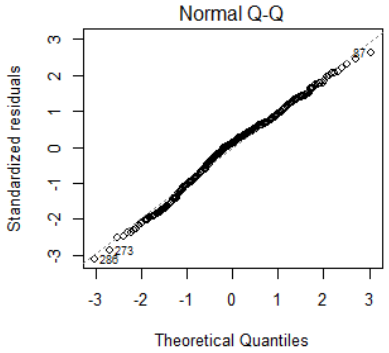


Source: https://stats.stackexchange.com/questions/58141/interpreting-plot-lm

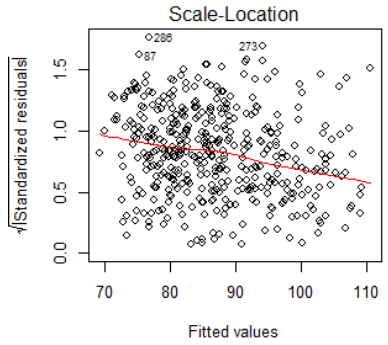
Casewise Diagnostics and Patterns



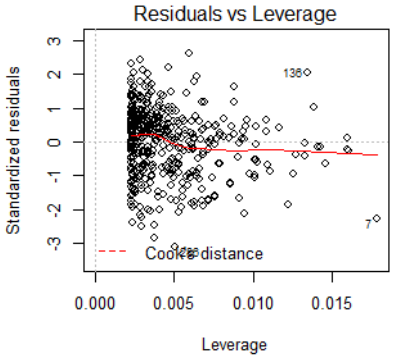
- Mean zero?
- Vertical patterns?
- Outliers?
- Functional dependence on $\hat{y}_i$?



- Normal?
- Outliers?
- Large enough sample?



- Constant variance?
- Vertical patterns?
- Outliers?
- Functional dependence on $\hat{y}_i$?



- NE & SE corners:
  - High leverage h_{ii}
  - High std resid r_i
- D_i > 0.5 or so?

- Don't automatically delete unusual or non-fitting cases
  - Discuss first with investigator; usually meaningful to him/her!
- We'll discuss ways to fix non-constant variance and functional patterns next time!

Summary

- HW02 due tonight, 1159pm
  - Quiz on Ch 3 (today in class)
  - HW03 due next Monday, 1159pm
- Monday: The standard R diagnostic plots:
  - Residuals,QQ plots, Scale-location, Leverage
  - Recommendations
- Wednesday: Transformations (& Examples…)
  - Intuitive / substantive theory-driven
  - Variance stabilization; by hand vs. Box-Cox
  - Perspective and recommendations