Linear and Generalized Linear Models (4433LGLM6Y)

Overview problems in linear models

Meeting 6

Vahe Avagyan

Biometris, Wageningen University and Research

Overview problems in linear models, diagnostics

- Errors in predictors
- Testing for lack of fit: regression vs ANOVA
- Leverages and hat matrix
- Outliers: residuals, standardized and studentized residuals

Overview problems in linear models, diagnostics

- Errors in predictors
- Testing for lack of fit: regression vs ANOVA
- Leverages and hat matrix
- Outliers: residuals, standardized and studentized residuals

Problems in Linear models: what can go wrong?

- What can go wrong?

- Recall the linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Potential problems (According to Faraway):
  - Data
    - Unusual observations
  - Systematic part
    - May not be correct
  - Random part
    - We do not have constant variance, uncorrelatedness, normal distribution.

Problems in Linear models

1. Data
    - Biased sample from population of interest.
    - Important predictors may have been missed.
    - Predictors may have been measured with error.
    - Observational data make causal conclusion problematic.
    - Range of data may limit predictions.
    - Data may contain unusual observations.

Problems in Linear models: what can go wrong?

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

2. Systematic (structural) part: $E(\mathbf{y}) = \mathbf{X}\beta$

- The model may be incorrect.
  - "All models are wrong, but some are useful". George Box:
- A linear model represents an approximation to a complex reality.
  - We hope that it is fair representation of reality.

Problems in Linear models: what can go wrong?

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

3. Error component. Recall: $\epsilon \sim N_n(0, \sigma^2\mathbf{I}_n)$.
   - Errors may be heterogeneous (i.e., unequal variance). 异方差
   - Errors may be correlated. 自相关
   - Errors may not be normally distributed.
     - In larger datasets this is not a big issue
       - E.g., $\hat{\beta}$'s are approximately normal due to CLT.

Diagnostics

- We will study most of the mentioned topics (but not all).
- Assumptions are checked using regression diagnostics.
- Diagnostic techniques can be graphical or numerical.
- Regression diagnostics may suggest improvements.
- Model building is iterative and interactive.

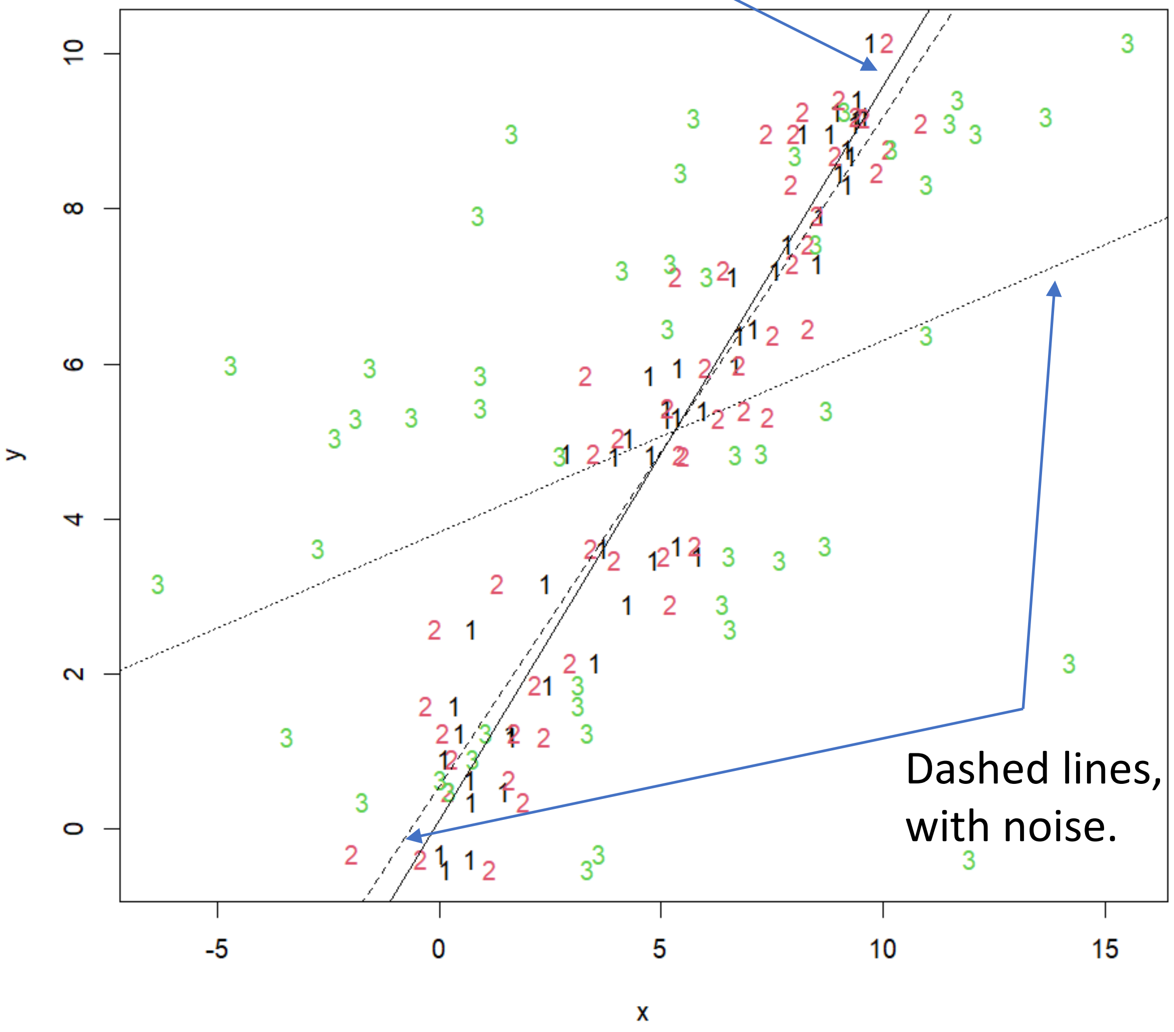
Errors in predictors: Simulated Example

```
> n <- 50; x <- 10* runif(n)
> eps <- rnorm(n)
> y <- 0 + x + eps
> # First model, without any
> # noise in the regressor
> model <- lm(y ~ x); coef(model)
(Intercept)           x
 0.09974288  0.94938496
```

```
> # Add some noise to the regressor
> x1 <- x + rnorm(n)
> model1 <- lm(y ~ x1); coef(model1)
(Intercept)          x1
  0.5371055   0.8646413
```

```
> # Add more noise
> x2 <- x + 5*rnorm(n)
> model2 <- lm(y ~ x2); coef(model2)
(Intercept)          x2
  3.8310088   0.2470816
> matplot(cbind(x, x1, x2), y,
+         xlab = "x", ylab = "y")
> abline(model)
> abline(model1, lty = 2)
> abline(model2, lty=3)
```

Solid line, model without noise.



Dashed lines, model with noise.

$y = 10 \cdot ru + eps$

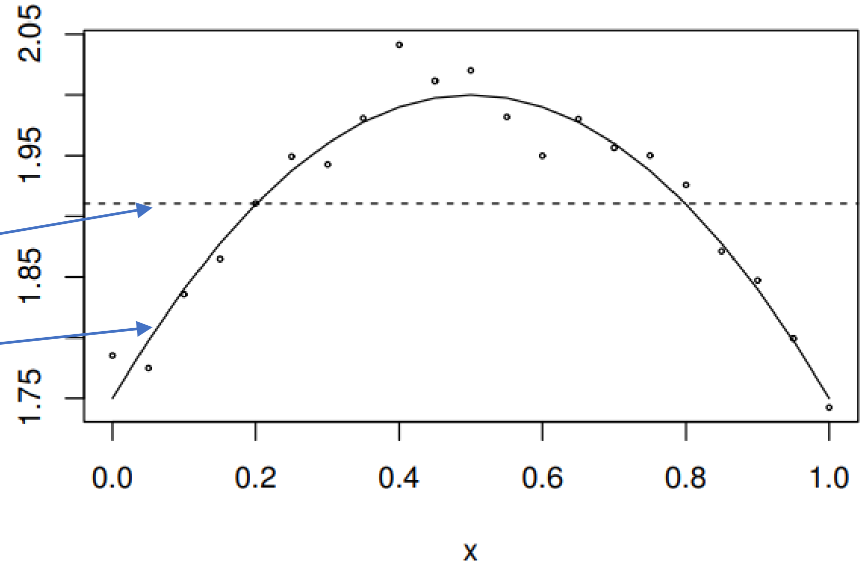
$X2 = 10 \cdot ru + 5 \times ru = 15\,ru$

Overview problems in linear models, diagnostics

- Errors in predictors
- Testing for lack of fit: regression vs ANOVA
- Leverages and hat matrix
- Outliers: residuals, standardized and studentized residuals

Testing for Lack of fit

- How to tell if a model fits the data?
  - Model is correct: $\hat{\sigma}_\epsilon^2$ is an unbiased estimate of σ_ϵ^2
  - model is too simple, $\hat{\sigma}_\epsilon^2$ will overestimate σ_ϵ^2.
  - is too complex, $\hat{\sigma}_\epsilon^2$ may underestimate σ_ϵ^2.



- So, for testing the lack of fit, we could compare $\hat{\sigma}_\epsilon^2$ with $\cancel{\hat{\sigma}_\epsilon^2}$. σ_ϵ^2
- Test of lack of fit: if σ_ϵ^2 is known, then

$$\frac{(n-p)\hat{\sigma}_\epsilon^2}{\sigma_\epsilon^2} \sim \chi^2_{n-p}$$

- Realistically, σ_ϵ^2 is unknown.
  - We need a model-free estimate of σ_ϵ^2.

Pure error variance

- Use repeated (independent) measurements
  - repeated values of y for one or more fixed x

- Pure error variance estimate:

$$\hat{\sigma}_{PE}^2 = SS_{PE}/df_{PE} = \sum_j \sum_i \left(y_{ij} - \bar{y}_j\right)^2 \ /df_{PE}$$

- Here, $df_{PE} = \sum_j (\text{number of replicates } - 1\ ) = n - \text{nr}$ groups.

- SS_{PE} can be seen as the within groups sum of squares from one-way ANOVA in which regressor X is treated as factor.

Testing for Lack of fit

- Hypothesis test:

$$H_0\text{: model fits adequately}$$

$$H_a\text{: model does not fit adequately}$$

- Lack of fit test is a comparison of regression models with ANOVA model.

	df	SS	MS	F
Lack of fit	$n-p-df_{PE}$	$RSS-SS_{PE}$	$\frac{RSS-SS_{PE}}{n-p-df_{PE}}$	Ratio of MS's
Pure Error	df_{PE}	SS_{PE}	SS_{PE}/df_{PE}	
Residual	$n-p$	RSS		

- *Note: Not rejecting* H_0 *does not necessarily mean that* H_0 *is true.*

Testing for Lack of fit: Example

- Iron corrosion

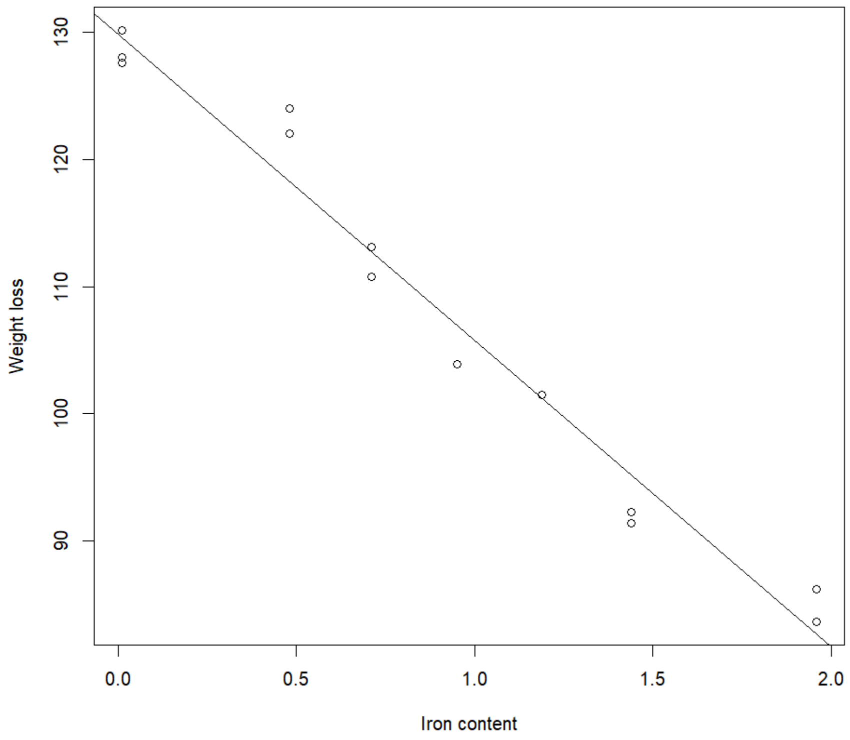
```
> arrange(corrosion, Fe)
     Fe  loss
1  0.01 127.6
6  0.01 130.1
11 0.01 128.0
2  0.48 124.0
7  0.48 122.0
3  0.71 110.8
9  0.71 113.1
4  0.95 103.9
5  1.19 101.5
8  1.44  92.3
12 1.44  91.4
10 1.96  83.7
13 1.96  86.2
```

```
> # Linear regression model
> g <- lm(loss ~ Fe, data = corrosion)
> summary(g)

Call:
lm(formula = loss ~ Fe, data = corrosion)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7980 -1.9464  0.2971  0.9924  5.7429 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  129.787      1.403   92.52  < 2e-16 ***
Fe           -24.020      1.280  -18.77 1.06e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.058 on 11 degrees of freedom
Multiple R-squared:  0.9697,	Adjusted R-squared:  0.967 
F-statistic: 352.3 on 1 and 11 DF,  p-value: 1.055e-09

> (rss <- sum((summary(g)$residuals)^2))
[1] 102.8502
> #An easier way of getting rss
> deviance(g)
[1] 102.8502
```

Assume, Fe is numerical, not a factor

Testing for Lack of fit: Example

```
> plot(corrosion$Fe,corrosion$loss,
+       xlab="Iron content",ylab="Weight loss")
> abline(g$coef)
```




```
> #ANOVA model with Fe factor.
> ga <- lm(loss ~ as.factor(Fe), data = corrosion)
> # RSS of the ANOVA model
> deviance(ga)
[1] 11.78167
> #Pure error variance estimate
> deviance(ga)/ga$df.residual
[1] 1.963611
> anova(g, ga)
Analysis of Variance Table

Model 1: loss ~ Fe
Model 2: loss ~ as.factor(Fe)
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     11 102.850
2      6  11.782  5    91.069 9.2756 0.008623 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We must conclude that there is a lack of fit.

Pure error sd is estimate: $\sqrt{\hat{\sigma}^2_{PE}} = \sqrt{\dfrac{11.78}{6}} = \sqrt{1.96} = 1.4 > 3.06$

Overview problems in linear models, diagnostics

- Errors in predictors
- Testing for lack of fit: regression vs ANOVA
- Leverages and hat matrix
- Outliers: residuals, standardized and studentized residuals

Outliers, Leverage, and Influence

- Unusual data are problematic in linear model's fit by least squares
- Regression outlier is an observation whose response-variable value is conditionally unusual given value of explanatory variable(s).
- An observation has high leverage if its regressor values are extreme so that it potentially has strong leverage (influence) on regression coefficients.
- An observation has high influence if it has both discrepancy (i.e., "outlyingness") and high leverage.

Influence on coefficients = Leverage $\times$ Discrepancy

Examples on simple linear regression

a) Low leverage, but regression outlier

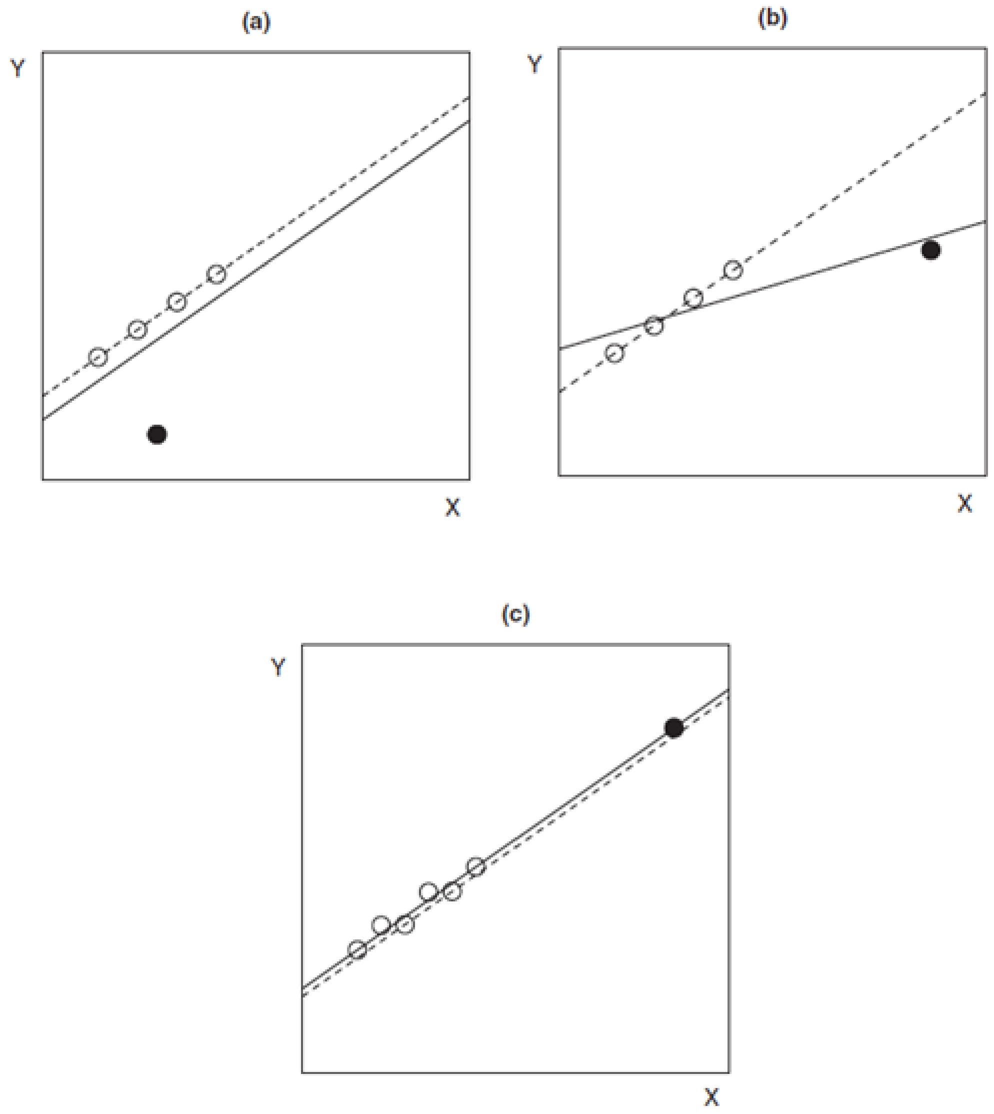
- Deletion of observation hardly has impact on slope, slightly affects the intercept.

b) High leverage, and regression outlier

- Deletion of observation will affect the slope and the intercept.

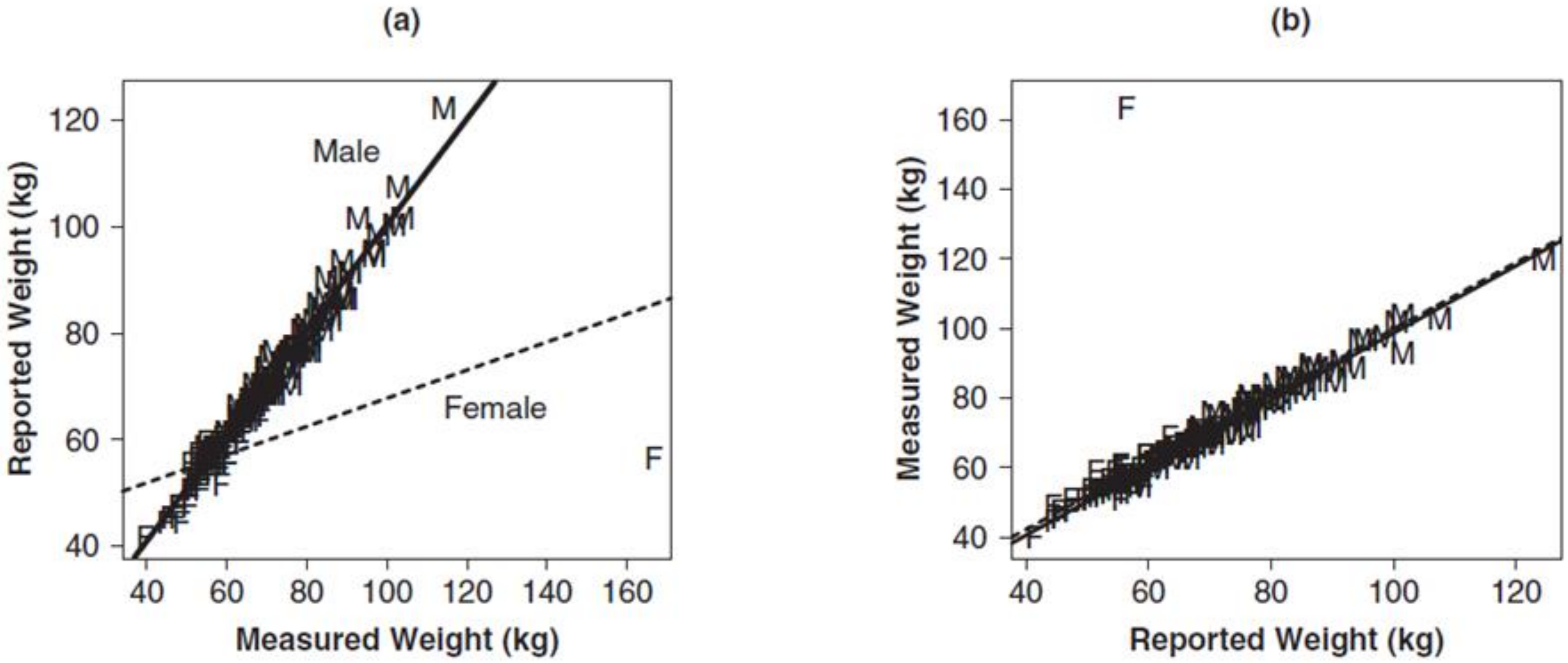
c) High leverage, but not a regression outlier

- Deletion will not change slope and intercept substantially.



Example on simple linear regression

- Example for Davis's data on reported and measured weight for women (F) and men (M).



Assessing Leverage: Hat-values

- Recall

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

- The fitted values are

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y},$$

- Define the $\mathbf{H}$ matrix as:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

- $\mathbf{H}$ depends only on the regressors, not on $\mathbf{y}$.
- $\mathbf{H}$ transforms $\mathbf{y}$ into $\hat{\mathbf{y}}$, i.e., $\hat{\mathbf{y}} = \overset{n\times n}{\mathbf{H}}\overset{n\times 1}{\mathbf{y}}.$
- Fitted values are : $\hat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{jj}Y_j + \ldots + h_{nj}Y_n.$

Assessing Leverage: Hat-values

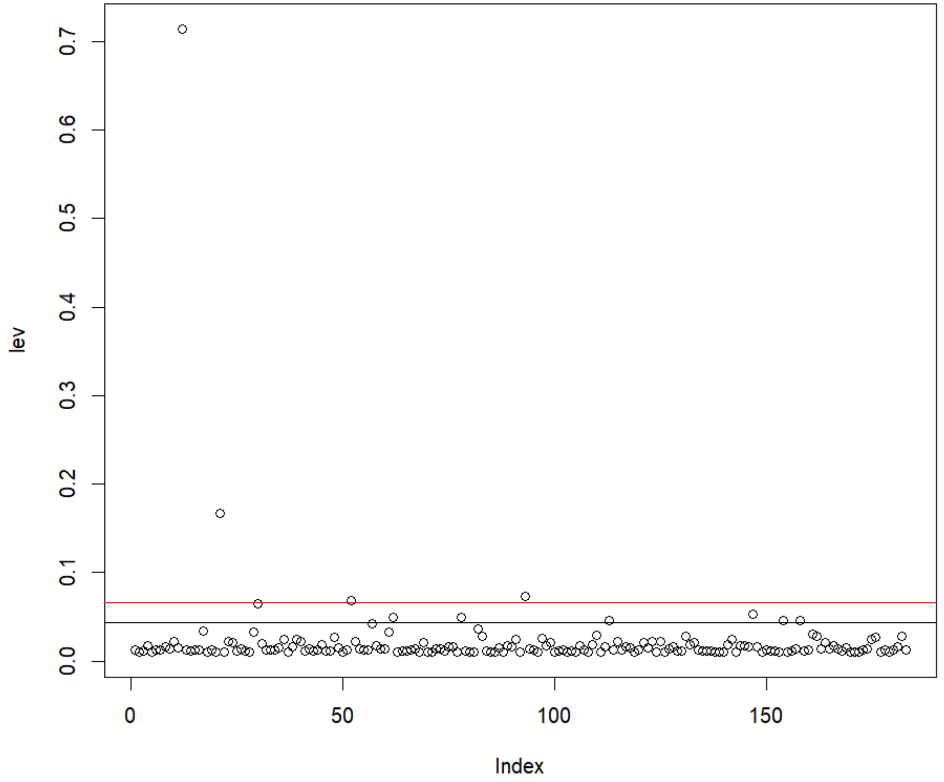
- Hat-values: $h_i \equiv h_{ii}$ is a measure of leverage in regression.
- Properties of $\mathbf{H}$ matrix:
  - Symmetric, i.e., $\mathbf{H}' = \mathbf{H}$
  - Idempotent, i.e., $\mathbf{H}^2 = \mathbf{H}$ 幂矩阵
  - $0 < h_i \leq 1.$
  - $\text{trace}(\mathbf{H}) = \sum h_i = k + 1$ (for regression model with k regressors). Or $\bar{h} = (k + 1)/n$.
- Common cut-offs: Hat values higher than $2 \times \bar{h}$ or $3 \times \bar{h}$ should be considered as high leverage:

Assessing Leverage: Example (Davis data)

- Davis data: $n = 183$, and $k = 3$ regressors. *What is the average leverage?*

??

```
> g1 <- lm(repwt ~ weight + factor(sex) + weight:factor(sex), dat
a=Davis)
> lev <- lm.influence(g1)$hat
> sort(lev,decreasing=T)[1:10] # 10 largest leverages
        12         21         97         54         30
0.71418565 0.16684054 0.07320771 0.06877588 0.06451113
       156         65         82        118        169
0.05254010 0.04912301 0.04895185 0.04569369 0.04569369
> # Alternative way of getting leverages
> X <- model.matrix(g1)
> lev2 <- hat(X)
> sort(lev2,decreasing=T)[1:10]
 [1] 0.71418565 0.16684054 0.07320771 0.06877588 0.06451113
 [6] 0.05254010 0.04912301 0.04895185 0.04569369 0.04569369
```



Overview problems in linear models, diagnostics

- Errors in predictors
- Testing for lack of fit: regression vs ANOVA
- Leverages and hat matrix
- Outliers: residuals, standardized and studentized residuals

$\vec{y} = X\vec{\beta}$

$\vec{z} = \text{proj}_X \vec{y}$

$\vec{z} = X\vec{\beta}$

$X^T(\vec{y}-\vec{z}) = \vec{0} \Rightarrow X^T(\vec{y} - X\vec{\beta}) = \vec{0} \Rightarrow X^T X\vec{\beta} = X^T\vec{y}$

Detecting Outliers: Residuals

- Remember the least square residuals

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\underbrace{\mathbf{X}}_{n\times p}\underbrace{\beta}_{p\times 1} + \underbrace{\epsilon}_{n\times 1}) - (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{X}\beta + \epsilon) = \epsilon - \mathbf{H}\epsilon = (\underbrace{\mathbf{I}_n - \mathbf{H}}_{n\times n})\underbrace{\epsilon}_{n\times 1}.$$

$\vec{\beta} = (X^TX)^{-1}X^T\vec{y}$

- The residuals do not have equal variance and are not uncorrelated ($\mathbf{e}$ vs ϵ).

$\vec{z} = X\vec{\beta} = X(X^TX)^{-1}X^T\vec{y}$

$$\underbrace{E(\mathbf{e})}_{n\times 1} = \mathbf{0} \text{ and } V(\mathbf{e}) = \underbrace{\sigma_\epsilon^2(\mathbf{I}_n - \mathbf{H})}_{n\times n}$$

- Single residual:

$$V(E_i) = \sigma_\epsilon^2(1 - h_i),$$

- A large leverage will make the variance of residual small.

Detecting Outliers: Standardized Residuals

- Standardized residuals (Fox) or (internally) studentized residuals (Faraway).

$$E_i' \equiv \frac{E_i}{S_E\sqrt{1-h_i}}$$

- These have variance 1 and give us some idea about the "outlyingness" of an observation.
- Rule of thumb: Values larger than 3 or smaller than -3 are unlikely to occur.
- E_i' does not follow t-distribution.
- Alternative, Externally studentized (jackknife) residuals:

$$E_i^* = E_i'\sqrt{\frac{n-k-2}{n-k-1-E_i'^{\,2}}}$$

- Rule of thumb: Values larger than 2 or smaller than -2 are unlikely to occur.

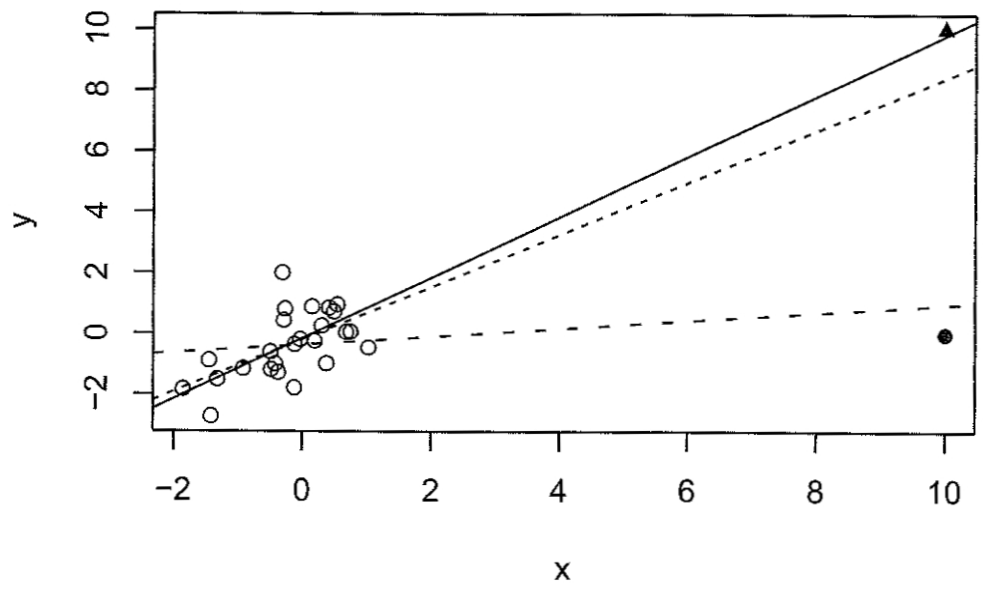
Problems with Standardized Residuals E_i'

- Outliers can conceal themselves.
- Example: 2 high leverage observations: ▲ and •
  - Solid line: including ▲ but excluding •.
  - dashed line: including •, excluding ▲;
  - dotted line: both excluded.

- This problem can not be solved with E_i' and E_i.

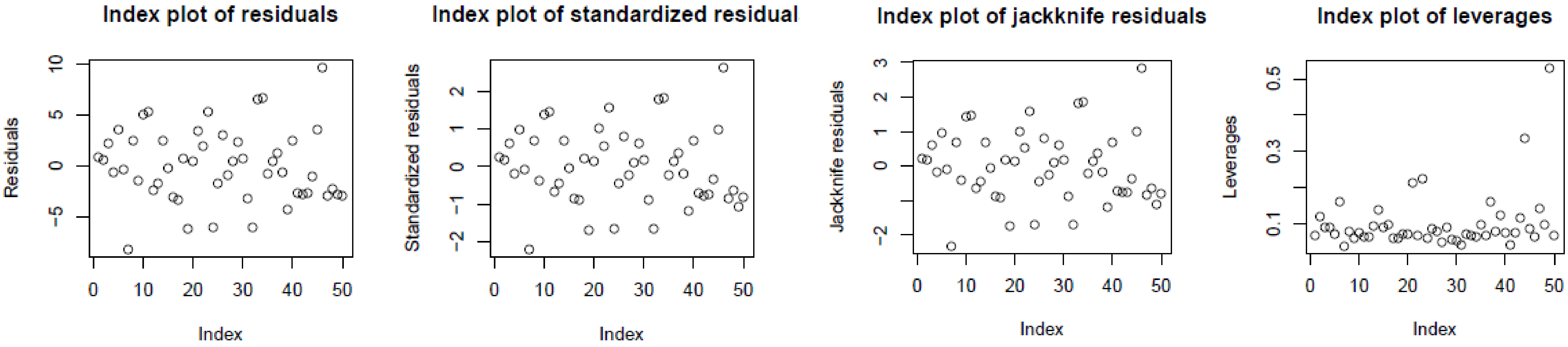
- Outlier tests can be done using E_i^* (see `outlierTest()`)
- If the model is correct:

$$E_i^* \sim t_{n-1-(k+1)}$$



Detecting Outliers: Example

```
> g <- lm(sr ~ pop15 + pop75 +dpi + ddpi, data=savings)
> plot(g$res, ylab="Residuals", main="Index plot of residuals")
> plot(rstandard(g), ylab="Standardized residuals", main="Index plot of standardized residuals")
> plot(rstudent(g), ylab="Jackknife residuals", main="Index plot of jackknife residuals")
> plot(lm.influence(g)$hat, ylab="Leverages", main="Index plot of leverages")
```



Some further remarks about outliers

- General remarks:
    - Two or more outliers next to each other can hide each other.
    - Outlier in one model may not be outlier in another when variables have been changed or transformed.
    - Error distribution may be non-normal, so that larger residuals may be expected.
    - Individual outliers much less of a problem in larger datasets: single point will not have leverage to affect the fit considerably. However, clusters of outliers may.

Some further remarks about outliers

- What to do about outliers?
  - Check the data-entry errors first.
  - Examine the physical context: what did happen? Discovery of outlier may be of great interest.
  - Exclude point from analysis, try reinclude later, compare results. Report honestly about the existence of outliers, even if not included in your model.
  - Robust regression may be preferred if outliers exist, which cannot be identified as mistakes or aberrations.
  - Don't exclude outliers in automated way.

Influential observations

- Influential point is one whose removal from dataset would cause large change in the fit.

- Measure of the influence: Cook's distance:

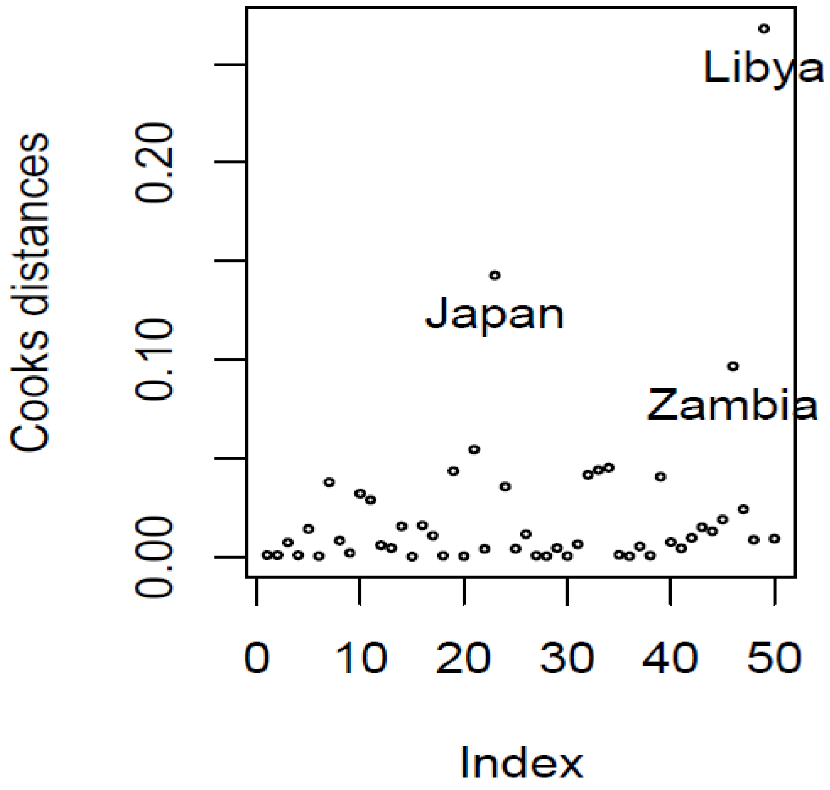
$$D_i = \frac{E_i'^{\,2}}{(k+1)} \times \frac{h_i}{1-h_i}$$

- Recall the formula: $\text{Influence on coefficients} = \text{Discrepancy} \times \text{Leverage}$

- Numerical cutoff: $D_i > \frac{4}{n-k-1}$.

Influential observations: Example

```
> g <- lm(sr ~ pop15 + pop75 +dpi + ddpi, data=savings)
> cook <- cooks.distance(g)
> range(cook)
[1] 4.736572e-05 2.680704e-01
```



- We can identify the largest three values.
- The cut-off here is 0.0888.