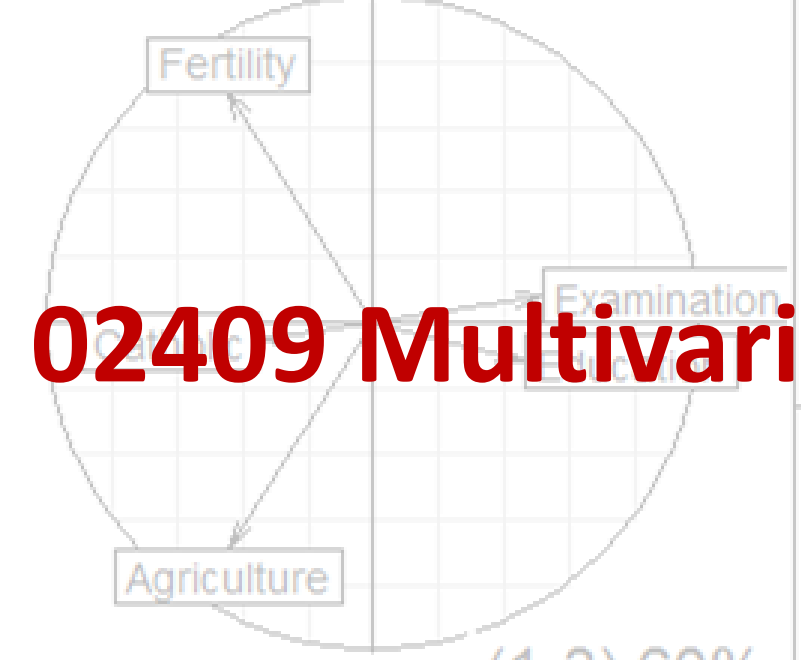
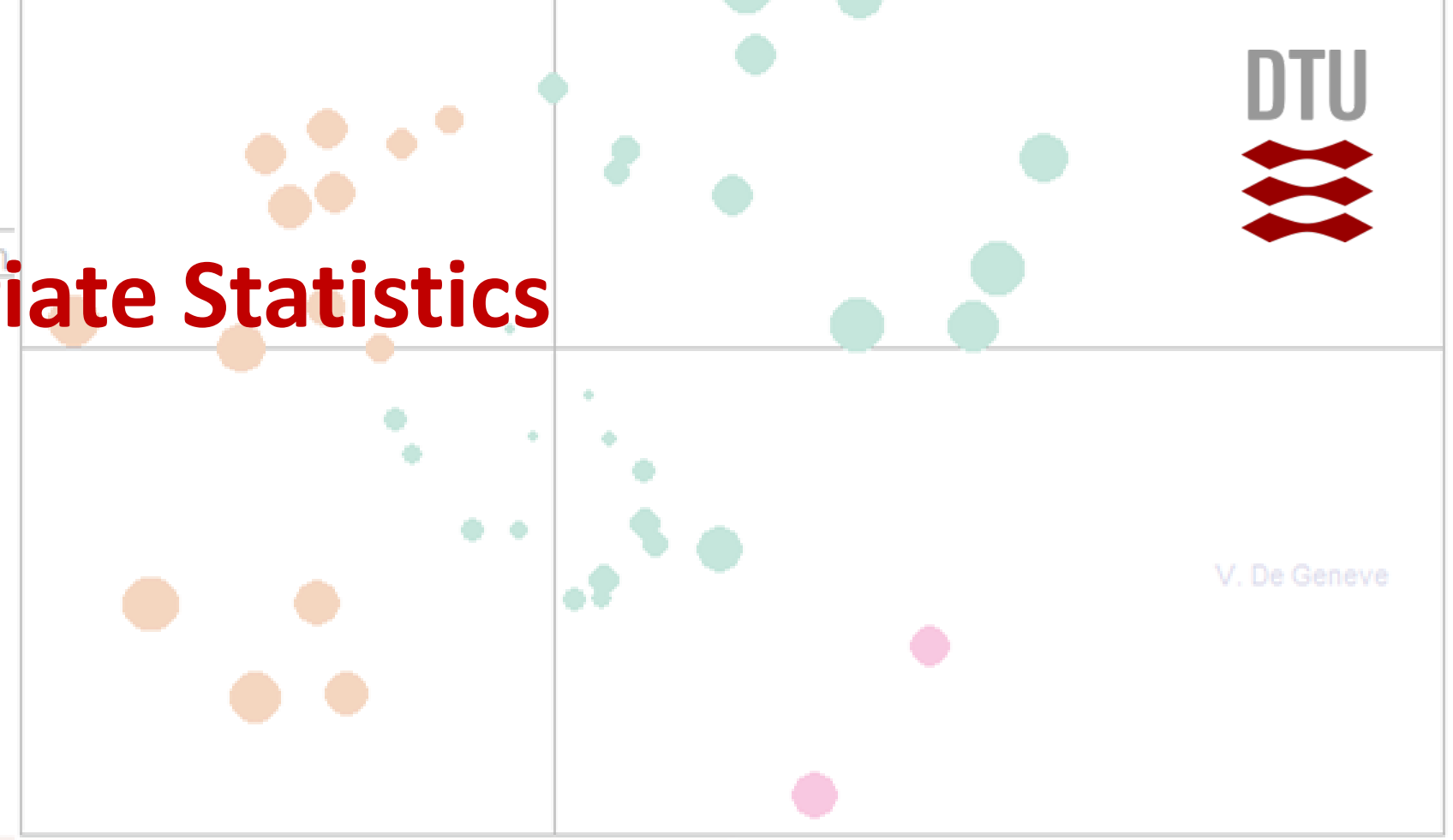


02409 Multivariate Statistics

Lecture I, November 2 2025

Anders Stockmarr

anst@dtu.dk



Clustering 4 groups

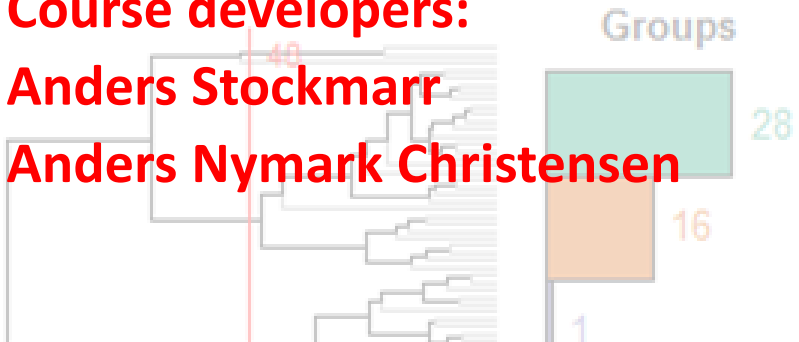
Factor 1 [41%]

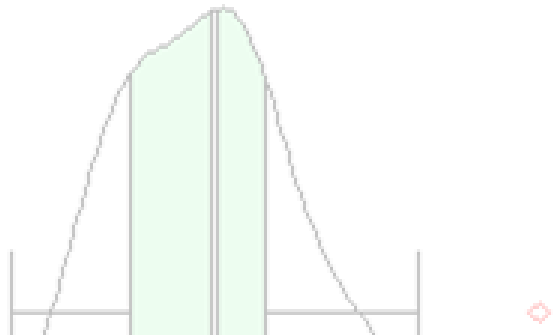
Factor 3 [19%]

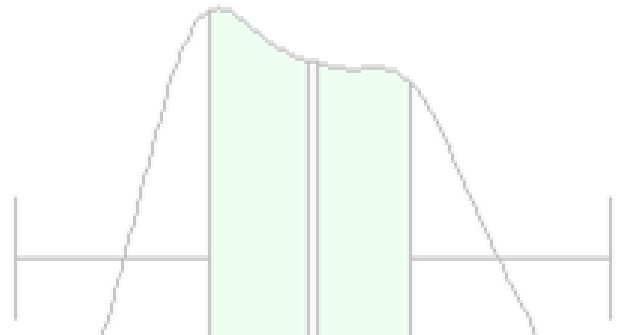
Course developers:

Anders Stockmarr

Anders Nymark Christensen







Agenda

- GLM in practice;
- Model selection;
- Outliers and influential observations

The General Linear Model - GLM

$\boldsymbol{x}$ and $\boldsymbol{\Sigma}$ assumed known:

$$\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma}) \quad \boldsymbol{\mu} = \mathrm{x}\,\boldsymbol{\theta}$$

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix}$$

or

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

i.e.

$$\boldsymbol{Y} = \boldsymbol{x\theta} + \boldsymbol{\varepsilon}$$

with

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}), \qquad \boldsymbol{x}, \boldsymbol{\Sigma} \text{ known}$$

Estimation in GLM I

Theorem 2.3

If Y is independent and identical distributed (iid), Σ is the identity matrix!

Let $\mathbf{x}$ and $\boldsymbol{\theta}$ be given as in the preceding section and let $\boldsymbol{Y} \sim \mathrm{N}_n(\mathbf{x}\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite. Then the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is given by $\mathbf{x}\hat{\boldsymbol{\theta}}$ being the projection (with respect to $\boldsymbol{\Sigma}$) onto M, $\hat{\boldsymbol{\theta}}$ is a solution to the so-called ***normal equation(s)***

$$(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\hat{\boldsymbol{\theta}} = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$

If Σ is the identity matrix, we simply have ordinary least squares

If $\mathbf{x}$ has full rank k, then

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y},$$

$$\Sigma = I \Rightarrow \boxed{\hat{\theta} = (x^T x)^{-1}x^T Y}$$

and being a linear combination of normally distributed variables $\hat{\boldsymbol{\theta}}$ is also normally distributed with parameters

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\theta}}) &= \boldsymbol{\theta} \\ \mathrm{D}(\hat{\boldsymbol{\theta}}) &= \sigma^2(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}. \end{aligned}$$

It is especially noted that $\hat{\boldsymbol{\theta}}$ is an unbiased estimate of $\boldsymbol{\theta}$.

Estimation in GLM II

Theorem 2.5

Let the situation be as above. The maximum likelihood estimator of σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}\|^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}).$$

The unbiased estimator of σ^2 is

If Σ is the identity matrix, we simply have the equations from ordinary least squares

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n - \mathrm{rk}(\mathbf{x})}\|\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}\|^2 \\ &= \frac{1}{n - \mathrm{rk}(\mathbf{x})}(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})\end{aligned}$$

where $\mathbf{x}\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\mathrm{E}(\mathbf{Y})$. The following holds

$$\hat{\sigma}^2 \sim \sigma^2\chi^2(n - \mathrm{rk}(\mathbf{x}))/(n - \mathrm{rk}(\mathbf{x}))$$

and $\hat{\sigma}^2$ is independent of the maximum likelihood estimator of the expected value and is therefore independent of $\hat{\boldsymbol{\theta}}$.

Example: Crime data

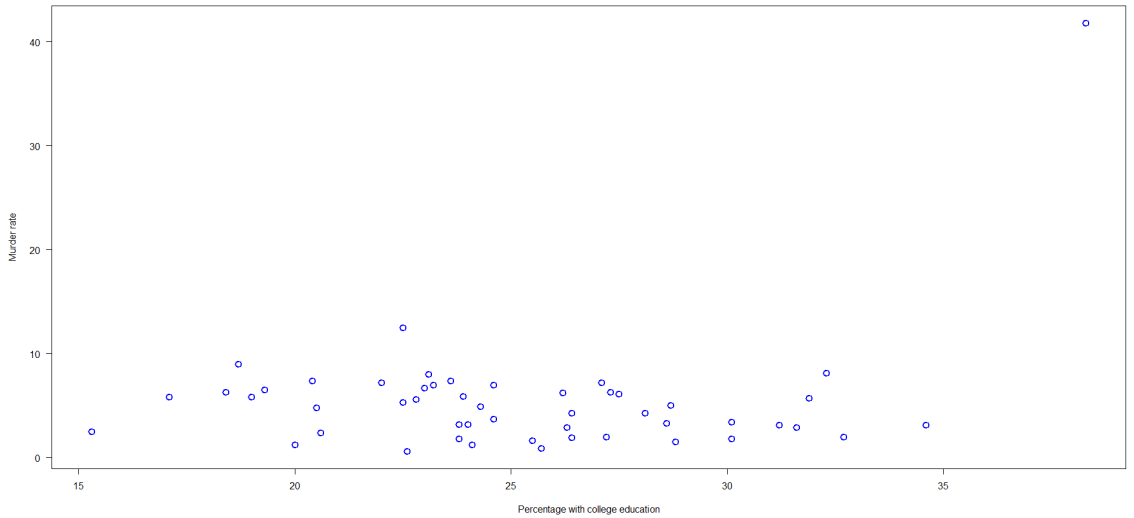
- Is higher education associated with higher murder rates?
- Crime data from 2003 for the US.

- **Murder rate**: The annual number of murders per 100,000 people in the
- population.
- **College**: Percentage of the adult residents who have a college
- education.

- Y is the murder rate; X is the explanatory variable college.

Example: Crime data

```
> crime <- read.delim("Data/us_statewide_crime.txt")
> crime<-crime[,c(1,3,6)]
> head(crime)
       State murder.rate college
1    Alabama         7.4    20.4
2     Alaska         4.3    28.1
3    Arizona         7.0    24.6
4   Arkansas         6.3    18.4
5 California         6.1    27.5
6   Colorado         3.1    34.6

> plot(crime$college, crime$murder.rate,
+        xlab="Percentage with college education",
+        ylab="Murder rate", las=1, cex = 1.5,
+        col = "blue", lwd = 2)
```



Example: Crime data

- We set up a linear model: With Y the murder rate,

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad \varepsilon_i \sim iid.\, N(0, \sigma^2),$$

Or

$$Y = X\theta + \varepsilon,$$

$$X = \begin{bmatrix} 1 & college_1 \\ \vdots & \vdots \\ 1 & college_{51} \end{bmatrix}, \theta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

Why is this a good idea?

Taylor expansion of real functions

- Unknown relationship between crime and college:

$$Y = f(X) + \varepsilon$$

Power series expansion:

$$f(x) = f(x_0) + \sum_{n=1}^{\infty} a_n (x - x_0)^n, \qquad a_n = \frac{f^{(n)}(x_0)}{n!}$$

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + r_1(x - x_0)$$

$$= \underbrace{f(x_0) - f'(x_0)x_0}_{\alpha} + \underbrace{f'(x_0)}_{\beta} x + r_1(x - x_0)$$

So that

$$Y = \alpha + \beta X + r_1(x - x_0) + \varepsilon = \alpha + \beta X + \tilde{\varepsilon}$$

IF the remainder term $r_1(x - x_0)$ is negligible compared to the residual ε. **Note:** $r_1(x - x_0) = o(x - x_0)$.

Taylor expansion of real functions

"**IF** the remainder term $r_1(x - x_0)$ is negligible compared to the residual ε"

- What if that is not the case?
- **Taylor expansion of order 2:**

$$Y = f(X) + \varepsilon$$

Power series expansion (higher dimensions of X):

$$f(x) = f(x_0) + < f'(x_0), x - x_0 > + \frac{1}{2}\|x - x_0\|^2_{f''(x_0)} + r_2(x - x_0)$$

$$Y = \alpha + \sum_{i=1}^{k} \beta_i X_i + \sum_{r,s=1}^{k} \beta_{rs} X_r X_s + \tilde{\varepsilon}$$

IF the remainder term $r_2(x - x_0)$ is negligible compared to the residual ε. **Note:** $r_2(x - x_0) = o\big((x - x_0)^2\big)$.

Example: Crime data

- We set up a linear model:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad \varepsilon_i \sim iid.N(0, \sigma^2),$$

Or

$$Y = X\theta + \varepsilon,$$

$$X = \begin{bmatrix} 1 & college_1 \\ \vdots & \vdots \\ 1 & college_{51} \end{bmatrix}, \theta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

- 1st order Taylor approximation to the relationship between murder rate and college.

Example: Crime data

- Parameter estimates:

Note that

$$X^T X = \begin{bmatrix} n & co. \\ co. & co^2. \end{bmatrix}$$

Where a dot indicates summation. Thus:

$$(X^T X)^{-1} = \frac{1}{nco^2. - co.^2} \begin{bmatrix} co^2. & -co. \\ -co. & n \end{bmatrix} = \frac{1}{SSD_{co}} \begin{bmatrix} \overline{co^2} & -\overline{co} \\ -\overline{co} & 1 \end{bmatrix}$$

where a overbar indicates average, and where $SSD_{co} = \sum_i (co_i - \overline{co})^2$.

SSD_{co} = **S**um of **S**quares of **D**eviations of `college`.

Example: Crime data

- Parameter estimates:

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} = (X^T X)^{-1} X^T Y = \frac{1}{SSD_{co}} \begin{bmatrix}\overline{co^2} & -\overline{co}\\ -\overline{co} & 1\end{bmatrix} \begin{bmatrix}cr.\\ < co, cr >\end{bmatrix}$$

$$= \ \ldots$$

$$= \begin{bmatrix}cr. - \hat{\beta}\overline{co}\\ \dfrac{< co - \overline{co}, cr - \overline{cr} >}{SSD_{co}}\end{bmatrix} = \begin{bmatrix}cr. - \hat{\beta}\overline{co}\\ \dfrac{SPD_{co,cr}}{SSD_{co}}\end{bmatrix}$$

with $SPD_{co,cr}$ = **S**um of **P**roducts of **D**eviations of `college` and `crime`.

Example: Crime data

- Parameter estimates in **R**:

```
> x <- crime$college
> y <- crime$murder.rate
> (beta <- sum((x-mean(x))*(y-mean(y)))/sum((x-mean(x))^2))
[1] 0.33307

> (alpha <- mean(y)- beta*mean(x))
[1] -3.058072

> Fitted <- alpha + beta*x; Resid <- y - Fitted
> sigma2 <- sum(Resid^2)/(length(y)-2)
> sqrt(sigma2)
[1] 5.6146
```

- Thus:

$$\hat{\alpha} = -3.058072, \qquad \hat{\beta} = 0.33307, \qquad \hat{\sigma} = 5.6146$$

Example: Crime data

- Parameter estimates in **R** using the `lm()` function:

```
> reg1 <- lm(murder.rate ~  college, data = crime)
> summary(reg1)

Call:
lm(formula = murder.rate ~ college, data = crime)

Residuals:
   Min     1Q Median     3Q    Max
-5.833 -3.368 -1.403  1.981 32.101

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.0581     4.3630  -0.701   0.4867
college       0.3331     0.1703   1.955   0.0563 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.615 on 49 degrees of freedom
Multiple R-squared:  0.07238,   Adjusted R-squared:  0.05345
F-statistic: 3.823 on 1 and 49 DF,  p-value: 0.05626
```

$\hat{\alpha}$

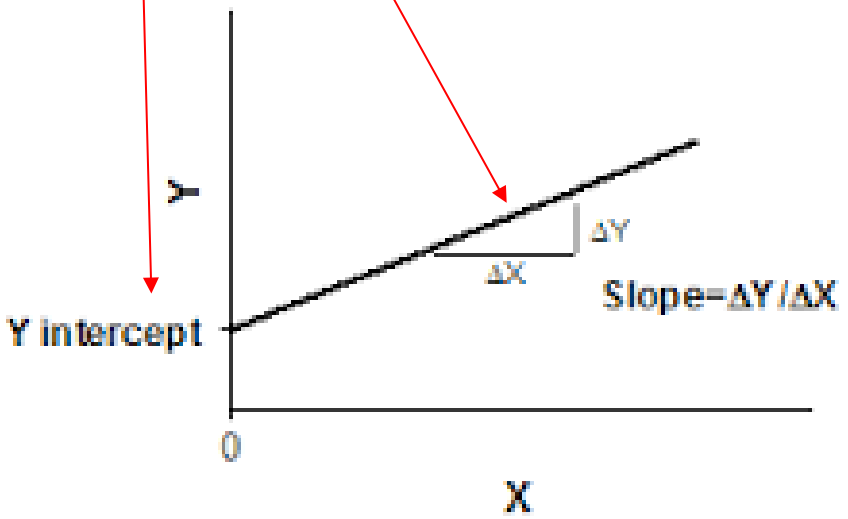
$\hat{\beta}$

$\hat{\sigma}$

Example: Crime data

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad \varepsilon_i \sim iid.N(0, \sigma^2)$$

The expression for a straight line:



The slope β is how much the murder rate increases when the college rate increase by one unit. $\hat{\beta} = 0.3331.$

Example: Crime data

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.0581     4.3630  -0.701   0.4867
college       0.3331     0.1703   1.955   0.0563 .
```

- Parameter estimates handing in **R**:

```
> confint(reg1)
                    2.5 %    97.5 %
(Intercept) -11.825900553 5.7097573
college      -0.009244278 0.6753843

> #Nice table
> tab <- cbind(coef(summary(reg1))[ , 1:2], "Lower" = confint(reg1)[ , 1],
+              "Upper" = confint(reg1)[ , 2])

> tab
             Estimate Std. Error         Lower     Upper
(Intercept) -3.058072  4.3630260 -11.825900553 5.7097573
college      0.333070  0.1703416  -0.009244278 0.6753843

> #Nice table with p-values
> data.frame(round(tab, 2),
+            "p-value" = format.pval(coef(summary(reg1))[ , 4], digits = 3, eps = 1e-3))
            Estimate Std..Error  Lower Upper p.value
(Intercept)    -3.06       4.36 -11.83  5.71  0.4867
college         0.33       0.17  -0.01  0.68  0.0563
```

Example: Crime data

- **Model check:**
  - Results can only be trusted if the model fits the data.

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad \varepsilon_i \sim iid.N(0, \sigma^2)$$

- **Model assumptions:**

1. Linear dependency of covariates;
2. Independence of residuals;
3. Normality of residuals;
4. Homogeneity of residuals (same σ^2 for all observations).

Example: Crime data

1. Linear dependency of covariates;
   - scatter plots of residuals vs. numerical explanatory variables;
2. Independence of residuals;
   - Context; cannot be checked with numerical methods in simple models;
3. Normality of residuals;
   - Quantile-Quantile plots of residuals versus the normal distribution;
4. Homogeneity of residuals (same σ^2 for all observations).
   - scatter plots of residuals vs. fitted values.

Example: Crime data

Quantile-quantile plots:

- Two distributions P and Q are the same if their distribution functions are the same:

$$\forall x \in \mathbb{R}: P\big((-\infty; x]\big) = Q\big((-\infty; x]\big)$$

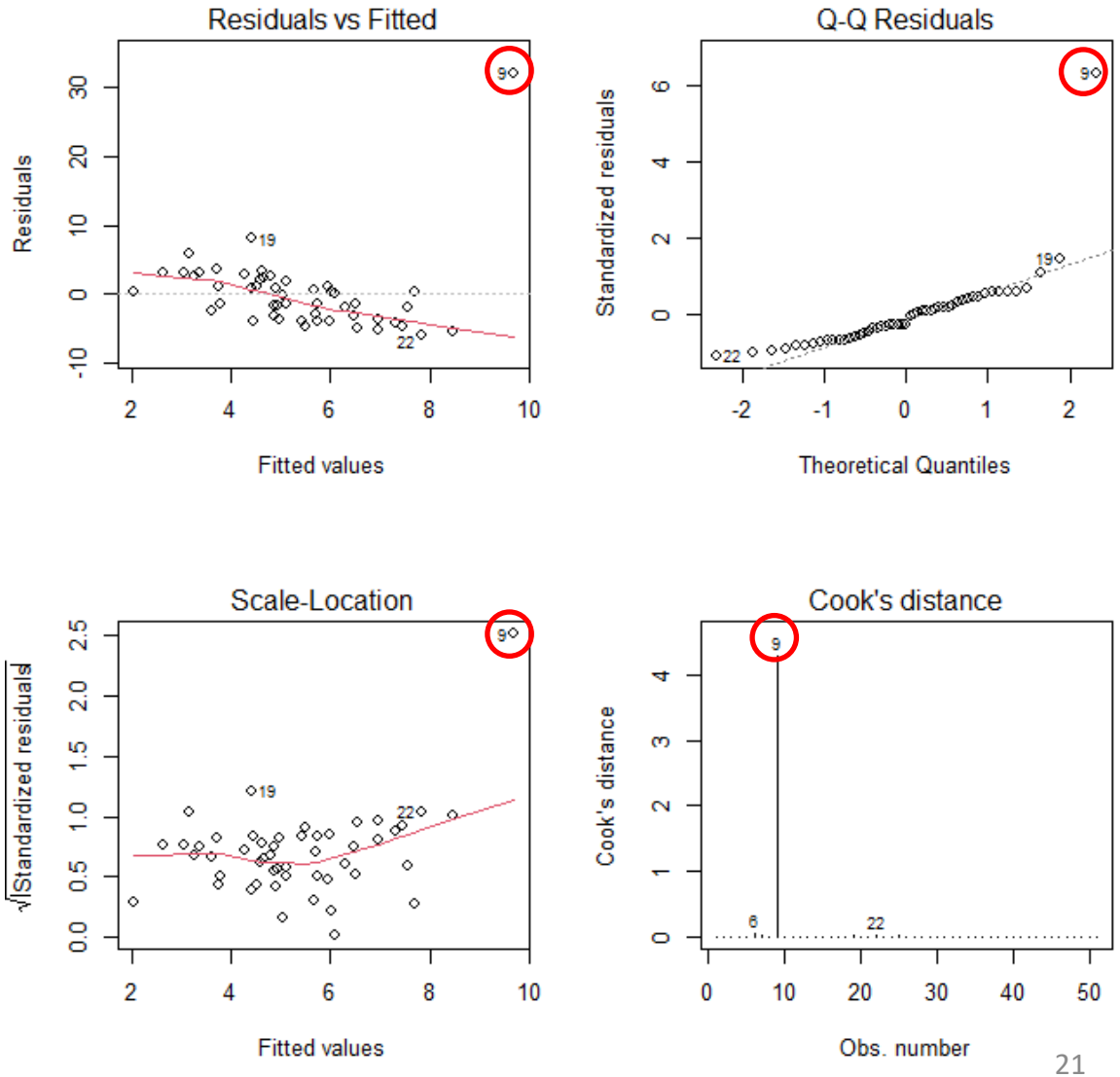
- Similar: Two distributions P and Q are the same if their inverse distribution functions (**quantile functions**) are the same:

$$\forall p \in [0; 1]: q_{P,p} = \inf_{x \in \mathbb{R}} \{x: P\big((-\infty; x]\big) \geq p\}$$
$$= \inf_{x \in \mathbb{R}} \{x: Q\big((-\infty; x]\big) \geq p\} = q_{Q,p}$$

Check of normality: Compare empirical quantiles with the quantiles of the corresponding normal distribution

Example: Crime data

```
par(mfrow = c(2, 2))
plot(reg1, which = 1:4)
par(mfrow = c(1, 1))
```



There is an issue with observation #9!

```
> crime[9,]
                  State murder.rate college
9 District of Columbia        41.8    38.3
```

Washington DC, the capital district of USA – not a state.

Example: Crime data

Outliers:

Incorrect data; OR data generated from a data generating mechanism deviating from the rest of the model.

Must be checked, corrected, or removed.

We want to uncover the underlying data generating mechanism; observations not complying with that will

a) disturb the interpretation and
b) not contribute with information.

Example: Crime data

Observation #9 is an outlier:

Observed value: 41.8.
Predicted value:

$$\hat{\alpha} + \hat{\beta} * 38.3 = -3.058072 + 0.33307 * 38.3 = 9.69851$$

Estimated residual: 41.8–9.69851 = 32.10149

$$\hat{P}\left(\max_{i=1,\ldots,51} |\varepsilon_i| \geq 32.10149\right) = 1 - \hat{P}\left(\max_{i=1,\ldots,51} |\varepsilon_i| < 32.10149\right)$$

$$= 1 - \hat{P}(|\varepsilon_i| < 32.10149)^{51} 1 - \left(1 - 2\Phi\left(\frac{-32.10149}{\hat{\sigma}}\right)\right)^{51}$$

$$= 0.0000006$$

Example: Crime data

- **Observation #9 is also the observation with the biggest residual;**
  The observation that contributes most to the estimated variation.

- **What if we estimated the variation disregarding the suspected outlier?**

```
>crime50 <- crime[-9, ]
>reg2 <- lm(murder.rate ~  college, data = crime50)

>(sigma.new2<-summary(reg2)$sigma^2)
[1] 2.464061

>1-(1-2*(1-pnorm(reg1$res[9]/sigma.new)))^51
9
0
```

ie., $< 0.000000000000000000001$. Same conclusion with stronger evidence.

No way that this can happen by chance

Example: Crime data

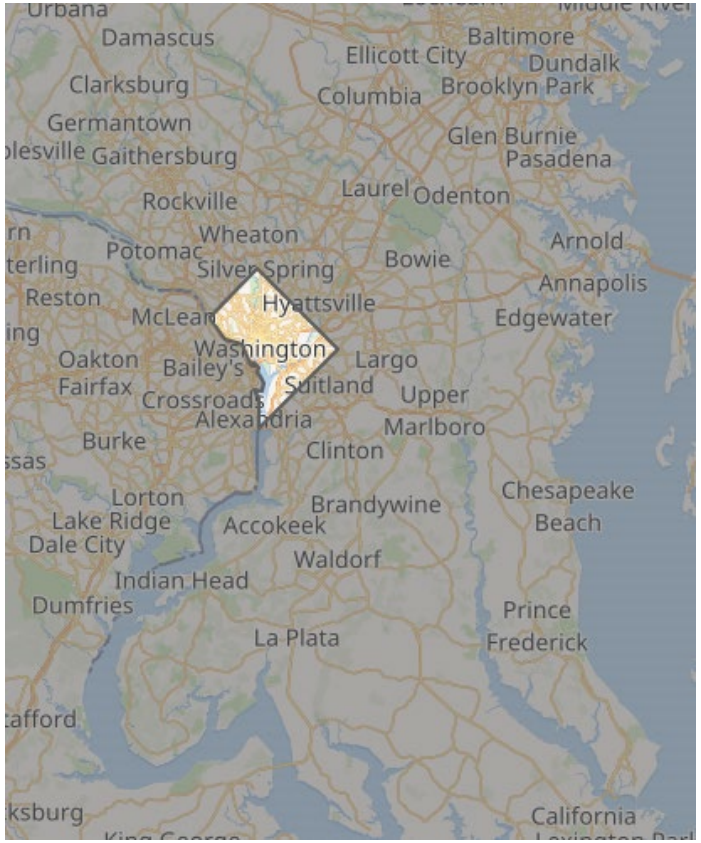
```
> crime[9,]
                  State murder.rate college
9 District of Columbia        41.8    38.3
```

- Washington DC, the capital district of USA – is a district, and NOT a state.
- As such, the district does not have the socioeconomy of a state – schools, welfare, etc works differently, for the invited clientele that works there.

- Only 30% of the workforce actually lives there!

- As such, **the data generating mechanism is different!**

- The inclusion on the list is likely for completion, but essentially irrelevant when it comes to the relation between college education and murder rate **at the state level**.



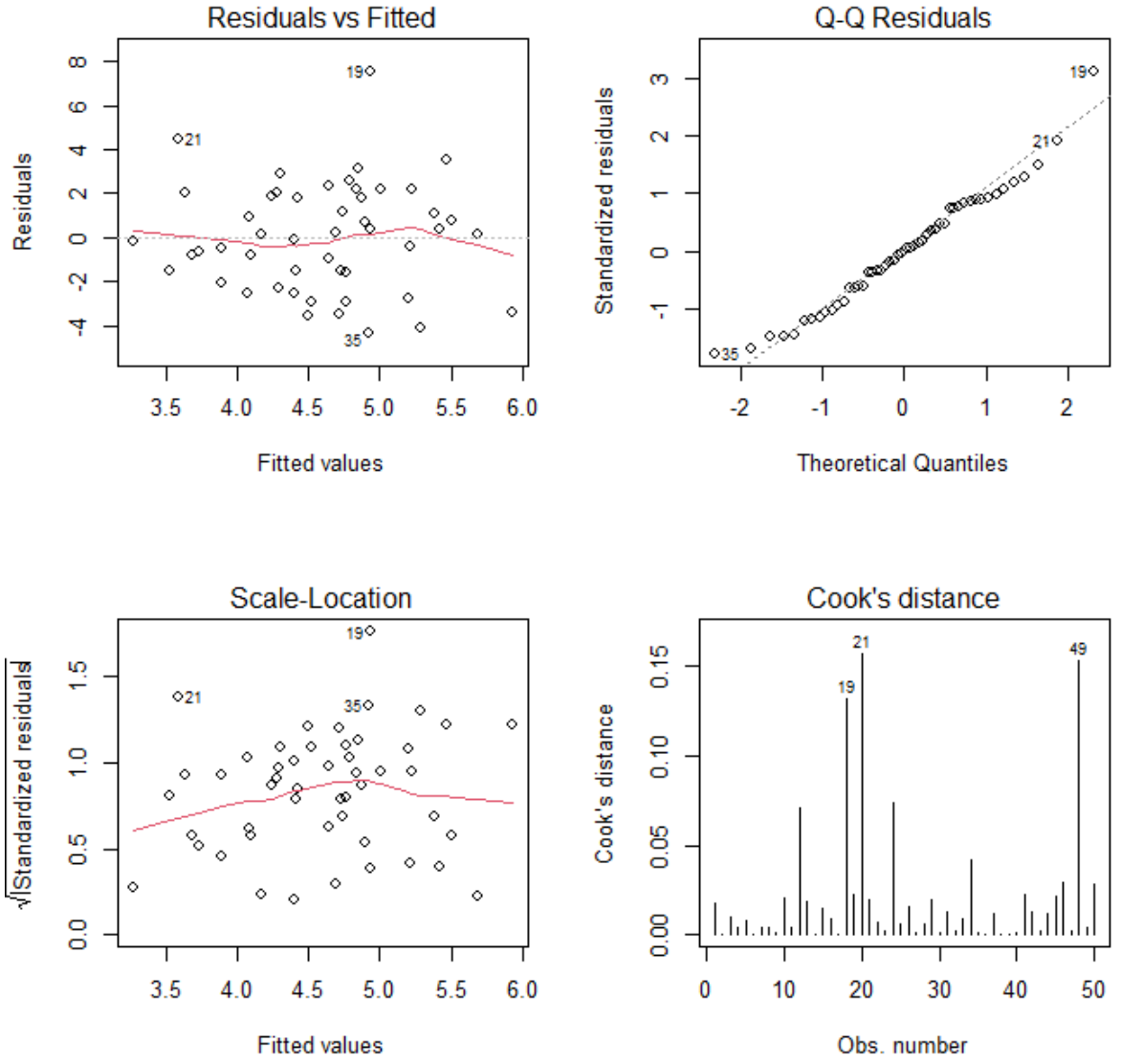
Picture: Wikipedia

We shall stick to STATES, and leave out Washington DC!

Example: Crime data

```
par(mfrow = c(2, 2))
plot(reg2, which = 1:4)
par(mfrow = c(1, 1))
```

- Much nicer; points towards a fitting model: Variance homogeneity and normality.

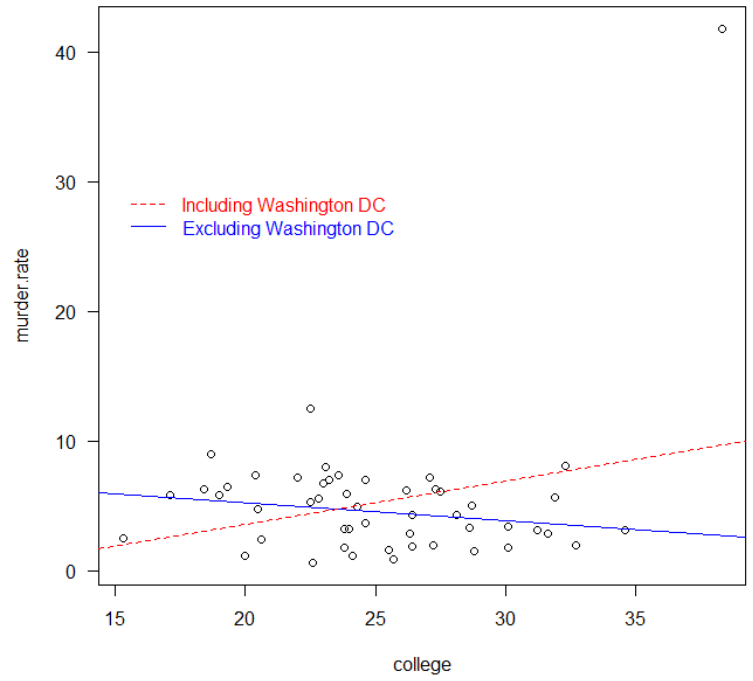


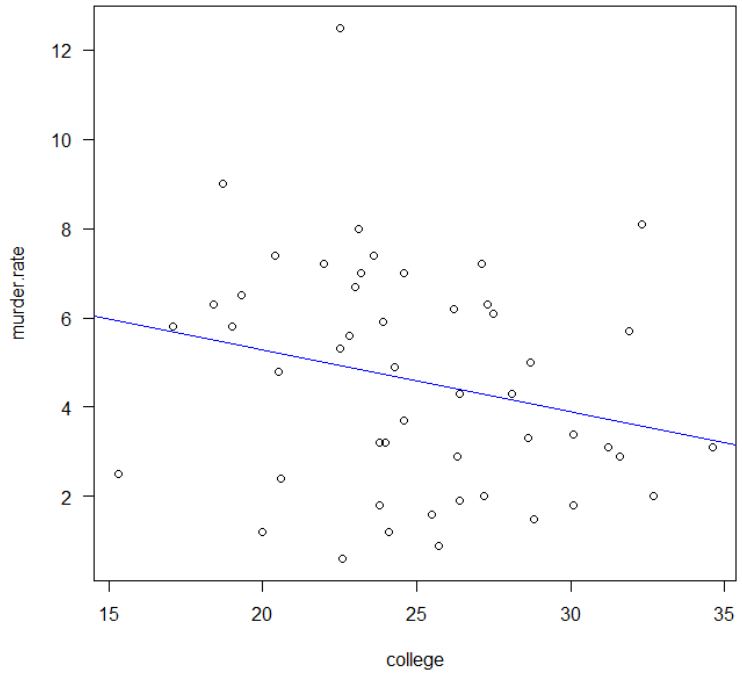
Example: Crime data

```
> summary(reg2)

Call:
lm(formula = murder.rate ~ college, data = crime50)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3255 -1.5533  0.0073  1.8450  7.5607

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.04159    2.06478   3.895 0.000304 ***
college     -0.13788    0.08163  -1.689 0.097687 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.464 on 48 degrees of freedom
Multiple R-squared:  0.0561,    Adjusted R-squared:  0.03644
F-statistic: 2.853 on 1 and 48 DF,  p-value: 0.09769
```

Example: Crime data

```
>plot(murder.rate ~ college, data = crime, las=1)
>abline(reg2,font=2,col="blue")
>abline(reg1,font=2,col="red",lty=2)

>legend(15,30,c("Including Washington DC","Excluding Washington DC"),
+       lty=c(2,1),col=c("red","blue"),text.col=c("red","blue"),bty="n")

>plot(murder.rate ~ college, data = crime50, las=1)
>abline(reg2,font=2,col="blue")
```

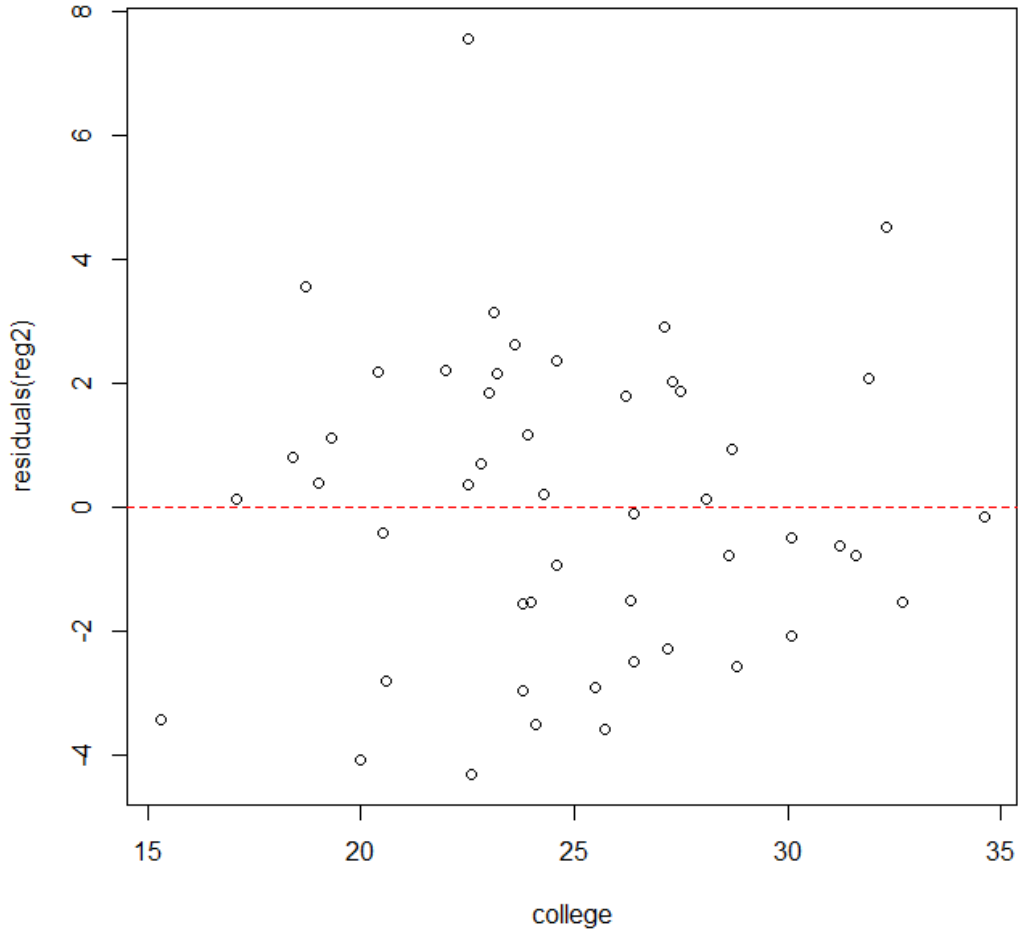




Example: Crime data

- Linearity: Plot residuals vs. numerical explanatory variables:

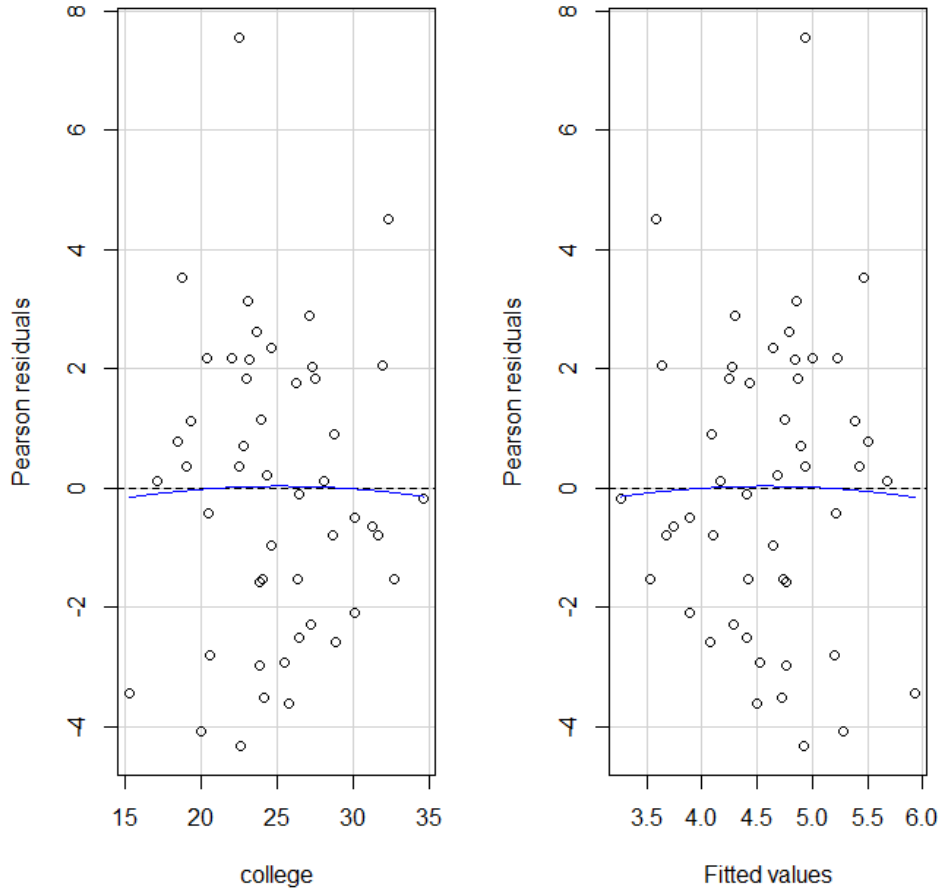
```
plot(residuals(reg2) ~ college, data=crime50)
abline(h = 0, lty = 2,col="red")
```



Example: Crime data

- Linearity: Plot residuals vs. numerical explanatory variables with estimated curvature:

```
> residualPlots(reg2)
            Test stat Pr(>|Test stat|)
college       -0.1277           0.8989
Tukey test    -0.1277           0.8984
```

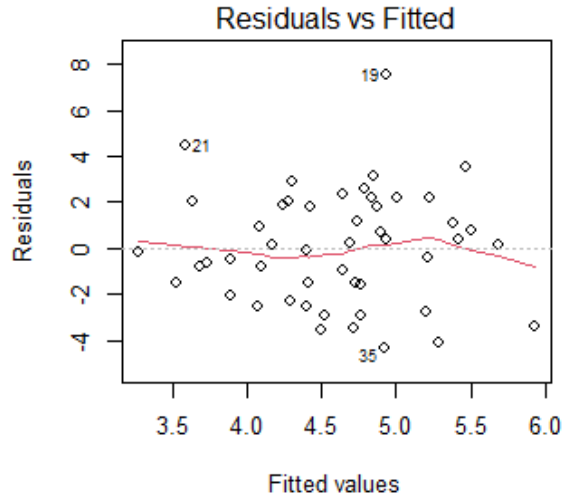


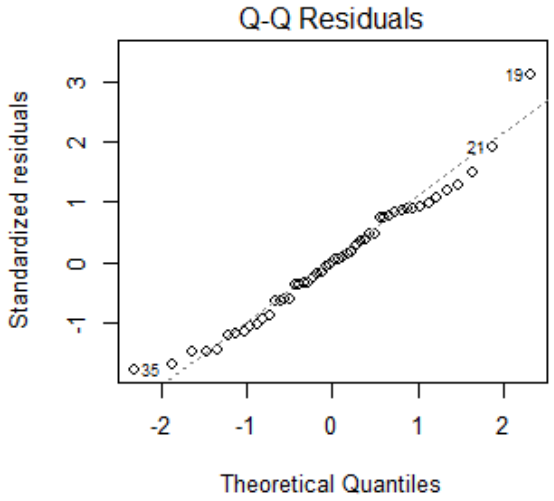
Example: Crime data

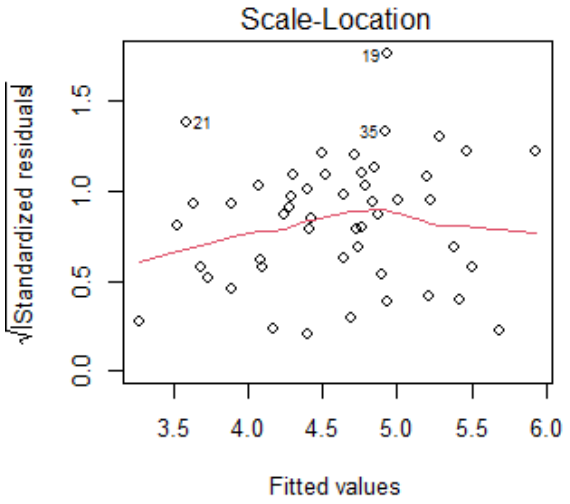
```
> # curvature college:
> reg3<-update(reg2,~.+I(college^2))
> drop1(reg3,test="F")
Single term deletions

Model:
murder.rate ~ college + I(college^2)
             Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>                    291.34 94.123
college       1  0.016451 291.35 92.125  0.0027 0.9591
I(college^2)  1  0.101079 291.44 92.140  0.0163 0.8989


> # curvature fitted values (Tukeys test):
> reg3<-update(reg2,~.+I(fitted(reg2)^2))
> drop1(reg3,test="F")
Single term deletions

Model:
murder.rate ~ college + I(fitted(reg2)^2)
                  Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>                        291.34 94.123
college            1   0.42886 291.76 92.196  0.0692 0.7937
I(fitted(reg2)^2)  1   0.10108 291.44 92.140  0.0163 0.8989
```

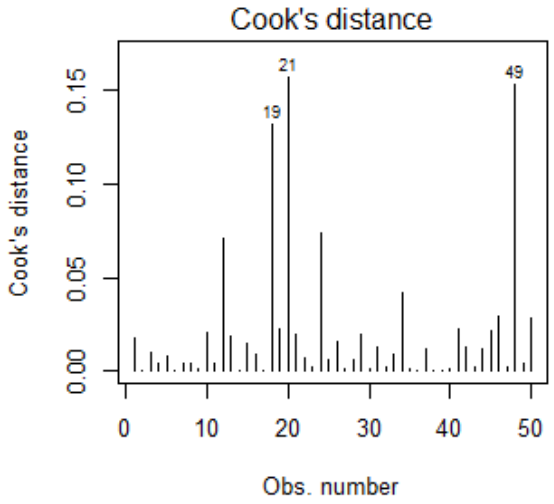
Example: Crime data

- Visual testing – the **EBT** test.
- Why not use one of the many numerical tests for normality?



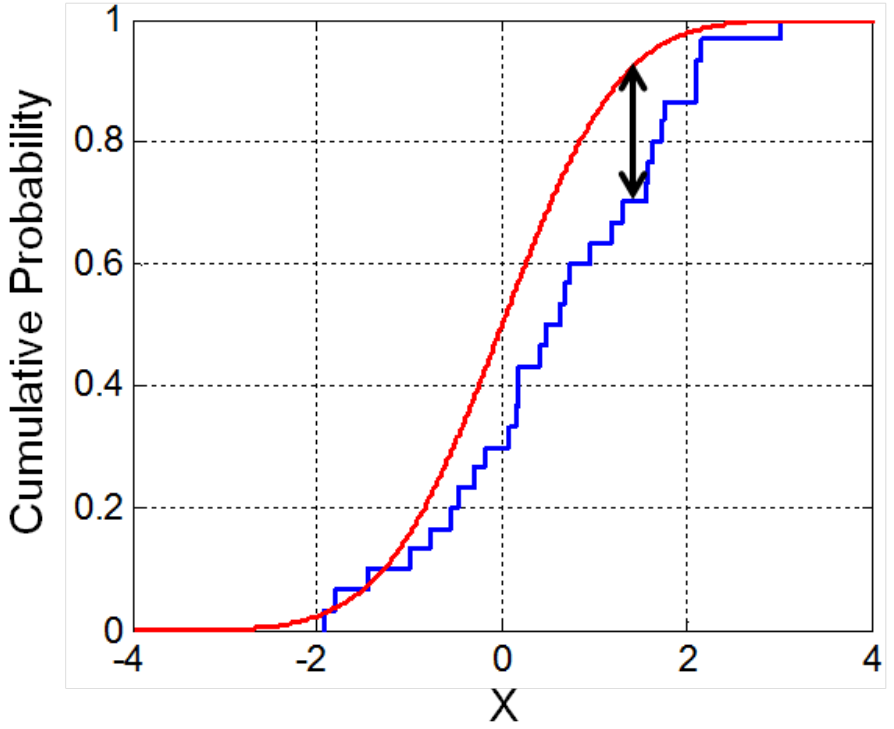






Example: Crime data

- The *Kolmogorov-Smirnov* test for normality:
- Compares (here) the distance between an empirical and a theoretical distribution function.
- Other tests are the Shapiro-Wilks test and the Anderson-Darling test.
- In **R**: `ks.test()`



Picture source: Wikipedia/Bscan

Tests For Normality

- The Kolmogorov-Smirnov test for normality:

```
> ks.test(rnorm(100),"pnorm")

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  rnorm(100)
D = 0.063339, p-value = 0.8172
alternative hypothesis: two-sided
```

- p=0.82, Test accepts – all good.

- Do normally distributed data follow an exponential distribution?

```
> ks.test(abs(rnorm(100)),"pexp")

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  abs(rnorm(100))
D = 0.1113, p-value = 0.1678
alternative hypothesis: two-sided
```

- p=0.17, Test accepts – not good.

Tests For Normality

- With $X \sim N(0,1), E[\,|X|\,] = \sqrt{\frac{2}{\pi}} \approx 0.8.$

  What happens if we pertube our data a little? Say, 1% on average:

$$Z = X + 0.008 * Y$$

With Y exponentially distributed (mean 1).

More than 99.99% of the variation comes from X.

```
> ks.test(rnorm(100)+0.008*rexp(100),"pnorm")

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  rnorm(100) + 0.008 * rexp(100)
D = 0.094395, p-value = 0.335
alternative hypothesis: two-sided
```

p=0.34; so little effect that it isn't recognized. All good.

Tests For Normality

- Do the same with many data, more than 100, say 100.000:

```
> ks.test(rnorm(100000)+0.01*rexp(100000),"pnorm")

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  rnorm(1e+05) + 0.008 * rexp(1e+05)
D = 0.0057518, p-value = 0.002676
alternative hypothesis: two-sided
```

- p=0.003 significant, even though we on average just pertube with 1% of the average value.

Even minor, irrelevant model aberrations are detected for large data;
For small data, many incorrect hypotheses are accepted.
Solution: Graphical tests and common sense.

Example: Crime data

- **Post hoc analysis – confidence and prediction intervals:**
- The confidence interval for the fitted line at co_0:

$$V(\hat{\alpha} + \hat{\beta} co_0) = V(\hat{\alpha}) + co_0^2 V(\hat{\beta}) + 2co_0 Cov(\hat{\alpha}, \hat{\beta})$$
$$= \sigma^2 \left(\frac{1}{n} + \frac{(co_0 - \overline{co})^2}{SSD_{co}} \right),$$

ie. confidence interval

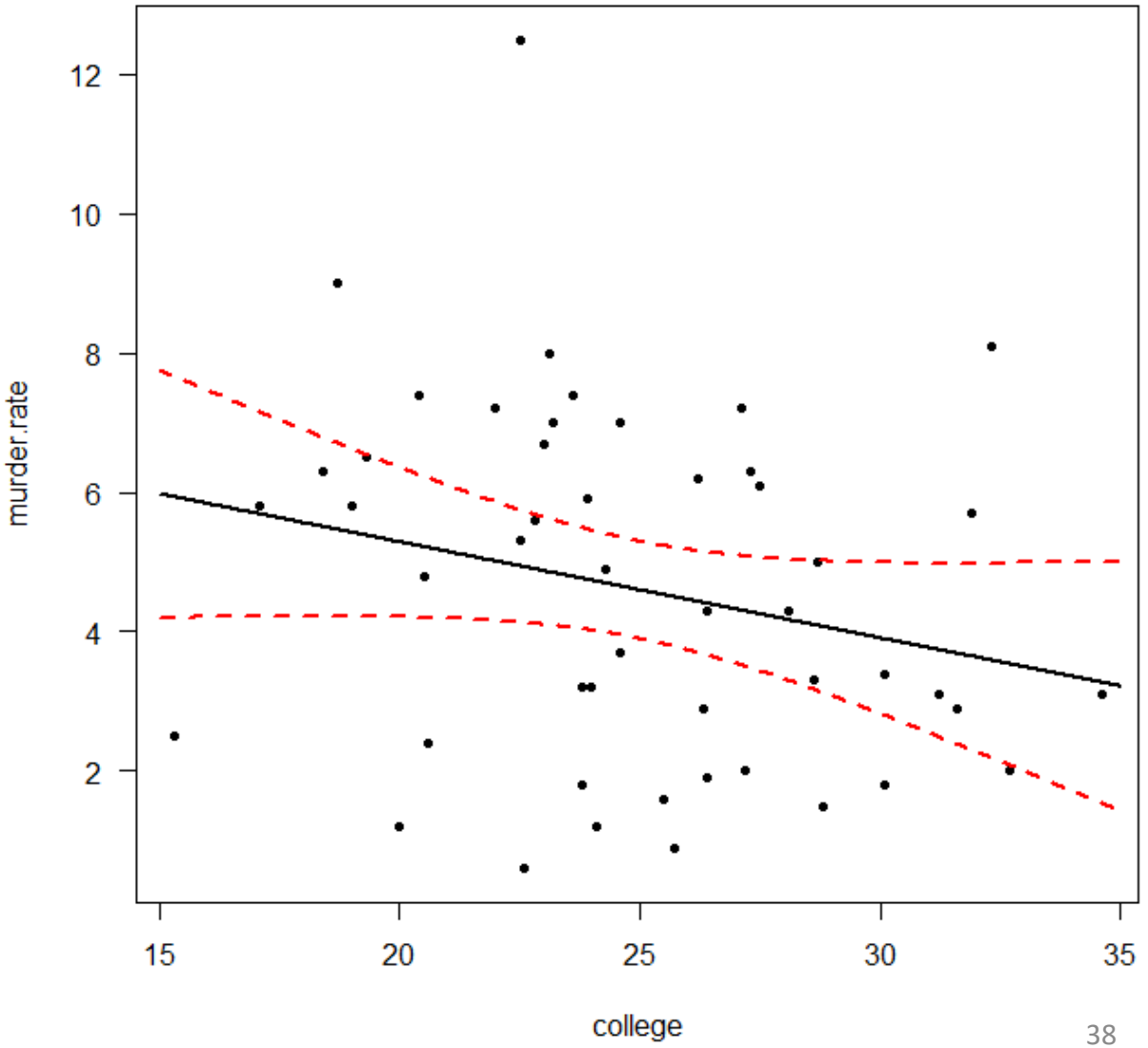
$$\hat{\alpha} + \hat{\beta} co_0 \pm t_{\frac{\alpha}{2}}(n-2)\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(co_0 - \overline{co})^2}{SSD_{co}}}$$

With $t_{\frac{\alpha}{2}}(n-2)$ the $\alpha/2$ percentile in a t-distribution with $n-2$ degrees of freedom.

- Here, $\alpha = 0.05$, $n = 50$, and thus $t_{\frac{\alpha}{2}}(n-2) =$ -2.010635.

Example: Crime data

```
>xval <- seq(from=15, to=35, length.out=500)
>newData <- data.frame(college=xval)
>Pred.ci <- predict(reg2, newdata=newData,
+                   interval="confidence",
+                   level=.95)

>plot(murder.rate ~ college, data = crime50, pch = 20, las = 1)
>lines(xval, Pred.ci[, "fit"], lwd=2)  ## or use: abline(reg2)
>lines(xval, Pred.ci[, "lwr"], lty=2, col="red", lwd=2)
>lines(xval, Pred.ci[, "upr"], lty=2, col="red", lwd=2)
```



Example: Crime data

- **Post hoc analysis – confidence and prediction intervals:**
- The prediction interval for a new observation with college rate co_0:

$$V\left(Y - \left(\hat{\alpha} + \hat{\beta} co_0\right)\right) = V(Y) + V\left(\hat{\alpha} + \hat{\beta} co_0\right)$$
$$= \sigma^2 \left(1 + \frac{1}{n} + \frac{(co_0 - \overline{co})^2}{SSD_{co}}\right),$$

ie. prediction interval

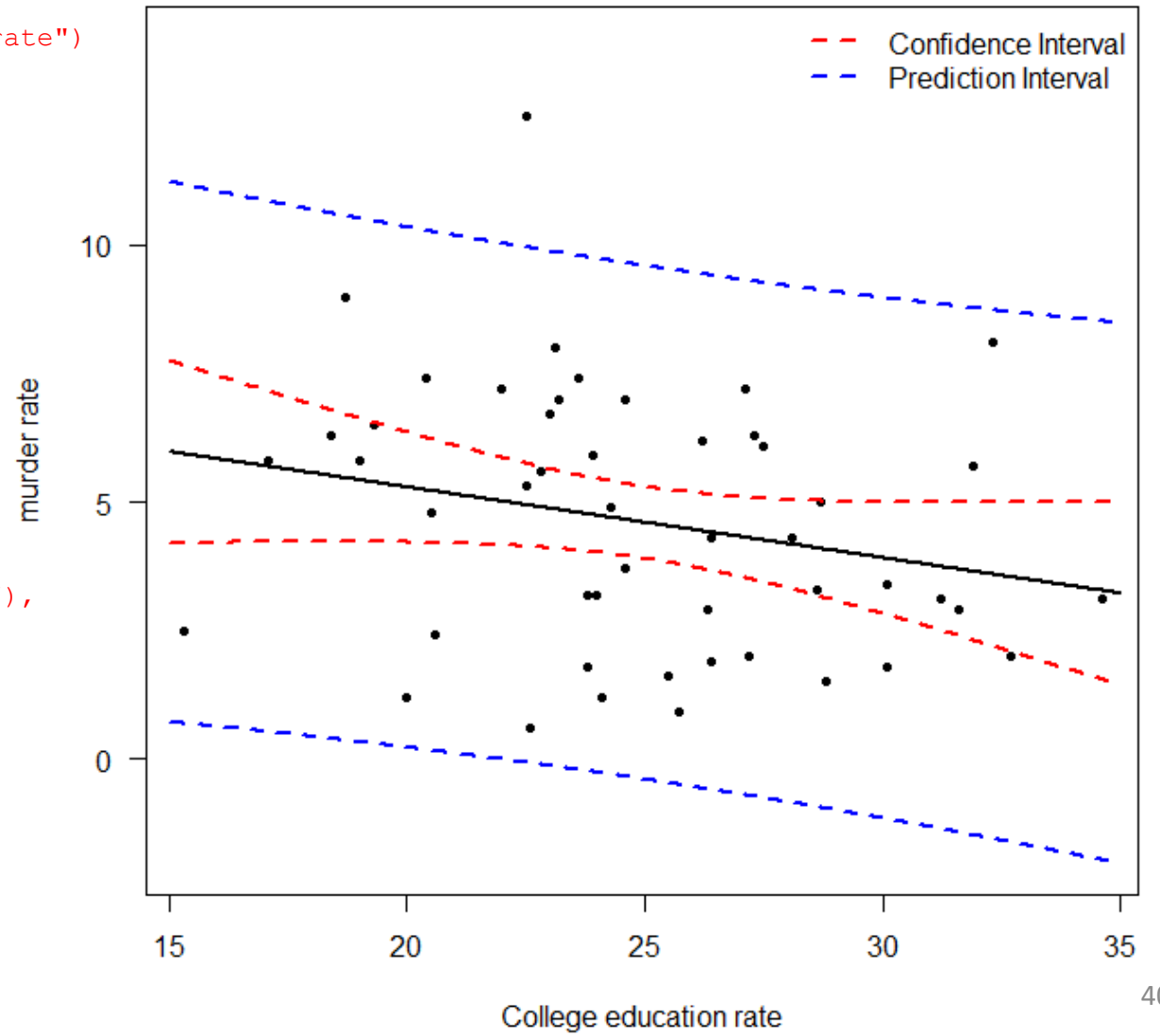
$$\hat{\alpha} + \hat{\beta} co_0 \pm t_{\frac{\alpha}{2}}(n-2)\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(co_0 - \overline{co})^2}{SSD_{co}}}$$

With $t_{\frac{\alpha}{2}}(n-2)$ the $\alpha/2$ percentile in a t-distribution with $n-2$ degrees of freedom.

- Wider than the confidence interval.

Example: Crime data

```
>plot(murder.rate ~ college, data = crime50, pch = 20, las = 1,
+     ylim=c(-2,14),xlab="College education rate",ylab="murder rate")
>lines(xval, Pred.ci[, "fit"], lwd=2)
>lines(xval, Pred.ci[, "lwr"], lty=2, col="red", lwd=2)
>lines(xval, Pred.ci[, "upr"], lty=2, col="red", lwd=2)

>## Prediction interval for a new observation:
>Pred.pi <- predict(reg2, newdata=newData,
>                   interval="prediction")

>## Add prediction intervals to plot:
>lines(xval, Pred.pi[, "lwr"], lty=2,
+      col="blue", lwd=2)
>lines(xval, Pred.pi[, "upr"], lty=2,
+      col="blue", lwd=2)
>legend("topright",c("Confidence Interval","Prediction Interval"),
+       lty=rep(2,1),lwd=rep(2,2),col=c("red","blue"),bty="n")
```



- Points outside PI: 1
- Expected: 2.5

Example: Crime data

- Influential observations.

- How much does the model change if an observation is left out?
- That is measured with the so-called

Cook's distance D_i :

- $D_i = \frac{\|\hat{Y} - \hat{Y}_{(i)}\|^2}{p\,\hat{\sigma}^2}$,

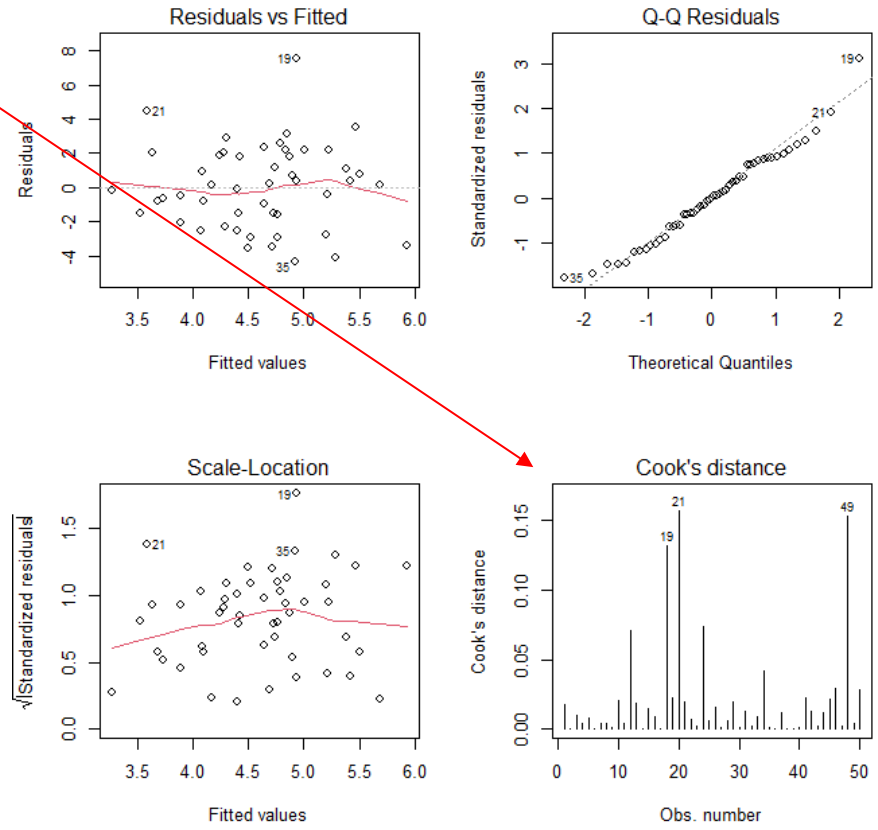
Where $\hat{Y}$ is the fitted values, $\hat{Y}_{(i)}$ is the fitted values when observation i is left out of the estimation, and p is the rank of the design matrix X.

- With h_{ii} the *leverage* of observation i, the diagnonal element of $P_M = X(X^TX)^{-1}X^T$:

$h_{ii} = X_i(X^TX)^{-1}X_i^T$;
$e_i = Y_i - \hat{Y}_i$,

a representation is

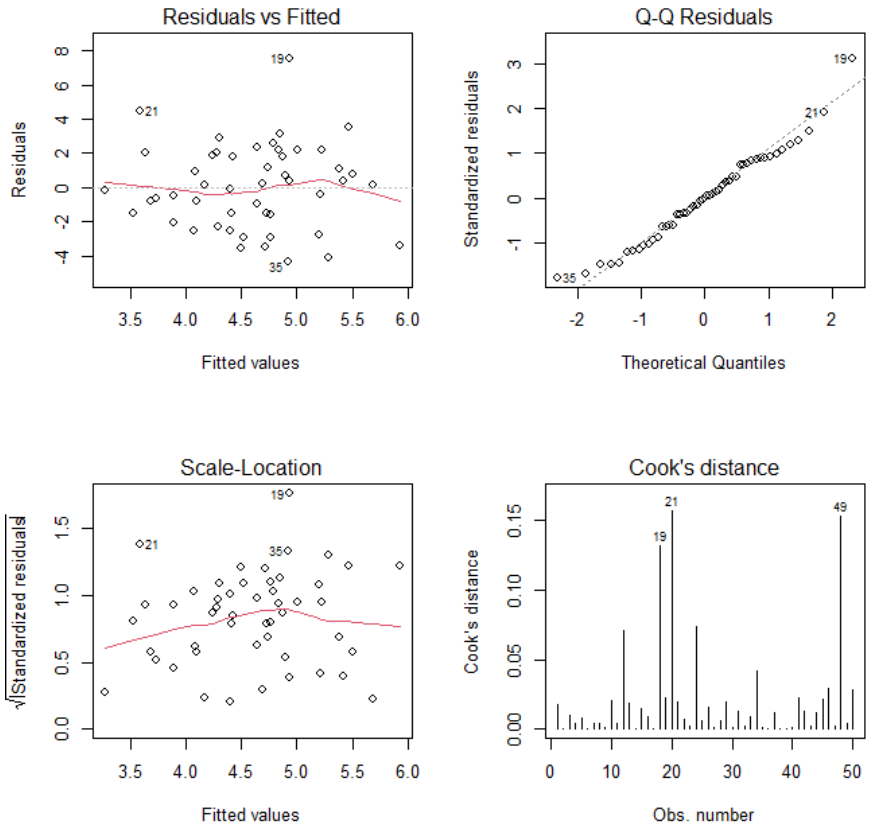
$$D_i = \frac{e_i^2}{p\hat{\sigma}^2}\frac{h_{ii}}{(1-h_{ii})^2}$$



Note that Cook's distance is high when both the leverage h_{ii} and the (squared) residual e_i^2 is high.

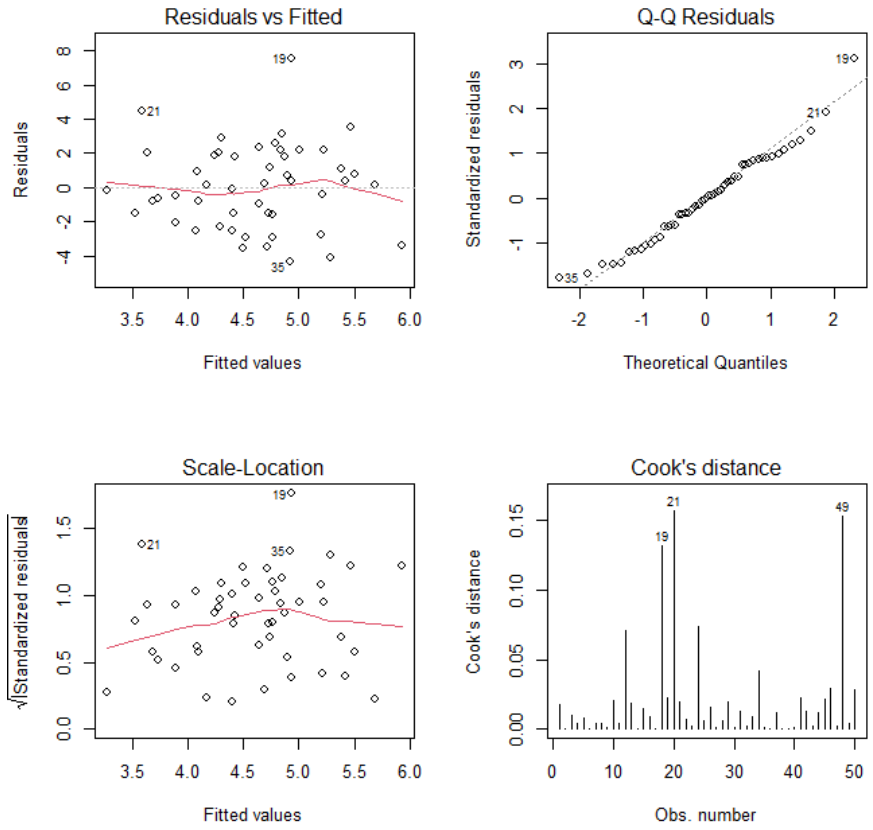
Example: Crime data

- When Cook's distance is high for an observation, results are affected by that single observation.
- If conclusions rely on the presence of single observations, the conclusions are *fragile*.
- We call an observation with a high Cook's distance for an *influential observation*.
- Influential observations should be investigated to see if they are indeed correct, and due to insight into the data generating mechanism; but they are valid data if they are not outliers, and important for the conclusion.
- In the presence of influential observations, reservations in the conclusion need to be made.



Example: Crime data

- Outliers are often influential observations.
- Outliers are **ALWAYS** corrected or removed;
- Influential observations which are not outliers are subject to investigation and possible correction, but they are **NEVER** removed.
- Reason: No indication of a different data generating mechanism.

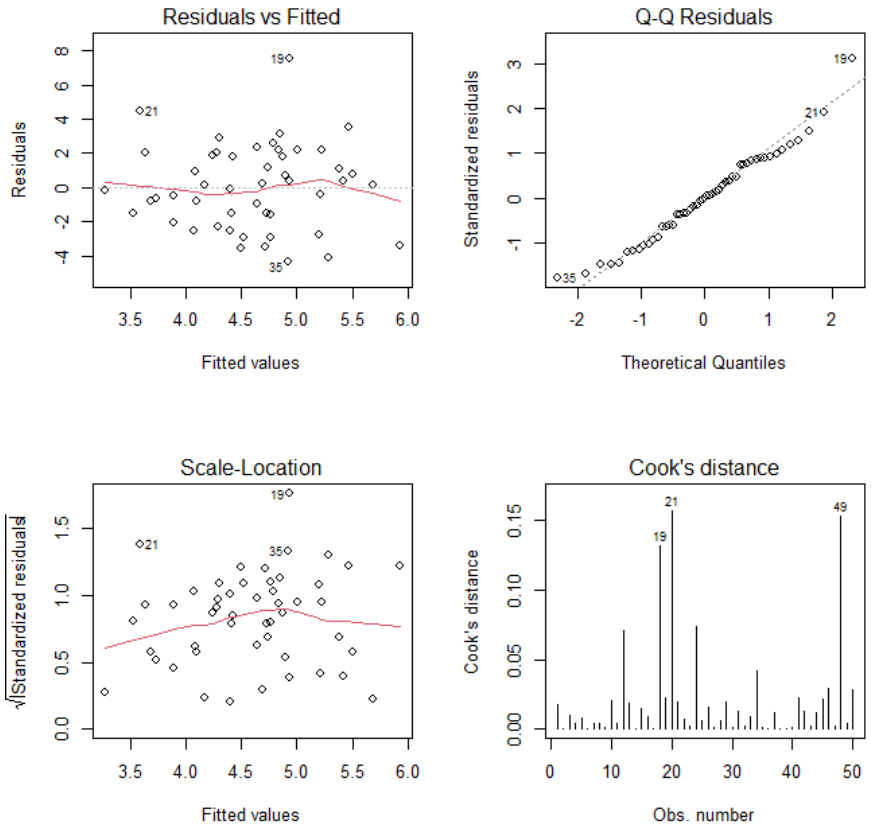


Example: Crime data

- When is Cook's distance big?
- No universal agreement about that.

- **Advise:** Do **not** spend time on size dependent formulas like $D > \frac{1}{n}$, $D > \frac{1}{n-p}$ etc.
- If $D > 1$ you should *always* investigate; if $D > \frac{1}{2}$ you should *likely* investigate; If $D < \frac{1}{2}$ *don't bother*.

- With the crime data, a $\frac{1}{n}$ or $\frac{1}{n-p}$ cut-off would leave 5 states for investigation. A $\frac{1}{2}$ cut-off leaves none. We will investigate none.



Example: Crime data

	Estimate	Std..Error	Lower	Upper	p.value
(Intercept)	8.04	2.06	3.89	12.19	<0.001
college	-0.14	0.08	-0.30	0.03	0.0977

The fitted model rataing murder rate in US states with college education rates has the form:

$$murder.rate_i = 8.04 - 0.14college_i + \varepsilon_i, \varepsilon_i \sim N(0, 2.46^2), i = 1, \dots, n$$

- Washington DC was left out of the data as it is not a state, and it exhibited a different data generating mechanism.

Example: Timber Hardness

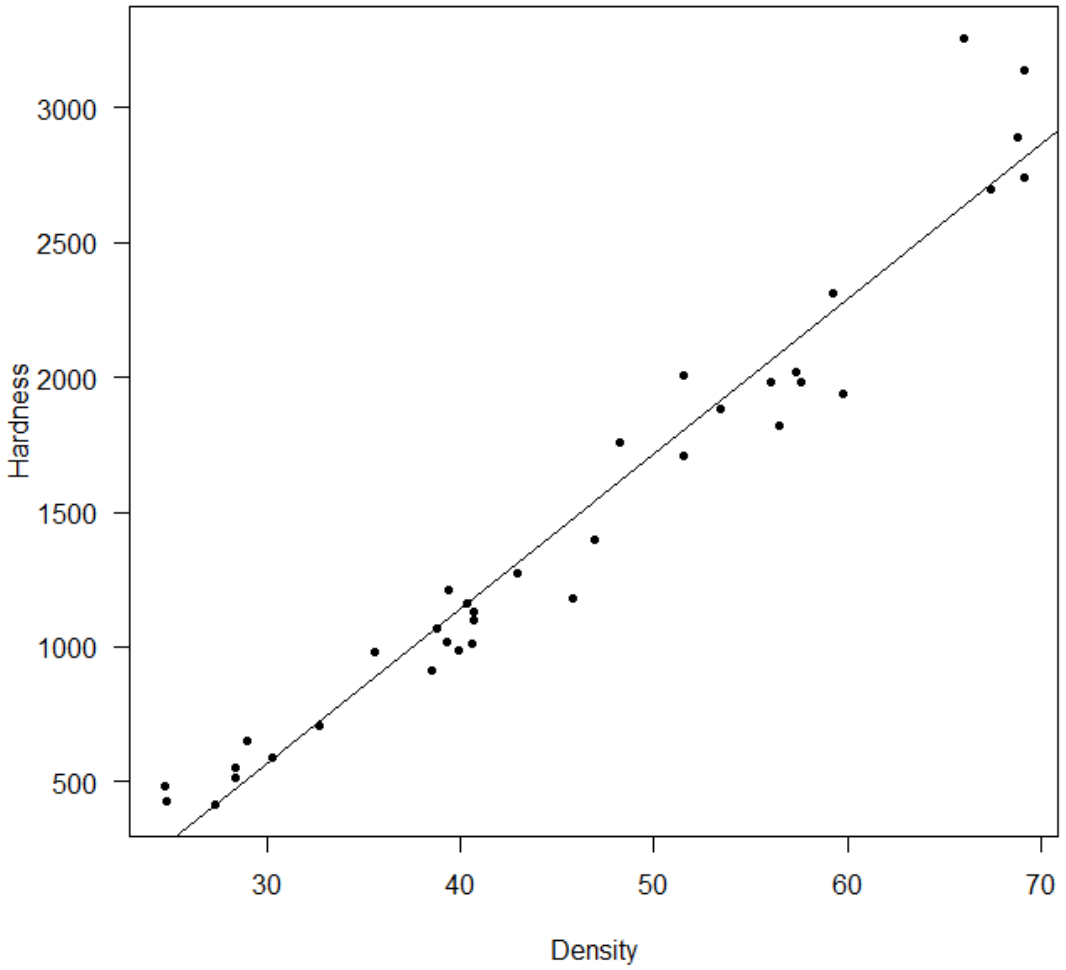
- 36 observations of timber hardness and density from trees in Australia.
- The aim of the study: To estimate parameters that determine the relationship between timber hardness and density, to predict (unknown) timber hardness in future samples based on the simpler measurement of timber density.
- First model: Simple regression.

$$Hardness_i = \alpha + \beta Density_i + \varepsilon_i, \varepsilon_i \, iid \sim N(0, \sigma^2), i = 1, \dots, 36.$$

- Does this model fit?

Example: Timber Hardness

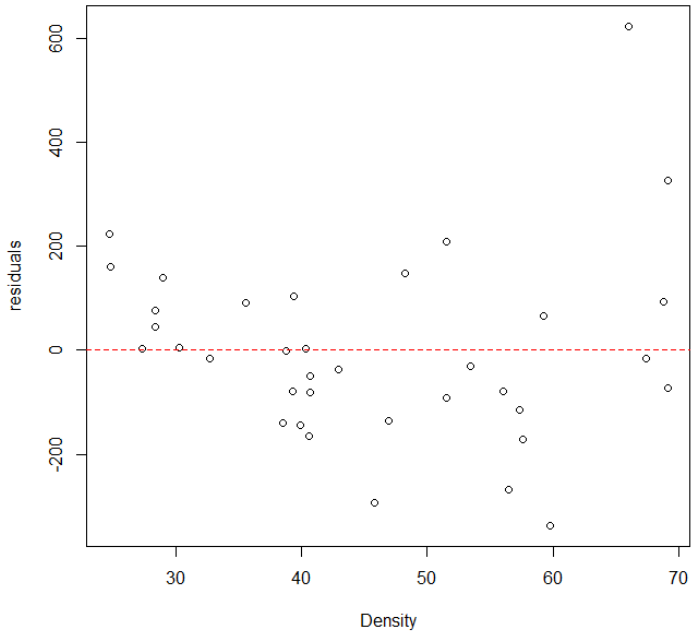
```
>janka <- read.table("Data/janka.txt", header=TRUE, quote="\"")
> names(janka) <- c("Density", "Hardness")
> summary(janka)
    Density          Hardness
 Min.   :24.70   Min.   : 413.0
 1st Qu.:37.77   1st Qu.: 962.8
 Median :41.80   Median :1195.0
 Mean   :45.73   Mean   :1469.5
 3rd Qu.:56.70   3rd Qu.:1980.0
 Max.   :69.10   Max.   :3260.0

> reg3 <- lm(Hardness ~ Density, data = janka)
> plot(Hardness ~ Density, data = janka, pch = 20, las = 1)
> abline(reg3)
```

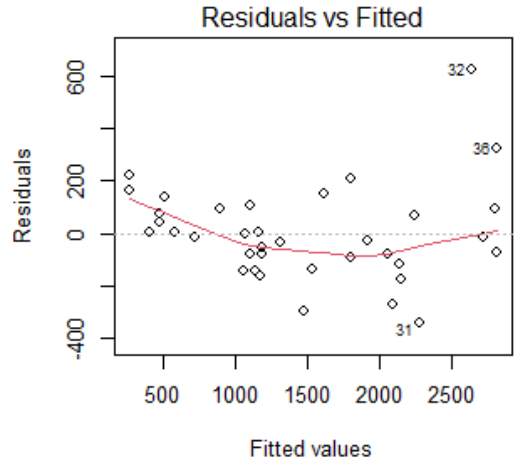


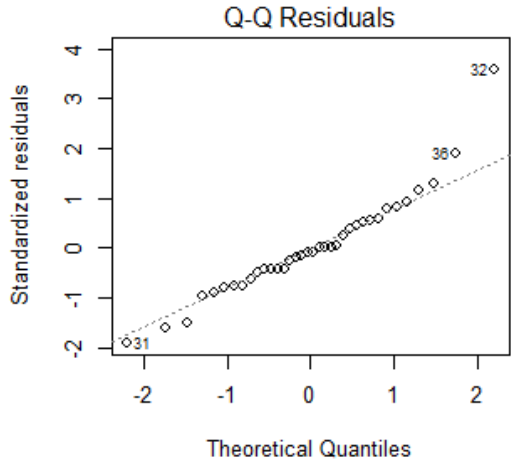
Example: Timber Hardness

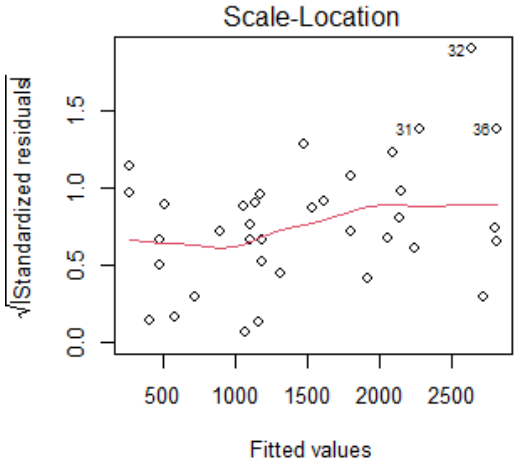
- Model fit:

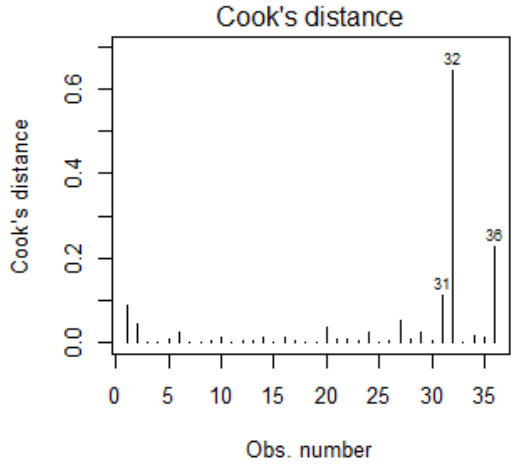
```
> par(mfrow=c(2, 2))
> plot(reg3, which=1:4)
> par(mfrow=c(1, 1))

> plot(residuals(reg3) ~ Density, data = janka,
+         ylab="residuals")
> abline(h = 0,col="red",lty=2)
```







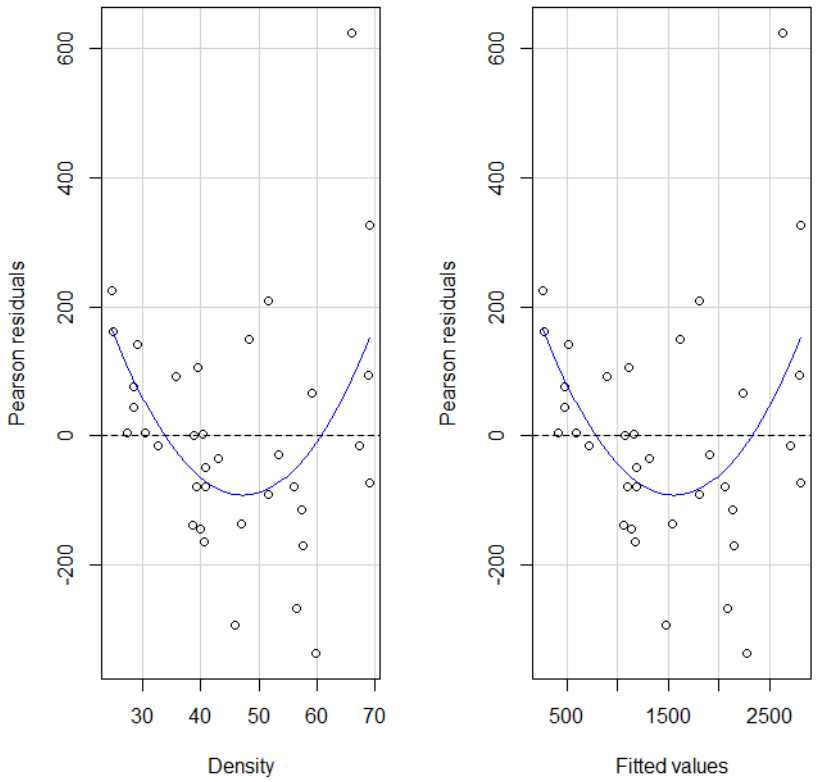




Example: Timber Hardness

- Model fit:

```
> residualPlots(reg3)
           Test stat Pr(>|Test stat|)
Density       3.2483          0.002669 **
Tukey test    3.2483          0.001161 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Example: Timber Hardness

- The lognormal distribution:

$$X = e^Z, Z \sim N(\mu, \sigma^2)$$

Coefficient of variation:

$$CV = \frac{\sqrt{V(X)}}{E(X)} = \frac{\sqrt{\left(e^{\sigma^2} - 1\right)e^{2\mu+\sigma^2}}}{e^{\mu+\frac{1}{2}\sigma^2}} = \sqrt{\left(e^{\sigma^2} - 1\right)}$$

Independent of μ when $\sigma^2$is fixed.

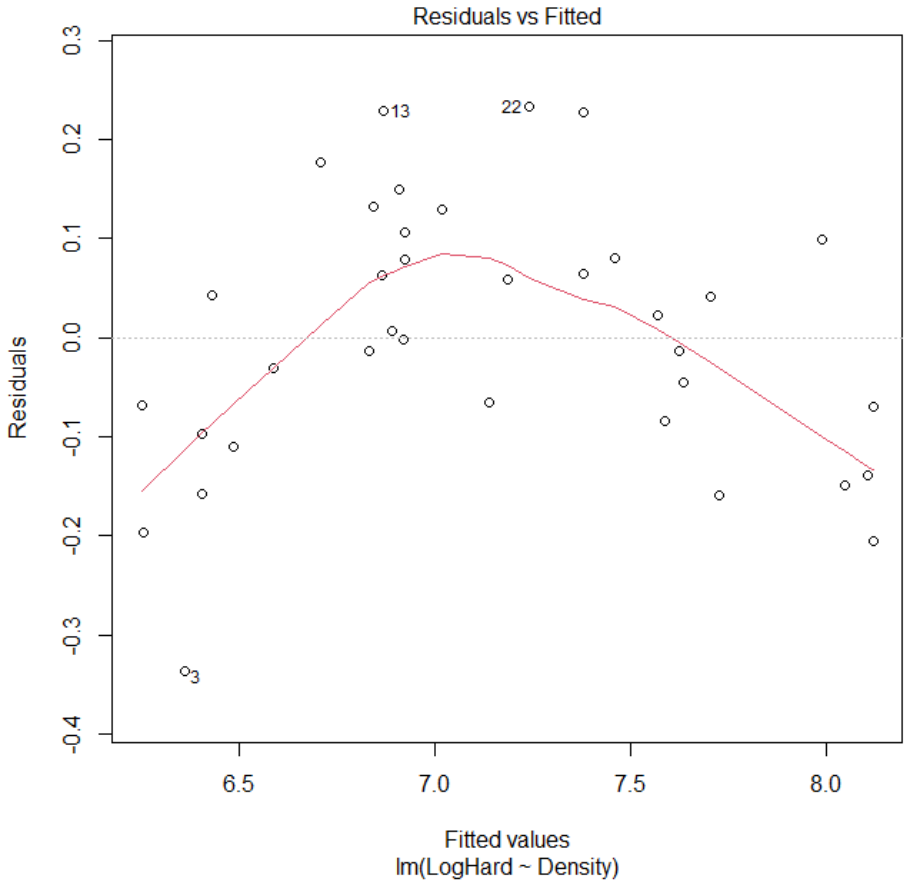
- For fixed σ^2, SD varies proportional with the mean!

Example: Timber Hardness

- **Data transformation:** Model $\log(Y)$:

```
janka$LogHard <- log(janka$Hardness)
reg4 <- lm(LogHard ~ Density, data = janka)
plot(reg4, which=1)
```

- No longer a trumpet shape, but still curvature.

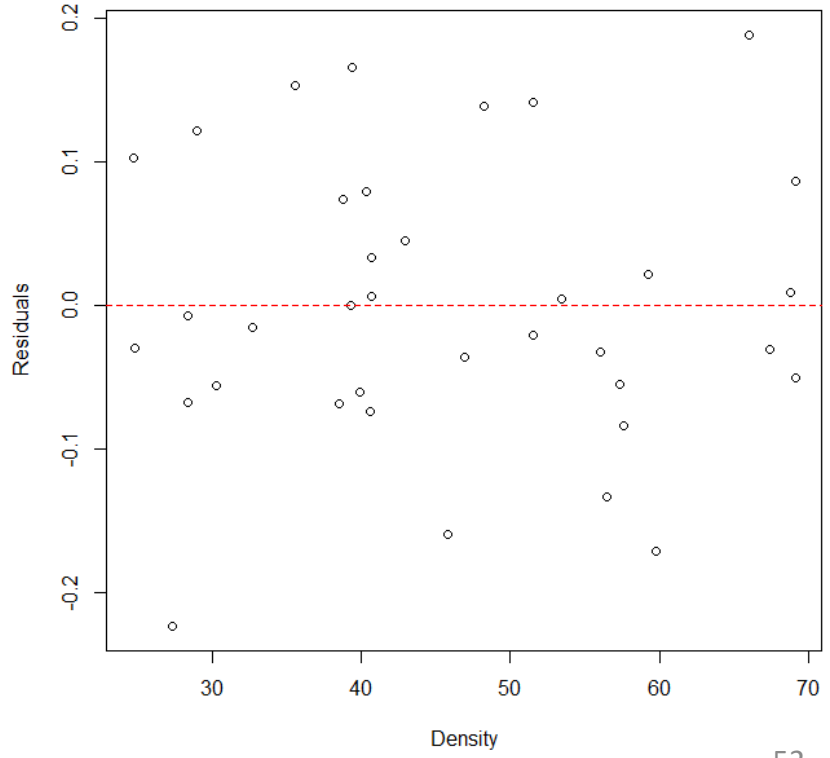


Example: Timber Hardness

- **2nd order Taylor approximation:**

$$log(Y_i) = \alpha + \beta_1 density_i + \beta_2 density_i^2 + \varepsilon_i, i = 1, \ldots, n$$

```
> reg5 <- lm(LogHard ~ Density + I(Density^2), data = janka)
> plot(residuals(reg5) ~ Density, data = janka,
+      ylab="Residuals")
> abline(h = 0,lty=2,col="red")
```

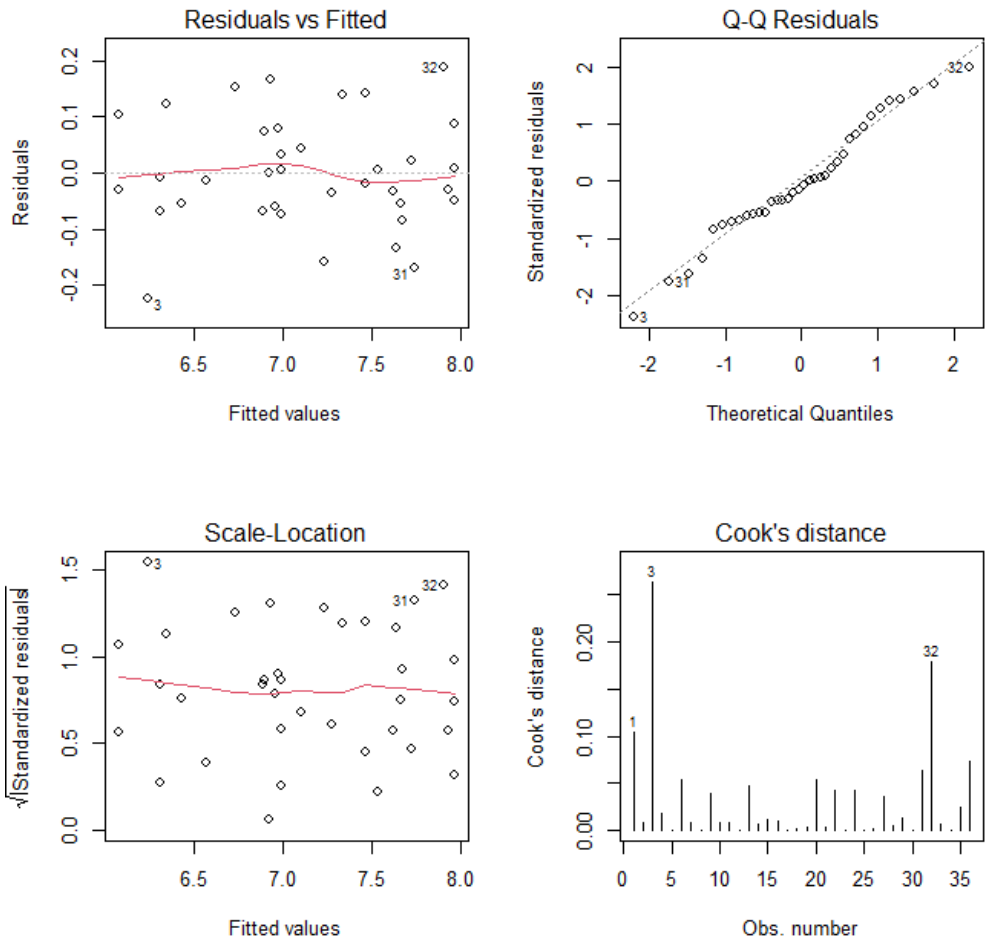


- Much better…

Example: Timber Hardness

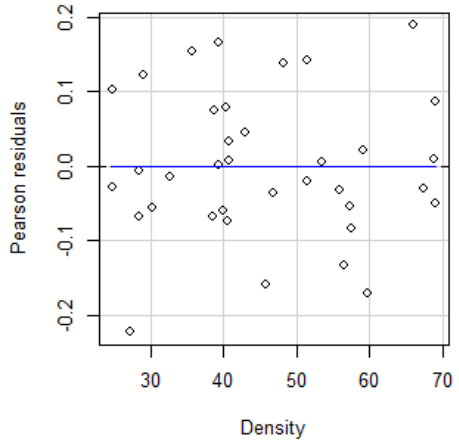
```
par(mfrow=c(2, 2))
plot(reg5, which=1:4)
par(mfrow=c(1, 1))
```

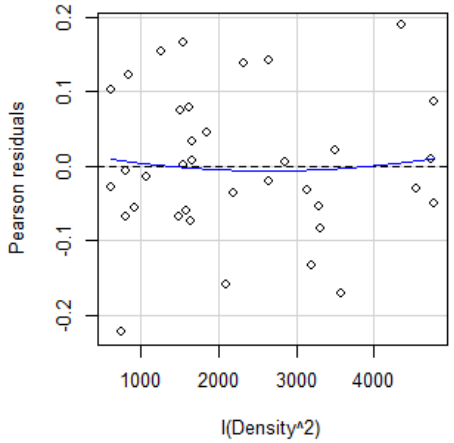
- Variance homogeneity, linearity (previous slide), normality and no appearant influential observations.

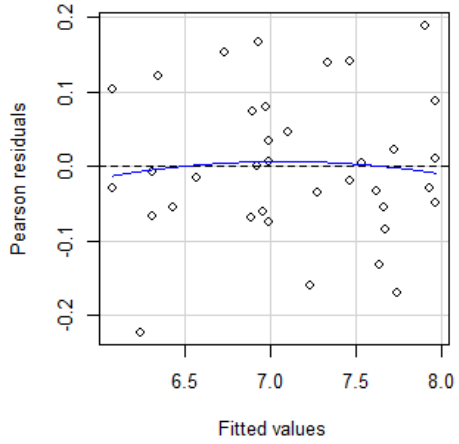


Example: Timber Hardness

```
> residualPlots(reg5)
               Test stat Pr(>|Test stat|)
Density          -2.4063           0.02206 *
I(Density^2)      1.3611           0.18300
Tukey test       -1.2757           0.20207
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```







- Looks okay, confirming the need for the square term.

Example: Timber Hardness

```
> summary(reg5)

Call:
lm(formula = LogHard ~ Density + I(Density^2), data = janka)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22331 -0.05708 -0.01104  0.07500  0.18871

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.138e+00  2.087e-01  19.828  < 2e-16 ***
Density       9.152e-02  9.305e-03   9.835 2.45e-11 ***
I(Density^2) -5.228e-04  9.764e-05  -5.354 6.49e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1008 on 33 degrees of freedom
Multiple R-squared:  0.9723,    Adjusted R-squared:  0.9706
F-statistic: 578.9 on 2 and 33 DF,  p-value: < 2.2e-16

> drop1(reg5,test="F")
Single term deletions

Model:
LogHard ~ Density + I(Density^2)
             Df Sum of Sq     RSS     AIC F value    Pr(>F)
<none>                    0.33510 -162.37
Density       1   0.98221 1.31731 -115.08  96.725 2.452e-11 ***
I(Density^2)  1   0.29111 0.62621 -141.86  28.668 6.486e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example: Timber Hardness

Final model:

$$log(Y_i) = \alpha + \beta_1 density_i + \beta_2 density_i^2 + \varepsilon_i, i = 1, \dots, n$$

On the original scale:

$$Y_i = a \cdot b_1^{density_i} \cdot b_2^{density^2} \cdot \widetilde{\varepsilon_i}, \mathrm{i} = 1, \dots, n$$

with

$$a = e^{\alpha}, b_1 = e^{\beta_1}, b_2 = e^{\beta_2}, log(\widetilde{\varepsilon_i})\ iid, \sim N(0, \sigma^2).$$

- Multiplicative model!

Example: Timber Hardness

Final model:

$$log(Y_i) = 4.14 + 0.0915 density_i - 0.000523 density_i^2 + \varepsilon_i, i = 1, \dots, n$$

On the original scale:

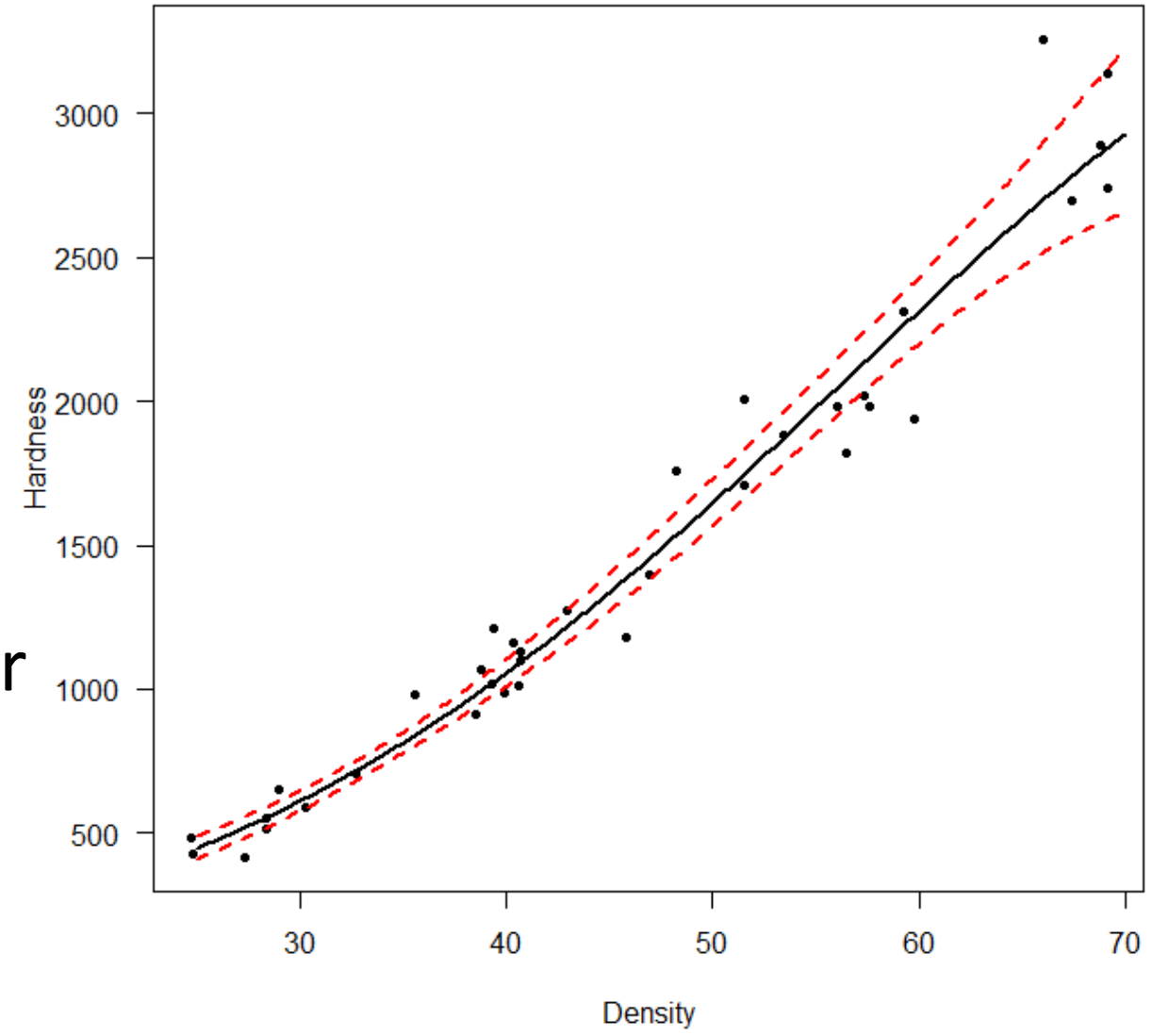
$$Y_i = 62.80 \cdot 1.0958^{density_i} \cdot 0.9994771^{\ density^2} \cdot \widetilde{\varepsilon_i}, \mathrm{i} = 1, \dots, n$$

With

$$62.80 = e^{4.14}, 1.0958 = e^{0.0915}, 0.9994 = e^{-0.000523}, log(\widetilde{\varepsilon_i})\ iid, \sim N(0, 0.1008^2).$$

Example: Timber Hardness

```
>xval <- seq(from=25, to=70, length.out=500)
>newData <- data.frame(Density=xval)
>Pred.ci <- predict(reg5, newdata=newData,
+                   interval="confidence",
+                   level=.95)

>plot(Hardness ~ Density, data = janka, pch = 20, las = 1)
>lines(xval, exp(Pred.ci[, "fit"]), lwd=2)
>lines(xval, exp(Pred.ci[, "lwr"]), lty=2, col="red", lwd=2)
>lines(xval, exp(Pred.ci[, "upr"]), lty=2, col="red", lwd=2)
```



- Confidence interval increases for large density values but not for small.

Example: Timber Hardness

- Note that

$$E\left(logN(\mu,\sigma^2)\right) = e^{\mu+\sigma^2/2}$$

Strictly, we should have

$$\hat{Y} = exp\left(\hat{Z} + V\left(\hat{Z}\right)/2\right)$$

and not just

$$\hat{Y} = exp\left(\hat{Z}\right)$$

as on the previous slide.

- But the difference is often immaterial, and indeed it is so here:

```
> my.se.fit<-predict(reg5, newdata=newData,
+                    interval="confidence",
+                    level=.95,se.fit=T)$se.fit

> summary(exp(my.se.fit^2/2))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.000   1.000   1.001
```

- Multiplication factor: 1.

Exercises

- exam 2009, problem 3 (testing).
- Exam 2012 problems 5.1-5.4 (estimation).
- Brain Exercise, including extra questions (model selection).
- Process Exercise (first look at interactions).