# Applied Statistical Regression
## HS 2010 – Week 04

*Marcel Dettling*

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, October 18, 2010

# Applied Statistical Regression
## HS 2010 – Week 04

## *Course Organization*

The exercises will be held on the days that were planned according to the schedule given on the organization sheet!

NEW: the exercise lessons will (until further notice) ALWAYS take place at the computer labs, i.e. in the following rooms:

| | |
|---|---|
| HG E27 | Ag – Go |
| HG E26.1 | Ha – Pa |
| HG E26.3 | Pe – Zh |

# Applied Statistical Regression
## HS 2010 – Week 04

## *Multiple Linear Regression*

The model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

Assumptions:

- $E[\varepsilon_i] = 0$ , i.e. the hyper plane is the correct fit
- $Var(\varepsilon_i) = \sigma_\varepsilon^2$ , constant scatter for the error term
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ , uncorrelated errors

# Applied Statistical Regression
## HS 2010 – Week 04

# *An Example*

| City | Mortality | JanTemp | JulyTemp | RelHum | Rain | Educ | Dens | NonWhite | WhiteCollar | Pop | House | Income | HC | NOx | SO2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Akron, OH | 921.87 | 27 | 71 | 59 | 36 | 11.4 | 3243 | 8.8 | 42.6 | 660328 | 3.34 | 29560 | 21 | 15 | 59 |
| Albany, NY | 997.87 | 23 | 72 | 57 | 35 | 11 | 4281 | 3.5 | 50.7 | 835880 | 3.14 | 31458 | 8 | 10 | 39 |
| Allentown, PA | 962.35 | 29 | 74 | 54 | 44 | 9.8 | 4260 | 0.8 | 39.4 | 635481 | 3.21 | 31856 | 6 | 6 | 33 |
| Atlanta, GA | 982.29 | 45 | 79 | 56 | 47 | 11.1 | 3125 | 27.1 | 50.2 | 2138231 | 3.41 | 32452 | 18 | 8 | 24 |
| Baltimore, MD | 1071.29 | 35 | 77 | 55 | 43 | 9.6 | 6441 | 24.4 | 43.7 | 2199531 | 3.44 | 32368 | 43 | 38 | 206 |
| Birmingham, AL | 1030.38 | 45 | 80 | 54 | 53 | 10.2 | 3325 | 38.5 | 43.1 | 883946 | 3.45 | 27835 | 30 | 32 | 72 |

# Applied Statistical Regression
## HS 2010 – Week 04

## *Properties of the Estimates*

Gauss-Markov-Theorem:

The regression coefficients are unbiased estimates, and they fulfill the optimality condition of minimal variance among all linear, unbiased estimators (*BLUE*).

- $E[\hat{\beta}] = \beta$
- $Cov(\beta) = \sigma_\varepsilon^2 \cdot (X^T X)^{-1}$
- $\hat{\sigma}_\varepsilon^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} r_i^2$ (note: degrees of freedom!)

# Applied Statistical Regression
## HS 2010 – Week 04

## *If the Errors are Gaussian…*

While all of the above statements hold for arbitrary error distribution, we obtain some more, very useful properties by assuming i.i.d. Gaussian errors:

- $\hat{\beta} \sim N\left(\beta, \sigma_{\varepsilon}^{2}(X^{T}X)^{-1}\right)$

- $\hat{y} \sim N(X\beta, \sigma_{\varepsilon}^{2}H)$

- $\hat{\sigma}_{\varepsilon}^{2} \sim \dfrac{\sigma_{\varepsilon}^{2}}{n-p}\chi_{n-p}$

*What to do if the errors are non-Gaussian?*

# Applied Statistical Regression
## HS 2010 – Week 04

## *Individual Parameter Tests*

If we are interested whether the j[th] predictor variable is relevant, we can test the hypothesis

$$H_0 : \beta_j = 0$$

against the alternative hypothesis

$$H_A : \beta_j \neq 0$$

We can derive the test statistic and its distribution:

$$T = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}_\varepsilon^2 (X^T X)^{-1}_{jj}}} \sim t_{n-(p+1)}$$

# *Individual Parameter Tests*

These tests quantify the effect of the predictor $x_j$ on the response Y after having subtracted the linear effect of all other predictor variables on Y.

Be careful, because of:

a) The *multiple testing problem*: when doing many tests, the total type II error increases. By how much: see blackboard

b) It can happen that all individual tests do not reject the null hypothesis, although some predictors have a significant effect on the response. Reason: correlated predictors!

# Applied Statistical Regression
## HS 2010 – Week 04

## *Global F-Test*

*Question*: is there *any* relation between predictors and response?

We test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

against the alternative

$$H_A : \beta_j \neq 0 \quad \text{for at least one j in 1,…, p}$$

The test statistic is:

$$F = \frac{n-(p+1)}{p} \cdot \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \sim F_{p,n-(p+1)}$$

# Applied Statistical Regression
## HS 2010 – Week 04

# *Partial F-Tests*

*Test the effects of p-q predictors simultaneously!*

We divide the model into 2 parts

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

So that we can test the hypotheses

$$H_0 : \beta_2 = 0 \text{ versus } H_A : \beta_2 \neq 0$$

We compute

$$SSR_{H_0} : \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2 \quad \text{and} \quad SSR_{H_A} : \sum_{i=1}^{n} (\hat{y}_i^{\%} - \bar{y}_i)^2$$

# *Partial F-Tests*

*Test the effects of p-q predictors simultaneously!*

The test statistic is

$$F = \frac{n-p-1}{p-q} \cdot \frac{SSR_{H_A} - SSR_{H_0}}{\sum_{i=1}^{n} (y_i - \hat{y}_i^{\%})^2} \sim F_{p-q,n-p-1}$$

Where do we need this?

- meteorological variables in the mortality dataset
- later, when we work with factor/dummy variables

# Applied Statistical Regression
## HS 2010 – Week 04

## *Coefficient of Determination*

The coefficient of determination, also called *multiple R-squared*, is aimed at describing the goodness-of-fit of the multiple linear regression model:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \in [0,1]$$

It shows the proportion of the total variance which has been explained by the predictors. The extreme cases 0 and 1 mean:…

# *Adjusted Coefficient of Determination*

If we add more and more predictor variables to the model, R-squared will always increase, and never decreases

*Is that a realistic goodness-of-fit measure?*

→ **NO, we better adjust for the number of predictors!**

$$adjR^2 = 1 - \frac{n-1}{n-(p+1)} \cdot \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \in [0,1]$$

# Applied Statistical Regression
## HS 2010 – Week 04

## *R-Output*

```
> summary(lm(Mortality~log(SO2)+NonWhite+Rain, data=mo…))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 773.0197    22.1852  34.844  < 2e-16 ***
log(SO2)     17.5019     3.5255   4.964 7.03e-06 ***
NonWhite      3.6493     0.5910   6.175 8.38e-08 ***
Rain          1.7635     0.4628   3.811 0.000352 ***
---
Residual standard error: 38.4 on 55 degrees of freedom
Multiple R-squared: 0.641,  Adjusted R-squared: 0.6214
F-statistic: 32.73 on 3 and 55 DF,  p-value: 2.834e-12
```

# Applied Statistical Regression
## HS 2010 – Week 04

## *Interpreting the Result*

*Does the SO2 concentration affect the mortality?*

→ Might be, might not be

→ There are only 3 predictors

→ We could suffer from confounding effects

→ Causality is always difficult, but…

The next step is to include all predictor variables that are present in the mortality dataset.

# Applied Statistical Regression
## HS 2010 – Week 04
## *More Predictors*

```
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.164e+03  2.939e+02   3.960 0.000258 ***
JanTemp     -1.669e+00  7.930e-01  -2.105 0.040790 *  
JulyTemp    -1.167e+00  1.939e+00  -0.602 0.550207    
RelHum       7.017e-01  1.105e+00   0.635 0.528644    
Rain         1.224e+00  5.490e-01   2.229 0.030742 *  
Educ        -1.108e+01  9.449e+00  -1.173 0.246981    
Dens         5.623e-03  4.482e-03   1.255 0.215940    
NonWhite     5.080e+00  1.012e+00   5.019 8.25e-06 ***
WhiteCollar -1.925e+00  1.264e+00  -1.523 0.134623    
Pop          2.071e-06  4.053e-06   0.511 0.611799    
House       -2.216e+01  4.040e+01  -0.548 0.586074    
Income       2.430e-04  1.328e-03   0.183 0.855617    
log(SO2)     6.833e+00  5.426e+00   1.259 0.214262    
---
Residual standard error:  36.2 on 46 degrees of freedom
Multiple R-squared: 0.7333,  Adjusted R-squared: 0.6637 
F-statistic: 10.54 on 12 and 46 DF,  p-value: 1.417e-09
```

# Applied Statistical Regression
## HS 2010 – Week 04

## *Some Thoughts on Collinearity*

a) With collinear predictors, inference (i.e. interpreting p-values from individual parameter tests and the global F-test) should be "handled with care"!

b) Drawing conclusions on causality should be left out.

c) However, the fitted values are not affected by this, and also prediction with a model fitted from collinear predictors is always fine.

Measuring collinearity: $VIF_j = \dfrac{1}{1-R_j^2}$

# Applied Statistical Regression
## HS 2010 – Week 04

# *Model Diagnostics*

Why do we need to do this?

**a) make sure that estimates and inference are valid**

- $E[\varepsilon_i] = 0$
- $Var(\varepsilon_i) = \sigma_\varepsilon^2$
- $Cov(\varepsilon_i, \varepsilon_j) = 0$
- $\varepsilon_i \sim N(0, \sigma_\varepsilon^2 I), \;\; i.i.d$

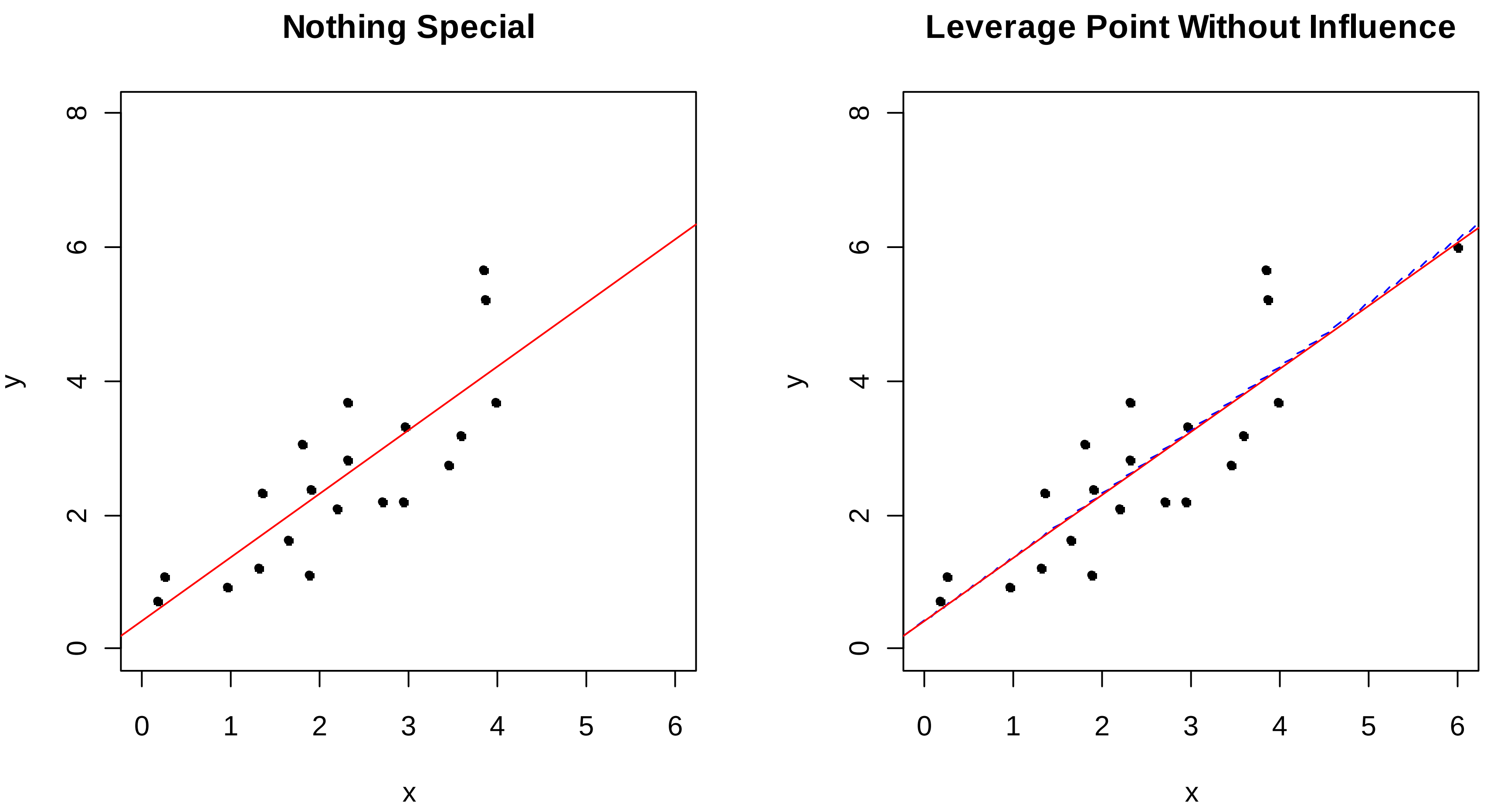
**b) improving the model (better fit, reliable conclusions)**

- variable transformations
- further predictors or interactions between them
- weighted regression or more general model

# *What Tools Do We Have?*

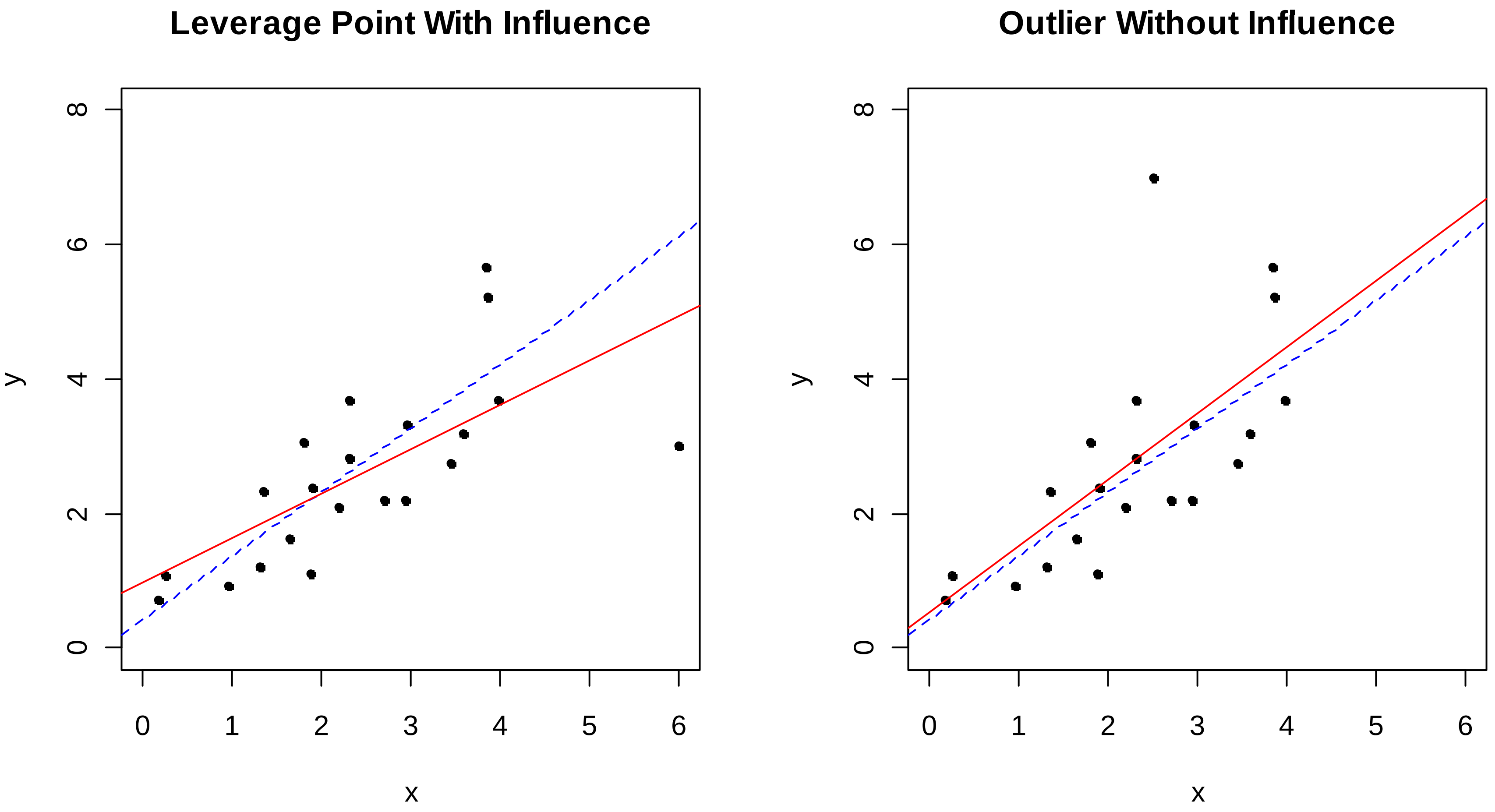
- Tukey-Anscombe plot
- Normal plot
- Scale-Location plot
- Serial Correlation plot
- Cook‘s Distance
- Leverage plot
- Residuals vs. predictors

# *Outliers and Influential Data Points*



# Applied Statistical Regression
## HS 2010 – Week 04

# *Outliers and Influential Data Points*



# Applied Statistical Regression
## HS 2010 – Week 04

## *How To Identify These Points?*

### 1) Poor man‘s approach

Redo the analysis n times by excluding each data point

### 2) Leverage

If we change $y_i$ by $\Delta y_i$, then $h_{ii}\Delta y_i$ is the change in $\hat{y}_i$
High leverage for a data point ($h_{ii} > 2(p+1)/n$) means that it forces the regression line to fit well to it.

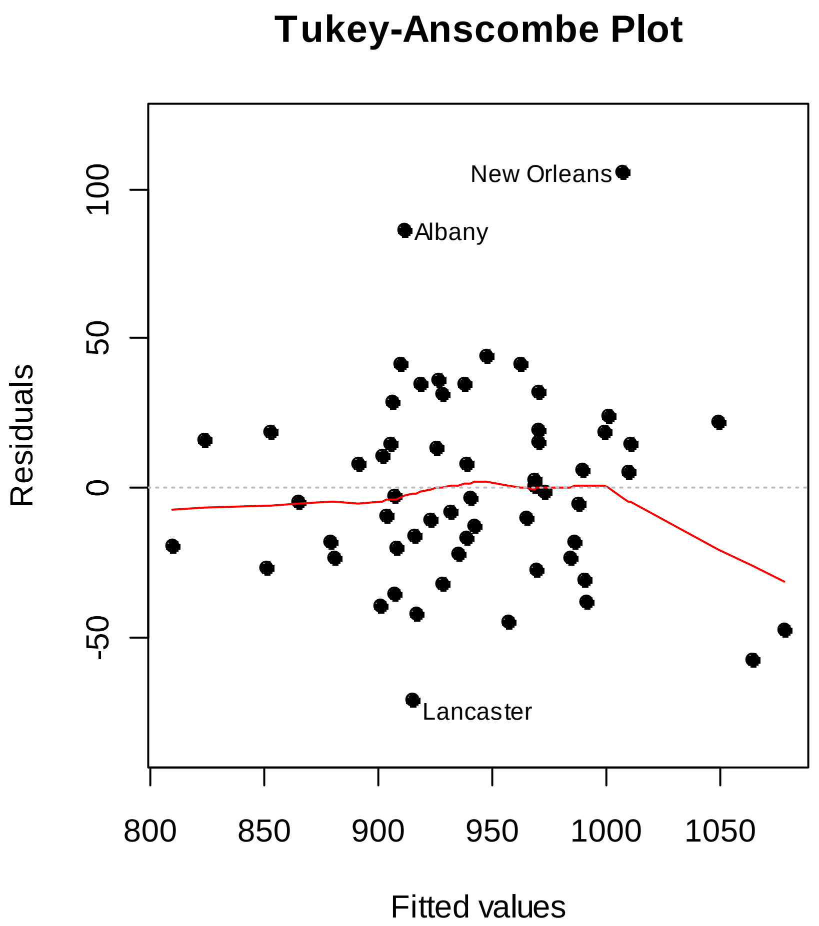
### 3) Cook‘s Distance

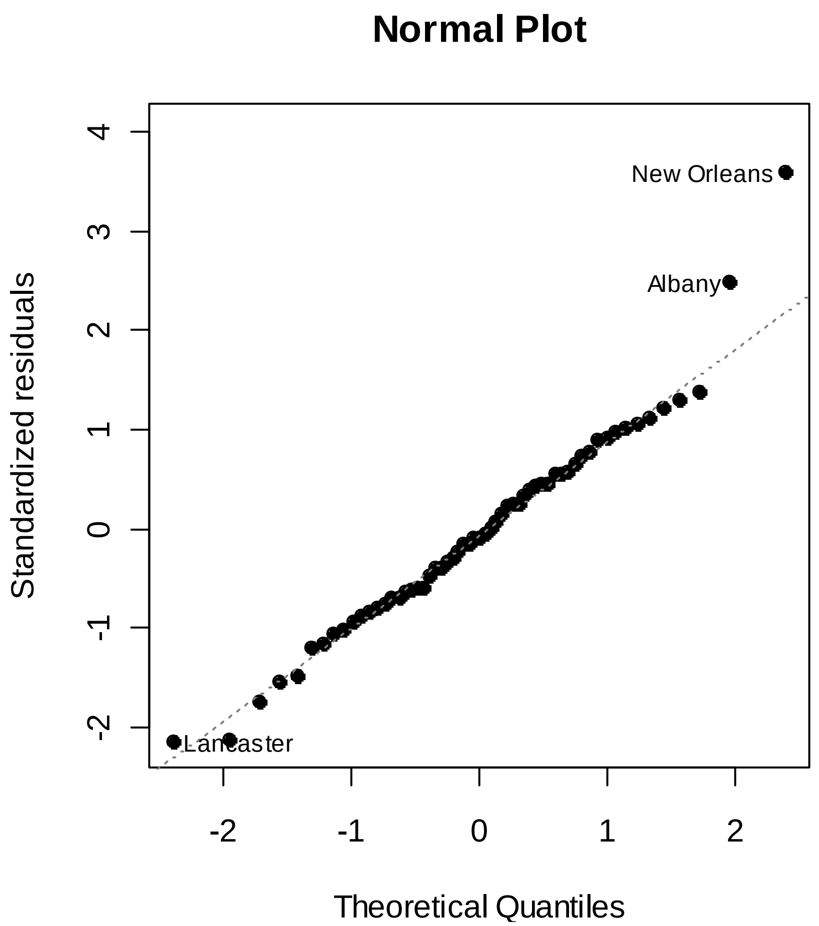
$$D_i = \frac{\sum(\hat{y}_j - y_{j(i)})^2}{(p+1)\sigma_\varepsilon^2} = \frac{h_{ii}}{1-h_{ii}} \cdot \frac{r_i^{*2}}{(p+1)}$$

Be careful if Cook‘s Distance > 1.

# Applied Statistical Regression
## HS 2010 – Week 04

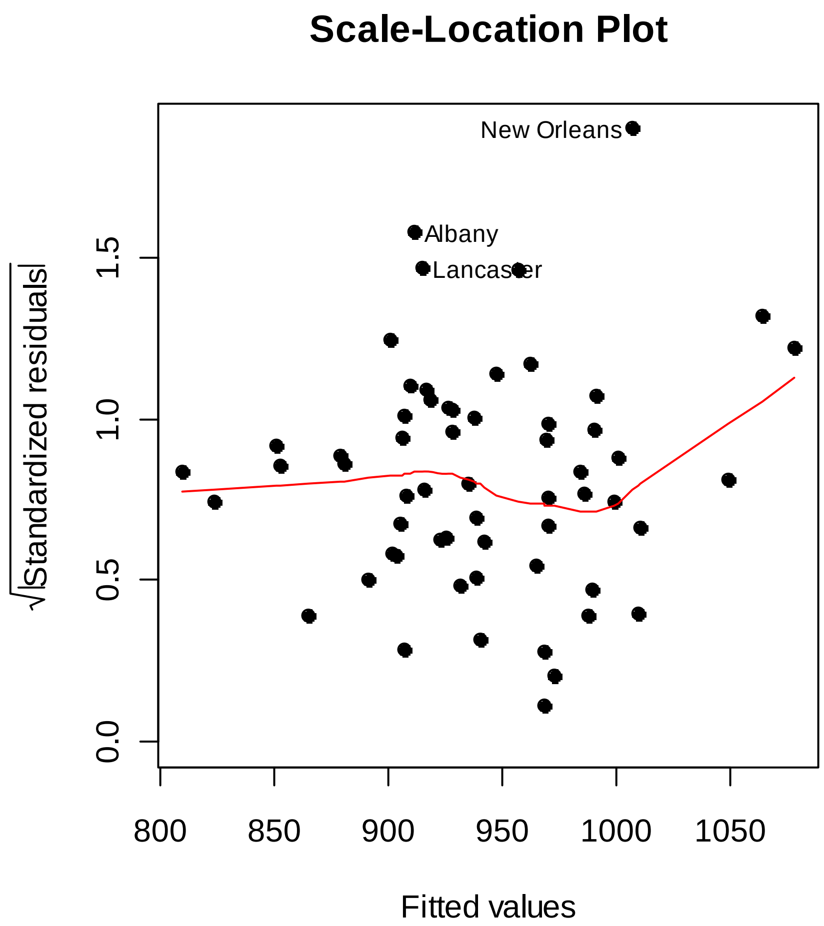
# *Model Diagnostics: Example*

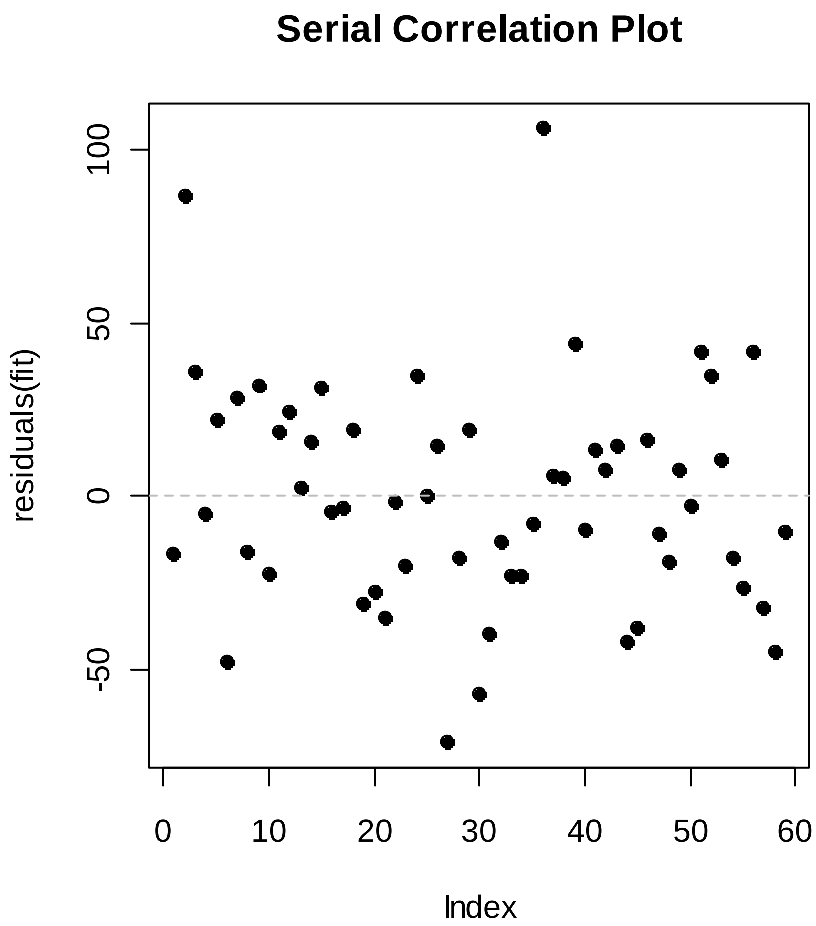


# Applied Statistical Regression
## HS 2010 – Week 04

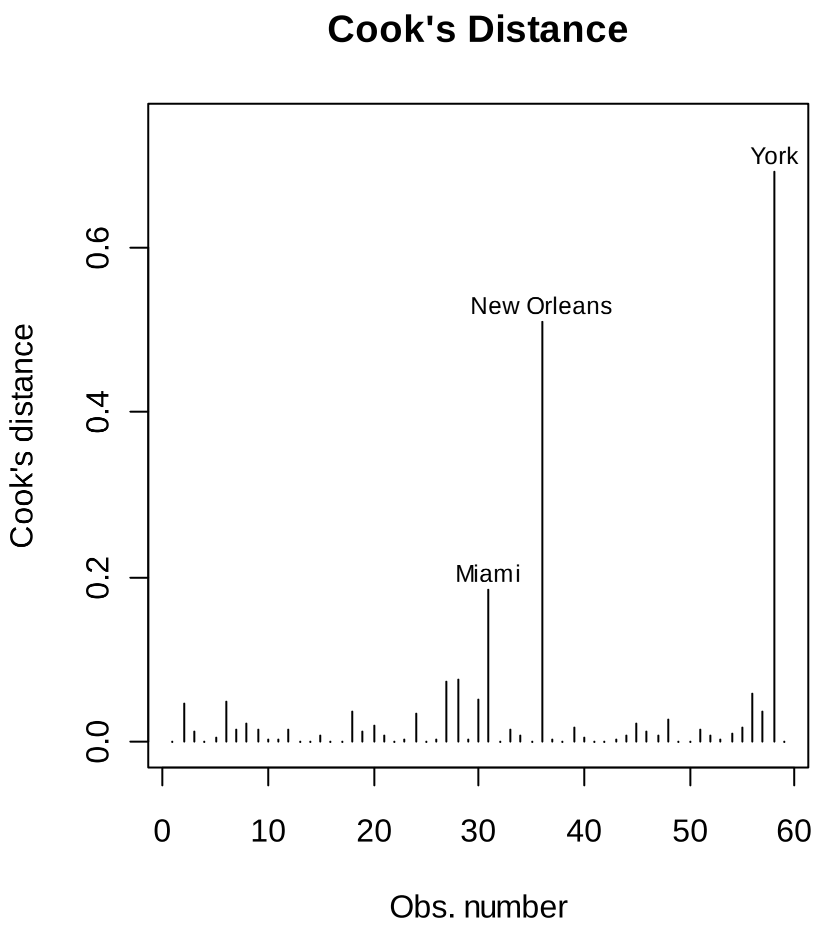
# *Model Diagnostics: Example*

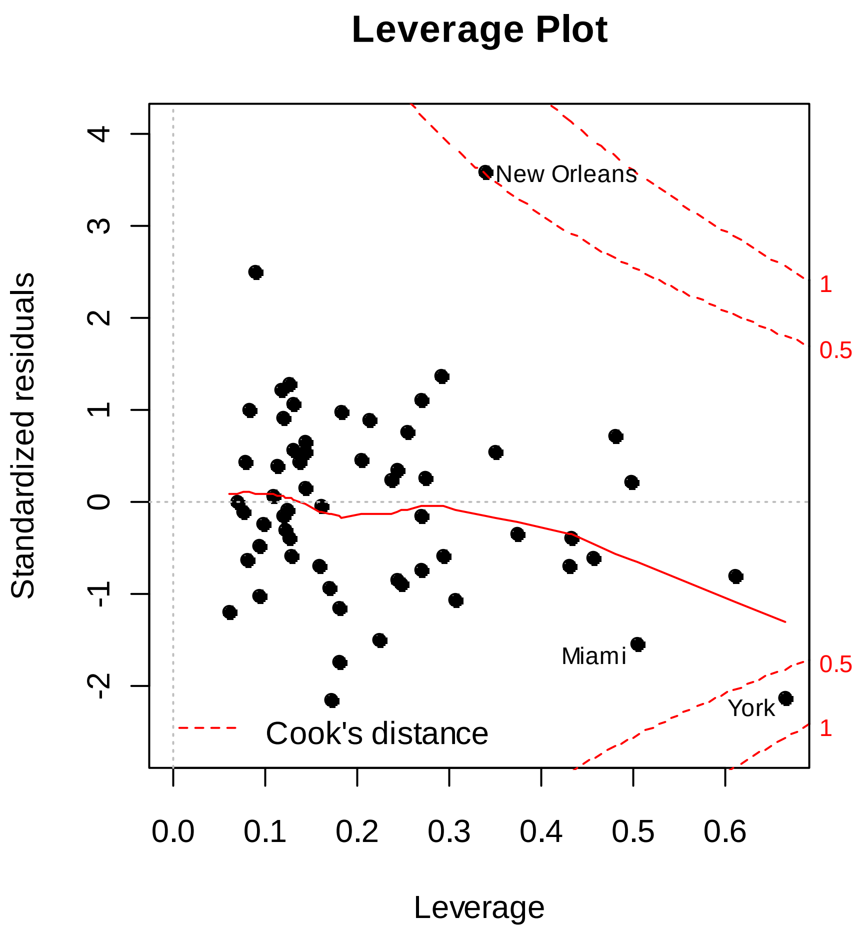


# Applied Statistical Regression
## HS 2010 – Week 04

# *Model Diagnostics: Example*



# *Model Diagnostics: Conclusions*

**Conclusions from the model diagnostics:**

- there are 2 influential data points: York and New Orleans
- they do not seem to be very strongly influential, but still:
- better to re-run the analysis without these and check results

**Results from that analysis:**

- log(SO2) is significant again!!!
- Residual standard error smaller
- Coefficient of determination higher
- Thus: better fit!

# Applied Statistical Regression
## HS 2010 – Week 04

# *Why Are They Influential?*