Zurich University of Applied Sciences

Applied Statistical Regression
HS 2010 – Week 05

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 25, 2010

Applied Statistical Regression
HS 2010 – Week 05

Mortality 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 05

Model Diagnostics

Why do we need to do this?

a) make sure that estimates and inference are valid

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

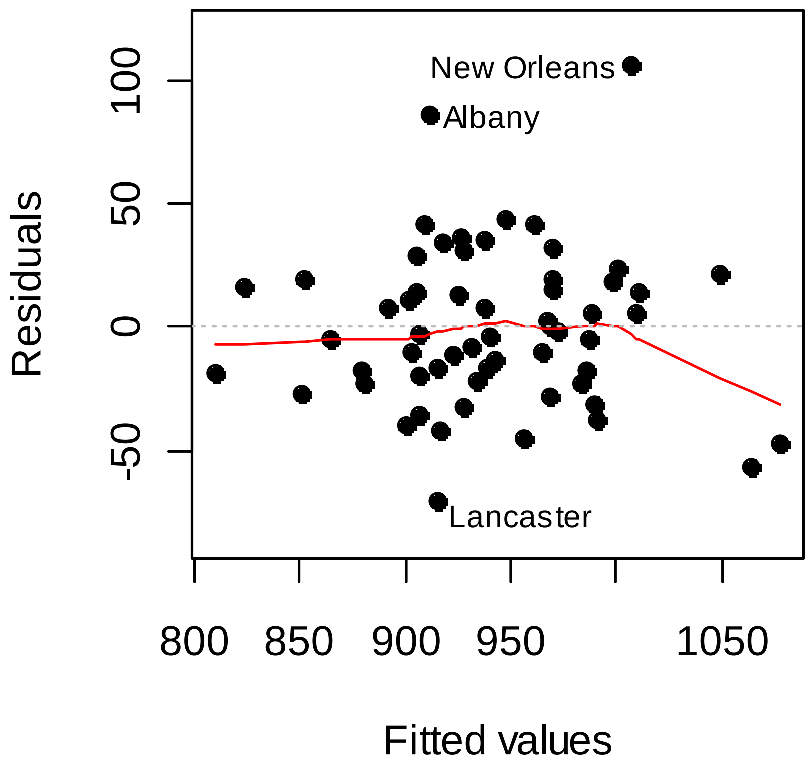
b) improving the model (better fit, reliable conclusions)

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

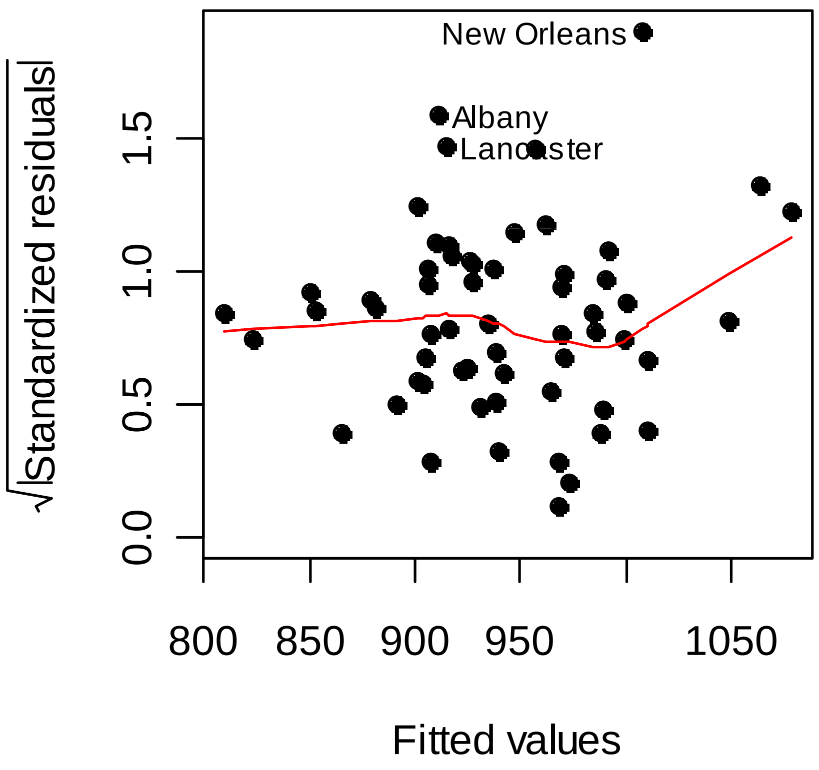
Applied Statistical Regression
HS 2010 – Week 05

Model Diagnostics: Structure and Variance

Tukey-Anscombe Plot



Scale-Location Plot



Model Diagnostics: Example

It is all in the residuals…

Important trickery:

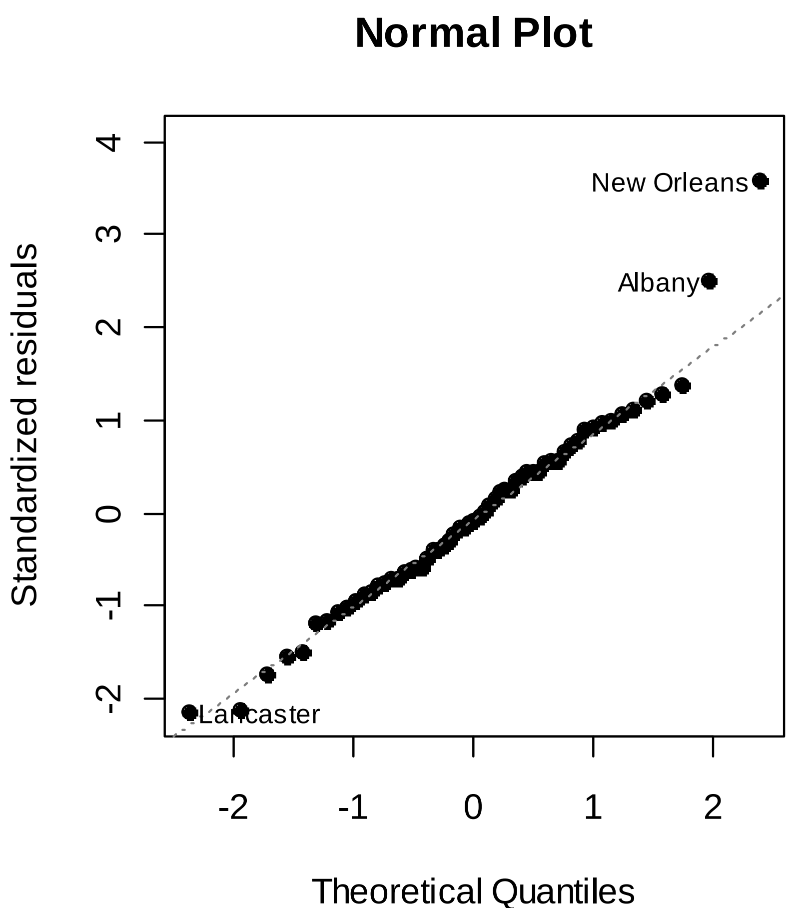
- Plot the residuals against some predictors

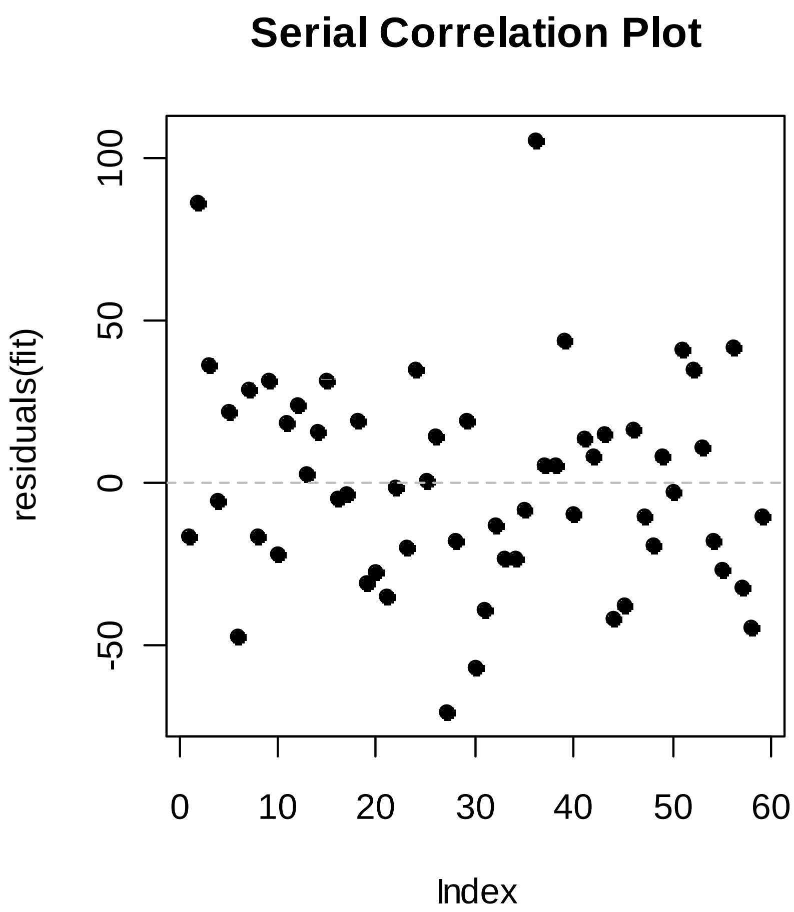
- These predictors can be in or out of the model

→ No matter what the predictor is, we must not see any structures in these plots. If we find some, that means the model is inadequate.

Applied Statistical Regression
HS 2010 – Week 05

Model Diagnostics: Normality/Correlation





Applied Statistical Regression
HS 2010 – Week 05

How To Identify Influential Data Points?

1) Poor man‘s approach

Redo the analysis n times by excluding each data point

2) Leverage

If we change y_i by Δy_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

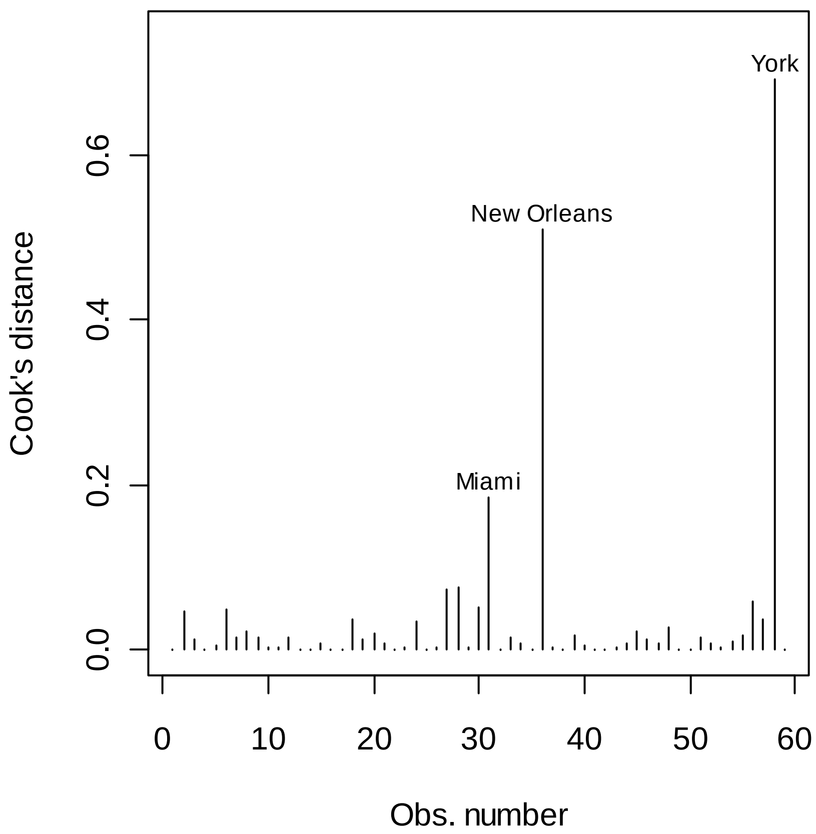
$$D_i = \frac{\sum (\hat{y}_j - y_{j(i)})^2}{(p+1)\sigma_\varepsilon^2} = \frac{h_{ii}}{1-h_{ii}} \cdot \frac{r_i^{*2}}{(p+1)}$$

Be careful if Cook‘s Distance > 1.

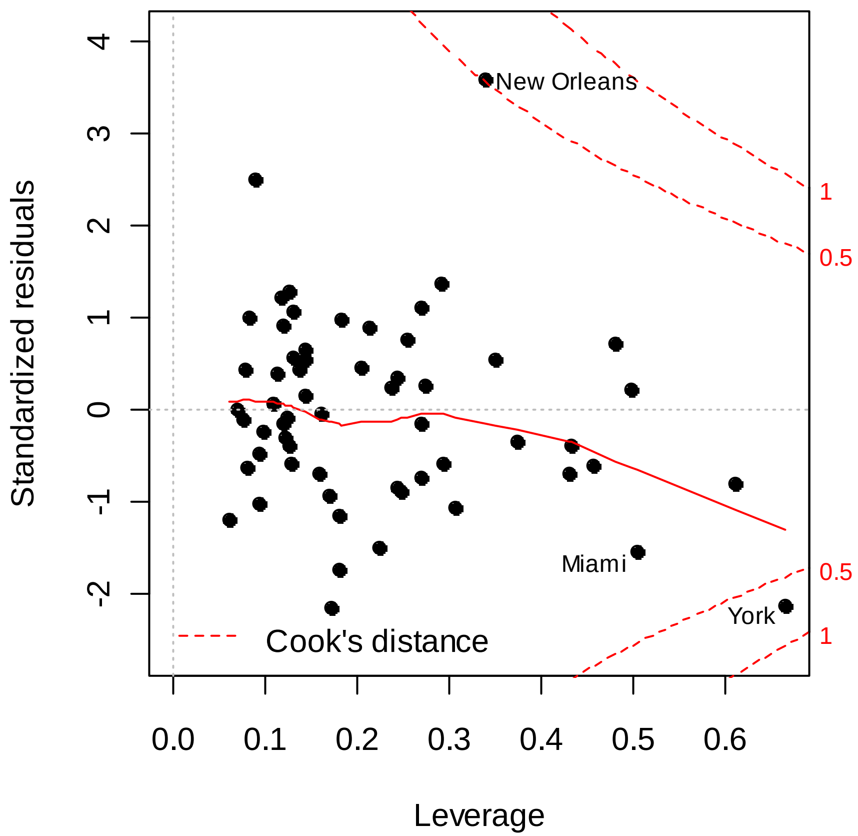
Applied Statistical Regression
HS 2010 – Week 05

Model Diagnostics: Example

Cook's Distance



Leverage Plot



Applied Statistical Regression
HS 2010 – Week 05

Model Diagnostics: Conclusions

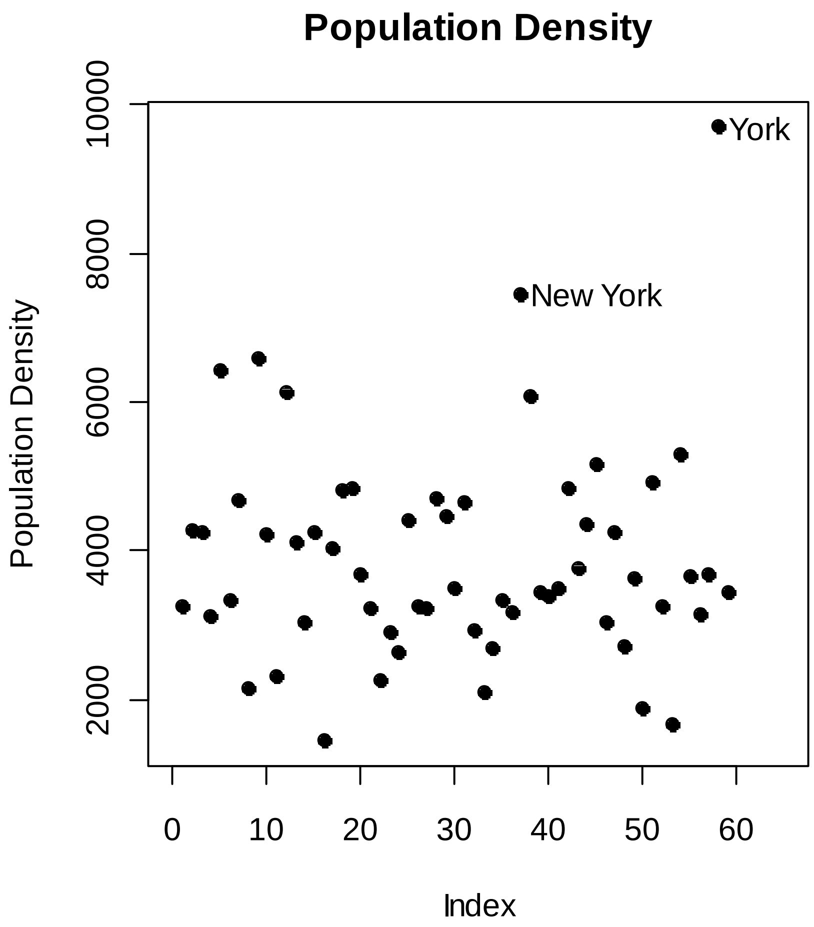
Conclusions from the model diagnostics:

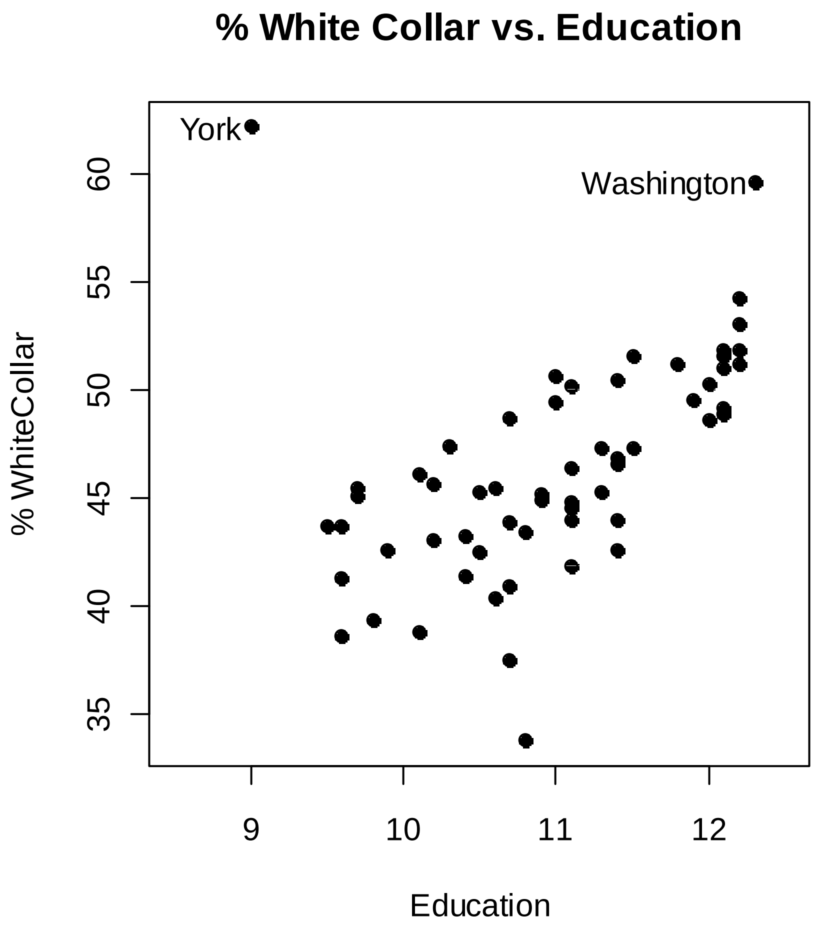
- there are 2 influential data points: York and New Orleans
- they do not seem to be very strongly influential, but still:
- better to re-run the analysis without these and check results

Results from that analysis:

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

Why Are They Influential?





Weighted Regression

We consider the model:

$$Y = X\beta + \varepsilon \text{, where } \varepsilon \sim N(0, \sigma_\varepsilon^2 \Sigma) \text{ with } \Sigma = I$$

→ generalized least squares …

→ weighted regression

$$\Sigma = diag\left(\frac{1}{w_1}, \frac{1}{w_2}, ..., \frac{1}{w_n}\right)$$

Weighted Regression

When, why and how:

- If the Y_i are means of observations, we choose $w_i = n_i$
- If the variance is proportional to a predictor: $w_i = 1/x_i$
- For observed non-constant variance: estimate from OLS!

The regression coefficients in weighted regression are obtained by minimizing the sum of weighted least squares:

$$\sum_{i=1}^{n} w_i r_i^2$$

Solution is still explicit and unique for full-ranked design.

Applied Statistical Regression
HS 2010 – Week 05

Robust Regression

How to deal with outliers?

- Check for typos first. Keep contact to the original data source.
- Examine the physical context – why did it happen? Outliers are often the most interesting data points!
- Exclude the outliers from the analysis, and re-fit the model. Always report the existence of outliers that were removed.
- If outliers are not mistakes but "occur naturally" due to long-tailed error distribution:
  → *do not exclude them from the analysis*
  → *run a robust regression*

Applied Statistical Regression
HS 2010 – Week 05

Robust Regression

```
> library(MASS)
> fit.rlm <- rlm(Mortality ~ JanTemp + … + log(SO2), data=…)
```

→ Relies on Huber's method

→ Downweights the effect of outliers

```
summary(fit.rlm)
Coefficients: Value Std. Error    t value
(Intercept) 945.4414  251.6184     3.7574
JanTemp       -1.2313   0.6788    -1.8139
log(SO2)      13.0484   4.6444     2.8095
---
Residual standard error: 30.17 on 46 degrees of freedom
```

Polynomial Regression

Polynomial Regression = Multiple Linear Regression !!!

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_d x^d + \varepsilon$$

Goals:

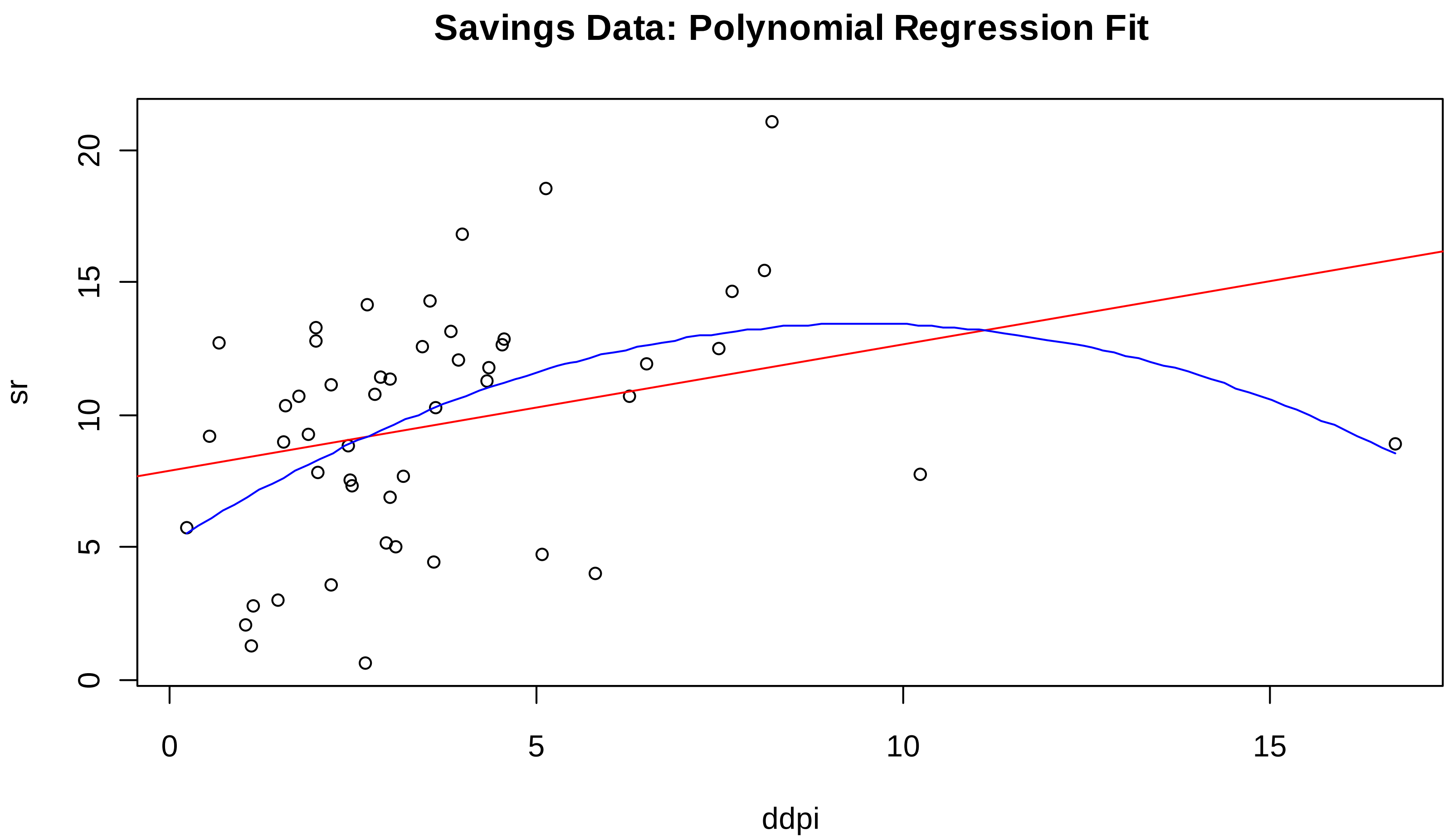
- fit a curvilinear relation
- improve the fit between x and Y
- determine the polynomial order d

Example:

- Savings dataset: personal savings ~ income per capita

Applied Statistical Regression
HS 2010 – Week 05

Polynomial Regression Fit



Applied Statistical Regression
HS 2010 – Week 05

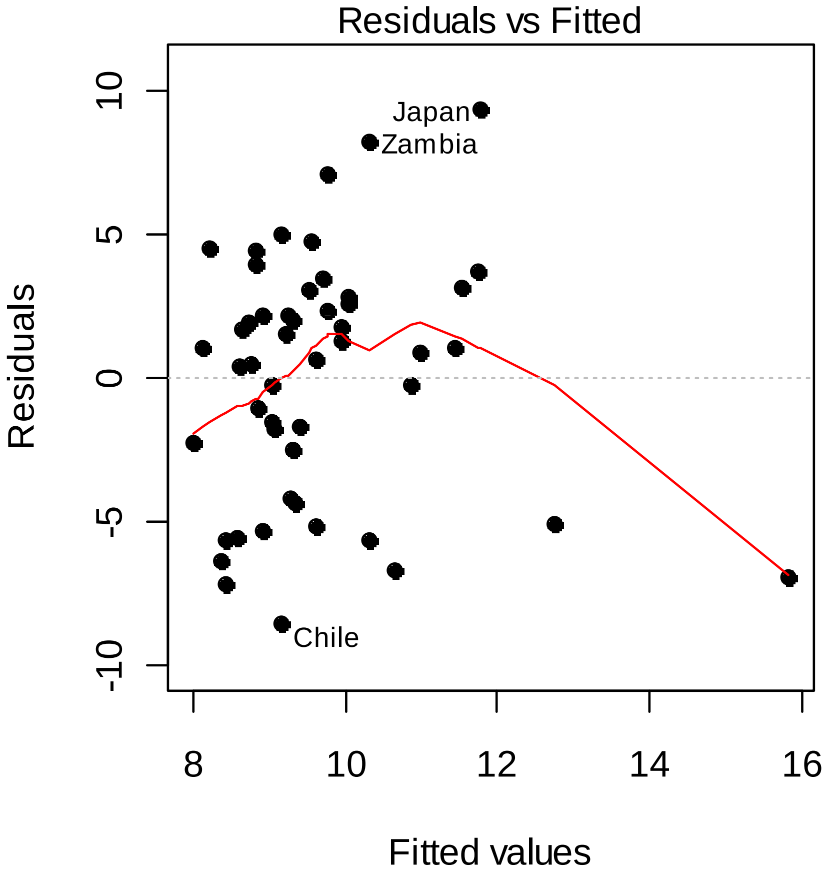
Polynomial Regression

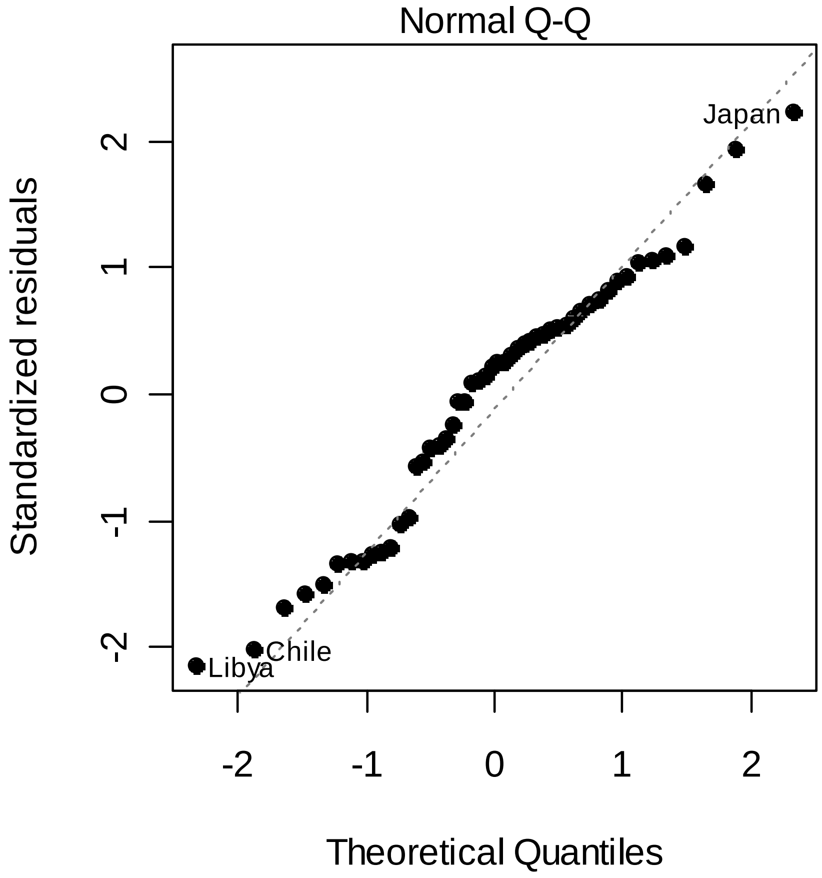
Output from the model with the linear term only:

```
> summary(lm(sr ~ ddpi, data = savings))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.8830     1.0110   7.797 4.46e-10 ***
ddpi          0.4758     0.2146   2.217   0.0314 *
---
Residual standard error: 4.311 on 48 degrees of freedom
Multiple R-squared: 0.0929, Adjusted R-squared: 0.074
F-statistic: 4.916 on 1 and 48 DF,  p-value: 0.03139
```

Applied Statistical Regression
HS 2010 – Week 05

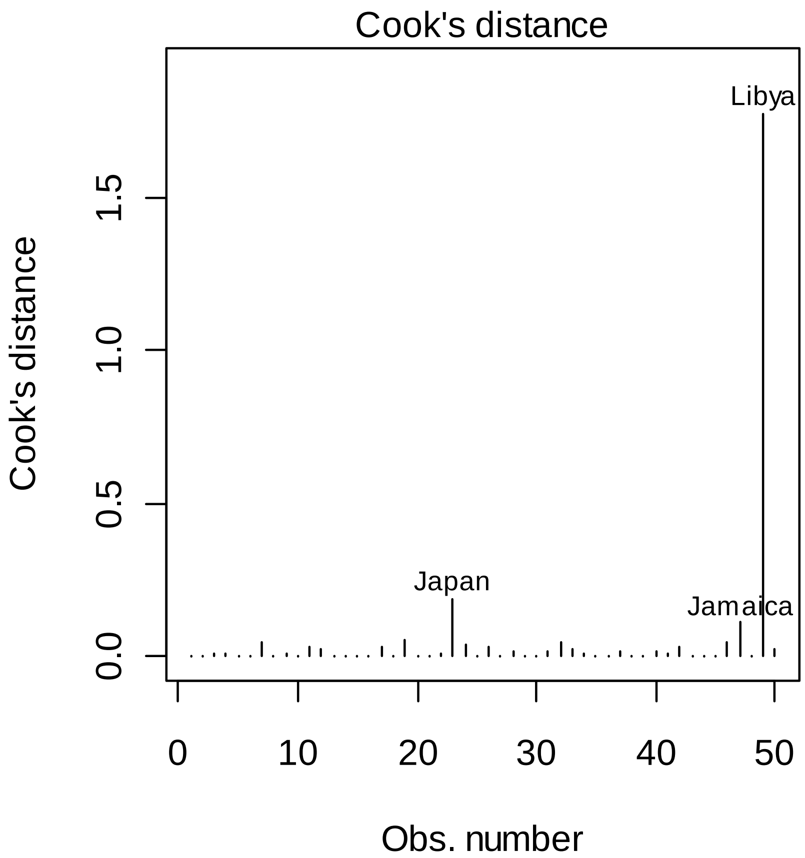
Diagnostic Plots 1

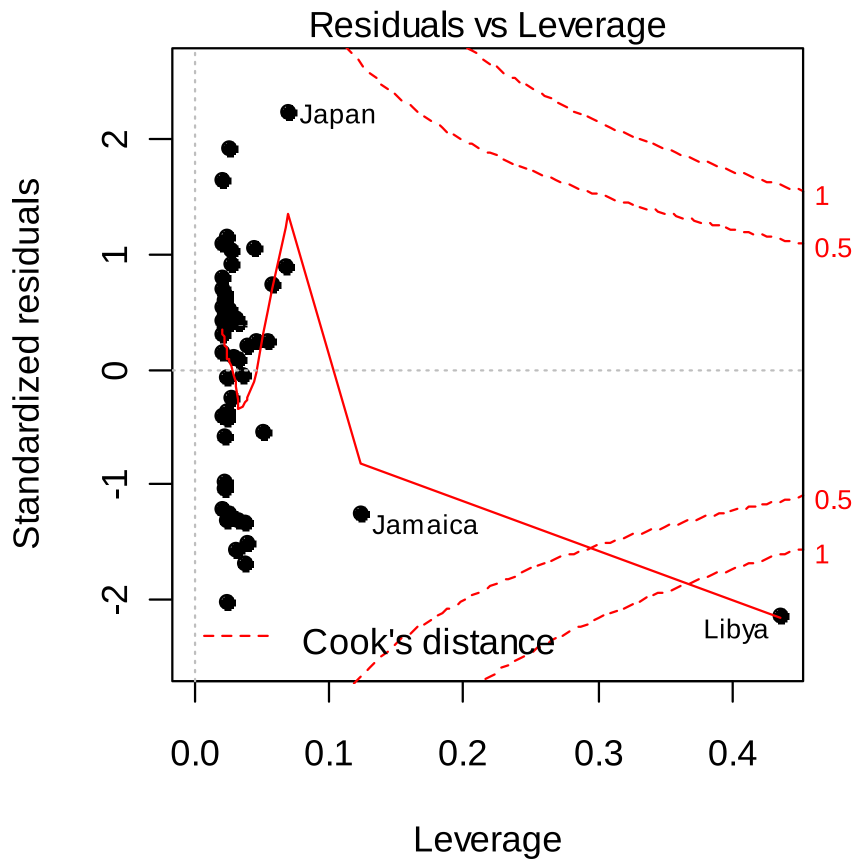




Applied Statistical Regression
HS 2010 – Week 05

Diagnostic Plots 2





Applied Statistical Regression
HS 2010 – Week 05

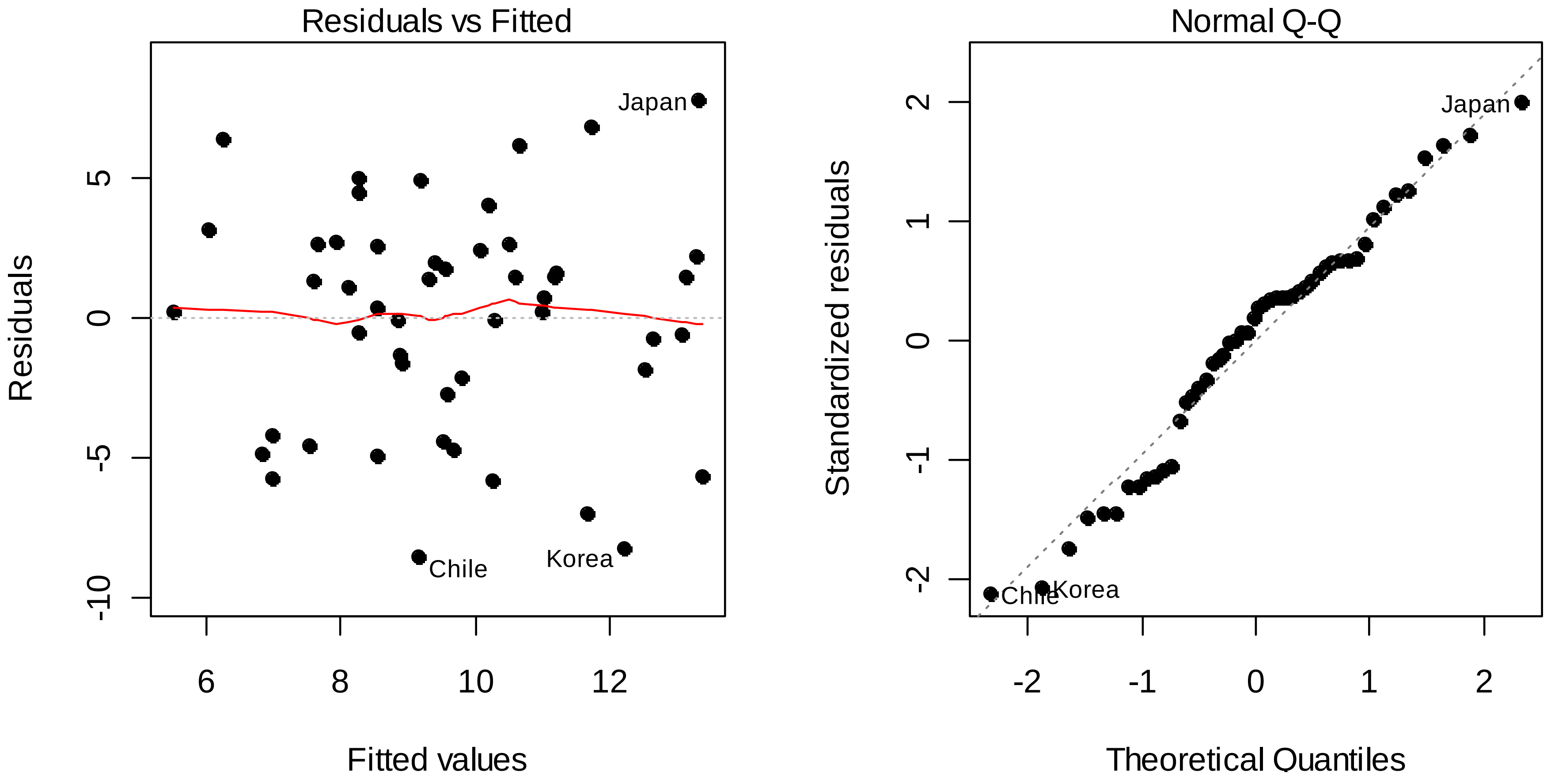
Quadratic Regression

Add the quadratic term: $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

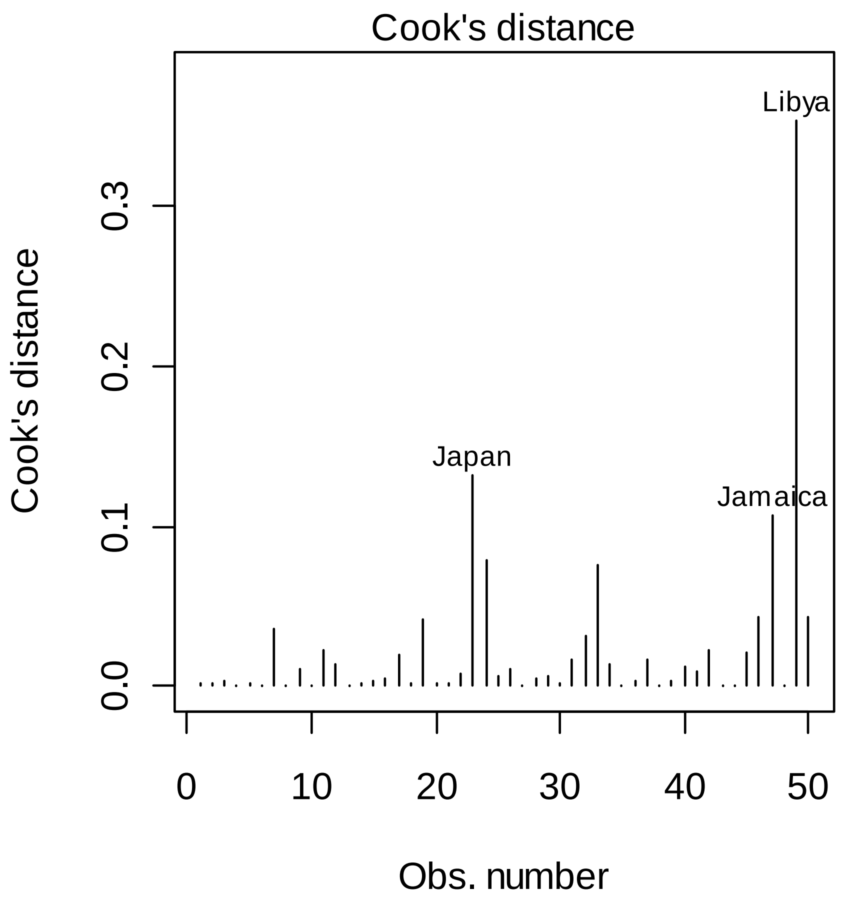
```
> summary(lm(sr ~ ddpi + I(ddpi^2), data = savings))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.13038    1.43472   3.576 0.000821 ***
ddpi         1.75752    0.53772   3.268 0.002026 **
I(ddpi^2)   -0.09299    0.03612  -2.574 0.013262 *
---
Residual standard error: 4.079 on 47 degrees of freedom
Multiple R-squared: 0.205,  Adjusted R-squared: 0.1711
F-statistic: 6.059 on 2 and 47 DF,  p-value: 0.004559
```

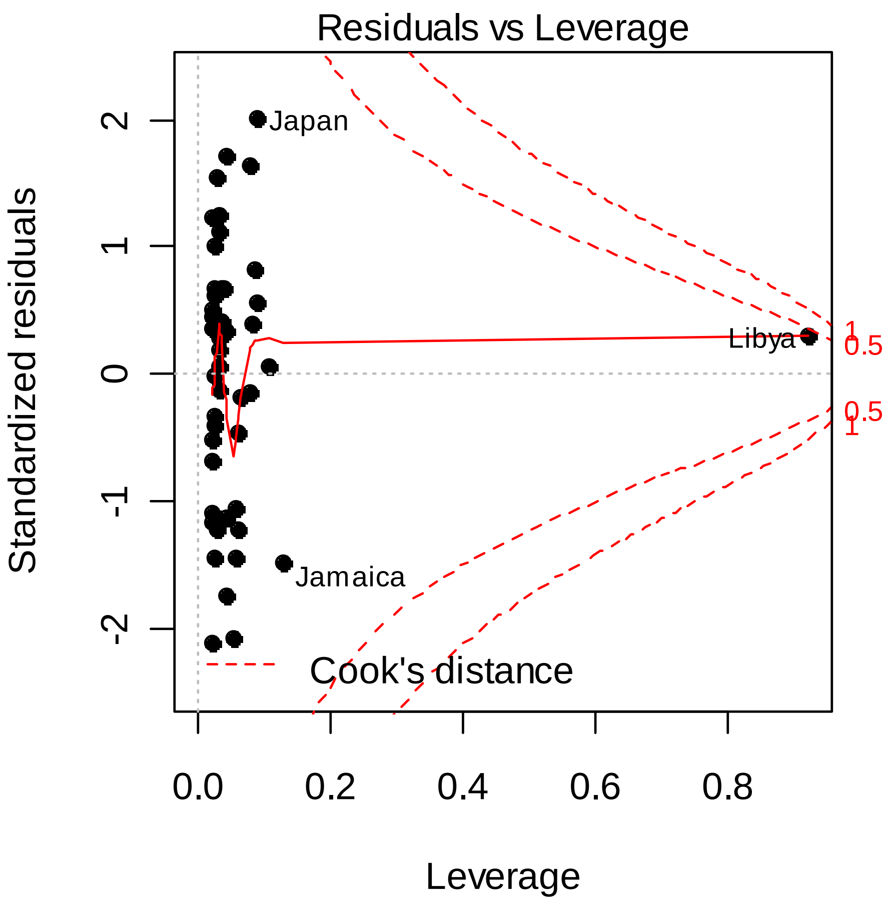
Applied Statistical Regression
HS 2010 – Week 05

Diagnostic Plots: Quadratic Regression



Diagnostic Plots: Quadratic Regression





Applied Statistical Regression
HS 2010 – Week 05

Cubic Regression

Add the cubic term: $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

```
> summary(lm(sr~ddpi + I(ddpi^2) + I(ddpi^3), data = savings)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.145e+00  2.199e+00   2.340   0.0237 *
ddpi         1.746e+00  1.380e+00   1.265   0.2123
I(ddpi^2)   -9.097e-02  2.256e-01  -0.403   0.6886
I(ddpi^3)   -8.497e-05  9.374e-03  -0.009   0.9928
---
Residual standard error: 4.123 on 46 degrees of freedom
Multiple R-squared: 0.205,     Adjusted R-squared: 0.1531
F-statistic: 3.953 on 3 and 46 DF,  p-value: 0.01369
```

Powers Are Strongly Correlated Predictors!

The smaller the x-range, the bigger the problem!

```
> cor(cbind(ddpi, ddpi2=ddpi^2, ddpi3=ddpi^3))
           ddpi     ddpi2     ddpi3
ddpi  1.0000000 0.9259671 0.8174527
ddpi2 0.9259671 1.0000000 0.9715650
ddpi3 0.8174527 0.9715650 1.0000000
```

Way out: use centered predictors!

$$z_i = (x_i - \overline{x})$$

$$z_i^2 = (x_i - \overline{x})^2$$

$$z_i^3 = (x_i - \overline{x})^3$$

Applied Statistical Regression
HS 2010 – Week 05

Powers Are Strongly Correlated Predictors!

```
> summary(lm(sr~z.ddpi+I(z.ddpi^2)+I(z.ddpi^3),dat=z.savings)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.042e+01  8.047e-01  12.946  < 2e-16 ***
z.ddpi       1.059e+00  3.075e-01   3.443  0.00124 **
I(z.ddpi^2) -9.193e-02  1.225e-01  -0.750  0.45691
I(z.ddpi^3) -8.497e-05  9.374e-03  -0.009  0.99281
```

→ Coefficients, standard error and tests are different
→ Fitted values and global inference remain the same
→ Not overly beneficial on this dataset!

→ **Be careful: extrapolation with polynomials is dangerous!**

Dummy Variables

So far, we only considered continuous predictors:

- temperature
- distance
- pressure
- …

It is perfectly valid to have categorical predictors, too:

- sex (male or female)
- status variables (employed or unemployed)
- working shift (day, evening, night)
- …

→ **Implementation in the regression with dummy variables**

Example: Binary Categorical Variable

The lathe dataset:

- Y lifetime of a cutting tool in a lathe
- x_1 speed of the machine in rpm
- x_2 tool type A or B

Dummy variable encoding:

$$x_2 = \begin{cases} 0 & tool \;\; type \;\; A \\ 1 & tool \;\; type \;\; B \end{cases}$$

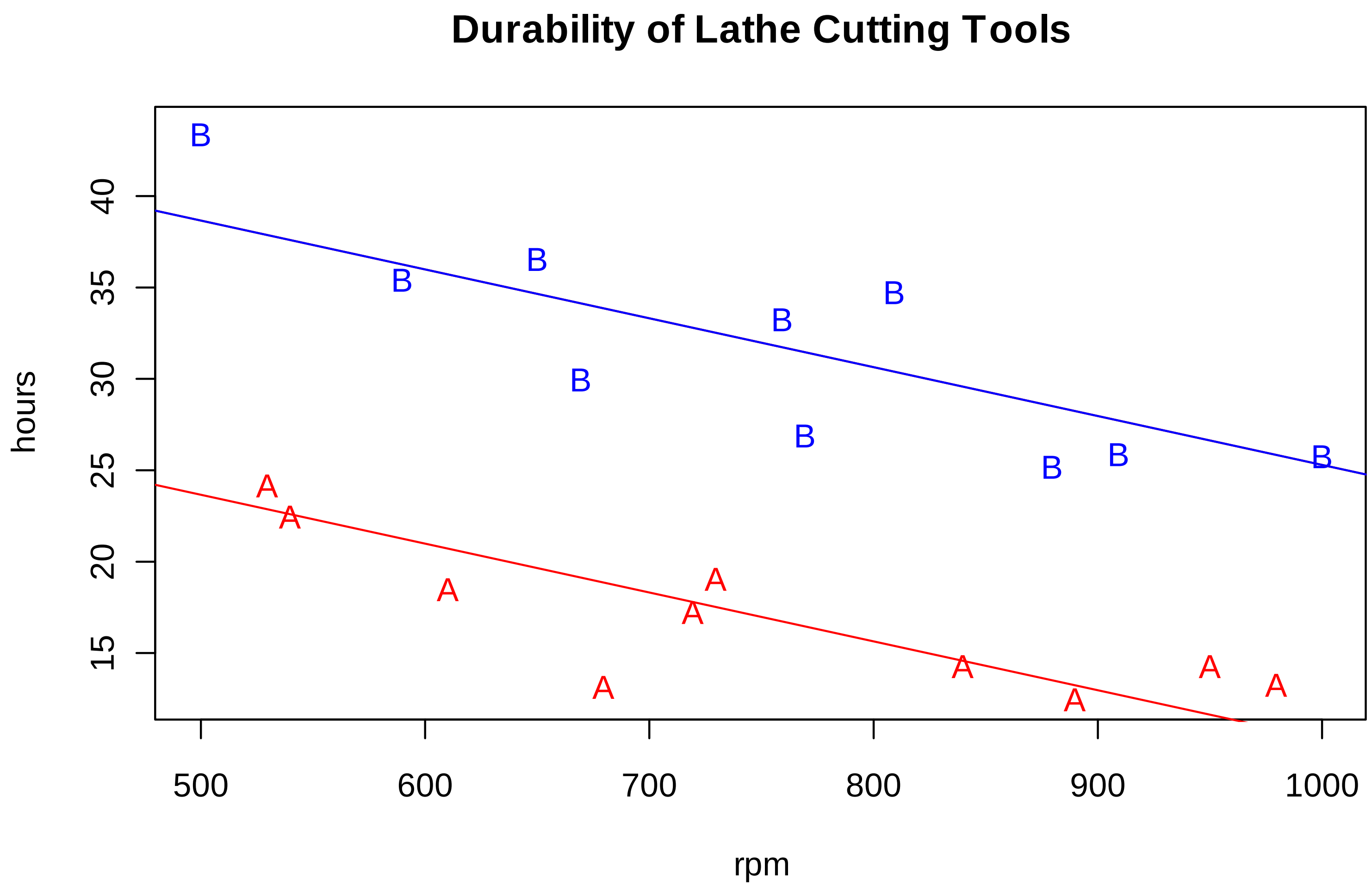
Interpretation of the Model

→ see blackboard…

```
> summary(lm(hours ~ rpm + tool, data = lathe))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.98560    3.51038  10.536 7.16e-09 ***
rpm         -0.02661    0.00452  -5.887 1.79e-05 ***
toolB       15.00425    1.35967  11.035 3.59e-09 ***
---
Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-squared: 0.9003,  Adjusted R-squared: 0.8886
F-statistic: 76.75 on 2 and 17 DF,   p-value: 3.086e-09
```

Applied Statistical Regression
HS 2010 – Week 05

The Dummy Variable Fit



A Model with Interactions

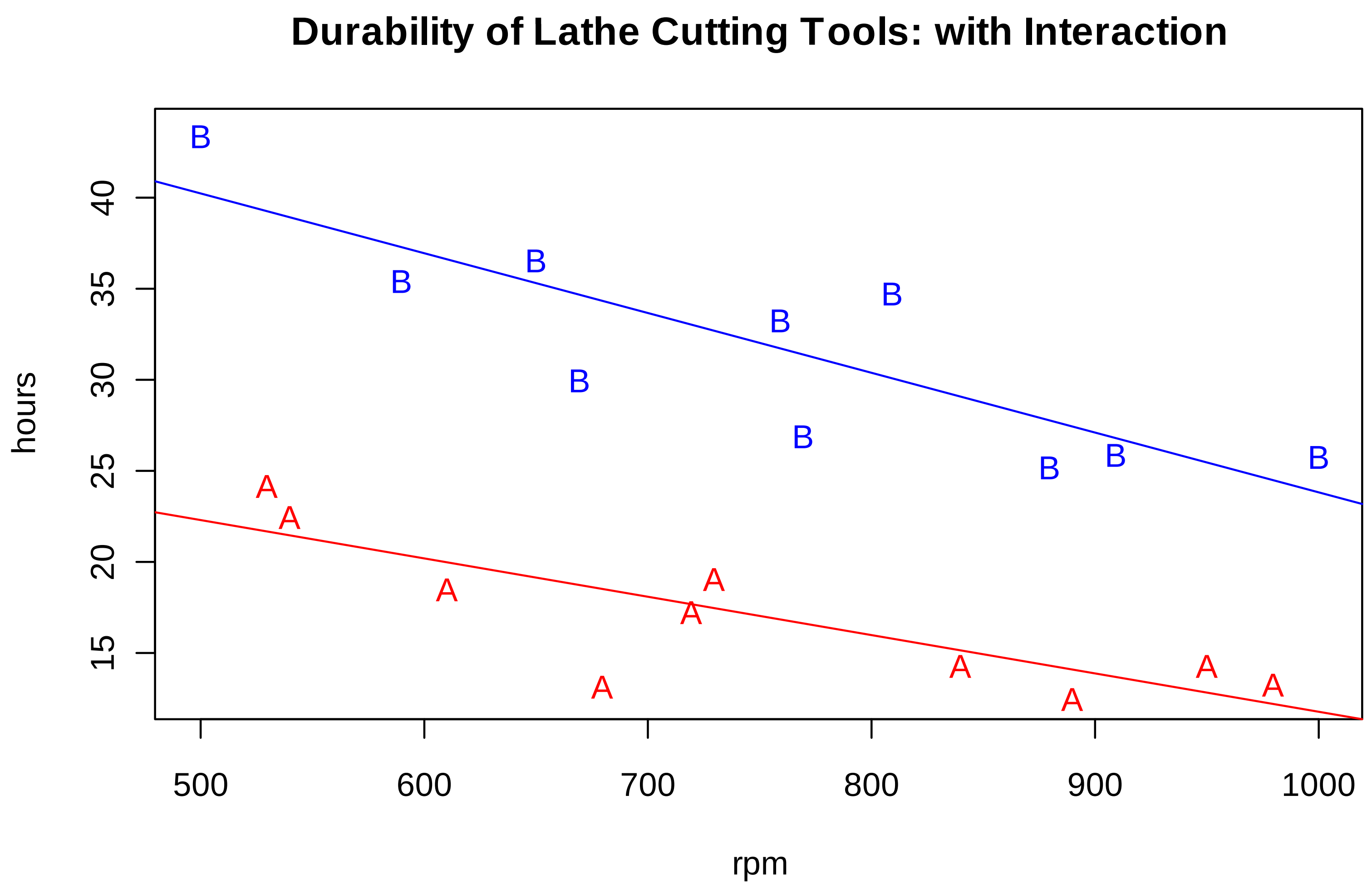
Question: do the slopes need to be identical?

→ with the appropriate model, the answer is no!

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

→ see blackboard for model interpretation…

Different Slope for the Regression Lines



Applied Statistical Regression
HS 2010 – Week 05

Summary Output

```
> summary(lm(hours ~ rpm * tool, data = lathe))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.774760   4.633472   7.073 2.63e-06 ***
rpm         -0.020970   0.006074  -3.452  0.00328 **
toolB       23.970593   6.768973   3.541  0.00272 **
rpm:toolB   -0.011944   0.008842  -1.351  0.19553
---
Residual standard error: 2.968 on 16 degrees of freedom
Multiple R-squared: 0.9105,  Adjusted R-squared: 0.8937
F-statistic: 54.25 on 3 and 16 DF,  p-value: 1.319e-08
```

Applied Statistical Regression
HS 2010 – Week 05

How Complex the Model Needs to Be?

Question 1: do we need different slopes for the two lines?

$$H_0 : \beta_3 = 0 \text{ against } H_A : \beta_3 \neq 0$$

→ individual parameter test for the interaction term!

Question 2: is there any difference altogether?

$$H_0 : \beta_2 = \beta_3 = 0 \text{ against } H_A : \beta_2 \neq 0 \ and\,/\,or \ \beta_3 \neq 0$$

→ this is a partial F-test
→ we try to exclude interaction and dummy variable together

R offers convenient functionality for these tests!

Applied Statistical Regression
HS 2010 – Week 05

Anova Output

Summary output for the interaction model

```
> fit1 <- lm(hours ~ rpm, data=lathe)
> fit2 <- lm(hours ~ rpm * tool, data=lathe)
> anova(fit1, fit2)
Model 1: hours ~ rpm
Model 2: hours ~ rpm * tool
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     18 1282.08
2     16  140.98  2    1141.1 64.755 2.137e-08 ***
```

→ no different slopes, but different intercept!

Categorical Input with More than 2 Levels

There are now 3 tool types A, B, C:

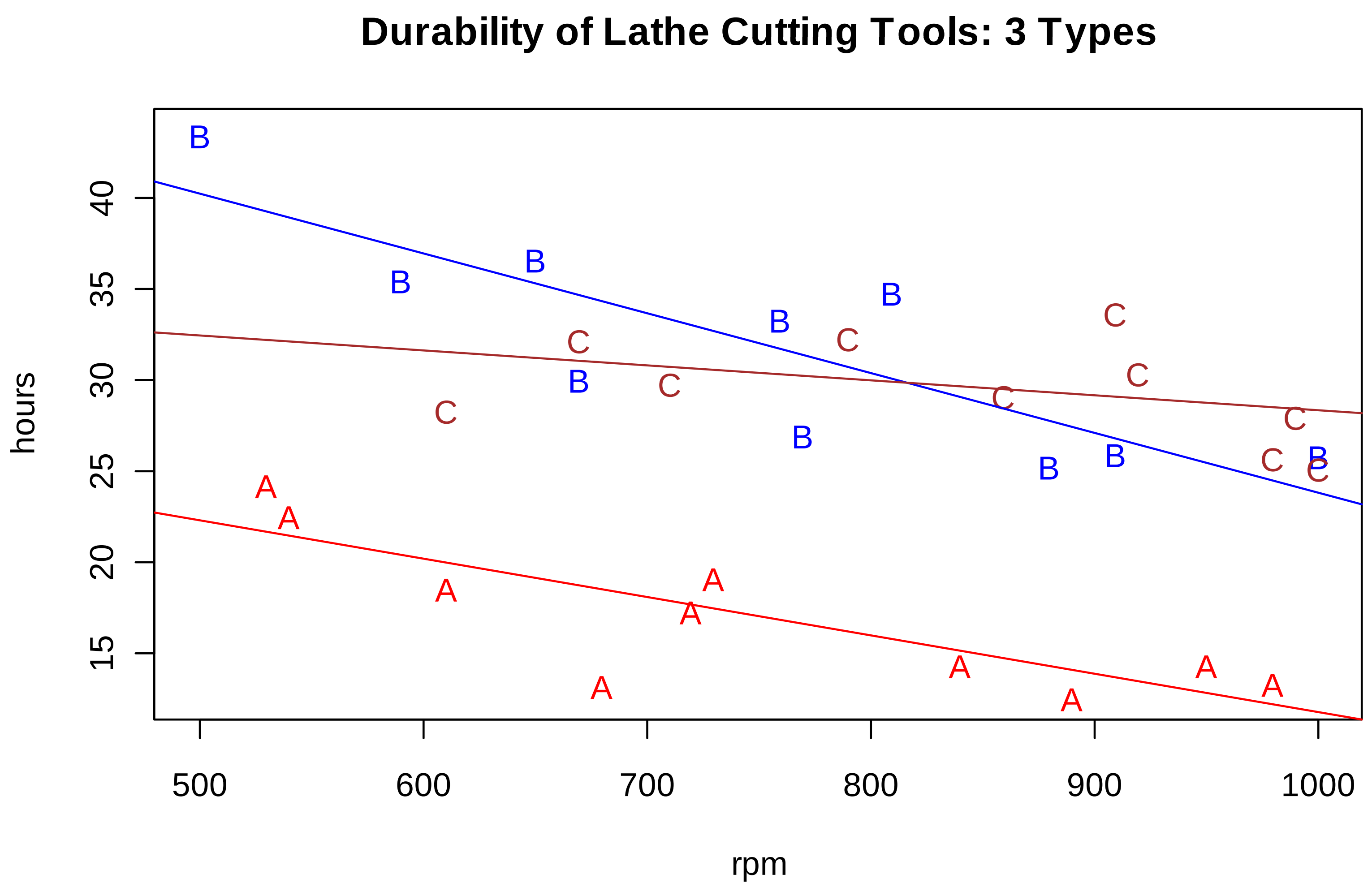
x_2	x_3	
0	0	*for observations of type A*
1	0	*for observations of type B*
0	1	*for observations of type C*

Main effect model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

With interactions: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \varepsilon$

Applied Statistical Regression
HS 2010 – Week 05

Three Types of Cutting Tools



Applied Statistical Regression
HS 2010 – Week 05
Summary Output

```
> summary(lm(hours ~ rpm * tool, data = abc.lathe)

Coefficients:Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.774760   4.496024   7.290 1.57e-07 ***
rpm         -0.020970   0.005894  -3.558  0.00160 **
toolB       23.970593   6.568177   3.650  0.00127 **
toolC        3.803941   7.334477   0.519  0.60876
rpm:toolB   -0.011944   0.008579  -1.392  0.17664
rpm:toolC    0.012751   0.008984   1.419  0.16869
---
Residual standard error: 2.88 on 24 degrees of freedom
Multiple R-squared: 0.8906,    Adjusted R-squared: 0.8678
F-statistic: 39.08 on 5 and 24 DF,  p-value: 9.064e-11
```

Inference with Categorical Predictors

Do not perform individual hypothesis tests on factors!

Question 1: do we have different slopes?

$$H_0: \beta_4 = 0 \ \ and \ \ \beta_5 = 0 \ \text{ against } \ H_A: \beta_4 \neq 0 \ \ and/or \ \ \beta_5 \neq 0$$

Question 2: is there any difference altogether?

$$H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \ \text{ against } \ H_A: any \ of \ \ \beta_2, \beta_3, \beta_4, \beta_5 \neq 0$$

→ Again, R provides convenient functionality

Anova Output

```
> anova(fit.abc)

Analysis of Variance Table
          Df  Sum Sq Mean Sq F value    Pr(>F)
rpm        1  139.08  139.08 16.7641  0.000415 ***
tool       2 1422.47  711.23 85.7321 1.174e-11 ***
rpm:tool   2   59.69   29.84  3.5974  0.043009 *
Residuals 24  199.10    8.30
```

→ strong evidence that we need to distinguish the tools!

→ weak evidence for the necessity of different slopes