Applied Statistical Regression
AS 2013 – Week 07

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 28, 2013

Applied Statistical Regression
AS 2013 – Week 07

Versatility of Multiple Linear Regression

Despite that we are using linear models only, we have a versatile and powerful tool. While the response is always a continuous variable, different predictor types are allowed:

- **Continuous Predictors**
  Default case, e.g. *temperature*, *distance*, *pH-value*, …
- **Transformed Predictors**
  For example: $log(x)$, $sqrt(x)$, $arcsin(\sqrt{x})$,...
- **Powers**
  We can also use: x^{-1}, x^2, x^3, ...
- **Categorical Predictors**
  Often used: *sex*, *day of week*, *political party*, …

Applied Statistical Regression
AS 2013 – Week 07

Categorical Predictors

The canonical case in linear regression are *continuous predictor variables* such as for example:

→ *temperature, distance, pressure, velocity, ...*

While in linear regression, we cannot have categorical response, it is perfectly valid to have *categorical predictors*:

→ *yes/no, sex (m/f), type (a/b/c), shift (day/evening/night), ...*

Such categorical predictors are often also called **factor variables**. In a linear regression, each level of such a variable is encoded by a dummy variable, so that $(\ell - 1)$ degrees of freedom are spent.

Regression with a Factor Variable

The lathe (*in German: Drehbank*) **dataset:**

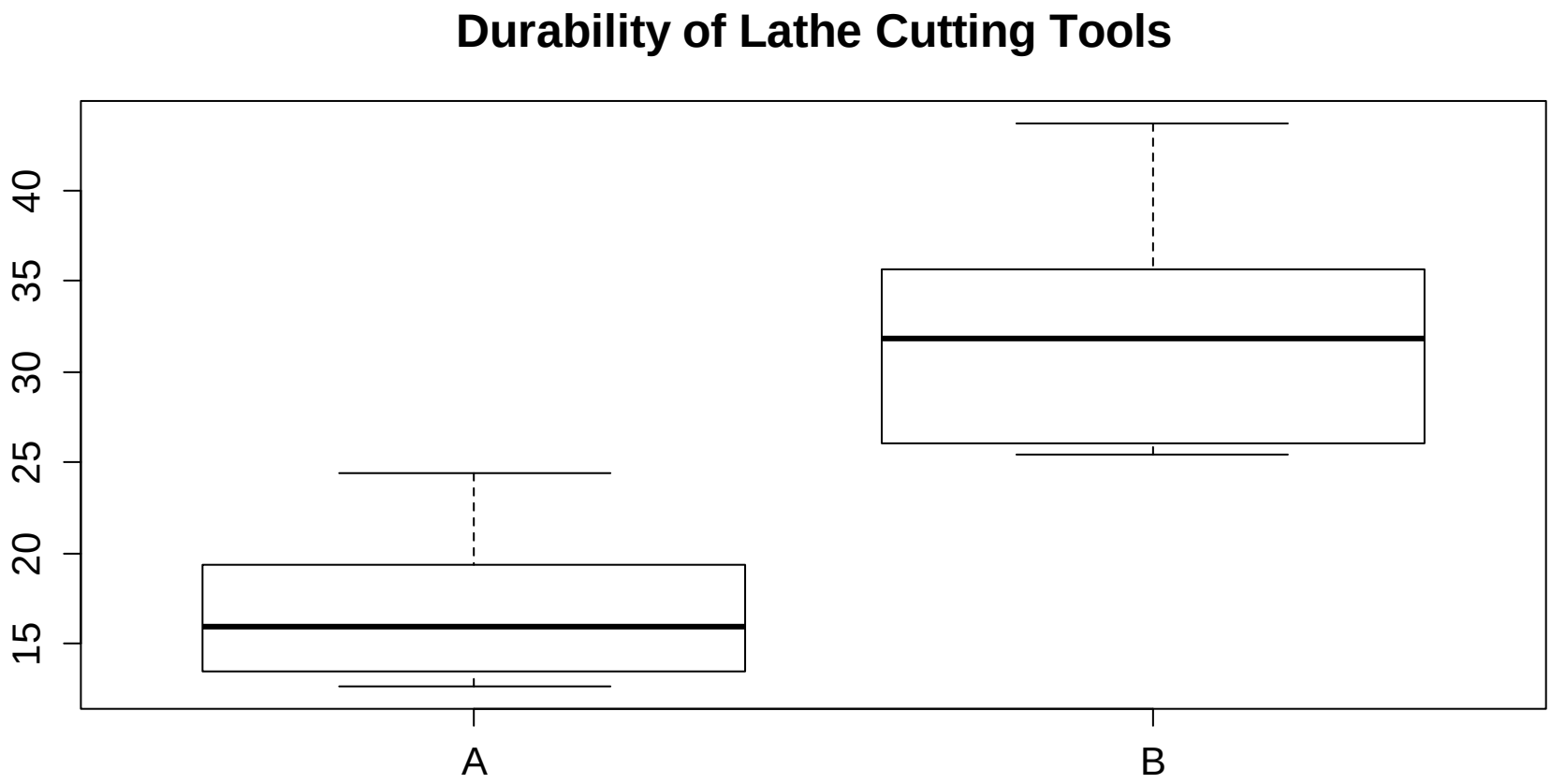
- y_i life time of cutting tool i
- x_i type of tool i, A or B

Dummy variable encoding:

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

Applied Statistical Regression
AS 2013 – Week 07

Typical Visualization of a Factor Model



Applied Statistical Regression
AS 2013 – Week 07

Interpretation of the Factor Model

→ **See blackboard…**

```
> summary(fit)

Call: lm(formula = hours ~ tool, data = lathe)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   17.110      1.628  10.508 4.14e-09 ***
toolB         14.818      2.303   6.435 4.68e-06 ***
---
Residual standard error: 5.149 on 18 degrees of freedom
Multiple R-squared: 0.697,  Adjusted R-squared: 0.6802
F-statistic: 41.41 on 1 and 18 DF,  p-value: 4.681e-06
```

Applied Statistical Regression
AS 2013 – Week 07

Another View: t-Test

→ **The 1-factor-model is a t-test for non-paired data!**

```
> t.test(hours ~ tool, data=lathe, var.equal=TRUE)

Two Sample t-test

data:  hours by tool
t = -6.435, df = 18, p-value = 4.681e-06
alternative hypothesis: true diff in means is not 0
95 percent confidence interval:
 -19.655814  -9.980186
sample estimates:
mean in group A mean in group B
         17.110          31.928
```

Example: Binary Categorical Variable

The lathe (*in German: Drehbank*) **dataset:**

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

Dummy variable encoding:

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

Applied Statistical Regression
AS 2013 – Week 07

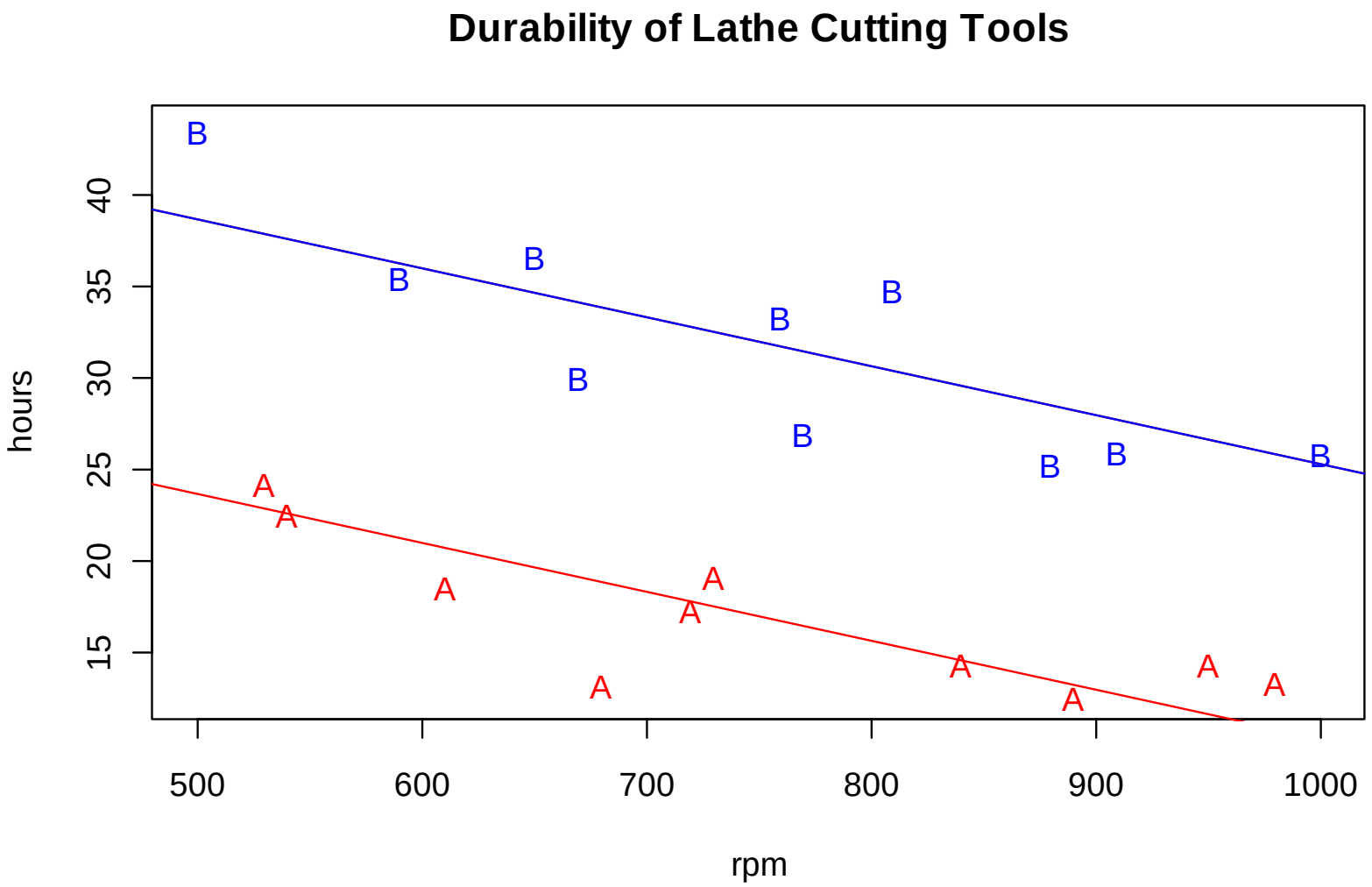
Interpretation of the Model

→ **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
AS 2013 – Week 07

The Dummy Variable Fit



Applied Statistical Regression
AS 2013 – Week 07

A Model with Interactions

Question: do the slopes need to be identical?

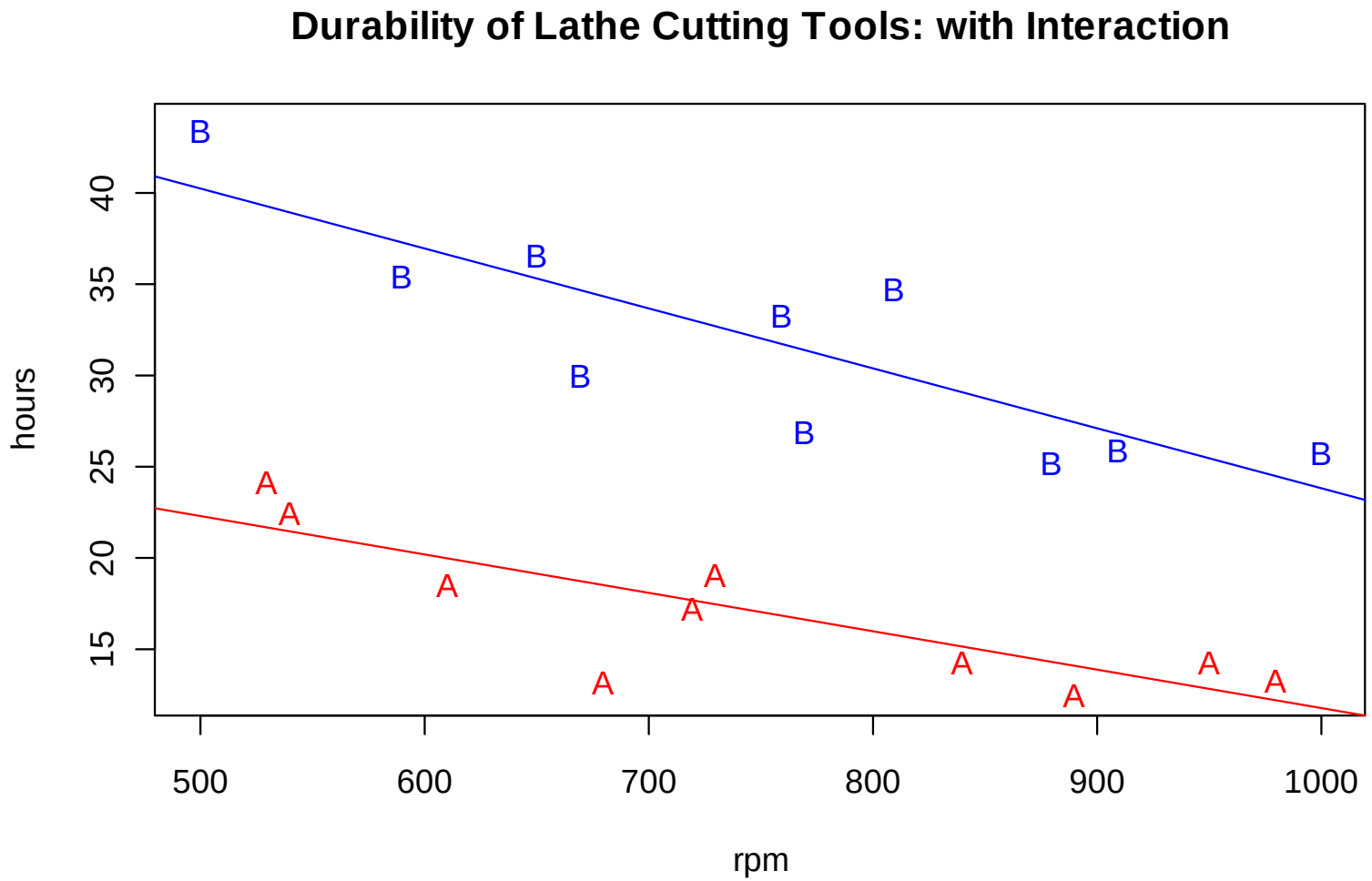
→ with the appropriate model, the answer is no!

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

→ **see blackboard for model interpretation…**

Applied Statistical Regression
AS 2013 – Week 07

Different Slopes for the Regression Lines



Applied Statistical Regression
AS 2013 – Week 07

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
AS 2013 – Week 07

How Complex the Model Needs to Be?

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

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

→ no, see individual test for the interaction term on previous slide!

Question 2: is there any difference altogether?

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

→ this is a hierarchical model comparison

→ we try to exclude interaction and dummy variable together

R offers convenient functionality for this test, see next slide!

Applied Statistical Regression
AS 2013 – Week 07

Testing the Tool Type Variable

Hierarchical model comparison with anova( ):

```
> fit.small <- lm(hours ~ rpm, data=lathe)
> fit.big <- lm(hours ~ rpm * tool, data=lathe)
> anova(fit.small, fit.big)
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 ***
```

→ The bigger model, i.e. making a distinction between the tools, is significantly better. The main effect is enough, though.

Applied Statistical Regression
AS 2013 – Week 07

Categorical Input with More Than 2 Levels

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

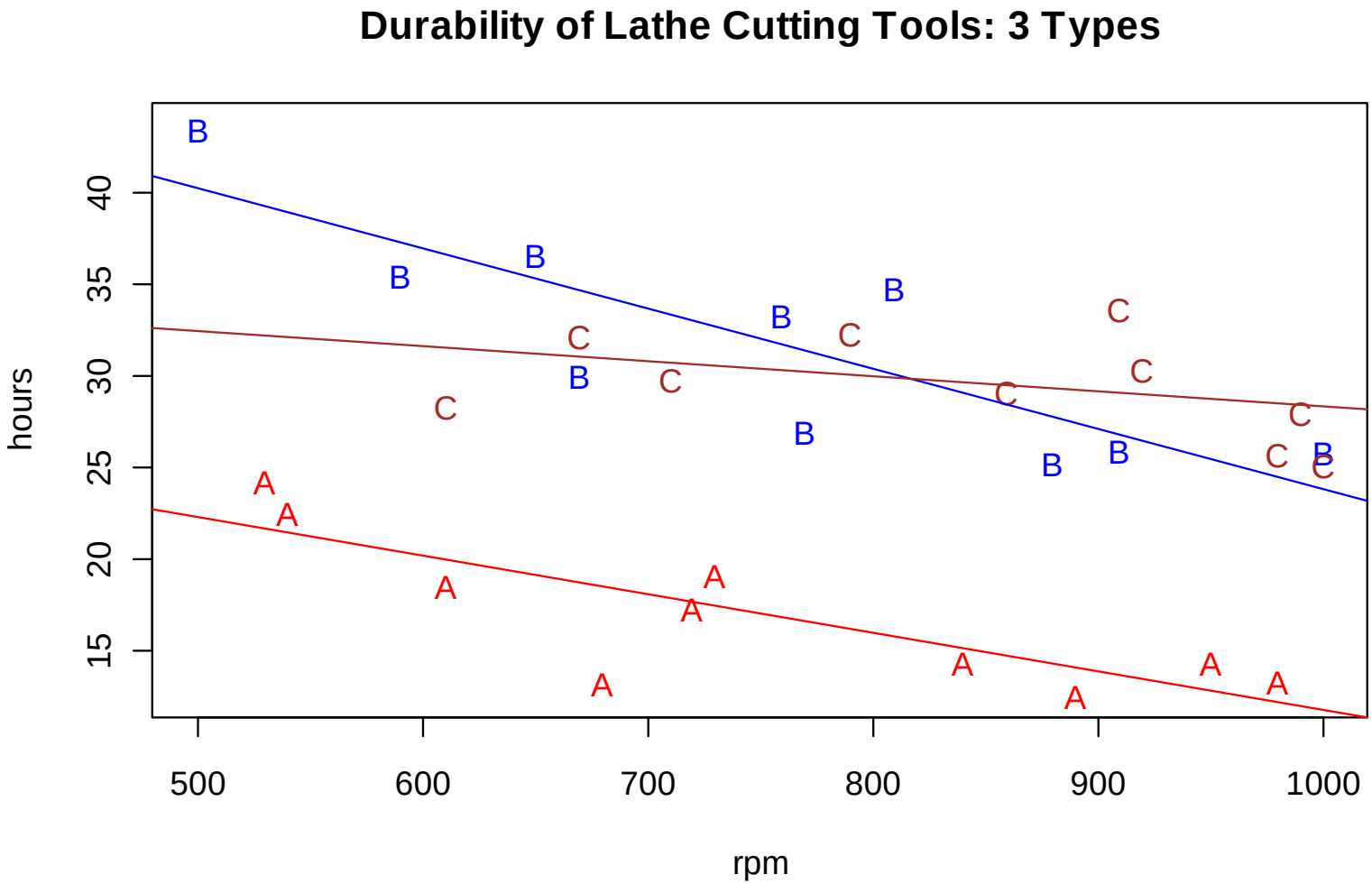
x_2	x_3	
0	0	*for observations of type A*
1	0	*for observations of type B*
0	1	*for observations of type C*

Main effect model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + E$

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

Applied Statistical Regression
AS 2013 – Week 07

Three Types of Cutting Tools



Applied Statistical Regression
AS 2013 – Week 07
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
```

This summary is of limited use for deciding about model complexity. We require hierarchical model comparisons!

Applied Statistical Regression
AS 2013 – Week 07

Inference with Categorical Predictors

Do not perform individual hypothesis tests on factors that have more than 2 levels, they are meaningless!

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( )`

Applied Statistical Regression
AS 2013 – Week 07

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
```

→ The interaction term is weakly significant. Thus, there is some weak evidence for the necessity of different slopes.

→ The p-value for the tool variable includes omitting interaction and main effect. Being strongly significant, we have strong evidence that tool type distinction is needed.