Topic 12

Analysis of covariance (ANCOVA)

Sergey Mastitsky ©
Klaipeda, 28-30 September 2011

ANCOVA

- A regression analysis that includes both numeric and nominal (factor) predictors is called analysis of covariance (ANCOVA)
- Although the same function, `lm()`, is used to fit the model, some specific details need to be examined

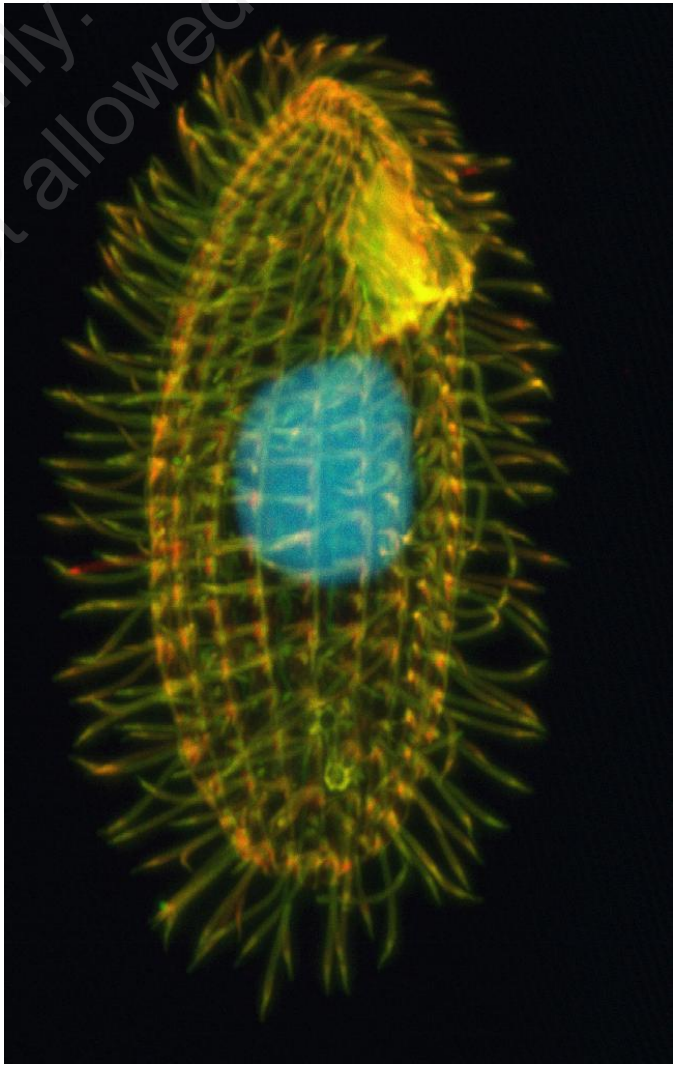
12. ANCOVA

12.1. Graphical presentation of ANCOVA-type data

An example: *Tetrahymena* size

```
> library(ISwR)
> data(hellung)
> head(hellung)
> help("hellung")
```

- Two groups of cell cultures: with glucose (`1`) and without glucose (`2`) in the growth medium
- Cell concentration (`conc`) and diameter (`diameter`) were measured at both conditions
- Does the `diameter~conc` relationship depend on glucose?



Author: Sergey Mastitsky

Summary of the `hellung` data

```
> summary(hellung)
```

```
    glucose             conc             diameter    
Min.   :1.000   Min.   : 11000   Min.   :19.20  
1st Qu.:1.000   1st Qu.: 27500   1st Qu.:21.40  
Median :1.000   Median : 69000   Median :23.30  
Mean   :1.373   Mean   :164325   Mean   :23.00  
3rd Qu.:2.000   3rd Qu.:243000   3rd Qu.:24.35  
Max.   :2.000   Max.   :631000   Max.   :26.30  
```

Recognized by R as a numeric variable – not good

Strongly right-skewed

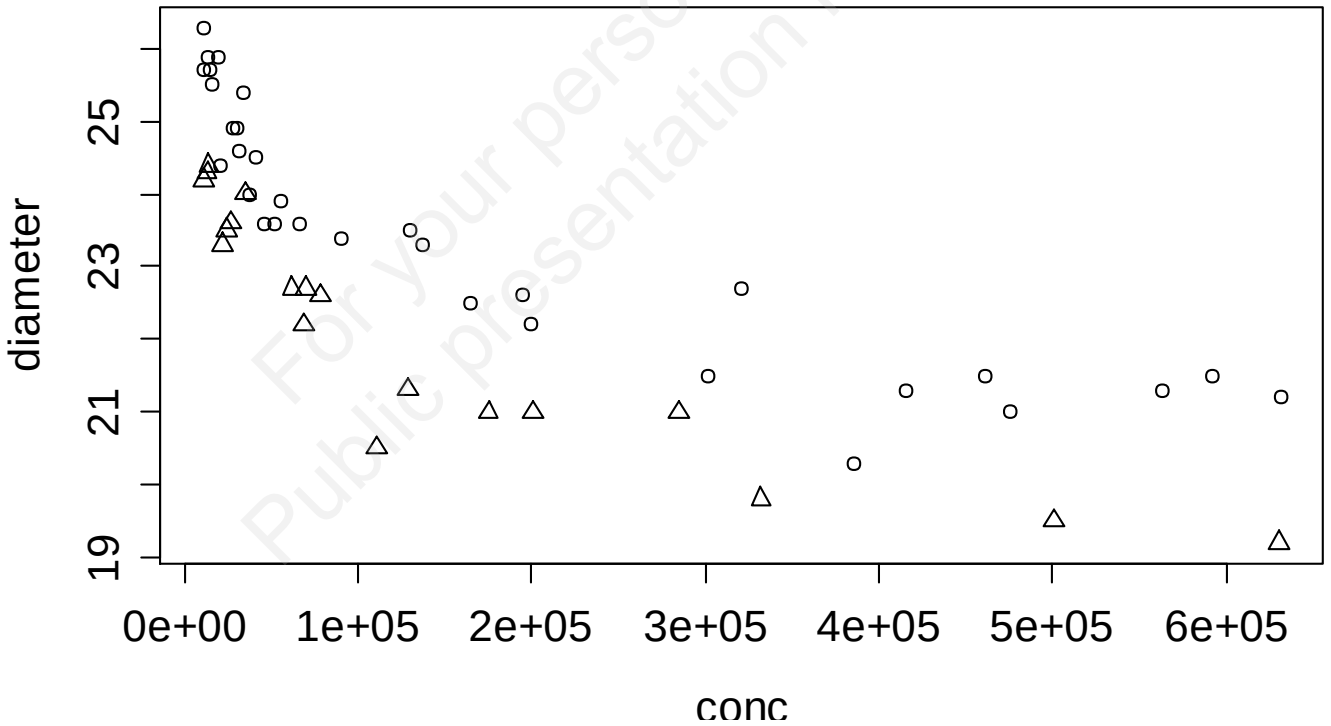
Converting glucose to factor (for convenience)

```
> hellung$glucose <-
  factor(hellung$glucose, labels =
  c("Yes", "No"))
> summary(hellung)
```

```
 glucose        conc              diameter
 Yes:32   Min.   : 11000   Min.   :19.20
 No :19   1st Qu.: 27500   1st Qu.:21.40
          Median : 69000   Median :23.30
          Mean   :164325   Mean   :23.00
          3rd Qu.:243000   3rd Qu.:24.35
          Max.   :631000   Max.   :26.30
```

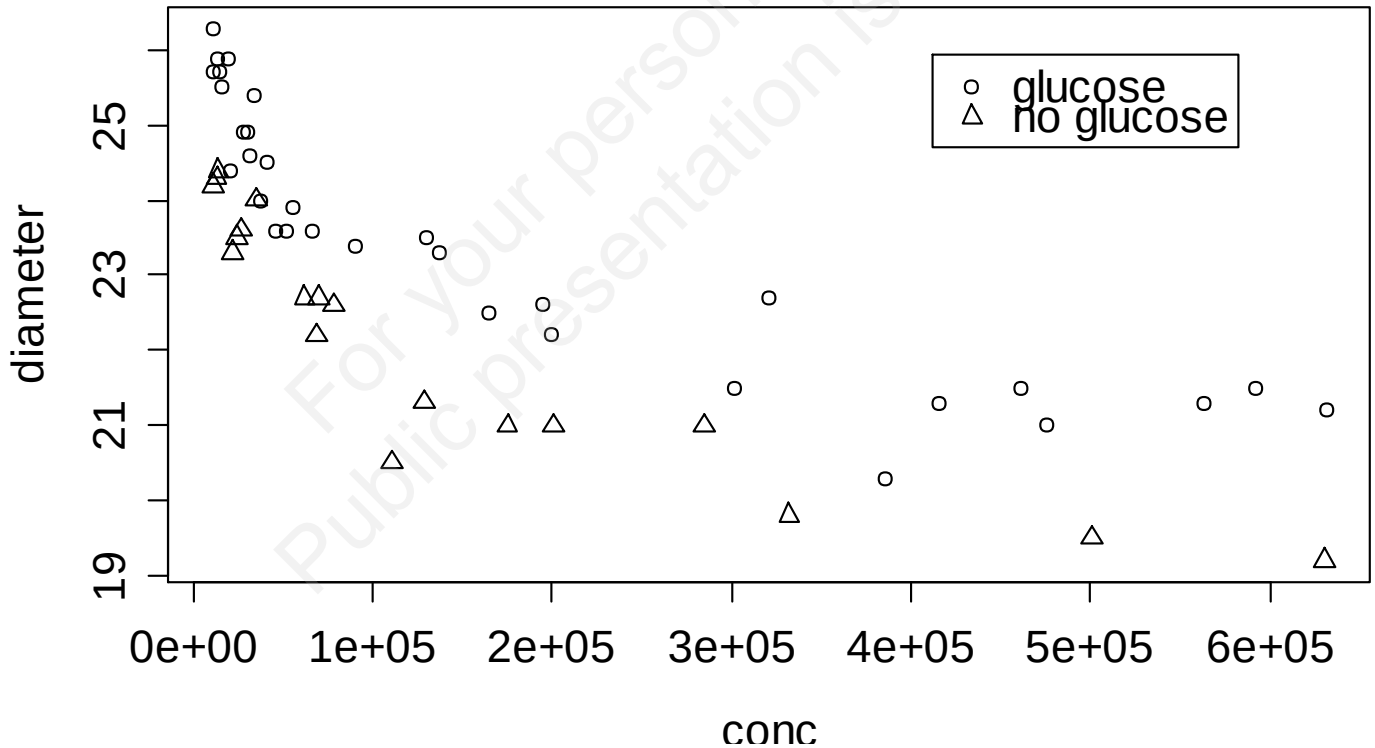
Plotting the `hellung` data

```
> attach(hellung)
> plot(conc, diameter,
  pch = as.numeric(glucose))
```



Inserting a legend

```
> legend(locator(), legend =
  c("glucose","no glucose"),
    pch = 1:2)
```

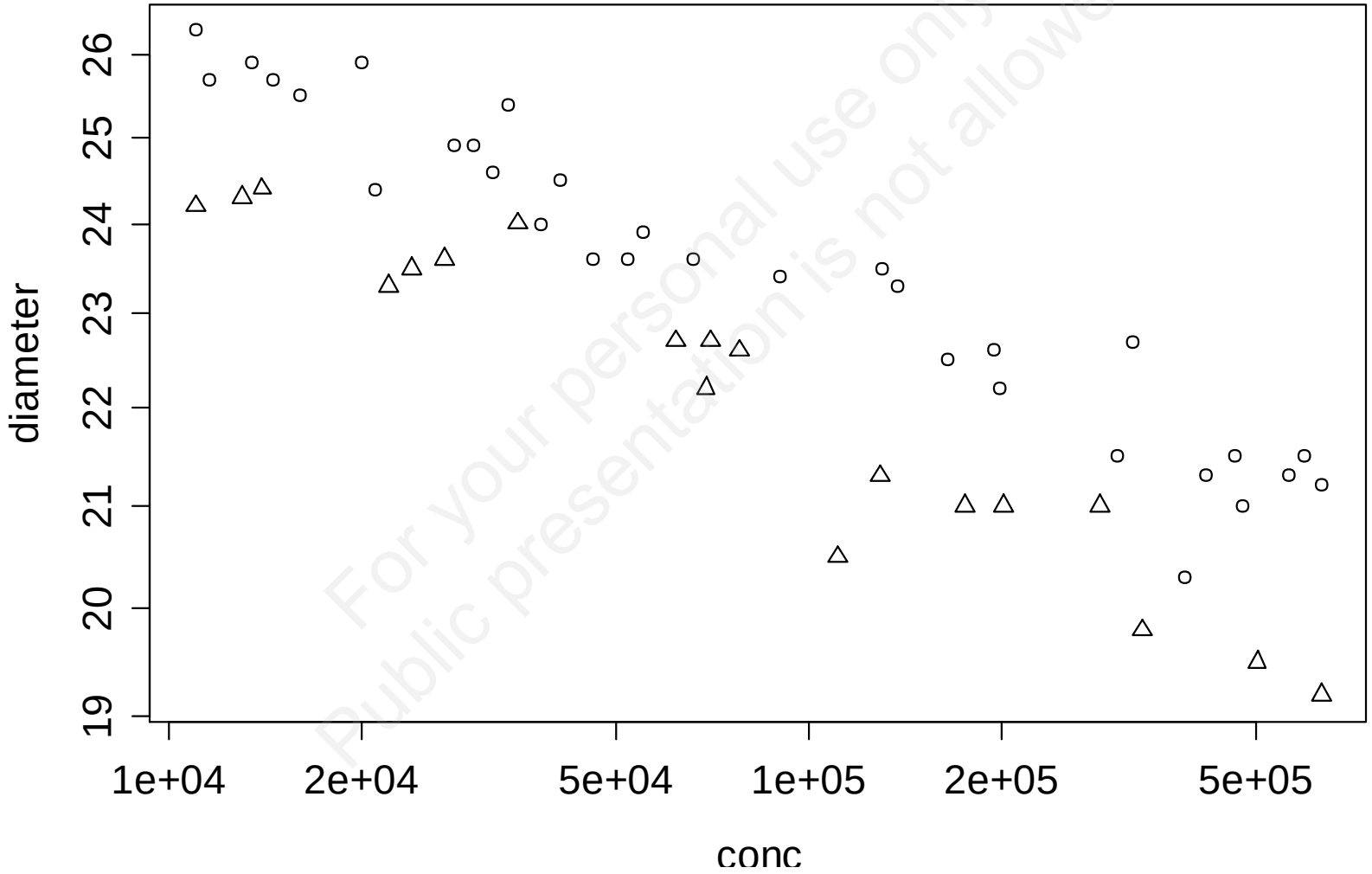


Transforming the data

- As there is a clear negative exponential-like relationship, it makes sense to log-transform the data (also to normalize):

```
> plot(conc, diameter,
    pch = as.numeric(glucose),
        log = "xy")
```

A plot of the log-transformed data

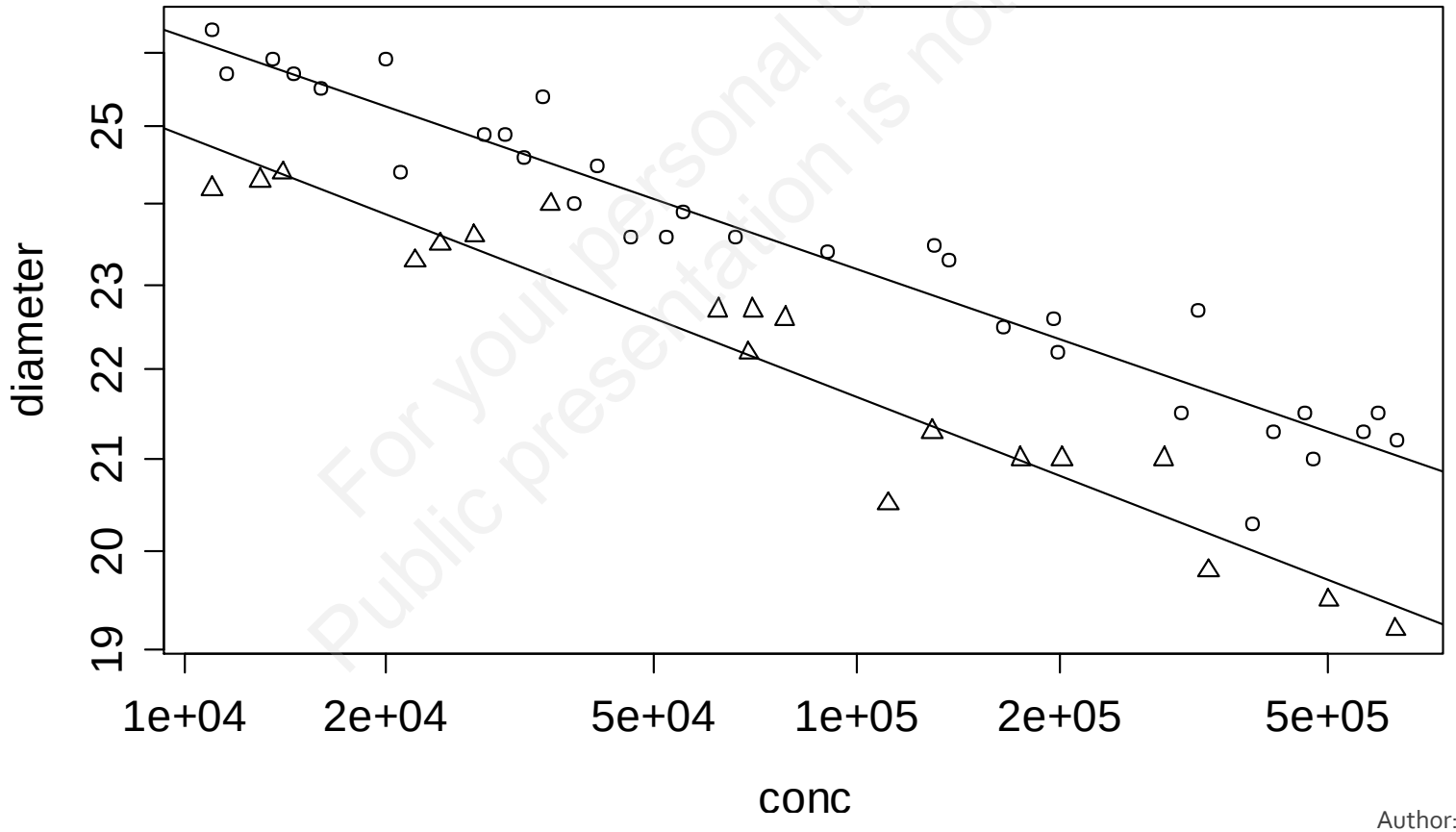


Plotting regression lines for each subset of the data

```
> tethym.gluc <-
    hellung[glucose == "Yes", ]
> tethym.nogluc <-
    hellung[glucose == "No", ]
> lm.nogluc <- lm(log10(diameter)~
  log10(conc), data = tethym.nogluc)
> lm.gluc <- lm(log10(diameter)~
  log10(conc), data = tethym.gluc)
```

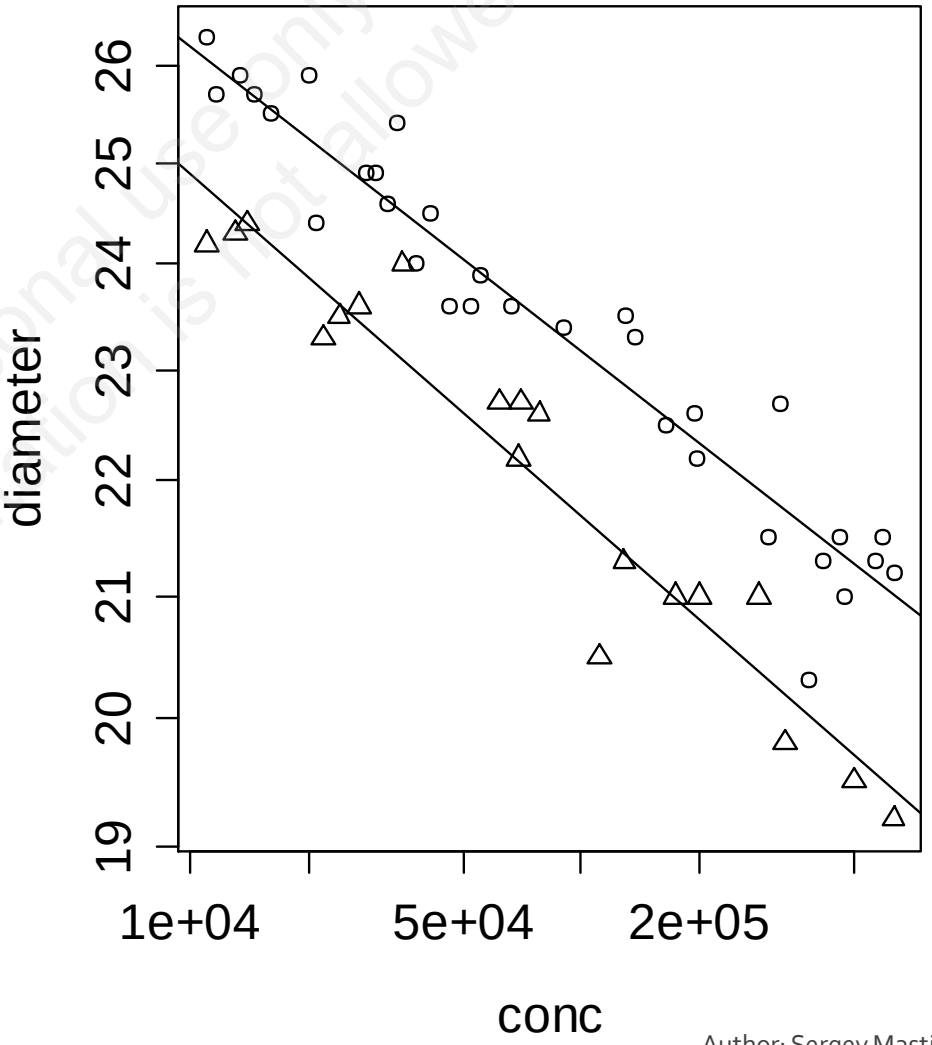
Plotting regression lines for each subset of the data

```
> abline(lm.nogluc)
> abline(lm.gluc)
```



The statistical question of ANCOVA

- The lines seems to be parallel, but not perfectly
- Is the difference in slopes statistically significant?



12. ANCOVA

12.2. Implementation of ANCOVA in R

Specifying ANCOVA in R

```
> AN1  <- lm(log10(diameter) ~
  log10(conc)*glucose)
> summary(AN1)
```

```
Call:
lm(formula = log10(diameter) ~ log10(conc) * glucose)

Residuals:
      Min        1Q    Median        3Q       Max
-0.026722 -0.004888  0.000056  0.003767  0.017608

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.631344   0.013879 117.543   <2e-16 ***
log10(conc)          -0.053196   0.002807 -18.954   <2e-16 ***
glucoseNo             0.003418   0.023695   0.144    0.886
log10(conc):glucoseNo -0.006480   0.004821  -1.344    0.185
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009059 on 47 degrees of freedom
Multiple R-squared: 0.9361, Adjusted R-squared: 0.9321
F-statistic: 229.6 on 3 and 47 DF,  p-value: < 2.2e-16
```


Author: Sergey Mastitsky

Interpretation of regression coefficients in AN1

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.631344	0.013879	117.543	<2e-16	***
log10(conc)	-0.053196	0.002807	-18.954	<2e-16	***
glucoseNo	0.003418	0.023695	0.144	0.886	
log10(conc):glucoseNo	-0.006480	0.004821	-1.344	0.185	

At cell concentration *C*, the expected value of the log cell diameter is the sum of:

- Intercept, 1.6313
- $-0.0532 \times \log_{10} C$
- 0.0034, but only for cultures w/o glucose
- $-0.0065 \times \log_{10} C$, but only for cultures w/o glucose

Interpretation of regression coefficients in AN1

	Estimate
(Intercept)	1.631344
log10(conc)	-0.053196
glucoseNo	0.003418
log10(conc):glucoseNo	-0.006480

Intercept and slope for cultures with glucose

Differences between the two groups in intercept and slope

Author: Sergey Mastitsky

Interpretation of regression coefficients in AN1

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.631344	0.013879	117.543	<2e-16	***
log10(conc)	-0.053196	0.002807	-18.954	<2e-16	***
glucoseNo	0.003418	0.023695	0.144	0.886	
log10(conc):glucoseNo	-0.006480	0.004821	-1.344	0.185	

- Thus, for cell cultures with glucose

$$\log_{10}D = 1.6313 - 0.0532 \times \log_{10}C$$

- For cell cultures without glucose

$$\log_{10}D = (1.6313+0.0034) - (0.0532+0.0064) \times \log_{10}C$$

Interpretation of regression coefficients in AN1

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.631344	0.013879	117.543	<2e-16	***
log10(conc)	-0.053196	0.002807	-18.954	<2e-16	***
glucoseNo	0.003418	0.023695	0.144	0.886	
log10(conc):glucoseNo	-0.006480	0.004821	-1.344	0.185	

Regression coefficient for cell cultures without glucose doesn't differ significantly from that in cultures with glucose, suggesting parallel lines

Author: Sergey Mastitsky

Fitting two parallel lines for the hellung data

```
> AN2  <- lm(log10(diameter) ~
  log10(conc) + glucose)

> summary(AN2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.642132   0.011417  143.83  < 2e-16 ***
log10(conc) -0.055393   0.002301  -24.07  < 2e-16 ***
glucoseNo   -0.028238   0.002647  -10.67 2.93e-14 ***
```

Interpretation of regression coefficients in AN2

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.642132	0.011417	143.83	< 2e-16	***
log10(conc)	-0.055393	0.002301	-24.07	< 2e-16	***
glucoseNo	-0.028238	0.002647	-10.67	2.93e-14	***

- For cell cultures with glucose

$$\log_{10}D = 1.6421 - 0.0554 \times \log_{10}C$$

- For cell cultures without glucose

$$\log_{10}D = (1.6421\text{-}0.0282) - 0.0554 \times \log_{10}C,$$

i.e. on the original scale, the cells in cultures w/o glucose are 6.3% smaller ($10^{-0.0282} = 0.937$)

Testing the variance assumption

- ANCOVA implies statistically equal variances in groups
- This assumption can be tested as follows:

```
> var.test(lm.gluc, lm.nogluc)

 F test to compare two variances

data:  lm.gluc and lm.nogluc
F = 0.8482, num df = 30, denom df = 17, p-value = 0.6731
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3389901 1.9129940
sample estimates:
ratio of variances
         0.8481674
```

The ANOVA table for AN2

```
> anova(AN2)

Analysis of Variance Table

Response: log10(diameter)
             Df   Sum Sq  Mean Sq F value    Pr(>F)
log10(conc)   1 0.046890 0.046890  561.99 < 2.2e-16 ***
glucose       1 0.009494 0.009494  113.78 2.932e-14 ***
Residuals    48 0.004005 0.000083
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```