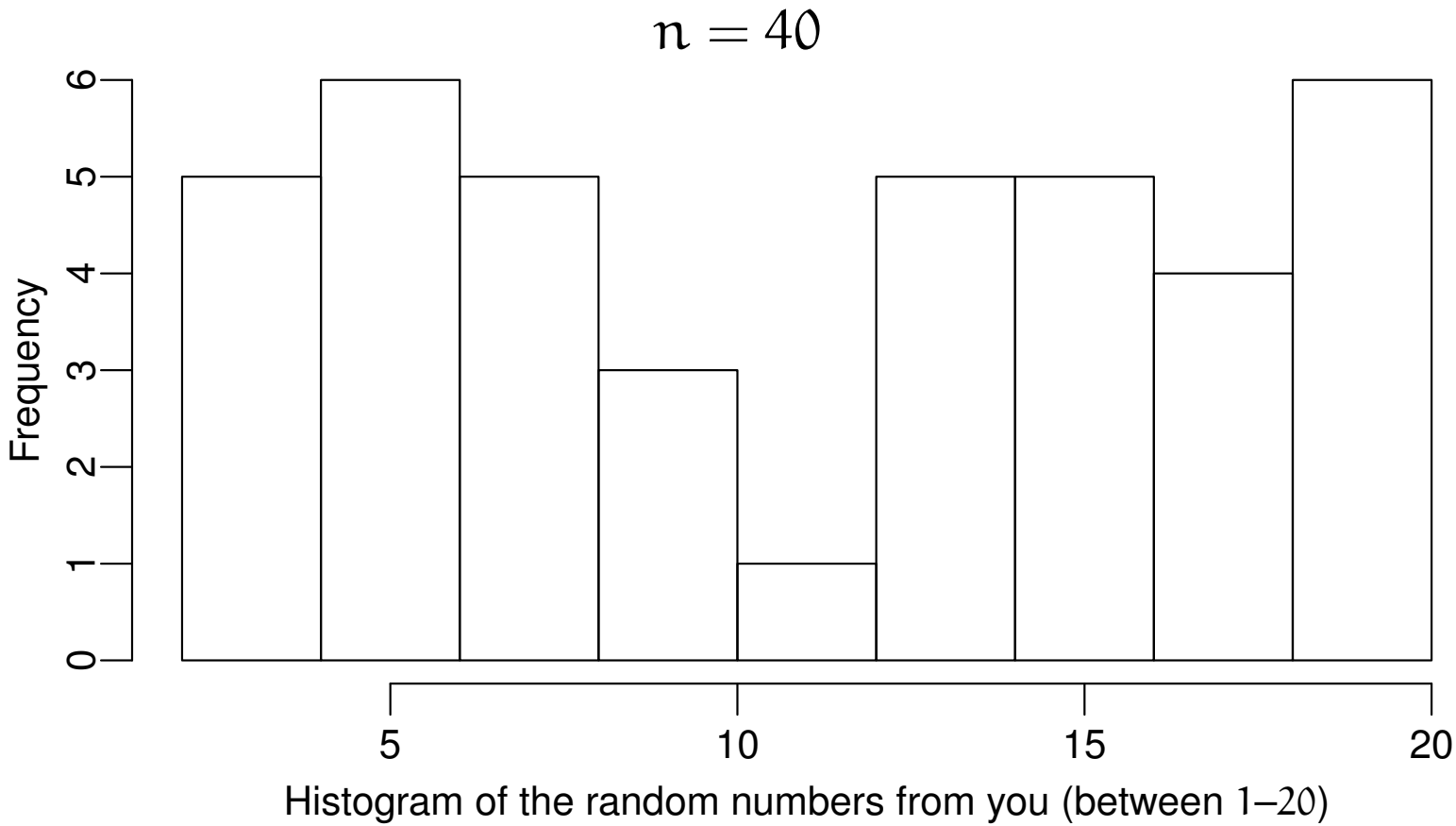
Statistics II
Single ANOVA

Çağrı Çöltekin

University of Groningen
Information Science

April 22, 2014

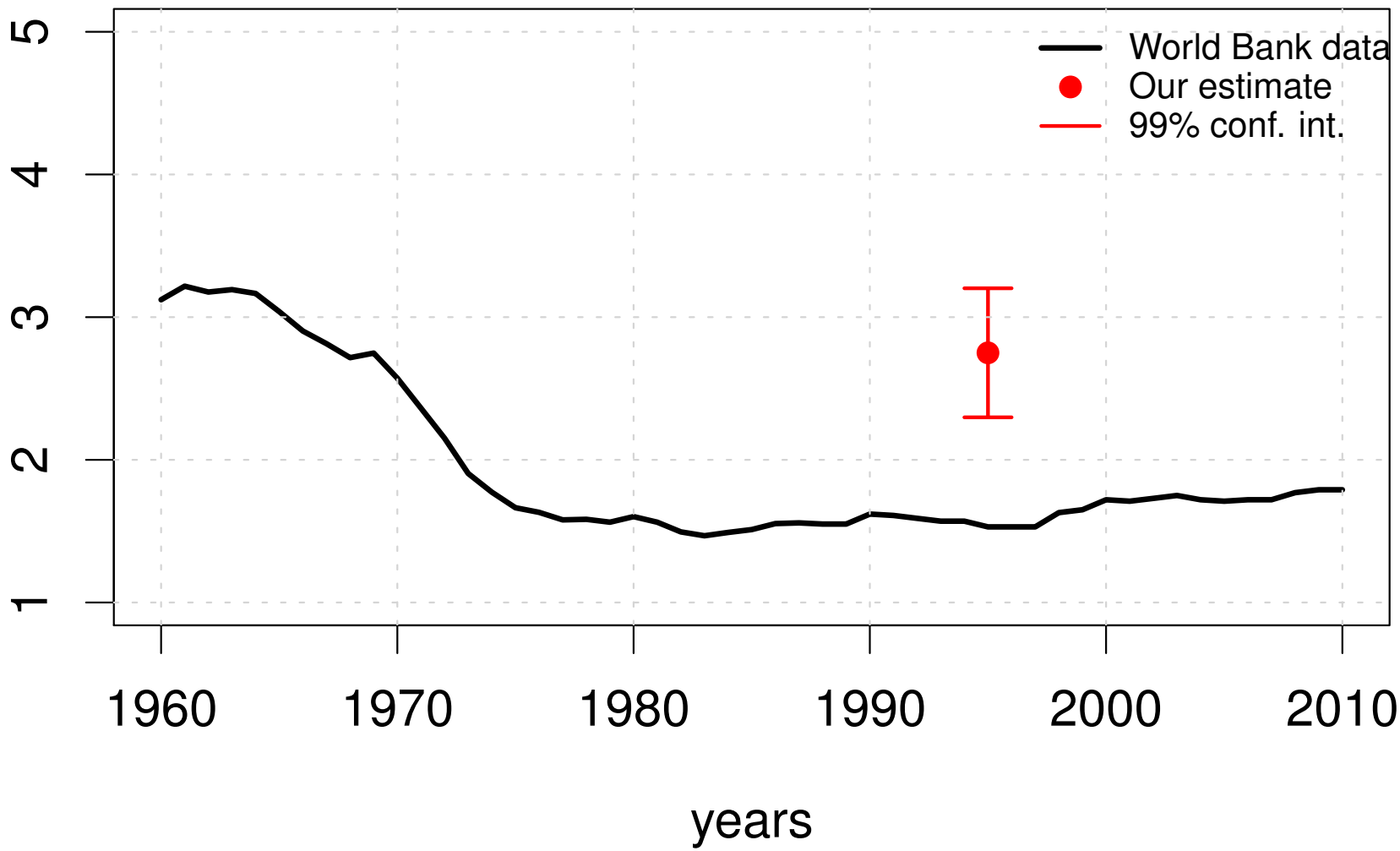
The random numbers from the survey



Class activity

- Your random numbers are in the bag (or 'urn').
- Pick four numbers randomly
- Write them down.
- Pass it to your neighbor.
- Calculate the mean of the numbers that you sampled.
- During the break, draw a box on the board based on your calculation.

Back to the birth-rate estimate



Multiple choice or open ended questions?

Last year the exam included two equally-weighted sections, one with 'multiple choice' and the other 'open ended' questions.
Do we need the open ended questions?

Multiple choice or open ended questions?

Last year the exam included two equally-weighted sections, one with 'multiple choice' and the other 'open ended' questions.
Do we need the open ended questions?

```
lm(formula = open ~ multiple.choice, data = exam)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7165    3.2795  0.523   0.602
multiple.choice 1.1084 0.1223  9.064 5.93e-14
---
Residual standard error: 6.76 on 81 degrees of freedom
Multiple R-squared: 0.5036, Adjusted R-squared: 0.4974
F-statistic: 82.16 on 1 and 81 DF, p-value: 5.933e-14
```

- What is the correlation between the two parts?

Multiple choice or open ended questions?

Last year the exam included two equally-weighted sections, one with 'multiple choice' and the other 'open ended' questions.
Do we need the open ended questions?

```
lm(formula = open ~ multiple.choice, data = exam)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7165    3.2795  0.523   0.602
multiple.choice 1.1084 0.1223  9.064 5.93e-14
---
Residual standard error: 6.76 on 81 degrees of freedom
Multiple R-squared: 0.5036, Adjusted R-squared: 0.4974
F-statistic: 82.16 on 1 and 81 DF, p-value: 5.933e-14
```

- What is the correlation between the two parts?
- Is the correlation significant?

Multiple choice or open ended questions?

Last year the exam included two equally-weighted sections, one with 'multiple choice' and the other 'open ended' questions.
Do we need the open ended questions?

```
lm(formula = open ~ multiple.choice, data = exam)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7165    3.2795  0.523   0.602
multiple.choice 1.1084 0.1223  9.064 5.93e-14
---
Residual standard error: 6.76 on 81 degrees of freedom
Multiple R-squared: 0.5036, Adjusted R-squared: 0.4974
F-statistic: 82.16 on 1 and 81 DF, p-value: 5.933e-14
```

- What is the correlation between the two parts?
- Is the correlation significant?
- How do you interpret the intercept?

Multiple choice or open ended questions?

Last year the exam included two equally-weighted sections, one with 'multiple choice' and the other 'open ended' questions.
Do we need the open ended questions?

```
lm(formula = open ~ multiple.choice, data = exam)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7165    3.2795  0.523   0.602
multiple.choice 1.1084 0.1223  9.064 5.93e-14
---
Residual standard error: 6.76 on 81 degrees of freedom
Multiple R-squared: 0.5036, Adjusted R-squared: 0.4974
F-statistic: 82.16 on 1 and 81 DF, p-value: 5.933e-14
```

- What is the correlation between the two parts?
- Is the correlation significant?
- How do you interpret the intercept?
- How do you interpret the slope?

Multiple choice or open ended questions?

Last year the exam included two equally-weighted sections, one with 'multiple choice' and the other 'open ended' questions.
Do we need the open ended questions?

```
lm(formula = open ~ multiple.choice, data = exam)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7165    3.2795  0.523   0.602
multiple.choice 1.1084 0.1223  9.064 5.93e-14
---
Residual standard error: 6.76 on 81 degrees of freedom
Multiple R-squared: 0.5036, Adjusted R-squared: 0.4974
F-statistic: 82.16 on 1 and 81 DF, p-value: 5.933e-14
```

- What is the correlation between the two parts?
- Is the correlation significant?
- How do you interpret the intercept?
- How do you interpret the slope?
- How should we check the model fit?

Another regression example: predicting shoe size from height

```
lm(formula = shoesize ~ height, data = survey)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.44853 2.64985 13.000 1.46e-15 ***
height       0.03092   0.01532  2.019  0.0506 .
---
Residual standard error: 2.758 on 38 degrees of freedom
Multiple R-squared: 0.09686, Adjusted R-squared: 0.07309
F-statistic: 4.075 on 1 and 38 DF, p-value: 0.05061
```

Let's try weight

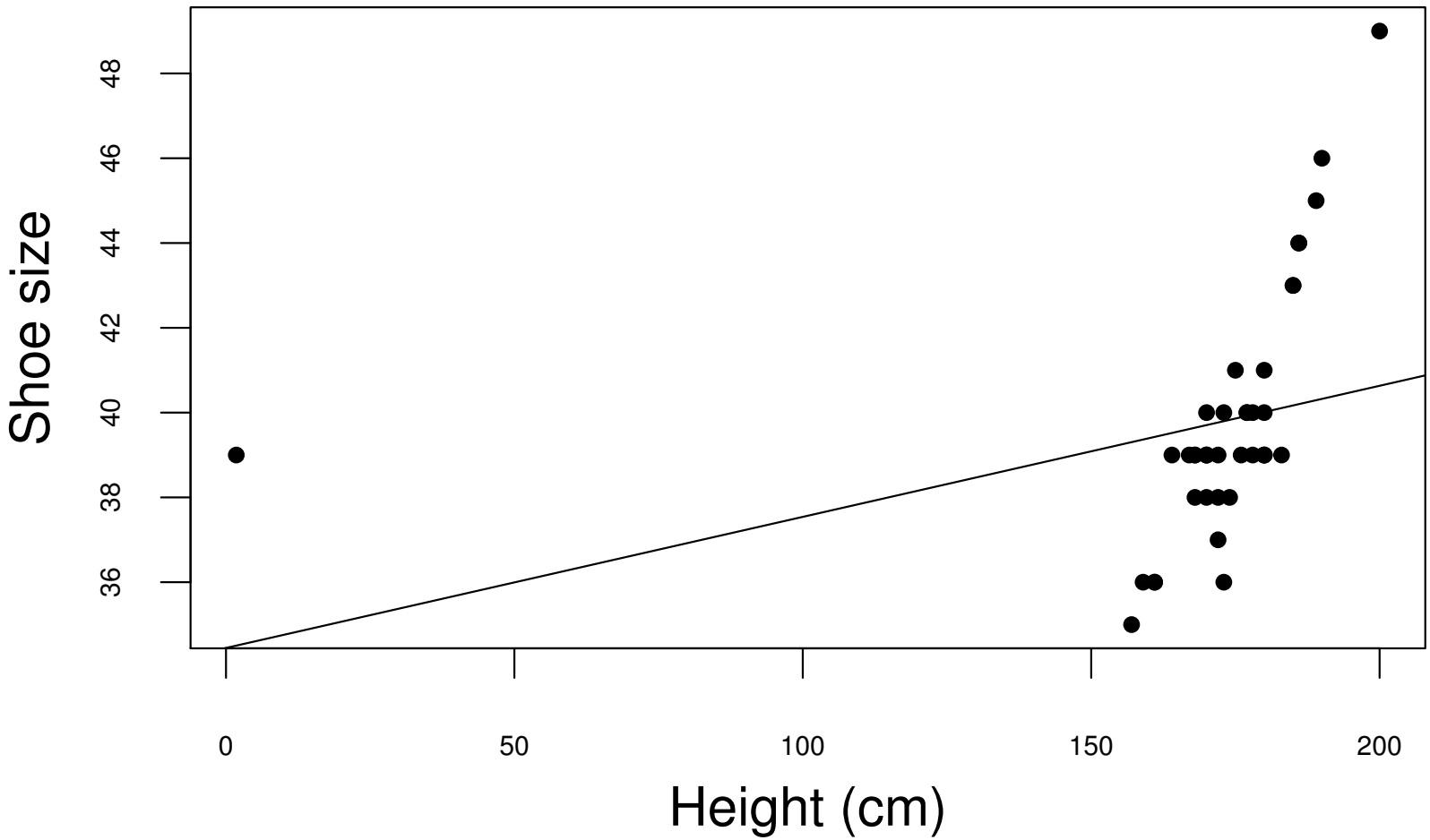
```
lm(formula = shoesize ~ weight, data = survey)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.79832 1.50902 18.421 < 2e-16 ***
weight       0.17914   0.02255  7.943 1.98e-09 ***
---
Residual standard error: 1.714 on 36 degrees of freedom
Multiple R-squared: 0.6367, Adjusted R-squared: 0.6266
F-statistic: 63.09 on 1 and 36 DF, p-value: 1.984e-09
```

Let's try weight

```
lm(formula = shoesize ~ weight, data = survey)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.79832 1.50902 18.421 < 2e-16 ***
weight       0.17914   0.02255  7.943 1.98e-09 ***
---
Residual standard error: 1.714 on 36 degrees of freedom
Multiple R-squared: 0.6367, Adjusted R-squared: 0.6266
F-statistic: 63.09 on 1 and 36 DF, p-value: 1.984e-09
```

Are you convinced?

Always plot your data



Once more with height (using the corrected data)

```
lm(formula = shoesize ~ height, data = survey)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.44831  4.49161 -1.881  0.0677 .
height       0.27520  0.02563 10.739 4.54e-13 ***
---
Residual standard error: 1.445 on 38 degrees of freedom
Multiple R-squared: 0.7522, Adjusted R-squared: 0.7456
F-statistic: 115.3 on 1 and 38 DF, p-value: 4.535e-13
```

Using both height and weight

In comparison to the single predictor models, do you expect changes in

- coefficient estimates?
- statistical significance?
- r^2?

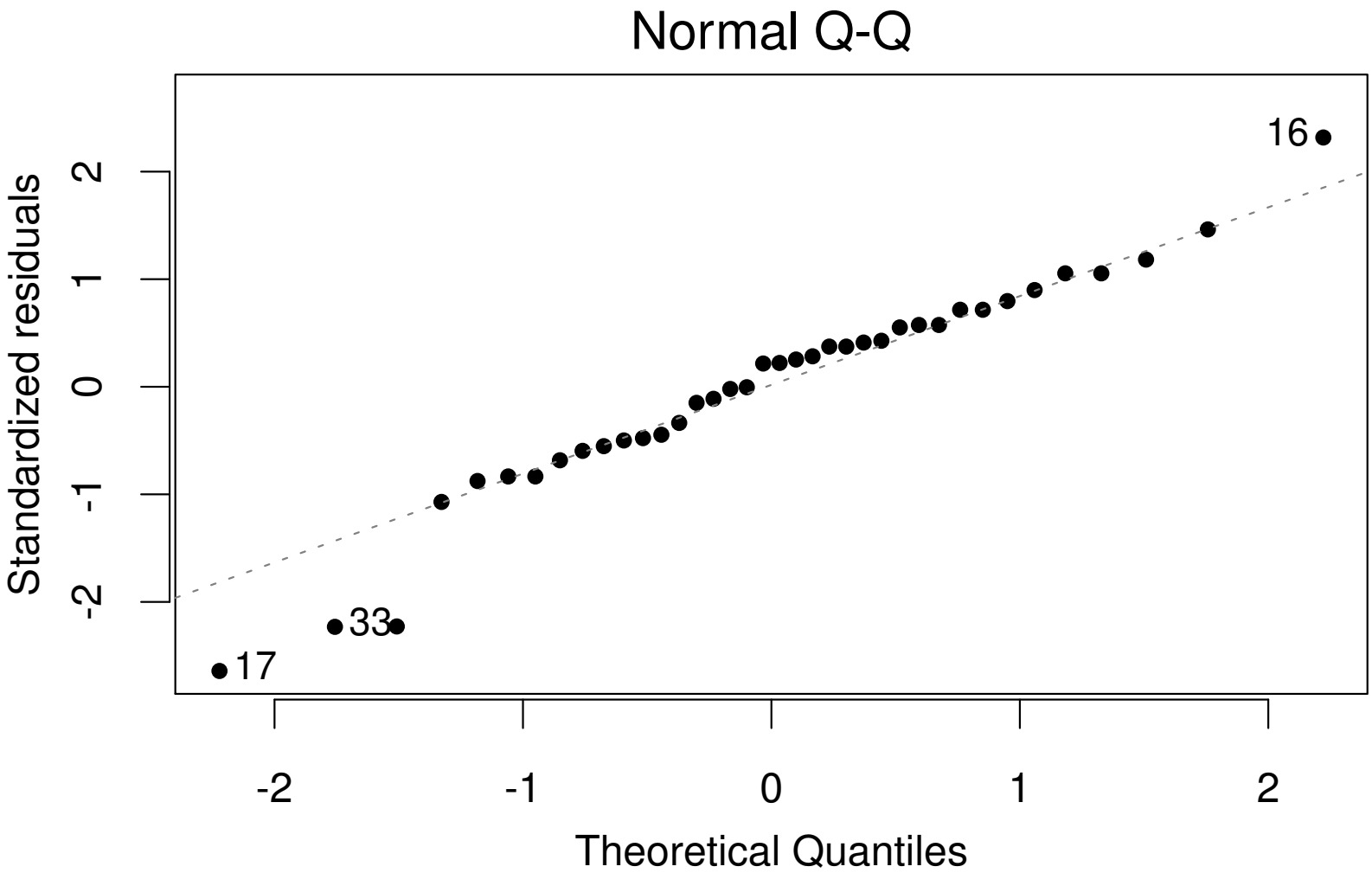
Using both height and weight

In comparison to the single predictor models, do you expect changes in

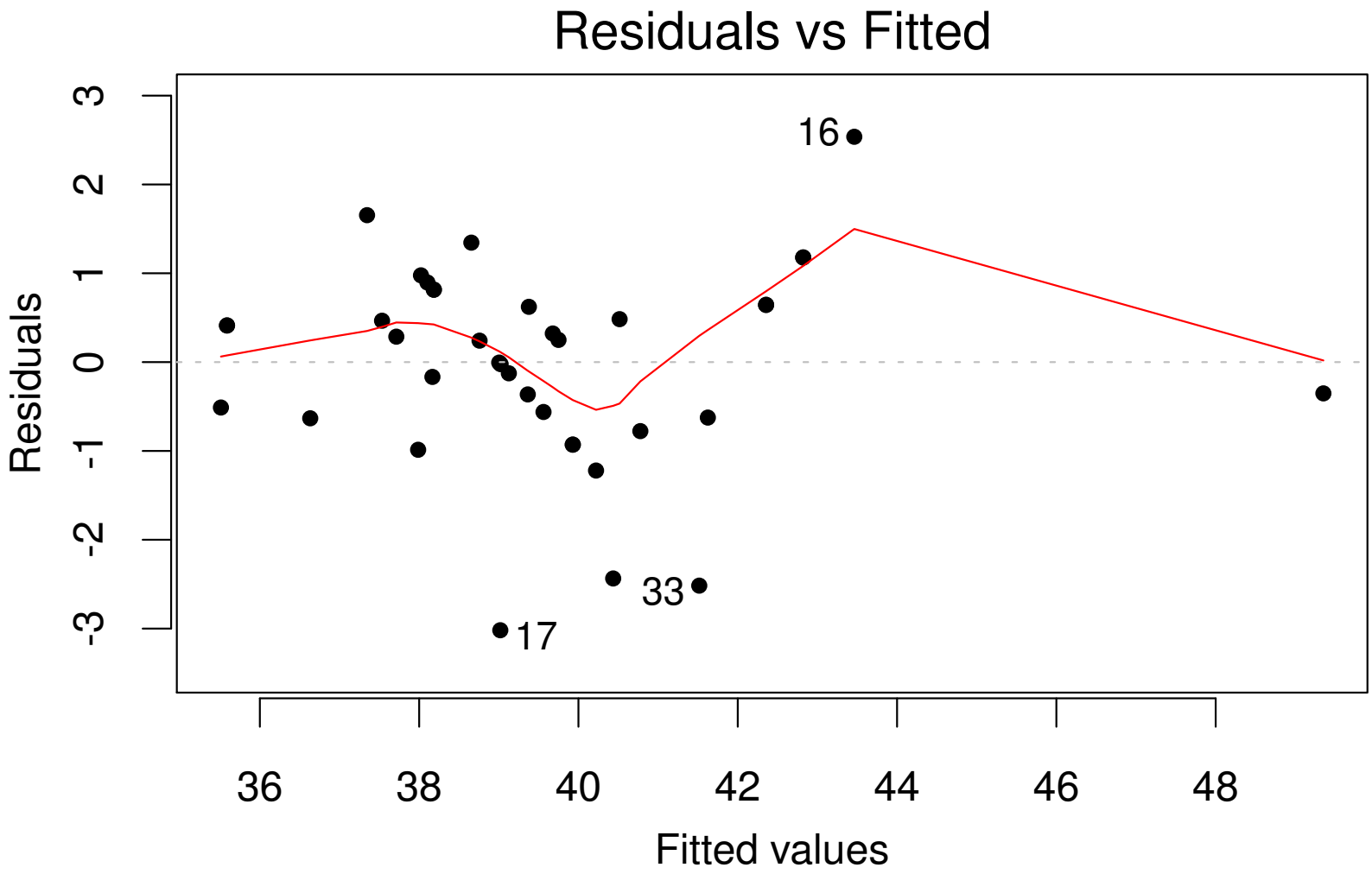
- coefficient estimates?
- statistical significance?
- r^2?

```
lm(formula = shoesize ~ height.fixed + weight, data = survey)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.27327   4.13571  0.308    0.76
height.fixed 0.18388  0.02778  6.618 1.20e-07 ***
weight       0.09419  0.01993  4.726 3.66e-05 ***
---
Residual standard error: 1.159 on 35 degrees of freedom
Multiple R-squared: 0.8386, Adjusted R-squared: 0.8294
F-statistic: 90.95 on 2 and 35 DF, p-value: 1.371e-14
```

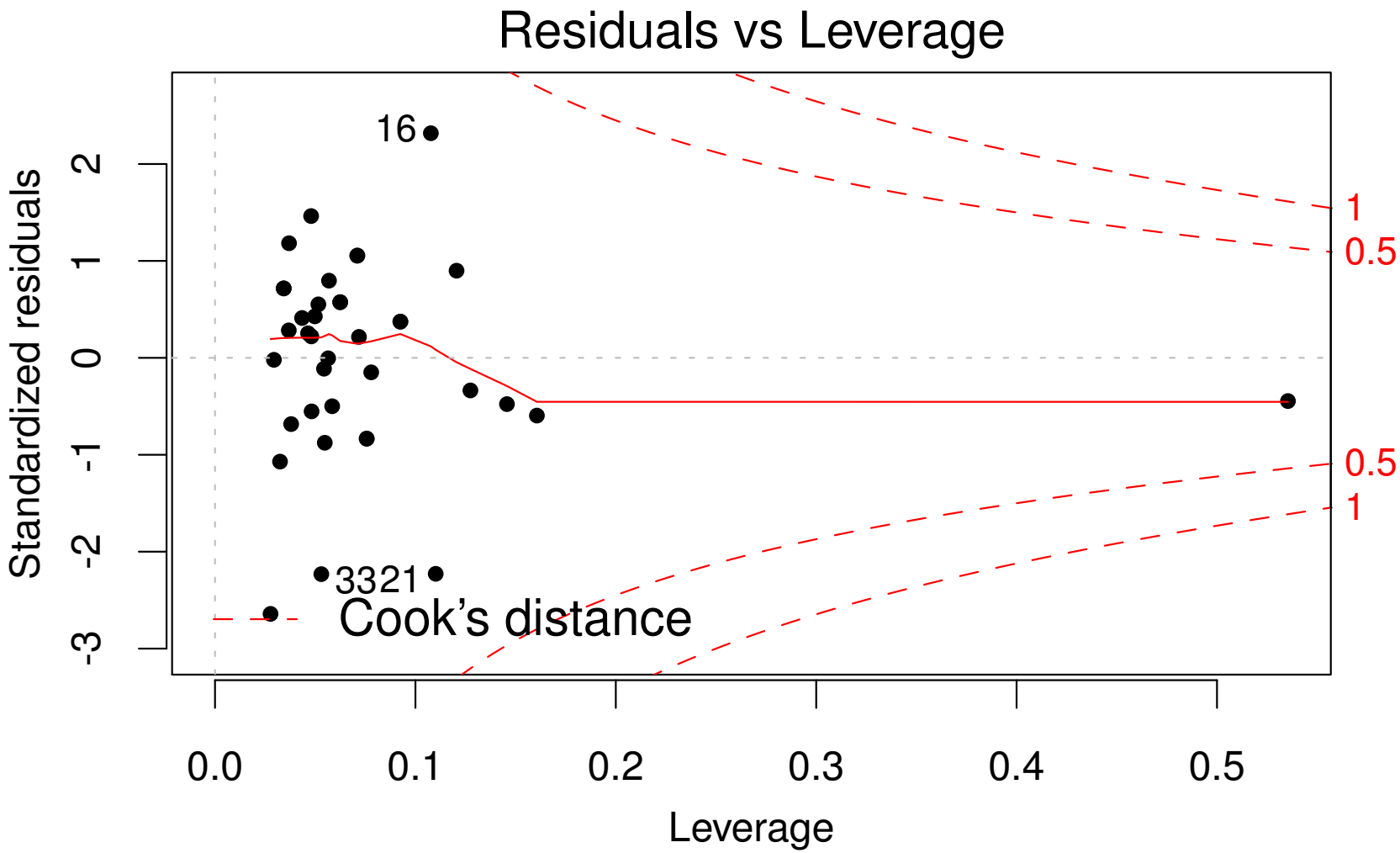
Model diagnostics: normality of residuals



Model diagnostics: residual distribution



Model diagnostics: influential observations



Hypothesis testing: a reminder

Aim: make inferences about the population based on a sample regarding a research question.

Procedure:

- Formulate your question as two explicit hypotheses:

  null hypothesis (H_0) is the formulation of the case where your expectations were wrong.

  alternative hypothesis (H_a) supports what you expect to find in the population.

- Set a probability level (α-level) at which to reject the H_0. Typical values are 0.05, 0.01, 0.001.
- Calculate the probability, p, of obtaining the sample you have, if H_0 was true.
- If $p < \alpha$, we reject the H_0, otherwise, we fail to reject the H_0.

Hypothesis testing: example

We want to know whether a new web-page design is easier to use based on evaluation from two groups of users, one using only the old design, the other only the new design.

Procedure:

- Formulate your question as two explicit hypotheses:

  H_0 the mean response score is the same for both groups ($\mu_1 = \mu_2$).

  H_a the mean response score differs for the groups ($\mu_1 \neq \mu_2$).
- We set $\alpha = 0.05$.
- The p-value for obtaining these samples given H_0 is true, can be calculated from a t-distribution (given the response scores for both groups are normally distributed, and the variances are similar).
- If $p < 0.05$, we reject the H_0, otherwise, we fail to reject the H_0.

Hypothesis testing: Type I and Type II errors

		Real world	
		H_0 is false	H_0 is true
Test decision	Reject H_0 ($p > \alpha$)	Correct decision True positive	Type I error False positive
	Fail to reject H_0 ($p \leq \alpha$)	Type II error False negative	Correct decision True negative

- Note that accepting H_0 means we will be wrong (committing a Type I error) with probability α.
- If we set $\alpha = 0.05$, and repeat an experiment multiple times, we expect to reject the null hypothesis once every 20 runs even it is true.

The inequalities are wrong. It should be 'reject H_0 if $p < \alpha$, fail to reject otherwise'.

Example problems for ANOVA

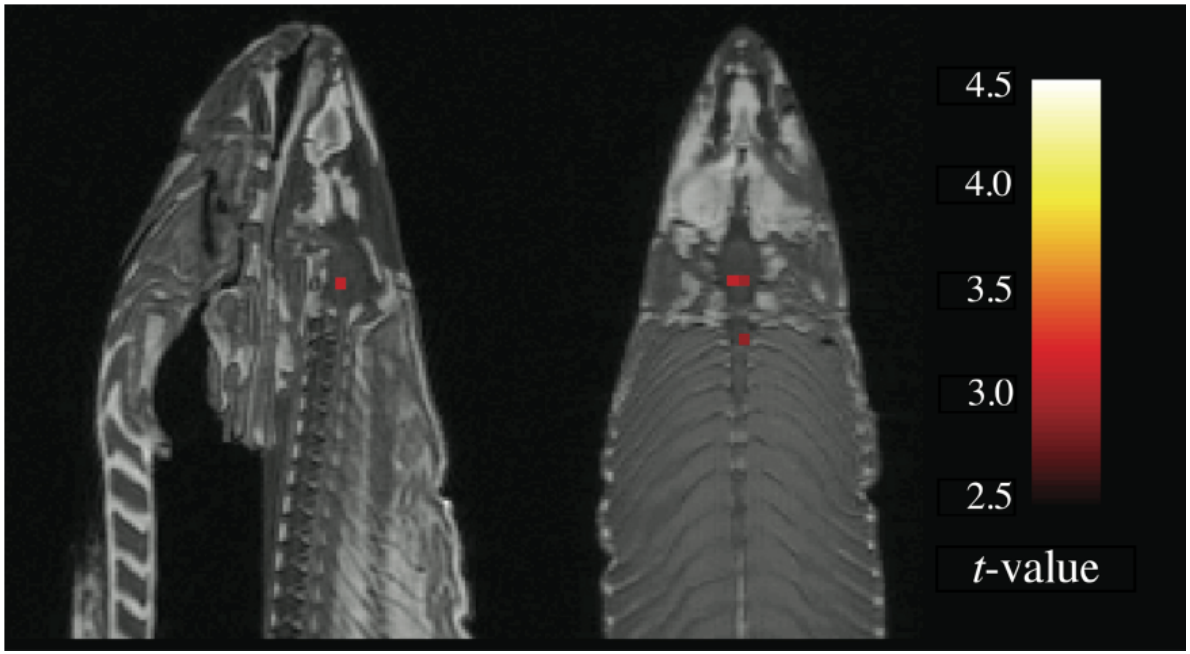
- Compare time needed for lexical recognition in
  1. healthy adults
  2. patients with Wernicke's aphasia
  3. patients with Broca's aphasia
- Effect of background color choice in a web site.
- Compare Dutch proficiency scores of second language learners based on their native language.

Why not multiple t-tests?

- Multiple comparisons over the same sample increases the chances of rejecting the null hypothesis (finding an effect where there is none).
- With $\alpha = 0.05$, if you do 20 different t-tests on the same sample, we expect to one of them being significant if the null hypothesis was true.
- We need
  3 comparisons 3 groups,
  6 comparisons for 4 groups,
  10 comparisons for 5 groups,
  45 comparisons for 10 groups,
  4950 comparisons for 100 groups.
- In general, for k groups, we need $\binom{k}{2}$ comparisons.

An extreme demonstration

finding emotional response in a dead salmon's brain activity



Subject One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at he time of scanning.

Task The task administered to the salmon involved completing an open-end mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Results Several active voxels were discovered in a cluster located within the salmon's brain cavity ...with a cluster-level significance of $p = 0.001$. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

* From the poster by Bennett et al. (2009).

ANOVA

- ANOVA (analysis of variance) is a method to compare means of more than two groups.
- For two groups the result is equivalent to t-test.
- ANOVA indicate whether there is any difference at all. For k groups:

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_k$$

- A limited number and type of comparisons can be carried out by specifying contrasts.
- Otherwise, post-hoc pairwise comparisons can be carried out using corrected α-levels.
- ANOVA is strongly related to regression (later today).

Logic of ANOVA

We want to know whether there are any differences between the means of k groups.

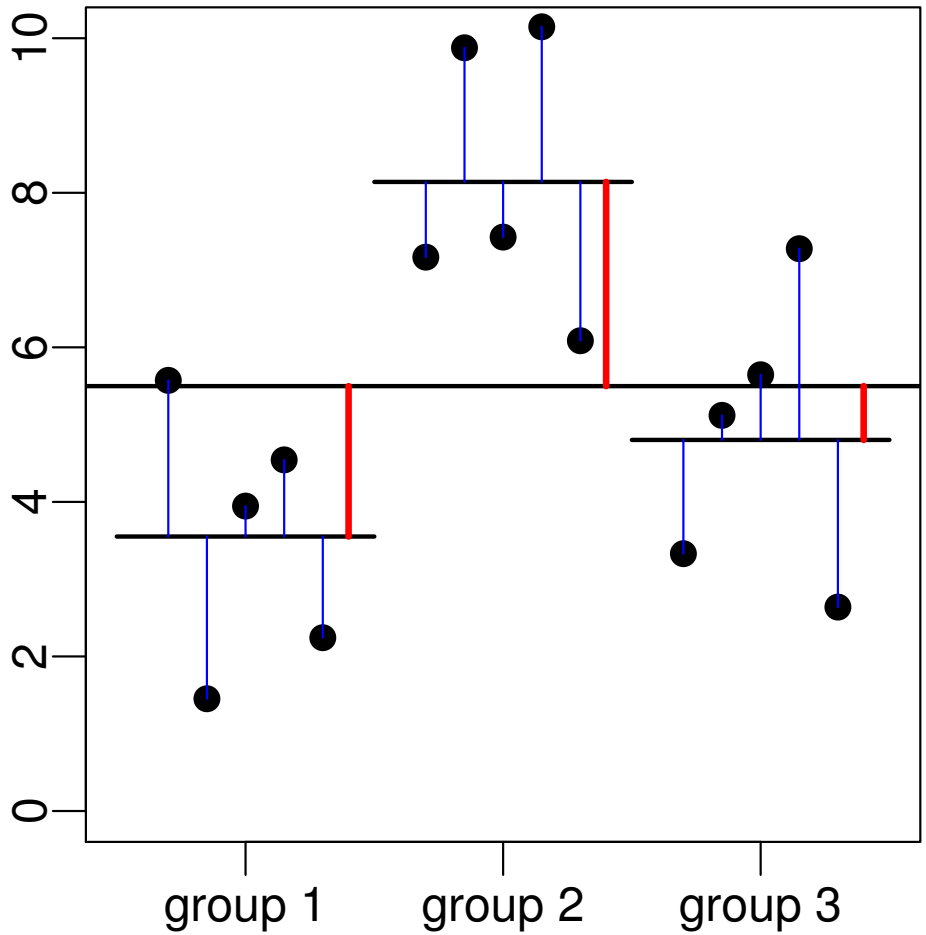
- If the variance between the groups is higher than the variance within the groups, there must be a significant group effect.
- Between group variance ($MS_{between}$, or MS_M or MS_G) is characterized by variance between the group means.
- Within group variance (MS_{within}, or MS_R or MS_E) is characterized by variance of data round the group means.

Then, the statistic of interest is

$$F = \frac{MS_{between}}{MS_{within}}$$

.

ANOVA: visualization



$$F = \frac{MS_{between}}{MS_{within}}$$

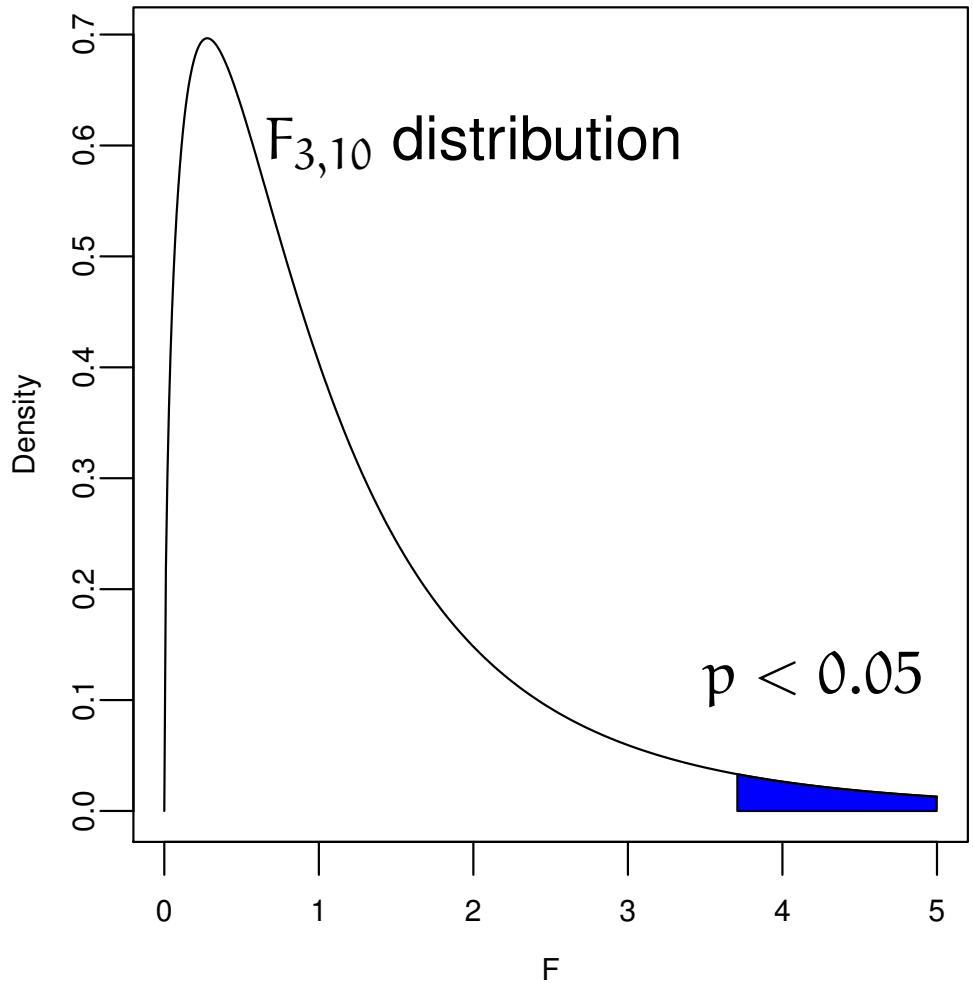
$$F = \frac{\frac{SS_{between}}{DF_{between}}}{\frac{SS_{within}}{DF_{within}}}$$

$$DF_{between} = k - 1$$
$$DF_{within} = n - k$$

where k is the number of groups, and n is the number of observations.

F-distribution

- Ratio of variances follows F distribution.
- F distribution has two parameters, $\mathrm{DF}_{numerator}$ and $\mathrm{DF}_{denominator}$.
- If variances are equal, we get $\mathrm{F} = 1$.
- In ANOVA, we get an effect if $MS_{between}$ is larger than MS_{within}.



ANOVA: assumptions

- All observations are independent.
- The response within each group follow an approximately normal distribution.
- The variances within each group are approximately the same.

ANOVA: example

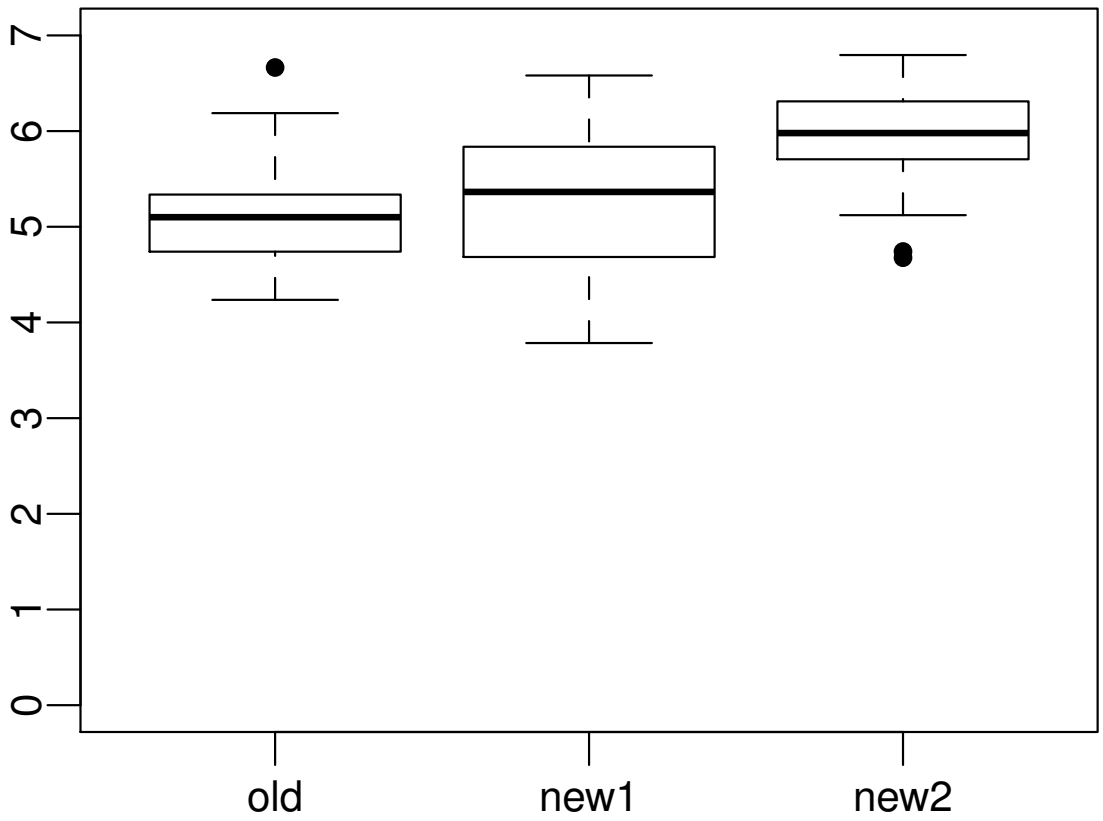
We have two new designs for company web site, want to know which one is easier to use. We test the new web site prototypes and the old one on three different group, and get their opinion via a questionnaire with 7-point scale. The data looks like:

	Old	New 1	New 2
	4.4	6.6	5.9
	5.8	6.2	4.9
	⋮	⋮	⋮
Mean	4.76	5.03	6.11
Variance	1.11	1.12	0.97

Note: rows in the table are not related!

Visualizing the data

Box-and-whisker plots (or box plots) are one of the best ways to visualize this type of data.



Note: the vertical bars are medians.

ANOVA results from software

```
Analysis of Variance Table
Response: ease
          Df Sum Sq Mean Sq F value Pr(>F)
design     2 10.796 5.3978 13.955 5.541e-06 ***
Residuals 87 33.652 0.3868
```

- There is a significant effect (for any conventional α level, the p-values is 0.0000055)
- but we do not know where the effect is.

Regression with categorical predictors: some terminology

- We take 'grouping' variables like design as categorical (or factor) variables.
- The values a categorical variable take are called levels.
- A categorical variable with k levels is converted to $k-1$ numeric variables, called 'indicator' or 'dummy' variables.

Regression with categorical predictors

- Consider the 'design' variable with three levels ('old', 'new 1', 'new 2'), we can code it as two variables, 'Contrast 1', 'Contrast 2' :

design	Contrast 1	Contrast 2
old	0	0
new 1	1	0
new 2	0	1

- Other coding options (contrasts) are possible. With some constraints, the inferences will not change.

An example with only two levels

Do men or women are better statisticians?

An example with only two levels

Do men or women are better statisticians?
Normally we would do a t-test:

```
          Two Sample t-test
data: stat1grade by gender
t = 0.2086, df = 37, p-value = 0.8359
alt. hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7779205 0.9564919
sample estimates:
mean in group F mean in group M
      7.375000      7.285714
```

Doing t-test with regression

- We have two levels of the predictor (`F` and `M`).
- We code `F` as 0 and `M` as 1.

$$y_i = a + b \times genderM_i + e_i$$

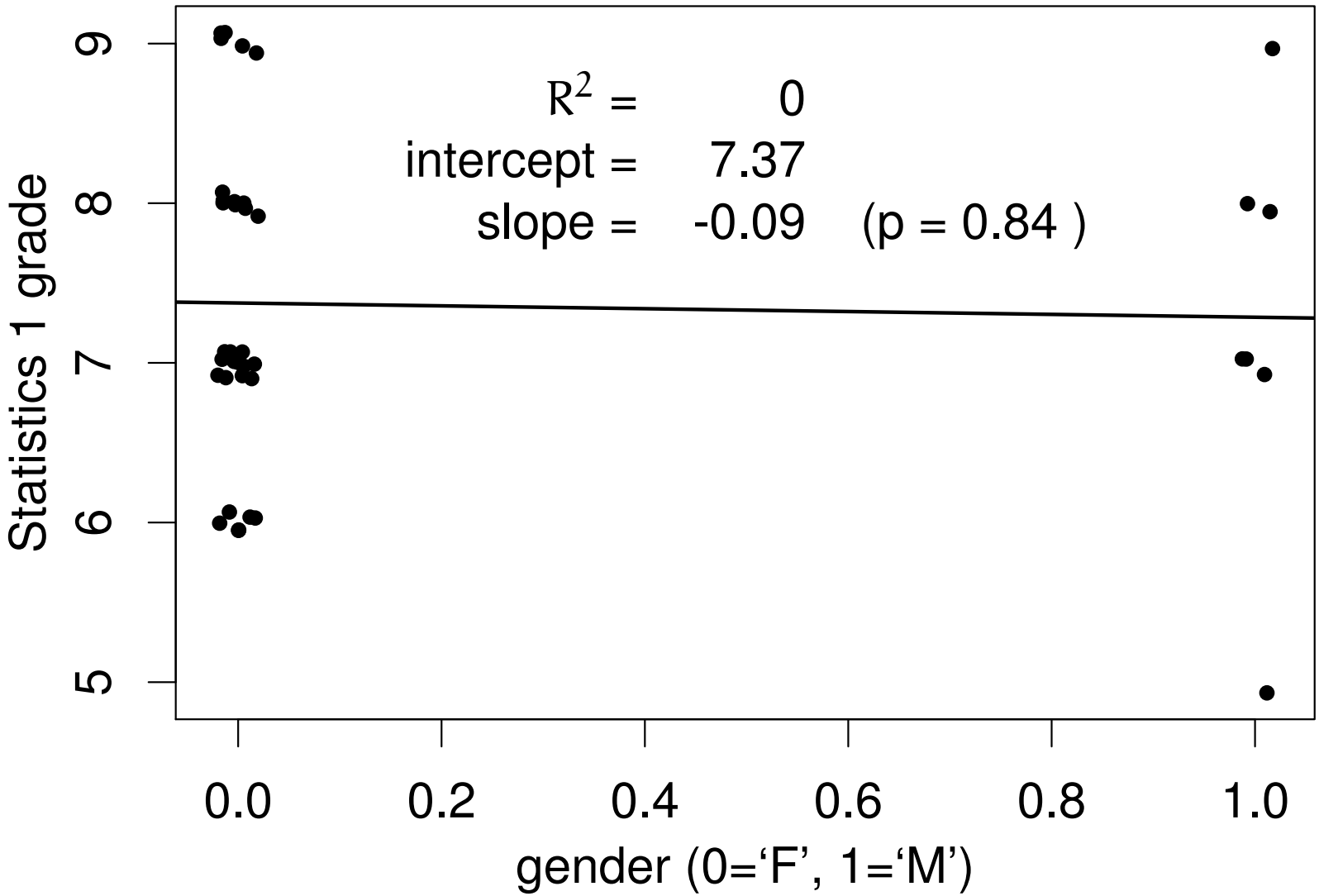
a (intercept) is the mean of level `F`.

b (slope) is the mean difference between `M` and `F`.

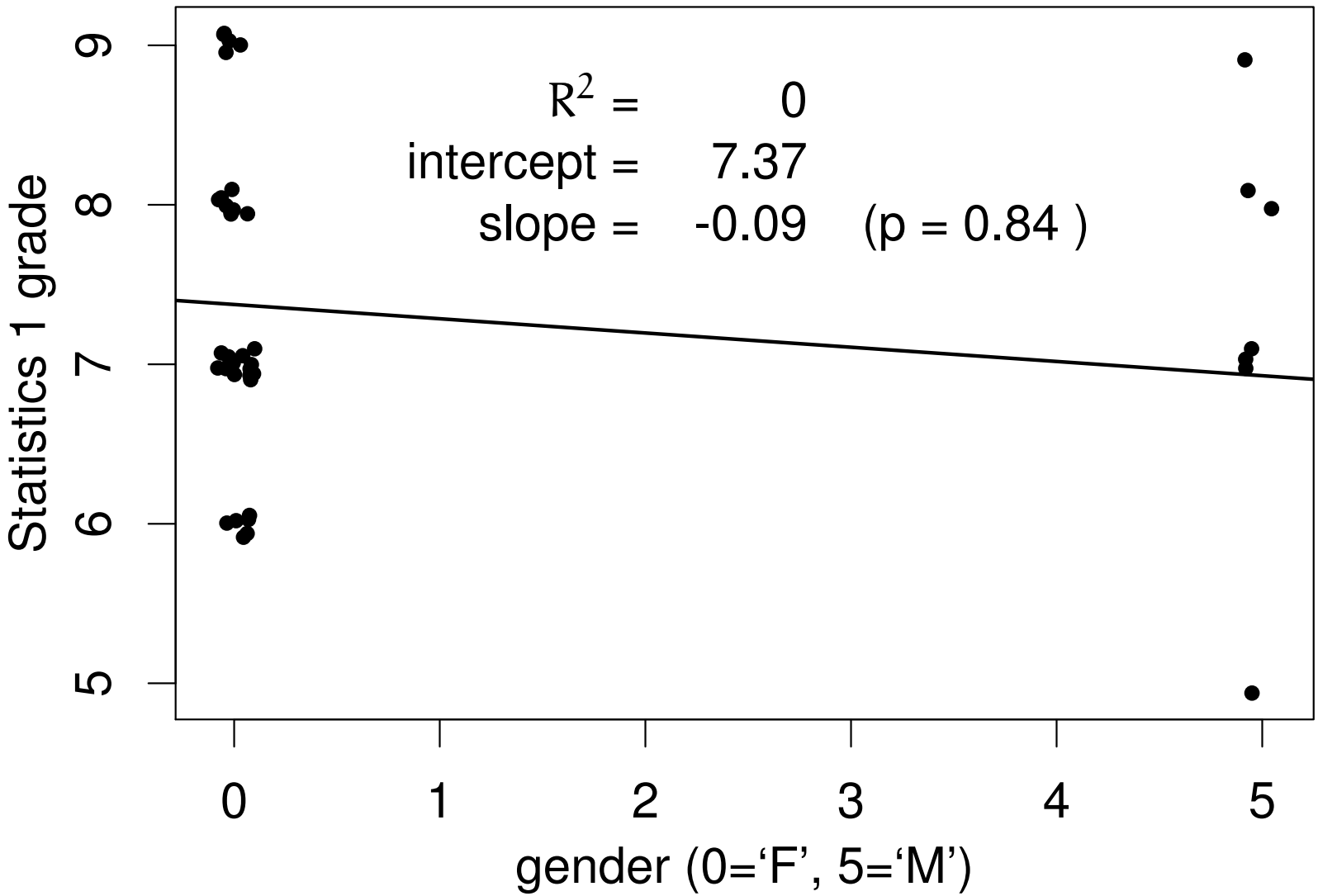
Doing t-test with regression: practice

```
lm(formula = stat1grade ~ gender, data = survey)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.37500  0.18133 40.673  <2e-16
genderM    -0.08929  0.42800 -0.209   0.836
---
Residual standard error: 1.026 on 37 degrees of freedom
Multiple R-squared: 0.001175, Adjusted R-squared: -0.02582
F-statistic: 0.04352 on 1 and 37 DF, p-value: 0.8359
```

T-test as regression: the picture



T-test as regression: the picture



ANOVA as regression

Back to our web-site design example.

$$y_i = a + b_1 \times x_{1,i} + b_2 \times x_{2,i} + e_i$$

- a (intercept) is the mean of old design.
- b_1 (slope of x1) is the mean difference between new 1 and old.
- b_2 (slope of x2) is the mean difference between new 2 and old.

ANOVA as regression: practice

Regression view:

```
lm(formula = ease ~ design, data = webdesign)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.1701    0.1135 45.531 < 2e-16
designnew1  0.1166    0.1606  0.726    0.47
designnew2  0.7860    0.1606  4.895 4.49e-06
---
Residual standard error: 0.6219 on 87 degrees of freedom
Multiple R-squared: 0.2429, Adjusted R-squared: 0.2255
F-statistic: 13.95 on 2 and 87 DF, p-value: 5.541e-06
```

ANOVA as regression: practice

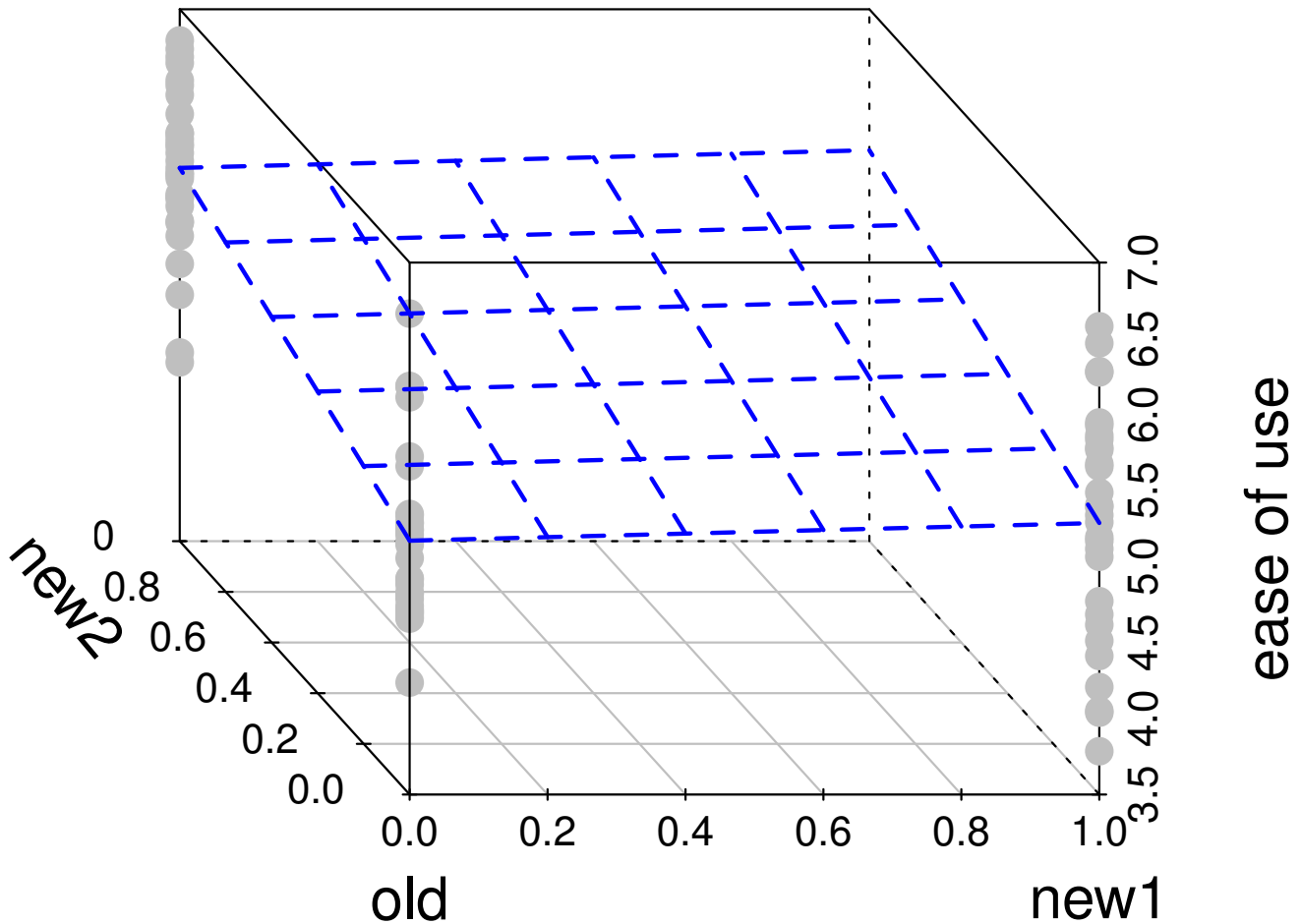
Regression view:

```
lm(formula = ease ~ design, data = webdesign)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.1701    0.1135 45.531 < 2e-16
designnew1  0.1166    0.1606  0.726    0.47
designnew2  0.7860    0.1606  4.895 4.49e-06
---
Residual standard error: 0.6219 on 87 degrees of freedom
Multiple R-squared: 0.2429, Adjusted R-squared: 0.2255
F-statistic: 13.95 on 2 and 87 DF, p-value: 5.541e-06
```

ANOVA view:

```
            Df Sum Sq Mean Sq F value Pr(>F)
design       2  10.80   5.398   13.96 5.54e-06 ***
Residuals 87  33.65   0.387
```

ANOVA as regression: the picture



Contrast coding

- For k levels (or groups) we have $k - 1$ coefficients.
- We can code some comparisons (contrasts) into these coefficients to test for differences between certain groups, or certain trends due to groups.
- If the contrasts does not inflate the t-value (does not cause additional Type I error) it is called an orthogonal contrast.

Example: different contrasts

We have 3 groups, so we can specify 2 contrasts. Two interesting questions to ask:

1. Are new designs (on average) better than the old one?
2. Are the new designs different?

design	x_1	x_2
old	-2	0
new 1	1	-1
new 2	1	1

Example: different contrasts

We have 3 groups, so we can specify 2 contrasts. Two interesting questions to ask:

1. Are new designs (on average) better than the old one?
2. Are the new designs different?

design	x_1	x_2
old	-2	0
new 1	1	-1
new 2	1	1

- A contrast is orthogonal if columns sum to 0 and product of rows sum to 0.
- Orthogonal contrasts do not increase Type I errors.

Post-hoc comparisons

- In most cases, you will have a specific hypothesis and a (small) set of comparisons to make.
- You can do pairwise comparisons once you found a significant ANOVA result.
- Every comparison you make increases finding a significant difference where there isn't any (Type I error).
- If you do multiple comparisons you need to correct for it.
- Correction is applied such that your α-level is adjusted (called family-wise error rate).

Post-hoc comparisons (2)

Remember: finding a significant difference means that there is a chance (for example, $p = 0.05$) of finding a difference when there is no difference (null hypothesis is true).

- The simplest (and most conservative) correction is called 'Bonferroni' correction, which is obtained by dividing α to number of comparisons. If you have $\alpha = 0.05$ and n comparisons your family-wise α should be $\frac{0.05}{n}$.
- Bonferroni correction is safe in all cases, but increases the Type II error rate.
- There are other multiple-comparison methods that are more powerful, but they typically apply only in specific cases.

Summary

- Single ANOVA is used when we have a single factor with more than two levels/groups.
- ANOVA tests whether there is a difference between means of the groups by comparing the variance within the groups, and variance of the means of the groups.
- ANOVA tests for 'any difference', you can inspect specific differences through planned contrasts, or post-hoc comparisons.
- ANOVA is a specific case of regression.

Next week: more ANOVA. Reading: Ch. 12 (13 in 4th edition), 'factorial ANOVA'.