Stat 411/511

Anova & Regression

Nov 31st 2015

Charlotte Wickham

stat511.cwick.co.nz

This week

Today: Lack of fit F-test

Weds: Review…email me topics, otherwise I'll go over some of last year's final exam questions.

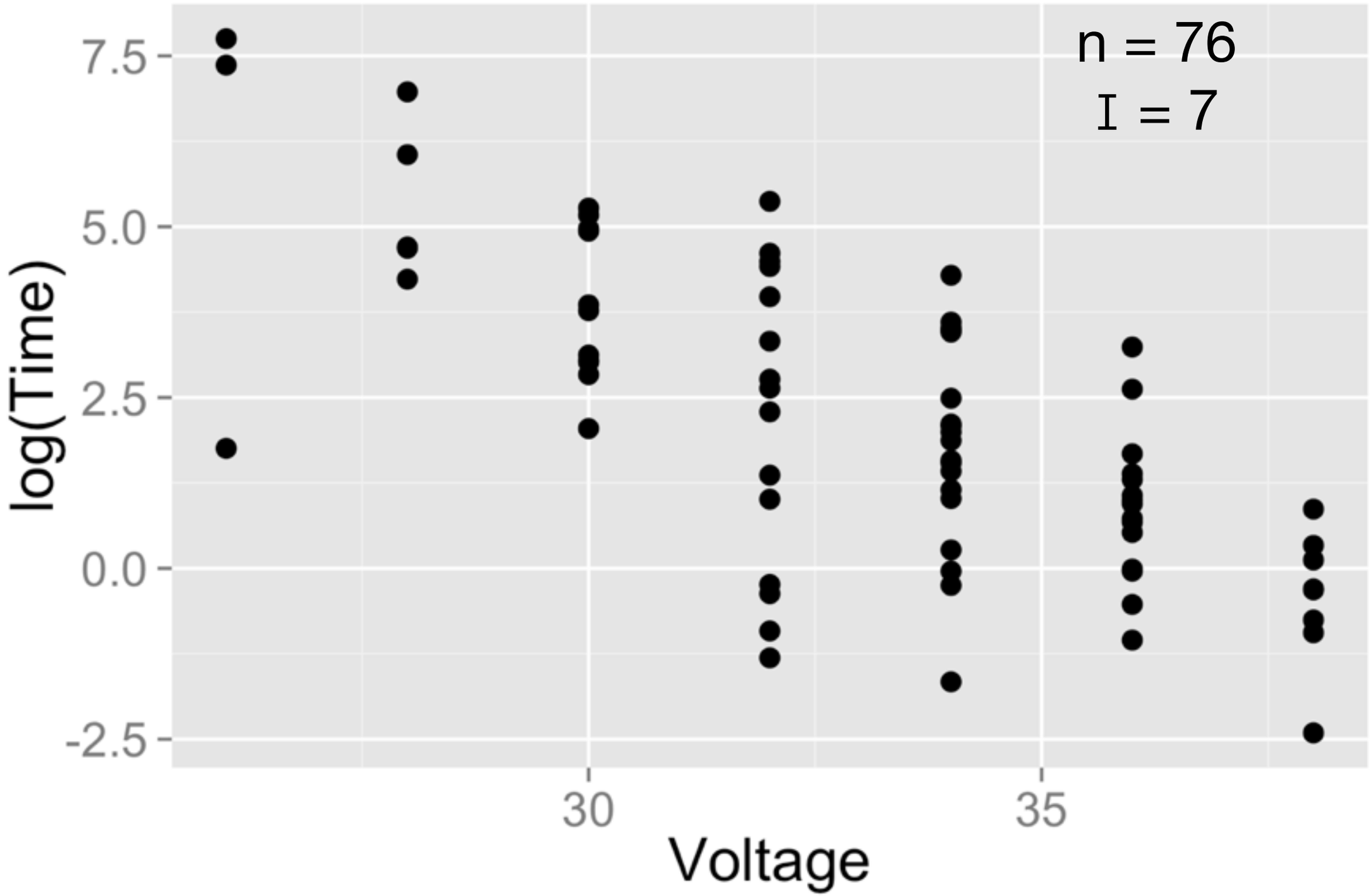
Fri: Office hours instead of lecture. Find me in Weniger 255 10-11am.

Finals week Office hours

Mon & Wed 10-11am in my office 255 Weniger

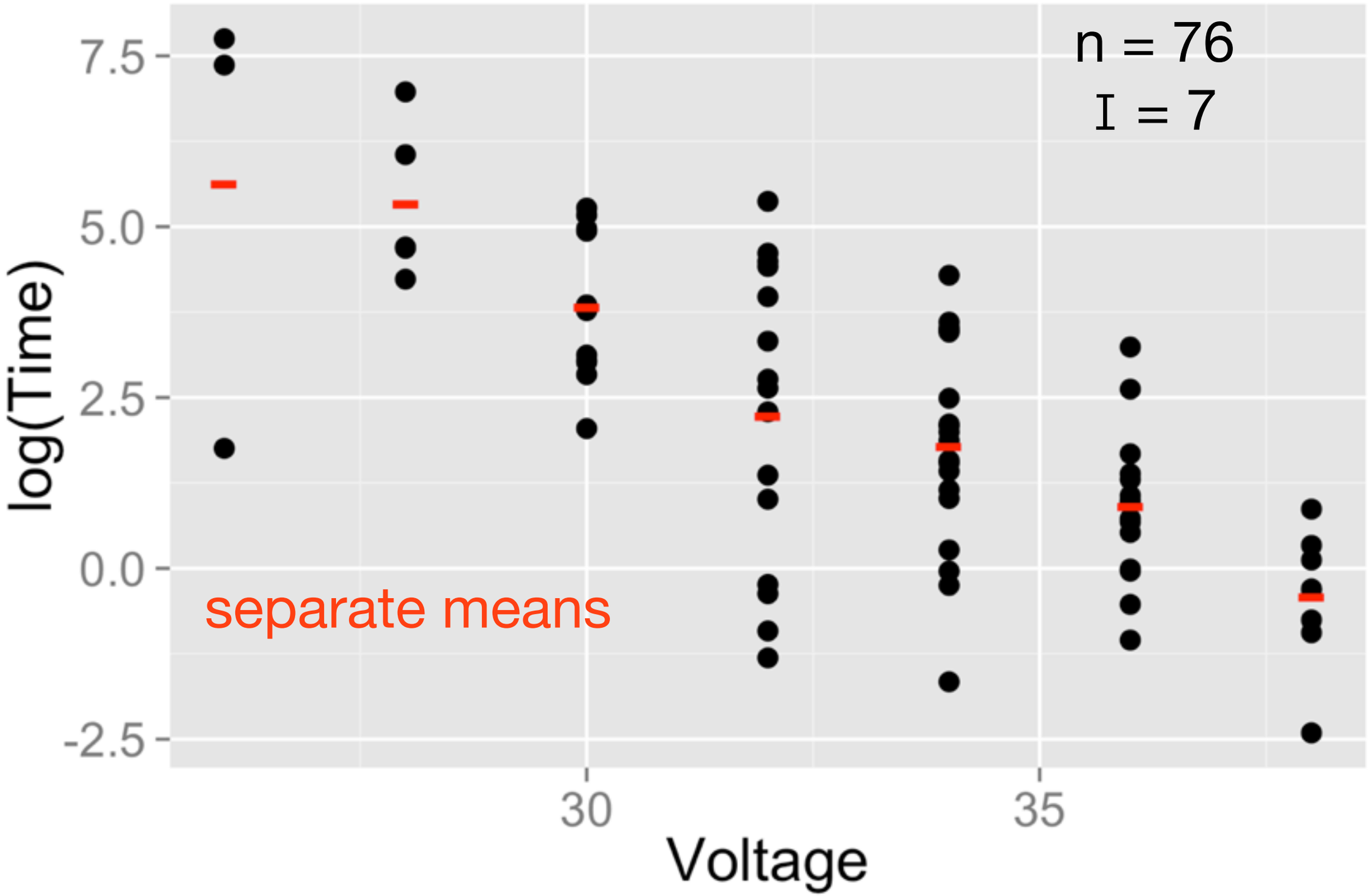
Insulating Fluid Case Study

Breakdown times for electrical insulating fluid at various voltages.



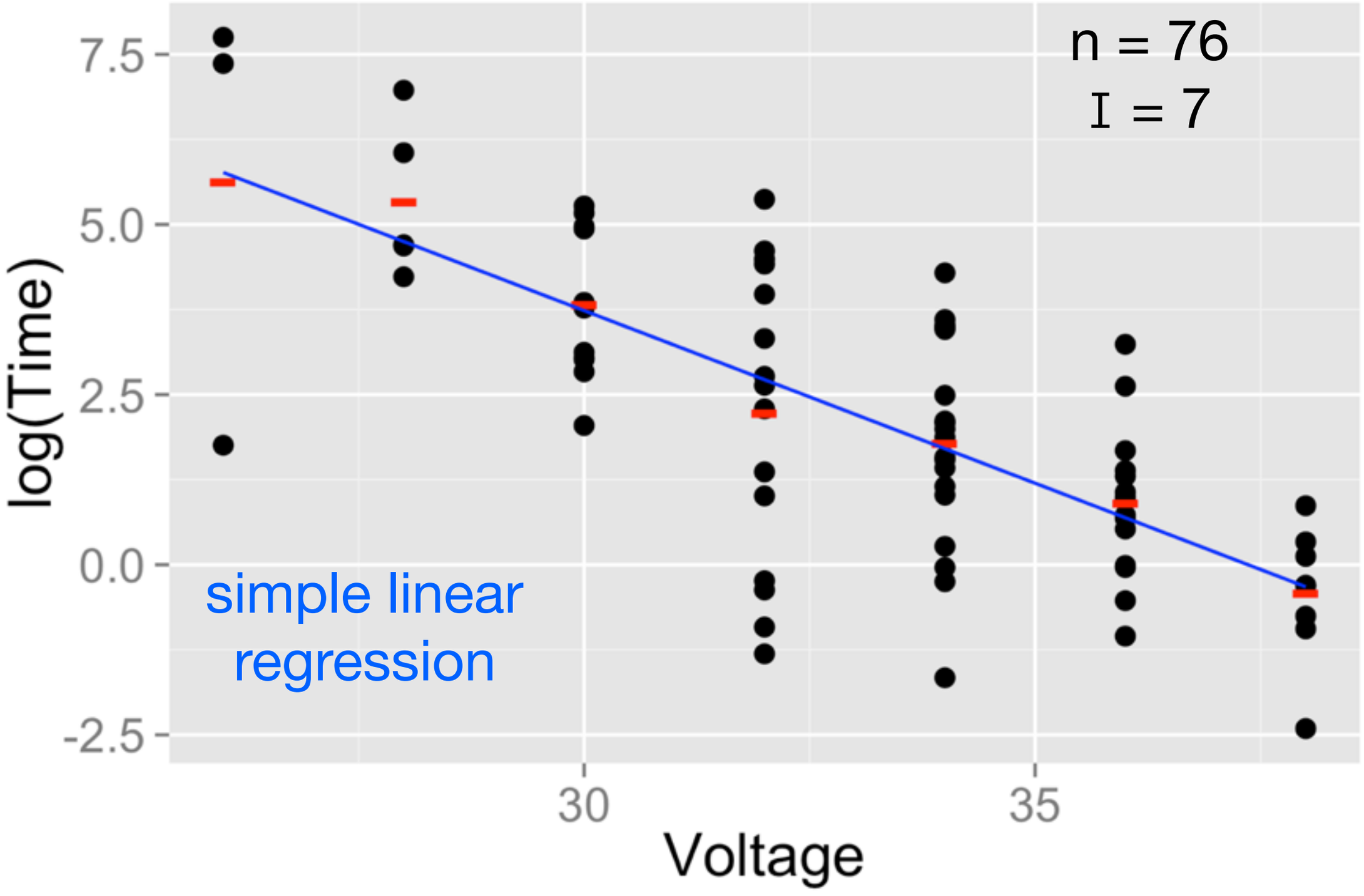
Insulating Fluid Case Study

Breakdown times for electrical insulating fluid at various voltages.



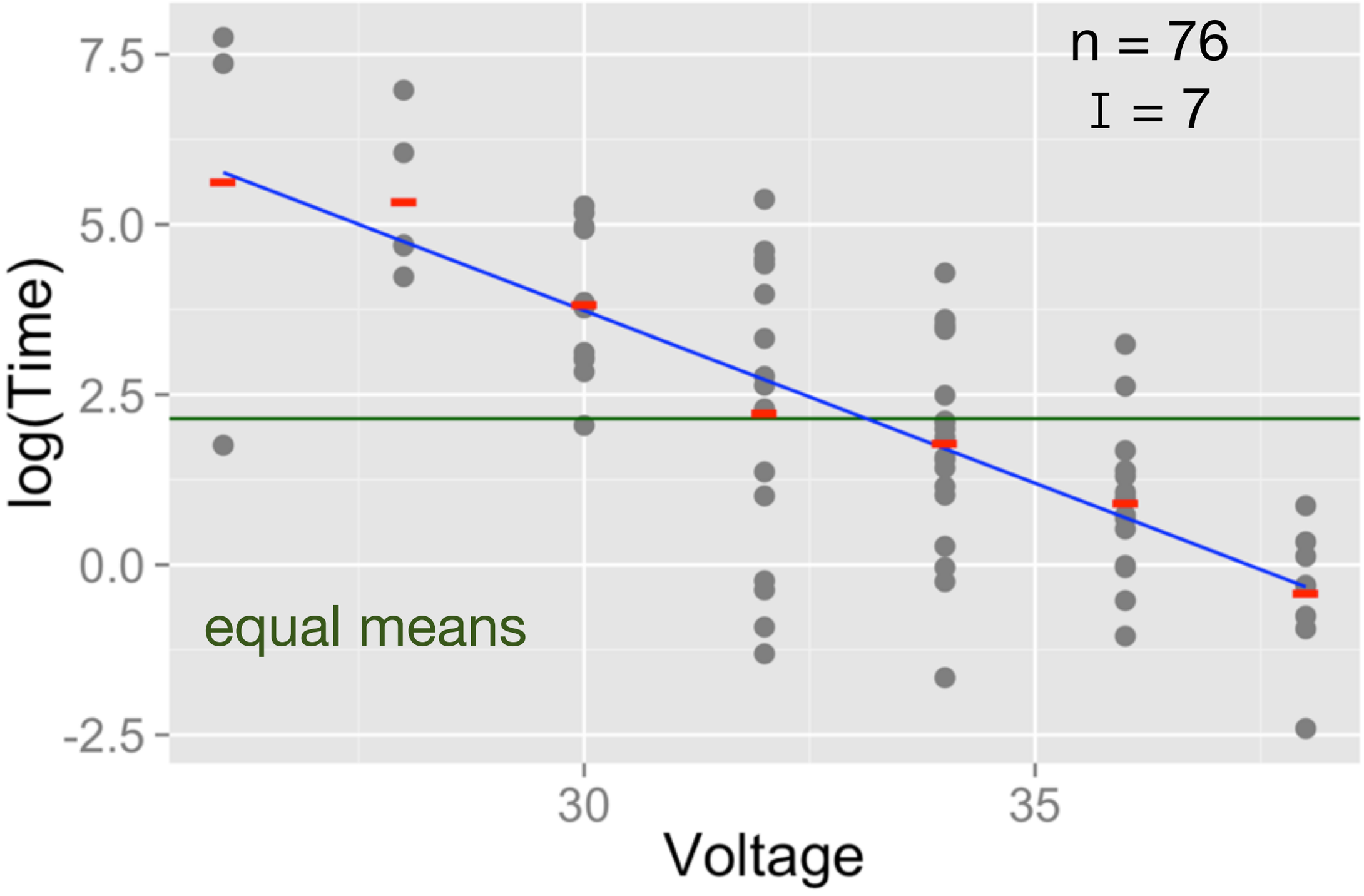
Insulating Fluid Case Study

Breakdown times for electrical insulating fluid at various voltages.



Insulating Fluid Case Study

Breakdown times for electrical insulating fluid at various voltages.



Comparing models

One way ANOVA

Equal means model

Regression model

Separate means model

Regression ANOVA

Lack of fit F-test

least complicated

most complicated

(only if there is more than one observation at each value of the explanatory)

All three comparisons are made with an Extra SS F-test

Review

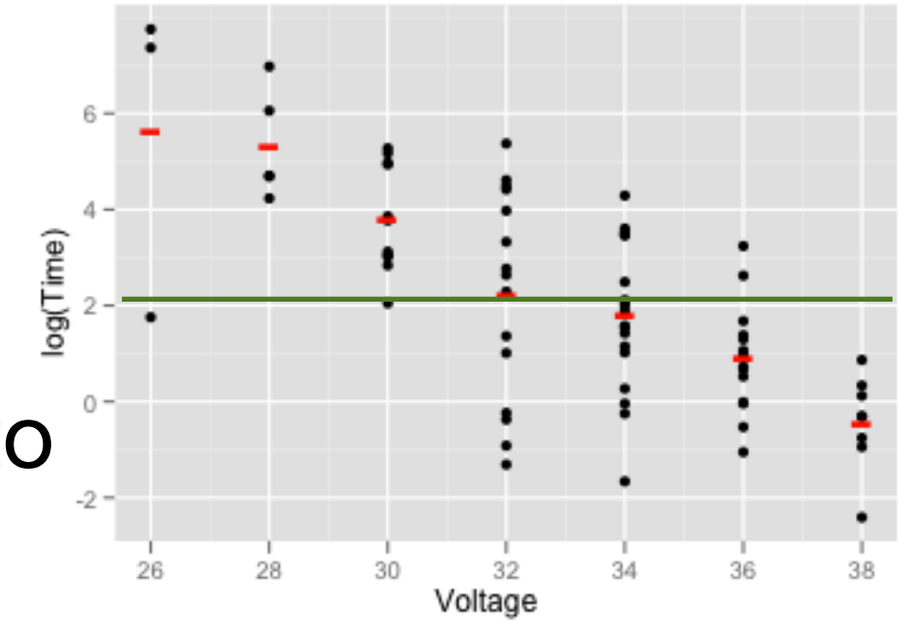
Extra SS F-test

Under the **null** hypothesis (reduced model is true) the F-statistic has an F-distribution with v_1 and v_2 degrees of freedom.

	Sum of squared residuals	d.f.	MSS	F	P-value
Extra	C: subtract A from B	F: subtract D from E v_1	G: divide C by F		I: divide G by H
Full model		v_2	H: divide A by D		
Reduced model					

One way ANOVA

Compares separate means model to equal means model



(B): ANALYSIS OF VARIANCE TABLE
FROM A ONE-WAY ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square	F-Statistic	p-value
Between Groups	196.4774	6	32.7462	13.00	<.0001
Within Groups	173.7484	69	2.5181		
Total	370.2258	75			

Equal means model

Residuals from separate means model

Residual sum of squares, separate-means model

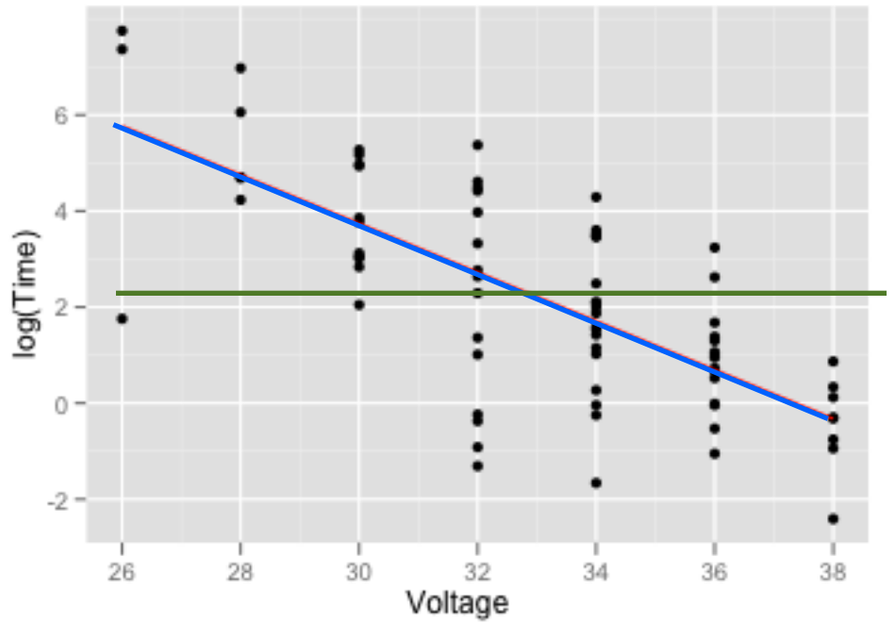
$\hat{\sigma}^2$ *in separate-means model*

compares separate-means and equal-means models

New!

Regression ANOVA

Compares regression model to equal means model



(A): ANALYSIS OF VARIANCE TABLE
FROM A SIMPLE LINEAR REGRESSION ANALYSIS

Source	Sum of Squares	df	Mean Square	F-Statistic	p-value
Regression	190.1514	1	190.1514	78.14	<.0001
Residual	180.0745	74	2.4334		
Total	370.2258	75			

Equal means model

Residuals from regression model

d.f. = n - 2

Residual sum of squares, regression model

$\hat{\sigma}^2$ *in regression model*

compares regression and equal-means models

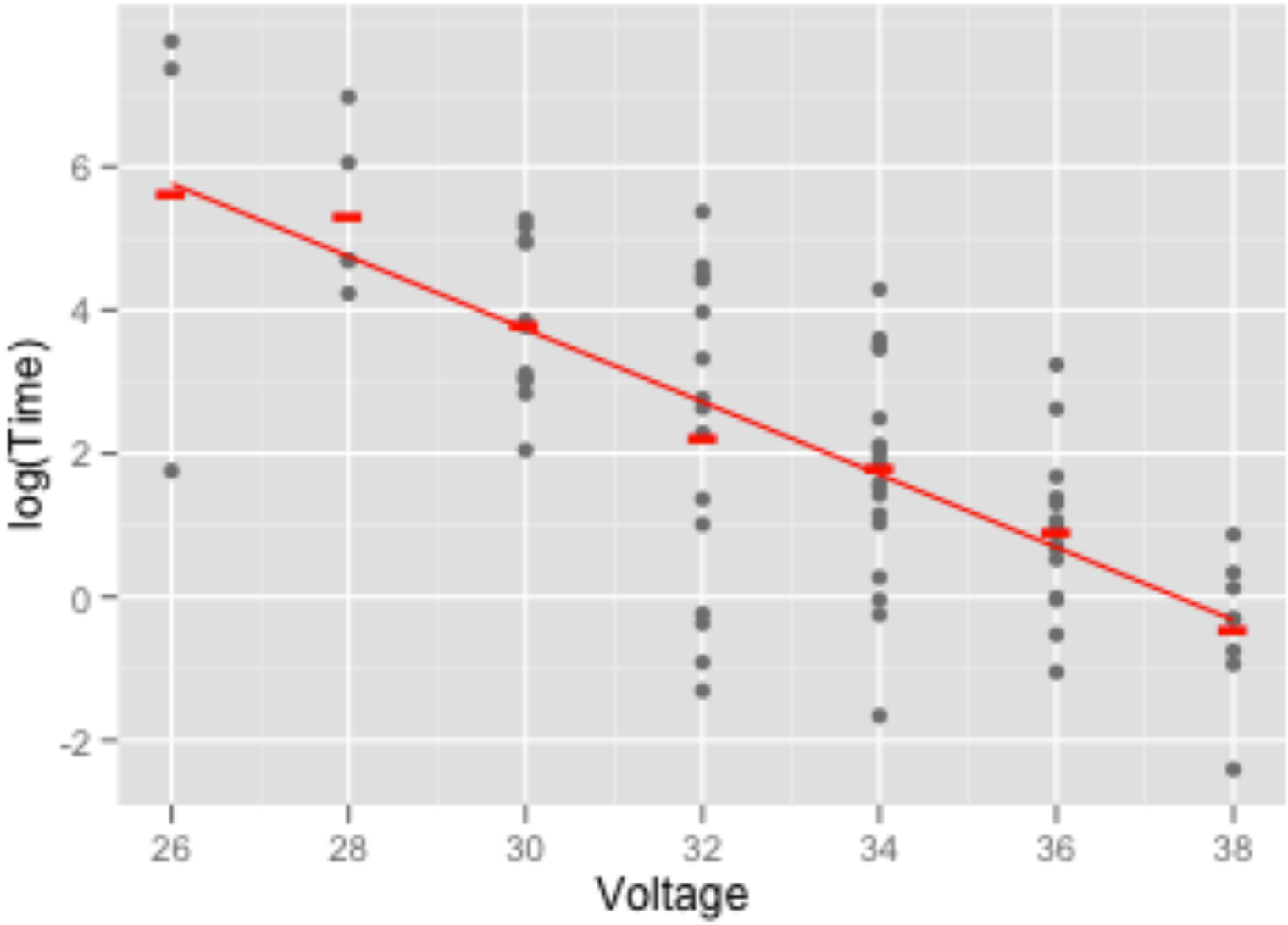
Equal means model: $\mu\{Y|X\} = \mu$

Regression model: $\mu\{Y|X\} = \beta_0 + \beta_1 X$

Saying the regression model doesn't fit any better than the equal means model, is the same as saying $\beta_1 = 0$.

The p-value in the regression ANOVA is the same as the p-value in the t-test that $\beta_1 = 0$.

Does the separate means model fit better than the regression model?



Does the decrease in residual sum of squares, justify the extra 5 parameters?

7 groups means versus 2 regression parameters

Lack of fit F-test

$$\text{F-statistic} = \frac{(RSS_{reg} - RSS_{separate\ means})/(df_{reg} - df_{separate\ means})}{\hat{\sigma}^2_{separate\ means}}$$

RSS = residual sum of squares

Compare to F-distribution with `I` - 2 and
n - `I` degrees of freedom

If **null is rejected**, **separate means model** is a better fit.
If we **fail to reject the null**, there is no evidence the separate means model fits any better than the **regression model**.

Your turn

Find the numbers needed to calculate the F-statistic.

(A): ANALYSIS OF VARIANCE TABLE FROM A SIMPLE LINEAR REGRESSION ANALYSIS

Source	Sum of Squares	df	Mean Square	F-Statistic	p-value
Regression	190.1514	1	190.1514	78.14	<.0001
Residual	180.0745	74	2.4334		
Total	370.2258	75			

Residual sum of squares, regression model

$\hat{\sigma}^2$ *in regression model*

compares regression and equal-means models

(B): ANALYSIS OF VARIANCE TABLE FROM A ONE-WAY ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square	F-Statistic	p-value
Between Groups	196.4774	6	32.7462	13.00	<.0001
Within Groups	173.7484	69	2.5181		
Total	370.2258	75			

Residual sum of squares, separate-means model

$\hat{\sigma}^2$ *in separate-means model*

compares separate-means and equal-means models

F-statistic =

$$\frac{(RSS_{reg} - RSS_{separate\ means})/(df_{reg} - df_{separate\ means})}{\hat{\sigma}^2_{separate\ means}}$$

Find the F-statistic

		Sum of squared residuals	d.f.	MSS	F	p-value
		C: subtract A from B	F: subtract D from E	G: divide C by F	I: divide G by H	use R: `1- pf(I, F, D)`
Full model	separate means model	A	D	H: divide A by D		
Reduced model	regression model	B	E			

There is evidence against the regression model (lack of fit F-test, p-value =).

Find the F-statistic

		Sum of squared residuals	d.f.	MSS	F	p-value
Extra		C: subtract A from B 180.0745-173.7484 = 6.33	F: subtract D from E 5	G: divide C by F 6.33/5 =1.2669	I: divide G by H 1.2669/2.5181 =0.5031	use R: 1- pf(I, F, D) 0.78
Full model	separate means model	A 173.7484	D 69	H: divide A by D 173.7484/69 = 2.5181		
Reduced model	regression model	B 180.0745	E 74			

There is **no** evidence against the regression model (lack of fit F-test, p-value = 0.78).

Another way to lay it out

Composite analysis of variance table with F-test for lack-of-fit

Source of Variation	Sum of Squares	df	Mean Square	F-Statistic	p-value
Between Groups	*196.4774*	*6*	*32.7462*	*13.00*	*<.0001*
Regression	190.1514	1	190.1514	75.51	<.0001
Lack of Fit	**6.3260**	**5**	**1.2652**	**0.50**	**.78**
Within Groups	*173.7484*	*69*	*2.5181*		
Total	370.2258	75			

by subtraction

LEGEND

Normal type items come from Regression Analysis (A)
Italicized items come from separate-means Analysis (B)
Bold face items are new and calculated here

```
sep_means <- lm(log(Time) ~ Group, data = case0802)
reg_fit <- lm(log(Time) ~ Voltage, data = case0802)
anova(reg_fit, sep_means)
```

```
Analysis of Variance Table

Model 1: log(Time) ~ Voltage
Model 2: log(Time) ~ Group - 1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     74 180.07
2     69 173.75  5    6.3259 0.5024 0.7734
```

R-squared

$$R^2 = \frac{(\text{Total sum of squares} - \text{RSS})}{\text{Total sum of squares}}$$

Measures the proportion of variation in the response that is explained by the regression model for the mean.

Is between 0 and 1.

Slope = 0

Perfect fit

```
> summary(reg_fit)

Call:
lm(formula = log(Time) ~ Voltage, data = case0802)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0291 -0.6919  0.0366  1.2094  2.6513

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.9555     1.9100   9.924 3.05e-15 ***
Voltage      -0.5074     0.0574  -8.840 3.34e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.56 on 74 degrees of freedom
Multiple R-squared: 0.5136,    Adjusted R-squared: 0.507
F-statistic: 78.14 on 1 and 74 DF,  p-value: 3.34e-13
```

Correlation

The sample correlation coefficient describes the degree of linear association between two variables. (see formula in Sleuth 7.5.4) or `cor` in R

```
with(case0802, cor(log(Time), Voltage))
```

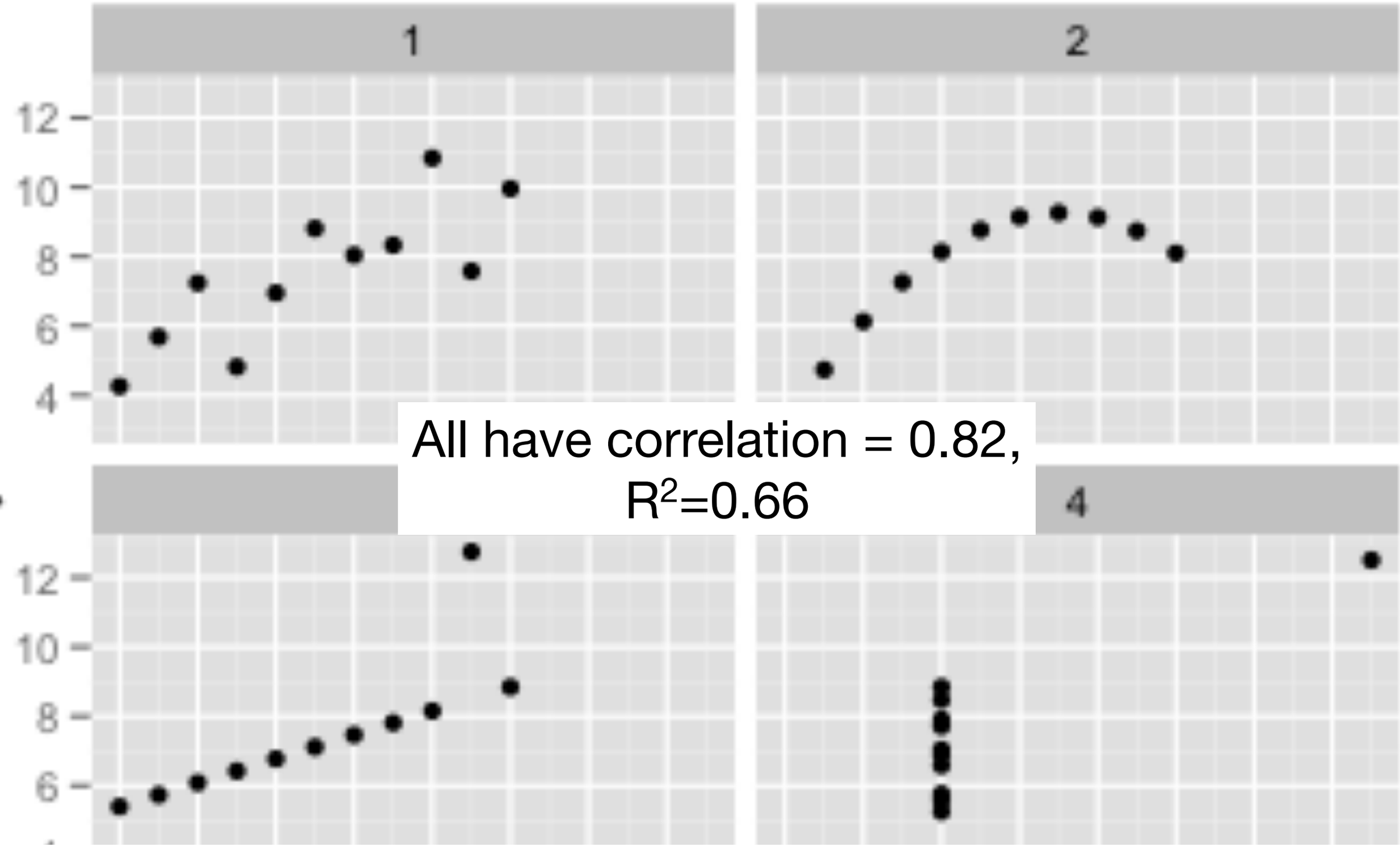
```
[1] -0.716665
```

In simple linear regression R^2 is the sample correlation squared, `(-0.716665)^2 = 0.5136087`

Correlation

Inference using correlation only makes sense if the data are pairs drawn from a population.

Simple linear regression doesn't make this assumption so don't use the correlation or R^2 for inference.


1
2
4
12
10
8
6
4
12
10
8
6
All have correlation = 0.82,
R^2=0.66
A high R^2 **does not** mean that
simple linear regression is appropriate
x

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x             0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6665, Adjusted R-squared: 0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576 *
x              0.500      0.118   4.239  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6662, Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0025     1.1245   2.670  0.02562 *
x             0.4997     0.1179   4.239  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6663, Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671  0.02559 *
x             0.4999     0.1178   4.243  0.00216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6667, Adjusted R-squared: 0.6297
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

The only way to tell if linear regression is appropriate is to examine the data (or residuals)