STATISTICAL METHODS IN R

Lubor Homolka

September 11th, 2018

Standard Inferential Tests

Regression Analysis

Time-Series Modelling

```
str(mtcars)
```

```
## 'data.frame': 32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 .
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
mtcars$mpg
```

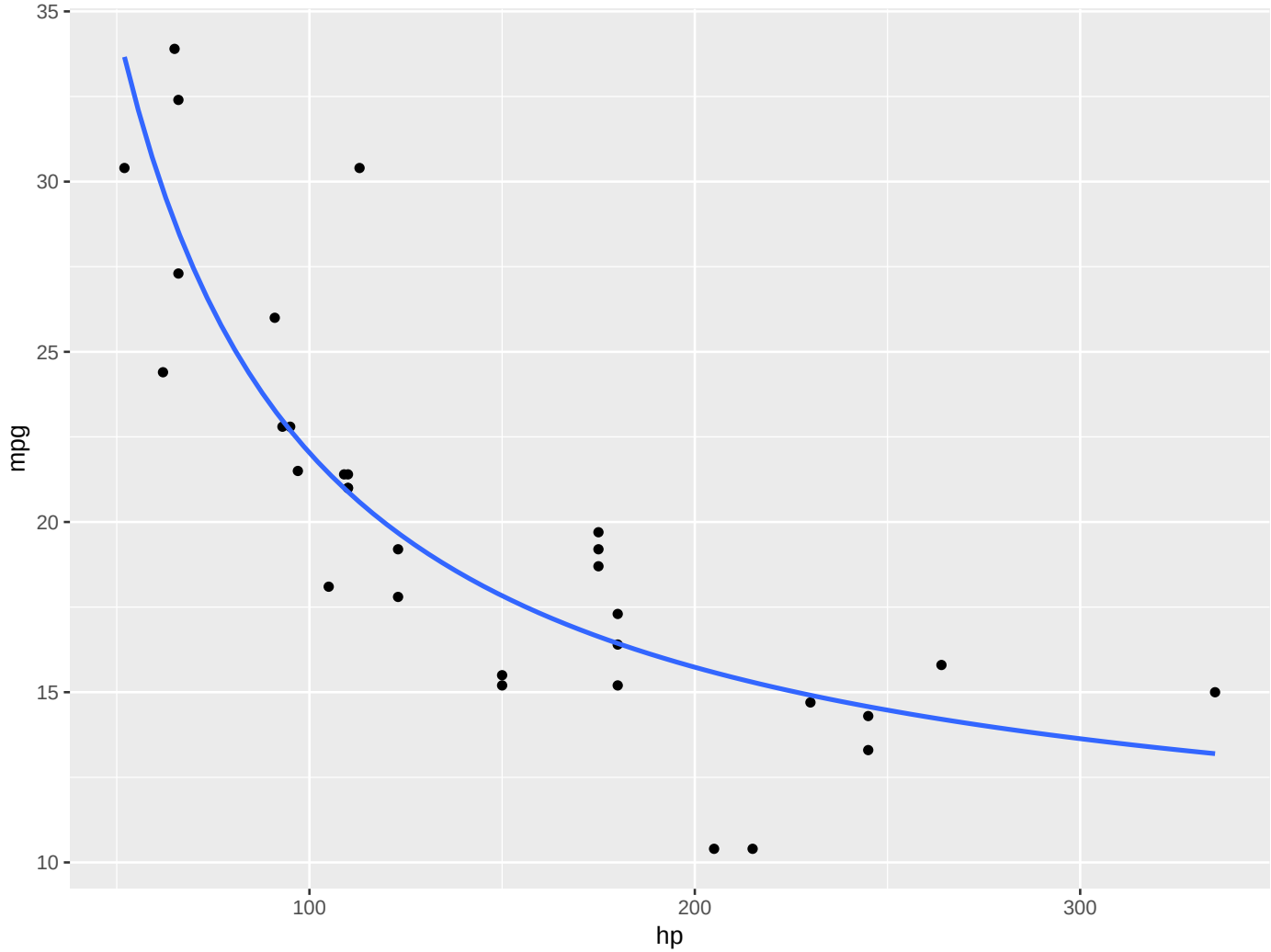
```
##  [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 1
## [15] 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 2
## [29] 15.8 19.7 15.0 21.4
```

Qualitative & Quantitative Research

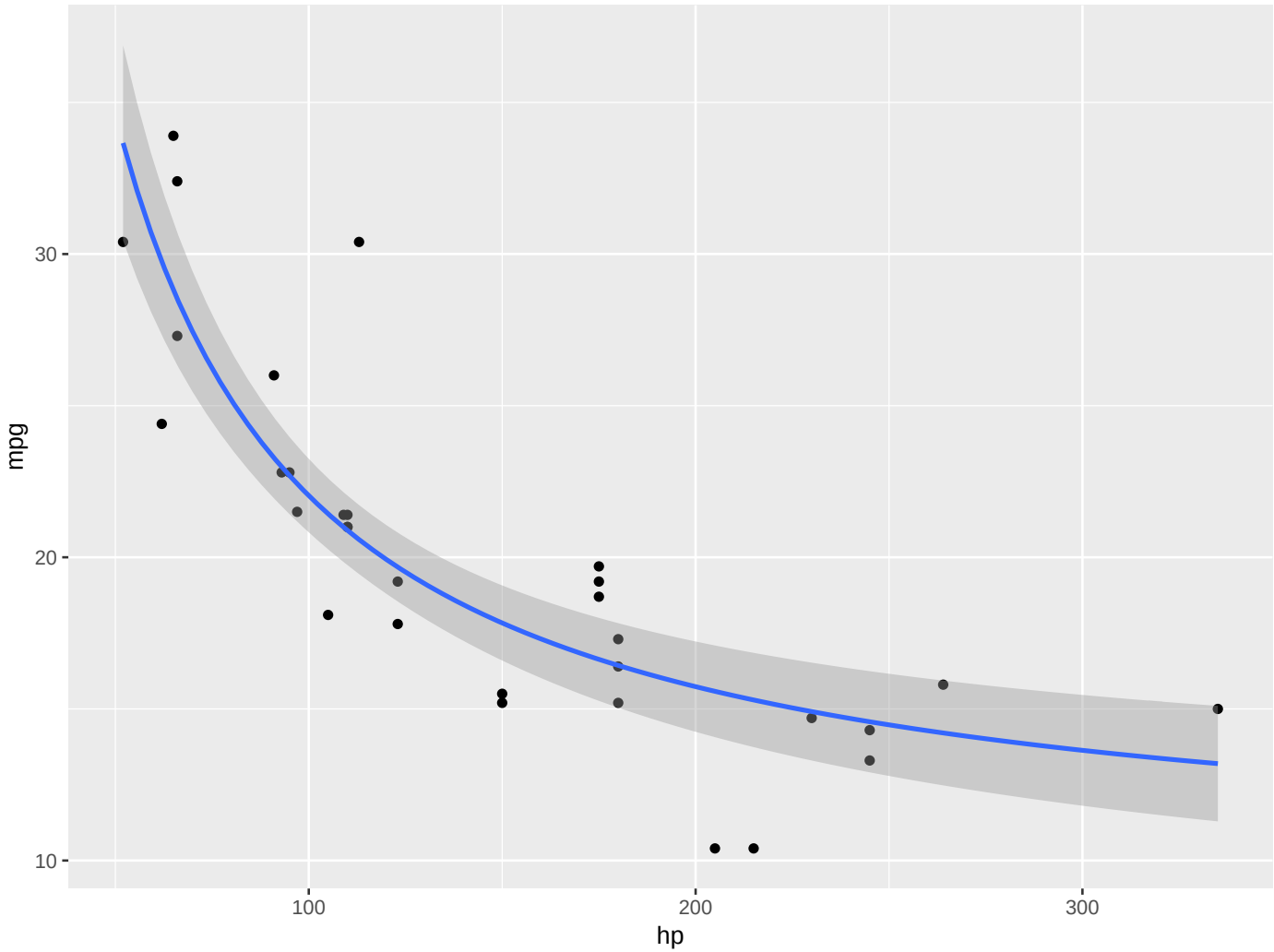
What's the difference?

(Statistical view)

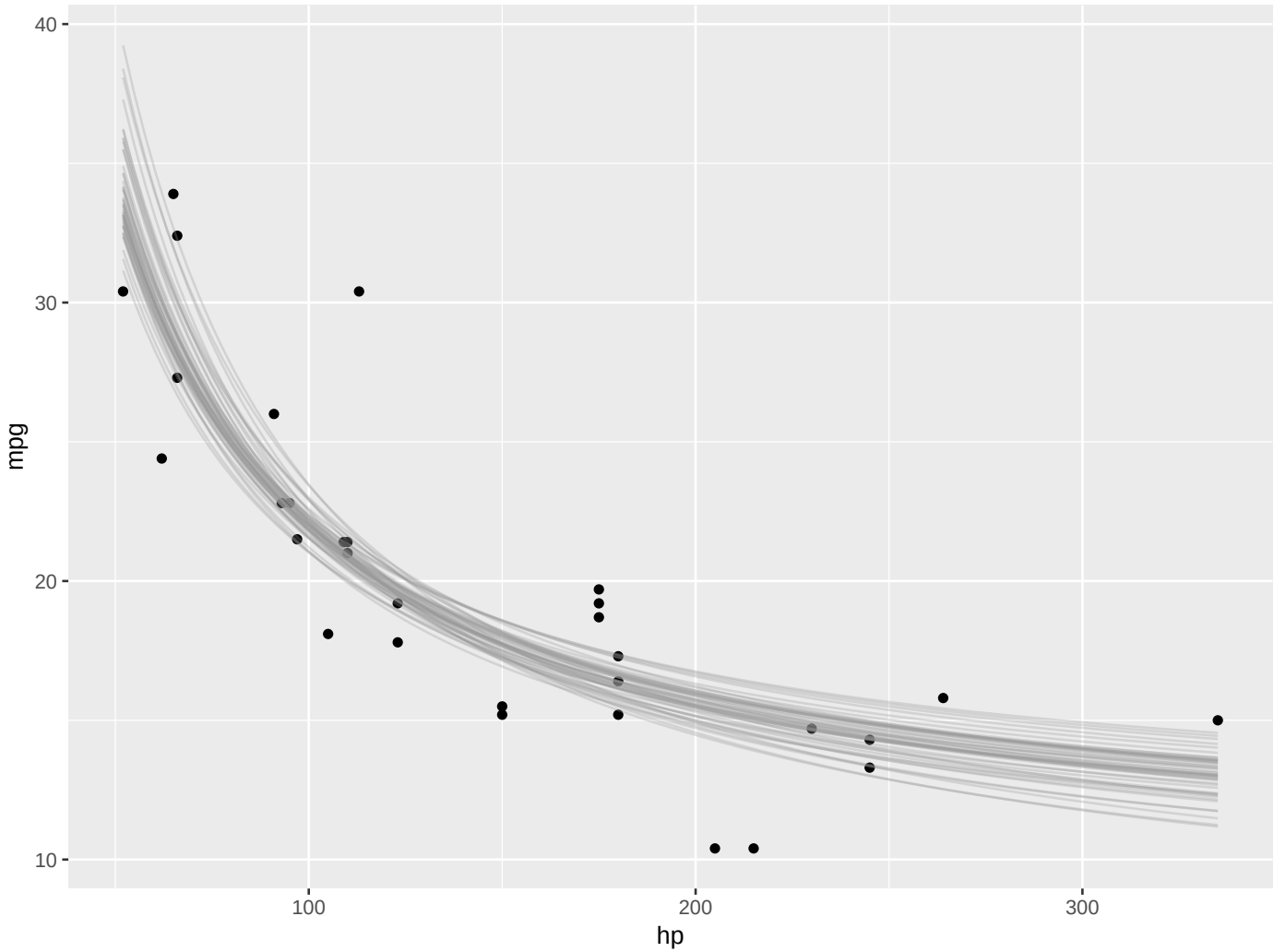
What qualitative researcher see...



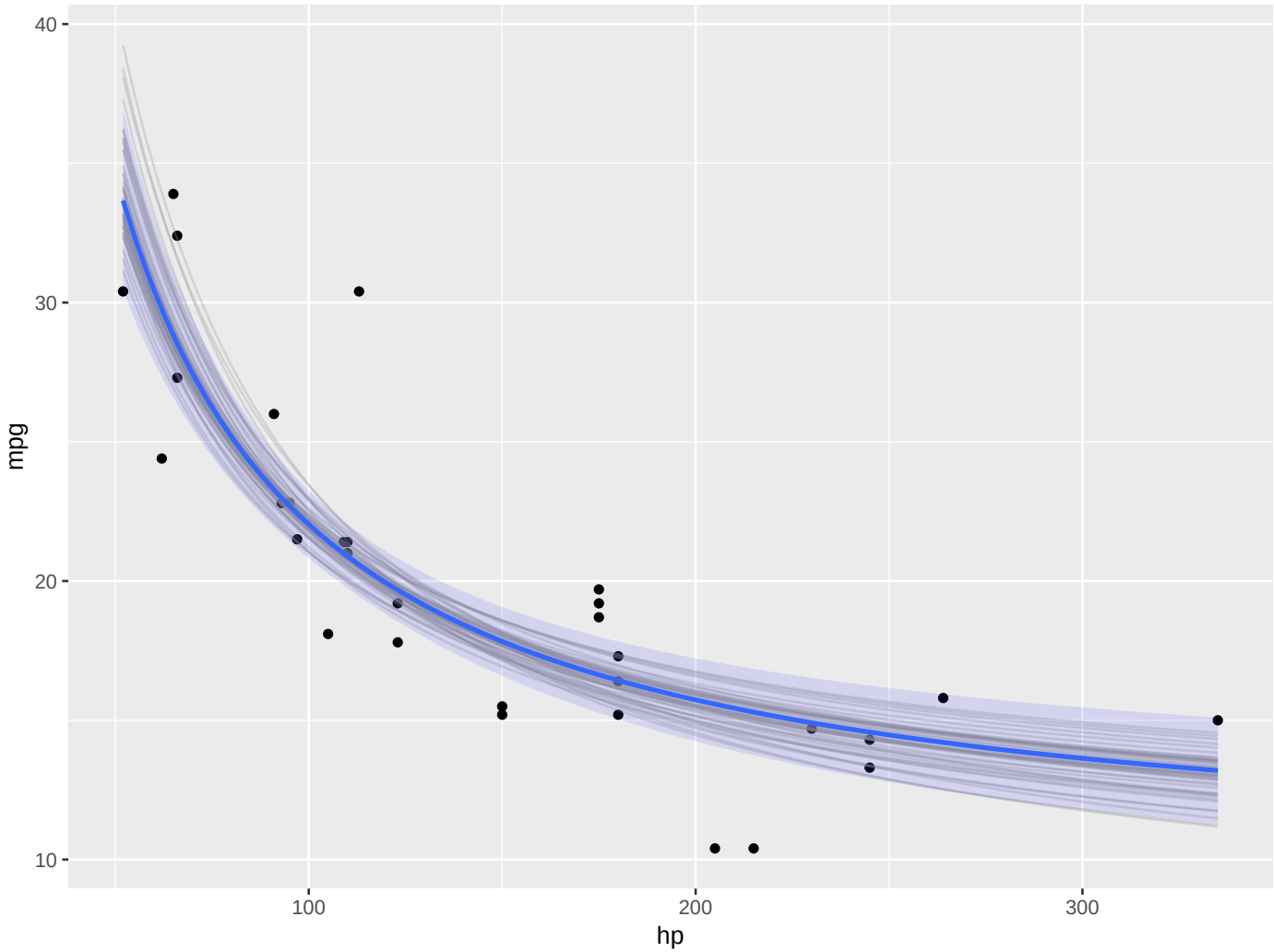
What Quant See..



If All Replications are observed



Do we need all the replications?



What is Statistical Inference?

Statistical inferrence is a process of estimating a population characteristic based on the sample statistics.

Sample	Population
$\bar{x}$	μ
$r_{x_1 x_2}$	$\rho_{x_1 x_2}$
$y = b_0 + b_1 x$	$y = \beta_0 + \beta_1 x$

Tools of Statistical Inference

- Point estimate
- Interval estimate $\rightarrow$ confidence intervals
- Null hypothesis testing (NHST)

Tools of Statistical Inference

- Point estimate
- Interval estimate $\rightarrow$ confidence intervals
- Null hypothesis testing (NHST)

NHST: We try to falsify statement H_0 by finding a counter-proof stated in the H_A.

Selected Test

- Test of Equal or Given Proportions
- Pearson's χ^2 Test for Count Data
- Pearson & Spearman correlation
- t-test

Test of Equal or Given Proportions

This test tests whether the characteristics of the population π can take some values.

Example: You want to prove that the proportion of companies which are ready to adopt a new technology is smaller than 50%.

Test of Equal or Given Proportions

This test tests whether the characteristics of the population π can take some values.

Example: You want to prove that the proportion of companies which are ready to adopt a new technology is smaller than 50%.

$H_0 : \pi = 0.5$ is what you want to disprove.

$H_A : \pi < 0.5$ is what you want to prove.

Test of Equal or Given Proportions

This test tests whether the characteristics of the population π can take some values.

Example: You want to prove that the proportion of companies which are ready to adopt a new technology is smaller than 50%.

$H_0 : \pi = 0.5$ is what you want to disprove.

$H_A : \pi < 0.5$ is what you want to prove.

You have asked 100 companies, 45 are ready.

What do all these number mean?

```
prop.test(x=45, n=100, alternative="less", p=0.5)

##
##  1-sample proportions test with continuity correction
##
## data:  45 out of 100, null probability 0.5
## X-squared = 0.81, df = 1, p-value = 0.1841
## alternative hypothesis: true p is less than 0.5
## 95 percent confidence interval:
##  0.000000 0.537017
## sample estimates:
##    p
## 0.45
```

Let's start with p-value

$$\text{p-value} = P(\text{Data}|H_0 = \text{TRUE})$$

Let's start with p-value

$$\text{p-value} = P(\text{Data}|H_0 = \text{TRUE})$$

which is not

$$P(H_0 = \text{TRUE}|\text{Data})$$

which is, what we want!

Let's start with p-value

$$\text{p-value} = P(\text{Data}|H_0 = \text{TRUE})$$

which is not

$$P(H_0 = \text{TRUE}|\text{Data})$$

which is, what we want!

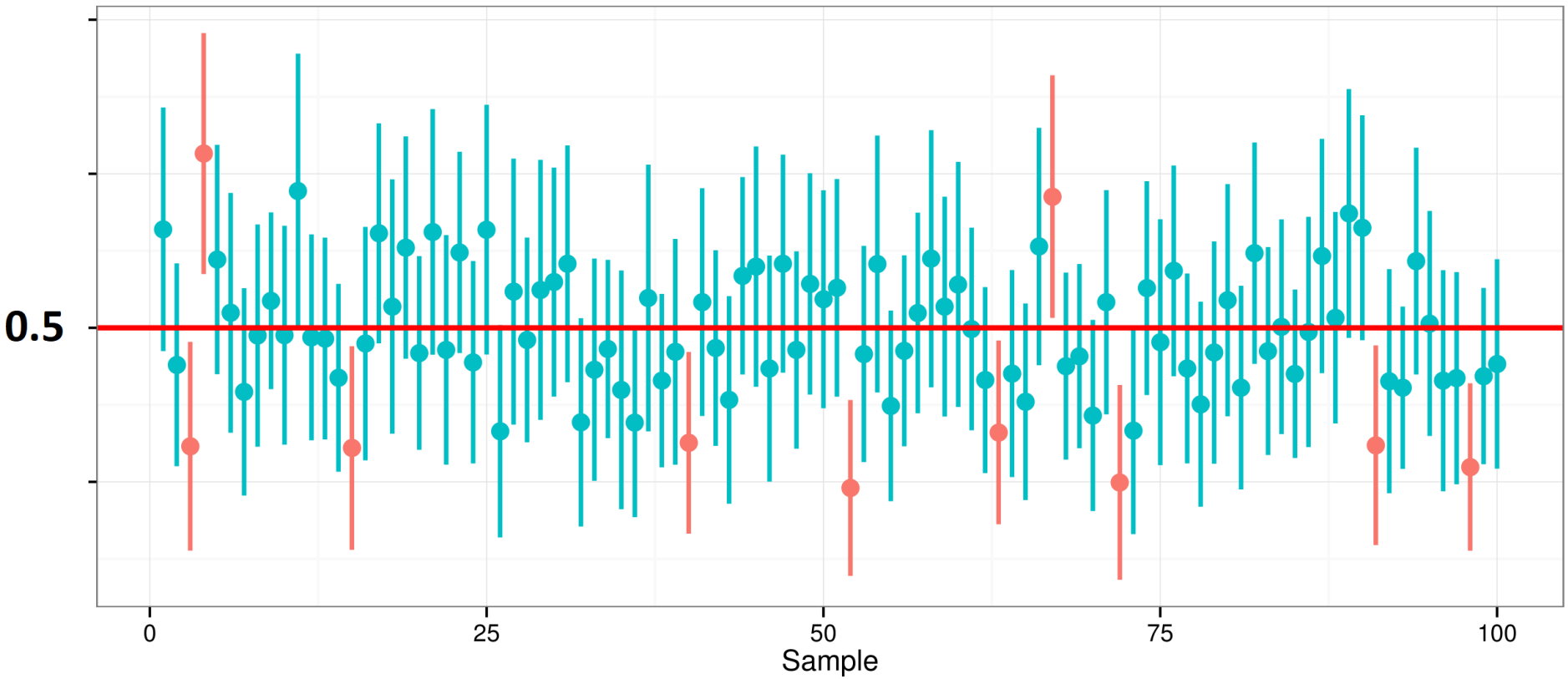
Data: possible results which would contradict H_0 and favour H_A more than the observed result.

Consider this Example with $H_A : \pi \neq 0.5$

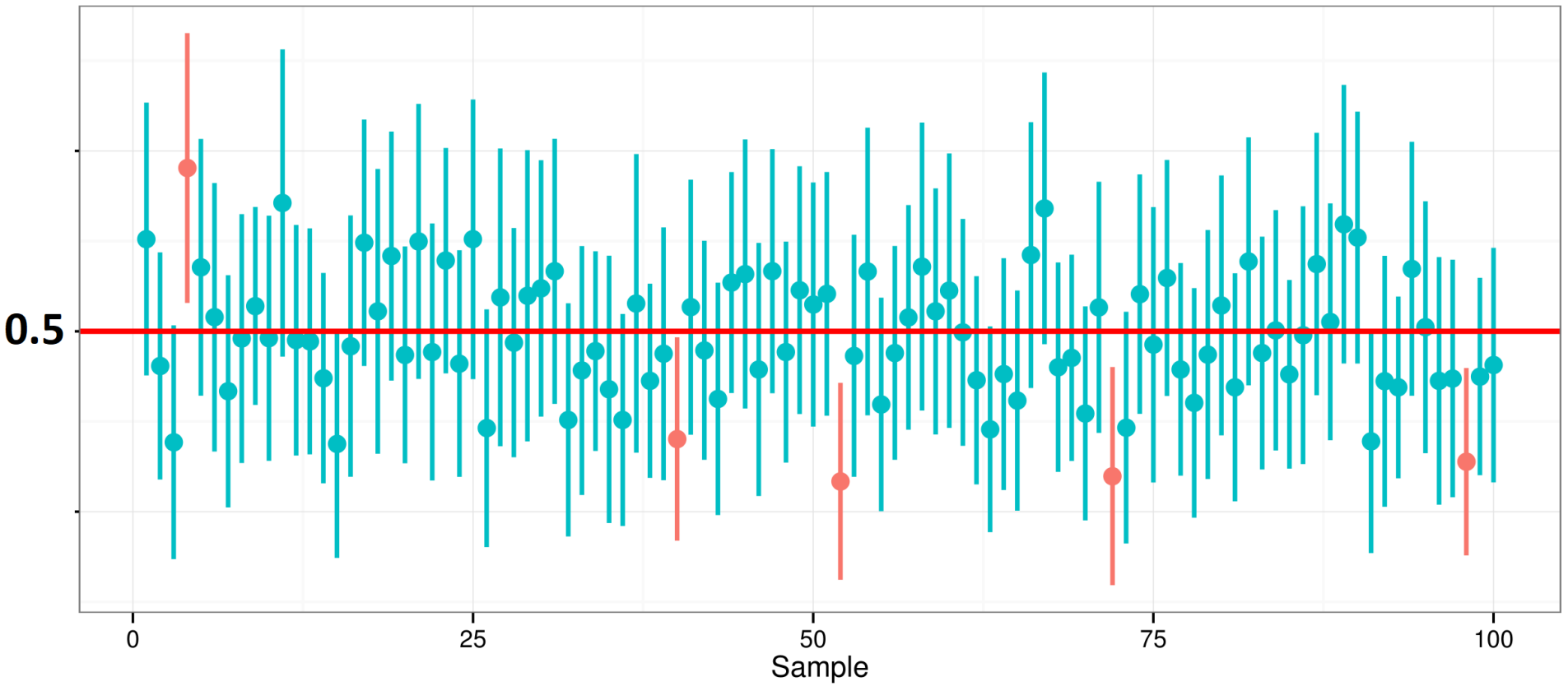
```
prop.test(x=45, n=100, alternative="two.sided", p=0.5)

##
##  1-sample proportions test with continuity correction
##
## data:  45 out of 100, null probability 0.5
## X-squared = 0.81, df = 1, p-value = 0.3681
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3514281 0.5524574
## sample estimates:
##    p
## 0.45
```

Let's focus on Confidence intervals

Confidence Interval with $\alpha = 0.9$



Confidence Interval with $\alpha = 0.95$



Test of Equal or Given Proportions

```
prop.test(x=45, n=100, alternative="less", p=0.5)

##
##  1-sample proportions test with continuity correction
##
## data:  45 out of 100, null probability 0.5
## X-squared = 0.81, df = 1, p-value = 0.1841
## alternative hypothesis: true p is less than 0.5
## 95 percent confidence interval:
##  0.000000 0.537017
## sample estimates:
##    p
## 0.45
```

Test of Equal or Given Proportions

```
prop.test(x=45, n=100, alternative="less", p=0.5)

##
##  1-sample proportions test with continuity correction
##
## data:  45 out of 100, null probability 0.5
## X-squared = 0.81, df = 1, p-value = 0.1841
## alternative hypothesis: true p is less than 0.5
## 95 percent confidence interval:
##  0.000000 0.537017
## sample estimates:
##    p
## 0.45
```

Don't forget that no evidence for $\pi < 0.5$ does not imply evidence for $\pi = 0.5$.

Test of Equal or Given Proportions

```
prop.test(x=450, n=1000, alternative="less", p=0.5)

##
##  1-sample proportions test with continuity correction
##
## data:  450 out of 1000, null probability 0.5
## X-squared = 9.801, df = 1, p-value = 0.0008721
## alternative hypothesis: true p is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.4764786
## sample estimates:
##    p
## 0.45
```

Test of Equal or Given Proportions

```
prop.test(x=40, n=100, alternative="less", p=0.5)

##
##  1-sample proportions test with continuity correction
##
## data:  40 out of 100, null probability 0.5
## X-squared = 3.61, df = 1, p-value = 0.02872
## alternative hypothesis: true p is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.4872158
## sample estimates:
##   p
## 0.4
```

Test of Equal or Given Proportions

This test tests whether k population π differ .

Example: You want to prove that the proportion of companies which are ready to adopt a new technology is smaller when the company is fully owned by Czech owners π_1, than proportion of international ready-companies π_2.

Test of Equal or Given Proportions

This test tests whether k population π differ .

Example: You want to prove that the proportion of companies which are ready to adopt a new technology is smaller when the company is fully owned by Czech owners π_1, than proportion of international ready-companies π_2.

$H_0 : \pi_1 = \pi_2$ is what you want to disprove.

$H_A : \pi_1 < \pi_2$ is what you want to prove.

Test of Equal or Given Proportions

This test tests whether k population π differ .

Example: You want to prove that the proportion of companies which are ready to adopt a new technology is smaller when the company is fully owned by Czech owners π_1, than proportion of international ready-companies π_2.

$H_0 : \pi_1 = \pi_2$ is what you want to disprove.

$H_A : \pi_1 < \pi_2$ is what you want to prove.

You have asked 100 CZ and 100 Int companies. 30CZ and 40Int companies are ready.

Test of Equal or Given Proportions

```
prop.test(x=c(30, 40), n=c(100, 100), alternative="less")

##
##  2-sample test for equality of proportions with continu
##  correction
##
## data:  c(30, 40) out of c(100, 100)
## X-squared = 1.7802, df = 1, p-value = 0.09106
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000  0.02034014
## sample estimates:
## prop 1 prop 2
##    0.3    0.4
```

Pearson's χ^2 Test for Count Data

This test tests associatio between two non-numeric variables.

Example: Example from Belas's Financial Markets book. Authors were intersted wheter Czech and Slovakian e-banking services are equally satisfied with the services.

Pearson's χ^2 Test for Count Data

This test tests associatio between two non-numeric variables.

Example: Example from Belas's Financial Markets book. Authors were intersted wheter Czech and Slovakian e-banking services are equally satisfied with the services.

H_0 : no assoc. between nationality and satisfaction

H_A : there is some assoc: this is what you want to prove.

Pearson's χ^2 Test for Count Data

This test tests associatio between two non-numeric variables.

Example: Example from Belas's Financial Markets book. Authors were intersted wheter Czech and Slovakian e-banking services are equally satisfied with the services.

H_0 : no assoc. between nationality and satisfaction

H_A : there is some assoc: this is what you want to prove.

```
head(bankingData)
```

```
##     country  satis
## 58       CZ    Yes
## 157      SK    Yes
## 81       CZ     No
## 174      SK     No
## 185      SK justOK
## 9        CZ    Yes
```

```
bankingData %>% table
```

```
##         satis
## country justOK No Yes
##      CZ     12 26  62
##      SK     16 23  61
```

```r
tab <- bankingData %>% table %>% t
chisq.test(tab)

##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 0.76323, df = 2, p-value = 0.6828
```

Conclusion: There is no evidence for statistically significant associaton.

Let's adjust the data to find association

```
tab2 <- matrix(c(12,35,40,16,23,61), ncol=3,
               byrow = TRUE)
tab2

##      [,1] [,2] [,3]
## [1,]   12   35   40
## [2,]   16   23   61

prop.table(tab2, 1) #1 means row margin

##           [,1]      [,2]      [,3]
## [1,] 0.137931 0.4022989 0.4597701
## [2,] 0.160000 0.2300000 0.6100000
```

```
chisq.test(tab2)

##
##  Pearson's Chi-squared test
##
## data:  tab2
## X-squared = 6.5484, df = 2, p-value = 0.03785
```

Conclusion: We have an evidence for statistically significant associaton.

```
chisq.test(tab2)

##
##  Pearson's Chi-squared test
##
## data:  tab2
## X-squared = 6.5484, df = 2, p-value = 0.03785
```

Conclusion: We have an evidence for statistically significant associaton. What caused the association? We need to inspect results:

```r
chTest <- chisq.test(tab2)
str(chTest)
```

```
## List of 9
##  $ statistic: Named num 6.55
##   ..- attr(*, "names")= chr "X-squared"
##  $ parameter: Named int 2
##   ..- attr(*, "names")= chr "df"
##  $ p.value  : num 0.0378
##  $ method   : chr "Pearson's Chi-squared test"
##  $ data.name: chr "tab2"
##  $ observed : num [1:2, 1:3] 12 16 35 23 40 61
##  $ expected : num [1:2, 1:3] 13 15 27 31 47 ...
##  $ residuals: num [1:2, 1:3] -0.284 0.265 1.543 -1.439 -
##  $ stdres   : num [1:2, 1:3] -0.422 0.422 2.541 -2.541 -
##  - attr(*, "class")= chr "htest"
```

Residuals is the key factor. It tells us what goes against our expectation stated in the H_0.

```
chTest$residuals

##             [,1]      [,2]       [,3]
## [1,] -0.2844735  1.543147 -1.0196109
## [2,]  0.2653392 -1.439351  0.9510297

residChi <- chTest$residuals %>% data.frame()
colnames(residChi) <- c("justOK", "No", "Yes")
rownames(residChi) <- c("CZ", "SK")
residChi

##        justOK        No        Yes
## CZ -0.2844735  1.543147 -1.0196109
## SK  0.2653392 -1.439351  0.9510297
```

Correlation Tests

Test whether there is a relationship between two independent variables.

Example: You have measured IQ of 50 respondents and their reaction time in seconds. You don't expect direct relations between variables. But you expect there is some degree of correlation.

Correlation Tests

Test whether there is a relationship between two independent variables.

Example: You have measured IQ of 50 respondents and their reaction time in seconds. You don't expect direct relations between variables. But you expect there is some degree of correlation.

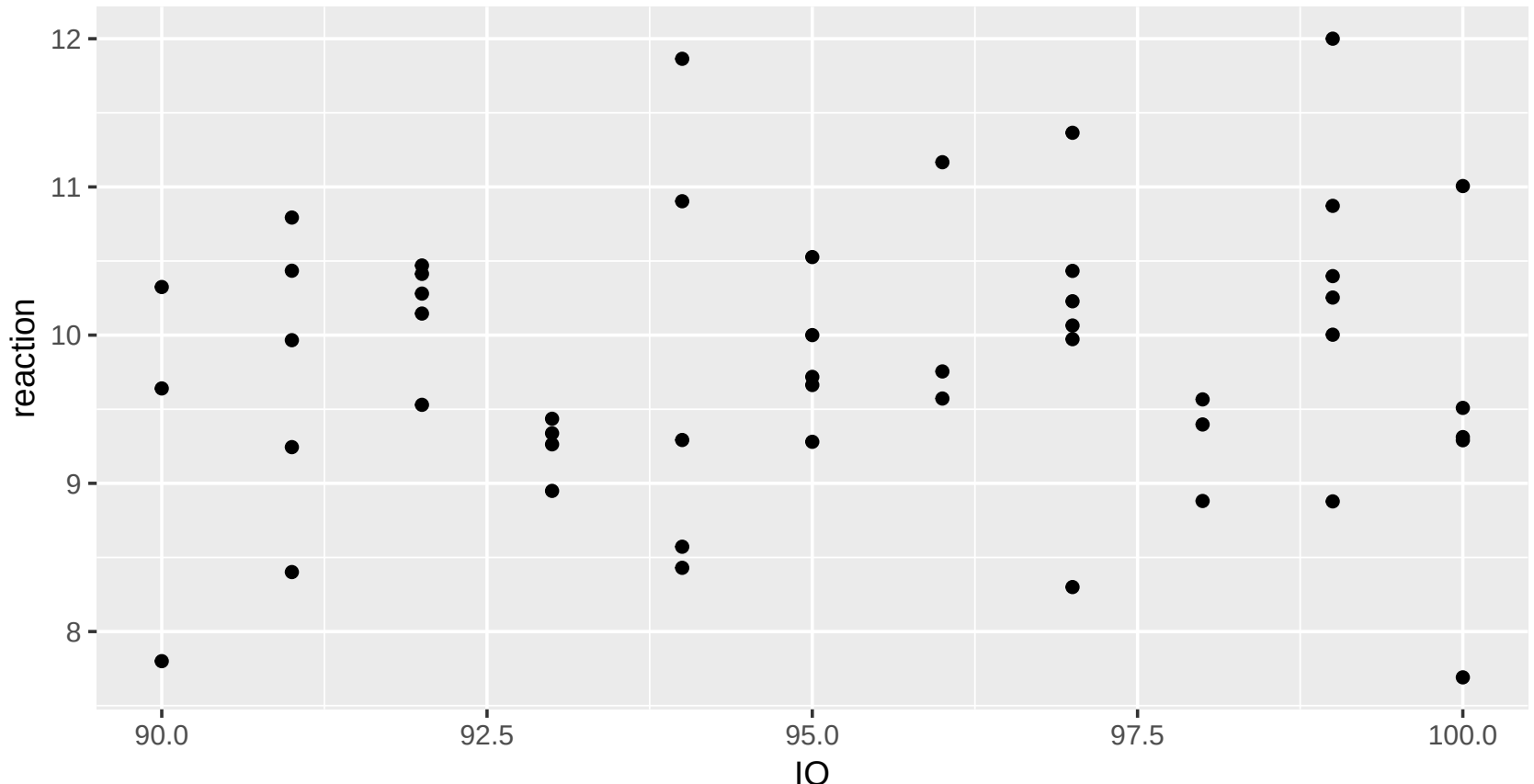
$H_0 : \rho = 0$ no correlation between variables

$H_A : \rho \neq 0$

```r
head(dataCor, 3)
```

```
##   IQ  reaction
## 1 90  7.800140
## 2 90 10.324620
## 3 90  9.640206
```

```r
ggplot(dataCor, aes(x=IQ, y=reaction)) + geom_point()
```



```r
cor(dataCor)
```

```
##                    IQ   reaction
## IQ         1.00000000 0.07261689
## reaction 0.07261689 1.00000000
```

```r
cor.test(dataCor$IQ, dataCor$reaction)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dataCor$IQ and dataCor$reaction
## t = 0.50444, df = 48, p-value = 0.6163
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  -0.2099750  0.3440112
## sample estimates:
##       cor
## 0.07261689
```

Non-parametric Correlation Coefficient

```
cor.test(dataCor$IQ, dataCor$reaction, method="spearman")

## Warning in cor.test.default(dataCor$IQ,
dataCor$reaction, method = "spearman"):  Cannot
compute exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  dataCor$IQ and dataCor$reaction
## S = 20034, p-value = 0.7935
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.03796654
```

t-test & Welch t-test (safer option)

Test whether there is a difference in mean values of 2 groups.

Example: Czech and International companies are competing at the same market. Do Int companies perform better than Czech companies (Measured by ROA)?

t-test & Welch t-test (safer option)

Test whether there is a difference in mean values of 2 groups.

Example: Czech and International companies are competing at the same market. Do Int companies perform better than Czech companies (Measured by ROA)?

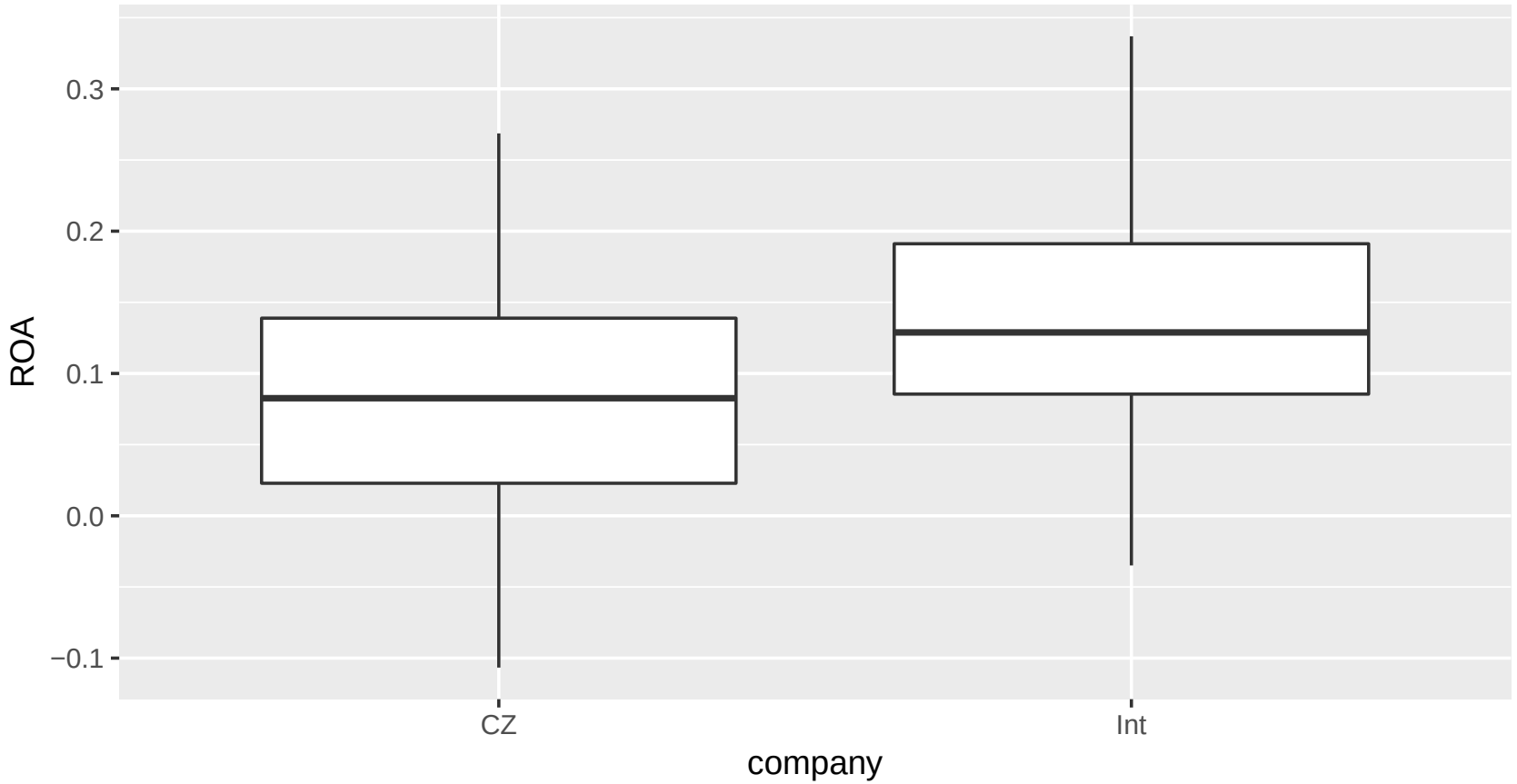
$H_0 : \mu_{CZ} = \mu_{Int}$: μ_{CZ} is a true mean value of Czech companies' ROA

$H_A : \mu_{CZ} < \mu_{Int}$

Create Data for t-test

```r
set.seed(123) # for reproducibility
company <- c(rep("CZ", 30), rep("Int", 40) )
ROA <- c(
 rnorm(30, mean=0.09, sd=0.1),
 rnorm(40, mean=0.12, sd=0.1) )

dataFin <- data.frame(company, ROA)
summary(dataFin)

##  company        ROA
##  CZ :30   Min.   :-0.10666
##  Int:40   1st Qu.: 0.04671
##           Median : 0.11274
##           Mean   : 0.11453
##           3rd Qu.: 0.17499
##           Max.   : 0.33690
```

```
ggplot(dataFin, aes(x=company, y=ROA) ) +
  geom_boxplot()
```



Welch t-test: μ of first group is **less** than second

```
t.test(dataFin$ROA ~ dataFin$company, alternative="less")

##
##  Welch Two Sample t-test
##
## data:  dataFin$ROA by dataFin$company
## t = -2.2853, df = 57.308, p-value = 0.01301
## alternative hypothesis: true difference in means is less than
## 95 percent confidence interval:
##         -Inf -0.01373392
## sample estimates:
##  mean in group CZ mean in group Int
##        0.08528962        0.13645247
```

When we just explore, no expectation

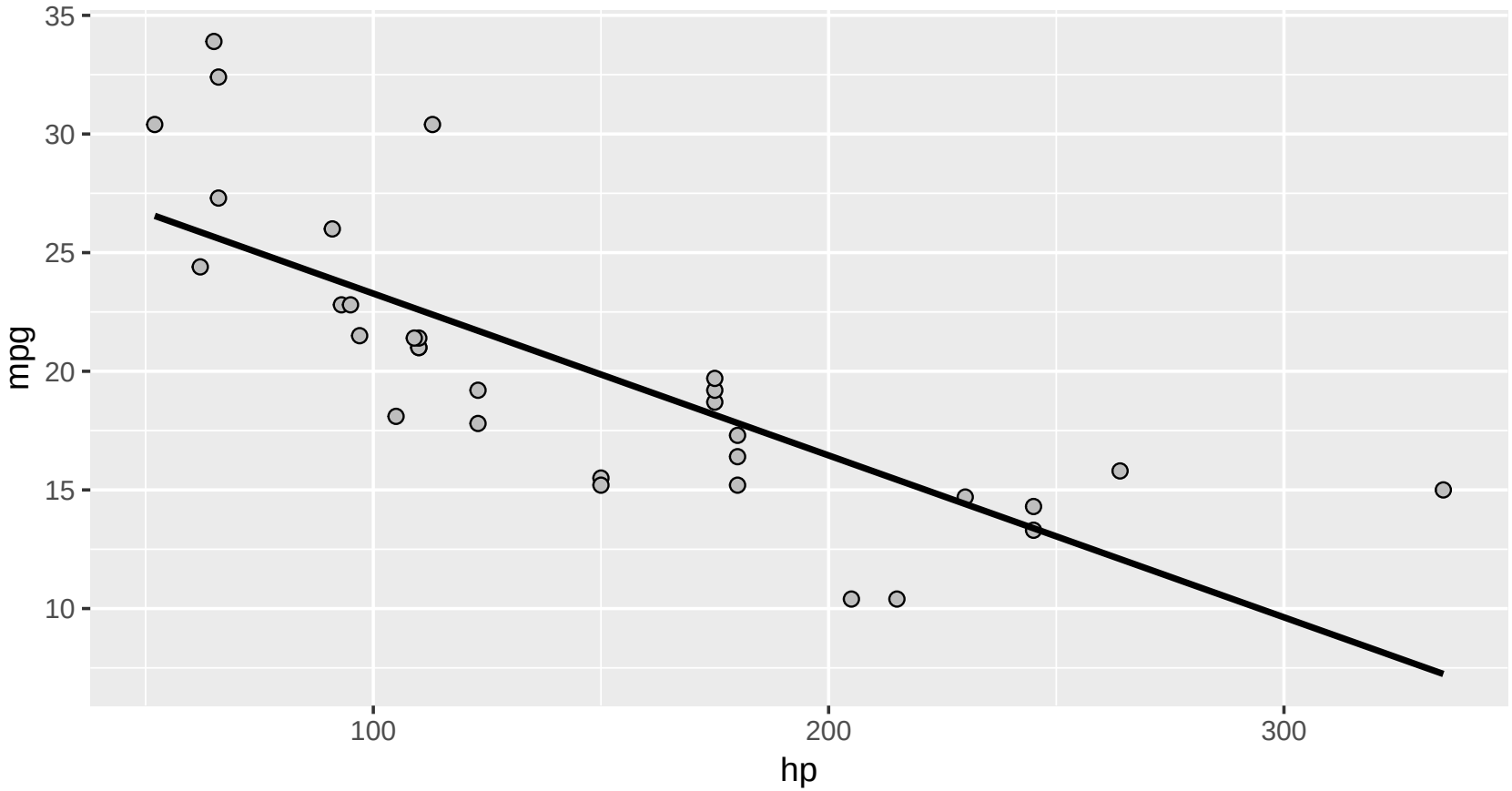
```
t.test(dataFin$ROA ~ dataFin$company)

##
##  Welch Two Sample t-test
##
## data:  dataFin$ROA by dataFin$company
## t = -2.2853, df = 57.308, p-value = 0.02601
## alternative hypothesis: true difference in means is not equal
## 95 percent confidence interval:
##  -0.095987460 -0.006338226
## sample estimates:
##  mean in group CZ mean in group Int
##        0.08528962        0.13645247
```

Welch t-test: μ of first group is **greater**

```
t.test(dataFin$ROA ~ dataFin$company, alternative="greater")

##
##  Welch Two Sample t-test
##
## data:  dataFin$ROA by dataFin$company
## t = -2.2853, df = 57.308, p-value = 0.987
## alternative hypothesis: true difference in means is greater t
## 95 percent confidence interval:
##  -0.08859177          Inf
## sample estimates:
##  mean in group CZ mean in group Int
##        0.08528962        0.13645247
```

Regression Analysis

```
library(ggplot2)
ggplot(mtcars, aes(x=hp, y=mpg) ) +
  geom_point(size=2, shape=21, fill="grey") +
  geom_smooth(method="lm", se=FALSE, color="black")
```



Purpose of Regression Analysis

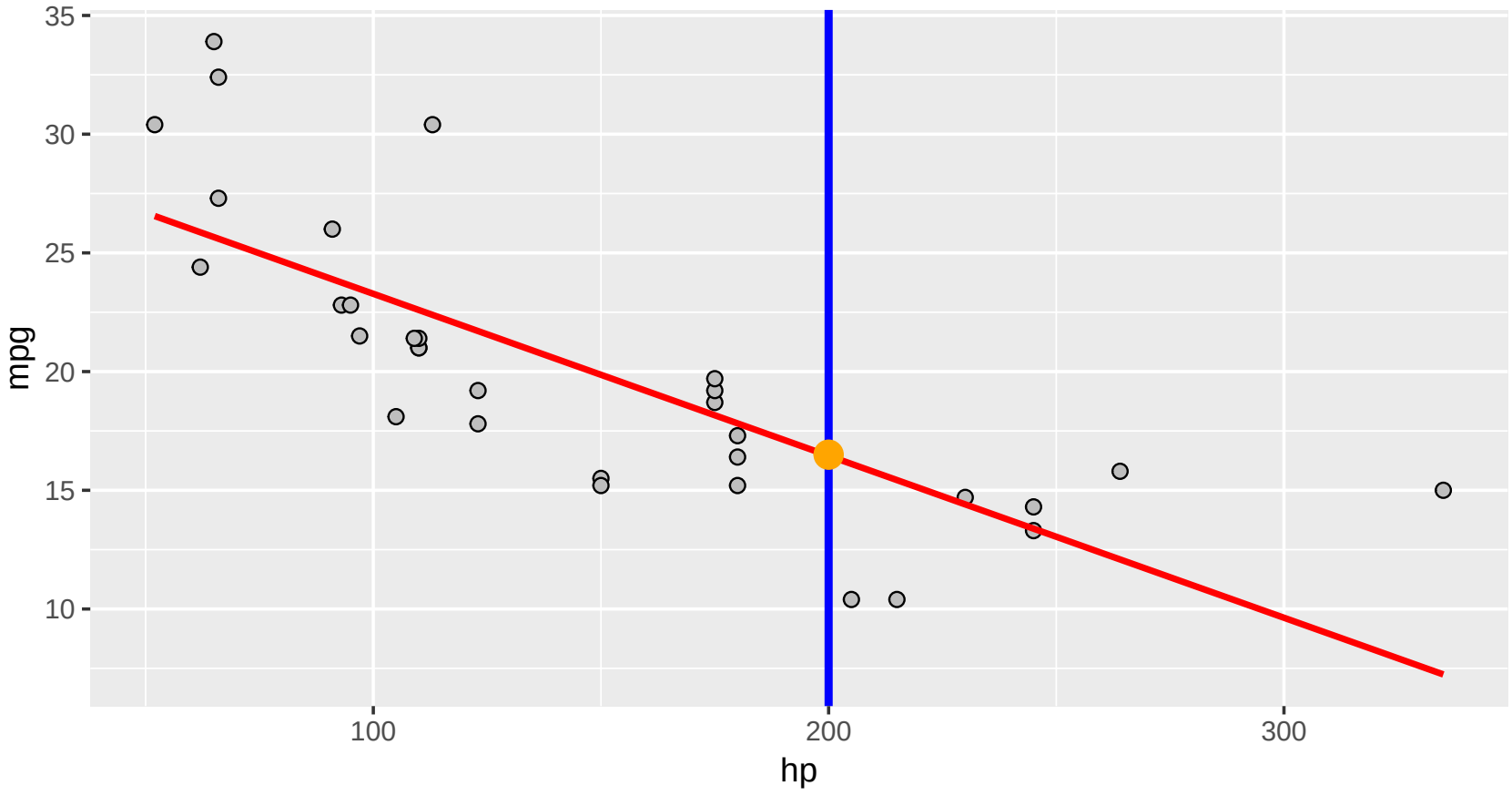
The main purpose is to create a model of conditional expectation of the dependent variable.

$$E[y|x] \rightarrow y = f(x; b) \tag{1}$$

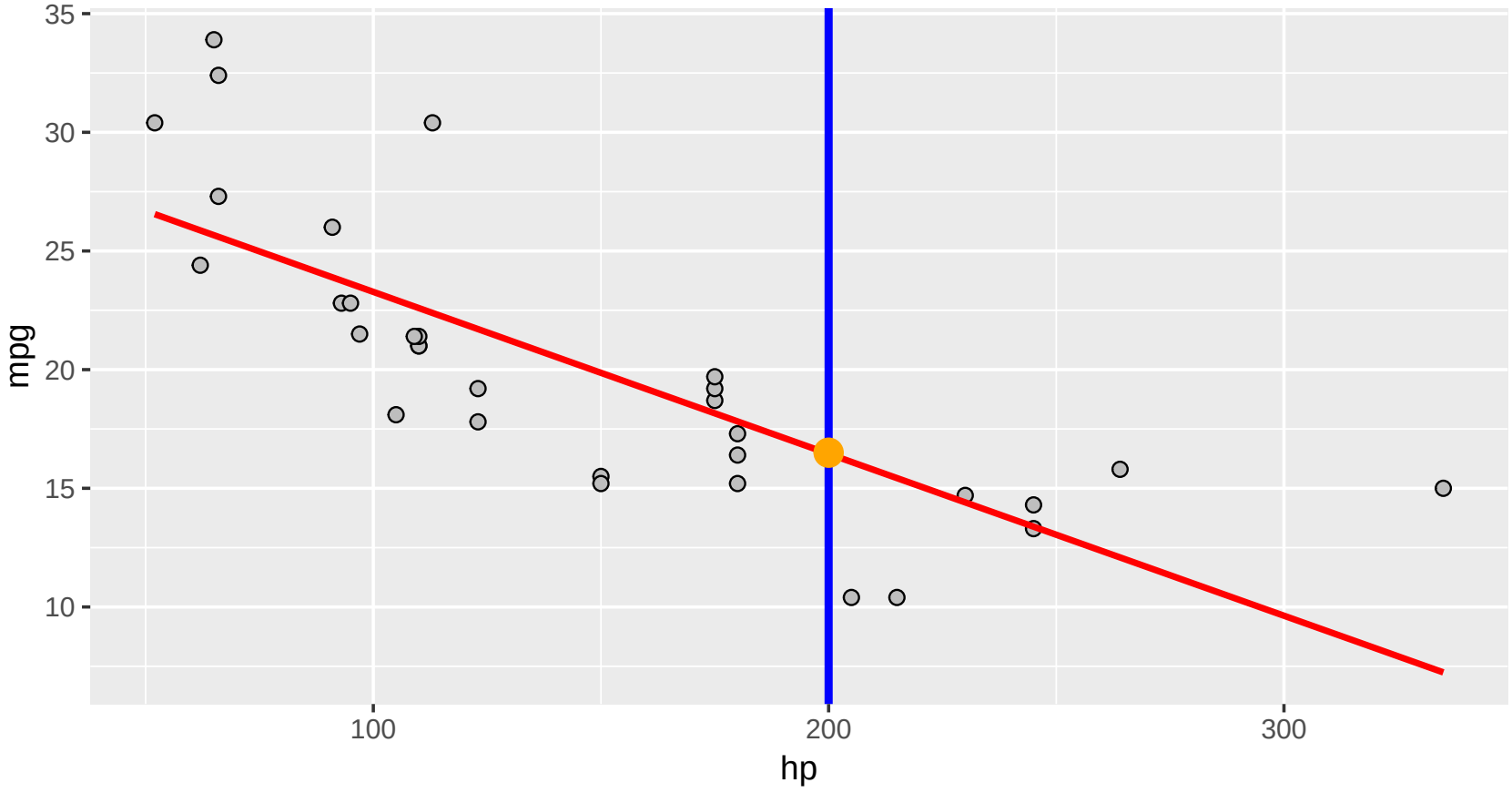
$f(\cdot)$ is a convenient function parametrised by parameters in vector b.

$$y = b_0 + b_1 x_1$$

$E[y|x]$



E[y|x]



Unit increase of x leads to change in y by b_1... is usually wrong interpretation. The most important is the data design.

Linearity in Parameters

- relation between β is additive
- β is not in a *functional* form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{2}$$

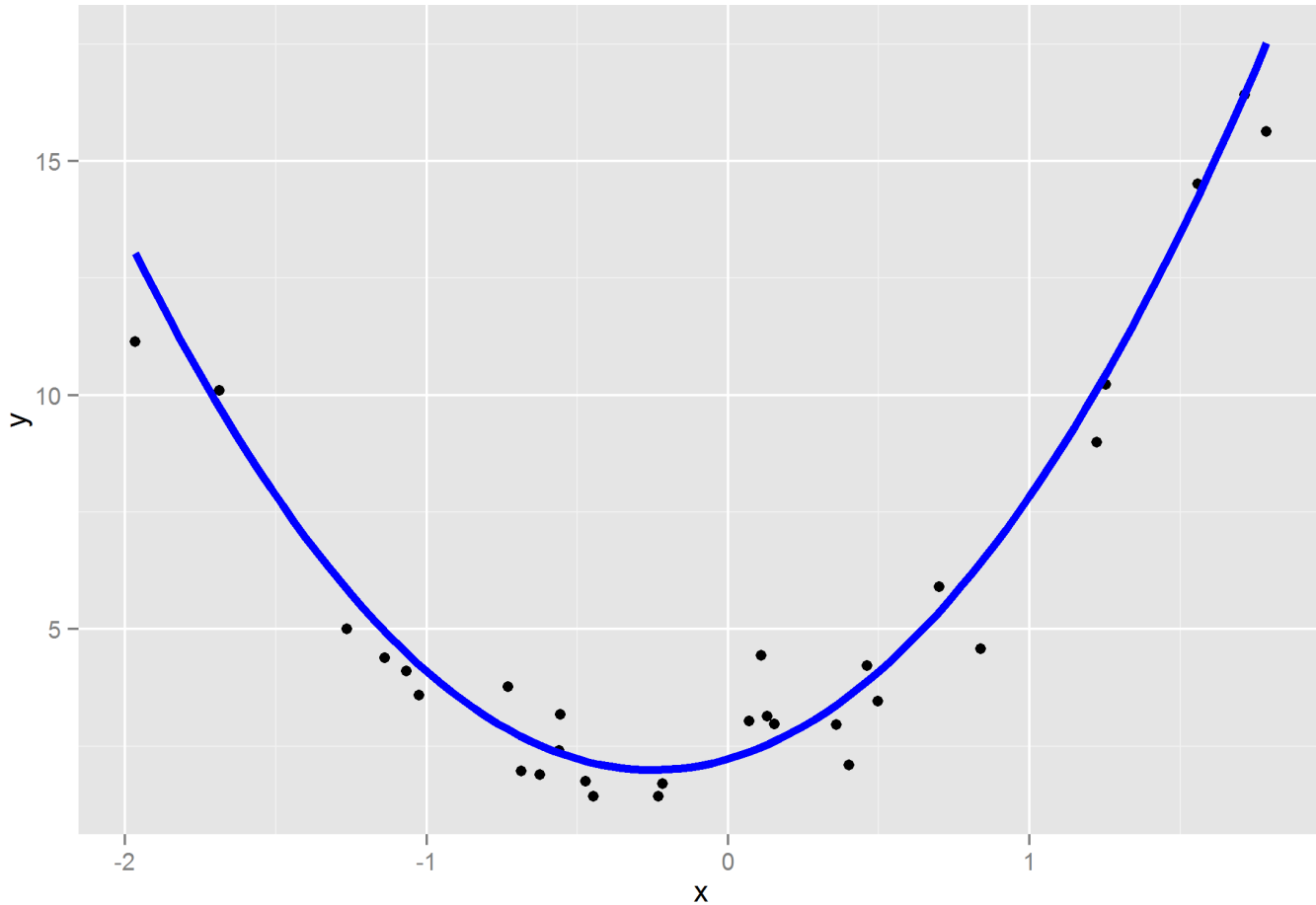
$$y = \beta_0 + \beta_1 \beta_2 x_1 + \beta_3 x_2 + \epsilon \tag{3}$$

$$y = \beta_0 + \beta_1^{\beta_2 x_1} + \epsilon \tag{4}$$

$$y = \beta_0 + \sqrt{\beta_1} x_1 + \beta_2 x_2 + \epsilon \tag{5}$$

Linear model in parameters II

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \tag{6}$$



Linear Model

If the model is linear in parameters, OLS method can be used to estimate parameters. In this case function `lm()` can be used:

```
fit1 <- lm(mpg ~ hp, data=mtcars)
```

Intercept (b_0) is added automatically.

Linear Model

If the model is linear in parameters, OLS method can be used to estimate parameters. In this case function `lm()` can be used:

```
fit1 <- lm(mpg ~ hp, data=mtcars)
```

Intercept (b_0) is added automatically. Consider model with intercept value only:

```
fit0 <- lm(mpg ~ 1, data=mtcars)
```

```
summary(fit0)
```

```
##
## Call:
## lm(formula = mpg ~ 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6906 -4.6656 -0.8906  2.7094 13.8094
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.091      1.065   18.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 6.027 on 31 degrees of freedom
```

```r
summary(fit0)
```

```
##
## Call:
## lm(formula = mpg ~ 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6906 -4.6656 -0.8906  2.7094 13.8094
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.091      1.065   18.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 6.027 on 31 degrees of freedom
```

```r
mean(mtcars$mpg)
```

```
## [1] 20.09062
```

```
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
## hp          -0.06823    0.01012  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024,Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

Polynomial Regression

There are two ways how to create polynomial (or log,...) regressions:

1. Updating data before `lm`

```
carsQuad <- mtcars %>% select(mpg, hp) %>%
  mutate(hp2 = hp^2)

lm(mpg ~ hp + hp2, data=carsQuad) %>% coef()
```

```
##   (Intercept)            hp           hp2
## 40.4091172029 -0.2133082599  0.0004208156
```

Polynomial Regression

There are two ways how to create polynomial (or log,...) regressions:

1. Updating data before `lm`

```
carsQuad <- mtcars %>% select(mpg, hp) %>%
  mutate(hp2 = hp^2)

lm(mpg ~ hp + hp2, data=carsQuad) %>% coef()
```

```
##   (Intercept)            hp           hp2
## 40.4091172029 -0.2133082599  0.0004208156
```

2. Inside of `lm`

```
lm(mpg ~ hp + I(hp^2), data=mtcars) %>% coef()
```

```
##   (Intercept)            hp        I(hp^2)
## 40.4091172029 -0.2133082599  0.0004208156
```

Coding of Variables

$x = 1$ means that highest achieved education level of a subject is a university degree. $x = 0$ indicates non-university degree.

$$\text{wage} = 20\,000 + 5\,000x$$

What if $x = 1$ would mean non-university degree?

Coding of Variables

$x = 1$ means that highest achieved education level of a subject is a university degree. $x = 0$ indicates non-university degree.

$$\text{wage} = 20\,000 + 5\,000x$$

What if $x = 1$ would mean non-university degree?

Redefine the model to have the same economic meaning.

Useful Packages

```
library(broom)
library(car)
```

R and Variable Coding

Recall there is a variable `cyl` which takes only values 4, 6, 8.

```
fit2 <- lm(mpg ~ cyl, mtcars)
tidy(fit2)

## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    37.9      2.07       18.3  8.37e-18
## 2 cyl            -2.88     0.322      -8.92 6.11e-10
```

What do you think about this model?

```
carData <- mtcars %>%
  mutate(cylinder = factor(cyl) )
```

R and Variable Coding

Recall there is a variable `cyl` which takes only values 4, 6, 8.

```
fit2.1 <- lm(mpg ~ cylinder, carData)
tidy(fit2.1)

## # A tibble: 3 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    26.7      0.972     27.4  2.69e-22
## 2 cylinder6      -6.92     1.56      -4.44 1.19e- 4
## 3 cylinder8     -11.6      1.30      -8.90 8.57e-10
```

This model is written as $y = b_0 + b_1 x_1 + b_2 x_2$ where $b_2 = -11.6$ corresponds to the effect of 8 cylinders on `mpg`.

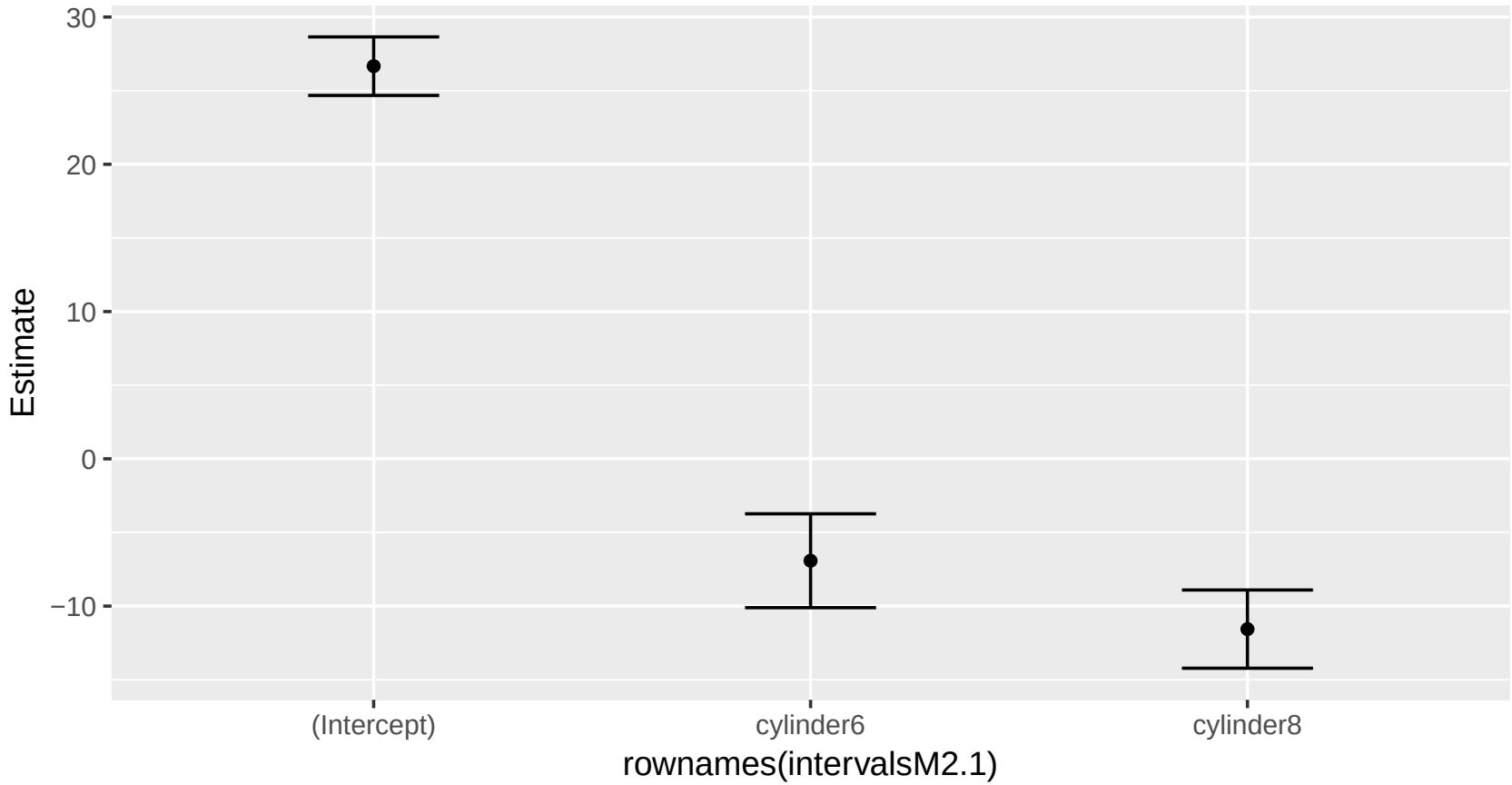
Following Tests – Wald test

```
linearHypothesis(fit2.1, "cylinder6 = cylinder8", test="F")

## Linear hypothesis test
##
## Hypothesis:
## cylinder6 - cylinder8 = 0
##
## Model 1: restricted model
## Model 2: mpg ~ cylinder
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 401.86
## 2     29 301.26  1    100.59 9.6835 0.004152 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Confidence Intervals

```
intervalsM2.1 <- data.frame(confint(fit2.1))
intervalsM2.1

##                  X2.5..    X97.5..
## (Intercept)   24.67608 28.651192
## cylinder6    -10.10796 -3.733599
## cylinder8    -14.21962 -8.907653

colnames(intervalsM2.1) <- c("Low", "Upp")
intervalsM2.1$Estimate <- coef(fit2.1)
intervalsM2.1

##                    Low       Upp   Estimate
## (Intercept)   24.67608 28.651192  26.663636
## cylinder6    -10.10796 -3.733599  -6.920779
## cylinder8    -14.21962 -8.907653 -11.563636
```

```
ggplot(intervalsM2.1, aes(
   x=rownames(intervalsM2.1), y=Estimate) ) +
  geom_point() +
  geom_errorbar(aes(ymin=Low, ymax=Upp), width=0.3)
```



Time-Series

Useful Packages

```
library(forecast) # Smoothing and many many more

library(vars) # Vector Autoregressive Models

library(urca) # unit-root and cointegration tests

library(xts) # Time-series objects

library(lubridate) # works with Date type
```

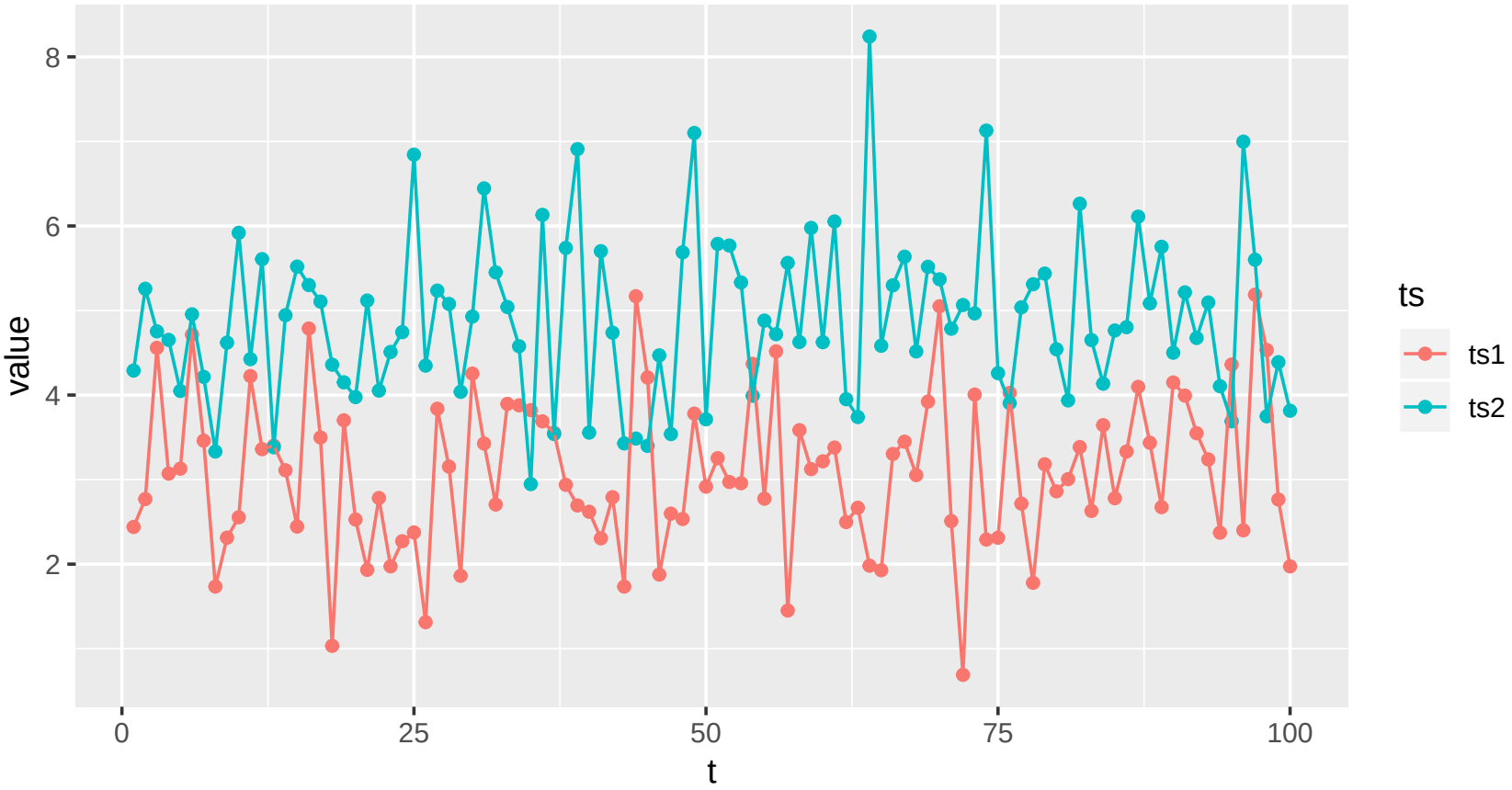
Revolution in progress! `library(tstibble)`, new version of `forecast` called `fable`.

Time Series

- Very different methodological approach (stationarity)
- Seasonality
- Time series is a set of autocorrelated values (unlike cross-sectional)

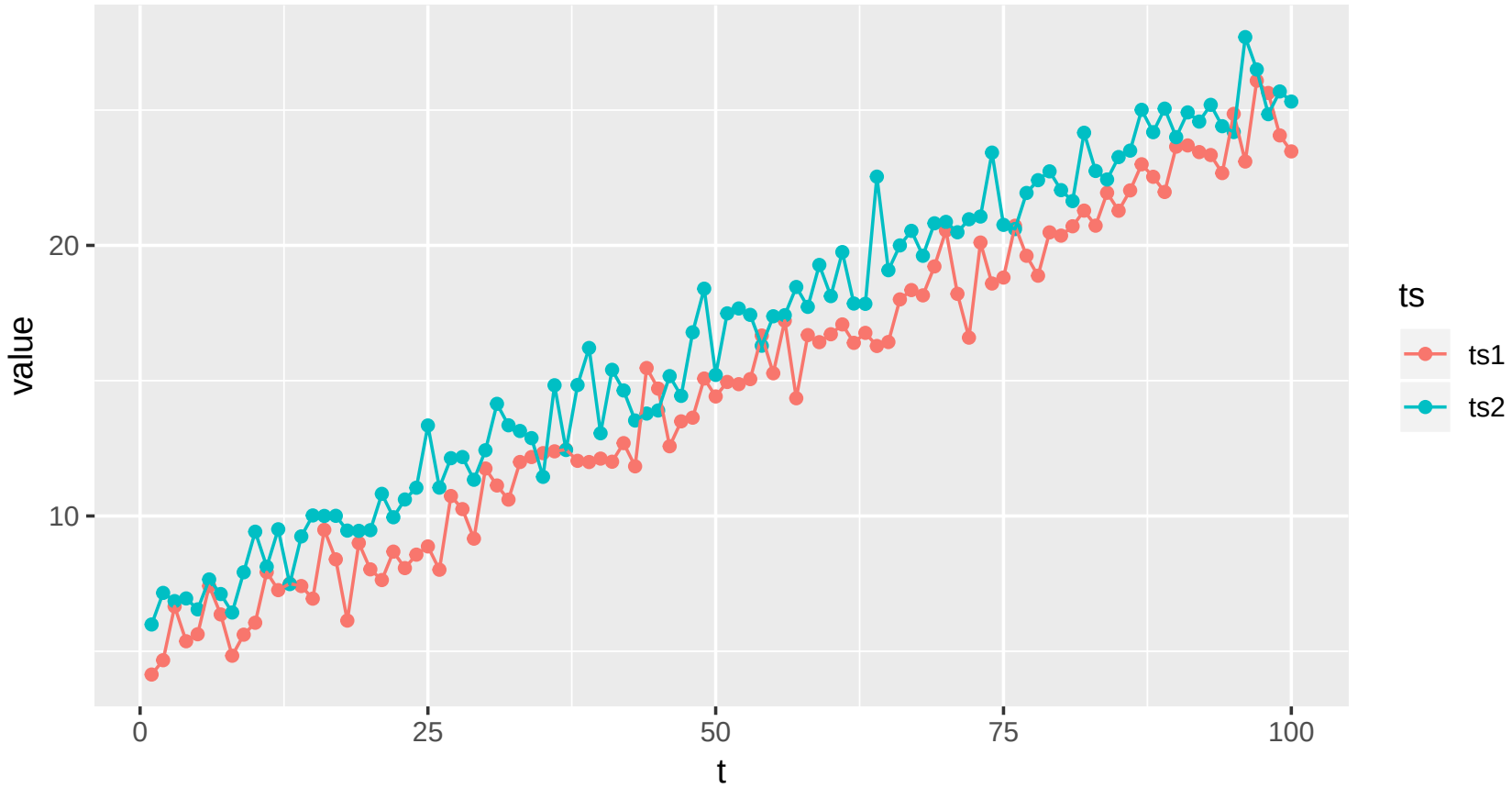
NEVER use standard Pearson Correlation test to analyse time-series!

R Demonstration of Spurious Correlation

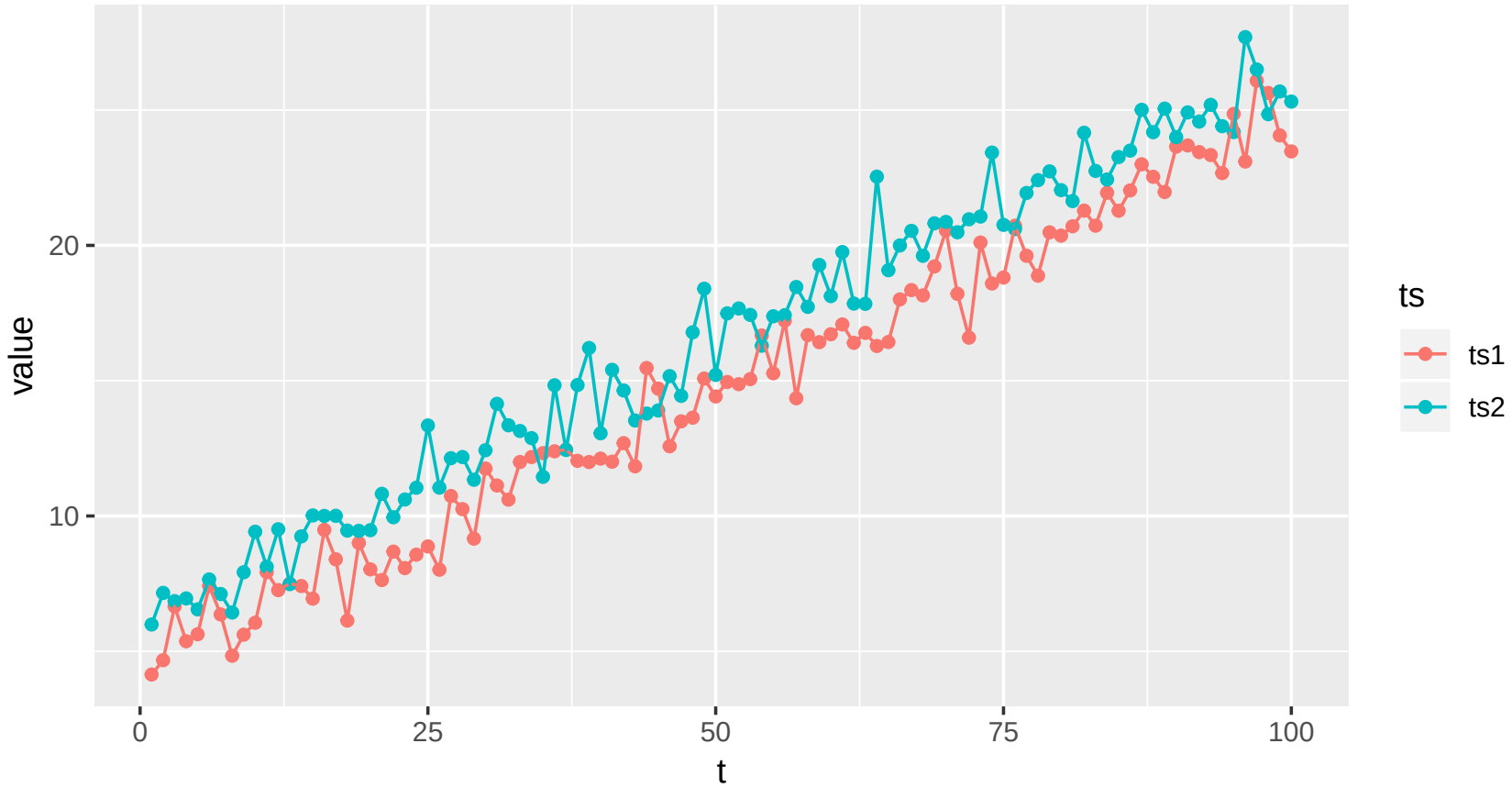



```
cor(df$ts1, df$ts2)

## [1] -0.04953215
```

```r
df2 <- df %>% mutate(
 ts1 = ts1 + 1.5 + 0.2*seq(100),
 ts2 = ts2 + 1.5 + 0.2*seq(100) )
```



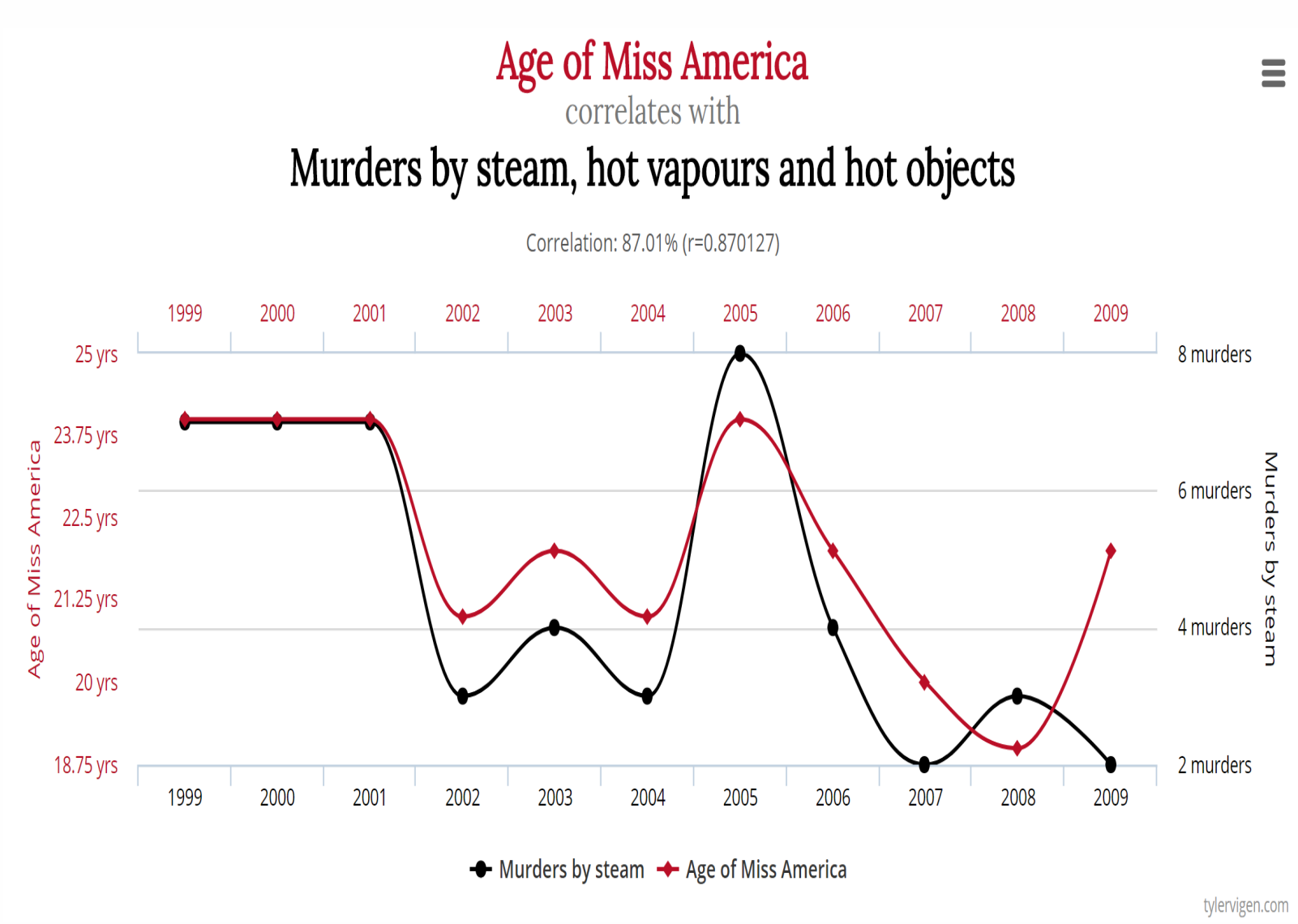
```r
df2 <- df %>% mutate(
 ts1 = ts1 + 1.5 + 0.2*seq(100),
 ts2 = ts2 + 1.5 + 0.2*seq(100) )
```




```r
cor(df2$ts1, df2$ts2)
```

```
## [1] 0.9738829
```

Spurious Correlation



Stationarity

Assumption that the time series has a constant mean and variance (weak form of definition).
This needs to be tested:

- stochastic stationarity (SS) - ADF test, PP test,...
- deterministic stationarity (DS) - KPSS test

It is important to distinguish between two types. SS time series can be made stationary by differencing, DS time series by removing trend (e.g., by using OLS).

```
library(tseries)
adf.test(df$ts1)

## Warning in adf.test(df$ts1):  p-value smaller than
printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  df$ts1
## Dickey-Fuller = -4.3961, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

```
adf.test(df2$ts1)

## Warning in adf.test(df2$ts1):  p-value smaller than
printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  df2$ts1
## Dickey-Fuller = -4.3961, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

Kwiatkowski et al. test

H_0: Time series is trend stationary:

```
kpss.test(df$ts1)$p.value
```

```
## [1] 0.1
```

```
kpss.test(df2$ts1)$p.value
```

```
## [1] 0.01
```

If rejected, de-trending will help.

Other Useful Functions

- stl – time-series decomposition
- ccf – cross-correlation
- Arima
- var

Thank you!

homolka@utb.cz