Lesson 4:
Bivariate Statistics

Recap: Graphs

- Categorical variables: bar plot, pie chart
- Continuous variables: histogram, boxplot, qq plot, spaghetti plot

Important commands (recap)

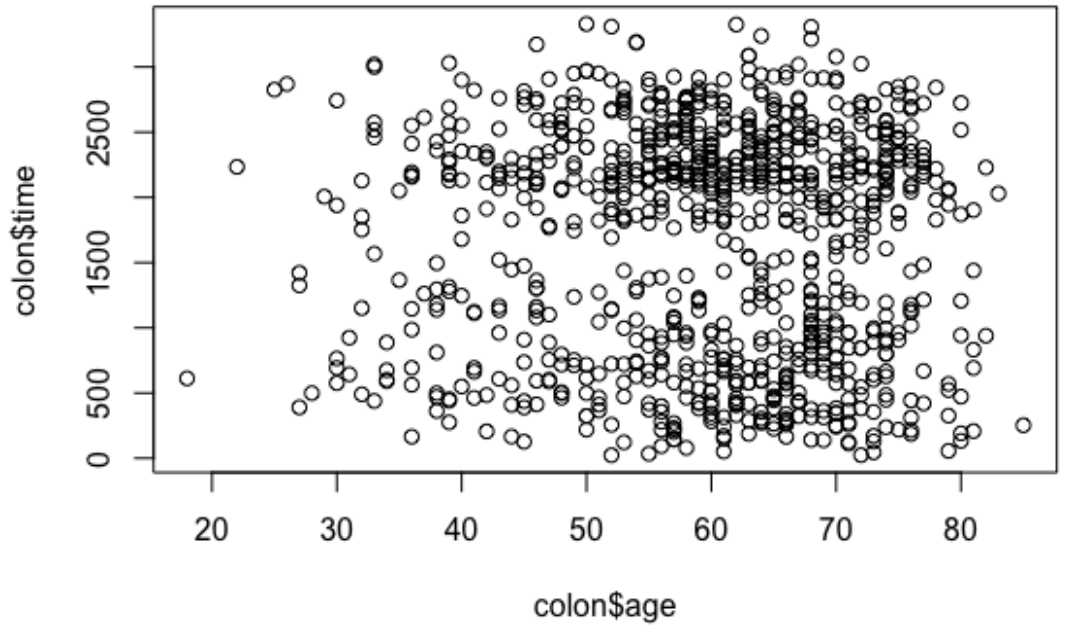
COMMAND	EFFECT
`boxplot(var)`	boxplot of a variable (without using `ggplot2`)
`ggplot(data = data, aes(x = var1, y = var2, fill = var_cat))`	package `ggplot2`, needs to be specified with geom layers (see below)
`coord_polar(start = 0, "y")`	make pie chart from data
`geom_bar(stat = "identity", position = position_dodge(), width = 1)`	bar plot (with bars arranged next to each other)
`geom_text()`	add text (such as a label) to the plot
`ggtitle()`	add a title to the plot
`theme()`	
`geom_histogram(aes(y = y), bins = 10)`	histogram with specified number of bins, `y = .. density ..` for density plot
`geom_density()`	add density curve to the histogram plot
`geom_boxplot()`	boxplot of the data
`geom_jitter()`	add data points to the boxplot

Bivariate Statistics

- scatterplot (graphs)
- correlational analysis
- linear regression

first: look at the data

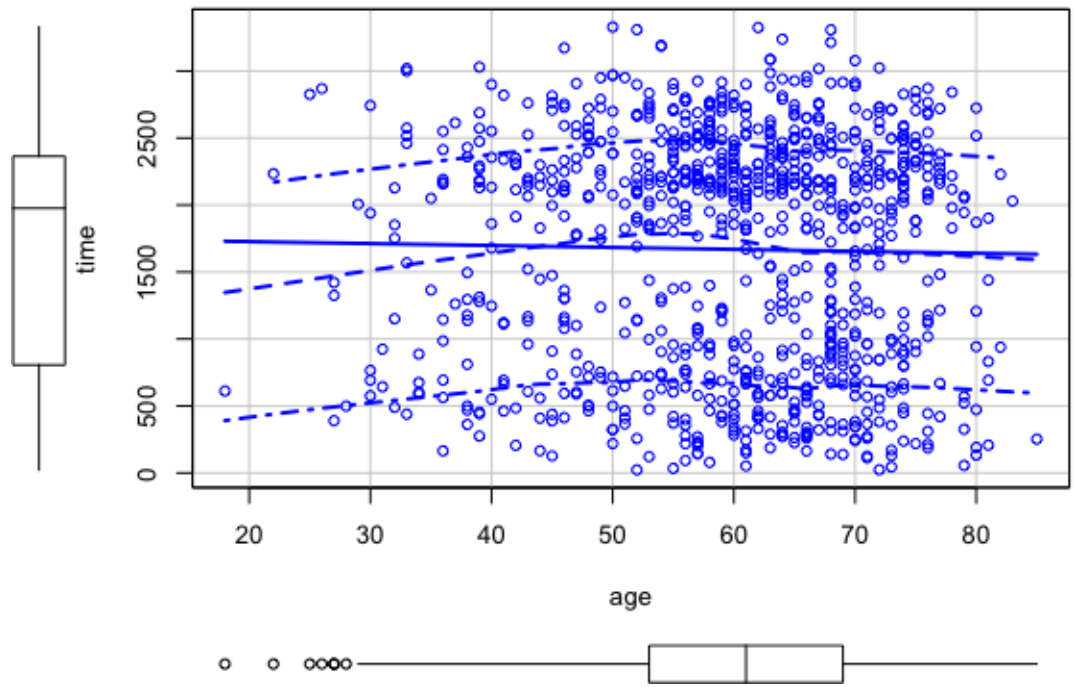
- create scatterplot for two variables
- use `plot(x, y)` (x: age; y: time)



first: look at the data

- create scatterplot for two variables
- more detailed: (package `car`)

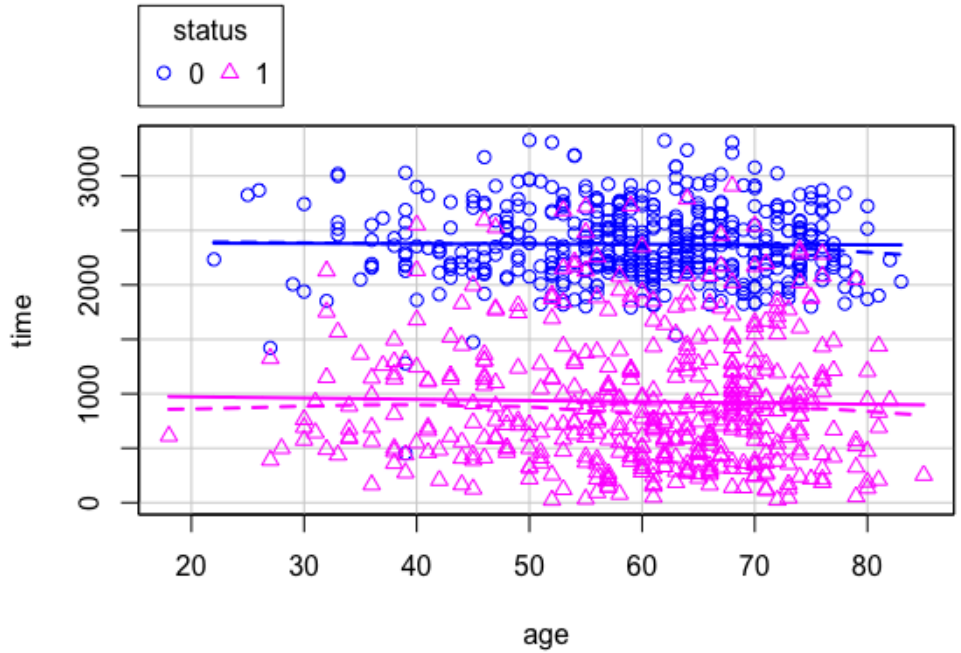
```
scatterplot(y ~ x, data = data)
```



first: look at the data

- create scatterplot for two variables
- split for categorical variable (here `status`)

```
scatterplot(y ~ x | cat_var, data = data)
```



correlational analysis

- use `cor()` from package `stats`

```
> cor(var_1, var_2)
[1] NA
```

- default: calculates pearson correlation with all values
- but: pearson only if data is normally distributed!
- default: uses all observations, even if NA

correlational analysis

- use `cor()` from package `stats`

```
> cor(var_1, var_2,
          method = “spearman”,
          use = “complete.obs”)
[1] 0.05
```

- spearman correlation is robust for not normally distributed data
- NA’s are removed from calculation (only complete observations used)

correlational analysis

- is correlation significant? (package `stats`)

```
cor.test(var_1, var_2,
     method = "pearson",
     alternative = "two.sided")
```

- use "`spearman`" if data is not normally distributed
- alternative can be "`less`" or "`greater`"

correlational analysis

- is correlation significant? (package `stats`)
- exemplary output for `cor.test(data$v1, data$v2)`

```
        Pearson's product-moment correlation

data:  data$v1 and data$v2
t = -0.60907, df = 927, p-value = 0.5426
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08421163  0.04437612
sample estimates:
        cor
-0.02000047
```

correlational analysis

- correlations for multiple variables (package `Hmisc`)

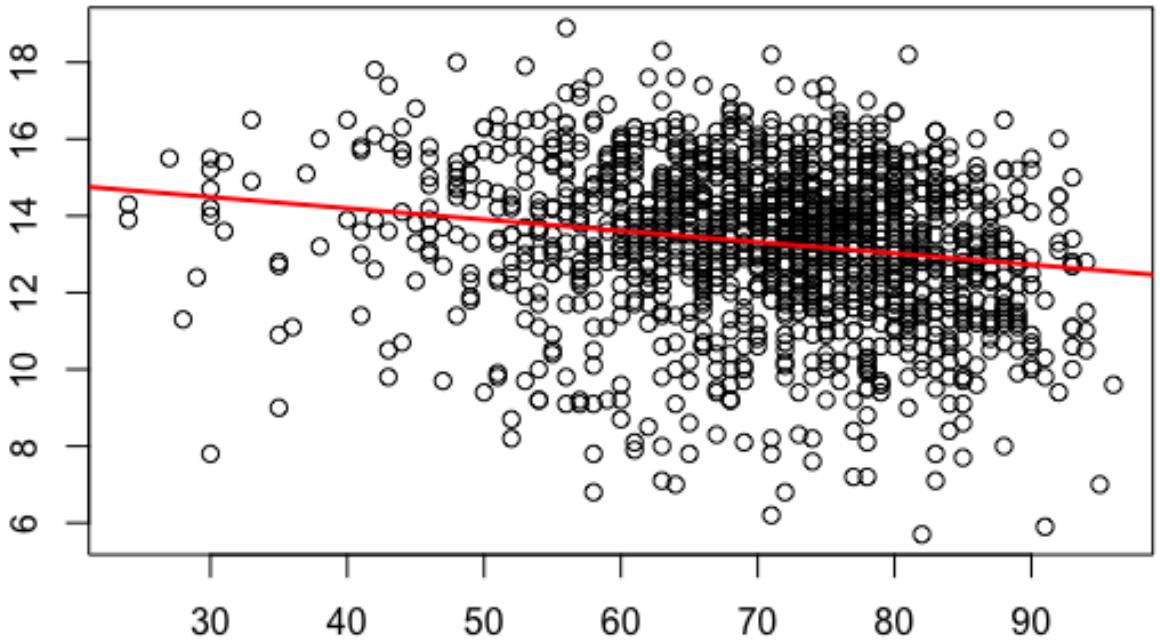
```
rcorr(as.matrix(data[, c(”var_1", ”var_2", ”var_3")]),
      type = "spearman")
```

- produces (spearman) correlation table for all 3 variables
- gives n with which correlations were calculated
- gives p values for correlations

linear regression

- add linear regression line to scatterplot

```
plot(y ~ x)
abline(lm(y ~ x))
```



linear regression

- output of statistical measures:

```
summary(lm(y ~ x))
```

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2647 -1.1385  0.1937  1.4290  5.2059

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.378579   0.315833  48.692  < 2e-16 ***
x           -0.029438   0.004416  -6.666  3.8e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.99 on 1369 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.03144, Adjusted R-squared:  0.03073
F-statistic: 44.44 on 1 and 1369 DF,  p-value: 3.796e-11
```

β for x and test of that β

R^2 for regression (variance measure)

linear regression

- multiple linear regression for more than one predictor
- add predictor variable to your model:

```
summary(lm(y ~ x1 + x2, data = data))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.840672   0.314422  50.380  < 2e-16 ***
x1          -0.028297   0.004336  -6.526 9.54e-11 ***
x2          -0.418480   0.046679  -8.965  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.927 on 1346 degrees of freedom
  (35 observations deleted due to missingness)
Multiple R-squared:  0.08627, Adjusted R-squared:  0.08491
F-statistic: 63.54 on 2 and 1346 DF,  p-value: < 2.2e-16
```

Important commands

COMMAND	EFFECT
plot(var1, var2)	plot scatterplot of two continuous variables
scatterplot()	package car, advances scatterplot
cor(var1, var2)	package stats, calculate correlation between two variables
cor.test()	package stats, test the correlation for significance
rcorr()	package Hmisc, calculates correlation also for more than one variable, adds p values
abline()	add linear regression line to scatterplot
lm()	package stats, calculates linear regression model
summary()	summarizes linear regression model