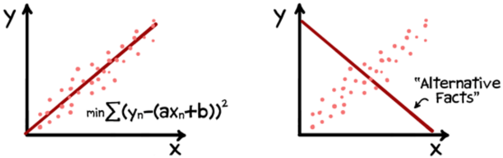
WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Correlation and Regression

Class 3: Marketing Research

Service and Digital Marketing
March 27, 2017

- Understand how to use regression analysis for statistical model building
- Conduct linear regression analysis, interpret the results and their statistical validity
- Be able to formulate a (multiple) regression model and how to assess prediction accuracy of the model
- Understand when and how to use correlation analysis
- R: working with lists, generic functions, formatting R-output



Overview of technique

- technique to analyze associative relationships between one metric **dependent** and one **independent** variable (**simple regression**) or more independent variables (**multiple regression**)
- to determine if the independent variable(s) explain a significant variation in the dependent variable
- to determine the structure or form of the relationship (mathematical equation)
- to predict the values of the dependent variable

The aim is to derive a **predictor formula** (the mathematical formulation) to model the relationship of two or more variables.

The **statistical model** describes the expected change of the dependent variable whenever the independent variable is altered.

Dependent and independent variables

Dependent (response) variable:

- represents the output or effect
- also known as a response variable, regressand, measured variable, responding variable, explained variable, outcome variable, experimental variable, and output variable
- for linear regression is of kind: continuous
- for generalized linear regression is of kind: continuous or discrete

Independent (explanatory) variable:

- represents the inputs tested to see if they are the cause of the effect
- also known as a predictor variable, regressor, controlled variable, manipulated variable, explanatory variable
- may be of kinds: continuous, binary/dichotomous, nominal categorical, ordinal categorical

Simple linear regression (1)

Simple linear regression

- we start formulating our **linear model** in a general form

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

data: we want to explain y_i (the **dependent variable**) by one explanatory variable x_i (the **independent variable**) both measured on $i = 1, \ldots, N$ observations

e_i: is denoting the error term (the residuum that we cannot explain by the model)

β_0 and β_1: are unknown and have to be estimated

- so that we get as final result the estimation equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Note: Another common notation is $y_i = \alpha + \beta_1 x_i + e_i$.

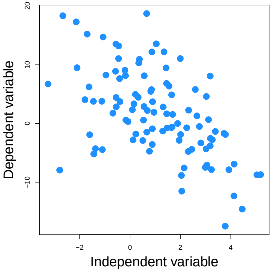
- The estimation equation

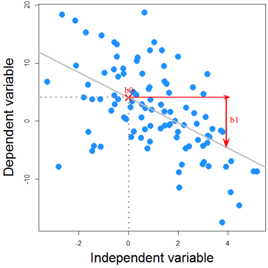
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- is the equation for a straight line

$$\begin{aligned} y &= kx + d \\ y &= d + kx \\ \hat{y}_i &= \underbrace{\hat{\beta}_0}_{\textit{Intercept}} + \underbrace{\hat{\beta}_1 y_i}_{\textit{Slope}} \end{aligned}$$

Thus, the **intercept** β_0 is a constant value (indicating the distance to the origin) and the **slope** β_1 indicates the expected change in y when x is changed by one unit.





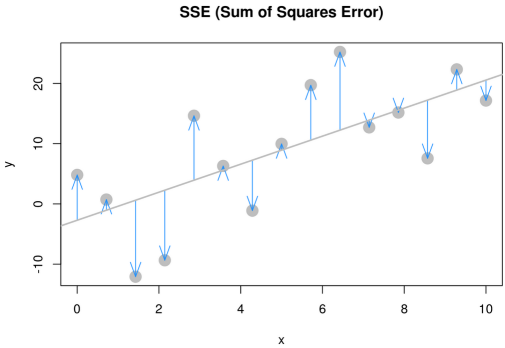
Parameter estimation

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unknown and have to be estimated from the sample observations using the general form $y_i = \beta_0 + \beta_1 x_i + e_i$
- By using the method of ordinary least squares (OLS) the **squared sum of the model errors** (= deviations of the data points from the line, also called **residuals**) is minimized

$$\min_{\beta_0,\beta_1} \sum_i e_i^2 =$$

$$\min_{\beta_0,\beta_1} \sum_i (y_i - \hat{y}_1)^2 =$$

$$\min_{\beta_0,\beta_1} \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$



First order conditions (leading to normal equations)

$$\frac{\partial \sum_i e_i^2}{\partial \hat{\beta}_0} = -2\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

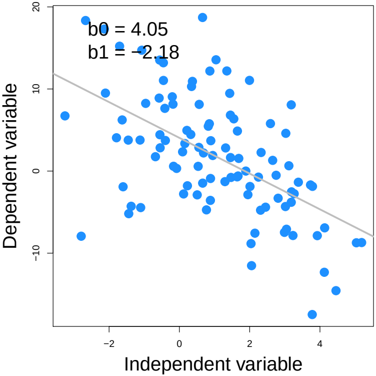
$$\frac{\partial \sum_i e_i^2}{\partial \hat{\beta}_1} = -2\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

$$\hat{\beta}_1 = \frac{\overbrace{n(\sum_i y_i x_i - \sum_i x_i \sum_i y_i)}^{\text{covariance of } x \text{ and } y:\ S_{xy}}}{\underbrace{n(\sum_i x_i^2 - (\sum_i x_i)^2)}_{\text{variance of } x:\ S_{xx}}} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2}$$

$$\hat{\beta}_0 = \frac{1}{n}\sum_i y_i - \hat{\beta}_1 \frac{1}{n} x_i = \bar{y} - \hat{\beta}_1 \bar{x}$$

Simple linear regression (5)

```
# DIY: obtaining the regression coefficients
> x <- bvn1[,1]; y <- bvn1[,2]
> Sxy <- sum((x - mean(x)) * (y - mean(y)))
> Sxx <- sum((x - mean(x)) ^ 2)
> Syy <- sum((y - mean(y)) ^ 2)
# covariance, variance of x, variance of y
> c(Sxy, Sxx, Syy)

 -735.2942  337.6921 5296.8487

> beta_1 <- Sxy / Sxx
> beta_0 <- mean(y) - beta_1 * mean(x)
> c(beta_0, beta_1)

 4.046775 -2.177410

# fast
> lm(y ~ x)

(Intercept)            x
      4.047       -2.177
```



Simple linear regression (6)

Assumptions

1. metric dependent variable
2. linear relationship
3. residuals: $e_i \sim N(0, \sigma^2 I)$
   3.1 independent and normally distributed
   $=$ no relationship between subsequent residuals
   3.2 **constant variance** (**homoscedasticity**)
   $=$ dispersion is the same across all observations
4. attention to **outliers**

Simple linear regression (7)

Example: Yelp Dataset (`yelp.csv`)

The Yelp dataset is a collection of millions of restaurant reviews, each accompanied by a 1-5 star rating. Sentiment analysis was performed on each review using the AFINN lexicon was to obtain a positivity score for each word, ranging from -5 (most negative) to 5 (most positive).

Research question: To what extent can we describe and predict a customer's rating based on their written opinion? Can we predict the positivity or negativity of someone's writing by counting words?

$$Av.Star\ Rating_i = \beta_0 + \beta_1 \times Positivity\ Score_i + e_i$$

$H_0 : \beta_1 = 0$; there is no linear relationship from X (positivity score) on Y (star rating)
$H_1 : \beta_1 \neq 0$; there is a positive or negative linear relationship from X on Y

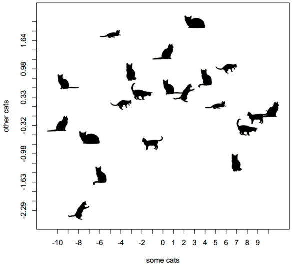
Note: The intercept (or constant) β_0 is usually not formally tested.

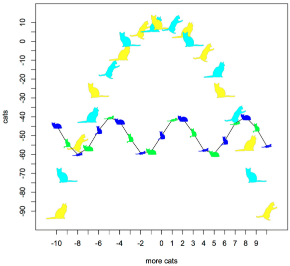
Assumption checks: linear relationship (1)

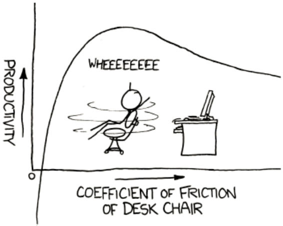
- **Assumption 1**: **metric** (or at least **interval**) scaled dependent variables: dependent variable: sentiment score (✓)
- **Assumption 2**: **linear relationships** between the dependent and the independent variables: inspect the scatterplot (✓)

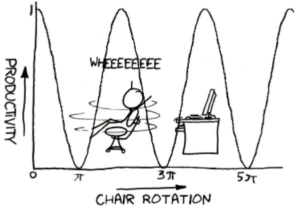
```
> yelp <- read.table("yelp.csv", header=T, sep=";", dec=".")
> head(yelp)
          word businesses reviews uses average_stars afinn_score
1      ability        527     552  570      3.927536           2
2       accept        773     835  910      2.937725           1
3     accepted        378     390  411      2.979487           1
4     accident        480     540  606      3.679630          -2
5 accidentally        283     295  299      3.166102          -2
6       active        296     346  398      3.933526           1

> plot(yelp$average_stars ~ jitter(yelp$afinn_score),
+ ylim=c(0,5), xlim=c(-5,5), pch=16, col=adjustcolor("dodgerblue",0.8),
+ cex=1.25, xlab="Positivity score", ylab="Average star rating")
```

Examples for (non-linear) relationships

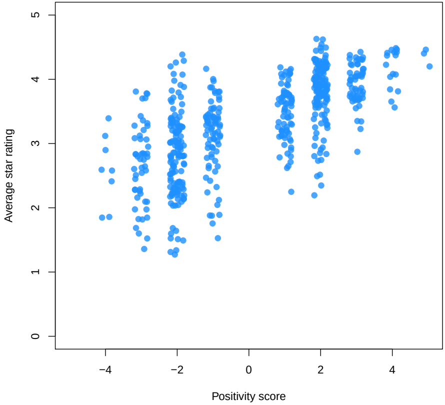








Assumption checks: linear relationship (2)



- the relationship looks linear (we continue with fitting the linear model)

Hypothesis testing (1)

```
# we can again fit the linear model using the function "lm()" (linear model)
> mod1 <- lm(yelp$average_stars ~ yelp$afinn_score)

Call:
lm(formula = yelp$average_stars ~ yelp$afinn_score)

Coefficients:
     (Intercept)  yelp$afinn_score
          3.3113            0.2227
```

Interpreation: up to a constant baseline of 3.31 for the average star rating, the rating increases about 0.22 if the positivity score increases by 1 point.

Test for significance of the β_1 coefficient

$$t = \frac{\beta_1}{SE_{\beta_1}}, df = n - 2$$

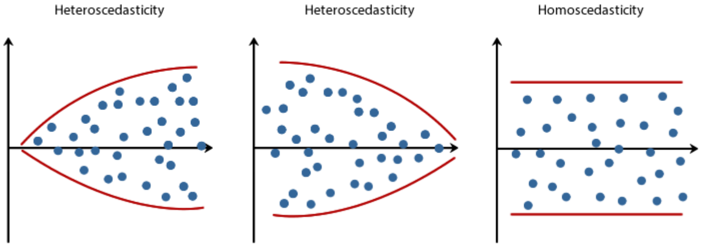
with SE_{β_1} being the standard error (SE) of the estimate β_1.
How to obtain the standard error?

Hypothesis testing (2)

```
# use "summary()" to obtain the t-statistics and the SE and
# the corresponding p-value for the regression coefficients
> summary(mod1)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.31126    0.02252  147.06   <2e-16 ***
yelp$afinn_score  0.22273    0.01031   21.61   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# summary() is a generic function (depending on the object class
# different summary output is returned)
> class(mod1)
"lm"
> class(yelp)
"data.frame"
```

Interpretation: There is a positive linear relationship from the positivity score on the average star rating ($p < .001$). The more positive the written text, the more positive is the average star rating.

Excursus: heteroscedasticity



Smaller predicted values would produce small residuals or larger predicted values would produce smaller residuals

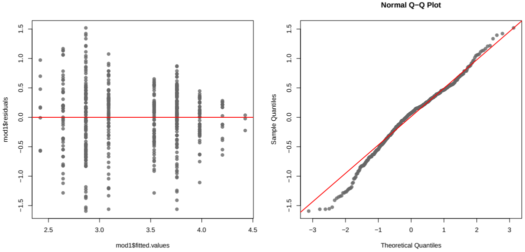
- heteroscedasticity has serious consequences for standard errors of the OLS estimator - they are wrong
- hypothesis tests are no longer valid, and predictions are inefficient
- appears within big data sets or when one uses grouped data

How to obtain the residuals?

```
# the object "mod1" is a named list object, thus containing more information
> str(mod1)

List of 12
 $ coefficients : Named num [1:2] 3.311 0.223
  ..- attr(*, "names")= chr [1:2] "(Intercept)" "yelp$afinn_score"
 $ residuals    : Named num [1:556] 0.171 -0.596 -0.555 0.814 0.3 ...
  ..- attr(*, "names")= chr [1:556] "1" "2" "3" "4" ...
 $ effects      : Named num [1:556] -78.749 -11.452 -0.552 0.781 0.268 ...
  ..- attr(*, "names")= chr [1:556] "(Intercept)" "yelp$afinn_score" "" "" ...
 $ rank         : int 2
 $ fitted.values: Named num [1:556] 3.76 3.53 3.53 2.87 2.87 ...
  ..- attr(*, "names")= chr [1:556] "1" "2" "3" "4" ...
  ...

# using "$" like in a named data.frame object to get the model residuals
> head(mod1$residuals)
         1          2          3          4          5          6
 0.1708182 -0.5962660 -0.5545033  0.8138218  0.3002938  0.3995355
```

Assumption checks: residuals $e_i \sim N(0, \sigma^2 I)$

```
# 1. constant variance
> plot(mod1$residuals ~ mod1$fitted.values, col=adjustcolor("grey40",0.8),
# cex=1.25, pch=16)
> abline(h=0, lwd=2, col="red")

# 2. normal distribution
> qqnorm(mod1$residuals, col=adjustcolor("grey40",0.8), cex=1.25, pch=16)
> qqline(mod1$residuals, col = "red", lwd=2)
```

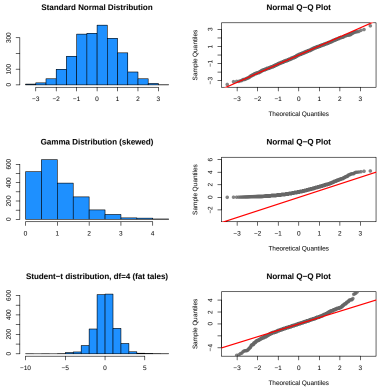



```
# also plot() is a generic function, thus fast:
> plot(mod1)
```

Excursus: Quantile comparison plots

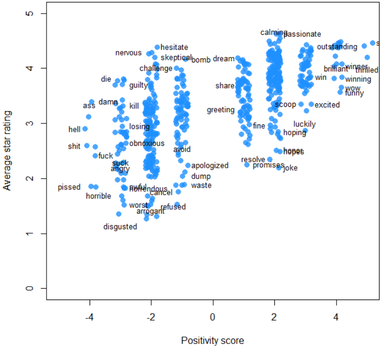
A **Q-Q plot** is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

- A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate).
- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$.

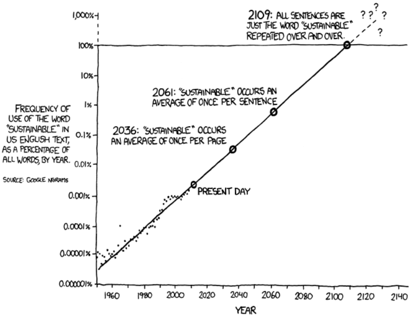


Assumption checks: residuals $e_i \sim N(0, \sigma^2 I)$

```
# word missclassification (using an interactive plot):
> plot(yelp$average_stars ~ jitter(yelp$afinn_score), ylim=c(0,5), xlim=c(-5,5),
+ pch=16, col=adjustcolor("dodgerblue",0.8), cex=1.25, xlab="Positivity score",
+ ylab="Average star rating")
> identify(yelp$afinn_score, jitter(yelp$average_stars), yelp$word, cex=0.8,
+ offset=1, pos=2)
```



Prediction



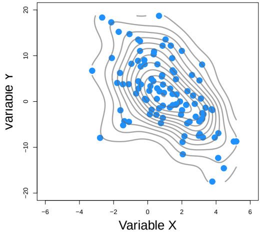
- $Av.Star\ Rating_i =$ $3.311 + 0.223 \times Positivity\ Score$
- What average star rating can be expected for a review including a word with a positivity score of 3?
- $Av.Star\ Rating_i =$ $3.311 + 0.223 \times 3 = 3.980$

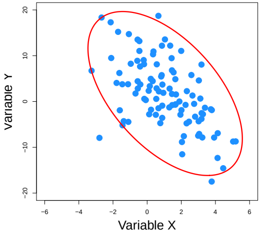
Note: Don't predict values that are outside the range of the data, predictions can be inaccurate (or suspicious).
Example: $Av.Star\ Rating_i = 3.311 + 0.223 \times (-20) = -1.143$

Overview of technique

- is a statistic that summarizes the strength of association between two metric variables
- indicates the degree to which the variation in X is related to the variation in Y
- there exist several correlation coefficients often denoted by ρ (population parameter) and r (sample statistic)

The aim is to derive a measure of how strong two variables are related





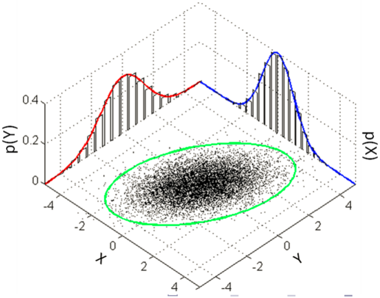
Pearson correlation coefficient (1)

- The **Pearson product-moment correlation coefficient** is the most popular measure of dependence between two variables
- It is only sensitive to a linear relationship (linear correlation, dependence) between two continuous variables X and Y
- It is a normalized measure, i.e., r_{xy} can get values between 1 ("perfect" positive linear relationship) and -1 ("perfect" negative linear relationship)

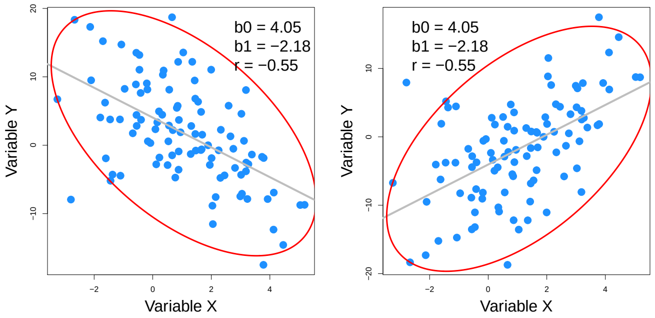
$$r_{xy} = \frac{s_{x,y}}{\sqrt{s_{x,x} s_{y,y}}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y}_i)}{\sqrt{\sum_i (x_i - \bar{x}^2) \sum_i (y_i - \bar{y})^2)}}$$

Assumptions

- linear relationship between X and Y
- metric (or at least interval) scaled dependent variables
- normally distributed values of X and Y



Pearson correlation coefficient (2)



positive relationship (left): small x values cause small y values and large x values cause large y values

negative relationship (right): small x values cause large y values and large x values cause small y values

Pearson correlation coefficient (3)

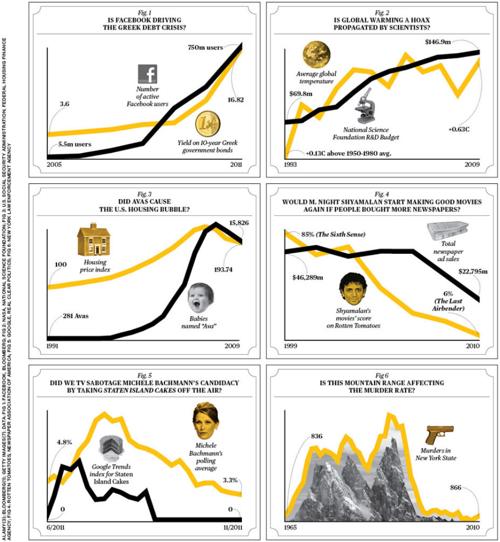
- Is also used in regression analysis as interpretation of the **model fit** (a value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points on a line)
- The higher the absolute value of r the better your linear model describes the data
- In our case there is a really high positive relationship between the *average star rating* and the *positivity score* ($r = 0.676$) indicating that our linear model

$$Av.Star\ \ Rating_i = 3.311 + 0.223 \times Positivity\ \ Score_i$$

describes the data well (= deviations from line are small).

```
cor(yelp$average_stars, yelp$afinn_score)
0.6762595
```

“Alternative” facts: correlation does not imply causation



Coefficient of determination R^2 (1)

- More common to use the R^2 coefficient to describe how well the regression line fits the data set (i.e., a **measure of model fit**).
- R^2 is defined by the **proportion of variability** in a data set that is accounted for by the statistical model (can be interpreted in terms of a percentage: how much variance of the data can be explained by the model).

$$\begin{aligned} y_i - \bar{y} &= (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \\ \underbrace{\sum_i (y_i - \bar{y})^2}_{SSTotal} &= \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{SSError} + \underbrace{\sum_i (y_i - \bar{y})^2}_{SSRegression} \end{aligned}$$

Thus, R^2 is given directly in terms of the explained variance. It compares the explained variance (variance of the models predictions) with the total variance of the data:

$$R^2 = \frac{SSRegression}{SSTotal} = 1 - \frac{SSError}{SSTotal}$$

- R^2 can get values between 0 (the regression line does not fit the data) and 1 (the regression line fits the data perfectly)

Coefficient of determination R^2 (2)

```
# obtain predicted values
y_hat <- lm(y ~ x)$fitted.values

SST   <- sum((y - mean(y)) ^ 2)
SSReg <- sum((y_hat - mean(y)) ^ 2)
SSE   <- sum((y - y_hat) ^ 2)
c(SST, SSReg, SSE)

     SST    SSReg      SSE
5296.849 1601.037 3695.812

R2 <- SSReg / SST
R2

0.3022621
```

Note: R^2 equals the **the square of the correlation coefficient** only for the **simple linear regression**.

```
> cor(y,x)^2
0.3022621
```

Relation to ANOVA (1)

The variance components are displayed in an analysis of variance table (c.f., Appendix 2 and Appendix Unit 3):

```
> anova(lm(yelp$average_stars ~ yelp$afinn_score))

Analysis of Variance Table

Response: yelp$average_stars
                  Df Sum Sq Mean Sq F value    Pr(>F)
yelp$afinn_score   1 131.16 131.159  466.87 < 2.2e-16 ***
Residuals        554 155.64   0.281
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F-test in the analysis of variance table checks if there are existing any predictors in the model that are meaningful. If the F-test yields a significant result ($p < .05$) there exists at least one predictor that is heaving a significant impact.

Multiple regression

- to develop a mathematical relationship between **two or more** independent variables and an (at least) interval scaled dependent variable.
- the general form is now described in the following way

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}_{n\times k} \times \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{k\times 1} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}_{n\times 1}$$

  with $i = 1, \dots, n$ and p the number of parameters to be estimated

- this can be summarized in matrix notation as

$$y = X\beta + e$$

- and the scalar notation is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + e_i$$

Multiple regression (2)

The OLS solution in this case is obtained as follows

$$e = y - X\hat{\beta}$$

squaring of two vectors is done by

$$\begin{aligned} e'e &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= y'y - 2\hat{\beta}X'y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

First order condition

$$\frac{\partial e'e}{\partial \beta} = -2X'Y + 2X'X\hat{\beta} = 0$$

Normal equation

$$(X'X)\hat{\beta} = X'y$$

which is solved as

$$\begin{aligned} (X'X)^{-1}(X'X)\hat{\beta} &= (X'X)^{-1}X'y \\ I\hat{\beta} &= (X'X)^{-1}X'y \\ \hat{\beta} &= (X'X)^{-1}X'y \end{aligned}$$

Multiple regression (3)

Assumptions

1. metric dependent variable
2. linear relationship
3. residuals: $e_i \sum N(0, \sigma^2 I)$
   3.1 independent and normally distributed
   $=$ no relationship between subsequent residuals
   3.2 **constant variance** (**homoscedasticity**)
   $=$ dispersion is the same across all observations
4. attention to **outliers**
5. **no multicollinearity** between the independent variables

Multiple regression (4)

Example: Yelp Dataset (cont.)

Some words, like "wtf", successfully predict a negative review, others, like "damn", are often positive (e.g. "the roast beef was damn good!"). Some of the words that AFINN most underestimated included "die" ("the pork chops are to die for!"), and one of the words it most overestimated was "joke" ("the service is a complete joke!")

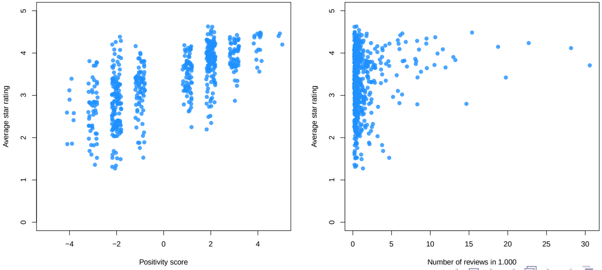
Research Question: can we account for misclassification by adding the word frequency in the number of reviews?

$$Av.Star\ Rating_i = \beta_0 + \beta_1 \times Positivity\ Score_i + \beta_2 \times No.\ reviews_i + e_i$$

$H_0 : \beta_i = 0$; there is no linear relationship from X_i on Y
$H_1 : \beta_i \neq 0$; there is a positive or negative linear relationship from X_i on Y

- **Assumption 2**: **linear relationsships** between the dependent and the independent variables (1: ✓, 2: “not exactly”)
  (Note: we divided the number of reviews by 1.000)

```
> plot(yelp$average_stars ~ jitter(yelp$afinn_score), ylim=c(0,5), xlim=c(-5,5),
+ pch=16, cex=1.25, col=adjustcolor("dodgerblue",0.8), xlab="Positivity score",
+ ylab="Average star rating")
> yelp$reviews2 <- yelp$reviews/1000
> plot(yelp$average_stars ~ yelp$reviews2, ylim=c(0,5), pch=16, cex=1.25,
+ col=adjustcolor("dodgerblue",0.8), xlab="Positivity score",
+ ylab="Number of reviews in 1.000")
```



Excursus: multicollinearity (1)

Multicollinearity is mathematically problematic because of $(X'X)^{-1}$

- If two variables are collinear they contain the same information about the dependent variable (i.e., two or more predictor variables in a multiple regression model are **highly correlated**), it may cause...
  1. wrong signs of the regression coefficients
  2. large changes in the estimated regression coefficients when a predictor variable is added or deleted.
  3. The regression coefficients are insignificant for the affected variables, but the F-test rejects the joint hypothesis (that those coefficients are all zero).

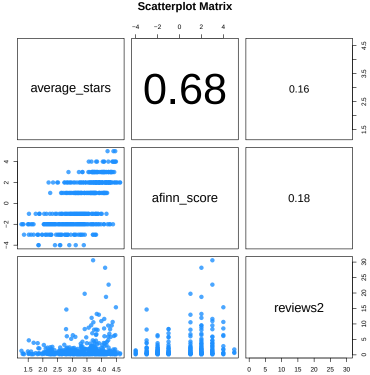
Some solutions in case of the presence of multicollinearity

- drop one of the variables in the regression
- conduct simple regressions instead of one multiple regression
- use factor scores which arise from exploratory factor analysis (uncorrelated by definition if orthogonal rotated)

Excursus multicollinearity (2)

- Multicollinearity can easily be checked by evaluating the correlation structure between the independent variables.
- In our case the variables are not correlated ($r = .18$).

```
> cor(cbind(yelp$reviews2,
+ yelp$afinn_score))

          [,1]      [,2]
[1,] 1.0000000 0.1822094
[2,] 0.1822094 1.0000000
```



Hypothesis testing

```
> mod2 <- lm(yelp$average_stars ~ yelp$afinn_score + yelp$reviews2)
> summary(mod2)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.297162   0.025104 131.342   <2e-16 ***
yelp$afinn_score 0.220308   0.010478  21.026   <2e-16 ***
yelp$reviews2    0.009448   0.007453   1.268    0.205
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Both relationships are positive but the impact of the number of reviews is not significant ($p = .205 > .05$).
- **General interpretation:** the number of reviews do not contribute to the model description accuracy of the average star rating.
- **Specific interpretation:** up to a constant baseline of 3.297 for the average star rating, the rating increases about 0.220 if the positivity score increases by 1 point, and increases about 0.009 for every 1.000 reviews added.

Adjusted R^2

- R^2 adjusted for the number of independent variables (= **model complexity**) in the equation and sample size (k is the number of coefficients, except the intercept)

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

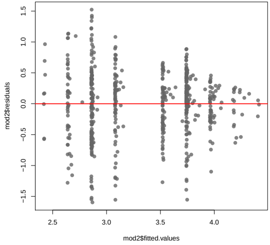
- R^2 increases only if the new terms added in the equation improve the model more than would be expected by chance
- adjusted R^2 can be negative (and it's value will always be less than or equal to R^2)

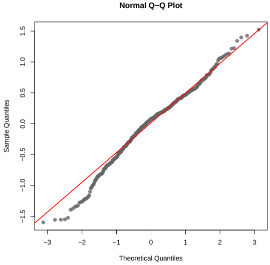
```
> summary(mod2)
Residual standard error: 0.5297 on 553 degrees of freedom
Multiple R-squared:  0.4589,    Adjusted R-squared:  0.4569
F-statistic: 234.5 on 2 and 553 DF,  p-value: < 2.2e-16

> summary(mod1)
Residual standard error: 0.53 on 554 degrees of freedom
Multiple R-squared:  0.4573,    Adjusted R-squared:  0.4563
F-statistic: 466.9 on 1 and 554 DF,  p-value: < 2.2e-16
```

The additional information did not improve the model fit although the multiple R^2 improved - **adjusted R^2s of the simple and the more complex models are almost equal**.

Assumption checks: residuals $e_i \sim N(0, \sigma^2 I)$

```
# 1. constant variance
> plot(mod2$residuals ~ mod2$fitted.values, col=adjustcolor("grey40",0.8),
+ cex=1.25, pch=16)
> abline(h=0, lwd=2, col="red")
# 2. normal distribution
> qqnorm(mod2$residuals, col=adjustcolor("grey40",0.8), cex=1.25, pch=16)
> qqline(mod2$residuals, col = "red", lwd=2)
```





Also the model residuals did not really change.

Relation to ANOVA (2)

Remember: If the F-test yields a significant result ($p < .05$) there exists at least one predictor that is heaving a significant impact.

```
> mod3 <- lm(yelp$average_stars ~ as.factor(yelp$afinn_score))
> anova(mod3)
                             Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(yelp$afinn_score)   8 132.52  16.565  58.735 < 2.2e-16 ***
Residuals                   547 154.27   0.282
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

one-factor ANOVA (= fixed effects model)

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip} + e_i$$

$$y_i = \begin{cases} \beta_0 & +\beta_1(x_{i1}=1) & & & & +e_i \\ \beta_0 & & +\beta_2(x_{i2}=1) & & & +e_i \\ \beta_0 & & & +\beta_3(x_{i3}=1) & & +e_i \\ \vdots & & & & \ddots & \vdots \\ \beta_0 & & & & +\beta_p(x_{ip}=1) & +e_i \end{cases}$$

Relation to ANOVA (3)

$$y = X\beta + e$$

X is a **design matrix** consisting of p **design vectors** of 0 (= effect is absent) and 1 (= effect is present). These design vectors are also called **dummy vectors**.

```
> model.matrix(mod3)
# rename the columns for a more compact output
> mm <- model.matrix(mod3)
> colnames(mm) <- paste("c",1:dim(mm)[2], sep="")
> head(mm)
  c1 c2 c3 c4 c5 c6 c7 c8 c9
1  1  0  0  0  0  1  0  0  0
2  1  0  0  0  1  0  0  0  0
3  1  0  0  0  1  0  0  0  0
4  1  0  1  0  0  0  0  0  0
5  1  0  1  0  0  0  0  0  0
6  1  0  0  0  1  0  0  0  0
...
attr(,"contrasts")
attr(,"contrasts")$`as.factor(yelp$afinn_score)`
[1] "contr.treatment"
```

Relation to ANOVA (4)

How does this work? Use the OLS solution

$$\widehat{\beta} = (X'X)^{-1}X'y$$

Note: matrix multiplication in R is carried out using $\% \star \%$.

```
> solve(t(mm) %*% mm) %*% t(mm) %*% yelp$average_stars

c1 2.58736717
c2 0.07852104
c3 0.23914525
c4 0.52963550
c5 0.89753056
c6 1.22717358
c7 1.33716560
c8 1.58118562
c9 1.76725046
```

Relation to ANOVA (5)

Contrast treatment: the mean of each group is

$$\beta_{0j} = \beta_0 + \beta_j$$

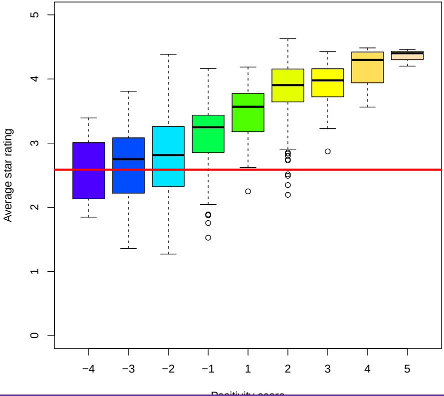
i.e., the difference of the overall mean and the group mean j.

We test the differences in the **group means** of the average star rating and the positivity score groups (starting with the lowest scoring group).

```
> summary(mod3)
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2.58737    0.18776  13.780  < 2e-16 ***
as.factor(yelp$afinn_score)-3  0.07852    0.20222   0.388  0.69796
as.factor(yelp$afinn_score)-2  0.23915    0.19324   1.238  0.21642
as.factor(yelp$afinn_score)-1  0.52964    0.19752   2.681  0.00755 **
as.factor(yelp$afinn_score)1   0.89753    0.19765   4.541 6.89e-06 ***
as.factor(yelp$afinn_score)2   1.22717    0.19305   6.357 4.35e-10 ***
as.factor(yelp$afinn_score)3   1.33717    0.20095   6.654 6.93e-11 ***
as.factor(yelp$afinn_score)4   1.58119    0.22996   6.876 1.69e-11 ***
as.factor(yelp$afinn_score)5   1.76725    0.35953   4.915 1.17e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Relation to ANOVA (6)

```
> boxplot(yelp$average_stars ~ as.factor(yelp$afinn_score),
+ col=topo.colors(9), ylim=c(0,5), xlab="Positivity score",
+ ylab="Average star rating")
> abline(h=2.58737, col="red", lwd=3)
```

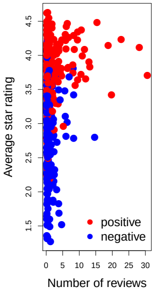


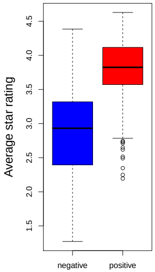

```
mean g=(-4) coefficients mean diff. group means
   2.587367   0.00000000   2.587367    2.587367
   2.587367  -0.07852104   2.665888    2.665888
   2.587367  -0.23914525   2.826512    2.826512
   2.587367  -0.52963550   3.117003    3.117003
   2.587367  -0.89753056   3.484898    3.484898
   2.587367  -1.22717358   3.814541    3.814541
   2.587367  -1.33716560   3.924533    3.924533
   2.587367  -1.58118562   4.168553    4.168553
   2.587367  -1.76725046   4.354618    4.354618
```

Relation to ANOVA (7)

Based on the result from the one-way ANOVA it might be sufficient to simply classify words into positive (positivity score >0) and negative (positivity score <0).

```
# build a color coding vector for each data point
> col. <- ifelse(yelp$afinn_score > 0, "red","blue")

# divide the plot into two parts
> par(mfrow=c(1,2))
> plot(yelp$average_stars ~ yelp$reviews2, col=col., pch=16,
+ cex=2, xlab="Number of reviews", ylab="Average star rating",
+ cex.lab=1.5)
> legend("bottomright", legend=c("positive","negative"),
+ col=c("red","blue"), pch=16, bty="n", cex=1.5)

# build a positive = 1, negative = 0 afinn score vector
> afin <- ifelse(yelp$afinn_score > 0, 1, 0)
> boxplot(yelp$average_stars ~ afin, col=c("blue","red"),
+ names=c("negative","positive"), ylab="Average star rating",
+ cex.lab=1.5)
```





Relation to ANOVA (7)

```
> mod3a <- lm(yelp$average_stars ~ afin)
> summary(mod3a)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.87070    0.03413   84.10   <2e-16 ***
afin         0.90544    0.04743   19.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The above regression model equals the simple **Student-t test**:

```
> t.test(yelp$average_stars ~ afin, var.equal=T)

t = -19.0912, df = 554, p-value < 2.2e-16

sample estimates:
mean in group 0 mean in group 1
       2.870700        3.776139
```

where the difference in group means equals the slope of the regression line.

See learning platform!

Submission deadline: 02. 04. at 23:00 pm

(via the learning platform `www.learn.wu.ac.at`)

Oral presentation of solutions on Monday!

(Recap: random selection of four students to present their solution).

Exam essentials

$$y_{ij} = \beta_0 + \beta_j + e_{ij}$$

with $e_{ij} \sim N(0, \sigma^2 I)$, $j = 1, \ldots, J$ the number of groups and $i = 1, \ldots, n_j$ the number of observations in group j.

$$\begin{aligned}
y_{ij} &= \bar{y} + (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j) \\
(y_{ij} - \bar{y}) &= (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j) \\
Dev.\ y_{ij}\ from\ overall &= (dev.\ group\ from\ overall) + (dev.\ y_{ij}\ from\ group) \\
(y_{ij} - \bar{y})^2 &= (\bar{y}_j - \bar{y})^2 + (y_{ij} - \bar{y}_j)^2 \\
\sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left((\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j) \right)^2 \\
\sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 &= \sum_{j=1}^{J} n_j (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \\
SS_{Total} &= SS_{Between} + SS_{Within}
\end{aligned}$$