Seminar 2

Martin Søyland

Disposisjon

1. Repetisjon
2. Korrelasjon
3. Bivariat OLS
4. Multipel OLS
5. Samspill og andregradsledd
6. Logistisk regresjon

Fokus: Regresjon og tolkning + litt visualisering!

Datasett

Sååå, samme datasett som sist. Husk at du må `setwd()` hver gang du åpner R (med mindre du bruker prosjekt). Data er passasjerer fra Titanic og variabler på om de overlevde, klasse, pris, osv. Dere kan enten laste ned data ved å skrive inn nettaddressen under i nettleseren og legge denne filen i mappen dere jobber fra:

```
setwd("~/Der/du/vil/jobbe/fra")


passengers <- read.csv("titanic.csv", stringsAsFactors = FALSE)
```

Jeg laster bare direkte inn fra linken. Legg merke til argumentet `stringsAsFactors = FALSE`. Dette står som default til `TRUE`. Argumentet konverterer alle variabler (kolonner) til klassen `factor()`, som er tilnærmet det samme som ordinalt målenivå – det vil vi ikke! Hvorfor vil vi ikke? Fordi vi vil ha lavest målenivå og heller sette det opp om vi finner ut at det gir mening, gitt data og det vi skal gjøre.

```
passengers <- read.csv("https://folk.uio.no/martigso/stv4020/titanic.csv", stringsAsFactors = FALSE)
class(passengers$Name)
```

```
## [1] "character"
```

Jobbe med variabler i dataset

Helt kort, noen av funksjonene vi gikk gjennom sist, som er viktige å bruke når man har et datasett man ikke kjenner.

```
class(passengers)
```

```
## [1] "data.frame"
```

```r
head(passengers)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp
## 1                            Braund, Mr. Owen Harris   male  22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
## 3                             Heikkinen, Miss. Laina female  26     0
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
## 5                           Allen, Mr. William Henry   male  35     0
## 6                                   Moran, Mr. James   male  NA     0
##   Parch           Ticket    Fare Cabin Embarked
## 1     0        A/5 21171  7.2500              S
## 2     0         PC 17599 71.2833   C85        C
## 3     0 STON/O2. 3101282  7.9250              S
## 4     0           113803 53.1000  C123        S
## 5     0           373450  8.0500              S
## 6     0           330877  8.4583              Q
```

```r
tail(passengers)
```

```
##     PassengerId Survived Pclass                                     Name
## 886         886        0      3     Rice, Mrs. William (Margaret Norton)
## 887         887        0      2                     Montvila, Rev. Juozas
## 888         888        1      1              Graham, Miss. Margaret Edith
## 889         889        0      3 Johnston, Miss. Catherine Helen "Carrie"
## 890         890        1      1                     Behr, Mr. Karl Howell
## 891         891        0      3                       Dooley, Mr. Patrick
##        Sex Age SibSp Parch     Ticket   Fare Cabin Embarked
## 886 female  39     0     5     382652 29.125              Q
## 887   male  27     0     0     211536 13.000              S
## 888 female  19     0     0     112053 30.000   B42        S
## 889 female  NA     1     2 W./C. 6607 23.450              S
## 890   male  26     0     0     111369 30.000  C148        C
## 891   male  32     0     0     370376  7.750              Q
```

```r
colnames(passengers)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
```

```
## [11] "Cabin"       "Embarked"
```

```
summary(passengers)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare            Cabin
##  Length:891         Min.   :  0.000   Length:891
##  Class :character   1st Qu.:  7.896   Class :character
##  Mode  :character   Median : 14.454   Mode  :character
##                     Mean   : 32.099
##                     3rd Qu.: 30.848
##                     Max.   :512.329
##                     NA's   :4
##    Embarked
##  Length:891
##  Class :character
##  Mode  :character
##
##
##
##
```

```
summary(passengers$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.42   20.12   28.00   29.70   38.00   80.00     177
```

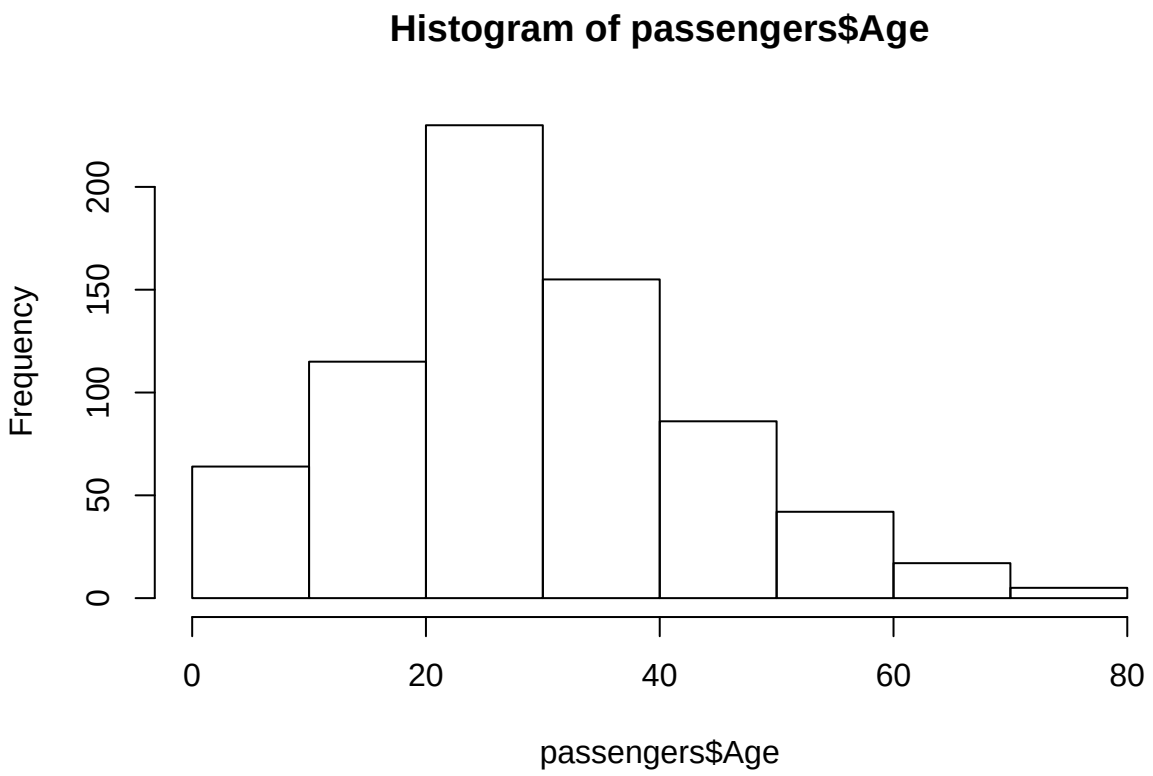
```
mean(passengers$Survived)
```

```
## [1] 0.3838384
```

```r
table(passengers$Pclass)
```

```
##
##   1   2   3
## 216 184 491
```

```r
hist(passengers$Age)
```



Vil også repetere litt på missing. Dette er viktig å forstå! Vi kan ikke bruke data vi ikke har...

```r
mean(passengers$Age)
```

```
## [1] NA
```

```r
table(is.na(passengers$Age))
```

```
##
## FALSE  TRUE
##   714   177
```

```r
mean(passengers$Age, na.rm = TRUE)
```

```
## [1] 29.69912
```

Litt omkoding

Ofte er vi heller ikke fornøyd med hvordan data er strukturert. Her er en av hovedfordelene med R; vi kan gjøre så og si hva som helst for å få dataene i det formatet vi ønsker. La oss si at vi, for eksempel, har en hypotese om at eldre personer hadde mindre sannsynlighet for å overleve enn yngre personer. Som dere husker fra forelesning :) kan det være lurt å sentrere variabler som alder fordi vi sjelden har et naturlig nullpunkt, som igjen gjør at konstantleddet i en evt regresjon ikke gir substansiell mening. La oss derfor sentrere alder:

```
median(passengers$Age, na.rm = TRUE)
```

```
## [1] 28
```

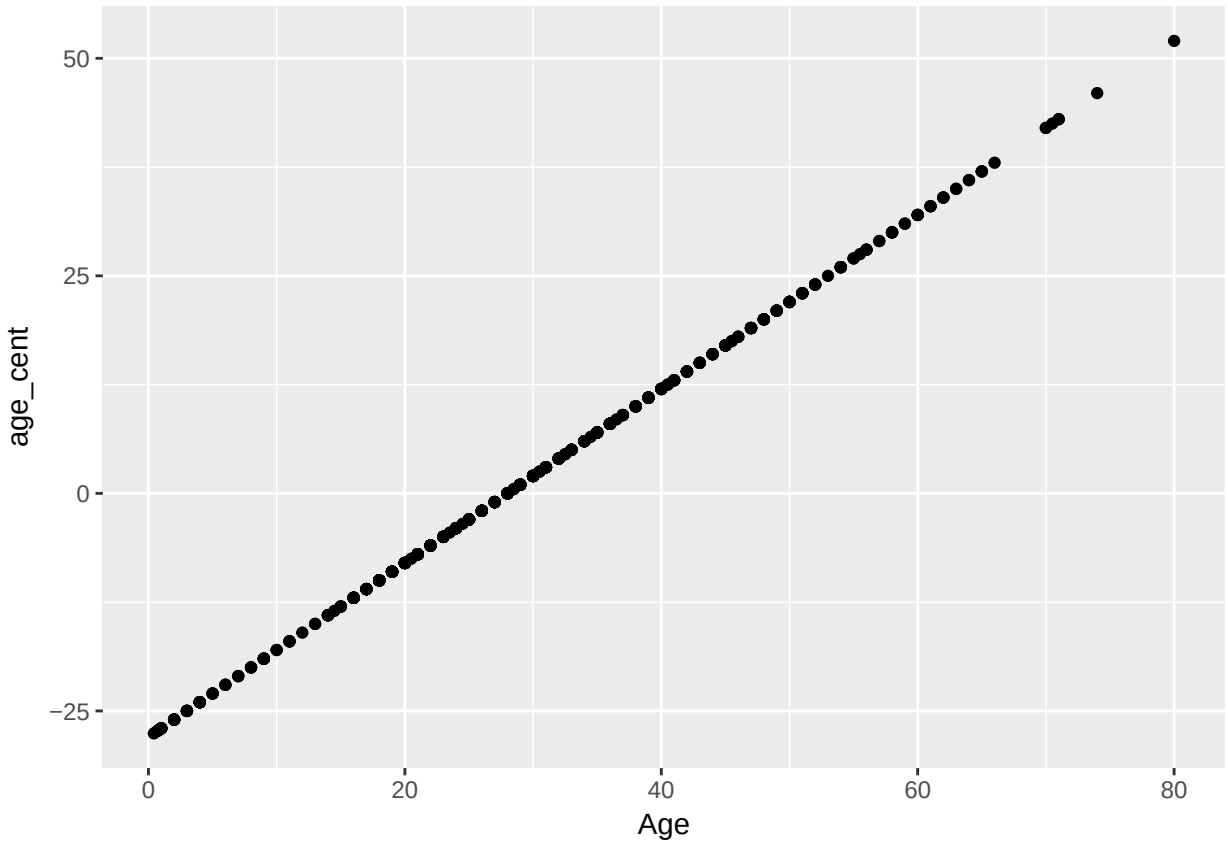
```
passengers$age_cent <- passengers$Age - median(passengers$Age, na.rm = TRUE)
```

Dette er en veldig god anledning til å se litt på **pakker**. R har nemlig et helt insane stort *open source* bibliotek med brukerlagede pakker alle har lov å bruke. Vi installerer en pakke med funksjonen `install.packages()` (husk å ha pakkenavnet i hermetegn her). Det er faktisk ikke nok å bare installer pakken, vi må også pakke den opp. Det gjør vi med `library()`. Pakken vil da være lastet *inn* til du avslutter R-sessionen du har åpen. Sååååå, "ggplot2" er en pakke for å lage grafikk, som vi kommer til å bruke mye (R har også en innebygd grafikk-fuksjon: `plot()`).

Nedenfor sjekker jeg om omkodingen vi gjorde er riktig. Syns dere det ser sånn ut?

```
# install.packages(ggplot2)
library(ggplot2)
ggplot(passengers, aes(x = Age, y = age_cent)) + geom_point()
```

```
## Warning: Removed 177 rows containing missing values (geom_point).
```



Vi kan også gjøre grafikken mye finere, men denne figuren vil ikke bli brukt i et evt paper. Så det er greit at den er litt quick and dirty. Kommer tilbake til det senere.

Korrelasjon

La oss også sjekke korrelasjonen mellom to av variablene våre. Her bruker vi funksjonen `cor()` for bare korrelasjonsestimat, og `cor.test()` for å se om estimatet er signifikant forskjellig fra null:

```
cor(passengers$age_cent, passengers$Survived)
```

```
## [1] NA
```

```
cor(passengers$age_cent, passengers$Survived, use = "complete.obs")
```

```
## [1] -0.07722109
```

```
cor.test(passengers$age_cent, passengers$Survived, use = "complete.obs")
```

```
##
##  Pearson's product-moment correlation
##
## data:  passengers$age_cent and passengers$Survived
## t = -2.0667, df = 712, p-value = 0.03912
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.149744955 -0.003870727
## sample estimates:
##         cor
## -0.07722109
```

Også her må vi håndtere missingverdier (ref første linje over). Men med korrelasjon er det, som dere vet, forskjellige måter å håndere missing på: pairwise og listwise exclusion. Dette er ikke viktig med korrelasjon mellom bare to variabler, men med flere variabler er det viktig:

```
cor(passengers[, c("age_cent", "Survived", "Fare")], use = "complete.obs")

cor(passengers[, c("age_cent", "Survived", "Fare")], use = "pairwise.complete.obs")
```

```
##             age_cent    Survived       Fare
## age_cent  1.00000000 -0.07692265 0.09638814
## Survived -0.07692265  1.00000000 0.27128592
## Fare      0.09638814  0.27128592 1.00000000
##             age_cent    Survived       Fare
## age_cent  1.00000000 -0.07722109 0.09638814
## Survived -0.07722109  1.00000000 0.25965960
## Fare      0.09638814  0.25965960 1.00000000
```

Bivariat OLS

OLS er veldig enkelt å kjøre i R (alle typer analyser er ganske enkle, egentlig). Vi bruker funksjonen `lm()`, som står for *linear model*. Her er avhengig variabel *Survived*, og uavhenig variabel *age_cent*. Vi skiller mellom AV og UV med en tilde: ~ . Sjekk ut hjelpefilen `?lm`.

```
pass_reg <- lm(Survived ~ age_cent, data = passengers)
summary(pass_reg)
```

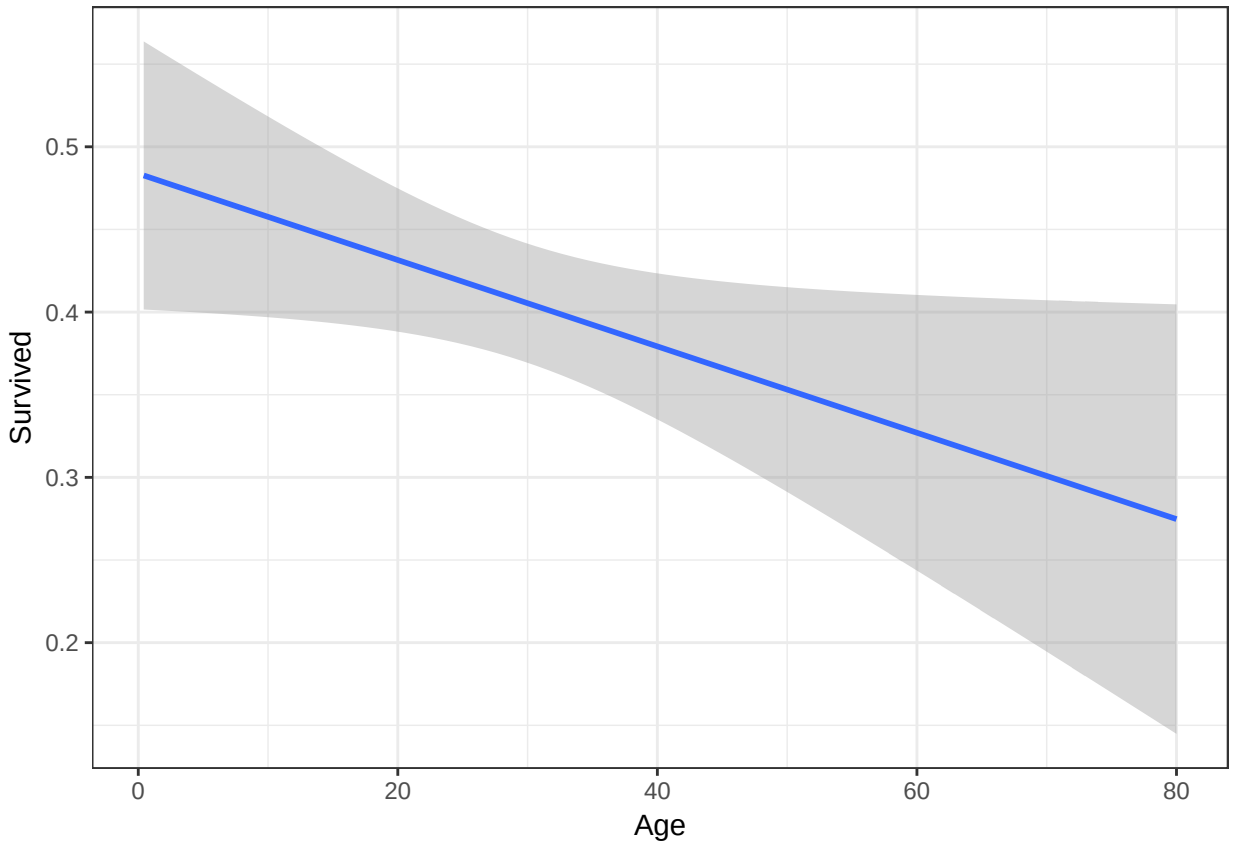
```
##
## Call:
## lm(formula = Survived ~ age_cent, data = passengers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4811 -0.4158 -0.3662  0.5789  0.7252
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.410601   0.018476  22.224   <2e-16 ***
## age_cent    -0.002613   0.001264  -2.067   0.0391 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4903 on 712 degrees of freedom
##   (177 observations deleted due to missingness)
## Multiple R-squared:  0.005963,   Adjusted R-squared:  0.004567
## F-statistic: 4.271 on 1 and 712 DF,  p-value: 0.03912
```

Her er det lurt å øve seg på å tolke hva resultatene betyr. Vi kan også gjøre en enkel visualisering med *ggplot* når vi har en binær regresjon.

```
# install.packages('ggplot')
library(ggplot2)
theme_set(theme_bw())
ggplot(passengers, aes(x = Age, y = Survived)) + geom_smooth(method = "lm")
```

```
## Warning: Removed 177 rows containing non-finite values (stat_smooth).
```



Kan dere tenke dere noen variabler som vi burde inkludere i denne regresjonen?

Multipel OLS

"Women and children", right:

```
pass_reg2 <- lm(Survived ~ age_cent + Sex, data = passengers)
summary(pass_reg2)
```

```
##
## Call:
## lm(formula = Survived ~ age_cent + Sex, data = passengers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7786 -0.2115 -0.1931  0.2471  0.8401
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.7547117  0.0256497  29.424   <2e-16 ***
## age_cent    -0.0009206  0.0010730  -0.858    0.391
## Sexmale     -0.5469036  0.0323428 -16.910   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4144 on 711 degrees of freedom
##   (177 observations deleted due to missingness)
## Multiple R-squared:  0.2911, Adjusted R-squared:  0.2891
## F-statistic:   146 on 2 and 711 DF,  p-value: < 2.2e-16
```

Og noen personer er viktiger enn andre. . . :

```
pass_reg3 <- lm(Survived ~ age_cent + Sex + factor(Pclass), data = passengers)
summary(pass_reg3)
```

```
##
## Call:
## lm(formula = Survived ~ age_cent + Sex + factor(Pclass), data = passengers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11410 -0.25081 -0.06422  0.23015  1.00676
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.972139   0.033730  28.821  < 2e-16 ***
## age_cent        -0.005460   0.001084  -5.039 5.96e-07 ***
## Sexmale         -0.479456   0.030718 -15.608  < 2e-16 ***
```

```
## factor(Pclass)2 -0.207747   0.041689  -4.983 7.86e-07 ***
## factor(Pclass)3 -0.406618   0.038288 -10.620  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3849 on 709 degrees of freedom
##   (177 observations deleted due to missingness)
## Multiple R-squared:  0.3902, Adjusted R-squared:  0.3867
## F-statistic: 113.4 on 4 and 709 DF,  p-value: < 2.2e-16
```

For å lage andregradsledd er det to alternativer, her er ett: (det andre er å bruke funksjonen `poly()`)

Andregradsledd (polynomer)

```
passengers$age_cent_andregrad <- passengers$age_cent^2

andregrads_reg <- lm(Survived ~ age_cent + age_cent_andregrad + Sex + factor(Pclass),
                     data = passengers)

# andregrads_reg <- lm(Survived ~ poly(age_cent, 2, raw = TRUE) + Sex + factor(Pclass),
#                      data = passengers[which(is.na(passengers$age_cent) == FALSE), ])

summary(andregrads_reg)
```

```
##
## Call:
## lm(formula = Survived ~ age_cent + age_cent_andregrad + Sex +
##     factor(Pclass), data = passengers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15891 -0.24944 -0.05217  0.23243  1.00981
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         9.564e-01  3.553e-02  26.920  < 2e-16 ***
## age_cent           -6.034e-03  1.158e-03  -5.211 2.47e-07 ***
## age_cent_andregrad  6.745e-05  4.820e-05   1.399    0.162
## Sexmale            -4.798e-01  3.070e-02 -15.629  < 2e-16 ***
## factor(Pclass)2    -2.042e-01  4.174e-02  -4.891 1.24e-06 ***
## factor(Pclass)3    -4.034e-01  3.833e-02 -10.523  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.3846 on 708 degrees of freedom
##   (177 observations deleted due to missingness)
## Multiple R-squared:  0.3919, Adjusted R-squared:  0.3876
## F-statistic: 91.24 on 5 and 708 DF,  p-value: < 2.2e-16
```

```
# plot(andregrads_reg)
```

Logistisk regresjon

Logistisk regresjon er veldig likt i oppbygning. Det er i familien **general linearized models** (`glm()`). Det viktige her er argumentet `family = "binomial"`, som spesifiserer at vi snakker om en binær avhengig variabel – kan også skrive `binomial(link = "logit")`.

```
pass_logit <- glm(Survived ~ age_cent + Sex + factor(Pclass),
                  data = passengers, family = "binomial")

summary(pass_logit)
```

```
##
## Call:
## glm(formula = Survived ~ age_cent + Sex + factor(Pclass), family = "binomial",
##     data = passengers)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7303  -0.6780  -0.3953   0.6485   2.4657
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       2.741425   0.266050  10.304  < 2e-16 ***
## age_cent         -0.036985   0.007656  -4.831 1.36e-06 ***
## Sexmale          -2.522781   0.207391 -12.164  < 2e-16 ***
## factor(Pclass)2 -1.309799   0.278066  -4.710 2.47e-06 ***
## factor(Pclass)3 -2.580625   0.281442  -9.169  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 647.28  on 709  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 657.28
##
```

```
## Number of Fisher Scoring iterations: 5
```

Neste gang:

- Mer wrangling
- Samspill
- Diagnostisering
- Plotte effekter med multipel regresjon
- Ønsker?

Bonus for LaTeX-elskere:

```
# install.packages("stargazer")
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
stargazer(pass_reg, pass_reg2, pass_reg3, pass_logit,
          star.cutoffs = c(.05, .01, .001),
          column.sep.width = ".01cm",
          no.space = FALSE,
          covariate.labels = c("Alder (sentrert)", "Kjønn (mann)",
                               "Klasse (2)", "Klasse (3)", "Konstantledd"),
          keep.stat = c("n", "rsq", "adj.rsq", "ll"))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: sø., sep. 10, 2017 - kl. 20.14 +0200

Table 1:

	Dependent variable:			
	Survived			
	OLS			*logistic*
	(1)	(2)	(3)	(4)
Alder (sentrert)	−0.003*	−0.001	−0.005***	−0.037***
	(0.001)	(0.001)	(0.001)	(0.008)
Kjønn (mann)		−0.547***	−0.479***	−2.523***
		(0.032)	(0.031)	(0.207)
Klasse (2)			−0.208***	−1.310***
			(0.042)	(0.278)
Klasse (3)			−0.407***	−2.581***
			(0.038)	(0.281)
Konstantledd	0.411***	0.755***	0.972***	2.741***
	(0.018)	(0.026)	(0.034)	(0.266)
Observations	714	714	714	714
R^2	0.006	0.291	0.390	
Adjusted R^2	0.005	0.289	0.387	
Log Likelihood				−323.642
Note:				*p<0.05; **p<0.01; ***p<0.001