oppgavene 4C1 og 4C3 fra Wooldridge

```
library(wooldridge)
library(car)
```

```
## Loading required package: carData
```

```
library(multcomp)
```

```
## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:wooldridge':
##
##     cement

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

Dette er viktig stoff og er dekket i avsnittene 4-2c, 4-4 og 4-5. Viser først eksemplene fra disse avsnittene løst vha. R.

Eksempler løst i R

Fra avsnitt 4-2c

Ex. 4.4

Modellen

$$\mathsf{log(crime)} = \beta_0 + \beta_1 \mathsf{log(enroll)} + u$$

skriver vi i R som "log(crime) ~ log(enroll)" konstantleddet kommer automatisk. Skulle vi ønske *uten* konstantledd (generelt frarådet) skriver vi "log(crime) ~ log(enroll) - 1"

```r
data(campus)
# cc crime campus, finnes også variablene lcrime og lenroll i datasettet der
# ln av variablene alt er tatt. I R er det like lett å bruke funksjonen selv
mod_cc <- "log(crime) ~ log(enroll)"
lm_cc <- lm(mod_cc, data=campus)
# uten konstantledd
# mod_cc <- "log(crime) ~ log(enroll) - 1"
```

```r
summary(lm_cc)
```

```
##
## Call:
## lm(formula = mod_cc, data = campus)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5136 -0.3858  0.1174  0.4363  2.5782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.6314     1.0335  -6.416 5.44e-09 ***
## log(enroll)   1.2698     0.1098  11.567  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8946 on 95 degrees of freedom
## Multiple R-squared:  0.5848, Adjusted R-squared:  0.5804
## F-statistic: 133.8 on 1 and 95 DF,  p-value: < 2.2e-16
```

Elastisiteten for `campus crime` mht. `enroll` (universitetsstørrelse) er altså estimert til $1,2698 \approx 1,27$, dvs. 1% økning i antall studenter gir 1,27% økning i kriminalitet. Et sentralt punkt her er om denne elastisiteten er signifikant større enn 1. Er den det vil vi ha en *relativ* økning i kriminalitet når størrelsen øker. Altså at et dobbelt så stort universitet vil ha *mer enn* dobbelt så høy kriminalitet.

Ønsker å test H0: $\beta_1 = 1$ mot H1: $\beta_1 > 1$. I summary ovenfor er det hypotesene H0: $\beta_1 = 0$ mot H1: $\beta_1 > 1$ som blir testet. Hva gjør vi? Vi regner ut ny t-verdi vha. $\frac{\text{estimat - verdi i hypotese}}{\text{standard error}}$, dvs $\frac{1,2698-1}{0,1098} = 2,457$.

Vi må så finne kritisk verdi eller p-verdi (husk ensidig H1 her)

```r
# obs. in campus
dim(campus)
```

```
## [1] 97  7
```

```r
num_obs <- dim(campus)[1]
num_obs
```

```
## [1] 97
```

Antall frihetsgrader (95 DF) kan vi også lese direkte ut fra siste linje i summary ovenfor.

p-verdi (For å se fordelinger kjent av base R kjør `?distributions` i Console)

```r
# p-verdi ensidig H1
pt(2.457, df=num_obs-2, lower.tail = FALSE)
```

```
## [1] 0.007912419
```

kritisk verdi, ensidig H1 ulike α

```r
# alpha lik 0,05 ensidig
qt(0.05, df = num_obs-2)
```

```
## [1] -1.661052
```

```r
# alpha lik 0,05 ensidig
qt(0.01, df = num_obs-2)
```

```
## [1] -2.366243
```

Vi ser at β_1 er signifikant forskjellig fra 1 på nivå $\alpha = 1\%$, altså har vi en overproporsjonal økning i kriminalitet når universitetsstørrelsen øker.

For ordens skyld: Wooldrige skriver en del om å lage konfidensintervall. For å finne konfidensintervall for modellen ovenfor gjør vi følgende

```r
# default 5%
confint(lm_cc)
```

```
##                 2.5 %    97.5 %
## (Intercept) -8.683207 -4.579534
## log(enroll)  1.051827  1.487693
```

```r
# 1%
confint(lm_cc, level=0.99)
```

```
##                  0.5 %    99.5 %
## (Intercept) -9.3481083 -3.914632
## log(enroll)  0.9812058  1.558315
```

Ex. 4.5

```r
# Boston housing data; hprice2. Se book or give command ?hprice2 in console
# for description of the variables
data(hprice2)
mod_hp <- "log(price) ~ log(nox) + log(dist) + rooms + stratio"
lm_hp <- lm(mod_hp, data=hprice2)
```

```r
summary(lm_hp)
```

```
##
## Call:
## lm(formula = mod_hp, data = hprice2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05890 -0.12427  0.02128  0.12882  1.32531
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.083861   0.318111  34.843  < 2e-16 ***
## log(nox)    -0.953539   0.116742  -8.168 2.57e-15 ***
## log(dist)   -0.134339   0.043103  -3.117  0.00193 **
## rooms        0.254527   0.018530  13.736  < 2e-16 ***
## stratio     -0.052451   0.005897  -8.894  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.265 on 501 degrees of freedom
## Multiple R-squared:  0.584,  Adjusted R-squared:  0.5807
## F-statistic: 175.9 on 4 and 501 DF,  p-value: < 2.2e-16
```

Now β_1 (coefficient estimate for log(nox) lik -0,9535) er priselastisiteten for boliger mht. nox utslipp. Vi ønsker å teste H0: $\beta_1 = -1$ mot H1: $\beta_1 \neq -1$. Vi benytter samme teknikk som ovenfor, men husker at nå er alternativ hypotese tosidig. Vi regner altså ut ny t-verdi vha. $\frac{\text{estimat - verdi i hypotese}}{\text{standard error}}$, dvs $\frac{-0{,}9535-(-1)}{0{,}1167} = \frac{0{,}0465}{0{,}1167} = 0,3985$. Finner p-verdi og kritiske t-verdier. Antall frihetsgrader er 501.

```r
# p-verdi tosidig H1
2*pt(0.3985, df=501, lower.tail=FALSE)
```

```
## [1] 0.6904314
```

Kritisk t-verdi 5% nivå

```r
# alpha lik 0,05 tosidig
qt(0.05/2, df = 501, lower.tail = FALSE)
```

```
## [1] 1.96471
```

Vi kan altså ikke forkaste H0, dvs. det er lite bevis for at β_1 er forskjellig fra -1.

Fra avsnitt 4-4

Testing av hypoteser med én lineær kombinasjon av parametre.

Modellen

$$\mathsf{log(wage)} = \beta_0 + \beta_1 \mathsf{jc} + \beta_2 \mathsf{univ} + \beta_3 \mathsf{exper} + u$$

```r
data(twoyear)
#variable lwage in dataset is log(wage)
mod_2y <- "lwage ~ jc + univ +exper"
lm_2y <- lm(mod_2y, data = twoyear)
```

```r
summary(lm_2y)
```

```
##
## Call:
## lm(formula = mod_2y, data = twoyear)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.10362 -0.28132  0.00551  0.28518  1.78167
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.4723256  0.0210602  69.910   <2e-16 ***
## jc          0.0666967  0.0068288   9.767   <2e-16 ***
## univ        0.0768762  0.0023087  33.298   <2e-16 ***
## exper       0.0049442  0.0001575  31.397   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4301 on 6759 degrees of freedom
## Multiple R-squared:  0.2224, Adjusted R-squared:  0.2221
## F-statistic: 644.5 on 3 and 6759 DF,  p-value: < 2.2e-16
```

Vi ser at det lønner seg både med junior college (jc) og college (univ), begge koeffisientene er signifikant forskjellig fra 0. Det vi er mest interessert i er om β_{univ} er *signifikant* større enn β_{jc}. Altså om det *lønner seg* å velge universitet fremfor junior college.

For å test trenger vi å beregne t-verdien $t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\mathsf{se}(\hat{\beta}_1 - \hat{\beta}_2)}$. Problemet er at vi *ikke* finner $\mathsf{se}(\hat{\beta}_1 - \hat{\beta}_2)$ i standard rapporten for regresjon.

NB! $\mathsf{se}(\hat{\beta}_1 - \hat{\beta}_2) \neq \mathsf{se}(\hat{\beta}_1) - \mathsf{se}(\hat{\beta}_2)$

Vi bruker derfor et «triks» der vi skriver om modellen slik at $\mathsf{se}(\hat{\beta}_1 - \hat{\beta}_2)$ vil bli rapportert i standard `summary` fra modellen.

Definerer en ny parameter $\theta_1 = \beta_1 - \beta_2$ som gir at $\beta_1 = \theta_1 + \beta_2$. Det vi ønsker å teste er H0: $\theta = 0$ mot H1: $\theta < 0$. Vi er nå på jakt etter $\mathsf{se}(\theta_1)$ som vil være lik $\mathsf{se}(\hat{\beta}_1 - \hat{\beta}_2)$. Vi kan da skrive om modellen som

$$\mathsf{log}(\mathsf{wage}) = \beta_0 + \beta_1 \mathsf{jc} + \beta_2 \mathsf{univ} + \beta_3 \mathsf{exper} + u = \beta_0 + (\theta_1 + \beta_2)\mathsf{jc} + \beta_2 \mathsf{univ} + \beta_3 \mathsf{exper} + u$$

som gir

$$\mathsf{log}(\mathsf{wage}) = \beta_0 + \theta_1 \mathsf{jc} + \beta_2(\mathsf{univ} + \mathsf{jc}) + \beta_3 \mathsf{exper} + u$$

Vi kan altså få tak i $\mathsf{se}(\hat{\beta}_1 - \hat{\beta}_2)$ ved å kjøre modellen

```r
# Legg merke til bruk av I() funksjonen. Denne trengs siden + har  en spesiell
# betydning i R sitt formula «språk». Inne i I() blir det summen av univ og exper for
# hver student
mod_2y_b <- "lwage ~ jc + I(univ + jc) + exper"
lm_2y_b <- lm(mod_2y_b, data = twoyear)
```

```
summary(lm_2y_b)
```

```
##
## Call:
## lm(formula = mod_2y_b, data = twoyear)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.10362 -0.28132  0.00551  0.28518  1.78167
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.4723256  0.0210602  69.910   <2e-16 ***
## jc           -0.0101795  0.0069359  -1.468    0.142
## I(univ + jc)  0.0768762  0.0023087  33.298   <2e-16 ***
## exper         0.0049442  0.0001575  31.397   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4301 on 6759 degrees of freedom
## Multiple R-squared:  0.2224, Adjusted R-squared:  0.2221
## F-statistic: 644.5 on 3 and 6759 DF,  p-value: < 2.2e-16
```

Da kan vi lese ut standard error for θ_1, som jo også er standard error for $(\hat{\beta}_1 - \hat{\beta}_2)$ som var det vi var på jakt etter. Da kan vi enkelt regne ut t-verdien

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\mathsf{se}(\hat{\beta}_1 - \hat{\beta}_2)} = \frac{-0,01018}{0,00694} \approx -1,467$$

Finner p-verdi (ensidig)

```
# p-verdi ensidig H1
pt(-1.467, df=6759)
```

```
## [1] 0.07121129
```

Vi kan altså *ikke* på 5% nivå konkludere med at et år utdanning på college gir signifikant høyere lønn enn et år på junior college. På 10% nivå derimot kan vi konkludere med at et år på college gir signifikant mer uttelling i lønn enn et år på junior college.

Teknikken ovenfor kan vi *alltid* få til i et statistikkprogram. Finnes imidlertid pakker/rutiner som forsøker å forenkle dette. To slike, *car* og *multcomp* er vist nedenfor. Begge bruker F-test (istdenfor t-test) som samsvarer mer med avsnitt 4-5, men konklusjonene blir de samme.

Med pakken car Enklere måte (bruker F-test jmf. avsnitt 4-5)

```
# vi har lastet car så linearHypothesis er tilgjengelig
linearHypothesis(lm_2y, "jc - univ = 0")
```

```
## Linear hypothesis test
##
## Hypothesis:
## jc - univ = 0
##
## Model 1: restricted model
## Model 2: lwage ~ jc + univ + exper
##
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1   6760 1250.9
## 2   6759 1250.5  1   0.39853 2.154 0.1422
```

```
pf(2.154, 1, Inf, lower.tail = FALSE)
```

```
## [1] 0.142199
```

Som en ser er resultatet ovenfor tosidig H1, ønsker en ensidig H1: $\beta_1 < \beta_2$ blir $p = 0,142199/2 \approx 0,071$.

Kritisk verdi 10%, 5% og 1% nivå

```
# alpha 10% ensidig
qt(0.1, df = 6759, lower.tail = FALSE)
```

```
## [1] 1.281677
```

```
# alpha 5% ensidig
qt(0.05, df = 6759, lower.tail = FALSE)
```

```
## [1] 1.645079
```

```
# alpha 1% ensidig
qt(0.01, df = 6759, lower.tail = FALSE)
```

```
## [1] 2.3269
```

Vi ser at vi kan forkast H0 på 10% nivå, men ikke på 5% nivå.

Med pakken `multcomp` Med pakken `multcomp` er det enkelt å formulere ensidige hypoteser også. Denne er kanskje den enkleste å bruke.

```
library(multcomp)
# Specify the linear hypothesis
glht_mod <- glht(
  model = lm_2y,
  linfct = c("jc - univ <= 0")
)

# Inspect summary
summary(glht_mod)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = mod_2y, data = twoyear)
##
## Linear Hypotheses:
##                 Estimate Std. Error t value Pr(>t)
## jc - univ <= 0 -0.010180   0.006936  -1.468  0.929
## (Adjusted p values reported -- single-step method)
```

$\text{Pr}(< t) = 1 - \text{Pr}(> t) = 1 - 0,929 = 0,071$

```
# Inspect confidence interval
confint(glht_mod)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = mod_2y, data = twoyear)
##
## Quantile = -1.6451
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                Estimate lwr      upr
## jc - univ <= 0 -0.01018 -0.02159      Inf
```

Gjennomgangseksemplet avsnitt 4-5

Baseball (mlb1)

```
data(mlb1)
mod_bb <- "log(salary) ~ years + gamesyr + bavg + hrunsyr + rbisyr"
lm_bb <- lm(mod_bb, data = mlb1)
```

Historien er da at variablene `bavg, hrunsyr, rbisyr` angir spillernes individuelle ferdigheter. Spørsmålet er om dette betyr noe for lønn eller om det bare er hvor lenge en har spillet (year) og gjennomsnittlig antall kamper per år en har fått spillet som bestemmer lønnsnivået.

```
summary(lm_bb)
```

```
##
## Call:
## lm(formula = mod_bb, data = mlb1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.02508 -0.45034 -0.04013  0.47014  2.68924
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.119e+01  2.888e-01  38.752  < 2e-16 ***
## years       6.886e-02  1.211e-02   5.684 2.79e-08 ***
## gamesyr     1.255e-02  2.647e-03   4.742 3.09e-06 ***
## bavg        9.786e-04  1.104e-03   0.887    0.376
## hrunsyr     1.443e-02  1.606e-02   0.899    0.369
## rbisyr      1.077e-02  7.175e-03   1.500    0.134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7266 on 347 degrees of freedom
## Multiple R-squared:  0.6278, Adjusted R-squared:  0.6224
## F-statistic: 117.1 on 5 and 347 DF,  p-value: < 2.2e-16
```

Det vi ønsker å teste er om individuelle ferdigheter er overflødig i modellen, dvs om $\beta_3 = 0$, $\beta_4 = 0$ og $\beta_5 = 0$.

Læreboken gjør dette «manuelt» vha. SSR fra restricted og unrestricted model.

```
mod_bb_r <- "log(salary) ~ years + gamesyr"
lm_bb_r <- lm(mod_bb_r, data = mlb1)
```

```
summary(lm_bb_r)
```

```
##
## Call:
## lm(formula = mod_bb_r, data = mlb1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.66858 -0.46412 -0.01177  0.49219  2.68829
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.223804   0.108312 103.625  < 2e-16 ***
## years        0.071318   0.012505   5.703  2.5e-08 ***
## gamesyr      0.020174   0.001343  15.023  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7527 on 350 degrees of freedom
## Multiple R-squared:  0.5971, Adjusted R-squared:  0.5948
## F-statistic: 259.3 on 2 and 350 DF,  p-value: < 2.2e-16
```

```
(ssr_u <- sum(residuals(lm_bb)^2))
```

```
## [1] 183.1863
```

```
(ssr_r <- sum(residuals(lm_bb_r)^2))
```

```
## [1] 198.3115
```

F verdien blir da (n=353 obs, k =5 og q = 3

```r
(F_bb <- (ssr_r -ssr_u)/ssr_u * ((353-5-1)/3))
```

```
## [1] 9.550254
```

F-verdien kan så sjekkes opp mot tabell eller

```r
pf(9.5503, 3, 347, lower.tail = FALSE)
```

```
## [1] 4.473429e-06
```

Kritisk verdi 1%

```r
qf(c(0.005, 0.995), 3, 347)
```

```
## [1] 0.02387536 4.35317230
```

Vi ser at vi kan forkaste hypotesen om at individuelle ferdigheter ikke har betydning for lønnen.

```r
linearHypothesis(lm_bb, "bavg + hrunsyr + rbisyr = 0")
```

Samme med bruk av `linearHypothesis`

```
## Linear hypothesis test
##
## Hypothesis:
## bavg  + hrunsyr  + rbisyr = 0
##
## Model 1: restricted model
## Model 2: log(salary) ~ years + gamesyr + bavg + hrunsyr + rbisyr
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    348 186.50
## 2    347 183.19  1    3.3149 6.2793 0.01267 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Specify the linear hypothesis
glht_mod <- glht(
  model = lm_bb,
  linfct = c("bavg + hrunsyr + rbisyr = 0")
)

# Inspect summary
summary(glht_mod)
```

Samme med bruk av `multcomp`

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = mod_bb, data = mlb1)
##
## Linear Hypotheses:
##                               Estimate Std. Error t value Pr(>|t|)
## bavg + hrunsyr + rbisyr == 0  0.02617    0.01045   2.506   0.0127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Den ene bruker t-verdi, den andre F-verdi men konklusjonen blir den samme.

Oppgavene

Oppgave 4C1

i. Tolkning β_1?

   Holder alle andre variabler enn expendA fast. Gir oss

   $$\Delta \mathsf{VoteA} = \beta_1 \mathsf{log(expendA)} = \frac{\beta_1}{100}(100\Delta \mathsf{log(expendA)}) \approx \frac{\beta_1}{100}\%\Delta \mathsf{expendA}$$

   Altså gir β_1 oss tilnærmet antall prosentpoeng økning i voteA når expendA øker med 1%. Altså antall prosentpoeng økning (f.eks fra 12,1% til 12,7%, dvs. 0,6 prosentpoeng økning) når vi øker expendA med 1% (f.eks fra 20 millioner til $20 \cdot 1,01 = 20,2$ millioner).

   Eksempeltallene er selvsagt tatt rett ut av løse luften som en illustrasjon. Et viktig poeng er at den første størrelsen er *prosentpoeng* mens den andre er en relativ størrelse (*prosentvis endring*). Disse to begrepene blandes ofte.

ii. H0: $\beta_1 = -\beta_2$ eller H0: $\beta_1 + \beta\_2 = 0$. Altså har vi at hvis expendA og expendB økes med samme prosentvise størrelse (f.eks fra 10 mill. til 10,5 mill. for A og fra 30 mill. til 31,5 mill. for B) så vil As andel av stemmene være uendret.

   ```
   # load dataset vote1 from wooldridge package
   data(vote1)
   mod1 <- "voteA ~ log(expendA) + log(expendB) + prtystrA"
   lm1 <- lm(mod1, data=vote1)
   ```

```
summary(lm1)
```

```
##
## Call:
## lm(formula = mod1, data = vote1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.3968  -5.4174  -0.8679   4.9551  26.0660
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  45.07893    3.92631   11.48   <2e-16 ***
## log(expendA)  6.08332    0.38215   15.92   <2e-16 ***
## log(expendB) -6.61542    0.37882  -17.46   <2e-16 ***
## prtystrA      0.15196    0.06202    2.45   0.0153 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.712 on 169 degrees of freedom
## Multiple R-squared:  0.7926, Adjusted R-squared:  0.7889 
## F-statistic: 215.2 on 3 and 169 DF,  p-value: < 2.2e-16
```

Vi ser fra summaryat de estimerte koeffisientene for `log(expendA)` er 6.08332 og -6.61542 for `log(expendB)`. En økning på 1% i expendA vil altså gi $6{,}1/100 = 0{,}0608$ prosentpoeng økning i andelen stemmer for kandidat A. Likeledes vil 1% økning i expendB, alle andre variabler holdt fast, gi en reduksjon på $6{,}62/100 = 0{,}0662$ prosentpoeng i kandidat As andel av stemmene.