Løsningsforslag eksamen MET4 - VÅR 25

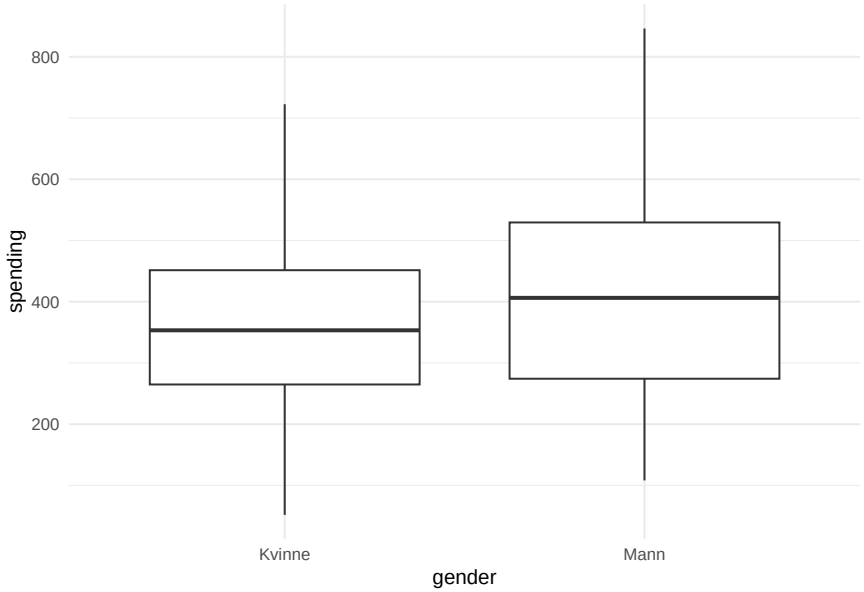
Til sensor: Hver deloppgave teller likt og gir maksimalt 10 poeng. Selv om R-kode gis i løsningsforslaget under skal kandidaten i utgangspunktet ikke legge ved noe R-kode i sin besvarelse, men det skal ikke gis noe trekk dersom de har gjort det. Det er også mulig å bruke R i Del 2, men da mer som en kalkulator og til å finne kritiske verdier. Vi gir ikke "stilpoeng" på hverken figurer eller ligninger. Så lenge sensor fortstår hva som menes/illustreres (og her er vi romslige) skal det gis full uttelling på den aktuelle delen av oppgaven.

Del 1 - Dataanalyse med R

Oppgave 1

a)

```
ggplot(survey, aes(x = gender, y = spending)) +
  geom_boxplot() +
  theme_minimal()
```



Det er særlig to iøyefallende karakteristikker ved denne figuren:

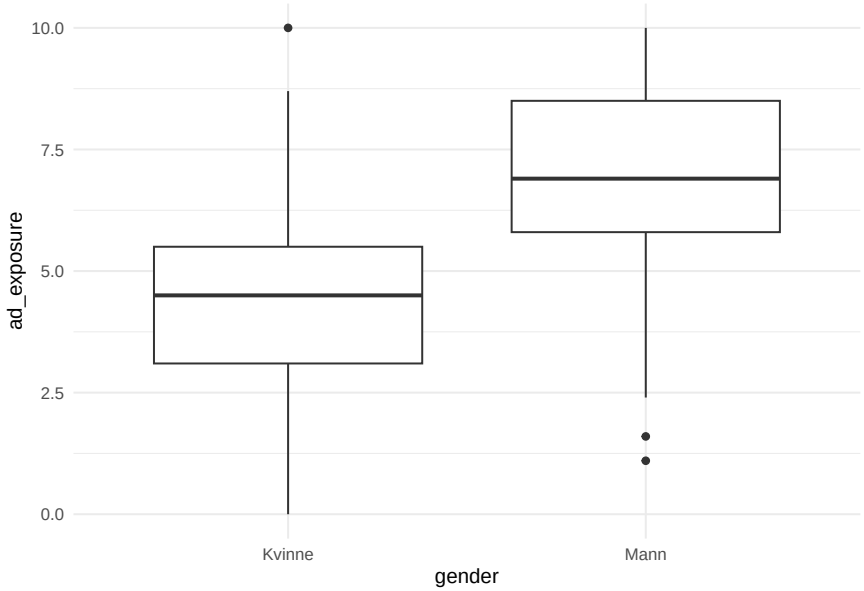
- Den viser at menn i gjennomsnitt bruker mer enn kvinner – både medianen og spesielt øvre kvartil er høyere for menn
- Spredningen er også tydelig større blant menn

Forskjellen i både nivå og variasjon indikerer at kjønn er assosiert med spending, men hva som forårsaker dette, utforskes videre i de neste oppgavene.

Til sensor: 4 Poeng for boksplott. 3 poeng for hver av de to karaktistikkene over. For den første karakteristikken godtar vi også kommentarer som svarer til at høyre hale er tyngre for menn. Dersom kandidaten bruker andre figurer/oppsummerende statistikk, men klarer å påpeke de to karaktistikkene over kan det gis opp til 7 poeng.

b)

```
ggplot(survey, aes(x = gender, y = ad_exposure)) +
  geom_boxplot() +
  theme_minimal()
```



Boksplottet viser at menn i gjennomsnitt rapporterer høyere reklameeksponering enn kvinner.

Dette gir en mulig forklaring på forskjellen i spending observert i oppgave (a): Det at menn er mer eksponert for reklame kan bidra til at menn bruker mer penger på tjenesten.

Det er også mulig at sammenhengen går motsatt vei: Menn som allerede har høy betalingsvilje eller interesse for tjenesten, kan være mer oppmerksomme på eller oppsøke reklame.

I videre analyser kan vi undersøke dette mer systematisk.

Til sensor: 3 Poeng for boksplott. 3 poeng for kommentar om forskjell og 4 poeng for en teori som ligner på en av forslagene over.

Oppgave 2

a)

Vi skal her teste nullhypotesen om lik varians i `spending` mellom menn og kvinner:

$$H_0 : \sigma^2_{\text{Mann}} = \sigma^2_{\text{Kvinne}} \quad \text{mot} \quad H_1 : \sigma^2_{\text{Mann}} \neq \sigma^2_{\text{Kvinne}}$$

der σ^2_{Mann} og σ^2_{Kvinne} er populasjonsvariansen til spending for hhv. menn og kvinner. Vi setter signfifikansnivået til $\alpha = 0.05$.

```r
spending_menn  <-
  survey %>%
  filter(gender == "Mann") %>%
  select(spending) %>%
  pull

spending_kvinner  <-
  survey %>%
  filter(gender == "Kvinne") %>%
  select(spending) %>%
  pull

var.test(spending_menn, spending_kvinner)
```

```
##
##  F test to compare two variances
##
## data:  spending_menn and spending_kvinner
## F = 1.4382, num df = 134, denom df = 164, p-value = 0.02671
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.042899 1.993855
## sample estimates:
## ratio of variances
##           1.438155
```

Siden p-verdien e lavere enn 0.05, forkaster vi nullhypotesen om lik varians. Vi konkluderer derfor med at det er en statistisk signifikant forskjell i variansen i `spending` mellom menn og kvinner.

Til sensor: 3 Poeng for riktig oppsatte hypoteser, 4 poeng for riktig utførelse av test og 3 poeng for riktig konklusjon. Merk: så lenge du forstår hva de mener skal det gis full pott. "H0: sigma_1^2 = sigma_2^2" er for eksempel ok.

---

b)

Vi skal her teste om gjennomsnittlig `spending` er forskjellig mellom menn og kvinner. Slik spørsmålet er formulert utfører vi en to-sidig test:

$$H_0 : \mu_{\text{Mann}} = \mu_{\text{Kvinne}} \quad \text{mot} \quad H_1 : \mu_{\text{Mann}} \neq \mu_{\text{Kvinne}}$$

der μ_{Mann} og μ_{Kvinne} er populasjonsgjennomsnitt til `spending` for hhv. menn og kvinner. Vi setter signifikansnivået til $\alpha = 0.05$.

```r
t.test(spending_menn, spending_kvinner, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  spending_menn and spending_kvinner
## t = 2.9871, df = 260.85, p-value = 0.003084
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  17.80980 86.70691
## sample estimates:
```

```
## mean of x mean of y
##  411.0930  358.8346
```

Vi bruker `var.equal = FALSE` fordi vi forkastet hypotesen om lik varians i oppgave 2a, men det er strengt tatt ikke nødvendig å bruke dette argumentet da dette er default verdi for argumentet `var.equal`.

Siden p-verdien er lavere enn 0.05 kan vi forkaste H_0. Det er forskjell i spending i de to gruppene.

Til sensor: 3 Poeng for riktig oppsatte hypoteser, 4 poeng for riktig utførelse av test og 3 poeng for riktig konklusjon. Merk: så lenge du forstår hva de mener skal det gis full pott. "H0: mu1 = mu2" er for eksempel ok.

---

c)

Vi skal her teste om det er en sammenheng mellom kjønn og reklamerespons:

$$H_0 : \text{Kjønn og reklamerespons er uavhengige kjennetegn} \quad \text{mot} \quad H_1 : \text{de er avhengige kjennetegn}$$

Vi setter signifikansnivået til $\alpha = 0{,}05$.

```r
# Krysstabell
krysstabell <- table(survey$gender, survey$promo_response)
krysstabell
```

```
##
##          Ja Nei
##   Kvinne 97  68
##   Mann   59  76
```

```r
# Kjikvadrattest
chisq.test(krysstabell)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  krysstabell
## X-squared = 6.1777, df = 1, p-value = 0.01294
```

Kjikvadrattesten gir en p-verdi som er lavere enn 0.05. Vi forkaster derfor nullhypotesen, og konkluderer med at det er en statistisk signifikant sammenheng mellom kjønn og reklamerespons.

Dette kan tyde på at reklamekampanjer har ulik effekt på kvinner og menn.

Til sensor: 3 Poeng for riktig oppsatte hypoteser, 4 poeng for riktig utførelse av test og 3 poeng for riktig konklusjon.

Oppgave 3

a)

```r
# Estimer enkel regresjonsmodell
modell_gender <- lm(spending ~ gender, data = survey)
summary(modell_gender)
```

```
##
## Call:
```

```
## lm(formula = spending ~ gender, data = survey)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -306.96 -113.11   -5.15  104.38  435.15
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   358.83      11.53  31.133  < 2e-16 ***
## genderMann     52.26      17.18   3.041  0.00256 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 148.1 on 298 degrees of freedom
## Multiple R-squared:  0.03011,    Adjusted R-squared:  0.02685
## F-statistic: 9.251 on 1 and 298 DF,  p-value: 0.002564
```

Modellen har to koeffisienter:

- **Intercept**: Estimatet for intercept (β_0) er 358.8. Dette representerer forventet forbruk for kvinner (referansekategorien).
- **genderMann**: Estimatet for denne koeffisienten er 52.3. Det betyr at menn er forventet å bruke 52.3 kroner mer enn kvinner. Denne forskjellen er signifikant på et 1% signfifikansnivå som er tråd med resultatet i oppgave 2b).

Til sensor: 4 Poeng for riktig modell, 3 poeng for riktig tolkning av hver av koeffisientene. Det er ingen stilpoeng for "fin" tabell. Så lenge det er riktig modell/tall/tolkning er det ok.

---

b)

```r
modell_flervariabel <- lm(spending ~ gender + income +
                            ad_exposure + loyalty_program, data = survey)
summary(modell_flervariabel)
```

```
##
## Call:
## lm(formula = spending ~ gender + income + ad_exposure + loyalty_program,
##     data = survey)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -344.28  -80.53   10.08   76.69  341.15
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     155.8929    34.7394   4.487 1.03e-05 ***
## genderMann       26.3776    17.1441   1.539    0.125
## income            1.1048     0.5437   2.032    0.043 *
## ad_exposure      15.1127     3.7143   4.069 6.07e-05 ***
## loyalty_program 165.4723    14.2110  11.644  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 120 on 295 degrees of freedom
```

```
## Multiple R-squared:  0.3696, Adjusted R-squared:  0.3611
## F-statistic: 43.25 on 4 and 295 DF,  p-value: < 2.2e-16
```

(i) Koeffisienten for `ad_exposure` er 15.1 og er signifikant forskjellig fra null.
Dette betyr at én enhet høyere reklameeksponering er assosiert med en forventet økning i `spending` på 15.1 kroner, kontrollert for kjønn, inntekt og lojalitetsprogram.

(ii) Koeffisienten for `genderMann` er betydelig mindre og er ikke lenger statistisk signfikant. Dette skyldes trolig at menn i gjennomsnitt har høyere inntekt og reklameeksponering (sistnevnte så vi i oppgave 1b), noe som forklarer deler av forskjellen i forbruk. Når vi kontrollerer for disse variablene reduseres den direkte effekten av kjønn i modellen.

Til sensor: 3 Poeng for riktig modell, 3 poeng for riktig tolkning i i) og 4 poeng for god tolkning i ii).

---

c)

```r
income_menn  <-
  survey %>%
  filter(gender == "Mann") %>%
  select(income) %>%
  pull

ad_exposure_menn  <-
  survey %>%
  filter(gender == "Mann") %>%
  select(ad_exposure) %>%
  pull

# Gjennomsnittsverdier
gj_income <- mean(income_menn)
gj_exposure <- mean(ad_exposure_menn)

# Prediksjonsdata
ny_kunde <- data.frame(
  gender = "Mann",
  income = gj_income,
  ad_exposure = gj_exposure,
  loyalty_program = 0
)

# Predikert spending
spending_uten_kampanje <- predict(modell_flervariabel, newdata = ny_kunde)
spending_uten_kampanje
```

```
##        1
## 354.7098
```

Vi predikerer forventet `spending` for en mannlig kunde med gjennomsnittlig inntekt og reklameeksponering, og som ikke er med i lojalitetsprogrammet.

Resultatet fra modellen er at forventet `spending` er 354.7 kroner.

Her kan noen kandidater ha tolket den siste setningen i oppgaveteksten som at en skal bruke gjennomsnittlige verdier blant kvinner og menn for de numeriske variablene. Vi godtar derfor også følgende prediksjon:

```r
# Gjennomsnittsverdier
gj_income <- mean(survey$income)
gj_exposure <- mean(survey$ad_exposure)

# Prediksjonsdata
ny_kunde <- data.frame(
  gender = "Mann",
  income = gj_income,
  ad_exposure = gj_exposure,
  loyalty_program = 0
)

# Predikert spending
spending_uten_kampanje <- predict(modell_flervariabel, newdata = ny_kunde)
spending_uten_kampanje
```

```
##        1
## 329.5664
```

Prediksjonen blir i så tilfelle 329.6.

Til sensor: 10 poeng for riktig input og prediksjon.

d)

Predikert spending for en kunde før kampanjen er:

$$\hat{y}_{\text{uten}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{gender} + \hat{\beta}_2 \cdot \text{income} + \hat{\beta}_3 \cdot \text{ad\_exposure} + \hat{\beta}_4 \cdot 0$$

Predikert spending for en kunde etter kampanjen er:

$$\hat{y}_{\text{kampanje}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{gender} + \hat{\beta}_2 \cdot \text{income} + \hat{\beta}_3 \cdot (\text{ad\_exposure} + 2) + \hat{\beta}_4 \cdot 1$$

Predikert økning i spending per individ på grunn av kampanjen (uavhengig av verdien til `gender`, `income` og `ad_exposure`) er derfor:

$$\Delta\hat{y} = \hat{y}_{\text{kampanje}} - \hat{y}_{\text{uten}} = 2\hat{\beta}_3 + \hat{\beta}_4$$

Så lenge vi holder `gender`, `income` og `ad_exposure` fast kan vi finne denne økningen ved å undersøke differansen mellom en prediksjon uten kampanje (f.eks den vi fant i oppgave c)) og en prediksjon med kampanje:

```r
# Ny situasjon etter kampanje
ny_kunde_kampanje <- data.frame(
  gender = "Mann",
  income = gj_income,
  ad_exposure = gj_exposure + 2,  # én ekstra enhet
  loyalty_program = 1
)

# Prediksjon etter kampanje
spending_med_kampanje <- predict(modell_flervariabel, newdata = ny_kunde_kampanje)
spending_med_kampanje
```

```
##        1
## 525.2642
```

```
# Differanse fra status quo (fra oppg. 3c)
økning <- spending_med_kampanje - spending_uten_kampanje
økning
```

```
##        1
## 195.6977
```

Alternativt kan vi regne ut differansen direkte fra koeffisientene:

```
# Hent koeffisientene
b3 <- coef(modell_flervariabel)["ad_exposure"]
b4 <- coef(modell_flervariabel)["loyalty_program"]

# Beregn total økning
økning <- 2*b3 + b4
økning
```

```
## ad_exposure
##    195.6977
```

Vi sammenligner så med kostnaden per kunde:

```
# Kostnad per kunde
kostnad <- 5 * 2 + 100

# Lønnsomt?
økning/kostnad
```

```
## ad_exposure
##     1.77907
```

Vi ser at forholdet mellom forventet økning i forbruk og kampanjekostnaden er 1.78 kr, som er lavere enn kravet på 5. **Kampanjen vurderes derfor ikke som lønnsom.**

Kommentar: Utregningen over gi gjennomsnittsøkningen i forbruk for en kunde som ikke er med i lojalitetsprogrammet. For en "gjennomsnittskunde" kunne vi i prinsippet brukt gjennomsnittsverdier for alle forklaringsvariablene, inkludert å sette `loyalty_program` lik andelen i datasettet som er med i lojalitetsprogrammet for den første prediksjonen over, og andelen pluss 1 i den andre prediksjonen. For kjønn kunne en gjort noe tilsvarende, men med en omkoding av variabelen siden denne er kategorisk. Konklusjonen vil likevel bli den samme.

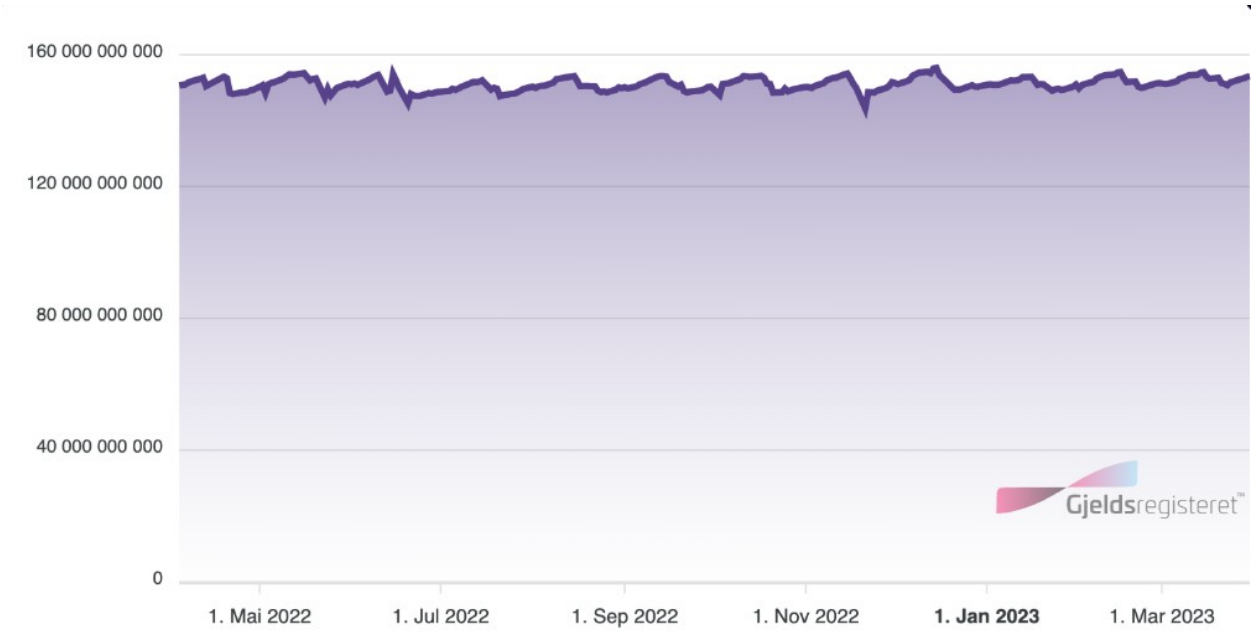
Til sensor: Man trenger ikke nødvendigvis vise fremgangsmåten med ligninger så lenge man på en god måte har forklart fremgangsmåten.

Del 2 - Regneoppgaver

Oppgave 4

Det visuelle inntrykket er at gjelden har økt kraftig, blant annet fordi y-aksen starter på et høyt nivå og dermed forsterker forskjellen. Den faktiske, relative endringen i forbruksgjeld er imidlertid beskjeden – bare noen få prosent. Vinklingen «Solid oppgang i forbruksgjelden» fremstår derfor som overdrevet når man vurderer både skalaen og utviklingen over tid.

INFO (ikke en del av løsningsforslaget): Vi kan se dette ved å gå direkte til kilden for disse dataene (gjeldsregisteret). Der finner vi en figur med en Y-akse som starter på null:



Figur 1: Brukt i E24-artikkelen "Solid oppgang i forbruksgjelden" (november 2023)

Til sensor: Det viktige poenget er forskjellen i start på Y-aksen.

Oppgave 5

a)

(i) Dersom deltakerne er helt ærlige, vil de kun rapportere at de gjettet riktig dersom det faktisk stemmer. Siden sannsynligheten for å gjette riktig i ett terningskast er $\frac{1}{6}$, forventer vi at både p_1 og p_2 er lik $\frac{1}{6}$:

$$p_1 = p_2 = \frac{1}{6}$$

(ii) Hvis deltakerne i begge grupper er like ærlige (eller like uærlige), vil begge gruppene ha samme grad av overrapportering utover $\frac{1}{6}$. Vi forventer da at andelen rapportert riktige gjetninger er lik i begge grupper:

$$p_1 = p_2$$

(iii) Hvis tillitsnudge-gruppen er mer ærlig enn kontrollgruppen, vil overrapporteringen være mindre i denne gruppen. Det betyr at andelen riktige rapporteringer vil være lavere – og nærmere $\frac{1}{6}$ – i tillitsnudge-gruppen enn i kontrollgruppen:

$$p_1 < p_2$$

b)

Basert på besvarelsen i a) skal vi teste følgende nullhypotese om lik andel:

$$H_0 : p_1 = p_2 \quad \text{mot} \quad H_1 : p_1 < p_2$$

Observerte andeler:

- $\hat{p}_1 = \frac{116}{400} = 0.29$
- $\hat{p}_2 = \frac{181}{400} = 0.4525$
- Samlet andel (pooled): $\hat{p} = \frac{297}{800} = 0.37125$

Testobservatoren for testen er gitt ved

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{400} + \frac{1}{400}\right)}} = \frac{0.29 - 0.4525}{0.034} \approx -4.78$$

Under H_0 er Z tilnærmet standardnormalfordelt og siden dette er en ensidig test med $H_1 : p_1 < p_2$ skal vi forkaste når Z blir veldig negativ. Den kritiske verdien vi må under er gitt av 0.01 percentilen i standard normalfordelingen, siden vi har et signifikansnivå på 1%. Denne finner vi i R med:

```
qnorm(0.01)
```

```
## [1] -2.326348
```

Siden $Z = -4.78$ er mindre enn -2.33, forkaster vi nullhypotesen. Det er statistisk støtte for at tillitsnudgen fører til lavere andel rapporteringer av riktige gjetninger. Dette tyder på at nudgen har gjort deltakerne mer ærlige (Se oppgave c)).

Til sensor: 4 poeng for hypotese og 6 poeng for riktig utførelse av test. Feil hypotese, men riktig gjennomføring gitt hypotese gir maksimalt 6 poeng. -3 poeng for regnefeil, dersom konklusjon gitt regnefeil er riktig. Ikke krav til kausal konklusjon her, selv om det er riktig.

---

c)

Dette er et kontrollert eksperiment der deltakerne ble tilfeldig fordelt til gruppene. Den eneste systematiske forskjellen mellom gruppene er tillitsnudgen, og denne er dermed den eneste kilden til den statistiske forskjellen vi så i rapporteringen. Gitt at eksperimentet er godt designet og randomiseringen har fungert som forutsatt, kan vi tolke den observerte forskjellen som en **kausal effekt**: Nudgen ser ut til å ha økt ærligheten i rapporteringen.

Det finnes mange situasjoner i praksis hvor en kunde må rapportere informasjon, og hvor vi er delvis avhengige av ærlig rapportering. Eksempler inkluderer reklamasjonsskjemaer, skattemeldinger, skademeldinger i forsikring og søknader om lån og kreditt. I slike tilfeller kunne man vist en melding som:

> *"Vi stoler på at kundene våre oppgir riktige opplysninger."*

før rapporteringen. Dette kan, på samme måte som i eksperimentet, bidra til mer ærlige svar.

Til sensor: 5 poeng for et noenlunde godt argument for kausalitet og 5 poeng for et eksempel.

Oppgave 6

(a)

For at en tidsserie Y_t skal være (svakt) stasjonær, må følgende tre betingelser være oppfylt:

1. $\mathbb{E}(Y_t) = \mu$, altså konstant forventning
2. $\text{Var}(Y_t) = \sigma^2$, altså konstant varians
3. $\text{Cov}(Y_t, Y_{t-k}) = \gamma_k$, altså at kovariansen mellom to tidspunkter kun avhenger av tidsforskjellen k

Basert på figuren i Vedlegg 2, ser det ut til at serien har en **stigende trend** som starter rundt år 2000. Det tyder på at forventningen μ ikke er konstant over tid, og at betingelse (1) er brutt.

Dette indikerer dermed at serien **ikke er stasjonær**.

Til sensor: Kandidaten trenger ikke å liste opp de tre punktene, så lenge vedkommende påpeker at tidsserien ikke ser ut til å ha en konstant forventning (eller gjennomsnitt), og dermed ikke er stasjonær.

(b)

Vi bruker formelen for enkel eksponentiell glatting:

$$S_t = wy_t + (1 - w)S_{t-1}$$

Vi får oppgitt at: $y_{2023} = 906$, $S_{2022} = 800$ og at $w = 0.3$.

Da er den nye glattede verdien for 2023:

$$S_{2023} = 0.3 \cdot 906 + 0.7 \cdot 800 = 271.8 + 560 = 831.8$$

Prediksjonen for 2024 med enkel eksponentiell glatting er:

$$\hat{y}_{2024} = S_{2023} = 831.8$$

Til sensor: 6 poeng for glatting og 4 for prediksjon. Regnefeil -3 poeng, gitt riktig formel.

(c)

Siden prediksjonen $\hat{Y}_{t+1} = S_t$ er et vektet gjennomsnitt av nåværende (Y_t) og tidligere (lavere) verdier av tidsrekken, vil den typisk være mindre enn den faktiske verdien Y_{t+1} når tidsrekken har en stigende trend. Modellen underpredikerer dermed systematisk – noe som også kommer tydelig frem i figuren i Vedlegg 2.

Til sensor: Et godt argument for underpredikering kreves for full pott. Kandidater som kun viser til figuren i Vedlegg 2, uten ytterligere forklaring, bør få maksimalt 6 poeng.

Oppgave 7

a)

Vi skriver for enkelhets skyld

$$p = P(\texttt{mistenkelig} \mid x_1 = \texttt{beløp},\ x_2 = \texttt{endringer},\ x_3 = \texttt{bilagsnummer},\ x_4 = \texttt{vedlegg},\ x_5 = \texttt{intern},\ x_6 = \texttt{refusjon})$$

Den estimerte modellen er da:

$$p = \frac{\exp(z)}{1+\exp(z)}, \quad z = -0.112 + 0.025x_1 + 0.595x_2 + 0.00083x_3 - 0.333x_4 + 0.870x_5 + 0.051x_6$$

Her er `intern` og `refusjon` indikatorvariabler for kategorien `bilagstype`, mens `faktura` er referansekategorien (baseline). Det betyr at bilag som er fakturaer ikke får noe tillegg i z utover konstantleddet.

Effekten av `vedlegg` er statistisk signifikant forskjellig fra null på 1% signifikansnivå ($p \approx 0.0017$). Ett ekstra `vedlegg` er assosiert med at oddsen for at et bilag er mistenkelig avtar med en multiplikativ faktor på $\exp(-0.333) = 0.7167$, det vil si en reduksjon på 28.3%. For denne tolkningen forutsetter vi at alle andre forklaringsvariabler holdes konstante.

Til sensor: 5 poeng for modell og 5 poeng for tolkning.

(b)

Vi skal beregne sannsynligheten for at et bilag med følgende verdier er mistenkelig:

- `beløp` = 30
- `endringer` = 2
- `bilagsnummer` = 800
- `vedlegg` = 1
- `bilagstype` = `faktura` (referansekategori)

Siden `faktura` er referansekategorien, settes indikatorvariablene `intern` og `refusjon` til null: $x_5 = 0$, $x_6 = 0$.

Vi setter inn verdiene i modellen:

$$z = -0.112 + 0.025 \cdot 30 + 0.595 \cdot 2 + 0.00083 \cdot 800 - 0.333 \cdot 1 + 0.870 \cdot 0 + 0.051 \cdot 0 = 2.159$$

Sannsynligheten er:

$$\hat{p} = \frac{\exp(2.159)}{1+\exp(2.159)} \approx \frac{8.66}{1+8.66} \approx 0.896$$

Siden $\hat{p} = 0.896 > 0.5$, klassifiseres bilaget som **mistenkelig**.

Til sensor: 7 poeng for sannsynlighet og 3 poeng for klassifisering. -3 for regnefeil. Rett klassifisering gitt regnefeil/misforståelse skal gis 3 poeng.

c)

For å klassifisere bilaget ved hjelp av en KNN-modell med $K = 5$, ser vi på de 5 bilagene i tabellen som ligger nærmest det aktuelle bilaget i oppgave b). Ut fra avtanden ser vi at dette er bilagene i de 5 øverste radene.

De fem nærmeste naboene har følgende verdier for `mistenkelig`:

- 1
- 0
- 1
- 1
- 0

Blant disse er det 3 av 5 som har `mistenkelig = 1`, og 2 som har `mistenkelig = 0`.

Flertallet er altså `mistenkelig = 1`.
Bilaget klassifiseres derfor som mistenkelig av KNN-modellen med $K = 5$.

Når det gjelder hvilken modell revisjonsteamet bør bruke for å forstå hva som kjennetegner juks, **vil logistisk regresjon være det beste valget.**

Logistisk regresjon gir:

- Tolkbare koeffisienter som viser hvordan hver forklaringsvariabel påvirker sannsynligheten for at et bilag er mistenkelig
- Mulighet til å kvantifisere effekter (for eksempel: ett ekstra vedlegg reduserer odds for mistanke med 28%)

KNN er en ikke-parametrisk metode som kan gi gode prediksjoner, men den er **vanskelig å tolke** uten å benytte seg av avansert metodikk.

Så gitt at den logistiske-regresjonsmodellen gir en god representasjon av dataene, vil anbefalingen være å bruke denne modellen til å forstå hva som kjennetegner juks.

Til sensor: 5 poeng for riktig klassifisering og 5 poeng for god argumentasjon for logistisk regresjon.