

# FORBEREDELSE TIL GRUPPEEKSAMEN I MET4 HØST 22

---

I dette dokumentet vil dere finne en innledning til problemstillingen som vi skal se på ved årets gruppeeksamen i MET4. Som vedlegg til dette dokumentet vil dere finne litt lesestoff som danner grunnlaget for problemstillingen, samt det eksakte datasettet som også blir utlevert med selve eksamensoppgavene 14. november 2022. Vi går også gjennom noen R-kommandoer som kan være nyttige for å jobbe med akkurat dette datasettet, samt en generell tilbakemelding på gruppeeksamen som ble gitt til hele kullet for et år siden.

---

## Innledning

Matsvinn utgjør en stor utfordring for målet om en mer bærekraftig verden. En studie fra 2011<sup>1</sup> anslår at årlig globalt matsvinn tilsvarer 30 % av den totale matproduksjonen. Ikke bare er dette dårlig utnyttelse av naturressurser, men utgjør også et globalt økonomisk tap på ca. 10 billioner kroner.

Det er et økende internasjonalt fokus på overvåking og reduksjon av matsvinn i verden. I 2016 sluttet alle FNs medlemsland seg opp om en rekke mål for bærekraftig utvikling<sup>2</sup>, blant annet at mengden matsvinn per innbygger skal halveres innen 2030. En stor utfordring med å følge opp et slikt mål er at vi på grunn av manglende data ikke har pålitelige metoder for å overvåke globalt matsvinn. En mulig tilnærming til dette problemet er å utnytte at matsvinn *kan* samvarierte med andre typer variabler som lettere lar seg måle. Eksempler på slike forklaringsvariabler er nedbør, temperatur, og utgifter knyttet til produksjon og oppbevaring. For eksempel har norske bønder nå store utgifter knyttet lagring av grønnsaker på kjøll på grunn av høye strømpriser. Tanken er at en statistisk modell som beskriver denne samvariasjonen kan brukes til å predikere matsvinn for steder eller perioder hvor man har data for forklaringsvariablene, men ikke selve matsvinnet.

Vi kan lese om en slik fremgangsmåte i en nylig utgitt artikkel av Mingione, Fabi og Lasino (2021) med tittel *Measuring and Modeling Food losses*.<sup>3</sup> Modellen som beskrives i artikkelen er svært sofistikert og langt utenfor pensum i MET4, men datasettet som brukes danner grunnlaget for denne hjemmeeksamen. Vi kan lese om problemet generelt og bli vist videre til aktuelle referanser i kapittel 1 og 2 i denne artikkelen. Artikkelen er lagt ved disse eksamensoppgavene sammen med en versjon av datasettet som ble brukt. Etter å ha satt working directory til mappen hvor datafilen ligger kan dere laste inn datasettet ved å kjøre

```
load("met4_h22.Rdata")
```

Hovedvariabelen i datasettet heter `loss` og er definert som andel kornavling tapt. Vi har en observasjon av andel tapt avling (`loss`) per land, per år, per korntype. Denne variabelen er målt for 15 korntyper i 68 land i perioden 1991 - 2014 (men vi har ikke observasjoner for *alle* kombinasjoner av korntype, land og år). Datasettet inneholder også en rekke andre variabler som ulike kilder hevder har en påvirkning på tap av korn. Se Tabell 1 for beskrivelse av de enkelte variablene i datasettet.

---

<sup>1</sup>Gustavsson et. al 2011, Global food losses and food waste. Technical report, <https://www.fao.org/3/mb060e/mb060e00.htm>

<sup>2</sup>Transforming our world: 2030 Agenda for Sustainable Development (2015), <https://sdgs.un.org/2030agenda>

<sup>3</sup>Mingione, Fabi og Lasino: *Measuring and Modeling Food Losses* Journal of Official Statistics (2021). Artikkelen er lagt ut på Canvas.

Table 1: Variabelbeskrivelser. For variabler som er 'standardisert' er det trukket fra gjennomsnittet og delt på standardavviket til variabelen. Det betyr at enheten på disse variablene er i standardavvik. Videre svarer verdien 0 til den gjennomsnittlige verdien på originalskalaen, slik at negative verdier kan tolkes som verdier mindre enn gjennomsnittet, mens positive verdier kan tolkes som verdier større enn gjennomsnittet.

Variabel	Beskrivelse
country	Landnavn. Totalt 68 land.
year	Årstall for observasjon. Observasjoner er gjort årlig i perioden 1991 - 2014.
crop	Kornstype. Totalt 15 typer.
loss	Andel av kornavling tapt i landet det aktuelle året og for den aktuelle korntypen.
temperature	Gjennomsnittlig temperatur i landet det aktuelle året (standardisert).
rain	Gjennomsnittlig nedbør i landet det aktuelle året (standardisert).
biofuel	Gjennomsnittlig pris på biobrensel i verden det aktuelle året (standardisert).
gas	Gjennomsnittlig pris på naturgass i verden det aktuelle året (standardisert).
GDPcontribution	Landbrukets bidrag til landets bruttonasjonalprodukt det aktuelle året (standardisert).
coal	Gjennomsnittlig pris på kull i verden det aktuelle året (standardisert).
lpi	Verdensbankens 'logistic performance index'. Se <a href="https://lpi.worldbank.org/">https://lpi.worldbank.org/</a> (standardisert).
input	Måler kostnader på forbruksvarer som er knyttet til landbruk i landet det aktuelle året (standardisert).
investment	Måler investeringer gjort i landbruk i landet det aktuelle året. (standardisert).
energy	Måler energipriser i landet det aktuelle året (standardisert).

## Noen nyttige R-kommandoer

Datasettet er et paneldatasett med en rad for hver kombinasjon av år og land og type avling. Denne strukturen gjør at vi ganske sikkert får nytte av `filter()`-funksjonen som vi så på i R-introduksjonen vår. Vi får ut alle observasjonene for Norge, for eksempel, ved å filtrere på `country`-variabelen:

```
wastedata %>% filter(country == "norway")

## # A tibble: 84 x 14
##   country crop    loss year temperature    rain biofuel    gas GDPcontribution
##   <chr>   <chr>   <dbl> <dbl>      <dbl>   <dbl>   <dbl>   <dbl>      <dbl>
## 1 norway barley 0.0453 1991      1.88    1.03   -0.284 -0.202      0.262
## 2 norway barley 0.041  1992      1.66    1.03   -0.284 -0.202      0.209
## 3 norway barley 0.0476 1993     -0.560 -0.746  -0.247 -0.202     -0.515
## 4 norway barley 0.0477 1994     -0.578 -0.744  -0.247 -0.202     -0.515
## 5 norway barley 0.0457 1995     -0.487 -0.805  -0.200 -0.202     -0.542
## 6 norway barley 0.044  1996     -0.388 -0.618  -0.209 -0.202     -0.525
## 7 norway barley 0.0452 1997     -0.506 -0.806  -0.237 -0.202     -0.525
## 8 norway barley 0.0404 1998     -0.573 -0.835  -0.213 -0.202     -0.525
## 9 norway barley 0.0481 1999     -0.410 -0.937  -0.169 -0.202     -0.540
## 10 norway barley 0.0436 2000     -0.268 -0.657  -0.182 -0.202     -0.540
## # ... with 74 more rows, and 5 more variables: coal <dbl>, lpi <dbl>,
## #   input <dbl>, energy <dbl>, investment <dbl>
```

Vi kan legge til flere filtreringskriterier i `filter()`-funksjonen, slik at vi for eksempel kan få ut alle observasjonene for *hvete* i Norge:

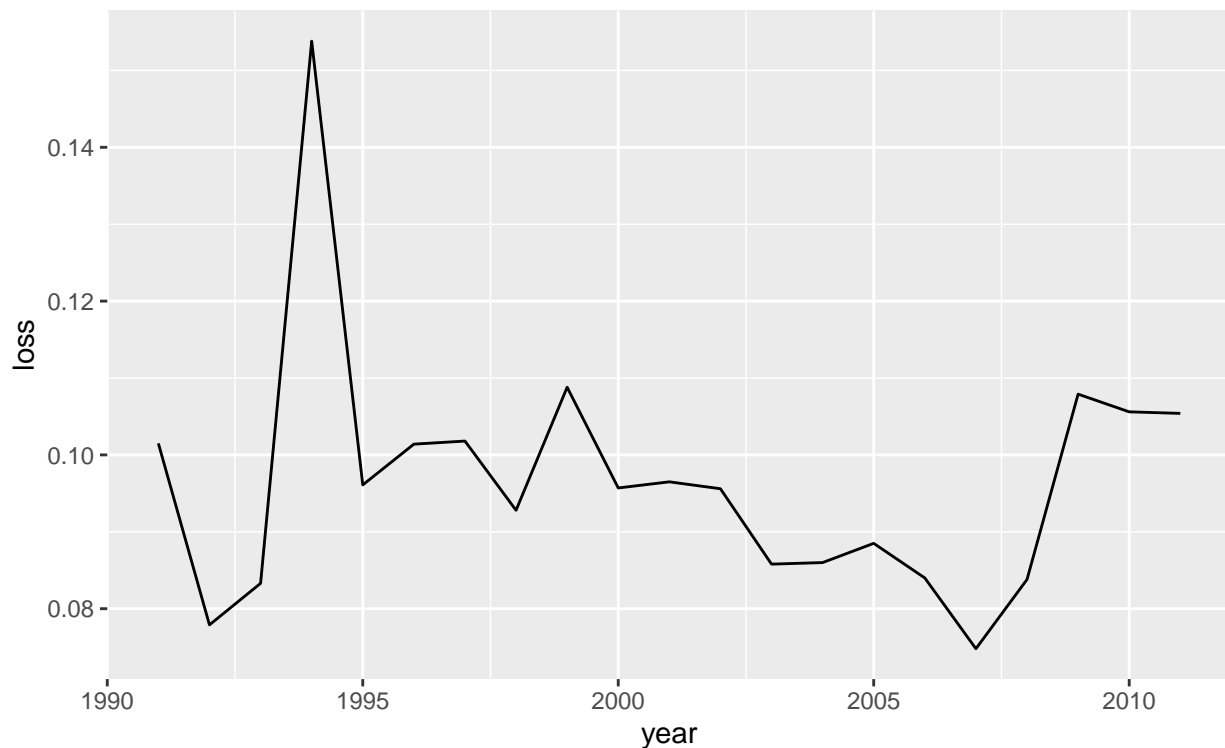
```
wastedata %>% filter(country == "norway", crop == "wheat")
```

```
## # A tibble: 21 x 14
```

```
##   country crop   loss year temperature   rain biofuel   gas GDPcontribution
##   <chr>   <chr> <dbl> <dbl>      <dbl>   <dbl>   <dbl>   <dbl>      <dbl>
## 1 norway  wheat 0.102 1991      1.88    1.03   -0.284 -0.202      0.262
## 2 norway  wheat 0.0779 1992      1.66    1.03   -0.284 -0.202      0.209
## 3 norway  wheat 0.0833 1993     -0.560 -0.746  -0.247 -0.202     -0.515
## 4 norway  wheat 0.154 1994     -0.578 -0.744  -0.247 -0.202     -0.515
## 5 norway  wheat 0.0961 1995     -0.487 -0.805  -0.200 -0.202     -0.542
## 6 norway  wheat 0.101 1996     -0.388 -0.618  -0.209 -0.202     -0.525
## 7 norway  wheat 0.102 1997     -0.506 -0.806  -0.237 -0.202     -0.525
## 8 norway  wheat 0.0928 1998     -0.573 -0.835  -0.213 -0.202     -0.525
## 9 norway  wheat 0.109 1999     -0.410 -0.937  -0.169 -0.202     -0.540
## 10 norway wheat 0.0957 2000     -0.268 -0.657  -0.182 -0.202     -0.540
## # ... with 11 more rows, and 5 more variables: coal <dbl>, lpi <dbl>,
## #   input <dbl>, energy <dbl>, investment <dbl>
```

Dette kan vi sende rett videre til `ggplot` ved hjelp av pipe-operatoren dersom vi ønsker å se utviklingen i svinn for hvete i Norge:

```
wastedata %>%
  filter(country == "norway", crop == "wheat") %>%
  ggplot(aes(x = year, y = loss)) +
  geom_line()
```



Når vi jobber med dette datasettet så kan vi være interessert i å regne ut gjennomsnitt av en variabel innad i grupper. For eksempel vil kanskje ha gjennomsnittlig matsvinn for hvert land i et gitt år. Da kan vi bruke to svært nyttige funksjoner i `dplyr`-pakken; `group_by()` og `summarise()`. Den første av disse funksjonene lager en gruppering i datasettet som alle påfølgende funksjoner i pipe-sekvensen vil respektere. `summarise()`-funksjonen er designet for å lage deskriptiv statistikk for et datasett. La oss demonstrere den raskt først. La oss si at vi ønsker å regne ut gjennomsnitt, standardavvik, min, og max for variabelen `loss`. Vi kan selvsagt kjøre `mean(wastedata$loss)`, `sd(wastedata$loss)` etc. hver for seg, men vi kan også få ut

en fin tabell ved hjelp av `summarise()` slik som her:

```
wastedata %>%
  summarise(Gjennomsnitt = mean(loss),
            Standardavvik = sd(loss),
            Minimum = min(loss),
            Maksimum = max(loss))

## # A tibble: 1 x 4
##   Gjennomsnitt Standardavvik   Minimum Maksimum
##   <dbl>         <dbl>       <dbl>    <dbl>
## 1     0.0567     0.0852 0.00000142    0.986
```

Men la oss gjøre det litt mer interessant. Vi ønsker å ta for oss kun observasjoner gjort i år 2000, og så ønsker vi *en linje med deskriptiv statistikk for matsvinn for hver land i verden*. Dette kan vi nok klare manuelt ved å ta tiden til hjelp og lage en tabell som over for hvert land i datasettet, men det er mye enklere hvis vi bare bruker `group_by()` til å fortelle at vi ønsker deskriptiv statistikk splittet opp på land:

```
wastedata %>%
  filter(year == 2000) %>%
  group_by(country) %>%
  summarise(Gjennomsnitt = mean(loss),
            Standardavvik = sd(loss),
            Minimum = min(loss),
            Maksimum = max(loss))

## # A tibble: 62 x 5
##   country      Gjennomsnitt Standardavvik Minimum Maksimum
##   <chr>         <dbl>         <dbl>    <dbl>    <dbl>
## 1 afghanistan    0.100             NA     0.100    0.100
## 2 albania        0.147             NA     0.147    0.147
## 3 argentina      0.0356          0.0238    0.0199    0.0776
## 4 armenia        0.254             NA     0.254    0.254
## 5 austria        0.0361          0.0156    0.0136    0.0569
## 6 azerbaijan     0.0384          0.0550     0.01    0.121
## 7 belgium        0.0149          0.00915   0.0101    0.0312
## 8 canada         0.00794          0.0117    0.0003    0.028
## 9 cyprus          0.017            0.0184    0.004     0.03
## 10 czechia       0.0061            0.00137   0.0042    0.0074
## # ... with 52 more rows
```

Den første funksjonen i sekvensen over er grei, den bare filtrerer ut alle observasjoner som ikke er gjort i år 2000. Den neste funksjonen *grupperer* observasjonene etter verdien av variabelen `country`. Denne funksjonen forandrer ikke på selve innholdet i datasettet, den bare legger til *metainformasjon* om at alle funksjoner som kommer etterpå skal respektere gruppetilhørigheten som den definerer. Med andre ord: alt som skjer i pipe-sekvensen etter `group_by(country)` skal skje gruppevis innen hvert land. Deretter bare slenger vi på hva vi ønsker å gjøre, i dette tilfellet lage et sammendrag av `loss`-variabelen som over, men nå med en linje per `country`.

Vi ser at vi for enkelte land mangler standardavviket. Det skyldes at vi bare har observert svinnet av en enkelt avling det året, og vi kan ikke ta standardavviket av et enkelt tall. Vi kan faktisk legge til en kolonne til med antall observasjoner innad i hver gruppe ved å bruke funksjonen `n()` i `summarise()`:

```
wastedata %>%
  filter(year == 2000) %>%
  group_by(country) %>%
  summarise(Gjennomsnitt = mean(loss),
```

```

Standardavvik = sd(loss),
Minimum = min(loss),
Maksimum = max(loss),
Antall = n())

```

```

## # A tibble: 62 x 6
##   country      Gjennomsnitt Standardavvik Minimum Maksimum Antall
##   <chr>          <dbl>          <dbl>    <dbl>    <dbl> <int>
## 1 afghanistan    0.100            NA      0.100    0.100     1
## 2 albania        0.147            NA      0.147    0.147     1
## 3 argentina      0.0356          0.0238    0.0199    0.0776     5
## 4 armenia        0.254            NA      0.254    0.254     1
## 5 austria        0.0361          0.0156    0.0136    0.0569     6
## 6 azerbaijan     0.0384          0.0550    0.01      0.121     4
## 7 belgium        0.0149          0.00915   0.0101    0.0312     5
## 8 canada         0.00794         0.0117    0.0003    0.028     5
## 9 cyprus          0.017           0.0184    0.004     0.03      2
## 10 czechia       0.0061          0.00137   0.0042    0.0074     4
## # ... with 52 more rows

```

Vi kan gruppere på flere variabler også. Vi kan for eksempel se for oss et plott av *gjennomsnittlig svinn per år per avling*. Her har vi to grupperingsvariabler (year og crop), men det er ikke noe problem:

```

wastedata %>%
  group_by(year, crop) %>%
  summarise(gjennomsnittlig_svinn = mean(loss))

```

```

## # A tibble: 331 x 3
## # Groups:   year [24]
##   year crop      gjennomsnittlig_svinn
##   <dbl> <chr>          <dbl>
## 1 1991 barley      0.0514
## 2 1991 buckwheat  0.0485
## 3 1991 green corn (maize) 0.222
## 4 1991 maize (corn) 0.0588
## 5 1991 millet    0.110
## 6 1991 mixed grain 0.120
## 7 1991 oats      0.0369
## 8 1991 quinoa    0.164
## 9 1991 rice      0.0383
## 10 1991 rye      0.0519
## # ... with 321 more rows

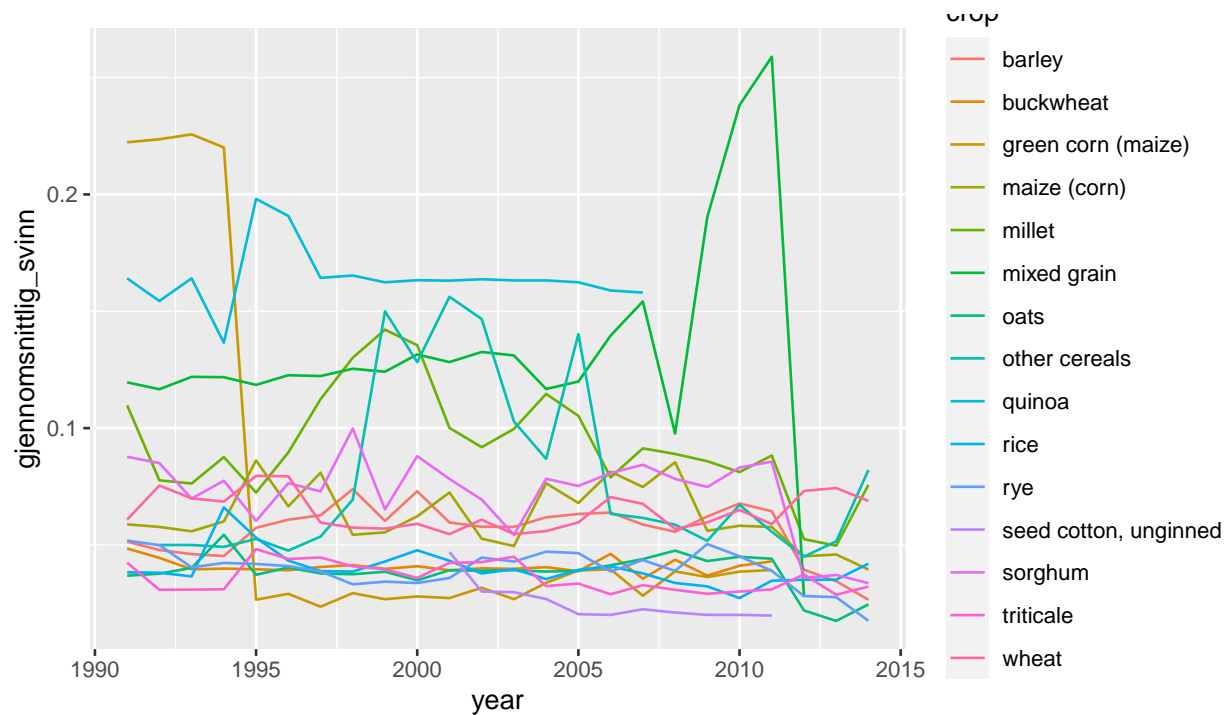
```

Da får vi en lang data frame med et gjennomsnitt per år, per avling. Denne kan sendes direkte videre til ggplot, der vi kan bruke for eksempel colour til å skille mellom de ulike typene avling:

```

wastedata %>%
  group_by(year, crop) %>%
  summarise(gjennomsnittlig_svinn = mean(loss)) %>%
  ggplot(aes(x = year, y = gjennomsnittlig_svinn, colour = crop)) +
  geom_line()

```



Hvis vi ønsker å gjøre klassiske hypotesetester på dette datasettet så kan det være nyttig å huske kommandoen `pull` for å gjøre om en data frame med en kolonne til en vektor. Svinnet til alle land i år 2000, for eksempel, kan hentes ut som følger:

```
wastedata %>%
  filter(year == 2000) %>%
  select(loss) %>%
  pull
```

```
## [1] 1.0010e-01 1.4660e-01 3.0500e-02 2.6800e-02 7.7600e-02 1.9900e-02
## [7] 2.3100e-02 2.5360e-01 3.6300e-02 3.8600e-02 4.7500e-02 2.3900e-02
## [13] 5.6900e-02 1.3600e-02 1.0000e-02 1.2600e-02 1.2090e-01 1.0300e-02
## [19] 1.2000e-02 1.0100e-02 1.0400e-02 3.1200e-02 1.0700e-02 4.0000e-04
## [25] 2.8000e-02 2.9000e-03 8.1000e-03 3.0000e-04 4.0000e-03 3.0000e-02
## [31] 6.1000e-03 7.4000e-03 6.7000e-03 4.2000e-03 3.0200e-02 3.0000e-02
## [37] 2.6700e-02 2.9800e-02 5.6200e-02 7.2400e-02 1.9000e-02 5.4800e-02
## [43] 6.7100e-02 1.1070e-01 1.4700e-02 1.6600e-02 7.8000e-03 1.3000e-02
## [49] 4.4000e-03 8.6000e-03 8.9000e-03 5.7400e-02 3.8030e-01 2.4700e-02
## [55] 2.9800e-02 2.4200e-02 2.6000e-02 2.0000e-02 2.4100e-02 2.8500e-02
## [61] 6.6300e-02 1.9600e-02 5.2500e-02 8.4000e-03 4.3000e-03 1.6150e-01
## [67] 3.3800e-02 1.5000e-02 5.6800e-02 5.4000e-02 4.4300e-02 3.9500e-02
## [73] 5.5600e-02 4.3480e-01 2.5400e-02 7.1700e-02 8.0000e-04 1.0000e-03
## [79] 2.2000e-02 7.4000e-03 6.0660e-01 1.5600e-02 2.3830e-01 2.0000e-03
## [85] 2.3300e-02 2.3800e-02 1.6300e-02 5.5800e-02 6.0900e-02 2.0930e-01
## [91] 9.4200e-02 4.8150e-01 6.6000e-02 1.0060e-01 9.9000e-02 1.0000e-01
## [97] 2.4300e-02 5.8200e-02 3.4500e-02 2.7800e-02 2.0100e-02 2.3200e-02
## [103] 4.2400e-02 2.0390e-01 2.6900e-02 4.3600e-02 2.5200e-02 6.2000e-02
## [109] 9.5700e-02 4.0900e-02 3.2300e-02 2.4120e-01 3.5200e-02 1.6330e-01
## [115] 1.7990e-01 1.0000e-02 1.4900e-02 6.6100e-02 4.0900e-02 6.0600e-02
## [121] 7.8500e-02 3.9500e-02 3.8900e-02 5.1500e-02 1.7000e-02 2.1000e-03
## [127] 2.1500e-02 5.6000e-03 1.5700e-02 1.1300e-02 2.1900e-02 1.5000e-02
## [133] 1.2820e-01 1.3700e-02 7.9000e-03 1.6800e-02 9.0700e-02 1.1350e-01
```

```
## [139] 4.0000e-02 1.5600e-02 2.3900e-02 6.4100e-02 4.2000e-03 4.1000e-03
## [145] 2.3000e-03 3.2000e-03 1.4880e-01 3.0200e-02 6.0000e-02 3.1000e-02
## [151] 6.4200e-02 2.7000e-02 2.9100e-02 6.2000e-03 7.1000e-03 3.4200e-02
## [157] 3.1200e-02 3.0900e-02 4.8300e-02 6.1100e-02 2.4080e-01 1.5000e-01
## [163] 4.0000e-02 1.9890e-01 1.4970e-01 4.0000e-02 1.5310e-01 9.0500e-02
## [169] 3.0000e-02 6.5000e-02 3.4000e-02 2.2300e-02 3.1000e-02 6.8000e-03
## [175] 4.7000e-03 1.9500e-02 1.0700e-02 1.1640e-01 1.4480e-01 1.6660e-01
## [181] 1.0000e-02 1.8523e-06
```

Til slutt kan vi nevne følgende distinksjon som kan være viktig når vi skal estimere regresjonsmodeller for dette datasettet. Legg merke til at tidsvariabelen `year` er *numerisk*. Hvis vi for eksempel ønsker å forklare variasjon i svinn ved hjelp av tid på følgende måte:

$$Y = \beta_0 + \beta_1 t + \beta_2 X + \epsilon,$$

der  $X$  er en eller annen forklaringsvariabel og altså der  $\beta_1$  representerer forventet økning i svinn for hvert år som går, så kan vi bare bruke `year` som forklaringsvariabel i regresjonen på vanlig måte ( $X$  er bare et eksempel, vi har ingen variabler med det navnet i datasettet vårt):

```
reg1 <- lm(loss ~ year + X, data = wastedata)
```

Dersom vi derimot ikke ønsker å anta at utviklingen i tid er lineær, kan vi heller estimere følgende modell:

$$Y = \phi_t + \beta_2 X + \epsilon,$$

der vi lar hvert år ha sin egen faste effekt  $\phi_t$  (med andre ord, vi antar ikke at utviklingen over tid er like stor hvert år), så må vi enten bruke syntaksen i `plm`-pakker for estimering av faste effekter, eller så kan vi bare legge til variabelen `year` på samme måte som over; men da må vi huske å gi beskjed om at dette ikke lenger skal behandles som en numerisk variabel men som en kategorisk variabel der hvert år skal ha sin egen dummyvariabel. Det kan vi enkelt gjøre slik:

```
reg2 <- lm(loss ~ as.factor(year) + X, data = wastedata)
```

Da blir derimot utskriften fra `summary()` veldig rotete siden den skriver ut estimert effekt  $\phi_t$  for alle år i datasettet. Det kan være en idé å presentere resultatet på en annen måte, siden det gjerne bare er  $\beta_2$  vi strengt tatt er interessert i.

## Retteskjema

Vi gjør oppmerksom på at vi kommer til å bruke retteskjemaet som er lagt ut i fillageret på Canvas vil bli brukt til å bedømme besvarelsene.

## Generell tilbakemelding etter Gruppeeksamen V21

Vi gav en generell tilbakemelding til kullet som tok gruppeeksamen i vårsemesteret 2021. Det kan være elementer der som er til hjelp også for dere inn mot eksamen. Tilbakemeldingen er gitt under:

---

Hei alle. Da er sensuren for den gruppebaserte hjemmeeksamen falt, og karakterene er publisert. Nivået ligger omtrent der det pleier med et snitt litt under B, som på alle måter må sies å være et godt resultat! En annen ting som er akkurat som det pleier er at ikke alle er fornøyd med karakteren sin. Vi har mottatt ganske mange forespørsler om begrunnelse, og de skal vi svare på i tur og orden, men jeg tenkte her å gi en mer generell tilbakemelding som alle kan dra nytte av i videre studier.

### Forståelse av den spesifikke situasjonen:

Noe av det vi er mest fornøyd med i årets gruppeeksamen er at dere i stor grad klarte å vise forståelse for den konkrete situasjonen. Begrepet “customer analytics” er svært viktig i store deler av det private næringslivet, og det er på mange måter et åpent spørsmål hvordan vi kan dra nytte av kundedata for å forstå hvorfor kunder sier opp avtalene sine, og forhindre at det skjer. Dette handler om mer enn å bare sette opp en logistisk regresjon og se hvor godt den predikerer, det handler også om å forstå at det finnes viktige og ikke-trivielle avveininger av tilfredshet og kostnader knyttet til en kampanje rettet mot risikokunder. Videre fikk vi kanskje erfare at statistikken i dette tilfellet ikke gav oss en vidundermedisin for prediksjon, og at det ikke er lett å slå den mest naive løsningen: å bare la det være og ikke ringe til noen. Mange grupper viste god forståelse for dette gjennom nøktern analyse av de statistiske resultatene, og har blitt belønnet for det gjennom karakteren. Andre klarte i mindre grad å vise selvstendige vurderinger, og virket i større grad å gå på autopilot. Det kan være med å forklare en karakter som kanskje ikke stod til forventningene.

### Kildekritikk:

Svært mange grupper presenterte følgende formulering av den logistiske regresjonsmodellen (eller en variant):  $\log(y_i) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$ . Dette er jo *galt*, og vi stusset lenge på hvor dette egentlig kom fra. Vi saumfarte videoer og forelesningsnotater i jakten etter synderen uten å finne en forklaring på at så mange grupper gjorde den identiske feilen. Besynderlig. Til slutt fant vi ut av det: eksempelbesvarelsen som fikk A har gjort akkurat det samme (rett før tabell 3). Ved nærmere ettersyn så viser det seg at flere grupper har lagt seg meget tett opp mot den besvarelsen, noe som kanskje er fristende gitt at eksamensoppgavene hadde store fellestrekk. Dette fremstår som svært lite sjarmerende for sensor, spesielt når man regelrett kopierer en feil. Hvis du følger deg truffet her, så vil jeg anbefale å ta med seg følgende læringspunkter i videre studier:

- Vi kommer ikke til å kjøre plagiatsaker på dette. Denne spesifikke feilen har heller ikke blitt brukt som eneste begrunnelse for å vippe noen ned fra en karakter til en annen. På samme måte som før er det totalinntrykket som til syvende og sist bestemmer karakteren.
- Det er ikke en unnskyldning at dette stod i en A-besvarelse; A betyr ikke perfekt, og i dette tilfellet kan det hende at feilen enten gikk under radaren til sensoren, eller så ble det satt på kontoen for “blingser” som vi selvsagt tillater i ellers sterke besvarelser.
- Vær kritisk til det du leser! Dobbelsjekk! Det er forskjell på kilder! Det finnes feil over alt! Dersom du skriver noe *galt* er det du som får svi for det, i alle fall dersom du ikke oppgir kilde.
- Lær deg bedre sitatskikk. En ting er å bli “inspirert”, men i min bunke var det flere besvarelser som absolutt burde referert til eksempelbesvarelsen, og som kanskje hadde måttet betale en mye høyere pris i et masterkurs for mangel på referanse. Ta dette som et vennlig råd på veien videre.

### Deskriptiv statistikk:

Den deskriptive statistikken var under pari i år. Det var mange grupper som ikke egentlig forholdt seg til resten av oppgaven, som handlet om klassifisering. Da forventer vi for eksempel tydelige oversikter om hvordan variablene fordeler seg i de to klassene. Det var også muligheter til å finne noen rariteter i datasettet (i.e. negative priser) som burde bli tatt hånd om eller i det minste kommentert. En stargazer-tabell over alle variablene og et histogram eller to over noen av variablene fremstår som helt isolert fra resten av oppgaven hvis ikke er noen videre diskusjon som er relevant for det som kommer etterpå. Da har man misset poenget med deskriptiv statistikk, og må naturlig nok betale for det.

### "Vi hadde alt rett, men fikk bare B (eller C)":

Dere er nå i alle fall over halvveis i bacheloren, og noen er allerede ferdige og skal videre på master. Enten vi liker det eller ikke, så må vi innse at det er på høy tid å oppdatere hvilke kriterier som gjelder for å hevde seg i toppen av karakterhierarkiet. MET4 er i grunn et godt bilde på en viktig overgang: vi jobber i større grad med *virkelige* problemer. Desverre har virkelige problemer en lei tendens til å være mye *vanskeligere*, og kanskje ikke på den måten at vi alltid trenger kompliserte modeller og lange formler. Nei, det er bare slik at den virkelige verden er mye mer nyansert enn fasiten bakerst i læreboken. Det finnes mange måter å angripe et problem på, men kanskje er det ingen måte som løser problemet fullstendig og for evig og alltid. Det er alltid noen nyanser som vi ikke klarer å ta hensyn

til, eller noen avveininger som gjør at vi kan komme et stykke på vei langs en dimensjon, men kanskje på bekostning av at et annet problem blir verre å hankses med. Med andre ord: det handler ikke bare om å regne riktig (selv om det er viktig, for all del), men om å jobbe seg frem til et svar som er så godt som mulig, som hjelper oss best mulig, og der vi samtidig har kontroll på hva vi ikke har klart å løse og hvilke konsekvenser det har.

Løsningsforslaget er ikke fullstendig. Det mangler fullstendig prosaen, argumentasjon for og mot, tolkning og diskusjon som trengs for å nå helt opp. Det inneholder noen kodelinjer som kan brukes som grunnlag for diskusjonen.

Universitets- og Høgskolerådet har generelle beskrivelser av karakterer for BØA-studiet (klikk for link) som kan være nyttige å kikke på. Blant annet ser vi der at man skal vise stor grad av selvstendighet for å få en A, og for B skal selvstendigheten og vurderingsevnen være meget god. De som lykkes godt i denne eksamensformen klarer nettopp dette, og det er klart at listen for de høyeste karakterene ligger høyere enn at man ikke har gjort direkte feil et sted. Dette er helt i tråd med Bloom's Taxonomy (klik for link) som viser at "remember" og "apply" er lavere læringsnivå enn for eksempel "analyze" og "evaluate". Jeg tror at ganske mange med stor fordel bør recalibrere hva som trengs for karakteruttelling i de videre studiene.

### **Noen leverer på et helt vilt høyt nivå**

Jeg må avslutte med skryt! Det er noen grupper som leverer besvarelser som er helt eksepsjonelt gode! Vi har flere eksempler på besvarelser som analyserer problemstillingen på en nær profesjonell måte, som kommer frem til innsikter og konklusjoner basert på datasettet som er nye og genuint interressante, og som viser mye kreativitet og initiativ. Det er åpenbart at mange har brukt tiden godt gjennom semesteret og opparbeidet seg en sterk statistisk intuisjon som har kommet godt med på eksamen. Toppnivået på NHH er ulikt alt jeg har sett som ekstern sensor ved andre institusjoner her i landet.

---