DTU Challenge – Multiple Regression

Vi er ofte interesseret i at se sammenhænge mellem forskellige observationer. Disse sammenhænge kan hjælpe os til at forstå naturvidenskabelige og samfundsmæssige sammenhænge. Den viden kan bruges når politikere og virksomhedsledere skal træffe beslutninger. Det kunne f.eks. være om der er en sammenhæng mellem antal år man uddanner sig og løn bagefter eller om vi får færre trafikofre når vi anlægger flere cykelstier. Det at finde sammenhænge kan også hjælpe os til at få en ide om fremtidig udvikling. Det kunne f.eks. være at få en ide om hvor mange elbiler der vil være om 5 eller 10 år på baggrund af salgstal for elbiler de seneste år. At finde sammenhænge (eller at finde ud af at der ikke er nogen sammenhæng imellem data) er et vigtigt redskab i at forstå nutid og fremtid, og indrette sine handlinger derefter.

Modul 1: Lineær Regression med 2 variable

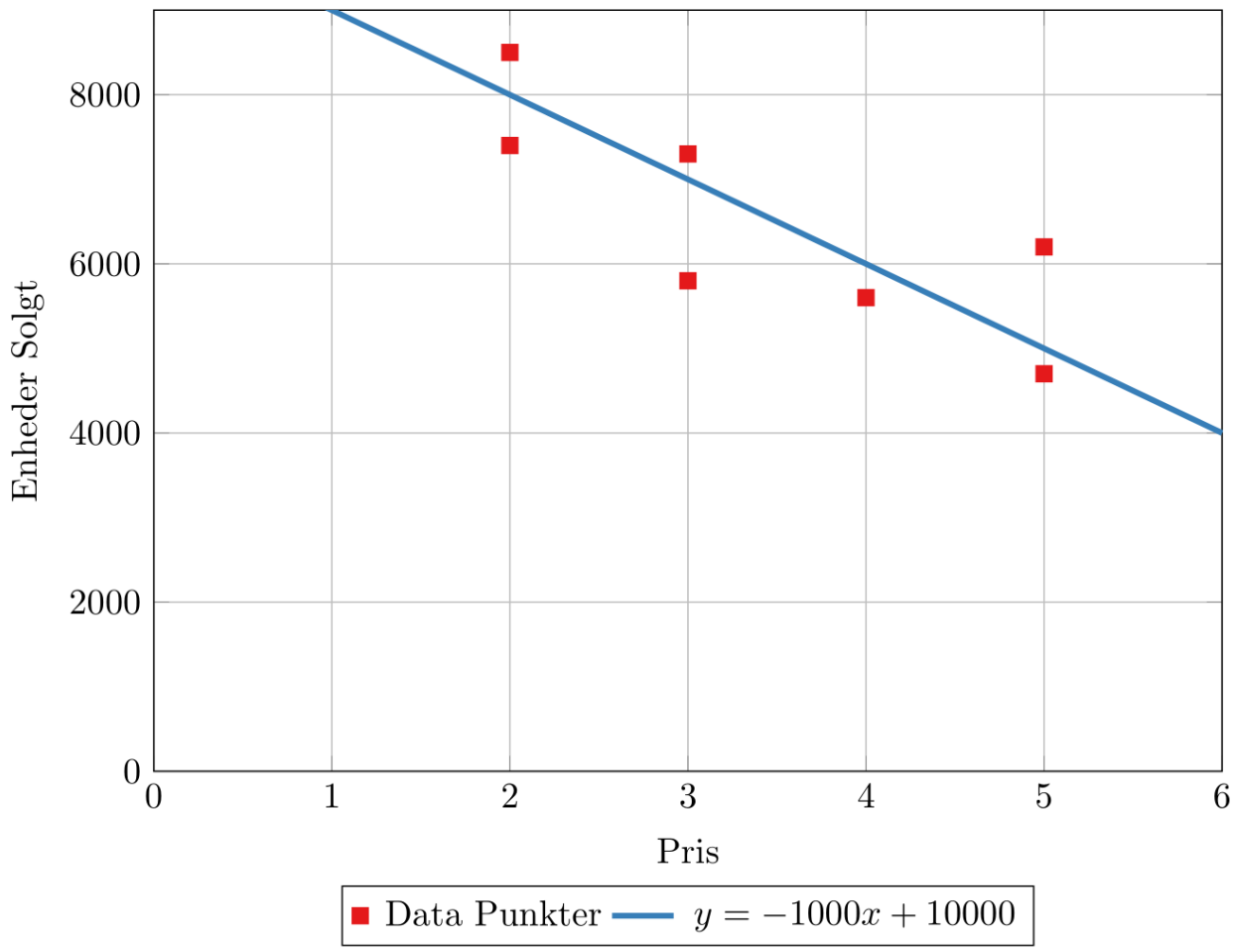
I lineær regressionsanalyse antager man at et antal observationer (data punkter) kan beskrives med en ret linje, derfor handler det om at finde den linje, der passer bedst til den data man vil beskrive. Vi vil i dette afsnit kigge på Lineær regression med 2 variable, og herefter udvide det til flere variable.

Eksempel:

Vi ser på en virksomhed, der sælger en række kasser. De har netop lavet en analyse af købsprisen i forhold til, hvor mange enheder af produktet de sælger. Hypotesen er, jo billigere pris jo flere køber produktet. Nedenstående tabel viser resultatet af analysen. Hvert af data punkterne kaldes også for observationer.

Data Punkt (i)	**Pris (x)**	**Enheder solgt (y)**
1	2 kr.	8500
2	5 kr.	4700
3	3 kr.	5800
4	2 kr.	7400
5	5 kr.	6200
6	3 kr.	7300
7	4 kr.	5600

Plottes disse punkter i 2d med prisen på x-aksen og antal solgte enheder på y-aksen fås nedenstående



De opdager, at punkterne ikke ligger en på en ret linje, men de opdager også, at der ser ud til at være en lineær tendens. Betragt linjen $y = -1000x + 10000$. Der ønskes nu et mål for hvor god denne linje passer til data.

Til dette kan man måle fejlen for hvert data punkt. Lad (x_i, y_i) være det i'te datapunkt. Inkluderes fejlene, ϵ_i, gælder det at

$$y_1 = -1000x_1 + 10000 + \epsilon_1 \Rightarrow 8500 = -1000 \cdot 2 + 10000 + \epsilon_1$$

$$y_2 = -1000x_2 + 10000 + \epsilon_2 \Rightarrow 4700 = -1000 \cdot 5 + 10000 + \epsilon_2$$

$$y_3 = -1000x_3 + 10000 + \epsilon_3 \Rightarrow 5800 = -1000 \cdot 3 + 10000 + \epsilon_3$$

Eller generelt

$$y_i = -1000x_i + 10000 + \epsilon_i$$

Vi kan derfor udregne fejlen ved

$$\epsilon_i = 1000x_i - 10000 + y_i$$

Bemærk at fejlen ved denne måde at regne på kan antage såvel positive såvel som negative værdier. Dermed kunne man opnå at summen af alle fejl blev lig med nul selv om alle punkter ligger langt fra linjen. Derfor bruger man i lineær regression den kvadrerede fejl. Disse værdier vil altid være ikke-negative og dermed vil summen af disse også være ikke-negativ. Som udtryk for hvor stor den overordnet fejl er, bruges fejlenes kvadrat traditionelt set. Her bruges SSE for det engelske 'Sum of Squared Errors'.

$$SSE = \sum_{i=1}^{n} \epsilon_i^2$$ Bemærk at hvis SSE er lige med nul, så ligger alle punkterne på linien.

Opgave:
Udregn SSE for ovenstående eksempel

Udregnes SSE her fås

$$SSE = 3830000$$

Betragt nu linjen $y = -850x_i + 9500$. Den passer bedre til data. For at teste denne påstand kan SSE udregnes, og sammenlignes. Jo mindre SSE er jo mindre er fejlene, og jo tættere vil punkterne ligge på linjen.

Opgave:
Hvilke af de to foreslået linjer passer bedst på data?

Man kan blive ved med at komme med nye linjer, og teste dem, men spørgsmålet er hvordan findes den linje der passer bedst på data? Som sagt, jo mindre SSE er jo mindre er fejlene, vi ønsker derfor at finde den linje der minimere SSE.

En generel linje kan beskrives som $y = ax + b$, og for hvert punkt:

$$y_i = ax_i + b + \epsilon_i \Leftrightarrow \epsilon_i = -ax_i - b + y_i$$

Hvor ϵ_i er afstanden fra punktet (x_i, y_i) til linjen. Bemærk her at alle x_i'erne er kendt og ligeledes er y_i 'erne.

Opskrives SSE nu mere generelt fås

$$SSE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (-ax_i - b + y_i)^2$$
$$= (-ax_1 - b + y_1)^2 + (-ax_2 - b + y_2)^2 + \ ... + (-ax_n - b + y_n)^2$$

Vi ved at en vi kan minimere en funktion ved at differentiere og sætte lig nul. Udfordringen her er at der er to variable, nemlig a og b. Det a og b der minimere ovenstående udtryk kan nu findes vha. teorien om differentiering af funktioner med 2 variable. Vi opfatter først SSE som funktion af a og differentiere efter a og dernæst opfatter vi SSE som en funktion hvor b er den variable størrelse. Her differentiere vi så efter b. At differentiere en funktion af to variable på denne møde kaldes partielt.

Lad os bruge dette til at finde de værdier for a og b som giver en minimal værdig af SSE og dermed den linje der er "tættest" på observationerne.

$$SSE = 92a^2 + 327920000 + 7b^2 + 48ab - 91000b - 296000a$$

Vi kan nu differentiere partielt mht a og vi får:

$$SSE'_a = 184a + 48b - 296000$$

Og hvis vi differentierer partielt mht. b får vi:

$$SSE'_b = 14b + 48a - 91000$$

Vi ved også at hvis SSE skal minimeres, skal der gælde:

$$SSE'_a = 0 \text{ og } SSE'_b = 0$$

Dette er to ligninger med to ubekendte (a og b), og ved at isolere f.eks. a i den ene ligning og indsætte i den anden ligning finder man frem til b, som så igen kan indsættes og vi har dermed udregnet værdierne af a og b.

Det giver i vores eksempel at $a = -823{,}53$ og $b = 9323{,}5$ minimerer SSE. Dermed er den rette linje der fitter dataen bedst $y = -823{,}53x + 9323{,}5$ med en r^2-værdi på $0{,}6536$

Note til Læren:

I denne lektion kan det evt. nævnes hvordan r^2 er defineret og udregnes. Lad $\bar{y}$ være den gennemsnitlige værdi af alle observationerne (6500 for ovenstående eksempel). Med det kan r^2 udregnes som

$$r^2 = 1 - \frac{SSE}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Hvor nævneren kaldes SST, for total sum of squares.

For de to eksempler fås henholdsvis

$$r^2 = 0{,}6200 \quad , \quad r^2 = 0{,}6478$$

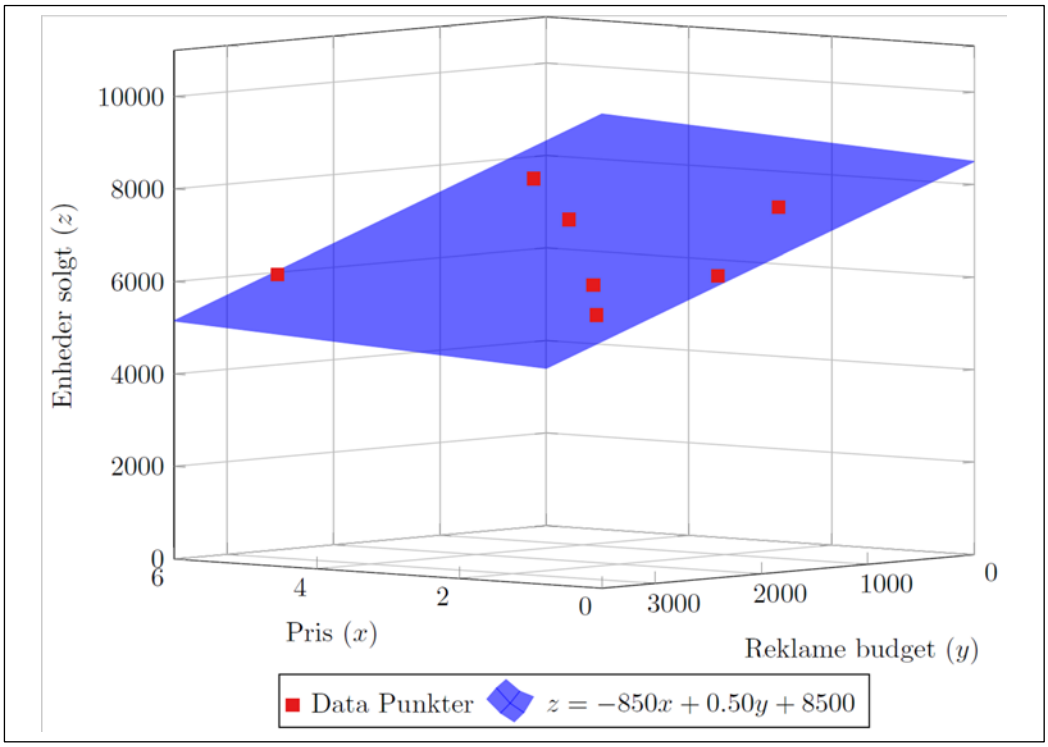
Modul 2: Lineær Regression med 3 variable

Inden for forskellige områder har man forskellige "krav" for hvor god en sammenhæng skal være for man acceptere analysen. Således vil man ofte i samfundsanalyser være tilfreds så snart man kan få r^2 op på 0,65a. Derimod vil man indenfor naturvidenskab og medicin kræve værdier omkring 0,95 eller mere for at godtage analysen.

En r^2 værdi på $0,6536$ er derfor ikke imponerende. Derfor overvejer virksomheden på om der er andre variable der kan influere antallet af enheder solgt. De finder frem til at reklamebudgettet brugt på at markedsføre produkterne også må influere på antallet af solgte enheder. Nedenstående tabel viser det nye data sæt med de 7 observationer udvidet med informationen om de reklamebudgetter der har været foretaget i forbindelse med salget.

Data Punkt (i)	Enheder Solgt (z)	Reklame Budget (y)	Pris (x)
1	8500	2.800 kr.	2 kr.
2	4700	200 kr.	5 kr.
3	5800	400 kr.	3 kr.
4	7400	500 kr.	2 kr.
5	6200	3.200 kr.	5 kr.
6	7300	1.800 kr.	3 kr.
7	5600	900 kr.	4 kr.

Nu svarer hvert data punkt til et punkt i et 3 dimensioneltdimensionalt rum med antal solgte enheder på z aksen og pris og reklamebudget på hhv. x og y aksen. Plottes dette fås følgende billede



Det foreslås nu at planen $z = -850x + 0.50y + 8500$ fitter data. På samme måde som før kan vi udregne fejlene. Lad (x_i, y_i, z_i) beskrive det i'te datapunkt, og lad ϵ_i beskrive fejlen forbundet med det i'te datapunkt og det estimerede plan

$$z_1 = -850x_1 + 0.50y_1 + 8500 + \epsilon_1 \Rightarrow 8500 = -850 \cdot 2 + 0.50 \cdot 2800 + 8500 + \epsilon_1$$

$$z_2 = -850x_2 + 0.50y_2 + 8500 + \epsilon_2 \Rightarrow 4700 = -850 \cdot 5 + 0.50 \cdot 200 + 8500 + \epsilon_2$$

$$z_3 = -850x_3 + 0.50y_3 + 8500 + \epsilon_3 \Rightarrow 5800 = -850 \cdot 3 + 0.50 \cdot 400 + 8500 + \epsilon_3$$

Eller mere generelt

$$z_i = ax_i + by_i + c + \epsilon_i$$

Isoleres ϵ_i heri fås

$$\epsilon_i = -ax_i - by_i - c + z_i$$

Som før kan SSE nu udregnes ved

$$SSE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (-ax_i - by_i - c + z_i)^2$$

Minimeres SSE fås det at $a = -835{,}722$, $b = 0{,}592228$ og $c = 8536{,}214$. Således er det planen $z = -835{,}722x + 0{,}592228y + 8536{,}214$ der fitter bedst til data. Igen kan vi udregne r^2 værdien. Den er $0{,}9617$. Dette indikere at den valgte plan er et meget godt estimat for sammenhængen mellem pris, antal solgte enheder og det brugte reklamebudget.

Opgave:
Hvor mange enheder kan virksomheden forvente at sælge hvis prisen på et nyt produkt sættes til 3 kroner, og reklamebudgettet til 2500?

Opgave:
For et andet nyt produkt er reklamebudgettet fastlagt til maksimalt 5000. Hvad skal prisen være for at maksimere den forventelige indkomst?

Denne teori kan let udvides til at gælde for et vilkårligt antal dimensioner. Hvis vi betragter n dimensioner kan vi opskrive ligningen

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \cdots \alpha_n x_n + \beta$$

Hvor x'erne er vores variable, og α'erne er deres koefficienterne, og β beskriver skæring med y-aksen. For at finde værdierne for α'erne og β kan vi opstille følgende ligninger

$$\begin{aligned} z_1 &= \alpha_1 x_{11} + \alpha_2 x_{12} + \cdots + \alpha_n x_{1n} + \beta + \epsilon_1 \\ z_2 &= \alpha_1 x_{21} + \alpha_2 x_{22} + \cdots + \alpha_n x_{2n} + \beta + \epsilon_2 \\ &\vdots \\ z_m &= \alpha_1 x_{m1} + \alpha_2 x_{m2} + \cdots + \alpha_n x_{mn} + \beta + \epsilon_m \end{aligned}$$

Hvor m er antallet af data punkter. Vi antager her at x-variablene er uafhængige af hinanden. SSE kan nu udregnes tilsvarende som før. Hvor vi i eksemplet ovenfor havde to ligninger med to ubekendte efter partiel differentation, får vi nu n ligninger med n ubekendte (α'erne).

Modul 3: Multivariabel lineær regression

Når vi går fra to variable til tre og flere variable bliver metoden omtalt som multivariable lineær regression. Formålet er stadig det samme, nemlig at finde en lineær sammenhæng mellem observationer, som nu ikke længere blot består af to variable, men tre eller flere.

Multivariabel lineær regression har mange styrker, men for at vurdere om det er det rette værktøj til en given problemstilling er det vigtigt at kende dets begrænsninger, og hvilke antagelser der må tages før det kan anvendes.

Antagelser:

- Det antages at der er en lineær sammenhæng mellem den afhængige variabel (i foregående eksempel; Enheder solgt) og de uafhængige variable (Pris og reklame budget)
- De uafhængige variable er uafhængige af hinanden.
- Variablene i modellen optræder udelukkende i første potens, og multipliceres ikke med hinanden. Modellen er altså lineær.

Grundstenen i en multivariabel lineær regression er et valg af et sæt beskrivende variable der beskriver den afhængige variable, herved forstås en *model*. Et valg af andre variable giver dermed en anden model.

Opgave:
Vi begynder nu at kigge på challenge casen. Hvilken model (og data) kan bruges til at beskrive antallet af passagerer på buslinjen 7B i København?

Når man vælger en model, er det vigtigt at overveje, om modellen passer og giver mening. Tilføjelsen af ekstra uafhængige variable vil altid resultere i en højere r^2 værdi. Derfor kan tilføjelsen af for mange uafhængige variable uden teoretisk begrundelse resultere i en model, der er overfittet. I det meste data er der en naturlig variation (støj), denne underliggende variation ønsker man ikke at repræsentere i en model. Man siger derfor at en model er overfittet, hvis den svarer for præcist til et datasæt, og en overfittet model vil derfor ikke kunne give pålidelige estimater på fremtidige observationer. Derfor er det ikke tilstrækkeligt at kigge på SSE for at vurdere om en model er bedre end en anden. I stedet måles en models kvalitets ved, hvor god den er til at forudsige kommende data. Til dette deler man normal vis sit datasæt op i to; et træningssæt og et testsæt.

Træningssæt:
Træningsdatasættet anvendes til at fitte parametrene i modellen, og træne din model.

Testsæt:
Efter modellen er fittet, skal den 'møde virkeligheden', og ved at anvende modellen på test sættet kan det måles hvor meget de estimerede værdier afviger fra de rent faktiske observerede værdier.

Den mest simple måde at opdele det fulde datasæt i et træningssæt og et testsæt er at holde noget data skjult under træningsprocessen, og derefter evaluere modellen på baggrund af det skjulte data. Normalvis bruges en fordeling af data på 70%/30% træning/test. Det er dog vigtigt at data udvælges på en sådan måde at træningssættet giver et repræsentativt billede af det fulde data. Typisk opnås dette ved at udvælge data tilfældigt.

En anden vigtig ting at vide om multivariable lineær regression, er hvordan kategoriske variable skal håndteres. En kategorisk variabel er en variabel som kun kan antage et fast antal mulige værdiger. I 7B ksemplet kan eksempler på kategoriseret data for eksempel være dag (som kun kan være en af værdierne mandag til søndag) og tidspunkt (som kun kan være et tidpunkt i løbet af et døgn fra 00:00 til 23:59).

Som eksempel, lad os kigge på nedenstående data, der beskriver antallet af brugere af 7B over en uge.

Data Punkt (i)	Antal brugere (y)	Dag (x)
1	15863	Mandag
2	14682	Tirsdag
3	15247	Onsdag
4	14953	Torsdag
5	13952	Fredag
6	8523	Lørdag
7	7569	Søndag

Her er variablen 'dag' en kategoriseret variable, som ikke lige umiddelbart er klar til indsættelse i en model. Før det kan bruges skal dagene omsættes til numeriske værdier.

I stedet kan der laves 7 nye variable, en for mandag, en for tirsdag osv. Disse variable kan nu bruges til at beskrive hvilken dag det er. For hvert datapunkt vil netop én af disse variable have værdien 1, resten vil være 0. Altså, vil ovenstående eksempel skrives på som vist herunder.

Data Punkt	Antal brugere	Mandag	Tirdag	Onsdag	Torsdag	Fredag	Lørdag	Søndag
1	15863	1	0	0	0	0	0	0
2	14682	0	1	0	0	0	0	0
3	15247	0	0	1	0	0	0	0
4	14953	0	0	0	1	0	0	0
5	13952	0	0	0	0	1	0	0
6	8523	0	0	0	0	0	1	0
7	7569	0	0	0	0	0	0	1

Opgave:
Nedenstående tabel viser antallet af brugere af 7B på i forskellige tidsrum i løbet af en eftermiddag.

Datapunkt	Antal brugere	Tidsrum
1	839	12:00-12:59
2	884	13:00-13:59
3	1151	14:00-14:59
4	1683	15:00-15:59
5	1720	16:00-16:59

6	1379	17:00-17:59

Omskriv dette datasæt således det ikke bruger kategoriseret variable.

Modul 4: Smart Mobility data tool

Vi vil i dette modul introducere værktøjet, som er tiltænkt til brug under udførelsen af casen. Vi vil give en introduktion til værktøjet, samt gå igennem eksempler der viser brugen af værktøjet.

Vi fortsætter med at se på antal passagerer på buslinjen 7B i København på flere måneder, samt antal hverdage og weekender.

Month (x_1)	Week days (x_2)	Weekend days (x_3)	Passengers (y)
1	21	10	372928
2	21	9	326168
3	21	10	368809
4	23	8	395092
5	21	9	402427
6	22	9	439148
7	22	8	436594
8	19	12	454490
9	22	9	448616
10	20	8	452250

Da det forventes at antallet af hverdage og weekenddage influerer antallet af passagerer på en måned, foreslås modellen

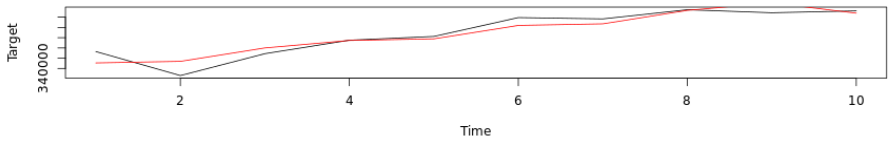
$$y = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \beta$$

For at estimere denne model bruges Smart Mobility data tool. Gå ind på http://www.dtuchallenge.dk og find datasættet *train-mini.csv*. Gå ind på https://man-vm-datachal1.win.dtu.dk/datachal2018

- Under *Browse* vælges training filen.
- For at estimere den ovenstående model, vælges alle kolonnerne under *Choose columns.*
- Tryk nu på *Estimate Linear Model*
- Herefter kan du øverste se et mål for den absolutte gennemsnitlige fejl (*Train Score (MAE)*).
- For nu at teste modellen på rigtigt ukendt data downloades og indlæses filen *test-mini.csv.*
- Tryk nu på knappen *Predict*

Den første del af outputtet viser Eksempel på træningsdataen. Dette indeholder et plot af target værdierne (sort linje), samt de af modellen estimerede værdier (rød linje). Tabellen nedunder viser et udsnit af det samme data.

Example of train data



Month	Weekday	Weekend	CheckIn	Predicted
1	21	10	372928	350752.59
2	21	9	326168	353827.72
3	21	10	368809	380016.20
4	23	8	395092	394552.98
5	21	9	402427	397723.13
6	22	9	439148	423864.10

Outputtet fra model estimatet ses under *Estimated Model Parameters.*

```
Call:
lm(formula = form, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-27660  -8271   2827   8417  22175

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -21138     232562  -0.091 0.930536
Month          14632       2290   6.389 0.000692 ***
Weekday        11509       8157   1.411 0.207931
Weekend        11557       7391   1.564 0.168921
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18790 on 6 degrees of freedom
Multiple R-squared:  0.8767,    Adjusted R-squared:  0.8151
F-statistic: 14.22 on 3 and 6 DF,  p-value: 0.003905
```

De værdier der minimerer SSE kan nu aflæses under *Estimate*, og det ses at værdierne er

$$\alpha_1 = 14632, \quad \alpha_2 = 11509, \quad \alpha_3 = 11557, \quad \beta = -21138$$

Yderligere ses det at $R^2 = 0.8767$.

Note til Læren:
For mere information om dette output se evt.

Til sidst i outputtet ses modelprædiktionerne på testdatasættet.

Opgave:
For det samme datasæt. Hvis vi kun tager måneden samt antal hverdage i måneden fås nedenstående model.

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \beta$$

Estimér nu denne model vha. Smart Mobility data tool. Hvilken af disse modeller er bedst til at forudsige kommende observationer?

Vi Introducerer nu Smart Mobility Challenge datasættet. Datasættet indeholder følgende kolonner

- **Date** (x_1): Datoen
- **Day** (x_2): Hvilken dag på ugen det er.
- **DayType** (x_3): Om det er hverdag, weekend, jul, påske, eller Skole ferie
- **StattutoryHoliday** (x_4): Om det er en helligdag.
- **Hour** (x_5): Hvilken time på dagnet det er. 1H svarer til 01:00 – 01:59.
- **Temperature** (x_6): Den gennemsnitlige temperatur i Københavns området i den pågældende time.
- **WindSpeedMS** (x_7): Den gennemsnitlige vindhastighed i Københavns i din pågældende time.
- **Condition** (x_8): Vejrforholdet, om det er klart, skyet , regnvejr eller snevejr.
- **TotalPassengerCount** (y): Antallet af passagere der er checket på 7B i den pågældende time.

Træningsdatasættet indeholder data for oktober 2016 til og med februar 2018. I Smart Mobility Challenge skal du lave en model, der på baggrund af dette træningssæt estimerer antallet af passagerer på 7B i marts 2018, givet at alle vejrinformationerne er kendte (temperatur, vindhastighed og vejrforhold).