Labbrapport i Statistik

Laboration 1

Kurskod: 732G53

Viet Tien Trinh
Duy Thai Pham

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet
2024-08-30

Innehåll

1. Introduktion

Denna rapport använder en försäljningsdata som handlar om månadsförsäljning hos en aktör och inkluderar alla variabler som kanske är relevanta till den beroende variabeln, vilket är månadsförsäjning. Dessa variabler är exempelvis annonsbudget, antal produkter, osv. Rapportens syftet är att skapa en linjär regression modell för alla variabler som kanske har ett samband med månadsförsäljning och kontrollera samt analysera om modellen är bra och lämplig.

2. Uppgifter

A) Utforska ert datamaterial med hjälp av beskrivande statistik och visualiseringar

i) Sammanställ en tabell med varje variabel och dess variabeltyp och skala

Tabell 2.1: Tabellen med de första 9 observationerna

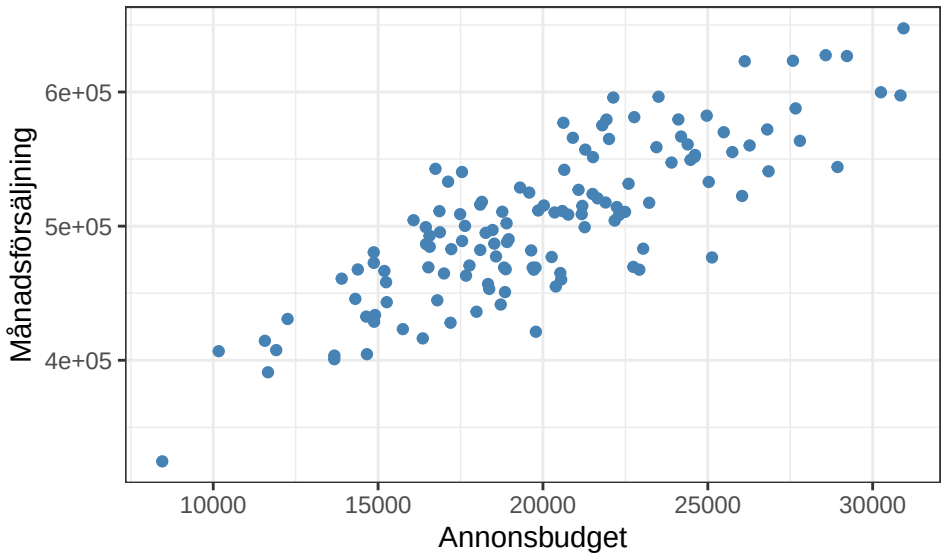
monthly_sales	ad_budget	num_products	store_size	days_open	store_type	region
427963.24	17197.62	644	1739.44	24	Online	Rural
450816.27	18849.11	545	1755.06	19	Online	Rural
563602.07	27793.54	504	2023.58	28	Online	Rural
510222.14	20352.54	457	2650.10	23	Brick-and-Mortar	Rural
541997.62	20646.44	294	3146.54	24	Brick-and-Mortar	Rural
627461.03	28575.32	613	2773.79	30	Brick-and-Mortar	Rural
508219.57	22304.58	353	1933.42	23	Brick-and-Mortar	Rural
403518.57	13674.69	573	1121.74	28	Brick-and-Mortar	Rural
484628.38	16565.74	690	1805.61	22	Online	Urban

monthly_sales: Månadsförsäljning i dollar.
ad_budget: Annonsbudget i dollar.
num_products: Antal produkter sålda per månad.
store_size: Butiks storlek i kvadratmeter.
days_open: Antal öppetdagar av butiken per månad.
store_type: Typ av butik (Online, Brick-and-Mortar).
region: Region (Urban, Rural).

Tabell 2.2: Tabellen över alla variabler och dess egenskaper

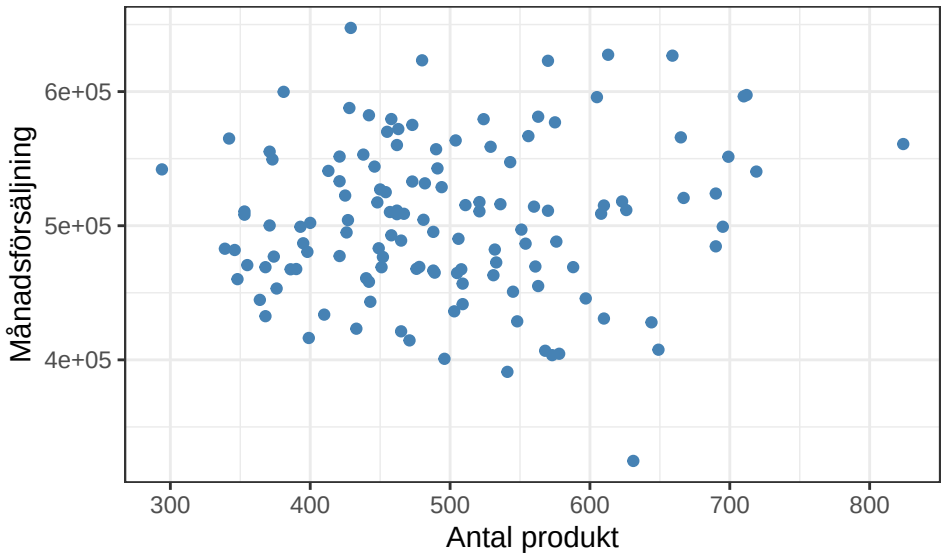
Variabler	Typ_Av_Variabler	Skala
Monthly_sales	Kontinuerligt kvantitativ	kvotskala
Ad_budget	Kontinuerligt kvantitativ	kvotskala
Num_product	Diskret kvantitativ	kvotskala
Store_size	Kontinuerligt kvantitativ	kvotskala
Days_open	Diskret kvantitativ	kvotskala
Store_type	Kvalitativ	nominalskala
Region	Kvalitativ	nominalskala

ii) Skapa diagram för parvisa samband mellan förklarande och responsvariabeln



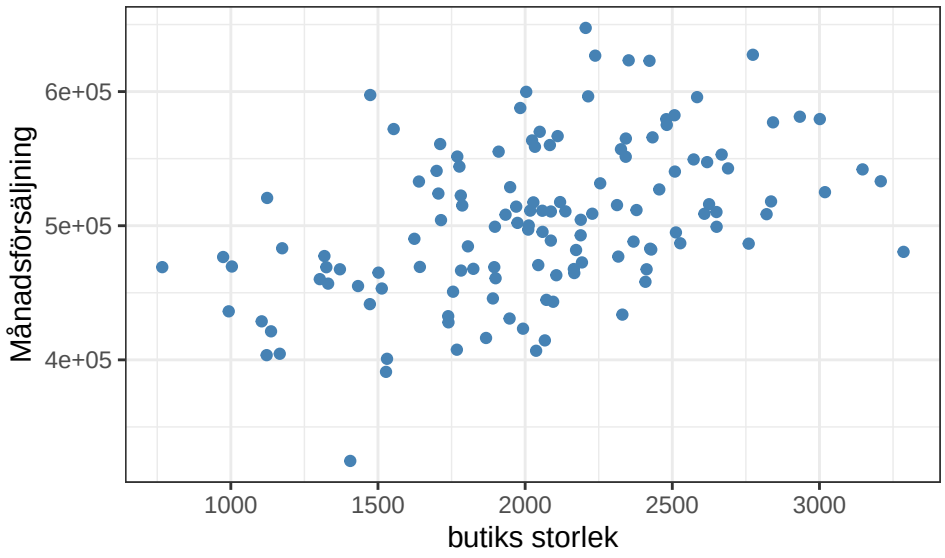
Figur 2.1: Spridningsdiagrammet mellan annonsbudget och månadsförsäljning

Sambandet mellan månadsförsäljning och annonsbudget är linjärt med en konstant förändring. I detta fall leder en konstant ökning av annonsbudget till en konstant ökning av månadsförsäljning. Trots att punkterna är måttligt utspridda, så följer majoriteten av punkterna denna trend, vilket tyder på ett starkt samband. Med hjälp av Pearson's korrelationstest i R får vi fram cor = 0.8232545, vilket bidrar till tidigare resonemang att det är ett starkt samband. Enligt figuren finns det ett extremvärde som avviker långt ifrån andra punkter men detta extremvärde verkar också följa trenden. Det är (x = 8454.156, y = 324647.7) .
Sammanfattningsvis är sambandet mellan månadsförsäljning och annonsbudget:
1. Linjärt,
2. Positivt,
3. Starkt,
4. Med ett extremvärde.



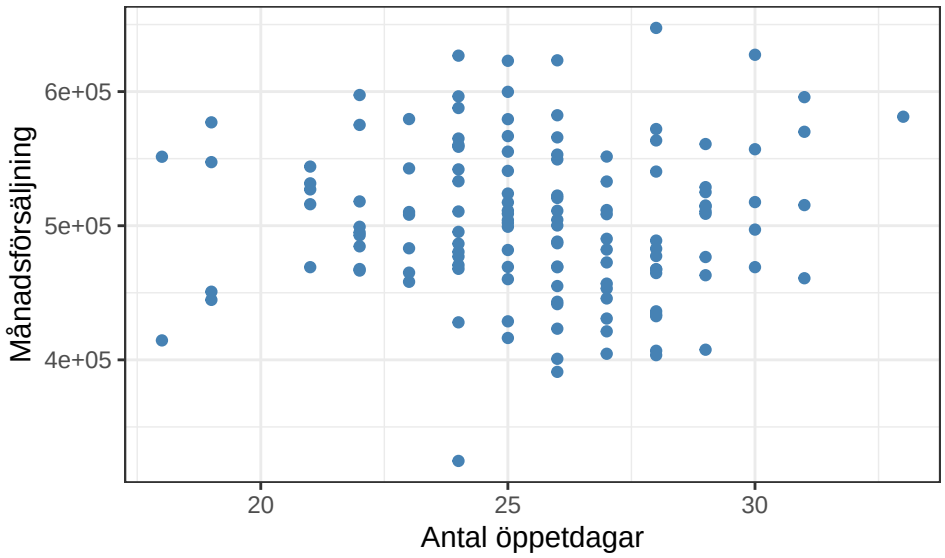
Figur 2.2: Samband mellan antal produkt och månadsförsäljning

Mellan antal produkt och månadsförsäljning verkar det inte finnas något samband. Observationer är mycket utspridda och det finns inget tydligt tecken på att sambandet kan vara linjärt, positivt eller negativt. Enligt figuren syns det även flertal extrempunkter som avviker ifrån samlingen av punkterna. Exempelvis har vi punkten (x = 631, y = 324647.7), (x = 824, y = 560901.6), eller (x = 429, y = 647469.5).
Sammanfattningsvis finns det inget samband mellan antal produkt och månadsförsäljning och utifrån spridningsdiagrammet mellan dem finns det ett flertal extremvärden.



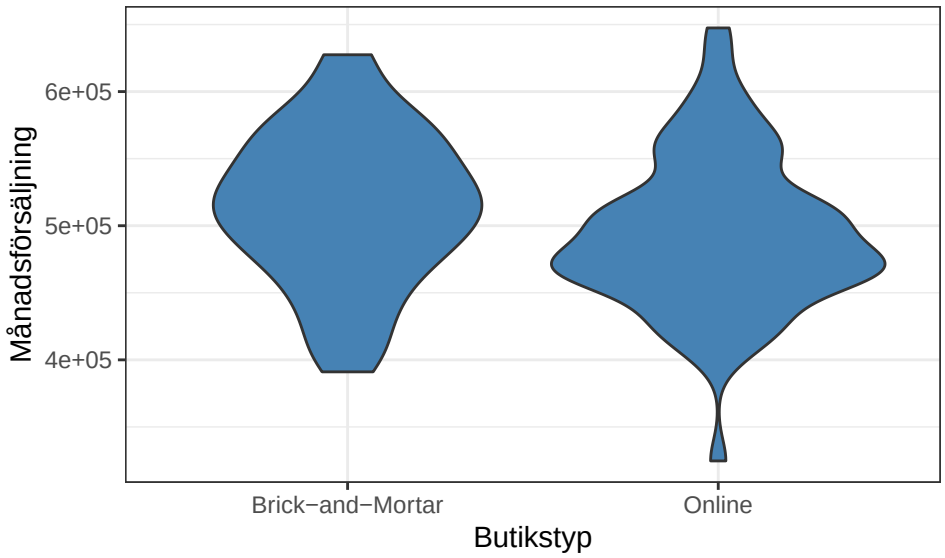
Figur 2.3: Samband mellan butiks storlek och månadsförsäljning

Sambandet mellan månadsförsäljning och butiks storlek är linjärt med en konstant förändring. I detta fall leder en konstant ökning av butiks storlek till en konstant ökning av månadsförsäljning. Mojaroteten av observationerna följer denna trenden, vilket tyder på att det finns ett linjärt positivt samband. Dock är punkterna måttligt utspridda och det finns punkter som avviker ifrån detta som exempelvis har mindre butiks storlek men samma månadsförsäljning som punkter med större butiks storlek. På grund av detta är sambandet medelstarkt. Med hjälp av Pearson's korrelationstest i R får vi fram cor = 0.4592714, vilket bidrar till tidigare resonemang att det är ett medelstarkt samband. Enligt figuren finns det några extremvärde som exempelvis (x = 1405.783, y = 324647.7), (x = 3285.729, y = 480511.8).
Sammanfattningsvis är sambandet mellan månadsförsäljning och butiks storlek:
1. Linjärt,
2. Positivt,
3. Medelstarkt,
4. Med några extremvärde.



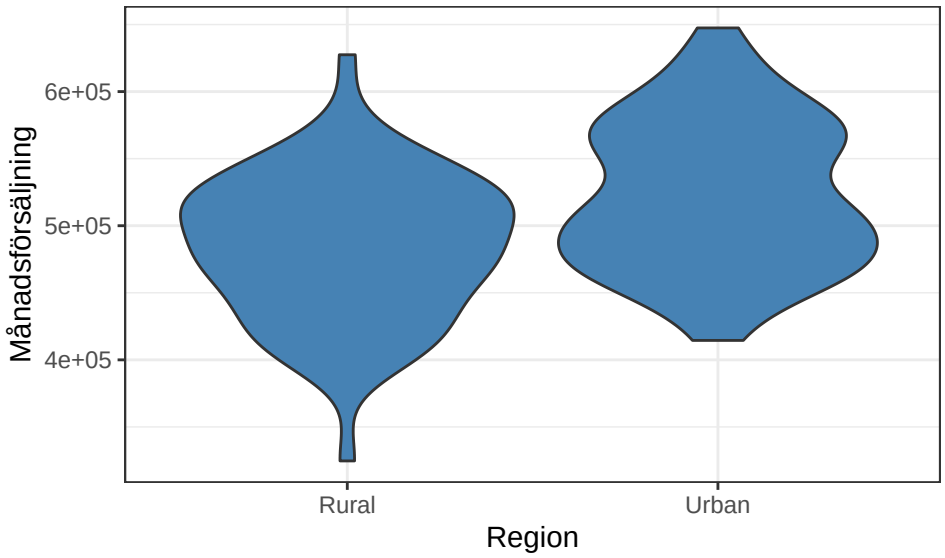
Figur 2.4: Samband mellan antal öppetdagar och månadsförsäljning

Mellan antal öppetdagar och månadsförsäljning verkar det inte finnas något samband. Fördelningen av observationer är utspridda och det finns inget tydligt tecken på att sambandet kan vara linjärt, positivt eller negativt. Enligt figuren syns det även flertal extrempunkter som avviker ifrån samlingen av punkterna. Exempelvis har vi punkten (x = 24, y = 324647.7), punkten (x = 33, y = 581276.9) eller punkten (x = 18, y = 414499.2). Sammanfattningsvis finns det inget samband mellan antal öppetdagar och månadsförsäljning. Utifrån spridningsdiagrammet mellan dem finns det ett flertal extremvärden.



Figur 2.5: Samband mellan butikstyp och månadsförsäljning

Figur ovan visar två annorlundar form och både är ganska symmetrisk. Med butikstyp “Online” har ett större månadsförsäljningstintervall än butikstyp “Offline” men man kan lätt hitta butikstyp “Offline” har månadsförsäljning i genomsnitt större än “Online”. Det verkar finnas ett samband mellan butikstyp och månadsförsäljning, vilket ska inkluderas i modellen.



Figur 2.6: Samband mellan region och månadsförsäljning

Figur ovan visar två annorlundar form och de är inte symmetrisk. Region “Rural” har ett större månadsförsäljningstintervall än region ”Urban” men region “Urban” har månadsförsäljning i genomsnitt större än “Ruval”. Med andra ord kan man säga att region ”Urban” har i genomsnitt flera observationer än region ”Rural” när det kommer till större månadsförsäljning. På grund av detta verkar det finnas ett samband mellan region och månadsförsäljning, vilket kommer att inkluderas i modellen.

iii) Motivera vilka variabler som bör inkluderas i regressionsmodellen och vilken struktur de bör ha

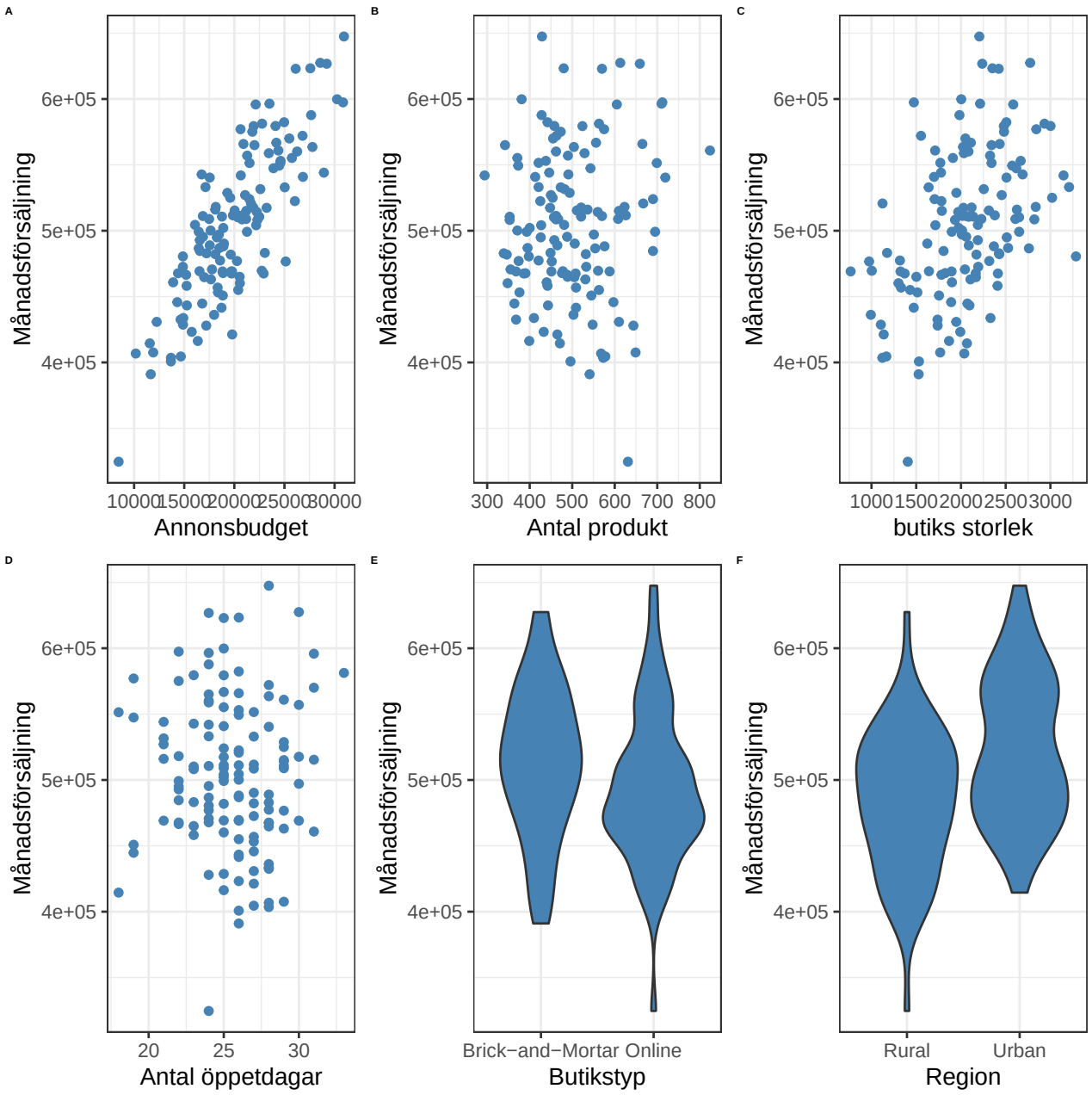
Med hjälp av Pearson's korrelationstest i R får vi fram cor = 0.08511079,

```
##
##  Pearson's product-moment correlation
##
## data:  data$ad_budget and data$monthly_sales
## t = 16.407, df = 128, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7586026 0.8718529
## sample estimates:
##       cor
## 0.8232545
```

vilken är correlation mellan antal produkt och månadsförsäljning 2.2.
Och cor = -0.005425397,

```
##
##  Pearson's product-moment correlation
##
## data:  data$days_open and data$monthly_sales
## t = -0.061382, df = 128, p-value = 0.9512
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1774457  0.1669166
## sample estimates:
##          cor
## -0.005425397
```

vilken är correlation mellan antal öppetdagar och månadsförsäljning 2.4.
Både correlation är mycket nära 0 och tyder på att sambandet är mycket svag. Med andra ord kan vi säga att det inte finns något samband mellan antal produkt, antal öppetdagar och månadsförsäljning så utslutas de 2 variabler från modellen. Man kan hänvisar figure 2.7.



Figur 2.7: Linjär regression

Data innehåller 2 kvalitativ variabler butikstyp och region men regressionsmodellen kan inte hantera kvalitativa variabler direkt så transformeras de 2 kvalitativ variabeln numerisk genom indekatorvariabler.

$$\text{Butikstyp} = \begin{pmatrix} 1 : \text{Online} \\ 0 : \text{Brick and Mortar} \end{pmatrix}$$

och

$$Region = \begin{pmatrix} 1 : \text{Urban} \\ 0 : \text{Ruval} \end{pmatrix}$$

Modellens struktur bli:

$$\text{Månadsförsäljning} = \beta_0 + \beta_1 * \text{annonsbudget} + \beta_2 * \text{butiks stolek} + \beta_3 * \text{Online} + \beta_4 * \text{Urban} + \text{E}$$

Eller kan beteckna den anpassade modellen med dess skattade parametrar enligt:

$$\widehat{Y}_i = \beta_0 + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \beta_3 * X_{3i} + \beta_4 * X_{4i}$$

Där:

$\widehat{Y}_i$: Månadsförsäljning eller responsvariabelns skattande för observation i
$\beta_1 - \beta_4$: Skattning av lutningsparametrar
β_0 : Skattning av interceptet

B) Anpassa en regressionsmodell enligt a.iii)

i) Sammanställ en tabell för modellens skattade koefficienter. Tolka parameterskattningarna för respektive parameter

Tabell 2.3: Tabellen för modellens skattande koefficienter.

	Skattning	Medelfel	t-värde	p-värder
(Intercept)	202455.9937	7631.6581	26.53	0.0000
ad_budget	10.1558	0.2825	35.95	0.0000
store_size	43.6639	2.4854	17.57	0.0000
store_typeOnline	-16929.0410	2512.6811	-6.74	0.0000
regionUrban	33287.2857	2483.1269	13.41	0.0000

Första värde av skattning parameter är skattning av interceptet eller β_0, dvs att om det finns ingen några förklarar variabler eller alla de värde bli 0, så kommer värde av responsvariabeln att bli 202455.9937 dollar.

Samtidigt, data innehåller två type av kategoriska variabler så skattning parametrar visa liten annorlunda mellan dem. Exempelvis man kan se med varje dolar av kvantitativet variabler annonsbudget ökar mer månadsförsäljning ökar 10.1558 dollar givet att alla andra variabler hålls konstansta. Det också gäller för variabeln butiks storlek.

Efter transformeras kvalitativ variabler numerisk genom indikatorsvariabler tolkas de inom sina grupp jämför med referensvariabler, till exempel har butikstyp “online” månadsförsäljning i genomsnitt mindre -16929.0410 än referenskategoris “offline” givet att alla andra variabler hålls konstansta.

ii) Beräkna en intervallskattning för den första och sista lutningsparametern i tabellen. Visa beräkningarna i ekvationsmiljön och tolka intervallet

Konfidensintervall ekvation för β :

$b_j \pm \mathrm{t}_{n-(k+1);1-\alpha/2} * S_{b_j}$

Där:
b_j : Skattning av lutningparametrar
$t_{n-(k+1);1-\alpha/2}$:
· Antal observation n = 130
· Antal förklarade variabler k = 4
· $\alpha = 0.05$
$S_{b_j} = \frac{S_{y|x}}{S_x * \sqrt{n-1}}$

```
## [1] "# Med hjälper från R"
##                     2.5 %       97.5 %
## (Intercept) 2.546175e+05 301590.59864
## ad_budget   9.311346e+00     11.58541
## regionUrban 2.195564e+04  42028.04071
```

Intervallsskattning för första lutningparameter annonsbudget med 95% säkerhet:
LCL: 9.31
UCL: 11.59

Intervallsskattning för sista lutningsparameter region med 95% säkerhet:
LCL: 21955.64
UCL: 42028.04

iii) Sammanställ en ANOVA-tabell och genomför ett test för hela modellens signifikans. Presentera hypoteserna, testvariabeln, kritiska värdet eller p-värdet, samt de tillhörande slutsatserna och tolkning

Tabell 2.4: ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ad_budget	1	291423901544.89	291423901544.89	1464.90	0.0000
store_size	1	68592330652.13	68592330652.13	344.79	0.0000
store_type	1	9355301705.87	9355301705.87	47.03	0.0000
region	1	35749897202.82	35749897202.82	179.70	0.0000
Residuals	125	24867142548.72	198937140.39		

Vi undersöker hypoteserna:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_a : Minst en av b_j *i* H_i är skild från 0

Med hjälp av tabellen ovan får vi fram testvariabeln.

$F_{test} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-(k+1)}} = \frac{MSR}{MSE} = 509.1073$

Vårt p-värdet är mindre än 5%. Allt detta innebär att Ho kan förkastas och minst en av variablerna har ett samband med responsvariabeln som är månadsförsäljning i vårt fall. Enligt Ha är minst en lutningsparameter signifikant och med andra ord, bland de variabler som ingår i modellen finns det minst en variabel som bidrar med förklarad variation. I jämförelsen med en modell med enbart $\overline{Y}$ som en baslinje kan vi säga att vår anpassad modell är bättre.

iv) Undersök om de två kvalitativa variablerna kan tas bort från modellen samtidigt. Presentera hypoteserna, testvariabeln, kritiska värdet eller p-värdet, samt de tillhörande slutsatserna och tolkning

Tabel 2.5: Bearbetad och förenklad ANOVA tabell

Tabell 2.5: Bearbetad och förenklad ANOVA tabell

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	4	405121431106	101280357776	509.1073	0
Residuals	125	24867142549	198937140	NA	NA

Tabell ovan visar förenklad ANOVA tabell av fullständigt förklarad variabler som finns i modellen.

Tabell 2.6: Bearbetad och förenklad ANOVA tabell

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	2	360016232197	180008116099	326.7152	0
Residuals	127	69972341457	550963319	NA	NA

Tabell ovan visar förenklad ANOVA tabell av 2 kvantitativ förklarad variabler som finns i modellen.

Vi undersöker hypoteserna:

$H_0 : \beta_{storetypeOnline} = \beta_{regionUrban} = 0$

H_a : Minst en av β_j i H_0 är skild från 0

Testvariabel: $F_{test} = \frac{\frac{(SSR_F - SSR_R)}{s}}{\frac{SSE_F}{n-(k+1)}} = \frac{\frac{(405121431106 - 360016232197)}{4-2}}{\frac{24867142549}{125}} = 113.36546$

P-värde med hjälp från R:

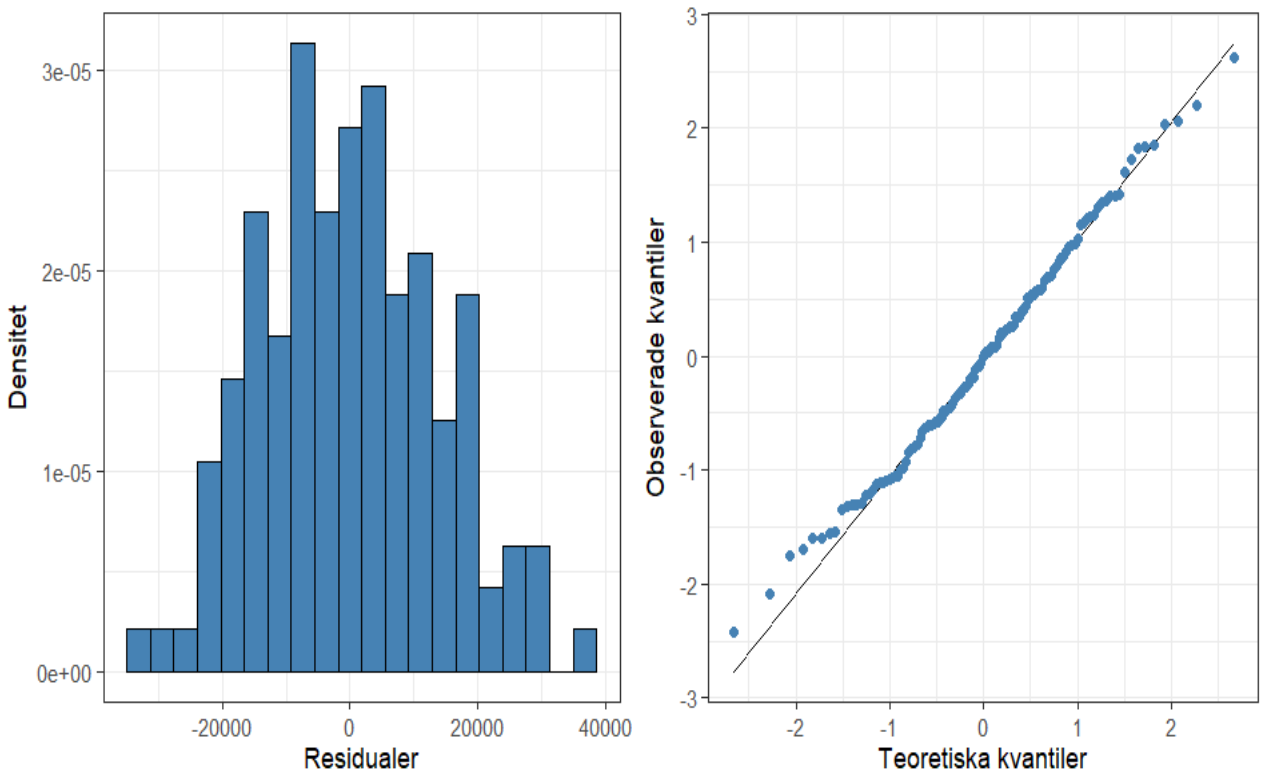
```
p_varde = 1 -pf(113,3,df1=2,df2=125)
p_varde

## [1] 8.295813e-10
```

Med hjälp av testvariabeln , kan vi räkna fram till att vårt p-värde är mindre än 5% signifikansnivå. Allt detta innebär att Ho kan förkastas och att minst en av de kvalitativa variablerna har ett samband med responsvariabeln som är månadsförsäljning i denna fall. Utifrån Ha kan man därmed säga att region och butikstyp tillsammans bidrar signifikant till att förklara variationen i månadsförsäljning.

C) Genomför en residualanalys på den valda modellen

i) Bedöm varje regressionsantagande med hjälp av lämpliga visualiseringar

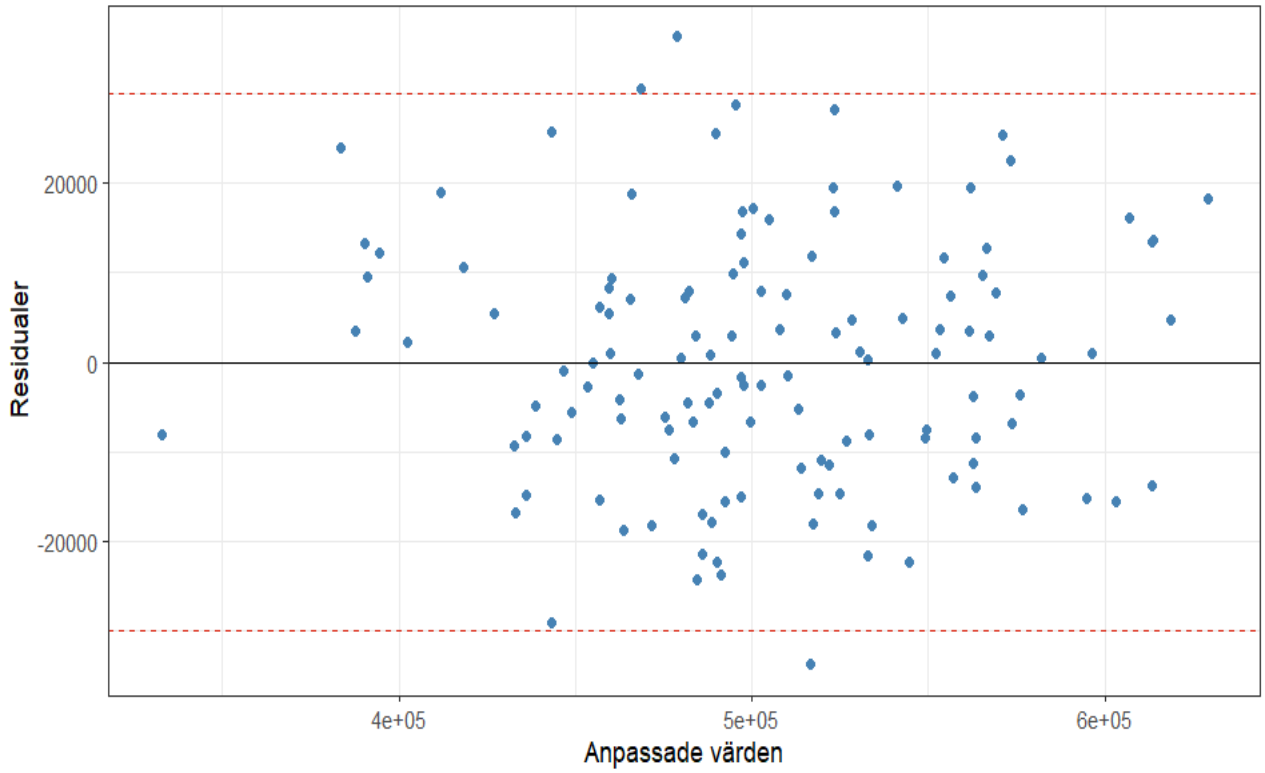
1. Normalfördelning



Figur 2.8: Normalfördelning

Syftet av detta avsnitt är att se normalfördelnings sysmetriska och centrerad runt noll, det härleds från residuals modellen E$\sim N(0, \sigma^2)$.Inte perfekt men histagrammet visas ganska symmetrisk och det har också klockliknade form med toppen av klocktornet centrerad runt väntevärder 0. Samtidigt på höger av figure 2.8
visar QQ-diagrammet främst observade följa inritade ljnie, även om detta minska i båda sidan men det är forfarande visar en perfekt normalfördelning totalt sett.

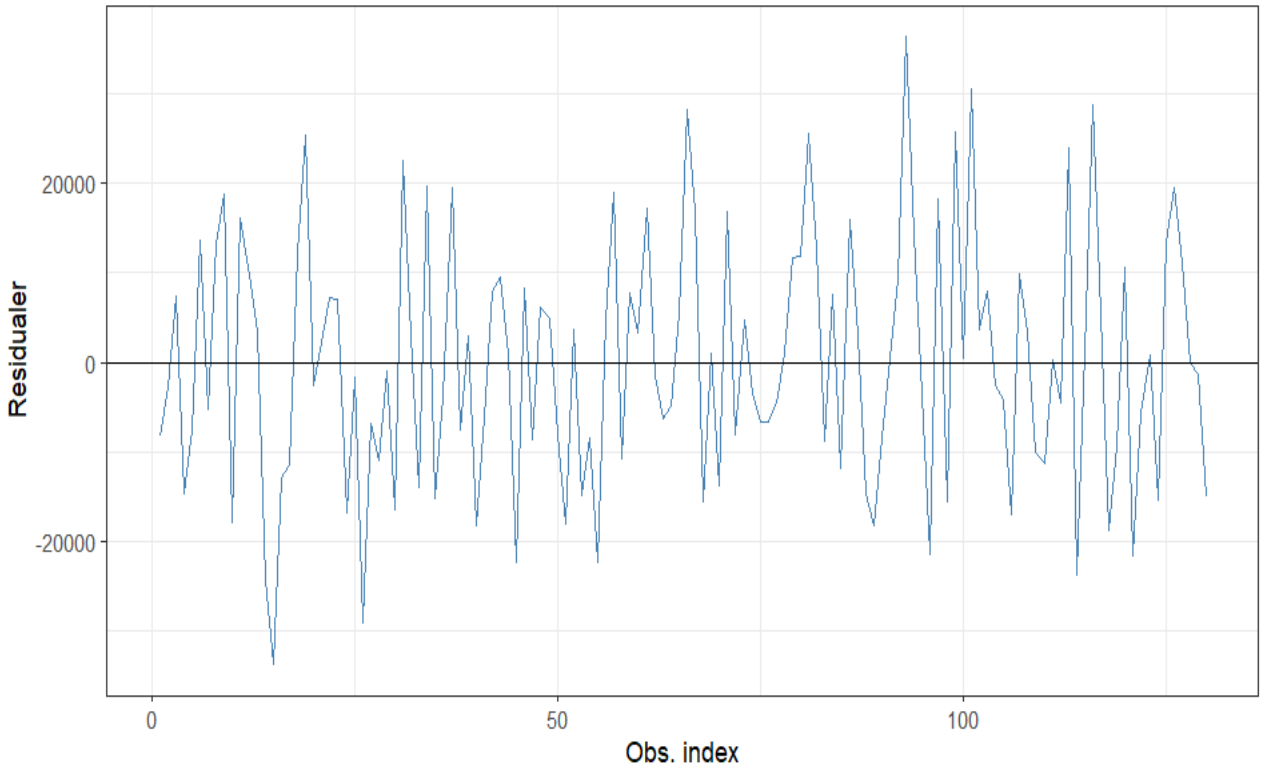
2. Homoskedasticitet



Figur 2.9: Residualernas spridning mot anpassade värden

Detta spridningsdiagram visar residualerna på y-axeln och de anpassade värdena (predicerade värden) på x-axeln. Man kan hitta det verka finns inte någon tydligt systematiskt mönster till exempel stigande eller fallande från vänster till höger och tvärtom i hur residualerna fördelar sig kring noll-linjen utan sprids slumpmässigt. och Här verkar det dock finns relativt jämn spridning, vilket indikerar att variansen i felen inte förändras systematiskt med de anpassade värdena, det vill säga att denna modellen har lika varians eller homoskedasticitet.

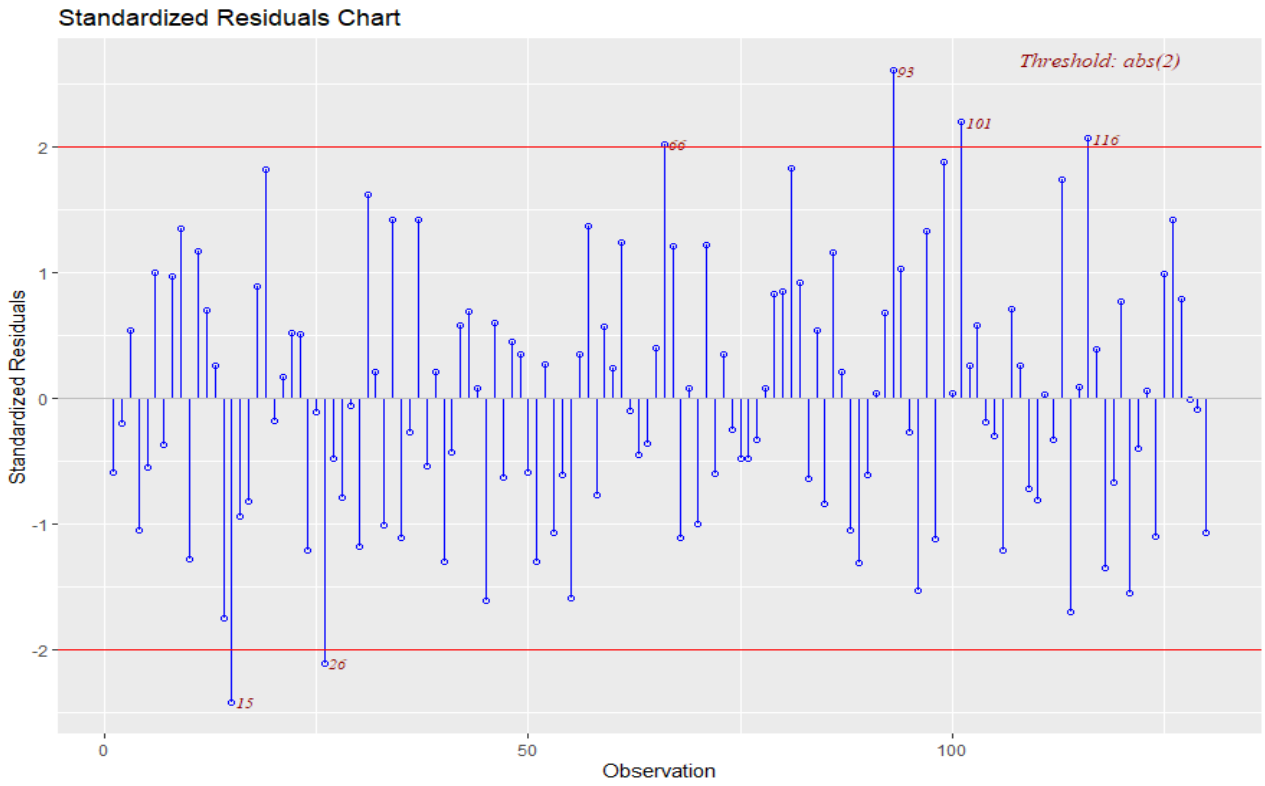
3. Oberoende



Figur 2.10: Residualer i observationsordning

Det verka visar inter någon tydligt systematiskt mönster i resudualerna, man kan säga att det är uppvisar slump. Det kan dras en slutsats att residualerna endast uppvisar oberoende och denna linjär regressions modell kan litas på.

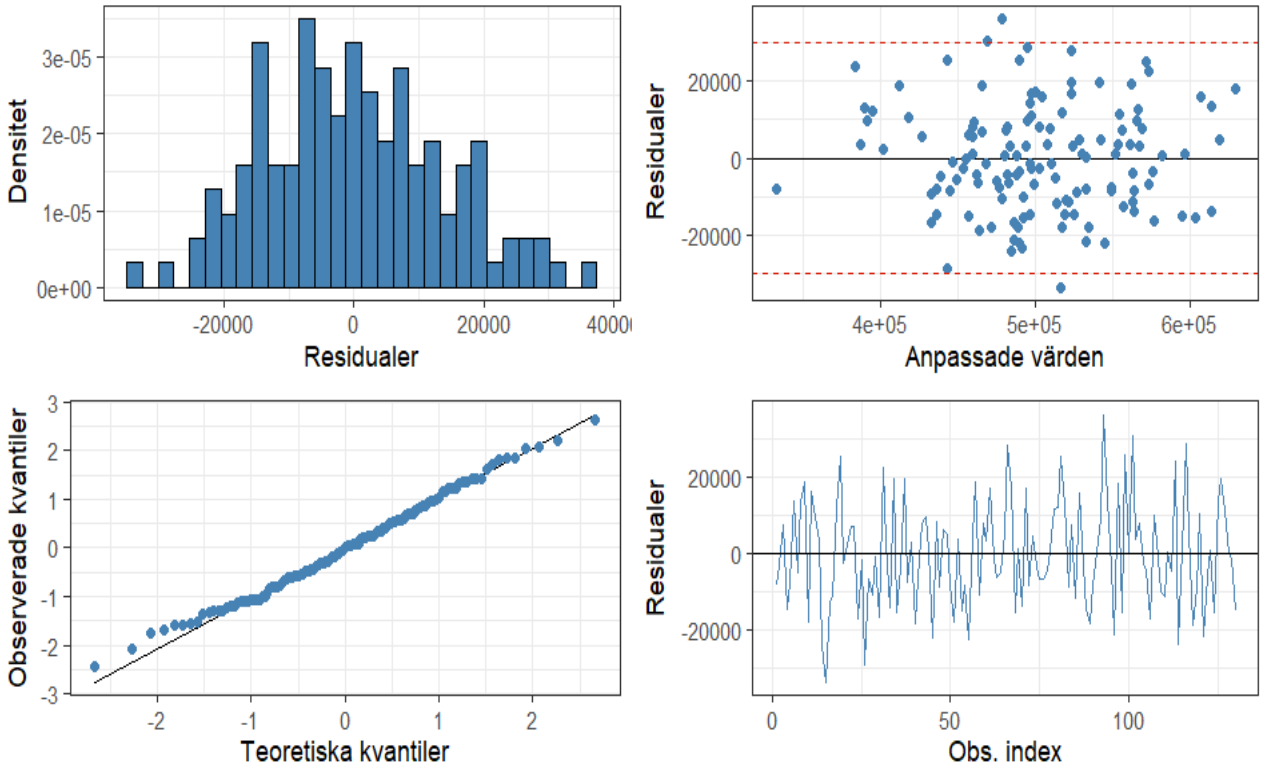
ii) Identifieras några extremvärden i ert data? Om ni gör det, identifiera vilken/vilka observationer som dessa är och använd visualiseringarna från a) för att identifiera för vilken förklarande eller responsvariabel observationen är ett extremvärde



Figur 2.11: Diagram över standardiserade residualer

I data finns det några värden som avviker lite ifrån de röda linjerna men dessa punkter ligger inte markant långt bort ifrån linjerna för att identifieras som extrempunkter. Med hjälp av residualanalys kan dessa observation identifieras, vilka är observationer nummer 15, 93.
Dock vid identifiering från visualiseringarna från a-uppgiften visar det sig att dessa observationer inte är extremvärdena. Detta bidrar till tidigare resonemanget att det inte finns några extremvärden som kan identifieras i residualdata.

Sammanfattning



Figur 2.12: Residualdiagrammen i en och samma bild

Sammanfattningvis figur 2.12 att residualerna uppfyller antagandet om normalfördelning med väntevärde 0, har homoskedasticitet och finns inte någon som kan förrändras systemastiskt med de anpassade värderna.

D) Utifrån tidigare resultat och beräknade utvärderingsmått, motivera ifall modellen är bra på att beskriva sambandet och om modellen är lämplig. Om ni inte anser modellen bra eller lämplig, beskriv (genomför inte) vilka förändringar ni hade gjort för att förbättra modellen

Förklaraingsgrader(R^2) :

$$R_a^2 = 1 - \frac{\frac{SSE}{n-(k+1)}}{\frac{\frac{SSR+SSE}{(n-1)}}{(n-1)}} = 1 - \frac{\frac{24867142549}{130-(4+1)}}{\frac{\frac{405121431106+24867142549}{(130-1)}}{(130-1)}} = 0.94031727194$$

Med 94% av den totala variationen som förklaras av modellens förklarande variabler.
Modellen är bra och lämplig på grund av att den uppfyller antaganden som slumpvariabel med ändlligt medelvärde och varians, oberoende observationer, linjärt samband för väntevärdet av Y|X, osv. Mer detaljerat uppfyller modellen feltermens antagande.

3. Lärdomar, problem, övriga kommentarer

Under detta kapitel lärt vi oss mycket om linjär modell, vi visste mer många begrepp, kunskap för hur bilda en linjär modell och hur kolla modellen är bra och lämpligt. Vi visste också mer många diagrammet som relevant om linjär modell, samatidigt tränna mer om Latex och hur skriver rapport med Latex i Overleaf. Förutom finns det också några problem när vi börja skriva rapporten, vi måste bekanta med Overleaf och är liten svårt i början. Och måste absorbera en ny mängd av kunskap samtidigt, så starten går något långsam.

Litteraturförteckning

1. hietalai.github.io
2. Koefficienttabell för Fit-regressionsmodell