Stockholm Universitet
Statistiska Institutionen
Maria Anna Di Lucca

Regressions- och tidsserieanalys

Tentamen

Måndagen den 9 januari, 2023

Godkända hjälpmedel

- Miniräknare och språklexikon.
- Formelsamling och statistiska tabeller delas ut vid tentamen.

Info

- Tentamen består av fem uppgifter, uppdelade i deluppgifter.
- För full poäng på en uppgift krävs tydliga, utförliga och välmotiverade lösningar.
- Resultatet meddelas senast den **2023-01-30**.
- Lösningsförslag till tentamensuppgifterna läggs ut på Athena strax efter tentamen.
- Använd minst fem värdesiffror i dina beräkningar (1,2345 och 1234,5 är exempel på sådana tal). I dina beräkningar från R-utskrifter får du utgå från det som är givet. Du kan dock avrunda ditt slutliga svar.

Gräns för godkänt

- 50 poäng av totalt 100.

OBS. Fx och F är underkända betyg som kräver omexamination. Studenter som får betyget Fx kan alltså inte komplettera för högre betyg.

Uppgift 1. (35 poäng)

Data från 16 offentliga elförsörjningsmyndigheter samlades in. En enkel linjär regressionsmodell används för att se om elproduktionen i miljoner kilowattimmar, Y, förklaras med energikostnaderna i US dollar, X, enligt:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

Nedan visas data

$$\sum x_i = 786.9990 \; \sum y_i = 78.9550 \; \sum x_i y_i = 5556.6630$$
$$\sum x_i^2 = 57416.8702 \; \sum y_i^2 = 552.3916$$

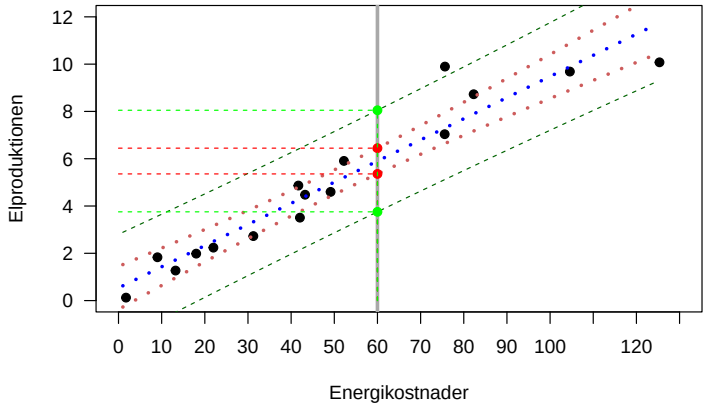


Figure 1: Samband mellan elproduktion och energikostnaderna

Vänligen lös och svara på följande deluppgifter (a-f):

a) Beräkna urvalskorrelationen och tolka resultat. (5p)

a) **Lösning**

För att beräkna urvalskorrelationen vet vi formeln:

$r_{xy} = \frac{s_{xy}}{s_x s_y}$ och då behöver vi urvalskovariansen och urvalsvarianser för x och y:

$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}$, s_x^2 och s_y^2

Vi använder summorna från tidigare för att beräkna urvalskovariansen

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{786.9990}{16} = 49.18744$ och

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{78.9550}{16} = 4.934687.$

$s_{xy} = \frac{5556.6630 - 16 * 49.18744 * 4.93469}{15} = \frac{1673.067}{15} = 111.5378$

För täljare behöver vi varianserna och sen roten ur för att få standardavvikelserna.

$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{57416.8702 - 16*49.18744^2}{15} = \frac{57416.8702 - 16*49.18744^2}{15} = 1247.093$

$s_y^2 = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1} = \frac{552.3916 - 16*4.93469^2}{15} = \frac{162.773}{15} = 10.85153.$

så $r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{111.5378}{\sqrt{10.85153*1247.093}} = \frac{111.5378}{116.3309} = 0.95879$

b) Testa om korrelationen är skild från noll på $\alpha = 0.05$. Utan beräkningar, kan man säga att lutningen är skild från noll? Varför? (5+3 p)

b) **Lösning**

Hypotesprövningen är huruvida om dessa två observerade variabler kommer ifrån respektive stokastiskt oberoende variabler

$H_0 : \rho_{xy} = 0$

$H_A : \rho_{xy} \neq 0$

Om nollhypotesen är sann då har vi för enkel regression $t \sim t_{n-2,\alpha/2}$.

Det statistiska testet är: $\frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$ och då är t från observationerna:

$\frac{0.95879\sqrt{16-2}}{\sqrt{1-0.95879^2}} = \frac{0.95879*3.741657}{\sqrt{0.08072174}} = \frac{3.58746}{0.2841157} = 12.62676$

Om man jämför t från observationerna med t-tabellen som är lika med 2.145 (t med 14 frihetsgrader och alpha delad med två på grund av tvåsidigt test) ser vi att vi kan förkasta nollhypotesen. Då vet vi att det finns ett samband mellan de två stokastiska variablerna, men vi vet inte om det är energikostnaderna som påverkar elproduktionen eller tvärtom därför att det är symmetriskt samband mellan x och y. Det är precis det som gäller skillnaden mellan lutningen och korrelationen. Lutningen visar om x påverkar y men det är säkert inte tvärtom och det kallas asymmetriskt samband.

Lutningen och urvalskorrelationen är numeriskt annorlunda men de är lika med formeln: $r_{xy} = b_1 \frac{s_x}{s_y}$ men också tvärtom: $b_1 = r_{xy} \frac{s_y}{s_x}$.

Om kvoten mellan standardavvikelserna är lika med 1, har vi två koefficienter som är lika. Dessutom har vi en kvot mellan standardavvikelserna som är hela tiden positiv, men meningen för urvalskorrelationen är annorlunda än skattade lutningen. När man testar lutningen, antar man att y förklaras med x men inte tvärtom. Istället, för urvalskorrelationen gäller hela tiden att x förklaras med hjälp av y men också tvärtom.

c) Beräkna och tolka förklaringsgraden R^2 i ord. (5p)

c) **Lösning**

För enkel regression kan man använda urvalskorrelationen i kvadrat för att beräkna förklaringsgraden och i detta fall har vi:

$R^2 = r_{xy}^2 = 0.95879^2 = 0.919278$ så 91 procent av elproduktionen förklaras med hjälp av electricitetskostnaderna. Det är ett väldigt högt förklarande resultat (SSR) som

finns väldigt litet kvar i SSE som att det inte behövs någon annan förklaringsvariabeln i den här modellen.

d) Skatta koefficienterna β_0 och β_1 med minsta-kvadrat-metoden. Tolka (endast) lutningen i ord. (5+3p)

d) **Lösning**

Man kan använda formeln som vi pratade om tidigare: $b_1 = r_{xy}\frac{s_y}{s_x}$ och då:

$b_1 = 0.95879\frac{3.294166}{35.3142} = 0.95879 * 0.09328163 = 0.08943749 \approx 0.08944$

$b_0 = \bar{y} - b_1\bar{x} = 4.934687 - 0.08944 * 49.18744 = 0.53537$

Lutningen är väldigt liten på grund av kvoten mellan standardavvikelser är nära till noll. Med 8 centsdollar ökar elproduktionen med 1 tusen kilowattimmar. Från Figur 1 ser man att det är ett positivt samband. Man skulle kunna förvänta sig högre värden för lutningen, till exempel 0.8 men det är på grund av olika enheter mellan x och y. Man ser att interceptet är väldigt nära till noll. Man skulle kunna testa för att se att intercepet är icke-signifikant men det är inget som efterfrågas på tentan.

e) Bestäm ett 95%-igt konfidensintervall för den genomsnittliga elproduktionen i kilowattimmar givet att energikostnaderna är lika med 60 (miljoner) dollar. (5p)

e) **Lösning**

Vi vill ha $\hat{\mu}_{Y|X=x} \pm t_{n-2,\alpha/2}\sqrt{s_e^2(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2})}$

Vi vill ha $\hat{\mu}_{Y|X=x} = b_0 + b_1 x = 0.53537 + 0.08944 * 60 = 5.90177$ och för att beäkna roten ur variansen behöver vi s_e^2. Vi vet från tidigare att $R^2 = 0.91928$ och då $SSE = SST(1 - R^2) = (n-1)s_y^2(1 - R^2) = (16-1)10.85153(1 - 0.91928) = 162.773 * 0.08072 = 13.13904$. Jag påminner om att $SST = (n-1)s_y^2$. Men vi syftar på $s_e^2 = MSE = \frac{SSE}{n-k-1} = \frac{13.13904}{14} = 0.9385$

Från tidigare har vi redan $t_{n-k-1,\alpha/2} = t_{14,0.025} = 2.145$

Så $\sqrt{s_e^2(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2})} = \sqrt{0.9385(\frac{1}{16} + \frac{(60-49.18744)^2}{(15)1247.093})} = \sqrt{0.9385(0.0625 + \frac{116.9115}{18706.4})} = \sqrt{0.9385(0.0625 + 0.0062498)} = \sqrt{0.9385(0.0687498)} = 0.2540112$

gränserna är då: $5.90177 \pm 2.145 * 0.2540112$ eller (5.356916; 6.446624)

Vi kan se dessa punkter i Figur 1.

f) Med hjälp av formelsamling och resultat i punkt e) beskriv Figur 1. Fundera sen över om du ser något i diagrammet som skulle kunna påverka en eventuell residualanalys, och på vilket sätt. (4p)

f) **Lösning**

Man kan se att konfidensintervall för den genomsnittliga prediktionen har inte med sig alla prickarna. Titta på figuren. Några observationer ligger nära övre gränsen

och en observation är utanför alla gränserna: den kommer att vara outlier och då kan påverka den residualanalysen. I texten kan man referera till formeln också och påpeka att i prediktionsvariansen har man bredare gränser: $(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2})$ runt residualsvariansen, s_e^2 men alla observationer kommer ändå inte att vara med och de kommer att påverka residualanalysen. Det efterfrågas inte på tentan, men man kan också testa med beräkningar även om det inte behövs:

$\sqrt{0.9385(1 + 0.0625 + 0.0062498)} = \sqrt{0.9385 * 1.06875} = \sqrt{1.003022} = 1.00151.$

gränserna är då: $5.90177 \pm 2.145 * 1.00151$ eller (3.753531; 8.049939). För att se skillnaden mellan olika resultat se de grönmarkerade prickarna för prognosen i Figur 1.

Uppgift 2. (30 poäng)

En linjär regressionsmodell med tre förklaringsvariabler, $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$, anpassades till ett datamaterial med $n = 18$ observationer. Den beroende variabeln är antal mord per 1 000 000 invånare per år i US (murders) och de förklarande variablerna är invånare (inhabitants, X_1), procent familjer med inkomst mindre än 5000 dollar per månad (incomes, X_2) och procentandelen arbetslösa (unemployed, X_3).

Modell 1 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$

```
## --------Modell 1--------------------

##
## Analysis of variance - ANOVA
## ------------------------------------------------
##       df      SS       MS      F     Pr(>F)
## Regr   2 1439.45 719.7254 78.394 1.1435e-08
## Error 15  137.71   9.1809
## Total 17 1577.16

## ------------------------------------------------
```

Modell 2 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2$

```
## --------Modell 2--------------------

##
## Analysis of variance - ANOVA
## ------------------------------------------------
##       df     SS       MS     F     Pr(>F)
## Regr   1 1338.1 1338.064 89.54 5.8911e-08
## Error 16  239.1   14.944
## Total 17 1577.2

## ------------------------------------------------
```

Modell 3 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_3 x_3$

```
## --------Modell 3--------------------

##
## Analysis of variance - ANOVA
## ---------------------------------------------
##       df      SS       MS      F     Pr(>F)
## Regr   1 1282.02 1282.017 69.498 3.2358e-07
## Error 16  295.15   18.447
## Total 17 1577.16

## ---------------------------------------------
```

Vänligen lös och svara på följande deluppgifter (a-f):

a) Baserat på ANOVA-tabellerna, beskriv vilken modell som är den bästa. Motivera ditt svar. (5p)

a) **Lösning**

Alla tre modeller är signifikanta om man tittar på p-värdet för F i ANOVA. Då är det viktigt att se vilken modell som lämnar mindre oförklarat. Den modellen som har de minsta residualerna är Modell 1. Det innebär att man vill ha två förklarande variabler för att ha en bättre skattning av den beroende variabeln. Modell 1 med inkomst och procentandelen av arbetslösa motiverar fler antal mord per 1 000 000 invånare per år i US.

b) Beräkna den justerade förklaringsgraden, R^2_{adj} från ANOVA-tabeller för att bekräfta att din valda modell i punkt a) stämmer överens med R^2_{adj}. (5p)

b) **Lösning**

För att bekräfta, numeriskt, att Modell 1 ger bättre resultat än de andra två enkla modellerna, beräknar vi den justerade förklaringsgraden, R^2_{adj}. För att beräkna det, använder vi förmeln:

$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$

Modell 1: $R^2_{adj} = 1 - \frac{137.71/15}{1577.16/17} = 1 - \frac{9.180667}{92.77412} = 1 - 0.0989572 = 0.9010428 \approx 0.90104$

Modell 2: $R^2_{adj} = 1 - \frac{239.1/16}{1577.16/17} = 1 - \frac{14.94375}{92.77412} = 1 - 0.1610767 = 0.8389233 \approx 0.83892$

Modell 3: $R^2_{adj} = 1 - \frac{295.15/16}{1577.16/17} = 1 - \frac{18.44687}{92.77412} = 1 - 0.1988364 = 0.8011636 \approx 0.80116$

Så har vi att det högsta värdet är när man väljer den första modellen.

c) Beskriv F-testet för din valda modell: förklara hypoteser, och vad det innebär att förkasta nollhypotesen. Genomför testet på signifikansnivån 5 procent. (5p)

c) **Lösning**

Vi refererar till den bästa modellen från punkt a) och b).

$H_0 : \beta_1 = \beta_2$

$H_A : \beta_j \neq 0$ åtminstone en förklaringsvariabel kan förklara och påverka den beroende variabeln.

Man kan direkt läsa p-värdet och ser att det är noll (1.1435e-08) och det är mindre än signifikansnivån, $0 < 0.05$. Vi kan förkasta nollhypotesen och modellen är i sin helhet signifikant. Man kan på ett annat sätt förklara med jämförelse mellan F från observationer och F från tabell. Man kan använda resultat: $F_{obs} = 78.394$ och som tumregel vet vi att det är större än 1 och då kan vi förkasta nollhypotesen. Vi kan också lägga till att den delen av regressionen (MSR) är betydligt större än residualdelen (MSE). Då kan vi säga att parametrar skiljer sig från noll och förklaringsvariabler förklarar den beroende variabeln. Om vi vill titta på F-tabellen kan vi säga att värdet är 3.68 baserade på frihetsgrader lika med 2 och 15.

d) I Tabell 2 finns delvis resultat av en regression mellan två förklarande variabler, beräkna VIF. Förklara om det är ett bra eller dåligt resultat och vilken av tidigare modeller: Modell 1, Modell 2 eller Modell 3 resultatet berör. (5p)

d) **Lösning**

Man kan beräkna VIF för multipla regressionsmodeller därför att för enkla modeller har man endast en förklarigsvariabel och då VIF är lika med 1. Det innebär att Tabell 2 refererar till Modell 1 som är den endast multipla modellen. Tabell 2 refereras till en regression mellan inkomst och procentandelen arbetslösa. Man kan beräkna VIF med följande formeln:

$VIF = \frac{1}{1-R^2} = \frac{1}{1-0.67588} = \frac{1}{0.32412} = 3.085277 \approx 3.0853$

Det är ett bra resultat därför att VIF är mindre än 10 och 5 då innebär att det inte finns problem med multikollinearitet.

Tabell 2 $\hat{x}_2 = \hat{\psi}_0 + \hat{\psi}_1 x_3$

```
## ----------------Tabell 2 -----------------------------------

##
## Measures of model fit
## ---------------------------------------------
## Root MSE       R2   R2-adj
##  1.95772  0.67588  0.65562
##
## Parameter estimates
## ---------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   4.7978    2.63602  1.8201 8.7506e-02
## unemployed    2.1691    0.37553  5.7762 2.8369e-05

## -------------------------------------------------------------
```

e) Fyll i följande R-utskrift för Modell 1. (5p)

```
## Parameter estimates
```

```
## ------------------------------------------------
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.7095 4.4823     -7.9667 0
## income      *        0.3869     4.141   9e-04
## unemployed  3.3926   *          3.3231  0.0046
```

e) **Lösning**

För att beräkna $\hat{\beta}_1$ använder vi $t_{obs} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$ och då: $\hat{\beta}_1 = t_{obs} * s_{\hat{\beta}_1} = 0.3869 * 4.141 = 1.602153$.

För att beräkna $s_{\hat{\beta}_2}$ kör vi: $s_{\hat{\beta}_2} = \frac{\hat{\beta}_2}{t_{obs}} = \frac{3.3926}{3.3231} = 1.020914$

f) Lista antaganden om feltermer. Beskriv sedan vilka antaganden man testar i Figur 2. (5p=3+2)

f) **Lösning**

Minnesregel "HEIL Gauss": H = feltermerna är ε_k är homoskedastiska dvs. variansen σ^2 är konstant; (E = existensantagandet, parametrarna i modellen är ändliga, inte oändliga, kanske lite för esoteriskt för den här kursen); I = feltermerna är sinsemellan oberoende (om X ska betraktas som en slumpvariabel så är de också oberoende av feltermerna); L = linjärt samband mellan x och y; Gauss = flertermerna är är normalfördelade. Ännu ett antagande för multipla modeller är att förklaringsvariablerna inte är linjärt bestämda av varandra dvs. ingen multikollinearitet föreligger.

I Figur 2 finns samband mellan residualer och regressorer. Man testar homogeneitet och linearitet. Homogeneitet innebär att feltermsvarians, σ^2_ε, är konstant. I diagrammet finns inget mönster eller grupper. I figuren ser man ingen kurvform som man kan relatera till kvadratiska termer och det innebär att det är ett linjärt samband mellan variablerna.

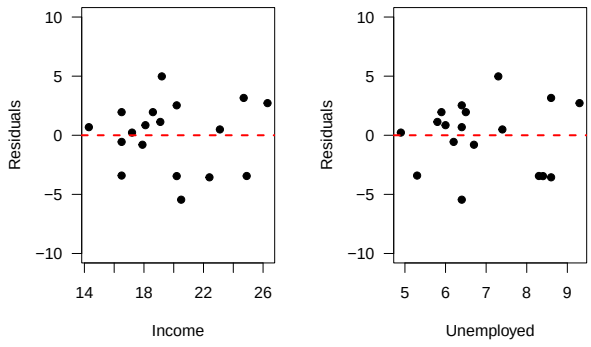


Figure 2: Regressorer och residualer

Uppgift 3. (15 poäng)

I nedanstående tabell visas kvartalsvisa observationer av produktionen (i miljoner kr) för ett företag från första kvartalet 2019 till det andra kvartalet 2021.

År	**Kvartal**	**Y**
2019	Kvartal 1	175
2019	Kvartal 2	75
2019	Kvartal 3	104
2019	Kvartal 4	56
2020	Kvartal 1	119
2020	Kvartal 2	9
2020	Kvartal 3	35
2020	Kvartal 4	26
2021	Kvartal 1	35
2021	Kvartal 2	5

Svara på följande frågor (a-c).

a) Skatta trenden för kvartalsdata med centrerade glidande medelvärden med säsongsvariation i en additiv eller multiplikativ modell. (5p)

Man kan välja mellan modeller. Om man tittar på Figur 3, kan man fundera mer att använda den multiplikativa modellen på grund av att det finns icke-ständiga val. Det finns för få observationer för att exkludera att det kan passa med den additiva modellen. Dock kan man använda en av dem. Vi kollar först med alla tre punkter med den additiva modellen och sen med den multiplikativa modellen

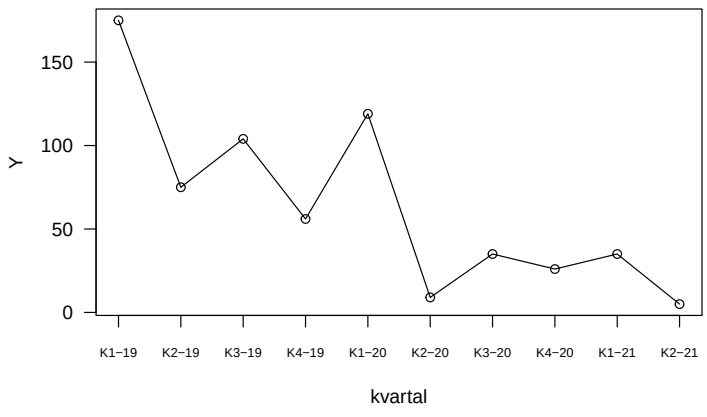


Figure 3: Kvartalsdata

a) **Lösning**

År	Kvartal	y	trend
2019	Kvartal 1	175	
2019	Kvartal 2	75	
2019	Kvartal 3	104	95.5
2019	Kvartal 4	56	80.25
2020	Kvartal 1	119	63.375
2020	Kvartal 2	9	51
2020	Kvartal 3	35	36.75
2020	Kvartal 4	26	25.75
2021	Kvartal 1	35	
2021	Kvartal 2	5	

År	Kvartal	y	trend	Y-T=S+E	s^+	Y-S=T+E
2019	Kvartal 1	175			54.375	120.625
2019	Kvartal 2	75			-43.250	118.250
2019	Kvartal 3	104	95.5	8.5	2.125	101.875
2019	Kvartal 4	56	80.25	-24.25	-13.250	69.250
2020	Kvartal 1	119	63.375	55.625	54.375	64.625
2020	Kvartal 2	9	51	-42	-43.250	52.250
2020	Kvartal 3	35	36.75	-1.75	2.125	32.875
2020	Kvartal 4	26	25.75	0.25	-13.250	39.250
2021	Kvartal 1	35			54.375	-19.375
2021	Kvartal 2	5			-43.250	48.250

Trenden om man väljer **den additiva modellen**:

Till exempel $T_3 = \frac{1}{8} * 175 + \frac{1}{4} * 75 + \frac{1}{4} * 104 + \frac{1}{4} * 56 + \frac{1}{8} * 119 = 95.5$

b) Skatta säsongkomponenterna för modellen i a). (5p)

b) **Lösning**

Om man väljer **den additiva modellen**, då följer det med **den additiva trenden** från punkt a). Till exempel

$S_1 = 55.625$, $S_2 = -42$, $S_3 = \frac{8.5-1.75}{2} = 3.375$, $S_4 = \frac{-24.25+0.25}{2} = -12$

$\bar{S} = \frac{55.625-42+3.375-12}{4} = 1.25$

$S_1^+ = S_1 - (\bar{S}) = 55.625 - (1.25) = 54.375$, $S_2^+ = -42 - 1.25 = -43.25$, $S_3^+ = 3.375 - 1.25 = 2.125$ och $S_4^+ = -12 - 1.25 = 13.25$.

Det gäller

$\sum S_j^+ = 54.375 - 43.25 + 2.125 - 13.25 \approx 0.$

c) Beräkna den skattade produktionen $\hat{y}_t = \hat{T}_t + \hat{S}_t$ eller $\hat{y}_t = \hat{T}_t \hat{S}_t$ med hjälp av resultat i punkt a) och b). (5p)

År	Kvartal	y	trend	Y-T=S+E	s^+	Y-S=T+E	yhat
2019	Kvartal 1	175			54.375	120.625	
2019	Kvartal 2	75			-43.250	118.250	
2019	Kvartal 3	104	95.5	8.5	2.125	101.875	97.625
2019	Kvartal 4	56	80.25	-24.25	-13.250	69.250	67
2020	Kvartal 1	119	63.375	55.625	54.375	64.625	117.75
2020	Kvartal 2	9	51	-42	-43.250	52.250	7.75
2020	Kvartal 3	35	36.75	-1.75	2.125	32.875	38.875
2020	Kvartal 4	26	25.75	0.25	-13.250	39.250	12.5
2021	Kvartal 1	35			54.375	-19.375	
2021	Kvartal 2	5			-43.250	48.250	

År	Kvartal	y	logy	trend	log(y)-log(T)
2019	Kvartal 1	175	2.24304		
2019	Kvartal 2	75	1.87506		
2019	Kvartal 3	104	2.01703	1.94989	0.06714
2019	Kvartal 4	56	1.74819	1.81386	-0.06567
2020	Kvartal 1	119	2.07555	1.63963	0.43592
2020	Kvartal 2	9	0.95424	1.53886	-0.58462
2020	Kvartal 3	35	1.54407	1.43077	0.1133
2020	Kvartal 4	26	1.41497	1.33243	0.08254
2021	Kvartal 1	35	1.54407		
2021	Kvartal 2	5	0.69897		

c) **Lösning**

Man kan beräkna som $\hat{Y}_t = \hat{T}_t + \hat{S}_t$ för **den additiva modellen**.

a) **Lösning**

Trenden om man väljer **den multiplikativa modellen** och följer exempel i airpassenger excelfilen, då behöver man logaritmera och sen beräkna:

Till exempel $T_3 = \frac{1}{8} * 2.24304 + \frac{1}{4} * 1.87506 + \frac{1}{4} * 2.01703 + \frac{1}{4} * 1.74819 + \frac{1}{8} * 2.07555 = 1.94989$

b) Skatta säsongkomponenterna för modellen i a). (5p)

b) **Lösning**

Om man väljer **den multiplikativa modellen**, då följer det med **den multiplikativa trenden** från punkt a). Till exempel

$S_1 = 0.43592$, $S_2 = -0.58462$, $S_3 = \frac{0.06714+0.1133}{2} = 0.09022$, $S_4 = \frac{-0.06567+0.08254}{2} = 0.008435$

$\bar{S} = \frac{0.43592-0.58462+0.09022+0.008435}{4} = -0.01251$

$S_1^+ = S_1 - (\bar{S}) = 0.43592 - (-0.01251) = 0.44843$, $S_2^+ = -0.58462 - (-0.01251) = -0.57211$, $S_3^+ = 0.09022 - (-0.01251) = 0.10273$ och $S_4^+ = 0.008435 - (-0.01251) =$

År	**log(y)**	**trend**	**log(Y)-log(T)**	s^+	10^S	**log(Y)-log(S)**
2019 K1	2.24304			0.4484312	2.8082208	1.79461
2019 K2	1.87506			-0.5721087	0.2678498	2.44717
2019 K3	2.01703	1.94989	0.06714	0.1027313	1.2668677	1.91430
2019 K4	1.74819	1.81386	-0.06567	0.0209462	1.0494125	1.72724
2020 K1	2.07555	1.63963	0.43592	0.4484312	2.8082208	1.62712
2020 K2	0.95424	1.53886	-0.58462	-0.5721087	0.2678498	1.52635
2020 K3	1.54407	1.43077	0.1133	0.1027313	1.2668677	1.44134
2020 K4	1.41497	1.33243	0.08254	0.0209462	1.0494125	1.39402
2021 K1	1.54407			0.4484312	2.8082208	1.09564
2022 K2	0.69897			-0.5721087	0.2678498	1.27108

År	**log(y)**	**trend**	$10^{(log(Y)-log(S))}$	10^T
2019 K1	2.24304		62.31750	
2019 K2	1.87506		280.00772	
2019 K3	2.01703	1.94989	82.09184	89.10252
2019 K4	1.74819	1.81386	53.36297	65.14184
2020 K1	2.07555	1.63963	42.37600	43.61441
2020 K2	0.95424	1.53886	33.60083	34.58279
2020 K3	1.54407	1.43077	27.62740	26.96311
2020 K4	1.41497	1.33243	24.77536	21.49958
2021 K1	1.54407		12.46350	
2022 K2	0.69897		18.66724	

0.020945.

Det gäller

$\sum S_j^+ = 0.448426427 - 0.572105528 + 0.102728888 + 0.020950212 \approx 0.$

10^{s^+} betyder att man beräknar $10^{S_1^+} = 10^{0.44843} = 2.80819$, $10^{S_2^+} = 10^{-0.57211} = 0.267852$, $10^{S_3^+} = 10^{0.10273} = 1.266861$ och $10^{S_4^+} = 0.020945 = 1.049422$

c) Beräkna den skattade produktionen $\hat{y}_t = \hat{T}_t + \hat{S}_t$ eller $\hat{y}_t = \hat{T}_t\hat{S}_t$ med hjälp av resultat i punkt a) och b). (5p)

c) **Lösning**

Man kan beräkna som $\hat{Y}_t = \hat{T}_t * \hat{S}_t$ för **den multiplikativa modellen**.

Observera att man kan beräkna säsongkomponenterna i den multiplikativa modellen med kvoten mellan data och trenden, i formeln: $S_t = \frac{y_t}{\hat{T}_t}$ exempelvis med $S_3 = \frac{104}{95.5} = 1.089005$ som övning 5.6 och 5.7 i boken.

För att beräkna grova säsongskomponenter S_j^+ använder vi följande:

$\bar{S}_1 = 1.87771$, $\bar{S}_2 = 0.17647$ $\bar{S}_3 = \frac{1.08901+0.95238}{2} = 1.020695$ och

År	Y	log(y)	trend	s^+	yhat
2019 K1	175	2.24304		0.4484312	
2019 K2	75	1.87506		-0.5721087	
2019 K3	104	2.01703	1.94989	0.1027313	112.8811
2019 K4	56	1.74819	1.81386	0.0209462	68.36066
2020 K1	119	2.07555	1.63963	0.4484312	122.47889
2020 K2	9	0.95424	1.53886	-0.5721087	9.26299
2020 K3	35	1.54407	1.43077	0.1027313	34.15869
2020 K4	26	1.41497	1.33243	0.0209462	22.56193
2021 K1	35	1.54407		0.4484312	
2022 K2	5	0.69897		-0.5721087	

År	Kvartal	y	trend	s	s^+	yhat
2019	Kvartal 1	175			191.18168	91.536
2019	Kvartal 2	75			17.96754	417.419
2019	Kvartal 3	104	95.5	1.08901	103.92350	100.074
2019	Kvartal 4	56	80.25	0.69782	86.92728	64.422
2020	Kvartal 1	119	63.375	1.87771	191.18168	62.244
2020	Kvartal 2	9	51	0.17647	17.96754	50.090
2020	Kvartal 3	35	36.75	0.95238	103.92350	33.679
2020	Kvartal 4	26	25.75	1.00971	86.92728	29.910
2021	Kvartal 1	35			191.18168	18.307
2021	Kvartal 2	5			17.96754	27.828

$\bar{S}_4 = \frac{0.69782+1.00971}{2} = 0.853765$

sen behöver vi: $\sum S_j = 1.87771 + 0.17647 + 1.020695 + 0.853765 = 3.92864$.

Nu kan vi använda formeln på 201: $\frac{\bar{S}_i}{\sum \bar{S}_i} * 400$ och vi har:

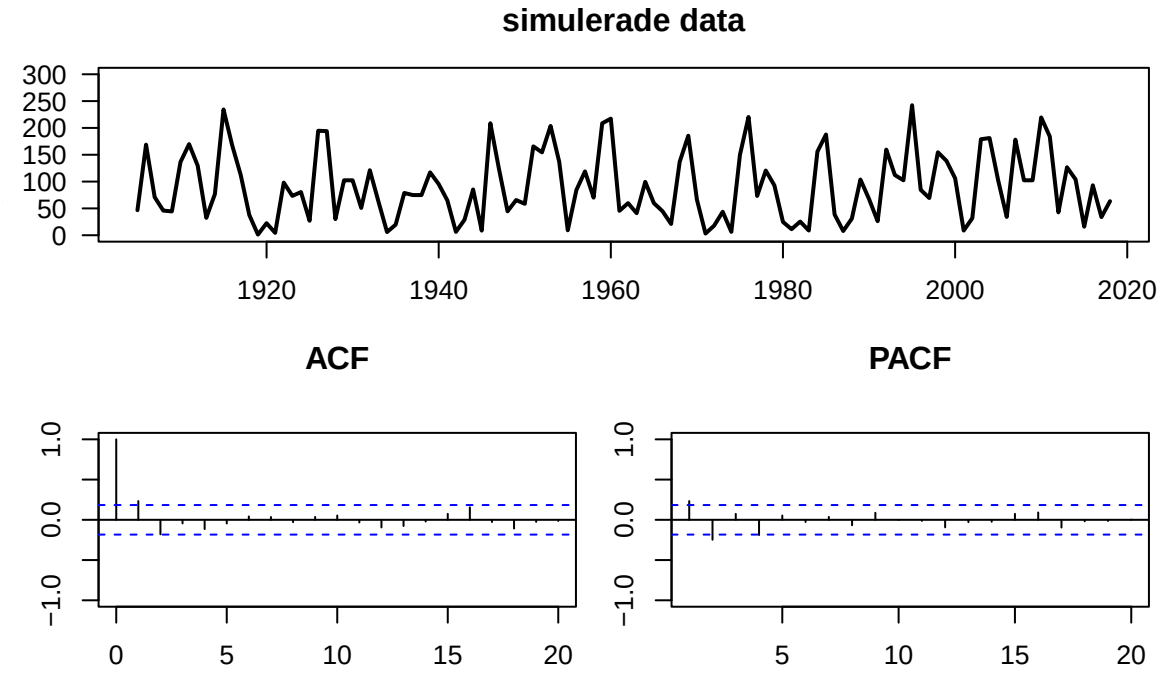
$\bar{S}_1^+ = \frac{1.87771}{3.92864} * 400 = 191.1817$, $\bar{S}_2^+ = \frac{0.17647}{3.92864} * 400 = 17.96754$

$\bar{S}_3^+ = \frac{1.020695}{3.92864} * 400 = 103.9235$, $\bar{S}_4^+ = \frac{0.853765}{3.92864} * 400 = 86.92728$

För att beräkna i punkt c) den skattade produktionen, kan man då beräkna som $\frac{\hat{y}_t}{S_t^+} * 100$, till exempel: $\frac{104}{103.92350} * 100 = 100.074$.

Uppgift 4. (10 poäng)

I den här uppgiften analyseras förändringar i simulerade antal lodjur fångade i Canada under de 114 åren 1905-2018. I Figuren 4.a nedan visas den ursprungliga tidserien, autokorrelationen och partiella autokorrelationen för datamaterialet.



Figur 4.a

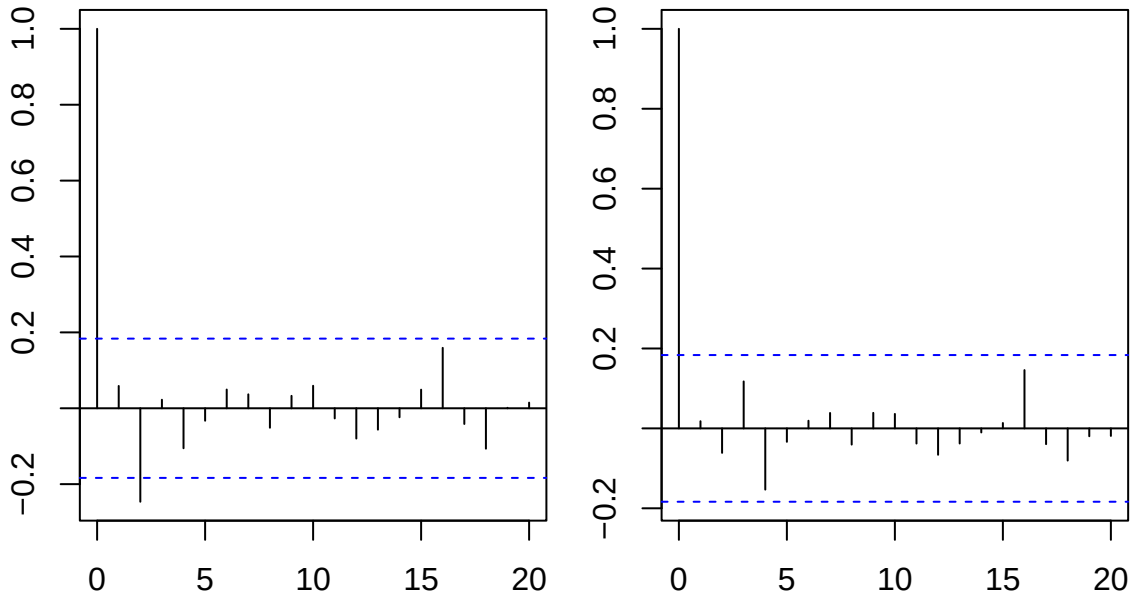
```
## ---------------------------AR(1)-------------------------------

##
## Parameter estimates
## ------------------------------------------------------------
##      Estimate Std. Error z-ratio Pr(>|z|)     2.5 %    97.5 %
## ar1   0.23246   0.090867  2.5582 0.010521  0.054359   0.41056
## mean 90.32334   7.486709 12.0645 0.000000 75.649393 104.99729

## ---------------------------AR(2) -------------------------------

##
## Parameter estimates
## ------------------------------------------------------------
##      Estimate Std. Error z-ratio  Pr(>|z|)    2.5 %     97.5 %
## ar1   0.29320   0.090751  3.2308 0.0012344  0.11533   0.471069
## ar2  -0.25241   0.091092 -2.7709 0.0055894 -0.43095  -0.073871
## mean 90.43414   5.822323 15.5323 0.0000000 79.02239 101.845893
```

ACF för residualer AR(1) **ACF för residualer AR(2)**



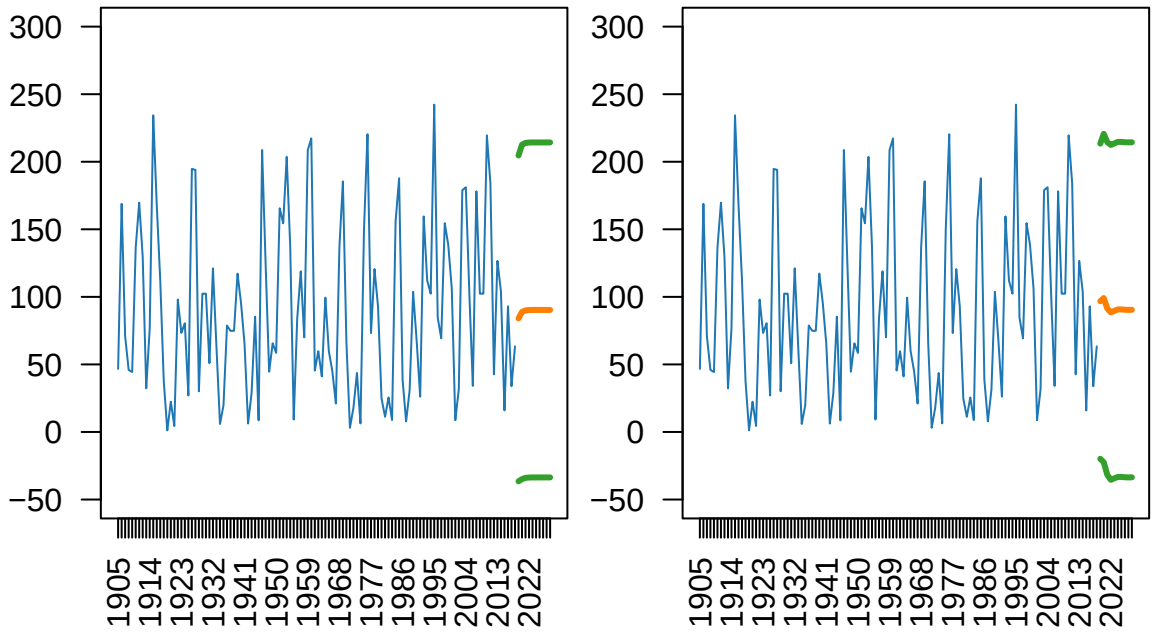
Figur 4.b

Lös följande deluppgifter (a-b):

a) Med hjälp av R-utskrifter och residualdiagram i Figur 4.b motivera om det är bättre med en AR(1) modell eller med en AR(2) modell. (5p)

a) **Lösning**

Om man tittar på skattade modellerna kan man använda båda två modeller, men om man observerar ACF för AR(1) ser man att det finns en residual som föreligger gränserna. Det är därför det är bättre med AR(2).



Figur 4.c

```
## -------------Tabell 4--------------

## Time Series:
## Start = 2019
## End = 2028
## Frequency = 1
##         AR(1)   AR(2)
## 2019 84.0615 96.7869
## 2020 88.8677 99.1241
## 2021  89.985 91.3785
## 2022       * 88.5176
## 2023 90.3051 89.6338
## 2024 90.3191 90.6833
```

```
## 2025 90.3224 90.7092
## 2026 90.3231      *
## 2027 90.3233 90.3699
## 2028 90.3233 90.4108
```

b) I nedanstående Figur 4c och Tabell 4 visas prognosticerade fångade lodjur för de tio åren 2019-2028 med AR(1) och respektive AR(2). I tabellen har ett par tal ersatts med "*". Glöm inte att i R-utskrifterna finns medelvärdet, inte interceptet. I modell AR(1) är interceptet lika med 69.32678 och i AR(2) är lika med 86.74533. (5p)

b) **Lösning**

För att beräkna prognos i AR(1), använder vi:

$\hat{y}_{2022} = b_0 + b_1 y_{2021}$

och då har vi $69.32678 + 0.23246 * 89.985 = 90.24469$

För att beräkna prognos i AR(2), kör vi:

$\hat{y}_{2026} = b_0 + b_1 y_{2025} + b_2 y_{2024}$

och i detta fall har vi: $86.74533 + 0.29320 * 90.7092 + (-0.25241)90.6833 = 90.4519$

Uppgift 5. (10 poäng)

Ett dataset i R innehåller information om förekomsten av diabetes hos n=392 patienter. En logistisk regression har använts för att modellera sannolikheten att en patient har diabetes, $Y = 1$ som en funktion av tre förklarande variabler: glukosnivån (glucose), X_1, Body Mass Index (BMI), X_2 och ålder (Age), X_3. BMI presenteras som en binär variabel. Det beskriver om en person är överviktig eller inte:

$x_2 = 1$, $BMI > 30$, patienten är överviktig

$x_2 = 0$, $BMI \leq 30$, patienten är inte överviktig

Resultatet från antal patienter med diabetes är i Tabell 5a och resultaten från en anpassad logistisk regression med diabetes som responsvariabel visas i Tabell 5b.

Vänligen lös och svara på följande deluppgifter (a-c):

a) Beräkna sannolikheten för en patient med diabetes givet att patienten har glukosen lika med 120, BMI är större än 30 och är 30 år gammal. (3p)

a) **Lösning**

$P(y|x_1 = 120, x_2 = 1, x_3 = 30) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = -7.805059 + (0.036475 * 120) + (1.174080 * 1) + (0.050789 * 30) = -0.73$

Det innebär att man är yngre och med låg glukosnivå, detta minskar sannolikheten att den personen har diabetes. Låta säga att istället än en person är 30 år gammal, då ser man att en person har samma glukosnivå och i övervikt men en person är 60 år gammal då ser man att sannolikheten att man har diabetes är högre, nästan 0.80.

$P(y|x_1 = 120, x_2 = 1, x_3 = 30) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = -7.805059 + (0.036475 * 120) + (1.174080 * 1) + (0.050789 * 60) = 0.79$

b) Tolka effekten av den förklarande variabeln BMI som en oddskvot (OR). (3p)

b) **Lösning**

Oddskvoten för BMI är lika med 3.23516654. Det innebär att sannolikheten för en person som är i övervikt är cirka tre gånger högre än för en person som är i normal vikt.

```
## ----------------------Tabell 5 a-----------------------
```

	neg	pos
Antal	262.0000	130.0000
Andel	0.6684	0.3316

```
## ----------------------Tabell 5 b-----------------------

##
## Parameter estimates
```

```
## ------------------------------------------------
##                 Estimate Std. Error z value   Pr(>|z|)
## (Intercept)   -7.805059  0.7832872 -9.9645 2.1799e-23
## glucose        0.036475  0.0049279  7.4019 1.3429e-13
## BMIöverviktig  1.174080  0.3161596  3.7136 2.0436e-04
## age            0.050789  0.0130973  3.8778 1.0539e-04
##
## Odds ratio estimates
## ------------------------------------------------
##                  Estimate Std. Error z value   Pr(>|z|)
## (Intercept)   0.00040767     2.1887 -9.9645 2.1799e-23
## glucose       1.03714871     1.0049  7.4019 1.3429e-13
## BMIöverviktig 3.23516654     1.3718  3.7136 2.0436e-04
## age           1.05210114     1.0132  3.8778 1.0539e-04
```

c) Figur 5 visar datamaterialet med ålder på x-axeln och den binära responsvariabeln om patienterna med diabetes, Y = 1, på y-axeln. Figuren visar också den anpassade logistiska regressionsmodellen

$$P(Y=1|x_1,x_2,x_3)=\frac{e^{\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_3}}{1+e^{\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_3}}$$

som en streckad linje och en heldragen linje. Förklara skillnaden mellan dessa två kurvor. (4p)

c) **Lösning**

Titta på Figur 5. Båda två kurvor representerar en skattning av den här modellen. Skillnaden är på grund av att BMI är en kvalitativ variabel som vi kodade om som noll och ett och det påverkar linjer. Dessutom, om BMI är lika med 1 innebär detta att man är i övervikt. Det innebär att man har en skattade koefficient till som kommer att vara i kurvan och det är högre i diagrammet och det är den rödafärgade kurvan. Istället, när BMI är noll försvinner den skattade koefficienten och det är därför i kombinationen med alla andra koefficienter i kurvan är under och den är grönfärgad.

$$P(Y=1|x_1,x_2=1,x_3)=\frac{e^{\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2*1+\hat{\beta}_3x_3}}{1+e^{\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2*1+\hat{\beta}_3x_3}}$$

Istället för den andra linjen har vi $x_2 = 0$ och det innebär att en koefficient försvinner och då blir numeriskt mindre och då den andra kurvan.

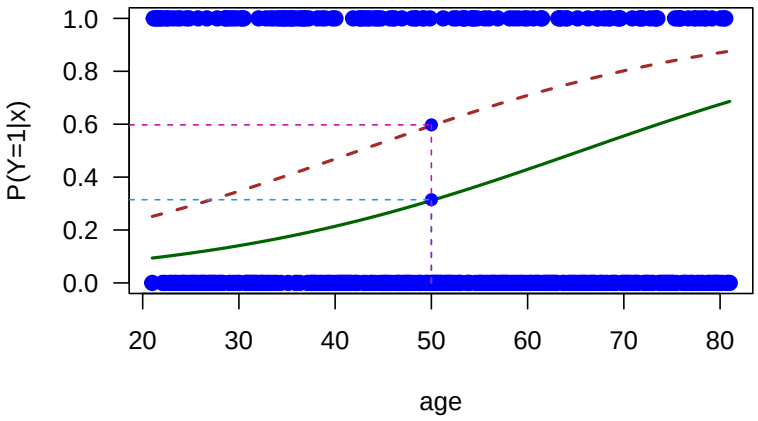
$$P(Y=1|x_1,x_2=0,x_3)=\frac{e^{\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2*0+\hat{\beta}_3x_3}}{1+e^{\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2*0+\hat{\beta}_3x_3}}$$

Till exempel, kan man beräkna för en person som är 50 år gammal och med glukos lika med 123. Glukosnivån är medelvärdet för glukosen i datasetet.Då ser vi resultat som

$$P(Y=1|x_1=123,x_2=1,x_3=50)=\frac{e^{\hat{\beta}_0+\hat{\beta}_1*123+\hat{\beta}_2*1+\hat{\beta}_350}}{1+e^{\hat{\beta}_0+\hat{\beta}_1*123+\hat{\beta}_2*1+\hat{\beta}_3*50}}=$$
$$\frac{e^{(-7.805059+(0.036475*123)+(1.174080*1)+(0.050789*50))}}{1+e^{(-7.805059+(0.036475*123)+(1.174080*1)+(0.050789*50))}}=$$
$$\frac{e^{0.394896}}{1+e^{0.394896}}=\frac{1.48423}{1+1.48423}=0.5974$$

$$P(Y=1|x_1=123,x_2=0,x_3=50)=\frac{e^{\hat{\beta}_0+\hat{\beta}_1*123+\hat{\beta}_3*50}}{1+e^{\hat{\beta}_0+\hat{\beta}_1*123+\hat{\beta}_3*50}}=$$
$$\frac{e^{-7.805059+(0.036475*123)+(1.174080*0)+(0.050789*50)}}{1+e^{-7.805059+(0.036475*123)+(1.174080*0)+(0.050789*50)}}=\frac{e^{-0.779184}}{1+e^{-0.779184}}=$$
$$\frac{0.4587802}{1+0.4587802}=0.3144958$$

I diagrammet har vi den streckade linjen för överviktiga patienter och sen den heldragna linjen för normalviktiga patienter. Mer i detalj, om en patient är 50 år gammal och med konstant glukosnivå, är sannolikheten att den patienten har diabetes 0.70 om den är överviktig. Om en patient är 50 år gammal och med konstant glukosnivå, är sannolikheten att den patienten har diabetes 0.31, om den är normalviktig. Så att vara i övervikt nästan dubblar sannolikheten att ha diabetes.



Figur 5