Labbrapport i Statistik

Laboration 2

732g43

Jakob Lindén

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet
22-11-2019

Innehåll

1. Uppgift 1

Uppgiften behandlar data om lägenhetspriset i miljontalskronor vilket är responsvariabeln y. Följande förklaringsvariabler finns med i datamaterialet:

- x_1 = area i kvadratmeter på lägenheten
- x_2 = antal rum i lägenheten
- x_3 = bostadsavgift per månad i tusentals kronor för lägenheten
- x_4 = antal trappor till lägenheten i bostadshuset
- x_5 = dummyvariabel som är lika med 1 om lägenheten såldes i region City
- x_6 = dummyvariabel som är lika med 1 om lägenheten såldes i Syd.

I följande uppgift antags följande Bayesianska linjära regressionsmodell för resposnvariabeln y utan förklaringsvariabler.

$$\begin{aligned} y_i|\mu_i, \sigma &\overset{iid}{\sim} N(\mu, \sigma) \\ \mu &\sim N(3, 10) \\ \ln\sigma &\sim N(0, 1) \end{aligned} \tag{1.1}$$

Med hjälp av kvadratisk approximation ska i uppgiften en Bayesiansk linjär regressionsmodell anpassas med den beroende variabeln y, alltså pris på lägenheter i miljoner. Inga förklarande variabler ska användas i modellen.

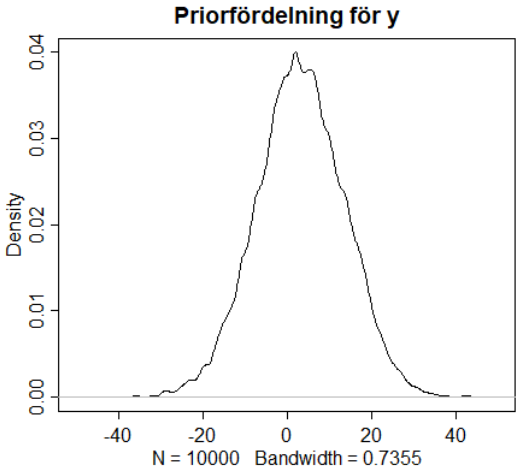
1.1 a)

```
load("C:/Users/Jakob/Desktop/Bayesian statistics/lab2/Data_Moment2.RData")

X <- as.data.frame(X)

set.seed(970922);y <- rnorm(n,A,B)

sampleMu <- rnorm(1e4,3,10)
sampleLogSigma <- rnorm(1e4,0,1)

sample_y <- rnorm(1e4,sampleMu,exp(sampleLogSigma))
dens(sample_y, main = "Priorfördelning för y", xlim = c(-50,50))
```

Priorfördelningen för y ska plottas och slutsats ska dras om fördelningen verkar rimlig utifrån hur y är definerad.

I figur 1.1 presenteras priorfördelningen för y. Fördelningen har ett medelvärde på ungefär 3, alltså så förväntas priset för en lägenhet ligga runt 3 miljoner dock med en väldigt stor osäkerhet. Detta är rimligt eftersom y är definerad med parametern μ som är normalfördelad med medelvärde 3 och standardavvikelse 10, samt parametern $ln\sigma$ som är normalfördelad med medelvärde 0 och standardavvikelse 1. Således är

mycket spridning att förvänta från hur y är definerad, detta blir alltså en mycket icke-informativ prior eftersom spridningen är så stor tänker vi oss att priset på en lägenhet kan ligga mellan -40 miljoner och 40 miljoner. Det kan tänkas konstigt att priorn kan anta negativa värden på en lägenhet, detta fungerar dock på ett bra sätt när vi vill ha en normalfördelad prior med denna fördelning.



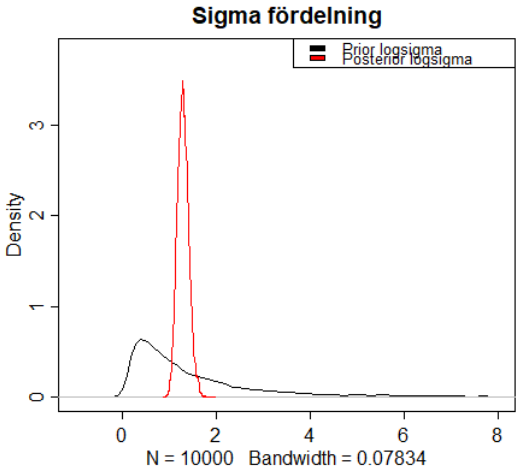
Figur 1.1

1.2 b)

```
flist <- alist(
  y/1000 ~ dnorm(mu, exp(logsigma)) ,
  mu ~ dnorm( 3 , 10 ) ,
  logsigma ~ dnorm ( 0 , 1 )
)

resNormal_logsigma <- map(flist, data=X)

precis(resNormal_logsigma)

postSamples_logsigma <- extract.samples(resNormal_logsigma , n=1e4) # posterior samples

dens(exp(sampleLogSigma), xlim = c(-1,8),ylim = c(0,3.8), main = "Sigma fördelning")
dens(exp(postSamples_logsigma[,2]),col="red",add=TRUE)
legend("topright", legend = c("Prior logsigma", "Posterior logsigma"), fill=c("black", "red"),
    cex = 0.8)

dens(sampleMu, ylim = c(0,2.5), xlim = c(-20,25), main = "Mu fördelning")
dens(postSamples_logsigma[,1],type="l",col="red",add=TRUE)
legend("topright", legend = c("Prior mu", "Posterior mu"), fill=c("black", "red"), cex = 0.8)
```

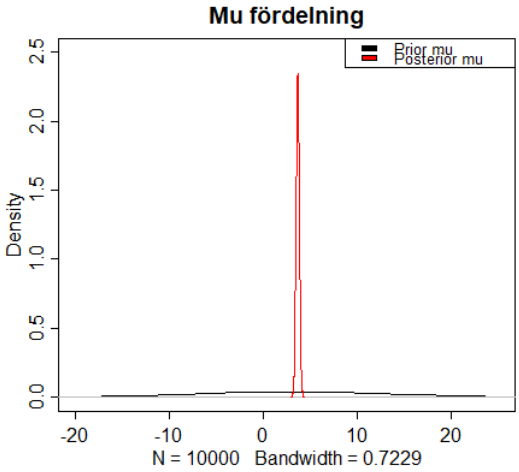
Genom att använda kvadratisk approximation ska en Bayesiansk linjär regerssionmodell med den beroende variabeln y anpassas utan förklaringsvariabler. Priorfördelningarna för μ och σ ska jämföras med deras respektive posteriorfördelningar.

I figur 1.2 redovisas posteriorfördelningen för σ mot priorfördelningen för σ. Det blir tydligt hur icke-informativ priorfördelningen för σ är här då den svarta kurvan har väldigt mycket mer spridning än den röda. Posteriorfördelningen ses ha ett mycket snävare fördelning alltså mindre spridning, detta är att förvänta eftersom vi kombinerat informationen från data med våran apriori kunskap i priorn.



Figur 1.2

I figur 1.3 redovisas posteriorfördelningen för μ mot priorfördelning för μ. Skillnaden i posterior och prior här blir så stor att det är svårt att visualisera. Detta eftersom spridningen i priorn är väldigt stor, medans spridningen i posteriorn är väldigt liten. I priorn var som tidigare redovisat medelvärdet alltså värdet med högst sannolikhet i fördelning ungefär 3 miljoner medans i posteriorn är medelvärdet ungefär 3.6 miljoner. Så det förväntade priset på en lägenhet i posteriorn och priorn är relativt nära varandra, men det finns betydligt mycket mer osäkerhet i priorn eftersom den är så icke-informativ.



Figur 1.3

1.3 c)

```
precis(resNormal_logsigma,prob=0.909)

postSamples_logsigma$sigma <- exp(postSamples_logsigma$logsigma)

Kovar <-cov(postSamples_logsigma[,c(1,3)])
cov2cor(Kovar)
```

En tabell med medelvärde, standardavvikelse och 90.9 procents kredibilitetsintervall för μ och σ ska redovisas och tolkas. Även korrelationsvärdet för μ och σ ska redovisas och tolkas.

Medelvärde och standardavvikelse samt ett 90.9 procents kredibilitets intervall för μ och σ redovisas i tabell 1.1. Medelvärdet för μ är 3.62 miljoner med en standardavvikelse på 0.17 miljoner. Kredibilitetsintervallet redovisar att med 90.9 procents säkerhet ligger μ mellan 3.33 och 3.9 miljoner. Medelvärdet för σ är 1.27 miljoner med en standardavvikelse på 1.09 miljoner. Kredibilitetsintervallet redovisar att med 90.9 procents säkerhet ligger σ mellan 1.09 och 1.49 miljoner.

Tabell 1.1

	Mean	**StdDev**	**4.55%**	**95.45%**
Mu	3.62	0.17	3.33	3.9
Sigma	1.27	1.09	1.09	1.49

I tabell 1.2 ses en korrelationsmatris mellan variablerna μ och σ. Vi kan se att korrelationen mellan μ och σ är 0.0066 vilket indikerar på väldigt låg korrelation mellan variablerna. Således när vi lär oss om μ så lär vi oss inget om σ och vice versa. Detta är ett rimligt resultat eftersom det är vanligt med låg korrelation mellan dessa variabler när man använder en Gaussisk modell av detta slag.

Tabell 1.2

	Mu	**Sigma**
Mu	1	0.0066
Sigma	0.0066	1

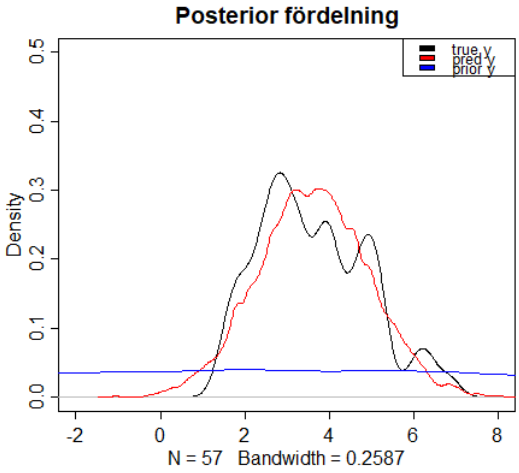
1.4 d)

```
yPred_1d <- rnorm(1e4,postSamples_logsigma[,1],exp(postSamples_logsigma[,2]))
dens(y/1000, xlim = c(-2,8), ylim = c(0,0.5), main="Posterior fördelning")
dens(yPred_1d,type="l",col="red",add=TRUE)
dens(sample_y, col="blue",add=TRUE)
legend("topright", legend = c("true y", "pred y", "prior y"), fill=c("black", "red","blue"),
    cex = 0.8)
```

En modellutvärdering ska göras genom att replikera data från posterior prediktiva fördelningen $p(\tilde{y}|y)$. Sedan ska priorfördelningen för y, posteriorprediktiva fördelningen $p(\tilde{y}|y)$ och fördelningen för de faktiskta värden på y plottas i samma graf. Slutsatser ska dras om skillnaderna/likheterna mellan dessa 3 fördelningar.

I figur 1.4 redovisas de 3 fördelningarna. Det syns tydligt att den posteriorprediktiva fördelningen $p(\tilde{y}|y)$ och fördelningen för de sanna värdena på y följer ungefär samma fördelning, vilket är att förvänta eftersom den posteriorprediktiva fördelningen är framtagen genom att sampla från modellen som kombinerat den eliciterade priorn med data. I figuren ses även priorfördelningen som har väldigt annorlunda utseende, detta pågrund av den prior som valts till modellen var mycket icke-informativ och således hade hög varians. Alla tre fördelningar har ett medelvärde runt 3 miljoner men där det tydligt går att se att priorn har en mycket högre variation.

Med hjälp av figur 1.4 kan vi dra slutsatsen att denna modell för data y verkar rimlig eftersom den posteriorprediktiva fördelningen till stordel följer fördelningen av y.



Figur 1.4

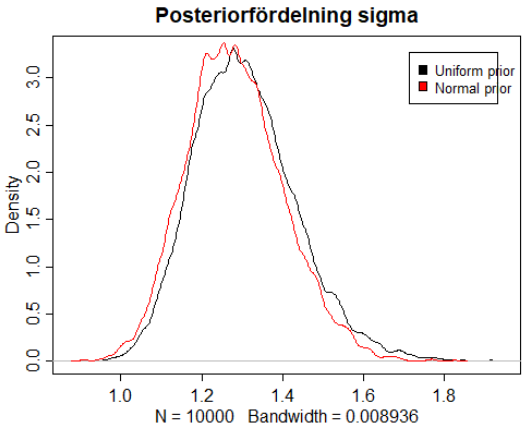
1.5 e)

```
f_sigma <- rchisq(1e4,nrow(X) - 1)

sigma2 <- ( (nrow(X) - 1) * var(y/1000) ) / f_sigma

mu_sigma2 <- rnorm(n = 1e4, mean(y/1000), sqrt(sigma2/nrow(X)))

dens(sqrt(sigma2), main="Posteriorfördelning sigma")
dens(exp(postSamples_logsigma[,2]),col="red",add=TRUE)
legend("topright", legend = c("Uniform prior", "Normal prior"), fill=c("black", "red","blue"),
    cex = 0.8)

dens(mu_sigma2, main="Betingad posterior mu")
dens(postSamples_logsigma[,1],type="l",col="red",add=TRUE)
legend("topright", legend = c("Uniform prior", "Normal prior"), fill=c("black", "red","blue"),
    cex = 0.8)
```

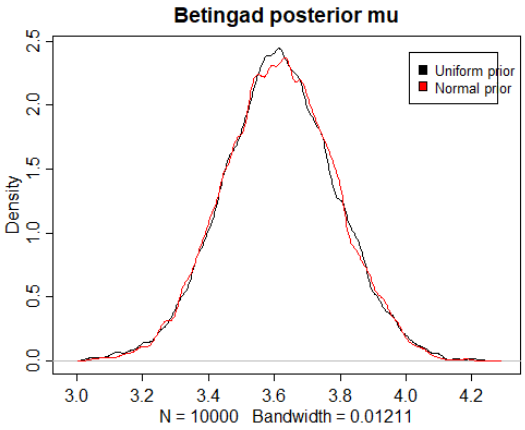
I denna uppgift ska en uniform prior för μ och $ln\sigma$ användas, alltså $p\left(\mu,\sigma^2\right) \propto \frac{1}{\sigma^2}$. Denna prior ger dom kända posteriorfördelningarna $p\left(\mu|\sigma^2,y_1,\ldots,y_n\right)$ och $p\left(\sigma^2|y_1,\ldots,y_n\right)$. Dessa posteriorfördelningar ska plottas mot posteriorfördelningarna från uppgift (b) och slutsatser ska dras om skillnader/likheter mellan dom.

I figur 1.5 ses posteriorfördelningen för σ följa ungefär samma fördelning med uniform prior och normalprior, detta är pågrund av att den valda normal priorn i uppgift (b) var så icke-informativ att det ger ungefär samma resultat som den helt icke-informativa uniforma priorn. I figur 1.6 ses samma resultat där den betingade posteriorn med uniform och normalprior har väldigt lik fördelning, som tidigare pågrund av att den valda normalpriorn var ungefär lika icke-informativ som en uniform prior.

Fördelen med att använda en sådanhär standard icke-informativ prior är att den betingade posteriorn $p\left(\mu,\sigma^2\right)$ och $p\left(\sigma^2|y_1,\ldots,y_n\right)$ följer kända fördelningar. Denna prior ger även acceptabla resultat om man har mycket data tillhands jämfört med antalet förklaringsvariabler. Om man dock har lite data eller många förklaringsvariabler så bör en rimligare prior specificeras.



Figur 1.5



Figur 1.6

2. Uppgift 2

I följande uppgift ska med hjälp av kvadratisk approxmation en Bayesiansk linjär regressionsmodell med den beroende variabeln y anpassas samt med den kontinuerliga förklaringsvariabel x (standardiserad) som har högst korrelation med den beroende variabeln y.

Den variabel som påvisar störst korrelation med den beroende variabeln y är variabeln x_2 alltså *antalrum*. Denna variabel kommer således användas för kommande analys.

Den antagna bayesianska linjära regressionmodellen för responsvariablen y med 1 standardiserad förklaringsvaribel är som följade:

$$\begin{aligned} y_i|\mu_i, \sigma, X &\overset{iid}{\sim} N(\mu_i, \sigma) \\ \mu_i &= \beta_0 + \beta_1 x_{1i} \\ \beta_0 &\sim N(3, 10) \\ \beta_j &\sim N(0, 10), j = 1 \\ \ln \sigma &\sim N(0, 1) \end{aligned} \tag{2.1}$$

2.1 a)

```
data <- data.frame(y/1000,scale(X[,1:4]),X[,5:6])
colnames(data) <- c("y","area","antal_rum","avgift","trappor","cityyes","sydyes")

cor(data)

flist <- alist(
  y ~ dnorm(mu, exp(logsigma)) ,
  mu <- b0 + b1*area ,
  b0 ~ dnorm( 3 , 10 ) ,
  b1 ~ dnorm( 0 , 10 ) ,
  logsigma ~ dnorm ( 0, 1 )
)

resNormal <- map(flist, data=data)
precis(resNormal,prob=0.909)
```

Ett 90.9 procent kredibilitetsintervall ska skapas och tolkas för β_1. Intervallet produceras i R där medelvärdet och standardavvikelsen för β_1 ses vara 0.85 respektive 0.13. Kredibilitetsintervallet ses ligga mellan 0.63 och 1.06, således med 90.9 procents sannolikhet ligger parametern mellan 0.63 och 1.06. Således kan slutsatsen dras att med 90.9 procents sannolikhet kommer förklaringsvariabeln *antalrum* påverka y på ett linjärt positivt sätt.

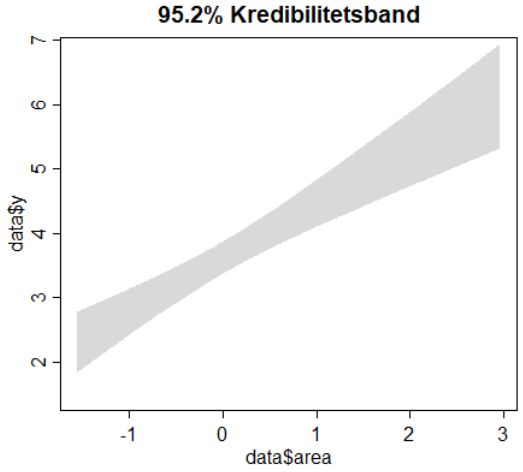
2.2 b)

```
postSamples_2a <- extract.samples(resNormal , n=1e4)
```

```
xSeq <- seq(from=min(data$area),to=max(data$area),by=0.01)
plot( data$y ~ data$area , type="n" , main = "95.2% Kredibilitetsintervall")

mu.ci <- sapply( xSeq , function(x) PI( postSamples_2a[,1] + postSamples_2a[,2]*x , prob=0.952)
    )
shade( mu.ci , xSeq)
```

Genom att använda en grid av värden på x från det lägsta till det högsta värdet på x ska ett 95.2 procent kredibilitetsintervall för μ som funktion av x skapas.

Kredibilitetsbandet i figur 2.1 visar på att när en lägenhetsarea ökar så ökar även priset. Det ses även att osäkerheten i intervallet är som minst då $y \approx 3$ och att osäkerheten ökar med störrevärden på y och ökar även för mindre värden på y. Det är rimligt att osäkerheten blir minst då $y \approx 3$ eftersom det är ungefär medelvärdet på y som tidigare redovisat, det finns således mest information runt detta värde och osäkerheten minskar i kredibilitetsintervallet.



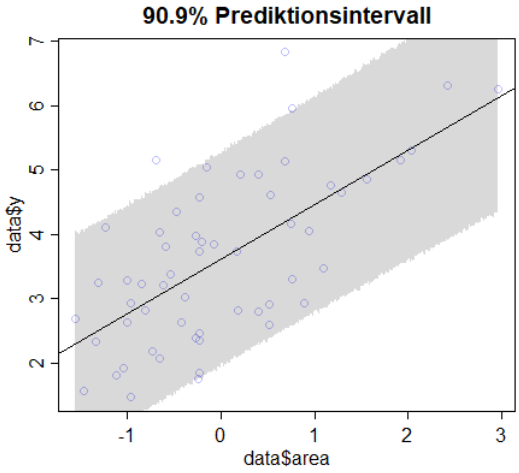
Figur 2.1

2.3 c)

```
sim.y <- matrix(0,nrow=1e4,ncol=length(xSeq))
sigma <- exp(matrix(postSamples_2a[,3],nrow=1000,ncol=1))
for ( i in 1:length(xSeq) ){
  mu <- postSamples_2a[,1] + postSamples_2a[,2]*xSeq[i]
  sim.y[,i] <- rnorm(1e4,mu,sigma)
}
y.PI <- apply( sim.y , 2 , PI , prob=0.909 )

plot( data$y ~ data$area , col=col.alpha(rangi2,0.5) , main = "90.9% Prediktionsintervall")
abline( a=coef(resNormal)["b0"] , b=coef(resNormal)["b1"] )
shade( y.PI , xSeq )
```

Ett 90.9 procent prediktionsintervall för y ska plottas som funktion av x genom att använda samma grid av värden på x som i uppgift (b).

Prediktionsintervallet plottas i figur 2.2 och kan tolkas som givet en *ny* lägenhets area predikteras den med 90.9 procents sannolikhet inom det gråa intervallet. Således går det att observera att en *ny* lägenhet som har stor area förväntas ha ett högre pris i genomsnitt.



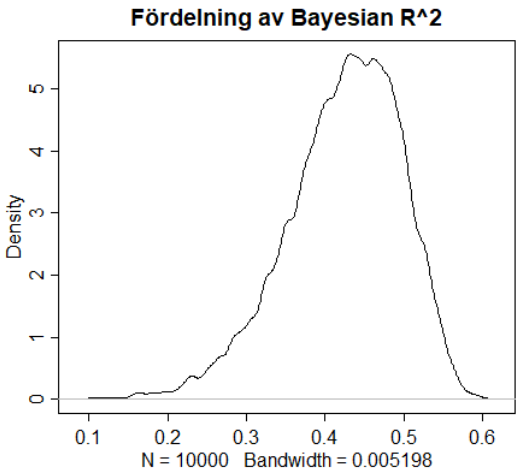
Figur 2.2

2.4 d)

```
RjSamples <- extract.samples(resNormal , n=1e4)

SSR_sample <- rep(NA, nrow(RjSamples))
SSE_sample <- rep(NA, nrow(RjSamples))
for (j in 1:nrow(RjSamples)) {
  SSR <- rep(NA, length(data$area))
  SSE <- rep(NA, length(data$area))
  for (i in 1:length(data$area)) {
  mu_i <- RjSamples[j,1] + RjSamples[j,2] * data$area[i]
  SSR[i] <- (mu_i - mean(data$y))^2
  SSE[i] <- (data$y[i] - mu_i)^2
  }
  SSR_sample[j] <- sum(SSR)
  SSE_sample[j] <- sum(SSE)
}

R2_bayes <- SSR_sample / (SSR_sample + SSE_sample)
dens(R2_bayes, main = "Fördelning av Bayesian R^2")
```

Den alternativa Bayesianska förklaringsgraden $BayesianR_j^2$ ska beräknas för den linjära regressionmodellen utifrån varje posteriordragning j på μ_{ij}, fördelningen ska plottas.

Fördelningen av $BayesianR_j^2$ presenteras i figur 2.3 där medelvärdet för den bayesianska förklaringsgraden ligger på 42 procent med en viss spridning. Den bayesianska förklaringsgraden kan tolkas som en databaserad skattning av andelen varians som förklaras för ny data. Således skattas med högst sannolikhet att 42 procent av variansen förklaras för ny data.


Fördelning av Bayesian R^2
Density
0
1
2
3
4
5
0.1
0.2
0.3
0.4
0.5
0.6
N = 10000 Bandwidth = 0.005198

Figur 2.3

3. Uppgift 3

I följande uppgift ska med hjälp av kvadratisk approximation en Bayesiansk linjär regressionsmodell anpassas med den beroende variabeln y och **alla** standardiserat förklaringsvariabler (dummy-variablerna standardiseras ej).

Den antagna bayesianska linjära regressionmodellen för responsvariablen y med 6 standardiserad förklaringsvaribel (ej dummyvariablerna) är som följade:

$$\begin{gathered} y_i|\mu_i, \sigma, X \overset{iid}{\sim} N\left(\mu_i, \sigma\right) \\ \mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} \\ \beta_0 \sim N(3, 10) \\ \beta_j \sim N(0, 10), j = 1, 2, 3, 4, 5, 6 \\ \ln \sigma \sim N(0, 1) \end{gathered} \tag{3.1}$$

3.1 a)

```
flist <- alist(
  y ~ dnorm(mu, exp(logsigma)) ,
  mu <- b0 + b1*area + b2*antal_rum + b3*avgift + b4*trappor + b5*cityyes + b6*sydyes,
  b0 ~ dnorm( 3 , 10 ) ,
  c(b1,b2,b3,b4) ~ dnorm( 0 , 10 ) ,
  c(b5,b6) ~ dnorm( 0 , 10 ) ,
  logsigma ~ dnorm ( 0, 1 )
)

resNormal <- map(flist, data=data)

postSamples <- extract.samples(resNormal , n=1e6)

Odds <- matrix(NA, ncol = 6, nrow = 1)
colnames(Odds) <- c("b1","b2","b3","b4","b5","b6")
for (i in 1:6) {
  ProbNegEff <- sum(postSamples[,i+1] < 0)/1e6
  ProbPosEff <- 1 - ProbNegEff
  Odds[,i] <- ProbNegEff/ProbPosEff
}

Odds
```

Oddset för positiv eller negativ lutning för respektive lutningsparameter ska beräknas och tolkas. I tabell 3.1 ses oddsen för parametrarna att bidra entydigt linjärt negativt till modellen givet att övriga är med i modellen. Parametrarna β_1,β_4 och β_6 ses ha 0 i odds, således påverkar förklaringsvariablerna x_1,x_4 och x_6 med 100 procent sannolikhet modellen entydigt linjärt positivt. Parametrarna β_2, β_3 och β_6 ses ha 1.68,2.42 och 0.27 i odds. Således är det 1.68 gånger med sannolikt att x_2 påverkar modellen på ett entydigt negativt sätt än ett entydigt positivt sätt. Samma tolkning gäller för x_3 och x_6. De förklaringsvariabler som verkar bidra till regressionsmodellen utifrån oddsen är x_1,x_4 samt x_5 eftersom dessa med 100 procent sannolikhet bidrar till modellen entydigt linjärt positivt.

Tabell 3.1

b1	b2	b3	b4	b5	b6
0	1.68	2.42	0	0	0.27

3.2 b)

```
precis(resNormal,prob=0.909)
```

Ett 90.9 procents kredibilitetsintervall för respektive lutningsparameter ska skapas om respektive förklaringsvariabel bidrar på ett entydigt linjärt positivt eller negativt sätt till modellen. Alla tolkningar nedan är givet att resterande förklaringsvariabler finns med i modellen.

Parametern β_1 ses ligga mellan 0.84 och 1.81 med 90.9 procents säkerhet. Således påverkar förklaringsvariabeln x_1 entydigt linjärt positivt till modellen. β_2 ligger mellan -0.47 och 0.32 med 90.9 procents sannolikhet och x_2 bidrar således inte till modellen på något entydigt linjärt positivt eller negativt sätt. β_3 ligger mellan -0.5 och 0.25 med 90.9 procents sannolikhet och x_3 påvisar således inget entydigt positivt eller negativt bidragande till modellen. β_4 ligger mellan 0.08 och 0.41 med 90.9 procents sannolikhet och x_4 bidrar således entydigt linjärt positivt till modellen. β_5 ligger mellan 0.92 och 2.03 med 90.9 procents sannolikhet och x_5 bidrar således entydigt linjärt positivt till modellen. β_6 ligger mellan -0.27 och 0.75 med 90.9 procents sannolikhet och x_6 bidrar således inte entydigt linjärt positivt eller negativt till modellen.

Tabell 3.2

	Mean	StdDev	4.55%	95.45%
b1	1.33	0.29	0.84	1.81
b2	-0.07	0.23	-0.47	0.32
b3	-0.12	0.22	-0.50	0.25
b4	1.48	0.10	0.08	0.41
b5	0.24	0.33	0.92	2.03
b6	0.24	0.30	-0.27	0.75

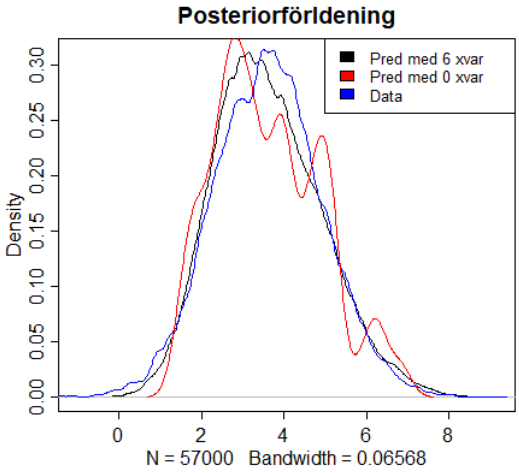
3.3 c)

```
postSamples <- extract.samples(resNormal , n=1e4)

mu_ij <- matrix(NA,ncol = length(data$y), nrow = nrow(postSamples))

for (i in 1:nrow(postSamples)) {
  for (j in 1:length(data$y)) {
    mu_ij[i,j] <- (postSamples[i,1] + postSamples[i,2] * data$area[j] + postSamples[i,3] *
        data$antal_rum[j]
               + postSamples[i,4] * data$avgift[j] + postSamples[i,5] * data$trappor[j]
               + postSamples[i,6] * data$cityyes[j] + postSamples[i,7] * data$sydyes[j])
  }
}


pred <- matrix(NA, ncol = length(data$y), nrow = nrow(postSamples))
for (i in 1:length(data$y)) {
  pred[,i] <- rnorm(n = 1e3, mean = mu_ij[,i], sd = exp(postSamples$logsigma))
}
```

```
pred_matrix <- matrix(pred, ncol=1)

dens(pred_matrix, main = "Posteriorförldening")
dens(yPred_1d,type="l",col="blue",add=TRUE)
dens(data$y,add=TRUE,col="red")
legend("topright", legend = c("Pred med 6 xvar", "Pred med 0 xvar", "Data"), fill=c("black",
    "red","blue"), cex = 0.8)
```

En modellutvärdering ska göras genom att replikera data från den posterior prediktiva fördelningen $p(\tilde{y}|y)$. Denna fördelning ska plottas tillsammans med den posterior prediktiva fördelningen $p(\tilde{y}|y)$ i uppgift 1(d) och även fördelningen för de faktiska värdena på y från uppgift 1(d).

I figur 3.1 redovisas de tre fördelningarna. Den posteriorprediktiva fördelningen för modellen med 6 förklaringsvariabler och 0 förklaringsvariabler ses båda anpassa data bra. Men modellen med 6 förklaringsvariabler ses ändå följa datas fördelning mer exakt vilket är förväntat vid fler antal förklaringsvariabler där 3 stycken av dom påverkade modellen på ett entydigt linjärt positivt sätt.



Figur 3.1