STA 360/602L: Module 6.3

Bayesian linear regression: weakly informative priors

Dr. Olanrewaju Michael Akande

Bayesian linear regression recap

- Sampling model:

$$\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_{n \times n}).$$

- Semi-conjugate prior for $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\beta}) = \mathcal{N}_p(\boldsymbol{\mu}_0, \Sigma_0).$$

- Semi-conjugate prior for σ^2:

$$\pi(\sigma^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

Full conditional

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2) = \mathcal{N}_p(\boldsymbol{\mu}_n, \Sigma_n),$$

where

$$\Sigma_n = \left[\Sigma_0^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}$$
$$\boldsymbol{\mu}_n = \Sigma_n\left[\Sigma_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{y}\right],$$

and

$$\pi(\sigma^2|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}) = \mathcal{IG}\left(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2}\right),$$

where

$$\nu_n = \nu_0 + n$$
$$\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right] = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + \text{SSR}(\boldsymbol{\beta})\right].$$

Weakly informative priors

- Specifying hyperparameters that represent actual prior information can be challenging, especially for β.

- It can therefore be desirable use weakly informative priors when possible. The Hoff book discusses a few different options, one of which is the Zellner's g-prior (there are other options but we will not cover them in this course).

- Note that we can also use Jefferys prior here to be completely non-informative.

- Zellner's g-prior is

$$\pi(\boldsymbol{\beta}|\sigma^2) = \mathcal{N}_p\left(\boldsymbol{\mu}_0 = \mathbf{0}, \Sigma_0 = g\sigma^2\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}\right)$$
$$\pi(\sigma^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$

for some positive value g, which is often commonly set to the sample size n.

Weakly informative priors

- Note that the g-prior uses a part of the data. As I have mentioned before, using your data to construct your prior is usually a no-no.

- However, the g-prior actually does not use the information in $\boldsymbol{y}$, the response variable of interest, just the information in $\boldsymbol{X}$.

- Observe that the prior specification actually looks like the conjugate prior we first used for the univariate normal model, that is, with

$$\begin{aligned}\sigma^2 &\sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)\\ \mu|\sigma^2 &\sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right).\end{aligned}$$

- Turns out that we also have conjugacy with the g-prior, so that we don't actually need Gibbs sampling to obtain posterior samples. $\pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2)$ takes the same form as before but now we also have $\pi(\sigma^2|\boldsymbol{y}, \boldsymbol{X})$.

Weakly informative priors

- With the g-prior, we have

$$\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2) = \mathcal{N}_p(\boldsymbol{\mu}_n, \Sigma_n)$$
$$\pi(\sigma^2|\boldsymbol{y},\boldsymbol{X}) = \mathcal{IG}\left(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2}\right)$$

where

$$\Sigma_n = \left[\Sigma_0^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\right]^{-1} = \left[\frac{1}{g\sigma^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\right]^{-1} = \frac{g}{g+1}\sigma^2\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}$$

$$\boldsymbol{\mu}_n = \Sigma_n\left[\Sigma_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{y}\right] = \frac{g}{g+1}\sigma^2\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}\left[\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{y}\right]$$
$$= \frac{g}{g+1}\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T\boldsymbol{y} = \frac{g}{g+1}\hat{\boldsymbol{\beta}}_{\text{ols}}$$

$$\nu_n = \nu_0 + n; \qquad \sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + \text{SSR}(g)\right],$$

where $\text{SSR}(g) = \boldsymbol{y}^T(\boldsymbol{I} - \frac{g}{g+1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T)\boldsymbol{y}$. See the Hoff book for the proof, and see homework for illustration.

Example

- Health plans use many tools to try to control the cost of prescription medicines.
- For older drugs, generic substitutes that are the equivalent to name-brand drugs are available at considerable savings.
- Another tool that may lower costs is restricting drugs that the physician may prescribe.
- For example if three similar drugs for treating the same condition are available, a health plan may require the physician to prescribe only one of them, allowing the plan to negotiate discounts based on a higher volume of sales.
- We have data from 29 health plans can be used to explore the effectiveness of these two strategies in controlling drug costs.
- The response is COST, the average cost of the prescriptions to the plan per day (in dollars).

Example

- Explanatory variables are:
  - RXPM: Average number of prescriptions per member per year
  - GS: Percent generic substitute used by the plan
  - RI: Restrictiveness Index, from 0 (no restrictions) to 100 (total restrictions on the physician)
  - COPAY: Average member copay on prescriptions
  - AGE: Average member age
  - F: percent female members
  - MM: Member months, a measure of the size of the plan
  - ID: an identifier for the name of the plan
- The data is in the file `costs.txt` on Sakai.
- For this illustration, we will restrict ourselves to GS and AGE. We will use the other variables later.

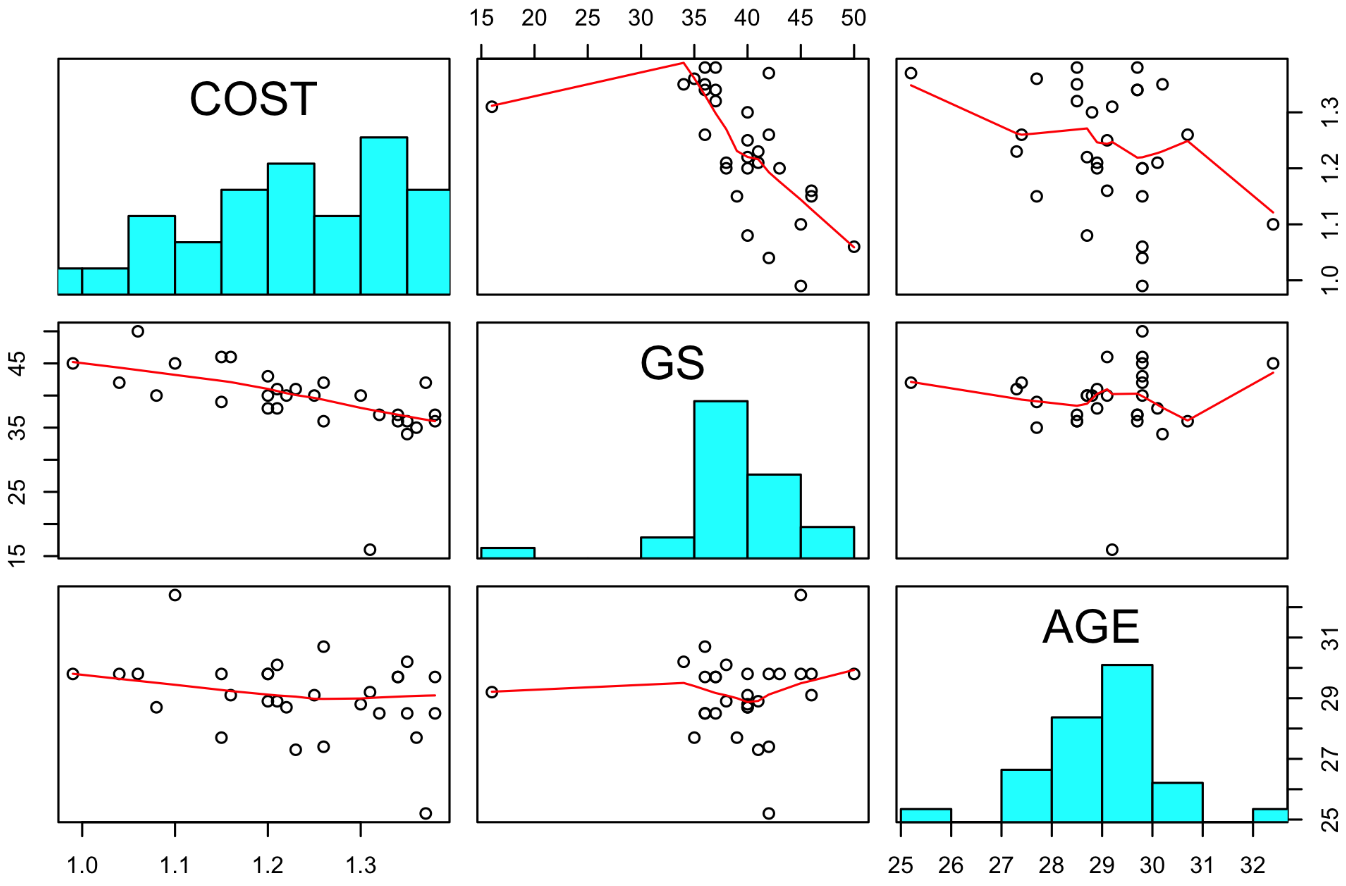
Data

```r
#require(lattice)
#library(pls)
#library(calibrate)
#library(mvtnorm)

###### Data
Data <- read.table("data/costs.txt",header=TRUE)[,-9]
head(Data)
```

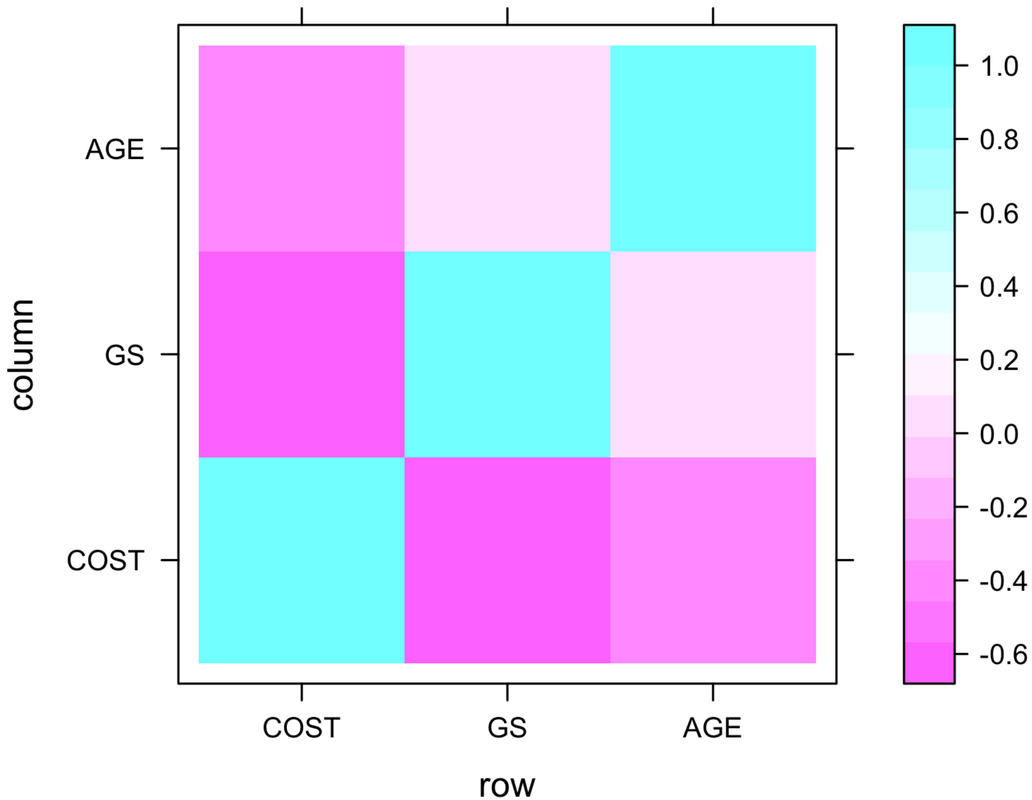
```
##   COST RXPM GS   RI COPAY  AGE    F      MM
## 1 1.34  4.2 36 45.6 10.87 29.7 52.3 1158096
## 2 1.34  5.4 37 45.6  8.66 29.7 52.3 1049892
## 3 1.38  7.0 37 45.6  8.12 29.7 52.3   96168
## 4 1.22  7.1 40 23.6  5.89 28.7 53.4  407268
## 5 1.08  3.5 40 23.6  6.05 28.7 53.4   13224
## 6 1.16  7.2 46 22.3  5.05 29.1 52.2  303312
```

Very basic EDA



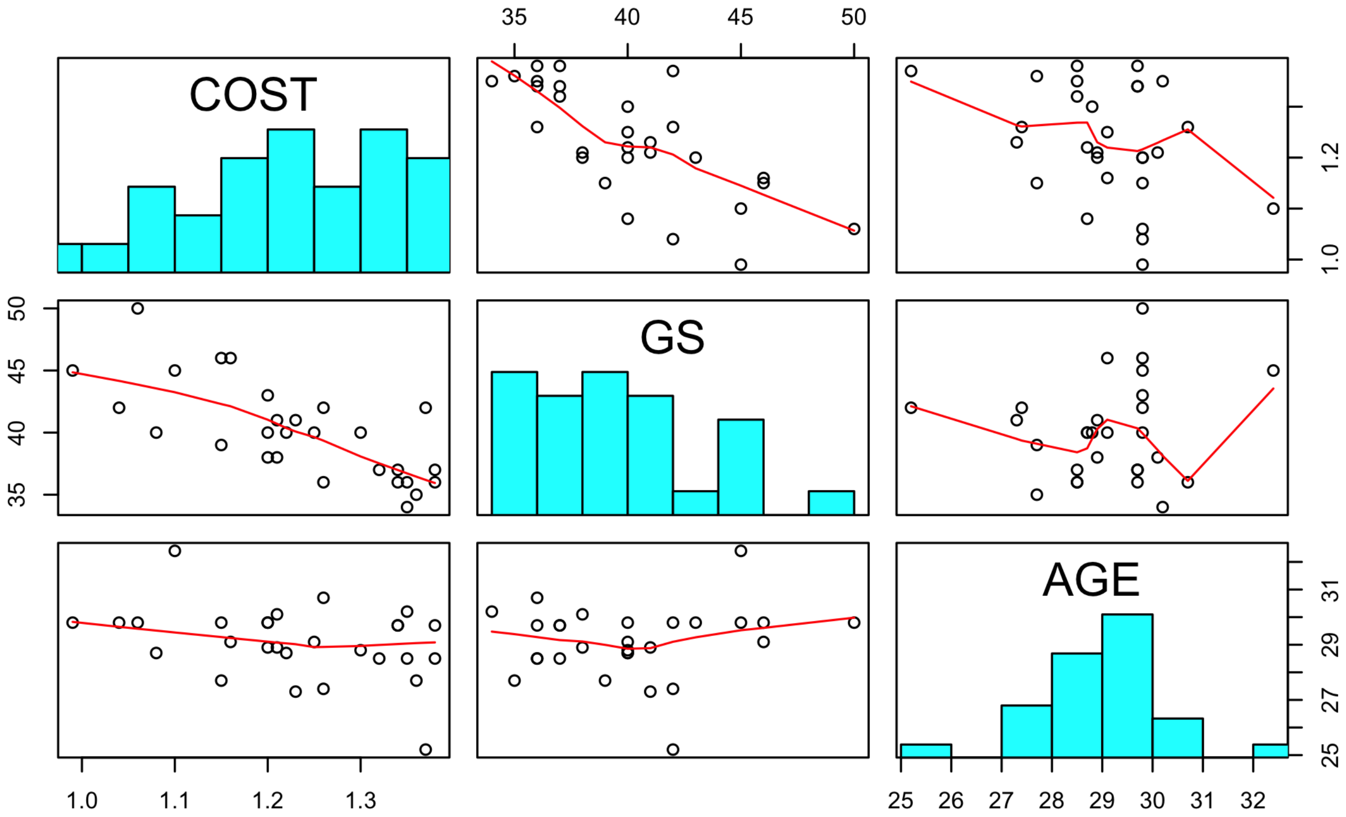
Very basic EDA

```
levelplot(cor(Data[,c("COST","GS","AGE")])) #Check correlation
```



Very basic EDA

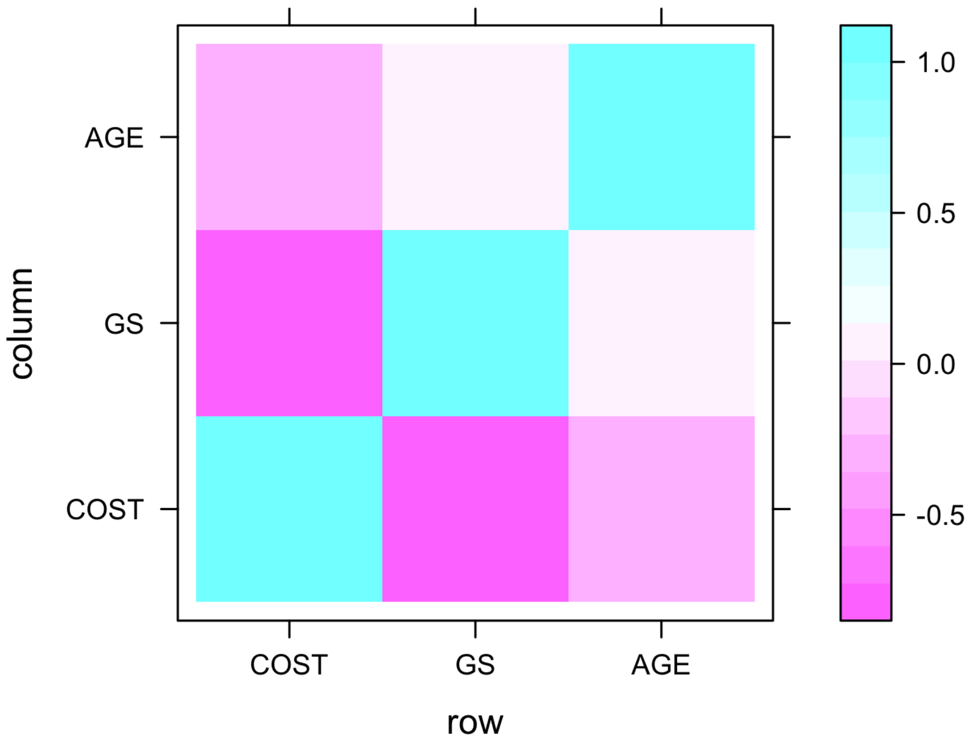
Without outlier:



Very basic EDA

Without outlier:

```
levelplot(cor(Data[-19,c("COST","GS","AGE")])) #Check correlation
```



Posterior computation

```
###### g-Prior: with g=n using full model
# Data summaries
X <- cbind(1,as.matrix(Data[-19,c("GS","AGE")])) #remove potential outlier
Y <- matrix(Data$COST[-19],ncol=1)
n <- length(Y)
p <- ncol(X)
g <- n

# OLS estimates
beta_ols <- solve(t(X)%*%X)%*%t(X)%*%Y
round(t(beta_ols),4)
```

```
##                 GS     AGE
## [1,] 2.7047 -0.02 -0.0231
```

```
SSR_beta_ols <- (t(Y - (X%*%beta_ols)))%*%(Y - (X%*%beta_ols))
sigma_ols <- SSR_beta_ols/(n-p)
sigma_ols
```

```
##              [,1]
## [1,] 0.005247074
```

```
# Hyperparameters for the priors
#sigma_0_sq <- sigma_ols
sigma_0_sq <- 1/100
nu_0 <- 1

# Set number of iterations
S <- 10000
```

Posterior computation

```
set.seed(1234)

# Sample sigma_sq
nu_n <- nu_0 + n
Hg <- (g/(g+1))* X%*%solve(t(X)%*%X)%*%t(X)
SSRg <- t(Y)%*%(diag(1,nrow=n) - Hg)%*%Y
nu_n_sigma_n_sq <- nu_0*sigma_0_sq + SSRg
sigma_sq <- 1/rgamma(S,(nu_n/2),(nu_n_sigma_n_sq/2))

# Sample beta
mu_n <- g*beta_ols/(g+1)
beta <- matrix(nrow=S,ncol=p)
for(s in 1:S){
  Sigma_n <- g*sigma_sq[s]*solve(t(X)%*%X)/(g+1)
  beta[s,] <- rmvnorm(1,mu_n,Sigma_n)
}

#posterior summaries
colnames(beta) <- colnames(X)
mean_beta <- apply(beta,2,mean)
round(mean_beta,4)
```

```
##                  GS     AGE
##  2.6057 -0.0193 -0.0221
```

```
round(apply(beta,2,function(x) quantile(x,c(0.025,0.975))),4)
```

```
##                        GS     AGE
## 2.5%   0.4392 -0.0432 -0.0935
## 97.5% 4.7903  0.0044  0.0460
```

What's next?

Move on to the readings for the next module!