Yuanjie Jin
12/5/2013

STA 721 Take-Home Data Analysis Report

Executive Summary

My data analysis reveals that the use of generic substitutes would be a more effective strategy in controlling drug costs than restricting physicians to prescribe drugs. Based on average settings for the other covariates, one percent increase in generic substitute use would save approximately 0.01 dollar in the average cost of the prescription per day, with the standard deviation of this saving being 0.0028. In other words, by increasing 1 percent in GS use, the projected saving in drug costs per day has around 95% chance to range from 0.0044 to 0.0156 dollars.

Introduction

Health care costing has been a concern of patients as well as insurance companies. To reduce costs of medicine, health plans utilize a variety of methods. For instance, generic substitutes with a lower price are used in place of name brand drugs. Another way is to restrict physicians' prescription of drugs when multiple drugs are available with similar effects. To study how effective these two strategies reduce the cost, 29 health plans across the US were collected. Each plan includes data for percentage of generic substitute used by the plan (GS), for Restrictiveness Index (RI, in a 0 to 100 scale, with 0 meaning no restrictions and 100 representing the total restrictions), and for several other factors. Based on the data, I seek to identify the relationship between drug costs and GS or RI, if any, and try to estimate the relative effect of either strategy on control of the drug costs.

Methods

By scanning the original data, I found that the predictor MM (member months) have much larger values than all the other predictions and the response. So a logarithmic transformation was taken on MM for a smaller scale. The full linear model for the response COST (Y) given all the potential predictors is then:

$$Y_i = \alpha + \beta_1 RXPM_i + \beta_2 GS_i + \beta_3 RI_i + \beta_4 COPAY_i + \beta_5 AGE_i + \beta_6 F_i + \beta_7 \log(MM_i) + \epsilon_i, \quad \epsilon_i \sim N\left(0, \frac{1}{\phi}\right)$$

where $i = 1, \cdots, 29$ is the index of health plans. This model can be rewritten in a matrix form as follows:

$$Y = 1\alpha + X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I) = N(0, \frac{I}{\phi})$$

where $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_{29} \end{bmatrix}$, X is the 29×7 design matrix for all predictors, $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_7 \end{bmatrix}$ and $\phi = \frac{1}{\sigma^2}$ is the precision.

The above full model was fitted using the invariant Zellner's g prior with g = 29, that is, $\beta|\phi \sim N(0, \frac{g(X^T X)^{-1}}{\phi})$ and $p(\alpha, \phi) \propto \frac{1}{\phi}$.

The model was also fitted by ridge, lasso and horseshoe regression. Let $Y^c = (1 - P_1)Y$ and X^c be the centered and standardized X matrix. Then, ridge regression seeks to minimize the penalized likelihood

$$\min_{\beta} \|Y^c - X^c\beta\| + k\|\beta\|^2$$

The best k was chosen based on the minimum GCV value.

In lasso regression, following penalized likelihood was minimized:

$$\min_{\beta} \|Y^c - X^c\beta\| + \lambda\|\beta\|_1$$

The best λ was chosen based on the minimum Mallows's C_p value.

In horseshoe regression, we assume priors $\beta \left| \phi \sim N\left(0, \frac{diag(\tau^2)}{\phi}\right), \tau_j^2 \right| \lambda \sim i.i.d. C^+(0,\lambda), \lambda \sim C^+\left(0, \frac{1}{\phi}\right)$ and $p(\alpha, \phi) \propto \frac{1}{\phi}$, where C^+ is the Half-Cauchy distribution.

Finally, Bayesian Model Averaging (BMA) using Zellner-Siow Cauchy prior (see R code in Appendix II for details) was performed to search for the best models and give some ideas how important each predictor is in terms of predicting the value of the response.

Results

Table 1 lists coefficient estimates given by the full model using the g prior. Note that the coefficient estimate of GS is -0.0108, with a standard deviation 0.0023, suggesting that increase in GS is effective in reducing the cost. On the other hand, the small value of coefficient estimate 0.0006 with its relative large standard deviation 0.0017 for RI indicates that RI might have little effect on the costs. This relationship of GS and RI to COST adjusting for the other variables are also visualized in the added-variable plots (Figure 1).

Figure 2 shows the diagnostic plots using Frequentist normal multiple linear regression for the full model. Note that case 19 is highly influential because it has a high leverage and a high Cook's distance (greater than 1). Thus, this case is very likely to be an outlier and needs to be checked. Further analysis may be necessary by redoing our model fitting without case 19.

The ridge regression gives -0.056 and 0.0079 as estimates of coefficients for GS and RI, respectively (see details in Table 2). The lasso regression provides -0.065 and 0.0082 as estimates of coefficients for GS and RI, respectively (see details in Table 3). From horseshoe regression, the posterior means for coefficients of GS and RI are -0.0564 and 0.0064, respectively. And their 95% posterior credible interval are (-0.0893, -0.0189) and (-0.0191, 0.0393), respectively (see details in Table 4). These results corroborate the fact that GS has a negative effect on COST, whereas RI has only a marginal effect on COST.

BMA using the Zellner-Siow Cauchy prior shows that the posterior inclusion probability of GS is very high (0.99, greater than 0.5, see Table 5 for details), indicating GS should be included in the model. The estimate of GS coefficient is -0.0099, with a standard deviation of 0.0028, suggesting it has a high probability to be negative (>0.95). Although the inclusion probability of RI is less than 0.5, it does not mean that it should be ruled out at this point because it has strong correlation with COPAY (see the

yellow-shaded area in Table 6). This is also confirmed by the variance inflation factors for both RI and COPAY, which are greater than 2 (see Table 7 for details). Therefore, RI is likely to be collinear with COPAY, whose inclusion probability is also less than 0.5 (Table 5). RI also has a positive correlation with AGE, to a lessor extent (Table 6). As a result, although the coefficient estimate of RI is only 0.0004 with a standard deviation of 0.0010 (Table 5), its real effect might be masked by COPAY and AGE.

Discussion

The full model using g prior inherits the potential issue from ordinary least squares (OLS) in that it assumes lack of multicollinearity between variables. However, as mentioned in Results section, RI is likely to have some collinearity with COPAY and potentially AGE. Also, as shown in Table 9, the smallest eigenvalue of $X^T X$ (0.0117) is very close to 0, where X is the whole design matrix including intercepts. This is additional evidence suggesting that some columns in X are nearly linearly dependent and that the assumptions for g-prior in full model might be violated. These collinearity issues can inflate the mean square error (MSE) of model estimates, causing results sensitive to small changes in the model or the data. Collinarity will also affect the results of BMA. For example, if two variables are highly positively correlated, then the inclusion probability of either variable might be compromised in BMA. However, multicollinearity does not reduce the predictive power or reliability of the model as a whole, so BMS is expected to be still good at predictions. In contrast, the ridge, lasso and horseshoe regression can largely avoid the problem of multicollinearity by providing a certain degree of shrinkage to stabilize the estimates.

The ridge and lasso regression have their own problem, though. Quantifying uncertainty is difficult for both methods. Thus, without uncertainty associated with the coefficient estimates (and also variable selection in lasso), the results from these methods should be taken with caution and should be compared with results from other methods that provide uncertainty measures. That is one primary reason I also include horseshoe regression here, whose Bayesian paradigm gives some uncertainty measure.

As mentioned above, BMA is expected to give good predictions by averaging over all possible models. Hence, BMA is likely to beat the single full model with g prior in terms of prediction. The Model Probabilities plot in Figure 3 shows that model space is enumerated since the cumulative posterior probability of included models sums to 1 eventually. From the Model Complexity plot shown in Figure 3, model 9 appears to be the highest probability model. However, its inclusion probability is not dramatically higher than the other 4 top models summarized in Table 8. Therefore, I would not recommend picking one model. This also justifies the use of BMA, since no single model stands out with extremely high inclusion probability. Although the shrinkage methods may also work fine, the estimates of coefficients are harder to interpret since they fit normalized data. For these reasons, BMA estimates are picked in executive summary.

On the whole, it seems that the different methods used in this analysis lead to similar conclusions. That is, GS has a negative effect on the COST. Although the multicollinearity with other variables may mask

the effect of RI, RI is likely to have less effect on the COST compared with RI. To sum up, GS is a more effective strategy in controlling the drug costs.

Appendix I. Tables and figures for this report.

Table 1. Summary of posterior coefficient estimates from the full model using g prior with g = 29

```
> summary(cost.newzlm)
Coefficients
                Exp.Val.     St.Dev.
(Intercept)  1.7631549474          NA
RXPM         0.0192869170 0.009027223
GS          -0.0107657006 0.002307369
RI           0.0006219003 0.001692654
COPAY        0.0136920307 0.015329409
AGE         -0.0414979608 0.012411899
F            0.0134718212 0.007903334
logMM        0.0155756948 0.006229202

 Log Marginal Likelihood:
17.343
 g-Prior: UIP
Shrinkage Factor: 0.967
```

Table 2. Summary of coefficient estimates for all variables by ridge regression. The first number is the estimate of the intercept.

```
> coef(cost.ridge.newrefit)
               X.newnormRXPM    X.newnormGS    X.newnormRI X.newnormCOPAY
   1.233448276    0.024114506   -0.056009673    0.007896759    0.013217060
  X.newnormAGE     X.newnormF X.newnormlogMM
  -0.043187418    0.024227048    0.028965398
```

Table 3. Summary of coefficient estimates for all variables by lasso regression

```
> coef.newlasso
        RXPM           GS           RI        COPAY          AGE            F
 0.033432423 -0.065391307  0.008186307  0.020339368 -0.055330564  0.024840646
       logMM
 0.035020715
```

Table 4. Summary of posterior estimates and 95% credible intervals for all coefficients by horseshoe regression.

coefficient	posterior mean estimate	95% posterior credible interval
intercept	1.234	(1.203, 1.264)

RXPM	0.0157	(-0.0090, 0.0501)
GS	-0.0564	(-0.0893, -0.0189)
RI	0.0064	(-0.0191, 0.0393)
COPAY	0.0068	(-0.0194, 0.0411)
AGE	-0.0352	(-0.0754, 0.0014)
F	0.0169	(-0.0071, 0.0505)
MM	0.0209	(-0.0048, 0.0544)

Table 5. Summary of posterior coefficient estimates and inclusion probabilities in BMA using the Zellner-Siow Cauchy prior

```
> coef.newbma

 Marginal Posterior Summaries of Coefficients:
           post mean   post SD     post p(B != 0)
Intercept   1.2334483   0.0150352   1.0000000
RXPM        0.0093675   0.0109899   0.5641398
GS         -0.0099171   0.0028080   0.9900319
RI          0.0004007   0.0010432   0.3192730
COPAY       0.0041144   0.0099556   0.3385025
AGE        -0.0291194   0.0179650   0.8437756
F           0.0064066   0.0092074   0.4749040
logMM       0.0097574   0.0087849   0.6849991
```

Table 6. Correlation matrix of the data

```
> health.newcor
            RXPM          GS          RI       COPAY         AGE           F
RXPM   1.0000000  0.25951942  0.1010409 -0.18341062  0.25993757 -0.1163729
GS     0.2595194  1.00000000  0.1030968 -0.03173971  0.06227775 -0.1829342
RI     0.1010409  0.10309677  1.0000000  0.75839780  0.42422592 -0.2165733
COPAY -0.1834106 -0.03173971  0.7583978  1.00000000  0.37308093 -0.1521896
AGE    0.2599376  0.06227775  0.4242259  0.37308093  1.00000000 -0.3654983
F     -0.1163729 -0.18293418 -0.2165733 -0.15218957 -0.36549825  1.0000000
logMM  0.1570407  0.04880672  0.1171163  0.08551532  0.29514206 -0.1969117
            logMM
RXPM   0.15704071
GS     0.04880672
RI     0.11711629
COPAY  0.08551532
AGE    0.29514206
F     -0.19691168
logMM  1.00000000
```

Table 7. Summary of variance inflation factors (VIF) for all variables fit by the full normal regression model

```
> vif(cost.newlm)
    RXPM       GS       RI    COPAY      AGE        F    logMM
1.391889 1.116544 2.822002 2.947653 1.556824 1.207203 1.115081
```

Table 8. Summary of top 5 highest probability models from BMA

```
> summary(cost.newbma)
     Intercept RXPM GS RI COPAY AGE F logMM        BF PostProbs     R2 dim
[1,]         1    1  1  0     0   1 0     1 1.0000000    0.0671 0.5702   5
[2,]         1    1  1  0     1   1 1     1 0.9326095    0.0625 0.6548   7
[3,]         1    1  1  0     1   1 0     1 0.9318087    0.0625 0.6136   6
[4,]         1    0  1  0     0   1 0     1 0.9202712    0.0617 0.5154   4
[5,]         1    1  1  0     0   1 1     1 0.8249693    0.0553 0.6094   6
      logmarg
[1,] 3.836766
[2,] 3.766998
[3,] 3.766139
[4,] 3.753679
[5,] 3.644357
```

Table 9. Summary of eigenvalues for $X^T X$, where X is the whole design matrix including intercepts

```
> eigen(t(X.newextend) %*% X.newextend)$values
[1] 1.626837e+05 4.330030e+03 7.372158e+02 1.453021e+02 7.505137e+01
[6] 4.962077e+01 1.723678e+01 1.172993e-02
```

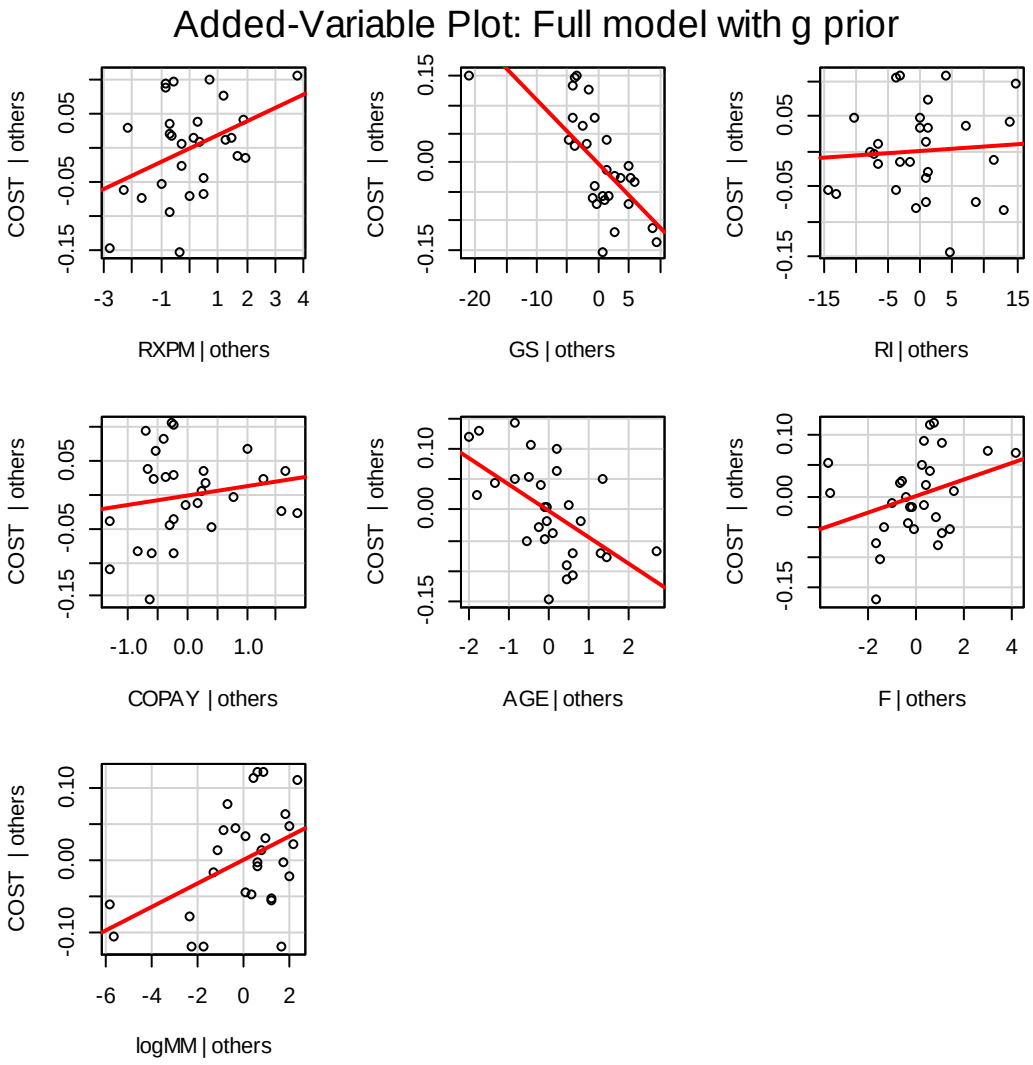

Figure 1. Added-Variable plot for the full model using g prior with g = 29

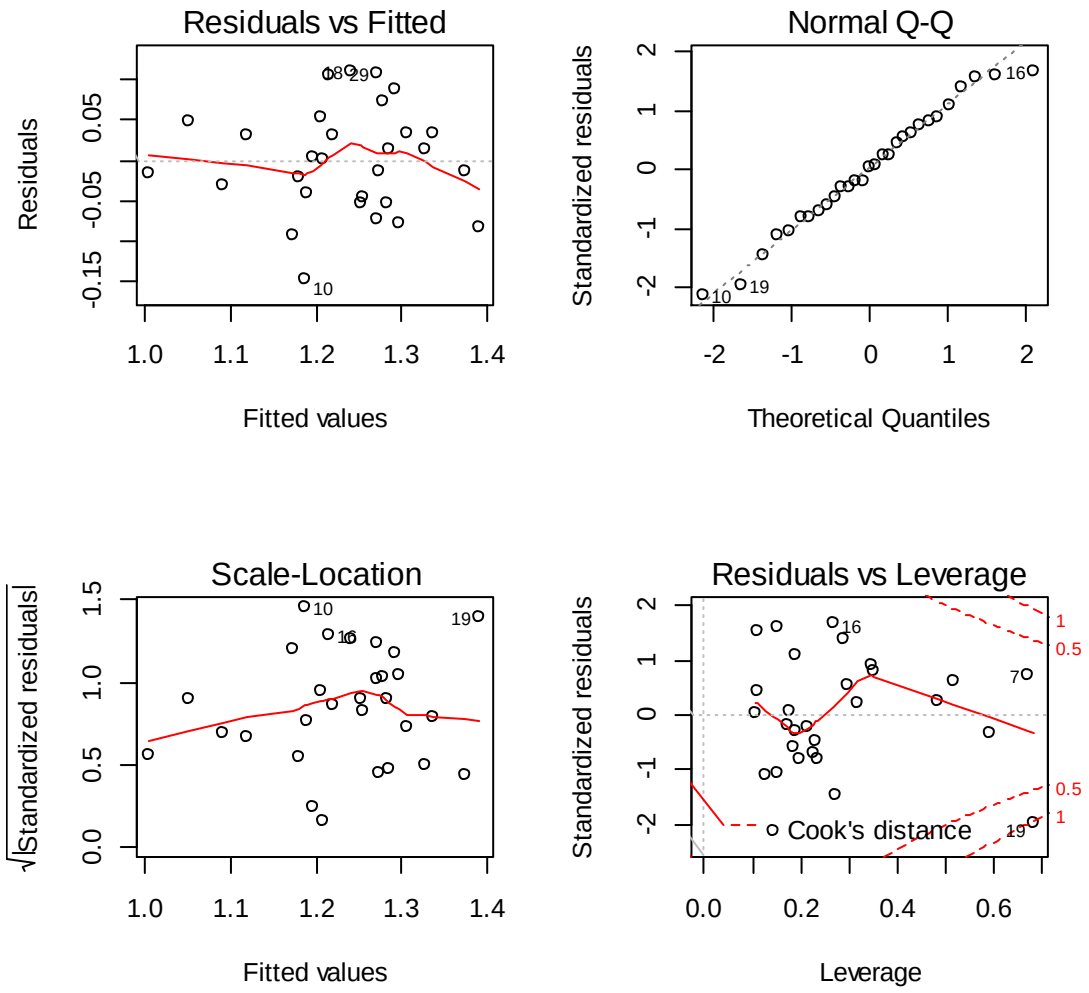

Figure 2. Diagnostic plots for the full Frequentist normal regression model

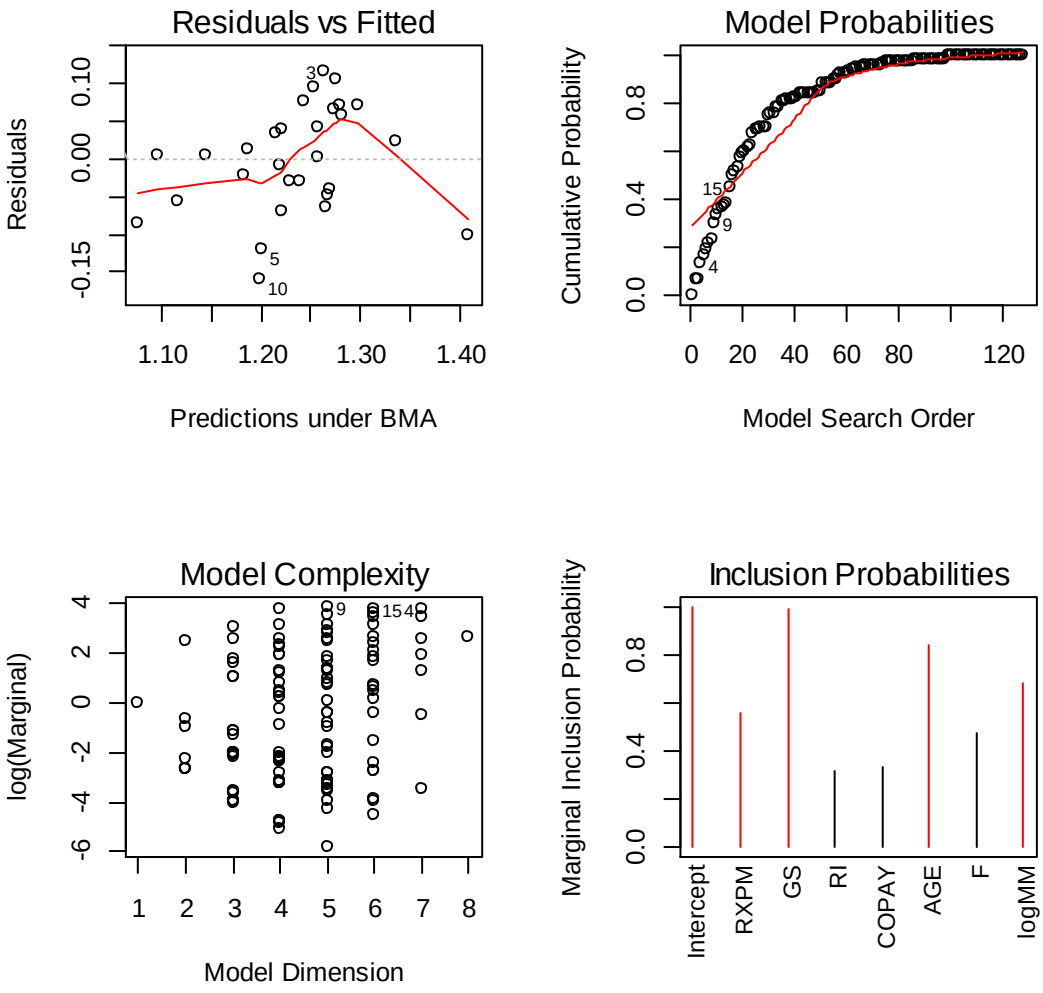

Figure 3. Four plots of BMA using the Zellner-Siow Cauchy prior. The top left is a plot of the residuals versus fitted values under BMA. The top right is a plot of the cumulative marginal likelihoods of models. The bottom left is a plot of log marginal likelihood versus model dimension and the bottom-right plot shows the posterior marginal inclusion probabilities of all variables.

Appendix II. R code

```
## load data and extract the response and the design matrix

health.data <- read.table("D:\\Course material\\Statistics and population genetics\\STA 721\\Take home Data analysis\\costs.txt", header = T)

names(health.data)

plot(health.data)

Y <- health.data[,1]

X <- data.matrix(health.data[,2:8])

X.norm <- scale(X)


## log transform the MM in original data

attach(health.data)

health.data$logMM <- log(MM)

detach(health.data)

health.newdata = data.frame(health.data[,c(1:7, 10)])


plot(health.newdata)

Y.new <- health.newdata[,1]

X.new <- data.matrix(health.newdata[,2:8])

X.newnorm <- scale(X.new)


## compute the eigenvalues of X^T X, where X here is the full design matrix including the intercept

X.newextend <- cbind(1,X.new)

eigen(t(X.newextend) %*% X.newextend)$values


## Frequentist normal regression and diagnostic plots
```

```
cost.newlm <- lm(COST ~ ., data = health.newdata)

summary(cost.newlm)

par(mfrow=c(2,2))

plot(cost.newlm)

vif(cost.newlm)


## correlation between variables

health.newcor <- cor(X.new)

health.newcor

library(lattice)

levelplot(health.newcor)


## Zellner's g prior with g = n using the full model

library(BMS)

cost.newzlm <- zlm(COST ~ ., data = health.newdata, g = "UIP")

summary(cost.newzlm)

coef(cost.newzlm)

## added-variable plots for g-prior regression

library(car)

avPlots(cost.newzlm, main="Added-Variable Plot: Full model with g prior")


## ridge regression

library(MASS)

cost.newridge = lm.ridge(Y.new ~ X.newnorm, lambda = seq(0, 10, 0.0001))

best.newlambda = as.numeric(names(which.min(cost.newridge$GCV)))
```

```
best.newlambda

cost.ridge.newrefit = lm.ridge(Y.new ~ X.newnorm, lambda = best.newlambda)

summary(cost.ridge.newrefit)

coef(cost.ridge.newrefit)

plot(cost.ridge.newrefit)


## lasso regression

library(lars)

cost.newlasso <- lars(X.newnorm, Y.new, type="lasso")

cost.newCp <- summary(cost.newlasso)$Cp

best.newCp <- (1:length(cost.newCp))[cost.newCp == min(cost.newCp)]

coef.newlasso = coef(cost.newlasso)[best.newCp,]

coef.newlasso

plot(cost.newlasso)


## horseshoe regression

library(monomvn)

cost.newbla <- blasso(X.newnorm, Y.new, T=11000, case = "hs", RJ = FALSE, normalize = F)

plot(cost.newbla, burnin = 1000)

## give posterior summaries using post-burnin MCMC samples

coef.newbla = c(mean(cost.newbla$mu[1001:11000]),apply(cost.newbla$beta[1001:11000,], 2, mean))

coef.newbla

quantile(cost.newbla$mu[1001:11000],c(.025,.975))

quantile(cost.newbla$beta[1001:11000,1],c(.025,.975))

quantile(cost.newbla$beta[1001:11000,2],c(.025,.975))
```

```
quantile(cost.newbla$beta[1001:11000,3],c(.025,.975))

quantile(cost.newbla$beta[1001:11000,4],c(.025,.975))

quantile(cost.newbla$beta[1001:11000,5],c(.025,.975))

quantile(cost.newbla$beta[1001:11000,6],c(.025,.975))

quantile(cost.newbla$beta[1001:11000,7],c(.025,.975))


## Bayesian model averaging with the Zellner-Siow Cauchy prior

library(BAS)

cost.newbma <- bas.lm(COST ~ ., data = health.newdata, prior="ZS-null", alpha=3, n.models=2^7,
update=50, initprobs="eplogp")

par(mfrow=c(2,2))

plot(cost.newbma, ask = F)

par(mfrow=c(1,1))

image(cost.newbma)

coef.newbma = coef(cost.newbma)

coef.newbma

par(mfrow=c(2,4))

plot(coef.newbma, subset=2:8,ask=F)

summary(cost.newbma)  ## show results with the top 5 models
```