Bayesian Linear Regression

Complete all **Exercises**, and submit answers to **Questions** on the Coursera platform.

Modeling Wages

In the field of labor economics, the study of income and wages provides insight about topics ranging from gender discrimination to the benefits of higher education. In this lab, we will analyze cross-sectional wage data in order to practice using Bayesian methods such as BIC and Bayesian Model Averaging to construct parsimonious predictive models.

Getting Started

Load packages

In this lab we will explore the data using the `dplyr` package and visualize it using the `ggplot2` package for data visualization. We also may use the MASS package to implement stepwise linear regression in one of the exercises. The data can be found in the companion package for this course, `statsr`.

Let's load the packages.

```
library(statsr)
library(MASS)
library(dplyr)
library(ggplot2)
library(BAS)
```

This is the first time we're using the `BAS` package. We will be using the `bas.lm` function from this package later in the lab to implement Bayesian Model Averaging. Please make sure that the version of `BAS` is 1.3.0 or greater.

The data

The data will be using in this lab were gathered as a random sample of 935 respondents throughout the United States. This data set was released as part of the series *Instructional Stata Datasets for Econometrics* by the Boston College Department of Economics (Wooldridge 2000).

Let's load the data:

```
data(wage)
```

variable	description
`wage`	weekly earnings (dollars)
`hours`	average hours worked per week
`IQ`	IQ score
`kww`	knowledge of world work score
`educ`	number of years of education
`exper`	years of work experience
`tenure`	years with current employer
`age`	age in years
`married`	=1 if married
`black`	=1 if black
`south`	=1 if live in south
`urban`	=1 if live in a Standard Metropolitan Statistical Area
`sibs`	number of siblings
`brthord`	birth order
`meduc`	mother's education (years)
`feduc`	father's education (years)
`lwage`	natural log of `wage`

Is this an observational study or an experiment?

Observational study

Experiment

Exploring the data

As with any new data set a good place to start is standard exploratory data analysis. We will begin with the `wage` variable since it will be the response variable in our models.

Which of the following statements is **false** about the distribution of `wage` ?

The median of the distribution is 905.

25% of respondents make more than 1160 dollars per week.

7 of the respondents make less than 300 dollars per week

`wage` is right-skewed, meaning that more respondents fall below the mean wage than above it.

```
# type your code for Question 2 here, and Knit
```

Since `wage` is our response, we would like to explore the relationship of the other variables as predictors.

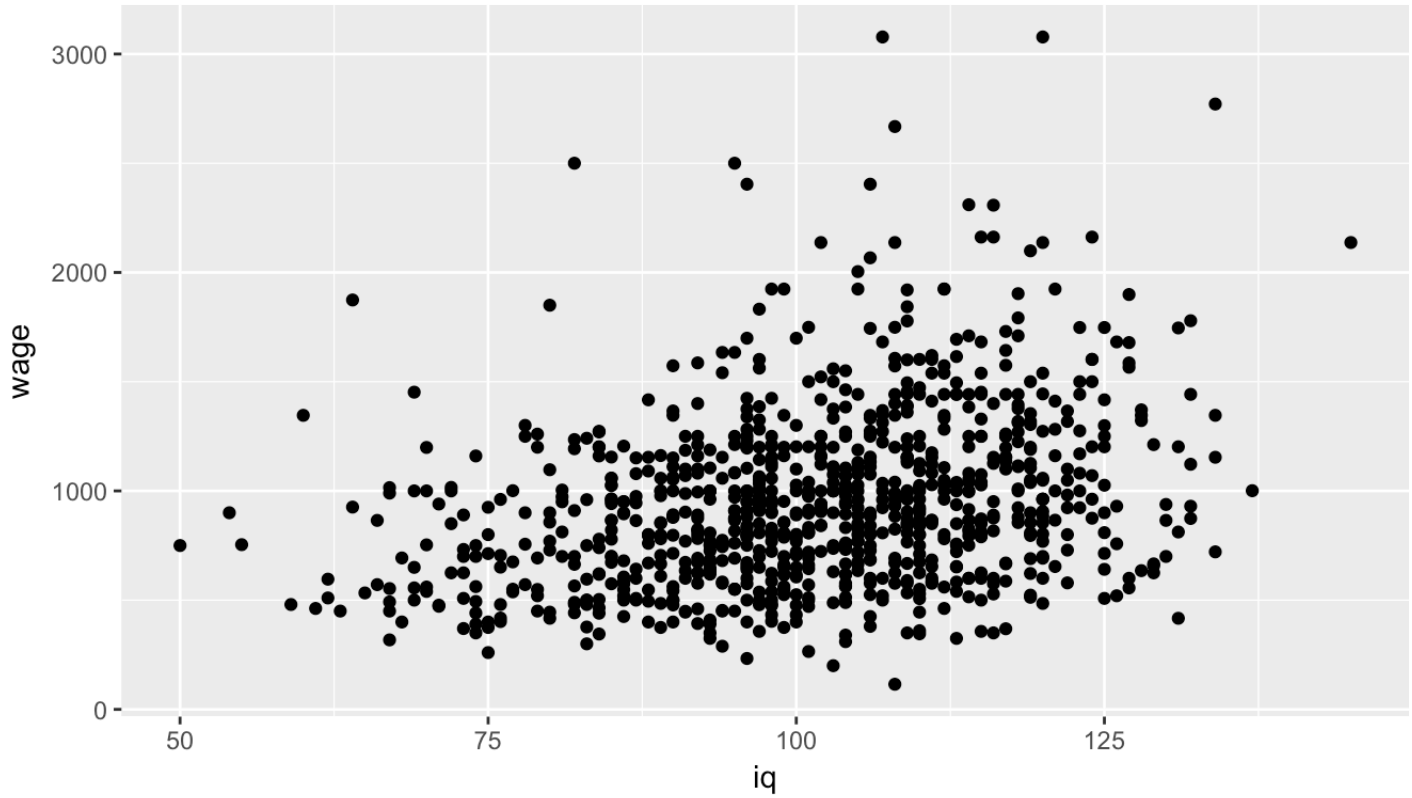
Exercise: Excluding `wage` and `lwage`, select two other variables that you think might be a good predictor of `wage`. Visualize their relationships with `wage` using appropriate plots.

```
# type your code for the Exercise here, and Knit
```

Simple linear regression

One possible, simplistic, explanation for the variation in wages that we see in the data is that smarter people make more money. The plot below visualizes a scatterplot between weekly wage and IQ score.

```
ggplot(data = wage, aes(x = iq, y = wage)) +
  geom_point()
```



This plot is rather noisy. While there may be a slight positive linear relationship between IQ score and wage, IQ is at best a crude predictor of wages. We can quantify this by fitting a simple linear regression.

```
m_wage_iq = lm(wage ~ iq, data = wage)
m_wage_iq$coefficients
```

```
## (Intercept)          iq
##  116.991565    8.303064
```

```
summary(m_wage_iq)$sigma
```

```
## [1] 384.7667
```

Recall from the lectures that under the model

$$wage_i = \alpha + \beta \cdot iq_i + \epsilon_i$$

if $\epsilon_i \sim N(0, \sigma^2)$ and the reference prior $p(\alpha, \beta, \sigma^2) \propto 1/\sigma^2$ is used, then the Bayesian posterior means and standard deviations will be equal to the frequentist estimates and standard errors respectively.

The Bayesian model specification assumes that the errors are normally distributed with a constant variance. As with the frequentist approach we check this assumption by examining the distribution of the residuals for the model. If the residuals are highly non-normal or skewed, the assumption is violated and any subsequent inference is not valid.

Examine the residuals of `m_wage_iq`. Is the assumption of normally distributed errors valid?

- Yes, since the distribution of the dependent variable (wage) is roughly normally distributed.
- Yes, since the distribution of the residuals of the model looks approximately normal.
- No, since the distribution of the residuals of the model is left-skewed.
- No, since the distribution of the residuals of the model is right-skewed.

```
# type your code for Question 3 here, and Knit
```

Exercise: Refit the model, this time using `educ` (education) as the independent variable. Does your answer to the previous exercise change?

```
# type your code for the Exercise here, and Knit
```

Variable transformation

One way to accommodate the right-skewness in the data is to (natural) log transform the dependent variable.

Note that this is only possible if the variable is strictly positive, since the log of negative value is not defined and $\log(0) = -\infty$. Let's try to fit a linear model with log-wage as the dependent variable. Question 4 will be based on this log transformed model.

```
m_lwage_iq = lm(lwage ~ iq, data = wage)
```

Exercise: Examine the residuals of this model. Is the assumption of normally distributed residuals reasonable?

```
# type your code for the Exercise here, and Knit
```

Recall that the posterior distribution of α and β given σ^2 is normal, but marginally follows a t distribution with $n - p - 1$ degrees of freedom. In this case, $p = 1$, since IQ is the only predictor of log-wage included in our model. Therefore both α and β will have a posteriors that follow a t distribution 933 degrees of freedom - since the df is so large these distributions will actually be approximately normal.

Under the reference prior $p(\alpha, \beta, \sigma^2) \propto 1/\sigma^2$, give a 95% posterior credible interval for β, the coefficient of IQ.

(0.00793, 0.00967)

(0.00709, 0.01050)

(0.00663, 0.01098)

(0.00010, 0.01750)

```
# type your code for Question 4 here, and Knit
```

Exercise: The coefficient of IQ is very small, which is expected since a one point increase in IQ score can hardly be expected to have a high multiplicative effect on wage. One way to make the coefficient more interpretable is to standardize IQ before putting it into the model. From this new model, an increase in IQ of 1 standard deviation (15 points) is estimated to increase wage by what percentage?

```
# type your code for the Exercise here, and Knit
```

Multiple linear regression

It is evident that wage can be explained by many predictors, such as experience, education, and IQ. We can include all relevant covariates in a regression model in an attempt to explain as much wage variation as possible.

```
m_lwage_full = lm(lwage ~ . - wage, data = wage)
```

The use of `.` in the `lm` tells R to include all covariates in the model which we then further modify with `-wage` which then excludes the `wage` variable from the model.

However, running this full model has a cost: we remove observations from our data since some measurements for (e.g. birth order, mother's education, and father's education) are missing. By default, the `lm` function does a complete-case analysis, and so it removes any observations with a missing (`NA`) value in one or more of the predictor variables.

Because of these missing values we must make an addition assumption in order for our inferences to be valid. This exclusion of rows with missing values requires that the data there is no systematic reason for the values to be missing, or in other words our data must be missing at random. For example, if all first-born children did not report their birth order, the data would not be missing at random. Without any additional information we will assume this is reasonable and use the 663 complete observations (as opposed to the original 935) to fit the model. Both Bayesian and frequentist methods exist to handle data sets with missing data, but they are beyond the scope of this course.

From the model, all else begin equal, who would you expect to make more: a married black man or a single non-black man?

The married black man

The single non-black man

```
# type your code for Question 5 here, and Knit
```

As you can see from a quick summary of the full linear model, many coefficients of independent variables are not statistically significant. In previous labs within this specialization, you selected variables based on Adjusted R^2. This module introduced the Bayesian Information Criterion (BIC), which is a metric that can be used for model selection. BIC is based on model fit, while simultaneously penalizing the number of parameters in proportion to the sample size. We can calculate the BIC of the full linear model using the command below:

```
BIC(m_lwage_full)
```

```
## [1] 586.3732
```

We can compare the BIC of the full model with that of a reduced model. Let's try to remove birth order from the model. To ensure that the observations remain the same, the data set can be specified as `na.omit(wage)`, which includes only the observations with no missing values.

```
m_lwage_nobrthord = lm(lwage ~ . -wage -brthord, data = na.omit(wage))
BIC(m_lwage_nobrthord)
```

```
## [1] 582.4815
```

As you can see, removing birth order from the regression reduces BIC, which we seek to minimize by model selection.

Elimination of which variable from the full model yielded the lowest BIC?

`brthord`

`sibs`

`feduc`

`meduc`

```
# type your code for Question 6 here, and Knit
```

Exercise: R has a function `stepAIC` that will work backwards through the model space, removing variables until BIC can be no longer be lowered. It takes as inputs a full model, and a penalty parameter k. Find the best model according to BIC (in which case $k = \log(n)$). Remember to use `na.omit(wage)` as your data set.

```
# type your code for the Exercise here, and Knit
```

Bayesian model averaging

Often, several models are equally plausible and choosing only one ignores the inherent uncertainty involved in choosing the variables to include in the model. A way to get around this problem is to implement Bayesian model averaging (BMA), in which multiple models are averaged to obtain posteriors of coefficients and predictions from new data. Dr. Merlise Clyde is the primary author of the R package `BAS` , which implements BMA. We can use this for either implementing BMA or selecting models. We start by applying BMA to the wage data.

```
wage_no_na = na.omit(wage)
bma_lwage = bas.lm(lwage ~ . -wage, data = wage_no_na,
                   prior = "BIC",
                   modelprior = uniform())
bma_lwage
```

```
##
## Call:
## bas.lm(formula = lwage ~ . - wage, data = wage_no_na, prior = "BIC",
##     modelprior = uniform())
##
##
##  Marginal Posterior Inclusion Probabilities:
## Intercept      hours         iq        kww       educ      exper
##   1.00000    0.85540    0.89732    0.34790    0.99887    0.70999
##    tenure        age   married1     black1     south1     urban1
##   0.70389    0.52468    0.99894    0.34636    0.32029    1.00000
##      sibs    brthord      meduc      feduc
##   0.04152    0.12241    0.57339    0.23274
```

```
summary(bma_lwage)
```

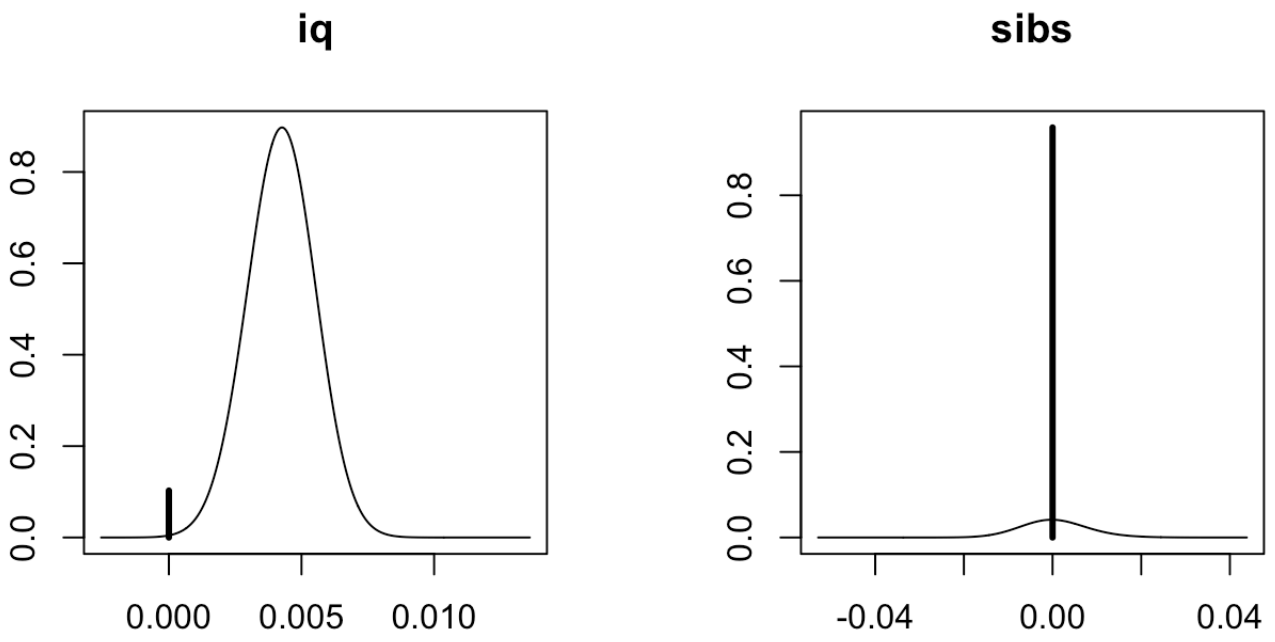
```
##           P(B != 0 | Y)    model 1       model 2       model 3
## Intercept    1.00000000     1.0000     1.0000000     1.0000000
## hours        0.85540453     1.0000     1.0000000     1.0000000
## iq           0.89732383     1.0000     1.0000000     1.0000000
## kww          0.34789688     0.0000     0.0000000     0.0000000
## educ         0.99887165     1.0000     1.0000000     1.0000000
## exper        0.70999255     0.0000     1.0000000     1.0000000
## tenure       0.70388781     1.0000     1.0000000     1.0000000
## age          0.52467710     1.0000     1.0000000     0.0000000
## married1     0.99894488     1.0000     1.0000000     1.0000000
## black1       0.34636467     0.0000     0.0000000     0.0000000
## south1       0.32028825     0.0000     0.0000000     0.0000000
## urban1       0.99999983     1.0000     1.0000000     1.0000000
## sibs         0.04152242     0.0000     0.0000000     0.0000000
## brthord      0.12241286     0.0000     0.0000000     0.0000000
## meduc        0.57339302     1.0000     1.0000000     1.0000000
## feduc        0.23274084     0.0000     0.0000000     0.0000000
## BF                   NA     1.0000     0.5219483     0.5182769
## PostProbs            NA     0.0455     0.0237000     0.0236000
## R2                   NA     0.2710     0.2767000     0.2696000
## dim                  NA     9.0000    10.0000000     9.0000000
## logmarg              NA -1490.0530 -1490.7032349 -1490.7102938
##                 model 4       model 5
## Intercept     1.0000000     1.0000000
## hours         1.0000000     1.0000000
## iq            1.0000000     1.0000000
## kww           1.0000000     0.0000000
## educ          1.0000000     1.0000000
## exper         1.0000000     0.0000000
## tenure        1.0000000     1.0000000
## age           0.0000000     1.0000000
## married1      1.0000000     1.0000000
## black1        0.0000000     1.0000000
## south1        0.0000000     0.0000000
## urban1        1.0000000     1.0000000
## sibs          0.0000000     0.0000000
## brthord       0.0000000     0.0000000
## meduc         1.0000000     1.0000000
## feduc         0.0000000     0.0000000
## BF            0.4414346     0.4126565
## PostProbs     0.0201000     0.0188000
## R2            0.2763000     0.2762000
## dim          10.0000000    10.0000000
## logmarg   -1490.8707736 -1490.9381880
```

Printing the model object and the summary command gives us both the posterior model inclusion probability for each variable and the most probable models. For example, the posterior probability that `hours` is included in the model is 0.855. Further, the most likely model, which has posterior probability of 0.0455,

includes an intercept, hours worked, IQ, education, tenure, age, marital status, urban living status, and mother's education. While a posterior probability of 0.0455 sounds small, it is much larger than the uniform prior probability assigned to it, since there are 2^{16} possible models.

It is also possible to visualize the posterior distribution of the coefficients under the model averaging approach. We graph the posterior distribution of the coefficients of `iq` and `sibs` below. Note that the subset command dictates which variable is plotted.

```
par(mfrow = c(1,2))
coef_lwage = coefficients(bma_lwage)
plot(coef_lwage, subset = c(3,13), ask=FALSE)
```



We can also provide 95% credible intervals for these coefficients:

```
confint(coef_lwage)
```

```
##                    2.5%        97.5%           beta
## Intercept  6.786258e+00 6.8401985740  6.8142970694
## hours     -9.322389e-03 0.0000000000 -0.0053079979
## iq         0.000000e+00 0.0062614447  0.0037983313
## kww        0.000000e+00 0.0082697580  0.0019605787
## educ       2.323135e-02 0.0666346854  0.0440707549
## exper      0.000000e+00 0.0210643255  0.0100264057
## tenure     0.000000e+00 0.0128809329  0.0059357058
## age        0.000000e+00 0.0253737564  0.0089659753
## married1   1.190280e-01 0.3011016699  0.2092940731
## black1    -1.914915e-01 0.0001255892 -0.0441863361
## south1    -1.039870e-01 0.0000000000 -0.0221757978
## urban1     1.338978e-01 0.2586975478  0.1981221313
## sibs       0.000000e+00 0.0000000000  0.0000218455
## brthord   -2.007178e-02 0.0000000000 -0.0019470674
## meduc      0.000000e+00 0.0226937234  0.0086717156
## feduc     -6.810365e-06 0.0157018490  0.0025125930
## attr(,"Probability")
## [1] 0.95
## attr(,"class")
## [1] "confint.bas"
```

For questions 7-8, we'll use a reduced data set which excludes number of siblings, birth order, and parental education.

```
wage_red = wage %>%
  select(-sibs, -brthord, -meduc, -feduc)
```

Based on this reduced data set, according to Bayesian model averaging, which of the following variables has the lowest marginal posterior inclusion probability?

kww

black

south

age

```
# type your code for Question 7 here, and Knit
```

True or False: The naive model with all variables included has posterior probability greater than 0.5. (Use a Zellner-Siow null prior for the coefficients and a Beta-Binomial (1,1) prior for the models)

True

False

```
# type your code for Question 8 here, and Knit
```

Exercise: Graph the posterior distribution of the coefficient of `age` , using the data set `wage_red` .

```
par(mfrow = c(1,1))
# type your code for the Exercise here, and Knit
```

Prediction

A key advantage of Bayesian statistics is prediction and the probabilistic interpretation of predictions. Much of Bayesian prediction is done using simulation techniques, some of which was discussed near the end of this module. This is often applied in regression modeling, although we'll work through an example with just an intercept term.

Suppose you observe four numerical observations of y, which are 2, 2, 0 and 0 respectively with sample mean $\bar{y} = 1$ and sample variance $s^2 = 4/3$. Assuming that $y \sim N(\mu, \sigma^2)$, under the reference prior $p(\mu, \sigma^2) \propto 1/\sigma^2$, our posterior becomes

$$\mu|\sigma^2, y \sim N(1, \sigma^2/4)$$

which is centered at the sample mean and

$$1/\sigma^2, y \sim Gamma(\alpha = 3/2, \beta = 4/2)$$

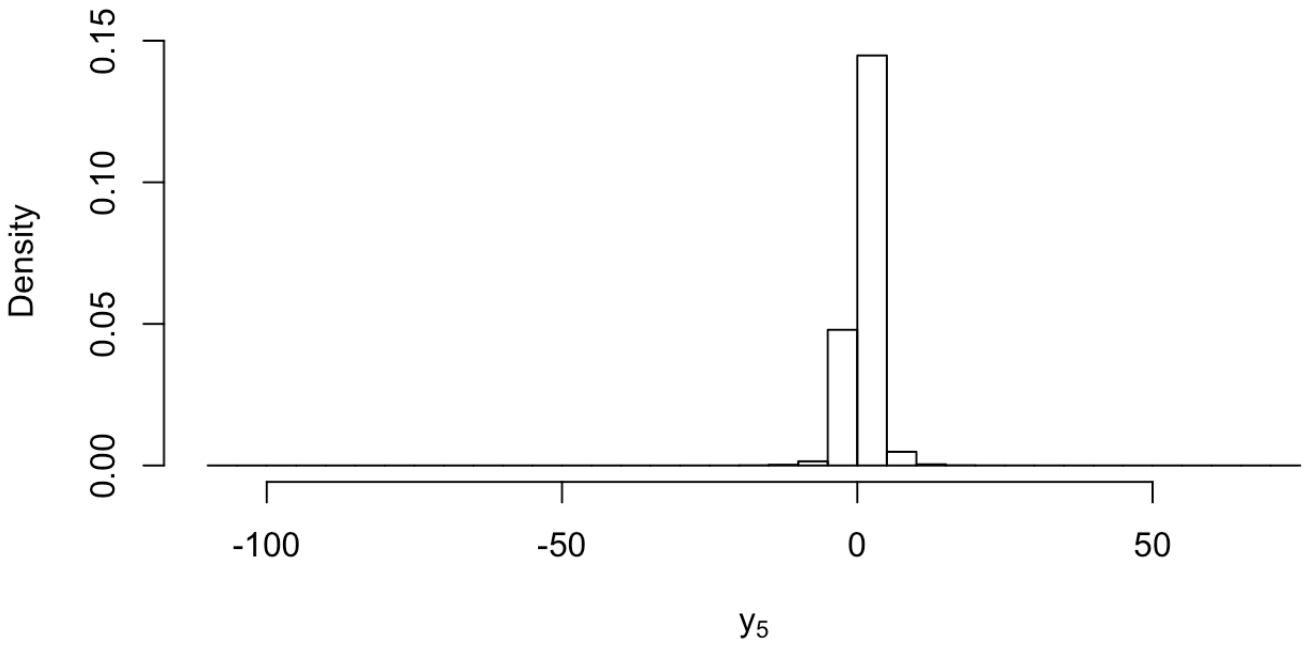
where $\alpha = (n-1)/2$ and $\beta = s^2(n-1)/2 = 2.$

To obtain the predictive distribution for y_5 , we can first simulate σ^2 from its posterior and then μ followed by y_5 . Our draws of y_5 will be from the posterior predictive distribution for a new observation. The example below draws 100,000 times from the posterior predictive distribution of y_5 .

```
set.seed(314)
N = 100000
phi = rgamma(N,3/2,2)
sigma2 = 1/phi
mu = rnorm(N, 1, sqrt(sigma2/4))
y_5 = rnorm(N, mu, sqrt(sigma2))
```

We can view an estimate of the predictive distribution, by looking at the a smoothed version of the histogram of the simulated data:

```
hist(y_5, prob=T, breaks=30, xlab=expression(y[5]), main="")
```



A 95% central credible interval for a new observation is the interval (L, U) where $P(Y_5 < L \mid Y) = .05/2$ and $P(Y_5 > U \mid Y) = .05/2)$. In this case L is the 0.025 quantile and U is the 0.975 quantile. We can obtain those values using the `quantile` function to find the sample quantiles for 0.025 and 0.975 of y_5 .

Estimate a 95% central credible interval for a new observation y_5

(-3.71, 5.73)

(-3.11, 5.13)

(-1.18, 3.19)

```
# type your code for Question 9 here, and Knit
```

Exercise: In the simple example above, it is possible to use integration to calculate the posterior predictive analytically. In this case, it is a scaled t distribution with 3 degrees of freedom $(n - 1)$ with mean 1 and scale = 5/3 ($s^2(1 + 1/n)$). Plot the empirical density of y alongside the actual density of the t-distribution. How do they compare?

```
# type your code for the Exercise here, and Knit
```

Prediction with BAS

Simulation is used in `BAS` to construct predictive intervals with Bayesian Model averaging, while exact inference is often possible with predictive intervals under model selection.

Returning to the wage data set, let's find predictive values under the best predictive model, the one that has predictions closest to BMA and corresponding posterior standard deviations.

```
BPM_pred_lwage =  predict(bma_lwage, estimator="BPM", se.fit=TRUE)
bma_lwage$namesx[BPM_pred_lwage$bestmodel+1]
```

```
##  [1] "Intercept" "hours"     "iq"        "kww"       "educ"
##  [6] "exper"     "tenure"    "age"       "married1"  "urban1"
## [11] "meduc"
```

We can compare this to the Highest probability model that we found earlier and the Median Probability Model (`MPM`)

```
MPM_pred_lwage =  predict(bma_lwage, estimator="MPM")
bma_lwage$namesx[MPM_pred_lwage$bestmodel+1]
```

```
##  [1] "Intercept" "hours"     "iq"        "educ"      "exper"
##  [6] "tenure"    "age"       "married1"  "urban1"    "meduc"
```

The `MPM` includes `exper` in addition to all of the variables as the `HPM` , while the `BPM` includes `kwh` in addition to all of the variables in the `MPM` .

Exercise: Using the reduced data, what covariates are included in the best predictive model, the median probability model and the highest posterior probability model?

Let's turn to see what characteristics lead to the highest wages with the `BPM` model.

```
opt = which.max(BPM_pred_lwage$fit)
t(wage_no_na[opt, ])
```

```
##         [,1]
## wage    "1586"
## hours   "40"
## iq      "127"
## kww     "48"
## educ    "16"
## exper   "16"
## tenure  "12"
## age     "37"
## married "1"
## black   "0"
## south   "0"
## urban   "1"
## sibs    "4"
## brthord "4"
## meduc   "16"
## feduc   "16"
## lwage   "7.36897"
```

A 95% credible interval for predicting log wages can be obtained by

```
ci_lwage = confint(BPM_pred_lwage, parm="pred")
ci_lwage[opt,]
```

```
##     2.5%    97.5%     pred
## 6.661871 8.056450 7.359160
```

To translated back to wages, we may exponentiate the interval

```
exp(ci_lwage[opt,])
```

```
##      2.5%     97.5%      pred
##  782.0124 3154.0718 1570.5169
```

to obtain a 95% prediction interval for the wages of an individual with covariates at the levels of the individual specified by `opt`.

If were to use BMA, the interval would be

```
BMA_pred_lwage =  predict(bma_lwage, estimator="BMA", se.fit=TRUE)
ci_bma_lwage = confint(BMA_pred_lwage, estimator="BMA")
opt_bma = which.max(BMA_pred_lwage$fit)
exp(ci_bma_lwage[opt_bma,])
```

```
##      2.5%     97.5%      pred
##  733.3446 2989.2076 1494.9899
```

Exercise: Using the reduced data, construct a 95% prediction interval for the individual who is predicted to have the highest predicted wages under the `BPM`.

References

Wooldridge, Jeffrey. 2000. *Introductory Econometrics- A Modern Approach*. South-Western College Publishing. http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.dta (http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.dta).