Complete all **Exercises**, and submit answers to **Questions** in the **Quiz: Week 4 of Lab** on Coursera.

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 selection methods such as BIC and Bayesian Model Averaging to construct parsimonious predictive models.

Getting Started

In this lab we will explore the data using the `dplyr` package and visualize it using the `ggplot2` package for data visualization. Both of these packages are part of the tidyverse. We will review simple linear regression using the `lm` function and how the output can be interpreted from a Bayesian perspective. We will also use the `broom` package to turn regression outputs to tidy data frames to help with diagnostic plotting. We will use the `stepAIC` function from the `MASS` package for model selection using step-wise selection using BIC. The `bas.lm` function from the `BAS` package later in the lab to implement Bayesian Model Averaging. Please make sure that the version of `BAS` is 1.4.9 or greater. The data can be found in the companion package for this course, `statsr`. Some learners may want to review material from the earlier courses in the specialization that covers EDA and regression if they are unfamiliar with `ggplot` basics or the `lm` function.

Load packages

Let's load the packages that we will be using:

```
library(MASS)
library(tidyverse)
library(statsr)
library(BAS)
library(broom)

options(width=100)
```

The data

The data we 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 [@Wooldridge2000].

Let's start by loading 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

variable	description
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? You may refer to http://study.com/academy/lesson/experiments-vs-observational-studies.html (http://study.com/academy/lesson/experiments-vs-observational-studies.html) for the definitions of the two.

- Observational study
- Experiment

Setting a seed

In this lab we will do some random generation, which means you should set a seed on top of your document. Setting a seed will cause R to sample the same sample each time you knit your document. This will make sure your results don't change each time you knit, and it will also ensure reproducibility of your work (by setting the same seed it will be possible to reproduce your results). You can set a seed like this:

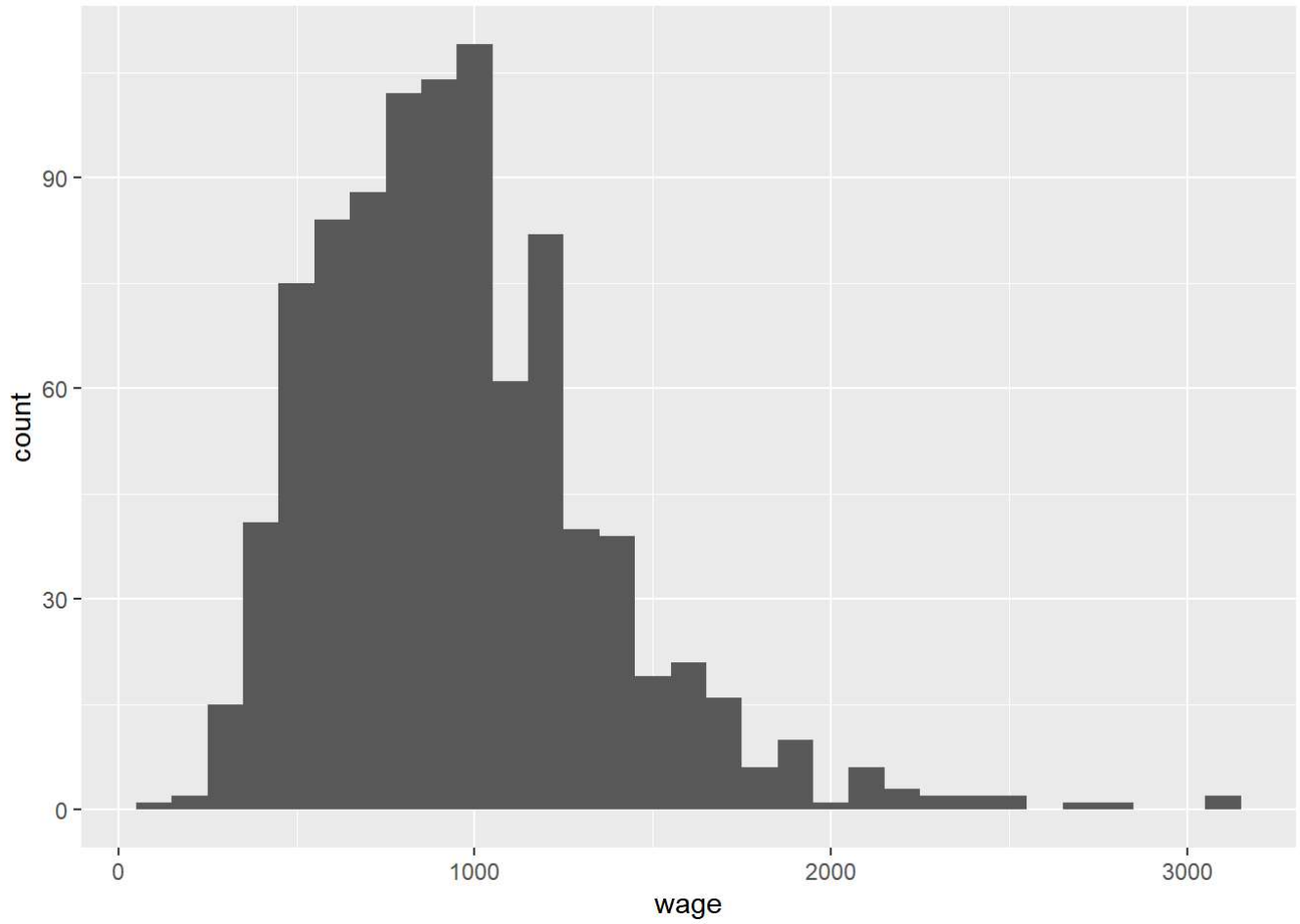
```
set.seed(18382)
```

The number above is completely arbitraty. If you need inspiration, you can use your ID, birthday, or just a random string of numbers. The important thing is that you use each seed only once. You only need to do this once in your R Markdown document, but make sure it comes before sampling.

Exploring the data

For a 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. We may use a histogram to visualize the distribution.

```
ggplot(data = wage, aes(x = wage)) +
  geom_histogram(binwidth = 100)
```



For numeric summary statistics, the `summary` function provides additional insights about the distribution of `wage`.

```
summary(wage$wage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     115     669     905     958    1160    3078
```

```
nrow(wage[wage$wage < 300,])
```

```
## [1] 6
```

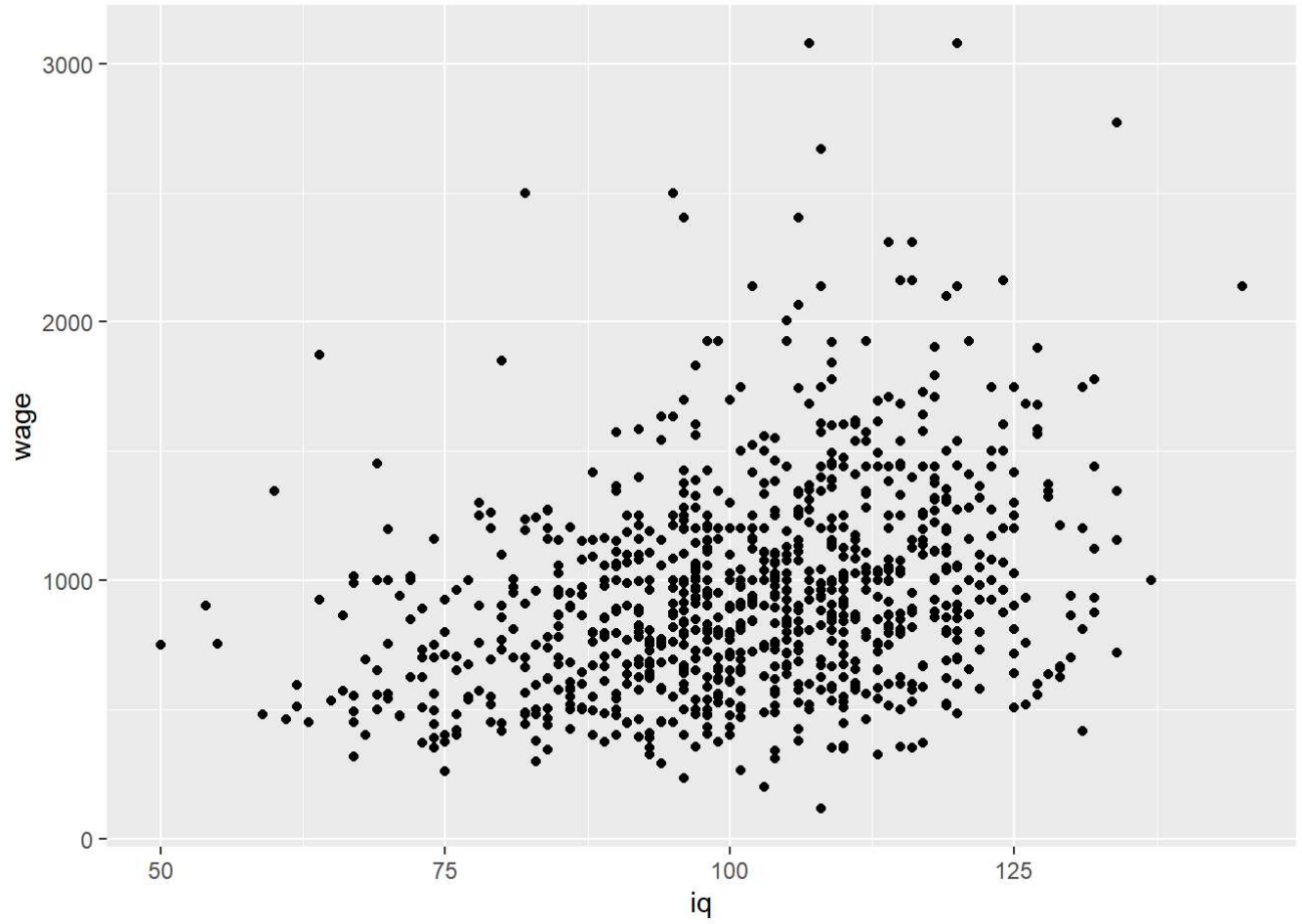
Which of the following statements is **false** about the distribution of weekly wages?

- The median of the distribution is 905.
- 25% of respondents make at least 1160 dollars per week.
- 10 of the respondents make strictly less than 300 dollars per week
- `wage` is right-skewed, meaning that more respondents have weekly wages below the mean weekly wage than above it.

Simple linear regression

Since `wage` is our response variable, we would like to explore the relationship between `wage` and other variables as predictors. 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()
```



There appears to be a positive relationship between IQ score and wage. We can quantify this by fitting a Bayesian simple linear regression

$$\text{wage}_i = \alpha + \beta \cdot \text{iq}_i + \epsilon_i$$

to the observed data using the reference prior. We can fit the model using the `lm` function:

```
m_wage_iq <- lm(wage ~ iq, data = wage)
```

and extract the summary statistics for the posterior distribution using the output from the `lm` by applyting the `tidy` function from the `broom` package.

```
tidy(m_wage_iq)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   117.      85.6        1.37 1.72e- 1
## 2 iq              8.30     0.836      9.93 3.79e-22
```

The first column displays the posterior means of the linear model's y-intercept and the regression coefficient of `iq`.

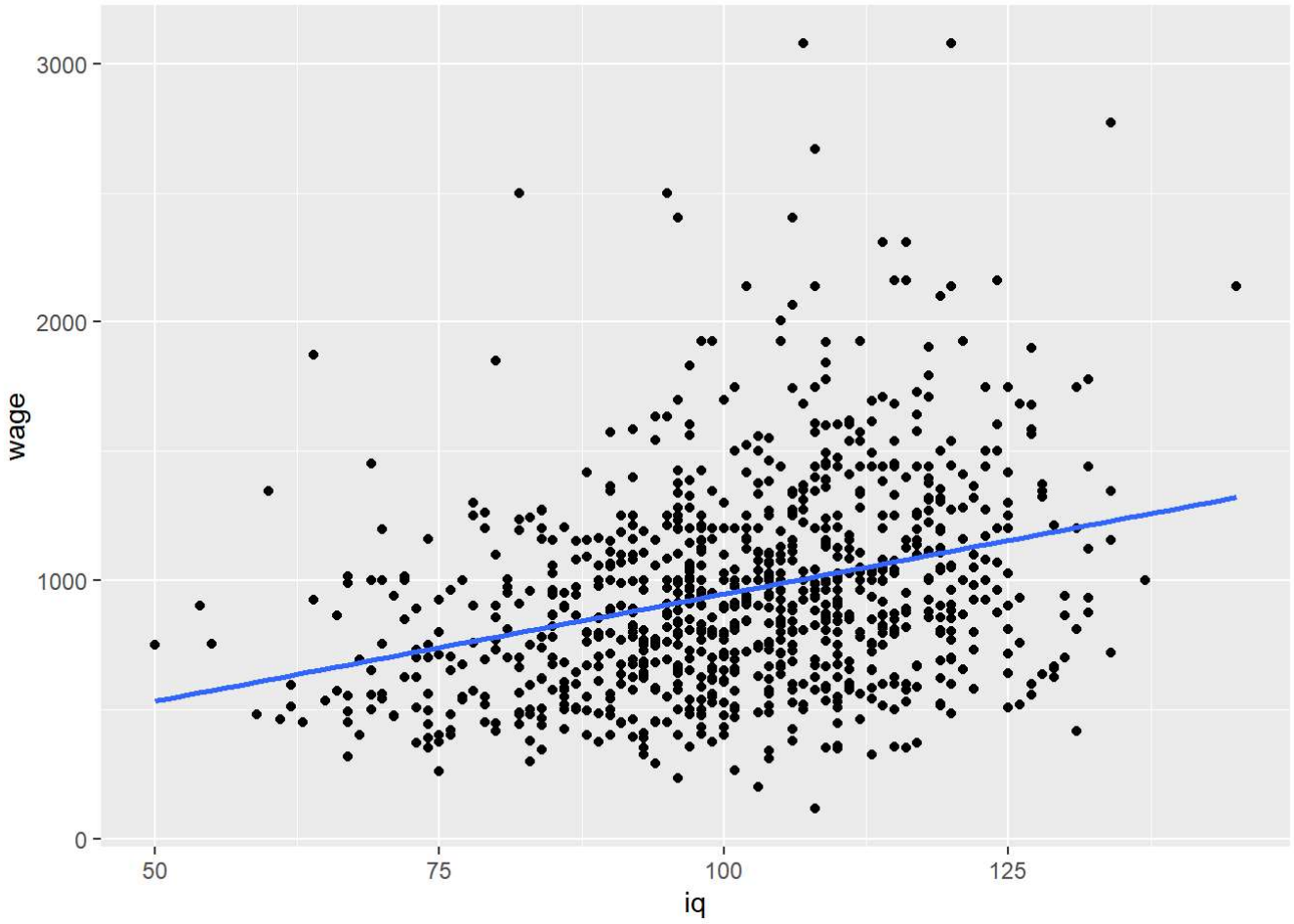
With this we can write down the posterior mean of the regression line

$$116.992 + 8.303 \times \text{IQ}$$

and create a scatterplot with the posterior mean for the regression line laid on top.

```
ggplot(data = wage, aes(x = iq, y = wage)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Under the assumption that the errors ϵ_i are independent and normally distributed with mean zero and an unknown variance σ^2, the posterior distributions for the intercept and slope will have a Student-t distribution under the reference prior with the posterior means and scales equal to the ordinary least squares estimates and standard errors respectively. We can create 95% credible intervals for the two parameters using the `confint` function:

```
confint(m_wage_iq)
```

```
##                2.5 %  97.5 %
## (Intercept) -51.081 285.064
## iq             6.662   9.944
```

Fit a new model that uses `educ` (education) to predict average weekly wages. Using the estimates from the R output, write the equation of the posterior mean of the regression line and obtain a 95% credible interval for the coefficients. What does the slope tell us in the context of the relationship between education and earnings?

- Each additional year of education increases weekly wages by $60.21.

- Each additional year of education increases weekly wages by $146.95.
- For each additional year of education, there is a 95% chance that average weekly wages will possibly decrease by $5.56 or increase by $299.47.
- For each additional year of education, there is a 95% chance that average weekly wages will increase by $49.04 to $71.39.

```
# Type your code for Question 3 here.
m_wage_educ <- lm(wage ~ educ, data = wage)
summary(m_wage_educ)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = wage)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -877.4 -268.6  -38.4  207.0 2148.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   146.95      77.71    1.89    0.059 .
## educ           60.21       5.69   10.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 382 on 933 degrees of freedom
## Multiple R-squared:  0.107,  Adjusted R-squared:  0.106
## F-statistic:  112 on 1 and 933 DF,  p-value: <2e-16
```

```
confint(m_wage_educ)
```

```
##              2.5 % 97.5 %
## (Intercept) -5.564 299.47
## educ        49.038  71.39
```

Model diagnostics

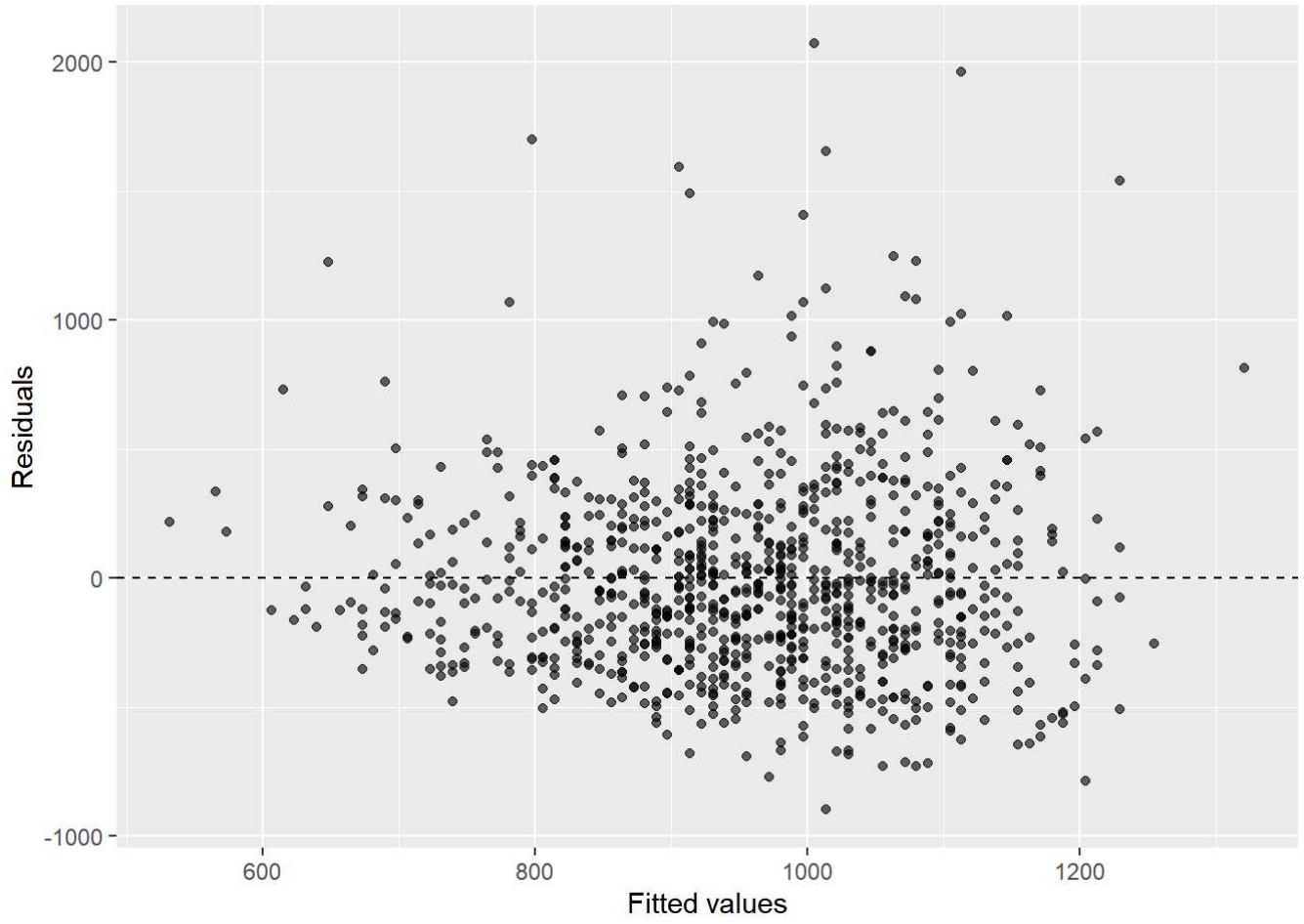
The Bayesian model specification assumes that the errors are normally distributed with a constant variance and that the mean expected weekly wages is linear in IQ. We can check these assumption by examining the distribution of the residuals for the model.

In order to do so we will use predicted values, residuals, and standardized residuals of the model we fit earlier. The `augment` function in the `broom` package is going to come in handy here as it takes in a model object (the output of an `lm`) and returns a data frame with columns correspinding to variables in the model as well as predicted values (`.fitted`), residuals (`.resid`), and standardized residuals (`.std.resid`), along with a few others.

```
m_wage_iq_aug <- augment(m_wage_iq)
```

Linearity and Constant Variance: You already checked if the relationship between weekly wages and IQ is linear using a scatterplot. We should also verify this condition with a plot of the residuals vs. fitted (predicted) values.

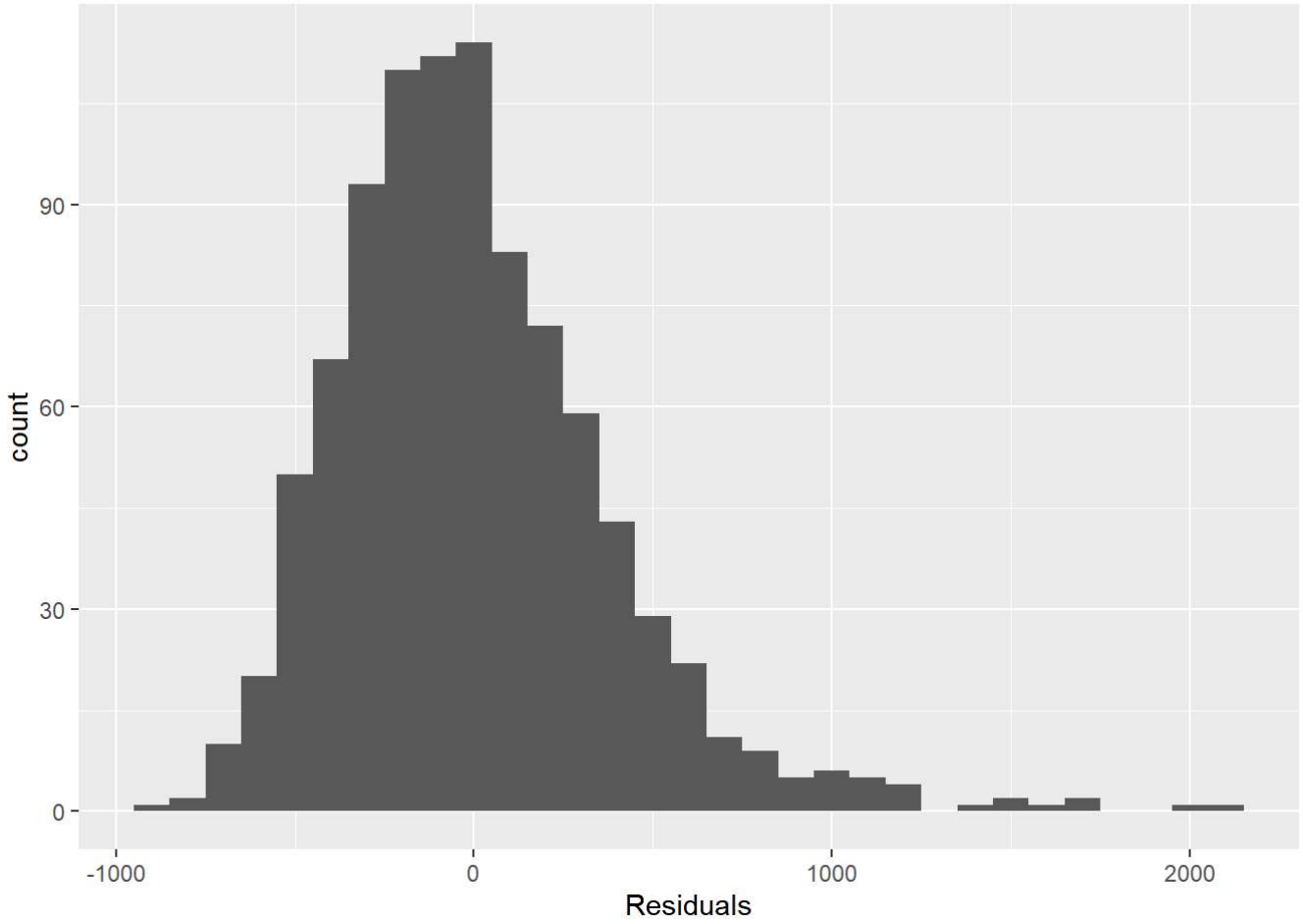
```
ggplot(data = m_wage_iq_aug, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted values", y = "Residuals")
```



Also note that we're getting fancy with the code here. We set the `alpha` level of our points to a value lower than 1 (`0.6` to be precise) in order to add plot the points with some transparency. This will allow us to more easily identify where the points are more dense vs. more sparse. Then, we overlay a horizontal dashed line at $y = 0$ (to help us check whether residuals are distributed evenly around 0 at each fitted value), and we also adjust the axis labels to be more informative.

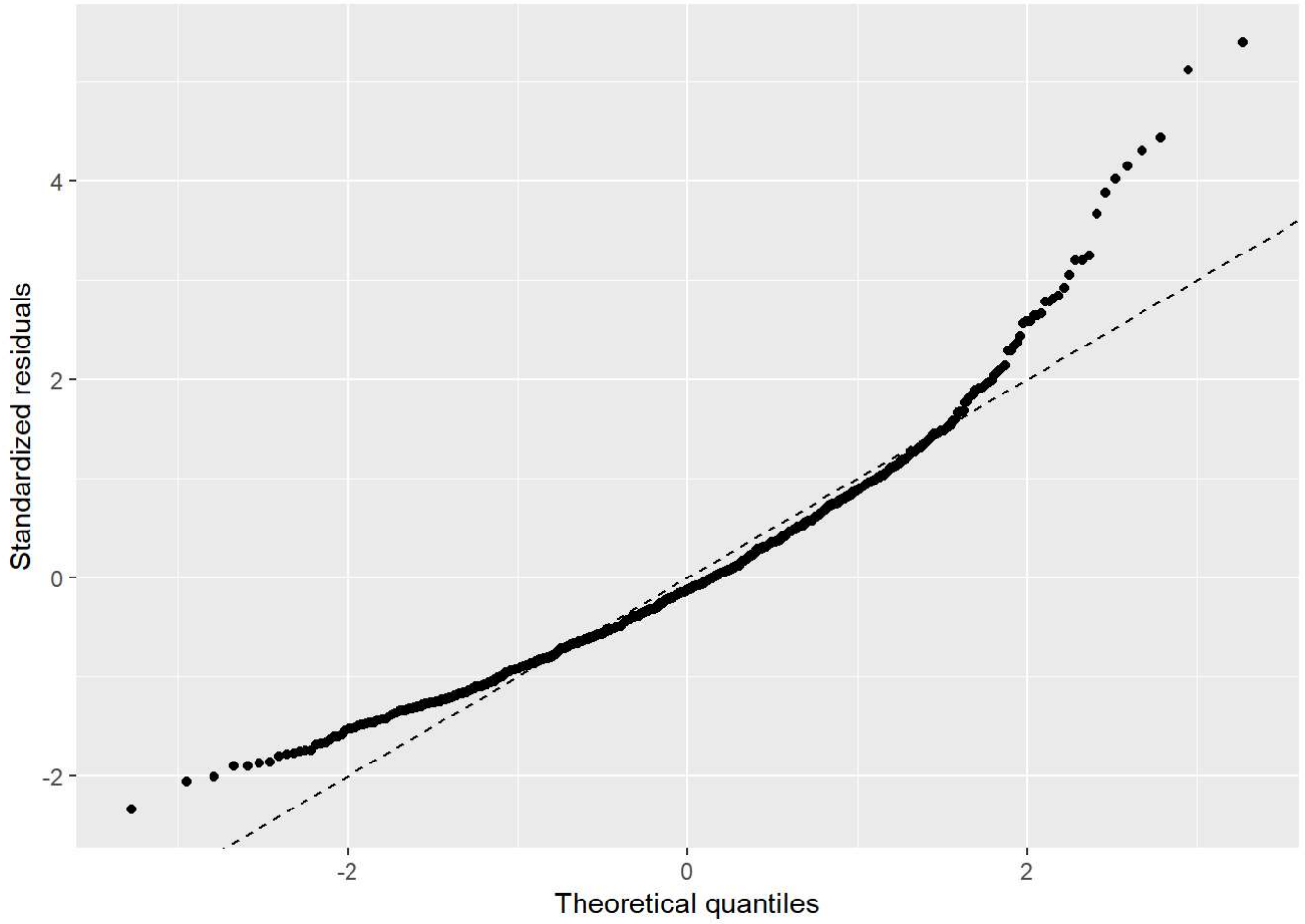
Normality: To check this condition, we can look at a histogram of residuals

```
ggplot(data = m_wage_iq_aug, aes(x = .resid)) +
  geom_histogram(binwidth = 100) +
  xlab("Residuals")
```



or a normal probability plot of the residuals

```
ggplot(m_wage_iq_aug) +
  geom_qq(aes(sample = .std.resid)) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  labs(x = "Theoretical quantiles", y = "Standardized residuals")
```



where we expect the points to be close to the dashed line, if the assumption of normality holds. Note that the y-axis in the plot uses standardized residuals, which are the residuals divided by their standard deviations so that they will have a normal distribution with mean zero and constant variance if the model holds.

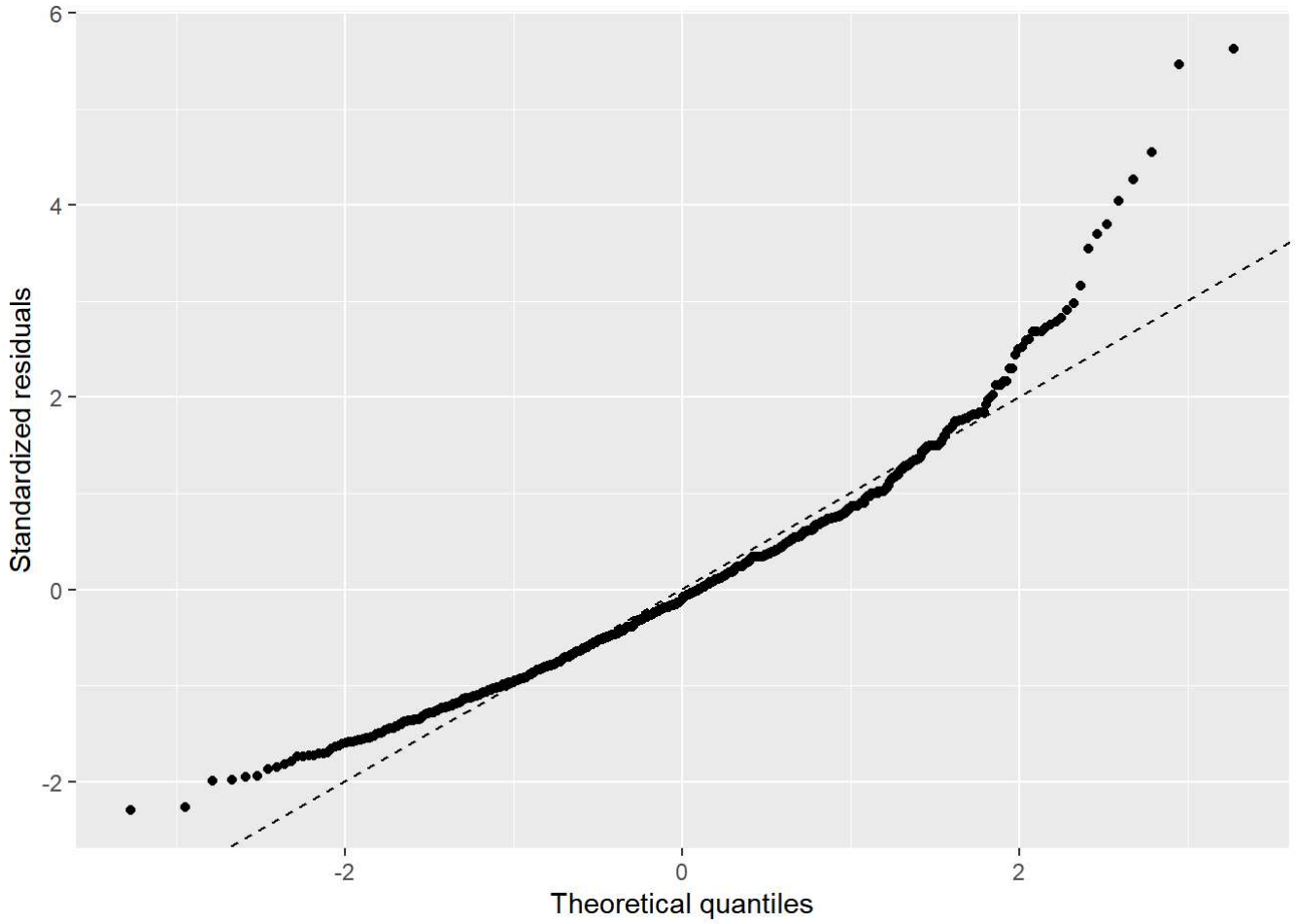
Which of the following statements about the residual plots are **false**?

- The residuals appear to be randomly distributed around 0.
- The residuals are strongly left skewed, hence the normal distribution of errors condition is not met.
- The variability of residuals appears to increase as the fitted increase, suggesting that the constant variance assumption does not hold.
- There are more individuals where the model under predicts weekly wages rather than over estimates weekly wages.

```
# Type your code for Question 4 here.
```

Refit the model by using `educ` (education) as the independent variable. Does your answer to the previous exercise change?

```
# Type your code for Exercise 1 here.
m_wage_educ_aug <- augment(m_wage_educ)
ggplot(m_wage_educ_aug) +
  geom_qq(aes(sample = .std.resid)) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  labs(x = "Theoretical quantiles", y = "Standardized residuals")
```



Linear Regression After Transforming `wage`

One way to accommodate the right-skewness in the residuals 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 $\ln(0) = -\infty$. Let us try to fit a linear model with log-wage (`lwage`) as the dependent variable. The next two questions will be based on this log transformed model.

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

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

```
# Type your code for Exercise 2 here.
```

Excluding `wage` and `lwage` , select two other variables that you think might be good predictors of `lwage` . Visualize their relationships with `wage` and check assumptions using appropriate plots.

```
# Type your code for  Exercise 3 here.
```

Outliers

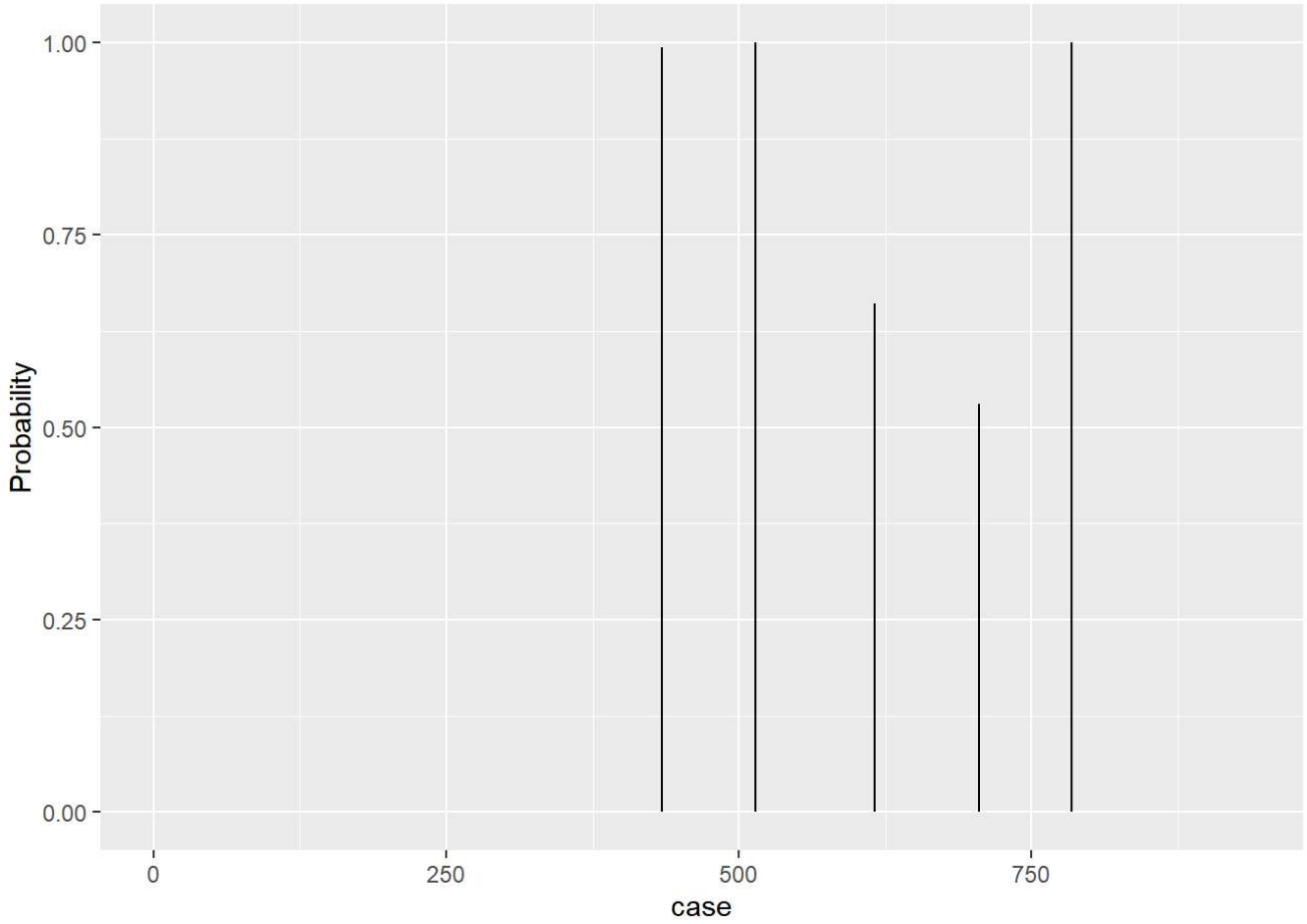
We declared observations to be outliers with respect to the population model if their deviation or error ϵ_i was more than $k = 3$ standard deviations above or below 0. Let's use the `Bayes.outlier` function from `BAS` , to calculate these probabilities for the model `m_lwage_iq` and plot them against the case number.

We start by calculating the probabilities,

```
outliers <- Bayes.outlier(m_lwage_iq, k = 3)
```

and then store the results in a data frame and plot them.

```
outliers_df <- data.frame(probability = outliers$prob.outlier,
                          case = 1:length(outliers$prob.outlier))
ggplot(outliers_df, aes(ymax = probability, x = case)) +
  geom_linerange(ymin = 0) +
  labs(y = "Probability")
```



To identify which cases have probabilities greater than 0.50 of being an outlier, we can use the `filter` function to return which cases have `probability > 0.50`.

```
outliers_df %>%
  filter(probability > 0.50)
```

```
##   probability case
## 1      0.9937  434
## 2      1.0000  514
## 3      0.6610  616
## 4      0.5310  705
## 5      1.0000  784
```

Using the definition of outlier above, which statement is **false**?

- Case 434 has a probability of close to 1 that it an outlier under the normal error model for regressing `lwage` on `iq`

- Case 514 has a probably of close to 1 that it an outlier under the normal error model for regressing `lwage` on `iq`
- Case 616 has a probably of close to 1 that it an outlier under the normal error model for regressing `lwage` on `iq`
- Case 784 has a probably of close to 1 that it an outlier under the normal error model for regressing `lwage` on `iq`

```
# Type your code for Question 5 here.
```

While being 3 standard deviations seems like an extremely unlikely event for a single observation, for large sample sizes, there is often a rather high probability that there will be at least one error ϵ_i that exceeds 3 standard deviations above or below zero *a priori*. We can calculate this as follows

```
# prob of a case being an outlier:
#   being below or above 3 standard deviations from 0
(prob_outlier <- pnorm(-3) + pnorm(3, lower.tail = FALSE))
```

```
## [1] 0.0027
```

```
# probability of a signle case not being an outler is therefore the complement
(prob_not_outlier <- 1 - prob_outlier)
```

```
## [1] 0.9973
```

```
# probability of no outliers in the sample of n assuming errors are independent a priori
n <- nrow(wage)
(prob_no_outliers <- prob_not_outlier^n)
```

```
## [1] 0.07984
```

```
# probability of at least one outlier in the sample is the complement of the
# probability of no outliers in the sample of n
1 - prob_no_outliers
```

```
## [1] 0.9202
```

With a sample size of 935 and using 3 standard deviations to define outliers, the chance of having at least one outlier in the sample is 92.02% so the fact that we did discover some outliers is not that surprising.

So instead of fixing the number of standard deviations to $k = 3$, an alternative is fix the prior probability of there being no outliers in the sample,

$$P(\text{no outliers in sample}) = P(\text{observation is not an outlier})^n = 0.95$$

which we can solve for

$$P(\text{observation is not an outlier}) = 0.95^{1/n}$$

and then solve for k using the normal quantile function.

```
n <- nrow(wage)
(prob_obs_not_outlier <- 0.95^(1/n))
```

```
## [1] 0.9999
```

```
(newk <- qnorm(0.5 + 0.5 * prob_obs_not_outlier))
```

```
## [1] 4.034
```

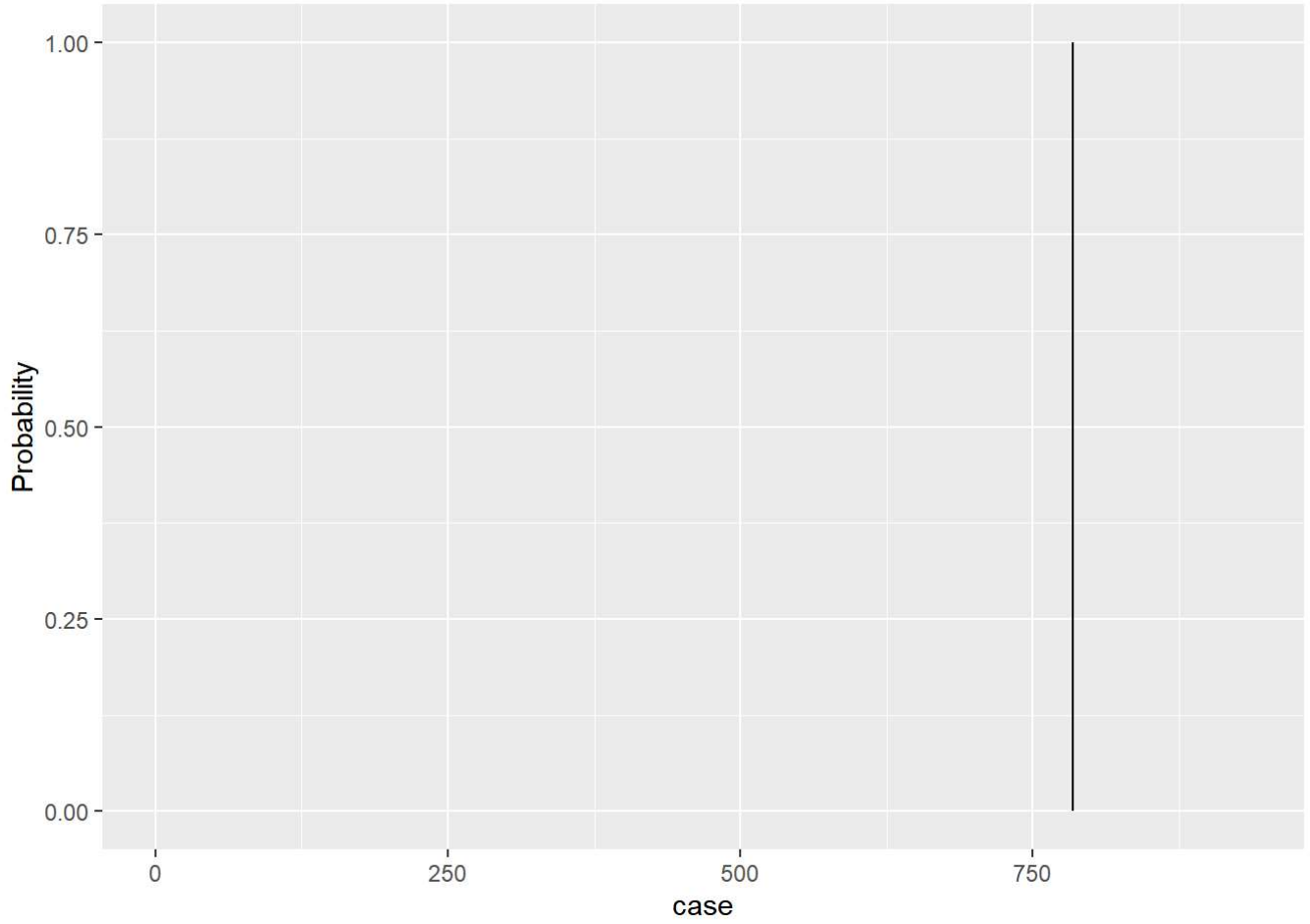
The function `Bayes.outlier` can also calculate `k` internally if we specify the prior probability of there being no outliers in the sample:

```
outliers <- Bayes.outlier(m_lwage_iq, prior.prob=0.95)
```

Use the new value of k to calculate the posterior probability of each observation being an outlier. Which observation has a posterior probability of being an outlier that exceeds the prior probability of being an outlier?

- Case 434
- Case 514
- Case 616
- Case 784

```
# Type your code for Qustion 6 here.
outliers_df <- data.frame(probability = outliers$prob.outlier,
                          case = 1:length(outliers$prob.outlier))
ggplot(outliers_df, aes(ymax = probability, x = case)) +
  geom_linerange(ymin = 0) +
  labs(y = "Probability")
```




```
outliers_df %>%
  filter(probability > 0.50)
```

```
##   probability case
## 1           1  784
```

Multiple linear regression

It is evident that wage can be explained by many predictors, such as experience, education, IQ, and so on. We can include all relevant covariates in a regression model in an attempt to explain as much wage variation as possible. In addition, sometimes outliers can be explained by changing the model by adding other predictors; let's take a look at multiple regression before removing any cases.

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

The use of `. - wage` in the `lm` function tells R to include all covariates in the model except the `wage` variable from the data set.

However, running this full model has a cost: we will remove observations from our data if some measurements in the variables (e.g. birth order, mother's education, and father's education) are missing. By default, the `lm` function does a complete-case analysis. 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 in the data there is no systematic reason for the values to be missing. 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 7 here.
summary(m_lwage_full)
```

```
##
## Call:
## lm(formula = lwage ~ . - wage, data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9689 -0.1946  0.0092  0.2240  1.3419
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.15644    0.22529   22.89  < 2e-16 ***
## hours       -0.00655    0.00193   -3.39  0.00075 ***
## iq           0.00319    0.00122    2.60  0.00943 **
## kww          0.00373    0.00239    1.56  0.11866
## educ         0.04127    0.00894    4.61  4.7e-06 ***
## exper        0.01075    0.00443    2.42  0.01563 *
## tenure       0.00710    0.00289    2.45  0.01440 *
## age          0.00911    0.00598    1.52  0.12806
## married1     0.20076    0.04600    4.36  1.5e-05 ***
## black1      -0.10514    0.05567   -1.89  0.05937 .
## south1      -0.04908    0.03075   -1.60  0.11102
## urban1       0.19566    0.03124    6.26  6.9e-10 ***
## sibs         0.00962    0.00788    1.22  0.22242
## brthord     -0.01846    0.01157   -1.60  0.11097
## meduc        0.00963    0.00617    1.56  0.11875
## feduc        0.00559    0.00540    1.04  0.30080
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.351 on 647 degrees of freedom
##   (272 observations deleted due to missingness)
## Multiple R-squared:  0.293,  Adjusted R-squared:  0.276
## F-statistic: 17.8 on 15 and 647 DF,  p-value: <2e-16
```

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 the values of Adjusted R^2. This module introduced the Bayesian Information Criterion (BIC), which is a criterion 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] 586.4
```

We can compare the BIC of the full model with that of a reduced model. Let us 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 in any variables in the data set.

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

```
## [1] 582.5
```

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

```
## [1] 582.5
```

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

```
## [1] 581.4
```

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

```
## [1] 581
```

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

```
## [1] 582.4
```

R has a function `stepAIC` from the `MASS` package that will work backwards through the model space, removing variables until the AIC score can be no longer lowered. It takes all inputs in the full model, and a penalty parameter k. The default setting is $k = 2$ for the AIC score. Find the best model according to BIC (in which case `k = log(n)` where n is the number of observations). Remember to use `na.omit(wage)` as your data set. You may type `?stepAIC` in the RStudio Console to get the use and examples of the function `stepAIC`.

```
# Type your code for Exercise 4 here.
stepAIC(lm(lwage ~ ., data = na.omit(wage)), k = 2)
```

```
## Start:  AIC=-2807
## lwage ~ wage + hours + iq + kww + educ + exper + tenure + age +
##     married + black + south + urban + sibs + brthord + meduc +
##     feduc
##
##           Df Sum of Sq  RSS   AIC
## - age      1       0.0  9.1 -2809
## - sibs     1       0.0  9.1 -2809
## - brthord  1       0.0  9.1 -2809
## - iq       1       0.0  9.1 -2809
## - kww      1       0.0  9.1 -2809
## - black    1       0.0  9.1 -2809
## - exper    1       0.0  9.1 -2808
## <none>                  9.1 -2807
## - meduc    1       0.0  9.2 -2807
## - feduc    1       0.0  9.2 -2807
## - educ     1       0.0  9.2 -2806
## - married  1       0.1  9.2 -2804
## - urban    1       0.1  9.2 -2802
## - south    1       0.1  9.2 -2802
## - hours    1       0.2  9.3 -2795
## - tenure   1       0.3  9.4 -2788
## - wage     1      70.5 79.6 -1374
##
## Step:  AIC=-2809
## lwage ~ wage + hours + iq + kww + educ + exper + tenure + married +
##     black + south + urban + sibs + brthord + meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
## - kww      1       0.0  9.1 -2811
## - sibs     1       0.0  9.1 -2811
## - brthord  1       0.0  9.1 -2811
## - iq       1       0.0  9.1 -2811
## - black    1       0.0  9.1 -2810
## <none>                  9.1 -2809
## - exper    1       0.0  9.2 -2809
## - meduc    1       0.0  9.2 -2809
## - feduc    1       0.0  9.2 -2808
## - educ     1       0.1  9.2 -2807
## - married  1       0.1  9.2 -2806
## - urban    1       0.1  9.2 -2804
## - south    1       0.1  9.2 -2803
## - hours    1       0.2  9.3 -2797
## - tenure   1       0.3  9.4 -2789
## - wage     1      70.7 79.9 -1373
##
## Step:  AIC=-2811
## lwage ~ wage + hours + iq + educ + exper + tenure + married +
##     black + south + urban + sibs + brthord + meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
## - iq       1       0.0  9.1 -2812
## - brthord  1       0.0  9.1 -2812
## - sibs     1       0.0  9.1 -2812
## - black    1       0.0  9.1 -2812
```

```
## - exper    1       0.0  9.2 -2811
## <none>                  9.1 -2811
## - meduc    1       0.0  9.2 -2810
## - feduc    1       0.0  9.2 -2810
## - educ     1       0.1  9.2 -2809
## - married  1       0.1  9.2 -2808
## - urban    1       0.1  9.2 -2806
## - south    1       0.1  9.2 -2805
## - hours    1       0.2  9.3 -2798
## - tenure   1       0.3  9.4 -2791
## - wage     1      71.4 80.5 -1370
##
## Step:  AIC=-2812
## lwage ~ wage + hours + educ + exper + tenure + married + black +
##     south + urban + sibs + brthord + meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
## - sibs     1       0.0  9.1 -2814
## - brthord  1       0.0  9.1 -2814
## - black    1       0.0  9.2 -2814
## - exper    1       0.0  9.2 -2813
## <none>                  9.1 -2812
## - meduc    1       0.0  9.2 -2812
## - feduc    1       0.0  9.2 -2812
## - married  1       0.1  9.2 -2809
## - educ     1       0.1  9.2 -2809
## - urban    1       0.1  9.2 -2807
## - south    1       0.1  9.3 -2806
## - hours    1       0.2  9.3 -2800
## - tenure   1       0.3  9.4 -2793
## - wage     1      72.4 81.5 -1363
##
## Step:  AIC=-2814
## lwage ~ wage + hours + educ + exper + tenure + married + black +
##     south + urban + brthord + meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
## - brthord  1       0.0  9.1 -2816
## - black    1       0.0  9.2 -2815
## - exper    1       0.0  9.2 -2814
## <none>                  9.1 -2814
## - meduc    1       0.0  9.2 -2814
## - feduc    1       0.0  9.2 -2813
## - educ     1       0.1  9.2 -2811
## - married  1       0.1  9.2 -2811
## - urban    1       0.1  9.2 -2809
## - south    1       0.1  9.3 -2807
## - hours    1       0.2  9.4 -2801
## - tenure   1       0.3  9.4 -2794
## - wage     1      72.5 81.6 -1365
##
## Step:  AIC=-2816
## lwage ~ wage + hours + educ + exper + tenure + married + black +
##     south + urban + meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
```

```
## - black    1       0.0  9.2 -2817
## - exper    1       0.0  9.2 -2816
## <none>                  9.1 -2816
## - feduc    1       0.0  9.2 -2815
## - meduc    1       0.0  9.2 -2815
## - married  1       0.1  9.2 -2813
## - educ     1       0.1  9.2 -2813
## - urban    1       0.1  9.2 -2811
## - south    1       0.1  9.3 -2809
## - hours    1       0.2  9.4 -2803
## - tenure   1       0.3  9.5 -2796
## - wage     1      72.7 81.9 -1365
##
## Step:  AIC=-2817
## lwage ~ wage + hours + educ + exper + tenure + married + south +
##     urban + meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
## - exper    1       0.0  9.2 -2817
## <none>                  9.2 -2817
## - feduc    1       0.0  9.2 -2817
## - meduc    1       0.0  9.2 -2816
## - educ     1       0.1  9.2 -2814
## - married  1       0.1  9.2 -2814
## - urban    1       0.1  9.2 -2813
## - south    1       0.1  9.3 -2809
## - hours    1       0.2  9.4 -2805
## - tenure   1       0.3  9.5 -2797
## - wage     1      73.8 82.9 -1358
##
## Step:  AIC=-2817
## lwage ~ wage + hours + educ + tenure + married + south + urban +
##     meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
## <none>                  9.2 -2817
## - feduc    1       0.0  9.2 -2817
## - meduc    1       0.0  9.2 -2816
## - educ     1       0.1  9.2 -2816
## - married  1       0.1  9.3 -2814
## - urban    1       0.1  9.3 -2813
## - south    1       0.1  9.3 -2809
## - hours    1       0.2  9.4 -2805
## - tenure   1       0.4  9.6 -2792
## - wage     1      75.9 85.1 -1343
```

```
##
## Call:
## lm(formula = lwage ~ wage + hours + educ + tenure + married +
##     south + urban + meduc + feduc, data = na.omit(wage))
##
## Coefficients:
## (Intercept)         wage        hours         educ       tenure     married1       south1
##    5.852655     0.000936    -0.002481     0.004547     0.004869     0.036821    -0.032715
##      urban1        meduc        feduc
##    0.026568     0.003436    -0.002995
```

```
stepAIC(lm(lwage ~ ., data = na.omit(wage)), k = log(ncol(wage) - 1))
```

```
## Start:  AIC=-2794
## lwage ~ wage + hours + iq + kww + educ + exper + tenure + age +
##     married + black + south + urban + sibs + brthord + meduc +
##     feduc
##
##           Df Sum of Sq  RSS   AIC
## - age      1       0.0  9.1 -2797
## - sibs     1       0.0  9.1 -2797
## - brthord  1       0.0  9.1 -2797
## - iq       1       0.0  9.1 -2796
## - kww      1       0.0  9.1 -2796
## - black    1       0.0  9.1 -2796
## - exper    1       0.0  9.1 -2796
## - meduc    1       0.0  9.2 -2794
## - feduc    1       0.0  9.2 -2794
## <none>                  9.1 -2794
## - educ     1       0.0  9.2 -2793
## - married  1       0.1  9.2 -2792
## - urban    1       0.1  9.2 -2790
## - south    1       0.1  9.2 -2789
## - hours    1       0.2  9.3 -2783
## - tenure   1       0.3  9.4 -2775
## - wage     1      70.5 79.6 -1361
##
## Step:  AIC=-2797
## lwage ~ wage + hours + iq + kww + educ + exper + tenure + married +
##     black + south + urban + sibs + brthord + meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
## - kww      1       0.0  9.1 -2799
## - sibs     1       0.0  9.1 -2799
## - brthord  1       0.0  9.1 -2799
## - iq       1       0.0  9.1 -2799
## - black    1       0.0  9.1 -2799
## - exper    1       0.0  9.2 -2797
## - meduc    1       0.0  9.2 -2797
## - feduc    1       0.0  9.2 -2797
## <none>                  9.1 -2797
## - educ     1       0.1  9.2 -2795
## - married  1       0.1  9.2 -2794
## - urban    1       0.1  9.2 -2792
## - south    1       0.1  9.2 -2792
## - hours    1       0.2  9.3 -2786
## - tenure   1       0.3  9.4 -2777
## - wage     1      70.7 79.9 -1362
##
## Step:  AIC=-2799
## lwage ~ wage + hours + iq + educ + exper + tenure + married +
##     black + south + urban + sibs + brthord + meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
## - iq       1       0.0  9.1 -2802
## - brthord  1       0.0  9.1 -2801
## - sibs     1       0.0  9.1 -2801
## - black    1       0.0  9.1 -2801
```

```
## - exper    1       0.0  9.2 -2800
## - meduc    1       0.0  9.2 -2800
## - feduc    1       0.0  9.2 -2799
## <none>                  9.1 -2799
## - educ     1       0.1  9.2 -2798
## - married  1       0.1  9.2 -2797
## - urban    1       0.1  9.2 -2795
## - south    1       0.1  9.2 -2794
## - hours    1       0.2  9.3 -2787
## - tenure   1       0.3  9.4 -2780
## - wage     1      71.4 80.5 -1359
##
## Step:  AIC=-2802
## lwage ~ wage + hours + educ + exper + tenure + married + black +
##     south + urban + sibs + brthord + meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
## - sibs     1       0.0  9.1 -2804
## - brthord  1       0.0  9.1 -2804
## - black    1       0.0  9.2 -2803
## - exper    1       0.0  9.2 -2802
## - meduc    1       0.0  9.2 -2802
## - feduc    1       0.0  9.2 -2802
## <none>                  9.1 -2802
## - married  1       0.1  9.2 -2799
## - educ     1       0.1  9.2 -2799
## - urban    1       0.1  9.2 -2797
## - south    1       0.1  9.3 -2796
## - hours    1       0.2  9.3 -2790
## - tenure   1       0.3  9.4 -2783
## - wage     1      72.4 81.5 -1353
##
## Step:  AIC=-2804
## lwage ~ wage + hours + educ + exper + tenure + married + black +
##     south + urban + brthord + meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
## - brthord  1       0.0  9.1 -2806
## - black    1       0.0  9.2 -2806
## - exper    1       0.0  9.2 -2805
## - meduc    1       0.0  9.2 -2804
## - feduc    1       0.0  9.2 -2804
## <none>                  9.1 -2804
## - educ     1       0.1  9.2 -2802
## - married  1       0.1  9.2 -2801
## - urban    1       0.1  9.2 -2800
## - south    1       0.1  9.3 -2798
## - hours    1       0.2  9.4 -2792
## - tenure   1       0.3  9.4 -2785
## - wage     1      72.5 81.6 -1355
##
## Step:  AIC=-2806
## lwage ~ wage + hours + educ + exper + tenure + married + black +
##     south + urban + meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
```

```
## - black    1       0.0  9.2 -2809
## - exper    1       0.0  9.2 -2807
## - feduc    1       0.0  9.2 -2807
## - meduc    1       0.0  9.2 -2807
## <none>                  9.1 -2806
## - married  1       0.1  9.2 -2804
## - educ     1       0.1  9.2 -2804
## - urban    1       0.1  9.2 -2802
## - south    1       0.1  9.3 -2800
## - hours    1       0.2  9.4 -2795
## - tenure   1       0.3  9.5 -2788
## - wage     1      72.7 81.9 -1356
##
## Step:  AIC=-2809
## lwage ~ wage + hours + educ + exper + tenure + married + south +
##     urban + meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
## - exper    1       0.0  9.2 -2810
## - feduc    1       0.0  9.2 -2809
## <none>                  9.2 -2809
## - meduc    1       0.0  9.2 -2809
## - educ     1       0.1  9.2 -2806
## - married  1       0.1  9.2 -2806
## - urban    1       0.1  9.2 -2805
## - south    1       0.1  9.3 -2801
## - hours    1       0.2  9.4 -2797
## - tenure   1       0.3  9.5 -2790
## - wage     1      73.8 82.9 -1351
##
## Step:  AIC=-2810
## lwage ~ wage + hours + educ + tenure + married + south + urban +
##     meduc + feduc
##
##           Df Sum of Sq  RSS   AIC
## <none>                  9.2 -2810
## - feduc    1       0.0  9.2 -2810
## - meduc    1       0.0  9.2 -2810
## - educ     1       0.1  9.2 -2809
## - married  1       0.1  9.3 -2807
## - urban    1       0.1  9.3 -2806
## - south    1       0.1  9.3 -2802
## - hours    1       0.2  9.4 -2798
## - tenure   1       0.4  9.6 -2785
## - wage     1      75.9 85.1 -1336
```

```
##
## Call:
## lm(formula = lwage ~ wage + hours + educ + tenure + married +
##     south + urban + meduc + feduc, data = na.omit(wage))
##
## Coefficients:
## (Intercept)         wage        hours         educ       tenure     married1       south1
##    5.852655     0.000936    -0.002481     0.004547     0.004869     0.036821    -0.032715
##      urban1        meduc        feduc
##    0.026568     0.003436    -0.002995
```

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 [@Clyde2018]. We can use this for either implementing BMA or selecting models.

We start by applying BMA to the wage data using all 15 potential predictors.

```
# Exclude observations with missing values in the data set
wage_no_na <- na.omit(wage)

# Fit the model using Bayesian linear regression, `bas.lm` function in the `BAS` package
bma_lwage <- bas.lm(lwage ~ . -wage, data = wage_no_na,
                    prior = "BIC",
                    modelprior = uniform())

# Print out the marginal posterior inclusion probabilities for each variable
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     tenure        age   m
arried1
##    1.0000     0.8554     0.8973     0.3479     0.9989     0.7100     0.7039     0.5247
0.9989
##    black1     south1     urban1       sibs    brthord      meduc      feduc
##    0.3464     0.3203     1.0000     0.0415     0.1224     0.5734     0.2327
```

```
# Top 5 most probably models
summary(bma_lwage)
```

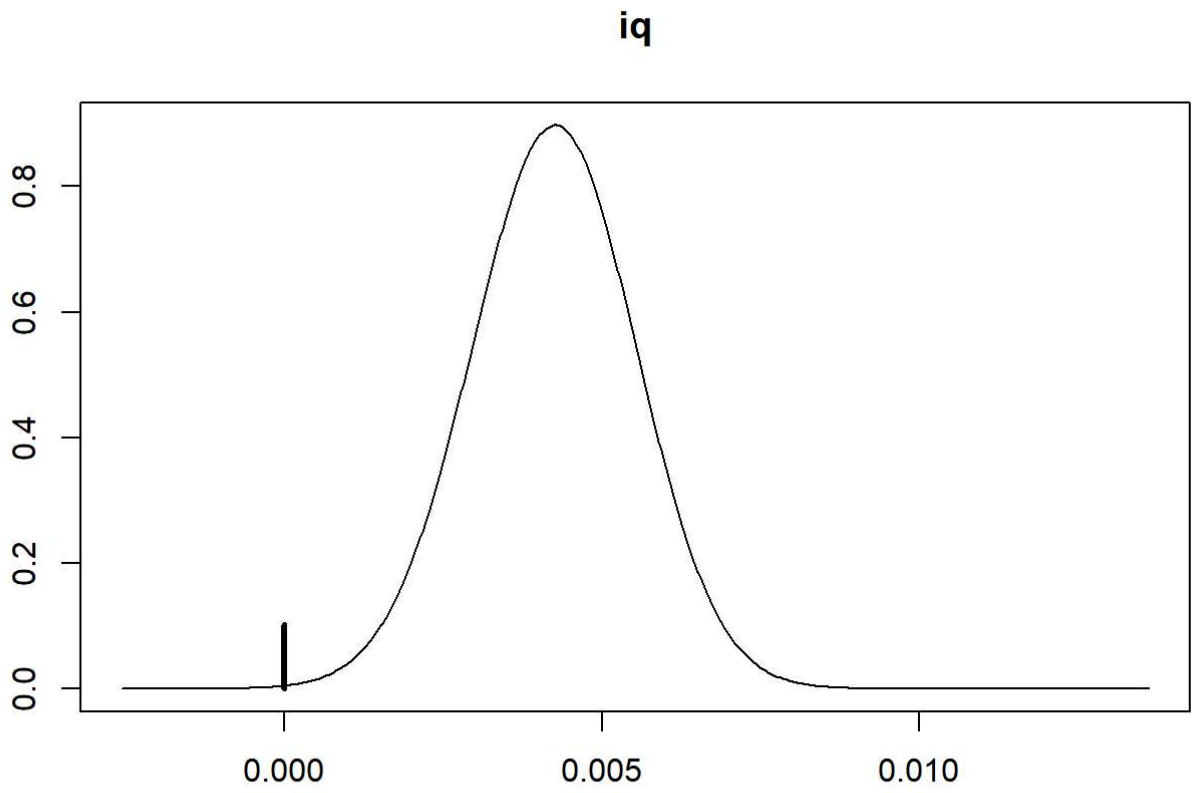
```
##           P(B != 0 | Y)   model 1    model 2    model 3    model 4    model 5
## Intercept       1.00000  1.00e+00     1.0000     1.0000     1.0000     1.0000
## hours           0.85540  1.00e+00     1.0000     1.0000     1.0000     1.0000
## iq              0.89732  1.00e+00     1.0000     1.0000     1.0000     1.0000
## kww             0.34790  0.00e+00     0.0000     0.0000     1.0000     0.0000
## educ            0.99887  1.00e+00     1.0000     1.0000     1.0000     1.0000
## exper           0.70999  0.00e+00     1.0000     1.0000     1.0000     0.0000
## tenure          0.70389  1.00e+00     1.0000     1.0000     1.0000     1.0000
## age             0.52468  1.00e+00     1.0000     0.0000     0.0000     1.0000
## married1        0.99894  1.00e+00     1.0000     1.0000     1.0000     1.0000
## black1          0.34636  0.00e+00     0.0000     0.0000     0.0000     1.0000
## south1          0.32029  0.00e+00     0.0000     0.0000     0.0000     0.0000
## urban1          1.00000  1.00e+00     1.0000     1.0000     1.0000     1.0000
## sibs            0.04152  0.00e+00     0.0000     0.0000     0.0000     0.0000
## brthord         0.12241  0.00e+00     0.0000     0.0000     0.0000     0.0000
## meduc           0.57339  1.00e+00     1.0000     1.0000     1.0000     1.0000
## feduc           0.23274  0.00e+00     0.0000     0.0000     0.0000     0.0000
## BF                   NA  1.00e+00     0.5219     0.5183     0.4414     0.4127
## PostProbs            NA  4.55e-02     0.0237     0.0236     0.0201     0.0188
## R2                   NA  2.71e-01     0.2767     0.2696     0.2763     0.2762
## dim                  NA  9.00e+00    10.0000     9.0000    10.0000    10.0000
## logmarg              NA -1.49e+03 -1490.7032 -1490.7103 -1490.8708 -1490.9382
```

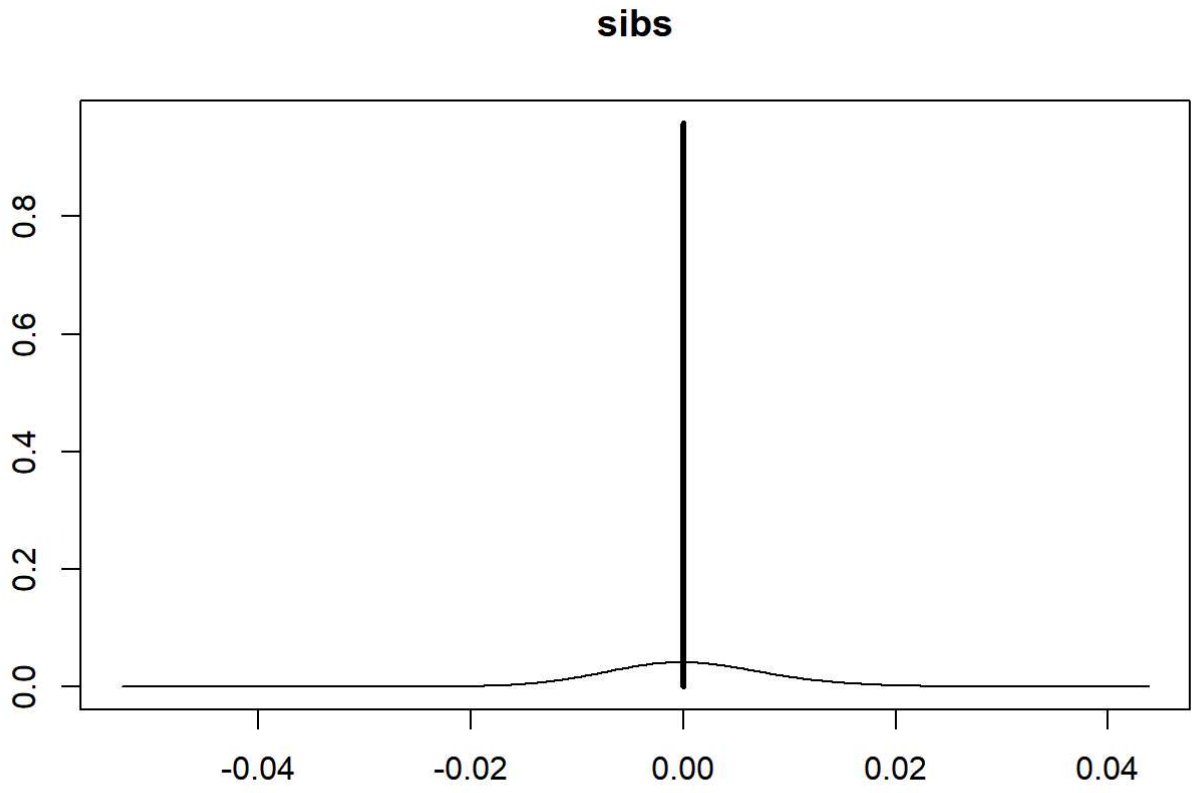
Printing the model object and the summary command gives us both the posterior model inclusion probability for each variable and the most probable models. For example, the posterior probability that `hours` is included in the model is 0.855. Further, the most likely model, which has posterior probability 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^{15} 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.

```
# Obtain the coefficients from the model `bma_lwage`
coef_lwage <- coefficients(bma_lwage)

# `iq` is the 3rd variable, while `sibs` is the 13th variable in the data set
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  6.787e+00 6.840625  6.814e+00
## hours     -9.285e-03 0.000000 -5.308e-03
## iq         0.000e+00 0.006246  3.798e-03
## kww        0.000e+00 0.008403  1.961e-03
## educ       2.209e-02 0.065713  4.407e-02
## exper      0.000e+00 0.021144  1.003e-02
## tenure     0.000e+00 0.012729  5.936e-03
## age        0.000e+00 0.025330  8.966e-03
## married1   1.142e-01 0.297275  2.093e-01
## black1    -1.924e-01 0.000000 -4.419e-02
## south1    -1.021e-01 0.000000 -2.218e-02
## urban1     1.336e-01 0.258637  1.981e-01
## sibs       0.000e+00 0.000000  2.185e-05
## brthord   -2.108e-02 0.000000 -1.947e-03
## meduc      0.000e+00 0.022668  8.672e-03
## feduc     -7.127e-06 0.015338  2.513e-03
## attr(,"Probability")
## [1] 0.95
## attr(,"class")
## [1] "confint.bas"
```

For Questions 9-10, we'll use a reduced data set which excludes wage, number of siblings, birth order, and parental education.

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

Let's use BMA with the Zellner-Siow prior on the regression coefficients:

```
bma_lwage_red <- bas.lm(lwage ~ ., data = wage_red,
                        prior = "ZS-null",
                        modelprior = uniform())
```

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 9 here.
coefficients(bma_lwage_red)
```

```
##
##  Marginal Posterior Summaries of Coefficients:
##
##  Using  BMA
##
##  Based on the top  2048 models
##            post mean  post SD   post p(B != 0)
## Intercept   6.77900    0.01181   1.00000
## hours      -0.00515    0.00201   0.94864
## iq          0.00322    0.00127   0.94757
## kww         0.00212    0.00253   0.50003
## educ        0.05166    0.00792   1.00000
## exper       0.01178    0.00436   0.95175
## tenure      0.01051    0.00249   0.99875
## age         0.00239    0.00488   0.26961
## married1    0.19736    0.03880   0.99997
## black1     -0.14715    0.04415   0.98661
## south1     -0.07444    0.03375   0.91571
## urban1      0.17909    0.02703   1.00000
```

True or **False**: The naive model with all variables included has posterior probability greater than 0.5.

- True
- False

```
# Type your code for Question 10 here.
which.max(lapply(bma_lwage_red$which, length))
```

```
## [1] 645
```

```
bma_lwage_red$which[645]
```

```
## [[1]]
##  [1]  0  1  2  3  4  5  6  7  8  9 10 11
```

```
bma_lwage_red$postprobs[645]
```

```
## [1] 0.05366
```

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

```
# Type your code for Exercise 5 here.
```

Because we have log transformed wage, interpretation of coefficients from the output is not as useful for understanding how the different predictors influence wages. Instead we can interpret coefficients after transforming back to the original units by exponentiation. The exponential of the posterior mean is not the same as the mean of the exponential, however, the median of wage can be found by exponentiating the median of log wage (i.e. the middle value is still in the middle with transformations that do not change the order of values).

Let's look at the coefficients and 95% credible intervals

```
coef(bma_lwage_red)
```

```
##
##  Marginal Posterior Summaries of Coefficients:
##
##  Using  BMA
##
##  Based on the top  2048 models
##            post mean  post SD   post p(B != 0)
## Intercept   6.77900    0.01181   1.00000
## hours      -0.00515    0.00201   0.94864
## iq          0.00322    0.00127   0.94757
## kww         0.00212    0.00253   0.50003
## educ        0.05166    0.00792   1.00000
## exper       0.01178    0.00436   0.95175
## tenure      0.01051    0.00249   0.99875
## age         0.00239    0.00488   0.26961
## married1    0.19736    0.03880   0.99997
## black1     -0.14715    0.04415   0.98661
## south1     -0.07444    0.03375   0.91571
## urban1      0.17909    0.02703   1.00000
```

```
coef(bma_lwage_red) %>%
  confint()
```

```
##                2.5%     97.5%      beta
## Intercept  6.755286  6.801300  6.779004
## hours     -0.008147  0.000000 -0.005153
## iq         0.000000  0.005072  0.003216
## kww        0.000000  0.006796  0.002125
## educ       0.036413  0.067237  0.051663
## exper      0.000000  0.017832  0.011777
## tenure     0.005741  0.015434  0.010513
## age       -0.000123  0.015410  0.002391
## married1   0.120885  0.272162  0.197358
## black1    -0.236653 -0.066288 -0.147152
## south1    -0.124118  0.000000 -0.074443
## urban1     0.125023  0.229867  0.179091
## attr(,"Probability")
## [1] 0.95
## attr(,"class")
## [1] "confint.bas"
```

The exponential transformation applied to coefficients has a multiplicative effect on the posterior median. What this means is that a one unit increase in predictor X_j leads to a $(\exp(\beta_j) - 1) \times 100$ percent increase in the median wage [@StatNews83]. For a factor or indicator variable like `urban`, the posterior median for wages for urban areas (`urban == 1`) will be $(\exp(0.1791) - 1) \times 100$ percent higher than in rural areas (`urban == 0`). Similarly we can use the same transformation with the credible interval. First, let's calculate the credible interval for the coefficient of `urban` and exponentiate it.

```
ci_urban <- coef(bma_lwage_red) %>%
  confint(parm = "urban1") %>%
  exp()

ci_urban
```

```
##         2.5% 97.5%  beta
## urban1 1.133  1.26 1.196
## attr(,"Probability")
## [1] 0.95
## attr(,"class")
## [1] "confint.bas"
```

Then, we can substract 1 from the bounds of the interval and multipl them by 100 to make them a bit more straightforward to interpret

```
(ci_urban - 1) * 100
```

```
##         2.5% 97.5%  beta
## urban1 13.29    26 19.61
## attr(,"Probability")
## [1] 0.95
## attr(,"class")
## [1] "confint.bas"
```

Based on this data, there is a 95% chance that median wages for urban areas are 13.38 to 25.96 times higher than in rural regions.

Find a 95% credible interval for the coefficient of `educ` and provide an interpretation.

```
# Type your code for Exercise 6 here
```

Prediction with BAS

Similar to last week's lab, we will be using Bayesian predictive distribution for predictions and interpretation of predictions. 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 us find the predictive values under the *Best Predictive Model* (`BPM`), the one which has predictions closest to BMA and corresponding posterior standard deviations.

```
BPM_pred_lwage <- predict(bma_lwage, estimator = "BPM", se.fit = TRUE)
variable.names(BPM_pred_lwage)
```

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

In the code above, the function `variable.names` can be used to extract the names of all of the predictors in the Best Probabilty model. This can be used to identify the variables in the *Highest Probability Model* (`HPM`)

```
HPM_pred_lwage <- predict(bma_lwage, estimator = "HPM")
variable.names(HPM_pred_lwage)
```

```
## [1] "Intercept" "hours"     "iq"        "educ"      "tenure"    "age"       "married1"  "u
rban1"
## [9] "meduc"
```

and the *Median Probability Model* (`MPM`)

```
MPM_pred_lwage <- predict(bma_lwage, estimator = "MPM")
variable.names(MPM_pred_lwage)
```

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

The `MPM` includes `exper` in addition to all of the variables in the *Highest Probability Model* (`HPM`), while the `BPM` includes `kww` in addition to all of the variables in the `MPM` .

Based on these results which covariates are included in **all** of the following: the best predictive model, the median probability model, and the highest posterior probability model?

- `kww` , `married` , `urban`
- `married` , `age` , `black`
- `black` , `south` , `married`
- `meduc` , `urban` , `married`

```
# Type your code for Question 11 here.
intersect(intersect(variable.names(BPM_pred_lwage), variable.names(HPM_pred_lwage)), variabl
e.names(MPM_pred_lwage))
```

```
## [1] "Intercept" "hours"     "iq"        "educ"      "tenure"    "age"       "married1"  "u
rban1"
## [9] "meduc"
```

Let us turn to examine what characteristics lead to the highest wages in the `BPM` model.

```
# Find the index of observation with the largest fitted value
opt <- which.max(BPM_pred_lwage$fit)

# Extract the row with this observation and glimpse at the row
wage_no_na %>%
  slice(opt) %>%
  glimpse()
```

```
## Rows: 1
## Columns: 17
## $ wage    <int> 1586
## $ hours   <int> 40
## $ iq      <int> 127
## $ kww     <int> 48
## $ educ    <int> 16
## $ exper   <int> 16
## $ tenure  <int> 12
## $ age     <int> 37
## $ married <fct> 1
## $ black   <fct> 0
## $ south   <fct> 0
## $ urban   <fct> 1
## $ sibs    <int> 4
## $ brthord <int> 4
## $ meduc   <int> 16
## $ feduc   <int> 16
## $ lwage   <dbl> 7.369
```

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
## 6.662 8.056 7.359
```

To translate this back to `wage` (recall that we regress `lwage`), we may exponentiate the interval to obtain a 95% prediction interval for the wages of an individual with covariates at the levels of the individual specified by `opt`.

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

```
##  2.5% 97.5%  pred
##   782  3154  1571
```

If we 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
##  747.4 3069.7 1495.0
```

Repeat these calculations for a 95% prediction interval for the individual who is predicted to have the highest predicted wages based on the best predictive model.

- [414, 1717]
- [782, 1571]

- [782, 3154]
- [706, 2950]

```
# Type your code for Question 12 here.
BPM_pred_lwage <- predict(bma_lwage, estimator = "BPM", se.fit = TRUE)
ci_bpm_lwage <- confint(BPM_pred_lwage, estimator = "BPM")
opt_bpm <- which.max(BPM_pred_lwage$fit)
exp(ci_bpm_lwage[opt_bpm, ])
```

```
##  2.5% 97.5%  pred
##   782  3154  1571
```


References