Projektbericht

Heehwan Soul, 885941

2024-02-02

1. Maximum-Likelihood-Schätzung

Eine Stichprobe aus der Poisson Verteilung mit λ = 4.

```
##   [1] 2 1 6 5 2 4 3 0 5 9 0 7 3 3 4 8 9 5 5 8 4 2 4 2 3 2 5 6 4 4 5 4 3 2 4 4 6
##  [38] 2 2 3 8 7 4 4 5 5 7 4 5 6 6 4 4 5 4 6 4 2 5 6 5 3 0 5 4 3 5 4 1 6 3 2 7 6
##  [75] 5 5 3 2 4 3 3 7 4 4 2 5 7 6 2 4 4 8 5 6 4 3 2 2 8 5
```

Die Herleitung der Schätzung in den wesentlichen Schritten

Seien $X_1, \ldots, X_n$ unabhängige Wiederholungen einer Poisson-verteilten Größe $Po(\lambda)$ mit zu schätzendem Wert λ. Die Realisationen seien $x_1, \ldots, x_n$. Damit erhält man die Likelihoodfunktion

$$L(\lambda) = f(x_1, \ldots, x_n \mid \lambda)$$

nach der Definition. Man kann diese Funktion so umformen, weil die Zufallsvariablen $X_1, \ldots, X_n$ $unabhängig$ sind.

$$= f(x_1 \mid \lambda) \cdots f(x_n \mid \lambda)$$

Man erhält dann die Log-Likelihood-Funktion

$$\ln(L(\lambda)) = \ln(f(x_1 \mid \lambda) \cdots f(x_n \mid \lambda)).$$

Man kann diesen Term durch die Produktregel von Logarithmus so umformen

$$= \ln(f(x_1 \mid \lambda)) + \cdots + \ln(f(x_n \mid \lambda)) = \sum_{i=1}^{n} \ln(f(x_i \mid \lambda).$$

Die funktion $f(x_i \mid \lambda) = e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$ nach der Defition von der Poisson Verteilung und kann man das oben ersetzen.

$$= \sum_{i=1}^{n} \ln(e^{-\lambda} \frac{\lambda^{x_i}}{x_i!})$$

Man kann diesen Term mit Eigenschaften von Logarithmus weiter einfachen.

$$= \sum_{i=1}^{n} (\ln(e^{-\lambda}) + ln(\lambda^{x_i}) - ln(x_i!)) = \sum_{i=1}^{n} (-\lambda + x_i \ln(\lambda) - \ln(x_i!))$$

Ableiten und Nullsetzen liefert

$$\frac{\partial \ln(L(\lambda))}{\partial \lambda} = \sum_{i=1}^{n}(-1 + \frac{x_i}{\lambda}) = 0$$

und damit

$$-n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0$$

und das heißt

$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}.$$

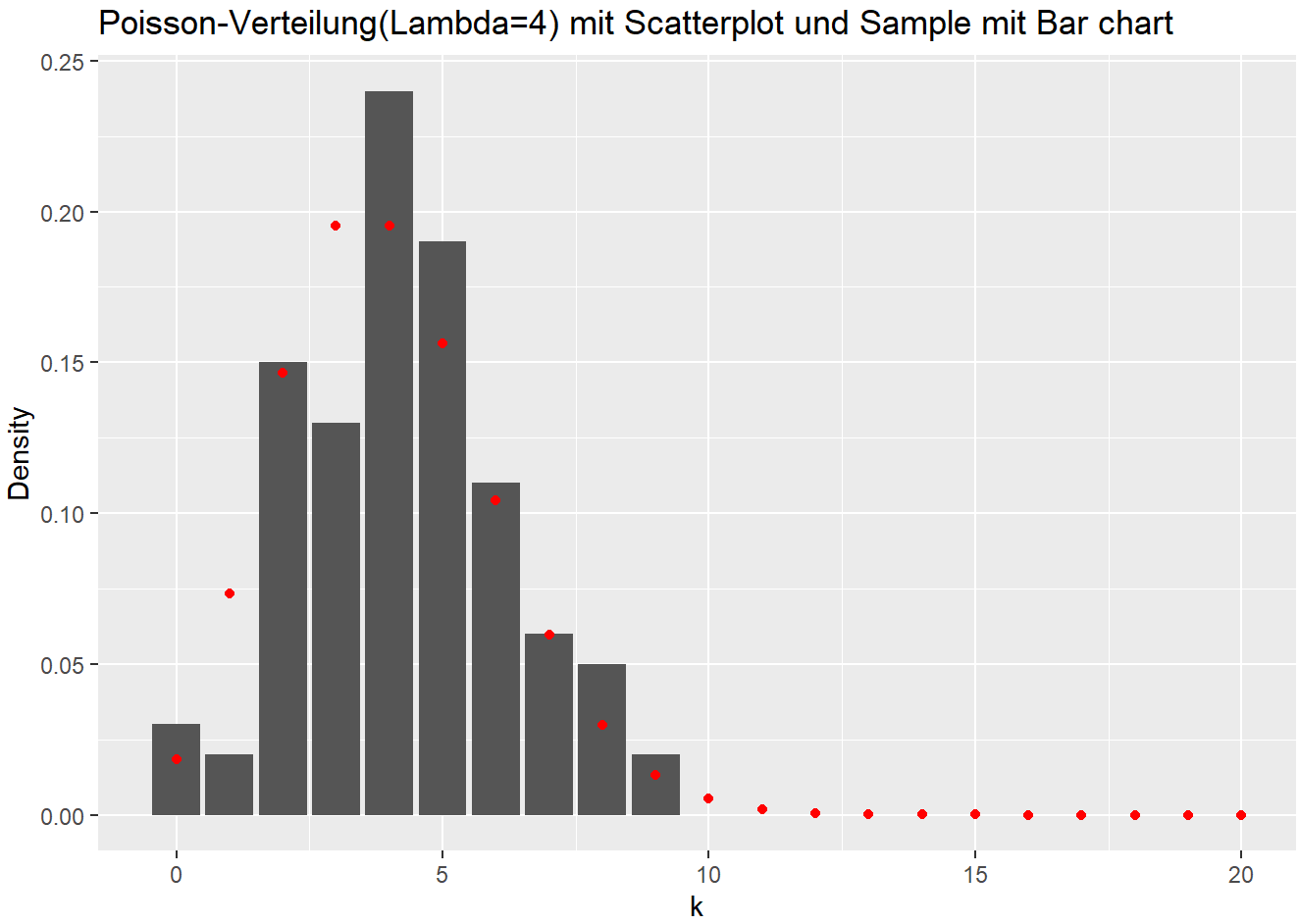
Wenn λ das arithmetische Mittel von alle Realisationen ist, der Wert von der Log-Likelihood-Funktion ist maximal und daraus folgt, dass der Wert von der Likelihood-Funktion auch maximal ist. Der Maximum Likelihood-Schätzer ist also in diesem Fall für jede Realisationsfolge identisch mit dem arithmetischen Mittle.

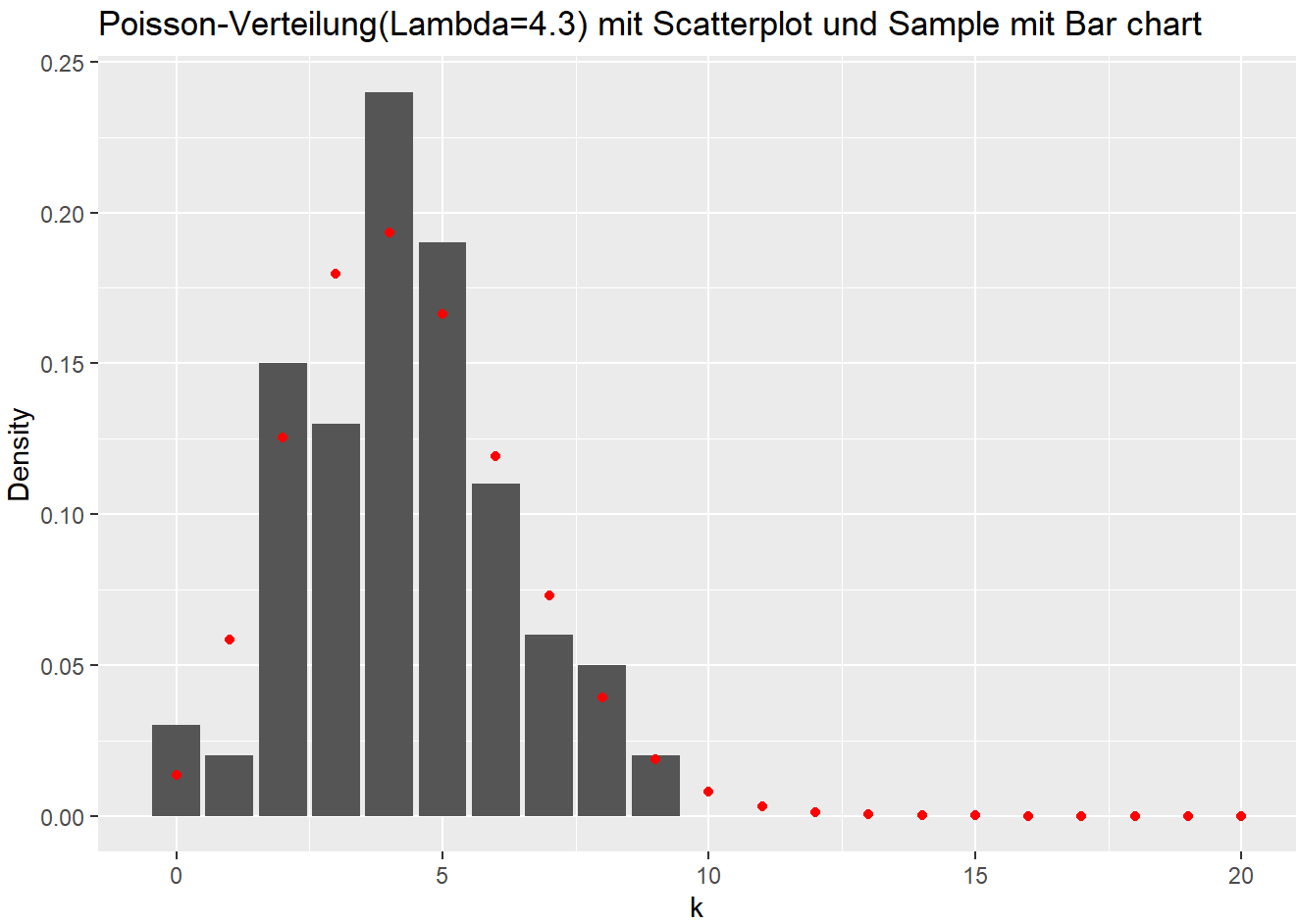
Reference: Fahrmeir, Statistik der Weg zur Datenanalyse

Wenden Sie die hergeleitete Schätzung auf Ihre Daten an

```
## [1] 4.28
```

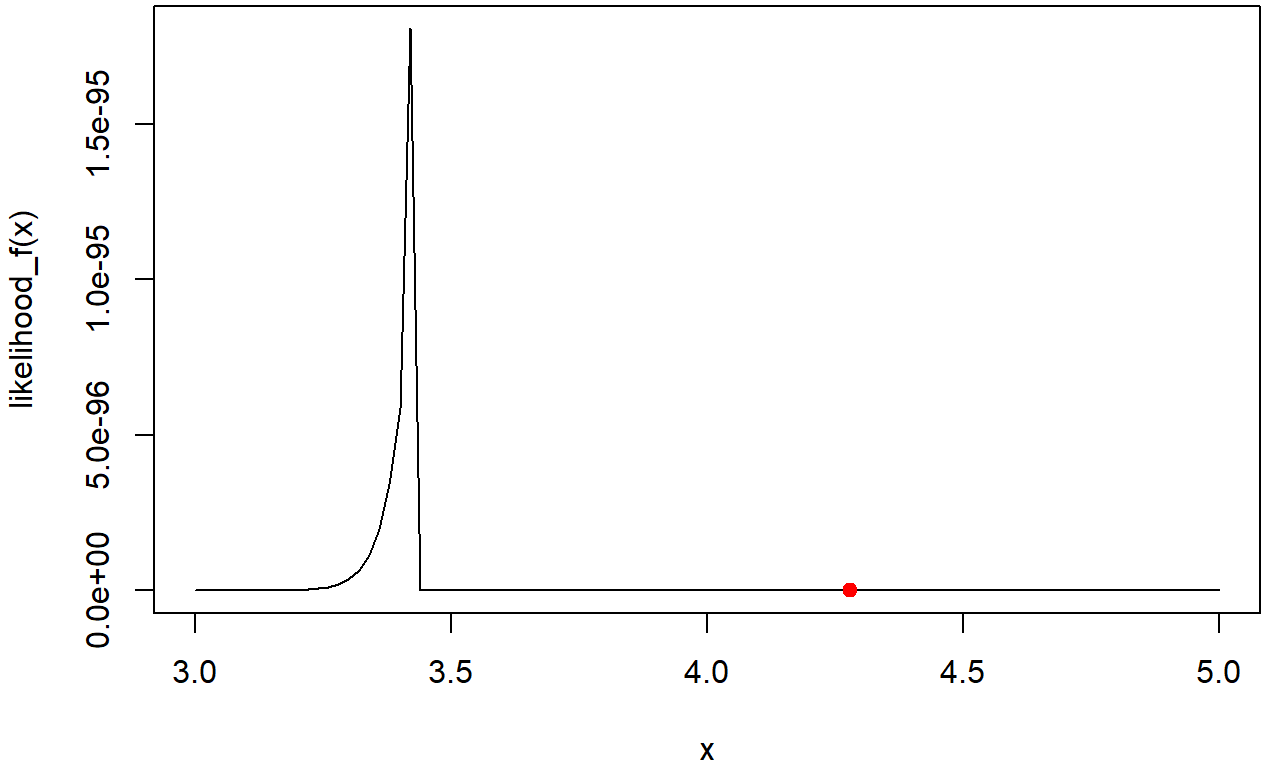
Der Maximum Likelihood-Schätzer ist 3.938333, obwohl das Sample aus der Poisson-Verteilung für $\lambda = 4$ kommt. Der Schätzer kann anders als 4 sein, sollte der Schätzer jedoch nähe von 4 sein.

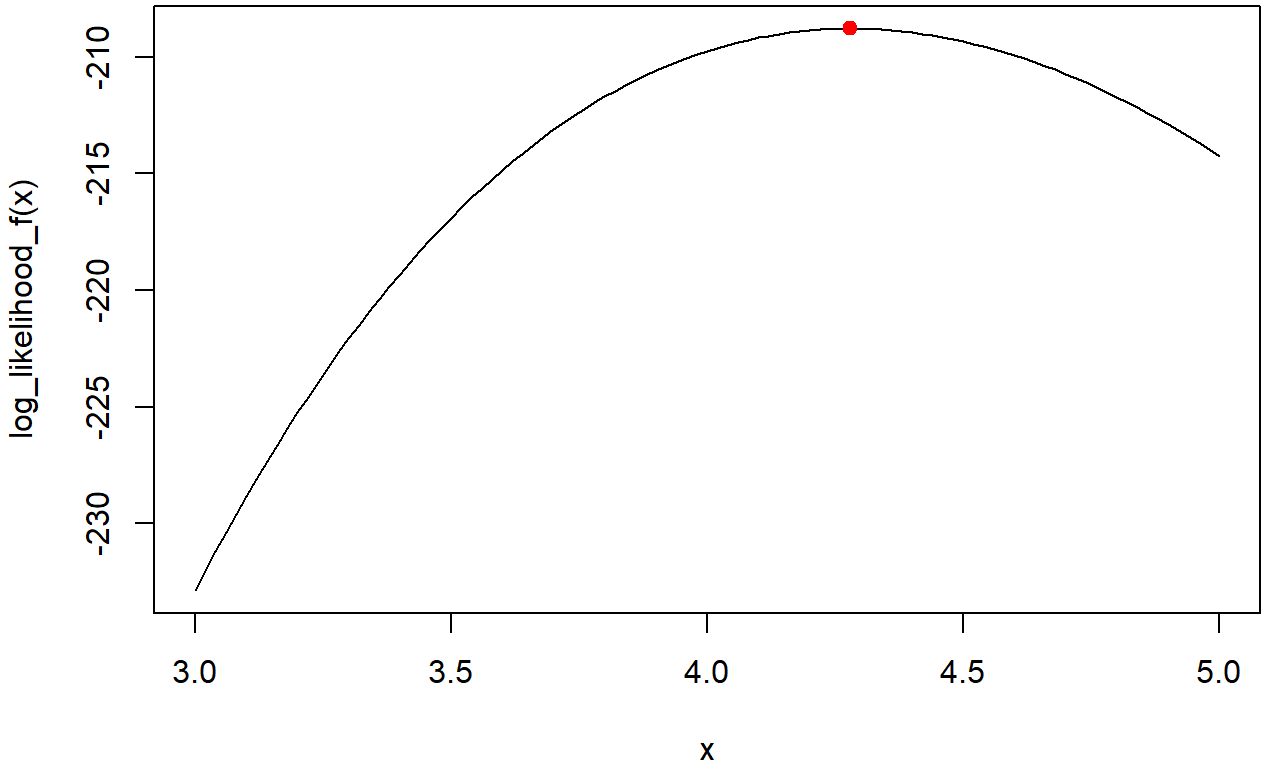

Poisson-Verteilung(Lambda=4) mit Scatterplot und Sample mit Bar chart
Density
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
k


Poisson-Verteilung(Lambda=4.3) mit Scatterplot und Sample mit Bar chart
Density
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
k

Man kann sehen, dass die Verteilung des Samples sehr änlich wie die Poisson-Verteilung für $\lambda = 4$ und $\lambda = 4.3$ ist. Nach der Maximum-Likelihood-Schätzung sollte die Verteilung der Poisson-Verteilung für $\lambda = 4.3$ näher von die Verteilung des Samples als die Verteilung der Poisson-Verteilung für $\lambda = 4$ sein, kann man jedoch durch die Grafiken nicht ganz klar das sehen.

Stellen Sie die sowohl die Likelihood- als auch die Log-Likelihood-Funktion (inklusive der eingezeichneten Schätzung) passend grafisch dar.





Auf

der zweiten Grafik kann man einfach sehen, dass der rote Punkt das Maximum der Log-Likelihood-Funktion ist. Auf der ersten Grafik ist jedoch der rote Punkt kein Maximum der Likelihood-Funktion. Die numerische Auswertungen dieser Likelihood-Funktion können viel unterschiedlich als die exakte Auswetungen sein, da diese Likelihood-Funktion ist eine Funktion mit Exponenten. Deswegen ist es in diesem Fall viel besser Log-Likelihood-Funktion zu nutzen.

2. Estimation Theory

In this section we are going to estimate a mean value of a distribution using a estimator. Suppose $X_1, \ldots, X_n$ is independent and identically distributed random variables(i.i.d.) with a Bernoulli distribution($Bernoulli(p)$). We are going to use the mean($\overline{X}$) as our estimator. Is it a good estimator? We can see if it is good or not through the concepts of bias, MSE and consistency.

Bias, MSE, Consistency

1. Bias: The bias of a estimator is

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

by Definition. Here $\hat{\theta}$ is a estimator and θ is parameter, which is estimated. In our case $\hat{\theta}$ is the mean($\overline{X}$) and θ is the mean of $Bernoulli(p)$, and we know that the mean of $Bernoulli(p)$ is p. So we can rewrite it as follows:

$$= E(\overline{X}) - p.$$

We can simplify this expression using the properties of the mean.

$$= E(\frac{X_1 + \cdots + X_n}{n}) - p$$

$$= \frac{1}{n} E(X_1 + \cdots + X_n) - p$$

$$= \frac{1}{n} E(\sum_{i=1}^{n} X_i) - p$$

$$= \frac{1}{n} \sum_{i=1}^{n} E(X_i) - p$$

$$= \frac{1}{n} \sum_{i=1}^{n} p - p$$

$$= \frac{1}{n} np - p = 0$$

This estimator is unbiased.

2. MSE: The MSE of a estimator is

$$MSE(\hat{\theta}) = E((\hat{\theta} - \theta)^2)$$

by Definition and it is as same as

$$= Var(\hat{\theta}) + Bias(\hat{\theta})^2.$$

We know that the bias of this estimator is 0, so we can rewrite it as follows:

$$= Var(\frac{X_1 + \cdots + X_n}{n}) + 0^2.$$

We can simplify this expression like above using the properties of variance and the fact that the variance of $Bernoulli(p)$ is $p(1-p)$.

$$= \frac{1}{n^2} Var(X_1 + \cdots + X_n)$$

$$= \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i)$$

$$\frac{1}{n^2} p(1-p)$$

3. Consistenty: The MSE of this estimator tends to zero as $n \to \infty$, therefore this estimator is consistent in mean square.

With the information above, we can say that this estimator is reasonable. Now let's do a simulation to check if this estimator works.

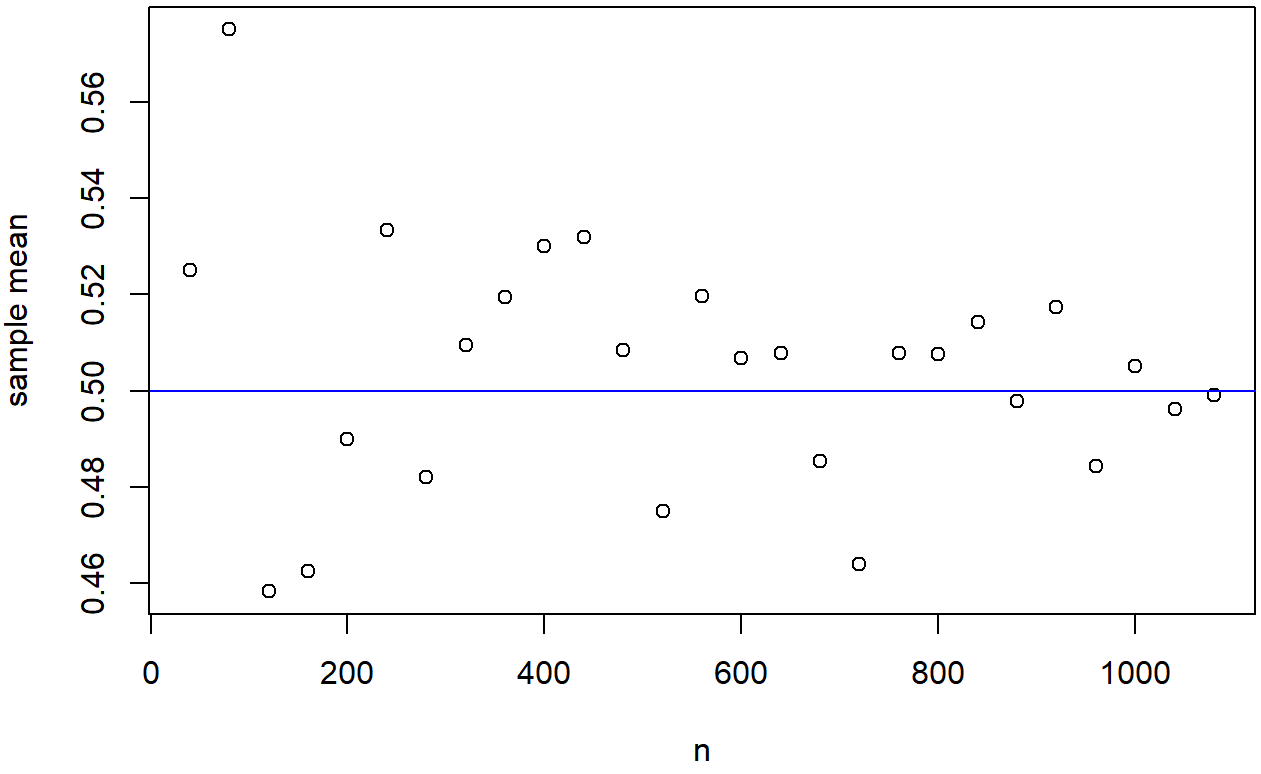
```
set.seed(885941) # damit sind die Ergebnisse immer wieder reproduzierbar

sample_bernoulli <- rbinom(10000, 1, 0.5) # we can use rbinom(), because Bernoulli distribution
is a special case of binomial distribution with size=1.

mean(sample_bernoulli)
```

```
## [1] 0.5035
```

The mean of our sample is 0.5056 and it is quite close to the mean of the Bernoulli distribution with p = 0.5. We can see that this estimator works well.

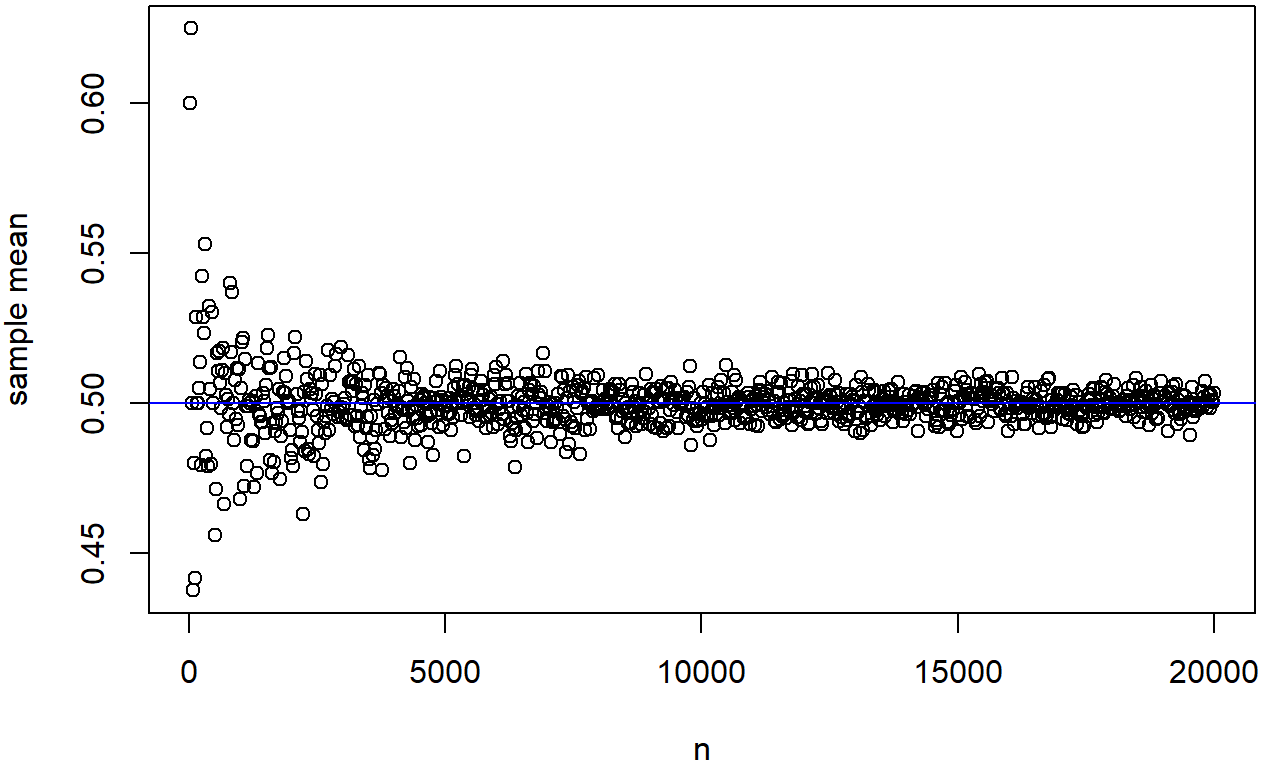
Sample mean converges to the true mean

Let's do simulations to see if sample mean converges to the true mean when n is getting bigger.




```
## integer(0)
```

The p value is 0.5 and the blue line on the graph is the p value which is also the mean of this Bernoulli distribution. We can find that the sample mean tends to close to 0.5 when n gets bigger. To see that trend clearly, let's use more data.



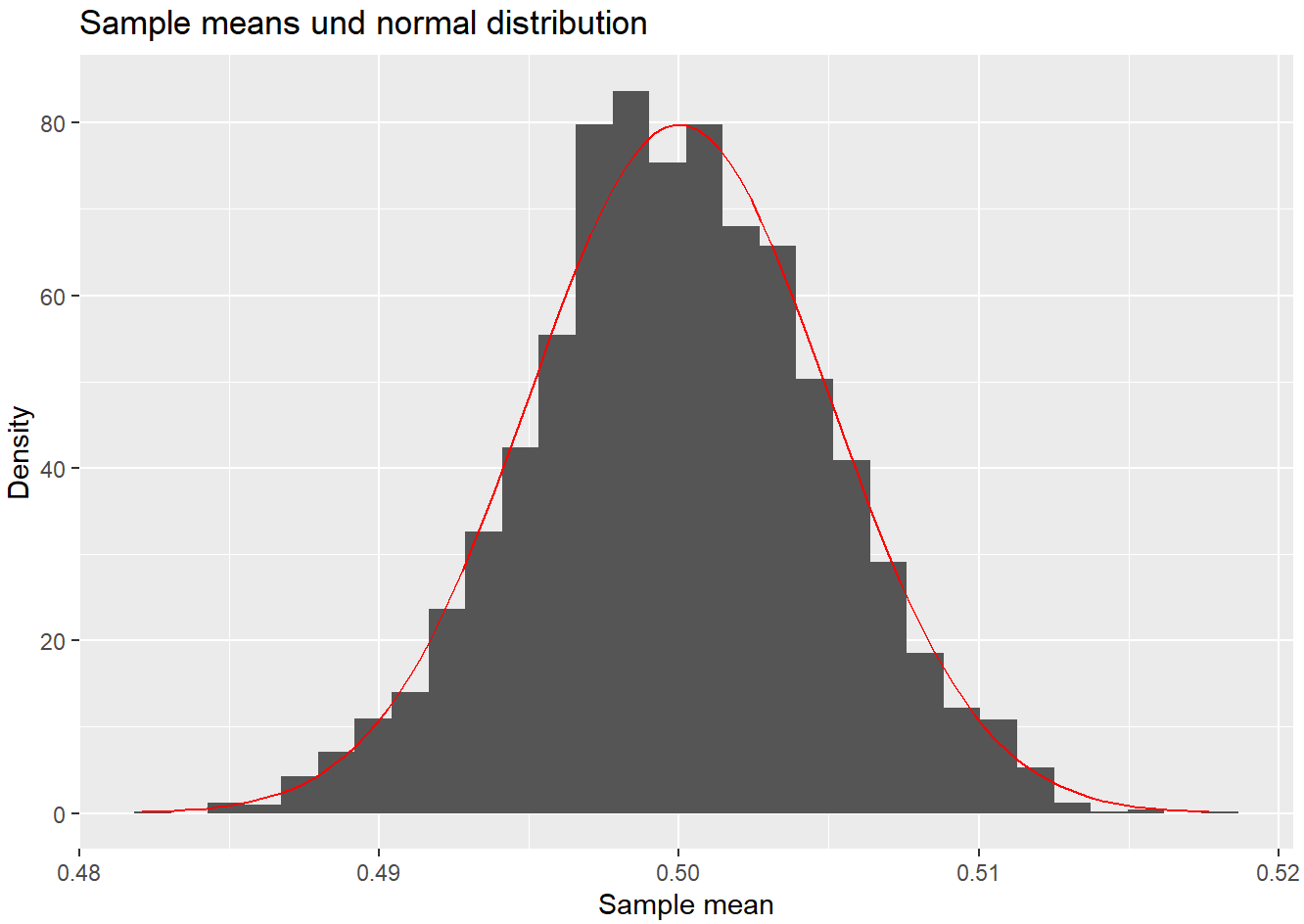

```
## integer(0)
```

We can see that the sample mean converges to the true mean(0.5) when n gets bigger.

The sample mean is approxiamtely normally distributed

We are going to do a similar simulation like above($Bernoulli(0.5)$), but this time with a fixed $n = 10000$, which is big enough, and compare the distribution of sample means with a normal distribution with mean $p = 0.5$ and standard deviation $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$. These mean and standard deviation are decided by the central limit theorem. The central limit Theorem says that for samples of size 30 or more, the sample mean is approximately normally distributed, with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\rho_{\overline{X}} = \frac{\rho}{\sqrt{n}}$, where n is the sample size, and the larger the sample size, the better the approximation.

Reference: LibreTexts Statistics, https://stats.libretexts.org/ (https://stats.libretexts.org/)



We can see that these two distributions are very similar.

3. Hypothesis Testing

We are going to do some hypothesis Testings(significance level of **0.05**) using the data set called `SwissLabor`. This data set has 7 variables, `participation`, `income`, `age`, `education`, `youngkids`, `oldkids`, `foreign` and all observations are female. This survey is executed in 1981. We are going to only use the data of 90% of all observations. Here is the description of the variables:

- participation: Factor. Did the individual participate in the labor force?
- income: Logarithm of nonlabor income.
- age: Age in decades (years divided by 10).
- education: Years of formal education.
- youngkids: Number of young children (under 7 years of age).
- oldkids: Number of older children (over 7 years of age).
- foreign: Factor. Is the individual a foreigner (i.e., not Swiss)?

```
## 'data.frame':    785 obs. of  7 variables:
##  $ participation: Factor w/ 2 levels "no","yes": 2 2 1 1 1 1 2 2 1 1 ...
##  $ income       : num  10.53 10.45 10.62 9.67 10.36 ...
##  $ age          : num  4.1 2.1 3.9 4.5 4.3 2.8 3.9 3.2 5.3 5.2 ...
##  $ education    : num  2 8 12 12 5 12 12 17 10 4 ...
##  $ youngkids    : num  0 0 0 0 0 3 0 1 0 0 ...
##  $ oldkids      : num  2 0 0 1 2 1 2 0 0 1 ...
##  $ foreign      : Factor w/ 2 levels "no","yes": 2 1 1 1 2 1 1 1 1 1 ...
```

Test between `income` and `foreign`

```
##
##  Welch Two Sample t-test
##
## data:  income by foreign
## t = 6.5562, df = 401.98, p-value = 8.489e-11
## alternative hypothesis: true difference in means between group no and group yes is greater th
an 0
## 95 percent confidence interval:
##  0.1388583       Inf
## sample estimates:
##  mean in group no mean in group yes
##          10.73771          10.55220
```

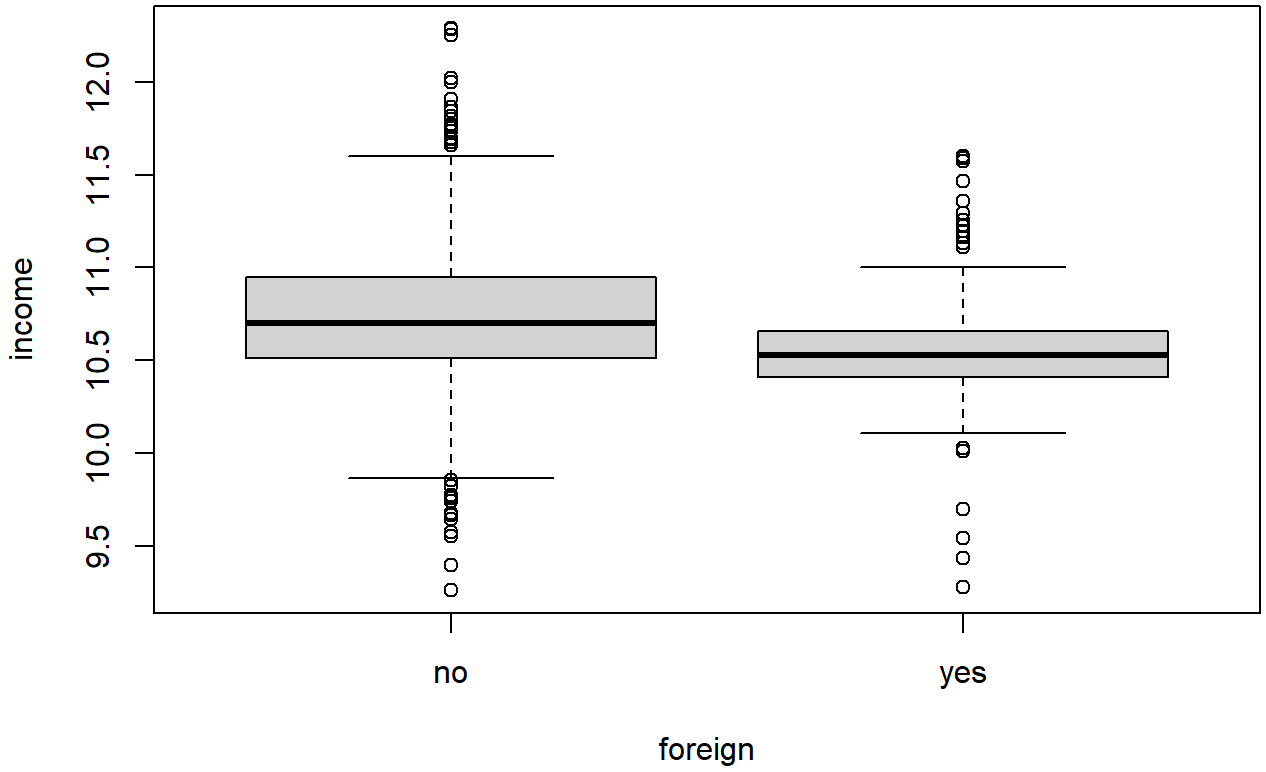
The null and alternative hypothesis for this test are as follows:

- H_0: The mean(income) in non-foreign group is less than or as same as the mean(income) in foreign group.
- H_1: The mean(income) in non-foreign group is greater than the mean(income) in foreign group.

The p-value = 8.489e-11 is smaller than 0.05, so we can reject the null hypothesis and accept the alternative hypothesis. Therefore, the mean(income) in non-foreign group is greater than the mean(income) in foreign group.

In my opinion, this are two reasons. First, foreigners could not earn money as much as citizens, so that foreigners could not invest as much as citizens. Second, it was difficult for foreigners to get age-related or hardship-related payments from the government.

In the following box plot we can also see that the result of this test trues in our data set.



Test between `income` and `participation`

```
##
##  Welch Two Sample t-test
##
## data:  income by participation
## t = 4.9082, df = 781.47, p-value = 5.595e-07
## alternative hypothesis: true difference in means between group no and group yes is greater th
an 0
## 95 percent confidence interval:
##  0.08828048        Inf
## sample estimates:
##  mean in group no mean in group yes
##          10.75238          10.61953
```

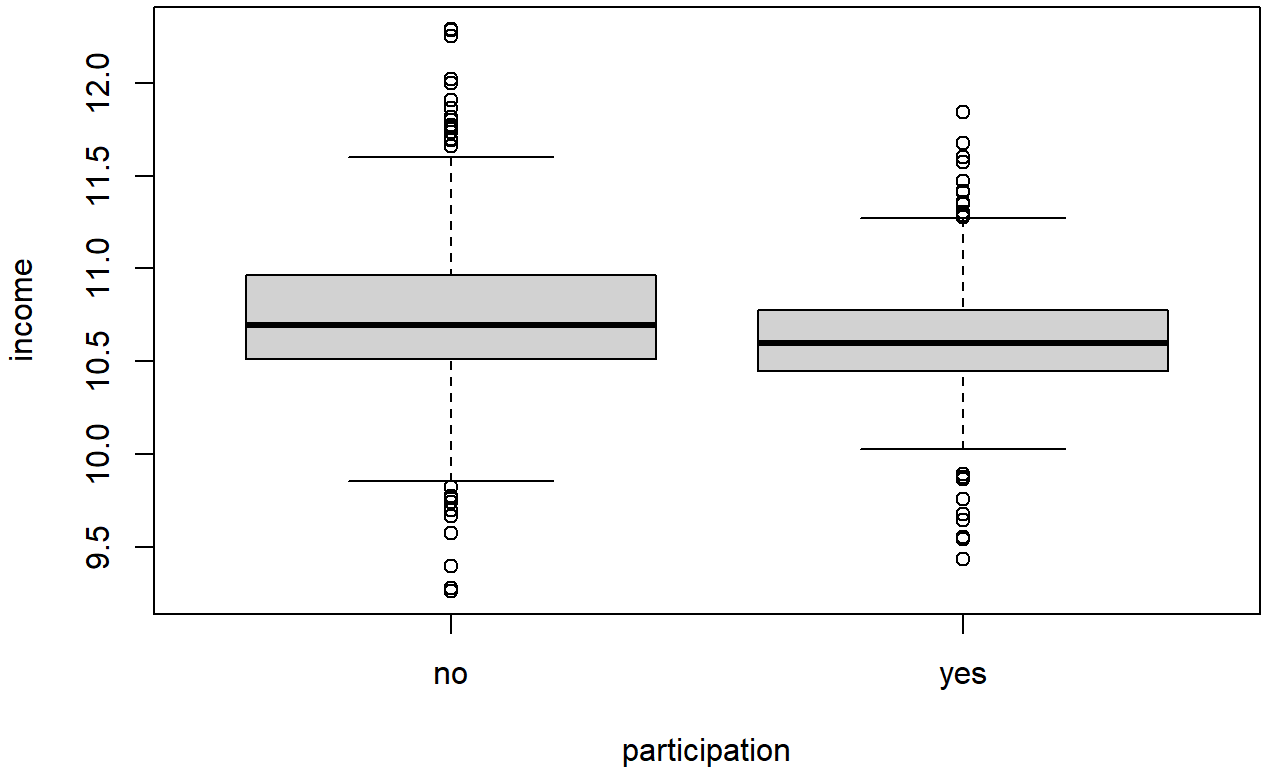
The null and alternative hypothesis for this test are as follows:

- H_0: The mean(income) in the group where people did not participate in the labor force is less than or as same as the mean(income) in the group where people participated in the labor force.
- H_1: The mean(income) in the group where people did not participate in the labor force is greater than the mean(income) in the group where people participated in the labor force.

The p-value = 5.595e-07 is smaller than 0.05, so we can reject the null hypothesis and accept the alternative hypothesis. Therefore, The mean(income) in the group where people did not participate in the labor force is greater than the mean(income) in the group where people participated in the labor force.

In my opinion, this is because the people who did not participate in the labor force could get more hardship-related payments(e.g., Medicaid & Unemployment) from the government.

In the following box plot we can also see that the result of this test trues in our data set.



Test between `income`, `youngkids` and `oldkids`

```
##
##  Welch Two Sample t-test
##
## data:  income_youngkids and income_oldkids
## t = -1.9766, df = 351.57, p-value = 0.04887
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1278033271 -0.0003201409
## sample estimates:
## mean of x mean of y
##  10.66716  10.73122
```

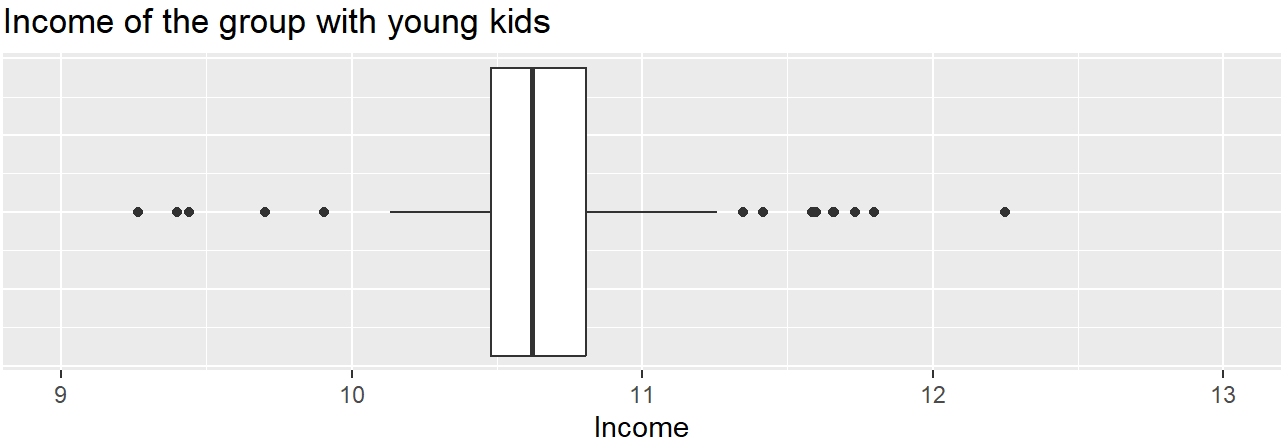
The null and alternative hypothesis for this test are as follows:

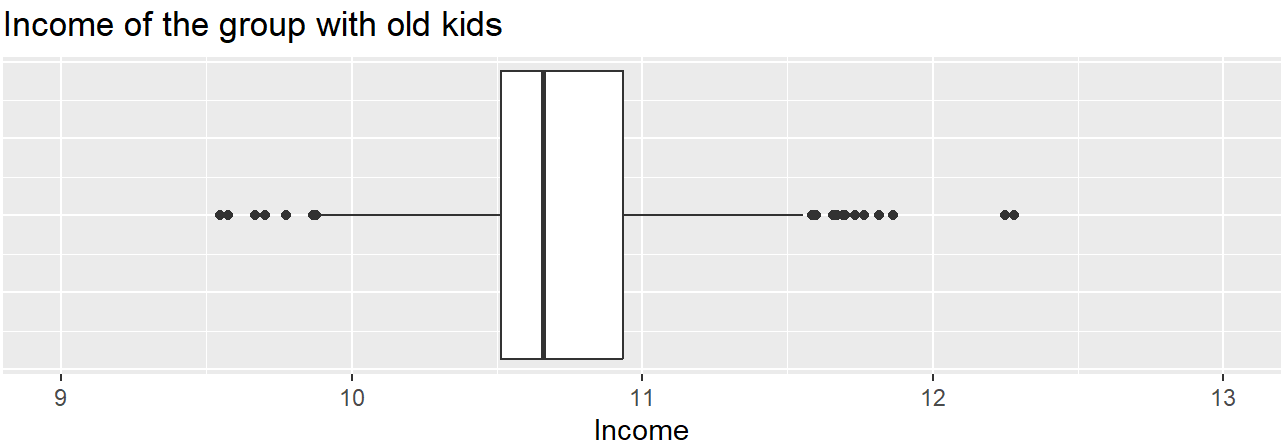
- H_0: There is no significant difference in the means of the two groups. One group had at least one young kid and the other group had at least one old kid.
- H_1: There is significant difference in the means of these two groups.

The p-value = 0.04887 is smaller than 0.05, so we can reject the null hypothesis and accept the alternative hypothesis. Therefore, there is significant difference in the means of these two groups.

In my opinion, it is more likely that the people who have at least one old kid are older than the people who have at least one young kid. It might affect the age-related payments(Medicare & Social Security) and investment income. The older people could have more work experience, so that they could earn more money and they could invest in something easier.

In the following box plots we can also see that the result of this test trues in our data set.





4. Regression Models

In this section, we are going to use the data set called `PSID1982` and only use the data of 90% of all observations. Here is the description of the variables:

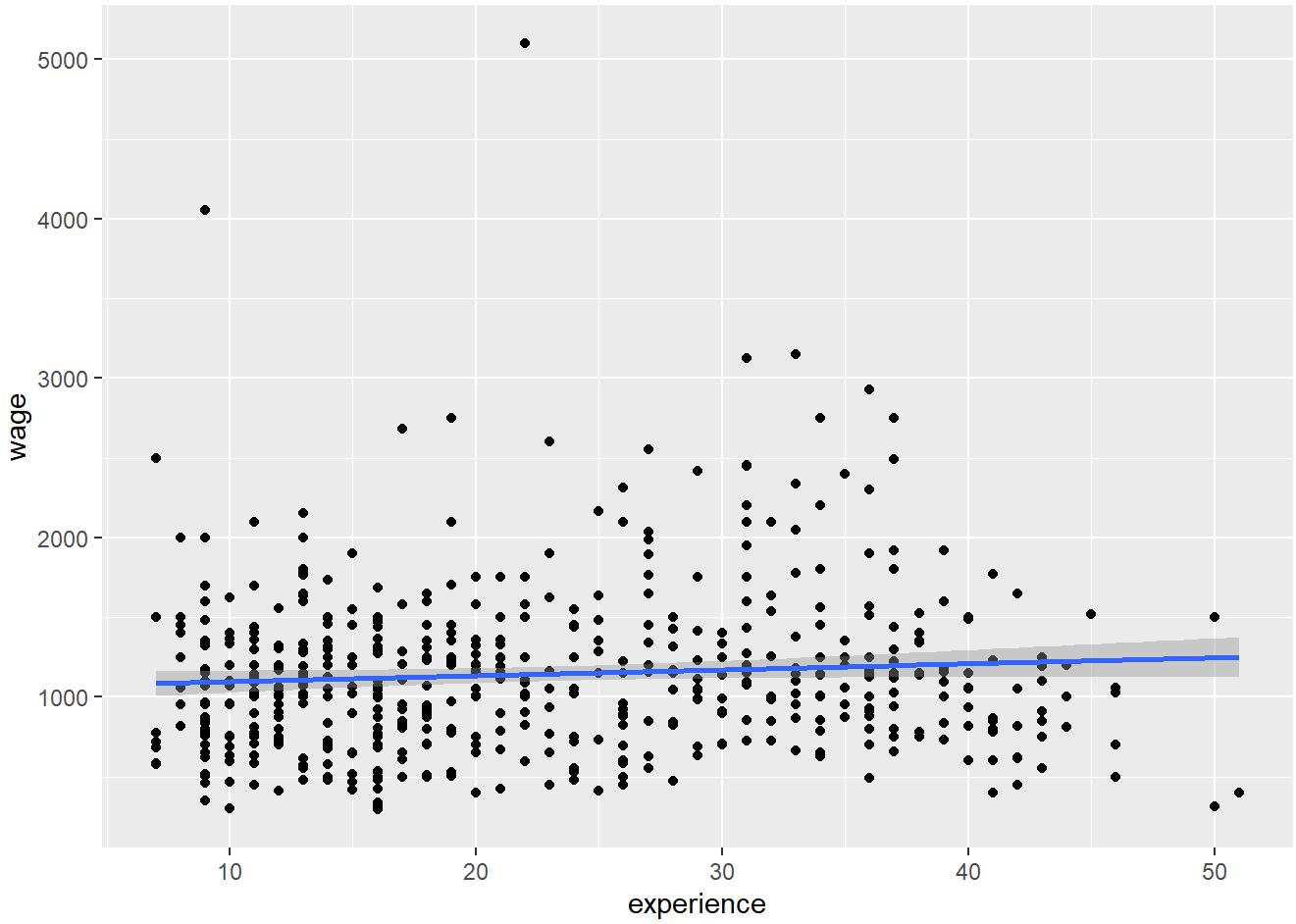
- experience: Years of full-time work experience.
- weeks: Weeks worked.
- occupation: factor. Is the individual a white-collar (“white”) or blue-collar (“blue”) worker?
- industry: factor. Does the individual work in a manufacturing industry?
- south: factor. Does the individual reside in the South?
- smsa: factor. Does the individual reside in a SMSA (standard metropolitan statistical area)?
- married: factor. Is the individual married?
- gender: factor indicating gender.
- union: factor. Is the individual’s wage set by a union contract?
- education: Years of education.

- ethnicity: factor indicating ethnicity. Is the individual African-American (“afam”) or not (“other”)?
- wage: Wage.

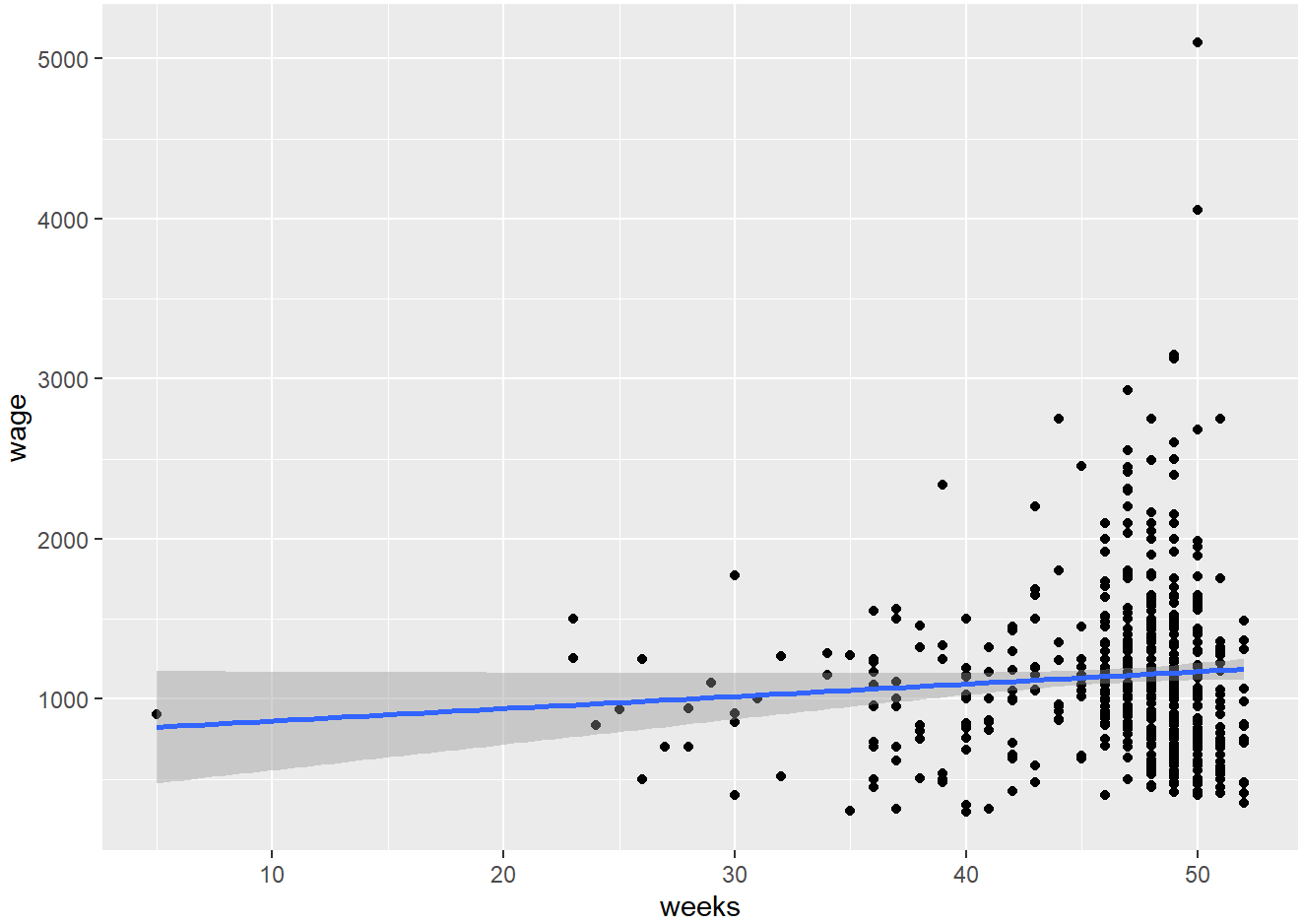
```
## 'data.frame':    536 obs. of  12 variables:
##  $ experience: int  33 9 19 30 15 14 16 27 29 14 ...
##  $ weeks     : int  41 48 49 49 48 45 49 47 47 49 ...
##  $ occupation: Factor w/ 2 levels "white","blue": 2 1 1 1 1 2 1 2 2 2 ...
##  $ industry  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 2 2 1 2 1 ...
##  $ south     : Factor w/ 2 levels "no","yes": 1 2 1 1 1 1 1 2 2 1 ...
##  $ smsa      : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 1 2 ...
##  $ married   : Factor w/ 2 levels "no","yes": 2 2 1 2 1 2 2 2 2 2 ...
##  $ gender    : Factor w/ 2 levels "male","female": 1 1 2 1 2 1 1 1 1 1 ...
##  $ union     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 1 1 1 2 ...
##  $ education : int  12 16 13 14 16 13 15 12 8 14 ...
##  $ ethnicity : Factor w/ 2 levels "other","afam": 1 1 1 1 1 2 1 1 2 1 ...
##  $ wage      : num  865 1350 775 900 1023 ...
```

Estimate the relationships between a dependent variable(`wage`) and independent variables

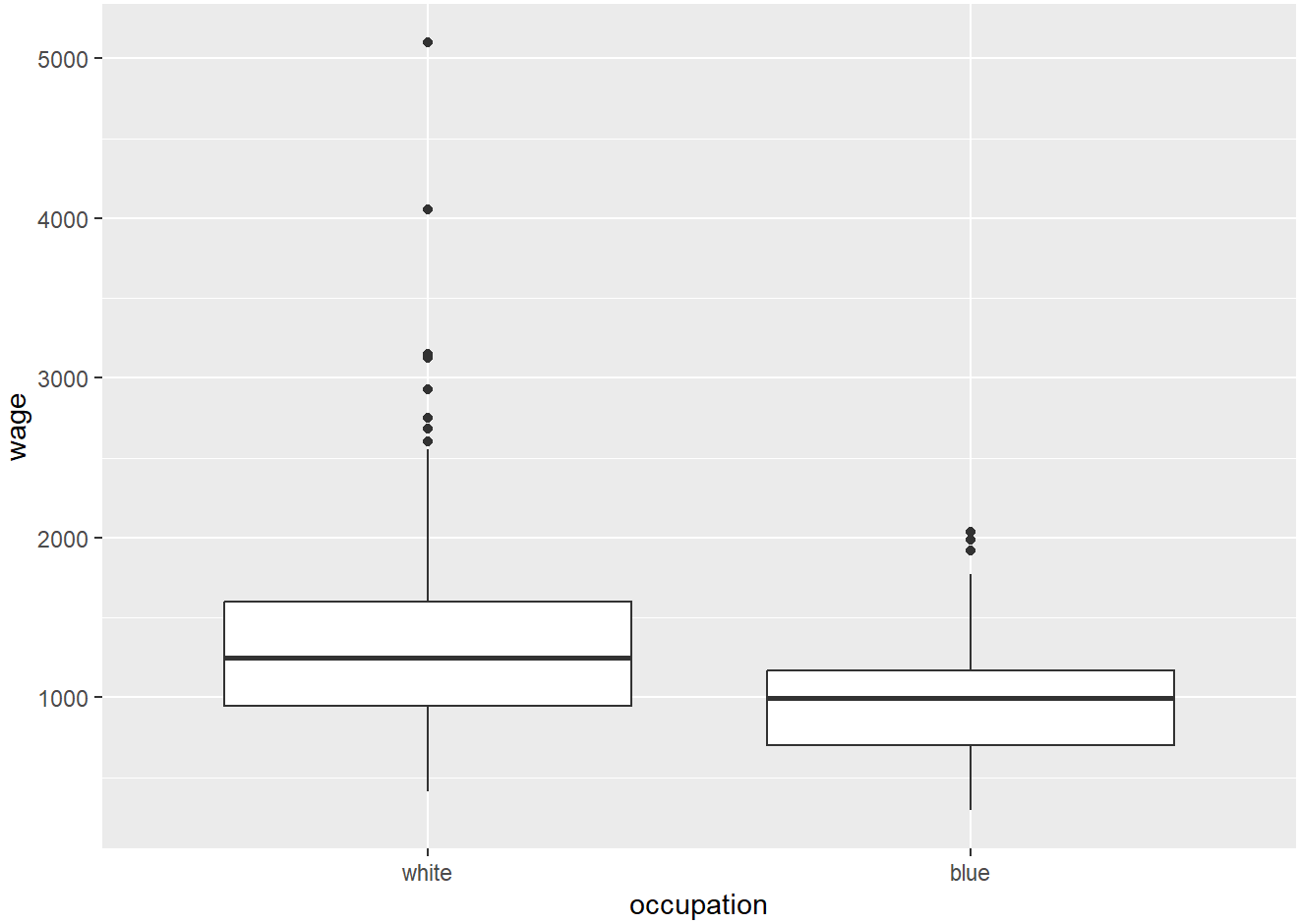
1. `wage` and `experience` : Surprisingly, `experience` seems to affect `wage` very slightly.



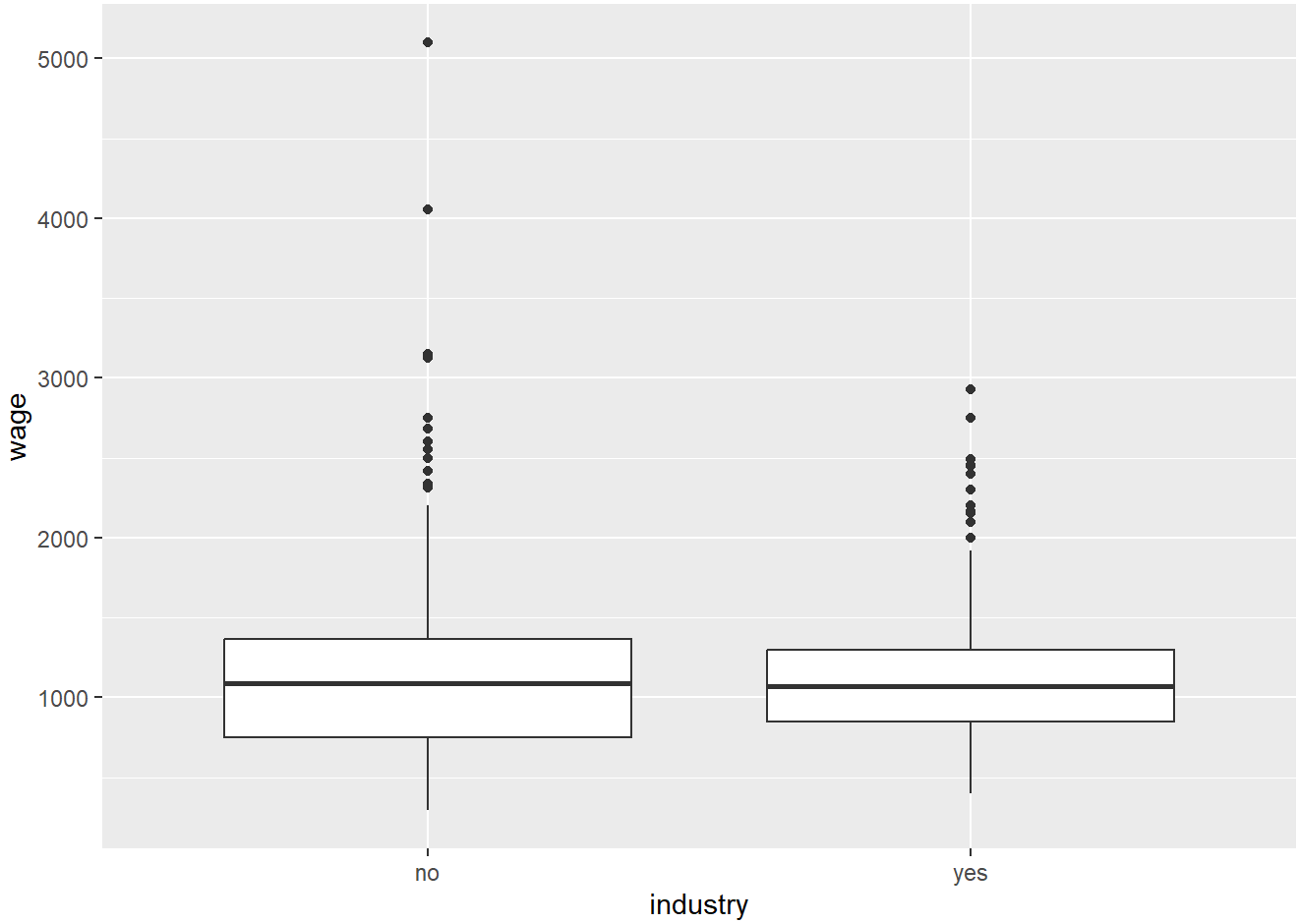
2. `wage` and `weeks` : `weeks` seems to affect `wage` slightly but more than `experience` affect `wage` .



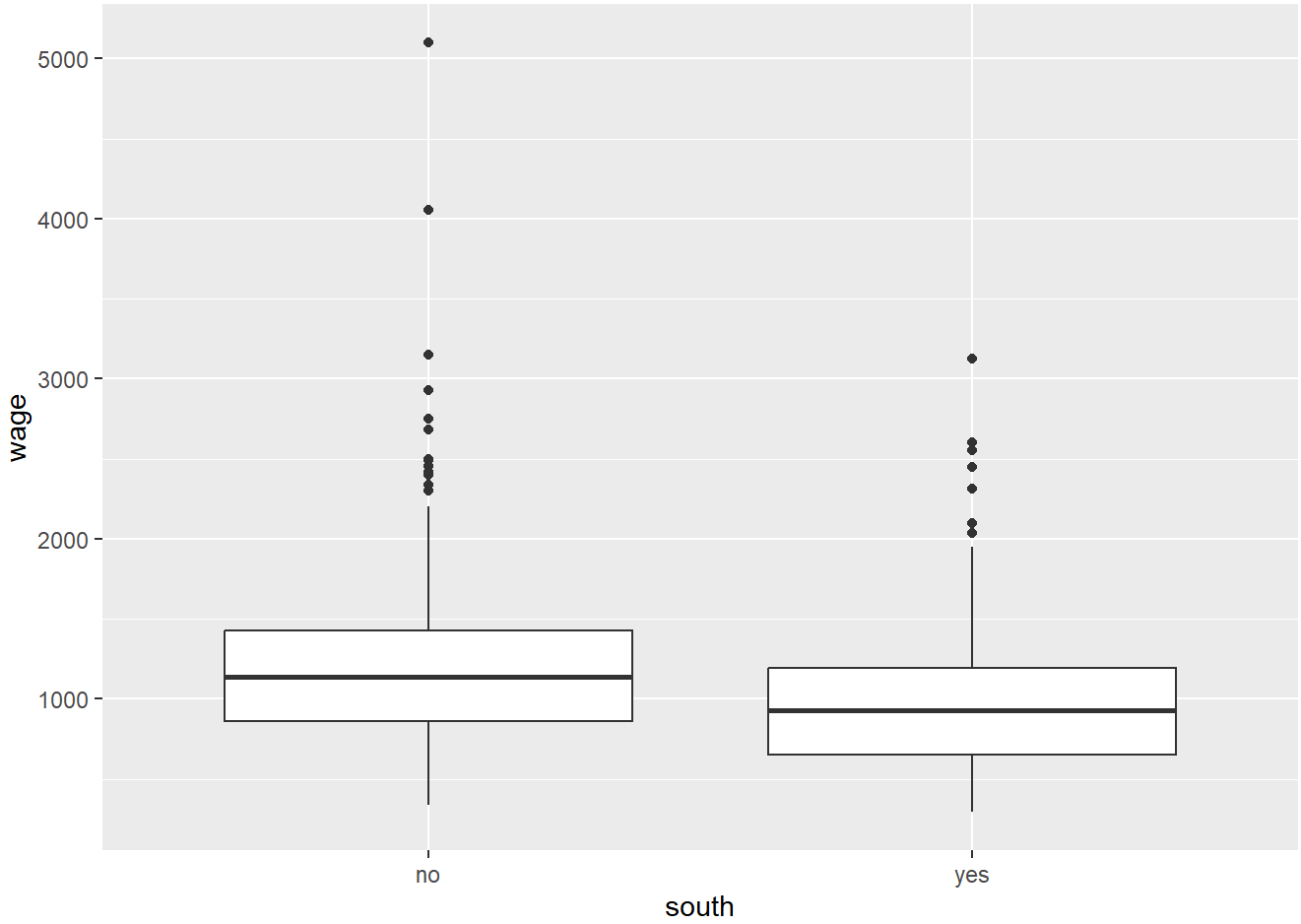
3. `occupation` and `wage` : In general, white-collar workers have higher wage.



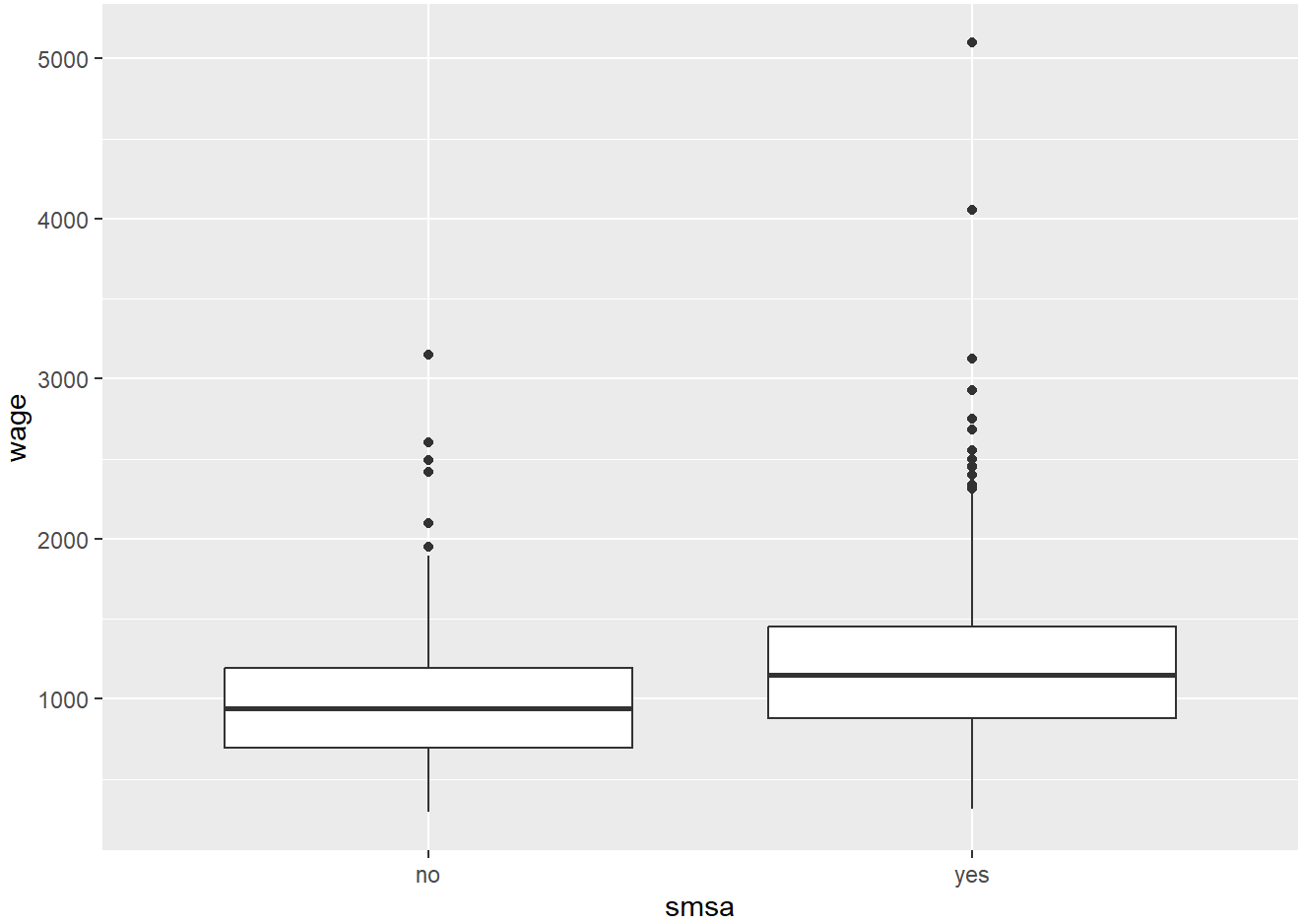
4. `industry` and `wage` : There is no significant difference in wage between the individual who works in an manufacturing industry(yes) and the individual who works in other industries.



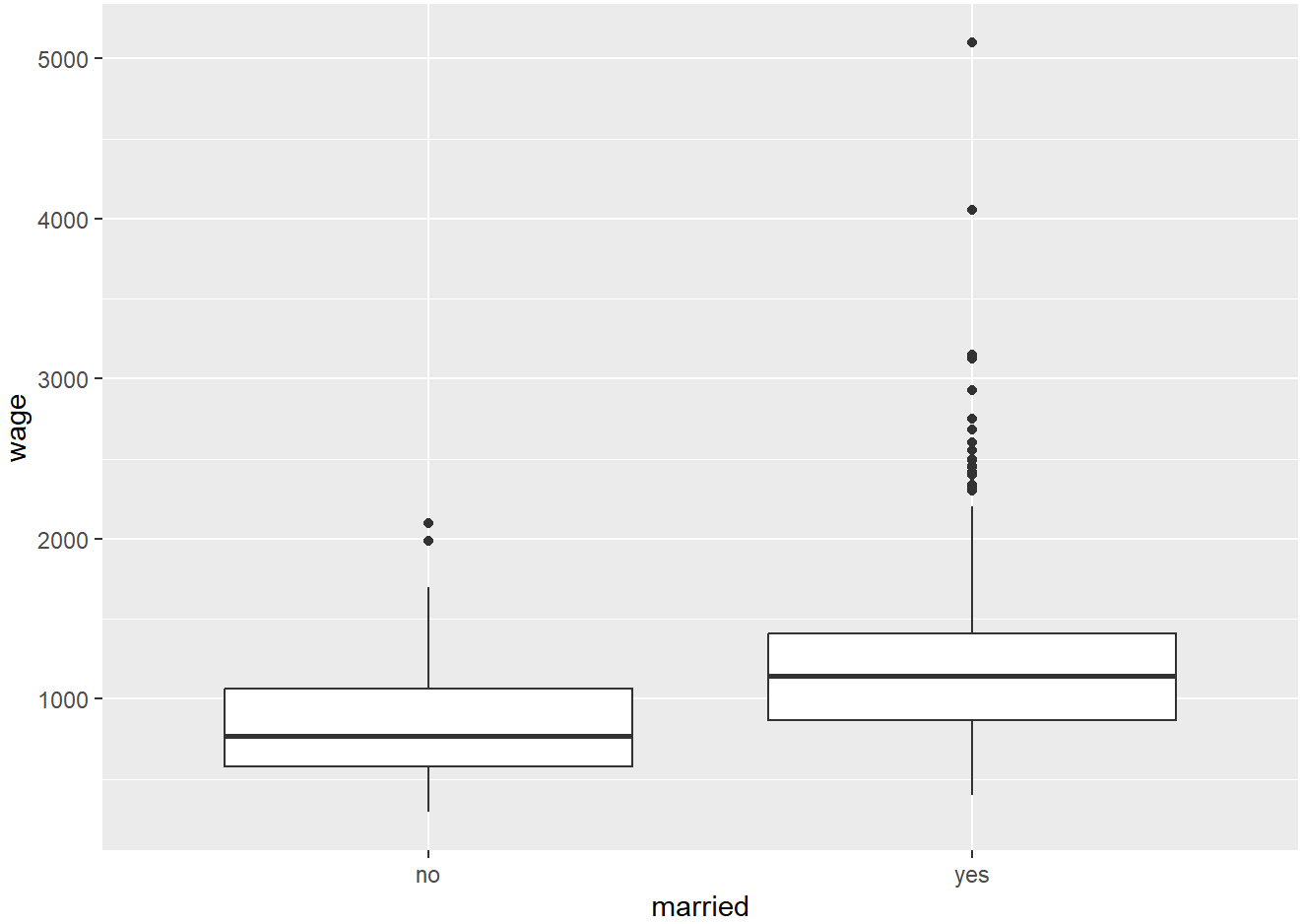
5. `south` and `wage` : The people who reside in the south tend to have lower wage than others.



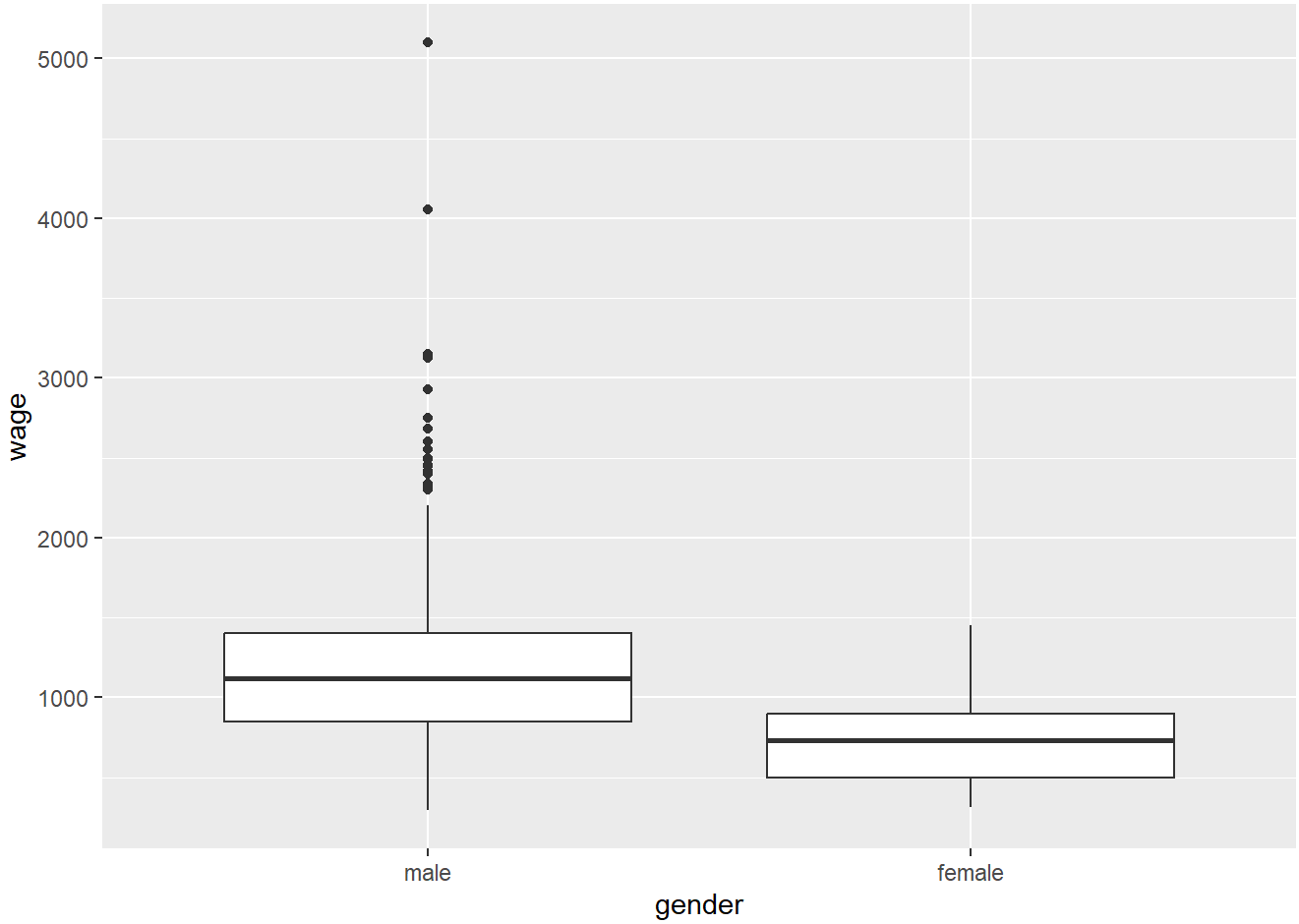
6. `smsa` and `wage` : The people who reside in a SMSA(standard metropolitan statistical area) tend to have higher wage than others.



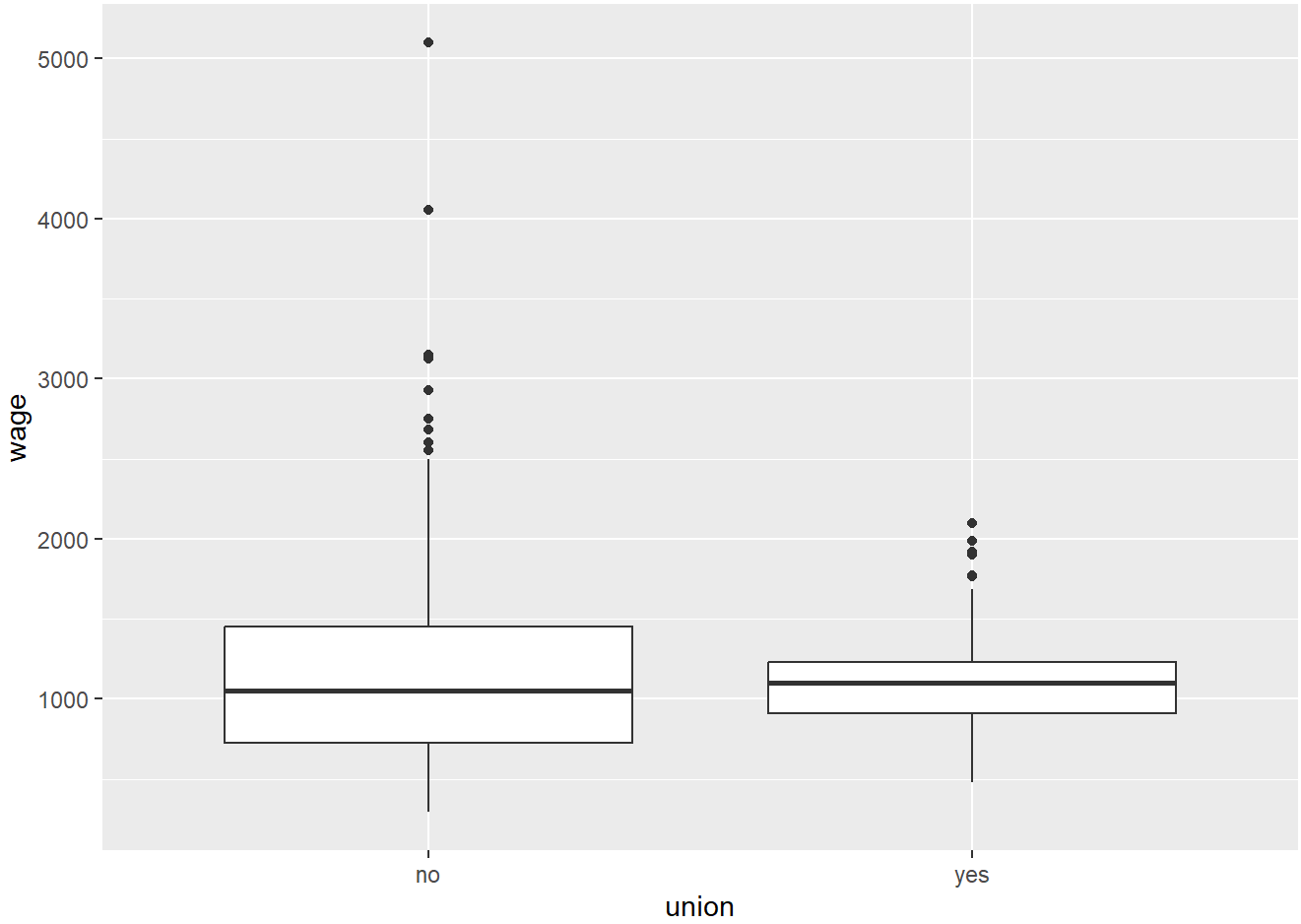
7. `married` and `wage` : In general, the married have higher wage than the unmarried.



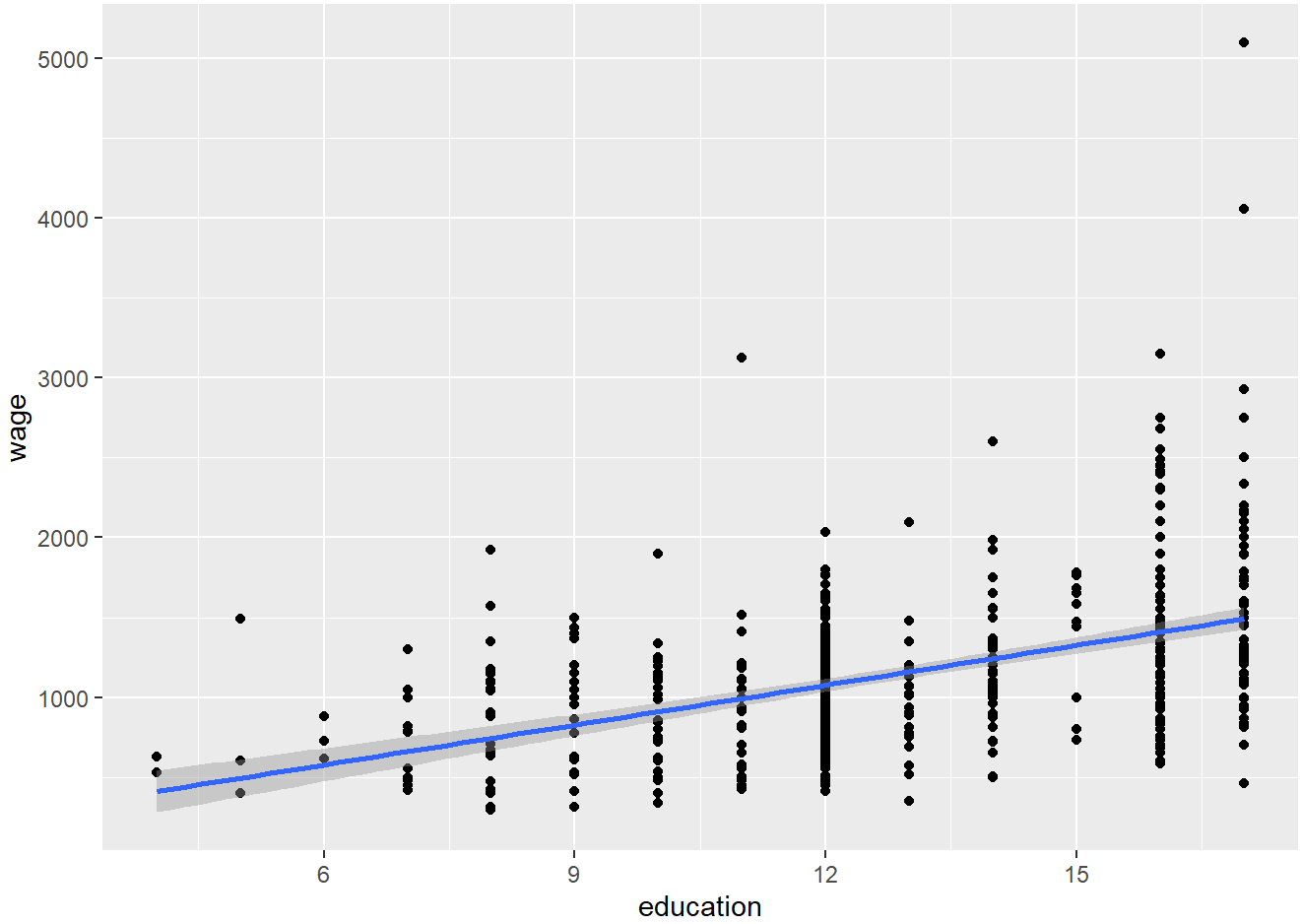
8. `gender` and `wage` : Males tend to have higher wage than females.



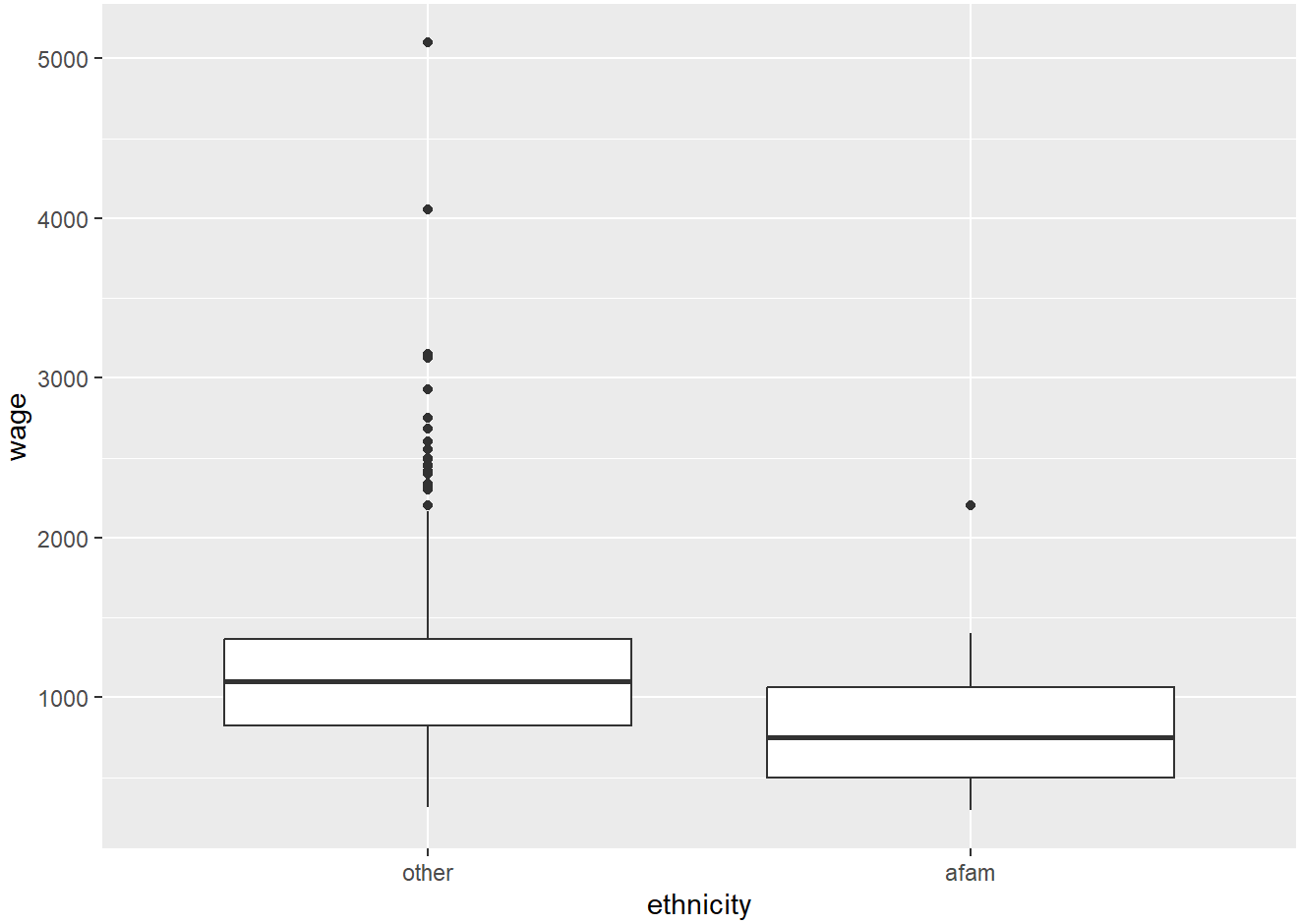
9. `union` and `wage` : There is no significant difference of wage between the individual whose wage is set by a union contract and the individual whose wage is not set by a union contract.



10. `education` and `wage` : In general, the longer a individual gets education, the higher wage a individual have.



11. `ethnicity` and `wage` : African-Americans(“afam”) tend to have lower wage than others.



Based on the information above, I am going to make a model `lm1` with variables: `weeks`, `occupation`, `south`, `smsa`, `married`, `gender`, `education`, `ethnicity`.

```
##
## Call:
## lm(formula = wage ~ weeks + occupation + south + smsa + married +
##     gender + education + ethnicity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1074.3  -273.4   -42.0   186.8  3462.6
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     221.508    223.549   0.991 0.322206
## weeks             3.221      3.532   0.912 0.362229
## occupationblue -179.299     48.937  -3.664 0.000274 ***
## southyes        -98.683     41.742  -2.364 0.018435 *
## smsayes         181.158     40.817   4.438 1.11e-05 ***
## marriedyes      189.323     68.364   2.769 0.005815 **
## genderfemale   -271.408     86.523  -3.137 0.001803 **
## education        52.022      8.780   5.925 5.65e-09 ***
## ethnicityafam  -163.879     75.156  -2.181 0.029661 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 428.2 on 527 degrees of freedom
## Multiple R-squared:  0.3406, Adjusted R-squared:  0.3306
## F-statistic: 34.02 on 8 and 527 DF,  p-value: < 2.2e-16
```

I am going to make new models `lm2` and `lm3`.

```
##
## Call:
## lm(formula = wage ~ weeks + I(weeks^2) + occupation + south +
##     smsa + married + gender + education + I(education^2) + ethnicity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1136.2  -263.0   -40.6   194.2  3399.9
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     439.5983   551.0580   0.798  0.42538
## weeks            31.6115    23.0097   1.374  0.17008
## I(weeks^2)       -0.3496     0.2904  -1.204  0.22919
## occupationblue -139.2877    50.1374  -2.778  0.00566 **
## southyes       -111.1560    41.5512  -2.675  0.00770 **
## smsayes         185.6263    40.4861   4.585 5.68e-06 ***
## marriedyes      184.2220    67.8109   2.717  0.00681 **
## genderfemale   -271.3482    86.0955  -3.152  0.00172 **
## education       -86.3530    46.8713  -1.842  0.06599 .
## I(education^2)    5.7160     1.9037   3.003  0.00280 **
## ethnicityafam  -161.4795    74.7361  -2.161  0.03117 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 424.5 on 525 degrees of freedom
## Multiple R-squared:  0.3544, Adjusted R-squared:  0.3421
## F-statistic: 28.82 on 10 and 525 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = wage ~ weeks + I(weeks^2) + I(weeks^3) + occupation +
##     south + smsa + married + gender + education + I(education^2) +
##     I(education^3) + ethnicity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1186.9  -259.3   -34.3   197.2  3389.5
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     282.81717 1033.82877   0.274  0.78453
## weeks          -116.58435   62.08705  -1.878  0.06097 .
## I(weeks^2)        4.55814    1.94028   2.349  0.01918 *
## I(weeks^3)       -0.04879    0.01916  -2.547  0.01116 *
## occupationblue -147.78496   49.90299  -2.961  0.00320 **
## southyes       -108.35505   41.27944  -2.625  0.00892 **
## smsayes         176.81424   40.32164   4.385  1.4e-05 ***
## marriedyes      173.82902   67.43381   2.578  0.01022 *
## genderfemale   -279.78499   85.54176  -3.271  0.00114 **
## education       326.45779  230.32667   1.417  0.15697
## I(education^2)  -32.03321   20.66655  -1.550  0.12175
## I(education^3)    1.08244    0.58948   1.836  0.06689 .
## ethnicityafam  -170.74620   74.27659  -2.299  0.02191 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 421.5 on 523 degrees of freedom
## Multiple R-squared:  0.3658, Adjusted R-squared:  0.3512
## F-statistic: 25.14 on 12 and 523 DF,  p-value: < 2.2e-16
```

I am going to make a new model `lm.all` which also uses all the other variables in the data set.

```
##
## Call:
## lm(formula = wage ~ . + I(weeks^2) + I(weeks^3) + I(education^2) +
##     I(education^3), data = PSID1982.90percent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1110.1  -261.5   -40.6   197.1  3395.7
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     104.49454 1028.61369   0.102  0.91912
## experience        5.80987    1.84198   3.154  0.00170 **
## weeks          -104.25578   61.97408  -1.682  0.09312 .
## occupationblue -149.93828   51.61253  -2.905  0.00383 **
## industryyes      43.58641   39.99816   1.090  0.27635
## southyes        -90.63584   41.79866  -2.168  0.03058 *
## smsayes         161.93254   40.39799   4.008 7.01e-05 ***
## marriedyes      126.87474   68.26012   1.859  0.06364 .
## genderfemale   -278.33033   85.39695  -3.259  0.00119 **
## unionyes         28.63037   44.13764   0.649  0.51684
## education       281.97431  229.03046   1.231  0.21882
## ethnicityafam  -174.15256   73.86271  -2.358  0.01875 *
## I(weeks^2)        4.05286    1.94404   2.085  0.03758 *
## I(weeks^3)       -0.04274    0.01927  -2.218  0.02701 *
## I(education^2)  -26.53894   20.57038  -1.290  0.19757
## I(education^3)    0.91242    0.58690   1.555  0.12064
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 418 on 520 degrees of freedom
## Multiple R-squared:  0.3799, Adjusted R-squared:  0.362
## F-statistic: 21.24 on 15 and 520 DF,  p-value: < 2.2e-16
```

Now we are going to use StepAIC function to identify the best subset of predictor variables for this model `lm.all` .

```
library(MASS)
lm.step <- stepAIC(lm.all)
```

```
## Start:  AIC=6485.85
## wage ~ experience + weeks + occupation + industry + south + smsa +
##     married + gender + union + education + ethnicity + I(weeks^2) +
##     I(weeks^3) + I(education^2) + I(education^3)
##
##                  Df Sum of Sq      RSS    AIC
## - union           1     73525 90939243 6484.3
## - industry        1    207500 91073219 6485.1
## - education       1    264868 91130586 6485.4
## - I(education^2)  1    290857 91156575 6485.6
## <none>                        90865718 6485.8
## - I(education^3)  1    422334 91288053 6486.3
## - weeks           1    494512 91360230 6486.8
## - married         1    603689 91469407 6487.4
## - I(weeks^2)      1    759467 91625186 6488.3
## - south           1    821622 91687340 6488.7
## - I(weeks^3)      1    859326 91725044 6488.9
## - ethnicity       1    971418 91837136 6489.5
## - occupation      1   1474726 92340445 6492.5
## - experience      1   1738440 92604159 6494.0
## - gender          1   1856235 92721954 6494.7
## - smsa            1   2807666 93673385 6500.2
##
## Step:  AIC=6484.28
## wage ~ experience + weeks + occupation + industry + south + smsa +
##     married + gender + education + ethnicity + I(weeks^2) + I(weeks^3) +
##     I(education^2) + I(education^3)
##
##                  Df Sum of Sq      RSS    AIC
## - industry        1    216747 91155990 6483.6
## - education       1    279267 91218510 6483.9
## - I(education^2)  1    303833 91243075 6484.1
## <none>                        90939243 6484.3
## - I(education^3)  1    434690 91373932 6484.8
## - weeks           1    527481 91466724 6485.4
## - married         1    629338 91568581 6486.0
## - I(weeks^2)      1    821876 91761119 6487.1
## - south           1    927369 91866611 6487.7
## - I(weeks^3)      1    947151 91886394 6487.8
## - ethnicity       1    981431 91920674 6488.0
## - occupation      1   1402434 92341677 6490.5
## - experience      1   1701276 92640519 6492.2
## - gender          1   1933452 92872695 6493.6
## - smsa            1   2948823 93888065 6499.4
##
## Step:  AIC=6483.56
## wage ~ experience + weeks + occupation + south + smsa + married +
##     gender + education + ethnicity + I(weeks^2) + I(weeks^3) +
##     I(education^2) + I(education^3)
##
##                  Df Sum of Sq      RSS    AIC
## - education       1    273812 91429803 6483.2
```

```
## - I(education^2)  1    303509 91459499 6483.3
## <none>                       91155990 6483.6
## - I(education^3)  1    436603 91592593 6484.1
## - weeks           1    588969 91744959 6485.0
## - married         1    672445 91828435 6485.5
## - I(weeks^2)      1    905021 92061011 6486.9
## - south           1   1034992 92190983 6487.6
## - I(weeks^3)      1   1038987 92194977 6487.6
## - ethnicity       1   1042454 92198444 6487.7
## - occupation      1   1297653 92453644 6489.1
## - experience      1   1782167 92938158 6491.9
## - gender          1   2034441 93190431 6493.4
## - smsa            1   2993829 94149819 6498.9
##
## Step:  AIC=6483.16
## wage ~ experience + weeks + occupation + south + smsa + married +
##     gender + ethnicity + I(weeks^2) + I(weeks^3) + I(education^2) +
##     I(education^3)
##
##                  Df Sum of Sq      RSS    AIC
## - I(education^2)  1     72189 91501992 6481.6
## <none>                       91429803 6483.2
## - I(education^3)  1    532060 91961862 6484.3
## - weeks           1    550165 91979968 6484.4
## - married         1    675948 92105751 6485.1
## - I(weeks^2)      1    855616 92285419 6486.2
## - I(weeks^3)      1    988500 92418303 6486.9
## - ethnicity       1   1014421 92444224 6487.1
## - south           1   1051223 92481026 6487.3
## - occupation      1   1198717 92628519 6488.1
## - experience      1   1865347 93295149 6492.0
## - gender          1   2008056 93437859 6492.8
## - smsa            1   3041070 94470872 6498.7
##
## Step:  AIC=6481.59
## wage ~ experience + weeks + occupation + south + smsa + married +
##     gender + ethnicity + I(weeks^2) + I(weeks^3) + I(education^3)
##
##                  Df Sum of Sq       RSS    AIC
## <none>                         91501992 6481.6
## - weeks           1    530434  92032427 6482.7
## - married         1    664154  92166146 6483.5
## - I(weeks^2)      1    836120  92338112 6484.5
## - I(weeks^3)      1    972881  92474873 6485.3
## - south           1    999713  92501706 6485.4
## - ethnicity       1   1025898  92527890 6485.6
## - occupation      1   1202222  92704214 6486.6
## - gender          1   2036806  93538798 6491.4
## - experience      1   2066348  93568340 6491.6
## - smsa            1   2990514  94492506 6496.8
## - I(education^3)  1   9438704 100940696 6532.2
```

```
##
## Call:
## lm(formula = wage ~ experience + weeks + occupation + south +
##     smsa + married + gender + ethnicity + I(weeks^2) + I(weeks^3) +
##     I(education^3), data = PSID1982.90percent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1083.9  -254.5   -35.7   190.8  3416.1
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1044.71155  657.24248   1.590 0.112542
## experience        6.19849    1.80191   3.440 0.000628 ***
## weeks          -106.99023   61.38721  -1.743 0.081942 .
## occupationblue -129.62825   49.40341  -2.624 0.008947 **
## southyes        -97.36320   40.69178  -2.393 0.017076 *
## smsayes         165.53600   40.00085   4.138 4.08e-05 ***
## marriedyes      132.61803   68.00139   1.950 0.051682 .
## genderfemale   -289.57658   84.78876  -3.415 0.000687 ***
## ethnicityafam  -178.50355   73.64519  -2.424 0.015695 *
## I(weeks^2)        4.20082    1.91977   2.188 0.029097 *
## I(weeks^3)       -0.04479    0.01898  -2.360 0.018622 *
## I(education^3)    0.13245    0.01801   7.352 7.56e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 417.9 on 524 degrees of freedom
## Multiple R-squared:  0.3756, Adjusted R-squared:  0.3625
## F-statistic: 28.65 on 11 and 524 DF,  p-value: < 2.2e-16
```

In-Sample

We are going to compare the values of AIC, BIC, RSS and R2 of all models that we created.

Modell	AIC	BIC	RSS	R2
lm1	8027.942	8070.783	96634617	0.3405652
lm2	8020.583	8071.993	94608244	0.3543932
lm3	8015.037	8075.015	92938158	0.3657899
lm.all	8008.949	8081.779	90865718	0.3799322
lm.step	8004.689	8060.383	91501992	0.3755903

In terms of AIC and BIC the model `lm.step` is the best and in terms of RSS and R2 the model `lm.all` is the best. Among `lm1`, `lm2` and `lm3`, `lm1` is the best model in terms of BIC, RSS and R2 and `lm3` is the best model in terms of AIC.

F-Test

We are going to perform F-test between `lm.step` and `lm.all` to see if there is a significant difference between these two models.

```
## Analysis of Variance Table
##
## Model 1: wage ~ experience + weeks + occupation + south + smsa + married +
##     gender + ethnicity + I(weeks^2) + I(weeks^3) + I(education^3)
## Model 2: wage ~ experience + weeks + occupation + industry + south + smsa +
##     married + gender + union + education + ethnicity + I(weeks^2) +
##     I(weeks^3) + I(education^2) + I(education^3)
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1    524 91501992
## 2    520 90865718  4    636274 0.9103 0.4576
```

The p value is much bigger than 0.05, so we cannot reject the null hypothesis. It means that there is no significant difference between these two models.

Out-of-Sample Validation

For out-of-sample validation we are going to separate our sample into two samples. one(for training) is 80% and the other one(for test) is 20%.

Estimate `lm1`, `lm2`, `lm3`, `lm.all`, `lm.step` only using the data for training(80%).

```
## Start:  AIC=5181.6
## wage ~ experience + weeks + occupation + industry + south + smsa +
##     married + gender + union + education + ethnicity + I(weeks^2) +
##     I(weeks^3) + I(education^2) + I(education^3)
##
##                  Df Sum of Sq      RSS    AIC
## - union           1      7583 70090557 5179.6
## - I(education^2)  1     98650 70181624 5180.2
## - education       1     99850 70182824 5180.2
## - I(education^3)  1    156752 70239726 5180.6
## - industry        1    280168 70363143 5181.3
## - weeks           1    292351 70375325 5181.4
## - married         1    305888 70388862 5181.5
## <none>                        70082974 5181.6
## - south           1    460017 70542991 5182.4
## - I(weeks^2)      1    460660 70543634 5182.4
## - I(weeks^3)      1    521893 70604867 5182.8
## - ethnicity       1    523847 70606821 5182.8
## - occupation      1   1327049 71410024 5187.6
## - experience      1   1908944 71991918 5191.1
## - gender          1   1970902 72053876 5191.5
## - smsa            1   2083100 72166074 5192.2
##
## Step:  AIC=5179.65
## wage ~ experience + weeks + occupation + industry + south + smsa +
##     married + gender + education + ethnicity + I(weeks^2) + I(weeks^3) +
##     I(education^2) + I(education^3)
##
##                  Df Sum of Sq      RSS    AIC
## - I(education^2)  1    101622 70192180 5178.3
## - education       1    103523 70194080 5178.3
## - I(education^3)  1    159634 70250192 5178.6
## - industry        1    284358 70374915 5179.4
## - weeks           1    302928 70393485 5179.5
## - married         1    310809 70401366 5179.5
## <none>                        70090557 5179.6
## - I(weeks^2)      1    481572 70572129 5180.6
## - south           1    485422 70575979 5180.6
## - ethnicity       1    521968 70612526 5180.8
## - I(weeks^3)      1    551710 70642267 5181.0
## - occupation      1   1357842 71448399 5185.9
## - experience      1   1902122 71992680 5189.1
## - gender          1   2005815 72096372 5189.8
## - smsa            1   2127508 72218065 5190.5
##
## Step:  AIC=5178.27
## wage ~ experience + weeks + occupation + industry + south + smsa +
##     married + gender + education + ethnicity + I(weeks^2) + I(weeks^3) +
##     I(education^3)
##
##                  Df Sum of Sq      RSS    AIC
## - education       1      1964 70194143 5176.3
```

```
## - weeks           1    291522 70483702 5178.0
## - industry        1    299678 70491858 5178.1
## - married         1    318496 70510675 5178.2
## <none>                        70192180 5178.3
## - I(weeks^2)      1    470172 70662352 5179.1
## - south           1    484565 70676745 5179.2
## - ethnicity       1    505544 70697724 5179.3
## - I(weeks^3)      1    543249 70735429 5179.6
## - I(education^3)  1    860559 71052739 5181.5
## - occupation      1   1291717 71483896 5184.1
## - gender          1   1979471 72171651 5188.2
## - experience      1   1992447 72184627 5188.3
## - smsa            1   2119069 72311249 5189.0
##
## Step:  AIC=5176.28
## wage ~ experience + weeks + occupation + industry + south + smsa +
##     married + gender + ethnicity + I(weeks^2) + I(weeks^3) +
##     I(education^3)
##
##                  Df Sum of Sq      RSS    AIC
## - weeks           1    295603 70489746 5176.1
## - industry        1    297731 70491875 5176.1
## - married         1    320249 70514392 5176.2
## <none>                        70194143 5176.3
## - I(weeks^2)      1    475051 70669195 5177.2
## - south           1    497547 70691690 5177.3
## - ethnicity       1    503584 70697727 5177.3
## - I(weeks^3)      1    547875 70742019 5177.6
## - occupation      1   1291873 71486016 5182.1
## - gender          1   1978997 72173140 5186.2
## - experience      1   2058181 72252324 5186.7
## - smsa            1   2153336 72347479 5187.2
## - I(education^3)  1   8633803 78827946 5224.0
##
## Step:  AIC=5176.08
## wage ~ experience + occupation + industry + south + smsa + married +
##     gender + ethnicity + I(weeks^2) + I(weeks^3) + I(education^3)
##
##                  Df Sum of Sq      RSS    AIC
## <none>                        70489746 5176.1
## - married         1    343186 70832932 5176.2
## - industry        1    352218 70841964 5176.2
## - I(weeks^3)      1    403781 70893527 5176.5
## - ethnicity       1    477373 70967120 5177.0
## - I(weeks^2)      1    485439 70975185 5177.0
## - south           1    516406 71006152 5177.2
## - occupation      1   1289624 71779371 5181.9
## - gender          1   1964436 72454182 5185.9
## - experience      1   2044963 72534709 5186.4
## - smsa            1   2250247 72739994 5187.6
## - I(education^3)  1   8723486 79213233 5224.1
```

Evaluate the values of test MSE using the data for test(20%).

The table comparing the models with respect to AIC, BIC, RSS and MSE is as follows:

Modell	AIC.train	BIC.train	RSS.train	MSE.train	MSE.test
lm1.train	6415.400	6456.015	74870693	174523.8	205676.9
lm2.train	6410.769	6459.507	73379441	171047.6	201091.5
lm3.train	6408.660	6465.521	72341905	168629.2	195412.1
lm.step.train	6395.534	6448.333	70489746	164311.8	200817.7

In terms of test MSE, the model `lm3` is the best.

Discussion

When I chose the independent variables for the model `lm1` , I did not include the variable `experience` , because it seemed like that the variable `experience` do not affect the dependent variable `wage` . But it was interesting that the variable `experience` was included with a very small p value, when I use stepAIC function to the model which has all the variables in the data set(PSID1982). In terms of the test MSE, the model `lm3` looks the best model now, but it would be much better to evaluate the test MSE also with other test samples.