Università Cattolica del Sacro Cuore

Professor Dimitris Fouskakis

Home Assignment

Applied Statistics and Big Data

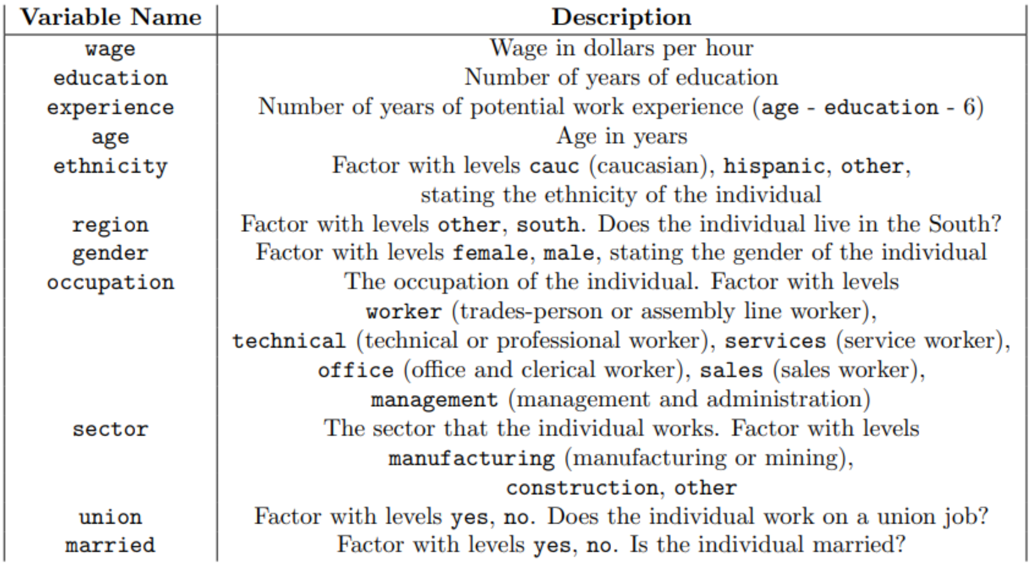
Chiara Lentsch

4901570

chiara.lentsch01@icatt.it

Exercise 1

We will use a sample of the Wage dataset, originating from the May 1985 Current Population Survey by the US Census Bureau. It contains data about 534 individuals (534 rows) and the following variables:

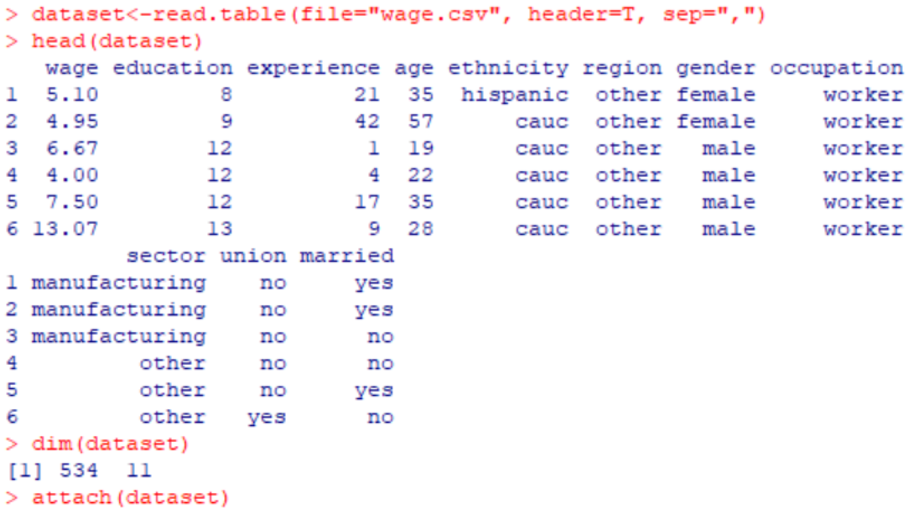


Variable Name	Description
wage	Wage in dollars per hour
education	Number of years of education
experience	Number of years of potential work experience (age - education - 6)
age	Age in years
ethnicity	Factor with levels cauc (caucasian), hispanic, other, stating the ethnicity of the individual
region	Factor with levels other, south. Does the individual live in the South?
gender	Factor with levels female, male, stating the gender of the individual
occupation	The occupation of the individual. Factor with levels worker (trades-person or assembly line worker), technical (technical or professional worker), services (service worker), office (office and clerical worker), sales (sales worker), management (management and administration)
sector	The sector that the individual works. Factor with levels manufacturing (manufacturing or mining), construction, other
union	Factor with levels yes, no. Does the individual work on a union job?
married	Factor with levels yes, no. Is the individual married?

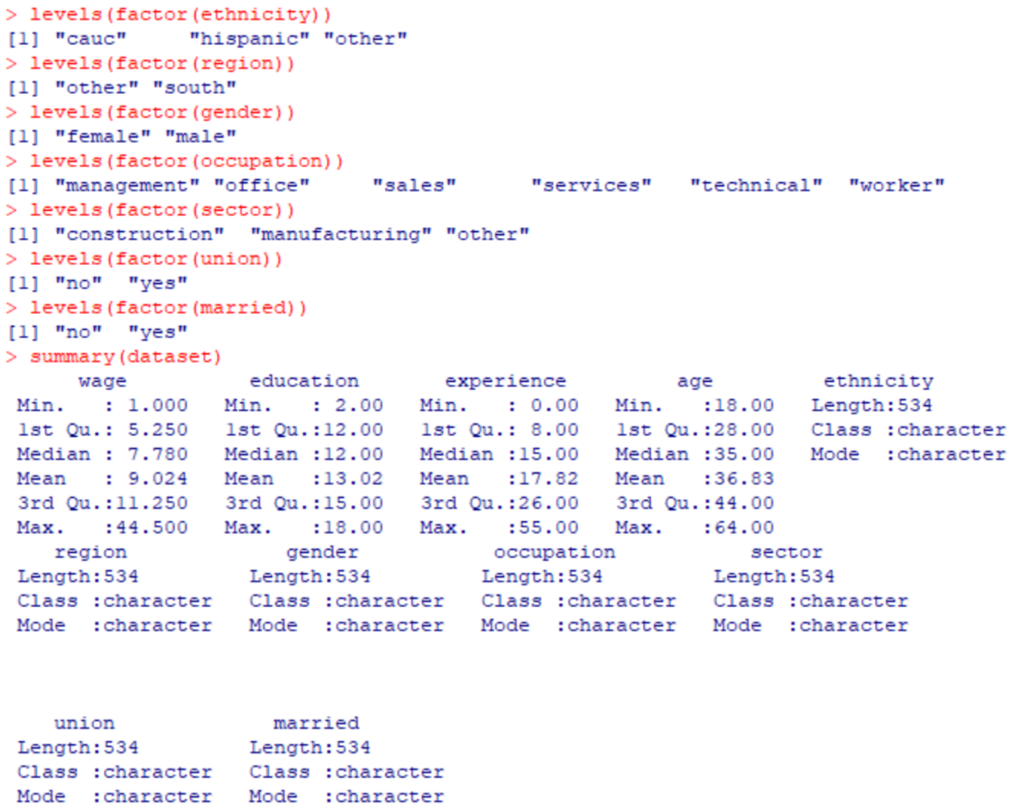
We start by loading our data (contained into wage.csv) into R making sure that we are working from the same directory where the file has been saved. Subsequently, we check the dimension to verify that the data have been read correctly, i.e., that the number of rows and columns is exact.

Since most of our variables are categorical-that is, they are not numerical, but they describe data that fit into categories- we need to factorize them. Converting categorical variables to factor variables allows us to use them in statistical modelling in a correct way. Before factorizing we attach our dataset to make objects accessible in R with fewer keystrokes.

Finally, applying the function summary() we get descriptive statistics of the dataset.

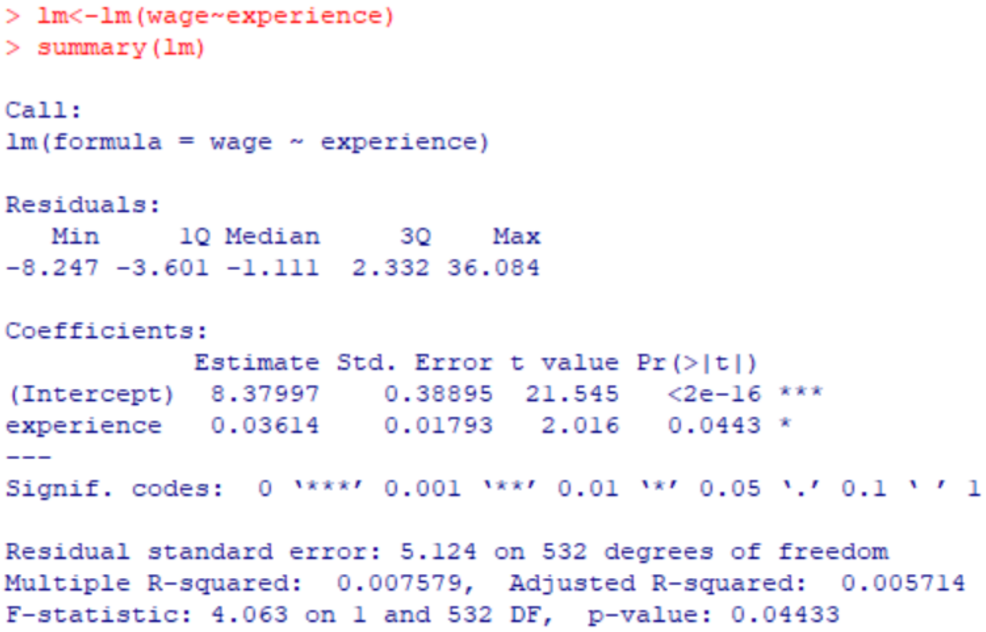


```
> dataset<-read.table(file="wage.csv", header=T, sep=",")
> head(dataset)
   wage education experience age ethnicity region gender occupation
1  5.10         8         21  35  hispanic  other female     worker
2  4.95         9         42  57      cauc  other female     worker
3  6.67        12          1  19      cauc  other   male     worker
4  4.00        12          4  22      cauc  other   male     worker
5  7.50        12         17  35      cauc  other   male     worker
6 13.07        13          9  28      cauc  other   male     worker
         sector union married
1 manufacturing    no     yes
2 manufacturing    no     yes
3 manufacturing    no      no
4         other    no      no
5         other    no     yes
6         other   yes      no
> dim(dataset)
[1] 534  11
> attach(dataset)
```

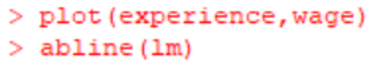


```
> levels(factor(ethnicity))
[1] "cauc"     "hispanic" "other"   
> levels(factor(region))
[1] "other" "south"
> levels(factor(gender))
[1] "female" "male"  
> levels(factor(occupation))
[1] "management" "office"     "sales"      "services"   "technical"  "worker"    
> levels(factor(sector))
[1] "construction"  "manufacturing" "other"        
> levels(factor(union))
[1] "no"  "yes"
> levels(factor(married))
[1] "no"  "yes"
> summary(dataset)
      wage          education       experience         age         ethnicity        
 Min.   : 1.000   Min.   : 2.00   Min.   : 0.00   Min.   :18.00   Length:534        
 1st Qu.: 5.250   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:28.00   Class :character  
 Median : 7.780   Median :12.00   Median :15.00   Median :35.00   Mode  :character  
 Mean   : 9.024   Mean   :13.02   Mean   :17.82   Mean   :36.83                     
 3rd Qu.:11.250   3rd Qu.:15.00   3rd Qu.:26.00   3rd Qu.:44.00                     
 Max.   :44.500   Max.   :18.00   Max.   :55.00   Max.   :64.00                     
    region             gender           occupation           sector         
 Length:534         Length:534         Length:534         Length:534        
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  


    union             married         
 Length:534         Length:534        
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

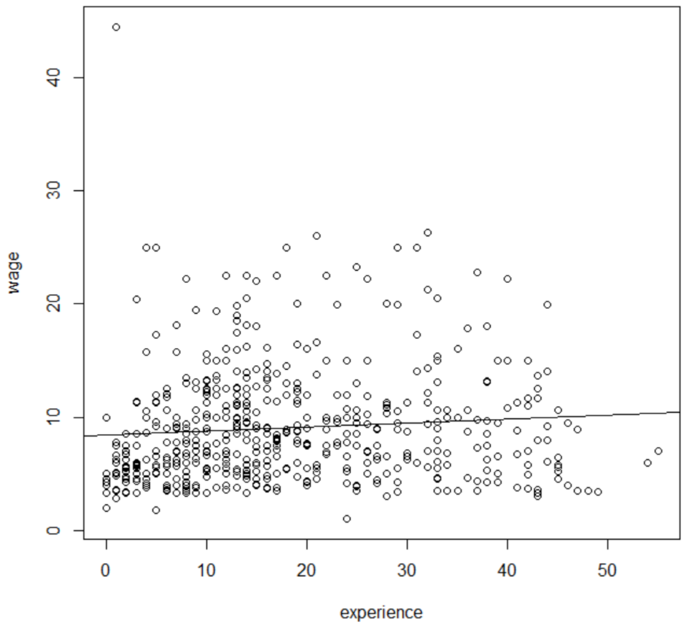
At this point we want to fit a simple linear regression model using wage as the response variable (Y) and experience as the explanatory variable (X). The aim is to create a model that we can use to predict Y from X. Then, we use summary() to obtain descriptive statistics and we construct the plot of the least squares fitted line.



```
> lm<-lm(wage~experience)
> summary(lm)

Call:
lm(formula = wage ~ experience)

Residuals:
   Min     1Q Median     3Q    Max 
-8.247 -3.601 -1.111  2.332 36.084 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.37997    0.38895  21.545   <2e-16 ***
experience   0.03614    0.01793   2.016   0.0443 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.124 on 532 degrees of freedom
Multiple R-squared:  0.007579,	Adjusted R-squared:  0.005714 
F-statistic: 4.063 on 1 and 532 DF,  p-value: 0.04433
```



```
> plot(experience,wage)
> abline(lm)
```

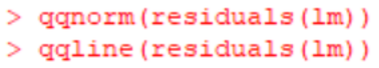


Linear regression implies the following assumptions:

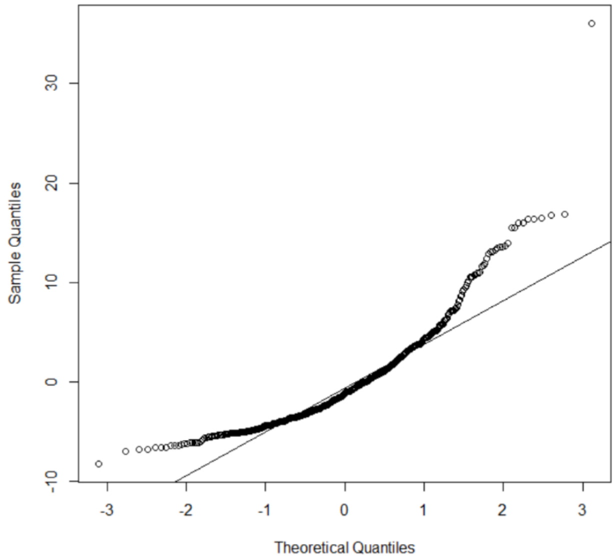
- **Linearity**: Y is a function of X and for simplicity we assume that their relationship is linear. Therefore, $Y \simeq \beta_0 + \beta_1 X$ where β_0 is the intercept and β_1 is the slope. To check this assumption, we plot our data and add the line described by the function as we did in the graph here above. We want a line which is close to all the points, i.e., the one that generates the smallest sum of errors.

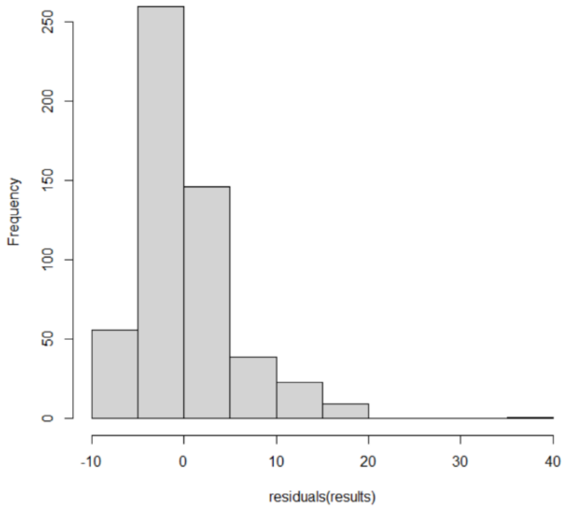
  From the plot we can notice that there is very slight linear association between experience and wage. Also, by looking at the summary we notice that the p-value is slightly lower than α with a 95% confidence interval. If we increase the confidence interval the p-value turns out to be higher than α. Therefore, this model can be used for predictions only for populations that are sufficiently similar to the one this sample was drawn from. Hence, we do not reject the assumption of linearity.

- **Normality**: regression model is composed of a systematic part and a random error ($Y=f(X)+\varepsilon$). We assume that the error follows a normal distribution since it is symmetric (i.e., the probability to get a negative error is the same to get a positive one) and its mean is 0. To check this assumption, we sketch a scatter plot of the residuals.



```
> qqnorm(residuals(lm))
> qqline(residuals(lm))
```



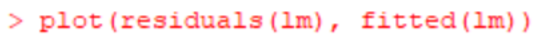


The points do not lie on the line; therefore, we assume that the distribution of residuals is not normal.

Using the function hist() we can plot an histogram of residuals through which we can state that the distribution is right-skewed due to the presence of outliers. Hence, we reject the assumption of normality.

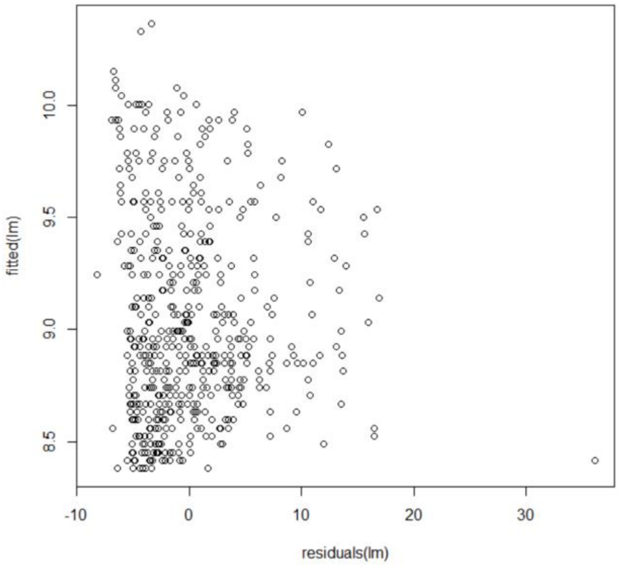
To solve this issue, we can use different strategies depending on the nature of the problem. For example, we might build a more complex model to address curvature, or we might apply a transformation to our data to address issues with normality, or we might analyse potential outliers, and then determine how to best handle these outliers.

- **Homoskedasticity**: the error is constant along the values of the dependent variable. The best way for checking this assumption is to make a scatterplot with the residuals against the fitted values.

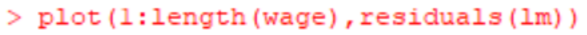


```
> plot(residuals(lm), fitted(lm))
```

From the graph we can notice that the points are disposed in a triangular shape whose base lies on the Y-axis. Therefore, the spread of the residuals decreases as the predicted values increase. This shows a case of heteroskedasticity; hence, we reject the homoskedasticity assumption.

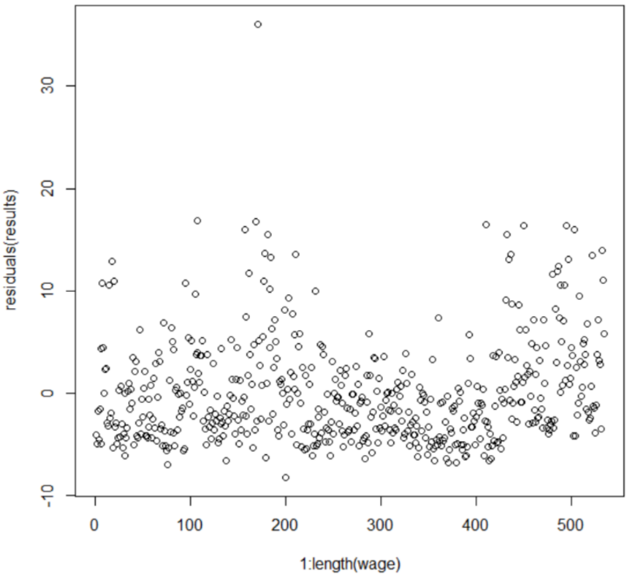


- **Independence**: there is not a relationship between the residuals and the Y variable; in other words, is independent of errors. To check this assumption, we need to make a scatterplot with the number of rows against the residuals.



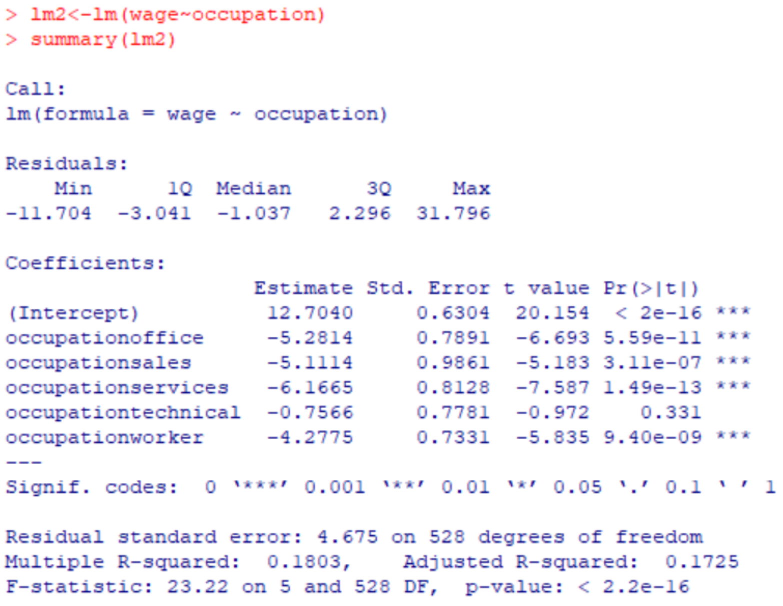
```
> plot(1:length(wage),residuals(lm))
```

There's no evidence of a systematic trend; hence, we do not reject the assumption of independence.



To deepen the understanding of the regression model, we want to interpret the estimated coefficients we saw in the summary (the estimated values for β_0 and β_1). The intercept (b_0) suggests that, if the years of experience of a US citizen are equal to 0, his/her expected wage will be equal to 8.38 dollars per hour. Instead, the coefficient of the experience variable (b_1) asserts that, if the experience of a US citizen increases by one year, his/her expected wage will increase by 0.4 dollars per hour.

At this point, we want to fit a simple linear regression model using wage as the response variable and occupation as explanatory variable.



```
> lm2<-lm(wage~occupation)
> summary(lm2)

Call:
lm(formula = wage ~ occupation)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.704  -3.041  -1.037   2.296  31.796 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           12.7040     0.6304  20.154  < 2e-16 ***
occupationoffice      -5.2814     0.7891  -6.693 5.59e-11 ***
occupationsales       -5.1114     0.9861  -5.183 3.11e-07 ***
occupationservices    -6.1665     0.8128  -7.587 1.49e-13 ***
occupationtechnical   -0.7566     0.7781  -0.972    0.331    
occupationworker      -4.2775     0.7331  -5.835 9.40e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.675 on 528 degrees of freedom
Multiple R-squared:  0.1803,    Adjusted R-squared:  0.1725 
F-statistic: 23.22 on 5 and 528 DF,  p-value: < 2.2e-16
```

Given that X is a categorical variable (occupation), R will automatically create dummies. A dummy variable is an artificial variable created to represent a categorical variable with two or more distinct categories. It takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.

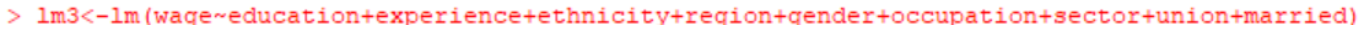
The number of dummy variables required to represent a particular categorical variable depends on the number of values that it can assume. If it can assume n different values, we need to define n - 1 dummy variables.

Since in our case the values that X can assume are six (worker, technical, services, office, sales, management), R will generate 5 dummies. Notice that there's no dummy variable to represent the "management" category because if for a certain citizen all dummies equal zero we know that he/she will work as a manager.

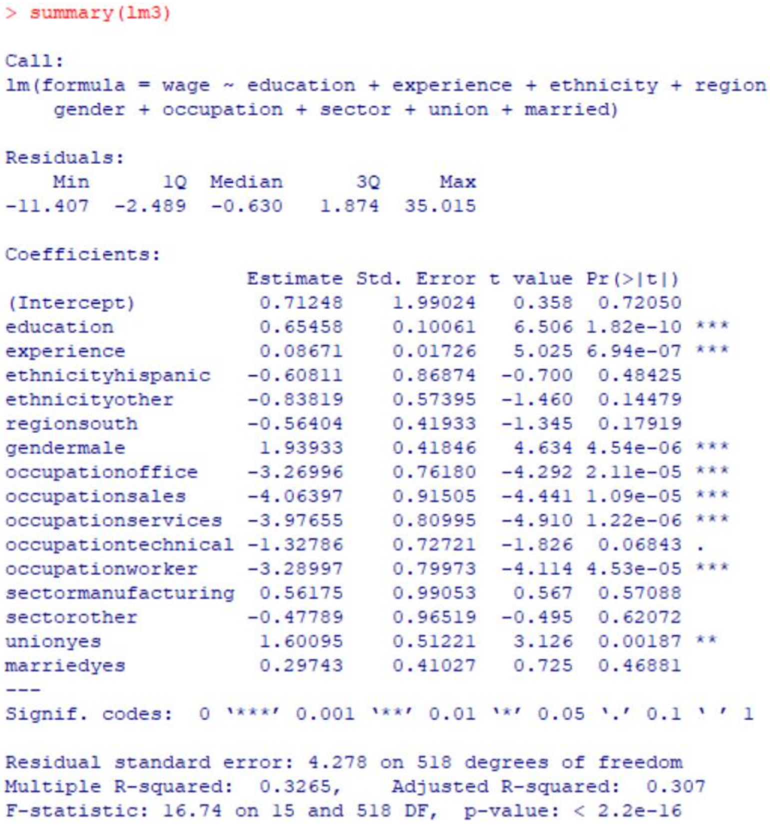
Hence, it is useful to interpret estimated coefficients:

- Intercept: the expected wage of US citizens employed in management and administration is 12.70 dollars per hour
- Occupationoffice: the US citizens employed in office and clerical work have a lower expected wage by 5.28 dollars per hour compared to the ones employed in management and administration
- Occupationsales: the expected wage of US citizens working in sales will be lower by 5.11 dollars per hour compared to the expected wage of US citizens working in management and administration
- Occupationservices: the US citizens employed in services have a lower expected wage by 6.17 dollars per hour compared to the ones employed in management and administration
- Occupationtechnical: the expected wage of US citizens which are technical or professional workers will be lower by 0.76 dollars per hour compared to the expected wage of US citizens working in management and administration
- Occupationworker: the US citizens working as trades-persons or assembly line workers have a lower expected wage by 4.28 dollars per hour compared to the ones employed in management and administration.

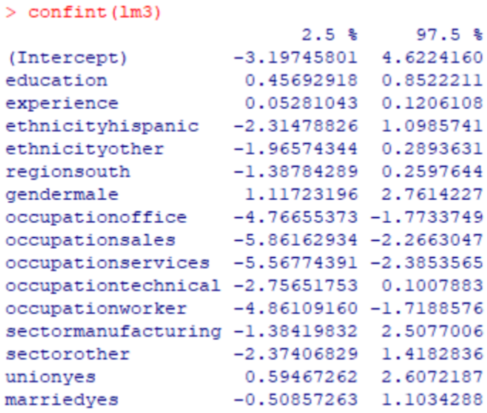
We now want to fit a multiple linear regression using wage as the response variable and all the rest (except age) as the explanatory variables.



```
> lm3<-lm(wage~education+experience+ethnicity+region+gender+occupation+sector+union+married)
```



```
> summary(lm3)

Call:
lm(formula = wage ~ education + experience + ethnicity + region +
    gender + occupation + sector + union + married)

Residuals:
    Min      1Q  Median      3Q     Max
-11.407  -2.489  -0.630   1.874  35.015

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.71248    1.99024   0.358  0.72050
education            0.65458    0.10061   6.506 1.82e-10 ***
experience           0.08671    0.01726   5.025 6.94e-07 ***
ethnicityhispanic   -0.60811    0.86874  -0.700  0.48425
ethnicityother      -0.83819    0.57395  -1.460  0.14479
regionsouth         -0.56404    0.41933  -1.345  0.17919
gendermale           1.93933    0.41846   4.634 4.54e-06 ***
occupationoffice    -3.26996    0.76180  -4.292 2.11e-05 ***
occupationsales     -4.06397    0.91505  -4.441 1.09e-05 ***
occupationservices  -3.97655    0.80995  -4.910 1.22e-06 ***
occupationtechnical -1.32786    0.72721  -1.826  0.06843 .
occupationworker    -3.28997    0.79973  -4.114 4.53e-05 ***
sectormanufacturing  0.56175    0.99053   0.567  0.57088
sectorother         -0.47789    0.96519  -0.495  0.62072
unionyes             1.60095    0.51221   3.126  0.00187 **
marriedyes           0.29743    0.41027   0.725  0.46881
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.278 on 518 degrees of freedom
Multiple R-squared:  0.3265,	Adjusted R-squared:  0.307
F-statistic: 16.74 on 15 and 518 DF,  p-value: < 2.2e-16
```



```
> confint(lm3)
                          2.5 %     97.5 %
(Intercept)         -3.19745801  4.6224160
education            0.45692918  0.8522211
experience           0.05281043  0.1206108
ethnicityhispanic   -2.31478826  1.0985741
ethnicityother      -1.96574344  0.2893631
regionsouth         -1.38784289  0.2597644
gendermale           1.11723196  2.7614227
occupationoffice    -4.76655373 -1.7733749
occupationsales     -5.86162934 -2.2663047
occupationservices  -5.56774391 -2.3853565
occupationtechnical -2.75651753  0.1007883
occupationworker    -4.86109160 -1.7188576
sectormanufacturing -1.38419832  2.5077006
sectorother         -2.37406829  1.4182836
unionyes             0.59467262  2.6072187
marriedyes          -0.50857263  1.1034288
```

Thus, it is useful to interpret the estimated coefficients also performing statistical inference about them considering the p-value and the confidence intervals shown here above:

- Intercept: the expected wage of a female US citizen with 0 years of education, 0 years of work experience, Caucasian ethnicity, which does not live in the south, is employed in management in the construction sector, is not working on a union job and is not married is 0.71 dollars per hour.

  We are 95% sure that the expected wage of a female US citizen with 0 years of education 0 years of work experience, Caucasian ethnicity, which does not live in the south, is employed in management in the construction sector, is not working on a union job and is not married, lies between -3.20 and 4.62 dollars per hour.

- Education: an additional year of education of a US citizen results to an increase in expected wage by 0.65 dollars per hour given that work experience, gender, ethnicity, region, occupation, sector, union job and marital status remain unchanged.

  The p-value of this variable is 0.000000000182 which is much smaller than α (=0.05); therefore, we have enough evidence to reject the null hypothesis H_0=0 (no linear relation between education and wage). Also, considering the confidence interval we can state that, we are 95% confident that as education increases by one year the expected wage can increase from 0.46 to 0.85 dollars per hour, given that all the other variables remain unchanged.

- Experience: an additional year of work experience of a US citizen results to an increase in expected wage by 0.09 dollars per hour given that education, gender, ethnicity, region, occupation, sector, union job and marital status remain unchanged.

  The p-value of this variable is 0.000000694 which is much smaller than α (=0.05); therefore, we have enough evidence to reject the null hypothesis H_0=0 (no linear relation between work experience and wage). Also, considering the confidence interval we can state that, we are 95% confident that as work experience increases by one year the expected wage can increase from 0.05 to 0.12 dollars per hour, given that all the other variables remain unchanged.

- Ethnicityhispanic: a US citizen of Hispanic ethnicity has a lower expected wage by 0.61 dollars per hour compared to a US citizen of Caucasian ethnicity given that both citizens have the same education, work experience, gender, region, occupation, sector, union job and marital status.

  The p-value of this variable is 0.48425 which is very close but still smaller than α (=0.05); therefore, we have enough evidence to reject the null hypothesis H_0=0. Also, considering the confidence interval we can state that, we are 95% confident that if a US citizen is Hispanic, he/she will have a change in expected wage from -2.31 to 1.10 dollars per hour compared to a citizen with Caucasian ethnicity, given that all the other variables remain unchanged. Notice that the confidence interval includes 0; thus, we cannot reject the null hypothesis since every value in the confidence interval is a plausible value of the parameter.

- Ethnicityother: a US citizen of neither Hispanic nor Caucasian ethnicity has a lower expected wage by 0.84 dollars per hour compared to a US citizen of Caucasian ethnicity given that both citizens have the same education, work experience, gender, region, occupation, sector, union job and marital status.

  The p-value of this variable is 0.14479 which is smaller than α (=0.05); therefore, we have enough evidence to reject the null hypothesis H_0=0. Also, considering the confidence interval we can state that, we are 95% confident that if a US citizen is neither Hispanic nor Caucasian, he/she will have a change in expected wage from -1.97 to 0.29 dollars per hour compared to a citizen with Caucasian ethnicity, given that all the other variables remain unchanged. Notice that the confidence interval includes 0; thus, we cannot reject the null hypothesis since every value in the confidence interval is a plausible value of the parameter.

- Regionsouth: a US citizen living in the south has a lower expected wage by 0.56 dollars per hour compared to a US citizen not living in the south given that both citizens have the same education, work experience, gender, ethnicity, occupation, sector, union job and marital status.

  The p-value of this variable is 0.17919 which is smaller than α (=0.05); therefore, we have enough evidence to reject the null hypothesis H_0=0. Also, considering the confidence interval we can state that, we are 95% confident that if a US citizen is not living in the south, he/she will have a change in expected wage from -1.39 to 0.26 dollars per hour compared to a citizen living in the south, given that all the other variables remain unchanged. Notice that the confidence interval includes 0; thus, we cannot reject the null hypothesis since every value in the confidence interval is a plausible value of the parameter.

- Gendermale: a male US citizen has a higher expected wage by 1.94 dollars per hour compared to a female US citizen given that both citizens have the same education, work experience, ethnicity, region, occupation, sector, union job and marital status.
  The p-value of this variable is 0.00000454 which is much smaller than α (=0.05); therefore, we have enough evidence to reject the null hypothesis H_0=0. Also, considering the confidence interval we can state that, we are 95% confident that if a US citizen is a male, he will have a change in expected wage from 1.12 to 2.76 dollars per hour compared to a female citizen, given that all the other variables remain unchanged.

- Occupationoffice: a US citizen employed in office work has a lower expected wage by 3.27 dollars per hour compared to a US citizen employed in management given that both citizens have the same education, work experience, gender, region, ethnicity, sector, union job and marital status.
  The p-value of this variable is 0.0000211 which is much smaller than α (=0.05); therefore, we have enough evidence to reject the null hypothesis H_0=0. Also, considering the confidence interval we can state that, we are 95% confident that if a US citizen is employed in office work, he/she will have a change in expected wage from -4.76 to -1.77 dollars per hour compared to a citizen working in management, given that all the other variables remain unchanged.

- Occupationsales: a US citizen employed in sales has a lower expected wage by 4.06 dollars per hour compared to a US citizen employed in management given that both citizens have the same education, work experience, gender, region, ethnicity, sector, union job and marital status.
  The p-value of this variable is 0.0000109 which is much smaller than α (=0.05); therefore, we have enough evidence to reject the null hypothesis H_0=0. Also, considering the confidence interval we can state that, we are 95% confident that if a US citizen is employed in sales, he/she will have a change in expected wage from -5.86 to -2.27 dollars per hour compared to a citizen working in management, given that all the other variables remain unchanged.

- Occupationservices: a US citizen employed in services has a lower expected wage by 3.98 dollars per hour compared to a US citizen employed in management given that both citizens have the same education, work experience, gender, region, ethnicity, sector, union job and marital status.

The p-value of this variable is 0.00000122 which is much smaller than α (=0.05); therefore, we have enough evidence to reject the null hypothesis H_0=0. Also, considering the confidence interval we can state that, we are 95% confident that if a US citizen is employed in sales, he/she will have a change in expected wage from -5.56 to -2.38 dollars per hour compared to a citizen working in management, given that all the other variables remain unchanged.

- Occuptiontechnical: a US citizen employed in technical work has a lower expected wage by 1.33 dollars per hour compared to a US citizen employed in management given that both citizens have the same education, work experience, gender, region, ethnicity, sector, union job and marital status.

  The p-value of this variable is 0.06843 which is smaller than α (=0.05); therefore, we have enough evidence to reject the null hypothesis H_0=0. Also, considering the confidence interval we can state that, we are 95% confident that if a US citizen is employed in technical work, he/she will have a change in expected wage from -2.76 to 1.10 dollars per hour compared to a citizen working in management, given that all the other variables remain unchanged. Notice that the confidence interval includes 0; thus, we cannot reject the null hypothesis since every value in the confidence interval is a plausible value of the parameter.

- Occupationworker: a US citizen employed as worker has a lower expected wage by 3.29 dollars per hour compared to a US citizen employed in management given that both citizens have the same education, work experience, gender, region, ethnicity, sector, union job and marital status.

  The p-value of this variable is 0.0000453 which is much smaller than α (=0.05); therefore, we have enough evidence to reject the null hypothesis H_0=0. Also, considering the confidence interval we can state that, we are 95% confident that if a US citizen is employed as a worker, he/she will have a change in expected wage from -4.86 to -1.72 dollars per hour compared to a citizen working in management, given that all the other variables remain unchanged.

- Sectormanufacturing: a US citizen working in the manufacturing sector has a higher expected wage by 0.56 dollars per hour compared to a US citizen working in construction sector given that both citizens have the same education, work experience, gender, region, occupation, ethnicity, union job and marital status.

  The p-value of this variable is 0.57088 which is higher than α (=0.05); therefore, we have enough evidence to not reject the null hypothesis H_0=0. Also, considering the confidence interval we can state that, we are 95% confident that if a US citizen is working in manufacturing sector, he/she will have a change in expected wage from -1.38 to 2.51 dollars per hour compared to a citizen working in construction sector, given that all the other variables remain unchanged. Notice that the confidence interval includes 0; thus, we confirm our decision of not rejecting H_0.

- Sectorother: a US citizen working neither in the manufacturing nor in the construction sector has a lower expected wage by 0.48 dollars per hour compared to a US citizen working in construction sector given that both citizens have the same education, work experience, gender, region, occupation, ethnicity, union job and marital status.
  The p-value of this variable is 0.62072 which is higher than α (=0.05); therefore, we have enough evidence to reject the null hypothesis $H_0=0$. Also, considering the confidence interval we can state that, we are 95% confident that if a US citizen is working neither in construction nor in manufacturing sector, he/she will have a change in expected wage from -2.37 to 1.42 dollars per hour compared to a citizen working in construction sector, given that all the other variables remain unchanged. Notice that the confidence interval includes 0; thus, we confirm our decision of not rejecting H_0.

- Unionyes: a US citizen working on a union job has a higher expected wage by 1.60 dollars per hour compared to a US citizen not working on a union job given that both citizens have the same education, work experience, gender, region, occupation, sector, ethnicity and marital status.
  The p-value of this variable is 0.00187 which is much smaller than α (=0.05); therefore, we have enough evidence to reject the null hypothesis $H_0=0$. Also, considering the confidence interval we can state that, we are 95% confident that if a US citizen work on a union job, he/she will have a change in expected wage from 0.59 to 2.61 dollars per hour compared to a citizen not working on a union job, given that all the other variables remain unchanged.

- Marriedyes: a married US citizen has a higher expected wage by 0.29 dollars per hour compared to a US citizen which is not married given that both citizens have the same education, work experience, gender, region, occupation, sector, union job and ethnicity.
  The p-value of this variable is 0.46881 which is close but still smaller than α (=0.05); therefore, we have enough evidence to reject the null hypothesis $H_0=0$. Also, considering the confidence interval we can state that, we are 95% confident that if a US citizen is married, he/she will have a change in expected wage from -0.51 to 1.10 dollars per hour compared to an unmarried citizen, given that all the other variables remain unchanged. Notice that the confidence interval includes 0; thus, we cannot reject the null hypothesis since every value in the confidence interval is a plausible value of the parameter.

The standard deviation of error (or residual standard error) is a measure of the variation of observed values from the regression line. The magnitude of the standard error should always be judged relative to the size of the y values in the sample data. In the model we built standard error = 4.278; it is quite high relative to wages in the \$1 - \$44.5 range. This suggests that sample means are widely spread around the population mean, so the sample may not closely represent the population.

The coefficient of determination, denoted as R^2, is the portion of the total variation of the response variable which is explained by the variation of the explanatory variables.

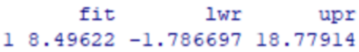
$$R^2 = \frac{SSR}{SST}$$

Its value must be as close as possible to 1 to have perfect linear regression. In our model R^2 =0.3265; this value suggests that the explanatory variables are not explaining much the variation of the response variable.

A more accurate tool is the Adjusted Coefficient of Determination (Adjusted R^2). It measures the proportion of variation explained by only those explanatory variables that really help in explaining the dependent variable. It penalizes you for adding variables that do not help in predicting the response variable. In our case R^2adjusted=0.307. This value confirms our statement for which the change in the response variable is not much explained by the variation of explanatory variables.

Based on the results of our last model, we want to produce a 98% confidence interval for the actual wage in dollars per hour for a male US citizen with 10 years of education, 15 years of experience, with Hispanic ethnicity, living in the South, employed as a worker in the manufacturing sector, has a union job and is married. To do so, we run the following function in R:

```
>predict(lm3, list(education=10, experience=15, ethnicity="hispanic", region="south", gender="male", occupation="worker", sector="manufacturing", union="yes", married="yes"), int="p", level=0.98)
```

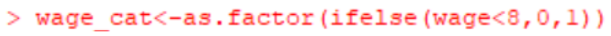


```
       fit       lwr      upr
1 8.49622 -1.786697 18.77914
```

Therefore, we are 98% confident that the actual wage of a male US citizen with 10 years of education, 15 years of work experience, Hispanic ethnicity, which lives in the south, is employed as a worker in the manufacturing sector, is working on a union job and is married lies between - 1.79 and 18.78 dollars per hour.

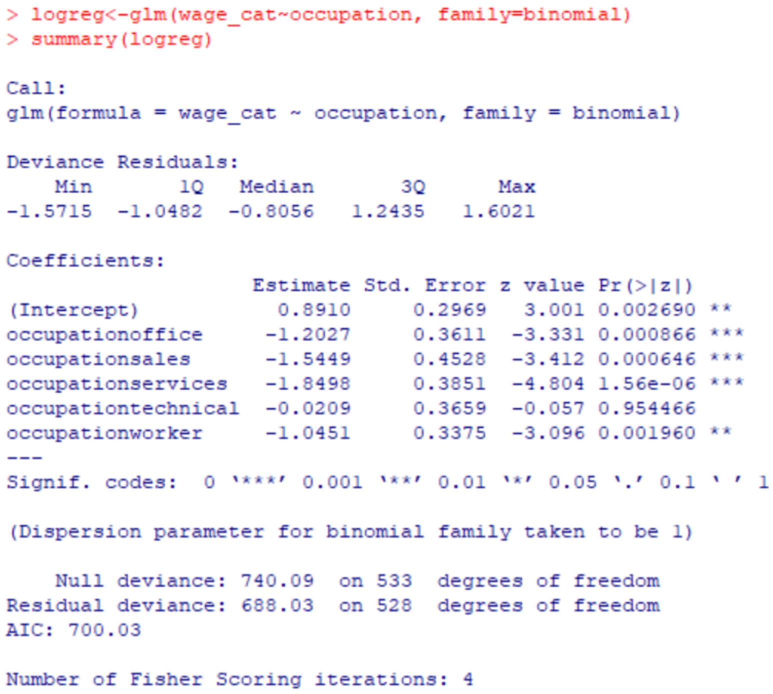
Exercise 2

Considering the data of exercise 1, we want to create a new categorical variable called wage_cat which takes the value 0 (low) when wage is less than 8 dollars per hour and the value 1 (high) in all other cases.



```
> wage_cat<-as.factor(ifelse(wage<8,0,1))
```

At this point, we fit a simple logistic regression model with wage_cat as a response variable and occupation as the explanatory variable. Subsequently, we produce the summary to estimate coefficients.



```
> logreg<-glm(wage_cat~occupation, family=binomial)
> summary(logreg)

Call:
glm(formula = wage_cat ~ occupation, family = binomial)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.5715  -1.0482  -0.8056   1.2435   1.6021  

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)            0.8910     0.2969   3.001 0.002690 ** 
occupationoffice      -1.2027     0.3611  -3.331 0.000866 ***
occupationsales       -1.5449     0.4528  -3.412 0.000646 ***
occupationservices    -1.8498     0.3851  -4.804 1.56e-06 ***
occupationtechnical   -0.0209     0.3659  -0.057 0.954466    
occupationworker      -1.0451     0.3375  -3.096 0.001960 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 740.09  on 533  degrees of freedom
Residual deviance: 688.03  on 528  degrees of freedom
AIC: 700.03

Number of Fisher Scoring iterations: 4
```

To interpret estimated coefficients, we use odds ratios. An odd ratio is a ratio of two odds which provides the relative change of the odds under two different conditions (for example X=0,1)

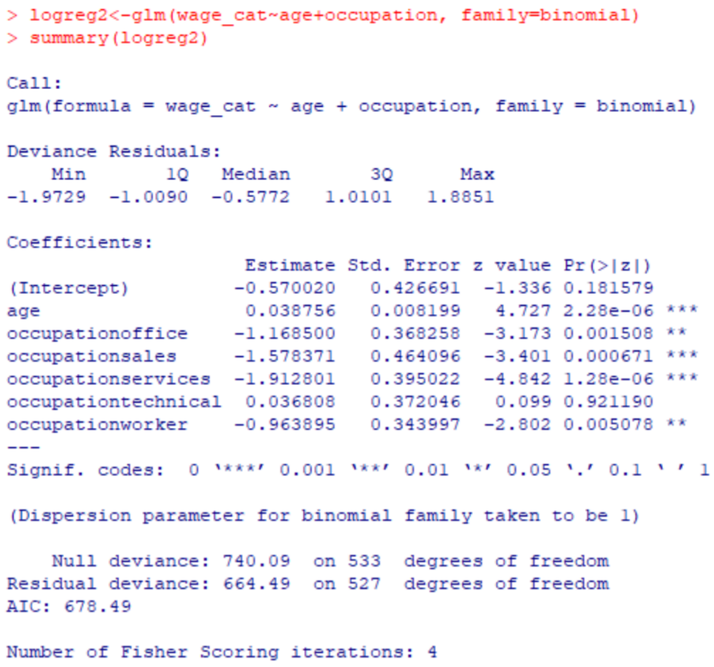
$$OR_{01} = \frac{\mathrm{odds}(\mathrm{Y} = 1|\mathrm{X} = 0)}{odds(Y = 1|X = 1)}$$

Through (OR-1)*100 we can obtain the percentage change in the odds for X=0 with the corresponding odds when X=1.

- Intercept: exp(0.8910)= 2.437566 is the odds of having a high wage for US citizens working in management and administration
- Occupationoffice: exp(-1.2027)= 0.3003821 expresses that the odds of having a high wage for a person employed in an office work are 70% lower than for a person employed in management and administration

- Occupationsales: exp(-1.5449)=0.2133332 expresses that the odds of having a high wage for a person employed in sales are 79% lower than for a person employed in management and administration
- Occupationservices: exp(-1.8498)= 0.1572686 expresses that the odds of having a high wage for a person employed in services are 84% lower than for a person employed in management and administration
- Occupationtechnical: exp(-0.0209)=0.9793169 expresses that the odds of having a high wage for a person employed in technical work are 2% lower than for a person employed in management and administration
- Occupationworker: exp(-1.0451)= 0.3516567 expresses that the odds of having a high wage for a person employed as a worker are 65% lower than for a person employed in management and administration

We now fit a multiple logistic regression model with wage_cat as a response variable and age and occupation as the explanatory variables. Subsequently, we produce the summary to estimate coefficients.



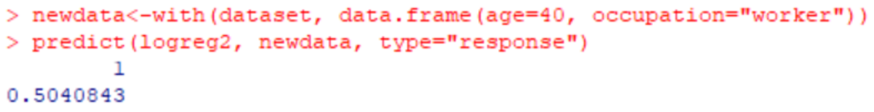
```
> logreg2<-glm(wage_cat~age+occupation, family=binomial)
> summary(logreg2)

Call:
glm(formula = wage_cat ~ age + occupation, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9729  -1.0090  -0.5772   1.0101   1.8851

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.570020   0.426691  -1.336 0.181579
age                  0.038756   0.008199   4.727 2.28e-06 ***
occupationoffice    -1.168500   0.368258  -3.173 0.001508 **
occupationsales     -1.578371   0.464096  -3.401 0.000671 ***
occupationservices  -1.912801   0.395022  -4.842 1.28e-06 ***
occupationtechnical  0.036808   0.372046   0.099 0.921190
occupationworker    -0.963895   0.343997  -2.802 0.005078 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 740.09  on 533  degrees of freedom
Residual deviance: 664.49  on 527  degrees of freedom
AIC: 678.49

Number of Fisher Scoring iterations: 4
```

Hence, it is useful to interpret estimated coefficients:

- Intercept: exp(-0.570020)= 0.5655141is the odds of having a high wage for a US citizen aged 0 and working in management and administration
- Age: exp(0.038756)= 1.039517 expresses that the odds of having a high wage increase by 4% when age increases by one year for the same occupation

- Occupationoffice: exp(-1.168500)= 0.3108328 expresses that the odds of having a high wage for a person employed in an office work are 69% lower than for a person employed in management and administration for the same age
- Occupationsales: exp(-1.578371)= 0.2063109 expresses that the odds of having a high wage for a person employed in sales are 79% lower than for a person employed in management and administration for the same age
- Occupationservices: exp(-1.912801)= 0.1476662 expresses that the odds of having a high wage for a person employed in services are 85% lower than for a person employed in management and administration for the same age
- Occupationtechnical: exp(0.036808)= 1.037494 expresses that the odds of having a high wage for a person employed in technical work are 4% higher than for a person employed in management and administration for the same age
- Occupationworker: exp(-0.963895)= 0.3814044 expresses that the odds of having a high wage for a person employed as a worker are 62% lower than for a person employed in management and administration for the same age

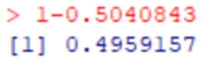
Based on the results of our last model, we want to estimate the probability of a low wage for a 40-year-old worker:



```
> newdata<-with(dataset, data.frame(age=40, occupation="worker"))
> predict(logreg2, newdata, type="response")
        1 
0.5040843 
```

R automatically computes the probability of success (Y=1), therefore the obtained result is the probability of a high wage for a 40-year-old worker (P(Y=1|age, occupation)). To get, instead, the probability of a low wage we need to proceed as follows:

1 - P(Y=1|age, occupation) = P(Y=0|age, occupation)



```
> 1-0.5040843
[1] 0.4959157
```

Hence, the probability of having a low wage for a 40-year-old worker is 49.59%.