Alessandro Secchi n. 944668

POLITECNICO
MILANO 1863

SCHOOL OF INDUSTRIAL AND INFORMATION ENGINEERING

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

BAYESIAN LEARNING AND MONTECARLO SIMULATION

HOMEWORK 2

MAY 2020

"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. The data can be found in the library statsr. Comment all the obtained results."

#Part 1 ----------

The variable of interest is lwage, the logarithm of the wages. Visualize the response with an histogram and add the density of a Normal with mean equal to the mean of the response, and standard deviation equal to the standard deviation of the response. Compute a summary of the dataset and control the correlations between variable. What assumptions do you need to compute a linear model? What is the distribution of the response?

We install the package required and we load the data, but before starting we use the command rm(list=ls()) in order to remove all the objects present in the workspace.

```
rm(list=ls())
install.packages("statsr")
library(statsr)
data(wage)
```

In order to include only the observations without missing values and to remove columns of factor variables we write:

```
wage <- wage[which(complete.cases(wage)),]
wage <- wage[,-c(1,9:12)]
wage
y <- wage$lwage
X <- as.matrix(wage[,1:11])
n <- nrow(wage)
```

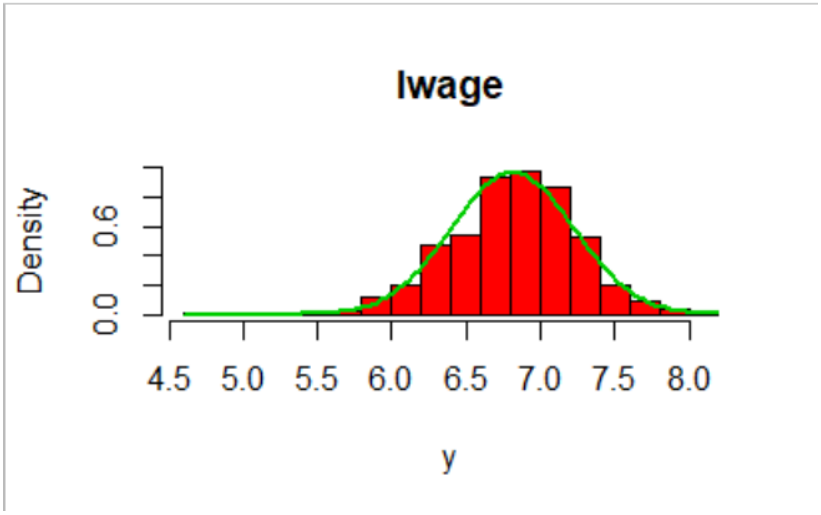
The first ten row of our new dataset are:

	hours	iq	kww	educ	exper	tenure	age	sibs	brthord	meduc	feduc	lwage
1	40	93	35	12	11	2	31	1	2	8	8	6.645091
2	40	108	46	14	11	9	33	1	2	14	14	6.715384
3	40	96	32	12	13	7	32	4	3	12	12	6.476973
4	40	74	27	11	14	5	34	10	6	6	11	6.331502
5	40	91	24	10	13	0	30	1	2	8	8	6.396930
6	45	111	37	15	13	1	36	2	3	14	5	7.050990
7	40	95	44	12	16	16	36	1	1	12	11	6.907755
8	43	132	44	18	8	13	38	1	1	13	14	6.835185
9	38	119	24	16	7	2	28	3	1	10	10	7.183871
10	40	118	47	16	9	9	34	1	1	12	12	7.491087

Lwage is our response or outcome while X are our covariates or explanatory variables.

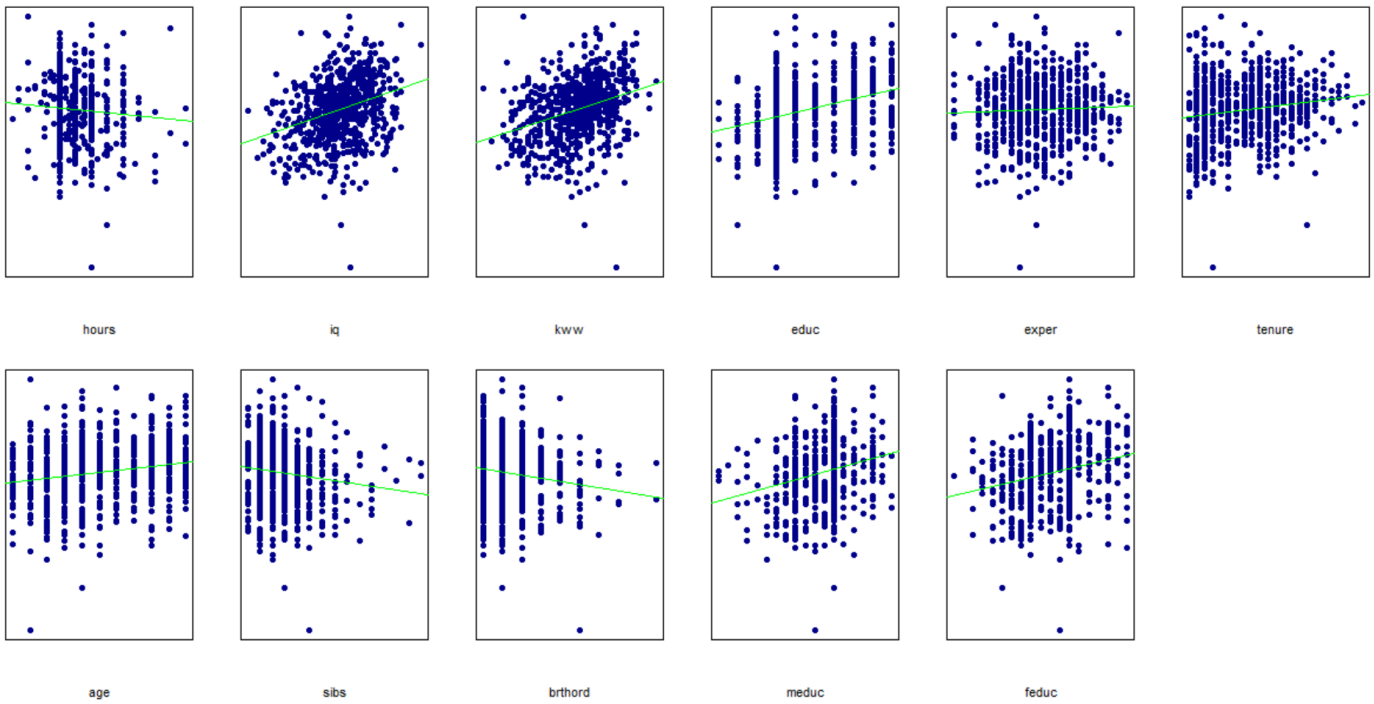
By plotting the histogram of the response with the density of a Normal with mean equal to the mean of the response, and standard deviation equal to the standard deviation of the response, we get:

```
hist(y, freq=F, col="2", main="lwage", ylim=c(0,1), breaks=20)
curve(dnorm(x, mean=mean(y), sd=sd(y)), col=3, lwd=2.5, add=T)
```



Now let's compute the correlation among variables and let's see the summary of the data set:

```
par(mfrow=c(2,6), mar=c(4.2,2,2,1.2))
for (j in 1:ncol(X)) {
  plot(X[,j], y, xlab=names(wage)[j],pch=19, col="blue4",xaxt="n", yaxt="n")
  abline(lm(y~X[,j]), col="green")
  }
```



Some notes: in correlations -1 means there is a perfect negative linear relationship, 0 means there is no linear relationship while +1 means there is a perfect positive linear relationship.

If the scatterplot doesn't indicate there's at least somewhat of a linear relationship, the correlation doesn't mean much. In fact correlation only applies to linear relationships and if a strong relationship exists but it's not linear, the correlation factor may be misleading, because in some cases a strong curved relationship exists. That's why it's critical to examine the scatterplot first (we will control the assumptions about the linearity).

We can have a summary of our data by using the command summary(wage):

```
     hours             iq             kww            educ           exper      
 Min.   :25.00   Min.   : 54.0   Min.   :13.00   Min.   : 9.00   Min.   : 1.0  
 1st Qu.:40.00   1st Qu.: 94.0   1st Qu.:32.00   1st Qu.:12.00   1st Qu.: 8.0  
 Median :40.00   Median :104.0   Median :37.00   Median :13.00   Median :11.0  
 Mean   :44.06   Mean   :102.5   Mean   :36.19   Mean   :13.68   Mean   :11.4  
 3rd Qu.:48.00   3rd Qu.:113.0   3rd Qu.:41.00   3rd Qu.:16.00   3rd Qu.:15.0  
 Max.   :80.00   Max.   :145.0   Max.   :56.00   Max.   :18.00   Max.   :22.0  
     tenure            age              sibs           brthord          meduc      
 Min.   : 0.000   Min.   :28.00   Min.   : 0.000   Min.   : 1.000   Min.   : 0.00  
 1st Qu.: 3.000   1st Qu.:30.00   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 9.00  
 Median : 7.000   Median :33.00   Median : 2.000   Median : 2.000   Median :12.00  
 Mean   : 7.217   Mean   :32.98   Mean   : 2.846   Mean   : 2.178   Mean   :10.83  
 3rd Qu.:11.000   3rd Qu.:36.00   3rd Qu.: 4.000   3rd Qu.: 3.000   3rd Qu.:12.00  
 Max.   :22.000   Max.   :38.00   Max.   :14.000   Max.   :10.000   Max.   :18.00  
     feduc           lwage      
 Min.   : 0.00   Min.   :4.745  
 1st Qu.: 8.00   1st Qu.:6.550  
 Median :11.00   Median :6.843  
 Mean   :10.27   Mean   :6.814  
 3rd Qu.:12.00   3rd Qu.:7.090  
 Max.   :18.00   Max.   :8.032  
```

A theoretical recap:

For a linear model we assume that:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$$

Where $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ is a random error normally distribuited. Y is called response or outcome while x are the covariates or explanatory variables.

Or more in general:

$$\begin{pmatrix} y_1 \\ \ldots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ 1 & x_{31} & x_{32} & \ldots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \ldots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \ldots \\ \epsilon_n \end{pmatrix}$$

In frequentist linear regression, the best explanation is taken to mean the coefficients β, that minimize the residual sum of squares (RSS). (the total of the squared differences between the known values (y) and the predicted model outputs ŷ)indicating an estimate).

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - \hat{y})^2 = \sum_{i=1}^{N} (y_i - \beta^T x_i)^2$$

The summation is taken over the N data points in the training set. The closed form solution for the model parameters, β, that minimize the error is known as the maximum likelihood estimate of β (the value that is the most probable given the inputs, X, and outputs, y).

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

This method of fitting the model parameters by minimizing the RSS is called Ordinary Least Squares (OLS).

Once we have β-hat, we can estimate the output value of any new data point by applying our model equation:

$$\hat{y} = \hat{\beta}^T X$$

In the Bayesian viewpoint, instead, we formulate linear regression using probability distributions rather than point estimates. The response y, is not estimated as a single value, but as a probability distribution:

$$y \sim N(\beta^T X, \sigma^2 I)$$

In contrast to OLS, we have a posterior *distribution* for the model parameters that is proportional to the likelihood of the data multiplied by the *prior* probability of the parameters.

Note: as the amount of data points increases the outputs for the parameters converge to the values obtained from OLS.

With multiple covariates the response has a multivariate normal distribution

$$y|\beta, \sigma^2, X \sim \mathcal{N}_n(X\beta, \sigma^2 I_n).$$

with mean and variance as:

$$\begin{aligned} \mathbb{E}[y_i|\beta, X] &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \\ Var(y_i|\sigma^2, X) &= \sigma^2 \\ Cov(y_i, y_j) &= 0 \quad \text{for } i \neq j \end{aligned}$$

Note: matrix X must be full rank and so the correlation among variables variables must be different from +-1(than means dependent or same information).

In order to carry out regression we could use the command “lm” used to fit linear models.

Simply by using the following commands:

```
lwage.lm=lm(y~X)

summary(lwage.lm)
```

We get:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.89465 -0.21419  0.00731  0.24213  1.25949

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.280010   0.225176  23.448  < 2e-16 ***
Xhours      -0.005898   0.002019  -2.921  0.00361 **
Xiq          0.003775   0.001236   3.053  0.00236 **
Xkww         0.005815   0.002472   2.352  0.01898 *
Xeduc        0.038872   0.009244   4.205 2.97e-05 ***
Xexper       0.012178   0.004628   2.631  0.00871 **
Xtenure      0.007340   0.003015   2.435  0.01517 *
Xage         0.008818   0.006220   1.418  0.15675
Xsibs        0.007572   0.008035   0.942  0.34636
Xbrthord    -0.019052   0.011981  -1.590  0.11228
Xmeduc       0.009270   0.006434   1.441  0.15012
Xfeduc       0.009929   0.005594   1.775  0.07641 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3667 on 651 degrees of freedom
Multiple R-squared:  0.222,     Adjusted R-squared:  0.2088
F-statistic: 16.88 on 11 and 651 DF,  p-value: < 2.2e-16
```

Let’s analyse this numbers:

The first section summarizes the residual (the error between the prediction of the model and the actual results). Smaller residuals are of course better. Note that the distribution of the residuals appears more or less symmetrical. That means that the points predicted by the model do not fall too far away from the actual observed points.

Then for each variable and the intercept, a weight is produced:

- The estimate is the weight given to the variable. In the simple regression case (for every one X increase, the model predicts an increase of the number of the estimate)
- The coefficient Standard Error measures the average amount that the coefficient estimates vary from the actual average value of our response variable. Useful for calculating the t-value (calculated by taking the coefficient divided by the Std. Error).
- The t-value is used to test whether or not the coefficient is significantly different from zero. If it isn’t significant, then the coefficient isn’t adding anything to the model and could be dropped or investigated further. (If it is significant it means we could reject the null hypothesis and so declare a relationship between response and covariate). Therefore the greater the magnitude of T, the greater the evidence against the null hypothesis.

- The *Pr(>t)* relates to the probability of observing any value equal or larger than *t*. A small p-value indicates that it is unlikely we will observe a relationship between the predictor and response variables. Note the "signif. Codes" associated to each estimate, three stars represent a highly significant p-value. So we can state that Xeduc seems the most correlated covariate to the response.
- The Residual Standard Error is measure of the *quality* of a linear regression fit.
- The Multiple R-Squared provides a measure of how well the model is fitting the actual data.
- The F-Statistic is a global test to check if the model has at least one significant variable (to see if there is a relationship between our predictor and the response variables). The further the F-statistic is from 1 the better it is).

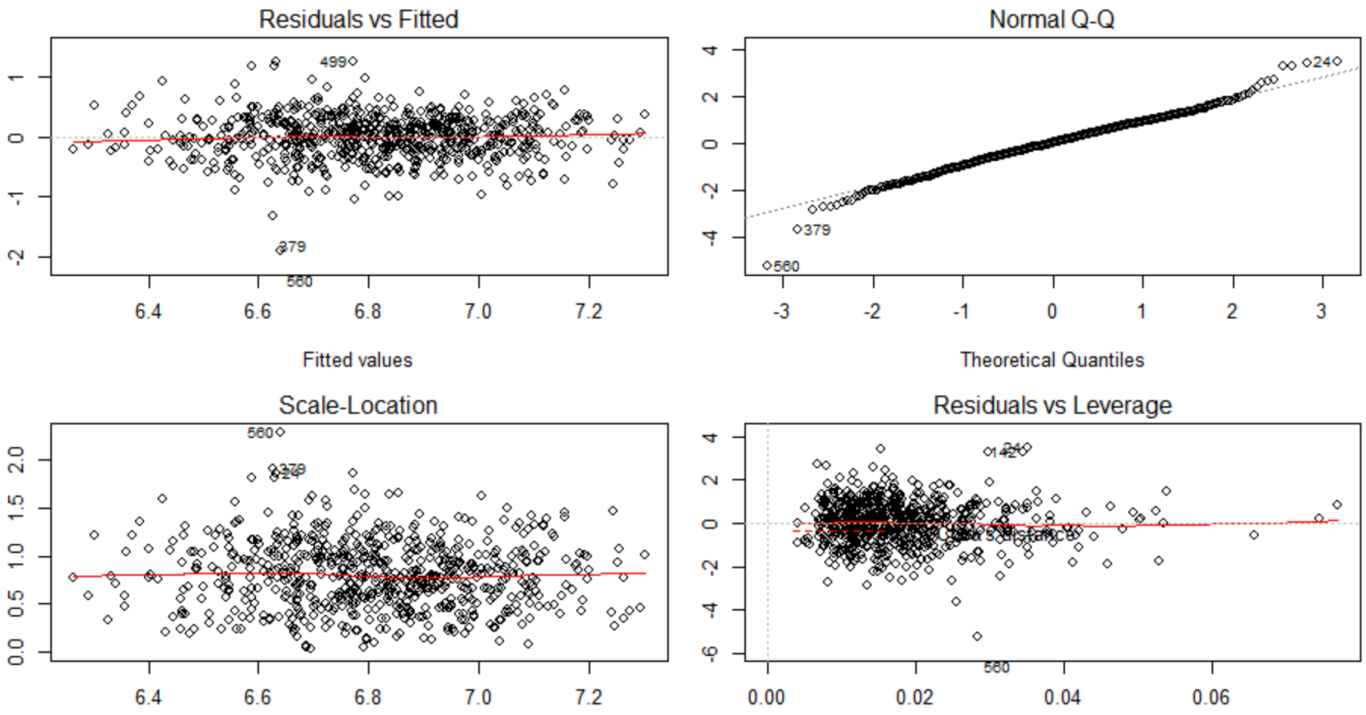
Linear regression model assumptions:

1. Linearity: the relationship between X and the mean of y is linear.
2. Homoscedasticity: the variance of residual is the same for any value of X.
3. Independence: observations are independent of each other.
4. Normality: For any fixed value of X, y is normally distributed.

We can study these assumptions by writing

```
par(mfrow=c(2,2))
plot(lwage.lm)
```

And we get:



Note: If the model does not meet the linear model assumption, we would expect to see residuals that are very large (big positive value or big negative value).

To assess if the homoscedasticity assumption is met we look to make sure that there is no pattern in the residuals and that they are equally spread around the y = 0 line.

Also from the third plot (scale-location plot) we can see if residuals are spread equally along the ranges of predictors. Since we find equally spread residuals around a horizontal line without distinct patterns we state that the assumption is met.

The Q-Q plot check the normality assumption by comparing the residuals to "ideal" normal observations. Observations lie well along the 45-degree line in the QQ-plot, so we may assume that normality of residuals holds here.

The fourth plot is a measure of the influence of each observation on the regression coefficients (it helps to identify influential data points on our model). We look for cases outside of a dashed line called “Cook’s distance”. When cases are outside of the Cook’s distance, they are influential to the regression results. In the plot we can’t see the Cook’s distance therefore it seems there aren’t “outlier” that are influential but we can see an extreme point (560), with a standardized residuals below -4.

Otherwise if we want to plot only the first graph we write.

```
resid = residuals(lwage.lm)
n = length(resid)
MSE = 1/ (n - 2) * sum((resid ^ 2))
MSE
# Combine residuals and fitted values into a data frame
result = data.frame(fitted_values = fitted.values(lwage.lm),
                    residuals = residuals(lwage.lm))
# Load library and plot residuals versus fitted values
library(ggplot2)
ggplot(data = result, aes(x = fitted_values, y = residuals)) +
  geom_point(pch = 1, size = 2) +
  geom_abline(intercept = 0, slope = 0) +
  xlab(expression(paste("fitted value ", widehat(wage)))) +
  ylab("residuals")
```

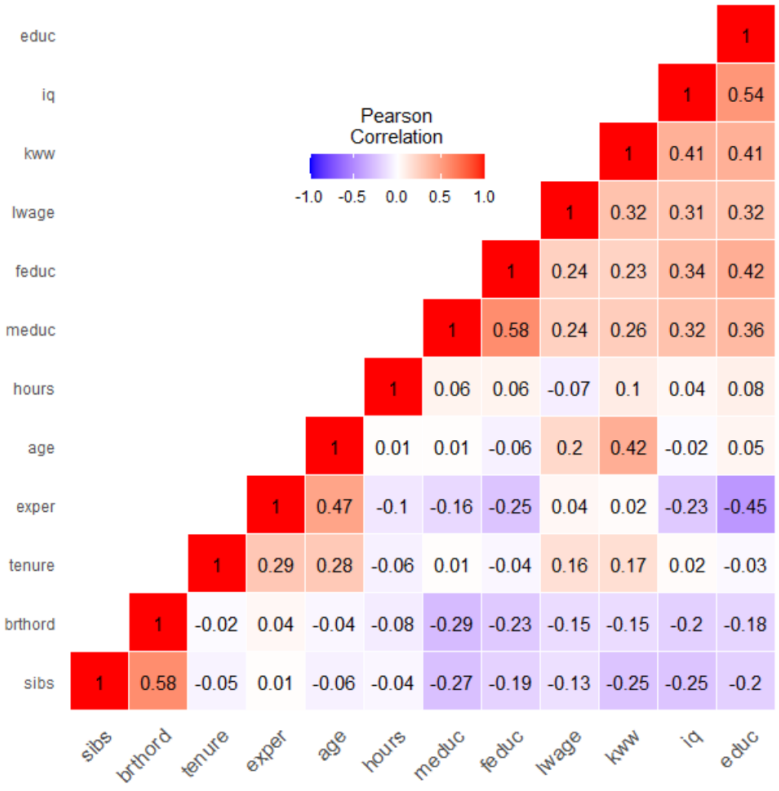
If we are curious to find the observation with the largest fitted value and the observation with the largest lwage:

```
which.max(as.vector(fitted.values(lwage.lm))) #16
which.max(wage$lwage) #499
```

If we want to have a graphical interpretation of the correlation among variables we can write:

```
cor(wage)
cormat <- round(cor(wage),2)
head(cormat)
library(reshape2)
melted_cormat <- melt(cormat)
# Get lower triangle of the correlation matrix
get_lower_tri<-function(cormat){
  cormat[upper.tri(cormat)] <- NA
  return(cormat)
}
# Get upper triangle of the correlation matrix
get_upper_tri <- function(cormat){
  cormat[lower.tri(cormat)]<- NA
  return(cormat)
}
```

```
reorder_cormat <- function(cormat){
  # Use correlation between variables as distance
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  cormat <-cormat[hc$order, hc$order]
}
# Reorder the correlation matrix
cormat <- reorder_cormat(cormat)
upper_tri <- get_upper_tri(cormat)
# Melt the correlation matrix
melted_cormat <- melt(upper_tri, na.rm = TRUE)
library(ggplot2)
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+  geom_tile(color = "white")+
   scale_fill_gradient2(low = "blue", high = "red", mid = "white",  midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Pearson\nCorrelation") +  theme_minimal()+
   theme(axis.text.x = element_text(angle = 45, vjust   = 1,   size = 12, hjust = 1))+  coord_fixed()
# Print the heatmap
ggheatmap +
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))
```



#Part 2 ----------

Looking at the summary, comment the iq's coefficient also referring to the response variable. Create a new data frame that allows you to build a second linear model (mod1)
using only scale iq as regressor. Make a plot of the response and iq (scaled), and add the regression line of the model on the same graph. From this new model, an increase in IQ of 1 unit is estimated to increase wage by what percentage?

Looking at the summary of the data we can see that Xiq estimate (beta) is positive. That means that an increase of iq leads an increase of lwage.

Moreover as we said before the more the stars beside the variable's p-value, the more significant the variable.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.89465 -0.21419  0.00731  0.24213  1.25949

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.280010   0.225176  23.448  < 2e-16 ***
Xhours      -0.005898   0.002019  -2.921  0.00361 **
Xiq          0.003775   0.001236   3.053  0.00236 **
Xkww         0.005815   0.002472   2.352  0.01898 *
Xeduc        0.038872   0.009244   4.205 2.97e-05 ***
Xexper       0.012178   0.004628   2.631  0.00871 **
Xtenure      0.007340   0.003015   2.435  0.01517 *
Xage         0.008818   0.006220   1.418  0.15675
Xsibs        0.007572   0.008035   0.942  0.34636
Xbrthord    -0.019052   0.011981  -1.590  0.11228
Xmeduc       0.009270   0.006434   1.441  0.15012
Xfeduc       0.009929   0.005594   1.775  0.07641 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3667 on 651 degrees of freedom
Multiple R-squared:  0.222,     Adjusted R-squared:  0.2088
F-statistic: 16.88 on 11 and 651 DF,  p-value: < 2.2e-16
```

The Null Hypothesis is that the coefficients associated with the variables is equal to zero. The alternate hypothesis is that the coefficients are not equal to zero. So we can state that for iq exists a relationship the response.

However in our case, both these p-Values are well below the 0.05 threshold, so we can conclude our model is statistically significant.

Now let's build a new linear model using only scaled iq as regressor:

```
IQ <- scale(wage$iq)
mod1 <- lm(y~IQ, data=wage)
mod1
summary(mod1)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.11742 -0.23547  0.02697  0.27216  1.18808

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.81430    0.01523  447.41  < 2e-16 ***
IQ           0.12788    0.01524    8.39 2.95e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3922 on 661 degrees of freedom
Multiple R-squared:  0.09624,   Adjusted R-squared:  0.09487
F-statistic: 70.39 on 1 and 661 DF,  p-value: 2.95e-16
```

Note that an increase by 1 of scaled IQ is associated to an increase by 0.128 of lwage. And if we want to estimate the increase of wage when IQ increases by 1 are:

$$lwage(b) - lwage(a) = \beta \rightarrow \ln\left(\frac{b}{a}\right) = \beta \rightarrow \frac{b}{a} = e^{\beta}$$

$\beta = 0.12788$, thus the expected increase of the wage is $e^{\beta} - 1 = 13.64\%$.

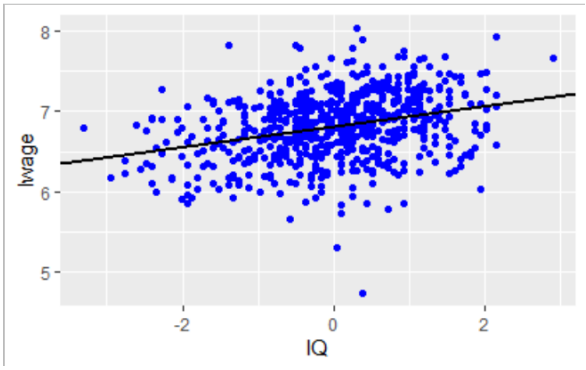
We can also consider the Mean Square Error, knowing that the degrees of freedom are n-2, as we have just 1 covariate.

$$MSE = \frac{1}{n-2}\sum_{i}^{n}(y_i - \hat{y}_i)^2$$

```
resid = residuals(mod1)
n = length(resid)
MSE = 1/ (n - 2) * sum((resid ^ 2))
MSE #0.1538
```

If we plot the correlation between the response and the covariate chosen we get:

```
library(ggplot2)
ggplot(data = wage, aes(x = IQ, y = lwage)) +
  geom_point(color = "blue") +
  geom_abline(intercept = beta[1], slope = beta[2], size = 1) +
  xlab("IQ")
```



Let's see also the relation between fitted values and the residuals of this new model:

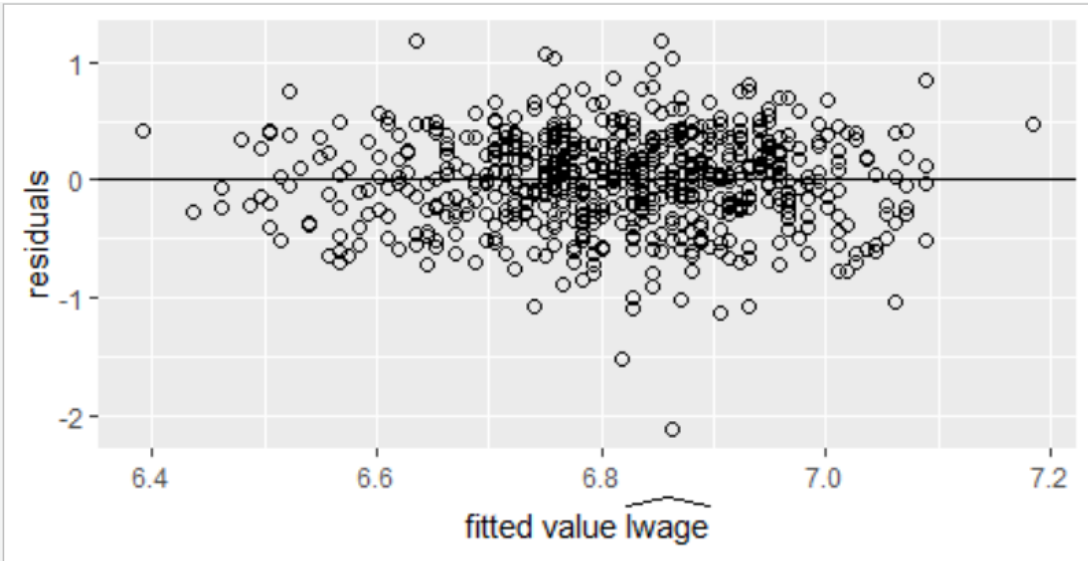
```
result = data.frame(fitted_values = fitted.values(mod1), residuals = residuals(mod1))

library(ggplot2)

ggplot(data = result, aes(x = fitted_values, y = residuals)) +  geom_point(pch = 1, size = 2) +

  geom_abline(intercept = 0, slope = 0) + xlab(expression(paste("fitted value ", widehat(lwage)))) +

  ylab("residuals")
```



#Part 3 ----------

Estimate the variance using the sample unbiased estimator of the variance, estimate the coefficient of all the regressors and estimate the variance of the coefficients.

The likelihood of the ordinary normal linear model is

$$f\left(y|X,\beta,\sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(y-X\beta)^t(y-X\beta)\right]$$

The MLE (maximum likelihood estimation) of β is solution of the least squares minimisation problem

$$\min_{\beta}(y-X\beta)^t(y-X\beta) = \min_{\beta}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_{i1}-\ldots-\beta_k x_{ik}\right)^2 ,$$

namely

$$\hat{\beta} = (X^tX)^{-1}X^ty$$

$\hat{\beta}$ is an unbiased estimator of β, i.e. $\mathbb{E}[\hat{\beta}|\beta] = \beta$

$Var(\hat{\beta}|\sigma^2, X) = \sigma^2(X^tX)^{-1}$

Unbiased estimator of σ^2

$$\hat{\sigma}^2 = \frac{1}{n-k-1}(y-X\hat{\beta})^t(y-X\hat{\beta}) = \frac{s^2}{n-k-1}$$

Therefore:

```
betahat=solve(t(X)%*%(X),t(X)%*%y) #estimate of the coefficients of the complete model

betahat1=solve(t(IQ)%*%(IQ),t(IQ)%*%y) #estimate of the coefficient of the iq model

# Sample unbiased estimate of the variance:

n <- nrow(wage)

p <- ncol(X)

s2 <- t(y-mod$coeff[1]*rep(1,n)-X%*%mod$coeff[2:12])%*%(y-mod$coeff[1]*rep(1,n)-
X%*%mod$coeff[2:12]) #leave out the intercept

s2_sample <- s2/(n-p-1) #estimation of sigma2 we get 0.1344

# Estimation of the variance of betahat

as.numeric(s2_sample)*(t(X)%*%X)

# Estimation of the variance of betahat

diag(as.numeric(s2_sample)*solve(t(X)%*%X))
```

And respectively we get:

#estimate of the coefficients of the complete model

```
hours    -0.04222622
iq        0.05543988
kww       0.04377875
educ      0.08673911
exper     0.05185947
tenure    0.03710738
age       0.02701057
sibs      0.01696828
brthord  -0.02834138
meduc     0.02617012
feduc     0.03264714
```

#estimate of the coefficient of the iq scaled model

```
0.1278764
```

Note that confint(mod):

```
                  2.5 %       97.5 %
(Intercept)  6.786336039  6.842258099
Xhours      -0.070614431 -0.013838002
Xiq          0.019784381  0.091095374
Xkww         0.007226557  0.080330946
Xeduc        0.046236382  0.127241847
Xexper       0.013160384  0.090558551
Xtenure      0.007179140  0.067035617
Xage        -0.010400499  0.064421636
Xsibs       -0.018389061  0.052325621
Xbrthord    -0.063338629  0.006655863
Xmeduc      -0.009496365  0.061836600
Xfeduc      -0.003474327  0.068768600
```

Only intercept, hours, iq, kww, educ, exper and tenure don't have beta = 0 in its confidence interval therefore the null hypothesis can be rejected.

And the estimation of the variance of betahat is:

```
       hours           iq          kww         educ        exper       tenure
0.0002090083 0.0003297163 0.0003465089 0.0004254574 0.0003884086 0.0002323003
         age         sibs      brthord        meduc        feduc
0.0003629841 0.0003242251 0.0003176546 0.0003299195 0.0003383905
```

#part4 ----------

CONJUGATE PRIOR. Consider the hyperparameters: $\tilde{\beta}$ = 0, M is an identity matrix divided by a coefficient c = 100, a = 1 and b = 1. Compute the posterior mean of sigma2, the posterior mean of beta2 (coefficient related to IQ) and posterior variance of the second component of beta. What are the main problems of this approach?

For a conjugate model we know that:

Conjugate prior for LM

$$\beta|\sigma^2, X \sim \mathcal{N}_{k+1}(\tilde{\beta}, \sigma^2 M^{-1}),$$

$$\sigma^2|X \sim \mathcal{IG}(a, b), \qquad a, b > 0,$$

where M $(k+1, k+1)$ positive definite symmetric matrix.

$$\tilde{\beta}_{updated} = (M + X^tX)^{-1}\{(X^tX)\hat{\beta} + M\tilde{\beta}\}$$

$$M_{updated} = (M + X^tX)^{-1}$$

$$a_{updated} = \frac{n}{2} + a$$

$$b_{updated} = b + \frac{s^2}{2} + \frac{(\tilde{\beta} - \hat{\beta})^t \left(M^{-1} + (X^tX)^{-1}\right)^{-1} (\tilde{\beta} - \hat{\beta})}{2}$$

Therefore after setting the coefficient and $\tilde{\beta}$ = 0 we compute:

```
a = 1
b = 1
c=100
# prior mean of sigma2:
mean.prior <- b/(a+b))
# prior variance of sigma2:
var.prior <- b^2/((a+b+1)^2*(a+b)))

# Posterior means of sigma2
M=1/c*diag(11)
T=solve(solve(M)+solve(t(X)%*%X))
(2*b+s2+t(betahat)%*%T%*%betahat)/(n+2*a-2) #we get 0.13501
# Posterior means of beta
b1beta=solve(M+t(X)%*%X)%*%t(X)%*%y
b1beta[2] #we get 0.055439
```

Note that if we compare b1beta[2] with betahat[2] they are almost identical because we have a large number of observations and the coefficient c is high.

Obviously, the choices of the hyperparameters a, b, beta_tilde, and M are important, but they are not always easy in practice. In particular, building prior beliefs on the correlation between the components of beta is often difficult. This is one of the reasons why M is frequently chosen as a diagonal matrix or even as a multiple of the identity matrix, M = Ik+1/c.
Note that conjugacy is a convenient tool mainly if we want translate prior information on beta.

```
# Posterior variances
T=solve(solve(M)+solve(t(X)%*%X))
E=as.numeric(2*b+s2+t(betahat)%*%T%*%betahat)/(n+2*a-2)*solve(M+t(X)%*%X)
E[2,2] #we get 0.0003311
```

The problem of this approach is that there is a dependence on parameter. But in the case we have a large number of observation even with c=0.1 the results don't really change.

Therefore if we don't have enough information about the prior is better to choose an improper prior, like the Zellner's G prior.

#part 5 ----------

ZELLNER'S G PRIOR. We want to use the Zellner's G prior approach. What are the main advantages respect than the approach of the previous point? Consider the same hyperparameter of the previous point (except for the distribution of sigma2), what represents 1/c? Compute the posterior mean and posterior variance of beta.

A conjugate prior shows its limitations in this setup since there is some lasting influence of the hyperparameters on at least the posterior variance. Therefore a more elaborate noninformative strategy is a noninformative prior analysis. But we first consider a middle-ground perspective where the prior information is available on beta only, adopting the so-called Zellner's G-prior, which somehow settles the problem of the choice of M.
Zellner's G prior allow the experimenter to introduce information about the location parameter of the regression but to bypass the most difficult aspects of the prior specification (the derivation of the prior correlation structure). Zellner's G-prior thus relies on a (conditional) Gaussian prior for beta and an improper (Jeffreys) prior for sig2. The experimenter thus restricts prior determination to the choices of beta_tilde and the constant c. (sig2 is an improper prior, but it is ok if it allows to obtain a finite distribution for the posterior).
"c" can be interpreted as a measure of the amount of information available in the prior relative to the sample. For instance, setting 1/c = 0.5 gives the prior the same weight as 50% of the sample.
Mind that the Zellner prior (improper prior) faces difficulties for model comparison as in variable selection.

Assume the ε are iid normal with zero mean and variance sigma2. Then the g-prior for beta is the multivariate normal distribution as:

$$\beta|\sigma^2, X \sim \mathcal{N}_{k+1}(\tilde{\beta}, c\sigma^2(X^tX)^{-1})$$
$$\sigma^2|X \sim \pi(\sigma^2) = \sigma^{-2} \qquad \text{Hyper-par: } c > 0.$$

The posterior has the following structure:

$$\beta|\sigma^2, y, X \sim \mathcal{N}_{k+1}\left(\frac{c}{c+1}(\frac{\tilde{\beta}}{c}+\hat{\beta}), \frac{\sigma^2 c}{c+1}(X^tX)^{-1}\right)$$

$$\sigma^2|y, X \sim \mathcal{IG}\left(\frac{n}{2}, \frac{s^2}{2}+\frac{1}{2(c+1)}(\tilde{\beta}-\hat{\beta})^tX^tX(\tilde{\beta}-\hat{\beta})\right)$$

And

$$\beta|y, X \sim \mathcal{T}_{k+1}\left(n, \frac{c}{c+1}\left(\frac{\tilde{\beta}}{c}+\hat{\beta}\right), \frac{c(s^2+(\tilde{\beta}-\hat{\beta})^tX^tX(\tilde{\beta}-\hat{\beta})/(c+1))}{n(c+1)}(X^tX)^{-1}\right).$$

Therefore we can text:

```
# posterior mean of beta for c=100 : c/(c+1) * beta   (beta_tilde = 0)
c/(c+1)*betahat #we get 0.05489

# posterior variance of beta for c=100
diag(c/(n*(c+1))*as.numeric(s2+t(betahat)%*%t(X)%*%X%*%betahat/(c+1))*solve(t(X)%*%X))
#we het 0.00032144
```

exact Bayesian inference (derive a close-form posterior). So we would need to use sampling methods such as MCMC to derive the posterior.

#part 6 ----------

Consider the model with only IQ as regressor (mod1). Compute the credible intervals, interpret the results for the response for an increase of 1 points for IQ. We want to predict the wage of a new person. Construct the current prediction by creating a sequence from the min of IQ (scaled) to the max, with a length.out = 100. Compute the lower and the upper bound for the mean and for the prediction. We want also to predict the wage of a person that is an outlier in our dataset:

l i b r a r y (car)
o u t l i e rTe s t (mod1)
Plot all the results pointing out the outlier. What amount of dollars should he receive?

In order to compute the credible intervals we can write:

```
confint(mod1)
```

And we obtain:

```
                2.5 %    97.5 %
(Intercept) 6.78439115 6.8442030
iqscaled    0.09794794 0.1578049
```

Note that these intervals coincide with the confidence intervals from the frequentist approach, but the interpretation is different. In Bayesian approach we say that there is a 95% chance that lwage will increase by 0.09 up to 0.157 for every additional 1 increase of IQ(scaled).

Now let's predict the wage of a new person.

From theory we know that the predictive prior is:

$$y_{new}|\sigma^2, \beta, X_{new} \sim \mathcal{N}_m(X_{new}\beta, \sigma^2 I_m).$$

And its predictor under squared error loss is the posterior predictive mean:

$$\begin{aligned} \mathbb{E}^\pi[y_{new}|\sigma^2, y, X, X_{new}] &= \mathbb{E}^\pi[\mathbb{E}^\pi(y_{new}|\beta, \sigma^2, y, X, X_{new})|\sigma^2, y, X, X_{new}] \\ &= \mathbb{E}^\pi[X_{new}\beta|\sigma^2, y, X, X_{new}] \\ &= X_{new}\frac{\tilde{\beta} + c\hat{\beta}}{c+1} \end{aligned}$$

In order to *compute the lower and the upper bound for the mean and for the prediction the code is:*

```
df <- data.frame(x=as.vector(IQ), y=y)
mod1 <- lm(y~x, data=df)
alpha = mod1$coefficients[1]
beta = mod1$coefficients[2]
new_x = seq(min(df$x), max(df$x),
        length.out = 100)
y_hat = alpha + beta * new_x
ymean = data.frame(predict(mod1,
                   newdata = data.frame(x = new_x),
                   interval = "confidence",
                   level = 0.95))
ypred = data.frame(predict(mod1,
                   newdata = data.frame(x = new_x),
                   interval = "prediction",
                   level = 0.95))
output = data.frame(x = new_x, y_hat = y_hat, ymean_lwr = ymean$lwr, ymean_upr = ymean$upr,
                ypred_lwr = ypred$lwr, ypred_upr = ypred$upr)
head(output)
```

```
          x    y_hat ymean_lwr ymean_upr ypred_lwr ypred_upr
1 -3.301155 6.392157  6.288932  6.495383  5.615228  7.169086
2 -3.238566 6.400161  6.298727  6.501595  5.623468  7.176854
3 -3.175976 6.408165  6.308519  6.507810  5.631703  7.184626
4 -3.113387 6.416168  6.318308  6.514029  5.639934  7.192403
5 -3.050798 6.424172  6.328093  6.520251  5.648160  7.200184
6 -2.988209 6.432176  6.337875  6.526476  5.656382  7.207969
```

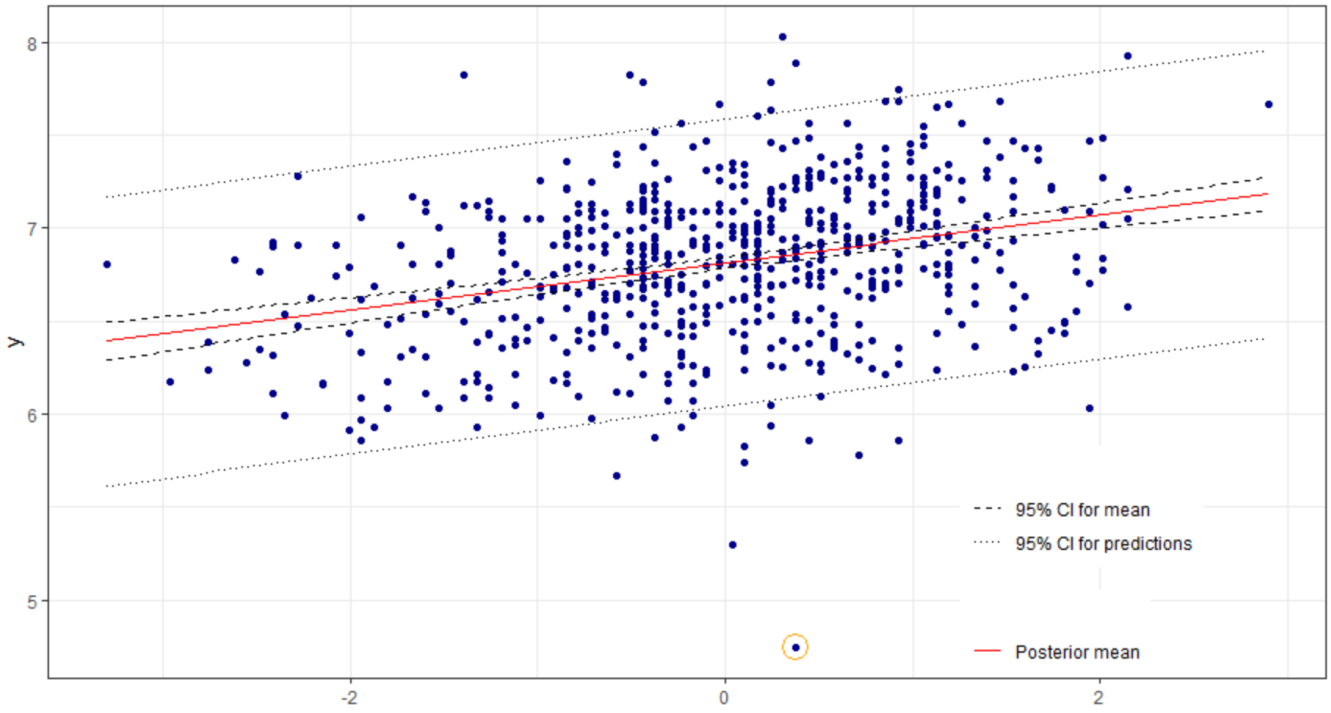
Lastly we want to find the outlier anche predict its wage:

```
# Extract potential outlier data point
library(car)
outlierTest(mod1)
outlier = data.frame(x = IQ[560], y = y[560])
outlier
```

From outlierTest(mod1) we get 560 (the same extreme point we get from the plot “Residuals vs Levarege” with a standardized residuals below -4). That plot helps us to find influential cases if any. Not all outliers are influential in linear regression analysis and since data have extreme values, they might not be influential to determine a regression line. That means, the results wouldn’t be much different if we either include or exclude them from analysis. They follow the trend in the majority of cases and they don’t really matter.

Now let’s plot everything using the following code:

```
plot1 = ggplot(data = df, aes(x = x, y = y)) + geom_point(color = "blue4")
plot2 = plot1 +
  geom_line(data = output, aes(x = new_x, y = y_hat, color = "first"), lty = 1) +
  geom_line(data = output, aes(x = new_x, y = ymean_lwr, lty = "second")) +
  geom_line(data = output, aes(x = new_x, y = ymean_upr, lty = "second")) +
  geom_line(data = output, aes(x = new_x, y = ypred_upr, lty = "third")) +
  geom_line(data = output, aes(x = new_x, y = ypred_lwr, lty = "third")) +
  scale_colour_manual(values = c("orange"), labels = "Posterior mean", name = "") +
  scale_linetype_manual(values = c(2, 3), labels = c("95% CI for mean", "95% CI for predictions")
 , name = "") +  theme_bw() +  theme(legend.position = c(1, 0), legend.justification = c(1.5, 0))
# Identify potential outlier
plot2 + geom_point(data = outlier, aes(x = x, y = y), color = "red", pch = 1, cex = 6)
```



And the prediction from that extreme point encircled is:

```
pred.560 = predict(mod1, newdata = df[560, ], interval = "prediction", level = 0.95)

out = cbind(df[560,]$x, pred.560)

colnames(out) = c("IQ", "prediction", "lower", "upper")

out
```

We get:

```
            IQ prediction    lower    upper
560 0.3757871   6.862351 6.091648 7.633055
```

So we would expect that a person with an IQ(scaled) of 0.3758 will have 6.8623 lwage (e^6.86 = 953.37 $).
So given our observed data, there is a 95% probability that the considered value falls within 6,091 and 7,633.
Note that by using the value from the dataset: wage$lwage[560] we obtain 4.745 (about 115$!).