Mohammad Javad Ranjbar

810101173

Homework 6

Statistical Inference, Fall 2022

Contents

Question 1:

We consider Y as $Y_i \sim N(E(Y_i), \sigma^2)$: $Y_i = E(Y_i) + e_i = \beta_0 + \beta_1 * x_{i1} + \cdots + \beta_k * x_{ik} + e_i$ and e_i are iid and $e_i \sim N(0, \sigma^2)$ and Y_i are independent.

The likelihood function would look like this:

$$\mathrm{L}(\beta, \sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}\Sigma(y_i - E(Y_i))^2} = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}Q(\beta)}$$

Since e^{-x} is a decreasing function the maximum of L is calculated by minimizing $Q(\beta)$. Therefore we take log and we can write the function like this:

$$\ln(\,\mathrm{L}(\hat{\beta}, \sigma)) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}Q(\hat{\beta})$$

now we want to maximize for σ:

$$\frac{\partial}{\partial\sigma}\ln(\,\mathrm{L}(\hat{\beta}, \sigma)) = -\frac{n}{\sigma} + \frac{2}{2\sigma^3}Q(\hat{\beta}) = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\,Q(\hat{\beta}) = \frac{1}{n}\mathrm{SS}_{\mathrm{res}}$$

Now we insert the answer in the original likelihood function:

$$\mathrm{L}(\hat{\beta}, \hat{\sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\hat{\sigma}} e^{-\frac{1}{2*Q(\hat{\beta})}Q(\hat{\beta})} = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{(n)^{\frac{n}{2}}}{(Q(\hat{\beta}))^{\frac{n}{2}}} e^{-\frac{n}{2}}$$

The likelihood ratio statistic is:

$$\Lambda(y) = \frac{L_w}{L_\Omega} = \frac{\frac{1}{(2\pi)^{\frac{n}{2}}} \frac{(n)^{\frac{n}{2}}}{(Q(\hat{\eta}))^{\frac{n}{2}}} e^{-\frac{n}{2}}}{\frac{1}{(2\pi)^{\frac{n}{2}}} \frac{(n)^{\frac{n}{2}}}{(Q(\hat{\beta}))^{\frac{n}{2}}} e^{-\frac{n}{2}}} = \frac{(Q(\hat{\beta}))^{\frac{n}{2}}}{(Q(\hat{\eta}))^{\frac{n}{2}}} = \left(\frac{\mathrm{SS}_{full}}{\mathrm{SS}_{\mathrm{red}}}\right)^{\frac{n}{2}} \xrightarrow{log} W = -2ln(\Lambda) = \mathrm{n}ln(\frac{\mathrm{SS}_{full}}{\mathrm{SS}_{\mathrm{red}}})$$

Logistic regression will reject H_0 if the W is large. This I equivalent to rejecting H_0 using f static:

We know $F = \frac{\mathrm{SS}_{\mathrm{red}} - \mathrm{SS}_{full}}{\mathrm{SS}_{full}} = \frac{\mathrm{SS}_{full}}{\mathrm{SS}_{\mathrm{red}}} - 1$

We conclude:

$$W = \mathrm{n}ln(f + 1)$$

Therefore, the likelihood ratio statistic for the multiple regression model is a monotonic function of the model F-statistic for the multiple regression model.

Question 2:

The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data. When estimating the model we minimize the residual sum of squares. In the unrestricted model we can always choose the combination of coefficients that the restricted model chooses. Hence the restricted model can never do better than the unrestricted one.

Question 3:

a)

$$b_1x + b_0 = y \Rightarrow b_1 = \frac{y - b_0}{x} \Rightarrow b_1 = \frac{\bar{y} - b_0}{\bar{x}} = \frac{3.9983 - 4.010}{-0.0883} = 0.13250283125$$

b) Based on the figure, it is unclear whether the two variables are correlated (I should mention that there is a slight positive correlation). However, based on the summary table as the $p - value = 0$. Therefore, the $p - value$ provides strong evidence that there is a positive relationship between beauty and score. In another word, the evidence of $p - value = 0$ reject the null hypothesis. Meaning that, Data rejects the hypothesis of beauty and score not being correlated.

Question 4:

a) $R = 0.636, s_y = 113, s_x = 99$

$$b_1 = R * \frac{s_y}{s_x} \Rightarrow b_1 = 0.636 * \frac{113}{99} = 0.726$$

Now the linear regression equation will be as followed:

$$b_1x + b_0 = y$$

We write the above equation for min value of each variable:

$$b_1\bar{x} + b_0 = \bar{y} \Rightarrow b_0 = 108 - 129 * 0.726 = 50.592$$

b) The slope means for each additional mile the travel time will be 0.726 minute longer.
The intercept means if the distance is zero, we have 50.592 minutes of travel time. Which does not make any sense in this context.

c) $R^2 = 0.404496$
$40.\ 4496\%$ of the variability in the travel time the is explained by this model. In this context, $40.\ 4496\%$ variability in the travel time can be explained by distance. There are other factor such as traffic, weather, etc. that can affect the travel time.

d) It takes 125.37 minutes to reach Los Angeles.

$$0.726 * x + 50.592 = y \Rightarrow 0.726 * 103 + 50.592 = 125.37$$

e) The actual time is 168 and the model has estimated 125.37. Therefore, the residual value is $168 - 125.37 = 42.63$ minutes. Which means it actually takes more time to get to Los Angeles. The residual value is positive meaning that the model underestimates the travel time.

f) Since the mean value of distance is 129 and it's deviation is 99. The 500 miles is more than 3*deviation away from the mean which is very far away. 500 miles will be outside of the realm of the original data and therefore, this model is not valid for it. If we use this model for 500 miles value extrapolation happens which is wrong.

Question 5:

The line passes through (0.25,0) and (0.75,7). Therefore we can calculate the slope by:

$$b_1 = \frac{7 - 0}{0.75 - 0.25} = 14$$

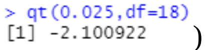
And the equation will be as followed:

$14 * (x - 0.25) = (y - 0) \Rightarrow 14x - 3.5 = y$

Question 6:

We have 20 students therefore df = 20 − 2 = 18

For the 95% confidence interval we have $t^*_{18} = 2.1$ ( )



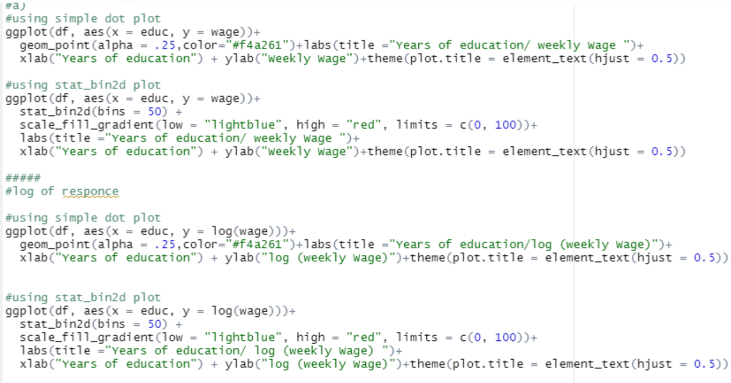
```
> qt(0.025,df=18)
[1] -2.100922
```

The confidence interval is calculated by below formula:

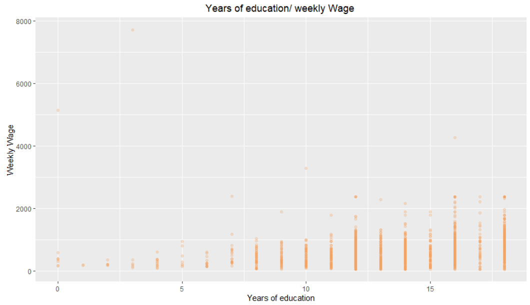
$$b_1 \pm t^*_{18} * SE_{b_1} \Rightarrow 0.164 \pm 2.1 * 0.057 = (0.0443, 0.2837)$$

Question 7:

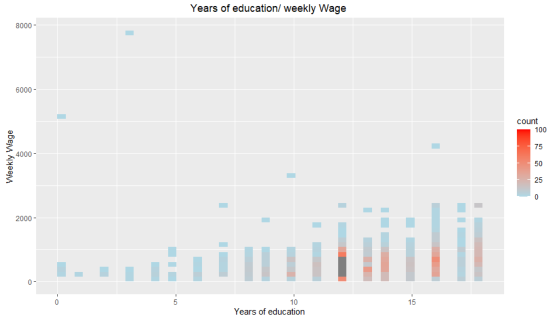
a) Our explanatory variable is years of education (eudc) and our response variable is wage:
We can use a simple scatter plot. In Which in order to avoid overplotting. We fade the dots in places that there is not lots of data present.



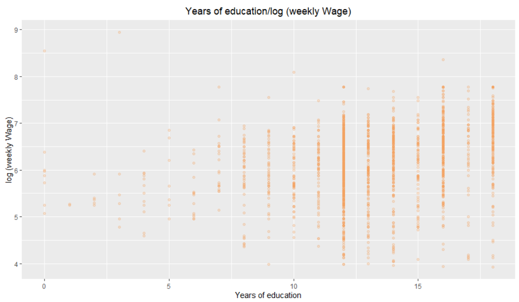
```
#a)
#using simple dot plot
ggplot(df, aes(x = educ, y = wage))+
  geom_point(alpha = .25,color="#f4a261")+labs(title ="Years of education/ weekly Wage ")+
  xlab("Years of education") + ylab("Weekly Wage")+theme(plot.title = element_text(hjust = 0.5))

#using stat_bin2d plot
ggplot(df, aes(x = educ, y = wage))+
  stat_bin2d(bins = 50) +
  scale_fill_gradient(low = "lightblue", high = "red", limits = c(0, 100))+
  labs(title ="Years of education/ weekly Wage ")+
  xlab("Years of education") + ylab("Weekly Wage")+theme(plot.title = element_text(hjust = 0.5))

#####
#log of responce

#using simple dot plot
ggplot(df, aes(x = educ, y = log(wage)))+
  geom_point(alpha = .25,color="#f4a261")+labs(title ="Years of education/log (weekly Wage)")+
  xlab("Years of education") + ylab("log (weekly Wage)")+theme(plot.title = element_text(hjust = 0.5))


#using stat_bin2d plot
ggplot(df, aes(x = educ, y = log(wage)))+
  stat_bin2d(bins = 50) +
  scale_fill_gradient(low = "lightblue", high = "red", limits = c(0, 100))+
  labs(title ="Years of education/ log (weekly Wage) ")+
  xlab("Years of education") + ylab("log (weekly Wage)")+theme(plot.title = element_text(hjust = 0.5))
```

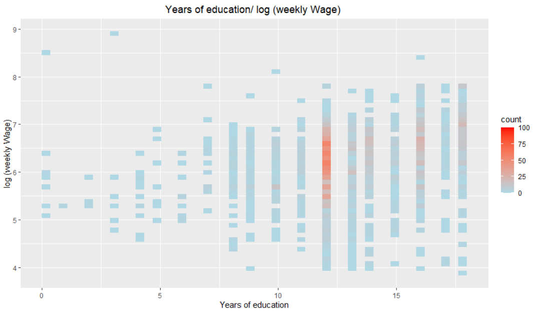


Or we can just stat_bin2d plot:

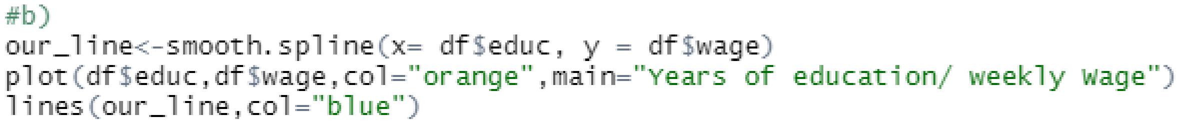


And the plots of log(weakly wage) are:

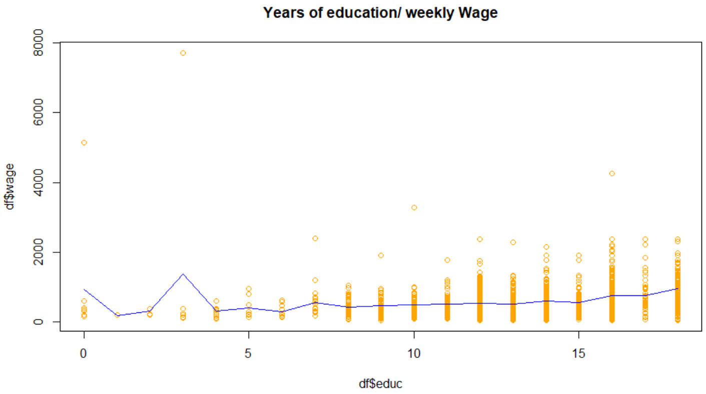




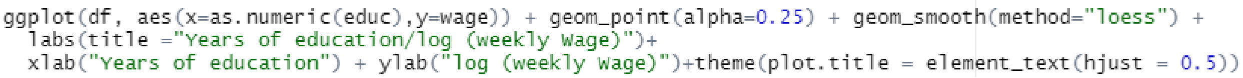
b) It appears the fit is close to linear. However, some outliers present in data have forced the fitted line to have some spikes in area near these values.



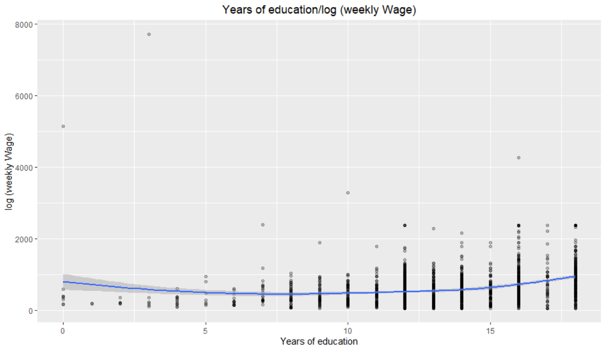
```
#b)
our_line<-smooth.spline(x= df$educ, y = df$wage)
plot(df$educ,df$wage,col="orange",main="Years of education/ weekly Wage")
lines(our_line,col="blue")
```



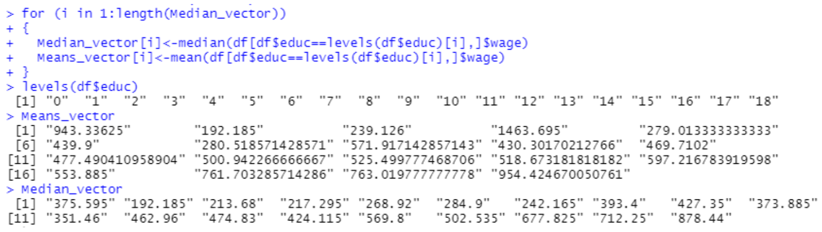
c) This fit handled the outliers better than smoothing splines. Also, it is still close to a linear line.



```
ggplot(df, aes(x=as.numeric(educ),y=wage)) + geom_point(alpha=0.25) + geom_smooth(method="loess") +
  labs(title ="Years of education/log (weekly Wage)")+
  xlab("Years of education") + ylab("log (weekly Wage)")+theme(plot.title = element_text(hjust = 0.5))
```

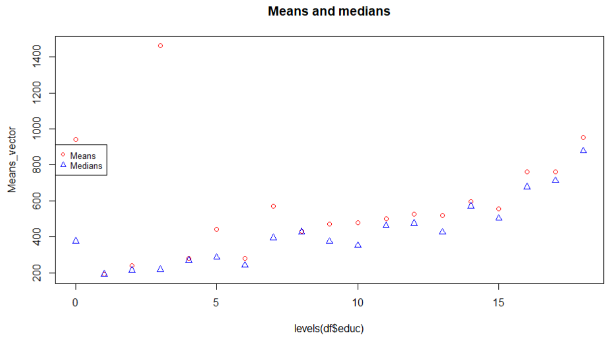


d) We calculate the means and median for each group:



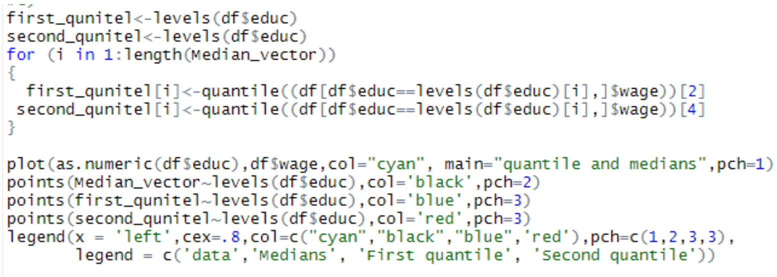
```
> for (i in 1:length(Median_vector))
+ {
+   Median_vector[i]<-median(df[df$educ==levels(df$educ)[i],]$wage)
+   Means_vector[i]<-mean(df[df$educ==levels(df$educ)[i],]$wage)
+ }
> levels(df$educ)
 [1] "0"  "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14" "15" "16" "17" "18"
> Means_vector
 [1] "943.33625"        "192.185"          "239.126"          "1463.695"         "279.013333333333"
 [6] "439.9"            "280.518571428571" "571.917142857143" "430.30170212766"  "469.7102"
[11] "477.490410958904" "500.942266666667" "525.499777468706" "518.673181818182" "597.216783919598"
[16] "553.885"          "761.703285714286" "763.019777777778" "954.424670050761"
> Median_vector
 [1] "375.595" "192.185" "213.68"  "217.295" "268.92"  "284.9"   "242.165" "393.4"   "427.35"  "373.885"
[11] "351.46"  "462.96"  "474.83"  "424.115" "569.8"   "502.535" "677.825" "712.25"  "878.44"
```

And the data would look like this:

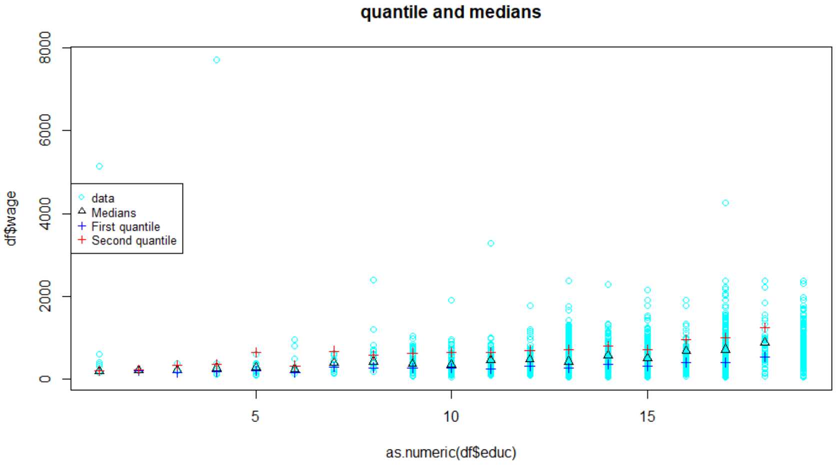


In this data, wage has some outliers which can affect the result significantly. Therefore, the median has showed a better result.

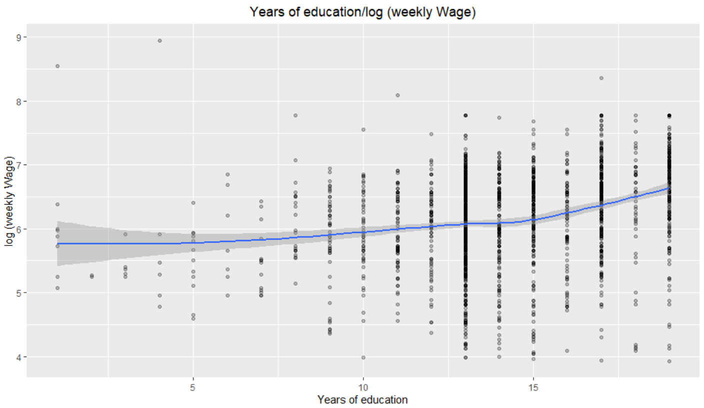
e)

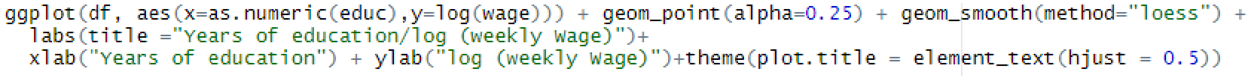


```
first_qunitel<-levels(df$educ)
second_qunitel<-levels(df$educ)
for (i in 1:length(Median_vector))
{
  first_qunitel[i]<-quantile((df[df$educ==levels(df$educ)[i],]$wage))[2]
 second_qunitel[i]<-quantile((df[df$educ==levels(df$educ)[i],]$wage))[4]
}

plot(as.numeric(df$educ),df$wage,col="cyan", main="quantile and medians",pch=1)
points(Median_vector~levels(df$educ),col='black',pch=2)
points(first_qunitel~levels(df$educ),col='blue',pch=3)
points(second_qunitel~levels(df$educ),col='red',pch=3)
legend(x = 'left',cex=.8,col=c("cyan","black","blue",'red'),pch=c(1,2,3,3),
       legend = c('data','Medians', 'First quantile', 'Second quantile'))
```



f) No it is not better, the original data had a close to linear fitting.

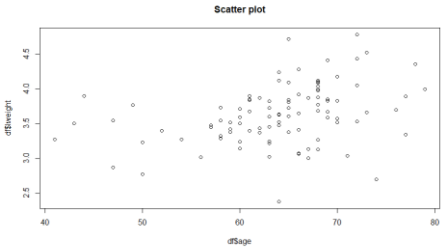


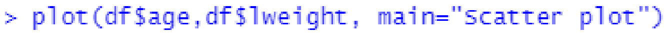


```
ggplot(df, aes(x=as.numeric(educ),y=log(wage))) + geom_point(alpha=0.25) + geom_smooth(method="loess") +
  labs(title ="Years of education/log (weekly wage)")+
  xlab("Years of education") + ylab("log (weekly wage)")+theme(plot.title = element_text(hjust = 0.5))
```

Question 8:

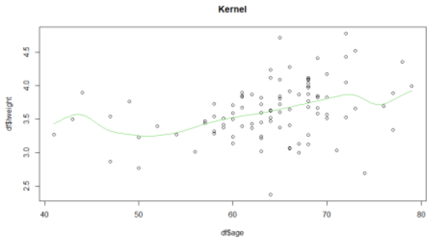
a) Based on the plot, it seems there is a positive relationship between age and lweight. Which, means with the increase in age the lweight increases too.

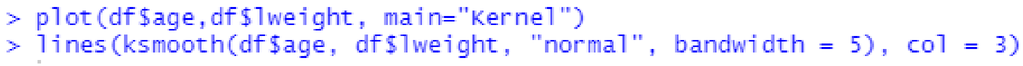




```
> plot(df$age,df$lweight, main="Scatter plot")
```

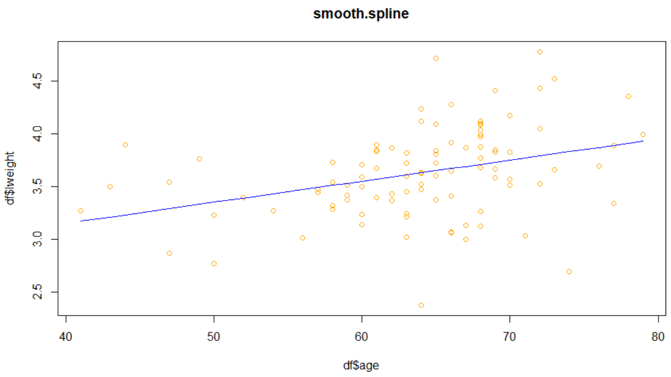
b) The outliers have affected the line. These outliers made the line deviate from a simple linear line. But in general their effects in this data are negligible.

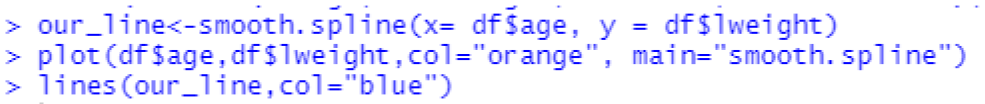




```
> plot(df$age,df$lweight, main="Kernel")
> lines(ksmooth(df$age, df$lweight, "normal", bandwidth = 5), col = 3)
```

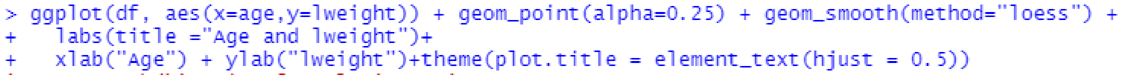
c) Based on the plot below, smoothing spline has determined the best fit is a linear fit.



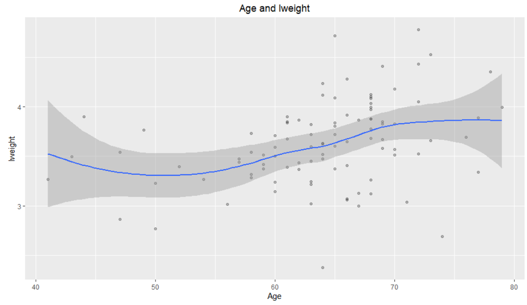


```
> our_line<-smooth.spline(x= df$age, y = df$lweight)
> plot(df$age,df$lweight,col="orange", main="smooth.spline")
> lines(our_line,col="blue")
```

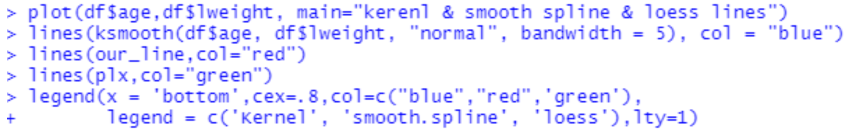
d) Yes, a linear fit seems plausible for this data from visually analysis. This seems reasonable because we could fit a linear line through the data that would be within the 95% confidence of the loess smoother and very close to the line itself.



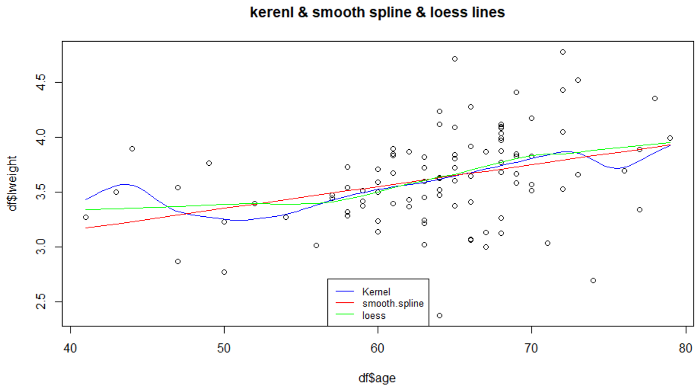
```
> ggplot(df, aes(x=age,y=lweight)) + geom_point(alpha=0.25) + geom_smooth(method="loess") +
+   labs(title ="Age and lweight")+
+   xlab("Age") + ylab("lweight")+theme(plot.title = element_text(hjust = 0.5))
```



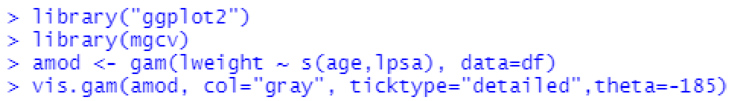
e) The lines are pretty similar and they all are close to a linear line. for some values some lines such as kernel have been more affected by outliers. But, in general they all are good fit for your data.



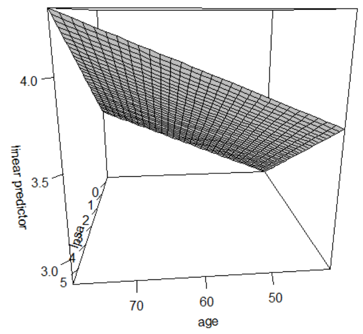
```
> plot(df$age,df$lweight, main="kerenl & smooth spline & loess lines")
> lines(ksmooth(df$age, df$lweight, "normal", bandwidth = 5), col = "blue")
> lines(our_line,col="red")
> lines(plx,col="green")
> legend(x = 'bottom',cex=.8,col=c("blue","red",'green'),
+        legend = c('Kernel', 'smooth.spline', 'loess'),lty=1)
```



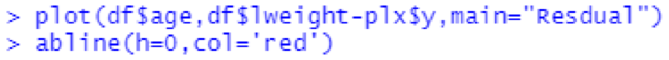
f)



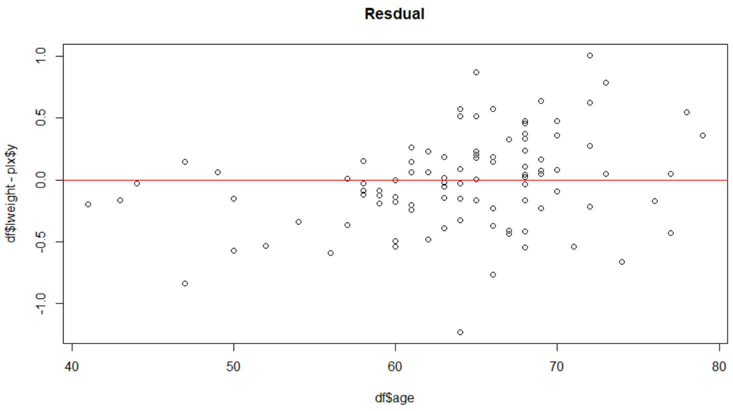
```
> library("ggplot2")
> library(mgcv)
> amod <- gam(lweight ~ s(age,lpsa), data=df)
> vis.gam(amod, col="gray", ticktype="detailed",theta=-185)
```



g) The residuals are disturbed with an almost constant variance around the zero. Which shows that our model is a good estimation for the explanatory variable.



```
> plot(df$age,df$lweight-plx$y,main="Resdual")
> abline(h=0,col='red')
```


Resdual
df$lweight - plx$y
df$age