MACQUARIE
University

Solution Assignment 1 (STAT811 and STAT711)

Benoit Liquet [*1,2] ***and Iris Jiang*** [†1,2]

[1]Macquarie University
[2]School of Mathematical and Physical Sciences

[*]benoit.liquet-weiland@mq.edu.au [†]iris.jiang@mq.edu.au

July 2023

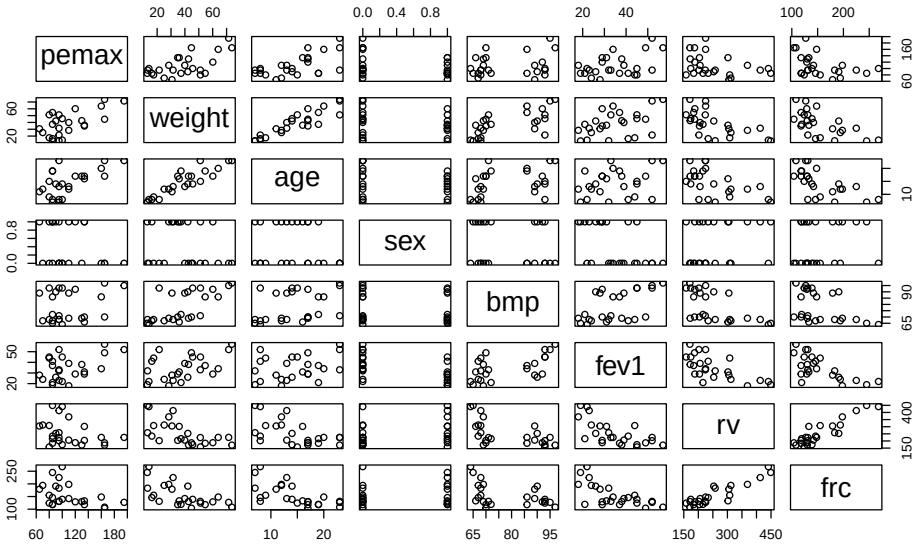
Abstract

This document presents solution of assignment 1. In red color the total marks for each question and in blue the detail of the attribution of the marks.

Question 1 [30 marks]

a) Examine first graphically and numerically correlation between variables. Comments your results. [3 marks]

[1 mark]: for pairwise plots



[1 mark]: for comments

This plot shows a clear linear relation between `age` and `weight`. This relation is not suprising as the study includes children and young adults.

A second clear relation is between `frc` and `rv` which was expected as the FRC is the sum of Expiratory Reserve Volume (ERV) and Residual Volume (RV).

This plot suggests also that `pemax` is link to the `weight` variable and also to the `age` variables.

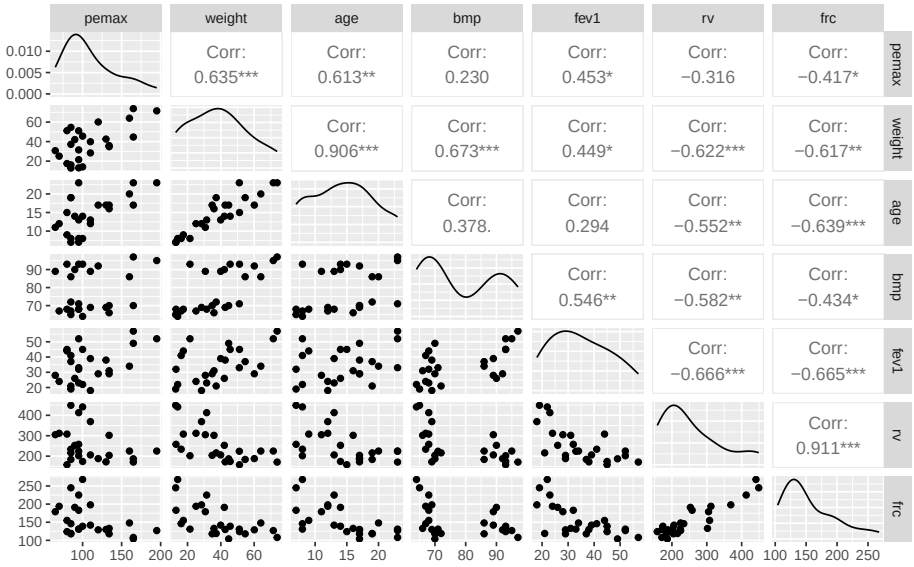
[1 mark]: for numerical correlation results and comments

The following correlation matrix confirms these results with high correlation between: (i) `weight` and `age` equals to 0.91, (ii) `frc` and `rv` equals to 0.91 , (iii) `pemax` and `weight` equals to 0.63, (iv) `pemax` and `age` equals to 0.61.

```
       pemax weight   age   sex   bmp  fev1    rv   frc
pemax   1.00   0.64  0.61 -0.29  0.23  0.45 -0.32 -0.42
weight  0.64   1.00  0.91 -0.19  0.67  0.45 -0.62 -0.62
age     0.61   0.91  1.00 -0.17  0.38  0.29 -0.55 -0.64
sex    -0.29  -0.19 -0.17  1.00 -0.14 -0.53  0.27  0.18
bmp     0.23   0.67  0.38 -0.14  1.00  0.55 -0.58 -0.43
fev1    0.45   0.45  0.29 -0.53  0.55  1.00 -0.67 -0.67
rv     -0.32  -0.62 -0.55  0.27 -0.58 -0.67  1.00  0.91
```

```
frc    -0.42  -0.62 -0.64  0.18 -0.43 -0.67  0.91  1.00
```

An alternative representation for quantitative variable is:



b) We are first interested in the relationship between `weight` and `pemax`. [6 marks]

- Write down the linear model and its assumptions to study the relationship between the response variable `pemax` and the explanatory variable `weight` based on the normal response distribution. [1 mark]

$$Pemax_i = \beta_0 + \beta_1 weight_i + \epsilon_i \quad i = 1, \ldots, 25$$

where $\epsilon_i \sim N(0, \sigma^2)$ independently, $(i = i = 1, \ldots, 25)$.

- Summarize your model fitting [1 mark].

Table 1 presents results of the fitting model including the 95% confidence interval of β_1: $CI_{0.95}(\beta_1) = [0.597, 1.776]$. This result shows a significant positive linear association between `weight` variable and the response variable `pemax` (pvalue= 0.000646 corresponding to the test of the hypothesis $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$).

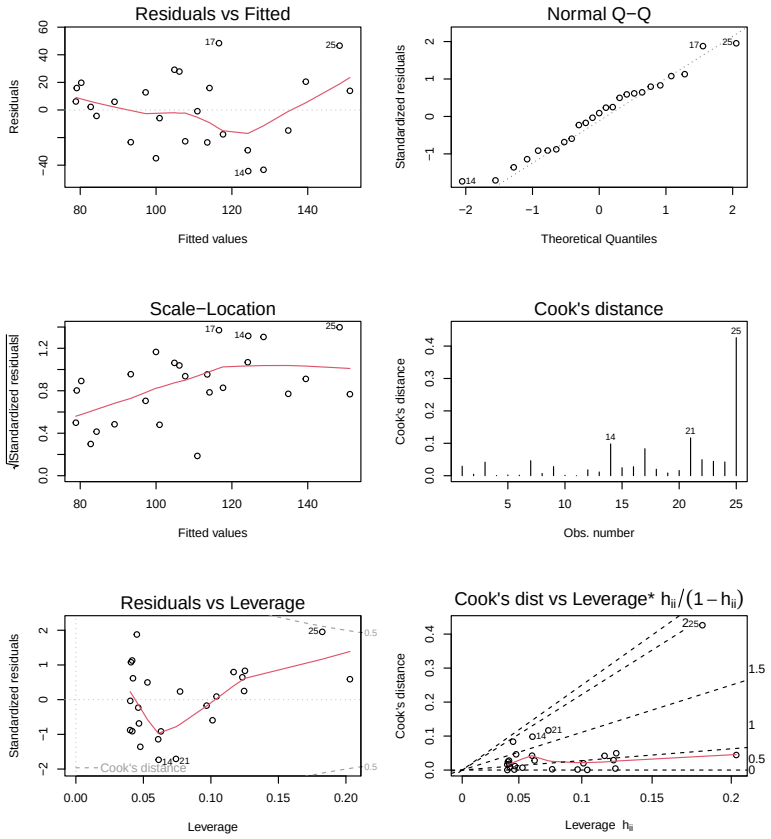
- Interpret the model. [1 mark]
The `weight` variable explains 40.35% (R^2) of the variance of the `pemax` variable. For a 1-kg increase in weight, expected pemax (maximum expiratory pressure) increases by 1.19 ($CI_{0.95}(\beta_1) = [0.60, 1.78]$).

Table 1: Regression Result: Pemax versus weight

	Dependent variable:
	pemax
weight	1.187*** (0.597, 1.776)
Constant	63.546*** (38.651, 88.440)
Observations	25
R^2	0.404
Adjusted R^2	0.378
Note:	*p<0.1; **p<0.05; ***p<0.01

- Present diagnostic checking of your model [3 marks]

[1 mark] for diagnostic plots:

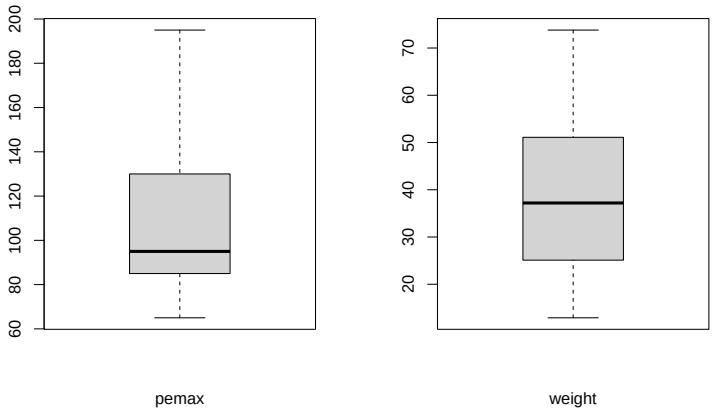


[2 marks] for comments:

- Residuals vs fitted values plot shows no trend, meaning homoscedasticity appears to be satisfied;

- Normal Q-Q plot of residuals is fairly close to a straight line, i.e. normality of residuals is approximately satisfied;
- No Cook's distances are close to the cutoff of 1 $\Rightarrow$ no influential observations;
- Leverage: $\frac{2p}{n} = \frac{2\times 2}{25} = 0.16$. Observations 23 and 25 stands out with high leverage and high Cook's distance:

```
     pemax weight
23     165   73.8
25     195   71.5
```



There are no values which seem extreme or incorrect, so we leave these observations.

c) We are now interested to include in the previous model the `sex` variable. Propose and analyse two different models to introduce the `sex` variable. [8 marks]

- Write down the two models equation [2 marks]

- (i) additive model [1 mark]

$$Pemax_i = \beta_0 + \beta_1 weight_i + \beta_2 sex_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ independently, $(i = 1, \ldots, 25)$. This additive model enables to explore the effect of `weight` on `Pemax` adjusted on `sex` variable. However, this additive model assume the same effect of `weight` on `Pemax` for male and female. An alternative model is to include an interaction term.

- (ii) interaction model [1 mark]

$$Pemax_i = \beta_0 + \beta_1 weight_i + \beta_2 sex_i + \beta_3 weight_i \times sex_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ independently, $(i = 1, \ldots, 25)$. The `sex` variable is codes as: 1 for female and 0 for male. Thus, this model could investigate a different effect of weight on pemax for the male and the female (two different slopes).

- Analyse the two proposed models (fit and interpretation) [3 marks] (1 mark for outputs and 2 marks for interpretation)

Table 2 presents the results of the two models.

Table 2: Regression Result: Pemax versus weight and sex

	Dependent variable:	
	pemax	
	Additive model	Interaction model
	(1)	(2)
weight	1.125*** (0.491, 1.759)	1.357*** (0.635, 2.079)
sex	−11.478 (−33.868, 10.913)	22.091 (−34.667, 78.848)
weight:sex		−0.924 (−2.363, 0.515)
Constant	70.972*** (40.975, 100.969)	61.360*** (28.225, 94.496)
AIC	239.24	239.2
R^2	0.433	0.477
Adjusted R^2	0.381	0.402
F Statistic	8.388*** (df = 2; 22)	6.385*** (df = 3; 21)
Note:		*p<0.1; **p<0.05; ***p<0.01

The additive model shows a significant effect of `weight` on `pemax` taking into account the adjustement on the gender ($\widehat{\beta_1} = 1.125$ and $CI_{0.95}(\beta_1) = [0.526, 1.724]$). In this model, the `sex` variable is not significatively associated to `Pemax`. [0.5 mark]

The interaction model does not show a significant effect of the interaction term ($CI_{0.95}(\beta_3) = [-2.363, 0.515]$).[0.5 mark]

For male (sex=0), the effect of weight on pemax is given by β_1. The estimated effect is $\widehat{\beta_1} = 1.357$. This effect is significant $CI_{0.95}(\beta_1) = [0.635, 2.079]$.[0.5 mark]

For female (sex=1), the effect of weight on pemax is given by $\beta_1 + \beta_3$. The estimated effect is $\widehat{\beta_1} + \widehat{\beta_3} = 0.433$. This effect is not significant $CI_{0.95}(\beta_1 + \beta_3) = [-0.812, 1.679]$. [0.5 mark]

- Choose one of the three statistical models (between the one from question 2 and the two from question 3). Justify your choice. [3 marks]

The additive model does not really improved the model performed in question b). Indeed, the effect of including the 'sex' variable is not statistically significant (pvalue=0.3) and the $R^2_{ajusted}$ has just sligthly been improved (37.8% versus 38.1% for the additive model). Moreover, the AIC criterion for including only `weight` variable is lower than the one for the additive model (238.49 versus 239.24). [1 mark]

The interaction term in the second proposed model is not significant. Thus, it was impossible to detect a difference in slope between female and male. The adjusted R^2 is sligthly better $R^2_{ajusted} = 0.402$ but the AIC has not been improved compared with a model without `sex`. Its value(AIC=239.2) is still higher than the model including only `weight` variable (AIC=238.49). [1 mark]

One can also compare the interaction model versus the model including only `weight` variable by testing:

$$H_0 : \beta_2 = \beta_3 = 0 \text{ versus } \quad H_1 : \ \exists j \in \{2, 3\}, \ \beta_j \neq 0.$$

The fisher partial test which is equivalent to the likelihood ratio test (for linear model) can be obtained using the following code:

```
anova(model1,model3)

Analysis of Variance Table

Model 1: pemax ~ weight
Model 2: pemax ~ weight + sex + weight * sex
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     23 16006
2     21 14033  2    1972.9 1.4762 0.2513
```

This result (pvalue > 0.05) does not show a signifcant result for introducing the `sex` variable in a model including `weight` variable. Thus our choice is to use the simple model including only the `weight` variable. [1 mark]

d) Construct a statistical model for the response variables `pemax` based on the normal response distribution and the `weight`, `bmp`, `fev1`, `rv`,`frc`. Note that `bmp` variable is one of the main explanatory variable of interest. [13 marks]

- Summarize your model fitting and selection process; [4 marks]

Initial screening: [1 mark]

Covariate	p-value
weight	0.00065
bmp	0.27
frc	0.038
fev1	0.02284
rv	0.124

Based on the above, all covariates are considered for inclusion even if the `bmp` variable gets a pvalue greater than 0.2. The `bmp` variable is an important variable to be considered in this study

Selection process: [2 marks]

We first examine a multivariate model including all variables. The result of this model are presented in table 3 (column noted (1)). In this model, the `rv` and `frc` variables do not present significant results (both pvalue greater than 0.05). A second model excluding the variable with the highest pvalue result is performed (pvalue for `frc` equals to 0.94). The results of this model are presented in table 3 (noted (2)). In this second model, the `rv` variable does not present a significant result and a model excluding this variable is then investigated. This final model is presented in table 3 and shows significant results for all variables.

Table 3: Regression Result: multivariate model

	Dependent variable:		
	pemax		
	(1)	(2)	(3)
weight	1.736*** (0.845, 2.626)	1.749*** (0.955, 2.543)	1.536*** (0.779, 2.294)
fev1	1.531** (0.213, 2.848)	1.548** (0.343, 2.753)	1.109** (0.039, 2.178)
rv	0.136 (−0.192, 0.464)	0.126 (−0.048, 0.299)	
frc	−0.025 (−0.679, 0.630)		
bmp	−1.351* (−2.748, 0.046)	−1.377** (−2.557, −0.198)	−1.465** (−2.670, −0.261)
Constant	64.186 (−50.376, 178.749)	63.947 (−47.187, 175.080)	126.334*** (54.130, 198.537)
AIC	235.6	233.6	234
R^2	0.614	0.614	0.570
Adjusted R^2	0.513	0.537	0.509
F Statistic	6.050*** (df = 5; 19)	7.957*** (df = 4; 20)	9.279*** (df = 3; 21)
Note:			*p<0.1; **p<0.05; ***p<0.01

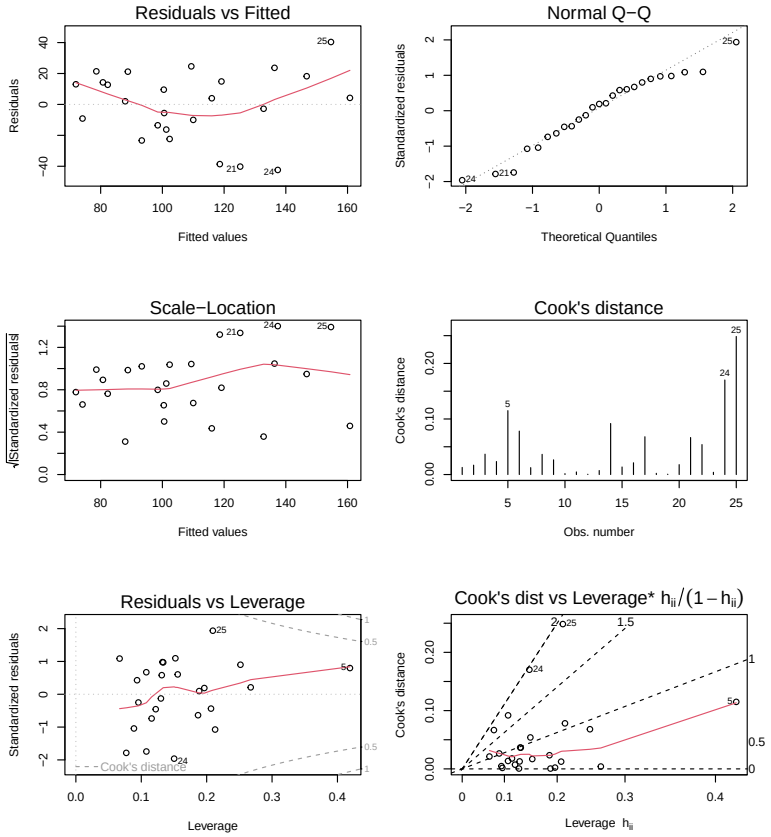
Reasonable final model: [1 mark]

$$pemax_i = \beta_0 + \beta_1 weight_i + \beta_2 fev1_i + \beta_3 bmp_i + \epsilon_i \ ,$$

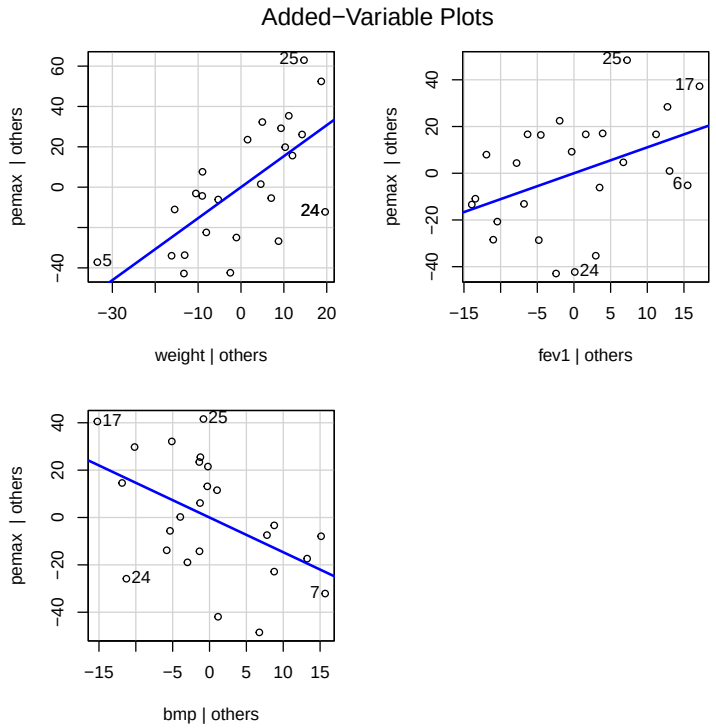
where $\epsilon_i \sim N(0, \sigma^2)$ independently, $i = 1, \ldots, 25$

- Present diagnostic checking of your final model; [5 marks]

For the following plots [1 mark]



- **Normality assumption:** residual plots in output show residuals are normally distributed (normal scores plot linear, OR histogram of residuals appears normal) [0.5 mark];
- **Constant variance assumption:** residuals against fitted values plot has no trend, meaning constant variance assumption is justified. [0.5 mark]
- **Mis-specification of systematic part of the model:**



The partial leverage (added-variable) plots confirm that all y-x relationships are linear (or approximately so).

The partial leverage (added-variable) plots confirm that all y-x relationships are linear (or approximately so).[1mark]

- **Detection of unusual observations:**
  Leverage statistics h_i:

  $2p/n = 2 \times 4/24 = 0.32$. The observation 5 has a notably high leverage ($h_5 = 0.42$). [1 mark]

  Cook's distances D_i:
  All of the D_i well below 1, so we conclude that even though one point has high leverage, none of them are influencing the model fit. [1 mark]

- Write down your final model equation; [1 mark]

  Final model equation:

  $$\widehat{pemax}_i = 126.33 + 1.536 weight_i + 1.109 fev1_{i2} - 1.465 bmp_i$$

- Interpret the model parameters. [3 marks]

  - $\hat{\beta}_1 = 1.536$: for every increase in weight of one kg, expected pemax increases by 1.536.
  - $\hat{\beta}_2 = 1.109$: for every increase in functional residual capacity one unit, expected expected pemax increases by 1.109.
  - $\hat{\beta}_3 = -1.465$: for every increase in body mass of one unit, expected pemax decreases by 1.465.

Question 2 [8 marks]

The Inverse Gaussian distribution is defined as

$$f(x;\mu,\gamma) = \sqrt{\frac{\gamma}{2\pi x^3}}\exp\left(-\frac{\gamma(x-\mu)^2}{2\mu^2 x}\right)$$

where $\mu > 0$ and $\gamma > 0$.

a. Show that the Inverse Gaussian distribution is a member of the exponential family. [5 marks]

By definition a random variable X with pdf $f(x,\theta,\phi)$ is a member of the exponential family distribution if the pdf is of the form

$$f(x;\theta,\phi) = \exp\left\{\frac{x\theta - b(\theta)}{a(\phi)} + c(x,\phi)\right\} \quad (1)$$

for particular functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. We first rewrite the pdf of the Inverse Gaussian distribution:

$$\begin{aligned} f(x;\mu,\gamma) &= \exp\left\{\log\left(\frac{\gamma}{2\pi x^3}\right)^{\frac{1}{2}}\right\}\exp\left\{-\frac{\gamma(x-\mu)^2}{2\mu^2 x}\right\} \\ &= \exp\left\{\frac{1}{2}\log\gamma - \frac{1}{2}\log 2\pi x^3 - \frac{\gamma}{2\mu^2}\frac{(x-\mu)^2}{x}\right\} \\ &= \exp\left\{-\frac{\gamma}{2\mu^2}x + \frac{\gamma}{\mu} + \frac{1}{2}\log\frac{\gamma}{2\pi x^3} - \frac{\gamma}{2x}\right\} \quad (2) \end{aligned}$$

Then by equating the log of the pdf of $f(x,\theta,\phi)$ (from (1)) and $f(x,\mu,\gamma)$ (from (2)), we get

$$\frac{x\theta - b(\theta)}{a(\phi)} + c(x,\phi) = -\frac{\gamma}{2\mu^2}x + \frac{\gamma}{\mu} + \frac{1}{2}\log\frac{\gamma}{2\pi x^3} - \frac{\gamma}{2x}$$

We can identify that:

$$\phi = \gamma,\ \ a(\phi) = \frac{1}{\phi},\ \ \theta = -\frac{1}{2\mu^2},\ \ b(\theta) = -\sqrt{-2\theta},\ \ \ c\left(x, \phi\right) = \frac{1}{2}\log\frac{\phi}{2\pi x^3} - \frac{\phi}{2x}$$

Thus, the Inverse Gaussian distribution is a member of the exponential family

b. Give the natural parameter and the nuissance parameter. [1 mark]
From previous question, the natural parameter is $\theta = -\frac{1}{2\mu^2}$ and the nuissance parameter $\phi = \gamma$

c. Hence, derive the mean and variance of Inverse Gaussian distribution. [2 marks]

The expectation of Y is given by:

$$\begin{aligned} E(Y) &= b'(\theta) \\ &= \frac{1}{\sqrt{-2\theta}} \\ &= \mu \end{aligned}$$

The variance of Y is given by:

$$\begin{aligned} V(Y) &= b''(\theta)\frac{1}{\phi} \\ &= \frac{1}{\sqrt{-2\theta}}\frac{1}{\phi} \\ &= (-2\theta)^{-\frac{3}{2}}\frac{1}{\phi} \\ &= \frac{\mu^3}{\gamma} \end{aligned}$$

Question 3 [12 marks]

Let consider a i.i.d sample of n observations $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i \in \Re^p$ is the vector of p covariates and $y \in \Re$ is the outcome. We model this data using the following linear model:

$$Y_i = \mathbf{x}_i^\top \beta + \epsilon_i \qquad \text{for} \qquad i = 1, \ldots, n,$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top$ is the p-dimensional regressor parameter and the ϵ_i are the noise of the model. We do not consider any intercept in the model.

a. (4 marks). We first consider here a Normal distribution for the noise $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Show, in this case, that the maximium likelihood estimator of the parameter β is also the solution of the squared error loss also know as least square error (LS) defined by

$$LS(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \beta)^2.$$

The likelihood for one observation i is given by the pdf of the $Y_i|X_i = x_i$ which follows a Normal distribution with mean $\mathbf{x}_i^\top \beta$ and variance σ^2

$$\mathcal{L}(\theta; y_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_1 - \mathbf{x}_1^\top \beta\right)^2}{2\sigma^2}\right)$$

where $\theta = (\beta, \sigma^2)$ Then, the likelihood of n observations is given by

$$\mathcal{L}(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^\top \beta\right)^2}{2\sigma^2}\right)$$

and so the log-likelihood:

$$\ln \mathcal{L}(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} \ln\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^\top \beta\right)^2}{2\sigma^2}\right)\right]$$

It simplify to:

$$\ln \mathcal{L}(\theta; y_1, \ldots, y_n) = -\sum_{i=1}^{n} \ln(\sqrt{2\pi\sigma^2}) - \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^\top \beta)^2}{2\sigma^2}$$

Maximizing the log-likelihood over the parameter β is equivalent to minimize the second term of the previous equation:

$$\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^\top \beta)^2}{2\sigma^2}$$

and so minimizing $\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^\top \beta)^2}{2\sigma^2}$ is equivalent to minimizing $\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^\top \beta)^2}{2}$ and so minimizing $LS(\beta)$.

b. (4 marks). We consider now a Laplace distribution for the noise $\epsilon_i \overset{i.i.d.}{\sim} L(0, \sigma)$. We remind you the pdf of the Laplace distribution with mean μ and variance $2\sigma^2$

$$f(z) = \frac{1}{2\sigma} e^{\left(-\frac{|z-\mu|}{\sigma}\right)}.$$

In this case, write down the likelihood of y_i and then the likelihood of the entire sample. Finally define the log-likelihood. Show, in the case of Laplace distribution error noise, maximizing the log-likelihood is equivalent that minimizing the absolute error loss (AL) defined by:

$$AL(\beta) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \mathbf{x}_i^\top \beta|.$$

There is no need to give in this case the obtained estimator as there is no closed form expression in this case even if an unique solution exist.

The likelihood for one observation i is given by the pdf of the $Y_i|X_i = x_i$ which follows a laplace distribution with mean $\mathbf{x}_i^\top \beta$ and variance σ^2

$$\mathcal{L}(\theta; y_1) = \frac{1}{2\sigma} \exp\left(-\frac{\left|y_1 - \mathbf{x}_1^\top \beta\right|}{\sigma}\right)$$

Then, the likelihood of n observations is given by

$$\mathcal{L}(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{2\sigma} \exp\left(-\frac{\left|y_i - \mathbf{x}_i^\top \beta\right|}{\sigma}\right)$$

and so the log-likelihood:

$$\ln \mathcal{L}(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} \ln\left[\frac{1}{2\sigma} \exp\left(-\frac{\left|y_i - \mathbf{x}_i^\top \beta\right|}{\sigma}\right)\right]$$

It simplify to:

$$\ln \mathcal{L}(\theta; y_1, \ldots, y_n) = -\sum_{i=1}^{n} \ln(2\sigma) - \sum_{i=1}^{n} \frac{|y_i - \mathbf{x}_i^\top \beta|}{\sigma}$$

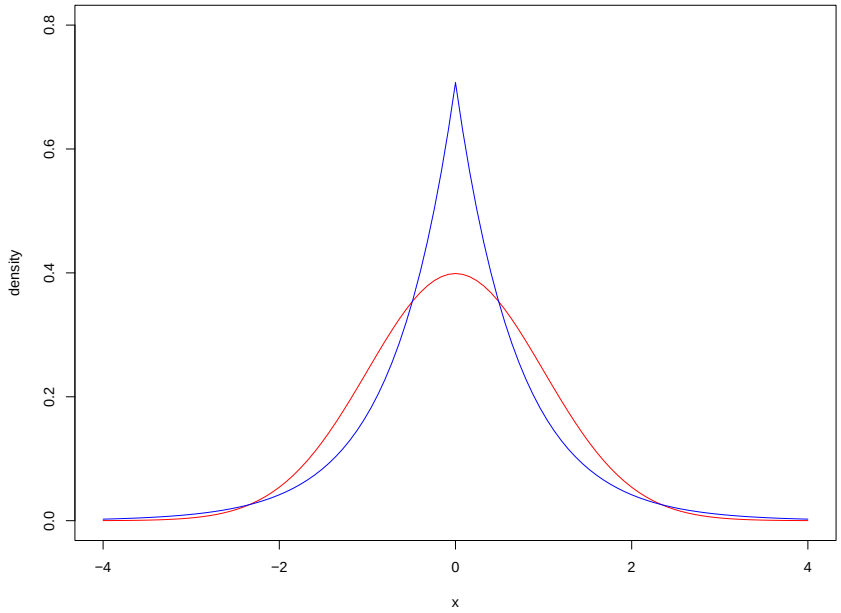
Maximizing the log-likelihood over the parameter β is equivalent to minimize the second term of the previous equation:

$$\sum_{i=1}^{n} \frac{|y_i - \mathbf{x}_i^\top \beta|}{\sigma}$$

and so minimizing $\sum_{i=1}^{n} \frac{|y_i - \mathbf{x}_i^\top \beta|}{\sigma}$ is equivalent to minimizing the absolute error loss (AL) $\sum_{i=1}^{n} \frac{|y_i - \mathbf{x}_i^\top \beta|}{n}$

c. (2 marks). Provide on the same plot the pdf of the Normal distribution with zero men and the pdf of the Laplace distribution. To facilitate the comparison you should match the variance of the two distribution.

The variance of the laplace distribution is given by $2\sigma^2$. For example consider a variance for the normal and the lablace distribution at 4.



d. (2 marks). The linear model using Laplace error noise is known to be more robust to outliers. Give some arguments for this claim using the previous plot.

The Laplace distribution has heavier tails than the Gaussian distribution.With these tails,large noise values,i.e.outliers,are more probable than with the Gaussian noise and hence the loss is typically more robust to extreme values.

Appendix R code for question 1

```
## load library
library(stargazer)
library(ISwR)
library(stats)
library(car)
data(cystfibr)
attach(cystfibr)


## Graphical exploration
plot(cystfibr[,c("pemax","weight","age","sex","bmp","fev1","rv","frc")])


## correlation
signif(cor(cystfibr[,c("pemax","weight","age","sex","bmp",
"fev1","rv","frc")]),digits=2)


## alternative representation
library(GGally)
ggpairs(cystfibr,columns=c("pemax","weight","age","bmp","fev1","rv","frc"))


## first model with weight
model1 <- lm(pemax~weight)
summary(model1)

stargazer(model1,omit.stat=c("LL","ser","f"), ci=TRUE, ci.level=0.95,
single.row=TRUE,title="Regression Result: Pemax versus weight"
,type='latex',header=FALSE)


## diagnostic plot
par(mfrow=c(3,2)) ### 3 rows and 2 columns for plots
plot(model1,which=1:6)


## extreme observation
cystfibr[25,c("pemax","weight")]
par(mfrow=c(1,2))
```

```
boxplot(pemax,xlab="pemax")
boxplot(weight,xlab="weight")


## model with sex
model2 <- lm(pemax~weight+sex)
model3 <- lm(pemax~weight+sex+weight*sex)

## to change the reference class
sex_f <- factor(sex,labels=c("male","female"))
model4 <- lm(pemax~weight+relevel(sex_f, ref = "female")+
weight*relevel(sex_f, ref = "female"))
model5 <- lm(pemax~weight+relevel(sex_f, ref = "male")+
weight*relevel(sex_f, ref = "male"))
summary(model4)
summary(model5)
confint(model4)
confint(model5)


stargazer(model2,model3,
title="Regression Result: Pemax versus weight and sex"
,omit.stat=c("LL","ser","n"), ci=TRUE, ci.level=0.95,
column.labels=c("Additive model","Interaction model"),
single.row=TRUE,type='latex',header=FALSE,
ci.custom = list(confint(model2),confint(model3)),
add.lines=list(c("AIC", round(AIC(model2),2),
round(AIC(model3),2))))


## Initial screening
model.w <- lm(pemax~weight)
model.bmp <- lm(pemax~bmp)
model.fev1 <- lm(pemax~fev1)
model.rv <- lm(pemax~rv)
model.age <- lm(pemax~age)
model.frc <- lm(pemax~frc)


## Selection process
model.mult1 <- lm(pemax~weight+fev1+rv+frc+bmp)
model.mult2 <- lm(pemax~weight+fev1+rv+bmp)
```

```
model.mult3 <- lm(pemax~weight+fev1+bmp)

stargazer(model.mult1,model.mult2,model.mult3,title="Regression Result:
multivariate model",omit.stat=c("LL","ser","n"),
ci=TRUE, ci.level=0.95,font.size = "footnotesize",
single.row=TRUE,type='latex',header=FALSE,
ci.custom = list(confint(model.mult1),confint(model.mult2),
confint(model.mult3)),
add.lines=list(c("AIC", round(AIC(model.mult1),2),
round(AIC(model.mult2),2),round(AIC(model.mult3)))))


## diagnostic plots
par(mfrow=c(3,2)) ### 3 rows and 2 columns for plots
plot(model.mult3,which=1:6)


## Partial leverage
avPlots(model.mult3)
```