# QUESTION 1

## (a)

To show the covariance covariance of b0 and b1; ,

[since b0= ]

[as E(b1)=β1]   
 [Since, Var(b1)=E(b12)-{E(b1)}2] [Since, Var(b1)=]

# (b)

The true model of the simulation is as mentioned on the simulation code; y ***<-1+2\*x + rnorm(n,0,2)*** So, the model is

**Y=1+2x+**

Similarly, the distribution of the random term is normal distribution with mean equaling to 0 and standard deviation equaling to 2. i.e.

Based on the true model, the theoretical values of covariance and correlation are given by

Covariance **=**  which from the code and R output is

***Covariance<- -(mean(x)/(sum((x-mean(x))^2) ))\*var(y)*** **-2.3329.**

The correlation(theoretical) is given by So, the correlation R code and the value are

***Correlation<- -(xbar)/(sqrt(xbar^2 +(Sxx/length(x))));* -0.8864.**

# (c)

|  |  |
| --- | --- |
|  | The plot on the left is the output from R which clearly shows that b1 and b0 are linearly and negatively corelated. Because, with the increasing values of b0 the values of b1 seem to be decreasing at a constant rate. The scattered points actually make a thick straight line which represents their linearity. So, to conclude these estimates do appear to be correlated. |

# (d)

To calculate the empirical values of b1 and b0 , I will use the values of the matrix estimates. So, using R codes as follows

***covariance1<-cov(estimates[,1],estimates[,2])***  and the value is equal to **-0.2767.**

Similarly for correlation, ***correlation1<-cor(estimates[,1],estimates[,2])*** and the value is **-0.8895.**

Comparing the theoretical values with the empirical values, I can see that the values do not match up.

# Question 2

# (a)

For the scatter plot and the identification of the high leverage points following codes are used.

***plot(Metab,Life,xlab = "Metabolic Rate",ylab = "Lifespan",main = "Scatter plot for quick Glance",pch=20,col="blue"),*** this corresponds to the first figure sown below.

For Identification, ***Index<-identify(Metab,Life,labels=CommonName)***  This corresponds to the second figure below.

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated | Chart, scatter chart  Description automatically generated |

From the R codes, we can find that the following output

|  |  |  |
| --- | --- | --- |
|  | **Commonname** | **Species** |
| **62** | **Asian elephant** | **Elephas maximum** |
| **84** | **Bottle-nosed whale** | **Hyperoodon ampullatus** |

# (b)

Most observations of **Life** and **Metab** fall near the 0 point on the plot i.e., left bottom. So, to have a closer glance at most data we must adjust the coordinate limits of the plot. **Metab** has 2 clear extreme values, so I will be sorting it and then I will use the third maximum value as my x-axis upper limit. Similarly, **Life** has one clear extreme value, and the values of **Life** are well readable. Hence, I will put my upper y-axis limit as 45. The code used and the output are as follows

***plot(Metab,Life,xlab = "Metabolic Rate",ylab = "Lifespan",main = "Scatter plot for quick Glance:A closer look",pch=20,col="blue",xlim = c(min(Metab), ((sort(Metab,decreasing = TRUE))[3)])), ylim = c(min(Life),45))***

Chart, scatter chart

Description automatically generated

For majority of observations, the relationship between **Life** and **Metab** does not look linear. It, in fact looks much more curvilinear.

# (C)

Now, to transform **Metab** using natural logarithm, the code is,

**TMetab<-log(Metab)**

After applying the natural logarithm, when we calculate the correlation between them, we find it to be a high figure at **0.7895**. when we plot **Life** against **TMetab**, we get a much nicer linear relationship between them. The scatter plot appears linear with positive correlation between **Life** and **TMetab.** However, there seems to be the presence of outliers.Similarly, the box plots show, TMetab does not have any outlier while Life has few outliers, and they lie towards the right of the median. This could potentially be a right skewed distribution. In addition, the variance of **Life** seems to increase with the increasing transformed **Metab** value. This could potentially be a violation of constant variance and the presence of influential points.

Chart, scatter chart

Description automatically generated

Then, in order to fit the model, We use the code **mymodel<-lm(Life~TMetab)**

This creates an object named **mymodel**, which is a linear regression model with **Life** as response variable and **TMetab** as the predictor variable. The fitted results are obtained by the code **summary(mymodel)** and are as follows.

**Call:**

**lm(formula = Life ~ TMetab)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-12.100 -4.599 -1.363 2.232 49.829**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) -12.1068 2.1287 -5.687 1.48e-07 \*\*\***

**TMetab 4.1742 0.3365 12.404 < 2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 7.915 on 93 degrees of freedom**

**Multiple R-squared: 0.6233, Adjusted R-squared: 0.6192**

**F-statistic: 153.9 on 1 and 93 DF, p-value: < 2.2e-16**

The four model diagnostic plots from R output are

Graphical user interface, chart, histogram

Description automatically generated

Now, when we look at the normal plots, the Lowess curve appears to be slightly curved. Although it is mostly closer to 0, it however spans from almost 10 to around negative 3 which is large. This is the violation of linearity. The variance seems to change with the fitted values. As we move further from the zero point, the variance of residuals seem to increase which violates the constant variance assumption. When we look at the normal Q-Q plot, the right tail looks skewed and as most observations do not fall on the straight line with lower tails also slightly non normal. We can conclude that the normality assumption is well violated here.

The 3rd plot shows 2 observations with relatively extremely high Cook’s distance; however, their Cook’s distances are still below 1. Similarly, the 4th plot shows few observations having higher leverage than **4/95=0.04211.** We must consider the fact that, observations with high leverage are potentially influential however to be influential an observation needs to be able to influence the model single handedly.

# (d)

We transform the response variable **Life** with natural logarithm as **NLife** and with square root as **SLife**. When the correlation coefficients of newly transformed response variable were calculated, the following result was obtained.

**Correlation NLife vs TMetab Correlation SLife vs TMetab**

**0.8825978 0.8776162**

Similarly, when the newly transformed response variables were scatter plotted against the transformed predictor, we found strong, and positively correlated linear curves. The square root transformation still seems to have outlier problem while the natural log transformation seems to fit the observations much better. I have noted that the natural log transformation could possibly have a slightly skewed left tail. However, the observations seem to be scattered with a constant variance and without any clear outlying, influential observations. While we could still proceed with more detailed exploratory data analysis and the model diagnostics, but by just looking at the scatter plot and through the comparison of their correlation coefficients with TMetab, I here conclude that model with natural logarithmic transformation of both predictor and response variable are much better suited for our observations. The mathematical form of my regression model is

**Fitted (** This equation can also be simply written as;

**Fitted (**

Chart, scatter chart

Description automatically generated

# (e)

Following the above procedures, the new model has been fitted and a part of the result is as follows.

summary(model2)

**Call:**

**lm(formula = NLife ~ TMetab)**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) -0.21231 0.13352 -1.59 0.115**

**TMetab 0.38215 0.02111 18.11 <2e-16 \*\*\***

After we have fitted the model, the mathematical form of the model is

**Fitted (**

Now, let us conduct model diagnostics,

Diagram

Description automatically generated

Looking at the diagnostic plots, we can see that the linearity assumption is fairly held. The Lowess curve is closer to 0 throughout the stretch. Few observations could potentially be outliers as depicted by the first plot. The residuals seem to have a constant variance(homoscedasticity) although, it does not appear to be perfectly constant. The normal QQ plot shows lighter tails and possibly skewed to the right. Similarly, most observations have cook’s distances below 0.2. However, few observations have relatively high cook’s distances. Observation 85,88 and 95 could potentially be outliers but they might not be influential as the leverage plots shows. Again, few observations have relatively high leverage. Observations 62 and 84 seem to be such points, however they do not possess the influential ability as their cook’s distance are not even comparable to observation 95.

# (f)

The estimated slope of our fitted model is **0.3822.** This means our response variable log (Life) increases by 0.3822 for a unit increase in our predictor variable log (Metab). i.e., for unit increase in natural logarithm of metabolic rate for different species, we can see an increase of 0.3822 in the natural logarithmic value of their life span.

For, the 95% confidence interval of the slope parameter, the code used is

**> Con\_Interval<-confint(model2,parm = 2, level = 0.95)**

**> Con\_Interval**

**2.5 % 97.5 %**

**TMetab 0.3402304 0.4240621**

So, the 95% confidence interval for the slope parameter is **(0.3402,0.4241)**.

# (g)

Using the ANOVA approach to test the significance of the model used in part (e), the hypothesis for the test is,

H0: β1=0 vs H1: β1≠0

**> anova(model2)**

**Analysis of Variance Table**

**Response: NLife**

**Df Sum Sq Mean Sq F value Pr(>F)**

**TMetab 1 80.797 80.797 327.77 < 2.2e-16 \*\*\***

**Residuals 93 22.925 0.247**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

The test statistics here is the F value which is 327.77.

Since, our significance level i.e. alpha(α) value is 5%, and the degrees of freedom for numerator is 1 while for the denominator is 93,so our rejection region is given by

**> crit\_val<-qf(0.05,df1=1,df2=93, lower.tail = F)**

**> crit\_val**

**[1] 3.943409**

So, our rejection region is F\* >3.9434. Similarly, the p value of the test is given by

**> pf(327.77,1,93,lower.tail = F)**

**[1] 3.060952e-32**

Since our P-value is way smaller than our significance level, we reject the null hypothesis and conclude that β1≠0. This is also proved by the critical value i.e. since our F-statistics is way bigger than our critical value, our test statistics falls on the rejection region advising us to reject the null hypothesis.

# (h)

The prediction interval for the mammal with 8000kj/ day of metabolic rate is given by,

**>Pre<-predict(model2,data.frame(TMetab=log(a)),interval ="prediction”,conf.level=0.9 )**

**> exp(Pre)**

**fit lwr upr**

**1 25.08107 9.228631 68.16397** So, the prediction interval has fitted value 25.0811 and the lower limit is 9.2286 while the upper limit is 68.1640. This means, based on our fitted model, any animal whose metabolic rate is 8000Kj per day is expected to live around 25 years while the expected life span of such a mammal could vary from 9 years to 68 years.

# (i)

As Kleiber’s law states that on average the metabolic rate of animal is proportional to its mass raised to the power of ¾. A simple linear regression model for this is

Metab=β0 + β1 \*Mass3/4+ Error term,

it can also be written as Metab= β0 + β1 \* MASS+ Error term, where MASS=Mass3/4.

Then, to test the adequacy of this theory I believe the following hypothesis should be employed.

Hypotheses **H0: β1 =0 vs H1 : β1≠0**

This hypothesis can be employed because, if the Kleiber’s law holds true for animals, then there should be a linear relationship between the mass and the metabolic rate of the animals. However, non-existence of such proportional relationship should indicate that, the law does not hold true for animals but instead they might have a different kind of relationship.

Let us first calculate the correlation between the metabolic rate and the mass raised to the power of ¾. We get the correlation to be **0.9945387**, which is an extremely strong and positive linear relationship.

So, now let us fit the model using the data set provided. Here, the predictor variable is Mass while the Metab (metabolic rate) is the response variable. Then the fitted model is

**Call:**

**lm(formula = Metab ~ mass)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-11712.7 -117.4 368.5 474.3 6598.5**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) -481.346 213.467 -2.255 0.0265 \***

**mass 395.016 4.299 91.895 <2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 1992 on 93 degrees of freedom**

**Multiple R-squared: 0.9891, Adjusted R-squared: 0.989**

**F-statistic: 8445 on 1 and 93 DF, p-value: < 2.2e-16**

Now let us test the aforementioned hypothesis.

Here the estimate of β1 is 395.016 while the standard error is 4.299. Under the significance level of 5% we have the following,

Hypotheses **H0 : β1 =0 vs H1 : β1≠0**

Test Statistics 395.016/4.299 = 91.88555478 = 91.8856 which has a t distribution with n-2=93 degrees of freedom.

So, the critical value is calculated as

**Lower Critical Value Upper Critical Value**

**-1.985802 1.985802**

Similarly, the P-value of our test statistics is **4.467589e-93** which is far smaller than 0.05.

So, from our empirical analysis, we see that the P value is extremely smaller than our significance level for the fitted model. In addition, our test statistics lies far away from the critical values on the rejection region. Hence, we reject our null hypothesis and conclude that the value of **β1**, which is the proportional constant in Kleiber’s law, is not equal to zero. Consequently, this tells us that, yes in fact there appears to be some linear relationship between the metabolic rate and the mass raised to the power of ¾. However, it’s necessary to conduct our model diagnostic which will tell us if in fact our model is appropriate or not.

Chart

Description automatically generated

If we look at the plots for exploratory data analysis, although there appears to be a strong linear relationship between those two variables, the linearity, constant variance seems to be violated. Even the distribution seems to be flat tailed on both tails. Some of the points appear to have high cook’s distance and some appear to be outliers. This tells us that although, we have high correlation between the variables, the model we have fit here is in fact not appropriate. So, to concludes, metabolic rate and mass seem to have some sort of relationship however that relationship cannot be explained by the model that we have fit here. Hence, to explain their relationship a better model is required. This also proves that the transformation of our predictor variable Mass to its ¾power causes the model to become severely inappropriate. So, a linear model cannot be used to test the Kielber’s law.

# Appendix