Question 1

The following model can be used to study whether campaign expenditures affect election outcomes:

voteA = β0 + β1 ln[expendA] + β2 ln[expendB] + β3prystrA + u

voteA = 45.1 + 6.08 ln[expendA] - 6.62 ln[expendB] + 0.152 prystrA + u n=173

1. What is the interpretation of β1?

Soln: When the campaign expenditure of candidate A increases/decreases by 1% then there is β1% increase/decrease in receiving vote, i.e., when there is 1% increase in expendA then there is 6.08% increase in voteA received.

2. In terms of the parameters, state the null hypothesis that a 1% increase in A's expenditures is offset by a 1% increase in B's expenditures.

Soln: Null hypothesis: H0: β1 = -β2 ➔ if there is 1% increase in expenditure A and 1% increase in expenditure B then percentage change in votes received is not changed

Alternate hypothesis: H1: β1≠ -β2 ➔ if there is 1% increase in expenditure A and 1% increase in expenditure B then percentage change in votes received is changed

3.Estimate the given model using the data in the vote1 table and report the results in usual form. Do A's expenditures affect the outcome? What about B's expenditures? Can you use these results to test the hypothesis in part 2?

Soln:

R-Code:

```
modlea <- lm(voteA~log(expendA)+log(expendB)+prtystrA,data=vote)
summary(modlea)
tidyw(modlea)
```

voteA = 45.1 + 6.08 ln[expendA] - 6.62 ln[expendB] + 0.152 prystrA + u   n=173, R-squared=0.7925

The coefficient of ln[expendA] and ln[expendB] are significant at 0.1% level and the t-statistics for expendA and expendB are 15.915 and -17.461 which is high. Since the coefficients of ln(expendA) and ln(expendB) are opposite sign we can test the hypothesis in part 2

4.Estimate a model that directly gives the t-statistic for testing the hypothesis in part 2

Soln:

R-code:

```
summary(lm(voteA~log(expendA)+I(log(expendB)-log(expendA))+prtystrA,data=vote))
```

ϴ1= β1 + β2 ➔ β1 = ϴ1-β2

voteA = β0 + (ϴ1-β2) ln[expendA] + β2 ln[expendB] + β3prystrA + u

voteA = β0 + ϴ1 ln[expendA] + β2 (ln[expendB]- ln[expendA]) + β3prystrA + u

voteA = 45.1 -0.534 ln[expendA] -6.62 β2 (ln[expendB]- ln[expendA]) + 0.152 prystrA + u

5. What do you conclude? (Use a two-sided alternative.)

Soln: Since ln(expendA) is not significant any all the levels we cannot reject the null hypothesis in part 2

Question 2

1. Using the model: ln[salary] = β0 + β1LSAT + β2GP A + beta3 ln[libvol] + β4 ln[cost] + β5rank + u, state and test the null hypothesis that the rank of law schools has no ceteris paribus effect on median starting salary

Soln:

R-code:

```
model <- lm(log(salary)~ LSAT + GPA +log(libvol) +log(cost) +rank ,data=lawsch)
tidy(model)
```

H0: $\beta 5=0$

log(salary)=8.34+ 0.00470 LSAT + 0.248 GPA +0.0950 log(libvol) +0.0376 log(cost) + -0.00332rank, data=lawsch)

Since the coefficient of rank are significant at 0.1% level (p=1.12e-16), we can reject the null hypothesis

2.Are features of the incoming class of students—namely, LSAT and GPA—individually or jointly significant for explaining salary? (Be sure to account for missing data on LSAT and GPA.)

Soln:

R Code:

```
lawsch2<-na.omit(lawsch) # removing the na values
modela <- lm(log(salary)~ LSAT+GPA+log(libvol)+log(cost)+rank,data=lawsch2)
tidy(modela)
summary(modela)
modelb <- lm(log(salary)~ log(libvol)+log(cost)+rank,data=lawsch2)
summary(modelb)
anova(modela,modelb)
```

I have omitted the na values in dataset

ln(salary) = 7.85+ 0.00683 LSAT+0.233 GPA+0.106 log(libvol)+ 0.0494 log(cost)- 0.00291 rank, data=lawsch2)

In the above model we see the LSAT has t-statistics = 1.23 and P value = 0.223 which is not significant and GPA with t-statistics =2.025 and p-value=0.0460 which is significant at 10%. The joint significance has the F-statistics = 7.6258 and the p-value = 0.0009052 which is significant at 1%.

Hence LSAT is not significant, GPA is significant and the jointly the model is significant.

3.Test whether the size of the entering class (clsize) and the size of the faculty (faculty) need to be added to this equation jointly

Soln:

R-code:

```
modelc <- lm(log(salary)~ LSAT + GPA +log(libvol) +log(cost) +rank+ clsize+faculty,data=lawsch2)
summary(modelc)
modeld <- lm(log(salary)~ LSAT + GPA +log(libvol) +log(cost) +rank,data=lawsch2)
summary(modeld)
anova(modelc,modeld)
```

By adding the class size and faculty size we can see slight increase in R-squared value. However, the p-value for the equation jointly is 0.1381 and F-statistics is 2.0282 which show the model is insignificant.

4.What factors might influence the rank of the law school that are not included in the salary regression?

Soln:

R-code:

```
lawsch[,.(salary,rank,east,west,north,south)]
modele <- lm(log(salary)~ rank+ north+south+east+west,data=lawsch)
tidy(modele)
```

Factors like the location of school i.e., east, west, north and south aren't significant, LSAT and GPA scores are good controlling variables for rank, if we have any other factors like gender which might control the rank, however these details are not given in dataset

Question 3

$\ln[price] = \beta0 + \beta1sqrf\ t + \beta2bdrms + u$

$\ln[price] = 4.77 + 0.000379sqrft + 0.0289bdrms+u$, n=88 , R-squared=0.5883

1.You are interested in estimating and obtaining a confidence interval for the percentage change in price when a 150-square-foot bedroom is added to a house. In decimal form, this is $\theta1 = 150\beta1 + \beta2$. Use the data in the hprice1 table to estimate $\theta1$

Soln:

R-code:

```
model1 <- lm(log(price)~I(sqrft-(150*bdrms))+bdrms,data=hprice)
tidy(model1,conf.int=TRUE)
```

$\beta0 + \beta1sqrf\ t + \beta2bdrms + u$, $\Theta1=150\ \beta1+ \beta2 \Rightarrow \beta2 = \Theta1\ -150\ \beta1$

$\ln[price]= \beta0 + \beta1sqrft + \beta2bdrms + u$

$\ln[price]=\beta0 + \beta1sqrft + (\Theta1\ -150\ \beta1)\ bdrms + u$

$\ln[price]= \beta0 + \beta1sqrft + (\Theta1\ bdrms\ -150\ \beta1\ bdrms) + u$

$\ln[price]= \beta0 + \beta1(sqrft-150\ bdrms) + \Theta1\ bdrms + u$

$\ln[price] = 4.77+ 0.000379(sqrft-150\ bdrms) + 0.0858\ bdrms +u$, n=88, R-squared=0.5883

$\Theta1=0.0858$ with t-statistics= 3.21 and p-value=0.00190, so the coefficient of adding 150 sqrft bedroom is statistically significant at 1% level. Confidence interval being low 0.0326 and high 0.139

2. Write $\beta2$ in terms of $\theta1$ and $\beta1$ and plug this into the ln[price] equation.

Soln:

R-code :

```
model2 <- lm(log(price)~I(sqrft-bdrms)+bdrms,data=hprice)
tidy(model2)
summary(model2)
```

$\Theta 1 = \beta 1 + \beta 2 \Rightarrow \beta 2 = \Theta 1 - \beta 1$

$\ln[price] = \beta 0 + \beta 1 sqrft + \beta 2 bdrms + u$

$\ln[price] = \beta 0 + \beta 1 sqrft + (\Theta 1 - \beta 1)\, bdrms + u$

$\ln[price] = \beta 0 + \beta 1 sqrft + (\Theta 1\, bdrms - \beta 1\, bdrms) + u$

$\ln[price] = \beta 0 + \beta 1 (sqrft - bdrms) + \Theta 1\, bdrms + u$

$\ln[price] = 4.77 + 0.000379(sqrft - bdrms) + 0.0293\, bdrms + u$, n= 88 and R-squared= 0.5883

3. Use part 2 to obtain a standard error for ˆθ1 and use this standard error to construct a 95% confidence interval.

Soln: Standard error for ˆθ1 = 0.0296 and 95% confidence interval is between -0.0296 and 0.0882

Question 4

$\ln[wage] = \beta 0 + \beta 1 educ + \beta 2 exper + \beta 3 tenure + u$

$\ln[wage] = 5.50 + 0.0749\, educ + 0.0153\, exper + 0.0134\, tenure + u$

1.Consider the standard wage equation $\ln[wage] = \beta 0 + \beta 1 educ + \beta 2 exper + \beta 3 tenure + u$.

State the null hypothesis that another year of general workforce experience has the same effect on ln[wage] as another year of tenure with the current employer

Soln: Null hypothesis $\Rightarrow$ H0: $\beta 2 = \beta 3$

2. Test the null hypothesis in part 1 against a two-sided alternative, at the 5% significance level

Soln:

R-code :

```
modela <- lm(log(wage)~educ+exper+I(exper+tenure),data=wage2)
tidy(modela,conf.int = T)
```

$\beta 2 = \beta 3 \Rightarrow \beta 2 - \beta 3 = 0 \Rightarrow \beta 2 - \beta 3 = \Theta 1 \Rightarrow \beta 2 = \Theta 1 + \beta 3$

$\ln[wage] = \beta 0 + \beta 1 educ + \beta 2 exper + \beta 3 tenure + u$

$\ln[wage] = \beta 0 + \beta 1 educ + (\Theta 1 + \beta 3)\, exper + \beta 3 tenure + u$

$\ln[wage] = \beta 0 + \beta 1 educ + \Theta 1\, exper + \beta 3\, exper + \beta 3 tenure + u$

$\ln[wage] = \beta 0 + \beta 1 educ + \Theta 1\, exper + \beta 3\, (exper + tenure) + u$

$\ln[wage] = 5.50 + 0.0749\, educ + 0.00195\, exper + 0.0134\, (exper + tenure) + u$

t critical value for two-sided alternative is 0.975

The coefficient of expenditure has t-statistics =0.412 and p-value= 0.681 is not significant at 5% level.

Hence, we cannot reject the null hypothesis in part 1

Question 5

The table 401Ksubs contains information on net financial wealth (nettfa) age of the survey respondent (age), annual family income (inc), family size (fsize), and participation in certain pension plans for people in the United States. The wealth and income variables are both recorded in thousands of dollars. For this question, use only the data for single-person households (so fsize = 1).

1.How many single-person households are there in the data set?

Soln:

R-code:

```
table(subs[,.(fsize==1)])
```

There are 2017 single person households in dataset

2.Use OLS to estimate the model nettfa = β0 + β1inc + β2age + u, and report the results using the usual format. be sure to use only the single-person households in the sample. Interpret the slope coefficients. Are there any surprises in the slope estimates?

Soln:

R-code:

```
subs1 <- subs %>% filter(fsize ==1)
model <- lm(nettfa~inc+age,data=subs1)
summary(model)
augment(model)
subs1$newy <- log(augment(model)$.resid^2)
model2 <- lm(newy~inc+age,data=subs1)
summary(model2)
subs1$h <- exp(augment(model2)$.fitted)
model3 <-lm(nettfa~inc+age,weights=1/h,data=subs1)
tidyw(model3)
```

nettfa = β0 + β1inc + β2age + u

nettfa = -43 + 0.799 inc +0.843 age +u, n=2017 and R-squared =0.12

When there is an increase of one dollar in the income then the net financial wealth will increase by 80 more cents and when there is an increase in age by 1 year the net financial wealth will increase by 843$. There is no surprise.

3. Does the intercept from the regression in part 2 have an interesting meaning? Explain

Soln:

R-code:

```
min(subs1$age)
min(subs1$inc)
```

The intercepts do not have interesting meaning here it gets the net financial wealth when the inc=0 and age=0, considering the dataset we cannot find anyone having these values.

4. Find the p-value for the test H0 : β2 = 1 against H1 : β2 < 1. Do you reject H0 at the 1% significance level?

Soln:

R-code:

```
tstat <-(0.84266-1)/0.09202
tstat
tstat <-(0.84266-0.5)/0.09202
```

tstat
pnorm(-abs(tstat))*2

H0: $\beta2 = 1$ ➔ the model in part 2. t-stats= -1.709846 and p-value = 0.0873 which is not significant at 1% level hence we keep the null hypothesis and reject the alternative hypothesis H1.

5.If you do a simple regression of nettfa on inc, is the estimated coefficient on inc much different from the estimate in part 2? Why or why not?
Soln:
R-code:
```
model1 <- lm(nettfa~inc,data=subs1)
summary(model1)
cor(subs1$inc,subs1$age)
```

nettfa = -10.571 + 0.821 inc + u , n=2017, R-squared=0.08267
There is a slight increase in the estimated coefficient value on inc. the difference is 0.0331. the correlation between age and income is 0.039 hence the difference.

Question 6
Use the data in the kielmc table, only for the year 1981, to answer the following questions. The data are for houses that sold during 1981 in North Andover, Massachusetts; 1981 was the year construction began on a local garbage incinerator.
1.To study the effects of the incinerator location on housing price, consider the simple regression model ln[price] = β0 + β1 ln[dist] + u, where price is housing price in dollars and dist is distance from the house to the incinerator measured in feet. Interpreting this equation causally, what sign do you expect for β1 if the presence of the incinerator depresses housing prices? Estimate this equation and interpret the results.
Soln:
R-code:
```
kielmc <- kie %>% filter(year==1981)
ln[price] = β0 + β1 ln[dist] + u
model <- lm(log(price)~ldist,data=kielmc)
tidy(model)
summary(model)
```

$\ln[price] = \beta0 + \beta1 \ln[dist] + u$
$Ln[price] = 8.05 + 0.365 \ln[dist] + u$, n=142 and R-squared=0.1803
If there is an increase in 1% of dist then there is 36.5% increase in price of the house. Which means the effect of dist on price i.e>, $\beta1 \geq 0$

2.To the simple regression model in part 1, add the variables ln[intst], ln[area], ln[land], rooms, bath, and age, where intst is distance from the home to the interstate, area is square footage of the house, land is the lot size in square feet, rooms is total number of rooms, baths is number of bath rooms, and

age is age of the house in years. Now, what do you conclude about the effects of the incinerator? Explain why 1 and 2 give conflicting results.
Soln:
R-code:
```
model2 <- lm(log(price)~log(dist)+log(intst)+log(area)+ log(land)+ rooms+ baths+ age,data=kielmc)
summary(model2)
```

ln[price]= β0 + β1ln[dist]+ β2 ln[intst]+ β3 ln[area]+ β4 ln[land]+ β5 rooms+ β6 baths+ β7 age + u
ln[price]= 7.59 + 0.055 ln[dist]-0.039 ln[intst]+ 0.32 ln[area]+ 0.076 ln[land]+ 0.043 rooms+ 0.166 baths - 0.0035 age + u, n=142, R-squared = 0.7475
we can see reduction in the value of dist from part 1 model compared to part 2 i.e., from 0.365 to 0.055 this is because of addition of other controlling variables like the number of bedrooms the area etc.

3. Add ln[intst] 2 to the model from part 2. Now what happens? What do you conclude about the importance of functional form?
Soln:
R-code:
```
model3 <- lm(log(price)~log(dist)+log(intst)+log(area)+ log(land)+ rooms+ baths+ age +
I(log(intst)^2),data=kielmc)
summary(model3)
```

ln[price]= β0 + β1ln[dist]+ β2 ln[intst]+ β3 ln[area]+ β4 ln[land]+ β5 rooms+ β6 baths+ β7 age + β8 ln[inststs]^2 + u
ln[price]= -3.318025+0.185256 ln[dist]+ 2.072959 ln[intst]+ 0.359352 ln[area]+ 0.091386ln[land]+ 0.038106 rooms+ 0.149533 baths -0.002927 age + -0.119329 ln[instst]^2 + u
After adding the ln[instst]^2 the distance from house to incinerator has become slightly significant, the distance to interstate inst has also become significant along with rooms and bathrooms. We can also see this increase in significance with dist and intst as there is a strong positive correlation between these to factors.

4. Is the square of ln[dist] significant when you add it to the model from part 3?
Soln:
R-code:
```
model4 <- lm(log(price)~log(dist)+log(intst)+log(area)+ log(land)+ rooms+ baths+ age +
I(log(intst)^2)+I(ldist^2),data=kielmc)
summary(model4)
```

Adding ln[dist]^2 doesn't make the model better as the coefficient is not significant even at 10% level.

Question 7
Use the data in the wage1 table for this exercise.
1.Use OLS to estimate the equation ln[wage] = β0 + β1educ + β2exper + β3exper2 + u and report the results using the usual format
Soln:

R-code:

```
model <- lm(log(wage)~educ+exper+I(exper^2),data=wage1)
summary(model)
```

ln[wage] = β0 + β1educ + β2exper + β3exper^2 + u

ln[wage] = 0.126 + 0.0906 educ + 0.0410 exper – 0.000712 exper^2 + u, n=526, R-squared=0.3003

2. Is exper2 statistically significant at the 1% level?

Soln: The coefficients of exper^2 with t-stats= -6.141 and p-value = 0.0000000016276762 is significant at 1% level

3. Using the approximation ∂ ln[wage]/ ∂exper ≈ β^ 2 + 2β^ 3exper, find the approximate return to the fifth year of experience. What is the approximate return to the twentieth year of experience?

Soln: for the return to 5^{th} year of experience ∂ ln[wage] ≈ 0.0410+2*(-0.000712) *4 =0.0035304

For the return to 20th year of experience ∂ ln[wage] ≈0.0410+2*(-0.000712) *19 = 0.013944

4. At what value of exper does additional experience actually lower predicted ln[wage]? How many people have more experience in this sample?

Soln:

R-code:

```
count_exp <- wage1 %>% filter(exper >=29)
nrow(count_exp)
```

lower predicted ln[wage] ≈ β^ 2 /2β^ 3exper ≈28.79 ≈ 29 years i.e., people in data set with at least 29 years of experience will lower the predicted ln[wage] we have 121 people with greater than or equal to 29 years of experience.

Question 8

Consider a model where the return to education depends upon the amount of work experience (and vice versa): log(wage) = β0 + β1educ + β2exper + β3educ × exper + u

ln[wage]= 5.95+0.0440 educ-0.0215 exper+0.00320 (educ*exper) +u

1.Show that the return to another year of education (in decimal form), holding experience fixed, is β1 + β3exper

Soln: By holding experience fixed above equation will be

∂ ln[wage]= β1 ∂(educ) + β3 ∂(educ) * exper

∂ ln[wage]= ∂(educ) (β1+ β3 * exper)

∂ ln[wage] / ∂(educ) = β1+ β3 * exper

2.State the null hypothesis that the return to education does not depend on the level of experience. What do you think is the appropriate alternative?

Soln:

R-code:

H0 : $\beta3=0$ is the null hypothesis and H1: $\beta3>0$ is the alternative hypothesis which states return to education depends on the level of experience

3. Use the data in the wage2 table to test the null hypothesis in 2 against your stated alternative

Soln:

R-code:

```
model <- lm(log(wage)~educ+exper,data=wage2)
tidy(model)
```

5.95+0.0440 educ-0.0215exper+0.00320(exper*educ) +u

t-stats =2.09 and p-value= 0.0365 of the interaction term is not significant even at 1% level hence we reject the H0 against the H1

4. Let θ1 denote the return to education (in decimal form), when exper = 10: θ1 = β1 + 10β3. Estimate θ1 and a 95% confidence interval for θ1. (Hint: Write β1 = θ1 − 10β3and plug this into the equation; then rearrange. This gives the regression for obtaining the confidence interval for θ1.)

Soln:

R-code:

```
modelb <- lm(log(wage)~educ+exper+I(educ*(-10+exper)),data=wage2)
tidy(modelb,conf.int = TRUE)
```

$\beta1 = \theta1 - 10\beta3$

$\log(wage) = \beta0 + \beta1 educ + \beta2 exper + \beta3 educ \times exper + u$

$\log(wage) = \beta0 + (\theta1 - 10\beta3)\ educ + \beta2 exper + \beta3 educ \times exper + u$

$\log(wage) = \beta0 + \theta1\ educ - 10\beta3\ educ + \beta2 exper + \beta3 educ \times exper + u$

$\log(wage) = \beta0 + \theta1\ educ + \beta2 exper + \beta3\ educ\ (-10 + exper) + u$

5.95+0.761 educ-0.0215exper+0.00320 educ(exper-10) + u

θ1 = 0.761 with 95% confidence interval between 0.0631 and 0.0891

Question 9

Use the data in the gpa2 table for this exercise.

1. Estimate the model sat = β0 + β1hsize + β2hsize^2 + u. where hsize is the size of the graduating class (in hundreds) and write the results in the usual form. Is the quadratic term statistically significant?

Soln:

R-code:

```
model <- lm(sat~hsize+I(hsize^2),data=gpa2)
tidy(model)
summary(model)
```

sat = β0 + β1hsize + β2hsize^2 + u
sat = 998 + 19.8 hsize -2.13 hsize^2 + u, n=935, R-squared= 0.1551
the coefficient of quadratic term is statistically significant at 1% level with t-stat = -3.90 and p-value=0.0000960

2. Using the estimated equation from part 1, what is the "optimal" high school size? Justify your answer

Soln: y=-b/2a which is the turning point in a parabola, -19.8/(2*-2.13) = 19.8/4.26 = 4.65 i.e., 465 will be the optimal high school size

3. Is this analysis representative of the academic performance of all high school seniors? Explain.

Soln: No this doesn't represent academic performance of all high school seniors as the data set doesnot specify the SAT scores of high school student it's just the collective combined score of student who took SAT exam.

4. Find the estimated optimal high school size, using ln(sat) as the dependent variable. Is it much different from what you obtained in part 2?

Soln:
R-code:

```
modela <- lm(log(sat)~hsize+I(hsize^2),data=gpa2)
tidy(modela)
```

ln(sat) = 6.9 + 0.0196 hsize -0.00209 Hsize^2 +u here the optimal high school size is 0.0196/(2*0.00209) = 0.0196/0.00418 = 4.68 i.e., 468 which is close to the value obtained in part 2

Question 10

Use the housing price data in the hprice1 table for this exercise.

1.Estimate the model ln[price] = β0 + β1 ln[lotsize] + β2 ln[sqrf t] + β3b drms + u and report the results in the usual OLS format

Soln:
R-code:

```
summary(lm(log(price)~log(lotsize)+log(sqrft)+bdrms,data=hprice))
```

ln[price] = β0 + β1 ln[lotsize] + β2 ln[sqrf t] + β3bdrms + u
Ln[price] = -1.29 + 0.168 ln(lotsize)+0.7 ln(sqrft)+0.037 ln(bdrms)+ u, n=88 R-squared = 0.643

2. Find the predicted value of price, when lotsize = 20 000, sqrf t = 2 500, and bdrms = 4

Soln:
R-code:

```
model <- lm(log(price)~log(lotsize)+log(sqrft)+bdrms,data=hprice)
predict_model <- predict(model,data.frame(lotsize = 20000,sqrft = 2500,bdrms = 4))
exp(predict_model)
```

Predicted value = 400.574 ➔ 400574$

3. For explaining variation in price, decide whether you prefer the model from part 1 or the model price = β0 + β1lotsize + β2sqrft + β3bdrms + u

Soln:

R-code:

```
summary(lm(price~lotsize+sqrft+bdrms,data=hprice,scipen=999))
```

price = -21.77 + 0.00206 lotsize + 0.1227 sqrft+ 13.85 bdrms +u , n=88 , R-squared =0.6724

For explaining variation in price we prefer model in part 1 which is easy to interpret as the dependent variables continuous type and log is better.