EX2

roi hezkiyahu

11 3 2022

Q.1

Question 1.

Suppose we have a linear model with an intercept, p explanatory variables and i.i.d. normal errors:

$$y=\beta_0+\beta_1 x_1+...+\beta_p x_p+\varepsilon$$

Transform the original response y to $y'=a(y-c)$, where a and c are fixed constants.

1. What will happen to the OLS estimates of β's and to the residual sum of squares (RSS) after this linear transformation?
2. Show that the F-statistics for testing $H_0:\beta_1=...=\beta_s=0 (1\le s\le p)$ will be the same in both cases.
3. Where did you use the assumption of normality for errors? How will you test the hypotheses H_0 (see above) when the distribution of errors ε is different from normal (at least asymptotically)?

a

$$y' = a(y-c)$$

$$\hat{\beta}' = (X^tX)^{-1}X^ty' = (X^tX)^{-1}X^t(a(y-c)) = a\hat{\beta} - ac((X^tX)^{-1}X^t)$$

b

$$\text{we saw in class that: } RSS = \varepsilon^t(I-H)\varepsilon \sim \sigma^2\chi^2_{n-p}$$

$$\text{and also: } \varepsilon^tH\varepsilon + \varepsilon^t(I-H)\varepsilon = \varepsilon^t\varepsilon \sim \chi^2_n \text{ and both are independent}$$

$$H\varepsilon = X(\hat{\beta}' - \beta) \Rightarrow \varepsilon^tH\varepsilon = (\hat{\beta}' - \beta)^tX^tX(\hat{\beta}' - \beta) \sim \sigma^2\chi^2_p$$

$$\text{thus we get: } \frac{(\hat{\beta}' - \beta)^tX^tX(\hat{\beta}' - \beta)}{\sigma^2 p} \Big/ \frac{RSS}{\sigma^2(n-p)} = \frac{(\hat{\beta}' - \beta)^tX^tX(\hat{\beta}' - \beta)}{RSS}\frac{n-p}{p} \sim \frac{\chi_p/p}{\chi_{n-p}/(n-p)} = F_{p,n-p}$$

c

i used the normality for the distribution of $\varepsilon^t(I-H)\varepsilon$ and $\varepsilon^tH\varepsilon$ if the errors are not distributed normally we can try and use a transformation as to make them normal and them use an F test

Q.2

Question 2.

Consider a simple linear regression with a single explanatory variable x. Show that

1. If all n observations x_i are equidistant from their average, than $h_{ii}=2/n$.
2. If all but one observation x_i's are identical, these will have $h_{ii}=1/(n-1)$, while for the remaining observation $h_{ii}=1$

1

$$y = ax + b + \varepsilon$$

$$x_i = x_j \quad \forall i,j$$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_j - \bar{x})^2}$$

$$x_i - \bar{x} = d \Rightarrow h_{ii} = \frac{1}{n} + \frac{d^2}{nd^2} = \frac{2}{n}$$

2

w.l.o.g we can assume i = 1 is the different x

$$x_j = w \quad \forall j > 1$$

$$\bar{x} = \frac{x_1}{n} + \frac{(n-1)}{n} w$$

$$x_1 - \bar{x} = x_1 - \frac{x_1}{n} - \frac{(n-1)}{n} w = \frac{(n-1)(w - x_1)}{n}$$

$$x_j - \bar{x} = w - \frac{x_1}{n} - \frac{(n-1)}{n} w = \frac{(w - x_1)}{n}$$

$$\sum_{j=1}^{n} (x_j - \bar{x})^2 = (x_1 - \bar{x})^2 + (n-1)(w - \bar{x})^2 = (\frac{(n-1)(w-x_1)}{n})^2 + (n-1)\frac{(w-x_1)^2}{n^2} = \frac{[(n-1)^2 - (n-1)](w-x_1)^2}{n^2} = \frac{(n-1)(w}{n}$$

$$h_{11} = \frac{1}{n} + \frac{(x_1 - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \frac{1}{n} + \frac{n-1}{n} = 1$$

$$h_{jj} = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2} = \frac{1}{n} + \frac{(\frac{(w-x_1)}{n})^2}{\frac{(n-1)(w-x_1)^2}{n}} = \frac{1}{n} + \frac{1}{n(n-1)} = \frac{1}{n-1}$$

Q.3

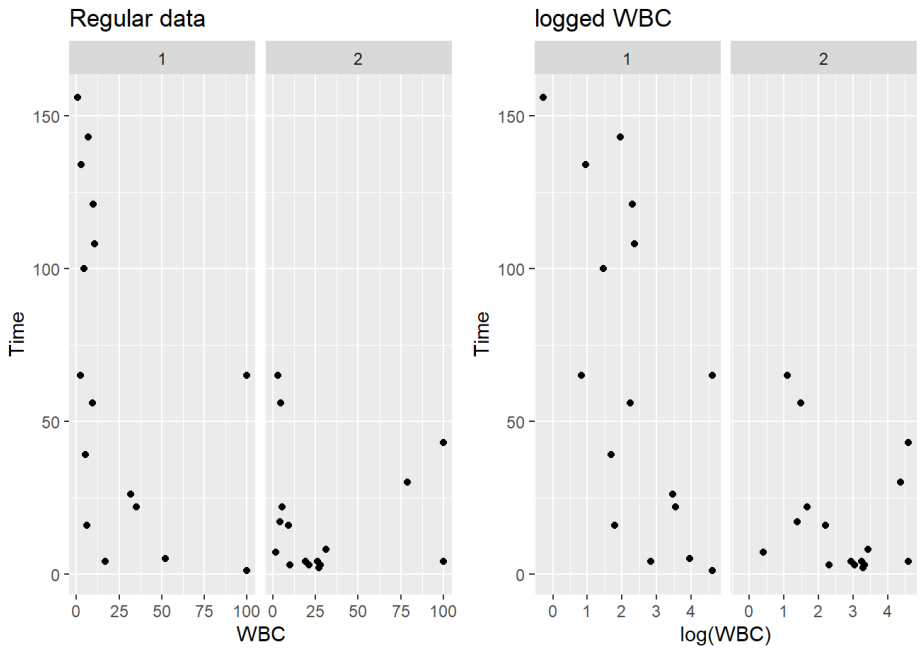
Question 3.

The file Feigl.dat gives the survival times (*Time*) in weeks from initial diagnosis of 33 patients with acute myelogeneous leukaemia, with two covariates: *WBC* (white blood cell count in thousands) and *AG*-factor at the time of diagnosis *(1=Pos, 2=Neg)*.

1. Plot *Time* against *WBC* for each level of *AG*. Does the plot indicate that the linear model will be appropriate? Try the effect of the log-transformations on *Time* and *WBC* on this plot.
2. Fit a full linear regression model (with iteraction) of *Time* on *WBC* and *AG*. Comment the results. Test for parallel regression. Does this model fit the data?
3. Re-fit the model on the log-log scale. Does the effect of *log(WBC)* on *log(Time)* depend on presence of *AG*-factor? Check the adequacy of the resulting model and try to think of possible reasons for problems you found (if any). Compare this model with that of the previous paragraph.

a

```
tbl <- as_tibble(read.table("Feigl.dat"))
colnames(tbl) <- c("Time","WBC","AG")
tbl <- tbl %>%
  mutate(AG = as.factor(AG))
g_reg <- ggplot(tbl,aes(WBC,Time))+
  geom_point()+
  facet_wrap(tbl$AG)+
  ggtitle("Regular data")
g_log <- ggplot(tbl,aes(log(WBC),Time))+
  geom_point()+
  facet_wrap(tbl$AG)+
  ggtitle("logged WBC")
g_reg + g_log
```



b

```
model <- lm(Time~WBC + AG + WBC:AG,data = tbl)
summary(model)
```

```
##
## Call:
## lm(formula = Time ~ WBC + AG + WBC:AG, data = tbl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.113 -15.922  -9.949  24.593  70.910
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  85.6887    11.6336   7.366 4.09e-08 ***
## WBC          -0.7986     0.2539  -3.145 0.003816 **
## AG2         -67.9454    17.0735  -3.980 0.000423 ***
## WBC:AG2       0.8052     0.3828   2.104 0.044206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.07 on 29 degrees of freedom
## Multiple R-squared:  0.429,  Adjusted R-squared:  0.3699
## F-statistic: 7.262 on 3 and 29 DF,  p-value: 0.00089
```

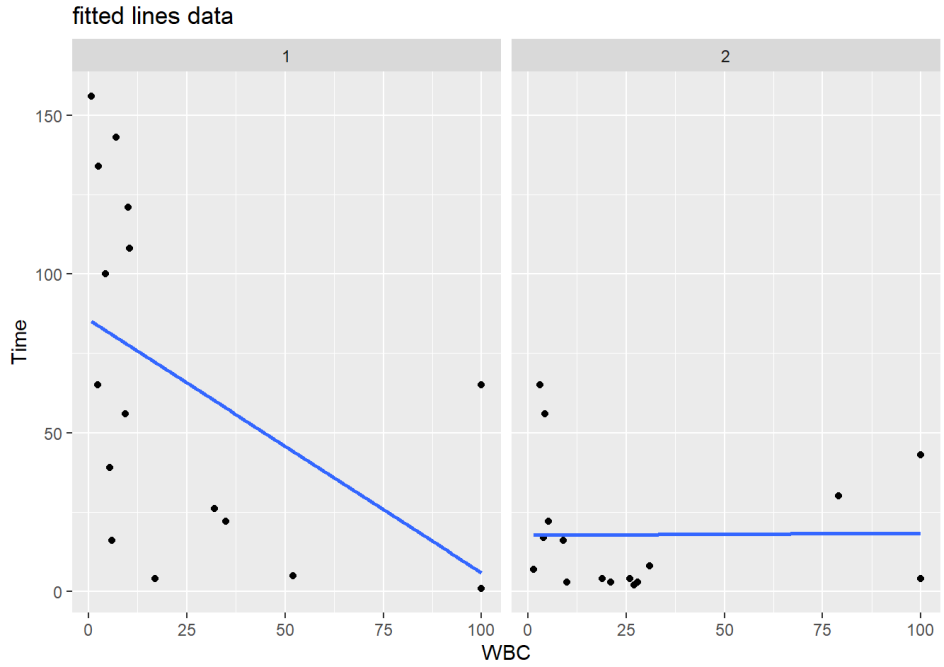
we can see that at least one of the covariates is significant according to the pvalue,lets check the interaction with a nested model anova:

```
model_wo_interaction <- lm(Time~WBC + AG,data = tbl)
anova(model_wo_interaction,model)
```

```
## Analysis of Variance Table
##
## Model 1: Time ~ WBC + AG
## Model 2: Time ~ WBC + AG + WBC:AG
##   Res.Df   RSS Df Sum of Sq     F  Pr(>F)
## 1     30 45937
## 2     29 39856  1    6081.4 4.425 0.04421 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

we can see that the interaction is significant therefore the lines are not parallel, but the low R squered suggests that the model does not fit the data well.

```
pred = data.frame(Time_pred = predict(model,tbl))
ggplot(tbl,aes(WBC,Time))+
  geom_point()+
  facet_wrap(tbl$AG)+
  geom_smooth(method = "lm",formula = y~x,se = FALSE)+
  ggtitle("fitted lines data")
```



we can see from the plots that the line does not fit the data very well

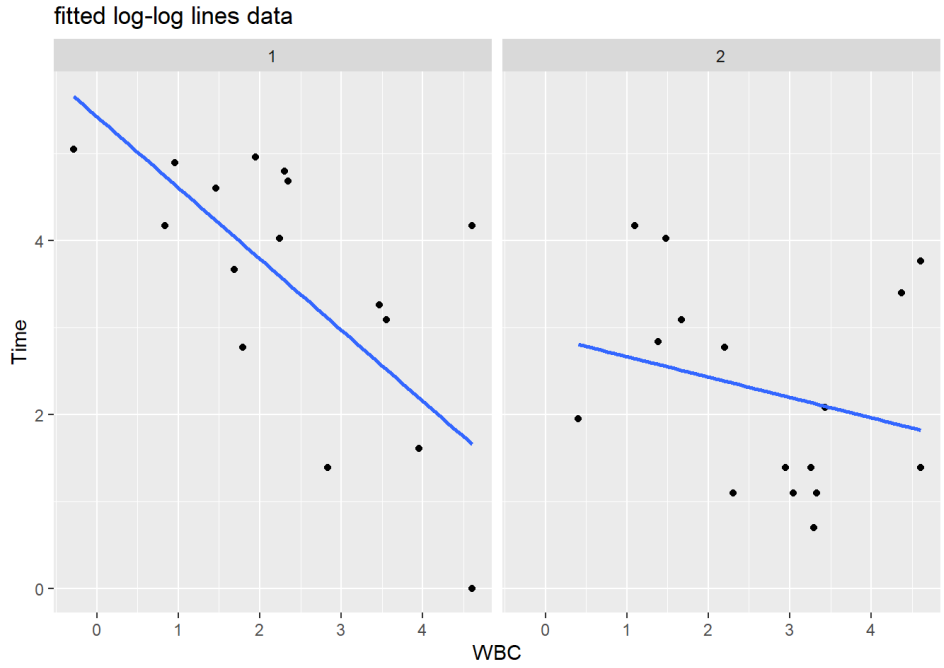
C

```
log_tbl <- tbl %>%
  mutate(across(c(Time,WBC),log))
log_log_model <- lm(Time~WBC + AG + WBC:AG,data = log_tbl)
summary(log_log_model)
```

```
##
## Call:
## lm(formula = Time ~ WBC + AG + WBC:AG, data = log_tbl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7221 -0.8597 -0.0164  0.6670  2.5151
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4254     0.6027   9.002 6.78e-10 ***
## WBC          -0.8178     0.2090  -3.914 0.000505 ***
## AG2          -2.5248     0.9445  -2.673 0.012208 *
## WBC:AG2       0.5834     0.3212   1.817 0.079640 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.203 on 29 degrees of freedom
## Multiple R-squared:  0.4433, Adjusted R-squared:  0.3857
## F-statistic: 7.697 on 3 and 29 DF,  p-value: 0.0006249
```

our pvalue is lower and out R squared is higher which suggestes that the model better fits the data, lets also see it visually

```
log_pred = data.frame(Time_pred = predict(log_log_model,log_tbl))
ggplot(log_tbl,aes(WBC,Time))+
  geom_point()+
  facet_wrap(log_tbl$AG)+
  geom_smooth(method = "lm",formula = y~x,se = FALSE)+
  ggtitle("fitted log-log lines data")
```



fit looks better in both groups

```
log_log_model_wo_interaction <- lm(Time~WBC + AG,data = log_tbl)
summary(log_log_model_wo_interaction)
```

```
##
## Call:
## lm(formula = Time ~ WBC + AG, data = log_tbl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1734 -0.9372  0.2012  0.6411  2.5761
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.8022     0.5142   9.340 2.19e-10 ***
## WBC          -0.5709     0.1647  -3.467  0.00161 **
## AG2          -0.9883     0.4361  -2.266  0.03081 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.249 on 30 degrees of freedom
## Multiple R-squared:  0.3799, Adjusted R-squared:  0.3386
## F-statistic: 9.191 on 2 and 30 DF,  p-value: 0.00077
```

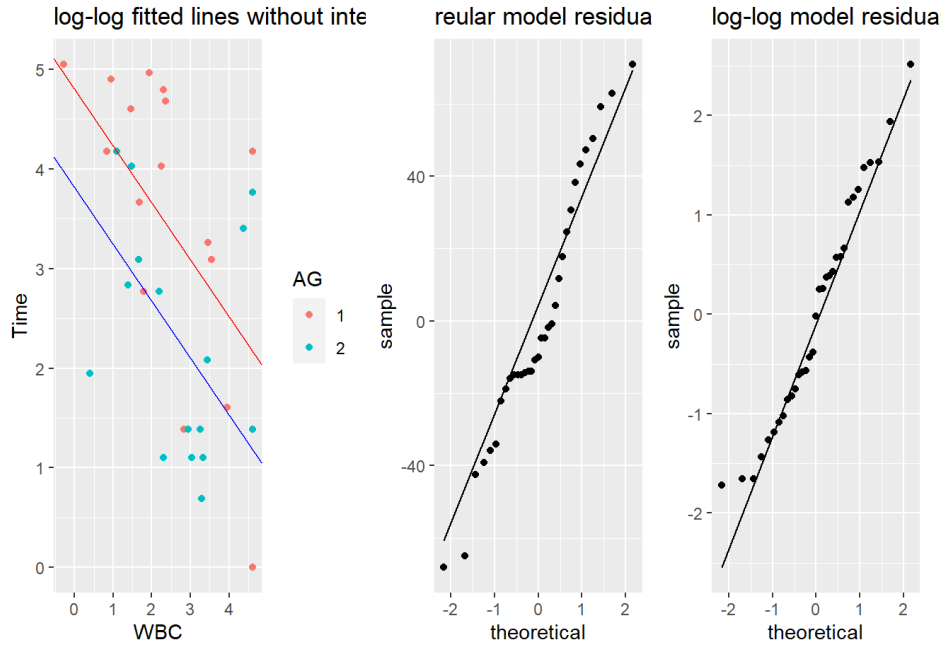
```
anova(log_log_model_wo_interaction,log_log_model)
```

```
## Analysis of Variance Table
##
## Model 1: Time ~ WBC + AG
## Model 2: Time ~ WBC + AG + WBC:AG
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     30 46.778
## 2     29 41.999  1    4.7789 3.2998 0.07964 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

the interaction no longer matters, we would expect that the interaction relevence to the model will stay the same, it could be that the not so normal errors in the regular model will cause this problem, lets comapre the qqplots and see how the non interaction model fits all the data

```
coef_log_log_wo_interaction = log_log_model_wo_interaction$coefficients
slope = coef_log_log_wo_interaction[2]
inter_a = coef_log_log_wo_interaction[1]
inter_b = inter_a + coef_log_log_wo_interaction[3]
g3 <- ggplot(log_tbl,aes(WBC,Time))+
  geom_point(aes(color = AG))+
  geom_abline(slope = slope,intercept = inter_a,color = "red")+
  geom_abline(slope = slope,intercept = inter_b,color = "blue")+
  ggtitle("log-log fitted lines without interaction")

g4 <- ggplot(tibble(res = model$residuals),aes(sample = res)) +
  stat_qq()+
  stat_qq_line()+
  ggtitle("reular model residual qq-plot")
g5 <- ggplot(tibble(res = log_log_model$residuals),aes(sample = res)) +
  stat_qq()+
  stat_qq_line()+
  ggtitle("log-log model residual qq-plot")
g3 + g4+g5
```



Q.4

Question 4.

The file Charges.dat contains data on the sex, the attending physician (*A*,*B* or *C*), severity of illness (*1-4*), total hospital charges (*Chrg*) and age for 49 patients, all of whom had an identical diagnosis, from Northwestern Memorial Hospital, Chicago.

1. Fit the main effect model expressing the charges against age and the other variables (don't forget first to express them as suitable indicator variables where necessary). Is the linear model adequate for this data?
2. Find the appropriate transformation of the dependent variable from the Box-Cox transformation family, re-fit the model and comment its adequacy.
3. Test the hypotheses that the attending physician has no effect on hospital charges (on the chosen scale).
4. Some feminist organizations claim that there is sexual discrimination in the hospital and women suffer from higher hospital charges. Does their claim have any statistical ground?
5. Point out influential observation(s) that strongly affected your model (if any). Remove them from the data and re-fit the model. Comment the results. Repeat Step 3 and Step 4.
6. Repeat Step 2 without influential observations you've found. Did you get the same scale for the response variable as before? Try to explain this phenomenon.
7. Are you completely satisfied with the resulting model(s)? If "yes", *mazal tov!*; if "no", give an idea(s) of improving it.

a

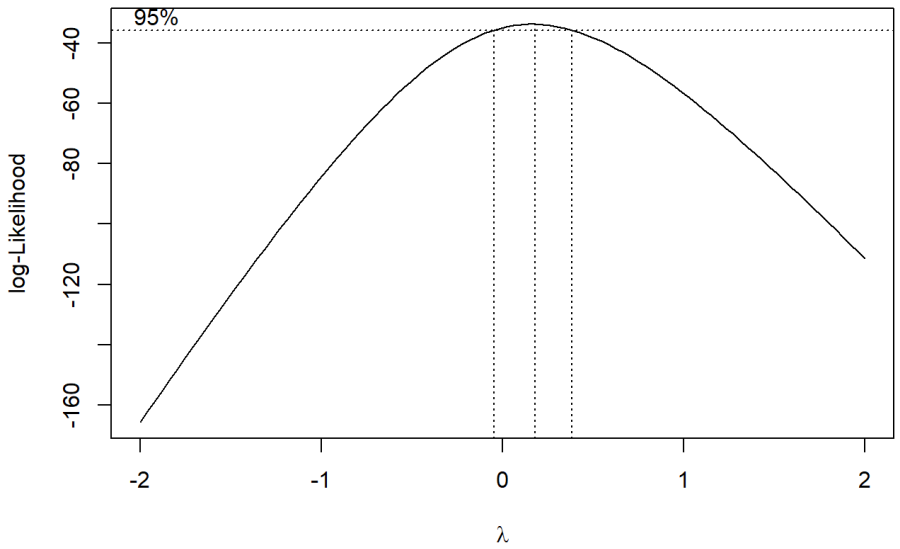
```
charges <- as_tibble(read.table("Charges.dat"))
colnames(charges) <- c("sex","physician","severity","Chrg","age")
charges <- charges %>%
  mutate(across(c(sex,physician,severity),as.factor))
lm_model <- lm(Chrg~.,data = charges)
summary(lm_model)
```

```
##
## Call:
## lm(formula = Chrg ~ ., data = charges)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7935  -2748   -785   1681  33225
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1742.54    4793.36  -0.364 0.718075
## sexM         -979.08    2133.96  -0.459 0.648796
## physicianB   1530.12    2368.62   0.646 0.521882
## physicianC   5332.34    2520.54   2.116 0.040503 *
## severity2    3613.57    2662.63   1.357 0.182161
## severity3   11703.53    2931.03   3.993 0.000264 ***
## severity4   18626.57    3315.32   5.618 1.51e-06 ***
## age           127.09      69.03   1.841 0.072859 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6464 on 41 degrees of freedom
## Multiple R-squared:  0.6779, Adjusted R-squared:  0.6229
## F-statistic: 12.33 on 7 and 41 DF,  p-value: 2.302e-08
```

the R squared of 0.6779 suggests that the model fits the data ok but not good enough

b

```
boxcox(lm_model)
```



we can see that $\lambda = 0$ is at the interval so i will use the ln transformation

```
log_model <- lm(log(Chrg)~.,data = charges)
summary(log_model)
```

```
##
## Call:
## lm(formula = log(Chrg) ~ ., data = charges)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54602 -0.17094 -0.06013  0.11014  0.83567
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.38832    0.23609  31.295  < 2e-16 ***
## sexM         0.05206    0.10511   0.495  0.62300
## physicianB   0.05961    0.11666   0.511  0.61213
## physicianC   0.34108    0.12415   2.747  0.00889 **
## severity2    0.45439    0.13114   3.465  0.00126 **
## severity3    0.83116    0.14436   5.757 9.57e-07 ***
## severity4    1.04479    0.16329   6.398 1.17e-07 ***
## age          0.02062    0.00340   6.065 3.50e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3184 on 41 degrees of freedom
## Multiple R-squared:  0.8398, Adjusted R-squared:  0.8124
## F-statistic:  30.7 on 7 and 41 DF,  p-value: 2.301e-14
```

we got R-squared: 0.8398 that's much better then out previous model

c

```
log_no_phys <- lm(log(Chrg)~.,data = dplyr::select(charges,-physician))
anova(log_no_phys,log_model)
```

```
## Analysis of Variance Table
##
## Model 1: log(Chrg) ~ sex + severity + age
## Model 2: log(Chrg) ~ sex + physician + severity + age
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     43 5.0335
## 2     41 4.1564  2   0.87716 4.3263 0.01974 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

we can reject the null at 95% confidence: thus conclude that physician has an effect on hospital charges

d

lets see if sex has an effect on hospital charges

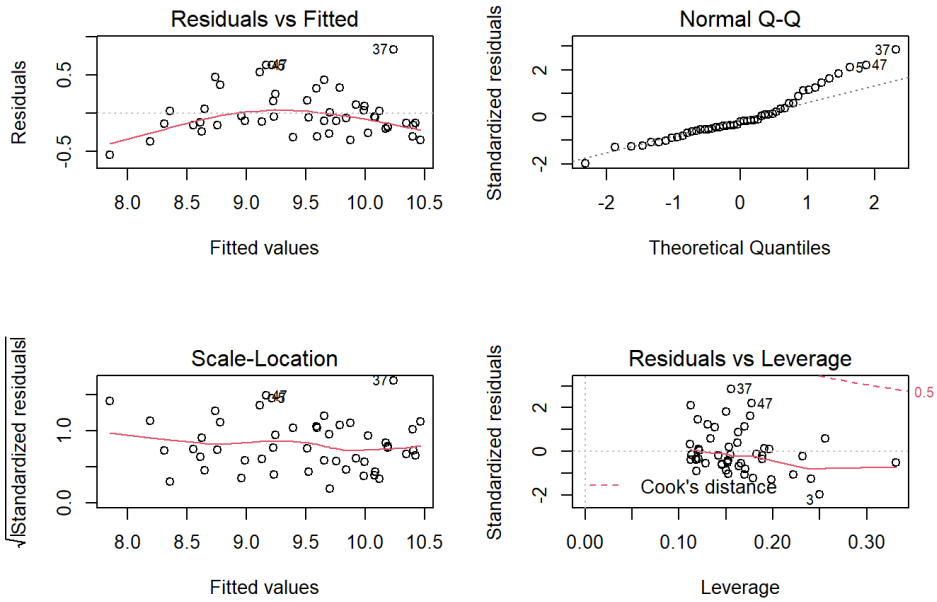
```
log_no_sex <- lm(log(Chrg)~.,data = dplyr::select(charges,-sex))
anova(log_no_sex,log_model)
```

```
## Analysis of Variance Table
##
## Model 1: log(Chrg) ~ physician + severity + age
## Model 2: log(Chrg) ~ sex + physician + severity + age
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     42 4.1812
## 2     41 4.1564  1  0.024875 0.2454  0.623
```

we can not reject the null at 95% confidence: thus we can not conclude that sex effect hospital charges

e

```
par(mfrow = c(2,2))
plot(log_model)
```




```
influential <- c(3,37,47)
glue("influential observations based on plots are: " ,glue_collapse(influential,sep=", "))
```

```
## influential observations based on plots are: 3, 37, 47
```

```
re_charge <- charges[-influential,]
re_log_model <- lm(log(Chrg)~.,data = re_charge)
re_log_no_phys <- lm(log(Chrg)~.,data = dplyr::select(re_charge,-physician))
anova(re_log_no_phys,re_log_model)
```

```
## Analysis of Variance Table
##
## Model 1: log(Chrg) ~ sex + severity + age
## Model 2: log(Chrg) ~ sex + physician + severity + age
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     40 2.7825
## 2     38 2.4144  2   0.36807 2.8964 0.06749 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

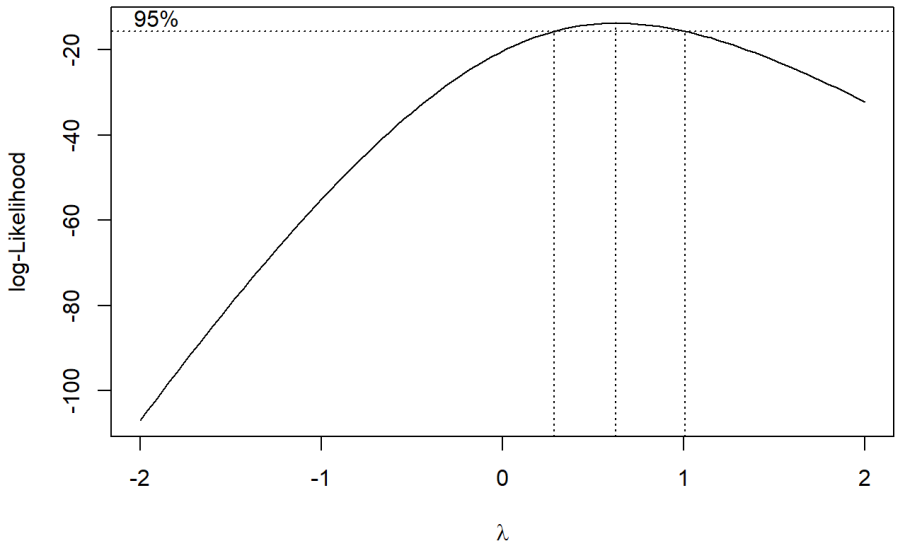
```
re_log_no_sex <- lm(log(Chrg)~.,data = dplyr::select(re_charge,-sex))
anova(re_log_no_sex,re_log_model)
```

```
## Analysis of Variance Table
##
## Model 1: log(Chrg) ~ physician + severity + age
## Model 2: log(Chrg) ~ sex + physician + severity + age
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     39 2.4168
## 2     38 2.4144  1  0.002398 0.0377  0.847
```

we can see that now we do not reject the null hypotheses for physician meaning it does not effect charges and we still don't reject the null for sex.

f

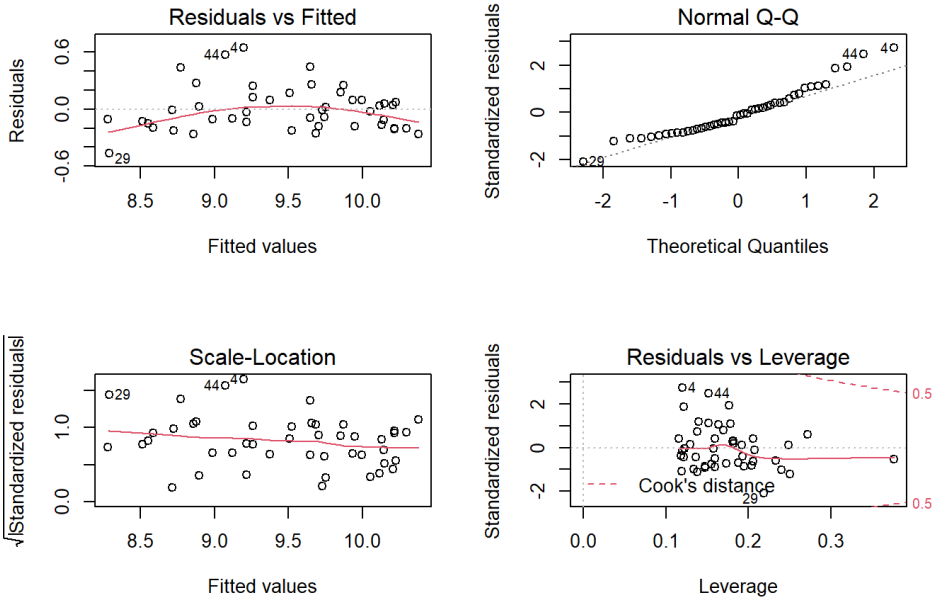
```
boxcox(lm(Chrg~.,data = re_charge))
```



the interval for λ moved. an explanation to this phenomenon can be that removing influential observations effects the likelihood function thus our λ changes.

g

```
par(mfrow = c(2,2))
plot(re_log_model)
```




```
summary(re_log_model)
```

```
##
## Call:
## lm(formula = log(Chrg) ~ ., data = re_charge)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46501 -0.17466 -0.02768  0.09462  0.64541
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.532956   0.203186  37.074  < 2e-16 ***
## sexM        0.016434   0.084594   0.194  0.84700
## physicianB  0.125850   0.094156   1.337  0.18929
## physicianC  0.241554   0.100807   2.396  0.02159 *
## severity2   0.340187   0.107810   3.155  0.00313 **
## severity3   0.798395   0.114930   6.947 2.90e-08 ***
## severity4   0.901634   0.134296   6.714 6.01e-08 ***
## age         0.019377   0.002943   6.585 9.01e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2521 on 38 degrees of freedom
## Multiple R-squared:   0.87,  Adjusted R-squared:  0.846
## F-statistic: 36.33 on 7 and 38 DF,  p-value: 6.207e-15
```

i think we can do better with the model, maybe try several lambdas in the interval and get a more suitable transformation (we can see that our log transformation could improve based on the qqplot) or maybe removing non significant covariants