Stat500(Section002): Homework #4

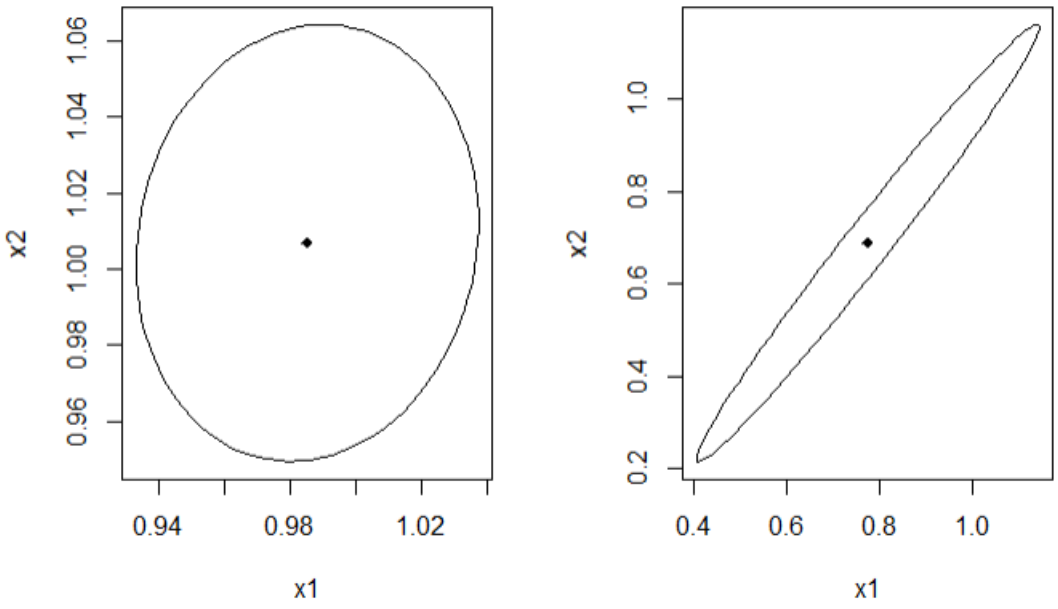
Due on Oct.13, 2021 at 3:00pm

Instructor:Naisyin Wang

Tiejin Chen **tiejin@umich.edu**

Problem 1

We use the code provided by the problem to get the following result:




```
[1] -0.09644917
       x1     x2
x1 0.0046 0.0005
x2 0.0005 0.0055
       x1     x2
x1 0.2644 0.3336
x2 0.3336 0.4326
```

For model B, the major axis and minor axis of the ellipse are not roughly parallel with the x-axis and y-axis corresponding to $\hat{\beta_{x1}}, \hat{\beta_{x2}}$. And -0.09644917 is the covariance of x_1, x_2 which is not 0. This shows that even x_1, x_2 are generated independently, they are still not independent mathmatically.And the rest of output is part of $(X_A^T X_A)^{-1}, (X_B^T X_B)^{-1}$. Where X_A is data matrix of A and X_B is of B.
Now let us explain this. Let us consider $Cov(\hat{\beta_{x1}}, \hat{\beta_{x2}})$, we have:

$$Cov(\hat{\beta_{x1}}, \hat{\beta_{x2}}) = \sigma^2 (X^T X)^{-1}_{(2,3)}$$

For model A, we have:

$$Cov(\hat{\beta_{x1}}, \hat{\beta_{x2}}) = \sigma^2 (X_A^T X_A)^{-1}_{(2,3)} = 0.0005\sigma^2$$

which will be very small and very close to 0.Which means the $\hat{\beta_{x1}}, \hat{\beta_{x2}}$ of model A is nearly orthogonal. Now let us see the situation of model B:

$$Cov(\hat{\beta_{x1}}, \hat{\beta_{x2}}) = \sigma^2 (X_B^T X_B)^{-1}_{(2,3)} = 0.3336\sigma^2$$

which is of course much larger than 0. Hence, we think $\hat{\beta_{x1}}, \hat{\beta_{x2}}$ of model B is far away from orthogonal. Thus, we should not expect the major axis and minor axis of the ellipse are roughly parallel with the x-axis and y-axis.
This phenomenon shows that the estimated coefficients and the corresponding inference are adjusted to the existence of other predictors.

Problem 2

Part a

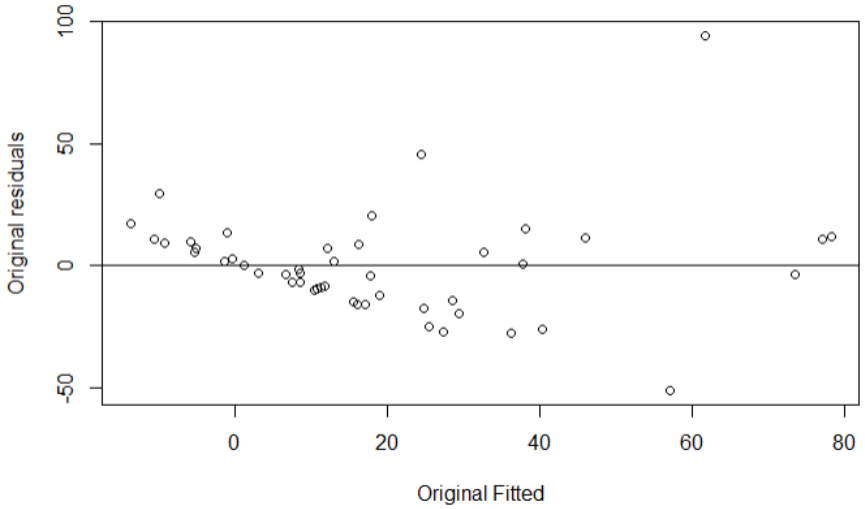
we use the following code to draw figures of Yorg:

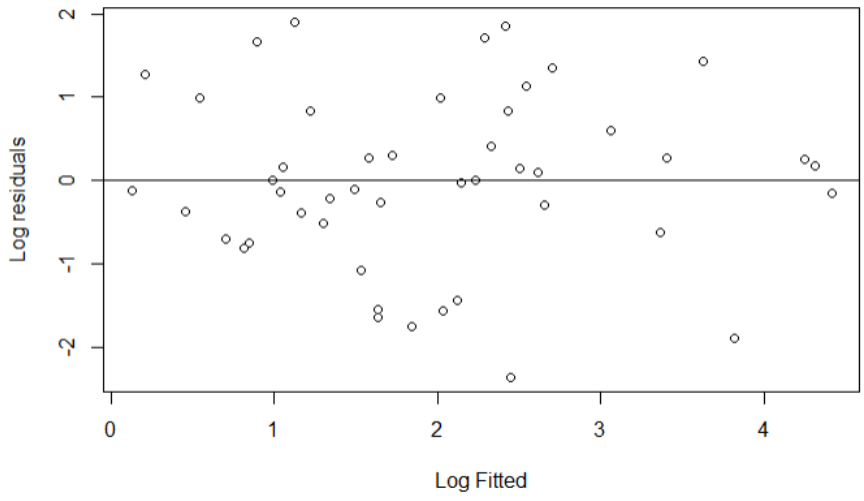
```
outorg = lm(gamble~sex+status+income+verbal,data=teengamb)
outlog = lm(log(gamble+1)~sex+status+income+verbal,data=teengamb)
plot(outorg$fitted,outorg$residual,
     xlab = "Original Fitted",ylab = 'Original residuals')
abline(h=0)
```

use the following code to draw figures of Ylog:

```
plot(outlog$fitted,outlog$residual,
     xlab = "Log Fitted",ylab = 'Log residuals')
abline(h=0)
```

We get the result:





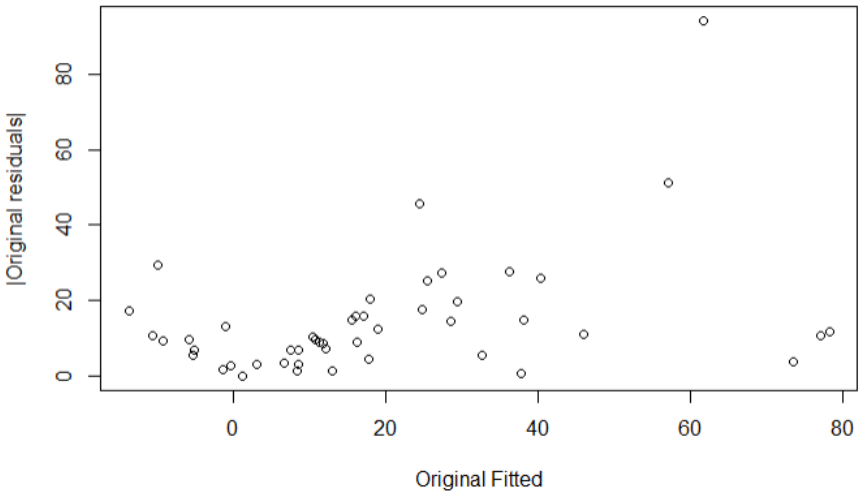
We can see that when original fitted value is less than 0, all the points is above 0, Hence it can not be centering at 0 when fitted values are less than 0. Also, the figure shows some pattern that when fitted value is larger than 40, it might not be centering at 0.
For Model B, we can see from the figure that it is centering at 0 in all areas.

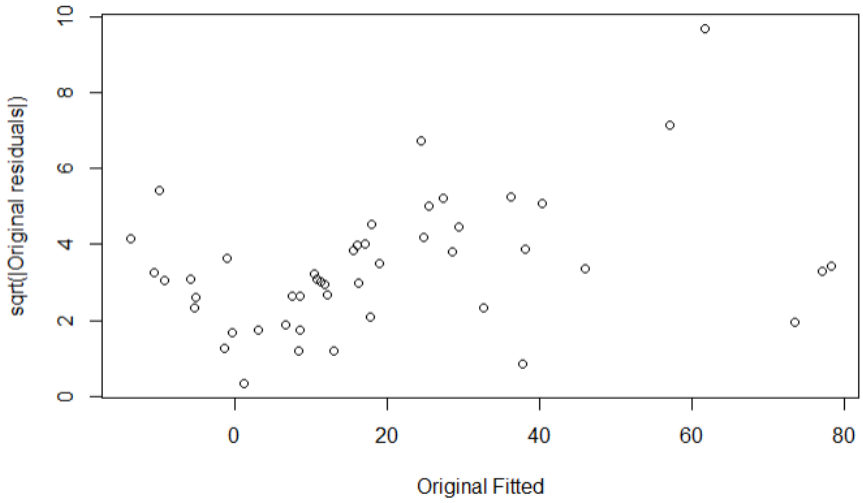
Part b

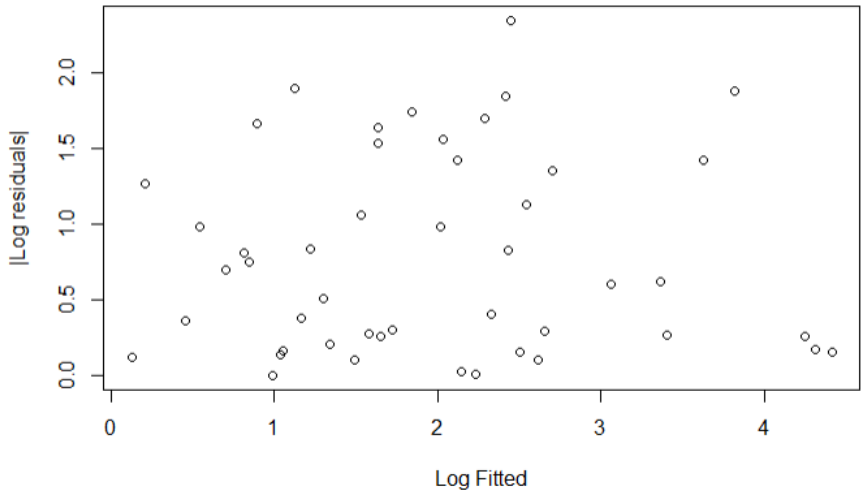
We can get some results from the figure in part a. However, to furthur consider this problem, we draw figure abs(residual) against fitting value and sqrt(abs(residual)) against fitting value. We use the following code:

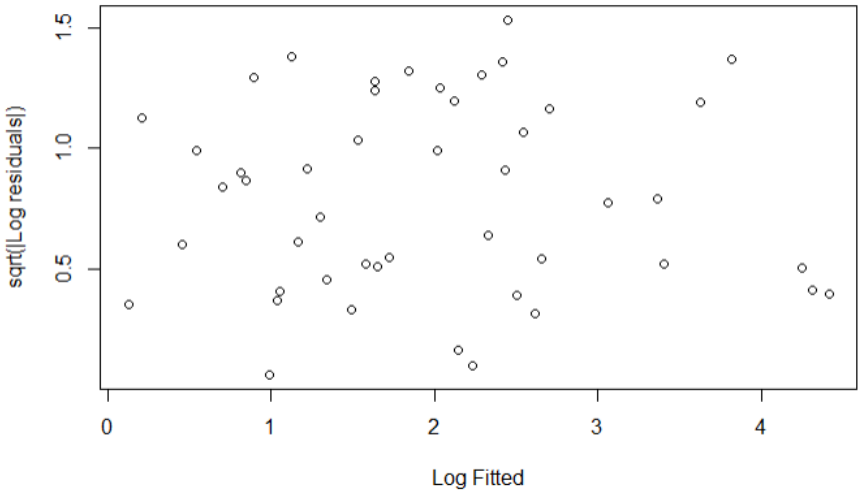
```
plot(outorg$fitted,abs(outorg$residual),
     xlab="Original Fitted",ylab='|Original residuals|')
plot(outorg$fitted,sqrt(abs(outorg$residual)),
     xlab="Original Fitted",ylab='sqrt(|Original residuals|)')
plot(outlog$fitted,abs(outlog$residual),
     xlab="Log Fitted",ylab='|Log residuals|')
plot(outlog$fitted,sqrt(abs(outlog$residual)),
     xlab="Log Fitted",ylab='sqrt(|Log residuals|)')
```

We get the result:




sqrt(|Original residuals|)
0
2
4
6
8
10
0
20
40
60
80
Original Fitted


|Log residuals|
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
Log Fitted


sqrt(|Log residuals|)
0.5
1.0
1.5
0
1
2
3
4
Log Fitted

We can see that, in the figure of model A, abs(residuals)(sqrt(abs(residuals))) and fitting value have some kind of linear relationship while abs(residuals)(sqrt(abs(residuals))) and fitting value of model B have no linear relationship. To check this, We calculate the correlation between them:

- Model A: the correlation bewteen abs(residuals) and fitting value is 0.3772323, the correlation bewteen sqrt(abs(residuals)) and fitting value is 0.355601.
- Model B: the correlation bewteen abs(residuals) and fitting value is 0.00548139, the correlation bewteen sqrt(abs(residuals)) and fitting value is -0.009811541.

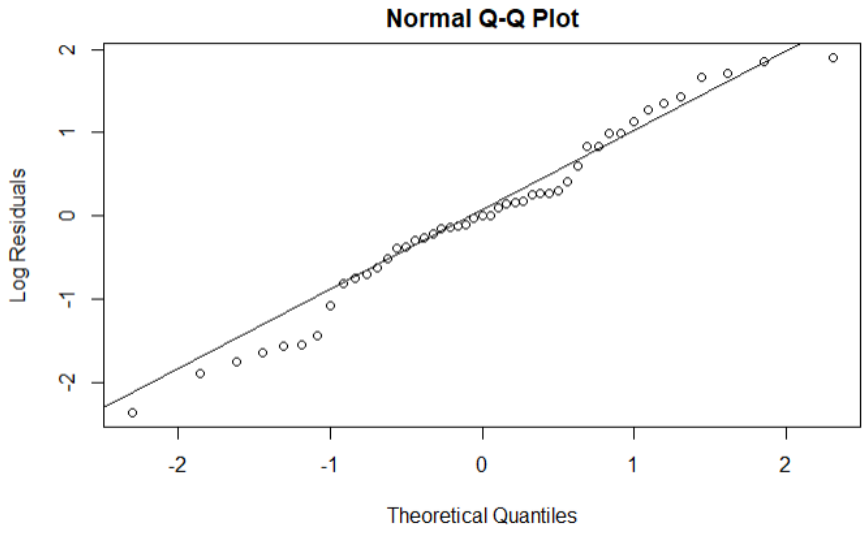
The result shows that the analysis we have above is right. And this shows that for model A, the constant variance assumption does not hold. And for model B, the constant variance assumption holds.

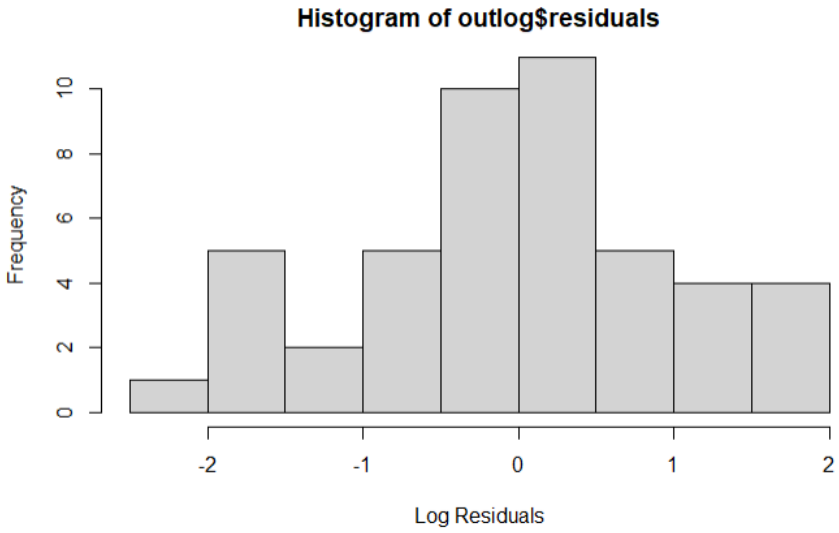
Part c

We only check the normality assumption for model B since model A does not hold constant variance assumption and its residuals are not centering at zero in all areas. We use the following code to draw the QQ-plot and histogram as well as use Shapiro-Wilk test:

```
qqnorm(outlog$residuals,ylab = 'Log Residuals')
qqline(outlog$residuals)
hist(outlog$residuals,xlab="Log Residuals")
shapiro.test(outlog$residuals)
```

And we get the result:






```
        Shapiro-Wilk normality test

data:  outlog$residuals
W = 0.97609, p-value = 0.4418
```

We find that the figure of residuals looks like a normal distribution in QQ-plot and histogram. And the p-value of Shapiro-Wilk test is 0.4418 which is larger than 0.05. Hence we will accept H_0: the distribution is a normal distribution. Overall, we think model B holds normality assumption.

Problem 3

We want to get interval of future observation. we use the following code to get result:

```
new.teengamb = data.frame(
  sex=0,status=40,income=2,verbal = 6
)
predict_log = predict(outlog,newdata =new.teengamb,interval = 'prediction')
prediction = exp(predict_log)-1
prediction
```

we get the result: the estimation is 4.875715. And the 95% prediction interval is (0.3981511,56.36328).