

# STATS 500 HW4

Minxuan Chen

2023-10-04

## Table of contents

Problem 1	1
(a)	1
(b)	5
Problem 2	6
(a)	6
(b)	7

Github repo: [https://github.com/PKUniiice/STATS\\_500](https://github.com/PKUniiice/STATS_500)

## Problem 1

(a)

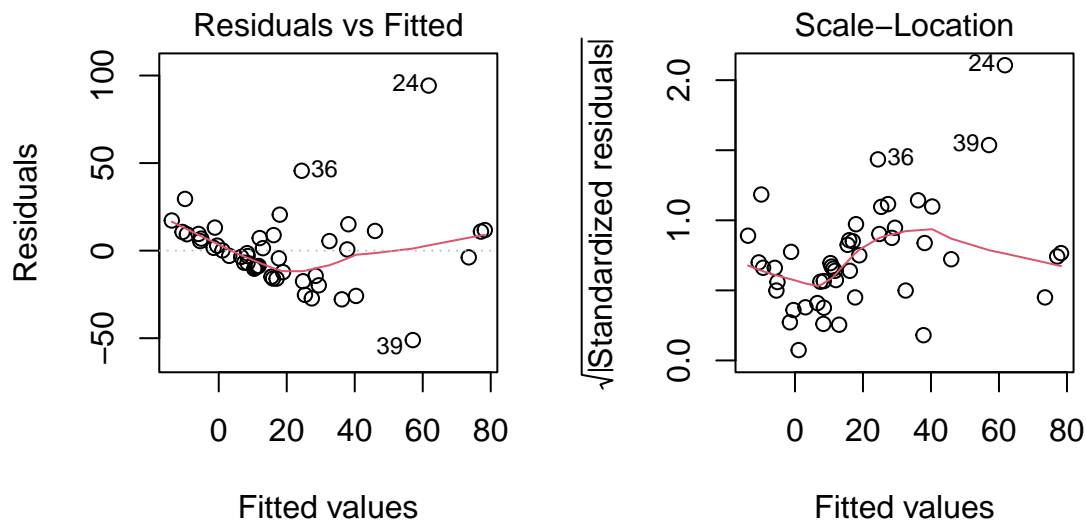
First, we perform diagnostics on the original model.

```
1 library(faraway)
2 data(teengamb)
3 # sex has already been encoded as 0,1, so we don't need to convert it to factor
4 m.ori <- lm(gamble ~ sex+status+income+verbal, data=teengamb)
5 #summary(m.ori)
```

- Check the constant variance assumption for the errors.

We can use both the **Residuals vs Fitted** and **Scale-Location** plots to check for variance issues. It's evident that, as  $x$  increases, the magnitude of residuals also increases. So we conclude that there is heteroscedasticity. Note that this violation may result in bias into all inferences. As a remedy, we choose to apply a transformation and proceed with the remaining diagnosis on the new model.

```
1 par(mfrow = c(1,2))
2 plot(m.ori, which=1)
3 plot(m.ori, which=3)
```



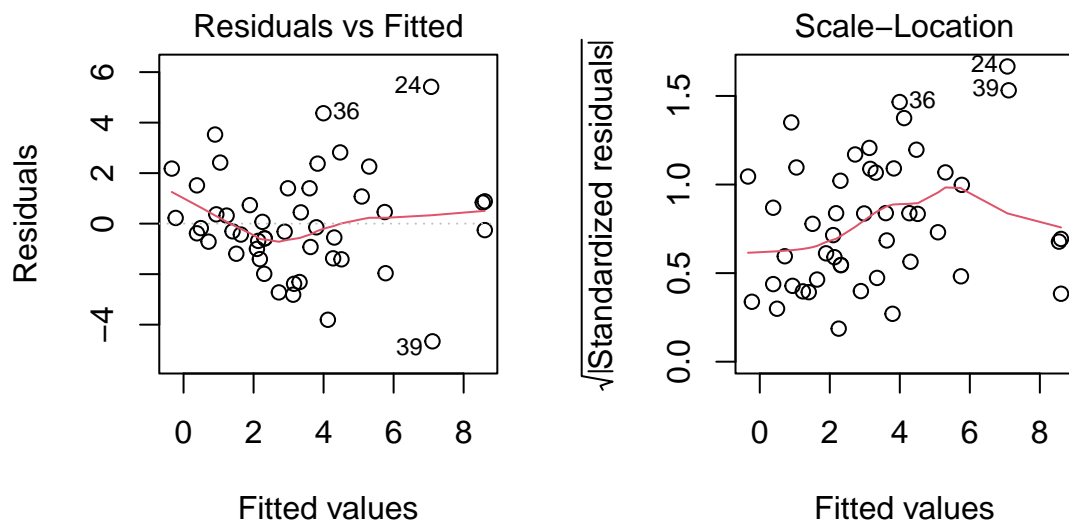
```
1 par(mfrow = c(1,1))
```

We take the square root of the response `gamble`.

```

1 #new model
2 m.new <- lm(sqrt(gamble) ~ ., data=teengamb)
3 #summary(m.new)
4
5 par(mfrow = c(1,2))
6 plot(m.new, which=1)
7 plot(m.new, which=3)

```



```

1 par(mfrow = c(1,1))

```

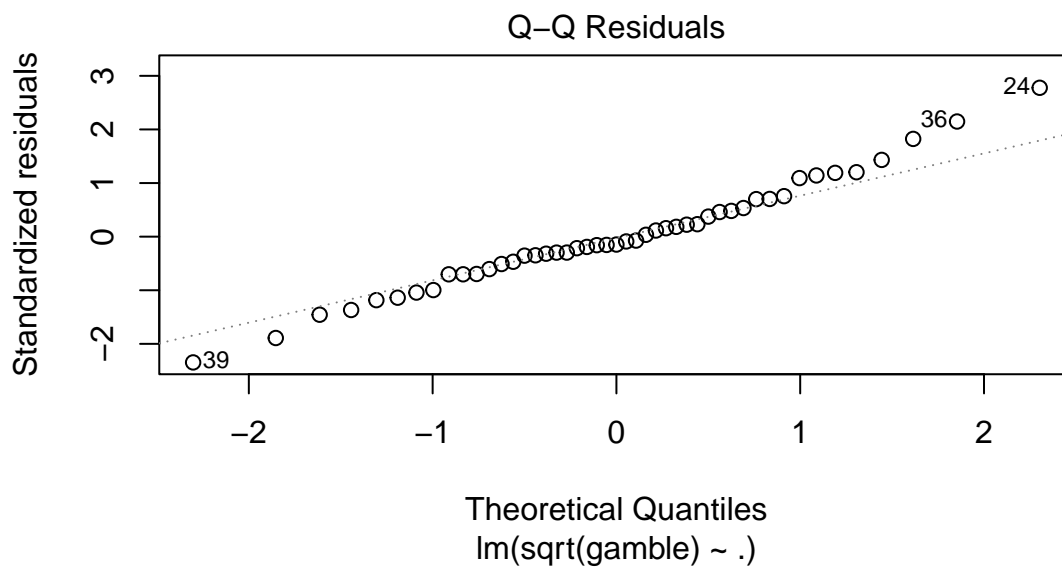
In the new model, we still observe some slight heteroscedasticity in residuals, although it is not as severe as in the original model.

- Check the normality assumption.

```

1 plot(m.new, which=2)

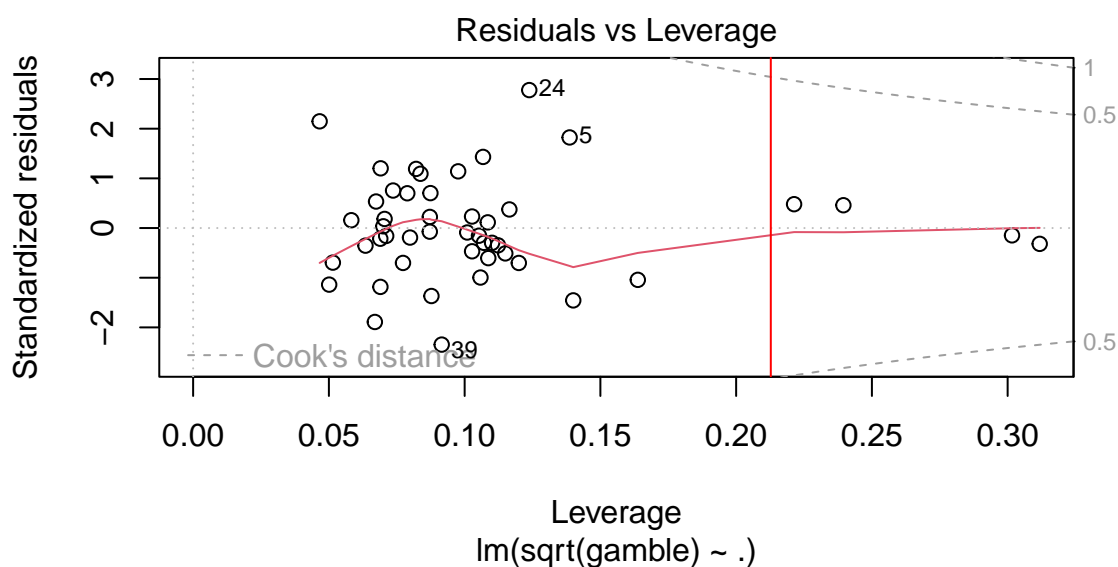
```



From cases No.39,36 and 24, we observe that the residuals are slightly heavy-tailed compared to a normal distribution. So, there is a slight violation of the normality assumption.

- Check for large leverage points.

```
1 plot(m.new, which=5)
2 abline(v=2*length(m.new$coefficients)/nrow(teengamb), col='red')
```



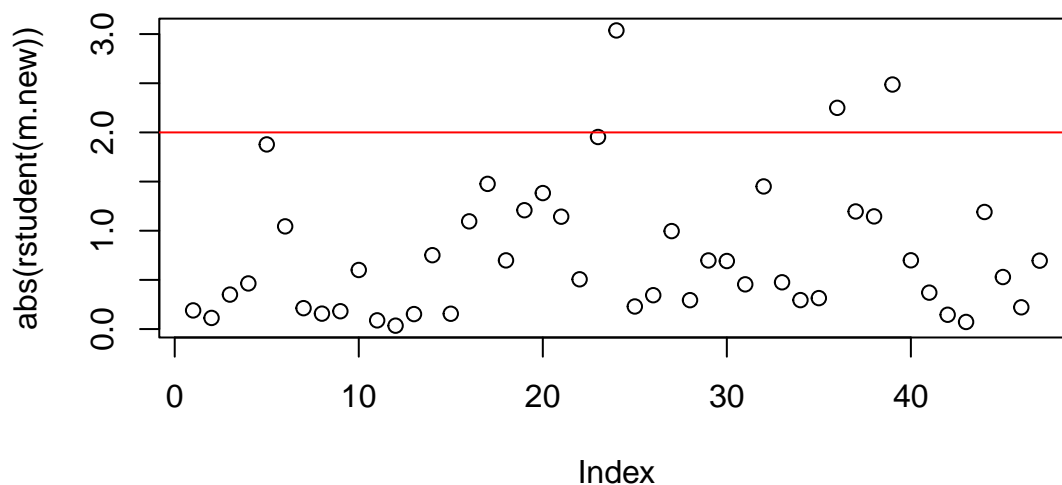
```
1 hatvalues(m.ori)>2*length(m.new$coefficients)/nrow(teengamb)
```

1	2	3	4	5	6	7	8	9	10	11	12	13
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
14	15	16	17	18	19	20	21	22	23	24	25	26
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
27	28	29	30	31	32	33	34	35	36	37	38	39
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
40	41	42	43	44	45	46	47					
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE					

When we add a vertical line  $x = \frac{2(p+1)}{n}$  on the **Residuals vs Leverage** plot, we can identify 4 points with high leverage. Upon direct calculation, case No.31, 33, 35 and 42 are large leverage points.

- Check for outliers.

```
1 plot(abs(rstudent(m.new)))
2 abline(h=2, col='red')
```



We can plot all absolute values of studentized deleted residuals. One empirical rule of outliers is

$$|t_i| > 2$$

We find it's likely that there are three outliers.

We can also perform a test.

```
1 library(car)
```

Loading required package: carData

Attaching package: 'car'

The following objects are masked from 'package:faraway':

logit, vif

```
1 outlierTest(m.new)
```

No Studentized residuals with Bonferroni  $p < 0.05$

Largest |rstudent|:

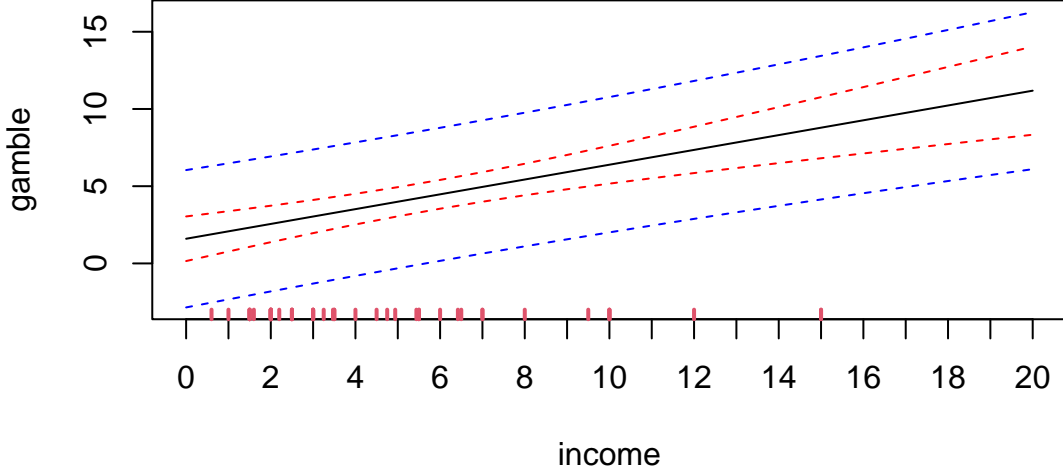
	rstudent	unadjusted p-value	Bonferroni p
24	3.037005	0.0041428	0.19471

The test result tells that there are no outliers.

## (b)

We generate the pointwise confidence/prediction band (ref:[click](#)).

```
1 newx <- data.frame(  
2     sex = 0,  
3     income = seq(0,20),  
4     status = 43,  
5     verbal = 7  
6 )  
7 conf <- predict(m.new, newdata=newx, interval="confidence")  
8 pred <- predict(m.new, newdata=newx, interval="prediction")  
9 matplot(newx$income, cbind(conf, pred[,2:3]),  
10     lty=c(1,2,2,2,2),  
11     col=c(1, 'red', 'red', 'blue', 'blue'), type="l",  
12     xlab="income", ylab="gamble", xaxt="n")  
13 axis(1, at = seq(0, 20))  
14 rug(teengamb$income,col=2, lwd=2)
```



From the rugplot, we observe that the majority of incomes fall within the range of  $[0, 10]$ . As a result, the model will provide confident inference results within this range. The width of the confidence and prediction bands further supports this observation. However, when income exceeds 10, there are only three data points available in the original dataset. Consequently, making inferences in this range is less likely to yield precise results.

## Problem 2

(a)

Note that if  $Z \sim \mathcal{N}(0, 1)$ , then  $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$  and the cdf of normal distribution is strictly monotonic. By definition

$$\begin{aligned} \mathbb{P}(X \leq F^{-1}(q)) &= q \\ &= \mathbb{P}(\mu + \sigma Z \leq F^{-1}(q)) \\ &= \mathbb{P}(Z \leq \frac{F^{-1}(q) - \mu}{\sigma}) \end{aligned}$$

i.e.

$$\mathbb{P}(Z \leq \frac{F^{-1}(q) - \mu}{\sigma}) = q$$

Use the definition of quantile function again

$$\frac{F^{-1}(q) - \mu}{\sigma} = \Phi^{-1}(q) \rightarrow F^{-1}(q) = \mu + \sigma \Phi^{-1}(q)$$

**(b)**

In fact, we need the converse version of result in (a). That is, for a r.v.  $X$  with cdf  $F$  and quantile function  $F^{-1}$ , if, for all  $p$ ,

$$F^{-1}(q) = \mu + \sigma\Phi^{-1}(q)$$

then  $X \sim N(\mu, \sigma^2)$

*Proof*

$$\mathbb{P}(X \leq F^{-1}(q)) = \mathbb{P}(X \leq \mu + \sigma\Phi^{-1}(q)) = \mathbb{P}\left(\frac{X - \mu}{\sigma} \leq \Phi^{-1}(q)\right) = q, \forall 0 < q < 1$$

Therefore,  $\Phi^{-1}(q)$  must be the quantile function of  $\frac{X - \mu}{\sigma}$ , i.e.

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) \rightarrow X \sim N(\mu, \sigma^2)$$

In QQ plots for normality test, we plot

$$\left\{y = r_{[i]}, x = \Phi^{-1}\left(\frac{i}{n+1}\right)\right\}$$

where  $n+1$  is the correction for continuous distribution.

After sorting  $r_i$  to  $r_{[i]}$ , this sequence consists of quantile points of the underlying distribution.

By identifying the corresponding quantiles in  $\Phi^{-1}$ , if we observe  $y - x$  forms a line or closely resembles a line, then we can conclude that the distribution of residual satisfies

$$F^{-1}(q) = \mu + \sigma\Phi^{-1}(q)$$

therefore, we conclude that the residuals follow a normal distribution.