STATS 500 HW4

Minxuan Chen

2023-09-29

Table of contents

Github repo: https://github.com/PKUniiiiice/STATS_500

Problem 1

(a)

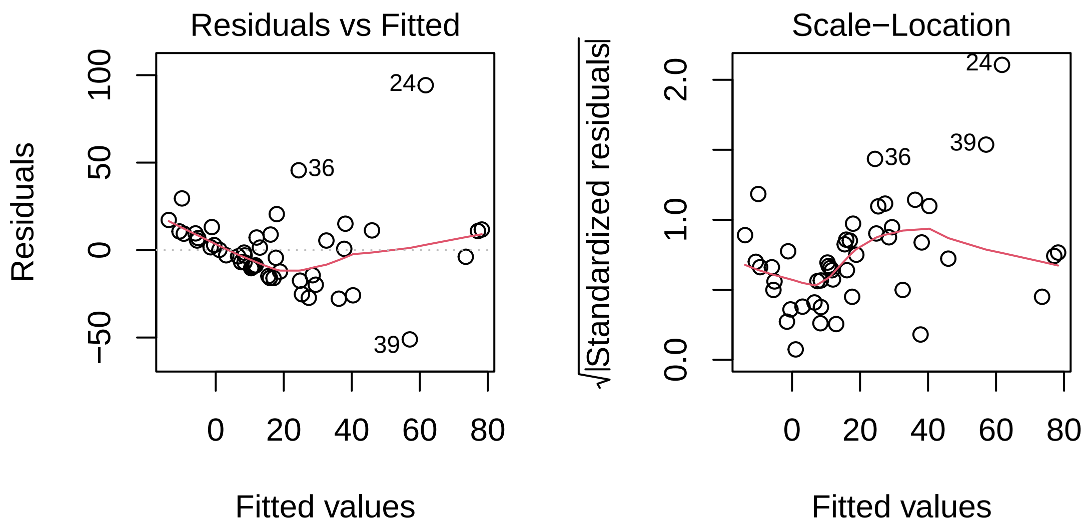
First, we perform diagnostics on the original model.

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

- Check the constant variance assumption for the errors.

We can use `Residuals vc Fitted` and `Scale-Location` plots together to check the variance. It's clear that, with x increase, the magnitude of residuals increase as well. So we conclude that there exists heteroscedasticity. Note that this violation may result in bias in all inferences, so we choose to make a transformation as remedy and continue the remaining diagnosis on the new model.

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

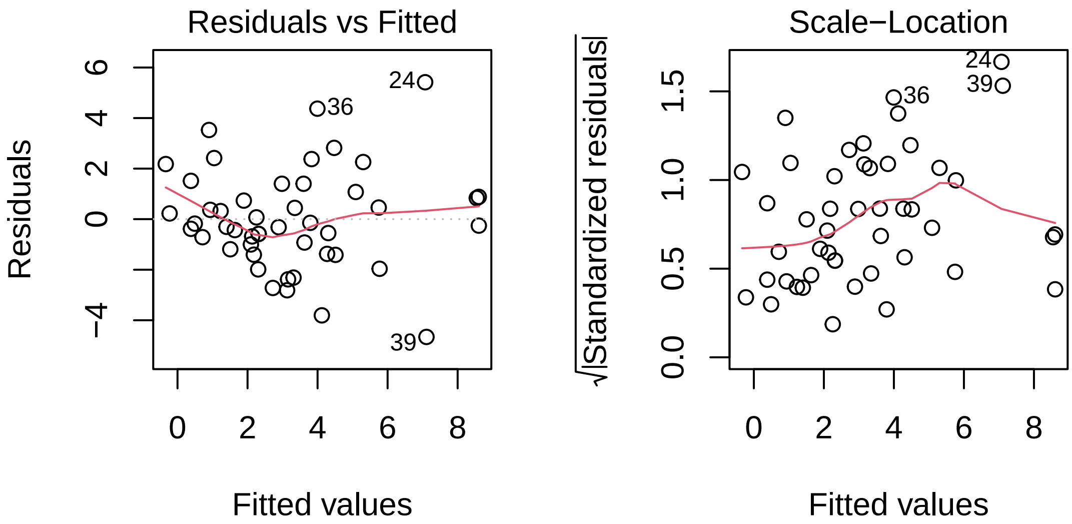



```r
par(mfrow = c(1,1))
```

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

```r
#new model
m.new <- lm(sqrt(gamble) ~ ., data=teengamb)
#summary(m.new)

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

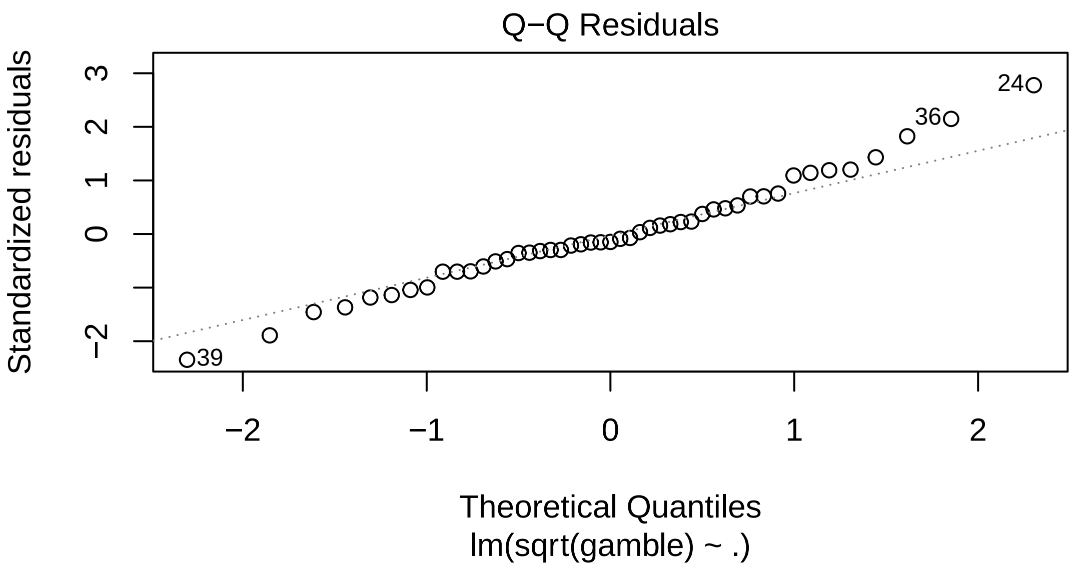



```r
par(mfrow = c(1,1))
```

In the new model, we observe that there still some slight heteroscedasticity in residuals, but not as severe as the original model.

- Check the normality assumption.

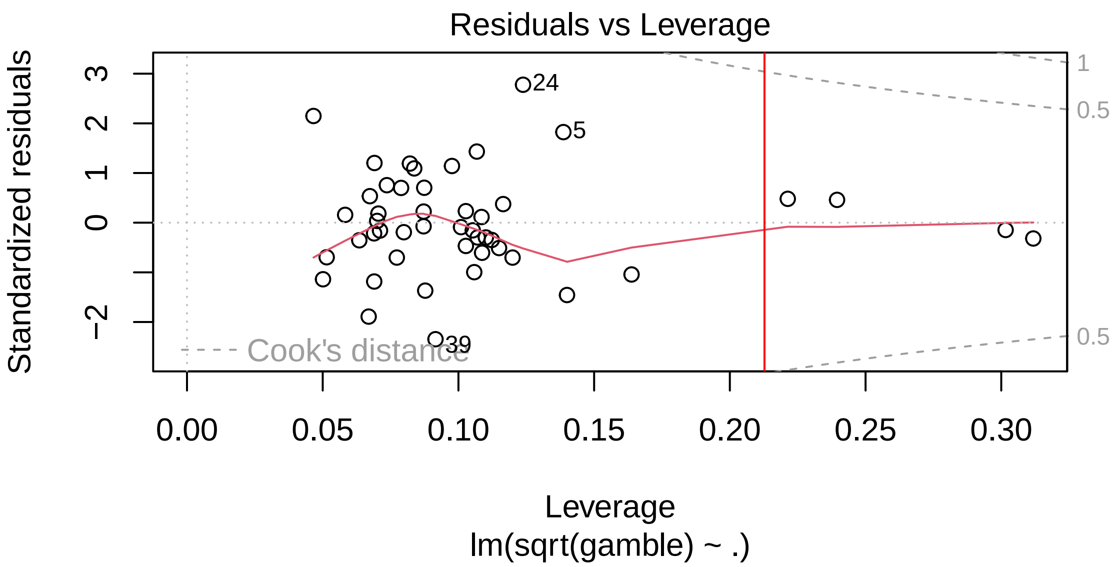
```r
plot(m.new, which=2)
```



From case No. 39,36 and 24, the residuals are slightly heavy-tailed comparing to normal distribution. So, the normality assumption is slightly violated.

- Check for large leverage points.

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



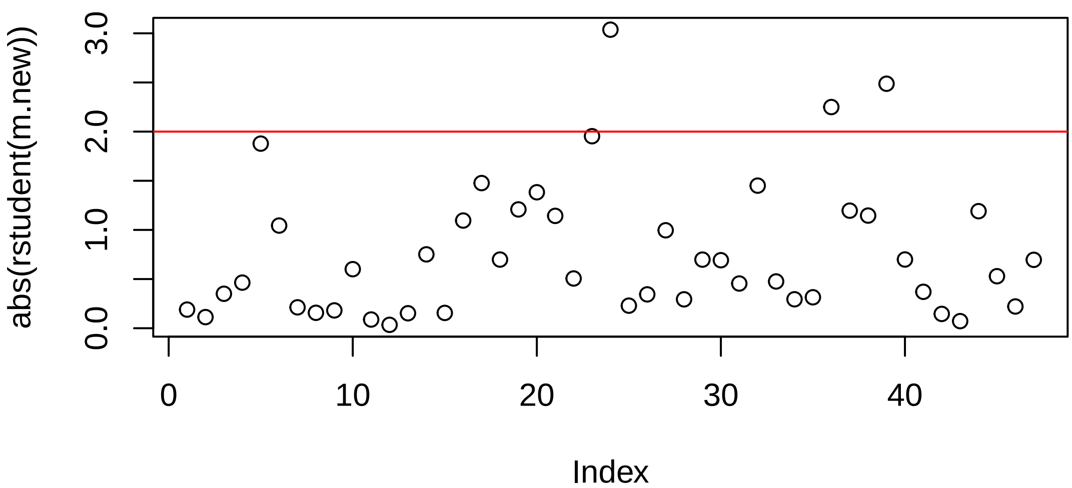
```r
hatvalues(m.ori)>2*length(m.new$coefficients)/nrow(teengamb)
```

```
    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 
```

We add a vertical line $x = \frac{2(p+1)}{n}$ at the `Residuals vs Leverage` plot, we can find 4 points with large leverage. From direct calculation, case No. 31, 33, 35 and 42 are large leverage points.

- Check for outliers.

```r
plot(abs(rstudent(m.new)))
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.

```r
library(car)
```

```
Loading required package: carData


Attaching package: 'car'


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

    logit, vif
```

```r
outlierTest(m.new)
```

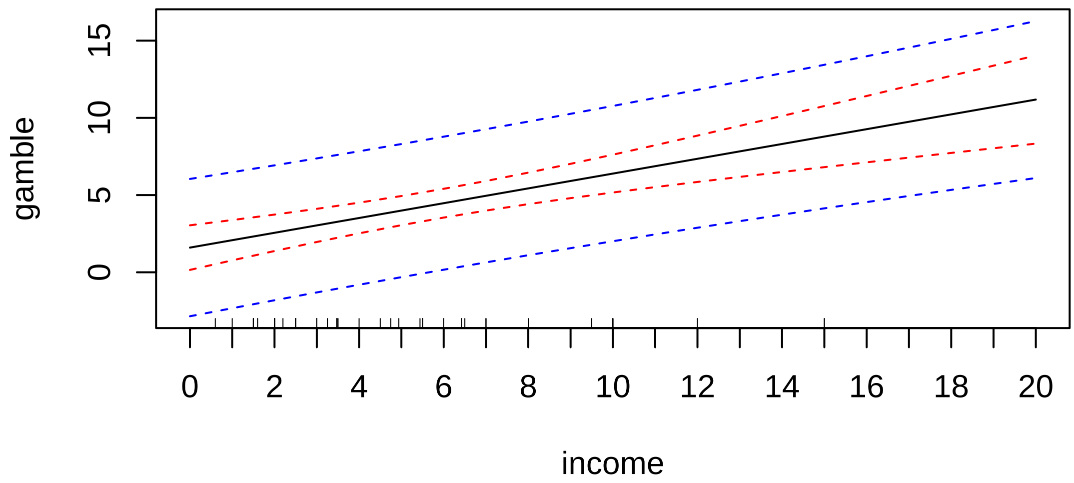
```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferroni p
24 3.037005          0.0041428      0.19471
```

The test result tells that there are no outliers.

(b)

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

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



As for this particular range, 43 and 7 are the median values of `status` and `verbal`, resp. The range of income between 0 to 20 covers all values in the original dataset. Under this setting, we can get a relative narrow confidence/prediction interval since the predictors are near their center in the dataset. In other words, we can get a relative precise result without worrying about issues due to extrapolation.

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 reverse version of result in (a). That is, for a r.v. X, it has 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

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

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

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

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 behind distribution.
By taking the corresponding quantile in Φ^{-1}, if we observe $y - x$ is a line or near a line, then we have the distribution of residual satisfies

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

therefore, we conclude that the residuals follow normal distribution.