Huiwen Chen ID: 02156341 unique name: huiwenc Department of Statistics

**STAT 500 HW4**

**1.Check the constant variance assumption for the errors. Modify the model if necessary (see below).**

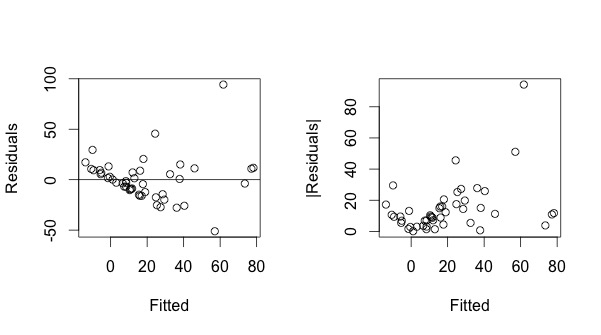


Figure 1

**Code:**

##plot residuals vs fitted values

result1 <- lm(gamble ~ sex+ status+ income+ verbal)

par(mfrow = c(1,2))

plot(result1$fitted, result1$residual, xlab= "Fitted", ylab= "Residuals")

abline(h = 0)

##plot absolute values of residuals vs fitted values

plot(result1$fitted, abs(result1$residual), xlab= "Fitted", ylab= "|Residuals|")

summary(lm(abs(result1$residual) ~ result1$fitted ))

Coefficients (regression of and )

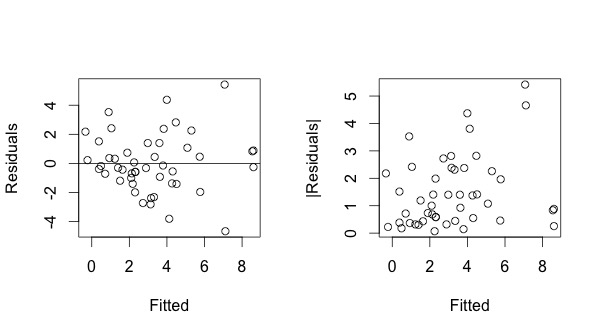
Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.3303 2.8789 3.241 0.00224 \*\*

result1$fitted 0.2645 0.0968 2.732 0.00895 \*\*

In left plot of figure1, we make a regression of and , it shows that the variance of and the scatter are not symmetrically distributed in the vertical direction, but displays a slightly downward trend. So it violates nonlinearity.

In the right plot of figure 2, we make a regression of and , we can see that the scatters are still not vertically symmetric. And since p-value of coefficients of this regression are both less than 0.05, the linear relationship between and is rather significant. So it violates constant variance.

Figure 2

**Codes:**

##square root transformation of y

result2 <- lm(sqrt(gamble) ~ sex+ status+ income+ verbal)

plot(result2$fitted, result2$residual, xlab= "Fitted", ylab= "Residuals")

abline(h = 0)

plot(result2$fitted, abs(result2$residual), xlab= "Fitted", ylab= "|Residuals|")

summary(lm(abs(result2$residual) ~ result2$fitted ))

Coefficients (regression of and )

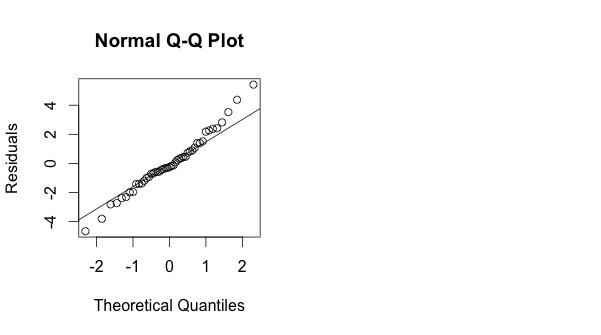
Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.01136 0.32365 3.125 0.00311 \*\*

result2$fitted 0.14957 0.08242 1.815 0.07623 .

In order to offset the influence of non-constant variance, we make a square root transformation of response and then regress and the new response. It shows in figure 2 that, both become more vertically symmetric. And the linear relationship between and is no more significant, as the p-value=0.07623 > 0.05. Then we may consider the variance of as constant.

1. **Check the normality assumption.**

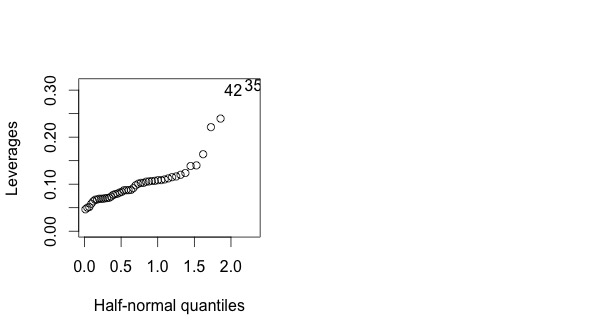


qqnorm (residuals (result2), ylab="Residuals")

qqline (residuals (result2))

In the plot of figure 3, we compare to “ideal” normal observations by Q-Q plot. qqline joining first and third quartiles is not influenced by outliers and approximately, except for slightly heavy tails. So can be considered normal.

1. **Check for large leverage points.**



sex status income verbal gamble

35 0 28 1.5 1 14.1

42 0 61 15.0 9 69.7

Figure 4

Half-normal plot in figure 4 shows that the #35 and #42 points diverge substantially from the rest of data, thus the two points have large leverages.

1. **Check for outliers.**

**Cod**e:

##problem 4

> ## compute (externally) studentized residuals

> ti <- rstudent(result2)

> max(abs(ti))

[1] 3.037005

> which(ti == max(abs(ti)))

24

24

> ## compute p-value

> 2\*(1-pt(max(abs(ti)),df = 47-5-1))

[1] 0.00414277

> ## compare to alpha/n

> 0.05/47

[1] 0.00106383

Since the p-value of the largest (externally) studentized residual is 0.00414277, which is larger than level 0.00106383, we conclude that the #24 point is not an outlier. Then no outlier can be seen in the regression model.

1. **Check for influential points.**

> ## Compute Cook’s distance

> cook <- cooks.distance(result2)

> halfnorm (cook, nlab=4, ylab = "Cook's distances")

> ## Compute changes in coefficients

> result.inf <- lm.influence(result2)

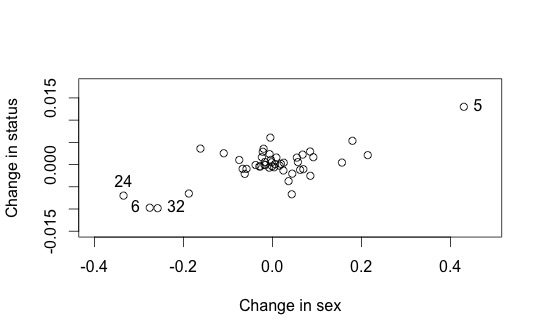
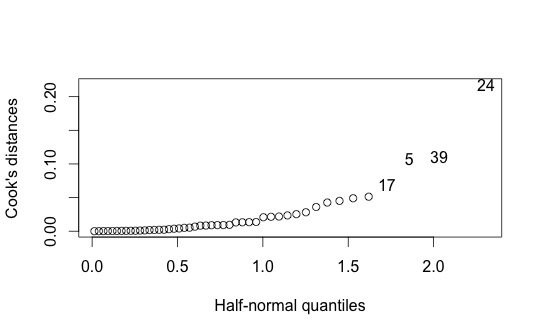
> plot(result.inf$coef[,2], result.inf$coef[,3], xlab="Change in sex", ylab="Change in status", xlim=c(-0.4, 0.48), ylim=c(-0.015, 0.018))

> ## interactive tool to identify points by clicking

> identify (result.inf$coef[, 2], result.inf$coef[, 3])

In the first plot of figure 5, #24, #5 and #39 points have larger cook’s distance from other points. The second plot of figure 5 shows the leaveout-one differences in the coefficients related to sex and status. We find that #24, #5, #32, # 6 points stick out on the plot. Then we examine the effects of removing #24 and #5 points below.

> summary(result2)



> result.24 <- lm(sqrt(gamble) ~ sex+ status+ income+ verbal, data = teengamb,subset = (row.names(teengamb) !="24" ))

> summary(result.24)

Call:

lm(formula = sqrt(gamble) ~ sex + status + income + verbal)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.97707 1.57947 1.885 0.06638 .

sex -2.04450 0.75416 -2.711 0.00968 \*\*

status 0.03688 0.02582 1.428 0.16057

income 0.47938 0.09418 5.090 7.94e-06 \*\*\*

verbal -0.42360 0.19950 -2.123 0.03967 \* Residual standard error: 2.084 on 42 degrees of freedom Multiple R-squared: 0.5646, Adjusted R-squared: 0.5231 F-statistic: 13.61 on 4 and 42 DF, p-value: 3.362e-07

Call:

lm(formula = sqrt(gamble) ~ sex + status + income + verbal, data = teengamb, subset = (row.names(teengamb) != "24"))

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.11915 1.47175 1.440 0.1575

sex -1.70997 0.69840 -2.448 0.0187 \*

status 0.04387 0.02372 1.849 0.0716 .

income 0.44312 0.08695 5.096 8.22e-06 \*\*\*

verbal -0.35706 0.18375 -1.943 0.0589 . Residual standard error: 1.906 on 41 degrees of freedom Multiple R-squared: 0.5503, Adjusted R-squared: 0.5065 F-statistic: 12.55 on 4 and 41 DF, p-value: 9.403e-07

> ##check for #5 point

> result.5 <- lm(sqrt(gamble) ~ sex+ status+ income+ verbal, data = teengamb,subset = (row.names(teengamb) !="5" ))

> summary(result.5)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.56531 1.56567 2.277 0.02806 \*

sex -2.47471 0.76745 -3.225 0.00248 \*\*

status 0.02388 0.02601 0.918 0.36397

income 0.47517 0.09150 5.193 6.01e-06 \*\*\*

verbal -0.40768 0.19395 -2.102 0.04174 \*

Residual standard error: 2.024 on 41 degrees of freedom Multiple R-squared: 0.5976, Adjusted R-squared: 0.5584 F-statistic: 15.22 on 4 and 41 DF, p-value: 1.041e-07

Comparing the data fit without #24 to the full data fit, we notice that the coefficient for sex increases about 15% and the verbal term is no longer significant.

In the data fit without #5, the coefficient for sex decreases about 20%, but the multiple R-squared increased slightly.