STAT 500: HW4

Jasmine Mou

10/12/2017

1. Using the `teengamb` dataset, fit a model with gamble as the response and the other variables as predictors. Answer the questions:

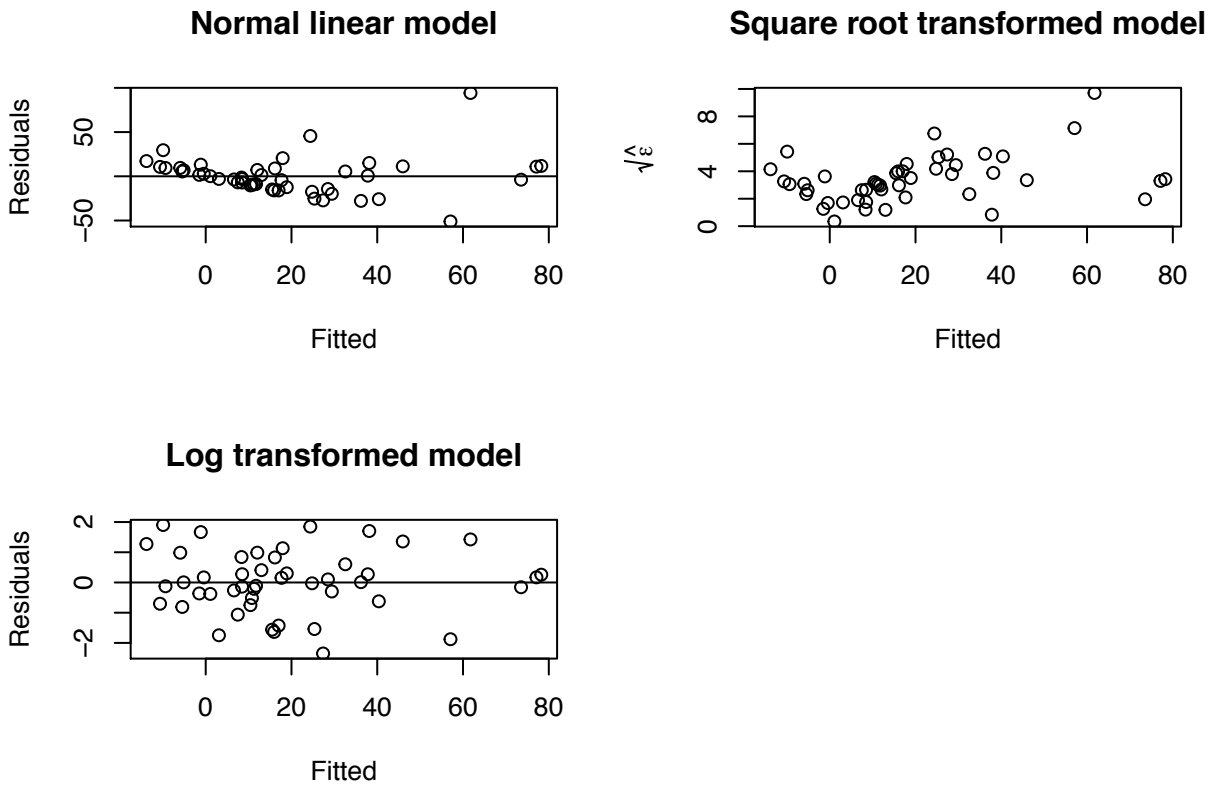
```
data(teengamb, package='faraway')
lm_teengamb <- lm(gamble~., data=teengamb)
```

(a) Check the constant variance assumption for the errors.

```
par(mfrow=c(2,2))
plot(fitted(lm_teengamb), residuals(lm_teengamb), xlab="Fitted", ylab="Residuals", main="Normal linear m
abline(h=0)

plot(fitted(lm_teengamb), sqrt(abs(residuals(lm_teengamb))), xlab="Fitted", ylab=expression(sqrt(hat(eps

lm_teengamb.log <- lm(log(gamble+1)~., data=teengamb)
plot(fitted(lm_teengamb), residuals(lm_teengamb.log), xlab="Fitted", ylab="Residuals", main="Log transfo
abline(h=0)
```

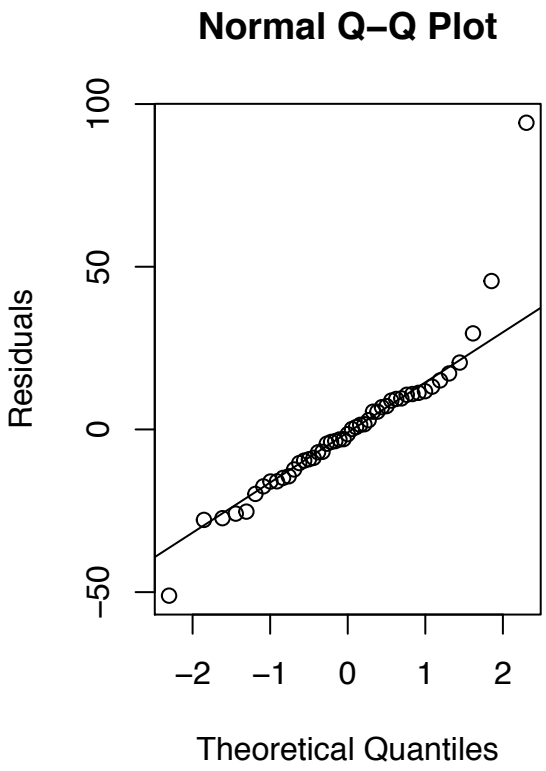


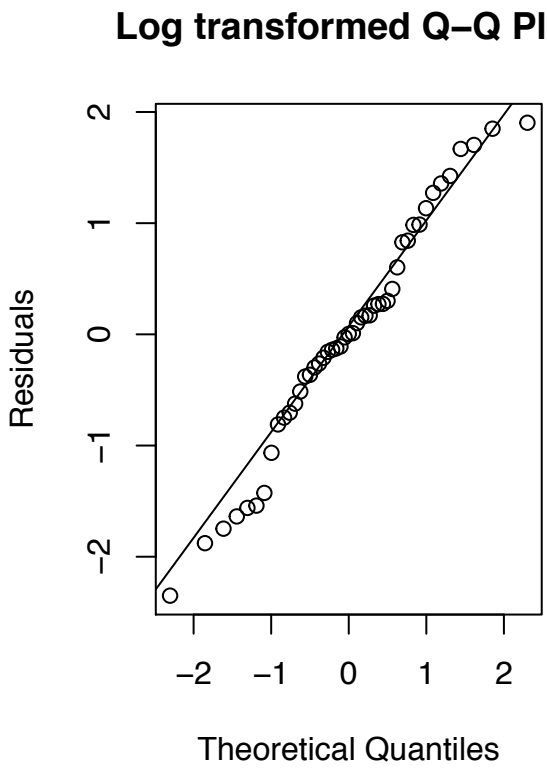
From the plot of normal linear model we can see the variance is not constant. Try two transformations and the log-transformation shows constant variance. Thus we will use the log transformed model from now on.

(b) Check the normality assumption. *Draw a Q-Q plot.*

```
par(mfrow=c(1,2))
qqnorm(residuals(lm_teengamb), ylab="Residuals", main="Normal Q-Q Plot")
qqline(residuals(lm_teengamb))
```

```
qqnorm(residuals(lm_teengamb.log), ylab="Residuals", main="Log transformed Q-Q Plot")
qqline(residuals(lm_teengamb.log))
```





For the first plot, we can observe a long-tailed distribution. The residuals are non normal. From the second plot we can observe a short-tailed distribution, which is not serious and can be ignored. Conduct a Shapiro-Wilk test for each model.

```
shapiro.test(residuals(lm_teengamb))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm_teengamb)
## W = 0.86839, p-value = 8.16e-05
```

```
shapiro.test(residuals(lm_teengamb.log))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm_teengamb.log)
## W = 0.97609, p-value = 0.4418
```

Since the first p-value is significantly small, we have enough evidence to reject the test's null hypothesis that the residuals of the normal linear model are normal. The second p-value > 0.5, meaning we cannot reject the hypothesis that the residuals of log-transformed model are normal. Thus using a log transformaion is appropriate here.

(c) Check for large leverage points.

```
n <- dim(teengamb)[1]
p <- dim(teengamb)[2] - 1
hatv <- hatvalues(lm_teengamb)
```

```
hatv[which(hatv>2*p/n)]
```

```
##        31        33        35        42
## 0.2395031 0.2213439 0.3118029 0.3016088
```

Thus these are the leverages that should be examined more closely.

(d) Check for outliers. *Compute and compare the studentized residuals with the Bonferroni critical value. For the normal model, there is one outlier for observation #24. For the log-transformed model there is no outlier observed.*

```
stud <- rstudent(lm_teengamb)
stud[which(abs(stud) > abs(qt(0.05/(n*2), n-1-p-1)))]
```

```
##       24
## 6.016116
```

```
stud.log <- rstudent(lm_teengamb.log)
stud.log[which(abs(stud.log) > abs(qt(0.05/(n*2), n-1-p-1)))]
```

```
## named numeric(0)
```

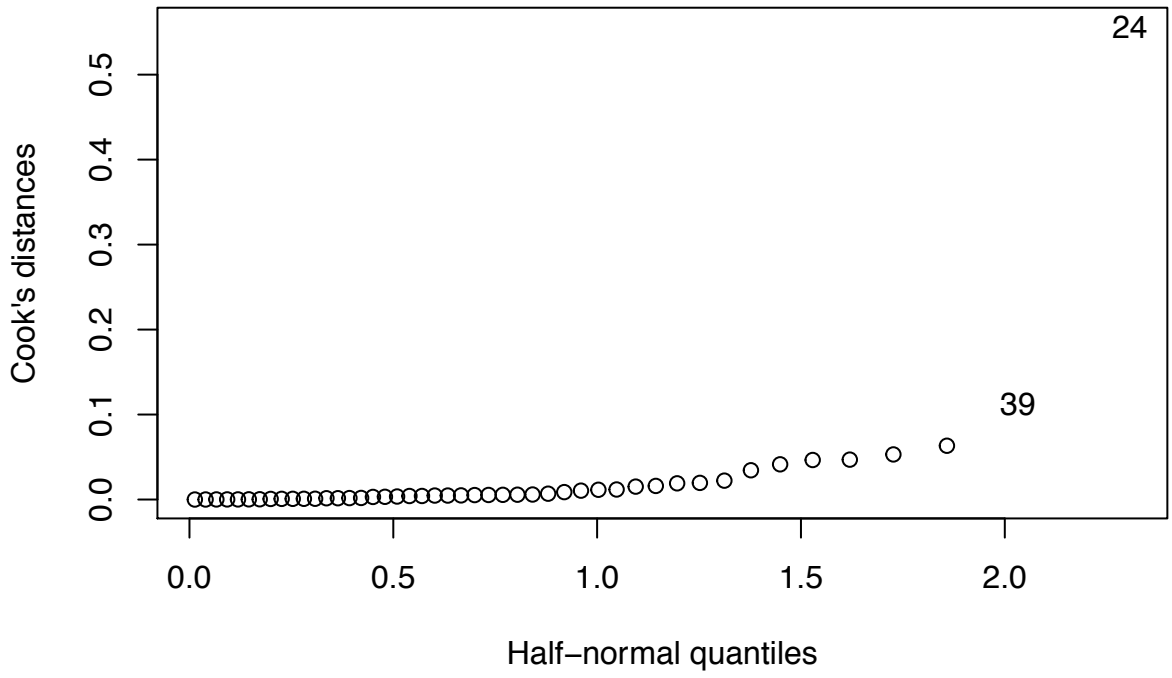
(e) Check for influential points. *Compute Cook statistics to find influential points.*

```
cook <- cooks.distance(lm_teengamb)
cook[which(cook>4/(n-p-1))]
```

```
##        24        39
## 0.5565011 0.1124498
```

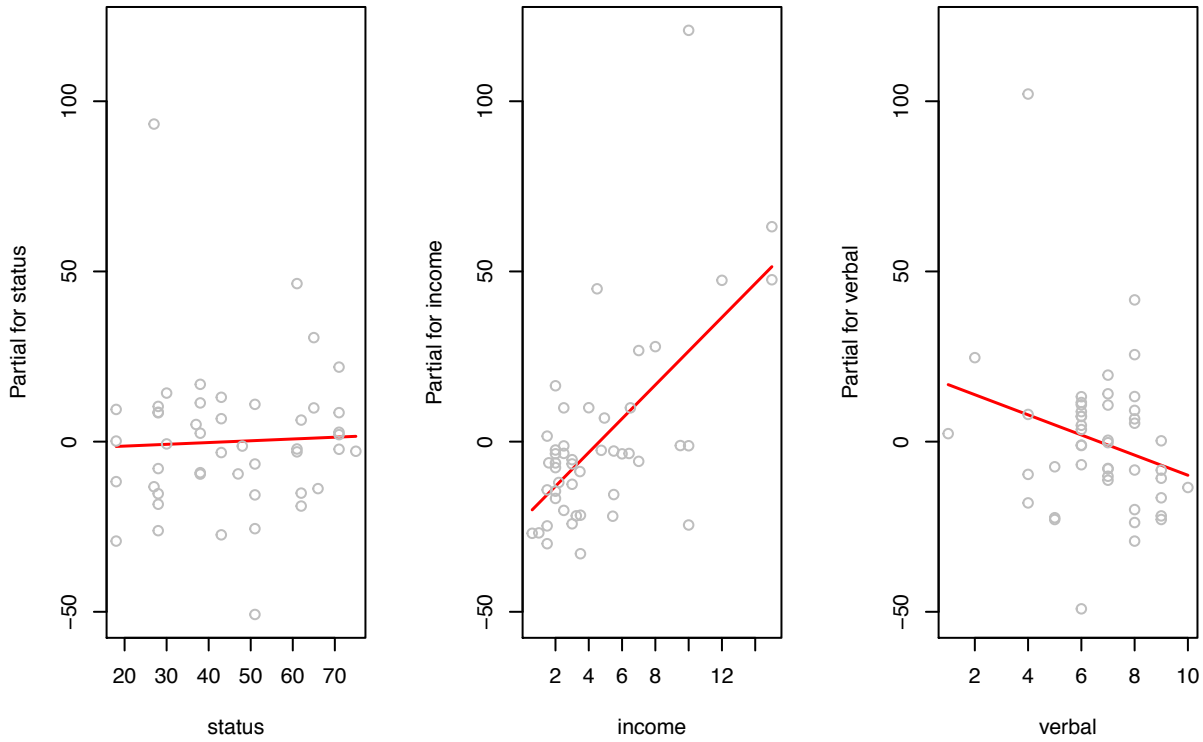
Thus the influential points are observations #24 and 39. Plot the half-normal quantiles plot of the Cook Statistics.

```
faraway::halfnorm(cook,2, ylab="Cook's distances")
```



(f) Check the structure of the relationship between the predictors and the response. *For structure between the predictors and the response, plot partial residual plots.*

```
par(mfrow=c(1,3))
termplot(lm_teengamb, partial.resid = TRUE, terms=c(2,3,4))
```



From the 3 plots, we can see the model doesn't have the severe non-linearity issue.

2. Use the `fat` data, fitting the model described in Section 4.2.

```
data(fat, package='faraway')
lm_fat <- lm(brozek ~ age + weight + height + neck + chest + abdom +
hip + thigh + knee + ankle + biceps + forearm + wrist, data=fat)
n <- dim(fat)[1]
p <- dim(fat)[2] - 1
```

(a) Compute the condition numbers and variance inflation factors. Comment on the degree of collinearity observed in the data.

```
compute_col <- function (x){
  e <- eigen(t(x) %*% x)
  p <- dim(x)[2]
  kappa <- sqrt(e$val[1]/e$val[p]) # condition numbers
  vifs <- faraway::vif(x) # VIF
  col <- list("condition_num" = kappa, "vif" = vifs)
  return(col)
}

x <- model.matrix(lm_fat)[,-1]
col = compute_col(x)
col$condition_num
```

```
## [1] 555.6707
```

```
col$vif
```

```
##       age    weight    height      neck     chest     abdom       hip
##  2.250450 33.509320  1.674591  4.324463  9.460877 11.767073 14.796520
##     thigh      knee     ankle    biceps   forearm     wrist
##  7.777865  4.612147  1.907961  3.619744  2.192492  3.377515
```

```
col$vif[which(col$vif>10)]
```

```
##   weight    abdom      hip
## 33.50932 11.76707 14.79652
```

The condition numbers is pretty large (>30). There are 3 large VIFs (>10) too. Thus there is much variance deflation.

(b) Cases 39 and 42 are unusual. Refit the model without these two cases and recompute the collinearity diagnostics. Comment on the differences observed from the full data fit.

```
lm_fat_b <- lm(brozek ~ age + weight + height + neck + chest + abdom + hip + thigh + knee + ankle + bic
x_b <- model.matrix(lm_fat_b)[,-1]
col_b = compute_col(x_b)
col_b$condition_num
```

```
## [1] 554.7978
```

```
col_b$vif
```

```
##       age    weight    height      neck     chest     abdom       hip
##  2.278191 45.298843  3.439587  3.978898 10.712505 11.967580 12.146249
##     thigh      knee     ankle    biceps   forearm     wrist
##  7.153711  4.441752  1.810253  3.409524  2.422878  3.263677
```

```
col_b$vif[which(col_b$vif>10)]
```

```
##   weight    chest    abdom      hip
## 45.29884 10.71251 11.96758 12.14625
```

The condition number is reduced a little bit (from 555.6707 to 554.7978). Yet there is one more large VIF compared to the previous model. Thus the refit doesn't reduce the collinearity issue.

(c) Fit a model with `brozek` as the response and just `age`, `weight` and `height` as predictors. Compute the collinearity diagnostics and compare to the full data fit.

```
lm_fat_c <- lm(brozek ~ age + weight + height, data=fat)
x_c <- model.matrix(lm_fat_c)[,-1]
col_c = compute_col(x_c)
col_c$condition_num
```

```
## [1] 22.6725
```

```
col_c$vif
```

```
##      age   weight   height
## 1.032253 1.107050 1.140470
```

```
col_c$vif[which(col_c$vif>10)]
```

```
## named numeric(0)
```

The condition number is reduced greatly and below 30. The VIFs are all pretty small now (<10). Thus the collinearity issue has been solved with this model compared to the full model. Let's also check the correlations of the length variables.

```
round(cor(x_c), 2)
```

```
##          age weight height
## age     1.00  -0.01  -0.17
## weight -0.01   1.00   0.31
## height -0.17   0.31   1.00
```

From the result we can see ***height*** *somehow correlates positively to* ***weight****.*

(d) Compute a 95% prediction interval for `brozek` for the median values of `age`, `weight` and `height`.

```
x_c_median <- apply(x_c,2,median)
pi_d <- predict(lm_fat_c, new=data.frame(t(x_c_median)), interval="prediction", level=0.95)
round(pi_d,2)
```

```
##     fit  lwr  upr
## 1 18.28 7.66 28.9
```

Thus the 95% P.I. for ***brozek*** *is (7.6596088, 28.9030384).*

(e) Compute a 95% prediction interval for `brozek` for `age`=40, `weight`=200 and `height`=73. How does the interval compare to the previous prediction?

```
y_e <- c(40,200,73)
names(y_e) <- c("age", "weight", "height")
pi_e <- predict(lm_fat_c, new=data.frame(t(y_e)), interval="prediction", level=0.95)
pi_e
```

```
##        fit      lwr      upr
## 1 20.47854 9.837784 31.11929
```

The 95% P.I. for ***brozek*** *for* ***age****=40,* ***weight****=200 and* ***height****=73 is (9.8377842, 31.1192922), which has a similar length of range but higher center compared to the previous prediction interval.*

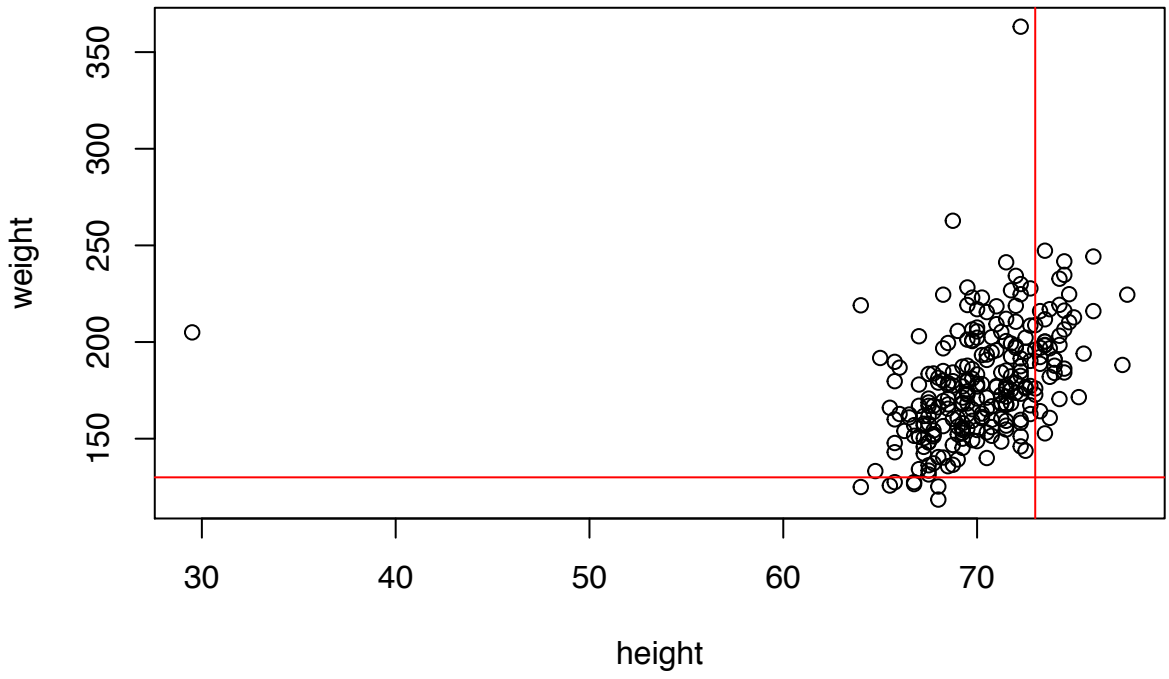
(f) Compute a 95% prediction interval for `brozek` for `age`=40, `weight`=130 and `height`=73. Are the values of predictors unusual? Comment on how the interval compares to the previous two answers.

```
y_f <- c(40,130,73)
names(y_f) <- c("age", "weight", "height")
pi_f <- predict(lm_fat_c, new=data.frame(t(y_f)), interval="prediction", level=0.95)
pi_f
```

```
##        fit       lwr     upr
## 1 7.617419 -3.101062 18.3359
```

The 95% P.I. for ***brozek*** *for* ***age****=40,* ***weight****=130 and* ***height****=73 is (0, 18.3358998. To discuss the predictors let's draw a plot for* ***weight*** *and* ***height*** *to check their relationship.*

```
plot(fat$height, fat$weight, xlab="height", ylab="weight")
abline(h=130, v=73, col='red')
```



The cross of horizontal and vertical line is the `weight` *and* `height` *predictors for this person, and we can see there are almost no points if the* `weight` *is lower and* `height` *is higher. Thus this combination of predictors looks unusual. The length of interval range is still similar to the previous two answers.*

Attachment: RMarkdown Codes

````
---
title: "STAT 500: HW4"
author: "Jasmine Mou"
date: "10/12/2017"
output:
  pdf_document: default
---
1. Using the `teengamb` dataset, fit a model with gamble as the response and the other variables as predictors. Answer the questions:
```{r}
data(teengamb, package='faraway')
lm_teengamb <- lm(gamble~., data=teengamb)
```
(a) Check the constant variance assumption for the errors.
```{r}
par(mfrow=c(2,2))
plot(fitted(lm_teengamb), residuals(lm_teengamb), xlab="Fitted",
ylab="Residuals", main="Normal linear model")
abline(h=0)

plot(fitted(lm_teengamb), sqrt(abs(residuals(lm_teengamb))),
xlab="Fitted", ylab=expression(sqrt(hat(epsilon))), main="Square root
transformed model")

lm_teengamb.log <- lm(log(gamble+1)~., data=teengamb)
plot(fitted(lm_teengamb), residuals(lm_teengamb.log), xlab="Fitted",
ylab="Residuals", main="Log transformed model")
abline(h=0)
```
*From the plot of normal linear model we can see the variance is not constant. Try two transformations and the log-transformation shows constant variance. Thus we will use the log transformed model from now on.*

(b) Check the normality assumption.
*Draw a Q-Q plot.*
```{r}
par(mfrow=c(1,2))
qqnorm(residuals(lm_teengamb), ylab="Residuals", main="Normal Q-Q Plot")
qqline(residuals(lm_teengamb))
qqnorm(residuals(lm_teengamb.log), ylab="Residuals", main="Log
transformed Q-Q Plot")
qqline(residuals(lm_teengamb.log))
```
*For the first plot, we can observe a long-tailed distribution. The residuals are non normal. From the second plot we can observe a short-tailed distribution, which is not serious and can be ignored. Conduct a Shapiro-Wilk test for each model.*
````

```{r}
shapiro.test(residuals(lm_teengamb))
shapiro.test(residuals(lm_teengamb.log))
```
Since the first p-value is significantly small, we have enough evidence to reject the test's null hypothesis that the residuals of the normal linear model are normal. The second p-value > 0.5, meaning we cannot reject the hypothesis that the residuals of log-transformed model are normal. Thus using a log transformaion is appropriate here.

(c) Check for large leverage points.
```{r}
n <- dim(teengamb)[1]
p <- dim(teengamb)[2] - 1
hatv <- hatvalues(lm_teengamb)
hatv[which(hatv>2*p/n)]
```
Thus these are the leverages that should be examined more closely.

(d) Check for outliers.
Compute and compare the studentized residuals with the Bonferroni critical value. For the normal model, there is one outlier for observation #24. For the log-transformed model there is no outlier observed.
```{r}
stud <- rstudent(lm_teengamb)
stud[which(abs(stud) > abs(qt(0.05/(n*2), n-1-p-1)))]
stud.log <- rstudent(lm_teengamb.log)
stud.log[which(abs(stud.log) > abs(qt(0.05/(n*2), n-1-p-1)))]
```

(e) Check for influential points.
Compute Cook statistics to find influential points.
```{r}
cook <- cooks.distance(lm_teengamb)
cook[which(cook>4/(n-p-1))]
```
Thus the influential points are observations #24 and 39. Plot the half-normal quantiles plot of the Cook Statistics.
```{r}
faraway::halfnorm(cook,2, ylab="Cook's distances")
```

(f) Check the structure of the relationship between the predictors and the response.
*For structure between the predictors and the response, plot partial residual plots. *
```{r}
par(mfrow=c(1,3))
termplot(lm_teengamb, partial.resid = TRUE, terms=c(2,3,4))
```

```
*From the 3 plots, we can see the model doesn't have the severe non-linearity issue.*


2. Use the `fat` data, fitting the model described in Section 4.2.
```{r}
data(fat, package='faraway')
lm_fat <- lm(brozek ~ age + weight + height + neck + chest + abdom + hip + thigh + knee + ankle + biceps + forearm + wrist, data=fat)
n <- dim(fat)[1]
p <- dim(fat)[2] - 1
```
(a) Compute the condition numbers and variance inflation factors. Comment on the degree of collinearity observed in the data.
```{r}
compute_col <- function (x){
  e <- eigen(t(x) %*% x)
  p <- dim(x)[2]
  kappa <- sqrt(e$val[1]/e$val[p]) # condition numbers
  vifs <- faraway::vif(x) # VIF
  col <- list("condition_num" = kappa, "vif" = vifs)
  return(col)
}

x <- model.matrix(lm_fat)[,-1]
col = compute_col(x)
col$condition_num
col$vif
col$vif[which(col$vif>10)]
```
The condition numbers is pretty large (>30). There are 3 large VIFs (>10) too. Thus there is much variance deflation.

(b) Cases 39 and 42 are unusual. Refit the model without these two cases and recompute the collinearity diagnostics. Comment on the differences observed from the full data fit.
```{r}
lm_fat_b <- lm(brozek ~ age + weight + height + neck + chest + abdom + hip + thigh + knee + ankle + biceps + forearm + wrist, data=fat, subset=c(-39,-42))
x_b <- model.matrix(lm_fat_b)[,-1]
col_b = compute_col(x_b)
col_b$condition_num
col_b$vif
col_b$vif[which(col_b$vif>10)]
```
*The condition number is reduced a little bit (from 555.6707 to 554.7978). Yet there is one more large VIF compared to the previous model. Thus the refit doesn't reduce the collinearity issue. *
```

(c) Fit a model with `brozek` as the response and just `age`, `weight` and `height` as predictors. Compute the collinearity diagnostics and compare to the full data fit.

```{r}
lm_fat_c <- lm(brozek ~ age + weight + height, data=fat)
x_c <- model.matrix(lm_fat_c)[,-1]
col_c = compute_col(x_c)
col_c$condition_num
col_c$vif
col_c$vif[which(col_c$vif>10)]
```

*The condition number is reduced greatly and below 30. The VIFs are all pretty small now (<10). Thus the collinearity issue has been solved with this model compared to the full model. Let's also check the correlations of the length variables. *

```{r}
round(cor(x_c), 2)
```

From the result we can see `height` somehow correlates positively to `weight`.

(d) Compute a 95% prediction interval for `brozek` for the median values of `age`, `weight` and `height`.

```{r}
x_c_median <- apply(x_c,2,median)
pi_d <- predict(lm_fat_c, new=data.frame(t(x_c_median)),
interval="prediction", level=0.95)
round(pi_d,2)
```

*Thus the 95% P.I. for `brozek` is (`r pi_d[2]`, `r pi_d[3]`). *

(e) Compute a 95% prediction interval for `brozek` for `age`=40, `weight`=200 and `height`=73. How does the interval compare to the previous prediction?

```{r}
y_e <- c(40,200,73)
names(y_e) <- c("age", "weight", "height")
pi_e <- predict(lm_fat_c, new=data.frame(t(y_e)), interval="prediction",
level=0.95)
pi_e
```

The 95% P.I. for `brozek` for `age`=40, `weight`=200 and `height`=73 is (`r pi_e[2]`, `r pi_e[3]`), which has a similar length of range but higher center compared to the previous prediction interval.

(f) Compute a 95% prediction interval for `brozek` for `age`=40, `weight`=130 and `height`=73. Are the values of predictors unusual? Comment on how the interval compares to the previous two answers.

```{r}
y_f <- c(40,130,73)
names(y_f) <- c("age", "weight", "height")
```

```
pi_f <- predict(lm_fat_c, new=data.frame(t(y_f)), interval="prediction",
level=0.95)
pi_f
```
*The 95% P.I. for `brozek` for `age`=40, `weight`=130 and `height`=73 is
(0, `r pi_f[3]`. To discuss the predictors let's draw a plot for
`weight` and `height` to check their relationship. *
```{r}
plot(fat$height, fat$weight, xlab="height", ylab="weight")
abline(h=130, v=73, col='red')
```
*The cross of horizontal and vertical line is the `weight` and `height`
predictors for this person, and we can see there are almost no points if
the `weight` is lower and `height` is higher. Thus this combination of
predictors looks unusual. The length of interval range is still similar
to the previous two answers.*