HW6

Andrew Shao

2024-11-06

(a)

1. The model is:

$$brozek = -46.216 + 0.646 \cdot chest$$

Since the p-value is extremely small ($7.373 \cdot 10^{-39}$), we conclude that these two variables are associated.

2. Pearson's correlation coefficient is 0.703 between the two variables.
The calculated p-value for T_{obs} is extremely small and close to 0, so we conclude that the two variables are significantly correlated at the $\alpha = 0.05$ level. The value is the same as the value for the slope.
The 95% confidence interval is [0.6344, 0.7604]

3. Spearman's correlation coefficient = 0.6731
Since the p-value is extremely close to 0, we can conlude that the two variables are significantly correlated at the $\alpha = 0.05$ level.

4. The two variables are correlated and the association between the two variables is positive and strong.

(b)

The model is:

$$brozek = -40.599 + 1.567 \cdot neck$$

Since the p-value is extremely small ($9.904 \cdot 10^{-17}$), we conclude that these two variables are associated.

Pearson's correlation coefficient is 0.491 between the two variables.
The calculated p-value for T_{obs} is extremely small and close to 0, so we conclude that the two variables are significantly correlated at the $\alpha = 0.05$ level. The value is the same as the value for the slope.
The 95% confidence interval is [0.3917, 0.5798]

Spearman's correlation coefficient = 0.491
Since the p-value is extremely close to 0, we can conlude that the two variables are significantly correlated at the $\alpha = 0.05$ level.

The two variables are correlated and the association between the two variables is positive and weaker than between `brozek` and `chest`.

(c)

I think `chest` has a stronger association because it is highly significant and the R-squared value is larger and both the pearson and spearman correlation coefficients are much larger than those for `neck`.

(d)

1. The model has 13 predictors: `age`, `weight`, `height`, `neck`, `chest`, `abdom`, `hip`, `thigh`, `knee`, `ankle`, `biceps`, `forearm`, and `wrist`. The four significant predictors at the $\alpha = 0.05$ level are `neck`, `abdom`, `forearm`, and `wrist`. The multiple R-squared value is 0.749 and the adjusted R-squared value is 0.7353. The F-statistic is 54.63 on 13 and 238 degrees of freedom with a p-value less than $2.2 \cdot 10^{-16}$.

2. The Breusch-Pagan p-value is 0.1107, meaning that there is not enough evidence to reject the null hypothesis at the $\alpha = 0.05$ level suggesting that the constant variance assumption holds.
The Q-Q plot residuals seem to generally follow the line which suggests that the residuals are indeed normally distributed. The Shapiro-Wilk normality test p-value is 0.2801 meaning there is not enough evidence to reject the null hypothesis at the $\alpha = 0.05$ level suggesting that the residual normality assumption holds.

3. In this specific model, `neck` is significantly associated with `brozek` while `chest` is not, which differs from my answer in (c) that `chest` had the stronger association. This could be due to collinearity between the predictors including `neck` and `chest`.

(e)

1. Many of the correlation coefficients are very close to 1 or -1 which suggests collinearity between predictors, which affects the significance of individual predictors within the model.

2. Most of the condition numbers are greater than 30 which suggests a high level of collinearity in the model which could cause imprecision in the estimation of β.

3. The VIF values that are greater than 10 are `weight`, `abdom`, `hip`, and `chest` which confirms the theory that collinearity is present among the variables.

(f)

1. Minimum: 0
Q1: 12.8
Median: 19
Mean: 18.94
Q3: 24.6
Maximum: 45.1
Standard deviation: 7.75

2. The data looks mostly normal, but the Q-Q plot deviates from the normal line at the tails which could indicate non-normality.

3. The Shapiro-Wilk p-value is 0.2747 meaning there is not enough evidence to reject the null hypothesis at the $\alpha = 0.05$ level suggesting that `brozek` is normal.

4. An error occurs because the some values for the response variable aren't positive (in this case, values of 0).

5. The plot shows that a lambda value close to 1 is within the 95% confidence interval, suggesting transformation might not be necessary since it is close to 1.

6. I would not recommend transformation since the Q-Q plot and Box-Cox plot do not suggest a need to transform; the best fit lambda is close to 1 and the data seems to be normal.

(g)

1. Of the fourteen predictors in the model, `age`, `abdom`, `abdom` squared, and `wrist` are significant. The R-squared value is 0.7573 and the F-statistic p-value is extremely small.

2. Of the fifteen predictors in the model, only `age` and `wrist` are significant. The R-squared value is 0.758 and the F-statistic p-value is extremely small.

3. I think adding the quadratic term was beneficial since it increased both R-squared and adjusted R-squared values slightly with the added term being statistically significant at the $\alpha = 0.05$ level. Adding the cubic term didn't change the R-squared values much and instead made both all the `abdom` terms insignificant. It seems the quadratic term captured some non-linearity within the data while adding the cubic term seem to add excess complexity to the model which could cause it to overfit.

Appendix

(a)

```r
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 4.3.3
```

```r
head(fat)
```

```
##   brozek siri density age weight height adipos  free neck chest abdom   hip
## 1   12.6 12.3  1.0708  23 154.25  67.75   23.7 134.9 36.2  93.1  85.2  94.5
## 2    6.9  6.1  1.0853  22 173.25  72.25   23.4 161.3 38.5  93.6  83.0  98.7
## 3   24.6 25.3  1.0414  22 154.00  66.25   24.7 116.0 34.0  95.8  87.9  99.2
## 4   10.9 10.4  1.0751  26 184.75  72.25   24.9 164.7 37.4 101.8  86.4 101.2
## 5   27.8 28.7  1.0340  24 184.25  71.25   25.6 133.1 34.4  97.3 100.0 101.9
## 6   20.6 20.9  1.0502  24 210.25  74.75   26.5 167.0 39.0 104.5  94.4 107.8
##   thigh knee ankle biceps forearm wrist
## 1  59.0 37.3  21.9   32.0    27.4  17.1
## 2  58.7 37.3  23.4   30.5    28.9  18.2
## 3  59.6 38.9  24.0   28.8    25.2  16.6
## 4  60.1 37.3  22.8   32.4    29.4  18.2
## 5  63.2 42.2  24.0   32.2    27.7  17.7
## 6  66.0 42.0  25.6   35.7    30.6  18.8
```

```r
## (a)(1) simple linear regression
lma <- lm(brozek ~ chest, data=fat);
summary(lma)
```

```
## 
## Call:
## lm(formula = brozek ~ chest, data = fat)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -13.8875  -3.8211  -0.2752   3.4950  13.8989 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -46.21636    4.18460  -11.04   <2e-16 ***
## chest         0.64622    0.04136   15.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.524 on 250 degrees of freedom
## Multiple R-squared:  0.494,  Adjusted R-squared:  0.492
## F-statistic: 244.1 on 1 and 250 DF,  p-value: < 2.2e-16
```

```
## Extract those information for the slope
summary(lma)$coefficients[2,]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 6.462223e-01 4.136018e-02 1.562426e+01 7.372549e-39
```

```
##  or the t-statistics and p-value for the slope
##  t-stat=1.562426e+01= 15.62426, p-value= 7.372549e-39
summary(lma)$coefficients[2,3:4]
```

```
##      t value     Pr(>|t|)
## 1.562426e+01 7.372549e-39
```

```
## (a)(2) Pearson's correlation
# (i) Pearson's correlation
r1 = cor(fat$brozek, fat$chest);
r1
```

```
## [1] 0.7028852
```

```
# (ii) hypothesis testing via Pearson's correlation
n= dim(fat)[1];
t.obs1 = r1* sqrt((n-2)/ (1-r1^2) );
t.obs1  ### compare with (i)
```

```
## [1] 15.62426
```

```
# p-value
pvalue1 = 2*(1-pt( abs(t.obs1), df= n-2 ));
pvalue1
```

```
## [1] 0
```

```
# (iii) 95% CI on Pearson's correlation
alpha = 0.05;
cutoffvalue = qnorm(1- alpha/2);
Zr1 = 0.5*log((1+r1)/(1-r1));
ZCI = Zr1 + c(-1, 1)* cutoffvalue /sqrt(n-3);
rho1.CI = (exp(2*ZCI) -1) / (exp(2*ZCI) +1);
rho1.CI
```

```
## [1] 0.6344161 0.7604106
```

```
### (a)(3) Spearman's Correlation
## (i) point estimate
rs1= cor(fat$brozek, fat$chest, method= "spearman");
rs1
```

```
## [1] 0.6730803
```

```r
## (ii) hypothesis testing
n= dim(fat)[1];
t.obs2 = rs1* sqrt((n-2)/ (1-rs1^2) );
t.obs2
```

```
## [1] 14.38991
```

```r
# p-value based on Spearman's correlation
pvalue2 = 2*(1-pt( abs(t.obs2), df= n-2 ));
pvalue2
```

```
## [1] 0
```

(b)

```r
# (b)(1) Simple Linear Regression Model for `brozek` and `neck`
lmb <- lm(brozek ~ neck, data=fat)
summary(lmb)
```

```
##
## Call:
## lm(formula = brozek ~ neck, data = fat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0076  -4.9450  -0.2405   5.0321  21.1344
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -40.5985     6.6857  -6.072 4.66e-09 ***
## neck          1.5671     0.1756   8.923  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.764 on 250 degrees of freedom
## Multiple R-squared:  0.2416, Adjusted R-squared:  0.2385
## F-statistic: 79.62 on 1 and 250 DF,  p-value: < 2.2e-16
```

```r
# Extract information for the slope of `neck`
summary(lmb)$coefficients[2,]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.567090e+00 1.756186e-01 8.923259e+00 9.904320e-17
```

```r
# t-statistics and p-value for the slope
summary(lmb)$coefficients[2, 3:4]
```

```
##      t value     Pr(>|t|)
## 8.923259e+00 9.904320e-17
```

```r
# (b)(2) Pearson's Correlation
# (i) Pearson's correlation coefficient
r2 <- cor(fat$brozek, fat$neck)
r2
```

```
## [1] 0.4914889
```

```
# (ii) Hypothesis testing for Pearson's correlation
n <- dim(fat)[1]
t.obs2 <- r2 * sqrt((n - 2) / (1 - r2^2))
t.obs2
```

```
## [1] 8.923259
```

```
# p-value
pvalue2 <- 2 * (1 - pt(abs(t.obs2), df = n - 2))
pvalue2
```

```
## [1] 0
```

```
# (iii) 95% Confidence Interval for Pearson's correlation
alpha <- 0.05
cutoffvalue <- qnorm(1 - alpha / 2)
Zr2 <- 0.5 * log((1 + r2) / (1 - r2))
ZCI2 <- Zr2 + c(-1, 1) * cutoffvalue / sqrt(n - 3)
rho2.CI <- (exp(2 * ZCI2) - 1) / (exp(2 * ZCI2) + 1)
rho2.CI
```

```
## [1] 0.3917062 0.5798451
```

```
# (b)(3) Spearman's Correlation
# (i) Spearman's correlation coefficient
rs2 <- cor(fat$brozek, fat$neck, method = "spearman")
rs2
```

```
## [1] 0.4913248
```

```
# (ii) Hypothesis testing for Spearman's correlation
t.obs3 <- rs2 * sqrt((n - 2) / (1 - rs2^2))
t.obs3
```

```
## [1] 8.919331
```

```
# p-value for Spearman's correlation
pvalue3 <- 2 * (1 - pt(abs(t.obs3), df = n - 2))
pvalue3
```

```
## [1] 0
```

(d)

```
#######Part (d) ##########
###  Lily's Model
modLily <- lm(brozek ~ age + weight+ height+ neck+ chest + abdom+ hip+
                thigh+ knee+ ankle+ biceps+ forearm+ wrist, data=fat);
summary(modLily)
```

```
##
## Call:
## lm(formula = brozek ~ age + weight + height + neck + chest +
##     abdom + hip + thigh + knee + ankle + biceps + forearm + wrist,
##     data = fat)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.264  -2.572  -0.097   2.898   9.327
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.29255   16.06992  -0.952  0.34225
## age           0.05679    0.02996   1.895  0.05929 .
## weight       -0.08031    0.04958  -1.620  0.10660
## height       -0.06460    0.08893  -0.726  0.46830
## neck         -0.43754    0.21533  -2.032  0.04327 *
## chest        -0.02360    0.09184  -0.257  0.79740
## abdom         0.88543    0.08008  11.057  < 2e-16 ***
## hip          -0.19842    0.13516  -1.468  0.14341
## thigh         0.23190    0.13372   1.734  0.08418 .
## knee         -0.01168    0.22414  -0.052  0.95850
## ankle         0.16354    0.20514   0.797  0.42614
## biceps        0.15280    0.15851   0.964  0.33605
## forearm       0.43049    0.18445   2.334  0.02044 *
## wrist        -1.47654    0.49552  -2.980  0.00318 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.988 on 238 degrees of freedom
## Multiple R-squared:  0.749,  Adjusted R-squared:  0.7353
## F-statistic: 54.63 on 13 and 238 DF,  p-value: < 2.2e-16
```

```
### model diagnostic
### (i) check equal variance assumption
library("lmtest")
```

```
## Warning: package 'lmtest' was built under R version 4.3.2
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

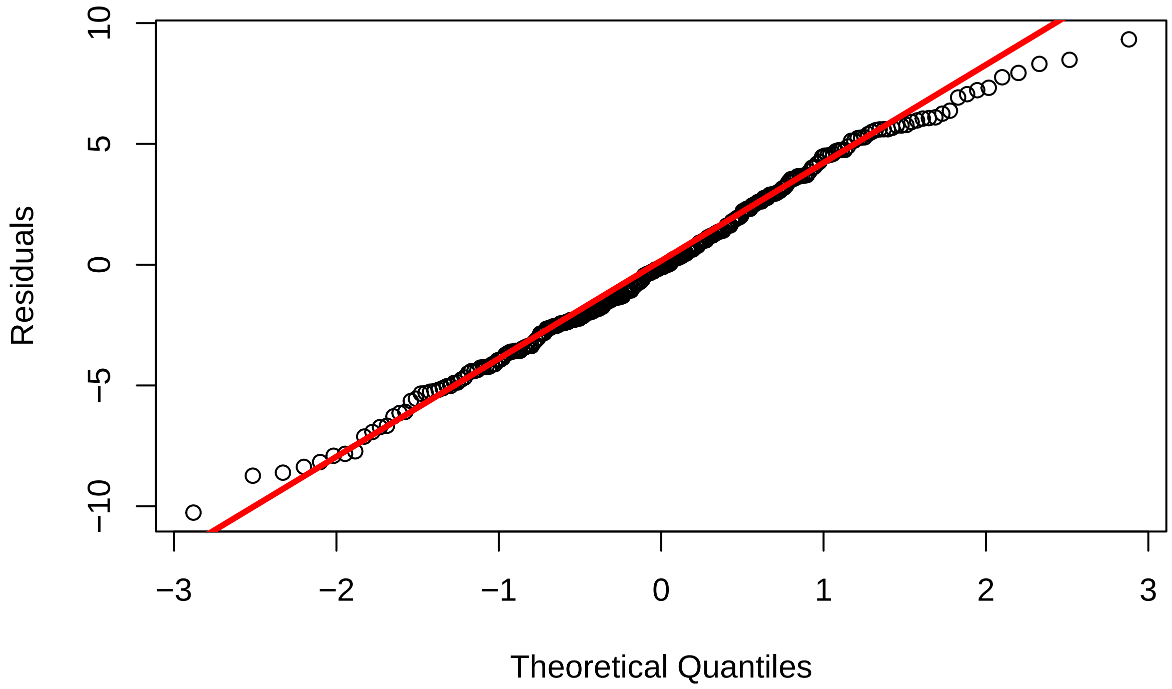
```
bptest(modLily)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modLily
## BP = 19.418, df = 13, p-value = 0.1107
```

```
### If you want to go above and beyond,
##   you can check the p-value of F-test when regressing
##   the absolute value of residuals on all X variables again
## If p-value >= 5%, then it is okay to accept the equal variance assumption
## If p-value < 5%, then we can run the weighted least square regression
lm.weight <- lm( abs(residuals(modLily))  ~ age + weight+ height+ neck+ chest + abdom+
↪  hip+
           thigh+ knee+ ankle+ biceps+ forearm+ wrist, data=fat);
summary(lm.weight)
```

```
##
## Call:
## lm(formula = abs(residuals(modLily)) ~ age + weight + height +
##     neck + chest + abdom + hip + thigh + knee + ankle + biceps +
##     forearm + wrist, data = fat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.781 -1.730 -0.285  1.393  6.470
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.09069    8.80802  -1.146   0.2531
## age           0.01827    0.01642   1.112   0.2671
## weight       -0.04476    0.02718  -1.647   0.1009
## height        0.06278    0.04874   1.288   0.1990
## neck          0.25208    0.11803   2.136   0.0337 *
## chest        -0.06308    0.05034  -1.253   0.2114
## abdom         0.02454    0.04389   0.559   0.5766
## hip           0.16113    0.07408   2.175   0.0306 *
## thigh        -0.03278    0.07329  -0.447   0.6551
## knee         -0.19103    0.12285  -1.555   0.1213
## ankle         0.25326    0.11244   2.252   0.0252 *
## biceps        0.14049    0.08688   1.617   0.1072
## forearm      -0.11631    0.10110  -1.150   0.2511
## wrist        -0.17917    0.27160  -0.660   0.5101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.186 on 238 degrees of freedom
## Multiple R-squared:  0.0839, Adjusted R-squared:  0.03386
## F-statistic: 1.677 on 13 and 238 DF,  p-value: 0.06664
```

```
### (ii) Check the normality assumption
qqnorm(residuals(modLily),ylab="Residuals",main="")
qqline(residuals(modLily), lwd=3,col="red")
```



```
shapiro.test(residuals(modLily))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modLily)
## W = 0.99297, p-value = 0.2801
```

(e)

```
#######Part (e): Collinearity ##########
### (e)(1) Pairwise correlation
## any pairs that have large correlation?
##   Say >= 0.90? >= 0.80? >= 0.70?

### There are two ways to extract X matrix
### The first one is to use "model.matrix" function in the regression model
##  here we assume to exclude the intercept for collinearity analysis
Lily.X <- model.matrix(modLily)[,-1];
round(cor(Lily.X),2)
```

```
##            age weight height neck chest abdom   hip thigh knee ankle biceps
## age       1.00  -0.01  -0.17 0.11  0.18  0.23 -0.05 -0.20 0.02 -0.11  -0.04
## weight   -0.01   1.00   0.31 0.83  0.89  0.89  0.94  0.87 0.85  0.61   0.80
## height   -0.17   0.31   1.00 0.25  0.13  0.09  0.17  0.15 0.29  0.26   0.21
## neck      0.11   0.83   0.25 1.00  0.78  0.75  0.73  0.70 0.67  0.48   0.73
## chest     0.18   0.89   0.13 0.78  1.00  0.92  0.83  0.73 0.72  0.48   0.73
## abdom     0.23   0.89   0.09 0.75  0.92  1.00  0.87  0.77 0.74  0.45   0.68
## hip      -0.05   0.94   0.17 0.73  0.83  0.87  1.00  0.90 0.82  0.56   0.74
## thigh    -0.20   0.87   0.15 0.70  0.73  0.77  0.90  1.00 0.80  0.54   0.76
## knee      0.02   0.85   0.29 0.67  0.72  0.74  0.82  0.80 1.00  0.61   0.68
## ankle    -0.11   0.61   0.26 0.48  0.48  0.45  0.56  0.54 0.61  1.00   0.48
## biceps   -0.04   0.80   0.21 0.73  0.73  0.68  0.74  0.76 0.68  0.48   1.00
## forearm -0.09   0.63   0.23 0.62  0.58  0.50  0.55  0.57 0.56  0.42   0.68
## wrist     0.21   0.73   0.32 0.74  0.66  0.62  0.63  0.56 0.66  0.57   0.63
##         forearm wrist
## age       -0.09  0.21
## weight     0.63  0.73
## height     0.23  0.32
## neck       0.62  0.74
## chest      0.58  0.66
## abdom      0.50  0.62
## hip        0.55  0.63
## thigh      0.57  0.56
## knee       0.56  0.66
## ankle      0.42  0.57
## biceps     0.68  0.63
## forearm    1.00  0.59
## wrist      0.59  1.00
```

```
### The second one is to create the X matrix by ourselves from the raw data
###     based on those 13 predictor variables
head(fat);
```

```
##   brozek siri density age weight height adipos  free neck chest abdom   hip
## 1   12.6 12.3  1.0708  23 154.25  67.75   23.7 134.9 36.2  93.1  85.2  94.5
## 2    6.9  6.1  1.0853  22 173.25  72.25   23.4 161.3 38.5  93.6  83.0  98.7
## 3   24.6 25.3  1.0414  22 154.00  66.25   24.7 116.0 34.0  95.8  87.9  99.2
## 4   10.9 10.4  1.0751  26 184.75  72.25   24.9 164.7 37.4 101.8  86.4 101.2
## 5   27.8 28.7  1.0340  24 184.25  71.25   25.6 133.1 34.4  97.3 100.0 101.9
## 6   20.6 20.9  1.0502  24 210.25  74.75   26.5 167.0 39.0 104.5  94.4 107.8
##   thigh knee ankle biceps forearm wrist
## 1  59.0 37.3  21.9   32.0    27.4  17.1
## 2  58.7 37.3  23.4   30.5    28.9  18.2
## 3  59.6 38.9  24.0   28.8    25.2  16.6
## 4  60.1 37.3  22.8   32.4    29.4  18.2
## 5  63.2 42.2  24.0   32.2    27.7  17.7
## 6  66.0 42.0  25.6   35.7    30.6  18.8
```

```
Lily.X2 <- fat[,c(4:6,9:18)]
round(cor(Lily.X2),2)
```

```
##           age weight height neck chest abdom   hip thigh knee ankle biceps
## age      1.00  -0.01  -0.17 0.11  0.18  0.23 -0.05 -0.20 0.02 -0.11  -0.04
## weight  -0.01   1.00   0.31 0.83  0.89  0.89  0.94  0.87 0.85  0.61   0.80
## height  -0.17   0.31   1.00 0.25  0.13  0.09  0.17  0.15 0.29  0.26   0.21
## neck     0.11   0.83   0.25 1.00  0.78  0.75  0.73  0.70 0.67  0.48   0.73
## chest    0.18   0.89   0.13 0.78  1.00  0.92  0.83  0.73 0.72  0.48   0.73
## abdom    0.23   0.89   0.09 0.75  0.92  1.00  0.87  0.77 0.74  0.45   0.68
## hip     -0.05   0.94   0.17 0.73  0.83  0.87  1.00  0.90 0.82  0.56   0.74
## thigh   -0.20   0.87   0.15 0.70  0.73  0.77  0.90  1.00 0.80  0.54   0.76
## knee     0.02   0.85   0.29 0.67  0.72  0.74  0.82  0.80 1.00  0.61   0.68
## ankle   -0.11   0.61   0.26 0.48  0.48  0.45  0.56  0.54 0.61  1.00   0.48
## biceps  -0.04   0.80   0.21 0.73  0.73  0.68  0.74  0.76 0.68  0.48   1.00
## forearm -0.09   0.63   0.23 0.62  0.58  0.50  0.55  0.57 0.56  0.42   0.68
## wrist    0.21   0.73   0.32 0.74  0.66  0.62  0.63  0.56 0.66  0.57   0.63
##         forearm wrist
## age       -0.09  0.21
## weight     0.63  0.73
## height     0.23  0.32
## neck       0.62  0.74
## chest      0.58  0.66
## abdom      0.50  0.62
## hip        0.55  0.63
## thigh      0.57  0.56
## knee       0.56  0.66
## ankle      0.42  0.57
## biceps     0.68  0.63
## forearm    1.00  0.59
## wrist      0.59  1.00
```

```
## these two methods should yield to same answers!


### (e)(2) Condition Numbers
Lily.X <- model.matrix(modLily)[,-1]
Lily.e <- eigen(t(Lily.X) %*% Lily.X)
Lily.e$val
```

```
##  [1] 1.959256e+07 6.418499e+04 3.059739e+04 5.704341e+03 2.803947e+03
##  [6] 1.934715e+03 1.030340e+03 6.376692e+02 5.280964e+02 4.318186e+02
## [11] 3.763758e+02 2.723663e+02 6.345357e+01
```

```
sqrt(Lily.e$val[1]/Lily.e$val)
```

```
##  [1]   1.00000  17.47144  25.30482  58.60610  83.59121 100.63222 137.89717
##  [8] 175.28623 192.61449 213.00748 228.15747 268.20620 555.67072
```

```
### (e)(3) VIF values
###  There are two ways. The first one is to use the function "vif" in the package
↪  "faraway"
require(faraway)
vif(Lily.X)
```

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

```
max(vif(Lily.X))
```

```
## [1] 33.50932
```

```
mean(vif(Lily.X))
```

```
## [1] 7.790078
```

```
### The second one is to compute VIF on your own
p = dim(Lily.X)[2];
VIF1 <- NULL;
for (i in 1:p){
 Rsqure.tmp <- summary(lm(Lily.X[,i] ~ Lily.X[,-i]))$r.squared;
 VIF1 <- cbind(VIF1, 1/(1-Rsqure.tmp));
}
VIF1
```

```
##         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
## [1,] 2.25045 33.50932 1.674591 4.324463 9.460877 11.76707 14.79652 7.777865
##           [,9]    [,10]    [,11]    [,12]    [,13]
## [1,] 4.612147 1.907961 3.619744 2.192492 3.377515
```

```
colnames(VIF1) <- colnames(Lily.X);
VIF1
```

```
##          age   weight   height     neck    chest    abdom      hip    thigh
## [1,] 2.25045 33.50932 1.674591 4.324463 9.460877 11.76707 14.79652 7.777865
##           knee    ankle   biceps  forearm    wrist
## [1,] 4.612147 1.907961 3.619744 2.192492 3.377515
```

```
## Check whether this is the same as the "vif" function
vif(Lily.X)
```

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

(f)

```
#######Part (f): Transforming the response ##########
require(MASS)
```

```
## Loading required package: MASS
```

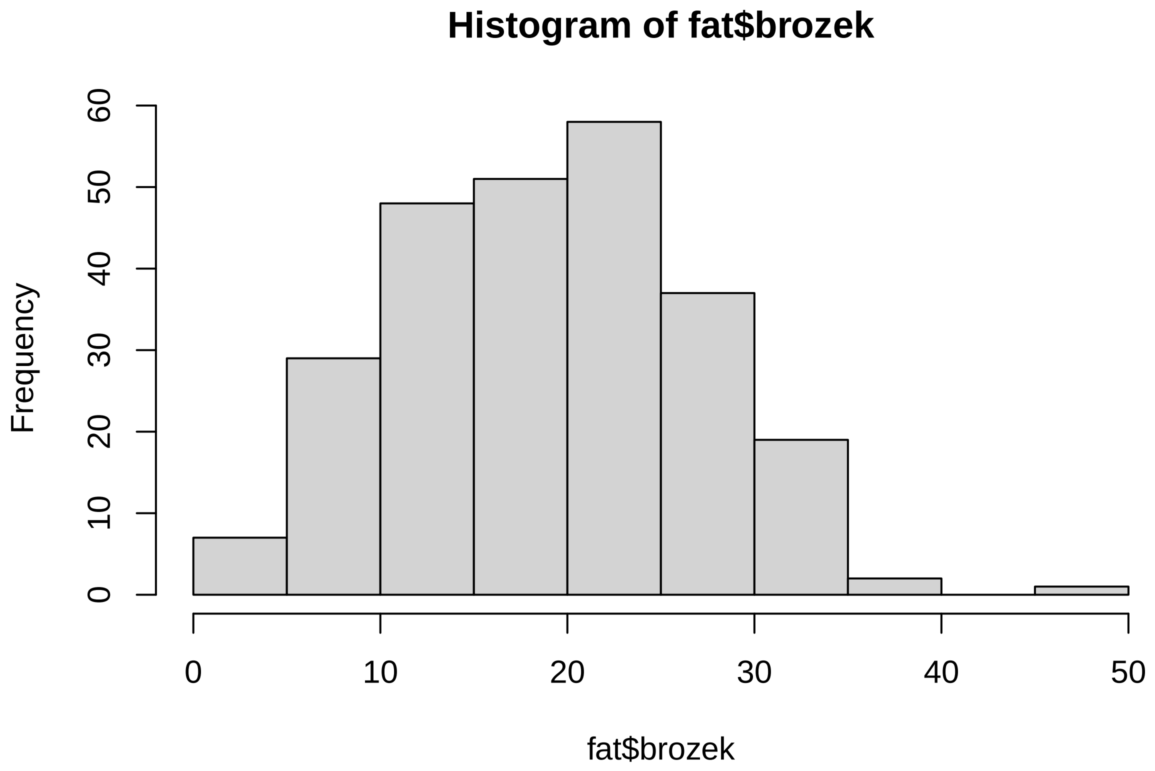
```
### 1. summary statistics for the response variable
summary(fat$brozek)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.80   19.00   18.94   24.60   45.10
```

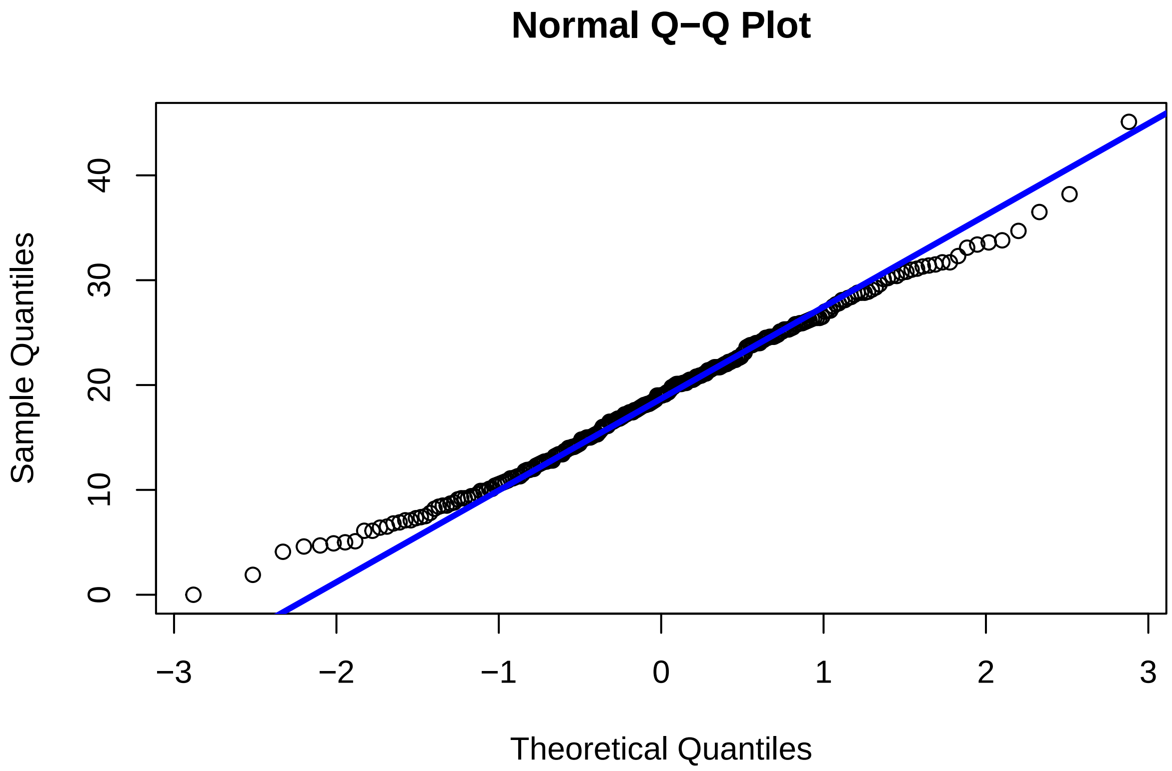
```
c(mean(fat$brozek), sd(fat$brozek) )
```

```
## [1] 18.938492  7.750856
```

```
### 2. plots
hist(fat$brozek)
```



```
qqnorm(fat$brozek)
qqline(fat$brozek, lwd=3, col="blue")
```

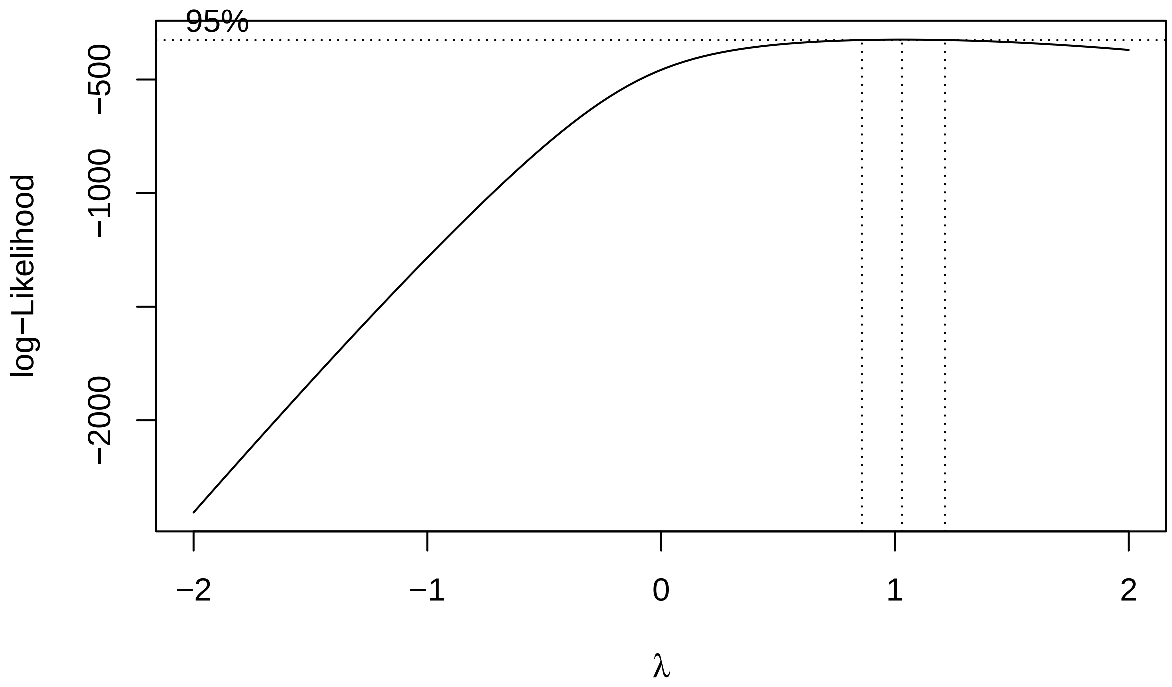


```
### 3. Test Normality assumption
shapiro.test(fat$brozek)
```

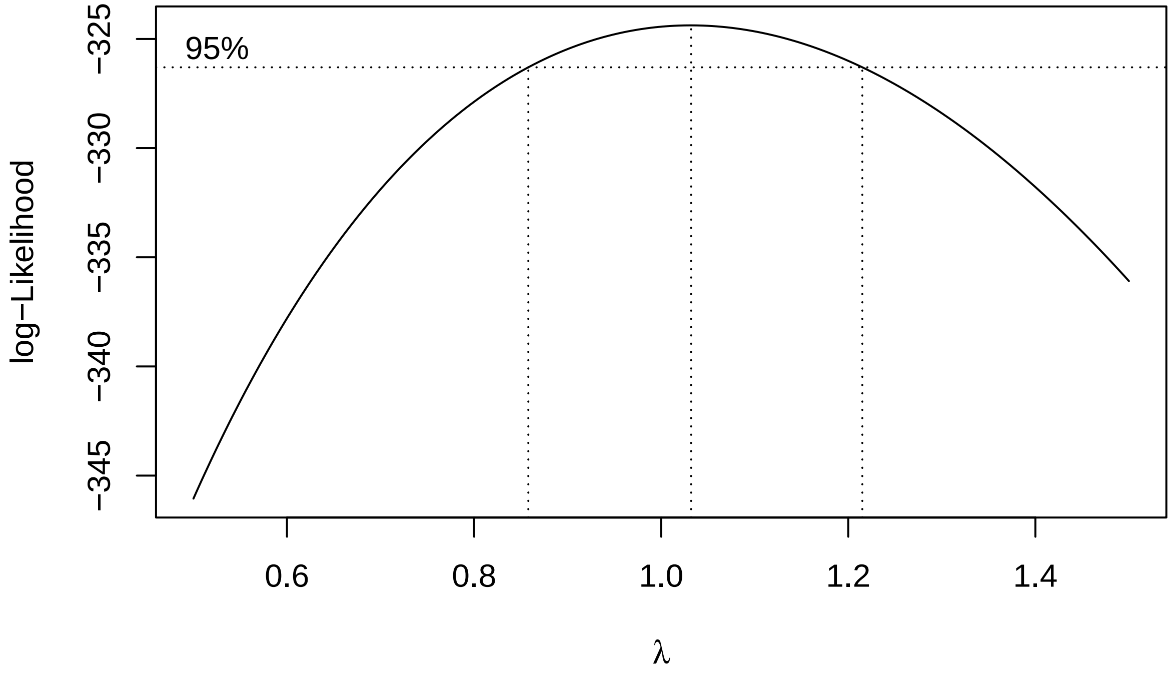
```
##
##  Shapiro-Wilk normality test
##
## data:  fat$brozek
## W = 0.99292, p-value = 0.2747
```

```
### 4. Box-Cox transformation
library(MASS)
# boxcox(modLily, plotit = T);

### 5. Making responses to be positive before using boxcox
modLily1 <- lm( I(0.1+brozek) ~ age + weight+ height+ neck+ chest + abdom+ hip+
                thigh+ knee+ ankle+ biceps+ forearm+ wrist, data=fat);
boxcox(modLily1, plotit = T);
```



```r
boxcox(modLily1, lambda= seq(0.5, 1.5, by=0.001))
```



(g)

```
#######Part (g): Transforming the predictor ##########
### 1. add quadatic function of "abdom"
modLily2 <- lm( brozek ~ age + weight+ height+ neck+ chest + abdom+ I(abdom^2) +
                  hip+thigh+ knee+ ankle+ biceps+ forearm+ wrist, data=fat);
summary(modLily2)
```

```
##
## Call:
## lm(formula = brozek ~ age + weight + height + neck + chest +
##     abdom + I(abdom^2) + hip + thigh + knee + ankle + biceps +
##     forearm + wrist, data = fat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0626  -2.8326  -0.0641   2.5456   8.9499
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -40.062052  18.064852  -2.218  0.02753 *
## age           0.059107   0.029538   2.001  0.04652 *
## weight       -0.042509   0.050625  -0.840  0.40193
## height       -0.117154   0.089551  -1.308  0.19206
## neck         -0.401973   0.212552  -1.891  0.05982 .
## chest        -0.051016   0.091007  -0.561  0.57562
## abdom         1.581316   0.256695   6.160 3.09e-09 ***
## I(abdom^2)   -0.003852   0.001352  -2.849  0.00477 **
## hip          -0.119470   0.136033  -0.878  0.38070
## thigh         0.162969   0.133966   1.217  0.22500
## knee         -0.167375   0.227527  -0.736  0.46268
## ankle         0.172261   0.202167   0.852  0.39503
## biceps        0.136916   0.156294   0.876  0.38191
## forearm       0.303322   0.187156   1.621  0.10641
## wrist        -1.574766   0.489489  -3.217  0.00148 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.93 on 237 degrees of freedom
## Multiple R-squared:  0.7573, Adjusted R-squared:  0.743
## F-statistic: 52.82 on 14 and 237 DF,  p-value: < 2.2e-16
```

```
### 2. add both quadatic and cubic function of "abdom"
modLily3 <- lm( brozek ~ age + weight+ height+ neck+ chest + abdom+ I(abdom^2) +
                  I(abdom^3) + hip+thigh+ knee+ ankle+ biceps+ forearm+ wrist, data=fat);
summary(modLily3)
```

```
##
## Call:
## lm(formula = brozek ~ age + weight + height + neck + chest +
##     abdom + I(abdom^2) + I(abdom^3) + hip + thigh + knee + ankle +
##     biceps + forearm + wrist, data = fat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0071  -2.8404  -0.1842   2.5369   9.0815
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.061e+00  5.558e+01   0.055  0.95613
## age          6.226e-02  2.981e-02   2.089  0.03780 *
## weight      -4.106e-02  5.069e-02  -0.810  0.41871
## height      -1.135e-01  8.973e-02  -1.264  0.20733
## neck        -3.743e-01  2.154e-01  -1.738  0.08350 .
## chest       -6.274e-02  9.218e-02  -0.681  0.49679
## abdom        2.627e-01  1.628e+00   0.161  0.87191
## I(abdom^2)   9.046e-03  1.578e-02   0.573  0.56698
## I(abdom^3)  -4.139e-05  5.045e-05  -0.820  0.41278
## hip         -1.079e-01  1.369e-01  -0.788  0.43128
## thigh        1.705e-01  1.344e-01   1.269  0.20572
## knee        -1.815e-01  2.283e-01  -0.795  0.42756
## ankle        1.873e-01  2.031e-01   0.922  0.35735
## biceps       1.427e-01  1.566e-01   0.911  0.36302
## forearm      2.779e-01  1.898e-01   1.464  0.14452
## wrist       -1.590e+00  4.902e-01  -3.243  0.00135 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.932 on 236 degrees of freedom
## Multiple R-squared:  0.758,  Adjusted R-squared:  0.7426
## F-statistic: 49.28 on 15 and 236 DF,  p-value: < 2.2e-16
```