Midterm

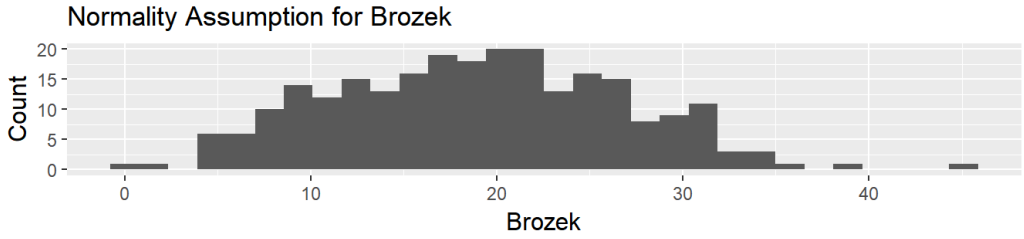
Chong Zhang

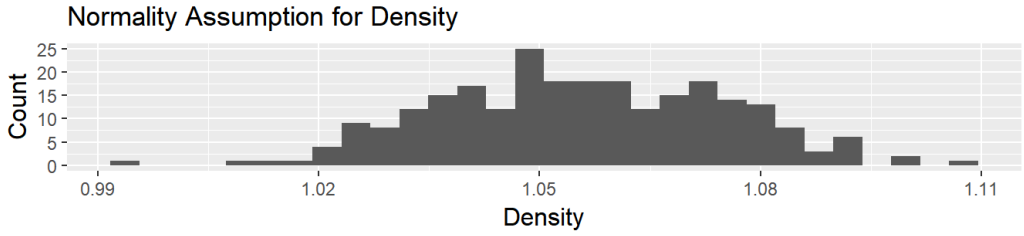
October 13, 2018

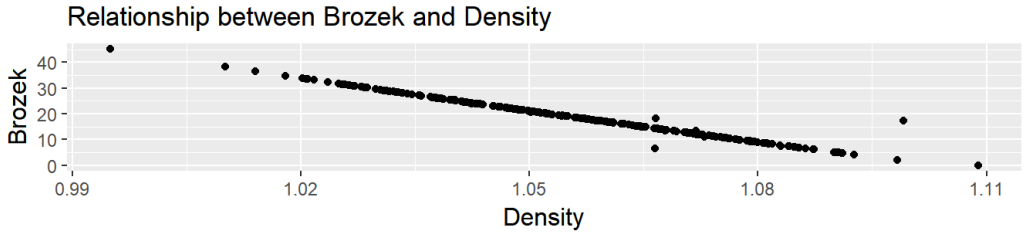
```
library(ggplot2)
library(dplyr)
library(gridExtra)
library(car)
data_table = read.table('fat.csv', header = T, sep = ',')
```

1. Perform exploratory data analysis: calculate the sample mean, sample variance, draw the histogram and test the normality assumption of the response variable brozek and one predicting variable density. What is the relationship between the two variables? Comment on the shape of the distribution for the two variables.

```
sample_mean_brozek = mean(data_table$brozek)
sample_mean_density = mean(data_table$density)
sample_variance_brozek = var(data_table$brozek)
sample_variance_density = var(data_table$density)
cor_brozek_density = cor(data_table$brozek,data_table$density)

p_brozek = ggplot(data= data_table,aes(brozek))+geom_histogram()+xlab('Brozek')+ylab('Count')+theme(axis.title =element_text(size=12))+
ggtitle('Normality Assumption for Brozek')
p_density = ggplot(data= data_table,aes(density))+geom_histogram()+xlab('Density')+ylab('Count')+theme(axis.title =element_text(size=12
))+ggtitle('Normality Assumption for Density')
p_cor = ggplot(data= data_table,aes(density,brozek))+geom_point()+xlab('Density')+ylab('Brozek')+theme(axis.title =element_text(size=12
))+ggtitle('Relationship between Brozek and Density')

grid.arrange(p_brozek, p_density,p_cor)
```







As we can see from the data the **sample mean for Brozek is** `18.9384921` , and **the sample variance for Brozek is** `60.0757635` . The **sample mean for denasity is** `1.0555738` , and **the sample variance for density is 0.00036**`. The **correlation between Brozek and Density is** `-0.9880867` . Since it is so close to -1, it means that there is an almost perfect linear relationship between Brozek and Density.

The histogram of both variables are shown above. It seems that both variables follow normal distribution.

2.Fit a multiple linear regression model named 'model1' using brozek as the response variable and all predicting variables described above. Calculate the mean square error of this model (i.e., MSE=). Is the model a good fit? Does the model have predictive power? Explain in detail.

```
model = lm(brozek~., data = data_table)
summary(model)
```

```
##
## Call:
## lm(formula = brozek ~ ., data = data_table)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11191 -0.04847  0.00277  0.04625  1.47542
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.1524013  4.1718589   2.913  0.00393 **
## siri         0.8884085  0.0111341  79.792  < 2e-16 ***
## density     -9.8456305  3.7471770  -2.627  0.00917 **
## age         -0.0005268  0.0012935  -0.407  0.68421
## weight       0.0084855  0.0036200   2.344  0.01991 *
## height      -0.0005459  0.0044439  -0.123  0.90234
## adipos      -0.0153248  0.0124778  -1.228  0.22062
## free        -0.0097388  0.0044270  -2.200  0.02880 *
## neck         0.0005002  0.0094279   0.053  0.95773
## chest        0.0021454  0.0043013   0.499  0.61840
## abdom        0.0014464  0.0044217   0.327  0.74388
## hip         -0.0044514  0.0058941  -0.755  0.45087
## thigh        0.0156926  0.0059507   2.637  0.00892 **
## knee        -0.0252126  0.0098531  -2.559  0.01113 *
## ankle        0.0027790  0.0089580   0.310  0.75667
## biceps      -0.0147134  0.0069201  -2.126  0.03454 *
## forearm      0.0149983  0.0080832   1.855  0.06478 .
## wrist        0.0326518  0.0218000   1.498  0.13554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1706 on 234 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 3.046e+04 on 17 and 234 DF,  p-value: < 2.2e-16
```

```
MSE = sum((model$fitted.values-data_table$brozek)^2)/length(data_table$brozek)
```

The mean square error of this model is 0.027026 . The R-square is **0.9995**, the adjusted-R-square is **0.9995**. Measn Square Error is an estimate of the standard deviation of the random component in the data. An MSE value closer to 0 indicates a fit that is more useful for prediction. R-square means the proportion of the variance in the dependent variable that is predictable from the independent variables. The closer R-square to 1, the better the model fits the data. Thus both MSE and R-square suggests that the model fit the data very nicely and have a strong predictive power.

3. Fit a multiple linear regression model named ‘model2’ using brozek as the response variable and only the significant predicting variables from model1 (with p-value<0.05). Write down the fitted equation and calculate the mean square error of this model. Use partial F-test to see whether the other predicting variables except for the variables from model2 contribute significant information to the response variable brozek.

```
model2 = lm(brozek~siri+density+weight+free+thigh+knee+biceps,data = data_table)
MSE_2 = sum((model2$fitted.values-data_table$brozek)^2)/length(data_table$brozek)
summary(model2)
```

```
##
## Call:
## lm(formula = brozek ~ siri + density + weight + free + thigh +
##     knee + biceps, data = data_table)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12961 -0.03491  0.00436  0.03831  1.56045
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.065121   4.078993   2.958   0.0034 **
## siri         0.896678   0.010507  85.344   <2e-16 ***
## density     -9.433031   3.690997  -2.556   0.0112 *
## weight       0.004512   0.002941   1.534   0.1263
## free        -0.004089   0.003883  -1.053   0.2934
## thigh        0.009431   0.004419   2.134   0.0338 *
## knee        -0.021243   0.009028  -2.353   0.0194 *
## biceps      -0.009345   0.006244  -1.497   0.1358
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.171 on 244 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 7.361e+04 on 7 and 244 DF,  p-value: < 2.2e-16
```

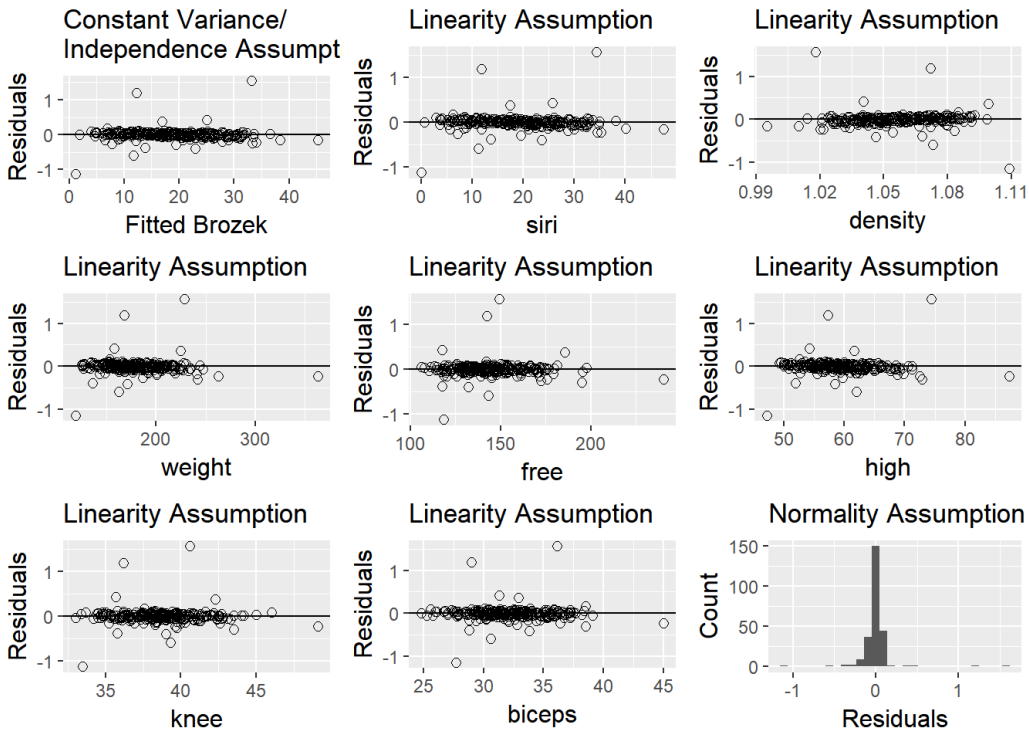
```
anova(model2,model)
```

```
## Analysis of Variance Table
##
## Model 1: brozek ~ siri + density + weight + free + thigh + knee + biceps
## Model 2: brozek ~ siri + density + age + weight + height + adipos + free +
##     neck + chest + abdom + hip + thigh + knee + ankle + biceps +
##     forearm + wrist
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    244 7.1370
## 2    234 6.8105 10   0.32652 1.1219 0.3464
```

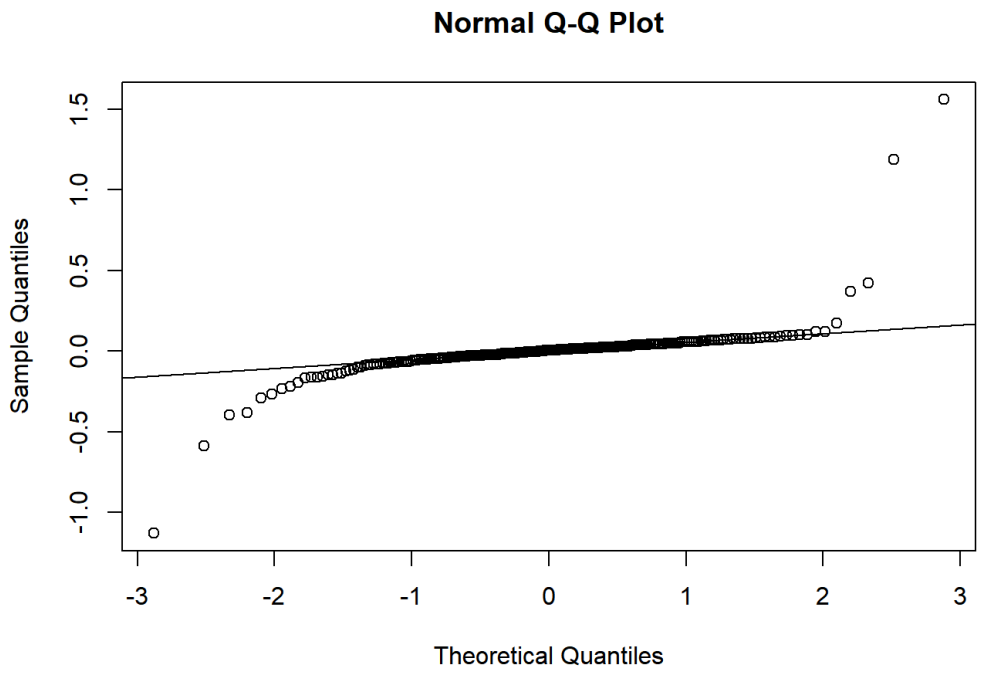
The fitted equation of model2 is
$brozek = 0.8967 * siri - 9.433 * density + 0.0045 * weight - 0.0041 * free + 0.0094 * thigh - 0.0212 * knee - 0.0093 * biceps.$
The mean square error of model2 is `0.0283216` Bease on the F-partial test using anova function in r, I found that the **p-value for the F-parital test is 0.3463**, which is bigger than 0.05. Thus it indicates that the other predicting variables except for the variables from model2 **didn't** contribute significant information to the response variable brozek.

4. Perform the residual analysis and check whether there is any multicollinearity among the predicting variables from model2.

```
p1=ggplot(data= model2, aes(x=model2$fitted.values, y=model2$residuals))+geom_point(shape=1, size=2)+xlab('Fitted Brozek')+ylab('Residuals')+theme(axis.title=element_text(size=12))+geom_hline(yintercept = 0)+ggtitle('Constant Variance/\nIndependence Assumption')

p2 =ggplot(data= data_table,aes(x=data_table$siri,y=model2$residuals))+geom_point(shape=1,size=2)+xlab('siri')+ylab('Residuals')+theme(axis.title = element_text(size=12))+geom_hline(yintercept=0)+ggtitle('Linearity Assumption')

p3 =ggplot(data= data_table,aes(x=data_table$density,y=model2$residuals))+geom_point(shape=1,size=2)+xlab('density')+ylab('Residuals')+theme(axis.title=element_text(size=12))+geom_hline(yintercept=0)+ggtitle('Linearity Assumption')

p4 =ggplot(data= data_table,aes(x=data_table$weight,y=model2$residuals))+geom_point(shape=1,size=2)+xlab('weight')+ylab('Residuals')+theme(axis.title = element_text(size=12))+geom_hline(yintercept = 0)+ggtitle('Linearity Assumption')

p5 =ggplot(data= data_table,aes(x=data_table$free,y=model2$residuals))+geom_point(shape=1,size=2)+xlab('free')+ylab('Residuals')+theme(axis.title =element_text(size=12))+geom_hline(yintercept = 0)+ggtitle('Linearity Assumption')

p6 =ggplot(data= data_table,aes(x=data_table$thigh,y=model2$residuals))+geom_point(shape=1,size=2)+xlab('high')+ylab('Residuals')+theme(axis.title =element_text(size=12))+geom_hline(yintercept = 0)+ggtitle('Linearity Assumption')

p7 =ggplot(data= data_table,aes(x=data_table$knee,y=model2$residuals))+geom_point(shape=1,size=2)+xlab('knee')+ylab('Residuals')+theme(axis.title =element_text(size=12))+geom_hline(yintercept = 0)+ggtitle('Linearity Assumption')

p8 =ggplot(data= data_table,aes(x=data_table$biceps,y=model2$residuals))+geom_point(shape=1,size=2)+xlab('biceps')+ylab('Residuals')+theme(axis.title =element_text(size=12))+geom_hline(yintercept = 0)+ggtitle('Linearity Assumption')

p9 = ggplot(data= data_table,aes(model2$residuals))+geom_histogram()+xlab('Residuals')+ylab('Count')+theme(axis.title =element_text(size=12))+ggtitle('Normality Assumption')

grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8, p9)
```




```
qqnorm(model2$residuals)
qqline(model2$residuals)
```




```
cor(data_table[c(2,3,5,8,13,14,15)])
```

```
##                siri      density      weight        free       thigh
## siri     1.00000000 -0.98778240  0.6124140  0.01937491  0.5596075
## density -0.98778240  1.00000000 -0.5940619 -0.00574871 -0.5530910
## weight   0.61241400 -0.59406188  1.0000000  0.79219519  0.8686935
## free     0.01937491 -0.00574871  0.7921952  1.00000000  0.6766805
## thigh    0.55960753 -0.55309098  0.8686935  0.67668053  1.0000000
## knee     0.50866524 -0.49504035  0.8531674  0.70362435  0.7991703
## ankle    0.26596977 -0.26489003  0.6136854  0.58294600  0.5397971
##                knee       ankle
## siri      0.5086652  0.2659698
## density -0.4950404 -0.2648900
## weight    0.8531674  0.6136854
## free      0.7036244  0.5829460
## thigh     0.7991703  0.5397971
## knee      1.0000000  0.6116082
## ankle     0.6116082  1.0000000
```

```
vif_thershold = max(10, 1/(1-summary(model2)$r.squared))
vif(model2)
```

```
##      siri   density    weight      free     thigh      knee    biceps
## 66.342568 42.342494 64.090266 42.998325  4.618385  4.068211  3.054303
```

Constant Variance Assumption: I plot the residuals against the fitted values. All the residuals are distributed around 0. It also seems that the variance is constant across all the fitted values. Thus the **Constant Variance Assumption holds**.

Independence Assumption It seems that there is no clusters of residuals which means that the error terms are not correlated. Thus **Independence Assumption hold**.

Linearity Assumption I plot the residuals against each predictiing variable. All the plots show residuals distribut equally around the y=0 line. Thus it indicates that **Linearity Assumption holds**.

Normality assumption I use the QQ plot to study the normality assumption. Plot the residuals against the theoretical quantiles. The residuals approximately follow a straight line with some exception at both ends. Thus it suggests that the **Normality assumption holds**.

I calculate the correlation between all the predicting variables in model2 as above. some of them have correlation close to 1, which indicating possible multicollinearity. However **Variance Inflation Factor** were calculated for each predicting variables in model2 as shown above. Since the thershold of VIF for model2 is `2112.7854422` , and the VIFs of each predicting variables are smaller than `2112.7854422` , thus VIF indicates there is no multicollinearity among the predicting variables from model2.

5. Compare model1 and model2, which model is better? Explain in detail.

Both model have very similar R-squared and adjusted-R-squared. However model2 has much fewer predicting variables which makes it more concise. Generally, a model with fewer predictors and about the same “explanatory power: R-squared” is better. A model with more variables could suffer from multicollinearity, because the more variables you include, the higher the chance some of those variables could be linear independent. The more predictors included in the model, the more variability could be introduced in. Also this could lead to overfitting, the model will be biased. A model with less variables requires less data to be collected. This saves time and money.