A1: Brain Size and Intelligence

Last name: Deng
First name: Christina (Qi)
Student ID: 1001142408
Course section: STA302H1F-L5101

Oct. 13, 2016

Q1: t-test for MRIcount between high and low intelligence groups

Type your concise and clear answer here.

- **Null hypothesis**: **H0**: the mean MRI count between the high intelligence and low intelligence groups are equal. **Ha**: true difference in means is not equal to 0. (H0 is not true)
- We use **Welch Two Sample t-test** /unequal variances t-test. (assume that variance of the MRI accounts in the two groups are unequal, follow normal distribution, and cases are independent)
- **t** = 1.53, **df** = 37.324, **significance level** = 0.05
- **p-value** = 0.1344 > 0.05, so **unable to reject** the null hypothesis.
- Since p-value is greater than significance, we are unable to reject the null hypothesis, it implies that there is no evidence exists for us to reject the null hypothesis.
- By t-test, we conclude that the difference in true means is equal to 0. Hence, we conclude that MRIcount is not affected by the intelligence level of a group.

Additional Information

- The **95 percent confidence interval** is [-11138.2, 79910.4]
- The **mean of x(high IQ MRIcount)** is 925948.1; the **mean of y(low IQ MRIcount)** is 891561.9.

Q2: correlation analysis among the MRI count and IQ variables

Correlations of the IQ measurements with MRI count (p-value for test of $\rho = 0$ is in brakets):

-	Full data	High-IQ group	low-IQ group
FSIQ	0.3576(0.0235)	0.5483(0.0123)	0.5273(0.0169)
VIQ	0.3375(0.0332)	0.4067(0.0752)	0.1464(0.5381)
PIQ	0.3868(0.0137)	0.2012(0.3948)	0.5862(0.0066)

From the correlation analysis, we conclude that

H0: there is no linear relationship between MRI and PIQ; Ha: H0 is false.

Evidence

p-value < alpha, reject H0;

p-value < 0.001, strong evidence against H0;

p-value < 0.05, moderate evidence against H0;

p-value < 0.1, week evidence against H0.

p-value > alpha, do not reject H0;

p-value > 0.1, no evidence against H0.

Correlation

0.2 <= |r| <= 0.4: weak correlation

0.4 < |r| <= 0.6: moderate correlation

0.6 < |r| <= 0.8: strong correlation

|r| > 0.8: very strong correlation.

Correlation between MRI count and 3 IQ variables in full data:

FSIQ and MRIcount in full data have moderate evidence against H0 and with weak correlation;

VIQ and MRIcount in full data have moderate evidence against H0 and with weak correlation;

PIQ and MRIcount in full data have moderate evidence against H0 and with weak correlation.

Correlation between MRI count and 3 IQ variables in high-IQ group:

FSIQ and MRIcount in High-IQ group have moderate evidence against H0 and with moderate correlation;

VIQ and MRIcount in High-IQ group have weak evidence against H0 and with moderate correlation;

PIQ and MRIcount in High-IQ group have no evidence against H0 and therefore no correlation.

(even r indicates weak correlation)

Correlation between MRI count and 3 IQ variables in Low-IQ group:

FSIQ and MRIcount in Low-IQ group have moderate evidence against H0 and with moderate correlation;

VIQ and MRIcount in Low-IQ group have no evidence against H0 and therefore no correlation;

(even r indicates weak correlation)

PIQ and MRIcount in Low-IQ group have moderate evidence against H0 and with moderate correlation.

Q3

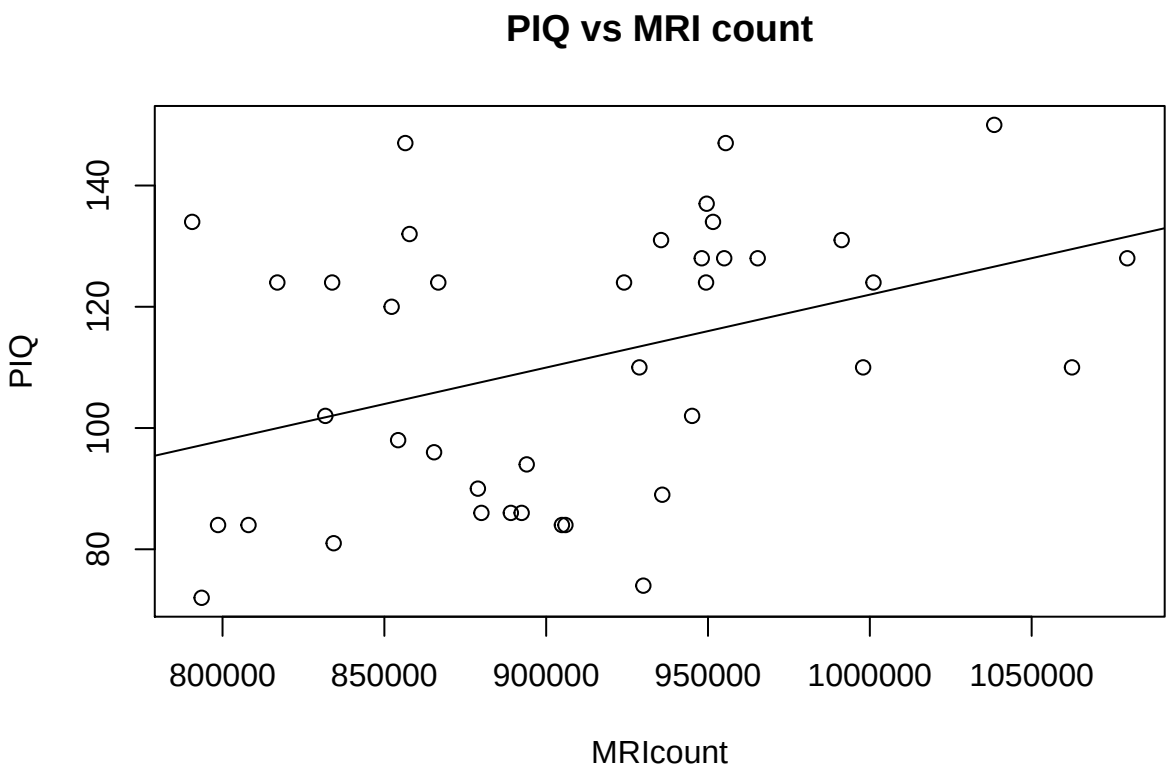
The first question tells us that MRIcount is not affected by the intelligence level of a group, but question 2 shows that MRIcount and intelligence level are somehow related since the correlations are different. The result of the t-test in question 1 does not quite agree with the relavant correlation result in question 2.

T-test in question 1 only tests the mean MRI count between the high and low IQ by dividing samples into two groups. By doing so, we lose significant amout of useful datas, which leads us to unaccurate results. However, question 2 strictly calculates the 9 correlations between MRIcount and 3 IQ scores, which gives us more precise results and significant interpretations.

Under the assumption that the data have a bivariate normal distribution, the correlation is the preferred analysis, apparently, statistics study is quite different than other study field such as medical science. Medicians like to dichotomized variables because it is more effective for them to proceed their research. The relationship they study is more obvious, and they often omit the information that are not that important. They care less about the numerical relationship and care more about the relationship in reality. However, statisticians have to utilize all the information to receive the most accurate results, they want to know exactly the relationship between each variable for each group. Thus, as a statistician, we more rely on the correlation results.

Q4(a) Scatter plot of PIQ versus MRI count

```
# scatter plot of PIQ versus MRI count
brain = read.table("/Users/christinadeng/Desktop/Fall 2016/STA 302/Assignemnt /Assignment 1/BrainData.csv",sep=" ",
#complete the following plot() command to get the scatter plot
plot(brain$MRIcount, brain$PIQ, main ="PIQ vs MRI count", xlab="MRIcount", ylab = "PIQ");
abline(lm(brain$PIQ~brain$MRIcount))
```



Q4(b) Regression analysis for two groups

Regression	R^2	Intercept (b_0)	Slope(b_1)	MSE	p-value for $H_0 : \beta_1 = 0$
High-IQ groups	0.04051	1.100e+02	2.265e-05	73.50501	0.3948
Low-IQ groups	0.3436	1.636e+00	1.003e-04	88.75023	0.006602

i.) By observation, we notice that the slopes are very small, even approching to zero. However, we cannot conclude that there is no relationship between PIQ and MRI count. Even though large increase of MRIcount results in smaller increase of PIQ than we are expected, there is still relationship between the two variables.

ii.) By comparing the value of R^2 for the two regression, Low-IQ groups will give us better fit if we use R^2 as the criteria, since the larger R^2 is preferred.

iii.) By comparing the value of MSE for the two regression, High-IQ groups will give us better fit if we use MSE as the criteria, since the smaller MSE is preferred.

iv.) Given only the values of R^2 and MSE, we would almost always choose MSE as our criteria. Through out research, statiscians claim that R-square has uncertainty, and MSE is more stable relatively. Even though higher R-square will give us better fit, but high R-square does not necessarily imply that you can make useful

prediction, nor you can conclude that estimated line is a good fit; and a low R-square does not necessarily imply that X and Y are not related or independent. R-square might be large but MSE may be still too large for inference to be useful in prediction, R-square might be small but MSE may still be small which is useful in prediction. Therefore, we should choose MSE as our criteria.

Apendix: Source R code

```
# ---------> complete and run the following code for this assignment  <-------
#
# R code for STA302 or STA1001H1F assignment 1
# copyright by Christina (Qi) Deng
# date: Oct. 04, 2016
#

## Load in the data set
brain = read.table("/Users/christinadeng/Desktop/Fall 2016/STA 302/Assignemnt /Assignment 1/BrainData.csv",sep=" ",

## create an indicator for high-IQ (value =1) and low-IQ (value=0)
highIQ = ifelse(brain$FSIQ>=130,1, 0)

## Q1: t-test on MRI count between high- and low IQ groups

# subsets by high and low intelligence group (IG)
highIG_dat <- subset(brain, highIQ==1)
lowIG_dat <- subset(brain, highIQ==0)

# mean MRI count between the high and low IG.
t.test(highIG_dat$MRIcount, lowIG_dat$MRIcount);


## Q2: correlation analysis
# cor.test() : missing value is suppressed, default setting:
#      mu = 0, alternative = c("two.sided"),  paired = FALSE,var.equal = FALSE
# - find correlation between MRI count and 3 IQ variables

cor.test(brain[,2], brain[,7])
cor.test(brain[,3], brain[,7])
cor.test(brain[,4], brain[,7])

# - find correlation between MRI count and 3 IQ variables in high-IQ group

cor.test(highIG_dat[,2], highIG_dat[,7])
cor.test(highIG_dat[,3], highIG_dat[,7])
cor.test(highIG_dat[,4], highIG_dat[,7])


# - find correlation between MRI count and 3 IQ variables in low-IQ group

cor.test(lowIG_dat[,2], lowIG_dat[,7])
cor.test(lowIG_dat[,3], lowIG_dat[,7])
cor.test(lowIG_dat[,4], lowIG_dat[,7])


## Q4:
# - Scatterplot of PIQ vs MRI count

plot(brain$MRIcount, brain$PIQ, main ="PIQ vs MRI count", xlab="MRIcount", ylab = "PIQ");
abline(lm(brain$PIQ~brain$MRIcount));

# - find R-square, b0, b1, MSE and p-value for b1 in high-IQ group

smart <- lm(highIG_dat$PIQ~highIG_dat$MRIcount)
summary(smart)
```

```r
(summary(smart)$sigma)^2

# - find R-square, b0, b1, MSE and p-value for b1 in low-IQ group

nonsmart <- lm(lowIG_dat$PIQ~lowIG_dat$MRIcount)
summary (nonsmart)
(summary(nonsmart)$sigma)^2
```

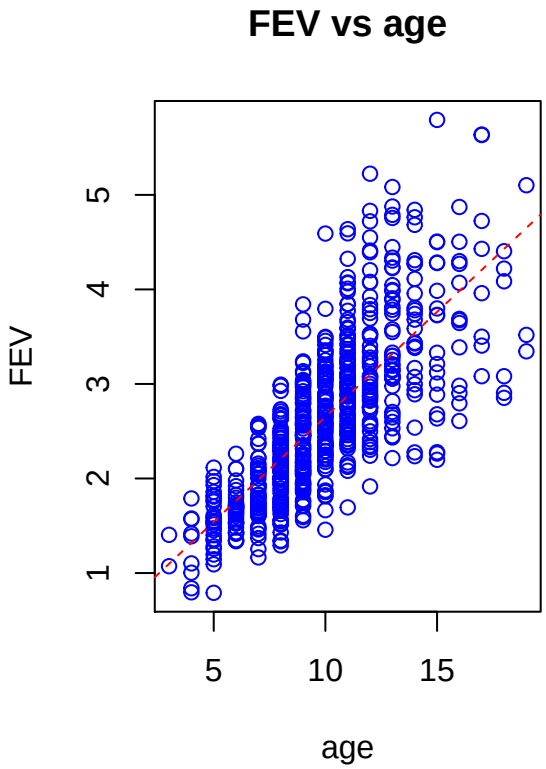
A2: Analysis to Forced Expiratory Volume data

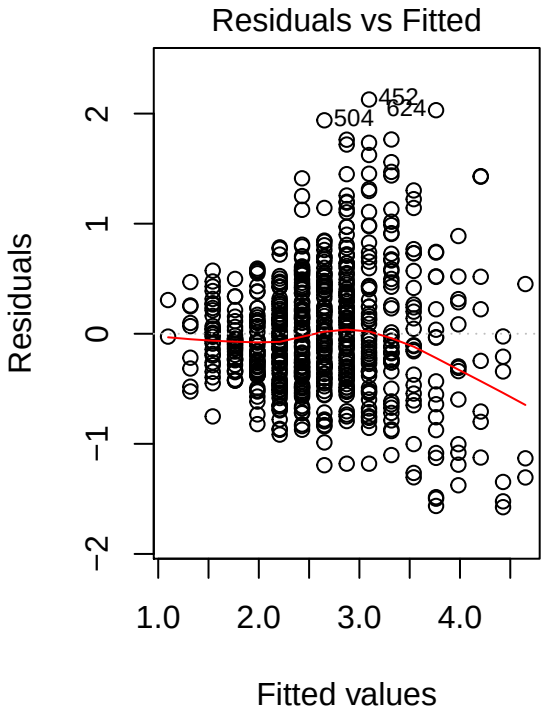
Last name: Deng
First name: Qi (Christina)
Student ID: 1001142408
Course section: STA302H1F-L5101

Nov. 6, 2016

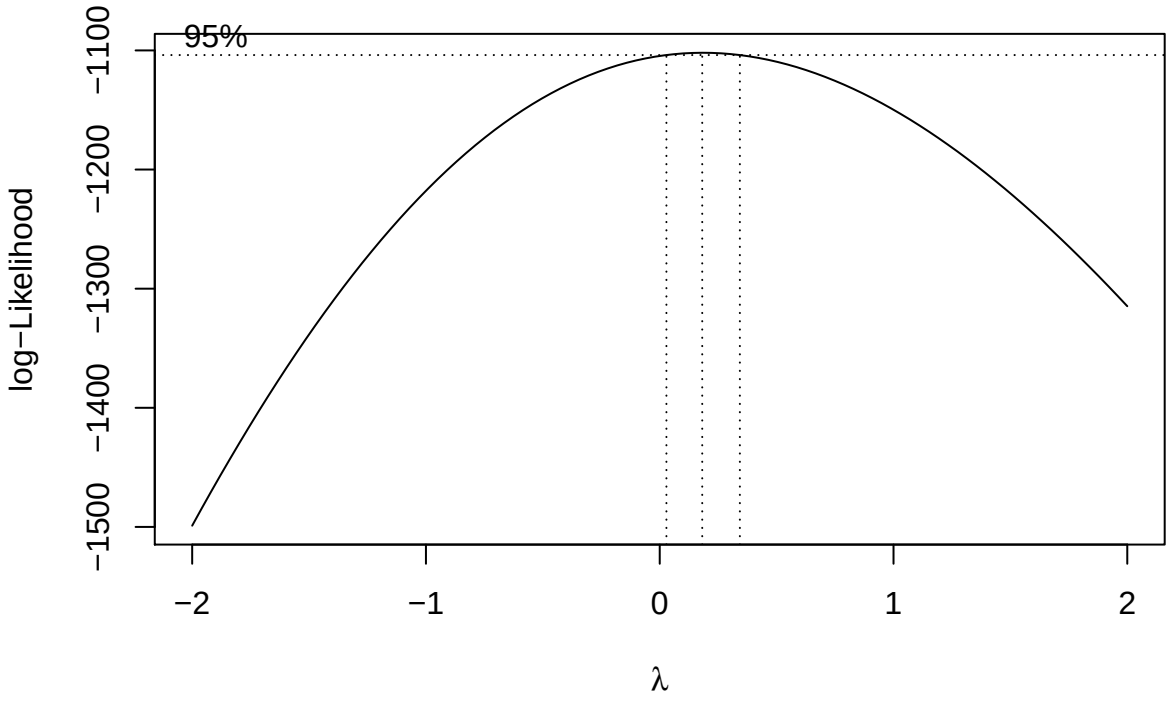
Q1: Fit a linear model to original data and looking for transformation using Box-Cox procedure

(a) From the two plots below, we first conclude the non-linearity of the regression function. The scatter plot has a actual curve that is different than the straight fitted line, since the points on scatter plot will form a concave upward sloping curve, while the residual plot has a clear pattern, which by definition clearly spot the **non-linearity**. Moreover, the residual plot does not spread out evenly across X, the residuals begin to spread wider along the x-axis, which means it **does not have constant variance**. In terms of **outliers**, point 504, 624, 452 may be consider as outlier observations.





(b) By definition, λ is the parameter to be determined from the data, The MLE is that value of λ for which SSE is a minimum. From the group below, the maximum point of the Box-Cox likelihood estimate occurs where λ is **close to zero**, which implies that **log** simple transformation seems the best. Also, by calculation, we accurately get the value of λ, which is **0.18**, it is close to zero, thus, we get the same conclusion as from the graph.



Q2: Fit a linear model wit transformed FEV and examine the residual plot of the fit.

(a) Estimated model

$$\widehat{log(FEV)} = 0.050596 + 0.087083 * \text{age}$$

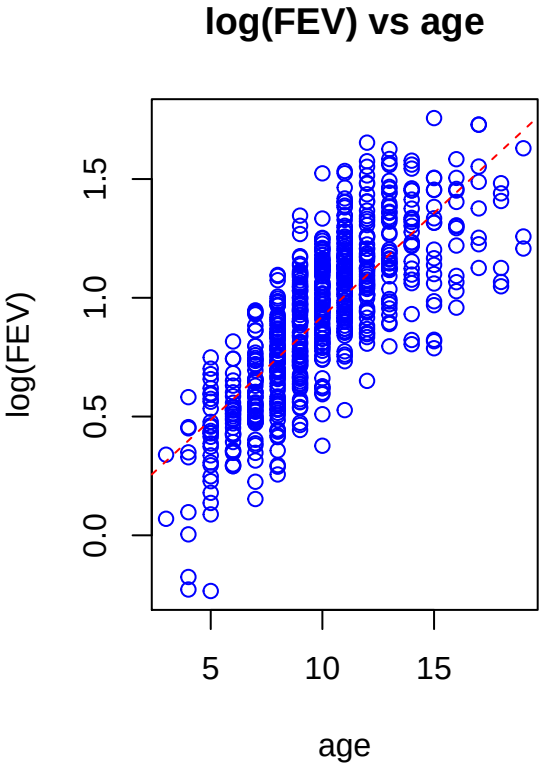
(b) The transformation has improved adherence to the constant variance assumption, but this linear model is not quite acceptable. By comparison of the two pairs of plots, mod 2's residual plot has pattern, I see a parabola in mod 2, so **non-linearity exists**. Error terms of mod 2 have variance that is **relatively more constant** since the residuals do not spread out nor gather along the x-axis.

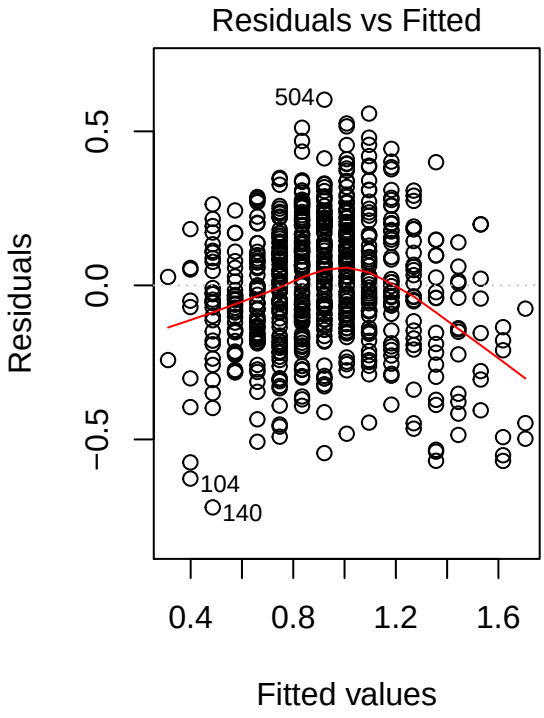
(c) B1 = 0.087083 > 0: As X increases by 1, expect Y to increase by (eˆ0.087083 - 1)*100% = 9.099%

(d)

1. age=8: 95% CI [2.070532 2.152692]; 95% PI [1.391573 3.203006]
2. age=17: 95% CI [4.431587 4.822374]; 95% PI [3.041955 7.025340]
3. age=21: 95% CI [6.148179 6.976410]; 95% PI [4.298236 9.979029]

But age of 21 is out of range, since max of age is 19. Thus, the CI and PI cannot be trusted.





Q3

(a) Estimated model

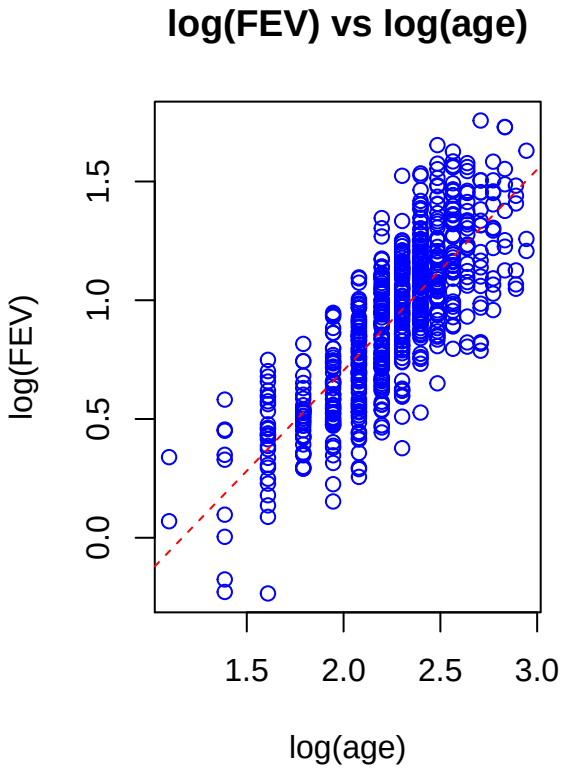
$$\widehat{log(FEV)} = -0.98772 + 0.84615log(\text{age})$$

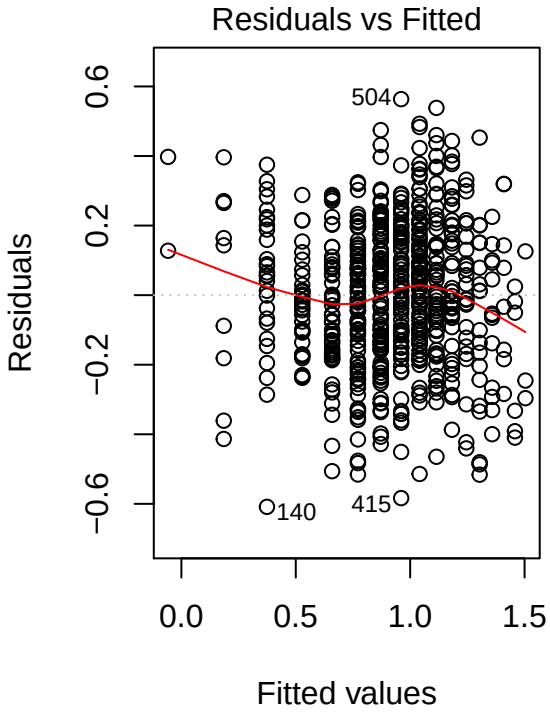
(b)

- The 95% confidence intervals for intercept (B0) in the transformed scale is [-1.1007528, -0.8746918]
- The 95% confidence intervals for slope (B1) in the transformed scale is [0.7963774, 0.8959283]

(c) B1 = 0.84615 > 0: As X doubles, expect Y to increase by (eˆ(0.84615log(2)) - 1)*100% = 79.770%

(d) By compare their scatter plots and residual plots, both models have non-linearity, so it is not efficient to compare their plots. Thus, to choose a better model, we will use **SSE** as the main criteria since it is a measure of the discrepancy between the data and an estimation model. Lower SSE is preferred. In our case, we will prefer model 3 than model 2. Model 2 has SSE as **29.316** and untransfered SSE as **241.2044**, while model 3's SSE is **26.773** and untransfered SSE as **210.0379**. Therefore, we will **choose model 3** as our transformed scale.





Q4: Source R code

```
# ---------> complete and run the following code for this assignment  <-------
#
#
# R code for STA302 or STA1001H1F assignment 2
# copyright by Christina Deng
# date: Nov. 6, 2016


## Load in the data set
a2 = read.table("/Users/christinadeng/Desktop/Fall 2016/STA 302/Assignemnt /Assignment 2/DataA2.txt",header=T)

## Q1: fit a linear model to FEV on age

mod1 <- lm(a2$fev~a2$age)

## ==> Q1(a) produce the scatter plot (FEV vs Age) and the residual plot with fitted value
# plot the scatter plot and residual plot in one panel
par(mfrow=c(1,2))
# make a scatter plot of the data and add regression line
plot(a2$age,a2$fev, type="p",col="blue",pch=21, main="FEV vs age",
     xlab = "age", ylab = "FEV")
abline(mod1,col="red",lty=2)
# to get the residual plot vs fitted value
plot(mod1,which=1)


##==> Q1(b): boxcox transformation
library(MASS)
# get the boxcox plot
boxcox(a2$fev~a2$age)
# find exact value of lambda
a=boxcox(a2$fev~a2$age, lambda=seq(-2, 4, 0.01))
Elambda= a$x[which.max(a$y)]
Elambda


## Q2
# fit a linear model to log(FEV) on age
m2 <- lm(log(fev) ~ age, data = a2)
# plot the scatter plot and residual plot in one panel
par(mfrow=c(1,2))
# make a scatter plot of the data and add regression line
plot(a2$age,log(a2$fev), type="p",col="blue",pch=21, main="log(FEV) vs age", xlab = "age", ylab = "log(FEV)")
abline(m2,col="red",lty=2)
# make a residual plot vs fitted value of mod 2 (m2)
plot(m2, which=1)
# get information on its slope and intercept
summary(m2)
# find 95% CI in untransformed scale
i <- predict(m2, interval = "confidence", newdata = data.frame(age = c(8, 17, 21)))
exp(i)
# find 95% PI in untransformed scale
j <- predict(m2,interval = "prediction", newdata = data.frame(age = c(8, 17, 21)))
exp(j)
```

```r
## Q3:
# fit a linear model to log(FEV) on log(age)
m3 <- lm(log(fev) ~ log(age),data=a2)
# plot the scatter plot and residual plot in one panel
par(mfrow=c(1,2))
# make a scatter plot of the data and add regression line
plot(log(a2$age),log(a2$fev), type="p",col="blue",pch=21, main="log(FEV) vs log(age)", xlab = "log(age)", ylab = "l
abline(m3,col="red",lty=2)
# make a residual plot vs fitted value of mod 3 (m3)
plot(m3, which=1)
# get CI of the regression coefficients.
confint(m3, level=0.95)
# get information on its slope and intercept
summary(m3)
# compare SSE of m2 and m3
anova(m2)
anova(m3)
untransfer2 = sum( (a2$fev-exp(m2$fitted))^2)
untransfer3 = sum( (a2$fev-exp(m3$fitted))^2)
untransfer2
untransfer3
```

A3: Determinants of Plasma Level

Last name: Deng
First name: Qi (Christina)
Student ID: 1001142408
Course section: STA302H1F-L5101

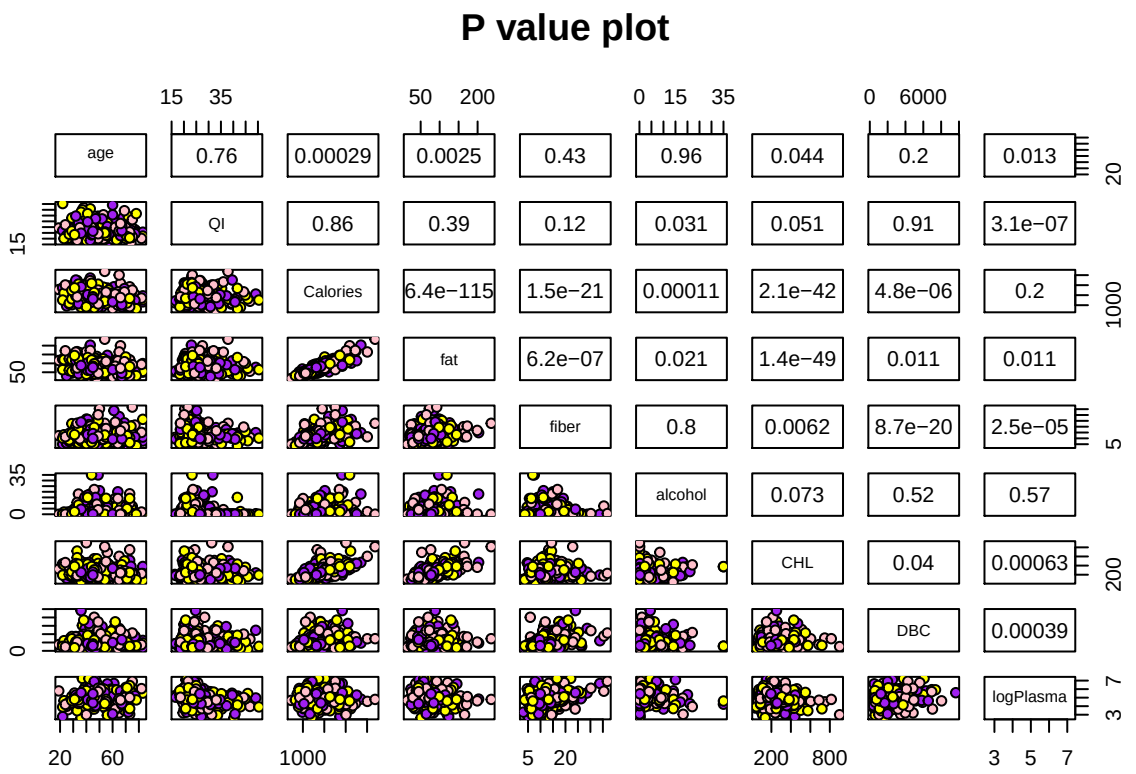
Dec. 1st, 2016

Q1: Observation of pairwise correlations and scatterplots

H0: Coefficient is 0; no correlation between the two variables.

Ha: Coefficient is not 0; there is correlation between the two variables .

++



++

By definition, if p-value < 0.01, the pairs of variables have **strong evidence** of a linear relationship; if $0.01 <$ p-value < 0.05, there is **moderate evidence**. Thus, by the graph "P value plot", we can conclude that:

strong evidence of a linear relationship:

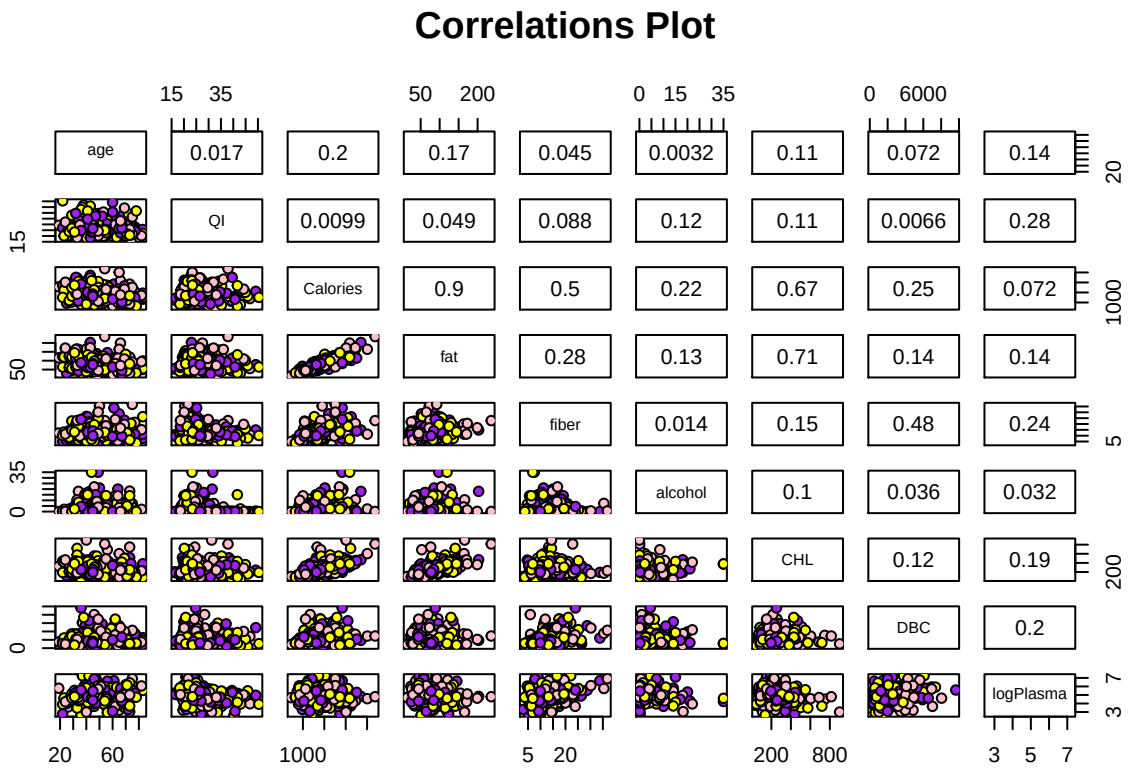
- Calories & age (p-value = 0.00029);
- Calories & fat (p-value=6.4e-115);
- Calories & alcohol (p-value = 0.00011);
- Calories & fiber (p-value = 1.5e-21);
- Calories & CHL (p-value = 2.1e-42);
- Calories & DBC (p-value = 4.8e-06)

- fat & fiber (p-value = 6.2e-07);
- fat & CHL (p-value = 1.4e-49);
- fat & age (p-value=0.0025);
- fiber & CHL (p-value = 0.0062);
- fiber & logPlasma (p-value = 2.5e-05);
- fiber & DBC (p-value = 8.7e-20);
- logPlasma & QI (p-value = 3.1e-07);
- logPlasma & CHL (p-value = 0.00063);
- logPlasma & DBC (p-value = 0.00039)

moderate evidence of a linear relationship:

- QI & alcohol (p-value = 0.031);
- fat & alcohol (p-value = 0.021);
- fat & DBC (p-value = 0.011);
- CHL & DBC (p-value = 0.04);
- CHL & age (p-value = 0.044);
- logPlasma & age (p-value = 0.013);
- logPlasma & fat (p-value = 0.011)

++



++

By our previous knowledge, we conclude that moderate correlation exists when $0.4 < |r| <= 0.6$; and have strong correlation if $|r| > 0.6$. Therefore, by observation of the pairwise correlations and scatterplots and P-value Plot above, we can conclude the following:

- Calories & Fat have **strong evidence** against H0 and with **strong correlation** (r=0.9); the hypothesis test agrees with the correlation

- Calories & CHL have **strong evidence** against H0 and with **strong correlation** (r=0.67)
- Fat & CHL have **strong evidence** against H0 and with **strong correlation** (r=0.71)
- Calories & Fiber **strong evidence** against H0 and with **moderate correlation** (r=0.5)
- Fiber & DBC **strong evidence** against H0 and with **moderate correlation** (r=0.48)
- Also, from the scatterplots, we observe that the greater the correlation, the more linear the regression line will be on the plot. The smaller the correlation, the linear pattern of the regression line is less obvious.

Q2: Fit the three regression equations

Fit the three regression equations

- Mod 1 (with Calories only):

$$\widehat{log(plasma)} = 5.107 - (8.586e-05)\text{Calories}$$

- Mod 2 (with calories & fat):

$$\widehat{log(plasma)} = 5.0273755 + 0.0003506 * \text{Calories} - 0.0090997 * \text{fat}$$

- Mod 3 (with calories & QI):

$$\widehat{log(plasma)} = 6.030 + (-8.252e-05)\text{Calories} - (3.550e-02)\text{QI}$$

A:

Mod 1 (with calories only) has p-value for coefficient of Calories as 0.201 and coefficient of Calories as -8.586e-05. Thus, we are fail to reject H0, which means that B1 = 0, so there is no correlation between Calories and logPlasma in Mod 1;

Mod 2 (with calories and fat) has p-value for coefficient of Calories as 0.02136 and coefficient of Calories as 0.0003506. Thus, we reject H0, which means that B1 is not 0, so there is correlation between Calories and logPlasma in Mod 2.

Mod 3 (with calories and Quetelet index) has p-value for coefficient of Calories as 0.201 and coefficient of Calories as -8.252e-05. Thus, we are fail to reject H0, which means that B1 = 0, so there is no correlation between Calories and logPlasma in Mod 3;

To summarize, from Q1, we conclude that pair of Calories & fat has strong evidence against H0 and with strong correlation, thus multicollinearity occurs and affects the p-value and coefficient of Calories. As we can see, Mod 2 has p-value for coefficient of Calories much lower than Mod 1, and Mod 2's coefficient of Calories is much higher than Mod 1's. This implies that adding fat into our model, correlation between Calories and logPlasma significantly enhance. Mod 3 has p-value and coefficient of Calories similiar with Mod 1, this happens since for pair of calories and QI, there is no evidence against H0, thus the two variables have no correlation, so that adding QI into the model does not influence p-value and coefficient of Calories that much. In general, the difference in (2) and (3) results in the difference of multicolinearity of the two pairs of variables.

Q3: Fit the regression with all 11 possible predictor variables, and find the important predictors of the log of plasma

```
commod <- lm(logPlasma ~ age + factor(gender) + factor(smoke) + QI + factor(Vitamin) + Calories + fat +
summary(commod)
```

```
##
## Call:
## lm(formula = logPlasma ~ age + factor(gender) + factor(smoke) +
##     QI + factor(Vitamin) + Calories + fat + fiber + alcohol +
##     CHL + DBC, data = a3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.91040 -0.36341  0.02106  0.40162  1.98162
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.376e+00  2.758e-01  19.490  < 2e-16 ***
## age               5.131e-03  2.952e-03   1.738   0.0832 .
## factor(gender)M  -2.585e-01  1.262e-01  -2.048   0.0414 *
## factor(smoke)1   -2.501e-01  1.162e-01  -2.152   0.0322 *
## QI               -3.179e-02  6.562e-03  -4.844 2.03e-06 ***
## factor(Vitamin)1  1.616e-01  8.080e-02   2.000   0.0463 *
## Calories         -7.220e-05  1.935e-04  -0.373   0.7094
## fat              -5.683e-04  3.126e-03  -0.182   0.8559
## fiber             2.643e-02  1.097e-02   2.410   0.0165 *
## alcohol           9.604e-04  8.671e-03   0.111   0.9119
## CHL              -5.221e-04  4.267e-04  -1.224   0.2220
## DBC               5.327e-05  2.978e-05   1.789   0.0747 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6738 on 303 degrees of freedom
## Multiple R-squared:  0.2308, Adjusted R-squared:  0.2029
## F-statistic: 8.265 on 11 and 303 DF,  p-value: 1.101e-12
```

A:

By definition, whichever variables have p-value < 0.05 are the significant ones. Thus, gender(p-value = 0.0414), smoke (p-value = 0.0322), QI (p-value = 2.03e-06), Vitamin (p-value = 0.0463) and fiber (p-value = 0.0165) are important predictors of the log of plasma. QI has outstandingly smaller p-value (with three stars), thus it is the most significant one.

Q4: Find a parsimonious model by applying stepwise procedure

Q:

What model does stepwise regression produce for this data? Are the indenpendent variables in the final model that seemed to be important in the previous question ?

A:

By stepwise method, we end up with model with variables QI + fiber + Calories + factor(smoke) + factor(Vitamin) + DBC + factor(gender) + age. This is where we reach the smallest AIC. These 8 variables contain all the 5 important predictors we find in Q3, which are QI, fiber, smoke, Vitamin and gender. These important predictors decrease the AIC of the model by making the model better off. Also, by checking the p-value of Calories, DBC and age (the three variables in the model but not indicate as significant ones in Q3), we notice that the p-value of DBC and age are also relatively small. Calories has larger p-value, but since it has strong correlation with fat as we shown in Q1, and by Q2, we shown that multicolinearity will likely to lower the p-value, thus, Calories is also included in the final model.

Q5: Source R code

```
# ---------> complete and run the following code for this assignment  <-------
#
#
# R code for STA302 or STA1001H1F assignment 3
# copyright by Christina Deng
# date: Nov. 30, 2016


## Load in the data set
a3 <- read.table("/Users/christinadeng/Desktop/Fall 2016/STA 302/Assignemnt /Assignment 3/a3data.txt",sep="",header
str(a3)
is.factor(a3$gender)              # TRUE
is.factor(a3$smoke)               # FALSE
a3$smoke = as.factor(a3$smoke)    # convert smoke to a factor variable
is.factor(a3$smoke)               # TRUE
a3$logPlasma <- log(a3$plasma)
head(a3)

## Q1:
# get the p-value of these pairs of variables
Find_Pvalue <- function(x, y, ...) {
  horizontal <- (par("usr")[1] + par("usr")[2]) / 2;
  vertical <- (par("usr")[3] + par("usr")[4]) / 2;
  text(horizontal, vertical, format(cor.test(x,y)$p.value, digits=2))
}
pairs(a3[, c(1,4,6:11,13)], main = "P value plot", pch = 21,
bg = c("pink","yellow","purple"), upper.panel=Find_Pvalue)

# get the correlation of these pairs of variables
Find_correlation <- function(x, y, ...) {
  horizontal <- (par("usr")[1] + par("usr")[2]) / 2;
  vertical <- (par("usr")[3] + par("usr")[4]) / 2;
  text(horizontal, vertical, format(abs(cor(x,y)), digits=2))
}
pairs(a3[, c(1,4,6:11,13)], main = "Correlations Plot", pch = 21,
bg = c("pink","yellow","purple"), upper.panel=Find_correlation)

# Strong correlation: Calories & Fat (0.9); Calories & CHL(0.67); Fat & CHL(0.71)
# Moderate correlation: Calories & Fiber (0.5); Fiber & DBC (0.48)


## Q2
# Fit the three regression equations
mod1 <- lm(logPlasma ~ Calories, data = a3)
mod2 <- lm(logPlasma ~ Calories + fat, data = a3)
mod3 <- lm(logPlasma ~ Calories + QI, data = a3)
summary(mod1)   #p-value:0.2011; coeff of Calories: -8.586e-05
summary(mod2)   #p-value:0.002872; coeff of Calories: 0.0003506
summary(mod3)   #p-value:-8.252e-05; coeff of Calories: 9.192e-07

## Q3:
# fit the complete modeal
commod <- lm(logPlasma ~ age + factor(gender) + factor(smoke) + QI + factor(Vitamin) + Calories + fat + fiber + alc
summary(commod)

## Q4
```

```r
# no predictor in the model
nullmod <- lm(logPlasma ~ 1, data = a3)
# with all predictors in the model
fullmod <- lm(logPlasma ~ age + factor(gender) + factor(smoke) + QI + factor(Vitamin) + Calories + fat + fiber + al
# stepwise method: apply both directions method
bothways = step(nullmod ,scope=list(lower=formula(nullmod),upper=formula(fullmod)), direction="both")
formula(bothways)
```