Homework2 – Multiple Regression

Xuan Yu

9/17/2018

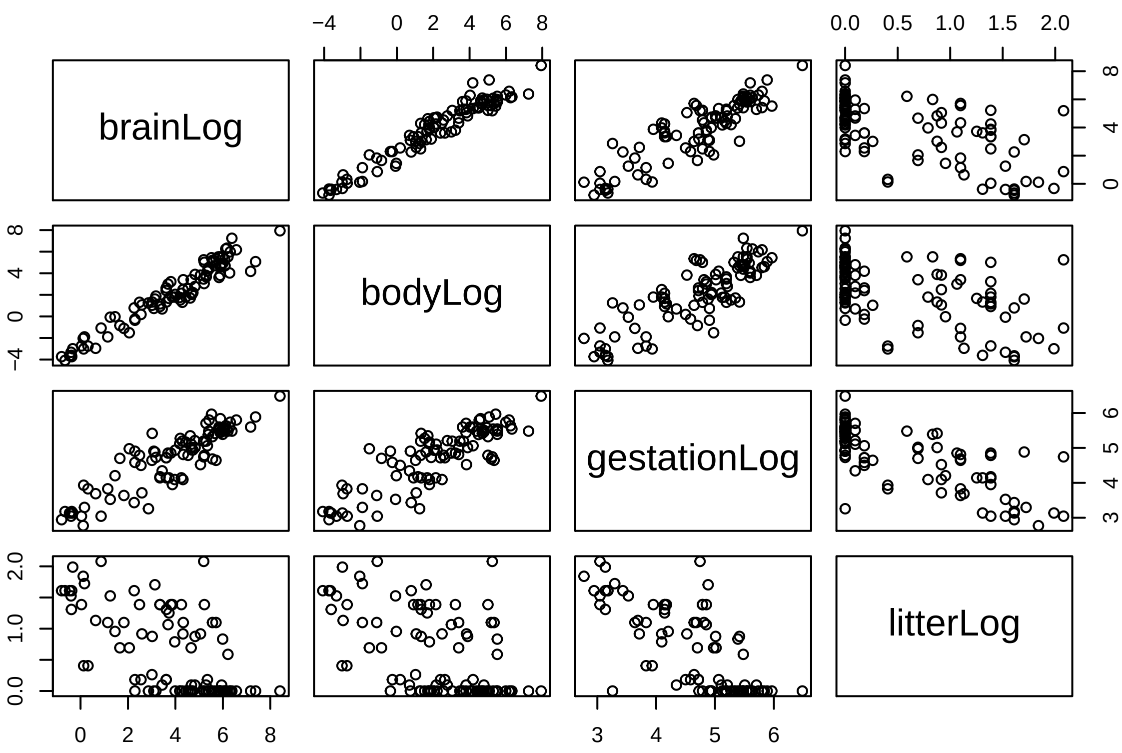
1. Brain Weights

Part A:

```
brain_data <- read.csv("/Users/xuanyu/Desktop/MIDS courses/data modeling/HW/HW2/Ex0912.csv")
brain_data$brainLog <- log(brain_data$Brain)
brain_data$bodyLog <- log(brain_data$Body)
brain_data$gestationLog <- log(brain_data$Gestation)
brain_data$litterLog <- log(brain_data$Litter)
```

Here's the matrix of data:

```
pairs(brain_data[,7:10])
```



Part B:

Here is the summary of the linear model and the confidence interval:

```
lm_brain <- lm(brainLog ~ bodyLog + gestationLog + litterLog, data = brain_data)
summary(lm_brain)
```

```
##
## Call:
```

```
## lm(formula = brainLog ~ bodyLog + gestationLog + litterLog, data = brain_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95415 -0.29639 -0.03105  0.28111  1.57491
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.85482    0.66167   1.292  0.19962
## bodyLog       0.57507    0.03259  17.647  < 2e-16 ***
## gestationLog  0.41794    0.14078   2.969  0.00381 **
## litterLog    -0.31007    0.11593  -2.675  0.00885 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4748 on 92 degrees of freedom
## Multiple R-squared:  0.9537, Adjusted R-squared:  0.9522
## F-statistic: 631.6 on 3 and 92 DF,  p-value: < 2.2e-16
```

```r
confint(lm_brain)
```

```
##                   2.5 %      97.5 %
## (Intercept)  -0.4593167  2.16896055
## bodyLog       0.5103490  0.63979373
## gestationLog  0.1383359  0.69754827
## litterLog    -0.5403124 -0.07982996
```

Part C:

The relationship between the log brain weight and litter size appear to be stronger than the relationship between log brain weight and log litter size.

2. Brain Weights Additional

Part D:

Here is the summary of the linear model and the confidence interval when litter size is on its natural scale:

```r
lm_brain_additional <- lm(brainLog ~ bodyLog + gestationLog + Litter, data = brain_data)
summary(lm_brain_additional)
```

```
##
## Call:
## lm(formula = brainLog ~ bodyLog + gestationLog + Litter, data = brain_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93895 -0.27922 -0.00929  0.28646  1.59743
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.82338    0.66206   1.244  0.21678
## bodyLog       0.57455    0.03264  17.601  < 2e-16 ***
## gestationLog  0.43964    0.13698   3.210  0.00183 **
```

```
## Litter       -0.11038    0.04227  -2.611  0.01053 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4756 on 92 degrees of freedom
## Multiple R-squared:  0.9535, Adjusted R-squared:  0.952
## F-statistic: 629.4 on 3 and 92 DF,  p-value: < 2.2e-16
```

```r
confint(lm_brain_additional)
```

```
##                  2.5 %      97.5 %
## (Intercept)  -0.4915254  2.13829063
## bodyLog       0.5097143  0.63937813
## gestationLog  0.1675856  0.71169994
## Litter       -0.1943220 -0.02643223
```

Part E:

Interpretaion:

1.Holding all other variables constant, A 10% increase in body weight will multiply brain weight by 1.7763243 ^ log(1.10) ??? 1.056287, i.e. we expect the brain weight to increase by about 5.6287%. Its 95% confidence interval is (1.6648155, 1.895302).

2.Holding all other variables constant, A 10% increase in gestation will multiply brain weight by 1.5521527 ^ log(1.10) ??? 1.042793, i.e. we expect the brain weight to increase by about 4.2793%. Its 95% confidence interval is (1.1824465, 2.037452).

3.Holding all other variables constant, each one unit increase of litter size multiplies brain weight by 0.8954963, i.e. we expect the brain weight to decrease by about 10.45%. Its 95% confidence interval is (-0.19432204, -0.02643223).

```r
exp(lm_brain_additional$coefficients)
```

```
##  (Intercept)      bodyLog gestationLog       Litter
##    2.2781930    1.7763243    1.5521527    0.8954963
```

```r
exp(confint(lm_brain_additional)[1:3,])
```

```
##                  2.5 %   97.5 %
## (Intercept)  0.6116926 8.484921
## bodyLog      1.6648155 1.895302
## gestationLog 1.1824465 2.037452
```

```r
confint(lm_brain_additional)[4,] #confidence interval for the litter.
```

```
##       2.5 %      97.5 %
## -0.19432204 -0.02643223
```

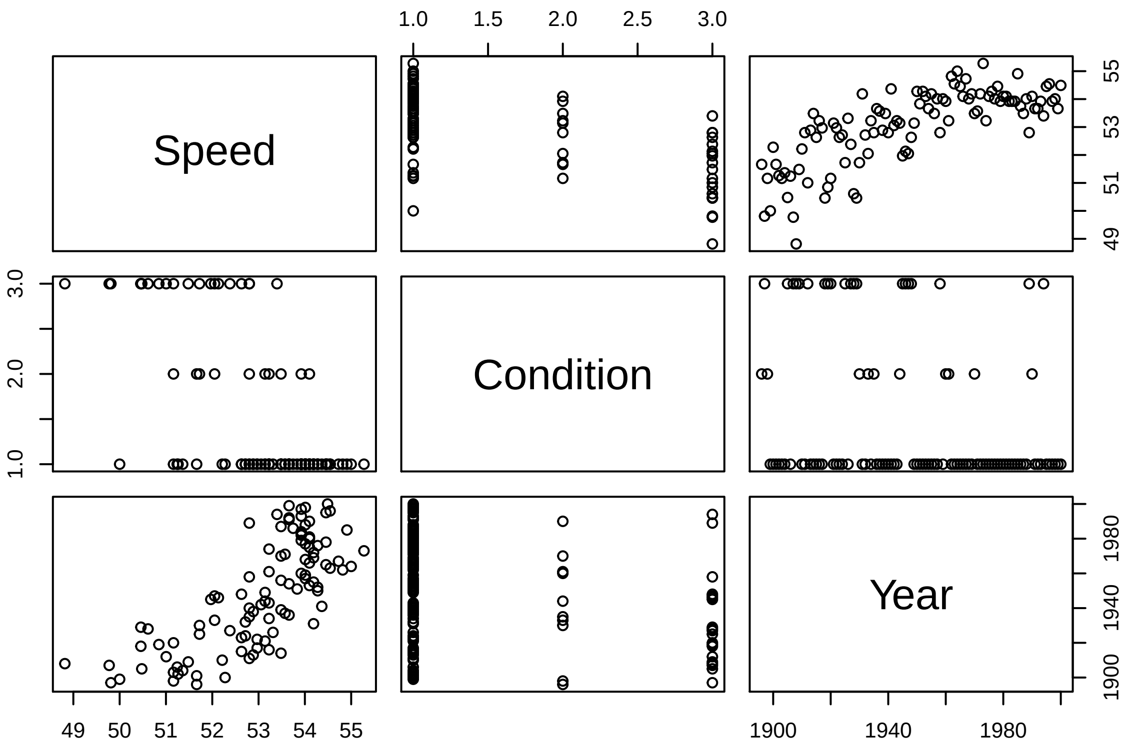
Part F:

I prefer the one in Part B. First, it's residual plots have better qualities. Second, it's easier to interpret with all variable being transformed, so interpretation of all variable are in the same format.

By the way, their R square are both large enough and are very close, so both models are fine for the R square.

3. Kentucky Derby

```r
kentucky_data <- read.csv("/Users/xuanyu/Desktop/MIDS courses/data modeling/HW/HW2/Ex0920.csv")
pairs(cbind(kentucky_data[5], kentucky_data[4], kentucky_data[2]))
```



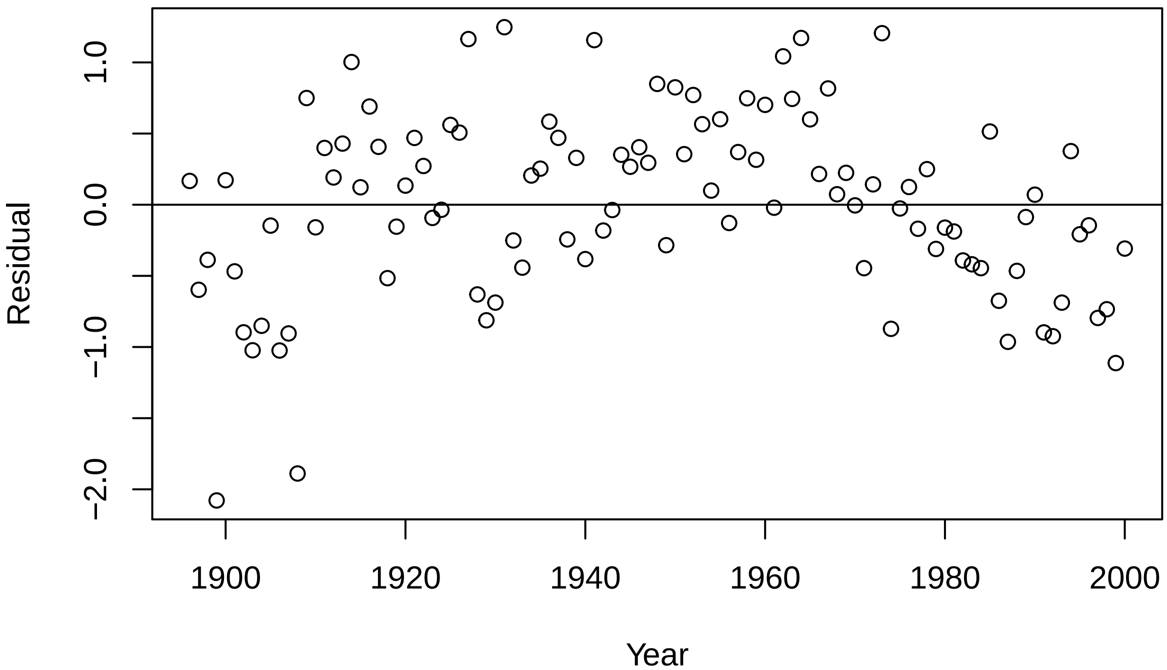
Some correlations were found among all variables.

Then do the multiple regression, we set the slowCond as the base case:

```r
n <- nrow(kentucky_data)
kentucky_data$slowCond <- rep(0, n)
kentucky_data$slowCond[kentucky_data$Condition == "slow"] = 1

kentucky_data$goodCond <- rep(0, n)
kentucky_data$goodCond[kentucky_data$Condition == "good"] = 1

kentucky_data$fastCond <- rep(0, n)
kentucky_data$fastCond[kentucky_data$Condition == "fast"] = 1

lm_kentucky <- lm(Speed ~ goodCond + fastCond + Year, data = kentucky_data)
```

We need to check the assumption, and found a quadratic trend in the residual~year scatter plot, so the model is not well-fitted. We need to consider transformations to improve the model:

```r
plot(y = lm_kentucky$residual, x = kentucky_data$Year, xlab = "Year", ylab = "Residual")
abline(0,0)
```



Then mean-center and transform the continuous predictor to improve interpretation of outputs:

```
kentucky_data$yearc <- kentucky_data$Year - mean(kentucky_data$Year)
kentucky_data$yearc2 <- kentucky_data$yearc ^ 2
```

Do the regression again:

```
lm_quadra_kentucky <- lm(Speed ~ goodCond + fastCond + yearc2 + yearc, data = kentucky_data)
```

This time, residual plots are randomly scattered, so this is the model we decied to use. Here is the relevant regression output:

```
summary(lm_quadra_kentucky)
```

```
##
## Call:
## lm(formula = Speed ~ goodCond + fastCond + yearc2 + yearc, data = kentucky_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60905 -0.30796 -0.02224  0.38851  1.10047
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.218e+01  1.405e-01 371.473  < 2e-16 ***
## goodCond     1.078e+00  2.136e-01   5.047 2.02e-06 ***
## fastCond     1.610e+00  1.439e-01  11.189  < 2e-16 ***
## yearc2      -4.214e-04  6.526e-05  -6.457 3.89e-09 ***
## yearc        2.693e-02  1.845e-03  14.596  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5492 on 100 degrees of freedom
## Multiple R-squared:  0.8365, Adjusted R-squared:  0.8299
## F-statistic: 127.9 on 4 and 100 DF,  p-value: < 2.2e-16
```

```r
confint(lm_quadra_kentucky)
```

```
##                     2.5 %        97.5 %
## (Intercept) 51.8983067264 52.4556424606
## goodCond     0.6541672128  1.5016797590
## fastCond     1.3244040523  1.8953027706
## yearc2      -0.0005508393 -0.0002918965
## yearc        0.0232687002  0.0305893943
```

Interpretation:

1.Holding all other variables constant, the winning time in good track condition is 0.6542 feet per second faster than that in the slow condition, and the winning time with fast track condition is 1.3244 feet per second faster than that in the slow condition.

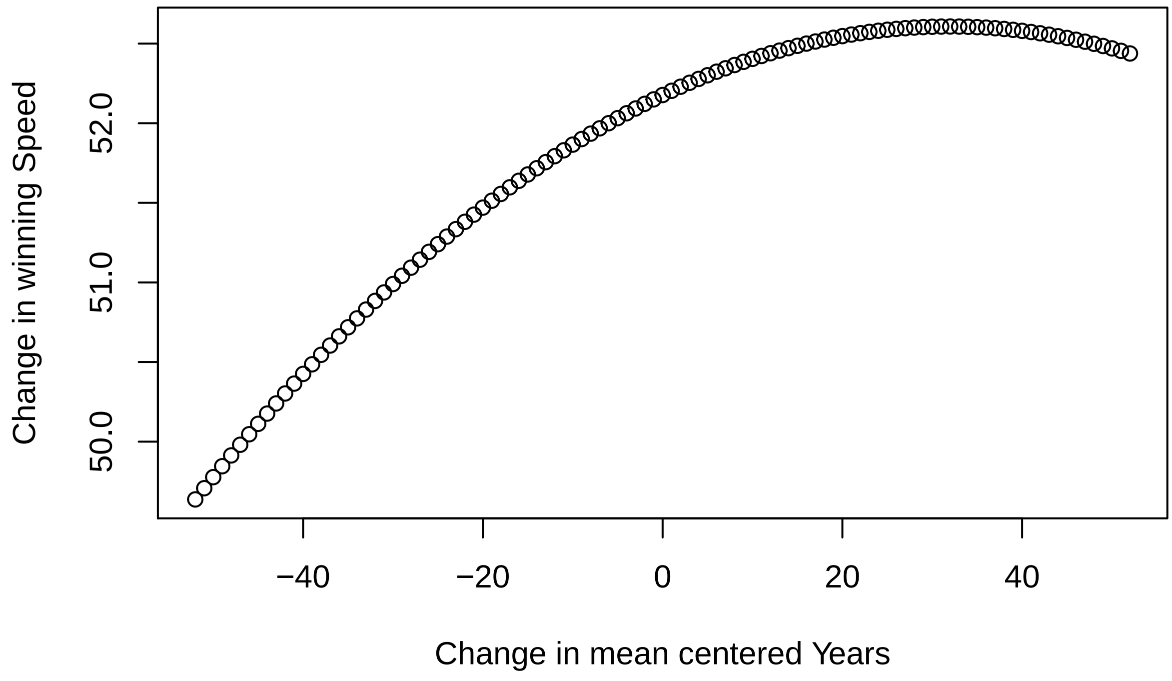
2.We use plot to interpret the Year variable.

Holding all other variables constant, the plot below describes the relation between winning speed and year in slow track condition:

```r
year_changes <- kentucky_data$yearc
coef(lm_quadra_kentucky)
```

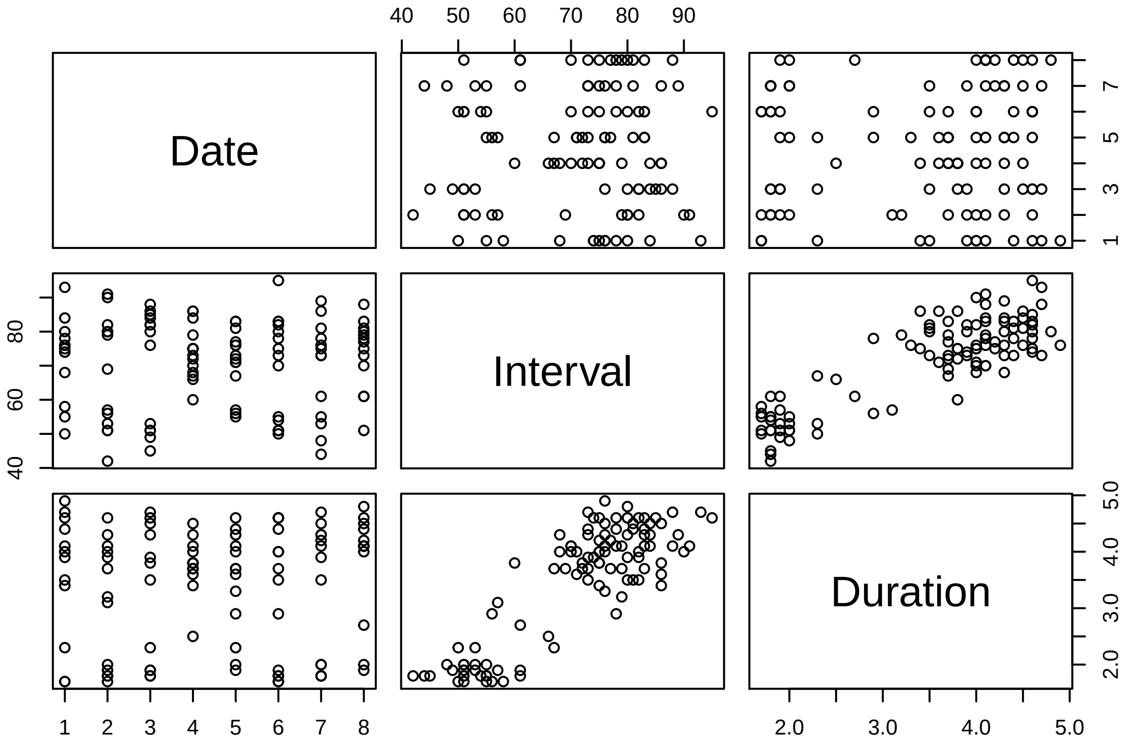
```
##   (Intercept)      goodCond      fastCond        yearc2         yearc
## 52.1769745935  1.0779234859  1.6098534114 -0.0004213679  0.0269290472
```

```r
winningSpeed <- coef(lm_quadra_kentucky)[4] * year_changes ^ 2 + coef(lm_quadra_kentucky)[5] * year_char
plot(x = year_changes, y = winningSpeed, xlab = "Change in mean centered Years", ylab = "Change in winni
```



4. Old Faithful

```r
OF_data <- read.csv("/Users/xuanyu/Desktop/MIDS courses/data modeling/HW/HW2/Ex1015.csv")
pairs(OF_data[,2:4])
```



Then do the multiple regression and the anova with Date:

```
lm_old <- lm(Interval ~ as.factor(Date) + Duration, data = OF_data)
summary(lm_old)
```

```
##
## Call:
## lm(formula = Interval ~ as.factor(Date) + Duration, data = OF_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3886  -4.7332  -0.5622   3.9759  15.9639
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        32.8770     3.0672  10.719   <2e-16 ***
## as.factor(Date)2    1.3275     2.7173   0.489    0.626
## as.factor(Date)3    0.7825     2.6994   0.290    0.773
## as.factor(Date)4    0.1625     2.6461   0.061    0.951
## as.factor(Date)5    0.2463     2.6459   0.093    0.926
## as.factor(Date)6    1.9918     2.6580   0.749    0.455
## as.factor(Date)7   -0.1700     2.7020  -0.063    0.950
## as.factor(Date)8   -0.6944     2.6957  -0.258    0.797
## Duration           10.8813     0.6622  16.431   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.866 on 98 degrees of freedom
## Multiple R-squared:  0.7408, Adjusted R-squared:  0.7196
## F-statistic:    35 on 8 and 98 DF,  p-value: < 2.2e-16
```

```r
anova(lm_old)
```

```
## Analysis of Variance Table
##
## Response: Interval
##                 Df  Sum Sq Mean Sq F value Pr(>F)
## as.factor(Date)  7   473.9    67.7   1.436 0.1996
## Duration         1 12727.9 12727.9 269.977 <2e-16 ***
## Residuals       98  4620.2    47.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We checked the assumption plots and found them to be randomly scattered, so this model is well-fitted.

Now we need to remove the Date variable to do the nested F test to determin whether Date variable is important:

```r
lm_noDate_old <- lm(Interval ~ Duration, data = OF_data)
anova(lm_old, lm_noDate_old)
```

```
## Analysis of Variance Table
##
## Model 1: Interval ~ as.factor(Date) + Duration
## Model 2: Interval ~ Duration
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     98 4620.2
## 2    105 4689.0 -7   -68.853 0.2086 0.9828
```

Conclusion: The F statistic is 0.2086 and the p value is 0.9828. The p value is much bigger than 0.05, which indicate that we didn't find obvious relationship between the interval variable and the date variable.

5. Wages and Race

```r
WR_data_raw <- read.csv("/Users/xuanyu/Desktop/MIDS courses/data modeling/HW/HW2/Ex1029.csv")
```

And we found negetive data in the Experience variable, which doesn't make sense, so we treat them as error and delete them:

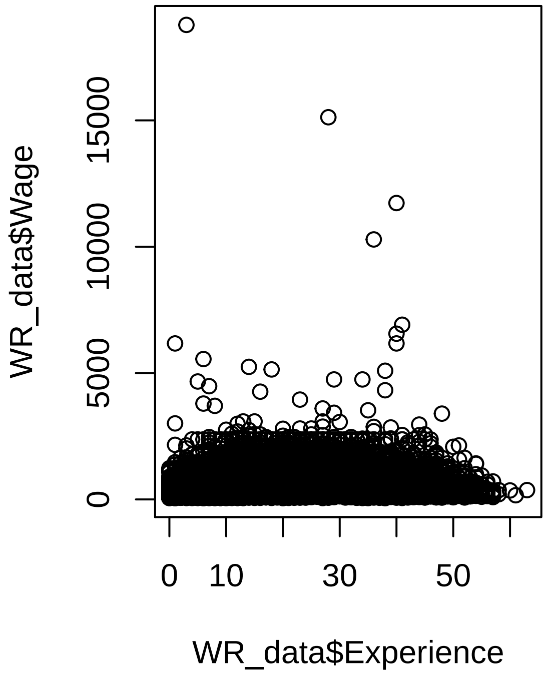
```r
summary(WR_data_raw)
```

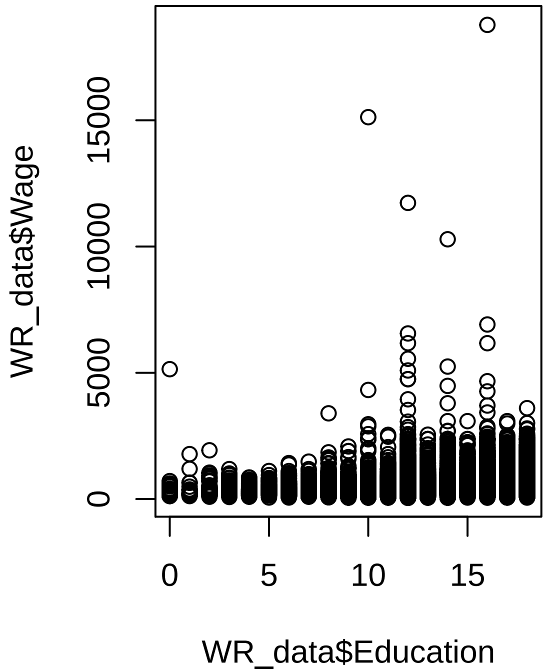
```
##        X              Wage            Education       Experience
##  Min.   :    1   Min.   :   50.39   Min.   : 0.00   Min.   :-4.00
##  1st Qu.: 6408   1st Qu.:  356.13   1st Qu.:12.00   1st Qu.: 9.00
##  Median :12816   Median :  567.23   Median :12.00   Median :16.00
##  Mean   :12816   Mean   :  640.16   Mean   :13.08   Mean   :18.59
##  3rd Qu.:19224   3rd Qu.:  826.21   3rd Qu.:16.00   3rd Qu.:27.00
##  Max.   :25631   Max.   :18777.20   Max.   :18.00   Max.   :63.00
##  Black        SMSA        Region
##  No :23643   No : 6591   MW:6226
##  Yes: 1988   Yes:19040   NE:5949
##                          S :7991
##                          W :5465
##
##
```

```
WR_data <- WR_data_raw[WR_data_raw$Experience >= 0,]
```

Plot the variables and we found quadratic trends in Experience variable and fan out trend in Education variables.

```
par(mfcol = c(1,2))
plot(WR_data$Experience, WR_data$Wage)
plot(WR_data$Education, WR_data$Wage)
```





Do the log transformation for Wage and quadratic transformation for Experience, then do the multiple regression, we set NEregion as the base case:

```
n <- nrow(WR_data)
WR_data$isBlack <- rep(0, n)
WR_data$isBlack[WR_data$Black == "Yes"] = 1
WR_data$isSMSA <- rep(0, n)
WR_data$isSMSA[WR_data$SMSA == "Yes"] = 1

WR_data$NEregion <- rep(0, n)
WR_data$NEregion[WR_data$Region == "NE"] = 1

WR_data$MWregion <- rep(0, n)
WR_data$MWregion[WR_data$Region == "MW"] = 1

WR_data$Sregion <- rep(0, n)
WR_data$Sregion[WR_data$Region == "S"] = 1

WR_data$Wregion <- rep(0, n)
WR_data$Wregion[WR_data$Region == "W"] = 1

WR_data$Education2 <- WR_data$Education ^ 2
WR_data$Experience2 <- WR_data$Experience ^ 2
WR_data$WageLog <- log(WR_data$Wage)
```

```r
lm_Log_WR <- lm(WageLog ~ Education + Experience2 + Experience + isBlack + isSMSA + MWregion + Sregion +
```

We checked the assumption plots and found them to be randomly scattered, and the R square is big enough, so this model is well-fitted.

Now we need to remove the Region variable to do the nested F test to determin whether Region variable is important:

```r
lm_Log_noRegion_WR <- lm(WageLog ~ Education + Experience2 + Experience + isBlack + isSMSA, data = WR_da
anova(lm_Log_WR, lm_Log_noRegion_WR)
```

```
## Analysis of Variance Table
##
## Model 1: WageLog ~ Education + Experience2 + Experience + isBlack + isSMSA +
##     MWregion + Sregion + Wregion
## Model 2: WageLog ~ Education + Experience2 + Experience + isBlack + isSMSA
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  25428 6629.9
## 2  25431 6666.7 -3   -36.841 47.099 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:

The F statistic is 47.099 and the p value is 2.2e-16. The p value is much smaller than 0.05, which indicate that we find obvious relationship between the Wage variable and the Region variable.

So we choose the model with Region variable as our final model, and here is the final regression output:

```r
summary(lm_Log_WR)
```

```
##
## Call:
## lm(formula = WageLog ~ Education + Experience2 + Experience +
##     isBlack + isSMSA + MWregion + Sregion + Wregion, data = WR_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7136 -0.2850  0.0349  0.3254  3.9057
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.464e+00  1.978e-02 225.681  < 2e-16 ***
## Education    8.862e-02  1.172e-03  75.597  < 2e-16 ***
## Experience2 -8.356e-04  1.958e-05 -42.681  < 2e-16 ***
## Experience   5.496e-02  9.112e-04  60.315  < 2e-16 ***
## isBlack     -2.352e-01  1.219e-02 -19.288  < 2e-16 ***
## isSMSA       1.648e-01  7.433e-03  22.167  < 2e-16 ***
## MWregion    -4.297e-02  9.374e-03  -4.584 4.58e-06 ***
## Sregion     -1.044e-01  8.931e-03 -11.695  < 2e-16 ***
## Wregion     -5.434e-02  9.667e-03  -5.621 1.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5106 on 25428 degrees of freedom
## Multiple R-squared:  0.3307, Adjusted R-squared:  0.3305
## F-statistic:  1570 on 8 and 25428 DF,  p-value: < 2.2e-16
```

```r
exp(lm_Log_WR$coefficients)
```

```
## (Intercept)   Education Experience2  Experience     isBlack      isSMSA
##  86.8242932   1.0926627   0.9991647   1.0564999   0.7904070   1.1791180
##    MWregion     Sregion     Wregion
##   0.9579356   0.9008216   0.9471120
```

```r
exp(confint(lm_Log_WR))
```

```
##                  2.5 %     97.5 %
## (Intercept) 83.5225964 90.2565080
## Education    1.0901550  1.0951762
## Experience2  0.9991264  0.9992030
## Experience   1.0546146  1.0583886
## isBlack      0.7717392  0.8095265
## isSMSA       1.1620638  1.1964225
## MWregion     0.9404947  0.9756999
## Sregion      0.8851902  0.9167291
## Wregion      0.9293355  0.9652285
```

Interpretation:

1.For the race variable, holding all other variables constant, wages of black employees tend to be 79.04% of that of nonblack employees, i.e. black people make 21.96% less money than nonblack people.

2.For education variable, holding all other variables constant, each one unit increase of education level multiplies brain weight by 1.0926627, i.e. we expect the brain weight to decrease by about 9.27%.

3.For region variable, we use NE region as the base case. Holding all other variables constant, the wage in MW region is 95.79% of that in the NE region, and the wage in S region is 90.08% of that in the NE region, and the wage in MW region is 94.71% of that in the NE region.

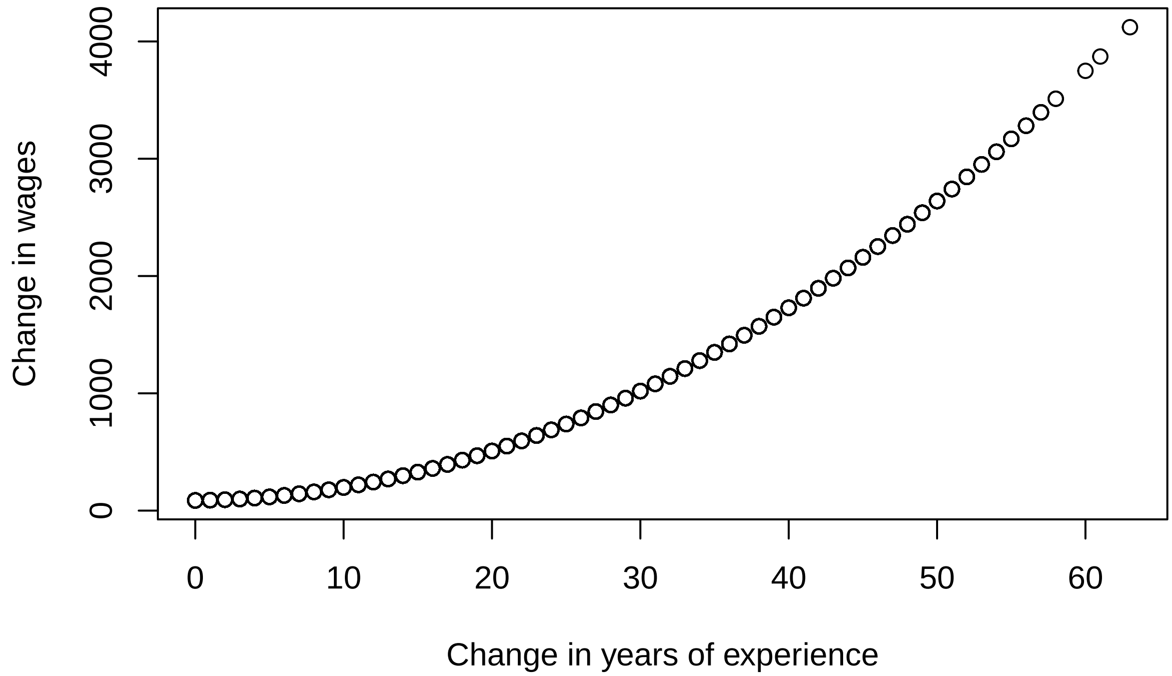
4.We use plot to interpret the Experience variable.

Holding all other variables constant and all at the base case, the plot below describes the relation between Wage and Experience:

```r
Exp_changes <- WR_data$Experience
exp(coef(lm_Log_WR))
```

```
## (Intercept)   Education Experience2  Experience     isBlack      isSMSA
##  86.8242932   1.0926627   0.9991647   1.0564999   0.7904070   1.1791180
##    MWregion     Sregion     Wregion
##   0.9579356   0.9008216   0.9471120
```

```r
avgWage <- exp(coef(lm_Log_WR))[3] * Exp_changes ^ 2 + exp(coef(lm_Log_WR))[2] * Exp_changes + exp(coef
plot(x = Exp_changes, y = avgWage, xlab = "Change in years of experience", ylab = "Change in wages")
```


Change in wages
0
1000
2000
3000
4000
0
10
20
30
40
50
60
Change in years of experience