Stats 101A Homework 3 (Lecture 1B)

Charles Liu (304804942)

February 13, 2020

Loading Necessary Packages/Data/:

Removing Missing Data:

Problem 1:

1a)

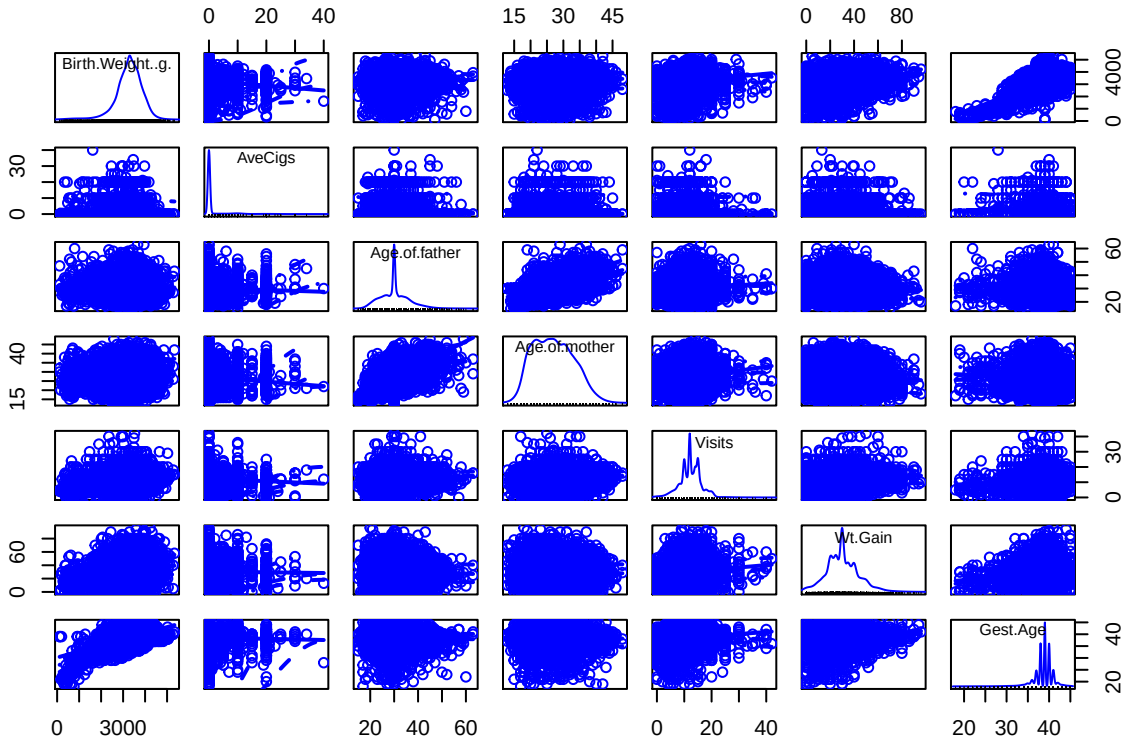
```
cormat <- round(cor(NCD[, c("Birth Weight (g)", "AveCigs", "Age of father",
                            "Age of mother", "Visits", 'Wt Gain', "Gest Age")],
                    use = "pairwise.complete.obs"), 4)
cormat
```

```
##                   Birth Weight (g) AveCigs Age of father Age of mother
## Birth Weight (g)            1.0000 -0.1073        0.0253        0.0840
## AveCigs                    -0.1073  1.0000       -0.0387       -0.0686
## Age of father               0.0253 -0.0387        1.0000        0.6556
## Age of mother               0.0840 -0.0686        0.6556        1.0000
## Visits                      0.1413 -0.0691        0.0862        0.1397
## Wt Gain                     0.1940 -0.0129       -0.0461       -0.0527
## Gest Age                    0.5965 -0.0311       -0.0436       -0.0372
##                    Visits Wt Gain Gest Age
## Birth Weight (g)   0.1413  0.1940   0.5965
## AveCigs           -0.0691 -0.0129  -0.0311
## Age of father      0.0862 -0.0461  -0.0436
## Age of mother      0.1397 -0.0527  -0.0372
## Visits             1.0000  0.1052   0.1325
## Wt Gain            0.1052  1.0000   0.1315
## Gest Age           0.1325  0.1315   1.0000
```

1b)

```
scatterplotMatrix( ~ `Birth Weight (g)` + AveCigs + `Age of father` +
                     `Age of mother` + Visits + `Wt Gain` + `Gest Age`,
                   data = NCD)
```

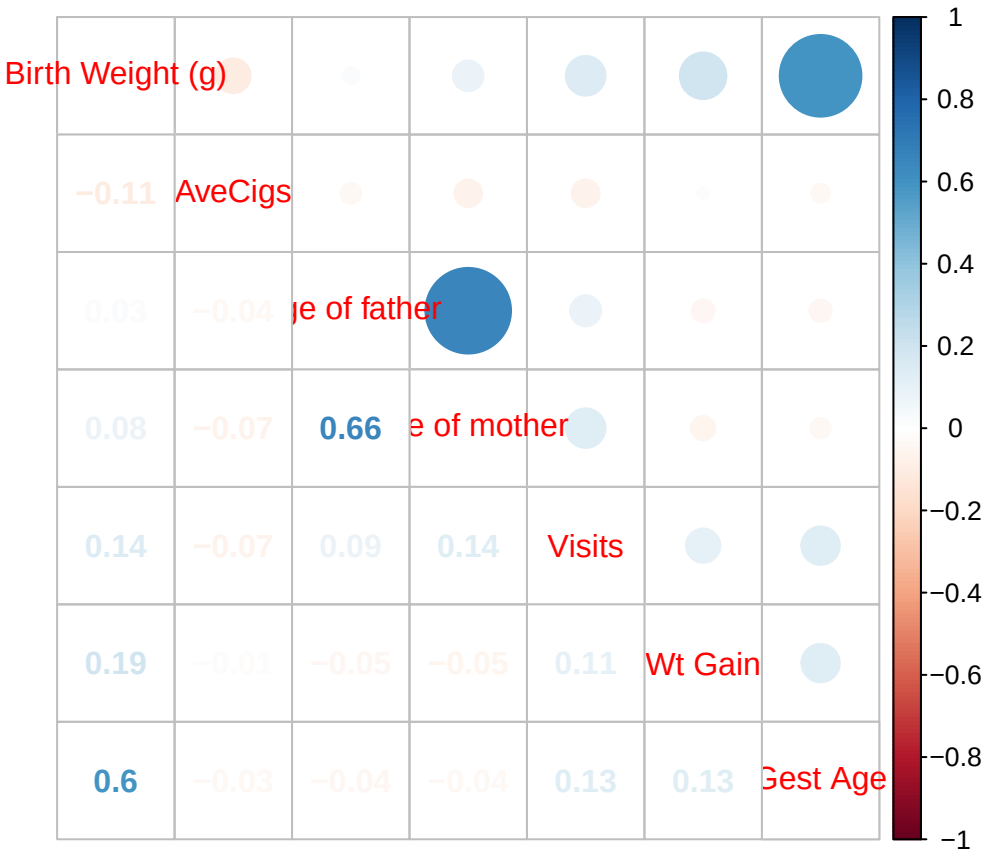
```
## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit negative part of the spread
```




```
# This portion of the code will take the longest
```

1c)

```
par(mfrow = c(1, 1))
corrplot.mixed(cormat)
```



What I noticed is that the largest circle (which means the higher the correlation is) is between "Age of Father" and "Age of Mother". This means they are highly correlated with each other. We can also see for "Visit", "Wt Gain", and "Gest Age" has a relatively smaller correlation, due to their circles being smaller. We do see that "Gest Age" and "Birth Weight (g)" have high correlation with each other because of the bigger circle.

Problem 2:

2a)

I will be choosing "Visits", "Wt Gain", and "Gest Age" for my Numerical Predictors.

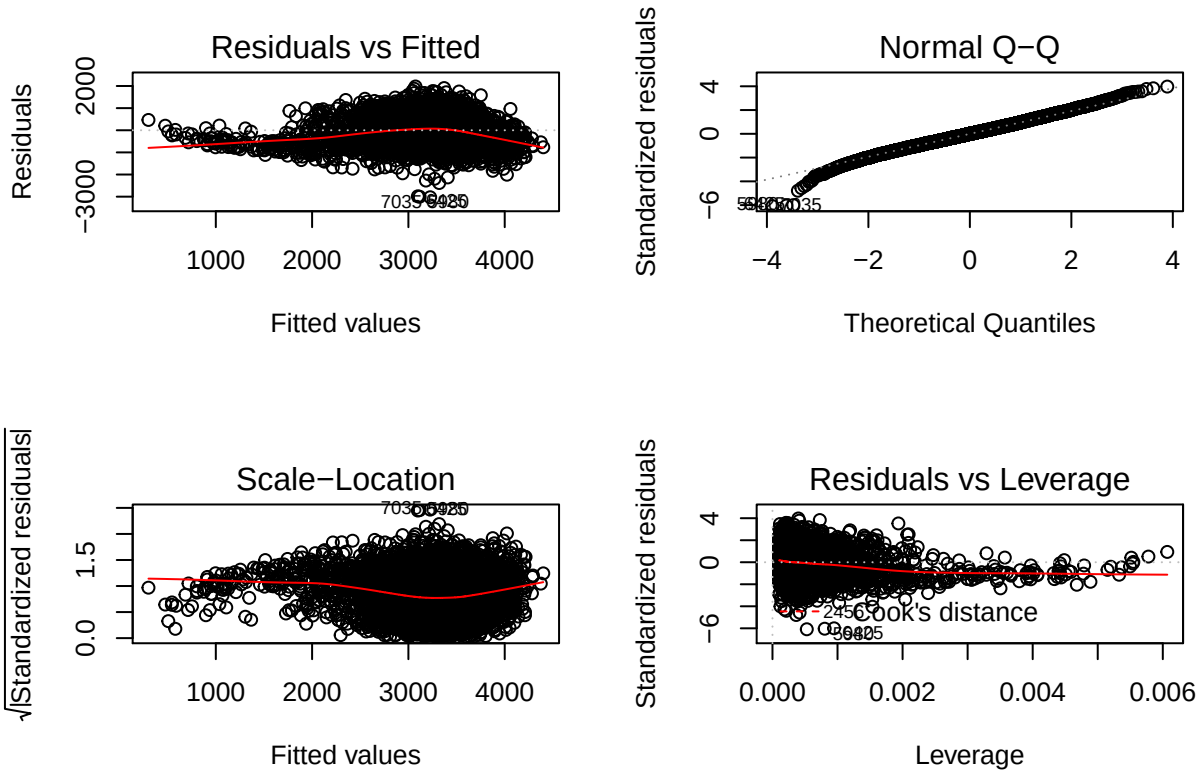
2b)

```
m1 <- lm(`Birth Weight (g)` ~ Visits + `Wt Gain` + `Gest Age`, data = NCD)
summary(m1)
```

```
##
## Call:
## lm(formula = `Birth Weight (g)` ~ Visits + `Wt Gain` + `Gest Age`,
##     data = NCD)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3027.46  -325.43    -0.29   319.98  1977.26
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2144.3973    71.3898 -30.038  < 2e-16 ***
## Visits          8.1458     1.2259   6.645 3.19e-11 ***
## `Wt Gain`       5.1290     0.3638  14.098  < 2e-16 ***
## `Gest Age`    133.5813     1.8721  71.356  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 497.6 on 9994 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.3722, Adjusted R-squared:  0.372
## F-statistic:  1975 on 3 and 9994 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(m1)
```



From our Summary, we can see that all our Predictors are significant. We would REJECT the Null Hypothesis. Our R-Squared (Multiple) is fairly good at 37.22% variation in our model. Checking our Plot and its Assumptions, we can see the "Residuals vs. Fitted" has randomness and proves it has non-linearity. Our "Normal Q-Q" plot is relatively a straight line, proving it follows a Normal Distribution. For "Scale-Location", we see they have about same variance, and "Leverages" plot tells us what the leverage values are. Overall, this satisfies our assumptions and is a rather good model to use.

2c)

```
m1_1 <- lm.beta(m1)
m1_1
```

```
##
## Call:
## lm(formula = `Birth Weight (g)` ~ Visits + `Wt Gain` + `Gest Age`,
##     data = NCD)
##
## Standardized Coefficients::
## (Intercept)      Visits   `Wt Gain`  `Gest Age`
##  0.00000000  0.05334635  0.11315756  0.57461260
```

From our lm.beta() using Model 1, we can see that the most important predictor from our three chosen is "Gest Age". It has the highest Standardized Coefficients at 0.57461260. Next most important predictor would be "Wt Gain" at 0.11315756, and then we have "Visits" at 0.05334635.

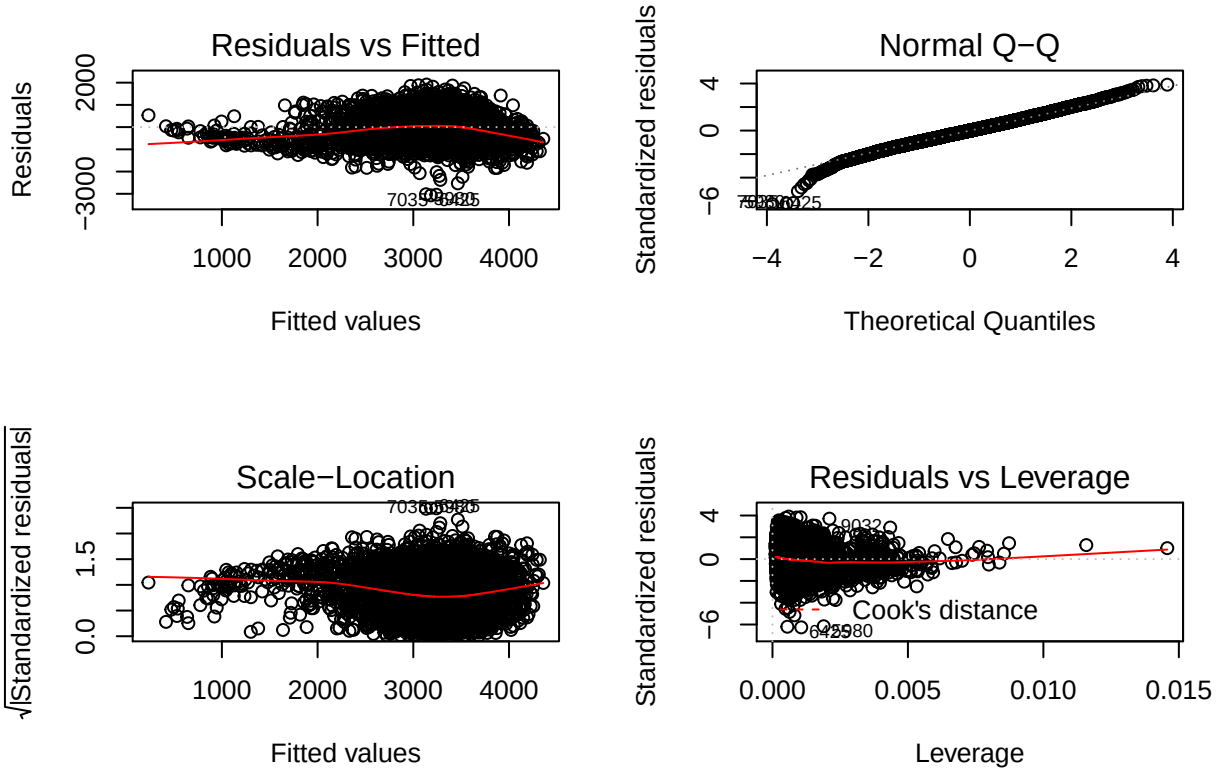
Problem 3:

3a)

```
m2 <- lm(`Birth Weight (g)` ~ AveCigs + `Age of father` + `Age of mother` +
           Visits + `Wt Gain` + `Gest Age`, data = NCD)
# We add 3 new Predictors: Visits, `Wt Gain`, `Gest Age`
```

3b)

```
par(mfrow = c(2, 2))
plot(m2)
```



Checking our Plot and its Assumptions, we can see the "Residuals vs. Fitted" has randomness and proves it has non-linearity. Our "Normal Q-Q" plot is relatively a straight line, proving it follows a Normal Distribution. For "Scale-Location", we see they have about same variance, and "Leverages" plot tells us what the leverage values are. Overall, this satisfies our assumptions and is a rather good model to use.

3c)

```
m2_1 <- lm.beta(m2)
m2_1
```

```
##
## Call:
## lm(formula = `Birth Weight (g)` ~ AveCigs + `Age of father` +
##     `Age of mother` + Visits + `Wt Gain` + `Gest Age`, data = NCD)
##
## Standardized Coefficients::
##     (Intercept)         AveCigs `Age of father` `Age of mother`
##      0.00000000     -0.07829084     -0.02900506      0.12089712
##          Visits       `Wt Gain`       `Gest Age`
##      0.03256638      0.11900368      0.57738410
```

From our lm.beta() using Model 2, we can see that the most important predictor from our three chosen is "Gest Age". It has the highest Standardized Coefficients at 0.57461260. Next most important predictor would be "Age of mother" (0.12089712) and then "Wt Gain" (0.11900368). The rest of the predictors have relatively lower Standardized Coefficients, making them not that important compared to the ones stated.

Problem 4:

4a)

```
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: `Birth Weight (g)` ~ Visits + `Wt Gain` + `Gest Age`
## Model 2: `Birth Weight (g)` ~ AveCigs + `Age of father` + `Age of mother` +
##     Visits + `Wt Gain` + `Gest Age`
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1   9994 2474890435
## 2   9991 2405207977  3  69682459 96.485 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

New predictors are significant since F-value is large, and the p-value is significant. Therefore, we reject Null Hypothesis.

4b)

Yes, it is worth to have the additional 3 predictors because it helps further explain our model. All 6 predictors, along with the F-value, are all statistically significant. This leads us to say these additional predictors are useful and worth having. We can also see that Model 2 has a better R-Squared (Multiple) than Model 1's. It is slightly better by about 1.8% increase in R-Squared.

4c)

```
summary(m2)$r.squared - summary(m1)$r.squared
```

```
## [1] 0.01767524
```

For Model 2, the R-Squared increased by about 1.767524% from Model 1.

4d)

The change is significant because Model 2 will always have the better R-Squared and be the better model since it helps give more explanation for our model. The significance for 6 predictors is better as they are all statistically significant. The improvement is very small for the R-Squared, but it is nonetheless better. All these significance lead us to say that more significant predictors lead to better fitted model. Overall, the change is significant.

Problem 5:

5a)

```
m3 <- lm(`Birth Weight (g)` ~ as.factor(`Gender of child`), data = NCD)
summary(m3)
```

```
##
## Call:
## lm(formula = `Birth Weight (g)` ~ as.factor(`Gender of child`),
##     data = NCD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3197.9  -303.7    61.0   377.3  2132.3
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3311.425      8.711 380.135   <2e-16 ***
## as.factor(`Gender of child`)2 -109.260     12.520  -8.727   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 625.6 on 9996 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.007561,   Adjusted R-squared:  0.007462
## F-statistic: 76.15 on 1 and 9996 DF,  p-value: < 2.2e-16
```

The variable "Gender of Child" does *NOT* have a decent R-Squared. It has R-Squared (Multiple) at 0.007561. Model 3 explains less than 1% variation of our model.

5b)

```
m4 <- lm(`Birth Weight (g)` ~ AveCigs + `Age of father` + `Age of mother` +
           Visits + `Wt Gain` + `Gest Age` + as.factor(`Gender of child`), data = NCD)
summary(m4)
```

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

```
## lm(formula = `Birth Weight (g)` ~ AveCigs + `Age of father` +
##     `Age of mother` + Visits + `Wt Gain` + `Gest Age` + as.factor(`Gender of child`),
##     data = NCD)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3099.21  -311.57    -3.03   310.19  1856.65
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2316.891     75.655 -30.624  < 2e-16 ***
## AveCigs                         -14.506      1.429 -10.151  < 2e-16 ***
## `Age of father`                  -2.861      1.062  -2.693   0.0071 **
## `Age of mother`                  11.876      1.059  11.211  < 2e-16 ***
## Visits                            5.165      1.216   4.248 2.18e-05 ***
## `Wt Gain`                         5.303      0.357  14.854  < 2e-16 ***
## `Gest Age`                      134.626      1.836  73.312  < 2e-16 ***
## as.factor(`Gender of child`)2  -116.981      9.762 -11.984  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 487.2 on 9990 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.3986, Adjusted R-squared:  0.3981
## F-statistic: 945.7 on 7 and 9990 DF,  p-value: < 2.2e-16
```

```
anova(m2, m4)
```

```
## Analysis of Variance Table
##
## Model 1: `Birth Weight (g)` ~ AveCigs + `Age of father` + `Age of mother` +
##     Visits + `Wt Gain` + `Gest Age`
## Model 2: `Birth Weight (g)` ~ AveCigs + `Age of father` + `Age of mother` +
##     Visits + `Wt Gain` + `Gest Age` + as.factor(`Gender of child`)
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1   9991 2405207977
## 2   9990 2371123246  1  34084730 143.61 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m4)$r.squared - summary(m2)$r.squared
```

```
## [1] 0.008645729
```

Adding the categorical predictor does improve our Model 2, but it only does it by a very small amount. We can see the F-value is higher and the p-value is statistically significant for all predictors. The R-Squared (Multiple) is increase by only 0.8645729%, which is a very small amount. Even though it is small, it still improves our model.