Assignment 2

Erik Pak

May 06, 2025

Table of Content:

Problem 1

The file bankingfull.txt attached to this assignment contains the full dataset. You analyzed a smaller set for Assignment 1. It provides data acquired from banking and census records for different zip codes in the bank's current market. Such information can be useful in targeting advertising for new customers or for choosing locations for branch offices. The data show

The data show:

- median age of the population (AGE)
- median years of education (EDUCATION)
- median income (INCOME) in $
- median home value (HOMEVAL) in $
- median household wealth (WEALTH) in $
- average bank balance (BALANCE) in $

Libraries

```
library(psych)         # used for describe
library(ggplot2)       # used for ggplot
library(ggpubr)        # combine scatter plots
library(QuantPsyc)     # normalize coefficients
library(equatiomatic)  # equation for a model
library(car)           # VIF for a model
library(corrplot)      # correlation plot
library(dplyr)         # using filter
```

Import text file

```
# set working directory
setwd("/Users/sir/Desktop/DePaul/Winter2023/DSC423/HW02")

# header in the bankfull.txt
bank <- read.table("Bankingfull.txt", header = TRUE)

# display using head()
head(bank)
```

```
##    Age Education Income HomeVal Wealth Balance
## 1 35.9      14.8  91033  183104 220741   38517
## 2 37.7      13.8  86748  163843 223152   40618
## 3 36.8      13.8  72245  142732 176926   35206
## 4 35.3      13.2  70639  145024 166260   33434
## 5 35.3      13.2  64879  135951 148868   28162
## 6 34.8      13.7  75591  155334 188310   36708
```

Descriptive Statistics

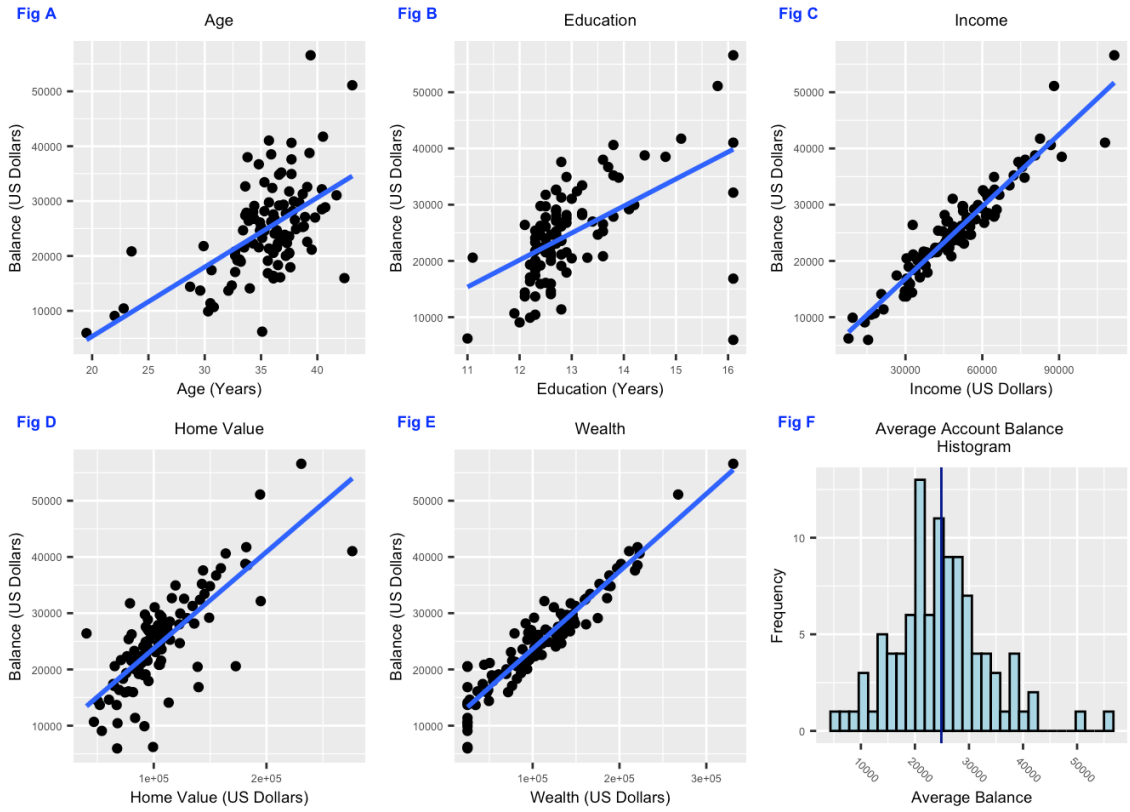
```
# descriptive stistics
describe(bank)
```

```
##           vars   n      mean       sd   median   trimmed      mad     min
## Age          1 102     35.45     3.89     36.1     35.86     2.52    19.5
## Education    2 102     12.98     1.01     12.7     12.81     0.44    11.0
## Income       3 102  48810.99 19361.88  47655.5  47832.16 18803.82  7741.0
## HomeVal      4 102 106844.73 38795.33  97743.5 102369.59 25262.76 40313.0
## Wealth       5 102 109025.76 59836.60 102348.0 104752.56 55767.26 24999.0
## Balance      6 102  24887.88  8697.81  24660.5  24546.44  6834.04  5956.0
##                max    range  skew kurtosis      se
## Age           43.1     23.6 -1.56     4.00    0.38
## Education     16.1      5.1  1.67     2.91    0.10
## Income    111548.0 103807.0  0.59     0.77 1917.11
## HomeVal   276139.0 235826.0  1.45     3.18 3841.31
## Wealth    331009.0 306010.0  0.73     0.82 5924.71
## Balance    56569.0  50613.0  0.59     1.27  861.21
```

Scatter Plots

```
# balance histogram
plot_hist_bal <- ggplot(bank, aes(x=Balance)) +
                  geom_histogram(bins = 30, color="black", fill="lightblue") +
                  geom_vline(aes(xintercept=mean(Balance)), col="darkblue") +
                  labs(title = "Average Account Balance \n Histogram",
                       x = "Average Balance", y = 'Frequency') +
                  # move the title text to the middle
                  theme(plot.title=element_text(hjust=0.5)) +
                  theme(text = element_text(size = 6))  +
                  theme(axis.title = element_text(size = 7)) +
                  theme(axis.text.x = element_text(angle = -45, hjust = .1))

# scatter plots
plot_age <- ggplot(bank, aes(x = Age, y = Balance)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Age",
                  x="Age (Years)", y = "Balance (US Dollars)") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_edu <- ggplot(bank, aes(x = Education, y = Balance)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Education",
                  x="Education (Years)", y = "Balance (US Dollars)") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_inc <- ggplot(bank, aes(x = Income, y = Balance)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Income",
                  x="Income (US Dollars)", y = "Balance (US Dollars)") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_hv  <- ggplot(bank, aes(x = HomeVal, y = Balance)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Home Value",
                  x="Home Value (US Dollars)", y = "Balance (US Dollars)") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_wlt <- ggplot(bank, aes(x = Wealth, y = Balance)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Wealth",
                  x="Wealth (US Dollars)", y = "Balance (US Dollars)") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
```

```
            theme(axis.title = element_text(size = 7))

# combine all plots
bank_com_plot <- ggarrange(plot_age, plot_edu, plot_inc, plot_hv, plot_wlt, ... =
plot_hist_bal,
                labels = c("Fig A", "Fig B", "Fig C", "Fig D", "Fig E", "Fig F"),
                font.label = list(size = 7, color = "blue"))
# plot all
bank_com_plot
```



Answer 1-a:

a) Create scatterplots to visualize the associations between bank balance and the other five variables. Discuss the patterns displayed by the scatterplot. Do the associations appear to be linear?

Fig A: This plot has linearity and two possible outliers on the top right side of the graph. Also, most points lumped between 35-40 age on a somewhat vertical and 30-35 age below the linear regression line on the scatter plot.

Fig B: This plot has less linearity than Fig A and four possible outliers on the graph's top and bottom right sides of the chart. Also, the points seem to have stretched "s" curve and the possibility of transformation on the education variable.

Fig C: It is an excellent linear relationship in this plot and the possibility of two outliers upper right side.

Fig D: This plot has a very good linear relationship, and the possibility of multiple outliers on the top right and lower left sides.

Fig E: This plot has an excellent linear relationship and the possibility of one or two outliers on the lower left side.

Correlation

```
# correlation values
corr.bank = cor(bank)

# plot correlation
corr_plot <- corrplot(corr.bank, method = 'square', addCoef.col = 'green',
                      type = 'lower',
                      title = "\n\n Correlation Plot Of Bank Data \n",)
```

Answer 1-b:

b) Compute correlation values of bank balance vs the other variables. Interpret the correlation values, and discuss which variables appear to be strongly associated.

Plot only lower portion of the correlation plot because upper half is same.

- Education & Age has a very weak positive relationship @ 0.17
- Income & Age has moderate positive relationship @ 0.48
- Income & Education has moderate positive relationship @ 0.58
- HomeVal & Age has a weak positive relationship @ 0.39
- HomeVal & Eduation has a strong positve relationship @ 0.75
- HomeVal & Income has a strong positve relationship @ 0.80
- Wealth & Age has moderate positive relationship @ 0.47
- Wealth & Education has moderate positive relationship @ 0.47
- Wealth & Income has a very strong positve relationship @ 0.95
- Wealth & HomeVal has a strong positve relationship @ 0.70
- Balance & Age has moderate positive relationship @ 0.57
- Balance & Education has moderate positive relationship @ 0.55
- Balance & Income has a very strong positve relationship @ 0.95
- Balance & HomeVal has a strong positve relationship @ 0.77
- Balance & Wealth has a very strong positve relationship @ 0.95

Multiple Linear Regression Model 1

```
# initial model
bank.model.1 <- lm(Balance ~ Education + Age + Income + HomeVal + Wealth, data=bank)

# display the actual equation for Model 1
equatiomatic::extract_eq(bank.model.1, use_coefs = FALSE)
```

$$\text{Balance} = \alpha + \beta_1(\text{Education}) + \beta_2(\text{Age}) + \beta_3(\text{Income}) + \beta_4(\text{HomeVal}) + \beta_5(\text{Wealth}) + \epsilon$$

```
# model 1 standardized coefficients
lm.beta(bank.model.1)
```

```
##  Education        Age     Income    HomeVal     Wealth
## 0.07186393 0.14239029 0.32572524 0.04095974 0.51136385
```

```
# model 1 summary
summary(bank.model.1)
```

```
##
## Call:
## lm(formula = Balance ~ Education + Age + Income + HomeVal + Wealth,
##     data = bank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5376.9 -1110.8   -77.2   872.3  7732.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.071e+04  4.261e+03  -2.514 0.013613 *
## Education    6.219e+02  3.190e+02   1.950 0.054135 .
## Age          3.187e+02  6.099e+01   5.225 1.01e-06 ***
## Income       1.463e-01  4.078e-02   3.588 0.000527 ***
## HomeVal      9.183e-03  1.104e-02   0.832 0.407505
## Wealth       7.433e-02  1.119e-02   6.643 1.85e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2056 on 96 degrees of freedom
## Multiple R-squared:  0.9469, Adjusted R-squared:  0.9441
## F-statistic: 342.4 on 5 and 96 DF,  p-value: < 2.2e-16
```

```
# model 1 ANOVA for Analysis of Variance
anova(bank.model.1)
```

```
## Analysis of Variance Table
##
## Response: Balance
##           Df     Sum Sq    Mean Sq  F value    Pr(>F)
## Education  1 2352558965 2352558965 556.7305 < 2.2e-16 ***
## Age        1 1734615269 1734615269 410.4948 < 2.2e-16 ***
## Income     1 2961454113 2961454113 700.8248 < 2.2e-16 ***
## HomeVal    1      68819      68819   0.0163    0.8987
## Wealth     1  186482706  186482706  44.1309 1.848e-09 ***
## Residuals 96  405664272    4225669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Variance Inflation Factors
vif.model.1 = vif(bank.model.1)
vif.model.1
```

```
## Education       Age    Income   HomeVal    Wealth
##  2.456706  1.342764 14.901724  4.382999 10.714276
```

Answer 1-c:

c) Fit a regression model of balance vs the other five variables (model M1). Compute the VIF statistics for each x-variable and analyze whether there is a problem of multicollinearity.

According to the Variance Inflation Factor (VIF), income and wealth have VIF over ten, meaning multicollinearity occurs in the first model. Therefore, in this model, we have Income & Wealth variables with VIF greater than ten. Also, the first model's summary report displayed t- statistics for education & HomeVal did not have significance to reject the null hypothesis. Also, in ANOVA, the HomeVal null hypothesis for the F statistics did not show enough evidence to reject the null hypothesis. Hence, we will remove one variable income and verify the t-statistic and F-statistic, including the VIF for the second model and determine the status of our second model.

Multiple Linear Regression Model 2

```
# model without Income
bank.model.2 <- lm(Balance ~ Education + Age + HomeVal + Wealth, data=bank)

# display the actual equation for Model 2
equatiomatic::extract_eq(bank.model.2, use_coefs = FALSE)
```

$$\text{Balance} = \alpha + \beta_1(\text{Education}) + \beta_2(\text{Age}) + \beta_3(\text{HomeVal}) + \beta_4(\text{Wealth}) + \epsilon$$

```
# model 2 standardized coefficients
lm.beta(bank.model.2)
```

```
##  Education        Age    HomeVal     Wealth
## 0.08902878 0.15227001 0.11085473 0.75821456
```

```
# model 2 summary
summary(bank.model.2)
```

```
##
## Call:
## lm(formula = Balance ~ Education + Age + HomeVal + Wealth, data = bank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7586.5 -1090.2    29.8   914.2  7670.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.187e+04  4.501e+03  -2.636  0.00976 **
## Education    7.704e+02  3.351e+02   2.299  0.02363 *
## Age          3.408e+02  6.428e+01   5.301 7.22e-07 ***
## HomeVal      2.485e-02  1.074e-02   2.314  0.02277 *
## Wealth       1.102e-01  5.317e-03  20.727  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2178 on 97 degrees of freedom
## Multiple R-squared:  0.9398, Adjusted R-squared:  0.9373
## F-statistic: 378.5 on 4 and 97 DF,  p-value: < 2.2e-16
```

```
# model 2 ANOVA for Analysis of Variance
anova(bank.model.2)
```

```
## Analysis of Variance Table
##
## Response: Balance
##           Df     Sum Sq    Mean Sq F value    Pr(>F)
## Education  1 2352558965 2352558965  496.01 < 2.2e-16 ***
## Age        1 1734615269 1734615269  365.73 < 2.2e-16 ***
## HomeVal    1 1055900748 1055900748  222.63 < 2.2e-16 ***
## Wealth     1 2037703795 2037703795  429.63 < 2.2e-16 ***
## Residuals 97  460065367    4742942
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Variance Inflation Factors
vif(bank.model.2)
```

```
## Education       Age   HomeVal    Wealth
##  2.415324  1.329055  3.696838  2.155681
```

Multiple Linear Regression Model 3

```
# model 3 with log transformation on education
bank.model.3 <- lm(Balance ~ log(Education) + Age + HomeVal + Wealth, data=bank)

# display the actual equation for Model 3
equatiomatic::extract_eq(bank.model.3, use_coefs = FALSE)
```

$$\text{Balance} = \alpha + \beta_1(\log(\text{Education})) + \beta_2(\text{Age}) + \beta_3(\text{HomeVal}) + \beta_4(\text{Wealth}) + \epsilon$$

```
# model 3 standardized coefficients
lm.beta(bank.model.3)
```

```
## log(Education)            Age        HomeVal         Wealth
##      0.0989804      0.1526561      0.1032532      0.7577386
```

```
# model 3 summary
summary(bank.model.3)
```

```
##
## Call:
## lm(formula = Balance ~ log(Education) + Age + HomeVal + Wealth,
##     data = bank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7217.1 -1090.6     8.6   927.5  7671.9
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.180e+04  1.152e+04  -2.761  0.00688 **
## log(Education)  1.175e+04  4.575e+03   2.569  0.01173 *
## Age             3.416e+02  6.376e+01   5.359 5.66e-07 ***
## HomeVal         2.315e-02  1.063e-02   2.178  0.03186 *
## Wealth          1.101e-01  5.280e-03  20.862  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2164 on 97 degrees of freedom
## Multiple R-squared:  0.9406, Adjusted R-squared:  0.9381
## F-statistic: 383.7 on 4 and 97 DF,  p-value: < 2.2e-16
```

```
# model 3 ANOVA for Analysis of Variance
anova(bank.model.3)
```

```
## Analysis of Variance Table
##
## Response: Balance
##                Df     Sum Sq    Mean Sq F value    Pr(>F)
## log(Education)  1 2465196734 2465196734  526.43 < 2.2e-16 ***
## Age             1 1680385448 1680385448  358.84 < 2.2e-16 ***
## HomeVal         1 1002955670 1002955670  214.17 < 2.2e-16 ***
## Wealth          1 2038066819 2038066819  435.22 < 2.2e-16 ***
## Residuals      97  454239473    4682881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Variance Inflation Factors
vif(bank.model.3)
```

```
## log(Education)            Age        HomeVal         Wealth
##       2.422591       1.324218       3.668555       2.152592
```

Answer 1-d1:

d) Apply your knowledge of regression analysis to define a better model M2, and answer the following questions:

d-1) Analyze the Coefficient of Determination R2 values and the adjusted adj-R2 values for both models M1 and M2. Which model has the largest adj-R2 value?

- M1: R-squared: 0.9469, Adjusted R-squared: 0.9441
- M2: R-squared: 0.9398, Adjusted R-squared: 0.9373
- **M3: R-squared: 0.9406, Adjusted R-squared: 0.9381**

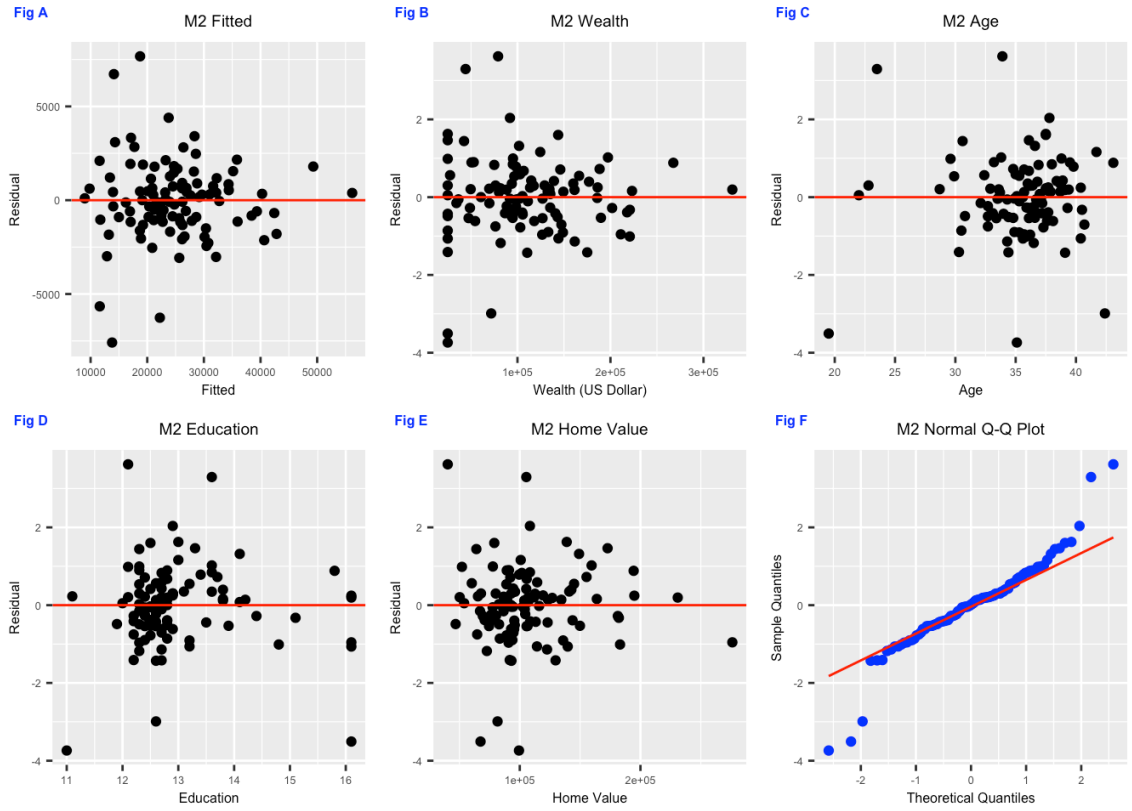
The first Model has the highest R-squared & Adjusted R-squared between the three models. However, M1 has multicollinearity, which weakens the regression model's statistical power. Therefore, I would choose M2 even having a slightly lower R-squared & Adjusted R-squared because M2 does NOT have multicollinearity according to the VIF test. Finally, I did run for fun one more Model for taking a

log of education to determine whether the performance would increase because the scatter plot somewhat displayed stretched “s” curve. I’ve created M3 for fun, and it is a slightly better model than the M2 without multicollinearity. Also, the t-statistic and F-statistic are all very significant for the new M3, and I would instead select M3 over M2 because of adjusted R-squared is higher than the M2.

Residual Analysis Model 2

```
# residual vs fitted
residual_plot <-  ggplot(bank.model.2, aes(x = .fitted, y = .resid)) +
                  geom_point() +
                  geom_hline(yintercept = 0, col = "red") +
                  labs(title="M2 Fitted",
                    x = "Fitted", y = "Residual") +
                  # move the title text to the middle
                  theme(plot.title=element_text(hjust=0.5))  +
                  theme(text = element_text(size = 6))  +
                  theme(axis.title = element_text(size = 6))

# wealth vs residuals
wealth_plot <-  ggplot(bank, aes(x = Wealth, y = rstandard(bank.model.2))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="M2 Wealth",
                    x = "Wealth (US Dollar)", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 6))

# age vs residuals
age_plot <-     ggplot(bank, aes(x = Age, y = rstandard(bank.model.2))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="M2 Age", x = "Age", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 6))

# Home Value vs residuals
hval_plot <-   ggplot(bank, aes(x = HomeVal, y = rstandard(bank.model.2))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="M2 Home Value", x = "Home Value", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 6))

# education vs residuals
educa_plot <-   ggplot(bank, aes(x = Education, y = rstandard(bank.model.2))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="M2 Education", x = "Education", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 6))

#create Q-Q plot
qq_plot <-  ggplot(bank.model.2, aes(sample=rstandard(bank.model.2))) +
            stat_qq(size=1.5, color='blue') +
            stat_qq_line(col = "red") +
            labs(title="M2 Normal Q-Q Plot",
              x = "Theoretical Quantiles", y = "Sample Quantiles") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5))  +
            theme(text = element_text(size = 6))  +
```

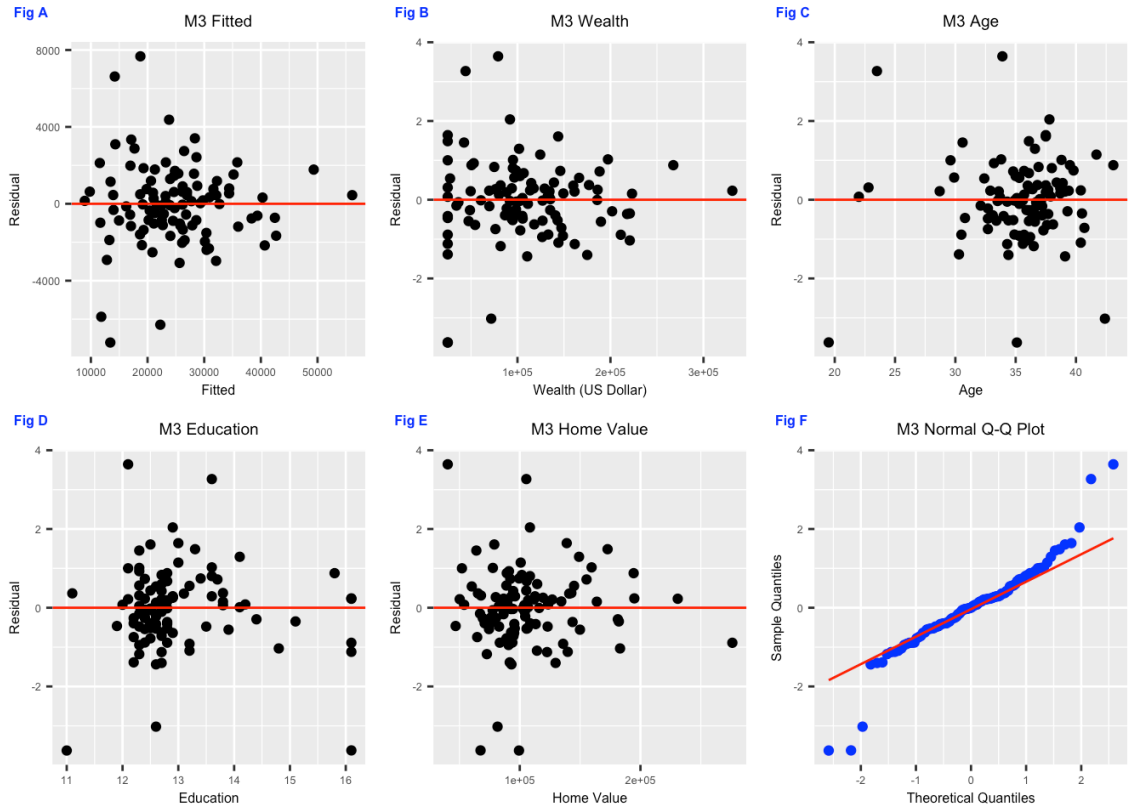
```
            theme(axis.title = element_text(size = 6))

# combine all four plots
combine_m2_plot <- ggarrange(residual_plot, wealth_plot, age_plot, educa_plot, hval_plot,qq_plot,
                  labels = c("Fig A", "Fig B", "Fig C", "Fig D", "Fig E", "Fig F"),
                  font.label = list(size = 6, color = "blue"))
# plot all
combine_m2_plot
```



Residual Analysis Model 3

```
# residual vs fitted
residual_plot3 <- ggplot(bank.model.3, aes(x = .fitted, y = .resid)) +
                  geom_point() +
                  geom_hline(yintercept = 0, col = "red") +
                  labs(title="M3 Fitted",
                    x = "Fitted", y = "Residual") +
                  # move the title text to the middle
                  theme(plot.title=element_text(hjust=0.5))  +
                  theme(text = element_text(size = 6))  +
                  theme(axis.title = element_text(size = 6))

# wealth vs residuals
wealth_plot3 <- ggplot(bank, aes(x = Wealth, y = rstandard(bank.model.3))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="M3 Wealth",
                    x = "Wealth (US Dollar)", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 6))

# age vs residuals
age_plot3 <-    ggplot(bank, aes(x = Age, y = rstandard(bank.model.3))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="M3 Age", x = "Age", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 6))

# education vs residuals
educa_plot3 <-  ggplot(bank, aes(x = Education, y = rstandard(bank.model.3))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="M3 Education", x = "Education", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 6))

# Home Value vs residuals
hval_plot3 <-  ggplot(bank, aes(x = HomeVal, y = rstandard(bank.model.3))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="M3 Home Value", x = "Home Value", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 6))

#create Q-Q plot
qq_plot3 <- ggplot(bank.model.3, aes(sample=rstandard(bank.model.3))) +
            stat_qq(size=1.5, color='blue') +
            stat_qq_line(col = "red") +
            labs(title="M3 Normal Q-Q Plot",
              x = "Theoretical Quantiles", y = "Sample Quantiles") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5))  +
            theme(text = element_text(size = 6))  +
```

```
            theme(axis.title = element_text(size = 6))

# combine all four plots
combine_m3_plot <- ggarrange(residual_plot3, wealth_plot3, age_plot3, educa_plot3, hval_plot3, qq_plot3,
                  labels = c("Fig A", "Fig B", "Fig C", "Fig D", "Fig E", "Fig F"),
                  font.label = list(size = 6, color = "blue"))
# plot all
combine_m3_plot
```



Answer 1-d2:

d-2) Create residual plots (standardized residuals vs predicted; standardized residuals vs x-variables; and normal plot of residuals). Analyze the residual plots to check if the regression model assumptions are met by the data.

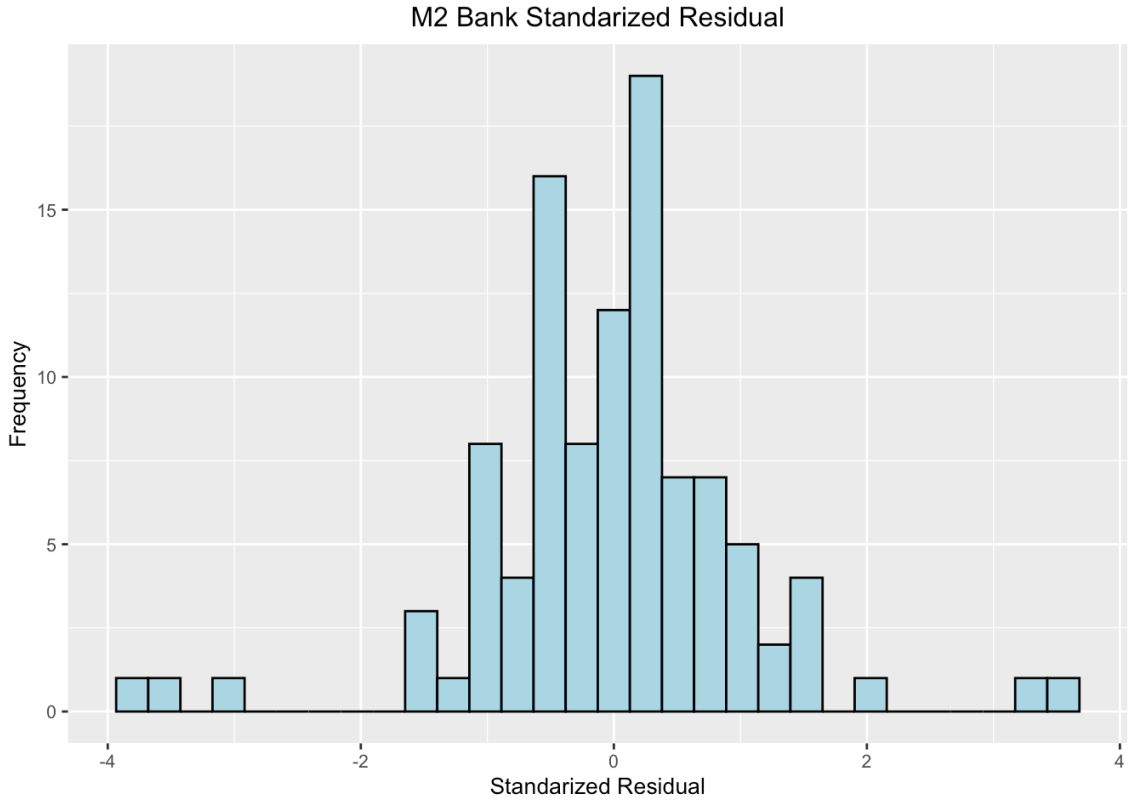
The plots of residuals against each independent variable and with the fitted values do indicate randomness for both M2 & M3. In the summary of the model, p-values are all below 0.05, and from the Analysis of Variance (ANOVA), p-values are below 0.05 as well. The Q-Q plot also displays that residuals are normally distributed, excluding five points that signal the possibility of outliers. Finally, there is an evidence that we could have five outliers in our dataset due to all the residual plots displaying the potential outliers.

Influential Points and Outliers

```
# outliers
bank_std_residual = data.frame(residual = rstandard(bank.model.2))

# display |standardized residuals| > 3
filter(bank_std_residual, abs(residual) > 3)
```

```
##      residual
## 38   3.294283
## 85  -3.507127
## 91   3.621609
## 102 -3.737934
```

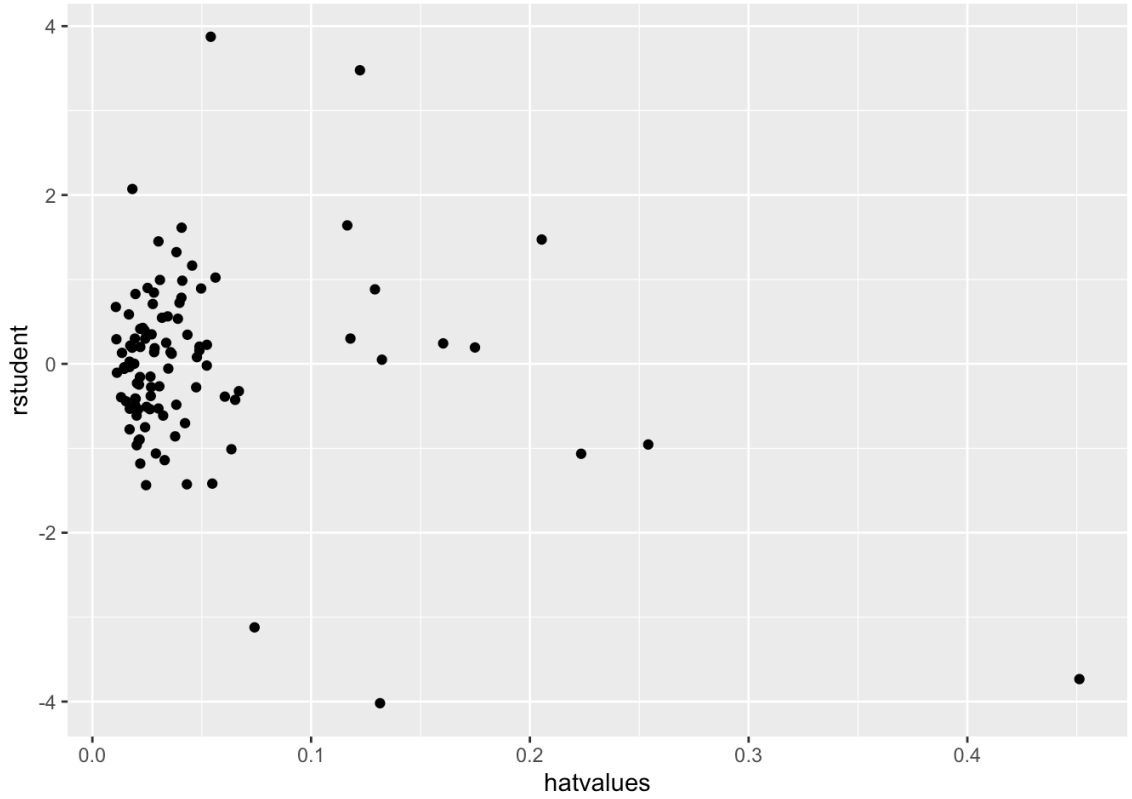
```
# historgram for outliers
ggplot(bank_std_residual, aes(x = residual)) +
  geom_histogram(bins=30, color="black", fill="lightblue") +
  labs(title = "M2 Bank Standarized Residual",
           x = "Standarized Residual", y = 'Frequency') +
      # move the title text to the middle
      theme(plot.title=element_text(hjust=0.5)) +
      theme(text = element_text(size = 10))  +
      theme(axis.title = element_text(size = 10))
```




```
# print out only observations that may be influential
summary(influence.measures(bank.model.2))
```

```
## Potentially influential observations of
##   lm(formula = Balance ~ Education + Age + HomeVal + Wealth, data = bank) :
##
##     dfb.1_  dfb.Edct dfb.Age dfb.HmVl dfb.Wlth dffit   cov.r   cook.d  hat
## 9   -0.07    0.08     0.15   -0.42     0.16    -0.56    1.35_*  0.06    0.25_*
## 12  -0.22    0.19     0.10   -0.10     0.13     0.34    1.16_*  0.02    0.13
## 15  -0.07    0.05     0.04    0.02    -0.06     0.11    1.25_*  0.00    0.16_*
## 21  -0.02    0.03    -0.01    0.00     0.05     0.09    1.27_*  0.00    0.17_*
## 38   0.41    0.11    -1.08_*  0.27    -0.18     1.30_*  0.66_*  0.30    0.12
## 59   0.42   -0.43    -0.14    0.07     0.31    -0.57    1.28_*  0.07    0.22_*
## 77   0.19   -0.30     0.04    0.65    -0.62     0.75_*  1.19_*  0.11    0.21_*
## 84   0.46   -0.18    -0.79    0.24     0.29    -0.88_*  0.70_*  0.14    0.07
## 85   1.36_* -2.62_*   1.50_*  1.67_*  -0.30    -3.39_*  0.97    2.02_*  0.45_*
## 91  -0.15    0.30     0.05   -0.76     0.39     0.93_*  0.54_*  0.15    0.05
## 98   0.06   -0.02    -0.10    0.02     0.00     0.11    1.19_*  0.00    0.12
## 100  0.01    0.00    -0.02    0.00     0.00     0.02    1.21_*  0.00    0.13
## 102 -1.05_*  1.27_*  -0.02   -1.20_*   0.88    -1.56_*  0.56_*  0.42    0.13
```

```
# plot of deleted studentized residuals vs hat values
student_hat <- data.frame(rstudent = rstudent(bank.model.2), hatvalues =  hatvalues(bank.model.2))

# plot rstudent vs hatvalues
student_hat_plot <- ggplot(student_hat, aes(x = hatvalues, y = rstudent)) + geom_point()

# plot
student_hat_plot
```



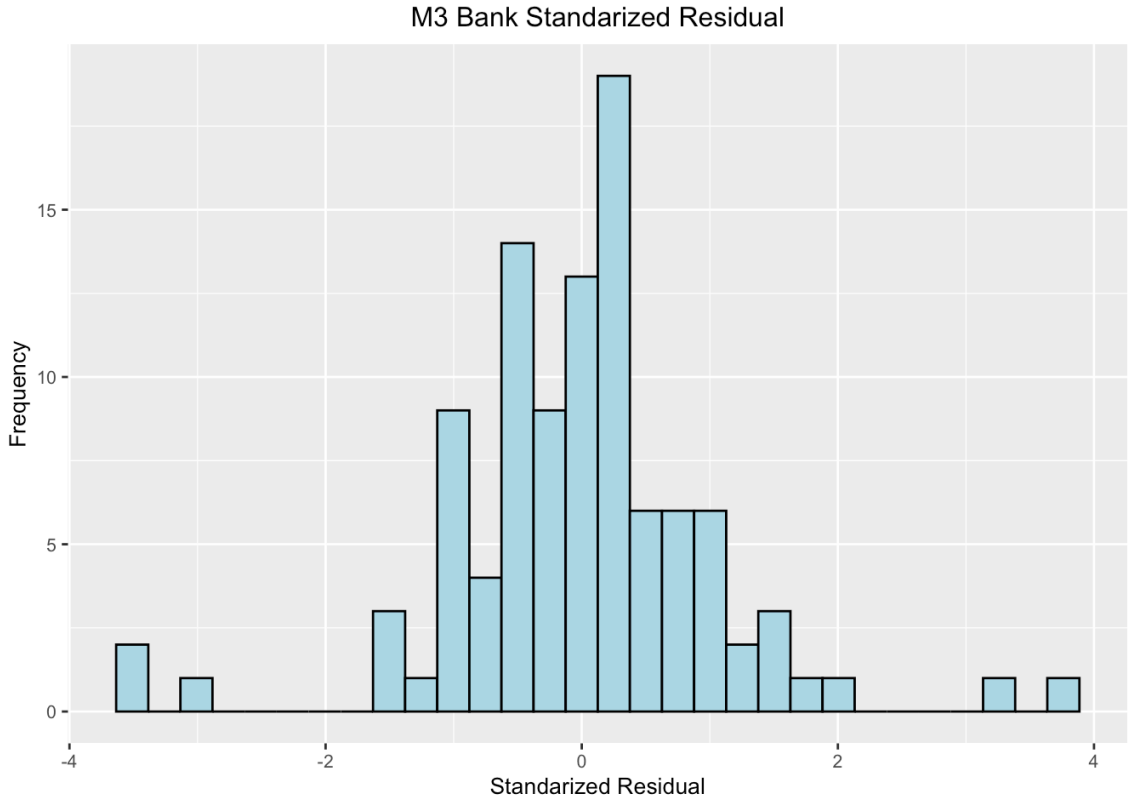
Influential Points and Outliers for Model 3

```
# M3 outliers
bank_std_residual3 = data.frame(residual = rstandard(bank.model.3))

# display |standardized residuals| > 3
filter(bank_std_residual3, abs(residual) > 3)
```

```
##      residual
## 38   3.268522
## 84  -3.021479
## 85  -3.624653
## 91   3.643096
## 102 -3.627714
```

```
# historgram for outliers
ggplot(bank_std_residual3, aes(x = residual)) +
  geom_histogram(color="black", fill="lightblue") +
  labs(title = "M3 Bank Standarized Residual",
           x = "Standarized Residual", y = 'Frequency') +
      # move the title text to the middle
      theme(plot.title=element_text(hjust=0.5)) +
      theme(text = element_text(size = 10))  +
      theme(axis.title = element_text(size = 10))
```

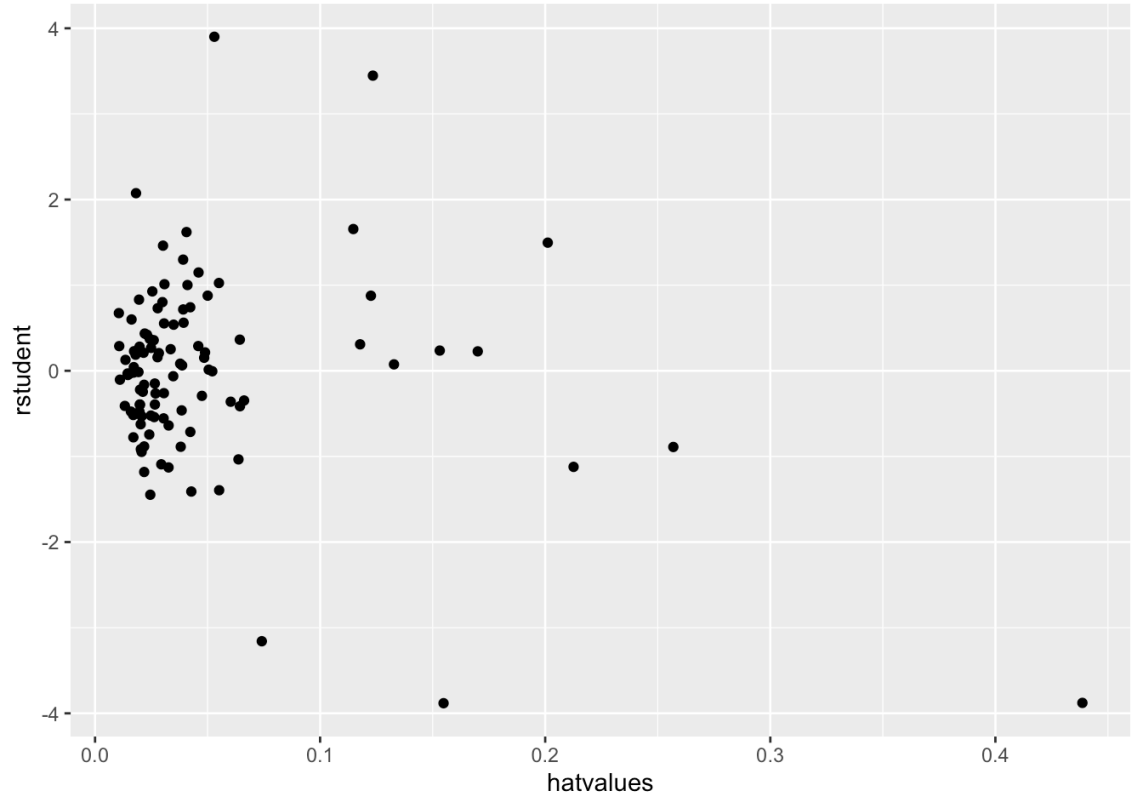
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# print out only observations that may be influential
summary(influence.measures(bank.model.3))
```

```
## Potentially influential observations of
##   lm(formula = Balance ~ log(Education) + Age + HomeVal + Wealth,      data = bank) :
##
##     dfb.1_  dfb.l(E) dfb.Age dfb.HmVl dfb.Wlth dffit   cov.r   cook.d  hat
## 9   -0.09    0.09     0.14   -0.40     0.15    -0.52    1.36_*  0.05    0.26_*
## 15  -0.06    0.05     0.04    0.02    -0.06     0.10    1.24_*  0.00    0.15_*
## 21  -0.02    0.02    -0.02    0.00     0.06     0.10    1.27_*  0.00    0.17_*
## 38   0.03    0.17    -1.07_*  0.23    -0.18     1.29_*  0.67_*  0.30    0.12
## 59   0.44   -0.43    -0.14    0.06     0.33    -0.58    1.25_*  0.07    0.21_*
## 77   0.25   -0.29     0.05    0.64    -0.62     0.75_*  1.17_*  0.11    0.20_*
## 84   0.29   -0.18    -0.80    0.25     0.30    -0.89_*  0.69_*  0.15    0.07
## 85   2.20_* -2.63_*   1.58_*  1.67_*  -0.27    -3.43_*  0.91    2.05_*  0.44_*
## 91  -0.22    0.27     0.04   -0.74     0.39     0.92_*  0.53_*  0.15    0.05
## 98   0.03   -0.02    -0.10    0.02     0.00     0.11    1.19_*  0.00    0.12
## 100  0.01    0.00    -0.03    0.00     0.00     0.03    1.21_*  0.00    0.13
## 102 -1.35_*  1.40_*  -0.01   -1.28_*   0.86    -1.66_*  0.60_*  0.48    0.15_*
```

```
# plot of deleted studentized residuals vs hat values
student_hat3 <- data.frame(rstudent = rstudent(bank.model.3), hatvalues =  hatvalues(bank.model.3))

# plot rstudent vs hatvalues
student_hat_plot3 <- ggplot(student_hat3, aes(x = hatvalues, y = rstudent)) + geom_point()

# plot
student_hat_plot3
```



Answer 1-d3:

d-3) Analyze if there are any outliers and influential points for your model. If so, what are your recommendations?

We have multiple tools for analyzing influential points in the model and retrieving various influential statistics commands are listed below for all the independent variables:

- dfbeta(model)
- covratio(model)
- dffits(model)
- cooks.distance(model)

We will take a pragmatic approach by examing High Cook's D distance (>1), High leverage hat hii value (> 0.5), High Deleted Studentized Residuals (outside (-3,3) band), and plot deleted studentized residuals vs. hat values to visualize potential, influential points.

We are checking the summary(influence.measures(model)), and this summary shows that there are 13 possible influential points in M2 & 12 possible influential points in M3, and we should examine each point individually for possible typos or outright errors. The second option is to remove these points and see if the model results change dramatically. Finally, there are four outliers from M2 and five from M3, which corresponds with the rstudent plot. At this point, I would choose M3 to predict because the M3 model is consistent with outliers and has one less possible influential point, but we required all these points from M3 to be verified from the source to validate outliers and influential points.

Standardized Coefficients

```
# model 2 standardized coefficients
lm.beta(bank.model.2)
```

```
##  Education        Age    HomeVal     Wealth
## 0.08902878 0.15227001 0.11085473 0.75821456
```

```
# model 3 standardized coefficients
lm.beta(bank.model.3)
```

```
## log(Education)           Age        HomeVal         Wealth
##      0.0989804      0.1526561      0.1032532      0.7577386
```

Answer 1-d4:

d-4) Compute the standardized coefficients and discuss which predictor has the strongest influence on balance?

The M2 & M3 standardized coefficients state that wealth has the most substantial factor in determining the balance.

Prediction

```
# predicted value
given_bank = data.frame(Age = c(34), Education = c(13) ,Income = c(64000), HomeVal = c(140000), Wealth =
c(160000))

# display the actual equation for Model 3
equatiomatic::extract_eq(bank.model.3, use_coefs = TRUE)
```

$$\widehat{\text{Balance}} = -31798.72 + 11750.84(\log(\text{Education})) + 341.64(\text{Age}) + 0.02(\text{HomeVal}) + 0.11(\text{Wealth})$$

```
# M2 & M3 - compute average balance with confidence interval @ 95%
predict(bank.model.2, given_bank, interval="confidence",level=0.95)
```

```
##     fit      lwr      upr
## 1 30848 30003.24 31692.76
```

```
#
predict(bank.model.3, given_bank, interval="confidence",level=0.95)
```

```
##        fit      lwr      upr
## 1 30821.32 29995.68 31646.95
```

Answer 1-e:

e) Use the fitted regression model from d) without removal of influential points to predict the average bank balance for a specific zip code area where there is a plan to open a new branch. Census data in that area show the following values: median age is 34 years, median education is 13 years, median income is $64,000, median home value is $140,000, median wealth is 160,000. (Note that you may not need all these values in your model). Provide predicted average bank balance and 95% confidence interval for your estimate.

M2:

- fit value: $30,848.00
- lower : $30,003.24
- upper : $31,692.76

M3:

- fit value: $30,821.32
- lower : $29,995.68
- upper : $31,646.95

The prediction range with M3 has a narrower gap in our prediction model which tells us it is a better model.

Problem 2

Analytics is used in many different sports and has become popular with the Money Ball movie. The pgatour2006.csv dataset contains data about 196 tour players in 2006. The variables in the dataset are:

- Player's name
- PrizeMoney = average prize money per tournament And a set of metrics that evaluate the quality of a player's game
- DrivingAccuracy = percent of times a player is able to hit the fairway with his tee shot
- GIR = percent of time a player was able to hit the green within two or less than par (Greens in Regulation)

- BirdieConversion = percentage of times a player makes a birdie or better after hitting the green in regulation
- PuttingAverage = putting performance on those holes where the green was hit in regulation
- PuttsPerRound = average number of putts per round (shots played on the green)

You are asked to build a model for PrizeMoney using the remaining predictors, and to evaluate the relative importance of each different aspects of a player's game on the average prize money.

import data set

```
# header in the bankfull.txt
golf <- read.csv("pgatour2006_small.csv", header = TRUE, sep = ",")

# display using str()
str(golf)
```

```
## 'data.frame':    196 obs. of  7 variables:
##  $ Name            : chr  "Aaron Baddeley" "Adam Scott" "Alex Aragon" "Alex Cejka" ...
##  $ PrizeMoney      : int  60661 262045 3635 17516 16683 107294 50620 57273 86782 23396 ...
##  $ DrivingAccuracy : num  60.7 62 51.1 66.4 63.2 ...
##  $ GIR             : num  58.3 69.1 59.1 67.7 64 ...
##  $ PuttingAverage  : num  1.75 1.77 1.79 1.78 1.76 ...
##  $ BirdieConversion: num  31.4 30.4 29.9 29.3 29.3 ...
##  $ PuttsPerRound   : num  28 29.3 29.2 29.5 28.9 ...
```

Descriptive Statistics

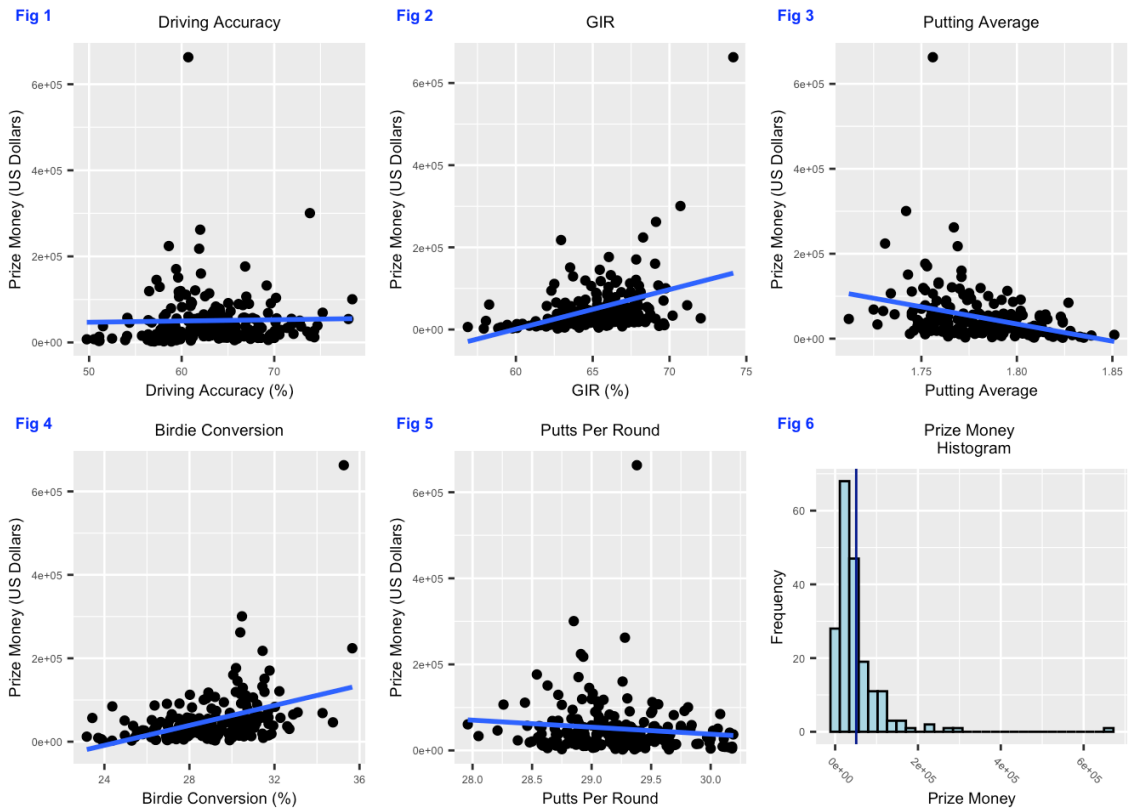
```
# descriptive stistics
describe(golf)
```

```
##                  vars   n     mean       sd   median  trimmed      mad     min
## Name*               1 196    98.50    56.72    98.50    98.50    72.65    1.00
## PrizeMoney          2 196 50891.17 63902.95 36644.50 40027.22 30153.12 2240.00
## DrivingAccuracy     3 196    63.38     5.41    63.24    63.31     5.37   49.75
## GIR                 4 196    65.19     2.72    65.35    65.27     2.61   56.87
## PuttingAverage      5 196     1.78     0.02     1.78     1.78     0.03    1.71
## BirdieConversion    6 196    28.98     2.21    29.01    29.01     2.28   23.17
## PuttsPerRound       7 196    29.20     0.44    29.19    29.19     0.42   27.96
##                      max     range  skew kurtosis      se
## Name*             196.00    195.00  0.00    -1.22    4.05
## PrizeMoney       662771.00 660531.00  5.29    42.57 4564.50
## DrivingAccuracy      78.43     28.68  0.09     0.03    0.39
## GIR                  74.15     17.28 -0.25     0.68    0.19
## PuttingAverage        1.85      0.14  0.16    -0.24    0.00
## BirdieConversion     35.66     12.49 -0.02     0.26    0.16
## PuttsPerRound        30.19      2.23  0.13    -0.10    0.03
```

Golf Scatter Plots

```
# prize money histogram
plot_hist_pri <- ggplot(golf, aes(x=PrizeMoney)) +
                  geom_histogram(color="black", fill="lightblue") +
                  geom_vline(aes(xintercept=mean(PrizeMoney)), col="darkblue") +
                  labs(title = "Prize Money \n Histogram",
                       x = "Prize Money", y = 'Frequency') +
                  # move the title text to the middle
                  theme(plot.title=element_text(hjust=0.5)) +
                  theme(text = element_text(size = 6))  +
                  theme(axis.title = element_text(size = 7)) +
                  theme(axis.text.x = element_text(angle = -45, hjust = .1))

# scatter plots
plot_acc <- ggplot(golf, aes(x = DrivingAccuracy, y = PrizeMoney)) +
            geom_point() + geom_smooth(method=lm, se=FALSE) +
            labs(title="Driving Accuracy",
                x ="Driving Accuracy (%)", y = "Prize Money (US Dollars)") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_gir <- ggplot(golf, aes(x = GIR, y = PrizeMoney)) +
            geom_point() + geom_smooth(method=lm, se=FALSE) +
            labs(title="GIR",
                x ="GIR (%)", y = "Prize Money (US Dollars)") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_put <- ggplot(golf, aes(x = PuttingAverage, y = PrizeMoney)) +
            geom_point() + geom_smooth(method=lm, se=FALSE) +
            labs(title="Putting Average",
                x ="Putting Average", y = "Prize Money (US Dollars)") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_bir <- ggplot(golf, aes(x = BirdieConversion, y = PrizeMoney)) +
            geom_point() + geom_smooth(method=lm, se=FALSE) +
            labs(title="Birdie Conversion",
                x ="Birdie Conversion (%)", y = "Prize Money (US Dollars)") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_pur <- ggplot(golf, aes(x = PuttsPerRound, y = PrizeMoney)) +
            geom_point() + geom_smooth(method=lm, se=FALSE) +
            labs(title="Putts Per Round",
                x ="Putts Per Round", y = "Prize Money (US Dollars)") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

# combine all plots
comb_golf_plot <- ggarrange(plot_acc, plot_gir, plot_put, plot_bir, plot_pur,
plot_hist_pri,labels = c("Fig 1", "Fig 2", "Fig 3", "Fig 4", "Fig 5", "Fig 6"),
```

```
                font.label = list(size = 7, color = "blue"))
# plot all
comb_golf_plot
```



Answer a:

a) Create scatterplots to visualize the associations between PrizeMoney and the other five variables. Discuss the patterns displayed by the scatterplot. Do the associations appear to be linear?

Fig 1: The driving accuracy does not appear linear but has very little linearity by closely examining the linear regression line. However, it seems odd that the regression line would have a positive slope since this could impact your score, and there is a conspicuous outlier in this plot.

Fig 2: The GIR appears to have some linear relationship, and according to the regression line, it does have a positive slope, with a conspicuous outlier sitting by itself, just like Fig 1.

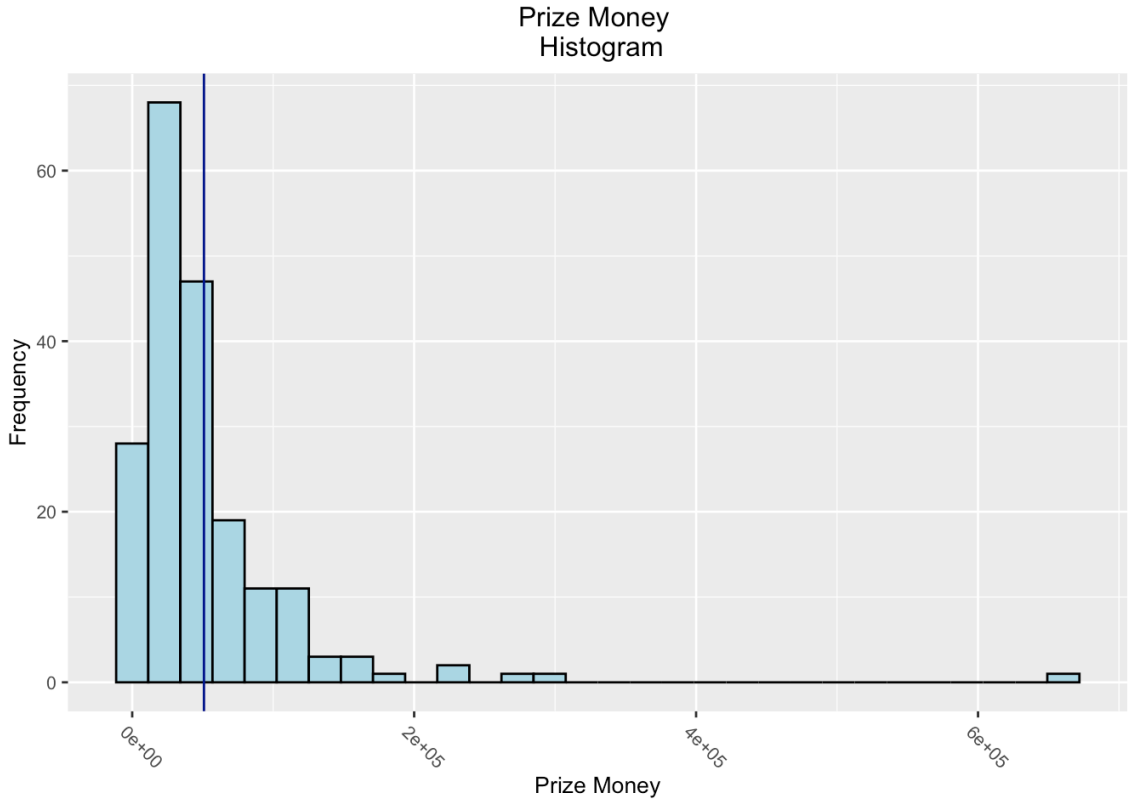
Fig 3: The putting accuracy has a slight linear relationship. According to the regression line, it is on a downward slope, which makes sense because reducing it would decrease your score. There is a point that looks like an outlier, like the previous two figures.

Fig 4: The birdie percentage has little linearity, and the regression line states it does have a positive linear relationship with the prize money. This makes sense because you want to reduce your score, and a birdie is considered -1 added to your total score.

Fig 5: The putts per round appear to have a slight linear relationship, and according to the regression line, it has a negative regression line. This makes sense because you are looking for the lowest score possible in golf, and fewer putts could result in a lower score.

Golf Histogram

```
# plot prize money histogram
ggplot(golf, aes(x=PrizeMoney)) +
      geom_histogram(bins=30, color="black", fill="lightblue") +
      geom_vline(aes(xintercept=mean(PrizeMoney)), col="darkblue") +
      labs(title = "Prize Money \n Histogram",
           x = "Prize Money", y = 'Frequency') +
      # move the title text to the middle
      theme(plot.title=element_text(hjust=0.5)) +
      theme(text = element_text(size = 10))  +
      theme(axis.title = element_text(size = 10)) +
      theme(axis.text.x = element_text(angle = -45, hjust = .1))
```



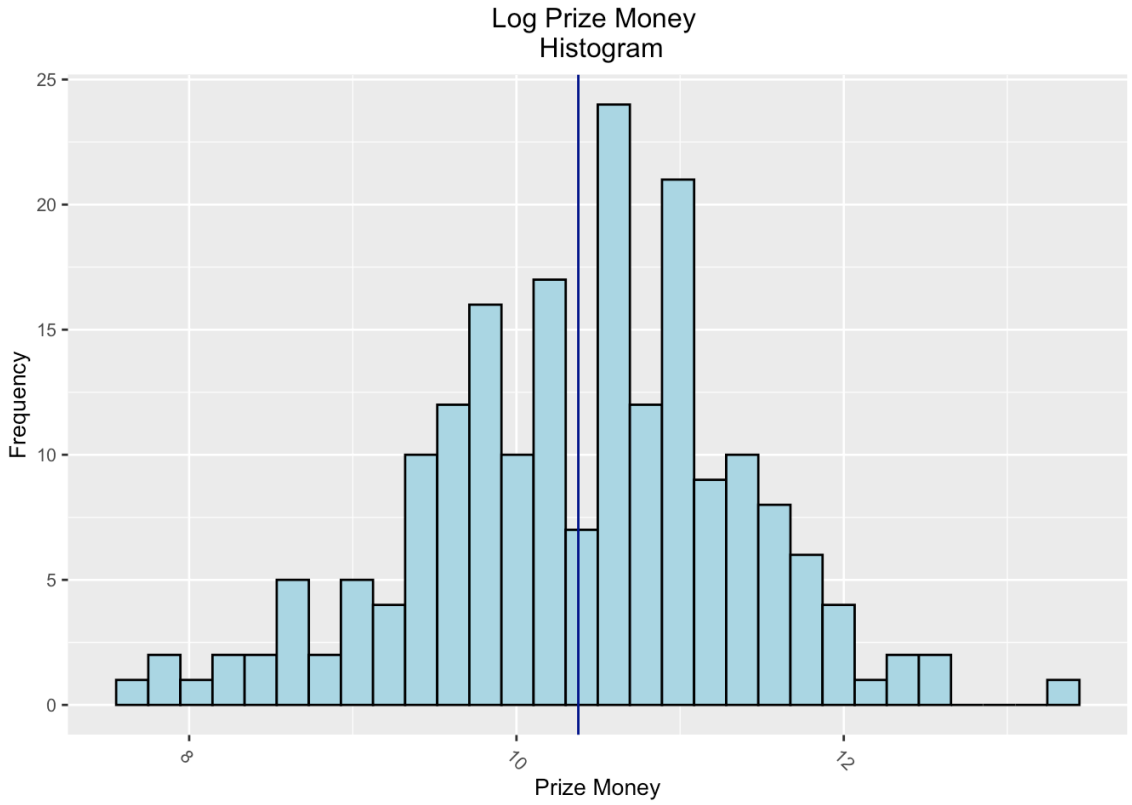
Answer b:

b) Analyze distribution of Prize Money, and discuss if the distribution is symmetric or skewed.

The prize money histogram is right heavily skewed and decreasing exponentially; therefore, transforming this value may provide a better model.

Log Transformation: Prize Money Histogram

```
# log transform into data frame
ln_Prize <- data.frame(PrizeMoney = log(golf$PrizeMoney))

# plot log prize money histogram
ggplot(ln_Prize, aes(x=PrizeMoney)) +
      geom_histogram(bins=30, color="black", fill="lightblue") +
      geom_vline(aes(xintercept=mean(PrizeMoney)), col="darkblue") +
      labs(title = "Log Prize Money \n Histogram",
           x = "Prize Money", y = 'Frequency') +
      # move the title text to the middle
      theme(plot.title=element_text(hjust=0.5)) +
      theme(text = element_text(size = 10))  +
      theme(axis.title = element_text(size = 10)) +
      theme(axis.text.x = element_text(angle = -45, hjust = .1))
```




```
# descriptive statistic
describe(ln_Prize)
```

```
##            vars   n  mean   sd median trimmed  mad  min  max range skew
## PrizeMoney    1 196 10.38 0.98  10.51   10.41 0.93 7.71 13.4  5.69 -0.2
##            kurtosis   se
## PrizeMoney     0.18 0.07
```

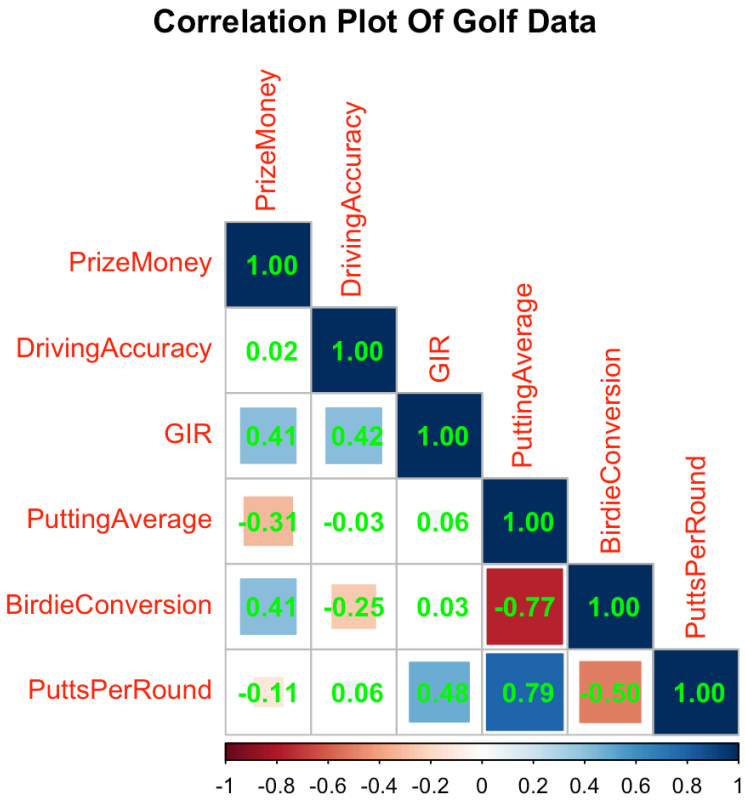
Answer b:

c) Apply a log transformation to PrizeMoney and compute the new variable ln_Prize=log(PrizeMoney). Analyze distribution of ln_Prize, and discuss if the distribution is symmetric or skewed.

By applying log transformation, the histogram of prize money is symmetric, and examing the mean and median values are nearly identical per describe function, which also states symmetry.

Correlation Answer b:

```
# correlation values except for name
corr.golf = cor(golf[, !names(golf) %in% c("Name")])

# plot correlation
corr_plot_golf <- corrplot(corr.golf, method = 'square', addCoef.col = 'green',
                     type = 'lower',
                     title = "\n\n Correlation Plot Of Golf Data \n",)
```



Multiple Linear Regression Model 1

```
# initial model
golf.model.1 <- lm(log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion + PuttsPer
Round, data=golf)

# display the actual equation for Model 1
equatiomatic::extract_eq(golf.model.1, use_coefs = FALSE)
```

$$\log(\text{PrizeMoney}) = \alpha + \beta_1(\text{DrivingAccuracy}) + \beta_2(\text{GIR}) + \beta_3(\text{PuttingAverage}) + \beta_4(\text{BirdieConversion}) + \beta_5(\text{PuttsPerRound}) + \epsilon$$

```
# model 1 standardized coefficients
lm.beta(golf.model.1)
```

```
##  DrivingAccuracy             GIR   PuttingAverage BirdieConversion
##     -0.004187957      0.746523426      0.220660141      0.342850182
##    PuttsPerRound
##     -0.545022608
```

```
# model 1 summary
summary(golf.model.1)
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage +
##     BirdieConversion + PuttsPerRound, data = golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.55696 -0.51250 -0.08005  0.45090  2.11898
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.2410192  7.1611241   1.151 0.251261
## DrivingAccuracy  -0.0007584  0.0116109  -0.065 0.947992
## GIR               0.2687898  0.0287938   9.335  < 2e-16 ***
## PuttingAverage    8.7467774  5.3734220   1.628 0.105228
## BirdieConversion  0.1523018  0.0408329   3.730 0.000253 ***
## PuttsPerRound    -1.2094847  0.2672761  -4.525 1.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6725 on 190 degrees of freedom
## Multiple R-squared:  0.5414, Adjusted R-squared:  0.5293
## F-statistic: 44.86 on 5 and 190 DF,  p-value: < 2.2e-16
```

```
# model 1 ANOVA for Analysis of Variance
anova(golf.model.1)
```

```
## Analysis of Variance Table
##
## Response: log(PrizeMoney)
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## DrivingAccuracy    1  6.184   6.184 13.6740 0.0002845 ***
## GIR                1 41.761  41.761 92.3464 < 2.2e-16 ***
## PuttingAverage     1 40.170  40.170 88.8272 < 2.2e-16 ***
## BirdieConversion   1  4.058   4.058  8.9744 0.0031030 **
## PuttsPerRound      1  9.260   9.260 20.4777 1.062e-05 ***
## Residuals        190 85.922   0.452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Variance Inflation Factors
vif.golf.model.1 = vif(golf.model.1)
vif.golf.model.1
```

```
##  DrivingAccuracy             GIR   PuttingAverage BirdieConversion
##         1.703301         2.649566         7.613214         3.500528
##    PuttsPerRound
##         6.009842
```

Multiple Linear Regression Model 2

```
# second model without putting average per correlation
golf.model.2 <- lm(log(PrizeMoney) ~ DrivingAccuracy + GIR + BirdieConversion + PuttsPerRound, data=golf)

# display the actual equation for Model 2
equatiomatic::extract_eq(golf.model.2, use_coefs = FALSE)
```

$$\log(\text{PrizeMoney}) = \alpha + \beta_1(\text{DrivingAccuracy}) + \beta_2(\text{GIR}) + \beta_3(\text{BirdieConversion}) + \beta_4(\text{PuttsPerRound}) + \epsilon$$

```
# model 2 standardized coefficients
lm.beta(golf.model.2)
```

```
##  DrivingAccuracy             GIR BirdieConversion    PuttsPerRound
##       -0.0231814       0.6976863       0.2452487       -0.3945729
```

```
# model 2 summary
summary(golf.model.2)
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ DrivingAccuracy + GIR + BirdieConversion +
##     PuttsPerRound, data = golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61656 -0.50888 -0.07585  0.45718  2.04271
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      16.680435   4.961087   3.362 0.000934 ***
## DrivingAccuracy  -0.004198   0.011466  -0.366 0.714698
## GIR               0.251206   0.026806   9.371  < 2e-16 ***
## BirdieConversion  0.108945   0.031083   3.505 0.000569 ***
## PuttsPerRound    -0.875615   0.172109  -5.088 8.64e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6754 on 191 degrees of freedom
## Multiple R-squared:  0.535,  Adjusted R-squared:  0.5253
## F-statistic: 54.94 on 4 and 191 DF,  p-value: < 2.2e-16
```

```
# model 2 ANOVA for Analysis of Variance
anova(golf.model.2)
```

```
## Analysis of Variance Table
##
## Response: log(PrizeMoney)
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## DrivingAccuracy    1  6.184   6.184  13.557 0.0003011 ***
## GIR                1 41.761  41.761  91.556 < 2.2e-16 ***
## BirdieConversion   1 40.484  40.484  88.756 < 2.2e-16 ***
## PuttsPerRound      1 11.806  11.806  25.883 8.635e-07 ***
## Residuals        191 87.120   0.456
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Variance Inflation Factors
vif.golf.model.2 = vif(golf.model.2)
vif.golf.model.2
```

```
##  DrivingAccuracy             GIR BirdieConversion    PuttsPerRound
##         1.646895         2.276642         2.011054         2.470661
```

Multiple Linear Regression Model 3

```
# second model without DrivingAccuracy per p-value on summary
golf.model.3 <- lm(log(PrizeMoney) ~ GIR + BirdieConversion + PuttsPerRound, data=golf)

# display the actual equation for Model 3
equatiomatic::extract_eq(golf.model.3, use_coefs = FALSE)
```

$$\log(\text{PrizeMoney}) = \alpha + \beta_1(\text{GIR}) + \beta_2(\text{BirdieConversion}) + \beta_3(\text{PuttsPerRound}) + \epsilon$$

```
# model 3 standardized coefficients
lm.beta(golf.model.3)
```

```
##              GIR BirdieConversion    PuttsPerRound
##        0.6816186        0.2578537       -0.3819335
```

```
# model 3 summary
summary(golf.model.3)
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ GIR + BirdieConversion + PuttsPerRound,
##     data = golf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6140 -0.5152 -0.0761  0.4540  2.0583
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       15.8102     4.3446   3.639 0.000352 ***
## GIR                0.2454     0.0216  11.360  < 2e-16 ***
## BirdieConversion   0.1145     0.0270   4.243 3.43e-05 ***
## PuttsPerRound     -0.8476     0.1538  -5.512 1.13e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6738 on 192 degrees of freedom
## Multiple R-squared:  0.5347, Adjusted R-squared:  0.5274
## F-statistic: 73.54 on 3 and 192 DF,  p-value: < 2.2e-16
```

```
# model 3 ANOVA for Analysis of Variance
anova(golf.model.3)
```

```
## Analysis of Variance Table
##
## Response: log(PrizeMoney)
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## GIR                1 47.760  47.760 105.182 < 2.2e-16 ***
## BirdieConversion   1 38.618  38.618  85.049 < 2.2e-16 ***
## PuttsPerRound      1 13.796  13.796  30.382 1.133e-07 ***
## Residuals        192 87.181   0.454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Variance Inflation Factors
vif.golf.model.3 = vif(golf.model.3)
vif.golf.model.3
```

```
##              GIR BirdieConversion    PuttsPerRound
##         1.485427         1.524123         1.981059
```

Multiple Linear Regression Model 4

```
# fourth model log(GIR)
# histogram plot with log transformation caused symmetrical distribution on GIR
golf.model.4 <- lm(log(PrizeMoney) ~ log(GIR) + BirdieConversion + PuttsPerRound, data=golf)

# display the actual equation for Model 4
equatiomatic::extract_eq(golf.model.4, use_coefs = FALSE)
```

$$\log(\text{PrizeMoney}) = \alpha + \beta_1(\log(\text{GIR})) + \beta_2(\text{BirdieConversion}) + \beta_3(\text{PuttsPerRound}) + \epsilon$$

```
# model 4 standardized coefficients
lm.beta(golf.model.4)
```

```
##         log(GIR) BirdieConversion    PuttsPerRound
##        0.6824429        0.2580527       -0.3808114
```

```
# model 4 summary
summary(golf.model.4)
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ log(GIR) + BirdieConversion +
##     PuttsPerRound, data = golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58562 -0.51930 -0.08077  0.46424  2.04488
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -34.63477    5.07089  -6.830 1.09e-10 ***
## log(GIR)          15.89123    1.39189  11.417  < 2e-16 ***
## BirdieConversion   0.11463    0.02693   4.256 3.25e-05 ***
## PuttsPerRound     -0.84508    0.15319  -5.517 1.11e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6725 on 192 degrees of freedom
## Multiple R-squared:  0.5365, Adjusted R-squared:  0.5293
## F-statistic: 74.09 on 3 and 192 DF,  p-value: < 2.2e-16
```

```
# model 4 ANOVA for Analysis of Variance
anova(golf.model.4)
```

```
## Analysis of Variance Table
##
## Response: log(PrizeMoney)
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## log(GIR)           1 48.194  48.194 106.566 < 2.2e-16 ***
## BirdieConversion   1 38.565  38.565  85.275 < 2.2e-16 ***
## PuttsPerRound      1 13.764  13.764  30.434 1.107e-07 ***
## Residuals        192 86.832   0.452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Variance Inflation Factors
vif.golf.model.4 = vif(golf.model.4)
vif.golf.model.4
```

```
##         log(GIR) BirdieConversion    PuttsPerRound
##         1.480190         1.522691         1.974022
```

Answer d-1:

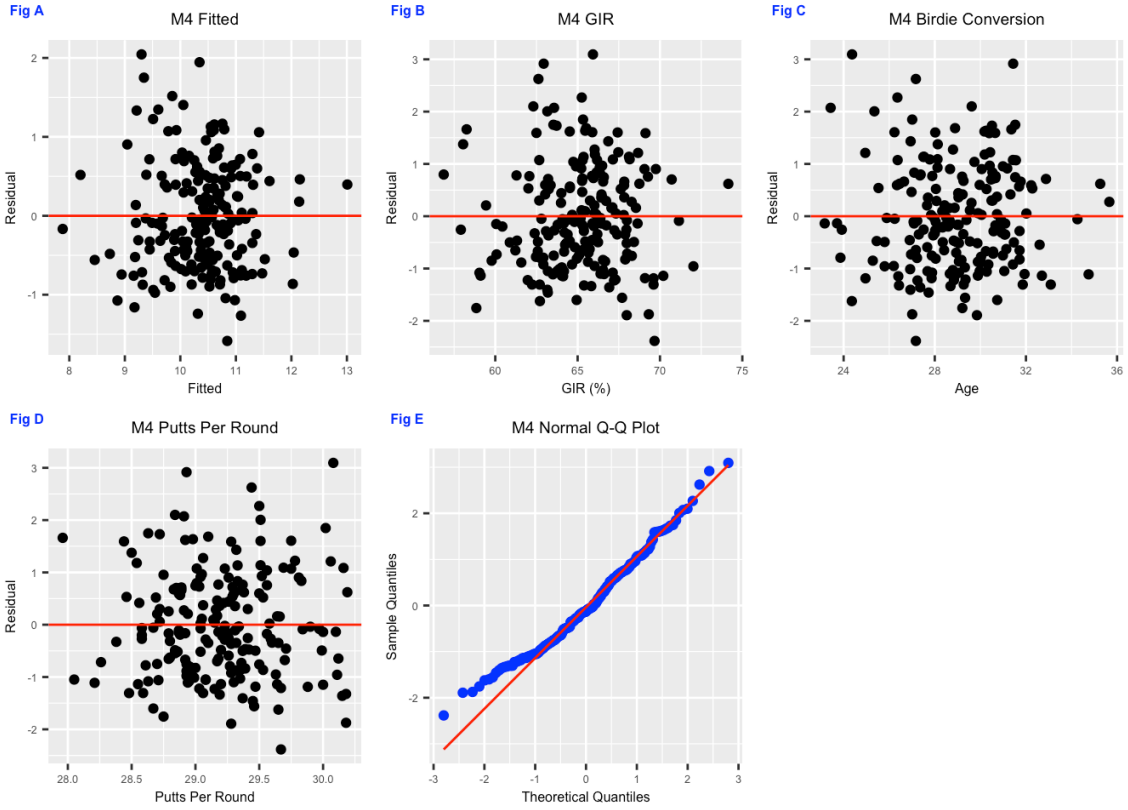
d) Fit a regression model of ln_Prize using the remaining predictors in your dataset. Apply your knowledge of regression analysis to define a valid model to predict ln_Prize. Hint: use scatterplots and correlation

d-1) If necessary remove not significant variables. Remember to remove one variable at a time (variable with largest p-value is removed first) and refit the model, until all variables are significant.

The first model containing all the variables is my base model and R-squared: 0.5414, Adjusted R-squared: 0.5293, but on the summary of the model for DrivingAccuracy & PuttingAverage, we are unable to reject the null hypothesis. I decided to remove PuttingAverage due to the high correlation with BirdieConversion & PuttsPerRound, which gave me R-squared: 0.535, Adjusted R-squared: 0.5253. The DrivingAccuracy was unable to reject the null hypothesis in the summary report. Therefore, I removed the variable, and according to the scatter plot, it didn't seem linear with the response variable. The third model produced R-squared: 0.5347, Adjusted R-squared: 0.5274, and all the p-value for the t-statistic & F-statistic were under 0.05. I plotted histograms of remaining independent variables and discovered that log transforming the GIR gave me a better histogram; therefore, I created a fourth model and **R-squared: 0.5365**, **Adjusted R-squared: 0.5293**, which turned out to be the best-performing model including all the variable are significant.

Residual Analysis Golf Plot Model 4

```
# residual vs fitted
resi_golf_plot <-  ggplot(golf.model.4, aes(x = .fitted, y = .resid)) +
                   geom_point() +
                   geom_hline(yintercept = 0, col = "red") +
                   labs(title="M4 Fitted",
                    x = "Fitted", y = "Residual") +
                   # move the title text to the middle
                   theme(plot.title=element_text(hjust=0.5))  +
                   theme(text = element_text(size = 6))  +
                   theme(axis.title = element_text(size = 6))

# GIR vs residuals
gir_plot <-  ggplot(golf, aes(x = GIR, y = rstandard(golf.model.4))) +
             geom_point() + geom_hline(yintercept = 0, col = "red") +
             labs(title="M4 GIR", x = "GIR (%)", y = "Residual") +
             # move the title text to the middle
             theme(plot.title=element_text(hjust=0.5))  +
             theme(text = element_text(size = 6))  +
             theme(axis.title = element_text(size = 6))

# Birdie Conversion vs residuals
birdie_plot <-  ggplot(golf, aes(x = BirdieConversion, y = rstandard(golf.model.4))) +
                geom_point() + geom_hline(yintercept = 0, col = "red") +
                labs(title="M4 Birdie Conversion", x = "Age", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 6))

# PuttsPerRound vs residuals
putts_plot <-   ggplot(golf, aes(x = PuttsPerRound, y = rstandard(golf.model.4))) +
                geom_point() + geom_hline(yintercept = 0, col = "red") +
                labs(title="M4 Putts Per Round", x = "Putts Per Round", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 6))

#create Q-Q plot
qq_golf_plot <-  ggplot(golf.model.4, aes(sample=rstandard(golf.model.4))) +
                stat_qq(size=1.5, color='blue') +
                stat_qq_line(col = "red") +
                labs(title="M4 Normal Q-Q Plot",
                  x = "Theoretical Quantiles", y = "Sample Quantiles") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 6))

# combine all four plots
combine_golf_plot <- ggarrange(resi_golf_plot, gir_plot, birdie_plot, putts_plot,        qq_golf_plot, la
bels = c("Fig A", "Fig B", "Fig C", "Fig D", "Fig E"),
                     font.label = list(size = 6, color = "blue"))
# plot all
combine_golf_plot
```



Answer d-2:

d-2) Analyze residual plots to check if the regression model is valid for your data.

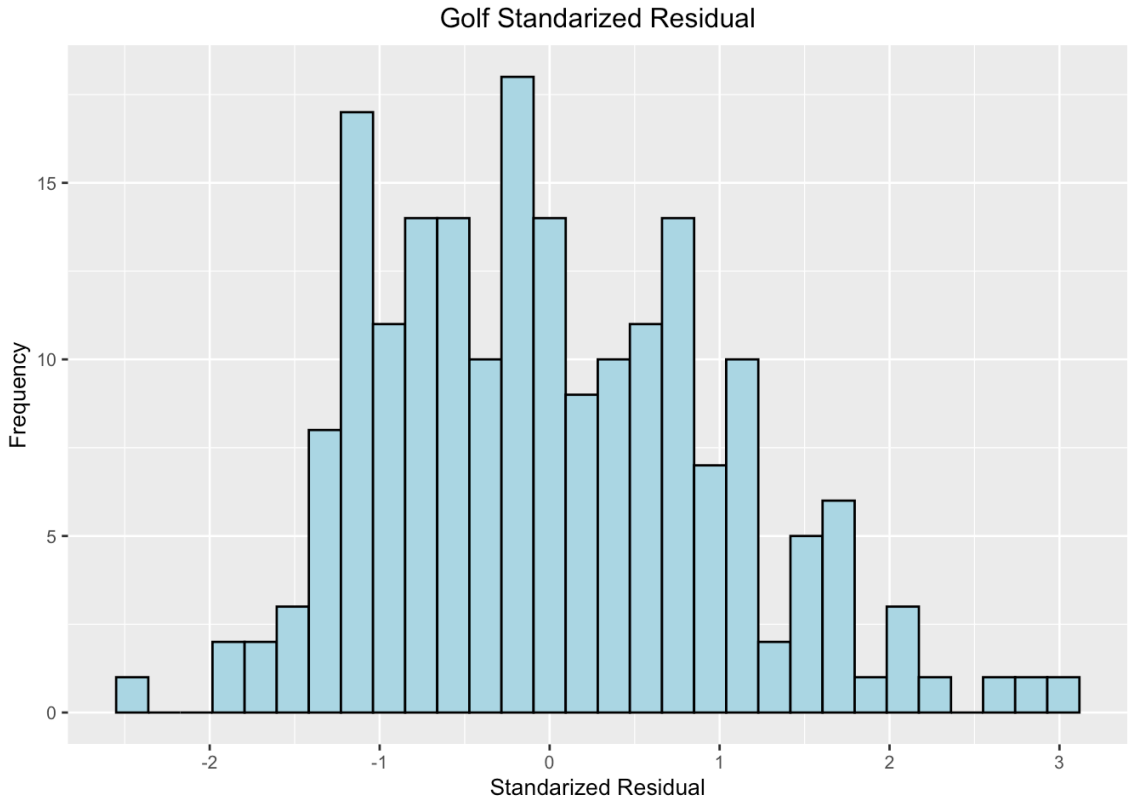
The plots of residuals against each independent variable and with the fitted values do indicate randomness. In the summary of the model, p-values are all below 0.05, and from the Analysis of Variance (ANOVA), p-values are below 0.05 as well. The Q-Q plot also displays that residuals are normally distributed, and all four residual plots show possible outliers on the top portion of the residual plots.

Influential Points and Outliers

```
# outliers |standardized residuals| > 3
golf_std_residual = data.frame(residual = rstandard(golf.model.4))

# display |standardized residuals| > 3
filter(golf_std_residual, abs(residual) > 3)
```

```
##     residual
## 185 3.094047
```

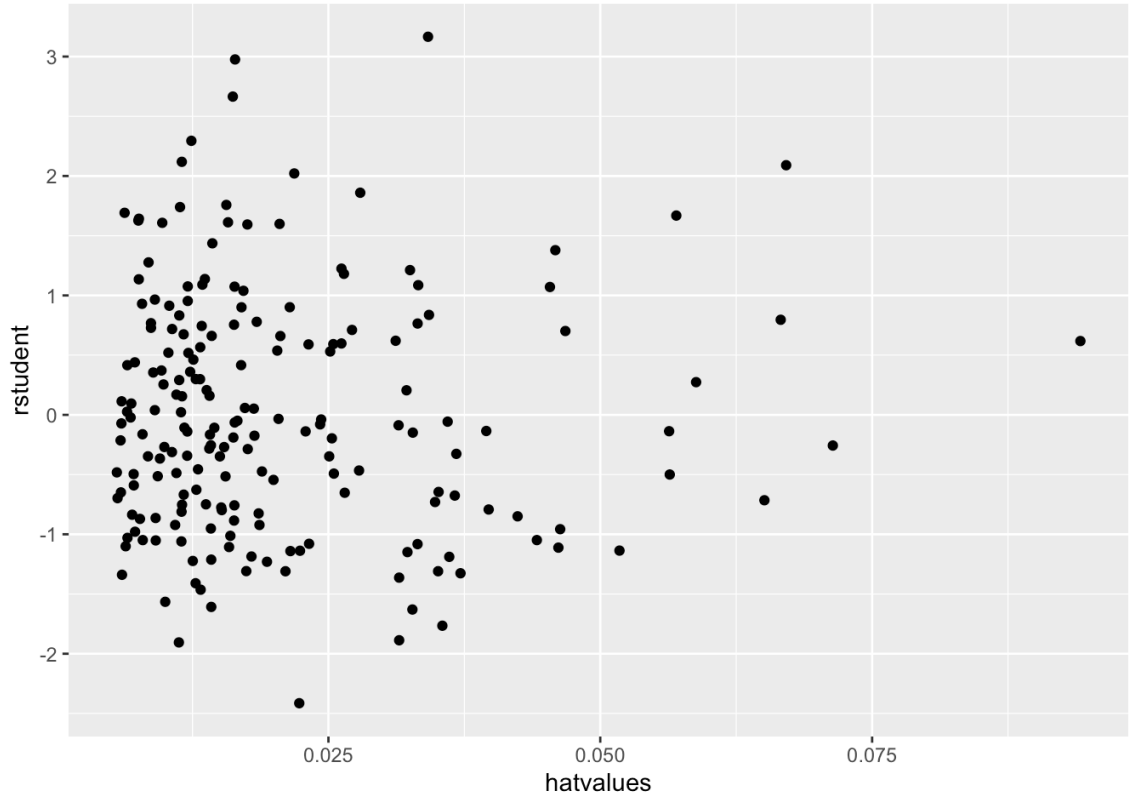
```
# historgram for outliers
ggplot(golf_std_residual, aes(x = residual)) +
  geom_histogram(bins=30, color="black", fill="lightblue") +
  labs(title = "Golf Standarized Residual", x = "Standarized Residual", y = 'Frequency') +
      # move the title text to the middle
      theme(plot.title=element_text(hjust=0.5)) +
      theme(text = element_text(size = 10))  +
      theme(axis.title = element_text(size = 10))
```




```
# print out only observations that may be influential
summary(influence.measures(golf.model.4))
```

```
## Potentially influential observations of
##   lm(formula = log(PrizeMoney) ~ log(GIR) + BirdieConversion +      PuttsPerRound, data = golf) :
##
##     dfb.1_ dfb.l(GI dfb.BrdC dfb.PtPR dffit   cov.r   cook.d hat
## 9    0.01   0.00    -0.15     0.01     0.26    0.93_*  0.02   0.01
## 40   0.24   0.14    -0.53    -0.36     0.56_*  1.00    0.08   0.07_*
## 47   0.24  -0.25     0.14     0.04    -0.36    0.93_*  0.03   0.02
## 63   0.11  -0.21     0.26     0.11     0.38    0.87_*  0.04   0.02
## 70  -0.05   0.05     0.03    -0.01    -0.07    1.10_*  0.00   0.07_*
## 96  -0.02   0.10    -0.09    -0.09    -0.12    1.08_*  0.00   0.06
## 113 -0.01  -0.01     0.03     0.02    -0.03    1.08_*  0.00   0.06
## 142 -0.03   0.01     0.05     0.01     0.07    1.08_*  0.00   0.06
## 172  0.00  -0.14     0.07     0.17    -0.19    1.08_*  0.01   0.07_*
## 177  0.00   0.00     0.01    -0.01    -0.03    1.06_*  0.00   0.04
## 178 -0.15   0.10     0.12     0.02     0.20    1.12_*  0.01   0.09_*
## 180  0.14  -0.23    -0.03     0.16     0.34    0.90_*  0.03   0.02
## 184  0.16  -0.19     0.01     0.06     0.21    1.08_*  0.01   0.07_*
## 185 -0.09  -0.08    -0.25     0.25     0.60_*  0.86_*  0.08   0.03
```

```
# plot of deleted studentized residuals vs hat values
student_hat <- data.frame(rstudent = rstudent(golf.model.4), hatvalues =  hatvalues(golf.model.4))

# plot rstudent vs hatvalues
student_hat_plot <- ggplot(student_hat, aes(x = hatvalues, y = rstudent)) + geom_point()

# plot
student_hat_plot
```



Answer d-3:

d-3 Analyze if there are any outliers and influential points. If there are points in the dataset that need to be investigated, give one or more reason to support each point chosen.

There is on outlier in the dataset according to standarized residual values greater than absolute value of 3. I've filtered rstandard(golf.model.4) > 3 & visualized the standardized reiduals for the outlier detection. Also, we could have used a boxplot to display the outliers.

We have multiple tools for analyzing influential points in the model and retrieving various influential statistics commands are listed below for all the independent variables:

- dfbeta(model)
- covratio(model)
- dffits(model)
- cooks.distance(model)

We will take a pragmatic approach by examing High Cook's D distance (>1), High leverage hat hii value (> 0.5), High Deleted Studentized Residuals (outside (-3,3) band), and plot deleted studentized residuals vs. hat values to visualize potential, influential points.

We are checking the summary(influence.measures(model)), which shows possible influential points in M4. Also, rstudent vs. hat value plot shows influential points using our guideline. Following the guidelines, our golf data has influential points that require investigation. Finally, the student and hat value plot also indicates two influential points top left that require our attention.

```
# display the actual equation for Model 4
equatiomatic::extract_eq(golf.model.4, use_coefs = FALSE)
```

$$\log(\text{PrizeMoney}) = \alpha + \beta_1(\log(\text{GIR})) + \beta_2(\text{BirdieConversion}) + \beta_3(\text{PuttsPerRound}) + \epsilon$$

Answer d-4:

d-4 Interpret the regression coefficients in the final model to answer the following question: How does an increase in 1% for GIR affect the average Prize money?

1% increase in GIR would increase by exp() on the average prize money while all other dependent variables are held constant.

Compute Prediction

```
# predicted value
given = data.frame(GIR = c(67), DrivingAccuracy = c(64) ,PuttingAverage = c(1.77), BirdieConversion = c(2
8), PuttsPerRound = c(29.16))

# display the actual equation for Model 4
equatiomatic::extract_eq(golf.model.4, use_coefs = TRUE)
```

$$\log(\widehat{\text{PrizeMoney}}) = -34.63 + 15.89(\log(\text{GIR})) + 0.11(\text{BirdieConversion}) - 0.85(\text{PuttsPerRound})$$

```
# second model - compute the average 95% prediction interval
prize_money <- predict(golf.model.4, given, interval="prediction",level=0.95)

# print actural prize
exp(prize_money)
```

```
##        fit      lwr      upr
## 1 46642.89 12275.43 177228.8
```

Answer d-5:

d-5 Compute the prediction and 95% prediction interval for average prize money for a player that has a GIR of 67%, driving accuracy of 64%, putting average of 1.77, Birdie Conversion of 28% and 29.16 average putts per round.

- fit : $46,642.89
- lwr : $12,275.43
- upr : $177,228.80