UNITED STATES MILITARY ACADEMY

HOMEWORK 1

MA478: GENERALIZED LINEAR MODELS

SECTION H2-4

COL NICHOLAS CLARK

By

CADET SAMIN KIM ’24, CO A2

WEST POINT, NEW YORK

30 JAN 2022

____ MY DOCUMENTATION IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT.

SK I DID NOT USE ANY SOURCES OR ASSISTANCE REQUIRING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT.

SIGNATURE: KIM.SAMIN.1594266885 Digitally signed by KIM.SAMIN.1594266885 Date: 2024.01.30 02:04:05 -05'00'

1) Design Matrix X, W

Column space $C(X) = C(W)$

$P_X = X(X^T X)^{-1} X^T$. $P_W = W(W^T W)^{-1} W^T$

$*$ $X^T X$ is invertible

Because $C(X) = C(W)$,

i) Y that $X = WY$ exists. $*$ $Y^T W^T W Y$ is also invertible.

$P_X = X(X^T X)^{-1} X^T = WY(Y^T W^T W Y)^{-1} Y^T W^T$ $\quad (Y^T W^T W Y)^{-1} = Y^{-1} (W^T W)^{-1} (Y^T)^{-1}$

$= WYY^{-1}(W^T W)^{-1}(Y^T)^{-1} Y^T W^T = W(W^T W)^{-1} W^T$

$P_X = W(W^T W)^{-1} W^T = P_W$

---

2) $E[Y_i] = \beta_0 + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2)$

$\Uparrow$

$E[Y_i] = \beta_0 + \beta_1(x_{1i}) + \beta_2(x_{2i})$

$x_{ni} = (x_{ni} - \bar{x}_n)$

This would shift the model by column average when the regression is plotted.

β_0 will no longer be the $E[Y]$ when $x_{ni} = 0$,

it will represent the $E[Y]$ when $x_{ni} = \bar{x}_n$

The β_1 and β_2 will not be changed.

This makes sense because subtracting by the column average does not change the column space.

Homework 1-3

Samin Kim

January 29th 2024

1 Introduction

The Indoor Obstacle Course Test (IOCT) is a test managed by the Department of Physical Education (DPE). It involves multiple evaluations of physical fitness, such as strength, endurance, balance, and etc. The IOCT is a requirement that all cadets must pass to graduate. Due to its significance, many cadets tried to find the association between the IOCT time and other variables that may impact the IOCT time. What are the variables that impact the IOCT Score? If there is, what is the relationship with IOCT Score?

2 Dataset

The "IOCT.csv" dataset has 384 observations with 9 variables. The 9 dataset variables are sex, height, weight, IOCT time, push-up score, sit-up score, 2 mi run score, Corps Squad athlete, and APFT Score. Of the observations, 280 were male, and 104 were female. As shown in figure 1, the majority of observations had an IOCT score of around 190. The higher IOCT score had less occurrence in the dataset. Figure 2,3, 4, and 5 shows the relationship between IOCT score and other variables. As shown in the scatter plots, the IOCT score had a significant difference in gender, with males having shorter IOCT scores than females. As shown in Appendix A, the APFT and run scores were negatively associated with the IOCT scores, while weight and height were positively associated with the score.

3 Linear Regression

Because the males had significantly larger observations than the females, I decided to filter the females when developing the linear regression models. Also, when accounting for males, any observation that had a higher than 300 IOCT score was considered an outlier. To predict the IOCT Score, I used the linear regression by the following equation:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$$

The y_i represents the predicted value of the IOCT Score, β_0 represents the IOCT score when the explanatory variables are 0, β_1 and β_2 are the change in IOCT Score for every unit increase in x_1 and x_2 respectively, and the ϵ_i represents the residuals. For this research, I compared four different models.

$$\text{Model 1:} y_i = \beta_0 + \beta_1(weight) + \epsilon_i$$

$$\text{Model 2:} y_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \epsilon_i$$

$$\text{Model 3:} y_i = \beta_0 + \beta_1(APFT) + \epsilon_i$$

$$\text{Model 4:} y_i = \beta_0 + \beta_1(APFT) + \beta_2(weight) + \epsilon_i$$

4 Results

The summary of analysis of all four models is shown in the following table:

	P-value	F-statistics	R squared	AIC	BIC
Model 1	0.0001135	15.34	0.05249	2520	2531
Model 2	3.576e-05	10.63	0.07176	2517	2531
Model 3	2.2e-16	82.94	0.2311	2462	2473
Model 4	2.2e-16	52.28	0.2755	2448	2462

Overall, all models showed low p-values. However, model 1 and model 2, which only included physical characteristics, have significantly lower F-statistics and R^2 values. Also, their AIC and BIC are higher than those from models 3 and 4. Model 3 only took the APFT score as an explanatory variable, and Model 4 took the APFT score and weight. Both models have low AIC and BIC values, but considering those two models are nested, model 3 has higher F-statistics than model 4. Based on the comparison of analysis, model 3 was the most fitted model.

5 Conclusion

Of multiple variables, the APFT variable was the variable that showed significant association with IOCT score. For every unit increase in APFT score, the IOCT Score decreased by -0.30934. Although the Model 3 has only APFT Score as an explanatory variable, we may conclude that push-up, sit-up, and run score are all associated with the IOCT score since the APFT score is the sum of three scores.

A APPENDIX

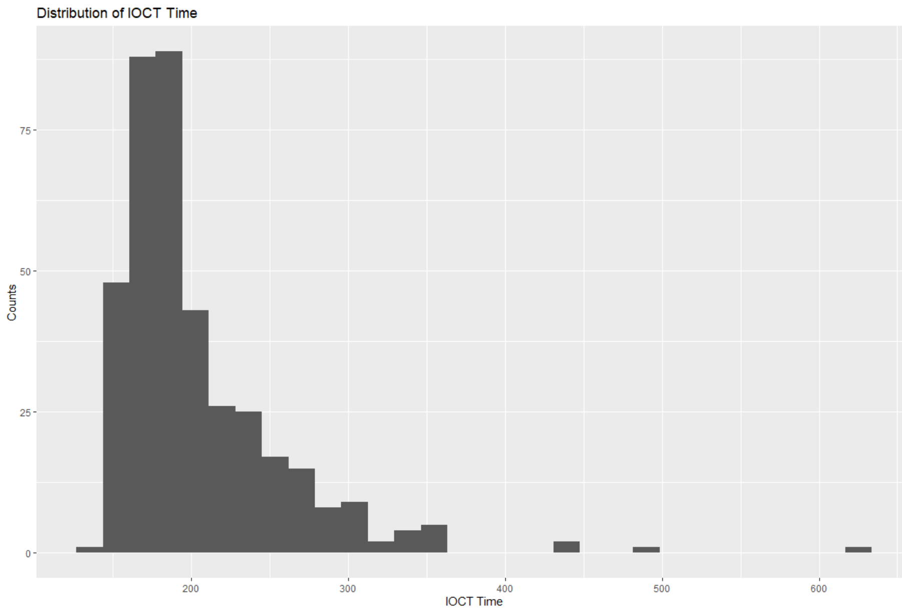


Figure 1: Histogram of IOCT Score

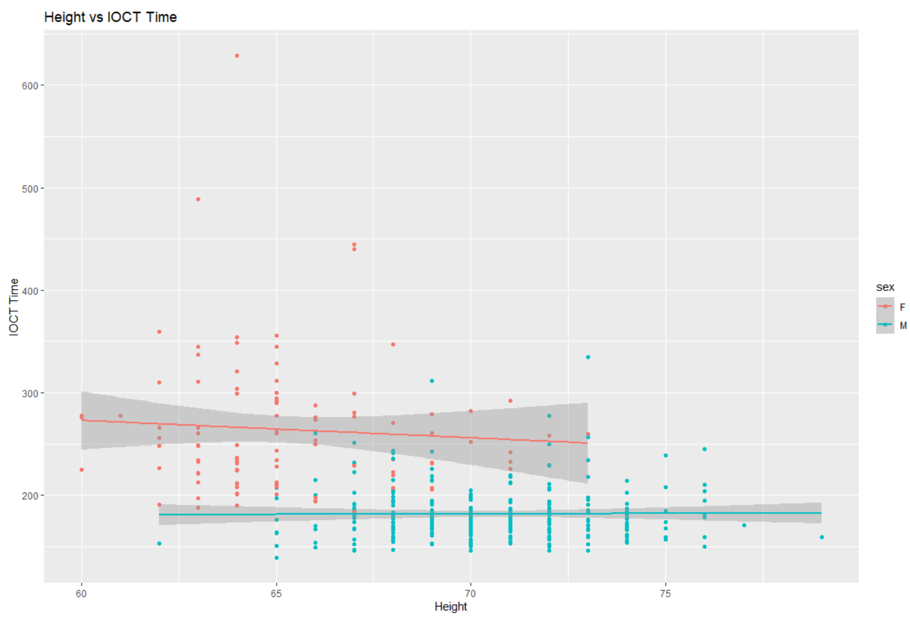


Figure 2: Scatter Plot of Height v. IOCT

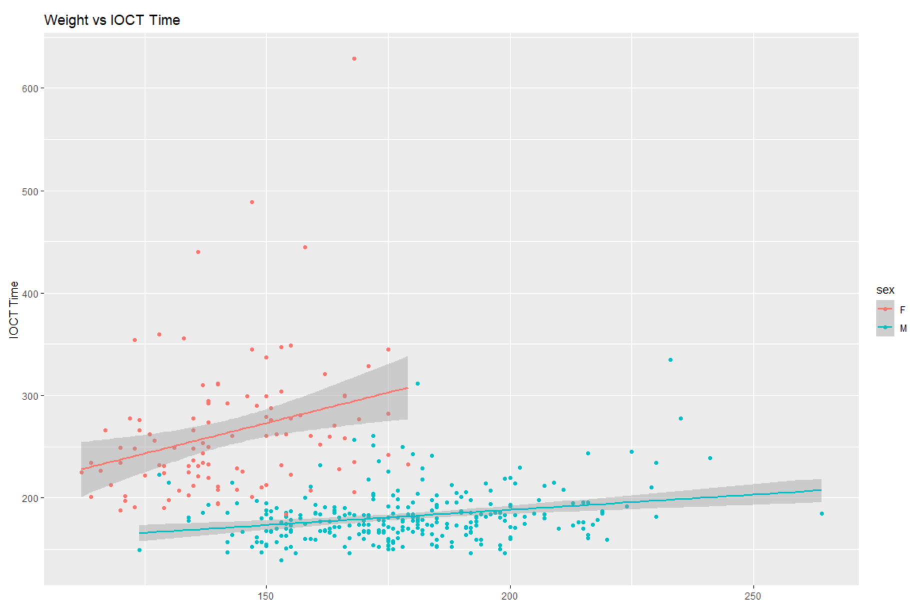


Figure 3: Scatter Plot of Weight v. IOCT

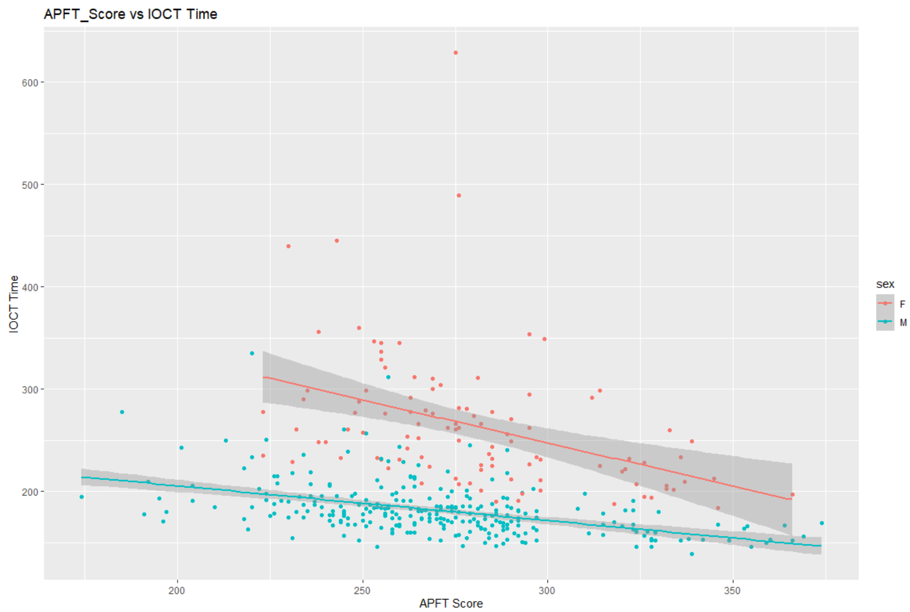


Figure 4: Scatter Plot of APFT Score v. IOCT

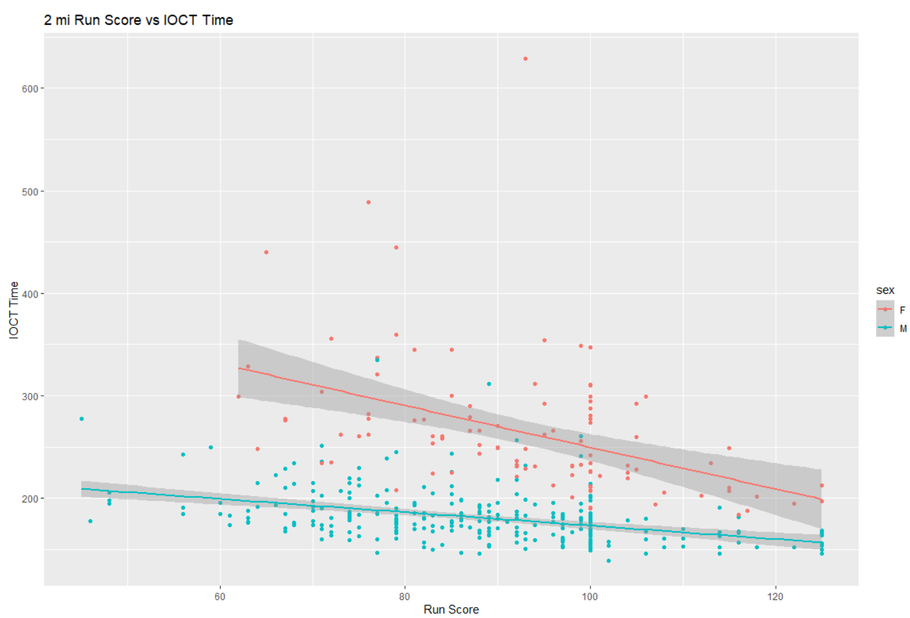


Figure 5: Scatter Plot of Run Score v. IOCT

Untitled

Samin Kim

2024-01-30

R Markdown

Loading libraries

```
library(dplyr)
```

```
##
## 다음의 패키지를 부착합니다: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## ── Attaching packages
## ───────────────────────────────────────────────────────
## tidyverse 1.3.2 ──
```

```
## ✔ ggplot2 3.3.6     ✔ purrr   0.3.5
## ✔ tibble  3.1.8     ✔ stringr 1.4.1
## ✔ tidyr   1.2.1     ✔ forcats 0.5.2
## ✔ readr   2.1.3
## ── Conflicts ──────────────────────────────────────────────────
───── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(janitor)
```

```
##
## 다음의 패키지를 부착합니다: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(ggplot2)
library(readr)
```

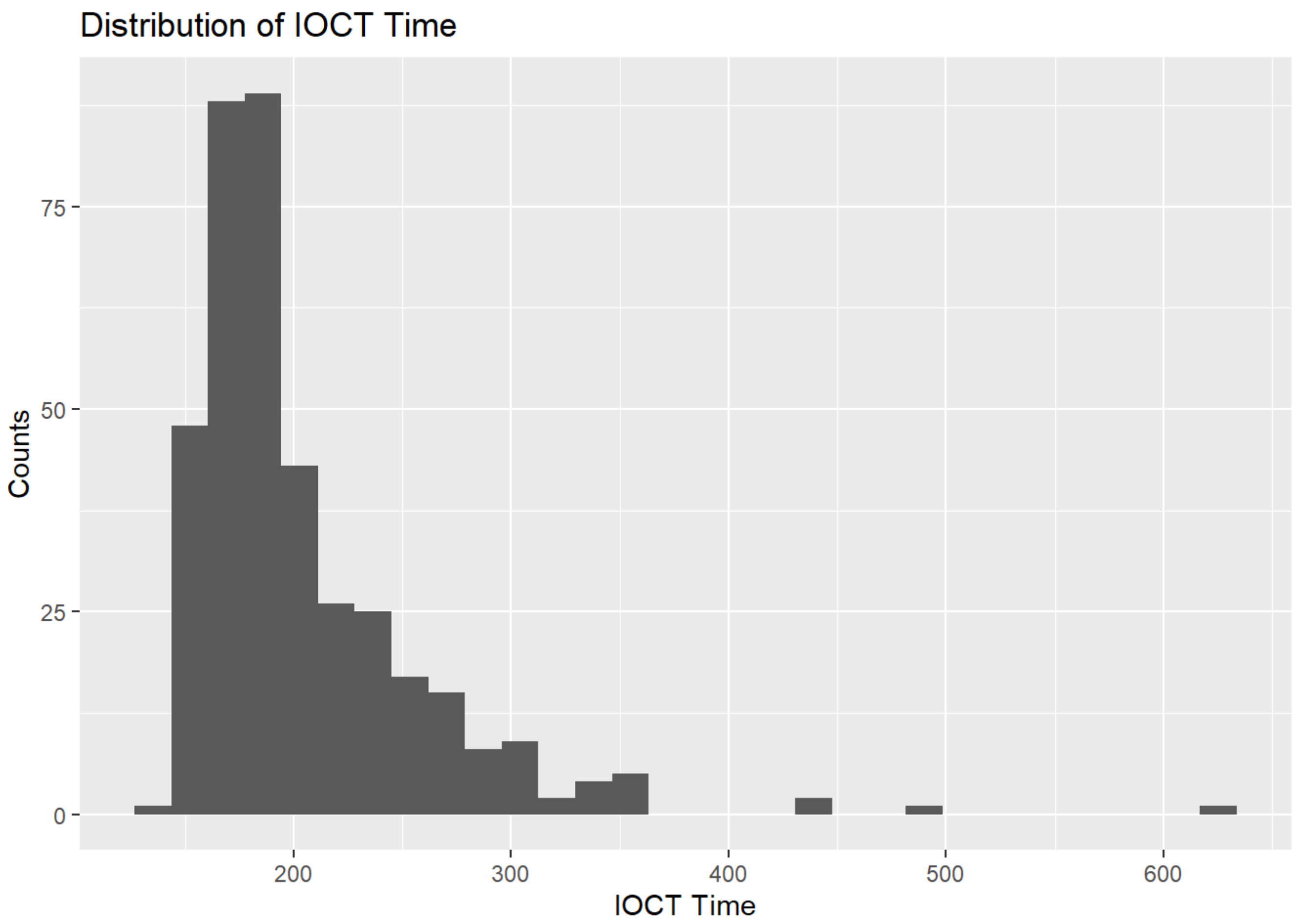
Read the File

Filtering The outliers

```
ic <- ioct %>%
  filter(IOCT_Time <300) %>%
  filter(sex == "M")
```

```
ioct %>%
  ggplot(aes(x = IOCT_Time)) +
  geom_histogram() +
  labs(title = "Distribution of IOCT Time")+
  xlab("IOCT Time") +
  ylab("Counts")
```

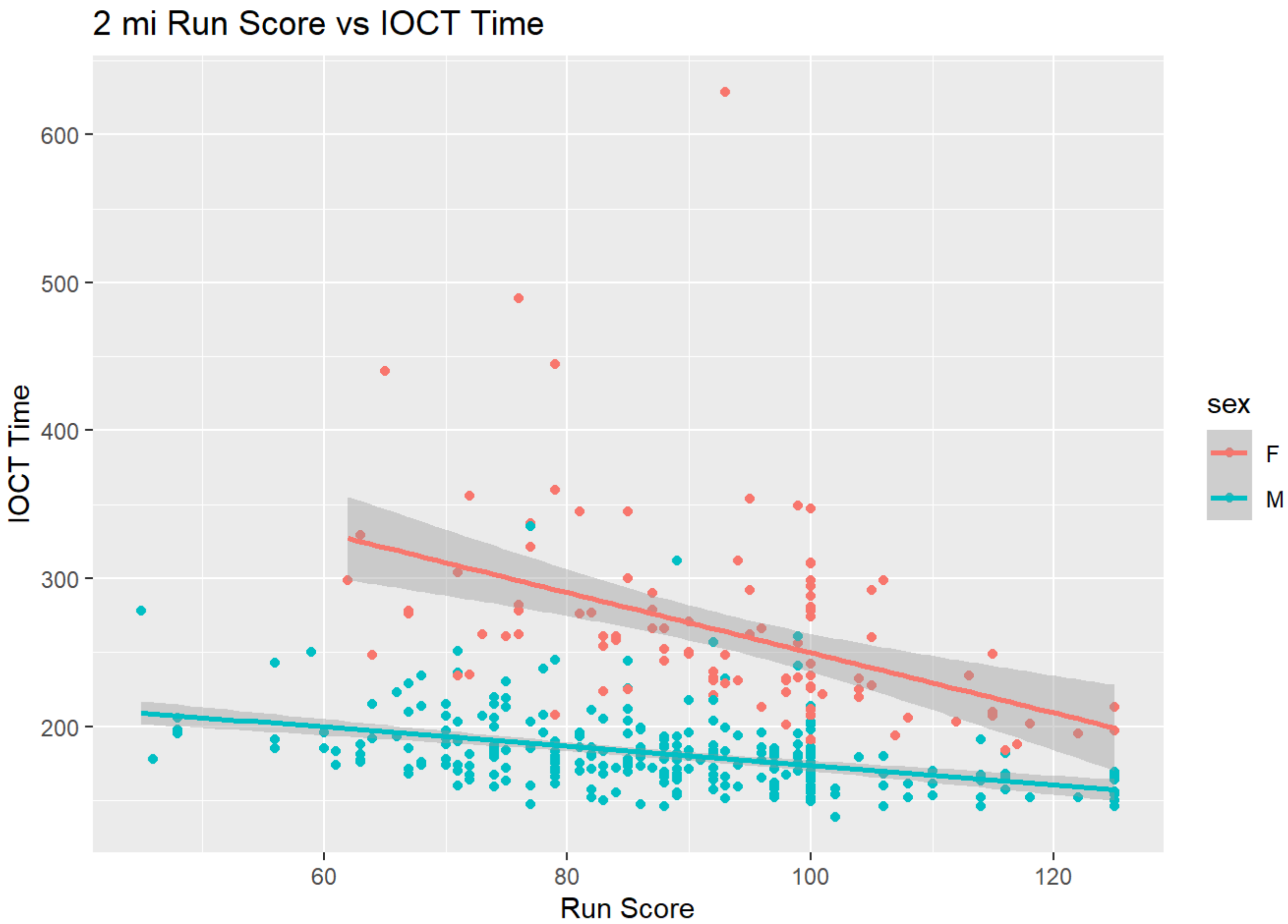
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Making Different Plots

```
ioct %>%
  ggplot(aes(x = run_score, y = IOCT_Time, color = sex)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "2 mi Run Score vs IOCT Time")+
  ylab("IOCT Time") +
  xlab("Run Score")
```

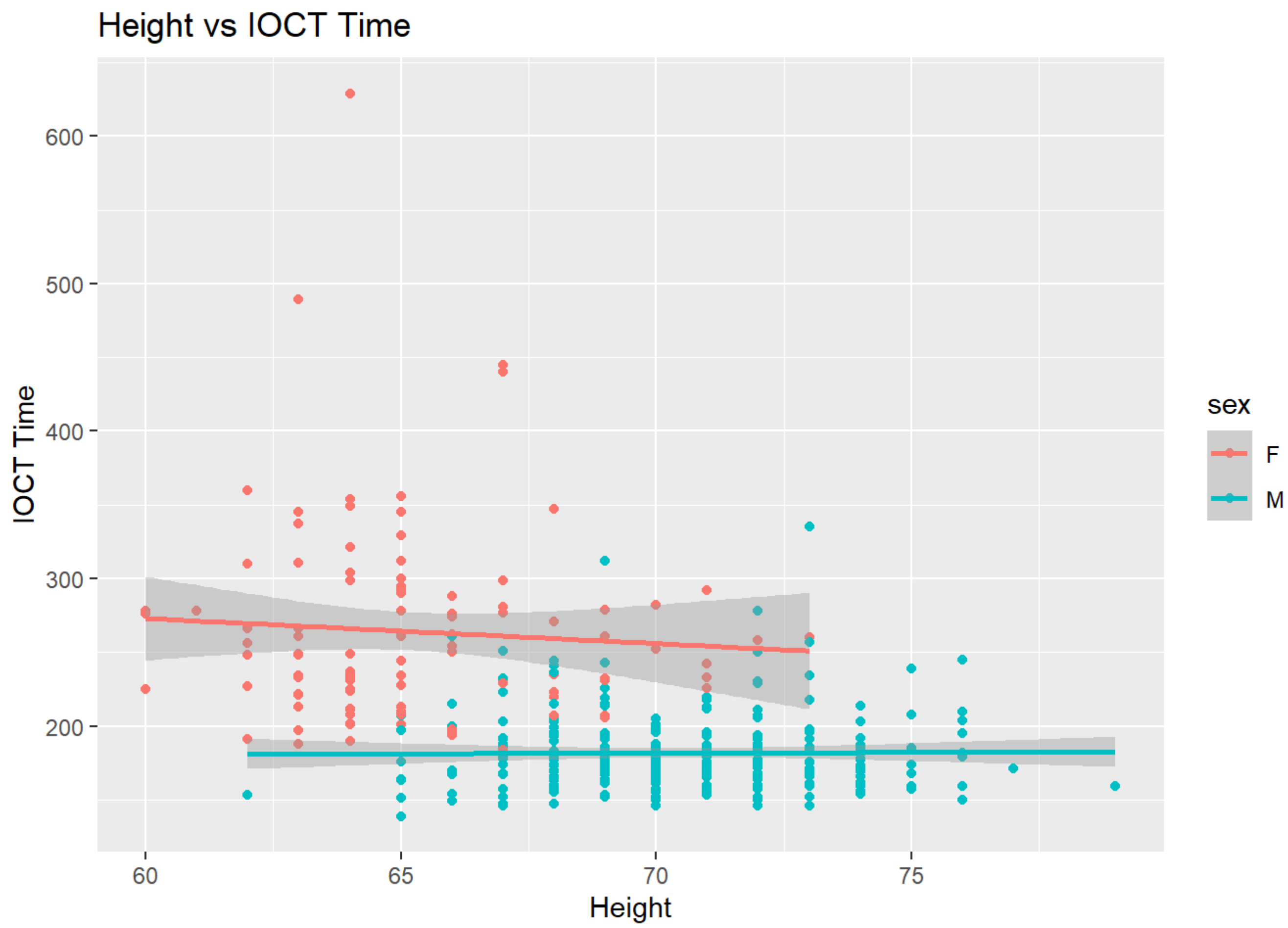
```
## `geom_smooth()` using formula 'y ~ x'
```



Height v. IOCT Time

```
ioct %>%
  ggplot(aes(x = height, y = IOCT_Time, color = sex)) +
  geom_point() +
  geom_smooth(method = "lm")+
  labs(title = "Height vs IOCT Time")+
  ylab("IOCT Time") +
  xlab("Height")
```

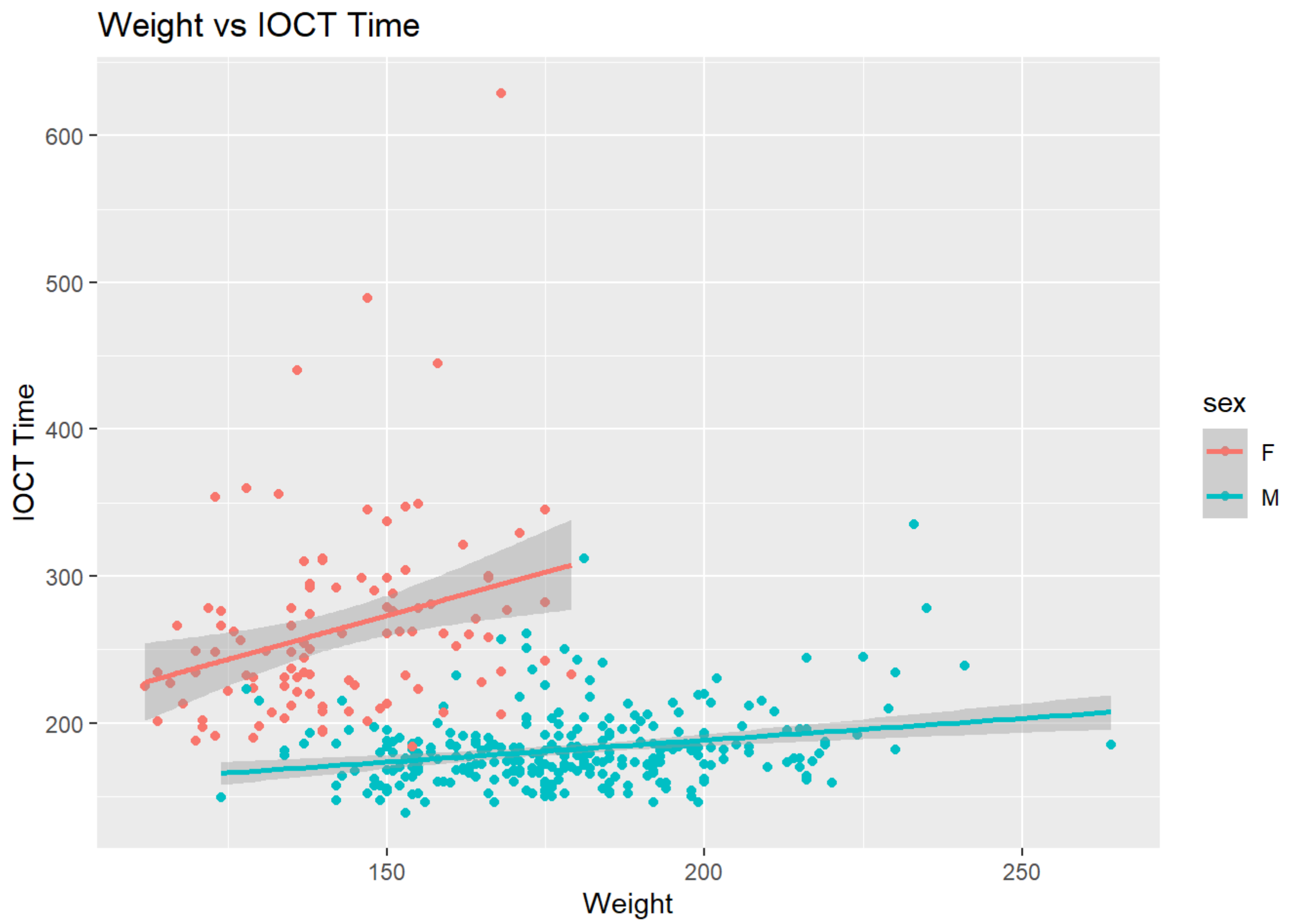
```
## `geom_smooth()` using formula 'y ~ x'
```



Weight V. IOCT Time

```
ioct %>%
  ggplot(aes(x = weight, y = IOCT_Time, color = sex)) +
  geom_point() +
  geom_smooth(method = "lm")+
  labs(title = "Weight vs IOCT Time")+
  ylab("IOCT Time") +
  xlab("Weight")
```

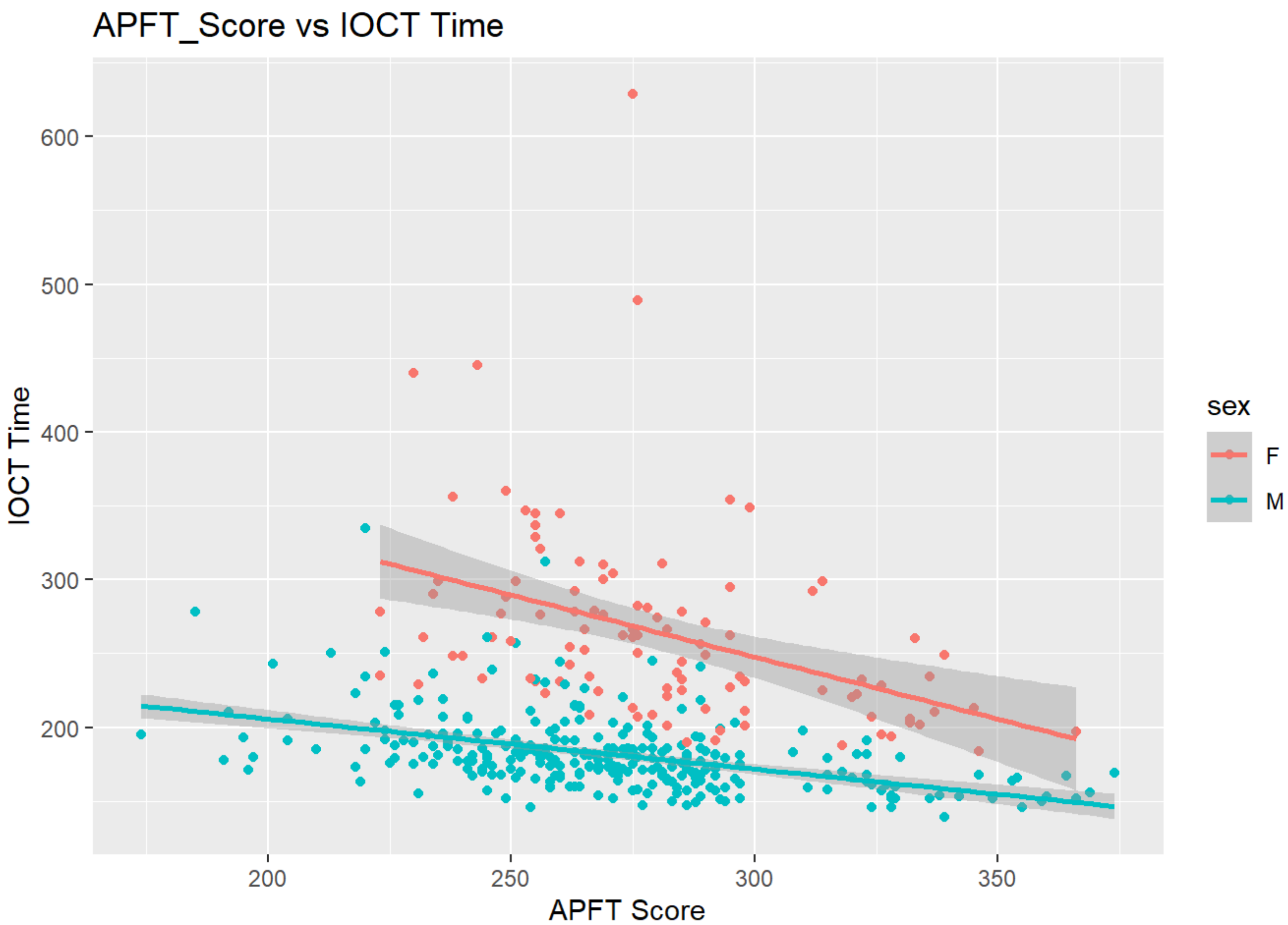
```
## `geom_smooth()` using formula 'y ~ x'
```



APFT v. IOCT TIme

```
ioct %>%
  ggplot(aes(x = APFT_Score, y = IOCT_Time, color = sex)) +
  geom_point() +
  geom_smooth(method = "lm")+
  labs(title = "APFT_Score vs IOCT Time")+
  ylab("IOCT Time") +
  xlab("APFT Score")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Liner Regression #1

```
model1 <- lm(IOCT_Time ~ weight, data = ic)
summary(model1)
```

```
##
## Call:
## lm(formula = IOCT_Time ~ weight, data = ic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.894 -14.300  -3.842  10.416  83.508
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 138.36496   10.93285  12.656  < 2e-16 ***
## weight        0.23884    0.06099   3.916 0.000113 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.36 on 276 degrees of freedom
## Multiple R-squared:  0.05264,    Adjusted R-squared:  0.04921
## F-statistic: 15.34 on 1 and 276 DF,  p-value: 0.0001135
```

```
model2 <- lm(IOCT_Time ~ height + weight, data = ic)
summary(model2)
```

```
##
## Call:
## lm(formula = IOCT_Time ~ height + weight, data = ic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.381 -13.838  -2.773  11.264  82.836
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 220.38768   36.12963   6.100 3.59e-09 ***
## height       -1.38104    0.58029  -2.380    0.018 *
## weight        0.32496    0.07048   4.611 6.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.18 on 275 degrees of freedom
## Multiple R-squared:  0.07176,    Adjusted R-squared:  0.06501
## F-statistic: 10.63 on 2 and 275 DF,  p-value: 3.576e-05
```

```
model3 <- lm(IOCT_Time ~ APFT_Score, data = ic)
summary(model3)
```

```
##
## Call:
## lm(formula = IOCT_Time ~ APFT_Score, data = ic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.026 -13.076  -3.524   7.798  72.141
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 265.97440    9.42394  28.223   <2e-16 ***
## APFT_Score   -0.31476    0.03456  -9.107   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.15 on 276 degrees of freedom
## Multiple R-squared:  0.2311, Adjusted R-squared:  0.2283
## F-statistic: 82.94 on 1 and 276 DF,  p-value: < 2.2e-16
```

```
model4 <- lm(IOCT_Time ~ APFT_Score + weight, data = ic)
summary(model4)
```

```
##
## Call:
## lm(formula = IOCT_Time ~ APFT_Score + weight, data = ic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.127 -13.301  -3.743   9.171  73.576
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 225.45002   13.46852  16.739  < 2e-16 ***
## APFT_Score   -0.30934    0.03363  -9.197  < 2e-16 ***
## weight        0.21955    0.05347   4.106 5.32e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.59 on 275 degrees of freedom
## Multiple R-squared:  0.2755, Adjusted R-squared:  0.2702
## F-statistic: 52.28 on 2 and 275 DF,  p-value: < 2.2e-16
```

```
aic1 <- AIC(model1)
bic1 <- BIC(model1)
aic2 <- AIC(model2)
bic2 <- BIC(model2)
aic3 <- AIC(model3)
bic3 <- BIC(model3)
aic4 <- AIC(model4)
bic4 <- BIC(model4)


print(c("Model1", aic1, bic1, "Model2", aic2, bic2, "Modle3", aic3, bic3, "Model4", aic4, bic
4))
```

```
##  [1] "Model1"           "2520.66792218962" "2531.55078553069" "Model2"
##  [5] "2517.00032046338" "2531.51080491814" "Modle3"           "2462.65310267819"
##  [9] "2473.53596601926" "Model4"           "2448.11324821215" "2462.62373266691"
```

```
anova(model1, model2)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	276	138041.8	NA	NA	NA	NA
2	275	135256.1	1	2785.773	5.66398	0.01799872

2 rows

```
anova(model3, model4)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	276	112041.6	NA	NA	NA	NA
2	275	105570.0	1	6471.59	16.85788	5.316446e-05

2 rows