UNITED STATES MILITARY ACADEMY

HOMEWORK 1

MA478

SECTION H2

COL. CLARK

BY

CADET AIMEE ROHAN’25, CO I3

WEST POINT, NEW YORK

29 JAN 2024

_ MY DOCUMENT IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT.

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

SIGNATURE:

Homework 1 continued

Aimee Rohan Ramirez

January 2024

1. Suppose X and W are two design matrices with n rows and $C(X) = C(W)$. Show that $PX = PW$.
Hint: If B is contained in Col(X), then $P_X B = B$. I got this during AI.

Since B is contained in Col(X) and $C(X) = C(W)$, we have $P_X B = B$.
Similarly, if P_W is contained in Col(X) (which is true since $C(X) = C(W)$), then $P_X P_W = P_W$.
Therefore, $PX = PW$.

What would this mean for $(PX - PW)^T(PX - PW)$? $(PX - PW)^T(PX - PW) = 0$.

2. Up to now, we have been considering models of the form $E[Y_i] = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$. What would happen if, instead, we fit the model:

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

where $\bar{x}_j$ denotes the column averages? Why does this make sense in light of what you showed above?

If the predictors x_1 and x_2 are already in the same column space (i.e., $C(X) = C(W)$), and you mean-center them by using $x_{1,i} - \bar{x}_1$ and $x_{2,i} - \bar{x}_2$, then the mean-centered terms $\bar{x}_1$ and $\bar{x}_2$ would become zeros in the model.
This simplifies the model to:

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

becomes $E[Y_i] = \beta_0$

Homework 1

Aimee Rohan Ramirez

January 2024

1 Introduction

This analysis aims to investigate the relationship between an individual's height and their performance in the Indoor Obstacle Course Test (IOCT). The primary question at hand is whether height significantly influences IOCT completion times. The goal is to build a linear regression model addressing the relationship between our possible predictors. The objective is to provide a clear and concise understanding of the impact of height on IOCT times.

2 Data exploration

The data consisted of 384 unit variables consisting of columns referring to their sex, height, weight, APFT score, IOCT time, and the individual event scores for each of the APFT events. The analysis began with a univariate examination of IOCT times, presenting the distribution through a histogram. The observed distribution revealed a rightward skewness, indicating that the majority of IOCT times are lower, with a tail extending towards higher values (Figure 1).

I performed a bivariate analysis examining the relationship between height and IOCT time, distinguishing data points by gender. The observed pattern reveals that, on average, females tend to have lower heights compared to males, and consequently, higher IOCT times (Figure 2). The overall trend in Figure 2 indicates a decrease towards the right, implying that taller individuals exhibit better IOCT times. Additionally, this pattern hints at a potential association with gender, suggesting that the relationship between height and IOCT time may be influenced by gender differences.

The final steps I have conducted is analyzing hidden relationships between my other predictors. When examining the correlation matrix with variables height, weight, and IOCT time,strong correlation of 0.7142015 between height and weight implies that these two variables share some degree of information (Figure 3). This implies a level of collinearity between these variables, indicating that a portion of the variability in one variable could be explained by the other. From seeing this, I made the decision to overall to not include weight in my models.

Lastly, I decided to see if there was an interaction between sex and height due clustering of males and females in Figure 2. The subsequent examination of the interaction plot (Figure 4) depicting IOCT time concerning sex and height revealed lines representing distinct gender groups displayed non-parallel slopes. This non-parallelism serves as a statistical indication of a discernible interaction effect between height and sex. As a result, this finding substantiates the inclusion of the interaction effect in the formulation of the three models.

3 Data preparation

Considering outliers in the data given, I used Cooks Distance to analyze if there were any outliers that would sway the data did have a lot of influence. CDT 113 had a score of 0.2377996. This score was not super close to 1 despite it looking far away so it is apparent that the point did not have enough sway on the data.

4 Models

Model 1: In this multiple linear regression model, we are exploring the relationship between IOCT Time, our response variable, and two predictor variables: height and sex. The model is formulated as:

$$IOCT\_Time = \beta_0 + \beta_1 \cdot height + \beta_2 \cdot sexM + \beta_3 \cdot (height \times sexM) + \varepsilon$$

The p values associated with using the summary of this model in Figure 5 suggested that there was not sufficient evidence to conclude a significant interaction effect.

To further elaborate on how well the model fit, I conducted residuals vs. fitted values plot(Figure 6). There was a presence of patterns or clusters in the residuals plot indicating that this model is not capturing certain aspects of the data. It could be that the relationship between the predictors and the response variable is not adequately described by a simple linear model.

Model 2: This model aims to predict the IOCT (Indoor Obstacle Test) time based on the values of the predictor variables. In addition to the main effects of height, sex, and APFT Score, the model considers the joint effect of height and sex, providing a more nuanced understanding of how these variables collectively influence IOCT time.

$$IOCT\_Time = \beta_0 + \beta_1 \cdot \text{height} + \beta_2 \cdot \text{sexM} + \beta_3 \cdot (\text{height} \times \text{sexM}) + \beta_4 \cdot \text{APFT\_Score} + \varepsilon$$

From conducting a summary on this data (Figure 7), height and the interaction term (height:sex_male) are not statistically significant (p >0.05), while sex and APFT Score are significant.This means that, according to the statistical tests, there is evidence to suggest that Gender (sex_male) and APFT score have a significant impact on IOCT times, while Height and the interaction term do not show significant associations in this model. Main takeaways from this model:

1. Gender - The estimated effect of being male (compared to being female) is a decrease of approximately 225.01 seconds in the IOCT time. This suggests that, on average, males have lower IOCT times than females.

2. APFT Score - For each one-unit increase in APFT score, the IOCT time is expected to decrease by approximately 0.46. This indicates a negative relationship, suggesting that higher APFT scores are associated with lower IOCT times, meaning individuals with higher APFT scores tend to have faster IOCT times.

3. Height- For each one-unit increase in height, IOCT time is expected to decrease by approximately 2.07 seconds. This suggests a negative relationship between height and IOCT time, meaning that taller individuals tend to have a faster IOCT time. However since the coefficient is not statistically significant, the effect of height is uncertain

The low p-value for the F-statistic indicates that the model, with all the predictors (height, sex, APFT score, and the interaction term), is statistically significant in explaining the variance in IOCT time.

In validating the findings with this model, a residuals vs. fitted plot was drawn from this model, Figure 8, where we saw more equal variance variance across the horizontal red line with only slight clustering. Since it is more spread out in comparison to model one it demonstrates that this model is capturing the variability in the data better.

Model 3: The gamma regression model is employed to model the distribution of IOCT times. The choice of the gamma distribution is motivated by the observed right-skewness in the data, where females tend to have lower heights, contributing to the skewness. The log link function ($\log(\mu)$) is used to ensure that the predicted mean (μ) is always positive.

The model includes the main effects of height, sexM (indicator variable for males), and their interaction. Additionally, the APFT_Score is included as a predictor. This model aims to capture the non-linear relationships between these variables and IOCT times, considering their joint effects.

$$\text{IOCT\_Time} \sim \Gamma(\mu),$$

$$\log(\mu) = \beta_0 + \beta_1 \cdot \text{height} + \beta_2 \cdot \text{sexM} + \beta_3 \cdot (\text{height} \times \text{sexM}) + \beta_4 \cdot \text{APFT\_Score}$$

From conducting a summary on this model, Figure 9, we found that being male (coded as 1) is associated with a decrease in the log of the mean IOCT time by approximately 0.926 units compared to being female (coded as 0). This effect is statistically significant (p-value = 0.0427). What was also significant was our APFT score as for each one-unit increase in APFT Score, the log of the mean IOCT time is expected to decrease by approximately 0.0022 units. This effect is statistically significant (p-value < 0.05).

The model suggests that, after adjusting for other variables, being male and having a higher APFT score are associated with a decrease in the log of the mean IOCT time. The impact of height is less clear in this model, and the interaction effect between height and gender is not statistically significant.

Finally, to confer using AIC (Akaike Information Criterion), it validates that our model does explain the variability in the data reasonably well despite it incorporating a penalty for the number of parameters in the model. From also using a residuals plot, I was also able to see equal variance with only slight clustering on the right.

5 Best Model chosen and why

The best model was model 2 as the model's simplicity, interpretability, and its ability to capture the essential relationships between IOCT time, height, gender, and APFT Score. Despite being less complex than the gamma model, the fitted residuals displayed a spread comparable to the more intricate model, indicating that this model adequately captures the underlying patterns in the data.

It is important to add the disclaimer that some effects particularly those related to height, were potentially masked by collinearity issues, specifically the strong correlation observed between height and weight. It is important to acknowledge that our model might not fully disentangle the unique effects of height and weight on IOCT time due to this collinearity.

This analysis suggests that while height alone may not be a decisive factor in determining IOCT time, its significance is contingent upon gender and APFT Score. The model highlights that gender and fitness level, play pivotal roles in predicting IOCT time. Therefore, the impact of height on IOCT time appears to be nuanced and intertwined with other factors, emphasizing the importance of considering multiple covariates in the analysis.

6 Appendices

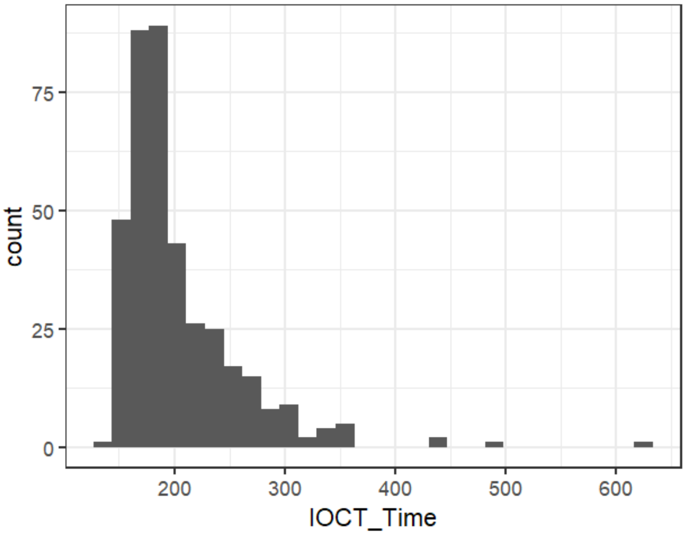


Figure 1: Univariate analysis of IOCT Times

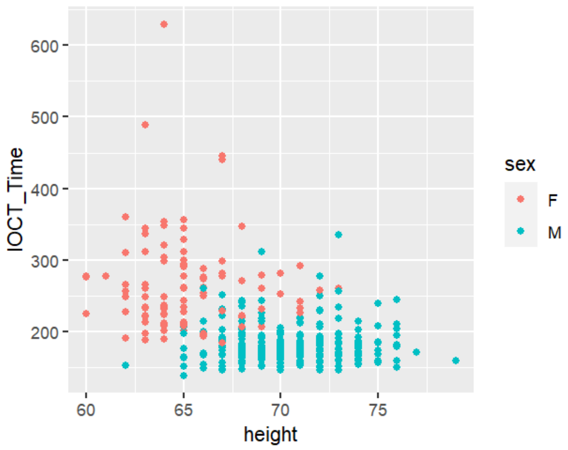
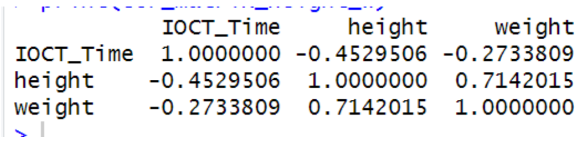


Figure 2: Bivariate analysis of height and sex



```
           IOCT_Time     height     weight
IOCT_Time  1.0000000 -0.4529506 -0.2733809
height    -0.4529506  1.0000000  0.7142015
weight    -0.2733809  0.7142015  1.0000000
>
```

Figure 3: Correlation matrix of height, sex, and IOCT time

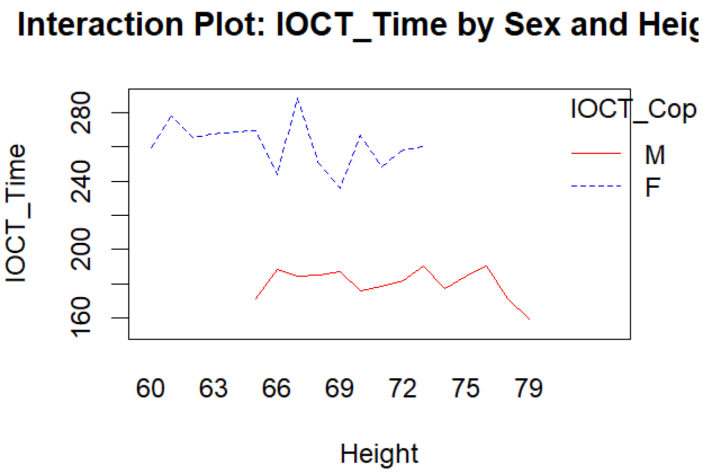
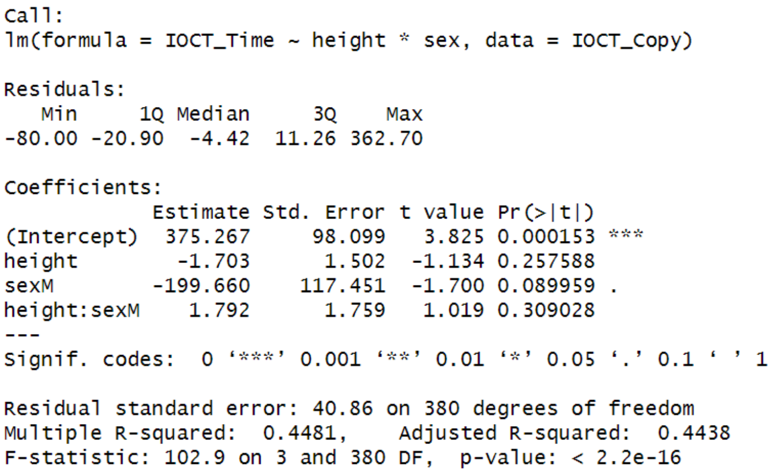


Figure 4: Interaction plot by sex and height



```
Call:
lm(formula = IOCT_Time ~ height * sex, data = IOCT_Copy)

Residuals:
   Min      1Q  Median      3Q     Max 
-80.00 -20.90  -4.42  11.26 362.70 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  375.267     98.099   3.825 0.000153 ***
height        -1.703      1.502  -1.134 0.257588    
sexM        -199.660    117.451  -1.700 0.089959 .  
height:sexM    1.792      1.759   1.019 0.309028    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.86 on 380 degrees of freedom
Multiple R-squared:  0.4481,	Adjusted R-squared:  0.4438 
F-statistic: 102.9 on 3 and 380 DF,  p-value: < 2.2e-16
```

Figure 5: Summary for Model 1

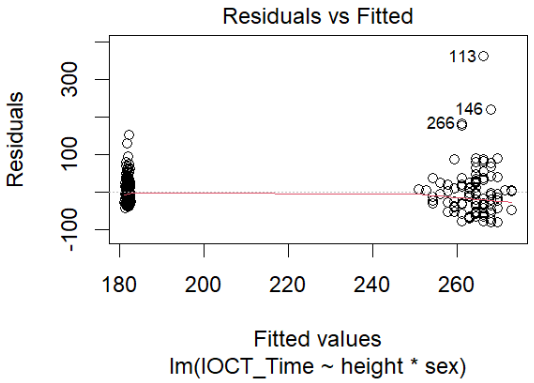
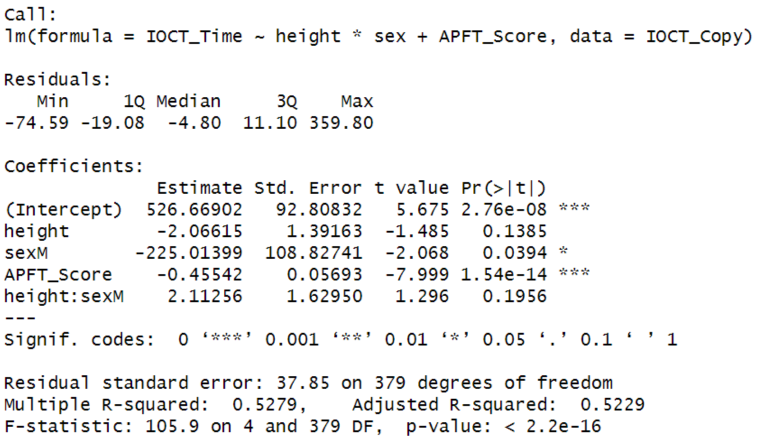


Figure 6: Residuals plot for Model 1



```
Call:
lm(formula = IOCT_Time ~ height * sex + APFT_Score, data = IOCT_Copy)

Residuals:
   Min      1Q  Median      3Q     Max 
-74.59 -19.08  -4.80  11.10 359.80 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  526.66902   92.80832   5.675 2.76e-08 ***
height        -2.06615    1.39163  -1.485   0.1385    
sexM        -225.01399  108.82741  -2.068   0.0394 *  
APFT_Score    -0.45542    0.05693  -7.999 1.54e-14 ***
height:sexM    2.11256    1.62950   1.296   0.1956    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37.85 on 379 degrees of freedom
Multiple R-squared:  0.5279,	Adjusted R-squared:  0.5229 
F-statistic: 105.9 on 4 and 379 DF,  p-value: < 2.2e-16
```

Figure 7: Summary Stats for model 2

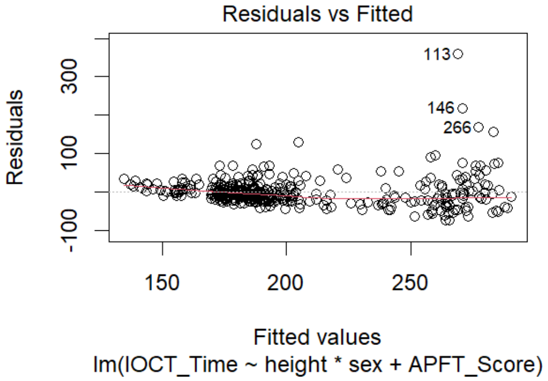
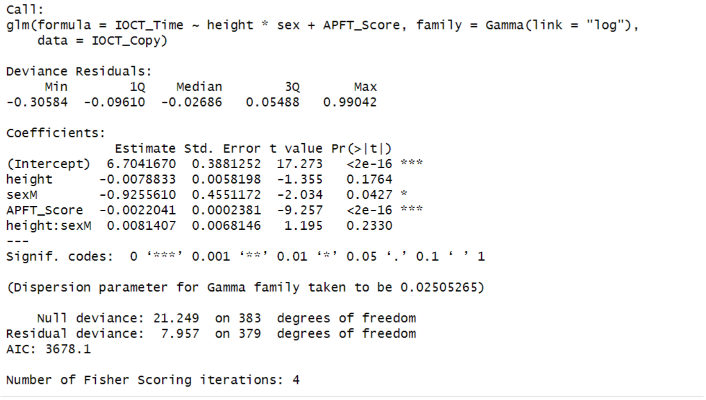


Figure 8: A demonstration of better fit using model 2



```
Call:
glm(formula = IOCT_Time ~ height * sex + APFT_Score, family = Gamma(link = "log"),
    data = IOCT_Copy)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.30584  -0.09610  -0.02686   0.05488   0.99042

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.7041670  0.3881252  17.273   <2e-16 ***
height      -0.0078833  0.0058198  -1.355   0.1764
sexM        -0.9255610  0.4551172  -2.034   0.0427 *
APFT_Score  -0.0022041  0.0002381  -9.257   <2e-16 ***
height:sexM  0.0081407  0.0068146   1.195   0.2330
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.02505265)

    Null deviance: 21.249  on 383  degrees of freedom
Residual deviance:  7.957  on 379  degrees of freedom
AIC: 3678.1

Number of Fisher Scoring iterations: 4
```

Figure 9: A demonstration of APFT scores being significant in model 3

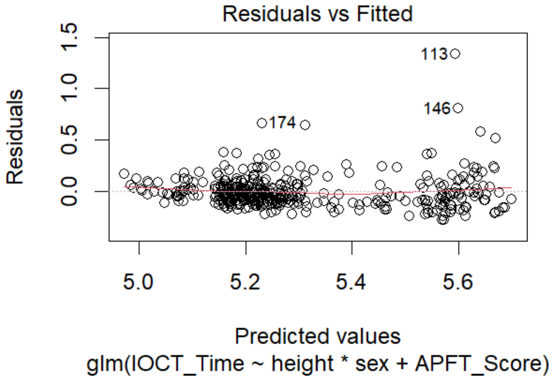


Figure 10: Residuals vs. Fitted plot for model 3 demonstrating a better fit