Exam II R

Filip Wilhelm Sjostrand

18/11-22

Problem 1

(a) Write down the estimated linear regression model for the full model.

$\hat{Y} = 4.562 + 0.002899X_1 + 0.00723X_2 - 6.854\times10^{-4}X_3 - 4.318\times10^{-4}X_4 - 0.5207X_{5,f} + 0.09429X_{6,S} + 0.2043X_{6,R}$

(b) Interpret b_1 and b_6 in terms of the problem.

b_1**:** Holding all other variables constant, when lean body mass increase by 1 kg, the red blood cell count increase by 0.002899 per liter, on average.

b_6**:** Holding all other variables constant, the average difference between a net sports player and swim sports player is 0.09429 per liter of red blood cell count.

(c) Test to see if we can drop the X_1, X_2, and X_3 from the model. State the hypotheses in terms of the betas, test statistic, p-value, and conclusion.

Hypothesis: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ VS. H_a : at least one $\beta_i \neq 0$

Test statistic: $F = 0.9425$

P-value= 0.4211

Conclusion: at 5% significance level we fail to reject H_0 and conclude that there is insufficient evidence that X_1, X_2, and X_3 has significant slopes. Thus, they can be dropped from the model.

(d) Based on your results from part (c), test to see if we can drop X_6 from the "best" model. State the hypotheses in terms of the betas, test statistic, p-value, and conclusion.

Hypothesis: $H_0 : \beta_6 = \beta_7 = 0$ VS. H_a : at least one $\beta_i \neq 0$

Test statistic: $F = 5.34$

P-value = 0.005516

Conclusion: at 5% significance level we reject H_0 and conclude that there is sufficient evidence that β_6 and/or β_7 has a significant slope. Thus, they cannot be dropped from the model.

(e) What is the additional reduction in error we expect to see when we add X_5 to a model which already includes X_4 and X_6.

Adding sex to a model which already includes plasma ferritins and sport reduces error by 39.04 %.

Problem 2

(a) Based on your "best" model above, predict the red blood cell count of a male subject who a ferritin level of 20ng and swims.

Answer: 5.03 per liter.

(b) Find the 95% confidence interval for only the value β_1, and interpret it in terms of the problem.

We are 95% confident that the true slope for Plasma ferritins is between -0.0014 and 0.0007. Thus, we cannot conclude that the slope is different from zero.

(c) Create two prediction intervals for the red blood cell count for a male who runs with a ferratin level of 15 and a female who swims with a ferratin level of 12 with overall/simultaneous/family-wise level 90%.

	2.5%	97.5%
Male, runs, 15ng	4.467788	5.790153
Female, swims, 12ng	3.783829	5.101680

(d) Interpret the intervals in (c) in terms of the problem.

We are overall 90% confident that a male who runs with 15 in ferratin level has a blood plasma count between 4.468 and 5.79 per liter, and a woman who swims with 12 in ferratin level has a bloom plasma count between 3.784 and 5.102 per liter.

Problem 3

(a) Fit the model which includes X_3 and X_6 and all interaction terms. Report back the estimated regression line.

$$\hat{Y} = 5.084 - 0.03544X_3 + 0.2214X_{6,S} - 0.0843X_{6,R} - 0.004685X_3X_{6,S} + 0.02358X_3X_{6,R}$$

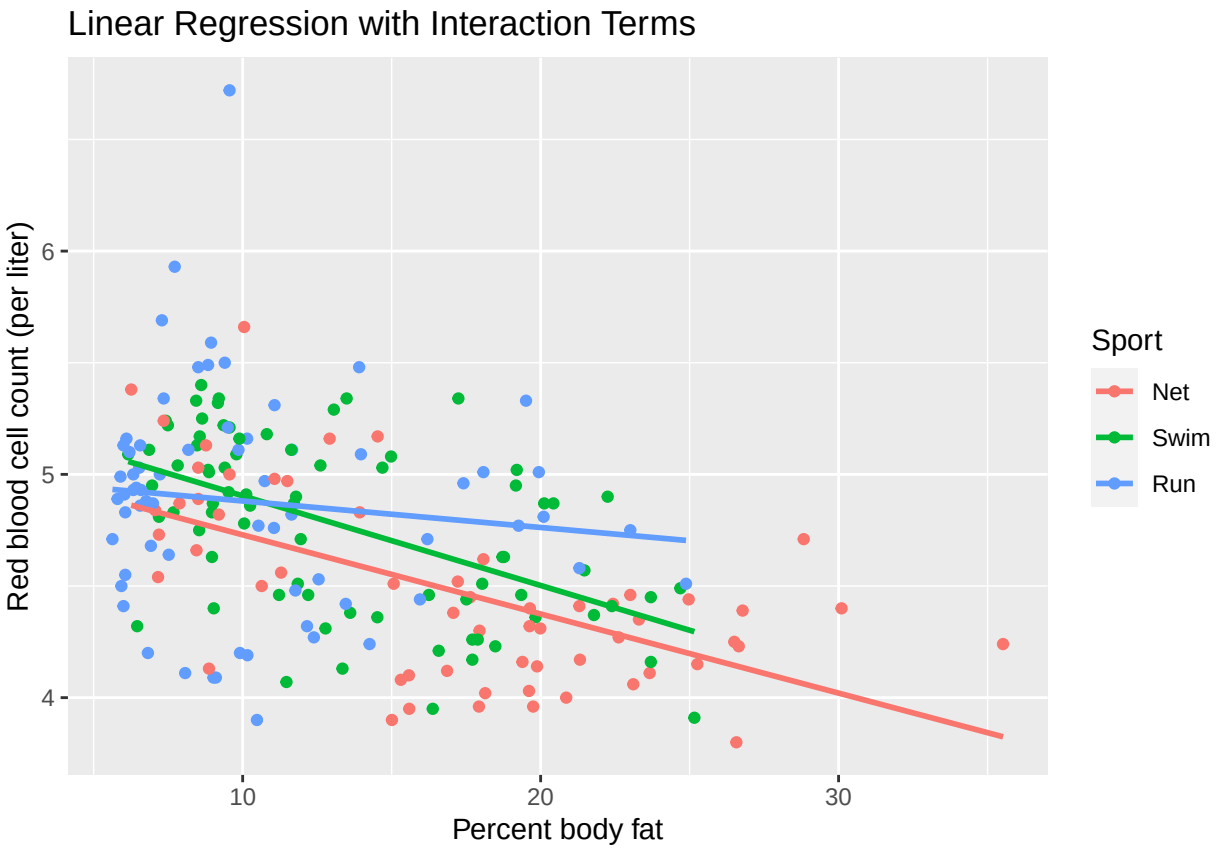
(b) Test to see if you can drop the interaction term from the model in (a), using $\alpha = 0.10$ State the null and alternative in terms of the betas, the p-value, and your conclusion in terms of the problem.

Hypothesis: $H_0 : \beta_4 = \beta_5 = 0$ VS. H_a : at least one $\beta_i \neq 0$

p-value $= 0.0804$

Conclusion: Given that p-value $< \alpha$, we reject H_0. Thus we conclude at 10% significance level that there is sufficient evidence that the interaction terms should not be dropped.

(c) Using ggplot2, plot the different lines suggested by the interaction term on the plot with Y and X_3.



(d) State the different models associated with the different levels of X_6. Explain the relationship.

Net: $\hat{Y} = 5.084 - 0.03544X_3$

Swim: $\hat{Y} = 5.3054 - 0.040125X_3$

Run: $\hat{Y} = 4.9997 - 0.01186X_3$

There is a negative relationship between body fat percentage and red blood cell count. On average, holding body fat constant, swimmers have a higher count of red blood cells compared to runners and net players, and runners have a higher than net players. However, per increase in body fat percentage, swimmers are expected to have the biggest decrease in red blood cell count on average while runners are expected to have the smallest.

(e) Find and interpret $R^2\{X_3X_6|X_3, X_6\}$

When we add interaction terms to a model that already contain body fat percentage and sports, the error is reduced by 2.539%.

Code Appendix

```
# Packages ----------
library(readr)
library(ggplot2)
library(dplyr)

# Get data ----------
df <- read_csv(
  file = "~/Downloads/athelete.csv",
  col_types = cols(
    sex = col_factor(levels = c("m", "f")),
    newsport = col_factor(levels = c("Net", "Swim", "Run"))
    )
  )



# -------------------- QUESTION I --------------------

# Full model ----------
model <- lm(formula = rcc ~ lbm + bmi + pcBfat + ferr + sex + newsport, data=df)

b0 <- signif(model$coefficients[1], 4)
b1 <- signif(model$coefficients[2], 4)
b2 <- signif(model$coefficients[3], 4)
b3 <- signif(model$coefficients[4], 4)
b4 <- signif(model$coefficients[5], 4)
b5 <- signif(model$coefficients[6], 4)
b6 <- signif(model$coefficients[7], 4)
b7 <- signif(model$coefficients[8], 4)

# Full vs reduced ----------
reduced <- lm(formula = rcc ~ ferr + sex + newsport, data=df)
sse_r <- anova(reduced)$`Sum Sq`[4]
df_r <- anova(reduced)$Df[4]

sse_f <- anova(model)$`Sum Sq`[7]
df_f <- anova(model)$Df[7]

Fstat <- ( (sse_r - sse_f) / (df_r - df_f) ) / (sse_f / df_f)

pvalue <- pf(Fstat, (df_r - df_f), df_f, lower.tail = FALSE)

# Full vs reduced ----------
full <- lm(formula = rcc ~ ferr + sex + newsport, data=df)
sse_f <- anova(full)$`Sum Sq`[4]
df_f <- anova(full)$Df[4]

reduced <- lm(formula = rcc ~ ferr + sex, data=df)
sse_r <- anova(reduced)$`Sum Sq`[3]
df_r <- anova(reduced)$Df[3]
```

```r
Fstat <- ( (sse_r - sse_f) / (df_r - df_f) ) / (sse_f / df_f)

pvalue <- pf(Fstat, (df_r - df_f), df_f, lower.tail = FALSE)

# Partial r-squared ----------
before <- lm(formula = rcc ~ ferr + newsport, data=df)
sse_b <- anova(before)$`Sum Sq`[3]

after <- lm(formula = rcc ~ ferr + sex + newsport, data=df)
sse_a <- anova(after)$`Sum Sq`[4]

part_r <- (sse_b - sse_a)/sse_b



# -------------------- QUESTION II --------------------

# Prediction ----------
best <- lm(formula = rcc ~ ferr + sex + newsport, data=df)
newdata <- data.frame(ferr = 20, sex = "m", newsport = "Swim")
pred <- predict(best, newdata)

# Confidence interval ----------
ci = confint(best, level = 0.95)

# Predictions----------
newdata = data.frame(
  ferr = c(15, 12),
  sex = c("m", "f"),
  newsport = c("Run", "Swim")
  )

preds <- predict(best, newdata)

# Find best multiplier ----------
Multiplier = function(n,p,g,alpha){
  Bon = qt(1-alpha/(2*g), n-p)
  WH = sqrt(p*qf(1-alpha,p,n-p))
  Sch = sqrt(g*qf(1-alpha,g, n-p))
  the.multipliers = round(c(Bon,WH,Sch), 4)
  names(the.multipliers) = c("Bonferroni", "WH", "Scheffe")
  return(the.multipliers)
}

bon <- Multiplier(nrow(df), 5, 2, 0.1)[1] #Bonferroni was the smallest

# Error ----------
se = predict(best, newdata, interval = 'prediction', se.fit = TRUE)$se.fit
MSE = sum(best$residuals^2)/ (length(best$residuals) - length(best$coefficients))
pred.se = sqrt(se^2 + MSE)

# Prediction intervals ----------
intervals <- cbind(preds - bon*pred.se, preds + bon*pred.se)
```

```r
colnames(intervals) <- c("2.5%", "97.5%" )
rownames(intervals) <- c("Male, runs, 15ng", "Female, swims, 12ng")



# -------------------- QUESTION III --------------------

# Interaction model ----------
int_model <- lm(formula = rcc ~ pcBfat*newsport, data=df)
coef <- int_model$coefficients
b0 <- signif(coef[1], 4)
b1 <- signif(coef[2], 4)
b2 <- signif(coef[3], 4)
b3 <- signif(coef[4], 4)
b4 <- signif(coef[5], 4)
b5 <- signif(coef[6], 4)

# Test of interaction ----------
noint_model <- lm(formula = rcc ~ pcBfat + newsport, data=df)
result <- anova(noint_model, int_model)
pval <- result$`Pr(>F)`[2]

int_model %>%
  ggplot(aes(x = pcBfat, y = rcc, group = newsport, color = newsport)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ylab("Red blood cell count (per liter)") +
  xlab("Percent body fat") +
  labs(color = 'Sport') +
  labs(title = "Linear Regression with Interaction Terms")


# Partial r-square ----------
sse_b <-  sum(summary(noint_model)$residuals^2)
sse_a <-  sum(summary(int_model)$residuals^2)
r <-  signif(((sse_b - sse_a)/sse_b)*100, 4)
```