Carson Slater STA 5381 Homework #3

Use the following information to answer the next 5 questions. To model the thrust of a jet turbine engine (y), five candidate regressors are recorded for $n = 25$ tests on a given jet. The explanatory variables are x_1 = primary speed of rotation, x_2 = secondary speed of rotation, x_3 = fuel flow rate, x_4 = pressure, and x_5 = exhaust temperature. Below is the ANOVA table for the full model.

```
Response: y
           Df Sum Sq Mean Sq F value   Pr(>F)
x1          1 13.045  13.045  2.6516 0.119912
x2          1 47.708  47.708  9.6975 0.005713 **
x3          1 11.871  11.871  2.4130 0.136828
x4          1 40.906  40.906  8.3149 0.009516 **
x5          1  0.402   0.402  0.0817 0.778160
Residuals 19 93.472   4.920
```

1. What are the test statistic and p-value needed to test the claim that any of the predictors are useful in explaining the thrust of the jet engine?

   (a) $TS = 2.6516$; p-value$= 0.1199$

   (b) $TS = 23.157$; p-value$= 1.782 \times 10^{-7}$

   (c) $TS = 0.0817$; p-value$= 0.7782$

   (d) $TS = 4.920$; p-value$= 0.0047$

   (e) $TS = 4.632$; p-value$= 0.0062$

2. What type of test would be appropriate to use to determine if x_5 can be removed from the model given that the other variables remain in the model?

   (a) t-test for $H_0 : \beta_5 = 0$

   (b) partial F-test for $H_0 : \beta_5 = 0$

   (c) overall F-test for $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.

   (d) nested F-test for $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

   (e) two of the above (circle which ones)

3. The researchers would like to test that the reduced model $y = \beta_0 + \beta_2 x_2 + \beta_4 x_4 + \epsilon$ is an adequate fit to the data over the full model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$. What is the appropriate test statistic, and what is its distribution?

   (a) $TS = 1.715$; F-distribution with 3 and 19 degrees of freedom.

   (b) $TS = 1.715$; F-distribution with 19 and 3 degrees of freedom.

   (c) $TS = 0.329$; F-distribution with 3 and 19 degrees of freedom.

   (d) None of (a) through (c) because the models are not nested.

   (e) Not enough information given to determine.

4. The multiple and adjusted R^2 values for a reduced model with x_5 removed are

```
Multiple R-squared: 0.5474, Adjusted R-squared: 0.4569
```

Based on these values, would you choose the full model or the reduced model without x_5?

(a) The full model because it has a higher R^2.

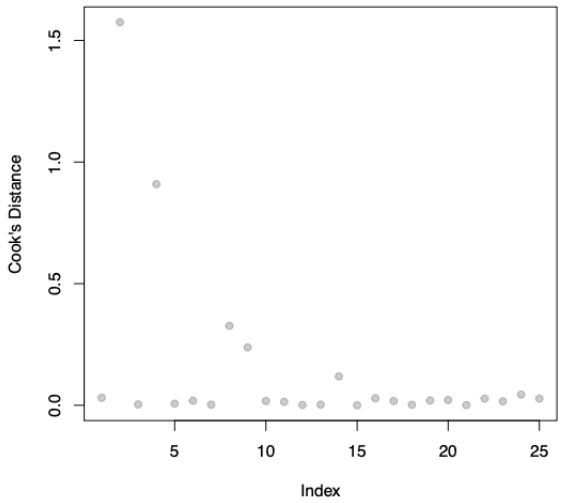
(b) The reduced model because it has a higher R^2.

(c) The full model because it has a higher R^2_{adj}.

(d) The reduced model because it has a higher R^2_{adj}.

(e) Not enough information to determine.

5. The following figure gives an index plot of Cook's distance for each observation in the full model.



Based on this figure, you can conclude the following:

(a) None of the observations have unusual values in their explanatory variables.

(b) At least two observations have high leverage.

(c) One of the observations has an influential effect on the least squares estimate of $\boldsymbol{\beta}$.

(d) Two of the observations have a really large residual.

(e) None of the observations are influential in estimating the full model.

6. Chapter 4, # 1 The website has changed. The 2005-2006 salary data are:

```
years <- c(0, 2, 4, 6, 8, 12, 17, 22, 28, 34) #Years of experience
size <- c(17, 33, 19, 25, 18, 60, 58, 31, 34, 19) #Number of people used to calculat
sal <- c(101300, 111303, 98000, 124000, 128475, 117410, 115825,
          134300, 128066, 164700) #3rd quartile of salaries
```

To see similar more recent data, you can go to `https://www.amstat.org/your-career/salary-information`. Let the weights based on the number of professors used to calculate the values. Add the following "Compute the prediction intervals for the 2005-2006 third quartile for the salary of full professors with 6 years of experience based on the least squares and the weighted least squares models and compare them."

```
stat_salary <- cbind(years, size, sal) |> as.data.frame()

ols <- lm(sal ~ years, data = stat_salary)
wls <- lm(sal ~ years, data = stat_salary, weights = size)

ols |> summary(); wls |> summary()


##
## Call:
## lm(formula = sal ~ years, data = stat_salary)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14150  -9430  -1428   9712  14370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 104352.9     5619.4  18.570 7.29e-08 ***
## years         1352.3      325.7   4.152   0.0032 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11320 on 8 degrees of freedom
## Multiple R-squared:  0.683,Adjusted R-squared:  0.6434
## F-statistic: 17.24 on 1 and 8 DF,  p-value: 0.0032
##
## Call:
## lm(formula = sal ~ years, data = stat_salary, weights = size)
##
## Weighted Residuals:
##    Min     1Q Median     3Q    Max
## -67520 -40994   4937  51648  87516
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 104759.0     5752.2   18.21 8.49e-08 ***
## years         1172.5      336.9    3.48  0.00832 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57620 on 8 degrees of freedom
## Multiple R-squared:  0.6022,Adjusted R-squared:  0.5524
## F-statistic: 12.11 on 1 and 8 DF,  p-value: 0.008323
```

```
new_data <- data.frame(years = c(6))

predict(ols, newdata = data.frame(years = new_data),
        interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 112466.4 84544.64 140388.3
```

```
predict(wls, newdata = data.frame(years = new_data),
        weights = c(25), interval = "prediction", level = 0.95)
```

```
##        fit      lwr    upr
## 1 111793.8 83478.58 140109
```

From predicting the salary for a full professor with 6 years of experience, the prediction interval for weighted least squares is marginally wider. Likewise the point estimate for the weighted least squares is slightly lower as well. The prediction interval for the OLS are (84544.64, 140388.3) and the prediction interval for WLS is (83478.58, 140109).

7. The following is artificial data. Do not standardize these values; just use them as given here.

```
x <- c(0, 0.5, 1, 10, 19, 19.5, 20)
y <- c(0.445, 1.206, 0.1, -2.198, 0.536, 0.329, -0.689)
```

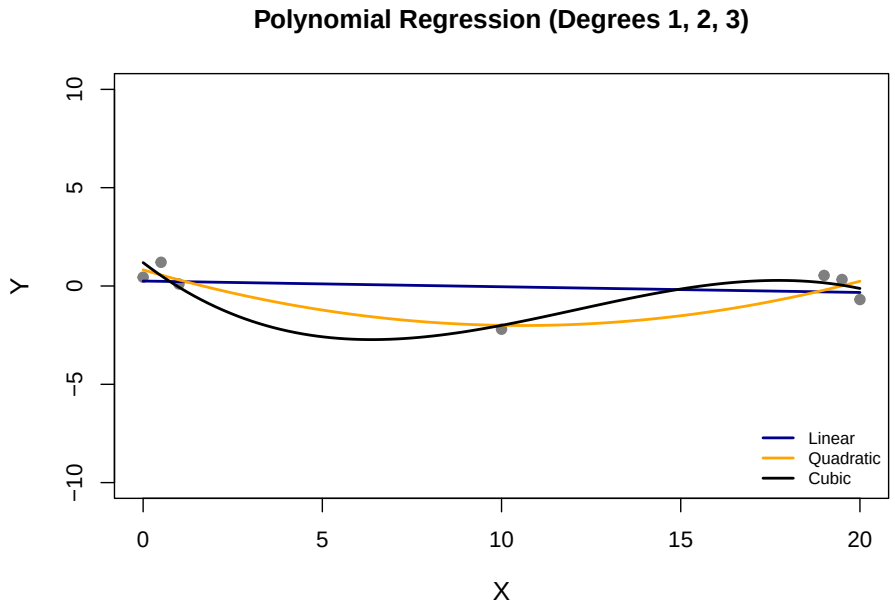
   (a) Use these values to fit a linear, quadratic, cubic, quartic, quintic, and hexic models. Give the estimated regression equations for each one.

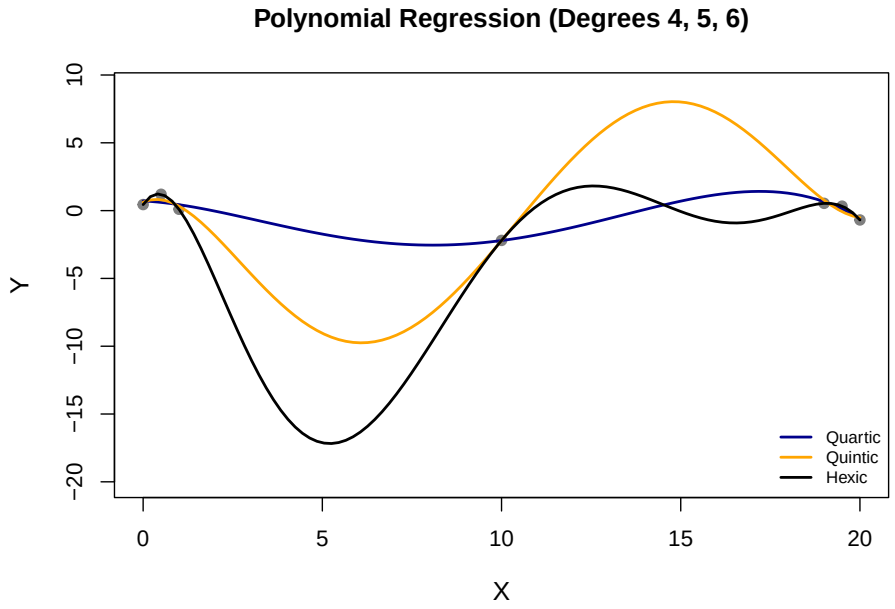
```
x <- c(0, 0.5, 1, 10, 19, 19.5, 20)
y <- c(0.445, 1.206, 0.1, -2.198, 0.536, 0.329, -0.689)

xy_df <- cbind(y, x) |> as.data.frame()

mods <- list()
for (i in 1:6) mods[[i]] <- lm(y ~ poly(x, degree = i, raw = TRUE),
                               data = xy_df)
mods <- mods |> setNames(c("Linear", "Quadratic",
                           "Cubic", "Quartic",
                           "Quintic", "Hexic"))
lapply(mods, summary) # got estimates from here
```

Model	Equation
Linear	$y = 0.25156 - 0.02903x$
Quadratic	$y = 0.821522 - 0.535658x + 0.025332x^2$
Cubic	$y = 1.188562 - 1.395530x + 0.148807x^2 - 0.004116x^3$
Quartic	$y = 0.7130112 - 0.1413714x - 0.1539149x^2 + 0.0198855x^3 - 0.0006000x^4$
Quintic	$y = 0.6226947 + 1.1444995x - 1.7196344x^2 + 0.3011085x^3 - 0.0177775x^4 + 0.0003435x^5$
Hexic	$y = 0.445 + 3.936x - 5.432x^2 + 1.263x^3 - 0.1174x^4 + 0.004876x^5 - 0.00007553x^6$

Table 1: Fitted regression models for different degrees of polynomial regression.

(b) Make a plot of y versus x. Overlay the fitted linear, quadratic, and cubic curves. Plot y vs x again, and overlay the fitted curves for the remaining 3 models.





(c) Given what you see in this example, explain what can be the danger in fitting very high order polynomials.

Higher order polynomials, in this example, tend to overfit to the points so much that they appear to miss the true trend of the data if there are far enough gaps in between the points.

8. Using the mercury data on the class website, check for multicollinearity among the predictors in the model $y = \beta_0 + \beta_1 \cdot \log(alkalinity) + \beta_2 \cdot \log(calcium) + \beta_3 \cdot pH + \epsilon$.

(a) Find the correlation matrix of the predictors.

```
mercury <- read.delim("mercury.txt",
                      sep ="", header = TRUE, dec =".") |>
  mutate("log_alkalinity" = log(alkalinity),
         "log_calcium" = log(calcium)) |>
  select(merc_ppm, log_alkalinity, log_calcium, pH)
cor(mercury)

##                  merc_ppm log_alkalinity log_calcium         pH
## merc_ppm        1.0000000     -0.7188845  -0.4903048 -0.6051330
## log_alkalinity -0.7188845      1.0000000   0.8270011  0.7951822
## log_calcium    -0.4903048      0.8270011   1.0000000  0.7049333
## pH             -0.6051330      0.7951822   0.7049333  1.0000000
```

(b) Find the determinant of the correlation matrix of the predictors.

```
cor(mercury |> select(!merc_ppm)) |> det()

## [1] 0.1139751
```

The determinant of the correlation matrix of the predictors is approximately 0.1139.

(c) Find the variance inflation factor of each predictor.

```
mercury_mod <- lm(merc_ppm ~ ., data = mercury)
vif(mercury_mod)

## log_alkalinity    log_calcium             pH
##       4.413849       3.226014       2.773142
```

	VIF
log_alkalinity	4.413849
log_calcium	3.226014
pH	2.773142

(d) Is there any indication of multicollinearity among these predictors? What types of problems could this cause in the model?
Although none of the variance inflation factors are above the typical rule of thumb threshold which is set to 5, we have reason to believe there exists multicollinearity in these data, as there exists high correlation between many of the variables, as well as a low determinant of the correlation matrix **R**. The pairs plot also demonstrates there exists linear associations between many of the predictor variables, which indicates multicollinearity.

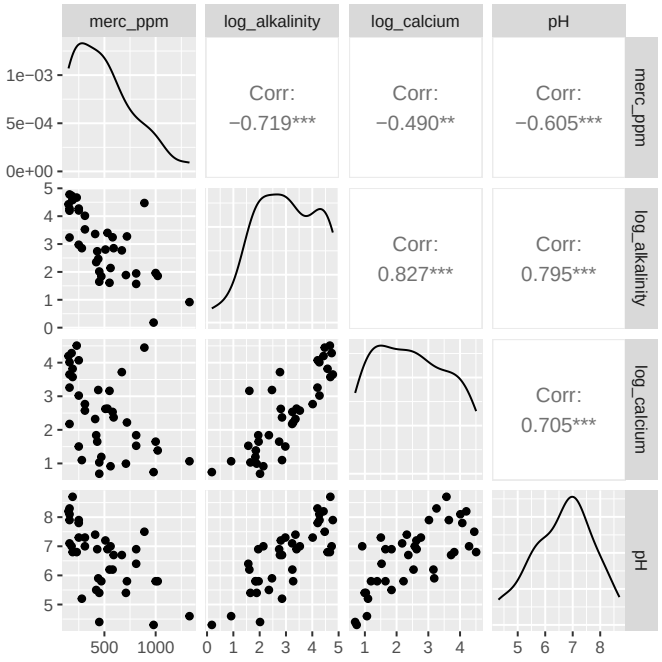


Figure 1: Pairs plot for mercury data.

(e) Assuming that there is no problem with multicollinearity, obtain 95% confidence intervals for each predictor's regression coefficient.

```
# 95% CI for each coefficient
confint(mercury_mod)

##                      2.5 %     97.5 %
## (Intercept)     684.90600 1677.48841
## log_alkalinity -340.81708 -101.54504
## log_calcium     -16.61096  192.41409
## pH             -140.45741   67.20644
```

	2.5 %	97.5 %
log_alkalinity	-340.81708	-101.54504
log_calcium	-16.61096	192.41409
pH	-140.45741	67.20644

(f) Assuming that there is no problem with multicollinearity, obtain the 95% prediction interval for a new observation whose alkalinity is 40, calcium is 20, and pH is 7.

```
# prediction interval
new_data <- data.frame(log_alkalinity = log(40),
                       log_calcium = log(20),
                       pH = 7)

predict(mercury_mod, newdata = new_data,
        interval = "prediction", level = 0.95)
```

```
##        fit       lwr    upr
## 1 372.2381 -50.40382 794.88
```

The 95% prediction interval for a new observation whose alkalinity is 40, calcium is 20, and pH is 7 would be (-50.40, 794.88).

9. The birth weight data given on Canvas is just a portion of data from a much larger study that includes all pregnancies that occurred between 1960 and 1967 at the Kaiser Foundation Health Plan in Oakland, CA. The data here are from one year of the study. It includes all 1236 male single births where the baby lived at least 28 days. The variable descriptions are given in Table **??**.

   (a) Plot birth weight against the gestation for individuals whose mother smoked and individuals whose mothers did not smoke. (You will notice some unusual values for gestation and smoking status. These observations should be removed first. Hint: There are 23 values.)

```
bw <- read.delim("birth_weight.txt",
                        sep ="", header = TRUE, dec =".") |>
  dplyr::filter(smoke <= 1, gestation < 999) |>
  mutate("smoke" = ifelse(smoke == 1, "Smoker", "Nonsmoker"),
         "parity" = as.factor(parity))
```

```
bw |> ggplot(aes(bwt, gestation)) +
  geom_point() +
  labs(title = "Birth Weight Against Gestation",
       x = "Birthweight",
       y = "Gestation") +
  facet_wrap(smoke ~ ., nrow = 2)
```

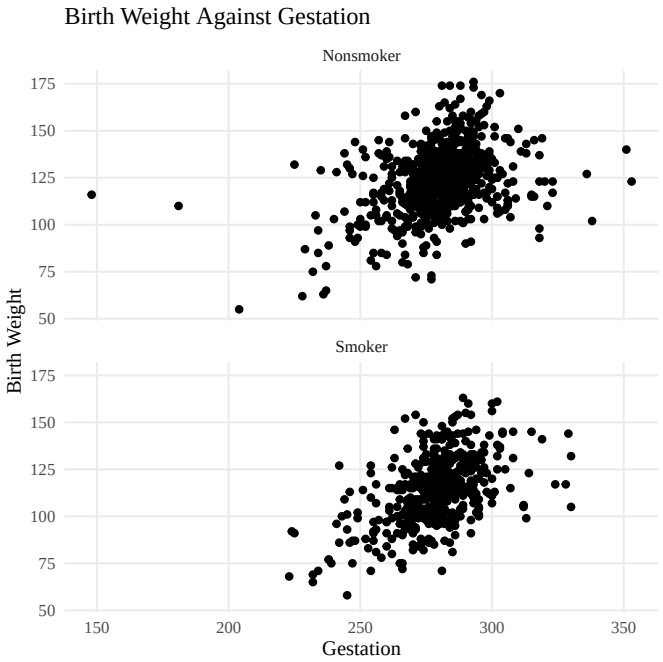


Figure 2: Birthweight against gestation, faceted by mother's smoker status.

(b) Write down a model for the birth weight dependent on the gestation, smoking status of the mother, and the interaction between these two variables.

$$birthweight = \beta_0 + \beta_1 gestation + \beta_2 smoke + \beta_3 gestation * smoke + \epsilon$$

(c) Test to determine if the interaction between gestation and smoking status is significant.

```
# model
bwt_mod <- lm(bwt ~ gestation*smoke, data = bw)
summary(bwt_mod)

##
## Call:
## lm(formula = bwt ~ gestation * smoke, data = bw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.001 -10.958  -0.073   9.927  50.555
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              21.98517   10.04233   2.189 0.028769 *
## gestation                 0.36107    0.03578  10.092  < 2e-16 ***
## smokeSmoker             -73.28346   16.89161  -4.338 1.55e-05 ***
## gestation:smokeSmoker     0.23388    0.06050   3.866 0.000117 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.1 on 1209 degrees of freedom
## Multiple R-squared:  0.2208,Adjusted R-squared:  0.2189
## F-statistic: 114.2 on 3 and 1209 DF,  p-value: < 2.2e-16
```

According to the model summary, we would reject the null hypothesis that the coefficient for the interaction term is zero for all reasonable α values (p-value . Hence, we could conclude that the coefficient for the interaction between gestation and smoker status is significant.

(d) Assuming the interaction remains, what are the estimated regression equations for smoking versus non-smoking mothers?

Model	Equation
Non-Smoker	$birthweight = 21.985 + 0.361 gestation$
Smoker	$birthweight = -51.298 + 0.594 gestation$

Table 2: Fitted regression models for different smoker status.

(e) Overlay the estimated regression lines on top of the observed data. Describe how birth weight changes for smoking versus non-smoking mothers.

```
bw |> ggplot(aes(gestation, bwt)) +
  geom_point(alpha = 0.4) +
  geom_function(fun = \(x) 21.985 + 0.361*x, color = "blue", lwd = 1) +
  geom_function(fun = \(x) -51.298 + 0.594*x, color = "red", lwd = 1) +
  labs(title = "Predicted Birth Weight Against Gestation",
       x = "Birthweight",
       y = "Gestation")
```

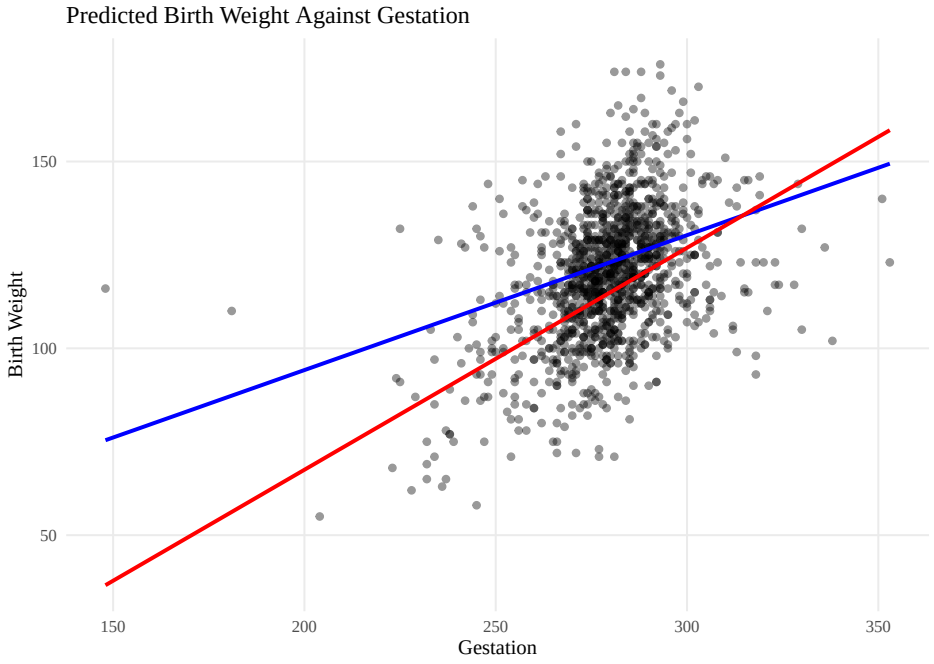


Figure 3: The fitted model of gestation on birthweight for smoking mothers (Red). The fitted model of gestation on birthweight for nonsmoking mothers (Blue).

Generally speaking, the expected birth weight for smoking mothers is lower across most viable values of gestation. For higher values of gestation, we see a higher birth weight for nonsmoking mothers.

10. Chapter 6, Problem 1
    We need to show that $\text{Var}(\hat{\mathbf{Y}}|\mathbf{X}) = \sigma^2\mathbf{H}$. So then we have that

$$\begin{aligned}
\text{Var}(\hat{\mathbf{Y}}|\mathbf{X}) &= \text{Var}(\mathbf{HY}|\mathbf{X}) \\
&= \mathbf{H}\,\text{Var}(\mathbf{Y}|\mathbf{X})\mathbf{H}' \\
&= \mathbf{H}\sigma^2\mathbf{I}\mathbf{H} \quad (\text{Because } \mathbf{H} = \mathbf{H}') \\
&= \sigma^2\mathbf{H} \quad (\text{Because } \mathbf{H} \text{ idempodent}).
\end{aligned}$$

11. Chapter 6, Problem 5

(a) A statistician from Australia has recommended to the analyst that they not transform any of the predictor variables but that they transform Y using the log transformation. Do you agree with this recommendation? Give reasons to support your answer.

```
pgatour <- read.csv("pgatour2006.csv", header = TRUE) |>
  select(PrizeMoney, DrivingAccuracy, GIR, PuttingAverage,
         BirdieConversion, SandSaves, Scrambling, PuttsPerRound)
```

```
GGally::ggpairs(pgatour |> select(!Name))
```

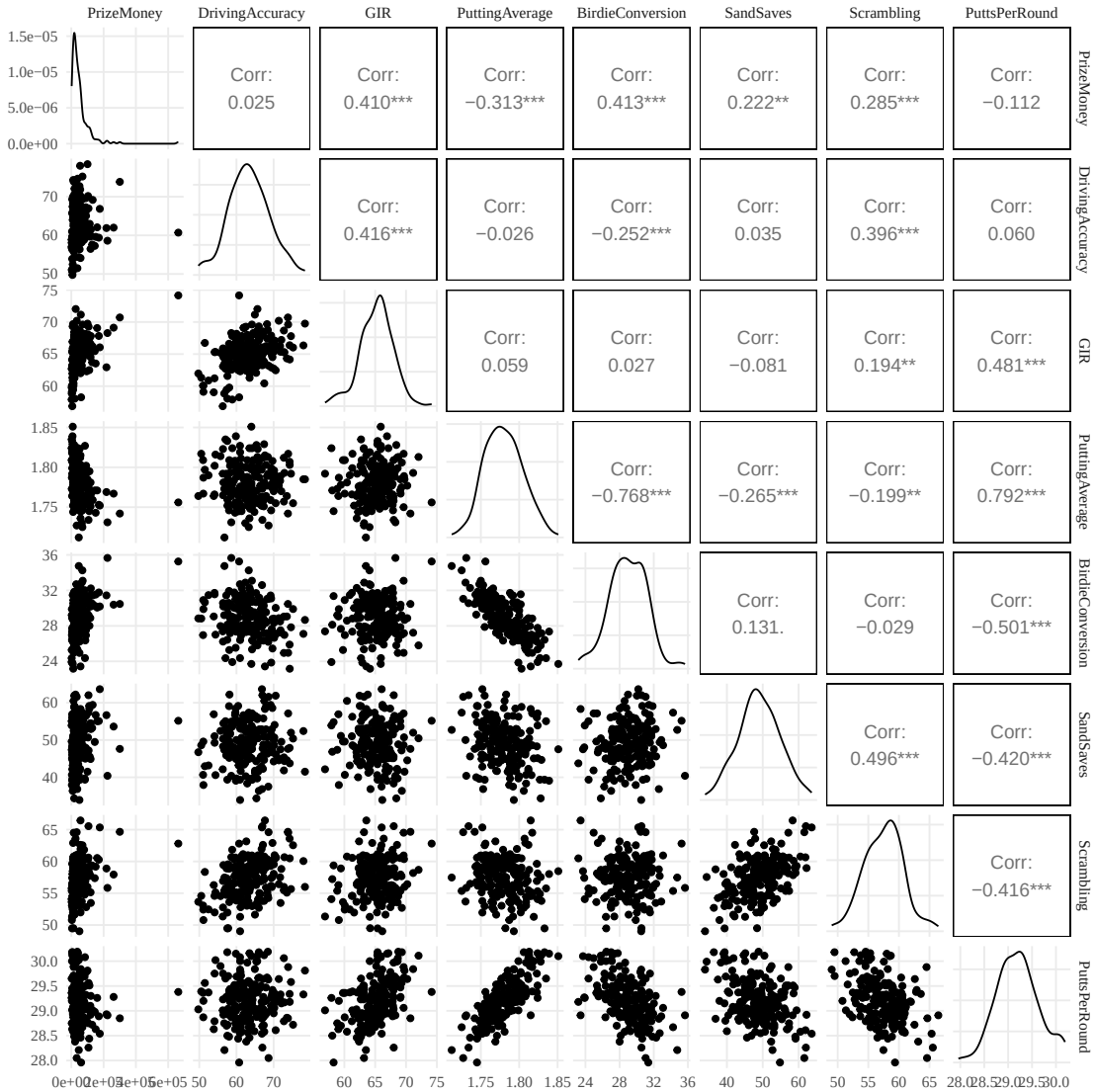


Figure 4: Pairs plot for PGA tour data.

The distribution of the outcome variable, Y, looks extremely right-skewed. Transforming the outcome to bring in the right tail of the distribution might alleviate the problems that come with the skewness. The first column in the pairs plot shows the

side effects of the this skewness, as it is hard to determine the relationship between any of the variables and the outcome of interest. We will log-transform prize money to reduce the skew of the data.

```
pgatour <- pgatour |>
  mutate("logPrizeMoney" = log(PrizeMoney)) |>
  select(!c(PrizeMoney))
```

```
GGally::ggpairs(pgatour |> select(!Name))
```

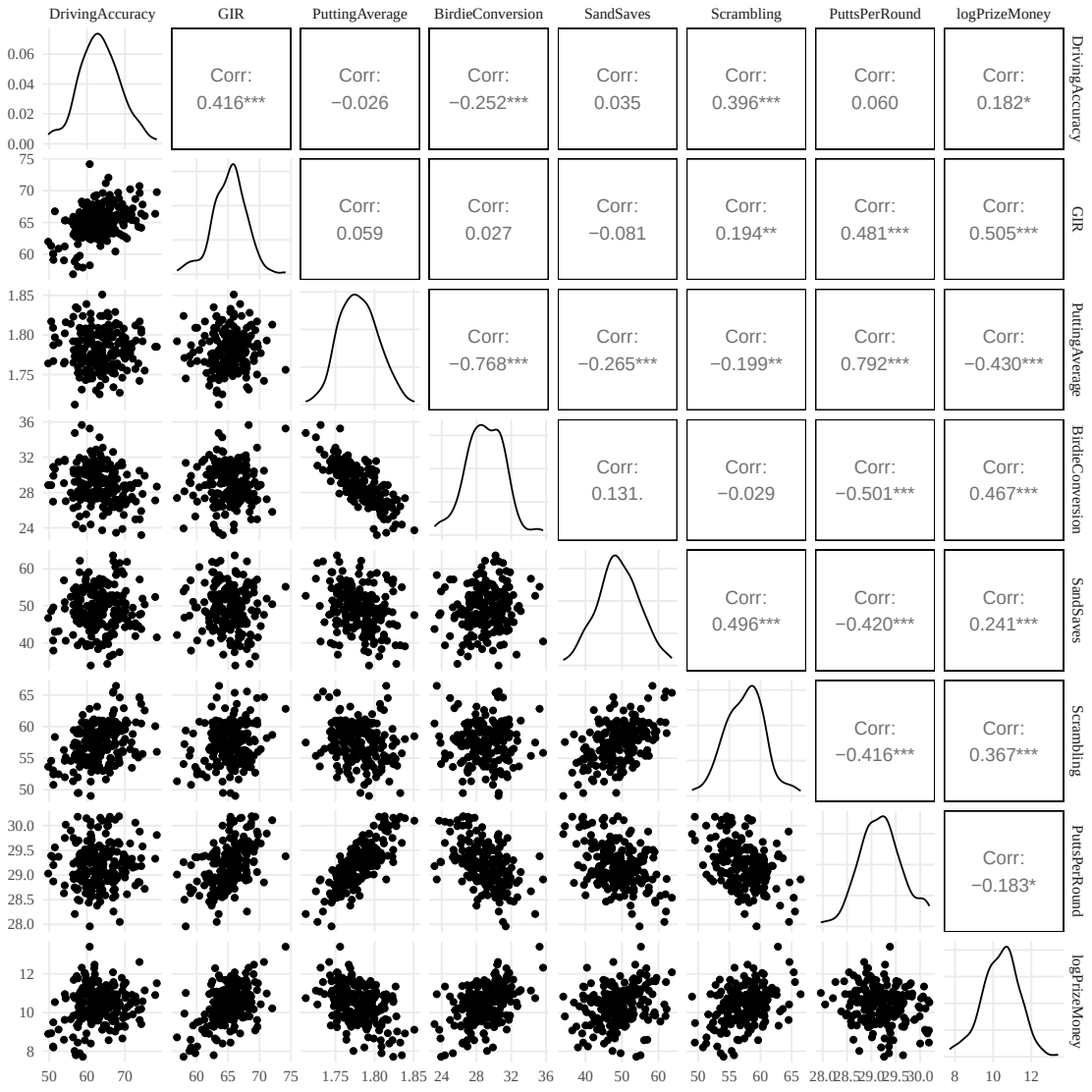


Figure 5: Pairs plot for PGA tour data with log-transformed outcome, prize money.

After log-transforming the outcome, the distribution of the prize money looks far less skewed to the right. Additionally, it is easier to see if there exists any linear relationships between the log-transformed outcome and any of its predictors as seen in the bottom row of the updated pairs plot.

(b) Develop a valid full regression model containing all seven potential predictor variables listed above. Ensure that you provide justification for your choice of full model, which includes scatter plots of the data, plots of standardized residuals, and any other relevant diagnostic plots.

```
pga_mod <- lm(logPrizeMoney ~ ., data = pgatour)
ri  <- rstandard(pga_mod)
fit <- pga_mod$fitted
```

```
par(mfrow = c(3,3))
for (i in 1:7) {
  plot(pgatour[, i], ri, pch = 19,
       col = scales::alpha("black", 0.4),
       xlab = paste(names(pgatour)[i]),
       ylab = "Standardized Residuals",
       ylim = c(-3.2, 3.2),
       main = paste("Residual Plot of ", names(pgatour)[i]),
       cex.lab = 1,
       cex.main = 1)
  abline(h = 0, col = "red", lty = 2)
}
plot(pgatour[, 8], ri, pch = 19,
     col = scales::alpha("black", 0.4),
     xlab = paste(names(pgatour)[i]),
     ylab = "Standardized Residuals",
     ylim = c(-3.2, 3.2),
     main = "Residual Plot of Fitted Values",
     cex.lab = 1,
     cex.main = 1)
abline(h = 0, col = "red", lty = 2)
qqnorm(ri, pch = 19, col = alpha("black", 0.2))
qqline(ri, col = "red")
```

```
coef(pga_mod)
```

```
##       (Intercept)  DrivingAccuracy              GIR   PuttingAverage
##        0.19430026      -0.00352986       0.19931087      -0.46630383
## BirdieConversion        SandSaves       Scrambling    PuttsPerRound
##        0.15734090       0.01517438       0.05151367      -0.34313140
```

```
summary(ri)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -2.638881 -0.744155 -0.140379 -0.001056  0.680080  3.309034
```

All standardized residual plots for the covariates appear to indicate there is no non-linear relationship being unaccounted for in the data. The only odd standardized

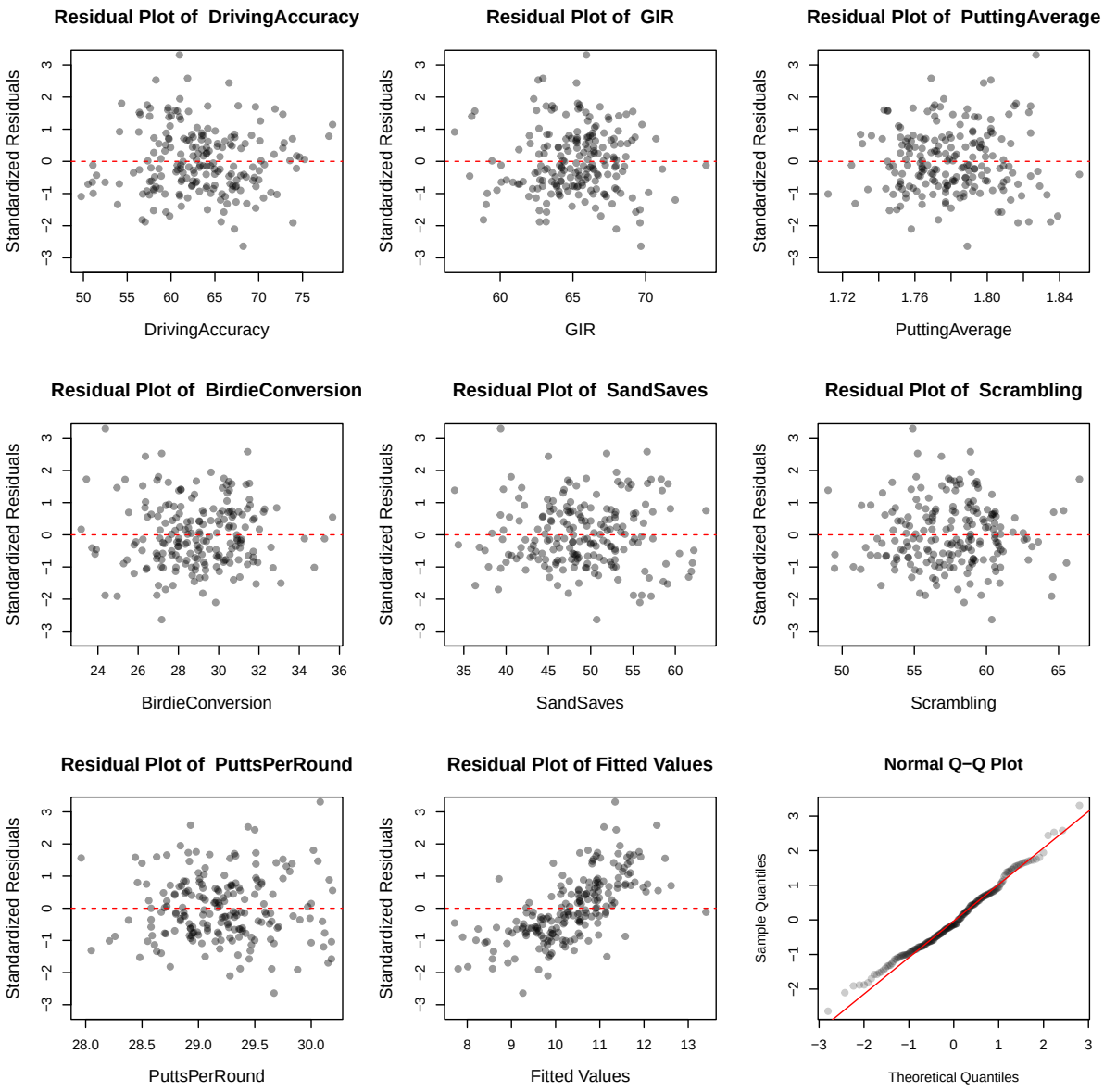


Figure 6: Plots to observe and confirm model assumptions, including standardized residual plots, and a QQ-plot to assess the normality of the standardized residuals.

residual plot that appears odd is the fitted values plot. This could indicate there exists some violation of an assumption of independence amongst predictors, as there did not appear to be noticeable heteroscedasticity in the pairs plots. The QQ-plot for the standardized residuals indicate there exists no deviation from a normal distribution for the residuals. In Figure 7, the covariates DrivingAccuracy, PuttingAverage, and PuttsPerRound appear to have no affect on the log prize money when all other variables are present. The added variable plots indicate that a linear model is appropriate, but that some of the covariates are not necessary.

(c) Identify any points that should be investigated. Give one or more reasons to support each point chosen.

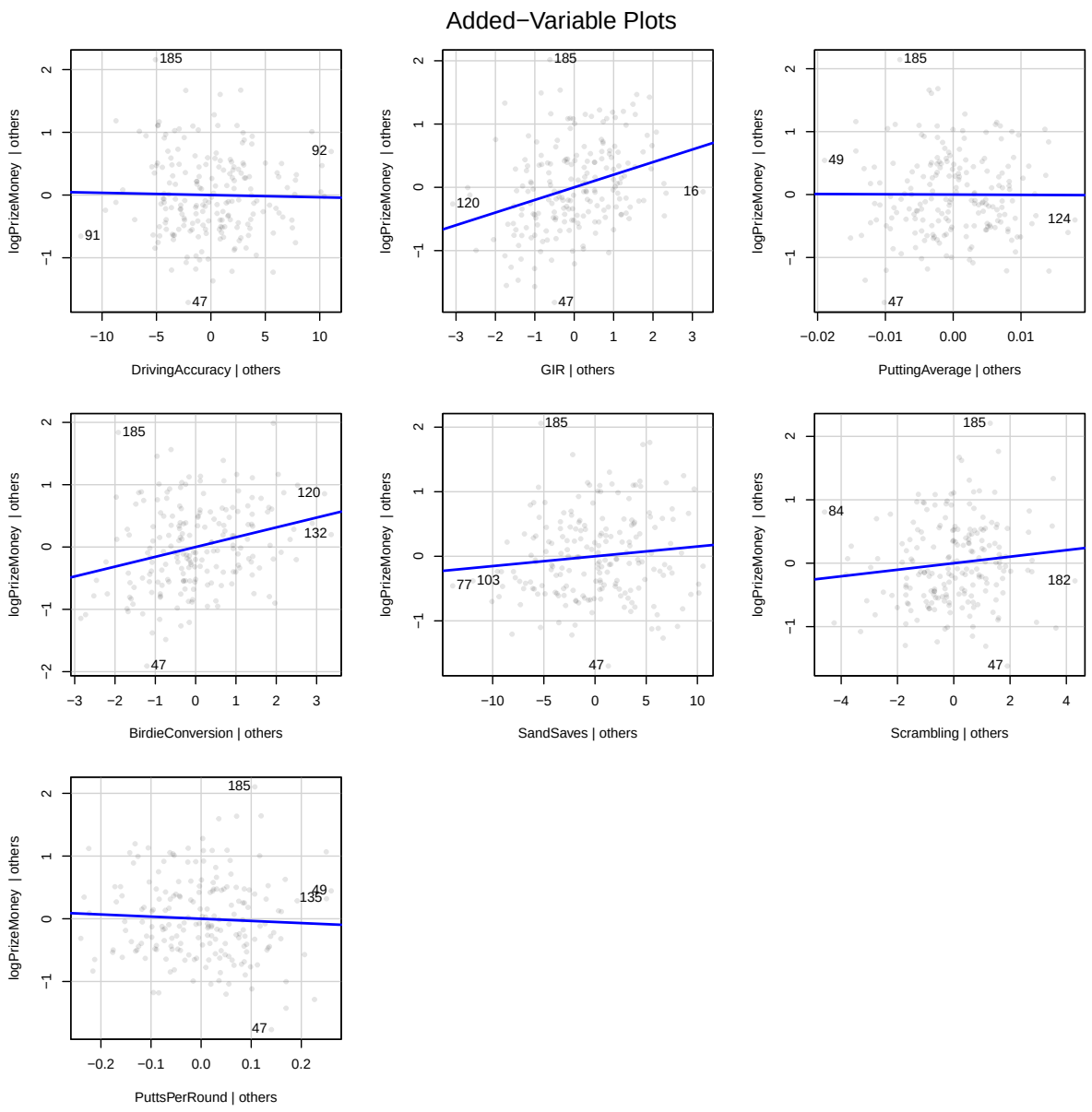


Figure 7: Added variable plots for the log-outcome model.

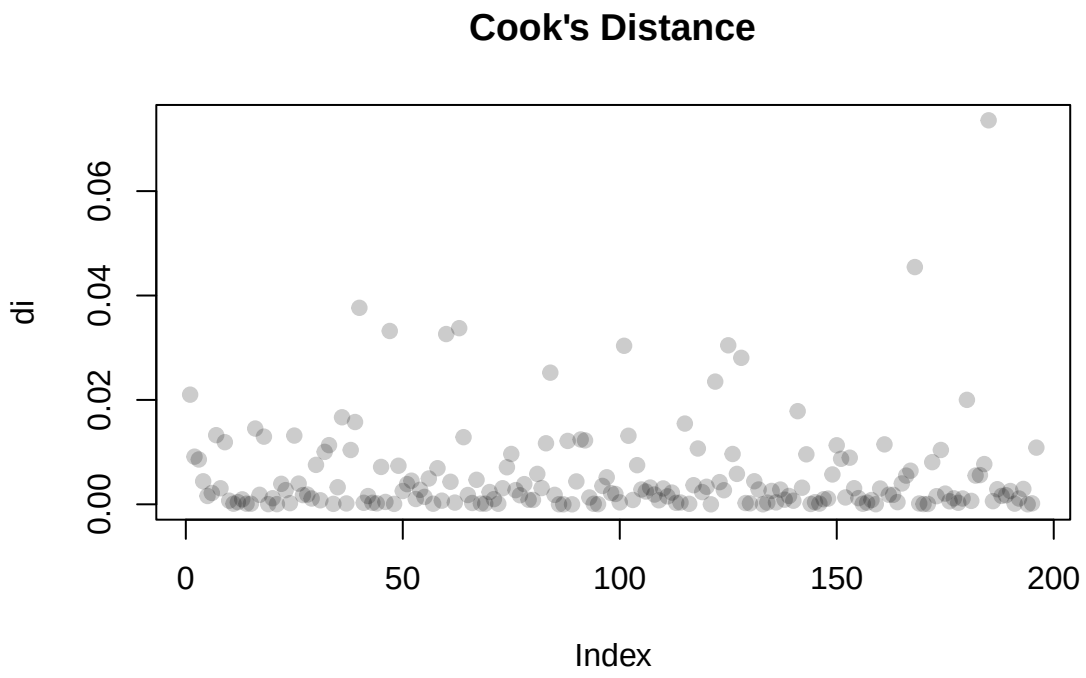


Figure 8: Cook's distance for each observation.

```
ri[which(abs(ri) >= 3)]

##      185
## 3.309034
```

The large residual is for observation 185. This observation also appears to be influential (large Cook's distance).

(d) Describe any weaknesses in your model.

```
vif(pga_mod)

##  DrivingAccuracy              GIR   PuttingAverage BirdieConversion
##         1.796616         6.294969        12.900789         3.511898
##        SandSaves       Scrambling    PuttsPerRound
##         1.461506         4.470203        19.355667
```

This model features what we would consider unnessecary predictors, though we were told to specify the model with all seven predictors. Adding covariates that do not necessarily need to be in the model can inflate the R^2 of the model, as well as the standard errors for the coefficient estimates. We also believe there exists multicollinearity within the model, as some of the VIF's of the covariates are high.

(e) The golf fan wants to remove all predictors with insignificant t-values from the full model in a single step. Explain why you would not recommend this approach.
The p-values obtained when a model is fitted are obtained conditional on every other variable being present in the model. If a golf fan wants to remove all predictors that

are insignificant, they could end up accidentally be omitting a necessary variable that does not end up being significant when the model is fitted with all seven variables.

(f) Find the partial R^2 and added variable plots for the predictors in this model. To see the AV plots, see part (c).

```
rsq::rsq.partial(pga_mod) |> as.data.frame() |>
  select(variable, partial.rsq) |>
  knitr::kable()
```

Variable	Partial R^2
DrivingAccuracy	0.0004780
GIR	0.0991461
PuttingAverage	0.0000243
BirdieConversion	0.0747308
SandSaves	0.0124375
Scrambling	0.0137768
PuttsPerRound	0.0027850

Table 3: Partial R^2 values for each variable.