STAT 5000 LAB #11
FALL 2024 DUE TUE, DEC 3RD NAME: _____

Directions: Complete the exercises below. When you are finished, turn in any required files online in Canvas, then check-in with the Lab TA for dismissal.

Multiple Linear Regression in R

Body fat is an important health measure, but accurately measuring body fat is not easy. The best method requires weighing someone underwater. A quicker, easier method, based on physical measurements, would be desirable. The file `bodyfat.txt` posted in Canvas includes data on body-fat (Y=`fat`) and physical measurements for 133 men, including `age` in years, `weight` in pounds, `height` in inches, and `neck`, `chest`, `abdomen`, `hip`, `thigh`, `knee`, `ankle`, `biceps`, `forearm`, and `wrist` circumference measurements in centimeters.

The R code needed to incorporate categorical explanatory variables, conduct model selection, diagnose multiple linear regression assumptions, and perform remedial measures is described below and the full script is saved in the `bodyfat_Lab11.r` file posted in Canvas.

- First, read in the data using the *Import Dataset* tool in R Studio:

```
library(readr)
bodyfat <- read_table("bodyfat.txt", col_types = cols(density = col_skip()))
```

  NOTE: You can skip the first column, labelled `density` because you won't use it.

- Explore the pairwise scatterplots and correlations of the data to determine whether a multiple linear regression model would be appropriate:

```
pairs(bodyfat)
cor(bodyfat)
```

- Let's start by fitting the full model with all possible variables:

```
full.fat <- lm(fat~., data=bodyfat)
summary(full.fat)
```

- We can assess the effect of the multicollinearity on the parameter estimates for this model by computing the variance inflation factor, or VIF, using the `vif()` function in the `car` package (if this is your first time using the package, you will need to install it first – `install.packages("car")`):

```
library(car)
vif.fat <- vif(full.fat)
barplot(vif.fat)
abline(h=10, col="red", lty=2)
```

  NOTE: The variables with VIF values larger than 10 indicate severe multicollinearity, so be cautious with inference or consider removing those variables from the model.

- Due to the large number of variables (some of which are correlated), we will perform a variety of model selection methods using the AIC (the built-in model selection function in R only allows this criteria to be used, unlike SAS), following the steps below.

- Fit the intercept-only null model:

```
null.fat <- lm(fat~1, data=bodyfat)
summary(null.fat)
```

- Conduct the backward selection process using the `step()` function, providing the full model, and including the `direction="backward"` option:

```
back.fat <- step(full.fat, direction="backward")
summary(back.fat)
```

- Conduct the forward selection process using the `step()` function, providing the intercept-only model, giving the full model as the largest model to consider using the `scope=formula(full.fat)` option, and including the `direction="forward"` option:

```
for.fat <- step(null.fat, scope=formula(full.fat), direction="forward")
summary(for.fat)
```

- To illustrate the mixed stepwise selection process, let's first fit a "really crappy" model (one that doesn't contain any of the terms selected by the previous methods) as the starting point for the `step()` function, giving the full model as the largest model and the intercept-only model as the smallest model to consider using the `scope=list(upper=formula(full.fat), lower=formula(null.fat))` option, and including the `direction="both"` option:

```
bad.fat <- lm(fat~height+knee+biceps+chest, data=bodyfat)
mix.fat <- step(bad.fat, direction="both",
                scope=list(upper=formula(full.fat), lower=formula(null.fat)))
summary(mix.fat)
```

- Conduct the all-possible-subsets selection procedure using the `regsubsets()` function in the `leaps` package (you may need to run `install.packages("leaps")` if this is your first time using this package) with the `method="exhaustive"` option:

```
library(leaps)
all.subsets <- regsubsets(fat~., data=bodyfat, method="exhaustive")
summary(all.subsets)
```

  In this package, you will see the results for the best model of each size (i.e. number of parameters), and then you can access some numeric model selection criteria, including the R^2 (`$rsq`), the adjusted R^2 (`$adjr2`), the residual sums of squares (`$rss`), the Mallow's C_p (`$cp`), and the BIC (`$bic`), to help you select among them. For example, you can access the R^2 value for each model using:

```
summary(all.subsets)$rsq
```

- The "best" model selected by each of the methods includes the quantitative `age` variable (in years). To practice incorporating categorical variables into the MLR in R, let's turn this into a categorical variable with levels `under39` (including men aged 22-38 years), `over52` (including men aged 53-81 years), and `mid` (including men aged 39-52 years). Note that these cutoff values were selected to correspond to the first and third quantile values.

```
ageCat <- vector(mode="character", length=length(bodyfat$age))
ageCat[bodyfat$age<39] = "under39"              # lower quartile
ageCat[bodyfat$age>52] = "over52"               # upper quartile
ageCat[bodyfat$age>38 & bodyfat$age<53] = "mid" # middle 50%
```

```
bodyfat = cbind(bodyfat, ageCat)
```

- You can proceed with conducting the MLR using this categorical version of the age variable:

```
cat.fat <- lm(fat ~ ageCat + weight + neck + abdomen + hip + thigh +
                ankle + forearm + wrist, data = bodyfat)
summary(cat.fat)
```

  NOTE: You should see that replacing the quantitative `age` variable with the categorical `ageCat` variable will reduce the model fit.

- Or you can code this variable using the 0-1 "dummy" variables, let `young` be 1 for the `under39` category and 0 otherwise, and let `older` be 1 for the `over52` category and 0 otherwise, so `mid` is the *baseline* category, and then perform the MLR:

```
young <- rep(0, length=length(bodyfat$age))
young[ageCat=="under39"] = 1
older <- rep(0, length=length(bodyfat$age))
older[ageCat=="over52"] = 1
bodyfat = cbind(bodyfat, young, older)
base.fat <- lm(fat ~ young + older + weight + neck + abdomen + hip + thigh +
                      ankle + forearm + wrist, data = bodyfat)
summary(base.fat)
```

  NOTE: This should result in the exact same model as before (provided R happens to choose the same baseline category as you did).

- Based on the model selection results, we'll stick with the model selected via backward elimination as the "best" model:

```
best.fat <- back.fat
summary(best.fat)
anova(best.fat)
```

- Now, we need to examine the diagnostic plots and measurements to determine whether the MLR model assumptions are met:
  - R will show you some residual plots by default using the `plot()` function on your model:

    ```
    plot(best.fat)
    ```

    But, there are many other diagnostic plots you could explore ...
  - You can plot the residuals against each of the explanatory variables to check for linearity and constant variance. An example is shown below:

    ```
    plot(bodyfat$age, best.fat$residuals, main="MLR for Body Fat Study",
          xlab="Age (years)", ylab="Residuals")
    ```
  - You can plot the residuals against the fitted values to further diagnose the equal variance assumption by looking for no trending and random-scatter around 0:

    ```
    plot(best.fat$fitted.values, best.fat$residuals, main="MLR for Body Fat Study",
          xlab="Fitted Values", ylab="Residuals")
    abline(h=0, col="red")
    ```

NOTE: If you think this assumption may be violated, you should consider using the weighted least squares procedure. You can do this in R using the `weights` option inside the `lm()` function.

– You can get a normal Q-Q plot of the residuals to check the normality assumption:

```
qqnorm(best.fat$residuals)
qqline(best.fat$residuals, col="red")
```

NOTE: If you think this assumption may be violated, you should consider transforming your data, checking for outliers, or using robust regression methods (which can be done using the `rlm()` function).

- Finally, we should check whether there are influential cases included in our dataset, but first we need to load the `MASS` package to access useful functions:

```
library(MASS)
```

– You can find which Studentized residuals exceed ± 2 using the `studres()` function:

```
stdresids <- studres(best.fat)
stdresids[which(abs(stdresids)>2)]
plot(best.fat$fitted.values, stdresids, main="MLR for Body Fat Study",
     xlab="Fitted Values", ylab="Studentized Residuals")
abline(h=0, col="gray")
abline(h=-2, col="red", lty=2)
abline(h=2, col="red", lty=2)
```

NOTE: These values indicate potential outliers (extreme Y values).

– You can find which observations have leverage exceeding $\pm 2(k+1)/n$ where k is the number of explanatory variable and n is the total number of observations using the `hatvalues()` function:

```
leverage <- hatvalues(best.fat)
leverage[which(abs(leverage)>(20/length(leverage)))]
plot(leverage, type = 'h', main="MLR for Body Fat Study",
     ylab="Leverage (hi)")
abline(h=(20/length(leverage)), col="red", lty=2)
```

NOTE: These values indicate potential leverage observations (extreme $\mathbf{x}$ values).

– You can find which observations have influence exceeding $\pm 2\sqrt{2/n}$ using the `cooks.distance()` function:

```
cooks <- cooks.distance(best.fat)
cooks[which(abs(cooks)>(2*sqrt(2/length(cooks))))]
plot(cooks, type = 'h', main="MLR for Body Fat Study",
     ylab="Cook's Distance (Di)")
abline(h=2*sqrt(2/length(leverage)), col="red", lty=2)
```

NOTE: These values indicate potential influence values (extreme $\mathbf{x}$ and Y values).

– You can also find potential influence values by finding the observations that cause a large difference in fitted values if they are excluded from the dataset, i.e. compute the DFFITS values using the `dffits()` function and find the values that exceed $\pm 2\sqrt{(k+1)/n}$:

```
dff <- dffits(best.fat)
dff[which(abs(dff) > 2*sqrt(20/length(dff)))]
plot(abs(dff), type = 'h', main="MLR for Body Fat Study",
     ylab="Absolute Value of DFFITS")
abline(h=2*sqrt(20/length(dff)), col="red", lty=2)
```

– Another method to find potential influence values is by finding observations that cause a large difference in estimated coefficients if they are excluded from the dataset, i.e. compute the DFBETA values using the `dfbetas()` function and find the values that exceed $\pm 2/\sqrt{n}$:

```
dfb <- dfbetas(best.fat)
dfb[which(abs(dfb) > 2/sqrt(length(dfb)))]
plot(dfb[,1], type = 'h', main="MLR for Body Fat Study",
     ylab="DFBETA", xlab="(Intercept)")
abline(h=2/sqrt(length(dfb)), col="red", lty=2)
abline(h=-2/sqrt(length(dfb)), col="red", lty=2)
```

NOTE: You should obtain this type of plot for each one of the parameters/coefficients in your model.

Assignment

Run the code you created in R for the body fat example to complete the following exercises:

1. Summarize your findings from examining the pairwise scatterplots and correlation matrix.
2. Discuss whether the VIFs indicate any explanatory variables exhibiting extreme multicollinearity.
3. Summarize the backward elimination method of model selection by providing:
   (a) an ordered list of which variable was removed from the model at each step;
   (b) a list of which variables remained in the final model;
   (c) a summary of the partial regression coefficients effects tests for the final model.
4. Summarize the forward selection method of model selection by providing:
   (a) an ordered list of which variable was added to the model at each step;
   (b) a list of which variables never entered the final model;
   (c) a summary of the partial regression coefficients effects tests for the final model.
5. Summarize the all-possible-subsets method of model selection by providing:
   (a) Which model would you choose based on the adjusted R^2 values?
   (b) Which model would you choose based on the Mallow's C_p criteria?
   (c) Which model would you choose based on the BIC values?
6. Interpret the *values* of the estimated regression coefficients in the context of the study for:
   (a) the two values corresponding to the categorical age variable;
   (b) one of the values corresponding to the quantitative variable of your choice.
7. Summarize your findings from examining all the residual plots used to diagnose the MLR model assumptions. Are there any assumptions that aren't met for this analysis?
8. Summarize your findings from examining the case diagnostic values/plots. Are there any outliers, leverage points, or influential observations?

Total: 50 points **# correct:** _____ **%:** _____