STAT 530 HW 3

Md Muhtasim Billah

3/4/2020

Question

Use the following data HW_3_2020.csv to answer the following questions. We are interested in modeling Calories consumed using the predictors: Weight, Height, Protein, Carbohydrates, Calcium, gender.

1. Fit the data and find estimates of the partial slopes.
2. Test for the significance of the slopes.
3. What is the sequential effect (extra sums of squares)
4. What is partial sum of squares?
5. Do the relevant diagnostics.
6. Is this a good model?
7. Comment on the following statements saying if it is TRUE or FALSE with reason:
    a. The best measure for model selection is the Adjusted R-square.
    b. Partial sums of squares are more useful than sequential sums of squares.
    c. If we have a categorical variable with 4 categories we will need 4 dummy variables to model this.

Answers

1. Fit the data and find estimates of the partial slopes.

```
#setting working directory
setwd("/Users/muhtasim/Desktop")
#importing data
mydata=read.csv("HW_3_2020.csv", header=T)
#data summary
summary(mydata)
```

```
##     subject        Weight.lbs    Height.Inches      Protein
##  Min.   : 1.00   Min.   : -1.0   Min.   :-1.00   Min.   : 21.39
##  1st Qu.:25.25   1st Qu.:165.0   1st Qu.:68.25   1st Qu.: 51.25
##  Median :49.50   Median :187.5   Median :70.00   Median : 74.32
##  Mean   :49.50   Mean   :185.9   Mean   :69.92   Mean   : 75.06
##  3rd Qu.:73.75   3rd Qu.:210.0   3rd Qu.:72.00   3rd Qu.: 94.65
##  Max.   :98.00   Max.   :280.0   Max.   :87.00   Max.   :142.83
##  Carbohydrates       Calcium           gender          Calories
##  Min.   : 62.17   Min.   : 161.0   Min.   :0.0000   Min.   : 492.4
##  1st Qu.:148.80   1st Qu.: 505.0   1st Qu.:1.0000   1st Qu.:1289.8
##  Median :210.34   Median : 794.3   Median :1.0000   Median :1740.9
```

```
##  Mean   :211.15   Mean   : 861.2   Mean   :0.7959   Mean   :1803.9
##  3rd Qu.:266.08   3rd Qu.:1045.5   3rd Qu.:1.0000   3rd Qu.:2262.9
##  Max.   :468.15   Max.   :2899.9   Max.   :1.0000   Max.   :3506.4
```

```
#fitting the data in a MLR model
fit = lm(Calories ~ Weight.lbs + Height.Inches + Protein + Carbohydrates + Calcium + gender, data = myda
#partial slopes for the predictors
round(fit$coefficients, 3)
```

```
##   (Intercept)    Weight.lbs Height.Inches       Protein Carbohydrates
##       -21.040         0.619        -0.577        13.421         4.550
##       Calcium        gender
##        -0.231       -23.693
```

From the summary of the data, we can see that for both the variables Weight.lbs and Heights.Inches, the Min. value is -1.00 which doesn't have any physical meaning. This unreasonable values might have been caused by an error during the data measurement. We will discard these data points and fit the model on the cleaned data.

```
#importing cleaned data
mydata1=read.csv("HW3_2020_clean.csv", header=T)
#summary of the data
summary(mydata1)
```

```
##     subject        Weight.lbs    Height.Inches      Protein
##  Min.   : 1.00   Min.   :125.0   Min.   :63.00   Min.   : 21.39
##  1st Qu.:25.50   1st Qu.:166.0   1st Qu.:69.00   1st Qu.: 50.49
##  Median :50.00   Median :188.0   Median :70.00   Median : 74.11
##  Mean   :49.79   Mean   :189.7   Mean   :70.66   Mean   : 74.80
##  3rd Qu.:74.50   3rd Qu.:210.0   3rd Qu.:72.00   3rd Qu.: 94.42
##  Max.   :98.00   Max.   :280.0   Max.   :87.00   Max.   :142.83
##  Carbohydrates       Calcium           gender          Calories
##  Min.   : 62.17   Min.   : 161.0   Min.   :0.0000   Min.   : 492.4
##  1st Qu.:148.81   1st Qu.: 497.3   1st Qu.:1.0000   1st Qu.:1279.3
##  Median :209.83   Median : 805.1   Median :1.0000   Median :1734.6
##  Mean   :211.24   Mean   : 862.2   Mean   :0.7895   Mean   :1798.9
##  3rd Qu.:266.19   3rd Qu.:1044.6   3rd Qu.:1.0000   3rd Qu.:2258.7
##  Max.   :468.15   Max.   :2899.9   Max.   :1.0000   Max.   :3506.4
```

```
#doing MLR
fit1 = lm(Calories ~ Weight.lbs + Height.Inches + Protein + Carbohydrates + Calcium + gender, data = my
#estimates of the partial slopes upto three decimal points
round(fit1$coefficients, 3)
```

```
##   (Intercept)    Weight.lbs Height.Inches       Protein Carbohydrates
##      -366.706         1.252         3.014        13.429         4.592
##       Calcium        gender
##        -0.237       -70.033
```

2. Test for the significance of the slopes.

```r
library(vars)
round(coeftest(fit1), 3)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -366.706    449.707  -0.815    0.417
## Weight.lbs        1.252      0.780   1.605    0.112
## Height.Inches     3.014      7.305   0.413    0.681
## Protein          13.429      1.412   9.511   <2e-16 ***
## Carbohydrates     4.592      0.400  11.466   <2e-16 ***
## Calcium          -0.237      0.066  -3.581    0.001 ***
## gender          -70.033     61.910  -1.131    0.261
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For testing the significance of the partial slopes,

The null hypothesis is, $H_0 : \beta_i = 0$.
The alternative hypothesis is, $H_a : \beta_i \neq 0$.
Level of significane, $\alpha = 0.05$.

From the t-test carried out for the partial slopes, we notice that the intecept is found not statistically significant (p-value = 0.417). This means, there is not enough statistical evidence to show that the coefficient differs from zero. The p-values for protein (< 0.001) and carbohydrates (<0.001) are indicating that there is sufficient statistical evidence to conclude that they are having significant effect on calory levels. The positive coefficients are also telling these effects are positive. Calcium has significant (p-value = 0.001) negative effect on calory. That means with the increase of calcium intake there is a significant decrease of calory is found. The only categirial predictor in the model is "gender" which has two categories, Male and Female. However, gender has no significant effect on the calory.

So, to summarize, only the predictors Protein, Carbohydrates and Calcium are statistically significant for an $\alpha = 0.05$. This means that there is enough statistical evidence to show that the coefficient is not zero i.e the response variable (Calories) is dependent on these three predictors at the population level. This also indicates that these three predictors are good addition to the model while the rest might not be. Thus, for a more precise model, dropping the other predictors might be considered.

3. What is the sequential effect (extra sums of squares)?

The extra sums of squares or the type-I sums of squares are calculates using SAS and provided in the Table 1 below.

Variables	Sequential Effects
Intercept	307420173
Weight	2247916
Height	976053
Protein	35015094
Carbohydrates	4455239
Calcium	410255
Gender	45343

Table 1: Sequential effects (Type-I sums of squares)

4. What is partial sum of squares?

The partial sums of squares or the type-II sums of squares are calculates using SAS and provided in the Table 2 below.

Variables	Partial Effect
Intercept	23561
Weight	91240
Height	6031.91
Protein	3205098
Carbohydrates	4658804
Calcium	454421
Gender	45343

Table 2: Partial effects (Type-II sums of squares)

5. Do the relevant diagnostics.

I) Overall significance

```
#summary of the fitted regression model
summary = summary(fit1)
summary
```

```
##
## Call:
## lm(formula = Calories ~ Weight.lbs + Height.Inches + Protein +
##     Carbohydrates + Calcium + gender, data = mydata1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -385.98 -122.97  -18.74  141.27  437.45
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -366.70646  449.70704  -0.815  0.41703
## Weight.lbs       1.25236    0.78046   1.605  0.11215
## Height.Inches    3.01398    7.30509   0.413  0.68091
## Protein         13.42889    1.41199   9.511 3.65e-15 ***
## Carbohydrates    4.59163    0.40044  11.466  < 2e-16 ***
## Calcium         -0.23660    0.06607  -3.581  0.00056 ***
## gender         -70.03331   61.90992  -1.131  0.26104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 188.2 on 88 degrees of freedom
## Multiple R-squared:  0.9326, Adjusted R-squared:  0.928
## F-statistic:   203 on 6 and 88 DF,  p-value: < 2.2e-16
```

From, the p-value of the fitted model, it can be said that it is statistically significant with an $\alpha = 0.05$. It means that the overall model is significant to study the functional relationship between the calory and the predictor variables. Also, from the high R-squared and Adjusted R-squared value, it can be concluded that

a high percentage of variance (~93%) was explained by this model which also indicates the model's goodness of fit.

II) Normality Test

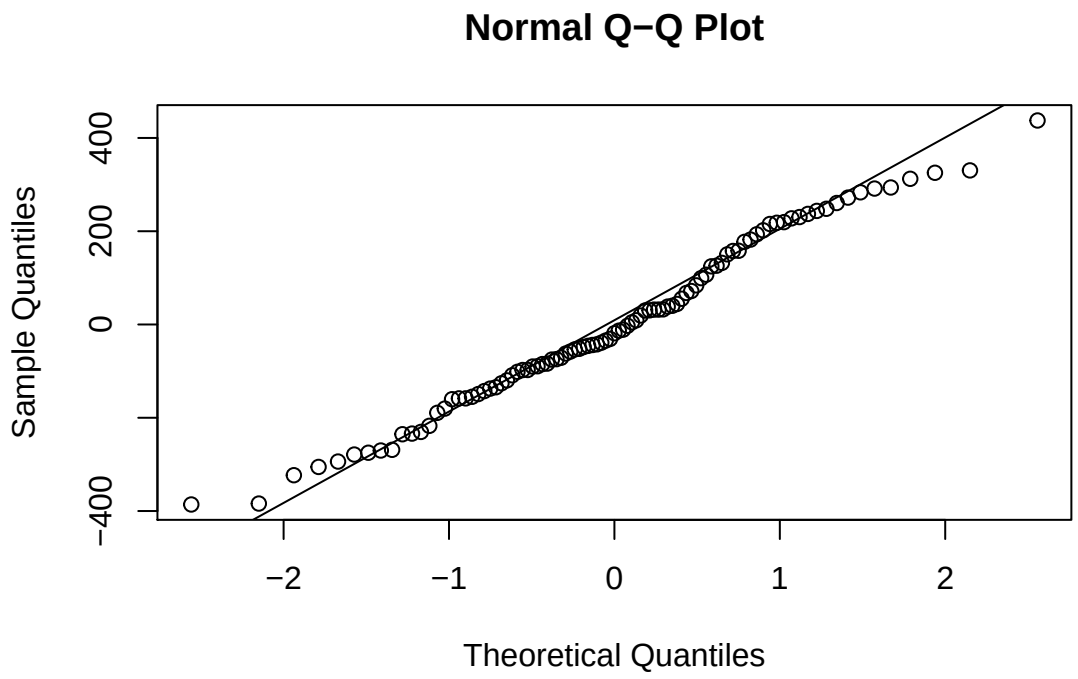
Shapiro-Wilk test: Since the p-value>$\alpha = 0.05$, the null is retained. Which means that the residuals belong to a normal distribution.

```
shapiro.test(fit1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit1$residuals
## W = 0.98469, p-value = 0.3357
```

Normal probability plot or the QQ plot: Though, the residuls deviate a bit from the line at the tails, most of the data are along the line which also indicates too a normal distribution of the residuals.

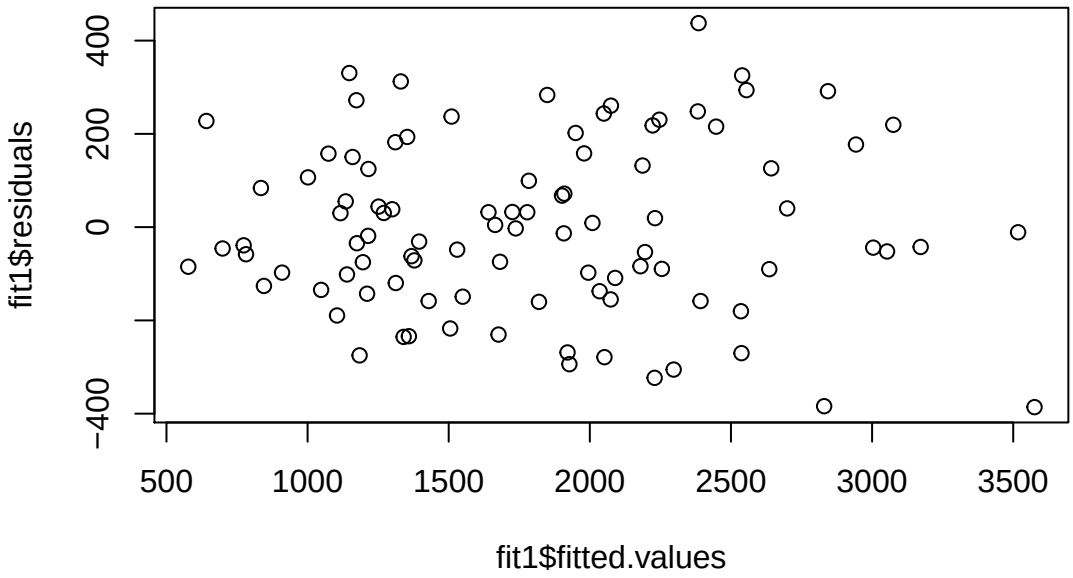
```
qqnorm(fit1$residuals)
qqline(fit1$res)
```



III) Constant Variance Test

Fitted vs Rsiduals plot: From the plot, we see a random scatter of the residuals rather than a noticeable pattern. This indicates to a constant variance of the data.

```
plot(fit1$fitted.values, fit1$residuals)
```



Breusch-Pagan test: From the test, it's evident that p-value$>\alpha = 0.05$ and so the null is retained. This also indicates that the data has constant variance.

```
lmtest::bptest(fit1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit1
## BP = 12.539, df = 6, p-value = 0.05098
```

IV) Autocorrelation Test

Durbin-Watson test: Since the p-value$>\alpha = 0.05$, the null is retained. The D-W value of 1.97 indicates that the fitted model residuals are free of autocorrelation.

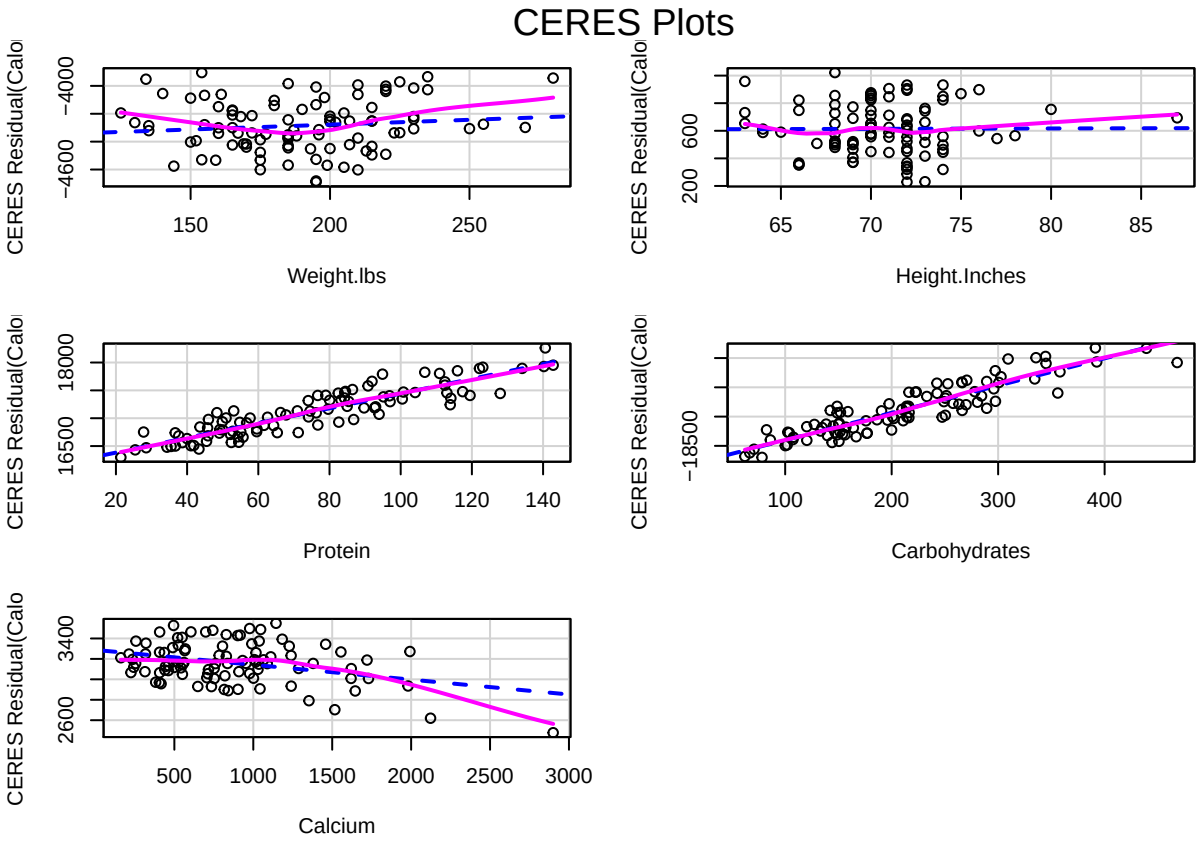
```
dwtest(fit1)
```

```
##
##  Durbin-Watson test
##
## data:  fit1
## DW = 1.9721, p-value = 0.4418
## alternative hypothesis: true autocorrelation is greater than 0
```

V) Linearity Test

Residuals vs the predictor variables' plots. It is evident that for Height, Protein and Carbohydrate, the residuals are very linear while for the rest of the predictors, it is less linear.

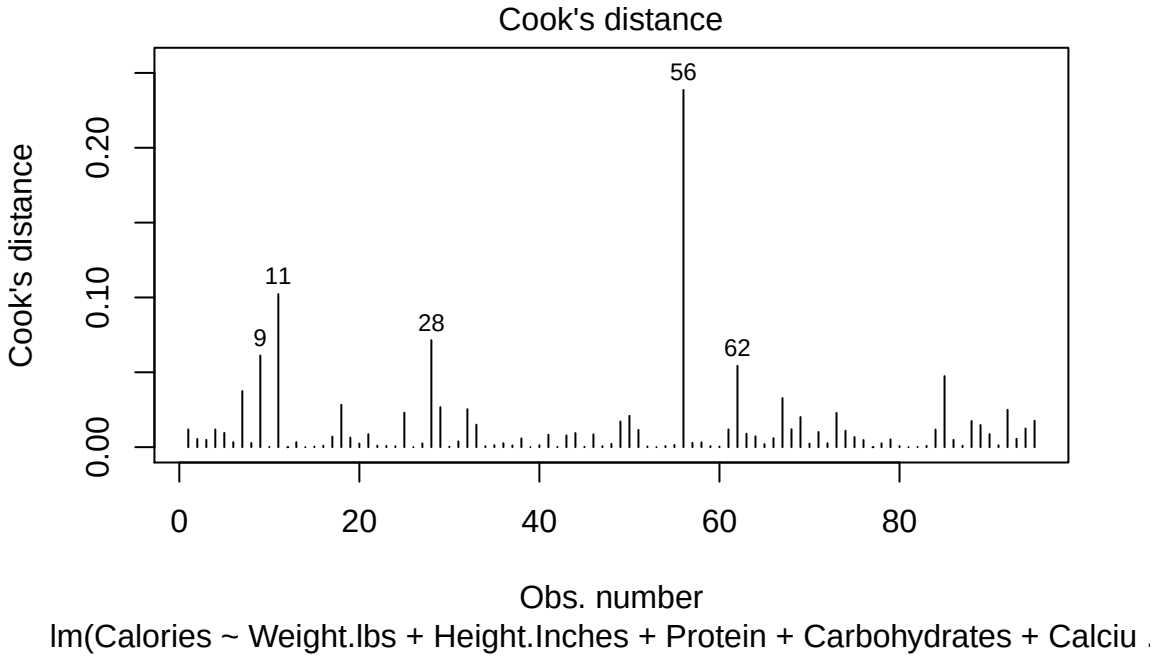
```
library(car)
attach(mydata1)

fit.l = lm(Calories ~ Weight.lbs + Height.Inches + Protein + Carbohydrates + Calcium , data = mydata1)

ceresPlots(fit.l)
```



VI) Potential Outliers

From the Cook's distance plot, we see that there is a huge outlier in the dataset (observation number 56) which can be an influencial point. For building a better model, observations number 11, 28, 62 can also be considered.

```
#cook's distance
plot(fit1, 4, id.n = 5)
```



6. Is this a good model?

From the overall diagnostics of the model, the R-squared values, the tests for normality, constant variance and autocorrelation it can be said that it is a good model.

7. Comment on the following statements saying if it is TRUE or FALSE with reason:

a. The best measure for model selection is the Adjusted R-square. **TRUE**

Reason: The issue with R^2 is that if a new predictor is added to the model, its value goes up whether it's adding any significant effect to the model or not. However, adjusted R^2 is a better measure which can remedy this issue. It takes into account the number of predictors in the model and penalizes for too many predictors. So, better information about the model is previded by adjusted R^2.

b. Partial sums of squares are more useful than sequential sums of squares. **TRUE**

Reason: Paritial sum of squares are more appropriate when there is no interaction among the predictors of the model, irrespective of the order in which they enter the model. So, as opposed to the sequential (extra) sums of square where the order is an issue, the partial sum of squares is a better measure.

c. If we have a categorical variable with 4 categories we will need 4 dummy variables to model this. **FALSE**

Reason: In general, to model a categorical variables with k levels, we require (k-1) dummy variables. Thus, if we have a categorical variable with 4 categories we will need 3 dummy variables to model this.