GR 8 Alnahdi, Darji, Varshney, Jin

38 ½ / 40

SCM 651 Fall 2018 Group Assignment 2

Due Date: Midnight, Tuesday, 10/16/2018, Total Points = 40

Please use the programming language R to complete this assignment. Copy and paste relevant parts of the R output and/or screen shots into a Word file to prepare the answers.

1.(20 points) Please use the data set 651F18 Orange Juice Homework 2.csv to do all parts of question 1.

This data set provides, for a random sample of 5780 cases drawn from the Dominicks data base, made available by University of Chicago, Kilts Center:

- MOVE: Number of units sold for three brands of orange juice: Florida's Natural Home-squeezed (FLNAT), Tree Fresh (TF), and Tropicana Grove Stand (TROPICANA), at a store in a given week.
- PRICE: Unit price of the brand.
- logMOVE: Natural logarithm of MOVE.
- logPRICE: Natural logarithm of price.
- BRAND
- Season
- Feat (1 if product is on sale, 0 if not)
- Demographic variables at the store location: AGE9, AGE60, EDUC, ETHNIC, INCOME, HSIZEAV, HH3PLUS, HH4PLUS, HHSINGLE, HHLARGE, HVAL150, HVAL200, MORTGAGE, NOCAR, NWHITE, SINGLE, POVERTY, RETIRED, SINGLE, UNEMP, WORKWOM

1(a)(4+3+3 = 10 points) Fit a regression model with dependent variable logMOVE and the following independent variables:

- BRAND
- logPRICE
- Interaction between BRAND and logprice
- Feat
- Season
- Demographic variables given in the data set.

1(a)(i) From the estimated parameters, what are the price elasticities of demand of the three brands?

```
Call:
lm(formula = logMOVE ~ BRAND + logPRICE + BRAND * logPRICE +
    Feat + Season + AGE9 + AGE60 + EDUC + ETHNIC + INCOME + HSIZEAVG +
    HH3PLUS + HH4PLUS + HHSINGLE + HHLARGE + HVAL150 + HVAL200 +
    MORTGAGE + NOCAR + NWHITE + SINGLE + POVERTY + RETIRED +
    UNEMP + WORKWOM, data = oj)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8048 -0.4487  0.0203  0.4549  3.3468

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -10.83734    4.36791  -2.481 0.013125 *
BRAND[T.TF]                  -0.54003    0.13661  -3.953 7.81e-05 ***
BRAND[T.TROPICANA]            1.06738    0.16127   6.619 3.95e-11 ***
logPRICE                     -3.31164    0.13166 -25.153  < 2e-16 ***
Feat                          0.23030    0.02409   9.559  < 2e-16 ***
Season[T.Spring]             -0.02445    0.02885  -0.847 0.396757
Season[T.Summer]             -0.02978    0.02849  -1.045 0.295895

Season[T.Winter]              0.04221    0.02909   1.451 0.146865
AGE9                         10.67244    2.23239   4.781 1.79e-06 ***
AGE60                         2.39892    1.45809   1.645 0.099972 .
EDUC                         -0.40258    0.34949  -1.152 0.249411
ETHNIC                       -1.50328    0.39754  -3.782 0.000157 ***
INCOME                        0.06580    0.19917   0.330 0.741123
HSIZEAVG                      8.52585    2.13839   3.987 6.77e-05 ***
HH3PLUS                      -4.68900    3.52432  -1.330 0.183417
HH4PLUS                     -11.28363    3.63435  -3.105 0.001914 **
HHSINGLE                      3.96397    2.69468   1.471 0.141337
HHLARGE                     -32.06710    6.78987  -4.723 2.38e-06 ***
HVAL150                       0.94689    0.17724   5.343 9.52e-08 ***
HVAL200                      -0.08945    0.22792  -0.392 0.694721
MORTGAGE                      0.15890    0.20936   0.759 0.447917
NOCAR                         1.70522    0.45465   3.751 0.000178 ***
NWHITE                        0.71943    0.39578   1.818 0.069153 .
SINGLE                        6.52714    0.91567   7.128 1.14e-12 ***
POVERTY                      -1.87986    1.63822  -1.148 0.251221
RETIRED                      -1.41060    1.82679  -0.772 0.440043
UNEMP                        -3.95899    2.67228  -1.482 0.138527
WORKWOM                      -4.82137    1.44372  -3.340 0.000845 ***
BRAND[T.TF]:logPRICE          0.40820    0.14540   2.807 0.005010 **
BRAND[T.TROPICANA]:logPRICE  -0.33128    0.16793  -1.973 0.048577 *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.768 on 5750 degrees of freedom
Multiple R-squared:  0.5283, Adjusted R-squared:  0.526
F-statistic: 222.1 on 29 and 5750 DF,  p-value: < 2.2e-16
```

Accordingly, the price elasticity for each brand is counted as follows:

FLNAT: **-3.31164**

TROPICANA: -3.31164+1.06738= **-2.24426**

TF: -3.31164-0.54003= **-3.85167**

You added coeffs of TF & TROPICANA.

1(a)(ii) Starting from the full model estimated, test the following hypothesis at a 99% level of confidence:

The price elasticity of demand is equal for the three brands.

Need to add coeffs of LogPrice * TF & LogPrice * TROPICANA

-1/2

```
> local({
+   .Hypothesis <- matrix(c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1), 2, 30, byrow=TRUE)
+   .RHS <- c(0,0)
+   linearHypothesis(LinearModel.3, .Hypothesis, rhs=.RHS)
+ })
Linear hypothesis test

Hypothesis:
BRAND[T.TF]:logPRICE = 0
BRAND[T.TROPICANA]:logPRICE = 0

Model 1: restricted model
Model 2: logMOVE ~ BRAND + logPRICE + BRAND * logPRICE + Feat + Season +
    AGE9 + AGE60 + EDUC + ETHNIC + INCOME + HSIZEAVG + HH3PLUS +
    HH4PLUS + HHSINGLE + HHLARGE + HVAL150 + HVAL200 + MORTGAGE +
    NOCAR + NWHITE + SINGLE + POVERTY + RETIRED + UNEMP + WORKWOM

  Res.Df    RSS Df Sum of Sq      F       Pr(>F)
1   5752 3410.3
2   5750 3391.7  2    18.612 15.777 0.0000001469 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the P value is **much smaller than 0.01**, so REJECT the null hypothesis, thus the price elasticity is not same for all three brands at 99% level of confidence.

1(a)(iii) Starting from the full model estimated, test the following hypothesis at a 99% level of confidence:

The price elasticity of demand is equal for Tree Fresh and Tropicana.

```
> local({
+   .Hypothesis <- matrix(c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,-1), 1, 30, byrow=TRUE)
+   .RHS <- c(0)
+   linearHypothesis(LinearModel.4, .Hypothesis, rhs=.RHS)
+ })
Linear hypothesis test

Hypothesis:
BRAND[T.TF]:logPRICE - BRAND[T.TROPICANA]:logPRICE = 0

Model 1: restricted model
Model 2: logMOVE ~ BRAND + logPRICE + BRAND * logPRICE + Feat + Season +
    AGE9 + AGE60 + EDUC + ETHNIC + INCOME + HSIZEAVG + HH3PLUS +
    HH4PLUS + HHSINGLE + HHLARGE + HVAL150 + HVAL200 + MORTGAGE +
    NOCAR + NWHITE + SINGLE + POVERTY + RETIRED + UNEMP + WORKWOM

  Res.Df    RSS Df Sum of Sq      F         Pr(>F)
1   5751 3409.3
2   5750 3391.7  1    17.616 29.865 0.00000004828 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the P value is **much smaller than 0.01**, so REJECT the null hypothesis, thus the price elasticity is not same for TF and TROPICANA at 99% level of confidence.

1(b)(5 points) From the output from the full model in 1(a)(i), identify the demographic variables that are not significant at a 90% level of confidence. At a 99% level of confidence, test the null hypothesis that none of these variables is significant.

At 90% level of confidence, the following demographic variables that are not significant: EDUC, INCOME, HH3PLUS, HHSINGLE, HVAL200, MORTGAGE, POVERTY, RETIRED, UNEMP

```
Linear hypothesis test

Hypothesis:
EDUC = 0
INCOME = 0
HH3PLUS = 0
HHSINGLE = 0
HVAL200 = 0
MORTGAGE = 0
POVERTY = 0
RETIRED = 0
UNEMP = 0

Model 1: restricted model
Model 2: logMOVE ~ BRAND + logPRICE + BRAND * logPRICE + Feat + Season +
    AGE9 + AGE60 + EDUC + ETHNIC + INCOME + HSIZEAVG + HH3PLUS +
    HH4PLUS + HHSINGLE + HHLARGE + HVAL150 + HVAL200 + MORTGAGE +
    NOCAR + NWHITE + SINGLE + POVERTY + RETIRED + UNEMP + WORKWOM

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   5759 3398.5
2   5750 3391.7  9    6.8022 1.2813 0.2413
```

Since the P value is **much greater than 0.01**, so CANNOT REJECT the null hypothesis. We cannot conclude none of the variables is significant.

1(c)(5 points) Based on the result of the hypothesis test in question 1(b), keep the full model or the restricted model as appropriate as the next model. Based on variance inflation factor (VIF), is there any problem with multi-collinearity with this model? If yes, then modify the model to reduce the effect of multi-collinearity. (To do this, identify groups of demographic variables that are strongly correlated with one another. Then, keep one variable from each group.) Fit the model and check if the problem of multi-collinearity is reduced.

VIF for the full model shows below:

	GVIF	Df	GVIF^(1/(2*Df))
BRAND	908.396783	2	5.489956
logPRICE	8.481980	1	2.912375
Feat	1.294578	1	1.137798
Season	1.083472	3	1.013452
AGE9	29.103878	1	5.394801
AGE60	79.711189	1	8.928112
EDUC	15.116887	1	3.888044
ETHNIC	55.784658	1	7.468913
INCOME	32.555378	1	5.705732
HSIZEAVG	2849.879032	1	53.384258
HH3PLUS	774.413378	1	27.828284
HH4PLUS	470.978588	1	21.702041
HHSINGLE	455.442956	1	21.341110
HHLARGE	394.914248	1	19.872449
HVAL150	18.362552	1	4.285155
HVAL200	17.246656	1	4.152909
MORTGAGE	8.813916	1	2.968824
NOCAR	33.896115	1	5.822037
NWHITE	55.788003	1	7.469190
SINGLE	35.205826	1	5.933450
POVERTY	52.275853	1	7.230204
RETIRED	82.343848	1	9.074351
UNEMP	36.907010	1	6.075114
WORKWOM	56.334651	1	7.505641
BRAND:logPRICE	678.412724	2	5.103563

Since we have multiple variables P value >0.05 and strongly correlated, we reduce the full model to model B.

```
lm(formula = logMOVE ~ BRAND + logPRICE + Feat + AGE9 + ETHNIC +
    HSIZEAVG + HH4PLUS + HHLARGE + HVAL150 + NOCAR + SINGLE +
    WORKWOM, data = oj)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0138 -0.4480  0.0208  0.4617  3.3884

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -3.60416    0.95798  -3.762 0.000170 ***
BRANDTF          -0.20027    0.03001  -6.673 2.74e-11 ***
BRANDTROPICANA    0.77765    0.02523  30.818  < 2e-16 ***
logPRICE         -3.11603    0.06038 -51.608  < 2e-16 ***
Feat              0.25643    0.02361  10.863  < 2e-16 ***
AGE9              3.70177    1.06801   3.466 0.000532 ***
ETHNIC           -0.88561    0.12676  -6.986 3.14e-12 ***
HSIZEAVG          5.46294    0.50107  10.903  < 2e-16 ***
HH4PLUS         -15.46418    1.67914  -9.210  < 2e-16 ***
HHLARGE         -16.81392    1.89758  -8.861  < 2e-16 ***
HVAL150           0.91361    0.06010  15.202  < 2e-16 ***
NOCAR             1.80118    0.24657   7.305 3.15e-13 ***
SINGLE            3.63382    0.55522   6.545 6.47e-11 ***
WORKWOM          -2.37155    0.35095  -6.758 1.54e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7719 on 5766 degrees of freedom
Multiple R-squared:  0.5223,    Adjusted R-squared:  0.5212
F-statistic: 484.9 on 13 and 5766 DF,  p-value: < 2.2e-16
```

VIF for the model B shows below:

	GVIF	Df	GVIF^(1/(2*Df))
BRAND	1.513678	2	1.109196
logPRICE	1.766118	1	1.328954
Feat	1.230449	1	1.109256
AGE9	6.595235	1	2.568119
ETHNIC	5.615731	1	2.369753
HSIZEAVG	154.925761	1	12.446918
HH4PLUS	99.538993	1	9.976923
HHLARGE	30.538884	1	5.526200
HVAL150	2.090401	1	1.445822
NOCAR	9.870323	1	3.141707
SINGLE	12.815321	1	3.579849
WORKWOM	3.295816	1	1.815438

Although all p value is smaller than 0.05, meaning significant in model B, we have some variables, such as HSIZEAVG, HH4PLUS and HHLARGE are strongly correlated. So we reduce the model again to model C.

```
Residuals:
    Min      1Q  Median      3Q     Max 
-5.2380 -0.4534  0.0214  0.4722  3.4361 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       7.33526    0.29206  25.116  < 2e-16 ***
BRANDTF          -0.20316    0.03036  -6.691 2.42e-11 ***
BRANDTROPICANA    0.77747    0.02554  30.442  < 2e-16 ***
logPRICE         -3.11392    0.06098 -51.067  < 2e-16 ***
Feat              0.25776    0.02386  10.803  < 2e-16 ***
AGE9             -1.57308    0.91317  -1.723 0.085005 .  
ETHNIC            0.02048    0.09934   0.206 0.836644    
HSIZEAVG         -0.37539    0.10692  -3.511 0.000450 ***
HVAL150           1.06644    0.05310  20.083  < 2e-16 ***
NOCAR             0.83080    0.22943   3.621 0.000296 ***
SINGLE           -0.44401    0.44288  -1.003 0.316123    
WORKWOM          -1.75765    0.34393  -5.110 3.32e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7814 on 5768 degrees of freedom
Multiple R-squared:  0.5102,	Adjusted R-squared:  0.5093 
F-statistic: 546.3 on 11 and 5768 DF,  p-value: < 2.2e-16
```

VIF for the model C shows below:

```
               GVIF Df GVIF^(1/(2*Df))
BRAND      1.511612  2        1.108817
logPRICE   1.757568  1        1.325733
Feat       1.226671  1        1.107552
AGE60      4.704531  1        2.168993
ETHNIC     3.365364  1        1.834493
HHSIZEAVG  6.883462  1        2.623635
HVAL150    1.592418  1        1.261910
NOCAR      8.338412  1        2.887631
SINGLE     7.956414  1        2.820712
WORKWOM    3.088553  1        1.757428
```

In this case, both model B and model C are better than full model, because they have less correlation between variables. However, by reducing the # of variables in the model, the Rsquare and adjust Rsquare become smaller.

2.(5 points) Please use the data set 651F18 Orange Juice Homework 2.csv to answer this question.

Fit a regression model with dependent variable log of move, and independent variables BRAND, Feat, logPRICE and BRAND*logPRICE. For the six cases given below, use R to construct 99% prediction intervals for logMOVE.

Case	BRAND	logPRICE	Feat
1	FLNAT	.9	1
2	FLNAT	1.0	0
3	TF	.55	1
4	TF	.75	0
5	TROPICANA	.80	1
6	TROPICANA	.95	0

Solution:

r-CODE:

lm2a<-lm(logMOVE~BRAND+Feat+logPRICE+BRAND*logPRICE,data=oj1)
summary(lm2a)
predict(lm2a,interval="prediction",level=.99,newdata=Book1)

Output:

```
> lm2a<-lm(logMOVE~BRAND+Feat+logPRICE+BRAND*logPRICE,data=oj1)
> summary(lm2a)

Call:
lm(formula = logMOVE ~ BRAND + Feat + logPRICE + BRAND * logPRICE,
    data = oj1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3859 -0.4938  0.0086  0.5080  3.2308

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.55183    0.14210  39.071  < 2e-16 ***
BRANDTF                 -0.19352    0.14776  -1.310   0.1904
BRANDTROPICANA           1.17171    0.17540   6.680 2.61e-11 ***
Feat                     0.30155    0.02593  11.629  < 2e-16 ***
logPRICE                -2.83936    0.14140 -20.080  < 2e-16 ***
BRANDTF:logPRICE         0.09249    0.15748   0.587   0.5570
BRANDTROPICANA:logPRICE -0.42307    0.18265  -2.316   0.0206 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8377 on 5773 degrees of freedom
Multiple R-squared:  0.4366,    Adjusted R-squared:  0.4361
F-statistic: 745.7 on 6 and 5773 DF,  p-value: < 2.2e-16

> predict(lm2a,interval="prediction",level=.99,newdata=Book1)
       fit       lwr      upr
1 3.297963 1.1384684 5.457457
2 2.712472 0.5533087 4.871636
3 4.149094 1.9895548 6.308634
4 3.298166 1.1389990 5.457333
5 4.415150 2.2555732 6.574726
6 3.624230 1.4650688 5.783391
```

3.(a)(4 points) Please perform logit analysis using Personal Loan as the dependent variable, and all the remaining variables as independent variables. (For education, include the two dummy variables GRAD and PROF.) Include **only main effects** in your model. Which variables are significant at a 90% level of confidence? Copy and paste screen shots from R analysis to support your answers.

Solution

GLM3a<-glm(PersonalLoan~Age+Experience+Income+Family+CCAvg+GRAD+PROF+Mortgage+SecuritiesAccount+

CDAccount+Online+CreditCard,data=Bank)

summary(GLM3a)

```
Deviance Residuals:
     Min        1Q     Median        3Q       Max
-0.79039  -0.13533  -0.03435   0.07122   1.05807

Coefficients:
                     Estimate  Std. Error t value     Pr(>|t|)
(Intercept)       -0.17725205  0.06969237  -2.543       0.0110 *
Age               -0.00534073  0.00273640  -1.952       0.0510 .
Experience         0.00588222  0.00273375   2.152       0.0315 *
Income             0.00306523  0.00009623  31.854      < 2e-16 ***
Family             0.03017912  0.00289139  10.438      < 2e-16 ***
CCAvg              0.01215732  0.00243823   4.986 0.0000006369 ***
GRAD               0.14539853  0.00817964  17.776      < 2e-16 ***
PROF               0.15441144  0.00816394  18.914      < 2e-16 ***
Mortgage           0.00006759  0.00003267   2.069       0.0386 *
SecuritiesAccount -0.05983812  0.01130047  -5.295 0.0000001240 ***
CDAccount          0.32609565  0.01568961  20.784      < 2e-16 ***
Online            -0.02752139  0.00673604  -4.086 0.0000446353 ***
CreditCard        -0.04380917  0.00748760  -5.851 0.0000000052 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for gaussian family taken to be 0.05249746)

    Null deviance: 433.92  on 4999  degrees of freedom
Residual deviance: 261.80  on 4987  degrees of freedom
AIC: -530.58

Number of Fisher Scoring iterations: 2
```

confint(GLM3a,level=0.9,type="LR")

```
> confint(GLM3a,level=0.9,type="LR")
                              5 %          95 %
(Intercept)       -0.29188579497 -0.0626183074
Age               -0.00984171512 -0.0008397473
Experience         0.00138559564  0.0103788462
Income             0.00290695137  0.0032235099
Family             0.02542321646  0.0349350305
CCAvg              0.00814678649  0.0161678612
GRAD               0.13194422027  0.1588528371
PROF               0.14098294150  0.1678399301
Mortgage           0.00001385491  0.0001213223
SecuritiesAccount -0.07842574124 -0.0412504974
CDAccount          0.30028854550  0.3519027636
Online            -0.03860118748 -0.0164415890
CreditCard        -0.05612516982 -0.0314931612
```

Variables significant at a 90% level of confidence according to the R Analaysis in the screenshots above are:

Income, Family, CCAvg, GRAD, PROF, CDAccount, SecuritiesAccount, Online and CreditCard

3(b)(4 points) From your answer to question 3(a), list the variables that are not significant at a 90% level of confidence. Using test linear hypothesis, test the null hypothesis that none of these variables is significant using a 99% level of confidence. Copy and paste R screenshots to support your answer and clearly state your conclusion.

Solution:

The variables that are not significant at a 90% level of confidence according to the results from question 3(a) are:

Age, Experience, Mortgage

Testing the null hypothesis that none of the above mentioned variables are significant:

```
> local({
+   .Hypothesis <- matrix(c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0), 3, 13, byrow=TRUE)
+   .RHS <- c(0,0,0)
+   linearHypothesis(GLM3ax, .Hypothesis, rhs=.RHS, test="Chisq")
+ })
Linear hypothesis test

Hypothesis:
Age = 0
Experience = 0
Mortgage = 0

Model 1: restricted model
Model 2: PersonalLoan ~ Age + Experience + Income + Family + CCAvg + GRAD +
    PROF + Mortgage + SecuritiesAccount + CDAccount + Online +
    CreditCard

  Res.Df Df  Chisq Pr(>Chisq)
1   4990
2   4987  3 3.5993     0.3092
```

```
> confint(GLM3ax,level=0.99,type="LR")
                           0.5 %       99.5 %
(Intercept)        -16.9665849540 -7.585022465
Age                 -0.2145188953  0.132595614
Experience          -0.1222938823  0.222529869
Income               0.0528536945  0.068159212
Family               0.4232712891  0.820794824
CCAvg                0.0505154881  0.277849907
GRAD                 3.2929144504  4.683721593
PROF                 3.3991425115  4.776239531
Mortgage            -0.0008298207  0.002235178
SecuritiesAccount   -1.6809979726 -0.128216365
CDAccount            2.9796154009  4.742307210
Online              -1.1918640237 -0.336446884
CreditCard          -1.6062695982 -0.505753483
```

From the screenshots above, we can deduce that Age, Experience and Mortgage are not significant at 99% level of confidence.

3(c)(7 points) Based on your conclusion for 3(b), use the reduced model (if you cannot reject null hypothesis) or the original model (if you can reject the null hypothesis) to answer this question. If you are using the reduced model, estimate it.

Briefly discuss how the predictors in this model affect the probability of Personal Loan. Identify which one among the continuous predictors has the highest effect on the probability of Personal

Loan. Identify which one among the 1/0 variables has the highest effect on the probability of personal loan.

(Hint: For a continuous predictor, compare how much the indicator function I changes if that predictor changes by one standard deviation. For a 1/0 variable, find how much I changes if that variable changes from 0 to 1.)

Solution:

GLM3b<-glm(PersonalLoan~Income+Family+CCAvg+GRAD+PROF+SecuritiesAccount+

CDAccount+Online+CreditCard,data=Bank)

summary(GLM3b)

```
Call:
glm(formula = PersonalLoan ~ Income + Family + CCAvg + GRAD +
    PROF + SecuritiesAccount + CDAccount + Online + CreditCard,
    data = Bank)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.79372  -0.13612  -0.03349   0.07087   1.05748

Coefficients:
                     Estimate  Std. Error t value      Pr(>|t|)
(Intercept)       -0.29586747  0.01194597 -24.767       < 2e-16 ***
Income             0.00310278  0.00009479  32.732       < 2e-16 ***
Family             0.02969212  0.00288685  10.285       < 2e-16 ***
CCAvg              0.01160466  0.00243490   4.766 0.00000193287 ***
GRAD               0.14321729  0.00813260  17.610       < 2e-16 ***
PROF               0.15046947  0.00790261  19.040       < 2e-16 ***
SecuritiesAccount -0.06101605  0.01130446  -5.398 0.00000007069 ***
CDAccount          0.32978272  0.01565696  21.063       < 2e-16 ***
Online            -0.02780050  0.00673977  -4.125 0.00003770269 ***
CreditCard        -0.04420399  0.00749134  -5.901 0.00000000386 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.05259411)

    Null deviance: 433.92  on 4999  degrees of freedom
Residual deviance: 262.44  on 4990  degrees of freedom
AIC: -524.38

Number of Fisher Scoring iterations: 2
```

The variables that have a positive significant affect on Perosnal Loan are :

Income, Family, CCAvg, GRAD, PROF, CDAccount

The variables that have a negative significant affect on Personal Loan are:

SecuritiesAccount, Online and CreditCard

-1/2

Need to tell which variables have highest effects on probability.

Running 1/0 model

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4818  -0.4786  -0.3296  -0.2544   2.8958

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.8856     0.1307 -22.074  < 2e-16 ***
PROF                1.3057     0.1391   9.384  < 2e-16 ***
GRAD                1.2458     0.1422   8.760  < 2e-16 ***
SecuritiesAccount  -1.2920     0.2055  -6.286 3.25e-10 ***
CDAccount           3.9494     0.2127  18.568  < 2e-16 ***
CreditCard         -1.1481     0.1553  -7.393 1.44e-13 ***
Online             -0.5292     0.1127  -4.698 2.63e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3162.0  on 4999  degrees of freedom
Residual deviance: 2616.1  on 4993  degrees of freedom
AIC: 2630.1
```

The variables affecting Personal Loan positively are:

PROF, GRAD, CDAccount

The variables affectint Personal Loan Negativiely are:

SecurtitiesAccount, CreditCard and Online

3(d)(5 points) Please perform a logit analysis with Personal Loan as the dependent variable and the following independent variables:

CCAvg, CDAccount , CreditCard, ED, Family, Income, Online, SecuritiesAccount,and the interaction of Income with each of ED, CD Account, Credit Card, Online and Securities Account.

(In this case, we will be able to check if the effect of income is "moderated," that is, affected by the level of ED, DC Account, etc. Here, ED, CD Account, etc., are called moderating variables.)

Focus on the interaction terms and identify the interactions that are not significant at a 90% level of confidence. At a 99% level of confidence, test the null hypothesis that none of these interaction terms is significant. Based on the result, keep the original model or the reduced model as appropriate. Briefly discuss how the effect of income is affected by the moderating variable.

Solution:

GLM3d <- glm(PersonalLoan ~ Income + Family + CCAvg + ED + SecuritiesAccount + CDAccount + Online + CreditCard + Income*ED + Income*CDAccount +

Income*CreditCard + Income*Online + Income*SecuritiesAccount, family=binomial(logit), data=Bank)

summary(GLM3d)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1408  -0.1591  -0.0514  -0.0035   3.4862

Coefficients:
                            Estimate  Std. Error z value     Pr(>|z|)
(Intercept)               -7.8648725   0.6173288 -12.740      < 2e-16 ***
Income                     0.0226901   0.0041195   5.557 0.0000000275 ***
Family                     0.8169216   0.0942480   8.668      < 2e-16 ***
CCAvg                      0.2091331   0.0517929   4.038 0.0000539382 ***
ED                       -10.6892661   1.1041484  -9.681      < 2e-16 ***
SecuritiesAccount         -1.8852091   1.0529051  -1.790       0.0734 .
CDAccount                  4.8925547   1.1567438   4.230 0.0000234115 ***
Online                    -0.8347757   0.6479559  -1.288       0.1976
CreditCard                -1.6638308   0.8750384  -1.901       0.0572 .
Income:ED                  0.1259426   0.0103406  12.179      < 2e-16 ***
Income:CDAccount          -0.0080032   0.0091572  -0.874       0.3821
Income:CreditCard          0.0037484   0.0069832   0.537       0.5914
Income:Online             -0.0007825   0.0051085  -0.153       0.8783
Income:SecuritiesAccount   0.0078930   0.0085382   0.924       0.3553
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3162.0  on 4999  degrees of freedom
Residual deviance:  549.4  on 4986  degrees of freedom
AIC: 577.4

Number of Fisher Scoring iterations: 9

> exp(coef(GLM3d))  # Exponentiated coefficients ("odds ratios")
           (Intercept)                 Income                 Family                  CCAvg                     ED    SecuritiesAccount
      0.00033339827           1.02315910195          2.26352107323          1.23262906753          0.00022278824        0.15175730324
             CDAccount                 Online             CreditCard              Income:ED       Income:CDAccount    Income:CreditCard
    135.29366125107           0.43397152663          0.18941198175          1.13421732735          0.59222873925        1.00375544202
         Income:Online Income:SecuritiesAccount
      0.99921752513           1.03792426055
```

According to the results above, the following interactions are not significant at 90% level of confidence:

```
Income:CDAccount
Income:CreditCard
Income:Online
Income:SecuritiesAccount
```

Following is the result of running null hypothesis test for the above insignificant interaction terms:

```
> local({
+   .Hypothesis <- matrix(c(0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1), 4, 14, byrow=TRUE)
+   .RHS <- c(0,0,0,0)
+   linearHypothesis(GLM.5, .Hypothesis, rhs=.RHS, test="Chisq")
+ })
Linear hypothesis test

Hypothesis:
Income:CDAccount = 0
Income:CreditCard = 0
Income:Online = 0
Income:SecuritiesAccount = 0

Model 1: restricted model
Model 2: PersonalLoan ~ Income + Family + CCAvg + ED + SecuritiesAccount +
    CDAccount + Online + CreditCard + Income * ED + Income *
    CDAccount + Income * CreditCard + Income * Online + Income *
    SecuritiesAccount

  Res.Df Df  Chisq Pr(>Chisq)
1   4990
2   4986  4 1.3001     0.8614
```

Running the reduced model

GLM.6 <- glm(PersonalLoan ~ Income + Family + CCAvg + ED + SecuritiesAccount + CDAccount + Online + CreditCard + Income * ED, family=binomial(logit), data=Bank)

summary(GLM.6)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1320  -0.1618  -0.0537  -0.0037   3.4936

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -7.822625   0.515658 -15.170  < 2e-16 ***
Income               0.022601   0.003089   7.316 2.56e-13 ***
Family               0.810191   0.093781   8.639  < 2e-16 ***
CCAvg                0.212874   0.051715   4.116 3.85e-05 ***
ED                 -10.691213   1.102557  -9.697  < 2e-16 ***
SecuritiesAccount   -0.962788   0.339646  -2.835  0.00459 **
CDAccount            3.942669   0.384271  10.260  < 2e-16 ***
Online              -0.922362   0.204460  -4.511 6.45e-06 ***
CreditCard          -1.214731   0.259601  -4.679 2.88e-06 ***
Income:ED            0.125879   0.010327  12.189  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3162.04  on 4999  degrees of freedom
Residual deviance:  853.67  on 4989  degrees of freedom
AIC: 875.67

Number of Fisher Scoring iterations: 9

> exp(coef(GLM.6))  # Exponentiated coefficients ("odds ratios")
      (Intercept)            Income            Family             CCAvg                ED SecuritiesAccount         CDAccount            Online        CreditCard
    0.00040056591     1.02285594898     2.24833793263     1.23722929950     0.00002274392     0.38182659915    51.55601917839     0.39757896434     0.29678991819
        Income:ED
    1.13414515440
```

When running the reduced model with splitting ED into GRAD and PROF, we see the effect of both these variables are similarly significant.

Need to tell how the effect of income is moderated by ~~inter~~ education.

−1/2

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4505  -0.2200  -0.1137  -0.0088   3.6024

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.474462   0.332276 -16.476  < 2e-16 ***
Income       0.021789   0.002353   9.261  < 2e-16 ***
GRAD        -9.412213   1.341865  -7.014 2.31e-12 ***
PROF        -9.666831   1.364517  -7.084 1.40e-12 ***
Income:GRAD  0.113072   0.012616   8.963  < 2e-16 ***
Income:PROF  0.113433   0.012569   9.024  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3162.0  on 4999  degrees of freedom
Residual deviance: 1093.3  on 4994  degrees of freedom
AIC: 1105.3

Number of Fisher Scoring iterations: 9
```