

Linear & Logistic Regression: Backward elimination

Farhat Jabeen

20/1/2022

Contents

Linear Regression	3
Model comparison:	7
Model interpretation:	7
Logistic Regression1:	7
Create a dummy variable for Gender.	7
Logistic Regression2:	9
Run a logistic regression with all the independent variables.	10
Step by step elimination	11
Transforming independent variables: Calculating log of Balance	13
Creating derived variables	14
Checking for multicollinearity	15

Load the required libraries.

```
library(c("lme4", "lmerTest", "emmeans", "car", "lattice", "ggplot2", "irr", "knitr", "languageR", "MASS"))

## Loading required package: lme4
## Loading required package: Matrix
## Loading required package: lmerTest
##
## Attaching package: 'lmerTest'
## The following object is masked from 'package:lme4':
##
##     lmer
## The following object is masked from 'package:stats':
##
##     step
## Loading required package: emmeans
## Loading required package: car
## Loading required package: carData
## Registered S3 methods overwritten by 'car':
##   method                               from
##   influence.merMod                     lme4
##   cooks.distance.influence.merMod      lme4
##   dfbeta.influence.merMod              lme4
##   dfbetas.influence.merMod             lme4
## Loading required package: lattice
```

```

## Loading required package: ggplot2
## Warning: package 'ggplot2' was built under R version 3.6.2
## Loading required package: irr
## Loading required package: lpSolve
## Loading required package: knitr
## Loading required package: languageR
## Loading required package: MASS
## Loading required package: Rmisc
## Loading required package: plyr
## Loading required package: dplyr
## Warning: package 'dplyr' was built under R version 3.6.2
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
## The following object is masked from 'package:MASS':
##
##     select
## The following object is masked from 'package:car':
##
##     recode
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
## Loading required package: MuMIn
## Loading required package: tidyr
##
## Attaching package: 'tidyr'
## The following objects are masked from 'package:Matrix':
##
##     expand, pack, unpack
## Loading required package: corpcor
## Warning: package 'corpcor' was built under R version 3.6.2
## [[1]]
## [1] TRUE
##
## [[2]]

```

```
## [1] TRUE
##
## [[3]]
## [1] TRUE
##
## [[4]]
## [1] TRUE
##
## [[5]]
## [1] TRUE
##
## [[6]]
## [1] TRUE
##
## [[7]]
## [1] TRUE
##
## [[8]]
## [1] TRUE
##
## [[9]]
## [1] TRUE
##
## [[10]]
## [1] TRUE
##
## [[11]]
## [1] TRUE
##
## [[12]]
## [1] TRUE
##
## [[13]]
## [1] TRUE
##
## [[14]]
## [1] TRUE
##
## [[15]]
## [1] TRUE
```

Linear Regression

Problem: Predicting company profits.

```
#Read in the dataset.
startup <- read.table("~/Desktop/dataScience/resources/P12-50-Startups.csv", sep = ",", header = T)

head(startup)
```

```
##   R.D.Spend Administration Marketing.Spend   State  Profit
## 1  165349.2      136897.80      471784.1 New York 192261.8
## 2  162597.7      151377.59      443898.5 California 191792.1
## 3  153441.5      101145.55      407934.5 California 191050.4
```

```
## 4 144372.4      118671.85      383199.6   New York 182902.0
## 5 142107.3      91391.77      366168.4 California 166187.9
## 6 131876.9      99814.71      362861.4   New York 156991.1
```

```
names(startup)
```

```
## [1] "R.D.Spend"      "Administration" "Marketing.Spend" "State"
## [5] "Profit"
```

create a dummy variable for state (California, New York)

```
startup$newyork <- 0
startup[startup$State == "New York",]$newyork <- 1
#I chose not to create the dummy for California as it can't be included in the analysis anyway.
```

Begin modeling with putting all the variables in the linear regression model.

Predicting profit based on how much companies spend on R&D, administration, marketing, and if their location plays a role in this.

```
profitAll <- lm(Profit ~ R.D.Spend + Administration + Marketing.Spend + newyork, data = startup)
```

```
#Calculate Type III Anova for significant p-value (0.05 in thsi instance).
Anova(profitAll, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: Profit
##
```

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	4.9756e+09	1	57.4168	1.431e-09 ***
R.D.Spend	2.7031e+10	1	311.9298	< 2.2e-16 ***
Administration	1.7979e+07	1	0.2075	0.6509
Marketing.Spend	2.1662e+08	1	2.4997	0.1209
newyork	2.1248e+07	1	0.2452	0.6229
Residuals	3.8996e+09	45		

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

```
summary(profitAll)
```

```
##
## Call:
## lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend +
##     newyork, data = startup)
##
## Residuals:
```

Min	1Q	Median	3Q	Max
-34163	-4312	113	6631	17916

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.042e+04	6.654e+03	7.577	1.43e-09 ***
R.D.Spend	8.080e-01	4.575e-02	17.662	< 2e-16 ***
Administration	-2.362e-02	5.186e-02	-0.455	0.651
Marketing.Spend	2.637e-02	1.668e-02	1.581	0.121
newyork	-1.332e+03	2.690e+03	-0.495	0.623

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9309 on 45 degrees of freedom
## Multiple R-squared:  0.951, Adjusted R-squared:  0.9467
## F-statistic: 218.4 on 4 and 45 DF,  p-value: < 2.2e-16
```

Remove the variable with the highest p-value (exceeding the a-level of 0.05) i.e. Administration.

```
profitA <- lm(Profit ~ R.D.Spend + Marketing.Spend + newyork, data = startup)
Anova(profitA, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: Profit
##              Sum Sq Df F value  Pr(>F)
## (Intercept)  2.1289e+10  1 249.9752 < 2e-16 ***
## R.D.Spend    3.0667e+10  1 360.0894 < 2e-16 ***
## Marketing.Spend 2.7864e+08  1   3.2718 0.07702 .
## newyork      2.6807e+07  1   0.3148 0.57749
## Residuals    3.9176e+09 46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(profitA)
```

```
##
## Call:
## lm(formula = Profit ~ R.D.Spend + Marketing.Spend + newyork,
##     data = startup)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -34332  -4681    100    5792   17834
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.772e+04  3.018e+03  15.811  <2e-16 ***
## R.D.Spend    8.003e-01  4.217e-02  18.976  <2e-16 ***
## Marketing.Spend 2.859e-02  1.581e-02   1.809   0.077 .
## newyork      -1.485e+03  2.646e+03  -0.561   0.577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9228 on 46 degrees of freedom
## Multiple R-squared:  0.9508, Adjusted R-squared:  0.9476
## F-statistic: 296.2 on 3 and 46 DF,  p-value: < 2.2e-16
```

Remove the next variable with the highest p-value i.e. newyork.

```
profitB <- lm(Profit ~ R.D.Spend + Marketing.Spend, data = startup)
Anova(profitB, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: Profit
##              Sum Sq Df F value    Pr(>F)
## (Intercept)  2.5595e+10  1 304.9767 < 2e-16 ***
## R.D.Spend    3.1149e+10  1 371.1616 < 2e-16 ***
## Marketing.Spend 3.1165e+08  1   3.7135 0.06003 .
## Residuals    3.9444e+09 47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(profitB)

##
## Call:
## lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = startup)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -33645  -4632   -414    6484   17097
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.698e+04  2.690e+03  17.464  <2e-16 ***
## R.D.Spend    7.966e-01  4.135e-02  19.266  <2e-16 ***
## Marketing.Spend 2.991e-02  1.552e-02   1.927    0.06 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9161 on 47 degrees of freedom
## Multiple R-squared:  0.9505, Adjusted R-squared:  0.9483
## F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

Remove the next variable with the highest p-value i.e. Marketing.Spend

```
profitC <- lm(Profit ~ R.D.Spend, data = startup)

Anova(profitC, type = "III")

## Anova Table (Type III tests)
##
## Response: Profit
##              Sum Sq Df F value    Pr(>F)
## (Intercept) 3.3097e+10  1  373.27 < 2.2e-16 ***
## R.D.Spend   7.5349e+10  1  849.79 < 2.2e-16 ***
## Residuals   4.2560e+09 48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(profitC)

##
## Call:
## lm(formula = Profit ~ R.D.Spend, data = startup)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -34351  -4626   -375    6249   17188
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.903e+04  2.538e+03  19.32  <2e-16 ***
## R.D.Spend   8.543e-01  2.931e-02  29.15  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9416 on 48 degrees of freedom
## Multiple R-squared:  0.9465, Adjusted R-squared:  0.9454
## F-statistic: 849.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

Model comparison:

Adjusted R-squared value shows that profitB model is better with a slightly high value. Although the p-value for “Marketing.Spend” (0.06) is slightly above the a-level, removing it lowers the Adjusted R-squared value. As its p-value is only marginally above the a-level, “Marketing.Spend” can be kept in the final model.

Model interpretation:

The summary of profitB model shows that higher expenditure on R&D leads to higher company profits. The spending on marketing also leads to an increase in profits.

Logistic Regression1:

Using their age and gender, is it possible to predict if someone visiting a website took an action or not.

```
email <- read.table("~/Desktop/dataScience/resources/P12-Email-Offer.csv", sep = ",", header = T)
head(email)
```

```
##   Age Gender TookAction
## 1  38 Female           0
## 2  32 Female           0
## 3  46  Male            1
## 4  34  Male            0
## 5  40  Male            0
## 6  37 Female           0
```

Look at the distribution of gender and if males/females differed in terms of taking action.

```
table(email$TookAction, email$Gender)
```

```
##
##      Female Male
## 0       35    25
## 1       15    25
```

Create a dummy varibale for Gender.

```
email$female <- "0"
email[email$Gender == "Male",]$female <- "1"

head(email)
```

```
##   Age Gender TookAction female
## 1  38 Female           0      0
## 2  32 Female           0      0
## 3  46  Male            1      1
## 4  34  Male            0      1
## 5  40  Male            0      1
## 6  37 Female           0      0
```

Run a logistic regression model with both the independent variables.

```
femaleAge.glm <- glm(TookAction ~ Age + female, data = email, family = "binomial")
summary(femaleAge.glm)
```

```
##
## Call:
## glm(formula = TookAction ~ Age + female, family = "binomial",
##      data = email)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96524  -0.09857  -0.00566   0.09864   2.67479
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -38.1520     9.9869  -3.820 0.000133 ***
## Age           0.8872     0.2318   3.828 0.000129 ***
## female1      4.4374     1.4919   2.974 0.002937 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 134.602  on 99  degrees of freedom
## Residual deviance:  29.613  on 97  degrees of freedom
## AIC: 35.613
##
## Number of Fisher Scoring iterations: 8
```

Both the variables are significant at the α -level of 0.05. Older people are more likely to take an action as compared with younger ones. Males took action significantly more often than females.

Interpreting model outcome in terms of probabilities.

Probability for Age.

```
plogis(0.88)
```

```
## [1] 0.7068222
```

Probability for Gender


```
plogis(4.43)
```

```
## [1] 0.9882258
```

Logistic Regression2:

Geo-demographic segmentation: Using independent variables to predict how likely is a customer to leave a (fictional) bank.

*#Following error message occurred: EOF within quoted string number of items read is not a multiple of t
#using quote = "" resolves this issue.*

```
exitData <- read.table("~/Desktop/dataScience/resources/P12-Churn-Modelling.csv", sep = ",", header = T
```

```
head(exitData)
```

```
##   RowNumber CustomerId Surname CreditScore Geography Gender Age Tenure
## 1         1   15634602 Hargrave         619    France Female  42      2
## 2         2   15647311   Hill         608    Spain Female  41      1
## 3         3   15619304   Onio         502    France Female  42      8
## 4         4   15701354   Boni         699    France Female  39      1
## 5         5   15737888 Mitchell        850    Spain Female  43      2
## 6         6   15574012    Chu         645    Spain   Male   44      8
##      Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
## 1         0.00             1          1              1      101348.88      1
## 2    83807.86             1          0              1      112542.58      0
## 3   159660.80             3          1              0      113931.57      1
## 4         0.00             2          0              0       93826.63      0
## 5   125510.82             1          1              1       79084.10      0
## 6   113755.78             2          1              0      149756.71      1
```

```
nrow(exitData)
```

```
## [1] 10000
```

Create dummies for gender and geography.

```
exitData$female <- 0
exitData[exitData$Gender != "Female",]$female <- 1

#keeping France as baseline, just due to alphabetical order.
exitData$Germany <- 0
exitData[exitData$Geography != "Germany",]$Germany <- 1

exitData$Spain <- 0
exitData[exitData$Geography != "Spain",]$Spain <- 1

head(exitData)
```

```
##   RowNumber CustomerId Surname CreditScore Geography Gender Age Tenure
## 1         1   15634602 Hargrave         619    France Female  42      2
## 2         2   15647311   Hill         608    Spain Female  41      1
## 3         3   15619304   Onio         502    France Female  42      8
## 4         4   15701354   Boni         699    France Female  39      1
## 5         5   15737888 Mitchell        850    Spain Female  43      2
## 6         6   15574012    Chu         645    Spain   Male   44      8
```

```
##      Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
## 1      0.00           1           1           1      101348.88           1
## 2  83807.86           1           0           1      112542.58           0
## 3 159660.80           3           1           0      113931.57           1
## 4      0.00           2           0           0       93826.63           0
## 5 125510.82           1           1           1       79084.10           0
## 6 113755.78           2           1           0     149756.71           1
##   female Germany Spain
## 1      0         1     1
## 2      0         1     0
## 3      0         1     1
## 4      0         1     1
## 5      0         1     0
## 6      1         1     0
```

Run a logistic regression with all the independent variables.

```
geodem.glm <- glm(Exited ~ CreditScore + Age + Tenure + Balance + NumOfProducts + HasCrCard + IsActiveMember + EstimatedSalary + female + Spain + Germany, family = "binomial", data = exitData)
summary(geodem.glm)
```

```
##
## Call:
## glm(formula = Exited ~ CreditScore + Age + Tenure + Balance +
##      NumOfProducts + HasCrCard + IsActiveMember + EstimatedSalary +
##      female + Spain + Germany, family = "binomial", data = exitData)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3097  -0.6589  -0.4560  -0.2697   2.9940
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.582e+00  2.641e-01  -9.777  < 2e-16 ***
## CreditScore  -6.683e-04  2.803e-04  -2.384   0.0171 *
## Age           7.271e-02  2.576e-03  28.230  < 2e-16 ***
## Tenure       -1.595e-02  9.355e-03  -1.705   0.0882 .
## Balance       2.637e-06  5.142e-07   5.128 2.92e-07 ***
## NumOfProducts -1.015e-01  4.713e-02  -2.154   0.0312 *
## HasCrCard     -4.468e-02  5.934e-02  -0.753   0.4515
## IsActiveMember -1.075e+00  5.769e-02 -18.643  < 2e-16 ***
## EstimatedSalary 4.807e-07  4.737e-07   1.015   0.3102
## female       -5.285e-01  5.449e-02  -9.699  < 2e-16 ***
## Spain        -3.522e-02  7.064e-02  -0.499   0.6181
## Germany      -7.747e-01  6.767e-02 -11.448  < 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: 10109.8  on 9999  degrees of freedom
## Residual deviance:  8561.4  on 9988  degrees of freedom
## AIC: 8585.4
##
```

```
## Number of Fisher Scoring iterations: 5
```

Step by step elimination

Remove the variable with the highest p-value (exceeding the a-level of 0.05) i.e. Spain

```
geodem.glmA <- glm(Exited ~ CreditScore + Age + Tenure + Balance + NumOfProducts + HasCrCard + IsActiveMember + EstimatedSalary + female + Germany, family = "binomial", data = exitData)

summary(geodem.glmA)
```

```
##
## Call:
## glm(formula = Exited ~ CreditScore + Age + Tenure + Balance +
##      NumOfProducts + HasCrCard + IsActiveMember + EstimatedSalary +
##      female + Germany, family = "binomial", data = exitData)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3099  -0.6584  -0.4559  -0.2691   2.9901
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.620e+00  2.533e-01 -10.342  < 2e-16 ***
## CreditScore  -6.666e-04  2.803e-04  -2.378   0.0174 *
## Age           7.272e-02  2.575e-03  28.238  < 2e-16 ***
## Tenure       -1.598e-02  9.354e-03  -1.708   0.0876 .
## Balance       2.637e-06  5.142e-07   5.129 2.91e-07 ***
## NumOfProducts -1.013e-01  4.713e-02  -2.149   0.0316 *
## HasCrCard     -4.493e-02  5.934e-02  -0.757   0.4489
## IsActiveMember -1.075e+00  5.768e-02 -18.640  < 2e-16 ***
## EstimatedSalary 4.813e-07  4.736e-07   1.016   0.3095
## female       -5.283e-01  5.449e-02  -9.697  < 2e-16 ***
## Germany      -7.629e-01  6.336e-02 -12.041  < 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: 10109.8  on 9999  degrees of freedom
## Residual deviance:  8561.6  on 9989  degrees of freedom
## AIC: 8583.6
##
## Number of Fisher Scoring iterations: 5
```

Remove HasCrCard

```
geodem.glmB <- glm(Exited ~ CreditScore + Age + Tenure + Balance + NumOfProducts + IsActiveMember + EstimatedSalary + female, family = "binomial", data = exitData)

summary(geodem.glmB)
```

```
##
## Call:
## glm(formula = Exited ~ CreditScore + Age + Tenure + Balance +
##      NumOfProducts + IsActiveMember + EstimatedSalary + female, family = "binomial", data = exitData)
```

```
##      Germany, family = "binomial", data = exitData)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3152  -0.6585  -0.4565  -0.2699   2.9859
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.654e+00  2.493e-01 -10.647 < 2e-16 ***
## CreditScore  -6.640e-04  2.803e-04  -2.369  0.0178 *
## Age           7.273e-02  2.575e-03  28.243 < 2e-16 ***
## Tenure       -1.615e-02  9.351e-03  -1.727  0.0842 .
## Balance       2.645e-06  5.141e-07   5.146 2.66e-07 ***
## NumOfProducts -1.013e-01  4.712e-02  -2.150  0.0315 *
## IsActiveMember -1.074e+00  5.767e-02 -18.631 < 2e-16 ***
## EstimatedSalary 4.818e-07  4.737e-07   1.017  0.3091
## female       -5.285e-01  5.449e-02  -9.700 < 2e-16 ***
## Germany      -7.619e-01  6.334e-02 -12.028 < 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: 10109.8  on 9999  degrees of freedom
## Residual deviance:  8562.2  on 9990  degrees of freedom
## AIC: 8582.2
##
## Number of Fisher Scoring iterations: 5
```

Remove EstimatedSalary

```
geodem.glmC <- glm(Exited ~ CreditScore + Age + Tenure + Balance + NumOfProducts + IsActiveMember + fem
summary(geodem.glmC)
```

```
##
## Call:
## glm(formula = Exited ~ CreditScore + Age + Tenure + Balance +
##      NumOfProducts + IsActiveMember + female + Germany, family = "binomial",
##      data = exitData)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3272  -0.6592  -0.4557  -0.2688   2.9787
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.605e+00  2.445e-01 -10.654 < 2e-16 ***
## CreditScore  -6.664e-04  2.803e-04  -2.378  0.0174 *
## Age           7.270e-02  2.575e-03  28.238 < 2e-16 ***
## Tenure       -1.598e-02  9.349e-03  -1.710  0.0873 .
## Balance       2.653e-06  5.140e-07   5.162 2.44e-07 ***
## NumOfProducts -1.005e-01  4.712e-02  -2.132  0.0330 *
## IsActiveMember -1.075e+00  5.766e-02 -18.644 < 2e-16 ***
```

```
## female          -5.290e-01  5.448e-02  -9.710  < 2e-16 ***
## Germany         -7.621e-01  6.334e-02 -12.031  < 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: 10109.8  on 9999  degrees of freedom
## Residual deviance:  8563.2  on 9991  degrees of freedom
## AIC: 8581.2
##
## Number of Fisher Scoring iterations: 5
```

Transforming independent variables: Calculating log of Balance

```
#+1 because there are 0s in the Balance column
exitData$logBalance <- log10((exitData$Balance) + 1)

head(exitData)
```

```
##   RowNumber CustomerId Surname CreditScore Geography Gender Age Tenure
## 1         1   15634602 Hargrave         619    France Female  42      2
## 2         2   15647311   Hill         608    Spain Female  41      1
## 3         3   15619304   Onio         502    France Female  42      8
## 4         4   15701354   Boni         699    France Female  39      1
## 5         5   15737888 Mitchell         850    Spain Female  43      2
## 6         6   15574012    Chu         645    Spain   Male   44      8
##   Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
## 1      0.00              1          1              1      101348.88      1
## 2  83807.86              1          0              1      112542.58      0
## 3 159660.80              3          1              0      113931.57      1
## 4      0.00              2          0              0       93826.63      0
## 5 125510.82              1          1              1       79084.10      0
## 6 113755.78              2          1              0      149756.71      1
##   female Germany Spain logBalance
## 1      0        1      1  0.000000
## 2      0        1      0  4.923290
## 3      0        1      1  5.203201
## 4      0        1      1  0.000000
## 5      0        1      0  5.098685
## 6      1        1      0  5.055977
```

Replace Balance with logBalance in the regression model.

```
geodem.glmD <- glm(Exited ~ CreditScore + Age + NumOfProducts + IsActiveMember + female + Germany + Tenure,
                  data = exitData, family = "binomial")

summary(geodem.glmD)
```

```
##
## Call:
## glm(formula = Exited ~ CreditScore + Age + NumOfProducts + IsActiveMember +
##      female + Germany + Tenure + logBalance, family = "binomial",
##      data = exitData)
##
```

```
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3104  -0.6586  -0.4553  -0.2679   2.9827
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.6382591  0.2480823 -10.635 < 2e-16 ***
## CreditScore  -0.0006749  0.0002803  -2.408  0.0160 *
## Age           0.0726550  0.0025745  28.221 < 2e-16 ***
## NumOfProducts -0.0950198  0.0475374  -1.999  0.0456 *
## IsActiveMember -1.0757759  0.0576457 -18.662 < 2e-16 ***
## female       -0.5267214  0.0544591  -9.672 < 2e-16 ***
## Germany      -0.7475955  0.0650514 -11.492 < 2e-16 ***
## Tenure       -0.0158791  0.0093463  -1.699  0.0893 .
## logBalance    0.0690263  0.0139592   4.945 7.62e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 10109.8  on 9999  degrees of freedom
## Residual deviance:  8565.1  on 9991  degrees of freedom
## AIC: 8583.1
##
## Number of Fisher Scoring iterations: 5
```

Creating derived variables

Balance divided by age to calculate accumulation of wealth over time in a customer's life.

```
exitData$wealthAcc <- exitData$Balance / exitData$Age
```

```
head(exitData)
```

```
##   RowNumber CustomerId Surname CreditScore Geography Gender Age Tenure
## 1         1   15634602 Hargrave         619    France Female  42      2
## 2         2   15647311 Hill         608    Spain Female  41      1
## 3         3   15619304 Onio         502    France Female  42      8
## 4         4   15701354 Boni         699    France Female  39      1
## 5         5   15737888 Mitchell       850    Spain Female  43      2
## 6         6   15574012 Chu          645    Spain Male    44      8
##   Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
## 1      0.00             1          1              1      101348.88      1
## 2  83807.86             1          0              1      112542.58      0
## 3 159660.80             3          1              0      113931.57      1
## 4      0.00             2          0              0       93826.63      0
## 5 125510.82             1          1              1       79084.10      0
## 6 113755.78             2          1              0      149756.71      1
##   female Germany Spain logBalance wealthAcc
## 1      0         1      1  0.000000      0.000
## 2      0         1      0  4.923290    2044.094
## 3      0         1      1  5.203201    3801.448
## 4      0         1      1  0.000000      0.000
## 5      0         1      0  5.098685    2918.856
```

```
## 6      1      1      0  5.055977  2585.359
```

Add variable for wealth accumulation to the model.

```
geodem.glmE <- glm(Exited ~ CreditScore + Age + NumOfProducts + IsActiveMember + female + Germany + Tenure + logBalance + wealthAcc, family = "binomial", data = exitData)
summary(geodem.glmE)
```

```
##
## Call:
## glm(formula = Exited ~ CreditScore + Age + NumOfProducts + IsActiveMember + female + Germany + Tenure + logBalance + wealthAcc, family = "binomial", data = exitData)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3047  -0.6585  -0.4550  -0.2698   2.9662
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.556e+00  2.582e-01  -9.898  < 2e-16 ***
## CreditScore  -6.756e-04  2.803e-04  -2.410  0.015958 *
## Age           7.067e-02  3.095e-03  22.836  < 2e-16 ***
## NumOfProducts -9.553e-02  4.756e-02  -2.009  0.044576 *
## IsActiveMember -1.073e+00  5.767e-02 -18.612  < 2e-16 ***
## female        -5.257e-01  5.447e-02  -9.651  < 2e-16 ***
## Germany       -7.463e-01  6.513e-02 -11.459  < 2e-16 ***
## Tenure        -1.593e-02  9.347e-03  -1.704  0.088415 .
## logBalance     9.509e-02  2.662e-02   3.572  0.000354 ***
## wealthAcc     -4.336e-05  3.779e-05  -1.147  0.251224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 10109.8  on 9999  degrees of freedom
## Residual deviance:  8563.8  on 9990  degrees of freedom
## AIC: 8583.8
##
## Number of Fisher Scoring iterations: 5
```

#not significant. Also, possible collinearity as balance and age are related with wealthAcc here.

Checking for multicollinearity

```
#Subset data from columns 7 to 19 to exclude irrelevant variables or the ones with factors.
colExitData <- exitData[,7:19]
head(colExitData)
```

```
##   Age Tenure  Balance NumOfProducts HasCrCard IsActiveMember
## 1  42      2    0.00             1         1             1
## 2  41      1 83807.86             1         0             1
## 3  42      8 159660.80             3         1             0
## 4  39      1    0.00             2         0             0
```

```
## 5 43      2 125510.82      1      1      1
## 6 44      8 113755.78      2      1      0
##      EstimatedSalary Exited female Germany Spain logBalance wealthAcc
## 1      101348.88      1      0      1      1      0.000000      0.000
## 2      112542.58      0      0      1      0      4.923290      2044.094
## 3      113931.57      1      0      1      1      5.203201      3801.448
## 4      93826.63      0      0      1      1      0.000000      0.000
## 5      79084.10      0      0      1      0      5.098685      2918.856
## 6      149756.71      1      1      1      0      5.055977      2585.359
```

Calculate collinearity: Alternative 1

Using cor()

```
round(cor(colExitData), 2)
```

```
##      Age Tenure Balance NumOfProducts HasCrCard
## Age      1.00 -0.01  0.03      -0.03      -0.01
## Tenure    -0.01  1.00 -0.01       0.01       0.02
## Balance   0.03 -0.01  1.00      -0.30      -0.01
## NumOfProducts -0.03  0.01 -0.30       1.00       0.00
## HasCrCard -0.01  0.02 -0.01       0.00       1.00
## IsActiveMember 0.09 -0.03 -0.01       0.01      -0.01
## EstimatedSalary -0.01  0.01  0.01       0.01      -0.01
## Exited     0.29 -0.01  0.12      -0.05      -0.01
## female    -0.03  0.01  0.01      -0.02       0.01
## Germany   -0.05  0.00 -0.40       0.01      -0.01
## Spain      0.00  0.00  0.13      -0.01       0.01
## logBalance 0.03 -0.01  0.94      -0.33      -0.02
## wealthAcc -0.25 -0.01  0.93      -0.28      -0.01
##      IsActiveMember EstimatedSalary Exited female Germany Spain
## Age      0.09      -0.01  0.29 -0.03 -0.05  0.00
## Tenure   -0.03      0.01 -0.01  0.01  0.00  0.00
## Balance  -0.01      0.01  0.12  0.01 -0.40  0.13
## NumOfProducts 0.01      0.01 -0.05 -0.02  0.01 -0.01
## HasCrCard -0.01     -0.01 -0.01  0.01 -0.01  0.01
## IsActiveMember 1.00     -0.01 -0.16  0.02  0.02 -0.02
## EstimatedSalary -0.01      1.00  0.01 -0.01 -0.01  0.01
## Exited    -0.16      0.01  1.00 -0.11 -0.17  0.05
## female     0.02     -0.01 -0.11  1.00  0.02 -0.02
## Germany    0.02     -0.01 -0.17  0.02  1.00 -0.33
## Spain     -0.02      0.01  0.05 -0.02 -0.33  1.00
## logBalance 0.00      0.01  0.12  0.01 -0.44  0.15
## wealthAcc -0.02      0.01  0.02  0.02 -0.35  0.13
##      logBalance wealthAcc
## Age      0.03     -0.25
## Tenure   -0.01     -0.01
## Balance   0.94      0.93
## NumOfProducts -0.33     -0.28
## HasCrCard -0.02     -0.01
## IsActiveMember 0.00     -0.02
## EstimatedSalary 0.01      0.01
## Exited     0.12      0.02
## female     0.01      0.02
## Germany   -0.44     -0.35
```



```
## Spain          0.15      0.13
## logBalance     1.00      0.87
## wealthAcc      0.87      1.00
```

#wealthAcc has high collinearity with Balance (0.93) and logBalance (0.87). Understandable!

Calculate collinearity: Alternative 2

Using corpcor package

```
cor2pcor(cov(colExitData))
```

```
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000000 -0.010596217  5.393986e-01 -0.038456102  0.002767123
## [2,] -0.0105962264  1.0000000000  1.320113e-02  0.007582190  0.021918773
## [3,]  0.5393985863  0.013201127  1.000000e+00  0.038209396 -0.005734808
## [4,] -0.0384561023  0.007582190  3.820940e-02  1.0000000000 -0.006046996
## [5,]  0.0027671234  0.021918773 -5.734809e-03 -0.006046996  1.0000000000
## [6,]  0.1306974492 -0.029332504 -4.471046e-02  0.012264209 -0.011415124
## [7,] -0.0080509947  0.007935803 -3.701465e-05  0.019484691 -0.009905243
## [8,]  0.1774696036 -0.015517917  4.997432e-02 -0.024889284 -0.005288715
## [9,] -0.0042980625  0.013736093  1.969893e-02 -0.020655075  0.005360705
## [10,] 0.0126525454 -0.007493160 -5.078621e-04 -0.152980539 -0.018428780
## [11,] -0.0009135134 -0.004243426 -8.264777e-03 -0.005755951  0.010319045
## [12,]  0.0066923788 -0.008434607  5.722448e-01 -0.160853721 -0.015984403
## [13,] -0.6993040895 -0.013467296  7.796735e-01 -0.034936751  0.012485159
##           [,6]      [,7]      [,8]      [,9]      [,10]
## [1,]  0.130697449 -8.050995e-03  0.177469604 -0.004298063  0.0126525455
## [2,] -0.029332503  7.935802e-03 -0.015517915  0.013736093 -0.0074931596
## [3,] -0.044710460 -3.701461e-05  0.049974323  0.019698926 -0.0005078621
## [4,]  0.012264209  1.948469e-02 -0.024889284 -0.020655075 -0.1529805392
## [5,] -0.011415124 -9.905243e-03 -0.005288715  0.005360705 -0.0184287798
## [6,]  1.0000000000 -8.336570e-03 -0.183530111  0.006582948 -0.0034749408
## [7,] -0.008336570  1.000000e+00  0.010529674 -0.006575613  0.0010614128
## [8,] -0.183530111  1.052967e-02  1.0000000000 -0.099506531 -0.1222298844
## [9,]  0.006582948 -6.575613e-03 -0.099506531  1.0000000000  0.0086409050
## [10,] -0.003474941  1.061413e-03 -0.122229884  0.008640905  1.0000000000
## [11,] -0.010090117  3.259705e-03 -0.003695600 -0.009824189 -0.2966226756
## [12,]  0.017716657  8.921042e-03  0.001100671 -0.013377858 -0.1971020489
## [13,]  0.047135430 -1.808001e-03 -0.050179540 -0.008135059  0.0219679169
##           [,11]      [,12]      [,13]
## [1,] -0.0009135134  0.0066923789 -0.6993040895
## [2,] -0.0042434258 -0.0084346066 -0.0134673030
## [3,] -0.0082647766  0.5722448017  0.7796734774
## [4,] -0.0057559514 -0.1608537209 -0.0349367507
## [5,]  0.0103190449 -0.0159844034  0.0124851593
## [6,] -0.0100901168  0.0177166571  0.0471354297
## [7,]  0.0032597055  0.0089210423 -0.0018080012
## [8,] -0.0036956000  0.0011006708 -0.0501795400
## [9,] -0.0098241889 -0.0133778577 -0.0081350586
## [10,] -0.2966226756 -0.1971020489  0.0219679168
## [11,]  1.0000000000 -0.0006985821  0.0112791586
## [12,] -0.0006985821  1.0000000000  0.0007911062
## [13,]  0.0112791586  0.0007911063  1.0000000000
```

Remove logBalance from the model due to collinearity with wealthAcc.

```
geodem.glmF <- glm(Exited ~ CreditScore + Age + NumOfProducts + IsActiveMember + female + Germany + Tenure + wealthAcc, family = "binomial", data = exitData)
summary(geodem.glmF)
```

```
##
## Call:
## glm(formula = Exited ~ CreditScore + Age + NumOfProducts + IsActiveMember + female + Germany + Tenure + wealthAcc, family = "binomial", data = exitData)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3224  -0.6600  -0.4569  -0.2701   2.9883
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.599e+00  2.577e-01 -10.088 < 2e-16 ***
## CreditScore  -6.719e-04  2.801e-04  -2.399 0.016434 *
## Age           7.583e-02  2.750e-03  27.579 < 2e-16 ***
## NumOfProducts -1.210e-01  4.711e-02  -2.569 0.010187 *
## IsActiveMember -1.079e+00  5.766e-02 -18.708 < 2e-16 ***
## female       -5.263e-01  5.443e-02  -9.670 < 2e-16 ***
## Germany      -8.082e-01  6.293e-02 -12.843 < 2e-16 ***
## Tenure       -1.580e-02  9.341e-03  -1.692 0.090664 .
## wealthAcc      7.075e-05  1.946e-05   3.636 0.000277 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 10109.8  on 9999  degrees of freedom
## Residual deviance:  8576.6  on 9991  degrees of freedom
## AIC: 8594.6
##
## Number of Fisher Scoring iterations: 5
```

Calculate log of wealthAcc.

```
exitData$logWealthAcc <- log10((exitData$wealthAcc) + 1)
head(exitData)
```

```
##   RowNumber CustomerId Surname CreditScore Geography Gender Age Tenure
## 1         1  15634602 Hargrave         619    France Female  42      2
## 2         2  15647311 Hill         608    Spain Female  41      1
## 3         3  15619304 Onio         502    France Female  42      8
## 4         4  15701354 Boni         699    France Female  39      1
## 5         5  15737888 Mitchell         850    Spain Female  43      2
## 6         6  15574012 Chu         645    Spain Male  44      8
##      Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
## 1         0.00           1           1              1       101348.88      1
## 2    83807.86           1           0              1       112542.58      0
## 3   159660.80           3           1              0       113931.57      1
```

```
## 4      0.00      2      0      0      93826.63      0
## 5 125510.82      1      1      1      79084.10      0
## 6 113755.78      2      1      0     149756.71      1
##   female Germany Spain logBalance wealthAcc logWealthAcc
## 1      0      1      1  0.000000      0.000  0.000000
## 2      0      1      0  4.923290    2044.094   3.310713
## 3      0      1      1  5.203201    3801.448   3.580063
## 4      0      1      1  0.000000      0.000  0.000000
## 5      0      1      0  5.098685    2918.856   3.465361
## 6      1      1      0  5.055977    2585.359   3.412689
```

Replace wealthAcc with logWealthAcc in the model.

```
geodem.glmG <- glm(Exited ~ CreditScore + Age + NumOfProducts + IsActiveMember + female + Germany + Tenure, data = exitData, family = "binomial")
summary(geodem.glmG)
```

```
##
## Call:
## glm(formula = Exited ~ CreditScore + Age + NumOfProducts + IsActiveMember +
##      female + Germany + Tenure + logWealthAcc, family = "binomial",
##      data = exitData)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3098  -0.6589  -0.4551  -0.2676   2.9867
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.6543746  0.2503565 -10.602 < 2e-16 ***
## CreditScore  -0.0006752  0.0002802  -2.409  0.0160 *
## Age           0.0733182  0.0025818  28.398 < 2e-16 ***
## NumOfProducts -0.0967189  0.0475513  -2.034  0.0420 *
## IsActiveMember -1.0763060  0.0576443 -18.672 < 2e-16 ***
## female       -0.5265970  0.0544528  -9.671 < 2e-16 ***
## Germany      -0.7515239  0.0650669 -11.550 < 2e-16 ***
## Tenure       -0.0158549  0.0093454  -1.697  0.0898 .
## logWealthAcc  0.0985483  0.0204776   4.812 1.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 10109.8  on 9999  degrees of freedom
## Residual deviance:  8566.4  on 9991  degrees of freedom
## AIC: 8584.4
##
## Number of Fisher Scoring iterations: 5
## Coefficient looks better for logWealthAcc.
```

The final model is left with all the variables that are significant at the a-level of 0.05.