Pair 2 – Assignment 1: Predictive Modeling

1. **Based on the background information on Vancity and RRSP (you can do extra outside research on the topics if you feel needed), answer the following questions (15 points)**
   (a) What factors influence whether a Vancity's non-RRSP customer will acquire a new RRSP during the 2015-2016 campain?
   (b) Connect the factors with the available data, form 4~6 hypotheses that can be tested using your predicted model and provide brief rationale for each hypothesis.

   Based on the information given, we presume that BALSAV, avginc1, age, BALMRGG and BALLOAN are factors that would influence whether a Vancity's non-RRSP customer will acquire a new RRSP during the 2015-2016 campaign.

   From these factors, we developed the following hypotheses:
   1. There is a negative linear relationship between purchasing an RRSP (APURCH) and the average monthly balance in one's savings account (BALSAV). According to Chart 1-1 and Chart 1-2 by Statistics Canada, RRSP contributions decreased as TFSA contributions increased.
   2. There is a positive linear relationship between purchasing an RRSP (APURCH) and the average employment income in one's postal code household area (avginc1). According to Chart 4 by Statistics Canada, the quartiles with higher income have more instances of contributing to an RRSP or TFSA.
   3. There is a positive non-linear relationship between purchasing an RRSP and age. We're assuming this because according to Statistics Canada (2021), citizens who are middle age are the more likely to make contributions than younger demographics.
   4. There is a negative linear non-linear relationship between purchasing an RRSP (APURCH) and the average monthly mortgage balance in the past 12 months (BALMRGG) because the more money put towards mortgage payments, the less money available for purchasing an RRSP.
   5. There's a negative relationship between purchasing an RRSP (APURCH) and personal loan balance (BALLOAN) because the more money put towards paying off loans, the less money available for purchasing an RRSP.

References

Statistics Canada (2021). "Registered retirement savings plan contributions, 2019." *Statistics Canada*. https://www150.statcan.gc.ca/n1/daily-quotidien/210309/dq210309c-eng.htm

2. **Examine the variables contained in the dataset carefully and decide (a) which variable is the target variable, (b) which variables must be excluded from the predictor variables before building your predictive models, briefly explain why. (5 points) (Please note that the multicollinearity problem will be addressed in the next question, please don't consider the correlated variables in this question).**

   a) Our target variable is APURCH, which provides whether or not an individual purchases an RRSP. This is our target variable because the aim of our model is to determine which Vancity members are likely to purchase an RRSP.

   b) We will exclude the following variables:
   - Unique: the Vancity member identification number is assigned at random, and it doesn't influence one's willingness to purchase an RRSP.
   - Pcode: Postal code alone is not relevant for predicting RRSP purchases, and there are other variables relating to living area (i.e. numrr1, avginc1, avginv1) that are more relevant for predicting RRSP purchases.

3. **Examine the correlations among predictor variables ("cor" command in R) (5 points).**
   (a) Are there any variables that you need pay attention to when you interpret your model because of the multicollinearity problem? If yes, list the variables.
   The variables were selected based on threshold value of 0.8
   - DUMNOMRGG and BALMRGG
   - NINDINC1 and numrr1
   - gendf and gendm

   (b) Are there any variables that you want to exclude from predictor variables list because of the multicollinearity problem? If yes, name the variables.
   - DUMNOMRGG
   - NINDINC1
   - gendf

4. Build a reasonably good model using logit regression without any variable transformation. (15 points) (A) Briefly explain the process that you derive the good model, for example, how the predictor variables in your model are decided) (B) Document the outputs in the process described above (i.e., logit outputs, liftcharts, …)

To develop a good model using logit regression, we used the backwards approach.

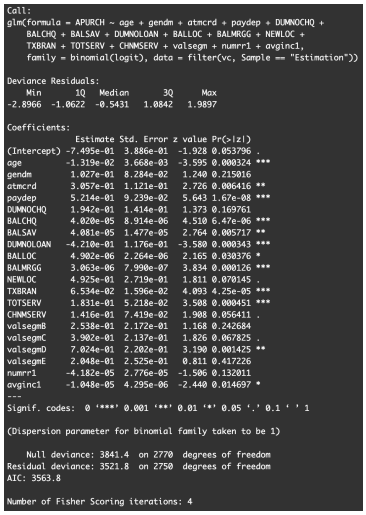
For our first logistic model (LM1), we included all the predictor variables excluding the three variables we excluded due to multicollinearity. We got the following output:

```
Call:
glm(formula = APURCH ~ age + gendm + atmcrd + paydep + DUMNOCHQ +
    BALCHQ + DUMNOSAV + BALSAV + TOTDEP + DUMNOLOAN + BALLOAN +
    DUMNOLOC + BALLOC + BALMRGG + NEWLOC + NEWMRGG + TXBRAN +
    TXATM + TXPOS + TXCHQ + TXWEB + TXTEL + TOTSERV + CHNMSERV +
    CHNMPRD + valsegm + numrr1 + avginc1 + avginv1, family = binomial(logit),
    data = filter(vc, Sample == "Estimation"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0040  -1.0599  -0.5449   1.0773   1.9918

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.247e+00  6.056e-01  -2.060 0.039393 *
age         -1.479e-02  3.832e-03  -3.861 0.000113 ***
gendm        1.103e-01  8.346e-02   1.322 0.186180
atmcrd       3.469e-01  1.169e-01   2.968 0.002999 **
paydep       5.572e-01  9.792e-02   5.690 1.27e-08 ***
DUMNOCHQ     2.339e-01  1.658e-01   1.411 0.158196
BALCHQ       3.391e-05  9.509e-06   3.566 0.000363 ***
DUMNOSAV     1.083e-01  1.260e-01   0.859 0.390158
BALSAV       3.996e-05  1.548e-05   2.581 0.009849 **
TOTDEP       1.880e-06  2.016e-06   0.932 0.351240
DUMNOLOAN   -3.926e-01  1.627e-01  -2.413 0.015809 *
BALLOAN      8.386e-07  1.134e-05   0.074 0.941048
DUMNOLOC     1.636e-01  1.331e-01   1.229 0.219149
BALLOC       5.622e-06  2.388e-06   2.354 0.018571 *
BALMRGG      3.152e-06  9.082e-07   3.471 0.000519 ***
NEWLOC       5.692e-01  2.814e-01   2.023 0.043087 *
NEWMRGG     -1.640e-01  3.401e-01  -0.482 0.629622
TXBRAN       6.417e-02  1.659e-02   3.868 0.000110 ***
TXATM       -2.213e-04  1.103e-02  -0.020 0.983990
TXPOS       -2.848e-03  3.964e-03  -0.719 0.472409
TXCHQ        1.393e-02  1.613e-02   0.864 0.387629
TXWEB       -1.426e-02  1.860e-02  -0.767 0.443293
TXTEL       -1.497e+00  1.153e+00  -1.299 0.193932
TOTSERV      2.670e-01  8.876e-02   3.008 0.002626 **
CHNMSERV     1.352e-01  9.299e-02   1.454 0.145872
CHNMPRD      8.230e-03  3.965e-02   0.208 0.835549
valsegmB     2.546e-01  2.208e-01   1.153 0.248795
valsegmC     4.072e-01  2.216e-01   1.837 0.066179 .
valsegmD     7.755e-01  2.392e-01   3.241 0.001189 **
valsegmE     2.920e-01  2.733e-01   1.068 0.285387
numrr1      -3.918e-05  2.785e-05  -1.407 0.159389
avginc1     -9.296e-06  6.000e-06  -1.549 0.121269
avginv1     -4.464e-06  1.891e-05  -0.236 0.813364
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3841.4  on 2770  degrees of freedom
Residual deviance: 3514.5  on 2738  degrees of freedom
AIC: 3580.5

Number of Fisher Scoring iterations: 4
```

For our second logistic model (LM2), we removed predictor variables that had a z-value greater than 0.5 as these higher values indicate statistical insignificance. The AIC reduced from 3580.5 to 3570.9 for our second model.We got the following output:

```
Call:
glm(formula = APURCH ~ age + gendm + atmcrd + paydep + DUMNOCHQ +
    BALCHQ + DUMNOSAV + BALSAV + TOTDEP + DUMNOLOAN + DUMNOLOC +
    BALLOC + BALMRGG + NEWLOC + TXBRAN + TXPOS + TXCHQ + TXWEB +
    TXTEL + TOTSERV + CHNMSERV + valsegm + numrr1 + avginc1,
    family = binomial(logit), data = filter(vc, Sample == "Estimation"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9756  -1.0592  -0.5415   1.0771   1.9980

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.186e+00  5.724e-01  -2.072 0.038257 *
age         -1.480e-02  3.818e-03  -3.876 0.000106 ***
gendm        1.105e-01  8.324e-02   1.328 0.184187
atmcrd       3.457e-01  1.151e-01   3.003 0.002673 **
paydep       5.573e-01  9.729e-02   5.728 1.01e-08 ***
DUMNOCHQ     2.336e-01  1.653e-01   1.413 0.157739
BALCHQ       3.389e-05  9.484e-06   3.573 0.000353 ***
DUMNOSAV     1.089e-01  1.259e-01   0.865 0.387098
BALSAV       3.968e-05  1.546e-05   2.567 0.010258 *
TOTDEP       1.702e-06  1.982e-06   0.859 0.390581
DUMNOLOAN   -3.967e-01  1.268e-01  -3.130 0.001750 **
DUMNOLOC     1.632e-01  1.329e-01   1.228 0.219262
BALLOC       5.484e-06  2.352e-06   2.332 0.019724 *
BALMRGG      2.993e-06  8.368e-07   3.576 0.000348 ***
NEWLOC       5.610e-01  2.810e-01   1.996 0.045908 *
TXBRAN       6.410e-02  1.655e-02   3.873 0.000108 ***
TXPOS       -2.878e-03  3.602e-03  -0.799 0.424336
TXCHQ        1.388e-02  1.583e-02   0.877 0.380527
TXWEB       -1.382e-02  1.855e-02  -0.745 0.456159
TXTEL       -1.475e+00  1.151e+00  -1.281 0.200027
TOTSERV      2.652e-01  8.858e-02   2.994 0.002752 **
CHNMSERV     1.407e-01  7.468e-02   1.884 0.059626 .
valsegmB     2.446e-01  2.176e-01   1.124 0.260907
valsegmC     3.826e-01  2.143e-01   1.786 0.074143 .
valsegmD     7.391e-01  2.224e-01   3.323 0.000891 ***
valsegmE     2.528e-01  2.563e-01   0.986 0.324050
numrr1      -3.865e-05  2.780e-05  -1.390 0.164442
avginc1     -1.029e-05  4.332e-06  -2.376 0.017485 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3841.4  on 2770  degrees of freedom
Residual deviance: 3514.9  on 2743  degrees of freedom
AIC: 3570.9
```

For our third to fifth logistic models, we removed the variable with the highest z-value from the previous model's output. We proceeded to create a new model with one less variable because the AICs of each model were lower than their respective preceding model. These were the outputs for the three models:



```
Call:
glm(formula = APURCH ~ age + gendm + atmcrd + paydep + DUMNOCHQ +
    BALCHQ + BALSAV + DUMNOLOAN + BALLOC + BALMRGG + NEWLOC +
    TXBRAN + TOTSERV + CHNMSERV + valsegm + numrr1 + avginc1,
    family = binomial(logit), data = filter(vc, Sample == "Estimation"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8966  -1.0622  -0.5431   1.0842   1.9897

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.495e-01  3.886e-01  -1.928 0.053796 .
age         -1.319e-02  3.668e-03  -3.595 0.000324 ***
gendm        1.027e-01  8.284e-02   1.240 0.215016
atmcrd       3.057e-01  1.121e-01   2.726 0.006416 **
paydep       5.214e-01  9.239e-02   5.643 1.67e-08 ***
DUMNOCHQ     1.942e-01  1.414e-01   1.373 0.169761
BALCHQ       4.020e-05  8.914e-06   4.510 6.47e-06 ***
BALSAV       4.081e-05  1.477e-05   2.764 0.005717 **
DUMNOLOAN   -4.210e-01  1.176e-01  -3.580 0.000343 ***
BALLOC       4.902e-06  2.264e-06   2.165 0.030376 *
BALMRGG      3.063e-06  7.990e-07   3.834 0.000126 ***
NEWLOC       4.925e-01  2.719e-01   1.811 0.070145 .
TXBRAN       6.534e-02  1.596e-02   4.093 4.25e-05 ***
TOTSERV      1.831e-01  5.218e-02   3.508 0.000451 ***
CHNMSERV     1.416e-01  7.419e-02   1.908 0.056411 .
valsegmB     2.538e-01  2.172e-01   1.168 0.242684
valsegmC     3.902e-01  2.137e-01   1.826 0.067825 .
valsegmD     7.024e-01  2.202e-01   3.190 0.001425 **
valsegmE     2.048e-01  2.525e-01   0.811 0.417226
numrr1      -4.182e-05  2.776e-05  -1.506 0.132011
avginc1     -1.048e-05  4.295e-06  -2.440 0.014697 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3841.4  on 2770  degrees of freedom
Residual deviance: 3521.8  on 2750  degrees of freedom
AIC: 3563.8

Number of Fisher Scoring iterations: 4
```

```
Call:
glm(formula = APURCH ~ age + atmcrd + paydep + DUMNOCHQ + BALCHQ +
    BALSAV + DUMNOLOAN + BALLOC + BALMRGG + NEWLOC + TXBRAN +
    TOTSERV + CHNMSERV + valsegm + numrr1 + avginc1, family = binomial(logit),
    data = filter(vc, Sample == "Estimation"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9033  -1.0618  -0.5472   1.0855   2.0097

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.890e-01  3.854e-01  -1.788 0.073837 .
age         -1.301e-02  3.664e-03  -3.551 0.000383 ***
atmcrd       3.064e-01  1.121e-01   2.733 0.006270 **
paydep       5.173e-01  9.230e-02   5.604 2.09e-08 ***
DUMNOCHQ     1.848e-01  1.411e-01   1.309 0.190483
BALCHQ       4.049e-05  8.917e-06   4.541 5.61e-06 ***
BALSAV       4.054e-05  1.471e-05   2.756 0.005859 **
DUMNOLOAN   -4.257e-01  1.175e-01  -3.623 0.000291 ***
BALLOC       4.989e-06  2.268e-06   2.200 0.027790 *
BALMRGG      3.156e-06  7.958e-07   3.966 7.30e-05 ***
NEWLOC       4.955e-01  2.719e-01   1.822 0.068428 .
TXBRAN       6.568e-02  1.597e-02   4.113 3.90e-05 ***
TOTSERV      1.823e-01  5.216e-02   3.494 0.000475 ***
CHNMSERV     1.405e-01  7.413e-02   1.895 0.058036 .
valsegmB     2.593e-01  2.171e-01   1.194 0.232406
valsegmC     3.956e-01  2.136e-01   1.852 0.064062 .
valsegmD     6.993e-01  2.202e-01   3.175 0.001496 **
valsegmE     2.058e-01  2.525e-01   0.815 0.415136
numrr1      -4.108e-05  2.777e-05  -1.479 0.139079
avginc1     -1.074e-05  4.290e-06  -2.503 0.012313 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3841.4  on 2770  degrees of freedom
Residual deviance: 3523.3  on 2751  degrees of freedom
AIC: 3563.3

Number of Fisher Scoring iterations: 4
```

```
Call:
glm(formula = APURCH ~ age + atmcrd + paydep + BALCHQ + BALSAV +
    DUMNOLOAN + BALLOC + BALMRGG + NEWLOC + TXBRAN + TOTSERV +
    CHNMSERV + valsegm + numrr1 + avginc1, family = binomial(logit),
    data = filter(vc, Sample == "Estimation"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8885  -1.0647  -0.5444   1.0842   1.9790

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.342e-01  3.663e-01  -1.458 0.144757
age         -1.336e-02  3.655e-03  -3.655 0.000257 ***
atmcrd       2.818e-01  1.105e-01   2.551 0.010742 *
paydep       5.049e-01  9.177e-02   5.502 3.75e-08 ***
BALCHQ       3.884e-05  8.764e-06   4.432 9.34e-06 ***
BALSAV       4.403e-05  1.478e-05   2.980 0.002885 **
DUMNOLOAN   -4.413e-01  1.169e-01  -3.775 0.000160 ***
BALLOC       4.904e-06  2.264e-06   2.166 0.030323 *
BALMRGG      3.176e-06  7.950e-07   3.995 6.48e-05 ***
NEWLOC       4.837e-01  2.714e-01   1.782 0.074756 .
TXBRAN       6.431e-02  1.591e-02   4.043 5.28e-05 ***
TOTSERV      1.616e-01  4.962e-02   3.257 0.001128 **
CHNMSERV     1.398e-01  7.404e-02   1.888 0.059020 .
valsegmB     2.594e-01  2.169e-01   1.196 0.231649
valsegmC     3.958e-01  2.134e-01   1.855 0.063633 .
valsegmD     6.957e-01  2.199e-01   3.163 0.001560 **
valsegmE     2.069e-01  2.522e-01   0.820 0.412108
numrr1      -4.034e-05  2.774e-05  -1.454 0.145960
avginc1     -1.086e-05  4.288e-06  -2.532 0.011326 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3841.4  on 2770  degrees of freedom
Residual deviance: 3525.1  on 2752  degrees of freedom
AIC: 3563.1

Number of Fisher Scoring iterations: 4
```

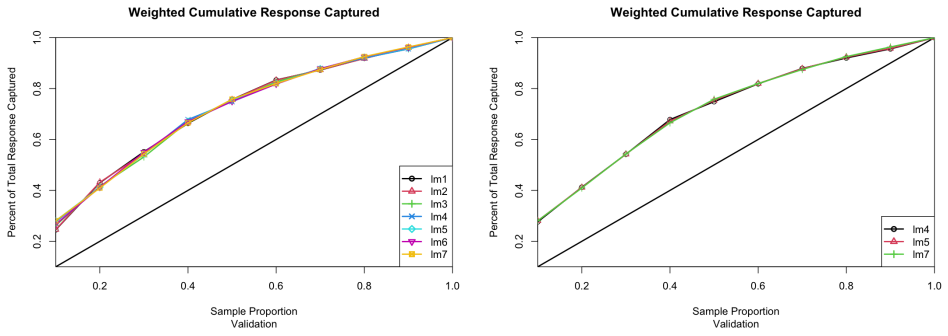
For our sixth logistic model (LM6), we removed the variable with the highest z-value from the previous model's output: 'numrr1.' We noticed the AIC for LM6 was 0.1 higher than LM5 at 3563.2 as shown:

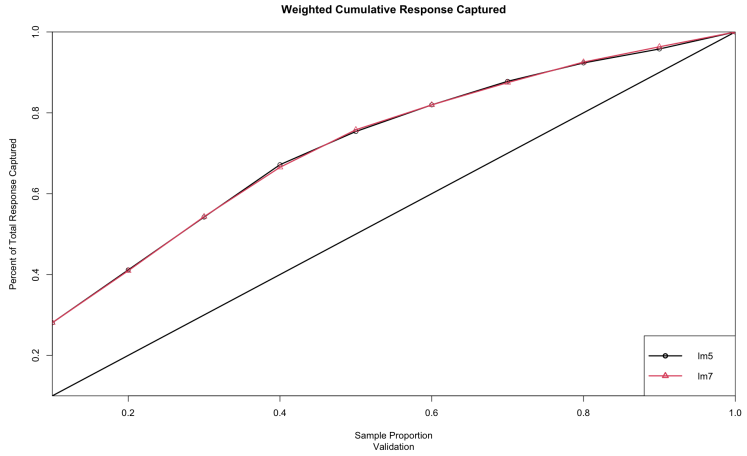
```
Call:
glm(formula = APURCH ~ age + atmcrd + paydep + BALCHQ + BALSAV +
    DUMNOLOAN + BALLOC + BALMRGG + NEWLOC + TXBRAN + TOTSERV +
    CHNMSERV + valsegm + avginc1, family = binomial(logit), data = filter(vc,
    Sample == "Estimation"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8761  -1.0647  -0.5579   1.0835   1.9853

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.506e-01  3.660e-01  -1.504 0.132484
age         -1.346e-02  3.653e-03  -3.685 0.000229 ***
atmcrd       2.861e-01  1.104e-01   2.592 0.009530 **
paydep       5.034e-01  9.173e-02   5.488 4.08e-08 ***
BALCHQ       3.901e-05  8.770e-06   4.448 8.65e-06 ***
BALSAV       4.412e-05  1.476e-05   2.989 0.002799 **
DUMNOLOAN   -4.386e-01  1.168e-01  -3.755 0.000173 ***
BALLOC       4.872e-06  2.259e-06   2.157 0.031043 *
BALMRGG      3.151e-06  7.937e-07   3.970 7.20e-05 ***
NEWLOC       4.897e-01  2.714e-01   1.804 0.071187 .
TXBRAN       6.431e-02  1.589e-02   4.046 5.21e-05 ***
TOTSERV      1.603e-01  4.957e-02   3.234 0.001219 **
CHNMSERV     1.388e-01  7.400e-02   1.876 0.060653 .
valsegmB     2.555e-01  2.168e-01   1.179 0.238469
valsegmC     3.928e-01  2.133e-01   1.842 0.065475 .
valsegmD     6.932e-01  2.198e-01   3.154 0.001610 **
valsegmE     2.002e-01  2.520e-01   0.794 0.426958
avginc1     -1.090e-05  4.292e-06  -2.541 0.011059 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3841.4  on 2770  degrees of freedom
Residual deviance: 3527.2  on 2753  degrees of freedom
AIC: 3563.2

Number of Fisher Scoring iterations: 4
```

To confirm that the model is fitting worse, for our seventh model (LM7), we again removed the variable with the highest z-value from the previous model's output: 'NEWLOC.' The AIC for LM7 was even higher than LM6 at 3564.6 as shown:

```
Call:
glm(formula = APURCH ~ age + atmcrd + paydep + BALCHQ + BALSAV +
    DUMNOLOAN + BALLOC + BALMRGG + TXBRAN + TOTSERV + CHNMSERV +
    valsegm + avginc1, family = binomial(logit), data = filter(vc,
    Sample == "Estimation"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9067  -1.0679  -0.5594   1.0825   1.9861

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.519e-01  3.658e-01  -1.509 0.131283
age         -1.357e-02  3.651e-03  -3.717 0.000201 ***
atmcrd       2.859e-01  1.104e-01   2.590 0.009593 **
paydep       5.059e-01  9.165e-02   5.520 3.38e-08 ***
BALCHQ       3.959e-05  8.774e-06   4.512 6.43e-06 ***
BALSAV       4.375e-05  1.473e-05   2.970 0.002978 **
DUMNOLOAN   -4.542e-01  1.165e-01  -3.899 9.67e-05 ***
BALLOC       4.998e-06  2.273e-06   2.199 0.027879 *
BALMRGG      3.237e-06  7.916e-07   4.089 4.33e-05 ***
TXBRAN       6.526e-02  1.590e-02   4.104 4.06e-05 ***
TOTSERV      1.636e-01  4.955e-02   3.303 0.000957 ***
CHNMSERV     1.745e-01  7.143e-02   2.442 0.014587 *
valsegmB     2.560e-01  2.167e-01   1.181 0.237457
valsegmC     3.938e-01  2.131e-01   1.848 0.064596 .
valsegmD     7.072e-01  2.194e-01   3.224 0.001266 **
valsegmE     2.109e-01  2.516e-01   0.838 0.401883
avginc1     -1.078e-05  4.286e-06  -2.516 0.011873 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3841.4  on 2770  degrees of freedom
Residual deviance: 3530.6  on 2754  degrees of freedom
AIC: 3564.6

Number of Fisher Scoring iterations: 4
```

Next, we created a few lift charts (as shown below) to compare the linear models. We initially included all the models, but gradually removed models that had a lower initial lift. We got the following lift charts:





Based on our lift charts, LM5 and LM7 have the highest initial lift that seem identical. Out of all our models, LM5 has the lowest AIC at 3563.1; whereas LM7 has an AIC of 3564.6. **Since LM5 has a lower AIC (3563.1), we chose it as our best fit model without any transformation for now.**

Our best fit logit regression model has the following predictor variables:

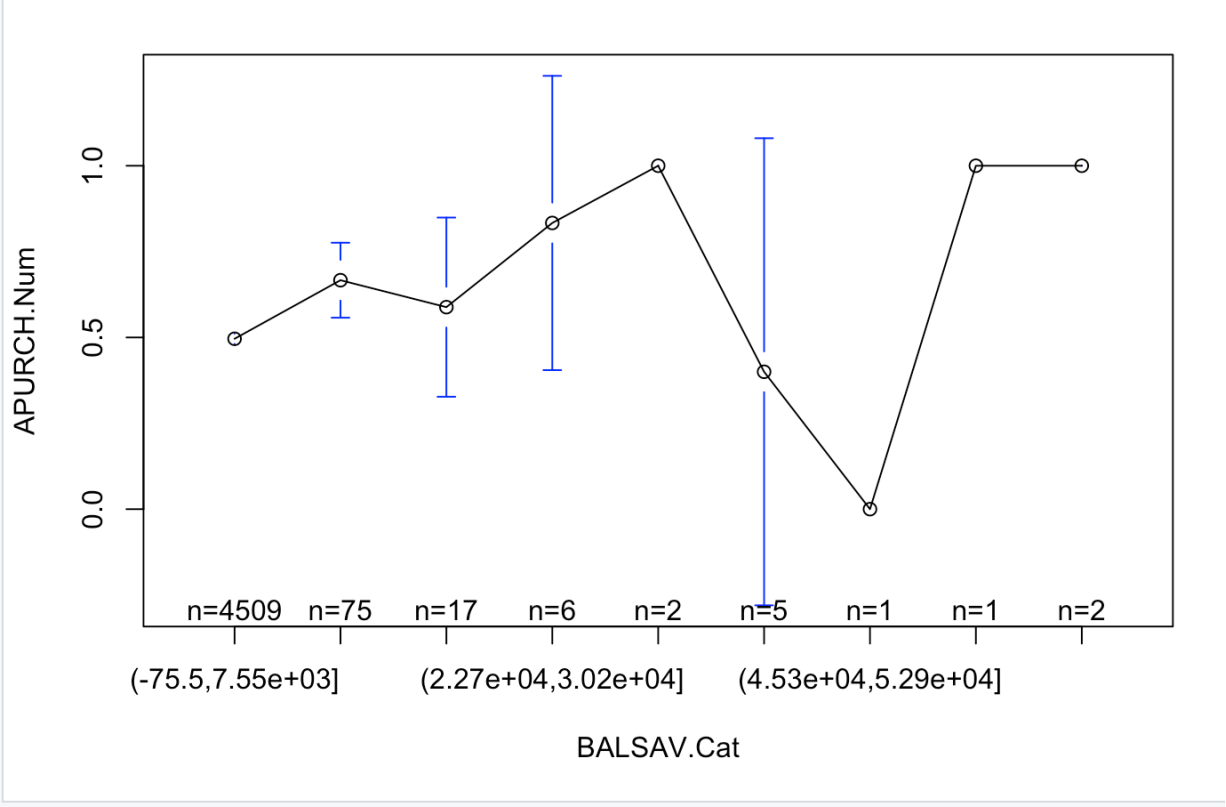
- age
- atmcrd
- paydep
- DUMNOCHQ
- BALCHQ
- BALSAV
- DUMNOLOAN
- BALLOC
- BALMRGG
- NEWLOC
- TXBRAN
- TOTSERV
- CHNMSERV
- valsegm
- numrr1
- avginc1

5. **Find the non-linear relationships between target variable and predictor variables. (10 points)**

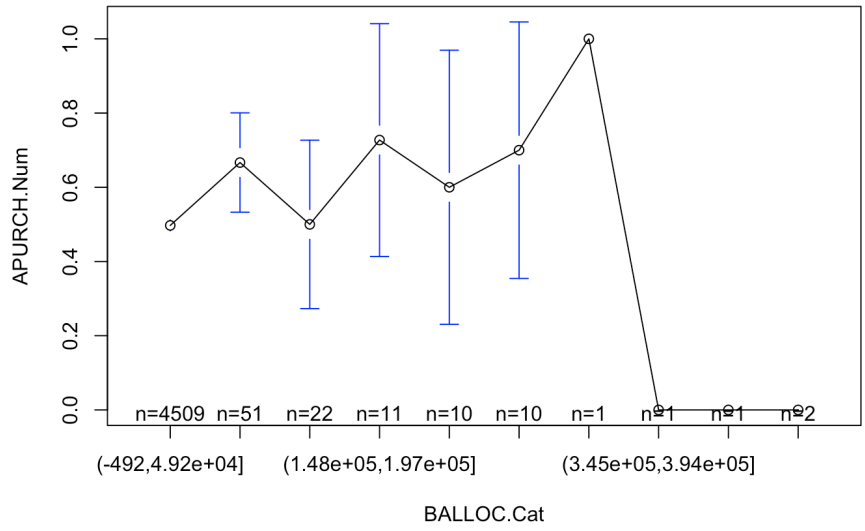
(a) Visually examine whether non-linear relationships exist between the target variable and predictor variables. (Hints: examine only those SIGNIFICANT CONTINUOUS predictor variables you have found in your BEST model in Question 4; use plot of means after binning a numeric variable).

BALCHQ,BALSAV,BALLOC,BALMRGG,TXBRAN,TOTSERV,avginc1 were the significant continuous predictor variables found in our best model.
Plot of means with binning works best when the data is fairly evenly distributed for predictor variables across its range of values but in case of BALSAV ,BALLOC, BALMRGG and TOTSERV the data is not evenly distributed and results in a large skew. However we were still able to draw useful conclusions with the help of Interval method of Plot of means which confirms non-linear relationship as shown in figures below .There is large skew e.g (A large number of customers have a low value, while a small number have a high value) in the predictor of interest)

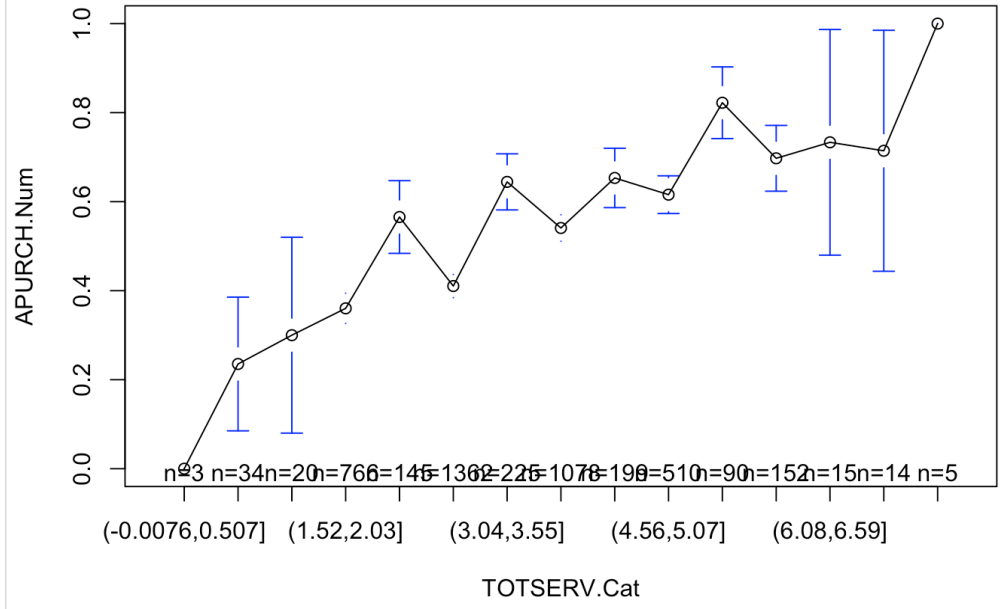
Non-linear(BALSAV)



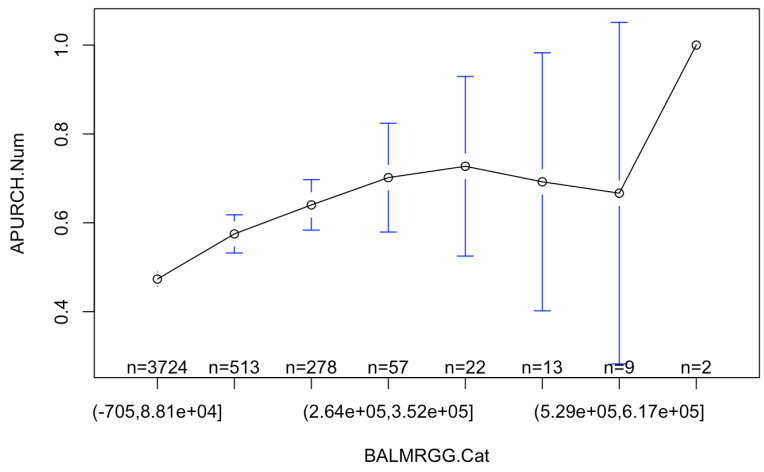
Non-Linear(BALLOC)



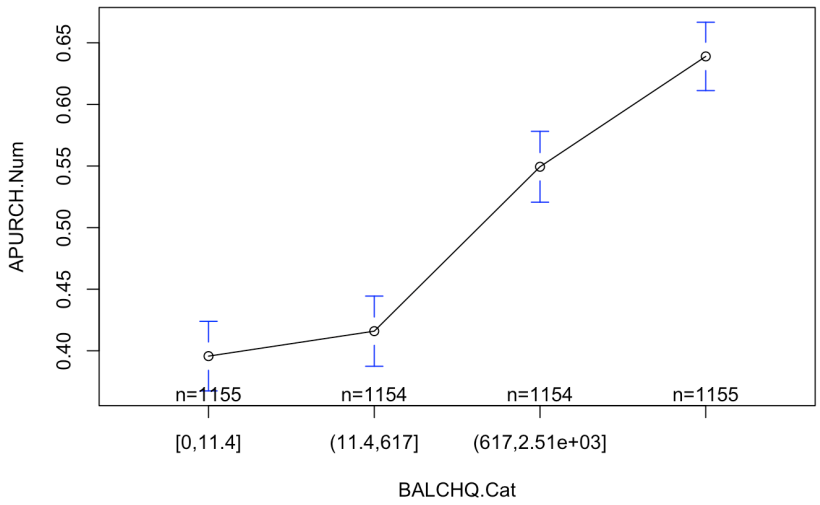
Non-linear(TOTSERV)

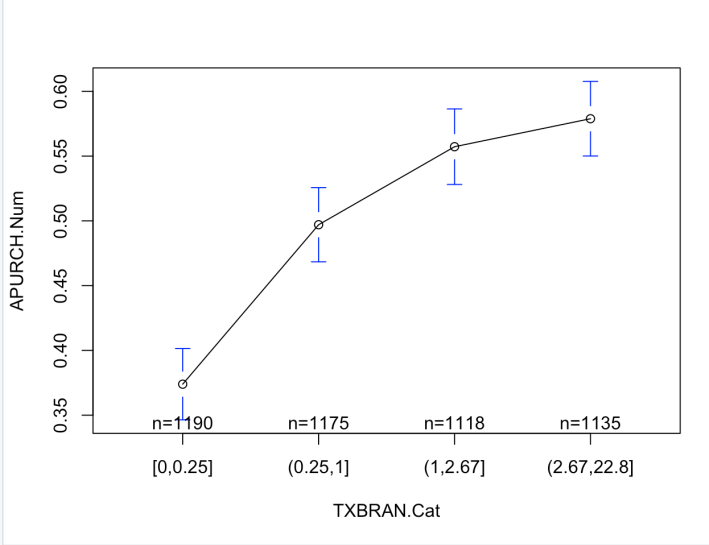


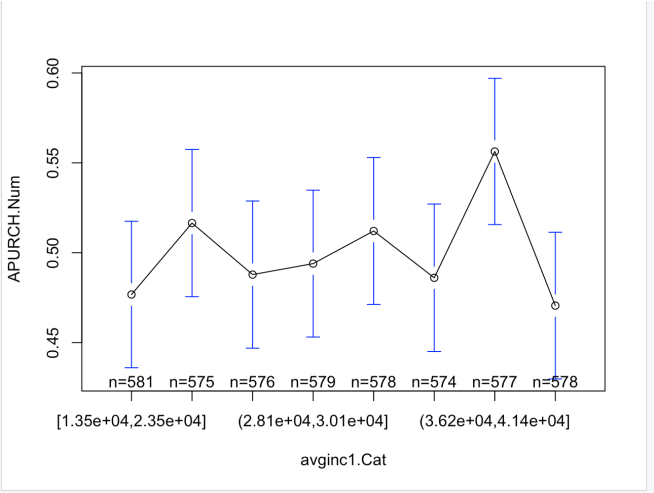
Nonlinear(BALMRGG)



However, incase of BALCHQ, TXBRAN and avginc1 the data was fairly evenly distributed. The Plot of means show non-linear relationship between target variable APURCH and three predictor variables (BALCHQ, TXBRAN and avginc1).







(b) If there is a non-linear relationship, describe the type of non-linear relationship (i.e., concave, convex, U-shape) and propose a proper transformation.

- APURCH and BALCHQ(Concaveup,Increasing,non-linear)
- APURCH and TXBRAN(Concavedown,Increasing,non-linear)
- APURCH and avginc1(Zigzag,non-linear)
- APURCH and BALSAV(Zigzag shaped,non-linear)
- APURCH and TORTSEV(Zigzag shaped,non-linear)
- APURCH and BALMRGG(Concave down initially but then changes to zigzag shape,non-linear)
- APURCH and BALLOC(Zigzag shaped,non-linear)

We propose LogTransformation,SquarerootTransformation as both of these can help in reducing the influence of extreme values,stabilize the variance and can improve the skewness as well. However, we will also be checking the ReciprocalTransformation and SquareTransformation to see if we can find any interesting results

6. Find the best overall model. (15 points)

(a) Build several non-linear models with the proper transformations proposed in Question 5,

Log Transformation

```
lm5_log <- glm(formula = APURCH ~ age + atmcrd + paydep + Log.BALCHQ + BALSAV +
          DUMNOLOAN + BALLOC + BALMRGG + NEWLOC + Log.TXBRAN +
TOTSERV + CHNMSERV +
           valsegm + numrr1 Log.avginc1,
         data = filter(vc, Sample == "Estimation"),
         family = binomial(logit))
```

SquarerootTransformation:

```
vc$sqrt.avginc1=sqrt(vc$avginc1)
vc$sqrt.TXBRAN=sqrt(vc$TXBRAN)
vc$sqrt.BALCHQ=sqrt(vc$BALCHQ)
vc$sqrt.BALSAV=sqrt(vc$BALSAV)
vc$sqrt.BALLOC=sqrt(vc$BALLOC)

lm5_sqrt <- glm(formula = APURCH ~ age + atmcrd + paydep + sqrt.BALCHQ + sqrt.BALSAV +
           DUMNOLOAN + sqrt.BALLOC + BALMRGG + NEWLOC+ sqrt.TXBRAN +
TOTSERV + CHNMSERV +
           valsegm + numrr1 + sqrt.avginc1,
         data = filter(vc, Sample == "Estimation"),
         family = binomial(logit))
```

SquareTransformation:

```
lm5_sqr <- glm(formula = APURCH ~ age + atmcrd + paydep + I(BALCHQ^2) + BALSAV +
           DUMNOLOAN + BALLOC + BALMRGG + NEWLOC + I(TXBRAN^2) + TOTSERV
+ CHNMSERV +
           valsegm + numrr1 I(avginc1^2),
         data = filter(vc, Sample == "Estimation"),
         family = binomial(logit))
```

ReciprocalTransformation:

```
lm5_rec <- glm(formula = APURCH ~ age + atmcrd + paydep + 1/BALCHQ + BALSAV +
           DUMNOLOAN + BALLOC + BALMRGG + NEWLOC+ 1/TXBRAN + TOTSERV +
CHNMSERV +
           valsegm + numrr1+ 1/avginc1,
         data = filter(vc, Sample == "Estimation"),
         family = binomial(logit))
```

(b) **Compare your non-linear models and your best model without variable transformations in Question 4.**

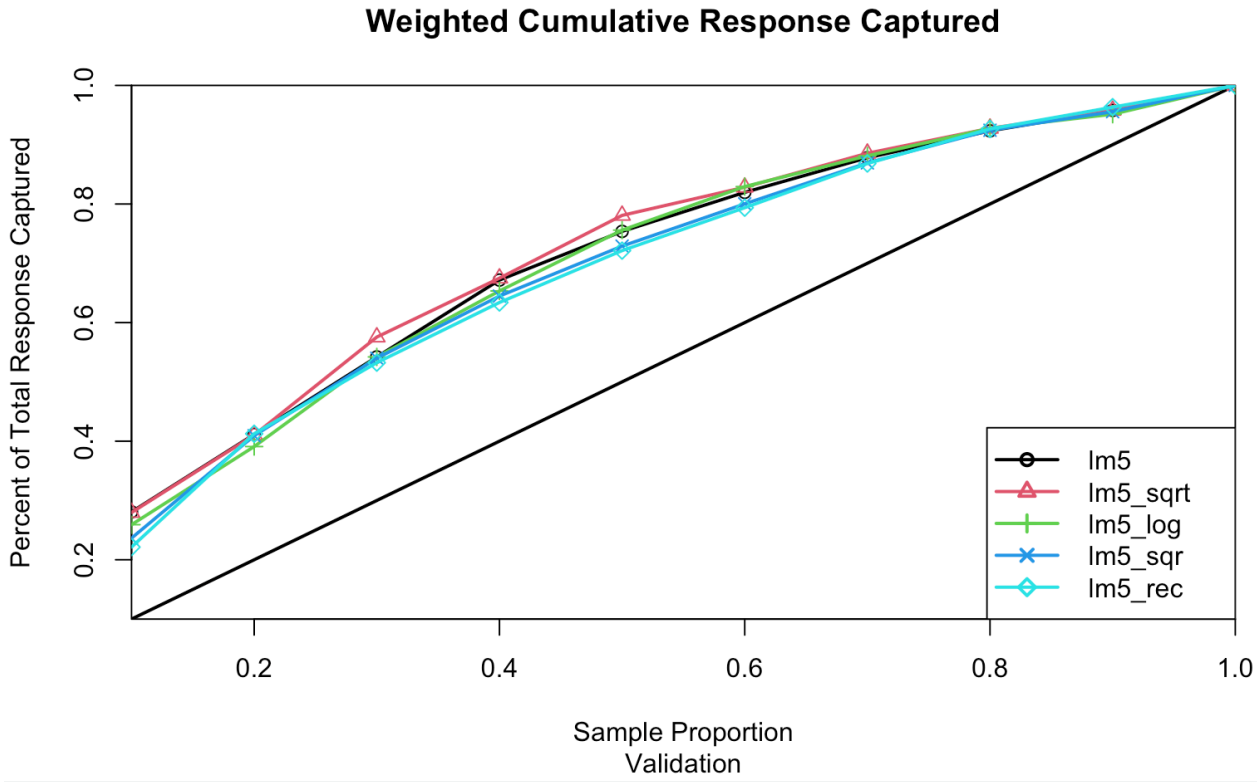
Our best model LM5(without transformation) had an AIC value of 3563.1 but using the squarerootTransformation on LM5 gives an AIC value of 3528.5 and is better than the model without transformation.

(a) Decide your final model using liftcharts on the Validation Sample. Please document your process with R outputs.

We are planning to go ahead with the model lm5_sqrt with square root transformation as our final model with an AIC value of 3528.5.

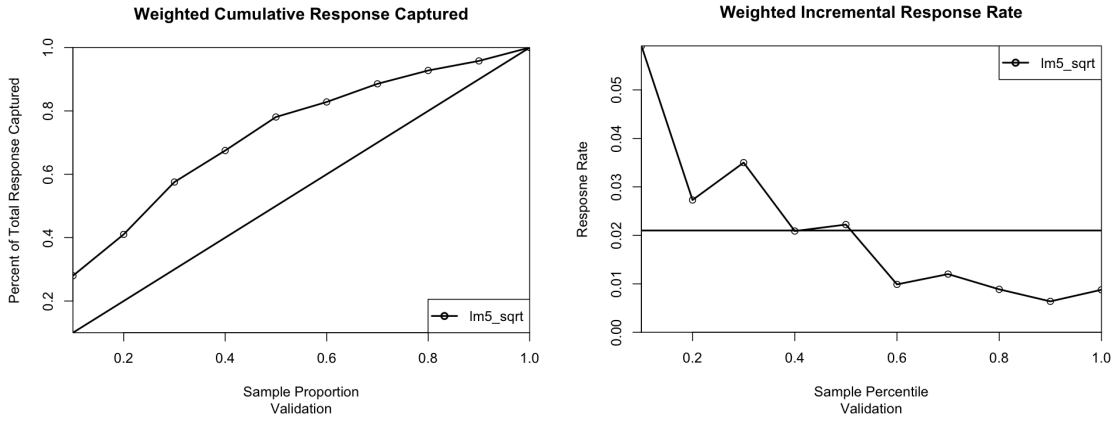
```
vc$sqrt.avginc1=sqrt(vc$avginc1)
vc$sqrt.TXBRAN=sqrt(vc$TXBRAN)
vc$sqrt.BALCHQ=sqrt(vc$BALCHQ)
vc$sqrt.BALSAV=sqrt(vc$BALSAV)
vc$sqrt.BALLOC=sqrt(vc$BALLOC)

lm5_sqrt <- glm(formula = APURCH ~ age + atmcrd + paydep + sqrt.BALCHQ + sqrt.BALSAV +
          DUMNOLOAN + sqrt.BALLOC + BALMRGG + NEWLOC+ sqrt.TXBRAN +
TOTSERV + CHNMSERV +
           valsegm + numrr1 + sqrt.avginc1,
         data = filter(vc, Sample == "Estimation"),
         family = binomial(logit))
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.600e-01  4.420e-01  -0.362 0.717408
age          -1.440e-02  3.700e-03  -3.893 9.90e-05 ***
atmcrd        2.217e-01  1.119e-01   1.982 0.047529 *
paydep        4.482e-01  9.310e-02   4.814 1.48e-06 ***
sqrt.BALCHQ   8.360e-03  1.297e-03   6.445 1.16e-10 ***
sqrt.BALSAV   7.898e-03  1.725e-03   4.578 4.69e-06 ***
DUMNOLOAN    -5.254e-01  1.190e-01  -4.414 1.01e-05 ***
sqrt.BALLOC   2.899e-03  8.396e-04   3.453 0.000554 ***
BALMRGG       3.350e-06  8.040e-07   4.167 3.08e-05 ***
NEWLOC        3.674e-01  2.742e-01   1.340 0.180283
sqrt.TXBRAN   2.073e-01  5.234e-02   3.960 7.49e-05 ***
TOTSERV       1.003e-01  5.099e-02   1.967 0.049199 *
CHNMSERV      1.236e-01  7.456e-02   1.657 0.097420 .
valsegmB      2.913e-01  2.177e-01   1.338 0.180965
valsegmC      4.365e-01  2.139e-01   2.040 0.041320 *
valsegmD      7.307e-01  2.208e-01   3.309 0.000937 ***
valsegmE      3.918e-01  2.566e-01   1.527 0.126729
numrr1       -4.009e-05  2.806e-05  -1.429 0.153058
sqrt.avginc1 -4.535e-03  1.700e-03  -2.668 0.007640 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3841.4  on 2770  degrees of freedom
Residual deviance: 3490.5  on 2752  degrees of freedom
AIC: 3528.5

Number of Fisher Scoring iterations: 4
```



Based on the lift chart, lm5_sqrt (model with square root transformation) looks the best.

(b) Create your final model's CUMULATIVE lift chart and INCREMENTAL lift chart on the validation sample. Copy and paste the lift charts in your document.



7. Use PLAIN language to interpret the coefficients output of your final model in Question 6. For example, who are more likely to be a new RRSP purchaser customer? (10 points)

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.600e-01  4.420e-01  -0.362 0.717408
age          -1.440e-02  3.700e-03  -3.893 9.90e-05 ***
atmcrd        2.217e-01  1.119e-01   1.982 0.047529 *
paydep        4.482e-01  9.310e-02   4.814 1.48e-06 ***
sqrt.BALCHQ   8.360e-03  1.297e-03   6.445 1.16e-10 ***
sqrt.BALSAV   7.898e-03  1.725e-03   4.578 4.69e-06 ***
DUMNOLOAN    -5.254e-01  1.190e-01  -4.414 1.01e-05 ***
sqrt.BALLOC   2.899e-03  8.396e-04   3.453 0.000554 ***
BALMRGG       3.350e-06  8.040e-07   4.167 3.08e-05 ***
NEWLOC        3.674e-01  2.742e-01   1.340 0.180283
sqrt.TXBRAN   2.073e-01  5.234e-02   3.960 7.49e-05 ***
TOTSERV       1.003e-01  5.099e-02   1.967 0.049199 *
CHNMSERV      1.236e-01  7.456e-02   1.657 0.097420 .
valsegmB      2.913e-01  2.177e-01   1.338 0.180965
valsegmC      4.365e-01  2.139e-01   2.040 0.041320 *
valsegmD      7.307e-01  2.208e-01   3.309 0.000937 ***
valsegmE      3.918e-01  2.566e-01   1.527 0.126729
numrr1       -4.009e-05  2.806e-05  -1.429 0.153058
sqrt.avginc1 -4.535e-03  1.700e-03  -2.668 0.007640 **
```

The coefficients output of the final model in Question 6 indicates the relationship between the predictor variables and the likelihood of a customer being a new RRSP purchaser. The intercept of the model represents the estimated log odds of a customer being a new RRSP purchaser when all other predictors are zero. The p-values associated with each coefficient indicate whether or not the coefficient is statistically significant in predicting the outcome.

Age, paydep, sqrt.BALCHQ, sqrt.BALSAV, DUMNOLOAN, sqrt.BALLOC, BALMRGG, sqrt.TXBRAN, TOTSERV, and valsegmD all have statistically significant coefficients. These coefficients are negative for age, indicating that older customers are less likely to be new RRSP purchasers. Paydep has a positive coefficient, suggesting that customers who use payroll deposit are more likely to be new RRSP purchasers.

The coefficients for the average monthly balance of the chequing account (sqrt.BALCHQ) and savings account (sqrt.BALSAV) are both positive, indicating that customers with higher average monthly balances in these accounts are more likely to be new RRSP purchasers. The coefficient for DUMNOLOAN is negative, indicating that customers without a personal loan are more likely to be new RRSP purchasers.

Customers with a higher average monthly balance in their line of credit (sqrt.BALLOC) and mortgage (BALMRGG) accounts are also more likely to be new RRSP purchasers, as indicated by the positive coefficients for these variables.

The coefficient for the number of in-branch transactions per month (sqrt.TXBRAN) is positive, suggesting that customers who conduct more in-branch transactions are more likely to be new RRSP purchasers. Similarly, the coefficient for TOTSERV is positive, indicating that customers who have more distinct services are more likely to be new RRSP purchasers.

Finally, the coefficient for the valsegmD variable is positive, indicating that customers in this value segment are more likely to be new RRSP purchasers compared to customers in (valsegmE).

8. **The costs and contribution of a RRSP purchaser are listed below: (not the true proprietary figures!)**
   - ***Contact cost (Mail and glossy brochure production cost): $2.75***
   - ***Number of potential contacts (members without an RRSP): 120,000***
   - ***Estimated average contribution from a single RRSP purchase: $180.00***

   Based on the CUMULATIVE lift chart numbers and these cost-contribution numbers, recommend the percentage of members (i.e., the top X%) who should be contacted to maximize profit after contact costs, and report what this expected profit is. Please attach your calculation table and formula. (15 points)

   To maximize profit after contact costs, Vancity should contact the top 50% of its members. This would result in a profit of approximately $188,808.

Number of potential contacts (members without an RRSP)	120,000
RRSP purchase rate	2.1%
Contact cost (Mail and glossy brochure production cost)	$2.75
Estimated average contribution from a single RRSP purchase	$180
The number of members who would purchase an RRSP	2520

Top % to Contact	The number of members to contact	The cost to contact top % members	The cumulative captured rate from the model	The expected number of members who would purchase an RRSP captured.	Expected $ Increase in RRSP Purchases	Expected Profit
10%	12000	$33,000	29%	730.8	$131,544	$98,544
20%	24000	$66,000	41%	1033.2	$185,976	$119,976
30%	36000	$99,000	58%	1461.6	$263,088	$164,088
40%	48000	$132,000	67%	1688.4	$303,912	$171,912
50%	60000	$165,000	78%	1965.6	$353,808	$188,808
60%	72000	$198,000	82%	2066.4	$371,952	$173,952
70%	84000	$231,000	89%	2242.8	$403,704	$172,704
80%	96000	$264,000	95%	2394	$430,920	$166,920
90%	108000	$297,000	97%	2444.4	$439,992	$142,992
100%	120000	$330,000	100%	2520	$453,600	$123,600

Calculation Formulas:

- The number of members who would purchase an RRSP = 120,000 * 2.1% = 2,520
- The number of members to contact = 120,000 * (Top % to Contact)
- The cost to contact top % member = $2.75 * (The number of members to contact)
- The cumulative captured rate from the model = from cumulative lift chart in Question 6B
- The expected number of members who would purchase an RRSP captured = 2,520 * (The cumulative captured rate from the model)
- Expected $ Increase in RRSP Purchases = $180 * (The expected number of members who would purchase an RRSP captured)
- Expected Profit = (The cost to contact top % member) - (Expected $ Increase in RRSP Purchases)

9. **Submit your final script (.R) for verification purpose. (10 points)**
   Submitted!