Final Project Analysis

CRYSTAL CONTRERAS

DSC 423 Spring 2021

\* Import dataset;

**DATA** bank;

INFILE "cleaned-bank.csv" FIRSTOBS=**2** DELIMITER=',';

INPUT i age job $ marital $ education $ default $ housing $ loan $ contact $ month $ day\_of\_week $ duration campaign pdays previous poutcome $ emp\_var\_rate cons\_price\_idx cons\_conf\_idx euribor3m nr\_employed y $;

\* Drop the index column from our data since it is not needed;

DROP i;

\* Create dummy variables for categorical attributes;

job1=(job='blue-collar');

job2=(job='services');

job3=(job='admin.');

job4=(job='self-employed');

job5=(job='technician');

job6=(job='management');

job7=(job='retired');

job8=(job='entrepreneur');

job9=(job='housemaid');

job10=(job='unemployed');

job11=(job='student');

marital1=(marital='married');

marital2=(marital='single');

marital3=(marital='divorced'); \* could also mean widowed;

ed0=(education='illiterate');

ed1=(education='basic.4y');

ed2=(education='basic.6y');

ed3=(education='basic.9y');

ed4=(education='high.school');

ed5=(education='professional.course');

ed6=(education='university.degree');

credit\_default=(default='yes'); \* 1 = yes, 0 = no;

housing\_loan=(housing='yes');

has\_loan=(loan='yes');

cellphone=(contact='cellular'); \* 0 = telephone;

\* survey conducted between March - December;

month3=(month='mar');

month4=(month='apr');

month5=(month='may');

month6=(month='jun');

month7=(month='jul');

month8=(month='aug');

month9=(month='sep');

month10=(month='oct');

month11=(month='nov');

month12=(month='dec');

day1=(day\_of\_week='mon');

day2=(day\_of\_week='tue');

day3=(day\_of\_week='wed');

day4=(day\_of\_week='thu');

day5=(day\_of\_week='fri');

prev\_outcome1=(poutcome='failure');

prev\_outcome2=(poutcome='nonexistent');

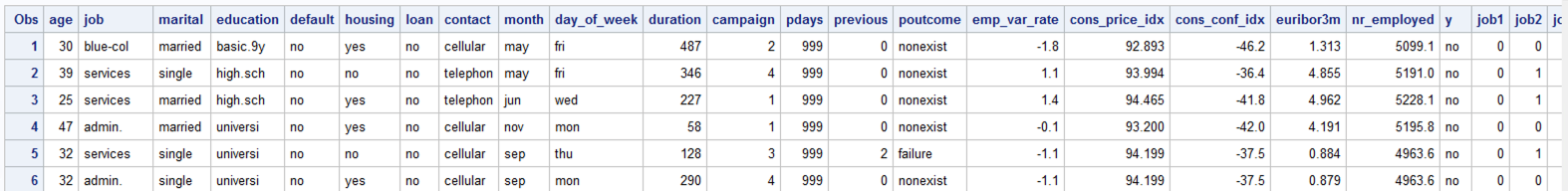
prev\_outcome3=(poutcome='success');

target=(y='yes');

**RUN**;

**PROC** **PRINT**;

**RUN**;



\* Create a copy of the dataset that omits duplicate attributes - the ones we created dummy variables for;

**DATA** bank\_new;

SET bank;

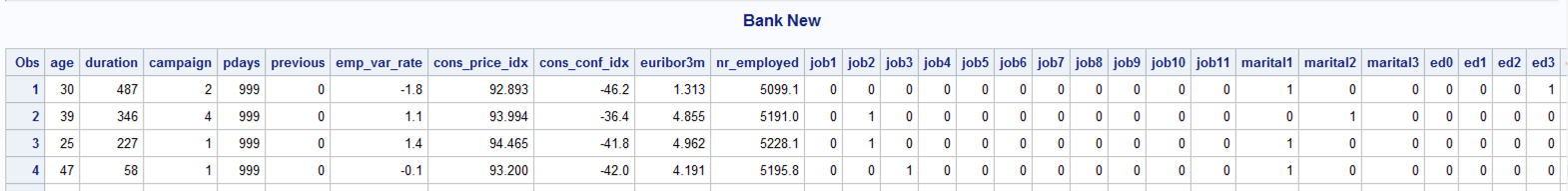
DROP job marital education month default housing loan contact day\_of\_week poutcome y;

**RUN**;

TITLE "Bank New";

**PROC** **PRINT**;

**RUN**;



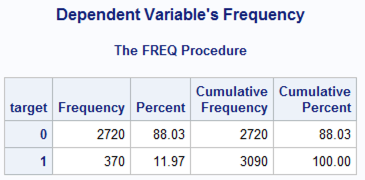
\* Check the frequency of our target variable to ensure we have enough samples of each side;

**PROC** **FREQ**;

TITLE "Dependent Variable's Frequency";

TABLES target;

**RUN**;



We have 2720 samples of 'no' = 88% Probability, and 370 samples of 'yes' = 12% Probability.

odds(y=1) = 0.12/0.88 => The odds that event Y = 1 occurs is 0.136 to 1. This means we have a higher chance of failure.

odds(y=0) = 0.88/0.12 => 7.33 to 1 (slide 8 of lecture 7)

I predict this will affect our Probability, threshold, train/test split.

1. Find out what percentage of each Y value we should have in relation to our dataset.

- Slide 35 of Lecture 7 states 10-30 cases per independent variable. 53 (w/dummy) \* 10 = 530 observations we should have.

" - Make sure it has enough observations for each case (1 & 0).

If there isn't enough samples or there are many cells with no response,

parameter estimates and standard errors are likely to be unstable & maximum likelihood estimation

(MLE) of parameters could be impossible to obtain."

2. Either remove some observations of 'no', add in more 'yes' samples, or both.

Check the following:

1. is the Y variable binary?
2. Is there enough samples?
   1. What is the sample size n?
   2. How many predictors are there (k)?
   3. How many observations have Y=1 & y=0?
   4. Do we have enough sample (sample/k=10+)?

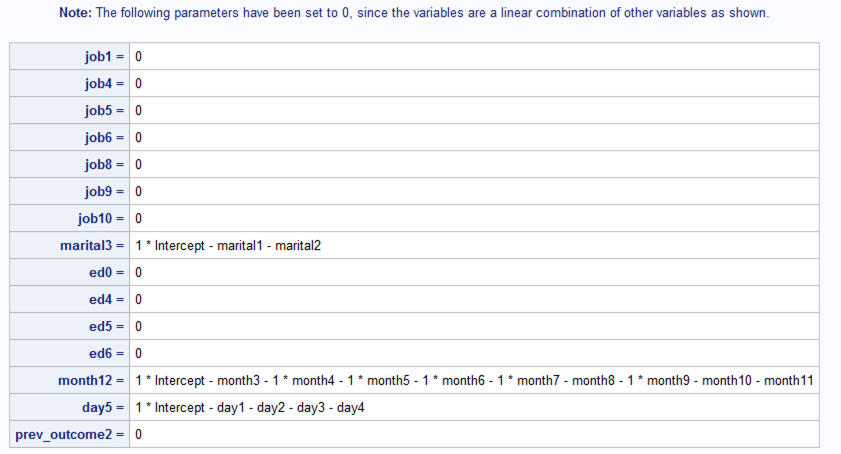
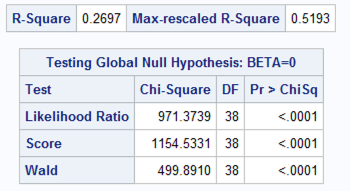
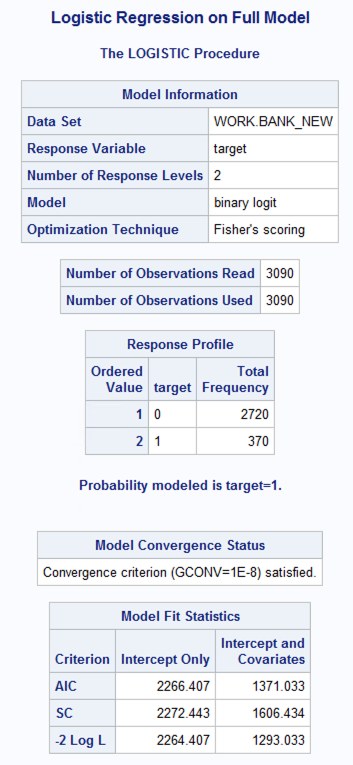
\* Run Logistic Regression on the full model;

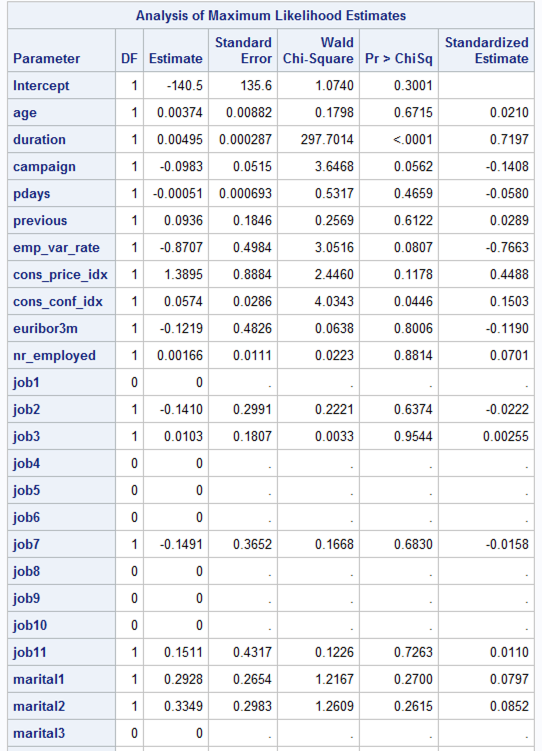
**PROC** **LOGISTIC**;

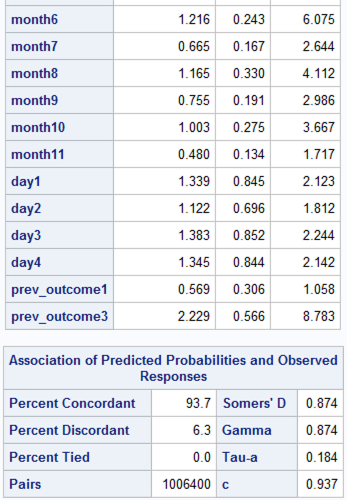
TITLE "Logistic Regression on Full Model";

MODEL target (event='1') = age duration campaign pdays previous emp\_var\_rate cons\_price\_idx cons\_conf\_idx euribor3m nr\_employed job1 job2 job3 job4 job5 job6 job7 job8 job9 job10 job11 marital1 marital2 marital3 ed0 ed1 ed2 ed3 ed4 ed5 ed6 credit\_default housing\_loan has\_loan cellphone month3 month4 month5 month6 month7 month8 month9 month10 month11 month12 day1 day2 day3 day4 day5 prev\_outcome1 prev\_outcome2 prev\_outcome3 / STB RSQUARE;

**RUN**;







Do we also need to write the general model equation?

Fit the full model with standardized coefficients.

1. Which variables have the highest influence?
   1. Duration, 0.7197
   2. Second highest influence: emp\_var\_rate, -0.7663
2. Which vars have sig effect on odd of Y=1?
   1. Credit\_default: -7.2688
   2. 2nd highest sig effect: month3: 2.078
3. Were any of the highest influential predictors insignificant?
   1. No. the top 2 predictors with the highest influence were not considered insignif.
4. Write full model equation

log(target=1/target=0) = -140.5 + 0.00374\*AGE + 0.00495\*DURATION – 0.0983\*CAMPAIGN – 0.00051\*PDAYS + 0.0936\*PREVIOUS – 0.8707\*EMP\_VAR\_RATE + 1.3895\*CONS\_PRICE\_IDX + 0.0574\*CONS\_CONF\_IDX – 0.1219\*EURIBOR3M + 0.00166\*NR\_EMPLOYED + 0\*JOB1 – 0.140\*JOB2 + 0.0103\*JOB3 + 0\*JOB4 + 0\*JOB5 + 0\*JOB6 -0.1491\*JOB7 + 0\*JOB8 + 0\*JOB9 + 0\*JOB10 + 0.1511\*JOB11 + 0.2928\*MARITAL1 + 0.3349\*MARITAL2 + 0\*MARITAL3 + 0\*ED0 – 0.3319\*ED1 + 0.02\*ED2 – 0.0675\*ED3 + 0\*ED4 + 0\*ED5 + 0\*ED6 – 7.2688\*CREDIT\_DEFAULT + 0.0448\*HOUSING\_LOAN – 0.1193\*HAS\_LOAN + 1.1239\*CELLPHONE + 2.0781\*MONTH3 – 0.3086\*MONTH4 - 0.7321\*MONTH5 + 0.1957\*MONTH6 – 0.4074\*MONTH7 + 0.1525\*MONTH8 – 0.2815\*MONTH9 + 0.003\*MONTH10 – 0.73\*MONTH11 + 0\*MONTH12 + 0.2922\*DAY1 + 0.1155\*DAY2 + 0.3240\*DAY3 + 0.2963\*DAY4 + 0\*DAY5 – 0.5637\*PREV\_OUTCOME1 + 0\*PREV\_OUTCOME2 + 0.8015\*PREV\_OUTCOME3

Where job1=1 when job='blue-collar',

job2=1 when job='services',

job3=1 when job='admin.',

job4=1 when job='self-employed',

job5=1 when job='technician',

job6=1 when job='management',

job7=1 when job='retired',

job8=1 when job='entrepreneur',

job9=1 when job='housemaid',

job10=1 when job='unemployed',

job11=1 when job='student',

marital1=1 when marital='married',

marital2=1 when marital='single',

marital3=1 when marital='divorced' (which could also mean widowed),

ed0=1 when education='illiterate',

ed1=1 when education='basic.4y',

ed2=1 when education='basic.6y',

ed3=1 when education='basic.9y',

ed4=1 when education='high.school',

ed5=1 when education='professional.course',

ed6=1 when education='university.degree',

credit\_default=1 when default='yes',

housing\_loan=1 when housing='yes',

has\_loan=1 when loan='yes',

cellphone=1 when contact='cellular',

cellphone=0 when contact=‘telephone’,

month3=1 when month='mar',

month4=1 when month='apr',

month5=1 when month='may',

month6=1 when month='jun',

month7=1 when month='jul',

month8=1 when month='aug',

month9=1 when month='sep',

month10=1 when month='oct',

month11=1 when month='nov',

month12=1 when month='dec',

day1=1 when day\_of\_week='mon',

day2=1 when day\_of\_week='tue',

day3=1 when day\_of\_week='wed',

day4=1 when day\_of\_week='thu',

day5=1 when day\_of\_week='fri',

prev\_outcome1=1 when poutcome='failure',

prev\_outcome2=1 when poutcome='nonexistent',

prev\_outcome3=1 when poutcome='success',

target=1 when y='yes'.

R-SQUARE = 0.2697 (Want highest)

AIC = 1371.033 (Want lowest)

SC = 1606.434 (Want lowest)

Likelihood Ratio (LR) = 971.3739 (Want highest)

For AIC & SC, we compare their ‘Intercept & Covariates’ values against each model.

LR used for GOF Test. Similar to F-test. Tests Global Null Hypothesis. P-value associated with it is < .0001. Reject H0. The predictors can significantly predict the outcome of Y=1. (Lecture 7, slide 32).

\* Explore data to check for multicollinearity amongst independent, non-categorical variables;

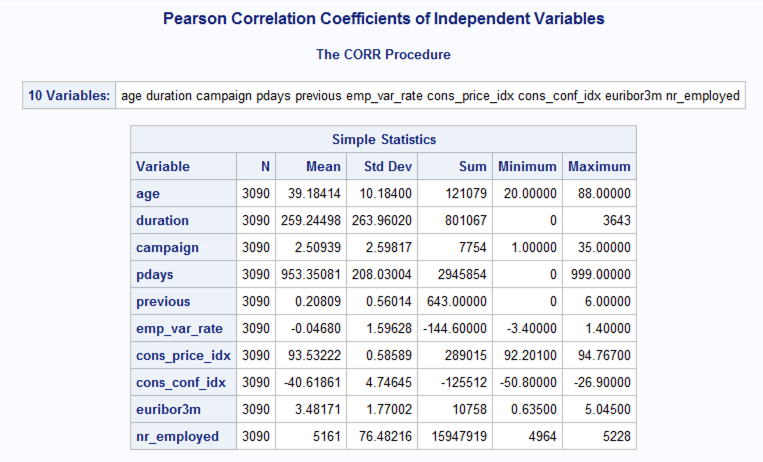
**PROC** **CORR**;

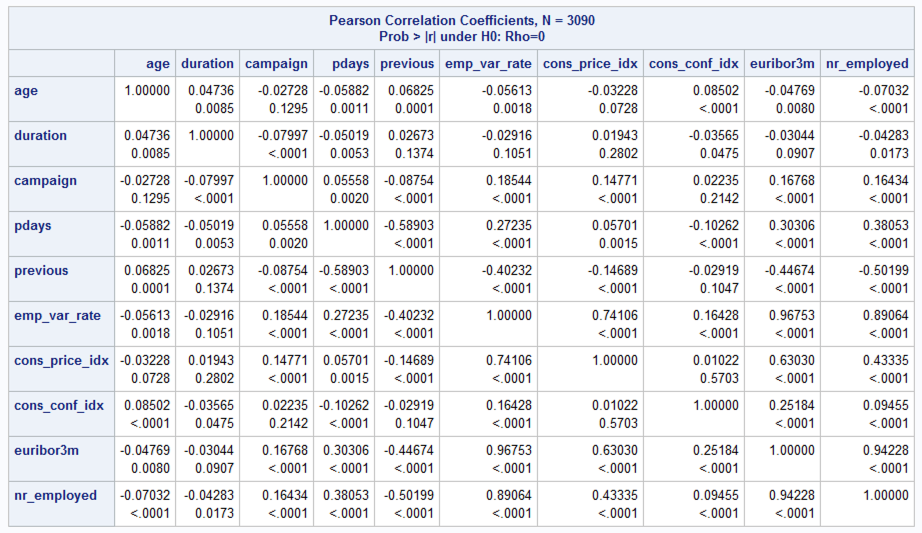
TITLE "Pearson Correlation Coefficients of Independent Variables";

VAR age duration campaign pdays previous emp\_var\_rate cons\_price\_idx cons\_conf\_idx euribor3m nr\_employed;

**RUN**;

Results:





EURIBOR3M & EMP\_VAR\_RATE have a Pearson correlation coefficient of 0.96753.

EURIBOR3M & NR\_EMPLOYED have a Pearson correlation coefficient of 0.94228.

NR\_EMPLOYED & EMP\_VAR\_RATE have a Pearson correlation coefficient of 0.89064.

\* Run a selection method for logistic regression;

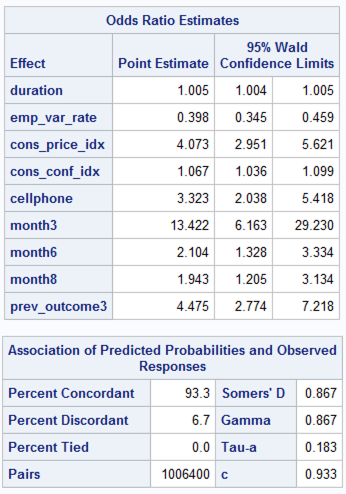
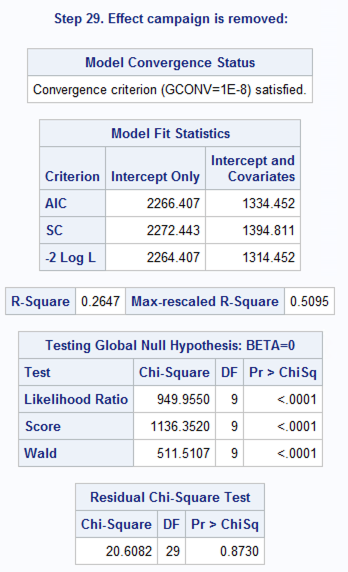
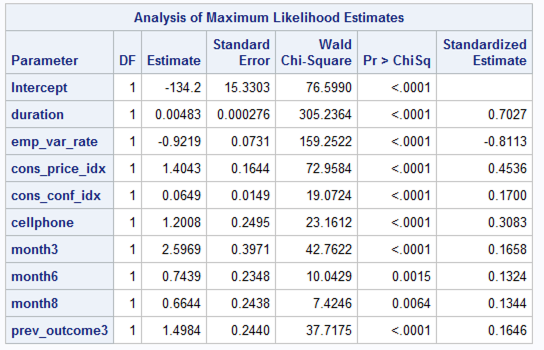
**PROC** **LOGISTIC**;

TITLE "Backwards selection method";

MODEL target (event='1') = age duration campaign pdays previous emp\_var\_rate cons\_price\_idx cons\_conf\_idx euribor3m nr\_employed job1 job2 job3 job4 job5 job6 job7 job8 job9 job10 job11 marital1 marital2 marital3 ed0 ed1 ed2 ed3 ed4 ed5 ed6 credit\_default housing\_loan has\_loan cellphone month3 month4 month5 month6 month7 month8 month9 month10 month11 month12 day1 day2 day3 day4 day5 prev\_outcome1 prev\_outcome2 prev\_outcome3 / SELECTION=BACKWARD RSQUARE STB;

**RUN**;

Backwards selection chose: DURATION, EMP\_VAR\_RATE, CONS\_PRICE\_IDX, CONS\_CONF\_IDX, CELLPHONE, MONTH3, MONTH6, MONTH8, PREV\_OUTCOME3

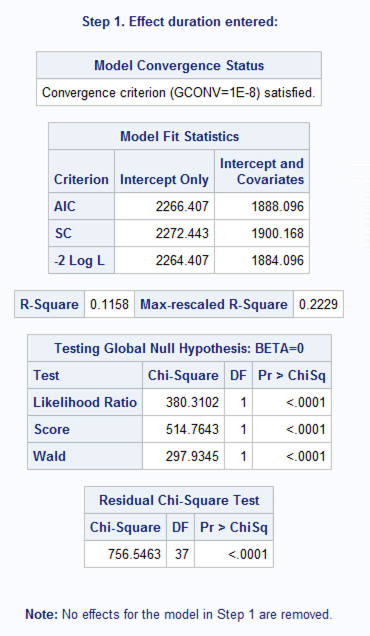
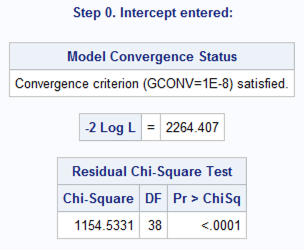
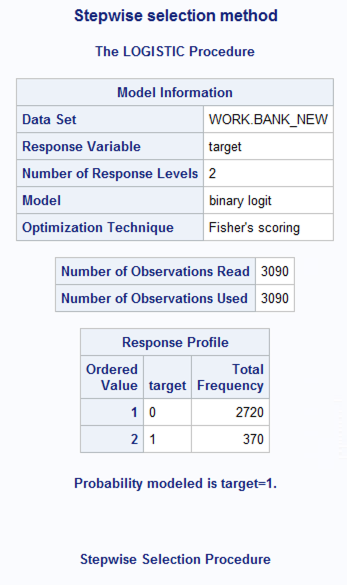
\* Run a selection method for logistic regression;

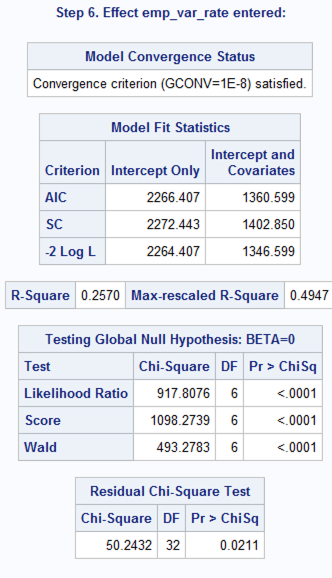
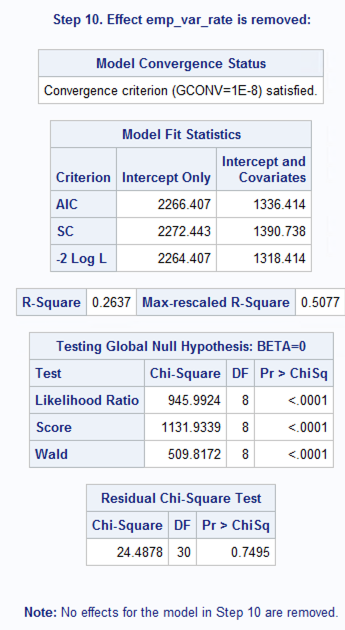
**PROC** **LOGISTIC**;

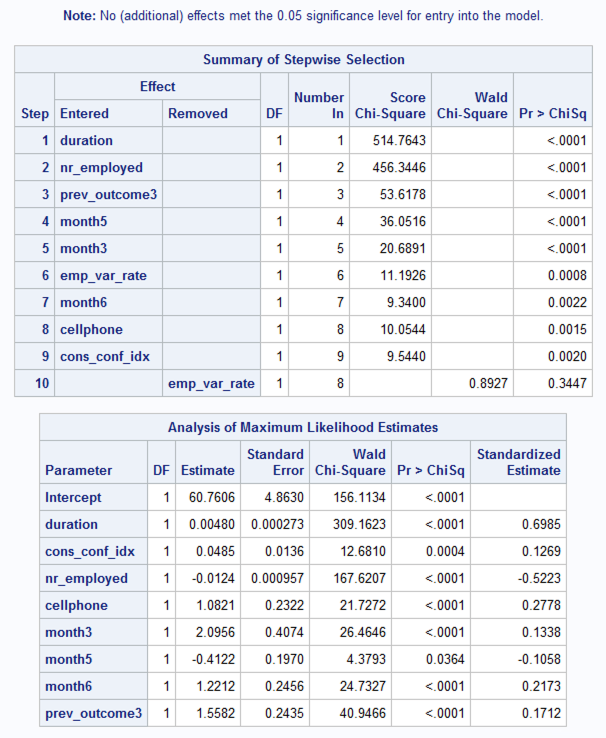
TITLE "Stepwise selection method";

MODEL target (event='1') = age duration campaign pdays previous emp\_var\_rate cons\_price\_idx cons\_conf\_idx euribor3m nr\_employed job1 job2 job3 job4 job5 job6 job7 job8 job9 job10 job11 marital1 marital2 marital3 ed0 ed1 ed2 ed3 ed4 ed5 ed6 credit\_default housing\_loan has\_loan cellphone month3 month4 month5 month6 month7 month8 month9 month10 month11 month12 day1 day2 day3 day4 day5 prev\_outcome1 prev\_outcome2 prev\_outcome3 / SELECTION=STEPWISE RSQUARE STB;

**RUN**;





Stepwise selection model chose: DURATION, CONS\_CONF\_IDX, NR\_EMPLOYED, CELLPHONE, MONTH3, MONTH5, MONTH6, PREV\_OUTCOME3

RSQUARE = 0.2637

Month5 (May) was seen as significant in the STEPWISE selection method.

This is biased because it's the month with the most calls.

Month3 (March) followed as significant. It has the 2nd lowest frequency (good thing),

but it's also the 1st month of the campaign. Finally, month6 (June) followed as significant.

This month may also be biased because it has the 2nd highest frequency of calls.

The stepwise selection method chose 8 out of the 53 attributes, but the R2 value was only 0.2637.

I want to test if removing some observations would increase the Rsquare value,

specifically observations that meet the following criteria:

- target = 0

- pdays = 999 (b/c 95% of our data is comprised of those values)

- month = 'may' (month w/the highest amount of calls)

Another option could be to add in more 'yes' samples from the full dataset to add

more variance and increase the probability that event Y will occur.

Use stepwise and backward selection methods. Compare the 2 methods.

1. Did both of them select the same predictors?

Backwards selection chose: DURATION, EMP\_VAR\_RATE, CONS\_PRICE\_IDX, CONS\_CONF\_IDX, CELLPHONE, MONTH3, MONTH6, MONTH8, PREV\_OUTCOME3

Stepwise selection model chose: DURATION, CONS\_CONF\_IDX, NR\_EMPLOYED, CELLPHONE, MONTH3, MONTH5, MONTH6, PREV\_OUTCOME3

Note:

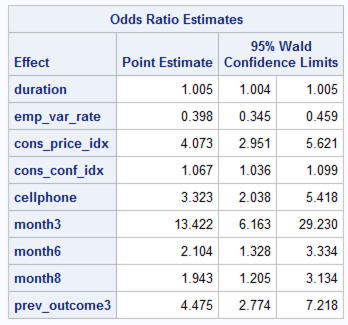
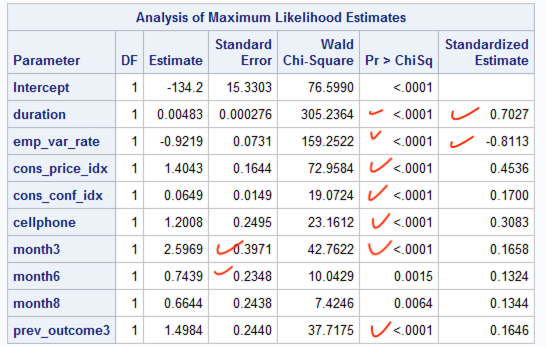
EURIBOR3M & EMP\_VAR\_RATE have a Pearson correlation coefficient of 0.96753.

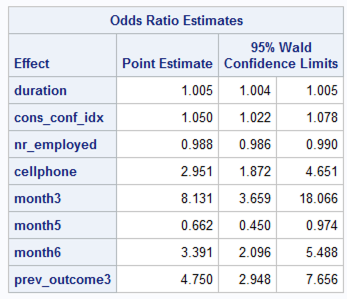
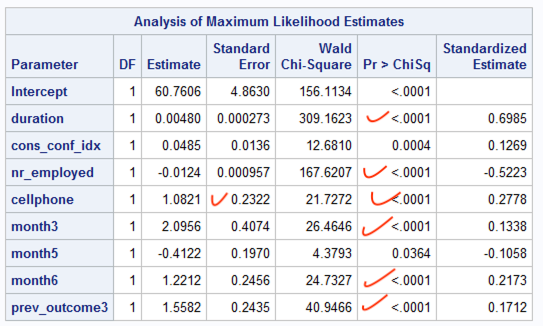
EURIBOR3M & NR\_EMPLOYED have a Pearson correlation coefficient of 0.94228.

NR\_EMPLOYED & EMP\_VAR\_RATE have a Pearson correlation coefficient of 0.89064.

1. Did they select the identical models (i.e. same parameter estimates, CI)?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter Estimates |  |  | Backwards | Stepwise |
| Intercept |  |  | -134.2 | 60.76 |
| DURATION |  |  | 0.00483 | 0.0048 |
| EMP\_VAR\_RATE |  |  | -0.9219 |  |
| CONS\_PRICE\_IDX | |  | 1.4043 |  |
| CONS\_CONF\_IDX | |  | 0.0649 | 0.0485 |
| CELLPHONE |  |  | 1.2008 | 1.0821 |
| MONTH3 |  |  | 2.5969 | 2.0956 |
| MONTH5 |  |  |  | -0.4122 |
| MONTH6 |  |  | 0.7439 | 1.2212 |
| MONTH8 |  |  | 0.6644 |  |
| NR\_EMPLOYED |  |  |  | -0.0124 |
| PREV\_OUTCOME3 | |  | 1.4984 | 1.5582 |

Backwards:  


Stepwise:  
 

Based on the following metrics, the variables selected by Backwards selection returned the better model to use:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Full Model | Backwards Selection Method | Stepwise Selection Method |
| highest | RSQUARE | 0.2697 | 0.2647 | 0.2637 |
| lowest | AIC | 1371.033 | 1334.452 | 1336.414 |
| lowest | SC | 1606.434 | 1394.811 | 1390.738 |
| highest | LR | 971.3739 | 1314.452 | 945.9924 |
|  |  |  |  |  |
| (Most Influential Predictors) Standardized Estimates for: |  |  |  |  |
| EMP\_VAR\_RATE |  | -0.7663 | -0.8113 |  |
| CONS\_PRICE\_IDX | | 0.4488 | 0.4536 | -0.1269 |
| MONTH5 |  | -0.1879 |  | -0.1058 |
| MONTH8 |  | 0.0308 | 0.1344 |  |
| NR\_EMPLOYED |  | 0.0701 |  | -0.5223 |

Fit the final model to check the following

\* Run Logistic Regression on the selected variables;

\* Check for multicollinearity - Pearson Correlation;

\* Requires CORRB option at the end of PROC LOGISTIC MODEL;

\* Check for outliers - Pearson or Deviance Residuals +-3;

\* Requires IPLOTS option;

\* Check for influential points - DFBetas;

\* Requires INFLUENCE option;

**PROC** **LOGISTIC**;

TITLE "Logistic Regression on Backward selection method's predictors";

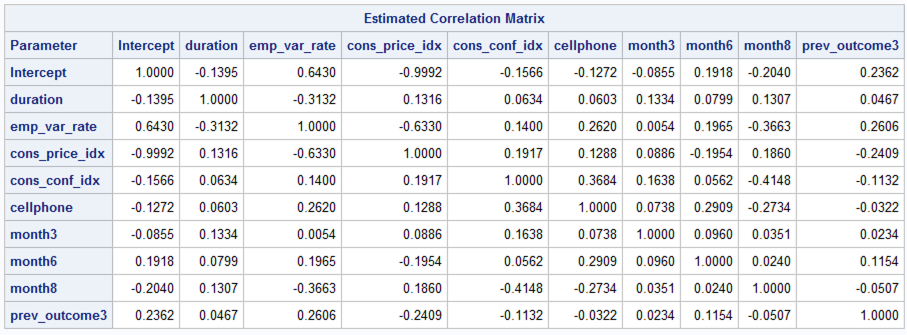
MODEL target (event='1') = duration emp\_var\_rate cons\_price\_idx cons\_conf\_idx cellphone month3 month6 month8 prev\_outcome3 / RSQUARE STB CORRB IPLOTS INFLUENCE;

**RUN**;

1. Are there multicollinearity among predictors that are in the final model?

In order to check for multicollinearity, we used the CORRB option in our Logistic Regression’s MODEL statement, which checks “the correlation of the coefficients of these variables in the model” (Slide 40, lecture 7).

The final model does not present multicollinearity among predictors. All the predictors selected have a correlation value of < 0.7.

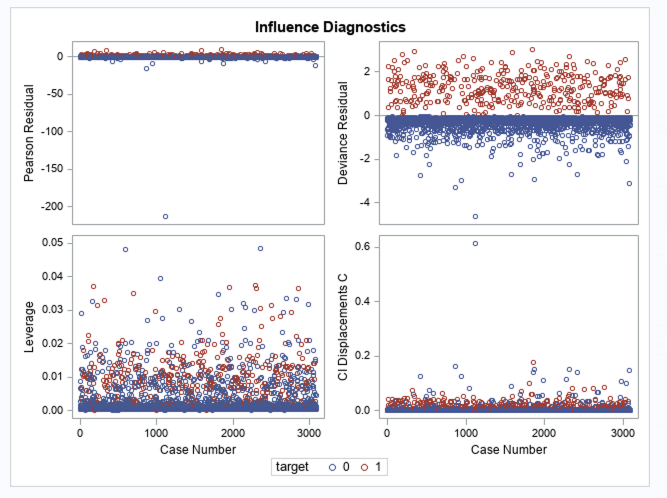


1. Are there influential points and outliers?

outliers:

Observation 1093, 1012, 1119, 1323, 1448, 1499, 1500, 1580, 1693, 1797, 3082

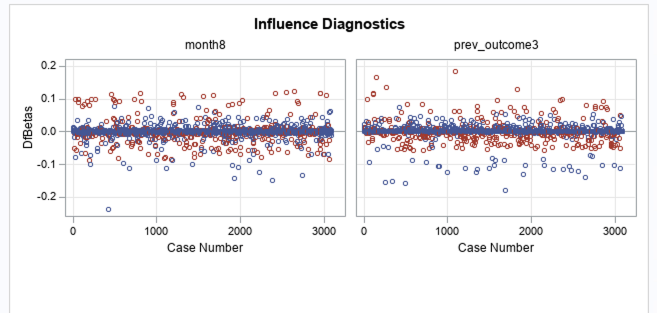
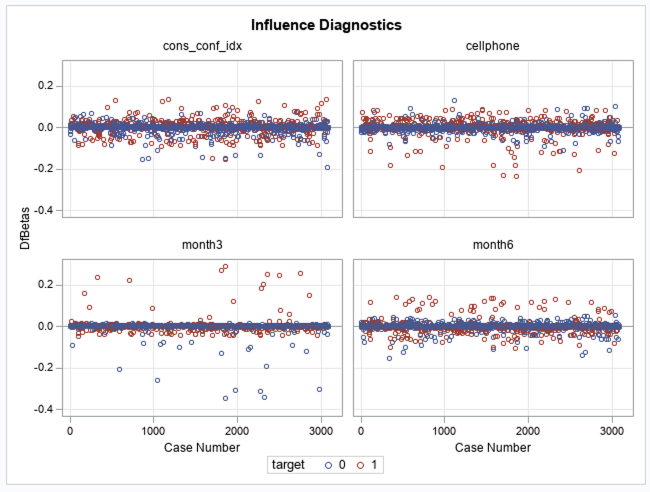
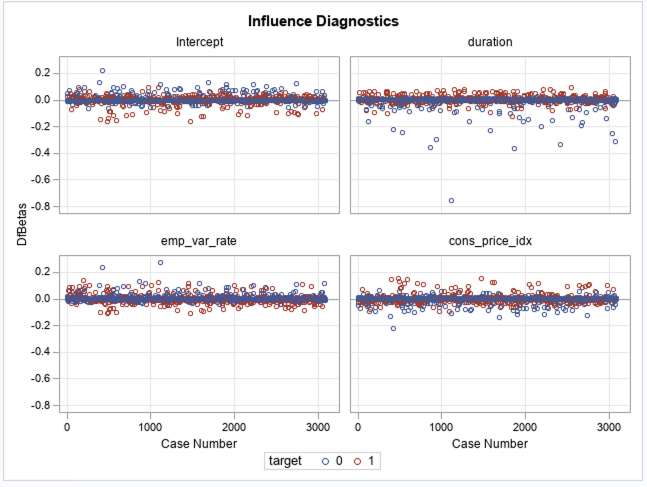
Found these based on Pearson Residuals & Deviance Residual values.



For influential points, we checked the DFBETA values.

n=3090 | Dfbeta | > 2/sqrt(n) | Dfbeta | > 2/sqrt(3090) ≈ 0.036

I think we have too many observations, which makes it harder to check for influential points. According to our formula, it would seem as though most of our observations would be considered influential points.



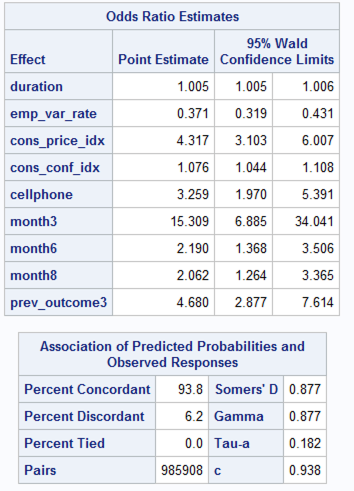
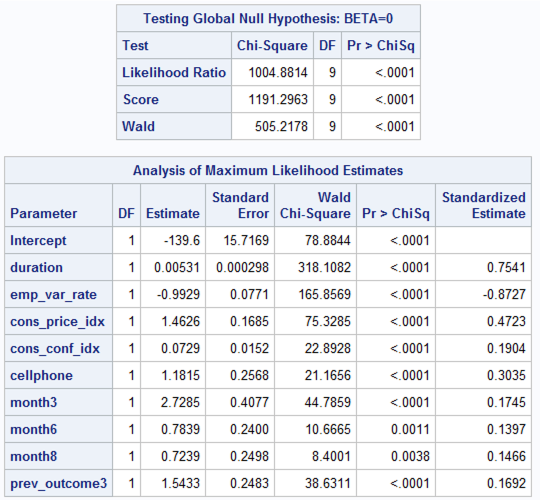
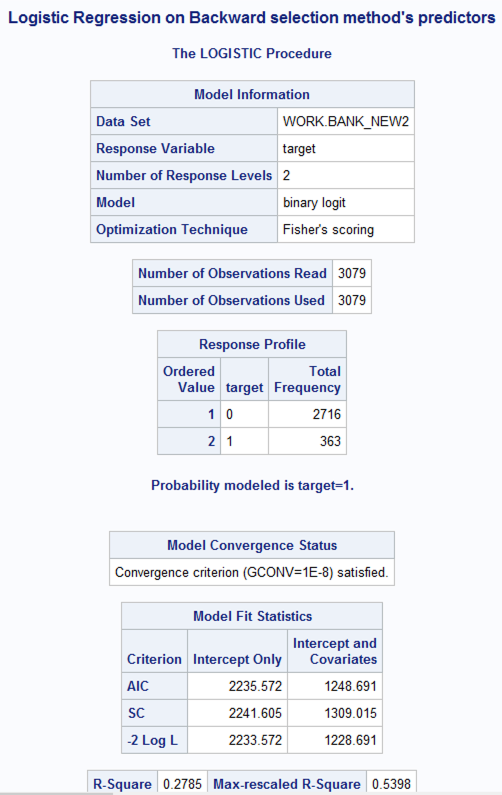
1. If you see collinearity among predictors what actions would you take?

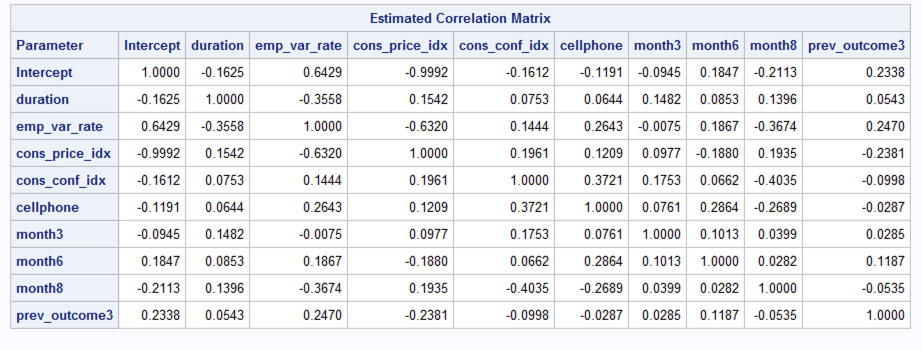
Since no collinearity existed among predictors, no actions were taken.

1. If you see influential points or outliers, what actions would you take?

Since we have an abundance of data points, we removed all *outliers* with Pearson & Deviance Residuals > 3.

New N = 3079, y=0 = 2716, 1 = 363. The RSQUARE, AIC, SC, & LR improved.

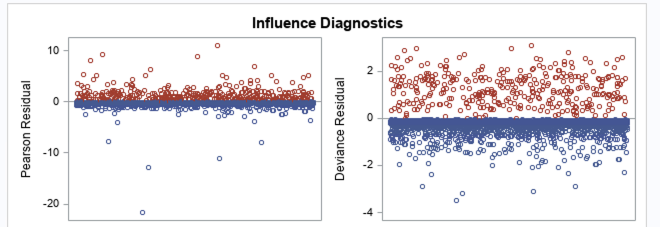




Still at least 5-8 outliers in observations:

2 that are <1000

Observations: 531, 712, 860, 940, 1835, 1860



Note: Don’t remove IPs for dummy variables, which in this case are MONTH3, MONTH6, MONTH8, & PREV\_OUTCOME3

For the final model

1. Write down the likelihood ratio test (null hypothesis, statistic and p-value)
2. Write down the final model equation
3. Ananlyze conditional effect of each varianle on the odds of Y=1. Use coefficient values and the confidence intervals to explain your answer.

R