Baruch College
STA 9750

BOOK BUYING PATTERNS

A Predictive Model for Residents in NY, PA and OH

Senem Acet Coskun

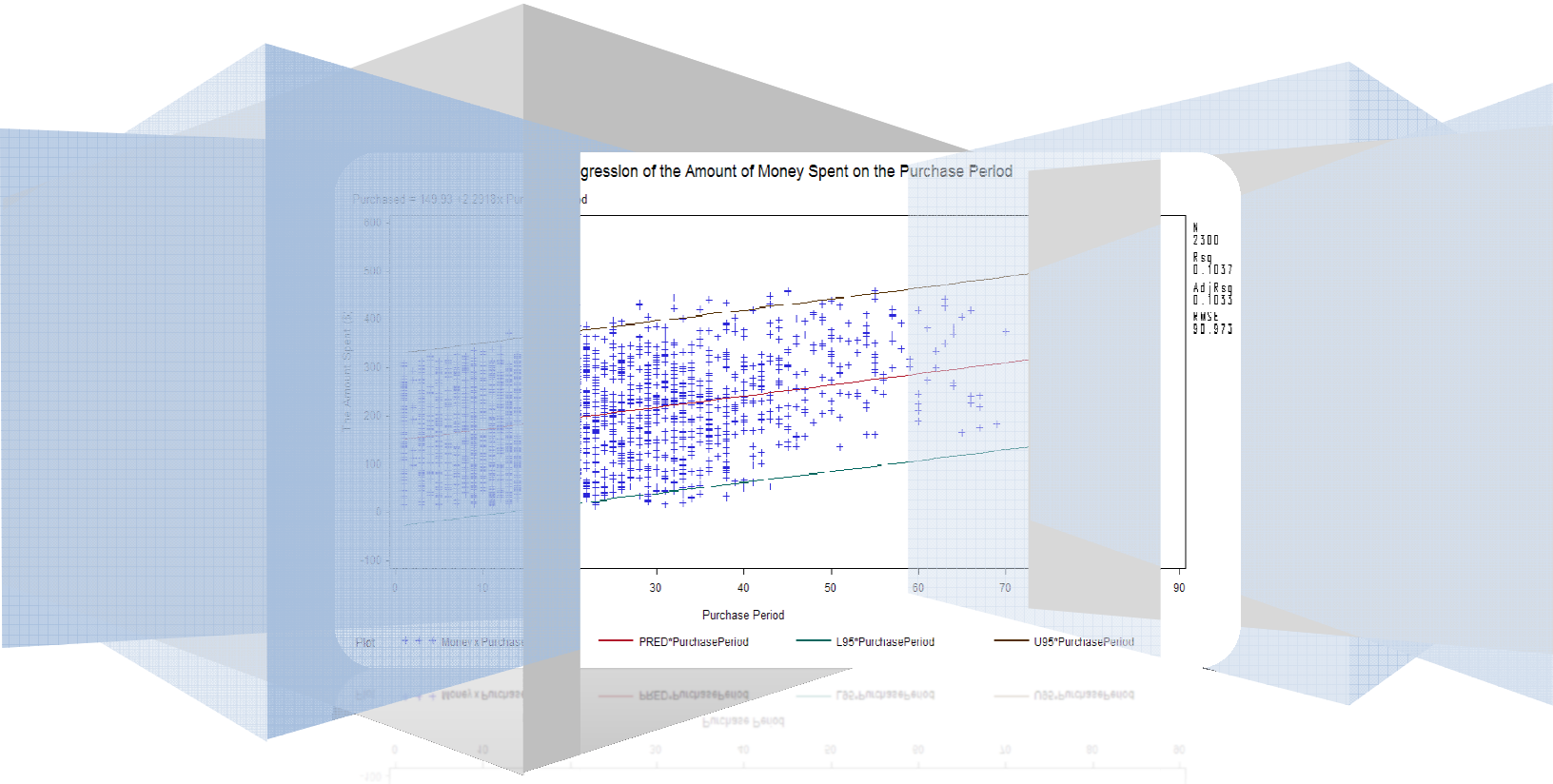


Table of Contents

Summary

In this project I intend to analyze book buying patterns of residents in NY, PA and OH by using the database of Bookbinders Club. BBBC has a database of more than 500,000 customers and recently made a direct marketing study which involved 20,000 customers. The data I am using in this project comes from the mentioned direct marketing study. Here, the problem that I am trying to solve is deriving a predictive model for buying books by using influential predictors such as purchase period, number of books purchase, and number of child books purchased…etc.

First I wanted to determine whether there is a difference in the average amount of money spent on BBBC books for different genders. This might help to select a gender oriented marketing strategy. However, the two-sample t-test revealed that both genders spend the same amount of money on average.

A regression of the time between first and last purchases on the amount of money spent suggested that there is a significant relationship between the two variables; however it is suspected that only one variable might not be enough to predict the amount of money spent. Hence a multiple linear regression is adopted to derive a predictive model by using all given parameters, except gender. The new model was promising in terms of having null residual plots and statistically significant parameter estimates. This part was included in the last section of the project as “something we haven’t done in class”.

A logistic regression was applied to see the probability of buying a particular book titled “The Art History of Florence” given the amount of money spent on BBBC books. The graph, unfortunately, didn’t give a good picture, however it was successfully driven that the probability of buying “The Art History of Florence” increased as the amount of money spent on BBBC books increased.

Finally, the dangers of regressing a random walk on time were shown by using Merck Pharmaceutical Company’s stock prices. The regression output suggested a very strong relationship between time and Merck’s stock prices, which is highly misleading. Scatter plot matrices revealed this misleading relationship very clearly.

This project was very helpful for me to understand how a logical and a statistically sound strategy could be driven from a data set. I would like to thank Prof. Lawrence Tatum for his endless efforts during the semester and my husband for his valuable discussions throughout the process and for his patience.

Why this topic?

I want to be a marketing strategist/ consultant who relies on statistics and quantitative models when making critical marketing decisions. I am also applying for a PhD in Quantitative Marketing for fall 2010. I will apply to Columbia, NYU, Rutgers and Baruch.

While making data set research for this project, I only looked for data that I can derive "naïve" marketing strategies out of it. Bookbinders Book Club (BBBC) data was just a perfect match for my intensions. The description of the dataset and the problem caught my interest immediately. It was indeed what I want to do in real life, when I have all the expertise and the experience. I didn't change the description of the problem as I think it's very nicely and neatly put.

As it is stated on the case study…

" About 50,000 new titles, including new editions, are published in the United States each year, giving rise to a $20+ billion book publishing industry. About 10 percent of the books are sold through mail order.

Book retailing in the 1970s was characterized by the growth of chain bookstore operations in concert with the development of shopping malls. Traffic in bookstores in the 1980s was enhanced by the spread of discounting. In the 1990s, the superstore concept of book retailing was responsible for the double-digit growth of the book industry. Generally situated near large shopping centers, superstores maintain large inventories of anywhere from 30,000 to 80,000 titles. Superstores are putting intense competitive pressure on book clubs, mail-order firms and retail outlets. Recently, online superstores, such as www.amazon.com, have emerged, carrying 1–2.5 million titles and further intensifying the pressure on book clubs and mail-order firms. In response to these pressures, book clubs are starting to look at alternative business models that will make them more responsive to their customers' preferences.

Historically, book clubs offered their readers continuity and negative option programs that were based on an extended contractual relationship between the club and its subscribers. In a continuity program, popular in such genres as children's books, a reader signs up for an offer of several books for a few dollars each (plus shipping and handling on each book) and agrees to receive a shipment of one or two books each month thereafter. In a negative option program, subscribers get to choose which and how many additional books they will receive, but the default option is that the club's selection will be

delivered to them each month. The club informs them of the monthly selection and they must mark "no" on their order forms if they do not want to receive it. Some firms are now beginning to offer books on a positive-option basis, but only to selected segments of their customer lists that they deem receptive to specific offers.

Book clubs are also beginning to use database marketing techniques to work smarter rather than expand the coverage of their mailings. BBBC is exploring whether to use predictive modeling approaches to improve the efficacy of its direct mail program. For a recent mailing, the company selected 20,000 customers in Pennsylvania, New York and Ohio from its database and included with their regular mailing a specially produced brochure for the book The Art History of Florence. This resulted in a 9.03 percent response rate (1806 orders) for the purchase of the book. BBBC then developed a database to calibrate a response model to identify the factors that influenced these purchases."

Data Sources:

DecisionsPro, Inc. provides software, training, and consulting services to improve marketing processes and decisions. (http://www.decisionpro.biz/). Marketing Engineering

Marketing Engineering - Computer Assisted Marketing Analysis and Planning is a combination of books, software, and business cases developed and written by Professors Gary L. Lilien and Arvind Rangaswamy. The books, software, and cases are intended to be used in conjunction with one another to provide theory (book), computer modeling techniques (software), and context-specific operations decisions and action (business cases) (http://www.mktgeng.com/index.cfm)

The Bookbinders Book Club (BBBC) data is one of free case studies for students in the website:

(http://www.mktgeng.com/student/downloads/datasets.cfm)

Variable Definitions

Choice: Whether the customer purchased the "The Art History of Florence". 1 corresponds to a purchase and 0 corresponds to a non-purchase.

Gender: 0 = Female and 1 = Male.

Amount purchased: Total money spent on BBBC books.

Frequency: Total number of purchases in the chosen period (used as a proxy for frequency.)

Last purchase (recency of purchase): Months since last purchase.

First purchase: Months since first purchase.

P_Child: Number of children's books purchased.

P_Youth: Number of youth books purchased.

P_Cook: Number of cookbooks purchased.

P_DIY: Number of do-it-yourself books purchased.

P_Art: Number of art books purchased.

Descriptive Statistics:

Variable: All variables in the data set

To analyze the buying patterns, a new variable (Purchase Period) was created to indicate the time between first and last purchase.

Descriptive Statistics for New Variable 'time'

The MEANS Procedure

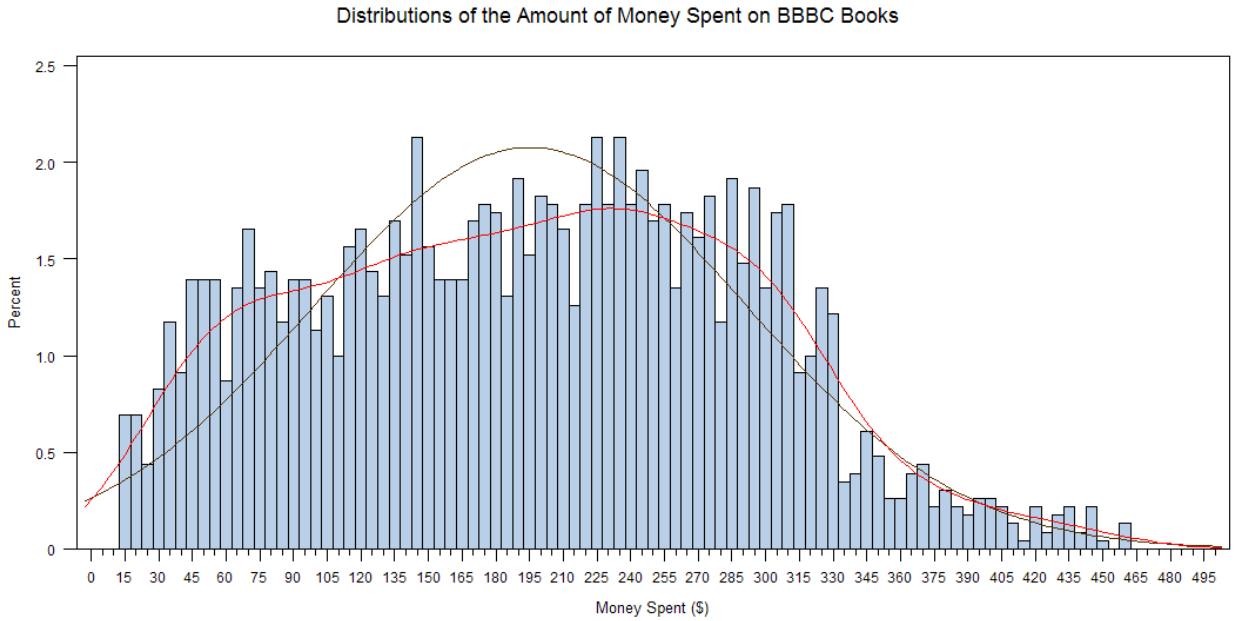
Variable	N	Mean	Std Dev	Std Error
Purchased	2300	195.3	96.1	2.0
Frequency	2300	13.3	8.2	0.2
LastPurchase	2300	3.1	2.9	0.1
FirstPurchase	2300	22.8	15.7	0.3
Childbook	2300	0.7	1.0	0.0
Youthbook	2300	0.3	0.6	0.0
Cookbook	2300	0.8	1.1	0.0
DIYbook	2300	0.4	0.7	0.0
Artbook	2300	0.3	0.6	0.0
PurchasePeriod	2300	19.8	13.5	0.3

Annotated SAS Program for the Descriptive Statistics

```
*reading data into SAS;
filename inf "C:\Data\book.csv";
data book;
infile inf firstobs=1 dlm=",";
input  Choice Gender Purchased Frequency LastPurchase  FirstPurchase Childbook
       Youthbook Cookbook DIYbook Artbook;


*descriptive statistics;
data book2;                              /* creates new data to add a new variable */
set book;                                   /* calls original data in action */
PurchasePeriod = FirstPurchase - LastPurchase;     /* new variable is created */
proc means Data=book2 N Mean std stderr maxdec=1;   /* descriptive statistics */
Title "Descriptive Statistics";
var Purchased Frequency LastPurchase FirstPurchase
Childbook     Youthbook     Cookbook      DIYbook
Artbook PurchasePeriod;                              /* variables of interest */
run;
```

Univariate Analysis and Histogram

Variable: Amount of Money Spent on BBBC Books.



The amount of money spent on BBBC books does not show a perfect normal distribution but it is safe to assume that it is normal as there is no other significant distribution pattern.

Histograms are sensitive to the number of bins or columns that are used in the display. Kernel density plot (red line), which approximates the probability density of the variable, is smooth and independent of the choice of origin, unlike histograms. Hence, I used kernel density plot together with normal density plot (black line) to illustrate the difference.

Univariate Output for the Amount of Money Spent on BBC Books

Distributions of the Amount of Money Spent on Books

The UNIVARIATE Procedure
Variable: Purchased

Moments

N	2300	Sum Weights	2300
Mean	195.276957	Sum Observations	449137
Std Deviation	96.0718761	Variance	9229.80538
Skewness	0.10777245	Kurtosis	-0.704408
Uncorrected SS	108925429	Corrected SS	21219322.6
Coeff Variation	49.1977537	Std Error Mean	2.0032371

Basic Statistical Measures

Location		Variability	
Mean	195.2770	Std Deviation	96.07188
Median	198.0000	Variance	9230
Mode	114.0000	Range	446.00000
		Interquartile Range	149.00000

Tests for Location: Mu0=0

Test		-Statistic-		-----p Value------	
Student's t	t	97.4807	Pr > |t|	<.0001
Sign	M	1150	Pr >= |M|	<.0001
Signed Rank	S	1323075	Pr >= |S|	<.0001

Tests for Normality

Test		--Statistic---		-----p Value------
Kolmogorov-Smirnov	D	0.042622	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.290403	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	9.2414	Pr > A-Sq	<0.0050

Quantiles (Definition 5)

Quantile	Estimate
100% Max	461.0
99%	421.0
95%	346.0
90%	315.0
75% Q3	268.0
50% Median	198.0
25% Q1	119.0
10%	63.0

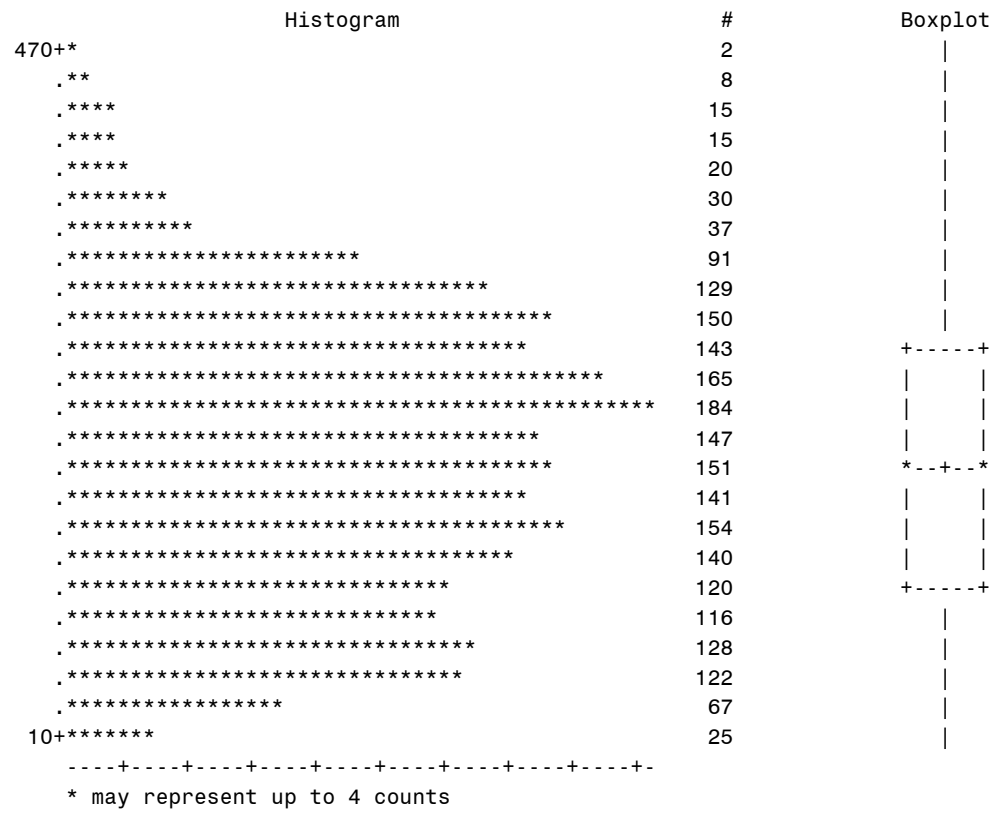
The UNIVARIATE Procedure
Variable: Purchased

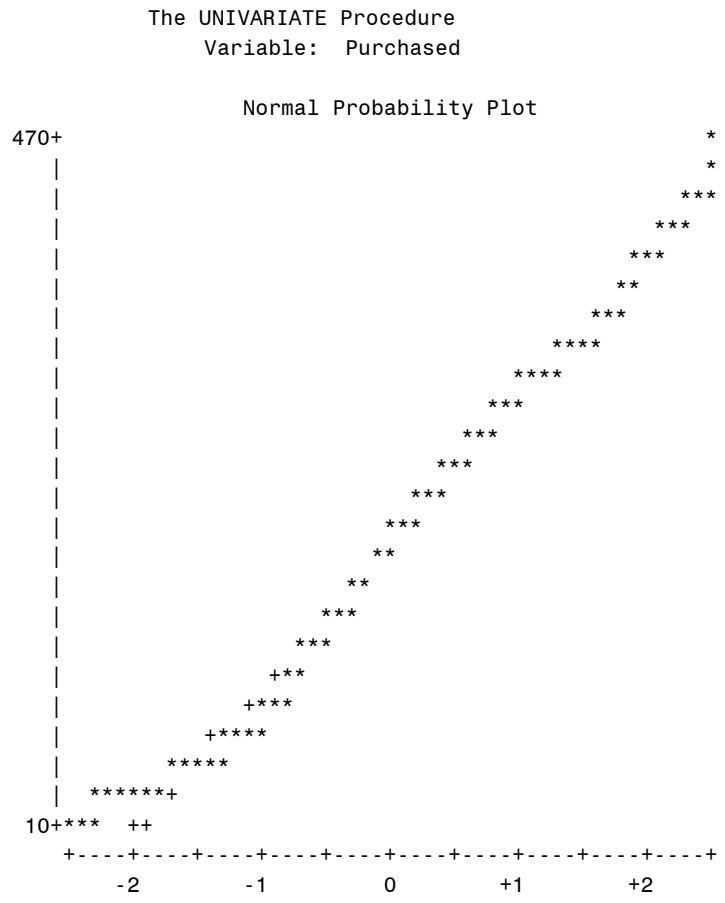
Quantiles (Definition 5)

Quantile	Estimate
5%	43.5
1%	18.5
0% Min	15.0

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
15	536	447	1569
15	278	450	1834
16	1985	458	260
16	1027	461	302
16	899	461	417



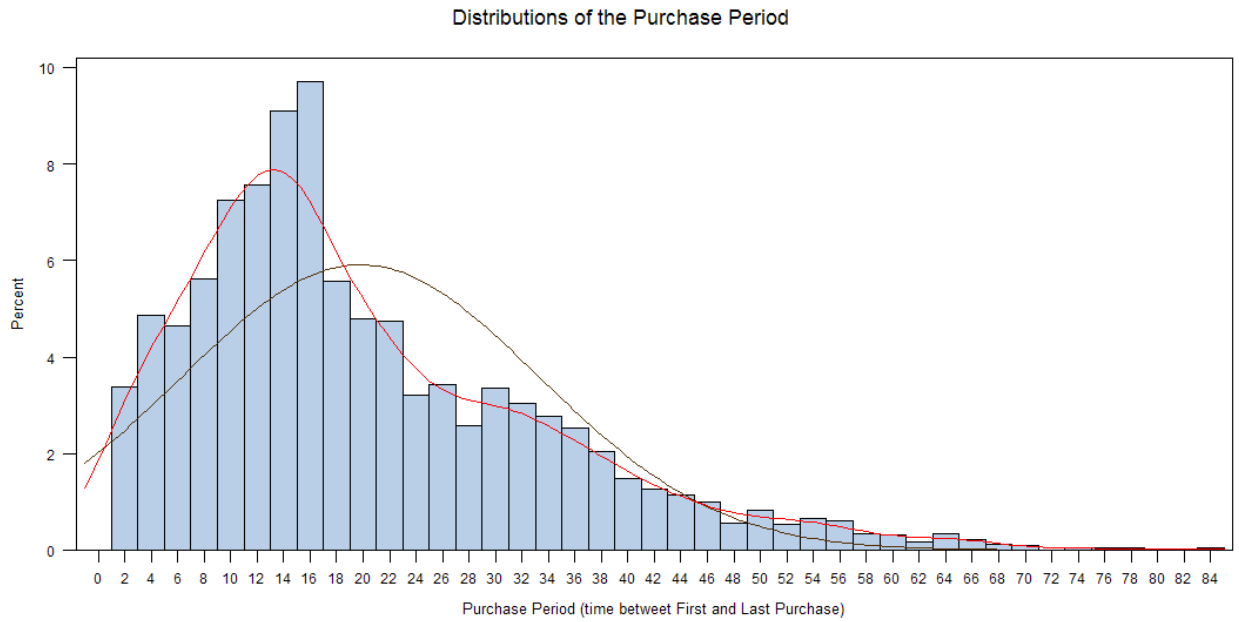


Annotated SAS program for Univariate Analysis and Histogram

```
*reading data into SAS;
filename inf "C:\Data\book.csv";
data book;
infile inf firstobs=1 dlm=",";
input  Choice Gender Purchased Frequency LastPurchase  FirstPurchase Childbook
      Youthbook Cookbook DIYbook Artbook;
*Proc print data=book; *run;                                    /* check point */


*histogram of Money Spent;
proc univariate data=book normal plot;                        /*univariate analysis*/
   title "Distributions of the Amount of Money Spent on Books";
   histogram / normal midpoints=0 to 500 by 5
   kernel(color=red);                        /*type of histogram and graph specifics */
   var  Purchased;                              /*variables of interest in histogram*/
run;
```

Univariate Analysis and Histogram

Variable: Purchase period (time between first and last purchase)



Clearly, the distribution of the purchase period is not a normal distribution. It looks more like a Gamma distribution (Γ (α=2, λ=2)).

Univariate Output for the Purchase Period for BBC Books

Distributions of the Purchase Period

The UNIVARIATE Procedure
Variable: PurchasePeriod

Moments

N	2300	Sum Weights	2300
Mean	19.7886957	Sum Observations	45514
Std Deviation	13.5005567	Variance	182.265031
Skewness	1.10841191	Kurtosis	1.16402603
Uncorrected SS	1319690	Corrected SS	419027.306
Coeff Variation	68.2235804	Std Error Mean	0.28150607

Basic Statistical Measures

Location		Variability	
Mean	19.78870	Std Deviation	13.50056
Median	16.00000	Variance	182.26503
Mode	15.00000	Range	83.00000
		Interquartile Range	17.00000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	70.29581	Pr > |t|	<.0001
Sign	M	1150	Pr >= |M|	<.0001
Signed Rank	S	1323075	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value------	
Kolmogorov-Smirnov	D	0.131808	Pr > D	<0.0100
Cramer-von Mises	W-Sq	8.958165	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	51.22486	Pr > A-Sq	<0.0050

Quantiles (Definition 5)

Quantile	Estimate
100% Max	84
99%	62
95%	46
90%	38
75% Q3	27
50% Median	16
25% Q1	10
10%	5

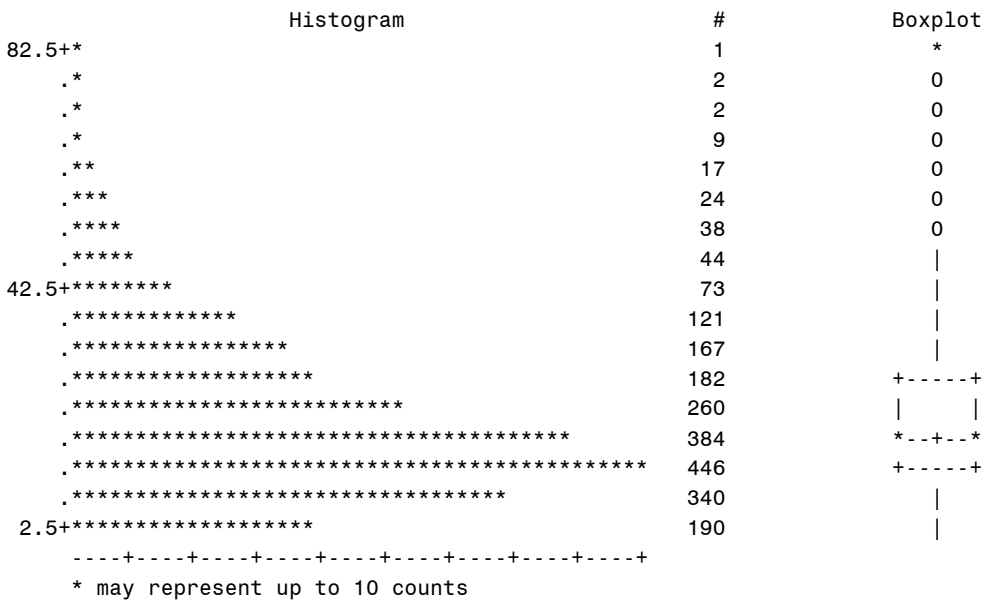
The UNIVARIATE Procedure
Variable: PurchasePeriod

Quantiles (Definition 5)

Quantile	Estimate
5%	3
1%	1
0% Min	1

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
1	2263	70	701
1	2261	73	1342
1	2237	76	164
1	2116	77	2097
1	2084	84	1967



```
                          The UNIVARIATE Procedure

                          Variable:  PurchasePeriod


                                Normal Probability Plot
                  82.5+                                                  *
                      |                                                  *
                      |                                                  *
                      |                                                  *
                      |                                                ****
                      |                                              ***
                      |                                            ****  +++
                      |                                          ** ++++
                  42.5+                                      ******++
                      |                                   ****+
                      |                                ****
                      |                            +****
                      |                          +****
                      |                       +*****
                      |                   ******
                      |             ********
                   2.5+*************+++
                      +----+----+----+----+----+----+----+----+----+----+
                          -2        -1         0        +1        +2
```

Annotated SAS program for Univariate Analysis and Histogram

```
*reading data into SAS;
filename inf "C:\Data\book.csv";
data book;
infile inf firstobs=1 dlm=",";
input  Choice Gender Purchased Frequency LastPurchase  FirstPurchase Childbook
       Youthbook Cookbook DIYbook Artbook;
*Proc print data=book; *run;                                        /* check point */

data book2;                                 /* creates new data to add a new variable */
set book;                                          /* calls original data in action */
PurchasePeriod = FirstPurchase - LastPurchase;          /* new variable is created */


*histogram of Purchase Period;
proc univariate data=book normal plot;                          /*univariate analysis*/
   title "Distributions of the Purchase Period";
   histogram / normal midpoints=0 to 500 by 5
   kernel(color=red);                       /*type of histogram and graph specifics */
   var  PurchasePeriod;                         /*variables of interest in histogram*/
run;
```

Two-sample t-test

a) I want to if the average amount of money spent on BBBC books is the same for different genders. I want to know this to determine which demographics I should target for the direct mailing campaign for the new coming book. Then my claim is that the average money spent is the same in different genders. To test this claim, I created two sample sets out of my data. The amount of money spent on BBBC books by Male and the amount of money spent on BBBC books by Female. Each sample set now has its own average and own standard deviation.

Two-sample t-test will compare these two sample averages ($\bar{x}_1$ and $\bar{x}_2$). Here is $\bar{x}$ is the average amount of money spent on BBBC books. If the difference between $\bar{x}_1$ and $\bar{x}_2$ is small enough to be explained by sampling variation, then the difference will not be statistically significant. Hence we will accept the null hypothesis that the average money spent is indeed the same for different genders.

The TTEST Procedure

Variable: Purchased

Gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	721	198.1	95.5136	3.5571	15.0000	458.0
1	1579	194.0	96.3274	2.4241	15.0000	461.0
Diff (1-2)		4.1806	96.0732	4.3182		

Gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		198.1	191.2	205.1	95.5136	90.8252	100.7
1		194.0	189.2	198.7	96.3274	93.0811	99.8101
Diff (1-2)	Pooled	4.1806	-4.2875	12.6487	96.0732	93.3744	98.9338
Diff (1-2)	Satterthwaite	4.1806	-4.2635	12.6247			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2298	0.97	0.3331
Satterthwaite	Unequal	1405.7	0.97	0.3316

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	1578	720	1.02	0.7960

The two sample t-test suggests that we should accept the null hypothesis. Hence, we conclude that the average spending is indeed the same for different genders.

b) Assuming that we have a population with true mean and true variance, we can create samples from that population with their own means and variances. Then, by taking the difference in sample means and sample variances, we create a new population of the differences of sample means and the differences of sample variances. This new population, as well, has its own mean and variance.

Two-sample t-test compares the means of two samples created from a population. In order to make the comparison, t-test normalizes the difference between two sample means by dividing them with the new created variance, which is the variance of new population of the difference of sample variances. The calculation yields a t-statistic in the output.

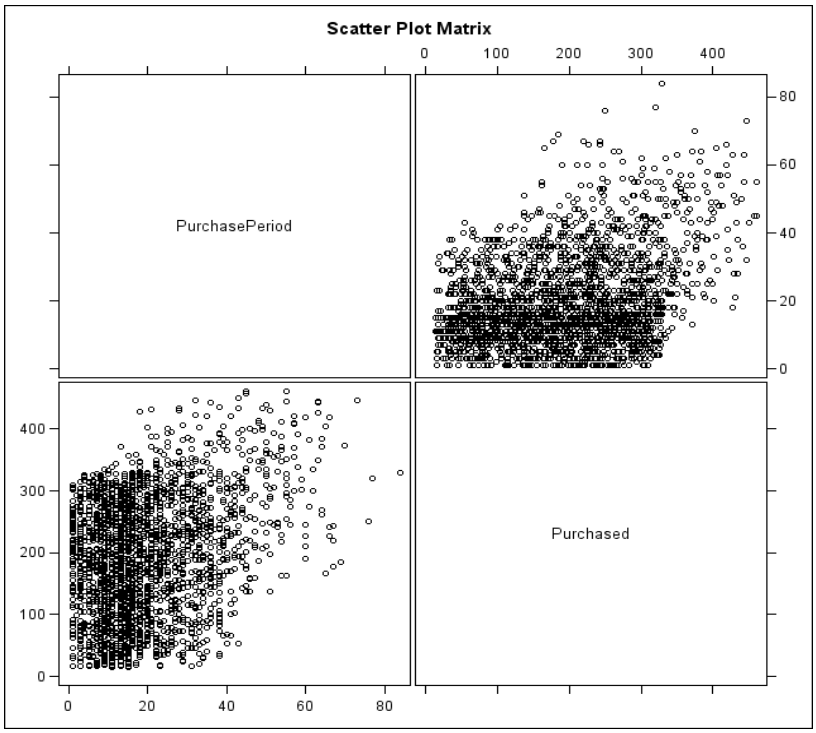
To make a decision of accepting or rejecting the null hypothesis, we compare the t-statistic with the t-distribution. The t distribution we use to compare has a $(n_1+n_2 - 2)$ degrees of freedom and 0 mean. The p-value given in the output, gives the probability of observing a t-stat as extreme or more extreme than the observed value if the null hypothesis were true.

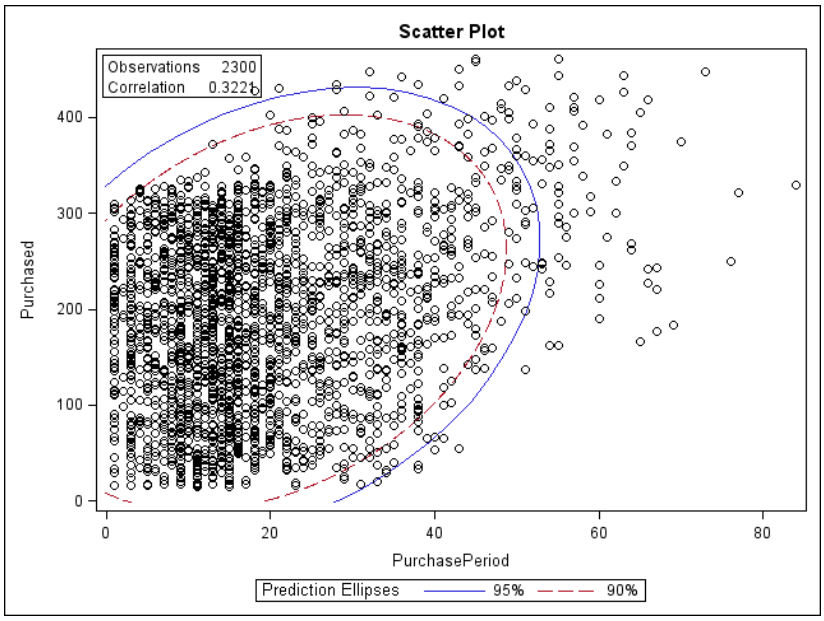
c) Annotated SAS program for two-sample t-test

```
*reading data into SAS;
filename inf "C:\Data\book.csv";
data book;
infile inf firstobs=1 dlm=",";
input  Choice Gender Purchased Frequency LastPurchase  FirstPurchase Childbook
       Youthbook Cookbook DIYbook Artbook;

data book2;                                  /* creates new data to add a new variable */
set book;                                          /* calls original data in action */
PurchasePeriod = FirstPurchase - LastPurchase;           /* new variable is created */

*t-test (to see if mean differs for different gender;
proc ttest data=book2;                                    /* two sample t-test procedure */
class Gender;                                        /* samples are created by Gender */
var Purchased; run;                          /* Average of interest is the money spent*/
```

Regression Analysis

a) Scatter plots (Amount of Money Spent vs. Purchase Period)





It is important to review the scatter plots before doing regression because an analyst should expect to see a linear relationship between variables. Here, scatter plots do not suggest a perfect liner relationship between the purchase period and the amount spent on BBBC books. Then, it's possible that the model should include more than one predictor for response.

b) Assuming that the Purchase Period is a good predictor for the Amount of Money Spent on BBBC books, my model should have the following mean function:

E (The Amount of Money Spent | Purchase Period) = $\beta_0 + \beta_1$(Purchase Period) + ε

This means that one unit change in the purchase period will change the expected amount of money spent on BBBC book by a factor of β_1.

The mean function was obtained by using the expected amount of money spent for each different purchase periods. Hence each person with the same purchase period has an expected average for the amount of money spending.

c)

The SAS System
The REG Procedure
Model: MODEL1
Dependent Variable: Purchased

Number of Observations Read 2300
Number of Observations Used 2300

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2200858	2200858	265.93	<.0001
Error	2298	19018465	8276.09435		
Corrected Total	2299	21219323			

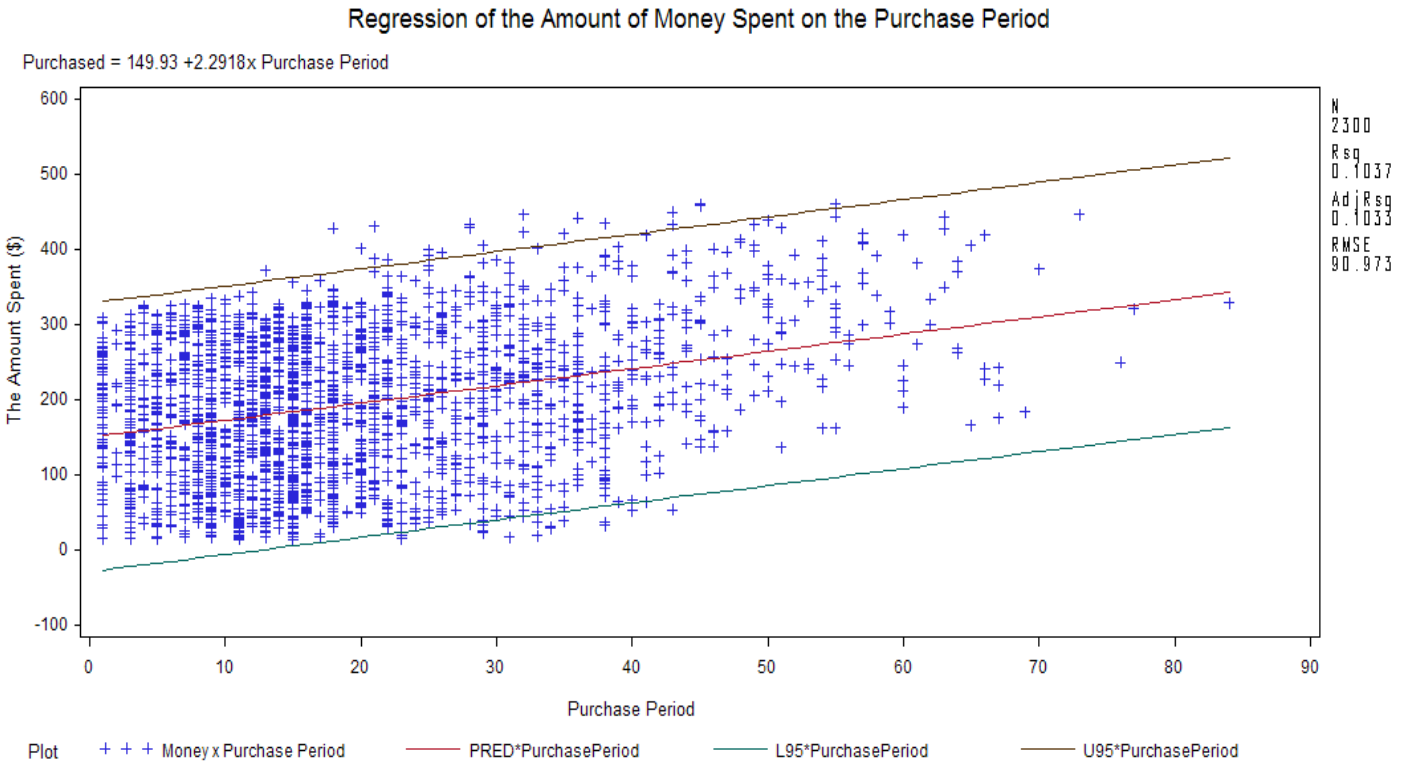
Root MSE	90.97304	R-Square	0.1037
Dependent Mean	195.27696	Adj R-Sq	0.1033
Coeff Var	46.58668		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	149.92542	3.36638	44.54	<.0001
PurchasePeriod	1	2.29179	0.14054	16.31	<.0001

The slope parameter has been estimated as 2.29. The p-value for the estimated slope parameter is very close to zero. This means that if the slope parameter (β_1) were equal to zero, the probability of observing a sample slope as for of farther from zero as 2.29 would be less than 0.01%. Using conventional decision point of 5%, we conclude that such a sample slope would be highly unlikely to occur if $\beta_1 = 0$. Then we reject the claim that $\beta_1 = 0$ and hence the relationship between the amount of money spent on BBBC books and the purchase period is statistically significant.

d)

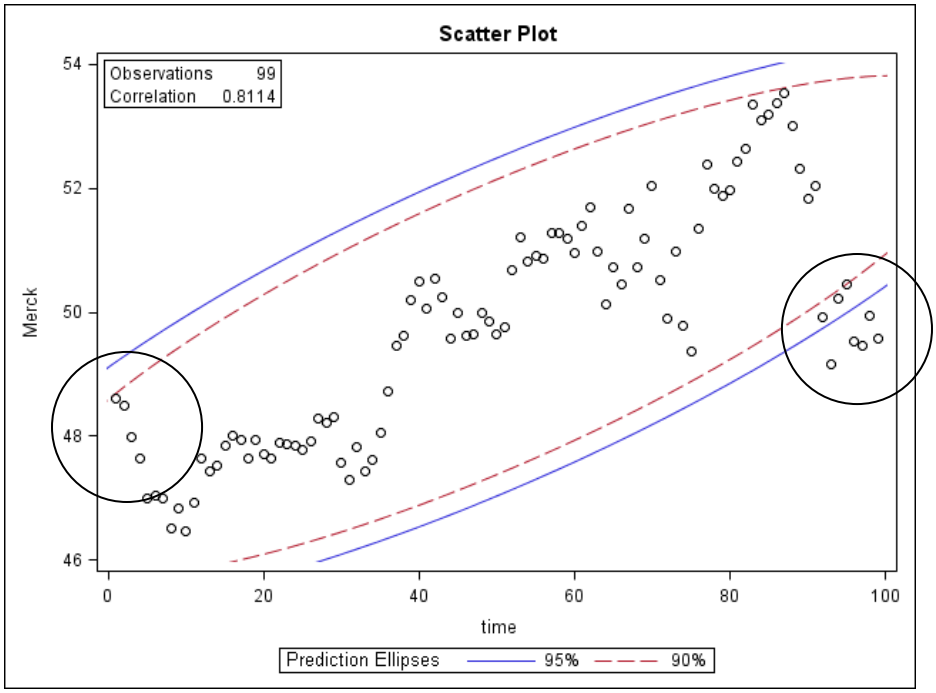


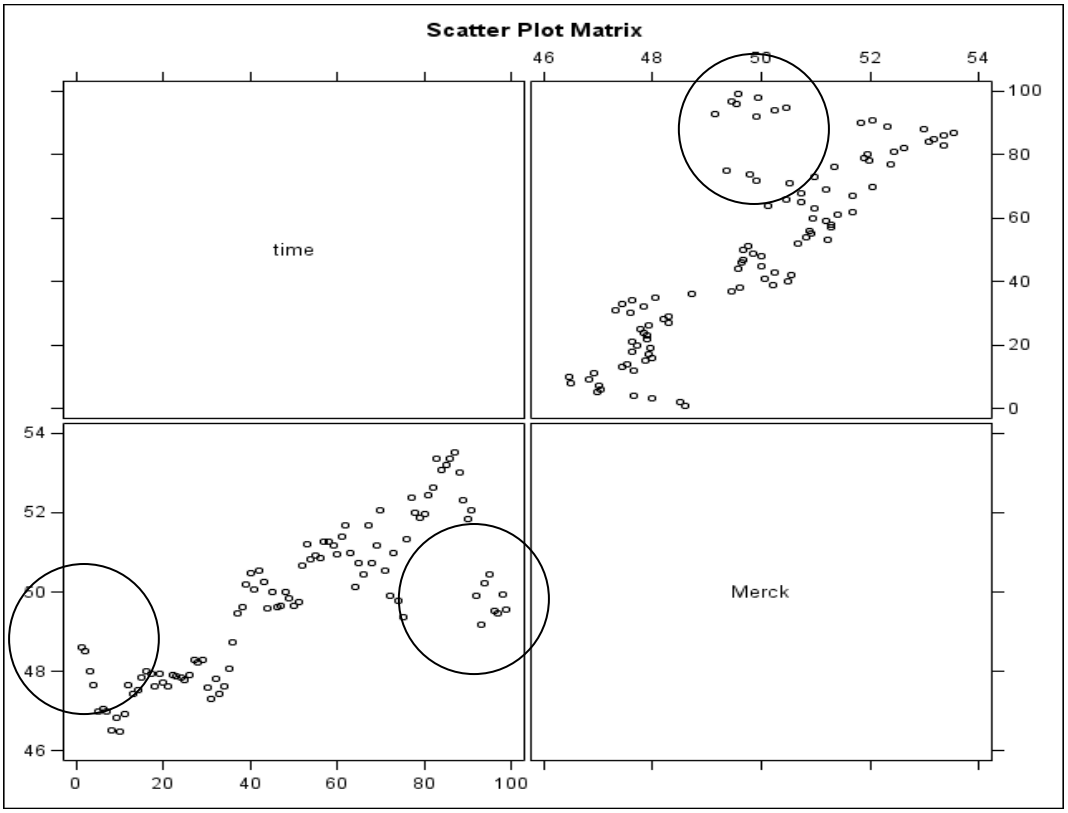
e) Annotated SAS program

```
*reading data into SAS;
filename inf "C:\Data\book.csv";
data book;
infile inf firstobs=1 dlm=",";
input  Choice Gender Purchased Frequency LastPurchase  FirstPurchase Childbook
       Youthbook Cookbook DIYbook Artbook;
data book2;                                  /* creates new data to add a new variable */
set book;                                      /* calls original data in action */
PurchasePeriod = FirstPurchase - LastPurchase;          /* new variable is created */

*scatterplot matrix;
ods graphics on;                                    /* turn Output Delivery System on */
   proc corr data=book2                                    /* correlation estimation */
             plots=scatter (alpha=.05 .1);              /* sets prediction intervals */
      var PurchasePeriod Purchased Frequency;                /* specifies variables */
    run;
ods graphics off;                                  /* turn Output Delivery System off */

*regression;
title "Regression of the Amount of Money Spent on the Purchase Period";
proc reg data=book2;                                                  /* regression */
model Purchased = PurchasePeriod; run;                          /* speficies model */
plot Purchased*PurchasePeriod / pred;  /*variables to plot and prediction intervals */
run;
```

Dangers of Regressing Random Walk

My data do not have random walk. Therefore, I used Merck's stock price and I regressed it on time to show the dangers of regressing a random walk on time. As it can be seen from the scatter plot matrices below, the relationship is highly misleading. A random walk cannot come back to itself however our graphs show the otherwise (see the circles).





The REG Procedure
Model: MODEL1
Dependent Variable: Merck

Number of Observations Read	99
Number of Observations Used	99

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	221.39564	221.39564	186.87	<.0001
Error	97	114.91912	1.18473		
Corrected Total	98	336.31476			

Root MSE	1.08845	R-Square	0.6583
Dependent Mean	49.74273	Adj R-Sq	0.6548
Coeff Var	2.18817		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	47.12626	0.22046	213.77	<.0001
time	1	0.05233	0.00383	13.67	<.0001

The regression output gives very high t-stat values as if the relationship between the time and the stock price of Merck is statistically important.

Logistic Regression

Recall that Choice represents whether the customer purchased the book titled "The Art History of Florence" or not. 1 corresponds to a purchase and 0 corresponds to a non-purchase. I want to see how the probability of purchase changes with the amount of money spent on BBBC books.

The LOGISTIC Procedure
Model Information

Data Set	WORK.BOOK
Response Variable	Choice
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	Choice	Total Frequency
1	1	204
2	0	2096

Probability modeled is Choice=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1379.744	1373.981
SC	1385.485	1385.462
-2 Log L	1377.744	1369.981

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.7632	1	0.0053
Score	7.7868	1	0.0053
Wald	7.7409	1	0.0054

The LOGISTIC Procedure
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.7615	0.1773	242.4934	<.0001
Purchased	1	0.00212	0.000764	7.7409	0.0054

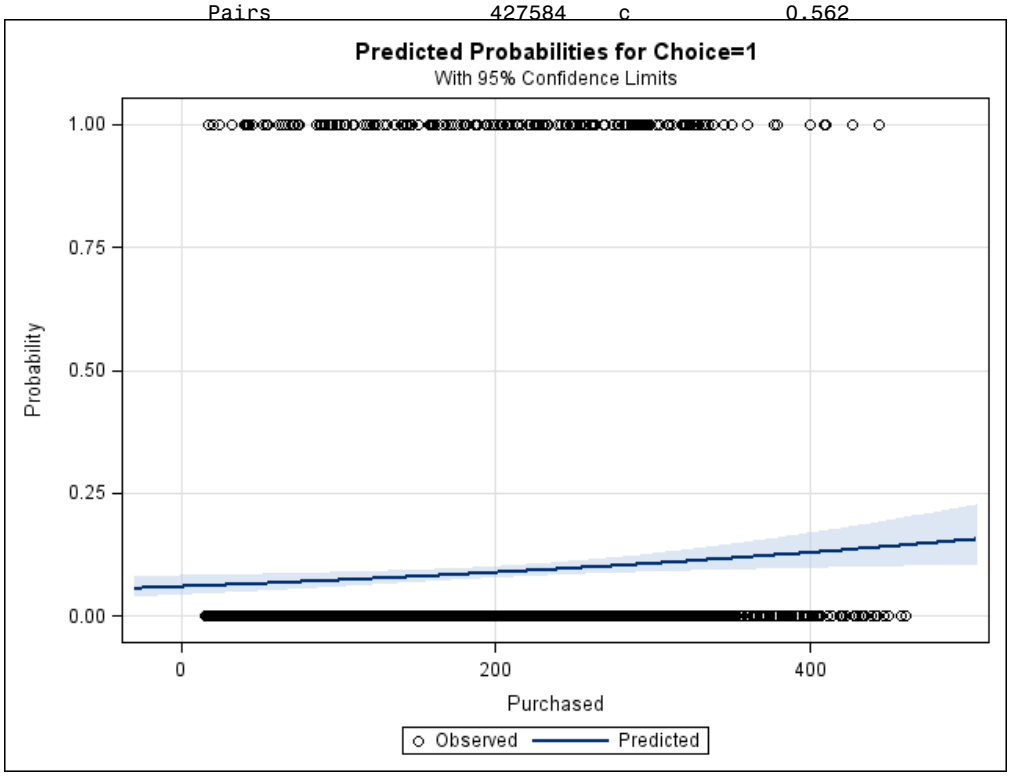
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Purchased	1.002	1.001	1.004

Association of Predicted Probabilities and Observed Responses

Percent Concordant	54.5	Somers' D	0.124
Percent Discordant	42.2	Gamma	0.128

Percent Tied	3.3	Tau-a	0.020
Pairs	427584	c	0.562



The logistic regression graph does not look as expected. This might be another supporting point for the claim that the response depends on more than one predictor. However, the interpretation of the output is that every 1 unit increase in the amount of money spent, the odds for buying the book increases with a factor of $e^{1.002}$.

Annotated SAS program for the Logistic Regression

```
*reading data into SAS;
filename inf "C:\Data\book.csv";
data book;
infile inf firstobs=1 dlm=",";
input  Choice Gender Purchased Frequency LastPurchase  FirstPurchase Childbook
       Youthbook Cookbook DIYbook Artbook;
data book2;                                  /* creates new data to add a new variable */
set book;                                      /* calls original data in action */
PurchasePeriod = FirstPurchase - LastPurchase;           /* new variable is created */

*logistic regression;
ods graphics on;                                    /* turns Output Display System on */
proc logistic descending plots (only) = effect (clbar); /* logistic regression and
                                                      graph with prediction interval */
model Choice = Purchased; run;                        /* sets variables in the model */
```

Multiple Regression

So far, I used a model with one response and one predictor variable. However, I was suspecting that the proposed model was not sufficient enough to state a functional relationship between response and predictor. It gave the clues of how response would behave in presence of particular predictors but didn't give a predictive model. Hence, I will employ a multiple regression analysis to see if I can find a predictive model by using all available predictors.

The REG Procedure

Number of Observations Read	2300
Number of Observations Used	2300

Descriptive Statistics

Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	2300.00000	1.00000	2300.00000	0	0
PurchasePeriod	45514	19.78870	1319690	182.26503	13.50056
Frequency	30600	13.30435	562776	67.70899	8.22855
Childbook	1676.00000	0.72870	3578.00000	1.02510	1.01247
Youthbook	788.00000	0.34261	1186.00000	0.39844	0.63122
Cookbook	1807.00000	0.78565	3993.00000	1.11932	1.05798
DIYbook	934.00000	0.40609	1542.00000	0.50575	0.71116
Artbook	759.00000	0.33000	1115.00000	0.37605	0.61323
Purchased	449137	195.27696	108925429	9229.80538	96.07188

Correlation

Variable	Purchase Period	Frequency	Childbook	Youthbook
PurchasePeriod	1.0000	0.5835	0.5000	0.3716
Frequency	0.5835	1.0000	-0.0149	-0.0149
Childbook	0.5000	-0.0149	1.0000	0.2796
Youthbook	0.3716	-0.0149	0.2796	1.0000
Cookbook	0.4702	-0.0205	0.2807	0.2416
DIYbook	0.3782	-0.0245	0.2304	0.2161
Artbook	0.3617	-0.0097	0.2592	0.1427
Purchased	0.3221	-0.0104	0.3009	0.2383

Correlation

Variable	Cookbook	DIYbook	Artbook	Purchased
PurchasePeriod	0.4702	0.3782	0.3617	0.3221
Frequency	-0.0205	-0.0245	-0.0097	-0.0104
Childbook	0.2807	0.2304	0.2592	0.3009
Youthbook	0.2416	0.2161	0.1427	0.2383
Cookbook	1.0000	0.2377	0.2190	0.2869
DIYbook	0.2377	1.0000	0.2102	0.2429
Artbook	0.2190	0.2102	1.0000	0.2323
Purchased	0.2869	0.2429	0.2323	1.0000

The REG Procedure
Model: FULL
Dependent Variable: Purchased

Number of Observations Read	2300
Number of Observations Used	2300

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	3925791	560827	74.33	<.0001
Error	2292	17293531	7545.17076		
Corrected Total	2299	21219323			

Root MSE	86.86294	R-Square	0.1850
Dependent Mean	195.27696	Adj R-Sq	0.1825
Coeff Var	44.48192		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	158.08858	4.05424	38.99	<.0001
PurchasePeriod	1	1.42308	0.31972	4.45	<.0001
Frequency	1	-1.39450	0.38515	-3.62	0.0003
Childbook	1	10.32082	2.34115	4.41	<.0001
Youthbook	1	11.83323	3.26920	3.62	0.0003
Cookbook	1	9.48416	2.17098	4.37	<.0001
DIYbook	1	10.92493	2.94038	3.72	0.0002
Artbook	1	12.47801	3.33942	3.74	0.0002

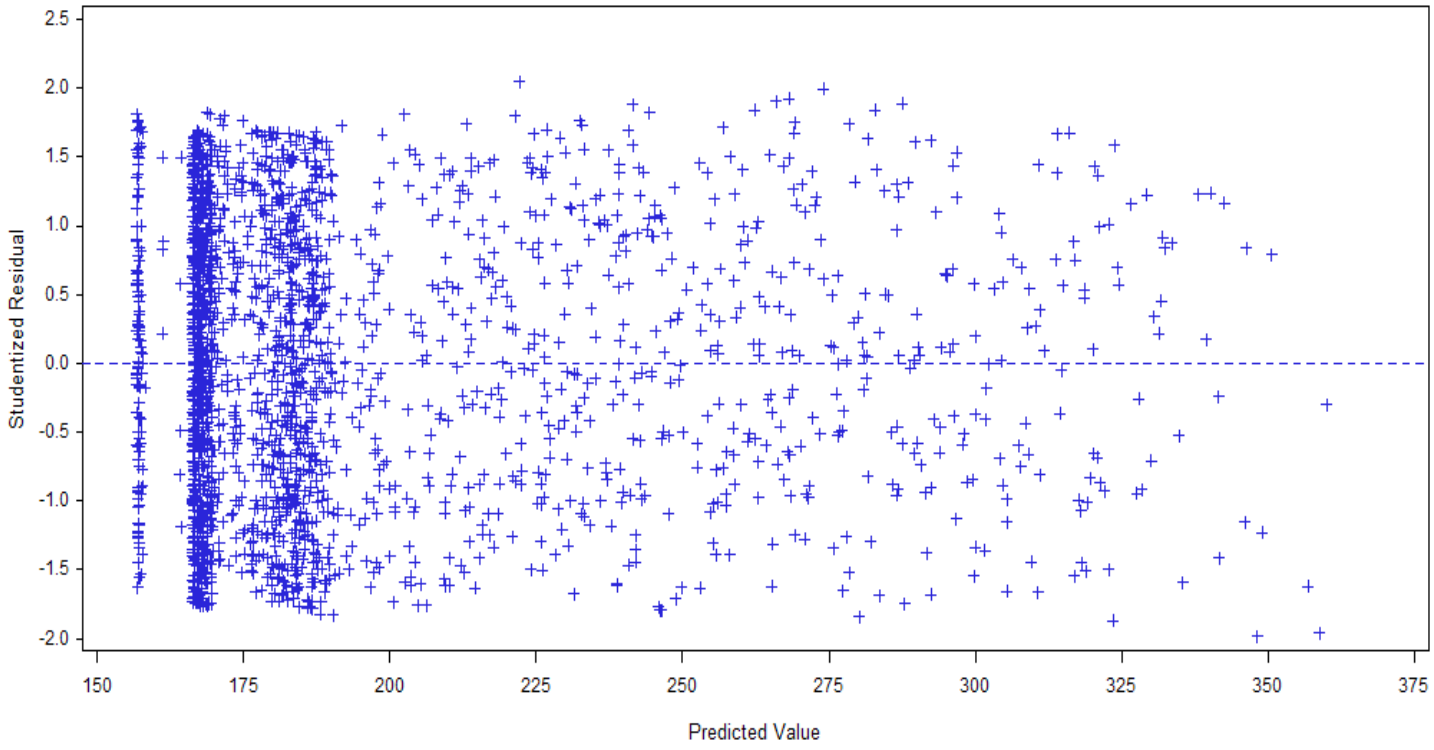
The coefficients of predictors seem statistically significant; hence the proposed model is successful.

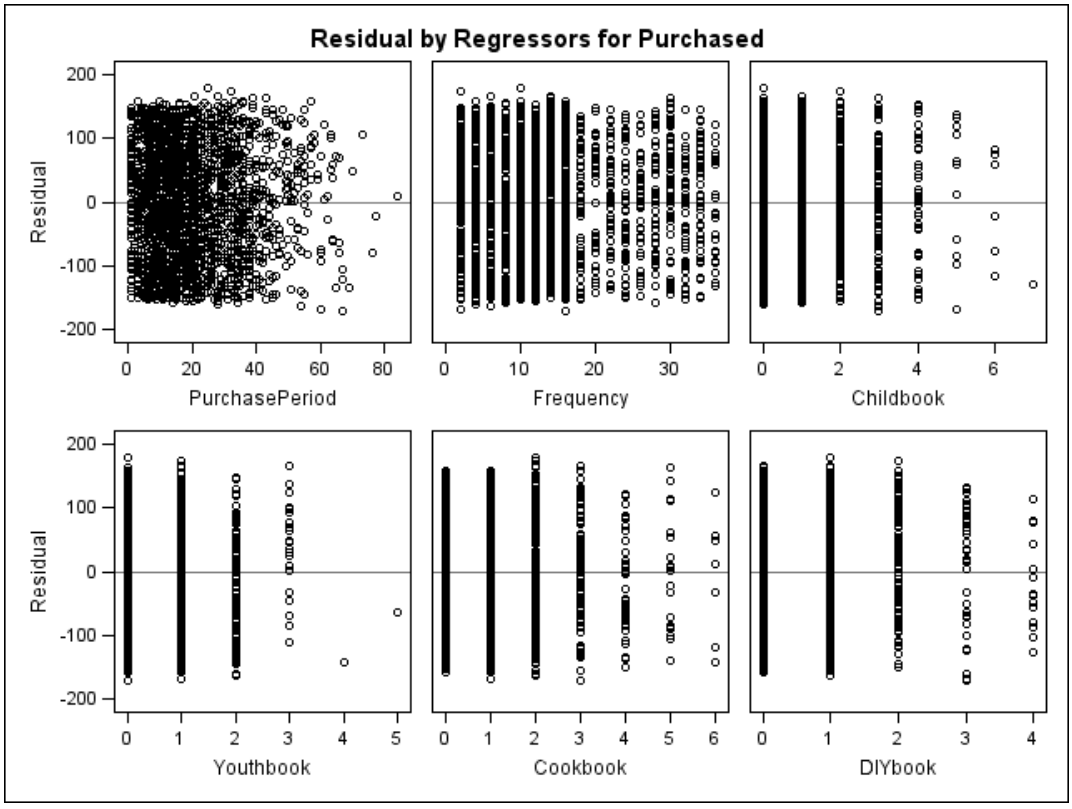
The predictive model is:

$E(Purchase|X=x) = 158.09 + 1.423*PurchasePeriod – 1.394*Frequency + 10.321*Childbook +$

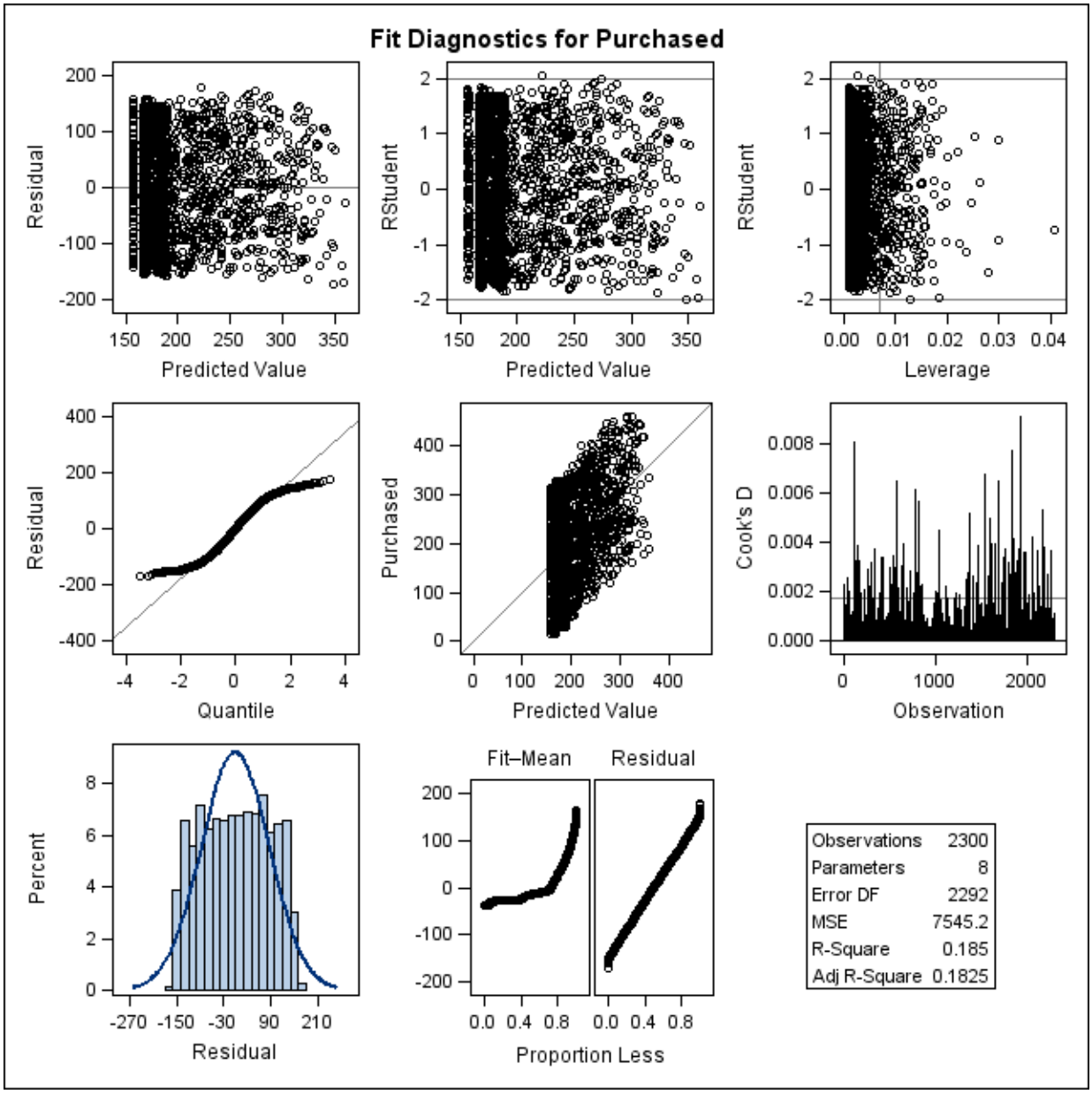
$10.925*DIYbook + 12.478*Artbook$

Purchased = 158.09 +1.4231PurchasePeriod -1.3945 Frequency +10.321Childbook +11.833Youthbook +9.4842 Cookbook +10.925DIYbook +12.478 Artbook





The residual plot gives a pretty good null plot, suggesting that the model is successful.



Annotated SAS Program

```
proc reg data=book2 simple corr;
FULL: model Purchased = PurchasePeriod  /* regression w/ all variables (FULL model)*/
      Frequency Childbook Youthbook Cookbook DIYbook Artbook
         plot student.*predicted.; /*plotting fitted vs predicted variables */
run;
```