**Neil Yetz**

**05/12/2016**

**ERHS 642: Applied Logistic Regression**

**Dr. Bachand**

**ERHS 642 Logistic Regression Spring 2016 Final Exam**

|  |
| --- |
| **Discussing any part of the exam with others is not permitted**.  Students caught cheating or attempting to cheat will be reported to the CSU Conflict Resolution and Student Conduct Services (CRSCS) office and will receive a grade of 0 on the exam. |

The website Mags4u.com is planning an e-mail marketing campaign. Emails will be sent to customers who have previously bought a magazine subscription at Mags4u.com and who have not opted out of receiving e-mails. Mags4u.com wants to target the ads to specific customers and needs to know which combinations of variables explain purchase probabilities for different magazines.

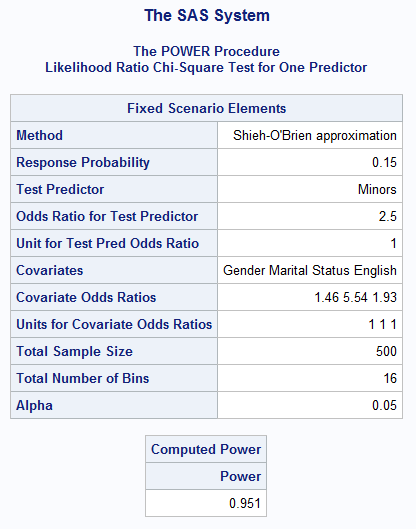
One of the magazines to be advertised is “Kid Creative” whose target audience are children. Mags4u.com sends “experimental” emails containing the ad for “Kid Creative” to 500 randomly selected customers who have previously bought a magazine subscription at Mags4u.com. For each customer, Mags4u.com records whether or not he/she buys a subscription for “Kid Creative” after receiving the ad. For each customer, Mags4u.com combines this information with data collected during the most recent prior subscription purchase and with third party data which Mags4u.com acquired. The following variables are available:

* Customer number (Obs\_No\_)
* Purchased “Kid Creative” (Buy = 1 if purchased “Kid Creative,” 0 otherwise)
* **Household Income (Income; rounded to the nearest $1,000)**
* Gender (Is\_Female = 1 if the person is female, 0 otherwise)
* Marital Status (Is\_Married = 1 if married, 0 otherwise)
* College Educated (Has\_College = 1 if one or more years of college education, 0 otherwise)
* Employed in a Profession (Is\_Professional = 1 if employed in a profession, 0 otherwise)
* Retired (Is\_Retired = 1 if retired, 0 otherwise)
* Not employed (Unemployed = 1 if not employed, 0 otherwise)
* **Length of Residency in Current City (Residence\_Length; in years)**
* Dual Income if Married (Dual\_Income = 1 if dual income, 0 otherwise)
* Children (Minors = 1 if children under 18 are in the household, 0 otherwise)
* Home ownership (Own = 1 if own residence, 0 otherwise)
* Resident type (House = 1 if residence is a single family house, 0 otherwise)
* Race (White = 1 if race is white, 0 otherwise)
* Language (English = 1 is the primary language in the household is English, 0 otherwise)
* Previously purchased a children’s magazine (Prev\_Child\_Mag = 1 if previously purchased a children’s magazine, 0 otherwise)
* Previously purchased a parenting magazine (Prev\_Parent\_Mag = 1 if previously purchased a parenting magazine, 0 otherwise)

(Note: Replicate is the number of your data set. For example, if your data set is kc\_1, then replicate=1)

1. Assume “Minors” is a risk factor of particular interest. Estimate the power to detect an OR of 2.5 for “Minors” in a model containing 3 additional dichotomous variables. Assume (i) that Minors=1 for 30% of controls but that the additional dichotomous variables are equal to 1 for only 5% of controls; (ii) that 15% of ad recipients buy “Kid Creative”; and (iii) that α=0.05. Note that controls are households who did not buy “Kid Creative” after receiving the ad. *(3 points)*

Table 1.01: SAS output for Power Analysis



**Power = 0.951**

1. Build a model based on goal 1. You must include the following steps (each step must be included but you can change the order or add other steps):

Table 2.01: Frequency tables of all variables separated by those who bought Kid Creative magazine and those that did not.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Level | BUY | | No Buy | |
|  |  | Frequency | Percent | Frequency | Percent |
| Gender | Otherwise | 35 | 37.63 | 200 | 49.14 |
|  | Female | 58 | 62.37 | 207 | 50.86 |
|  |  |  |  |  |  |
| Marital Status | Otherwise | 32 | 34.41 | 304 | 74.69 |
|  | Married | 61 | 65.59 | 103 | 25.31 |
|  |  |  |  |  |  |
| College Educated | Otherwise | 52 | 55.91 | 313 | 76.9 |
|  | one or more years of college education | 41 | 44.09 | 94 | 23.1 |
|  |  |  |  |  |  |
| Employed in a profession | Otherwise | 42 | 45.16 | 292 | 71.74 |
|  | Employed in a profession | 51 | 54.84 | 115 | 28.26 |
|  |  |  |  |  |  |
| Retired | Otherwise | 88 | 94.62 | 382 | 93.86 |
|  | Retired | 5 | 5.38 | 25 | 6.14 |
|  |  |  |  |  |  |
| Not Employed | Otherwise | 92 | 98.92 | 391 | 96.07 |
|  | Not Employed | 1 | 1.08 | 16 | 3.93 |
|  |  |  |  |  |  |
| Dual Income If Married\* | Otherwise | 21 | 12.80 | 43 | 26.22 |
|  | Dual Income | 40 |  | 60 | 36.59 |
|  |  |  |  |  |  |
| Children | Otherwise | 56 | 60.22 | 262 | 64.73 |
|  | Children under 18 are in the household | 37 | 39.78 | 145 | 35.63 |
|  |  |  |  |  |  |
| Home Ownership | Otherwise | 27 | 29.03 | 297 | 72.97 |
|  | Own residence | 66 | 70.97 | 110 | 27.03 |
|  |  |  |  |  |  |
| Resident Type | Otherwise | 15 | 16.13 | 155 | 38.08 |
|  | Residence is a single family house | 78 | 83.87 | 252 | 61.92 |
|  |  |  |  |  |  |
| Race | Otherwise | 12 | 12.9 | 139 | 34.15 |
|  | White | 81 | 87.1 | 268 | 65.85 |
|  |  |  |  |  |  |
| English | Otherwise | 5 | 5.38 | 42 | 10.32 |
|  | The Primary Language in the Household is English | 88 | 94.62 | 365 | 89.68 |
|  |  |  |  |  |  |
| Table 2.01: Continued |  |  |  |  |  |
| Previously purchased a children's magazine | Otherwise | 80 | 86.02 | 387 | 95.09 |
|  | Previously purchased a children's magazine | 13 | 13.98 | 20 | 4.91 |
|  |  |  |  |  |  |
| Previously purchased a parenting magazine | Otherwise | 83 | 89.25 | 373 | 91.65 |
|  | Previously purchased a parenting magazine | 10 | 10.75 | 34 | 8.35 |

Dual Income if married N = 164. This variable is dependent on if participant was married or not.

* 1. Assessment of the frequencies of the categorical study variables stratified by the outcome variable; description of sparse cells and how you are dealing with them *(2 points)*

Summary for Table 1:

* Gender
  + No Sparse cells
  + Female more likely to buy Magazine than Otherwise
* Marital Status
  + No Sparse Cells
  + Married more likely to buy Magazine than Otherwise
* College Educated
  + No Sparse Cells
  + Otherwise more likely NOT to buy magazine than one or more years of college education.
* Employed in a profession
  + No Sparse Cells
  + Otherwise more likely to buy magazine than Employed in a profession
* Retired
  + One Sparse Cell; Not including in Main effects model
  + Can’t combine categories
* Not Employed
  + One Sparse Cell; Consider not including in main effects model
  + Can’t combine categories
* Dual income if Married
  + No Sparse cells
  + Otherwise more likely to NOT buy magazine
* Children
  + No Sparse Cells
* Home Ownership
  + No Sparse cells
* Resident Type
  + No Sparse Cells
  + More likely to be a residence is a single house
* Race
  + One potential Sparse Cell; continue to include
  + More likely to be white
* English
  + One Sparse Cell;
* Previously purchased a parenting magazine
  + One potential sparse cell; continue to include
  + More likely to be otherwise
* Previously purchase a parenting magazine
  + One potential sparse cell; continue to include
  1. Assessment of the descriptive statistics of the continuous study variables stratified by the outcome variable *(1 point)* Assessment of the extreme observations of the continuous study variables stratified by the outcome variable *(1 point)*

Description of unusual values and how you are dealing with them *(1 point)*

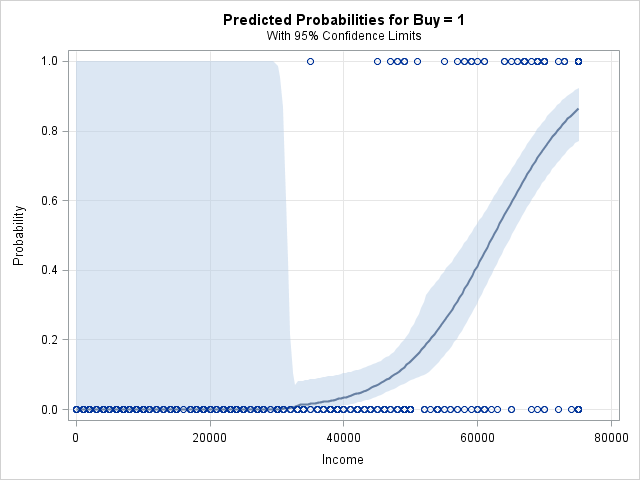
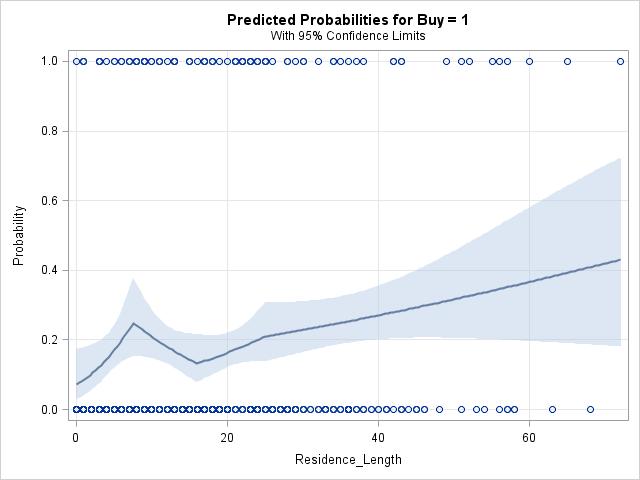
Table 2.02: Descriptive statistics of Continuous variables separated by if they bought Kid Cretaive Magazine or did not.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Household Income | | Length of Residency | |
|  | Buy | No Buy | Buy | No Buy |
| Mean | 68720.43 | 27587.22 | 27.7 | 17.23 |
| Std | 9089.37 | 18998.75 | 15.93 | 13.03 |
| Min | 35000 | 0 | 0 | 0 |
| 1st Quartile | 66000 | 10000 | 9 | 7 |
| Median | 75000 | 25000 | 19 | 16 |
| 3rd Quartile | 75000 | 42000 | 28 | 24 |
| Maximum | 75000 | 75000 | 72 | 68 |
| Lowest Values | 3500, 45000, 47000, 48000, 48000 | 0, 0, 0, 0, 0 | 0, 1, 1, 1, 3 | 0, 0, 0, 0, 0 |
| Highest Values | 75000, 75000, 75000, 75000, 75000 | 75000, 75000, 75000, 75000, 75000 | 56, 57, 60, 65, 72 | 57, 58, 58, 63, 68 |

Table 2 Summary

* Household income
  + Lots of Zero’s. May need to consider separating 0 from other income.
  + Capped at 75,000?
  + 0’s may be outliers
  + Keep for now
* Length of Residency
  + Lots of Zero’s. May need to consider separating 0 from other income.
  + 0’s may be outliers.
  + Keep for now
  1. Univariate scale assessment (spline plots and fp) including conclusions *(4 points)*
* There were 93 study participants that bought the magazine and 407 that did not
* Least Frequent outcome divided by 10 = 93/10 ≈ 9
* Model should contain no more than 9 covariates.

Figure 2.01: Univariate scale assessments for Household income and Residence Length

Residence length relatively linear.

Consider transforming Income

For now, create the following 2-level categorical variable.

Incomea =0 If income <= 40000

Incomea=1 If income > 40000

Table 2.03: Deviance and p-values, and suggested transformations from fp method for Income and Length of Residency

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Transformation | Deviance | Comparison | P-value |
| Income | Linear | 186.808 | Best 1-power vs. linear | 0.4508 |
|  | Best 1-power | 186.24 | Best 2-power vs. linear | 0.8931 |
|  | Best 2-power and x | 186.194 | Best 2-power vs. 1-power | 0.9773 |
|  |  |  |  |  |
| Length of residency | Linear | 465.024 | Best 1-power vs. linear | 0.4198 |
|  | Best 1-power x3 | 464.373 | Best 2-power vs. linear | 0.652 |
|  | Best 2-power and x3 | 463.405 | Best 2-power vs. 1-power | 0.6164 |

No recommended transformations

* 1. Univariate statistical significance assessment including conclusions *(2 points)*

Table 2.04: Univariate logistic regression assessment of variabels. Shaded areas indicate statistical significance at 0.25 level.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 1.601 | 1.008 | 2.542 | 0.046 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 5.626 | 3.472 | 9.116 | <0.0001 |
|  |  |  |  |  |  |
| College Educated | one or more years of college education vs Otherwise | 2.625 | 1.641 | 4.2 | <0.0001 |
|  |  |  |  |  |  |
| Employed in a profession | Employed in a profession vs Otherwise | 3.083 | 1.943 | 4.893 | <0.0001 |
|  |  |  |  |  |  |
| Retired | Retired vs Otherwise | 0.868 | 0.323 | 2.331 | 0.7791 |
|  |  |  |  |  |  |
| Not Employed | Not Employed vs Otherwise | 0.266 | 0.035 | 2.029 | 0.2012 |
|  |  |  |  |  |  |
| Dual Income If Married | Dual Income vs Otherwise | 1.365 | 0.707 | 2.635 | 0.3536 |
|  |  |  |  |  |  |
| Children | Children under 18 are in the household vs Otherwise | 1.194 | 0.752 | 1.895 | 0.4524 |
|  |  |  |  |  |  |
| Home Ownership | Own residence vs Otherwise | 6.599 | 4.009 | 10.861 | <0.0001 |
|  |  |  |  |  |  |
| Resident Type | Residence is a single family house vs Otherwise | 3.198 | 1.777 | 5.756 | <0.0001 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 3.5 | 1.846 | 6.636 | 0.0001 |
|  |  |  |  |  |  |
| Language | Primary language of the House is English vs otherwise | 2.025 | 0.779 | 5.268 | 0.148 |
| Table 2.04: Continued |  |  |  |  |  |
| Previously purchased a children's magazine | Previously purchased a children's magazine vs Otherwise | 3.144 | 1.502 | 6.581 | 0.0024 |
|  |  |  |  |  |  |
| Previously purchased a parenting magazine | Previously purchased a parenting magazine vs Otherwise | 1.322 | 0.628 | 2.782 | 0.4623 |
|  |  |  |  |  |  |
| IncomeA | >40000 vs <=40000 | 254.702 | 35.067 | >999.999 | <0.0001 |
|  |  |  |  |  |  |
| Income | 5000 | 2.134 | 1.829 | 2.49 | <0.0001 |
|  |  |  |  |  |  |
| Length of Residency | 1 | 1.022 | 1.007 | 1.038 | 0.0051 |

Notes: Keeping income as continuous. Dropping Income categorical variables IncomeA due to wide confidence intervals in order to keep model stability.

Significant variables (p<0.25):

* Gender
* Marital Status
* College Educated
* Employed in a profession
* Not employed
* Home ownership
* Resident Type
* Race
* Previously purchased a children’s magazine
* Income
* Length of Residency

Non-significant variables (p≥0.25):

* Retired
* Children
* Previously purchased a parenting magazine
* Dual income if married

* 1. Multivariate statistical significance/confounding assessment including your reasoning along the way and conclusions *(4 points)*

Table 2.05: Multivariate logistic regression results – Initial main effects model (contains variables significant at the 0.25 level in univariate analyses)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 8.886 | 3.036 | 26.011 | <0.0001 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 3.143 | 1.067 | 9.257 | 0.0377 |
|  |  |  |  |  |  |
| College Educated | one or more years of college education vs Otherwise | 0.497 | 0.183 | 1.345 | 0.1687 |
|  |  |  |  |  |  |
| Table 2.05: Continued |  |  |  |  |  |
| Employed in a profession | Employed in a profession vs Otherwise | 1.066 | 0.409 | 2.777 | 0.8965 |
|  |  |  |  |  |  |
| Home Ownership | Own residence vs Otherwise | 2.319 | 0.702 | 7.657 | 0.1674 |
|  |  |  |  |  |  |
| Resident Type | Residence is a single family house vs Otherwise | 0.333 | 0.088 | 1.262 | 0.1057 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 3.230 | 1.043 | 10.007 | 0.0421 |
|  |  |  |  |  |  |
| English |  | 3.166 | 0.583 | 17.184 | 0.1818 |
|  |  |  |  |  |  |
| Previously purchased a parenting magazine | Previously purchased a parenting magazine vs Otherwise | 2.306 | 0.612 | 8.690 | 0.2172 |
|  |  |  |  |  |  |
| Income | 5000 | 2.951 | 2.223 | 3.918 | <0.0001 |
|  |  |  |  |  |  |
| Length of Residency | 1 | 0.995 | 0.967 | 1.024 | 0.7241 |

Non-significant at P<0.05 = College educated, Employed in a profession, Home ownership, Resident type, Previously purchased a parent magazine, Length of Residency. Remove one by one.

Table 2.06: Mulitvariate logistic regression results, Employed in a profession variable removed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 8.897 | 3.039 | 26.048 | <0.0001 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 3.145 | 1.069 | 9.257 | 0.0375 |
|  |  |  |  |  |  |
| College Educated | one or more years of college education vs Otherwise | 0.505 | 0.192 | 1.327 | 0.1656 |
|  |  |  |  |  |  |
| Home Ownership | Own residence vs Otherwise | 2.337 | 0.713 | 7.665 | 0.1613 |
|  |  |  |  |  |  |
| Resident Type | Residence is a single family house vs Otherwise | 0.329 | 0.087 | 1.238 | 0.1002 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 3.179 | 1.056 | 9.568 | 0.0397 |
|  |  |  |  |  |  |
| English | English vs. Other | 3.243 | 0.621 | 16.946 | 0.1631 |
|  |  |  |  |  |  |
| Previously purchased a parenting magazine | Previously purchased a parenting magazine vs Otherwise | 3.243 | 0.621 | 16.946 | 0.2181 |
|  |  |  |  |  |  |
| Table 2.06 continued |  |  |  |  |  |
| Income | 5000 | 2.959 | 2.232 | 3.922 | <0.0001 |
|  |  |  |  |  |  |
| Length of Residency | 1 | 0.994 | 0.967 | 1.023 | 0.7009 |

Odds Ratios appear unchanged. Little evidence of confounding. Removing Employed as a profession.

Table 2.07: Mulitvariate logistic regression results, Length of residency variable removed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 9.134 | 3.134 | 26.623 | <0.0001 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 3.151 | 1.072 | 9.261 | 0.0369 |
|  |  |  |  |  |  |
| College Educated | one or more years of college education vs Otherwise | 0.501 | 0.191 | 1.316 | 0.1607 |
|  |  |  |  |  |  |
| Home Ownership | Own residence vs Otherwise | 2.231 | 0.700 | 7.112 | 0.1747 |
|  |  |  |  |  |  |
| Resident Type | Residence is a single family house vs Otherwise | 0.333 | 0.088 | 1.262 | 0.0999 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 3.193 | 1.061 | 9.607 | 0.0389 |
|  |  |  |  |  |  |
| English | English vs. Other | 3.206 | 0.617 | 16.647 | 0.1658 |
|  |  |  |  |  |  |
| Previously purchased a parenting magazine | Previously purchased a parenting magazine vs Otherwise | 2.321 | 0.622 | 8.654 | 0.2099 |
|  |  |  |  |  |  |
| Income | 5000 | 2.964 | 2.235 | 3.932 | <0.0001 |

Odds Ratios appear unchanged. Little evidence of confounding. Removing Length of Residency.

Table 2.08: Mulitvariate logistic regression results, Previously purchased a parenting magazine removed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 8.737 | 3.046 | 25.062 | <0.0001 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 3.091 | 1.063 | 8.984 | 0.0382 |
|  |  |  |  |  |  |
| College Educated | one or more years of college education vs Otherwise | 0.540 | 0.208 | 1.405 | 0.2068 |
|  |  |  |  |  |  |
| Home Ownership | Own residence vs Otherwise | 2.254 | 0.714 | 7.117 | 0.1658 |
|  |  |  |  |  |  |
| Table 2.08: continued |  |  |  |  |  |
| Resident Type | Residence is a single family house vs Otherwise | 0.311 | 0.085 | 1.131 | 0.0761 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 3.326 | 1.105 | 10.015 | 0.0326 |
|  |  |  |  |  |  |
| English | English vs. Other | 3.012 | 0.605 | 15.003 | 0.1784 |
|  |  |  |  |  |  |
| Income | 5000 | 2.895 | 2.206 | 3.801 | <0.0001 |

Odds Ratios appear unchanged. Little evidence of confounding.Previously purchased parenting magazine

.

Table 2.09: Mulitvariate logistic regression results, Language removed

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 8.494 | 3.003 | 24.025 | <0.0001 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 3.253 | 1.125 | 9.403 | 0.0294 |
|  |  |  |  |  |  |
| College Educated | one or more years of college education vs Otherwise | 0.592 | 0.232 | 1.507 | 0.2715 |
|  |  |  |  |  |  |
| Home Ownership | Own residence vs Otherwise | 2.317 | 0.738 | 7.280 | 0.1502 |
|  |  |  |  |  |  |
| Resident Type | Residence is a single family house vs Otherwise | 0.287 | 0.080 | 1.033 | 0.0560 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 3.856 | 1.310 | 11.353 | 0.0143 |
|  |  |  |  |  |  |
| Income | 5000 | 2.822 | 2.167 | 3.676 | <0.0001 |

Odds Ratios appear unchanged. Little evidence of confounding. Removing English.

Table 2.10: Mulitvariate logistic regression results, College Educated removed

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 8.445 | 3.014 | 23.664 | <0.0001 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 3.237 | 1.120 | 9.354 | 0.0300 |
|  |  |  |  |  |  |
| Home Ownership | Own residence vs Otherwise | 2.015 | 0.655 | 6.199 | 0.2218 |
|  |  |  |  |  |  |
| Resident Type | Residence is a single family house vs Otherwise | 0.324 | 0.091 | 1.147 | 0.0805 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 4.024 | 1.359 | 11.916 | 0.0119 |
|  |  |  |  |  |  |
| Income | 5000 | 2.723 | 2.123 | 3.493 | <0.0001 |

Odds Ratios appear unchanged. Little evidence of confounding. Removing College educated.

Table 2.11: Mulitvariate logistic regression results, House Ownership removed

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 8.440 | 3.036 | 23.461 | <0.0001 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 4.487 | 1.763 | 11.421 | 0.0016 |
|  |  |  |  |  |  |
| Resident Type | Residence is a single family house vs Otherwise | 0.428 | 0.132 | 1.385 | 0.1565 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 4.182 | 1.425 | 12.272 | 0.0092 |
|  |  |  |  |  |  |
| Income | 5000 | 2.697 | 2.114 | 3.442 | <0.0001 |

Odds Ratios appear unchanged. Marital status OR jumped quite a bit May be evidence of confounding. Leaving out for now. May add in later when assessing interactions.

Table 2.12: Mulitvariate logistic regression results, House Ownership removed

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 8.440 | 3.036 | 23.461 | <0.0001 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 4.487 | 1.763 | 11.421 | 0.0016 |
|  |  |  |  |  |  |
| Resident Type | Residence is a single family house vs Otherwise | 0.428 | 0.132 | 1.385 | 0.1565 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 4.182 | 1.425 | 12.272 | 0.0092 |
|  |  |  |  |  |  |
| Table 2.12: Continued |  |  |  |  |  |
| Income | 5000 | 2.697 | 2.114 | 3.442 | <0.0001 |

Odds Ratios appear unchanged. Marital status OR jumped quite a bit May be evidence of confounding. Leaving out for now. May add in later.

Table 2.13: Mulitvariate logistic regression results, House Ownership removed

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 7.871 | 2.898 | 21.378 | <0.0001 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 3.672 | 1.521 | 8.863 | 0.0038 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 4.452 | 1.522 | 13.025 | 0.0064 |
|  |  |  |  |  |  |
| Income | 5000 | 2.579 | 2.056 | 3.234 | <0.0001 |

Odds Ratios appear unchanged. No evidence of confounding. Keeping out.

Removing final significant variables to analyze Impact:

Table 2.14: Mulitvariate logistic regression results, Gender removed

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Marital Status | Married vs Otherwise | 4.006 | 1.751 | 9.165 | 0.0010 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 4.094 | 1.513 | 11.076 | 0.0055 |
|  |  |  |  |  |  |
| Income | 5000 | 2.237 | 1.872 | 2.672 | <0.0001 |

Odds Ratios appear unchanged. No evidence of confounding. Keeping Putting back in.

Table 2.15: Mulitvariate logistic regression results, Marital Status removed

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 8.444 | 3.173 | 22.470 | <0.0001 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 4.093 | 1.432 | 11.702 | 0.0085 |
|  |  |  |  |  |  |
| Income | 5000 | 2.494 | 2.019 | 3.082 | <0.0001 |

Odds Ratios appear unchanged. No evidence of confounding. Putting back in.

Table 2.16: Mulitvariate logistic regression results, Race removed

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 7.459 | 2.845 | 19.558 | <0.0001 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 3.403 | 1.445 | 8.012 | 0.0051 |
|  |  |  |  |  |  |
| Income | 5000 | 2.508 | 2.025 | 3.106 | <0.0001 |

Odds Ratios appear unchanged. No evidence of confounding. Putting back in.

Table 2.17: Mulitvariate logistic regression results, Income removed

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 1.477 | 0.896 | 2.435 | 0.1262 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 5.789 | 3.528 | 9.501 | <0.001 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 3.698 | 1.894 | 7.219 | <0.0001 |

Odds Ratios appear unchanged. All other variables VERY effected by income. Consider multivariate fp transformation to assess. Gender seems heavily influenced by Income. Will keep income in for now. May be evidence of interaction later.

Adding Minors as it seems an essential to have Minors to buy a magazine centered on kids.

Table 2.18: Mulitvariate logistic regression results, House Ownership removed

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 7.871 | 2.898 | 21.378 | <0.0001 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 3.672 | 1.521 | 8.863 | 0.0038 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 4.452 | 1.522 | 13.025 | 0.0064 |
|  |  |  |  |  |  |
| Income | 5000 | 2.579 | 2.056 | 3.234 | <0.0001 |
|  |  |  |  |  |  |
| Children | Children under 18 are in the household vs others | 1.974 | 0.809 | 4.820 | 0.1352 |

* 1. If your main effects model includes continuous variables: Multivariate scale assessment (fp or design variable plots) including conclusions *(-1 point if needed but missing)*

Table 2.19: Multivariate fp method transformation test on Income variable results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Transformation | Deviance | Comparison | P-value |
| Income | Linear | 143.990 | Best 1-power vs. linear | N/A |
|  | Best 1-power N/A | N/A | Best 2-power vs. linear | 0.9953 |
|  | Best 2-power and x2 | 143.921 | Best 2-power vs. 1-power | N/A |

No suggested transformations

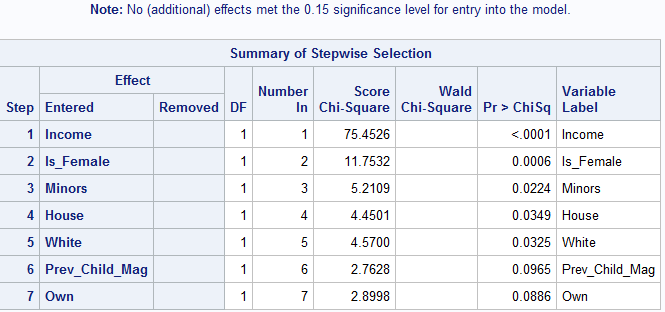
* 1. Presentation of the final main effects model (no interpretation of the results necessary) *(1 point)*

Table 2.20: Final Model before automated selection model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 7.871 | 2.898 | 21.378 | <0.0001 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 3.672 | 1.521 | 8.863 | 0.0038 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 4.452 | 1.522 | 13.025 | 0.0064 |
|  |  |  |  |  |  |
| Income | 5000 | 2.579 | 2.056 | 3.234 | <0.0001 |
|  |  |  |  |  |  |
| Children | Children under 18 are in the household vs others | 1.974 | 0.809 | 4.820 | 0.1352 |

1. Perform best subsets selection to determine if you missed any important model covariates (main effects only). You can use collapsed variables and scale assessment results from question 2. If necessary, make changes to your final model from question 2. *(4 points)*

Table 3.01: Results from automated best subsets selection procedure



Yes, this has changed my model. I have decided to add “owning a house” (Own) and “Type of House” (House) because those seem intuitively predictive of getting a magazine after looking at this created model.

**New Final Main effects model:**

Table 3.02: Final main effects model after automated best subsets selection.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-Value |
| Gender | Female vs Otherwise | 9.825 | 3.303 | 29.224 | <0.0001 |
|  |  |  |  |  |  |
| Marital Status | Married vs Otherwise | 2.564 | 0.851 | 7.719 | 0.0941 |
|  |  |  |  |  |  |
| Race | White vs Otherwise | 5.489 | 1.787 | 16.854 | 0.0029 |
|  |  |  |  |  |  |
| Income | 5000 | 2.969 | 2.233 | 3.949 | <0.0001 |
|  |  |  |  |  |  |
| Children | Children under 18 are in the household vs others | 3.801 | 1.315 | 10.987 | 0.0137 |
|  |  |  |  |  |  |
| Home Ownership | Own residence vs Otherwise | 3.594 | 1.025 | 12.598 | 0.0456 |
|  |  |  |  |  |  |
| Resident Type | Residence is a single family house vs Otherwise | 0.154 | 0.037 | 0.644 | 0.0104 |

1. Starting with your main effects model (from question 2 or, if you made changes, from question 3), perform a complete assessment of interactions. YOUR FINAL MODEL MUST INCLUDE AT LEAST ONE INTERACTION TERM. *(4 points)*

Table 4.01: p-values of every possible interaction through multivariate logistic regression with P<0.15 highlighted to indicate significant

|  |  |
| --- | --- |
| Interactions | P-value |
| Is\_Female\*Is\_Married | 0.9935 |
| Is\_Female\*White | 0.7641 |
| Is\_Female\*Minors | 0.6852 |
| Is\_Female\*own | 0.3987 |
| Is\_Female\*house | 0.2351 |
| Is\_Female\*Income | **0.1154** |
|  |  |
| Is\_Married\*White | 0.2630 |
| Is\_Married\*Minors | 0.2652 |
| Is\_Married\*Own | 0.6747 |
| Is\_Married\*House | 0.4098 |
| Is\_Married\*Income | 0.3378 |
|  |  |
| White\*Minors | 0.5615 |
| **White\*own** | **0.1383** |
| White\*house | 0.4446 |
| White\*Income | 0.4449 |
|  |  |
| Minors\*Own | 0.4677 |
| Minors\*House | 0.9559 |
| Minors\*Income | 0.6243 |
|  |  |
| Own\*House | 0.5988 |
| Own\*Income | 0.8968 |
|  |  |
| House\*Income | 0.9859 |

Considering interactions at p<0.15 as significant and including in interaction

Significant interactions = Being Female\*Income and Being white\*Owning a house

1. Present your final model: Present two tables, one containing the model coefficients, their standard errors and p-values and the other containing appropriate ORs and 95% confidence intervals. *(4 points)*

Table 5.01: Final model with coefficients with significant interaction terms included.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Coefficients | Standard Error | Wald Chi-Square | p-Value |
| Intercept | -14.7990 | 2.7902 | 28.1307 | <0.0001 |
| Is\_Female | -0.0213 | 3.1100 | 0.0000 | 0.9945 |
| Is\_Married | 0.9694 | 0.5731 | 2.8610 | 0.0908 |
| White | 1.4320 | 0.9173 | 2.4370 | 0.1185 |
| Minors | 1..4210 | 0.5607 | 6.4231 | 0.0113 |
| Own | 0.9882 | 1.1332 | 0.7605 | 0.3832 |
| House | -139822 | 0.7540 | 6.9117 | 0.0086 |
| Income | 0.000198 | 0.000039 | 26.2930 | <0.0001 |
| Is\_Female\*Income | 0.000037 | 0.000049 | 0.5574 | 0.4553 |
| White\*Own | 0.4701 | 1.1523 | 0.1664 | 0.6833 |

Table 5.02: Odds ratios, 95% CI’s and p-values for final model with included interactions.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95% CI | | P-value |
| Married | Married vs. otherwise | 2.636 | 0.857 | 8.107 | 0.0908 |
|  |  |  |  |  |  |
| Children | Have children under the age of 18 vs Otherwise | 4.141 | 1.380 | 12.427 | 0.0113 |
|  |  |  |  |  |  |
| Resident Type | Residence is a single family house vs Otherwise | 0.138 | 0.031 | 0.604 | 0.0086 |
|  |  |  |  |  |  |
| Race\*own a house | White yes vs. otherwise at own=1 | 6.7002 | 1.6150 | 27.7965 | 0.0088 |
| White yes vs. otherwise at own=0 | 4.1873 | 0.6936 | 25.2793 | 0.1185 |
|  |  |  |  |  |  |
| Female\*Income | Is\_Female yes vs otherwise, Income=25,000 | 2.4579 | 0.0584 | 103.5 | 0.6375 |
| Is\_Female yes vs otherwise, Income=50,000 | 6.1717 | 1.2662 | 30.0822 | 0.0243 |
| Is\_Female yes vs otherwise, Income=75,000 | 15.4967 | 2.9872 | 80.3913 | 0.0011 |
|  |  |  |  |  |
| Income+10,000 at Is\_Female=Yes | 10.4159 | 4.9431 | 21.9479 | <0.0001 |
| Income+10,000 at Is\_Female=Otherwise | 7.2071 | 3.3876 | 15.3329 | <0.0001 |

Model seems relatively stable. There are a few wide confidence intervals in Married, Children Race\*Own a house and Female\*Income

1. Interpret the ORs and 95% confidence intervals. *(2 points)*

* **Married**
  + If you are married, are 2.6 times as likely to buy Kid Creative Magazine as if you are otherwise not married. Although this Odds ration can span 0.857 to 8.107, therefore should be interpreted with caution and is not a significant finding.
* **Children** 
  + People who Have children under the age of 18 are 4.1 times as likely to buy Kids Creative Magazine as those that otherwise do not have children. That Odd Ratio can span from 1.4 to 12.4 and should be interpreted with some caution.
* **Resident Type**
  + Living in a residence that is a single family house makes a person less likely to buy Kids creative magazine. They are 0.14 times as likely (or rather, 86 times less likely) to buy Kid Creative magazine. This odds ratio can span from 0.03 to 0.60, and can be interpreted relatively safely.
* **Race by Owning a House**
  + Having race as a predictor for buying Kids Creative magazine is dependent on whether an individual’s owns a house or not.
  + An individuals who is white and does NOT owns a house is 6.7 times as likely to by Kids Creative Magazine as someone who is of a difference race and owns a house. The odds ratio can span from 1.6 to 27.8 and therefore should be interpreted with caution
  + An individuals who is white and owns a house is 4.8 times as likely to by Kids Creative Magazine as someone who is of a difference race and does NOT owns a house. The odds ratio can span from 0.69 to 25.28 and therefore should be interpreted with caution. This finding was not significant
* **Female by Income**
  + The odds of being female and buying Kids creative magazine is dependent on the current household income.
    - Females at an income level of $25,000 are 2.5 times as likely to buy Kid Creative Magazine as those that are otherwise NOT female at the same income level of $25,000. This odds ratio can span anywhere from 0.05 to 103.5 therefore should be interpreted with extreme caution. Furthermore, this finding was not significant.
    - Females at an income level of $50,000 are 6.1 times as likely to buy Kid Creative Magazine as those that are otherwise NOT female at the same income level of $50,000. This odds ratio can span anywhere from 1.26 to 30.0 therefore should be interpreted with extreme caution.
    - Females at an income level of $75,000 are 15.50 times as likely to buy Kid Creative Magazine as those that are otherwise NOT female at the same income level of $75,000. This odds ratio can span anywhere from 2.99 to 80.39 therefore should be interpreted with extreme caution.
    - A female is 10.4 times as likely to buy Kid Creative Magazine for every $10,000 rise in household income. This odds ratio can span from 4.9 to 21.9 and therefore should be interpreted with caution.
    - Someone who is NOT female is 7.2 times as likely to buy Kid Creative Magazine for every $10,000 rise in household income. This odds ratio can span from 3.4 to 15.3 and therefore should be interpreted with caution.
    - As can be seen from the trend, There is a clear upward trend in the odds of a female buying a magazine if they have a higher income status.

1. Recall that Mags4u.com wants to target ads to specific customers. Describe the customers who should receive the ad for “Kid Creative”. *(2 points)*

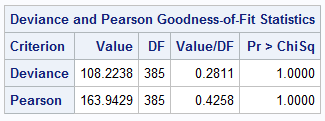
Based on the model created above, customers that should receive the ad for Kids creative magazine should be white females of higher income status ($50,000 or more) that have children and who own a multi-family house. It does not appear to matter too much if they are married or not, but definitely worth it to target those that are married as well.

1. For your final model (from question 5), use the Pearson chi-square test, deviance test, Hosmer-Lemeshow test and Osius-Rojek test to assess overall model fit. Comment on the appropriateness of each test and draw conclusions regarding model fit. *(5 points)*

N = 500

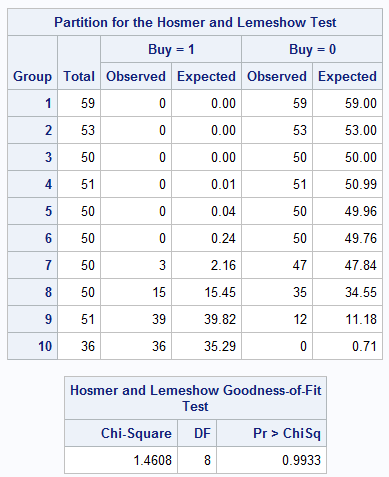
Number of model covariates (J) = 395.

Table 8.01: Results from Deviance and Pearson chi-square GOF tests on the finals model.



According to the Deviance and Pearson chi-square tests, This model is fitting nearly perfect (p=1.00). Given our covariate pattern is relatively low (J = 395) to our number of participants (n=500). The deviance and Pearson Chi-Square are actually the best GOF tests to go by. To be honest, it seems like this model is TOO fit.

Table 8.02: Results from Hosmer and Lemeshow GOF tests on the finals model.



According to the Hosmer and Lemeshow GOF test, This model is fitting nearly perfect (p=0.9933). Given our covariate pattern is relatively low (J = 395) to our number of participants (n=500). The Hosmer and Lemeshow test actually is NOT the best GOF tests to go by. Therefore, this rssult should be interpreted with caution as it is meant for models that have a large number of covariate patterns in comparison to the number of participants.

**Osius-Rojek Test**

Table 8.03: P-value results from the Osius-Rojek test on the Final model.



According to the Osius-Rojek GOF test, This model is fitting nearly perfect (p=0.978). Given our covariate pattern is relatively low (J = 395) to our number of participants (n=500). The Osius-Rojek test actually is NOT the best GOF tests to go by. Therefore, this result should be interpreted with caution as it is meant for models where the covariate pattern and number of participants are relatively close and for large sample sizes.

1. Use the Stukel test to test the tails assumption. Draw conclusions. *(2 points)*

Table 9.01: Model coefficients from the Stukel test with upper and lower tail coefficients boxed in red.

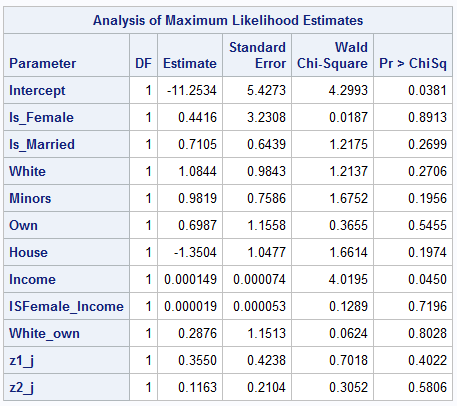


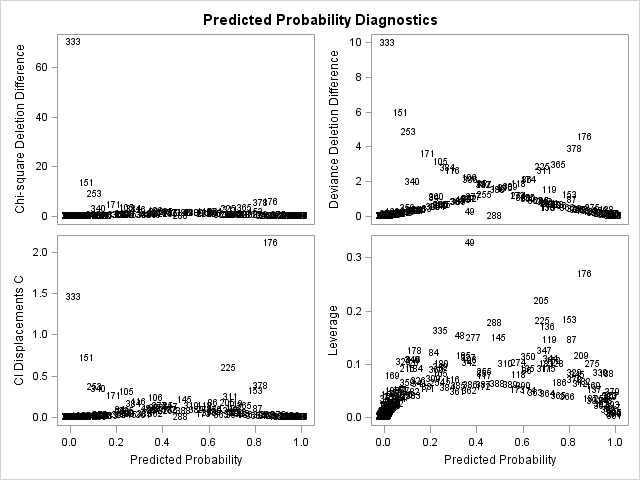
Table 9.02: Overall p-value from Stukel test assessing overall fit of the tails to the model



According to the Stukel Test for tail assumptions, our model does not appear to have any extreme outliers in the tails. Our upper tail is fitting very good (p=0.40) as well as our lower tail (p=0.58). Lastly, when looking at the overall p-value, both tails are fitting very good as well (p=0.61). All tails appear to be fitting adequately and correctly to this model.

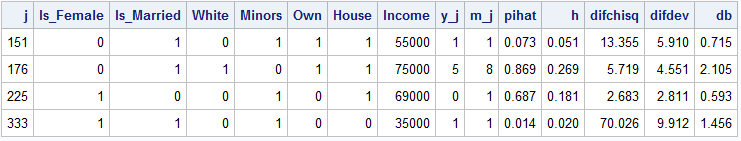
10.

* 1. Use all 4 logistic regression diagnostics to identify outliers. *(4 points)*

Figure 10.01: Diagnostic graphs for all 4 logistic regression diagnostics to identify outliers 

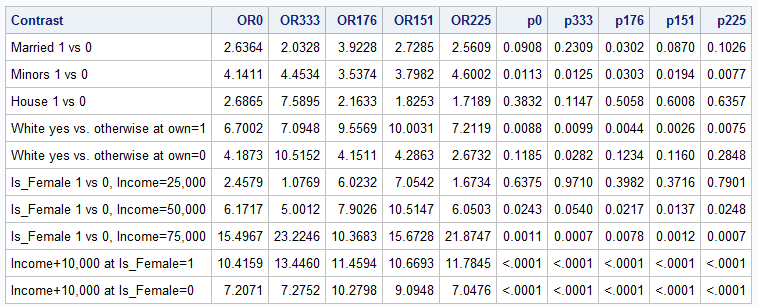
* 1. Describe why the outliers are outliers in the model. *(2 points)*

Table 10.02: Covariate patterns of selected outliers with outcomes and diagnostic results.



* **Outlier Covariate pattern 151**
  + The individual fulfilling this covariate pattern did not have 2 of risk factors associated with this did not have 2 of the risk factors associated with the furthermore, both of these predictors are associated with an interaction. Even though they had a very low probability to by Kid Creative magazine, they ended up buying it. Therefore, this appears as an outlier.
* **Outlier Covariate pattern 176**
  + This covariate pattern had had 8 individuals. They were very likely to have the outcome of buing Kid Creative magazine, but 3 of them did not. The predictive factors associated with the high probability are being married, being white owning a house, and they had the highest income range. Therefore, because not every individual end up buying the Kid creative magazine, it is appearing as an outlier.
* **Outlier Covariate Pattern 225**
  + For this covariate pattern, the one person that fulfilled it had a relatively good probability of buying Kid Creative magazine, yet, did not. They had a higher income, had kids under the age of 18, and were female. All of which are great predictors to hae someone buy Kid Creative magazine. The fact they did not buy the magazine makes them seem like an outlier
* **Outlier Covariate Pattern 333**
  + For this covariate pattern, the person that ended up buying Kid Creative magazine even though they had a very low probability of buying it. Their Income was on the lower end, they were not white and they did not own a house. All of which are predictive of an individual not buying the Kid Creative magazine, thus leading to it seeming as an outlier.
  1. Show the effect deletion of the outliers has on the model ORs and p-values. *(2 points)*

Table 10.03: Change in ORs and p-values when selected outliers are removed



* 1. Propose how to deal with the outliers. *(1 point)*

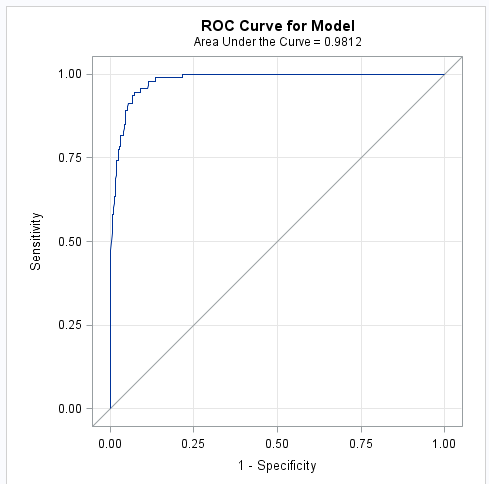
Personally, I do not like the idea of removing outliers at all. Removing the outliers honestly just seems like you are trying to sway the data in the direction you want. But, there are definitely instances where I am sure that the outliers are stopping you from seeing the true effect that is there.

Looking at the above table though, removing any of the outlier covariate patterns does nto seem to have too much of an effect on any of the odds ratios or p-values. Of all of the, covariate pattern 225 seems to insight the most change – but not enough for us to come to change any conclusions about our model. Therefore, I have decided to keep all outliers in as they do not appear to have too much of an effect on our earlier conclusions.

11. Keeping all outliers in the data set, determine how well your model predicts the outcome.

1. Plot the ROC curve and determine the area under the ROC curve. *(1 point)*

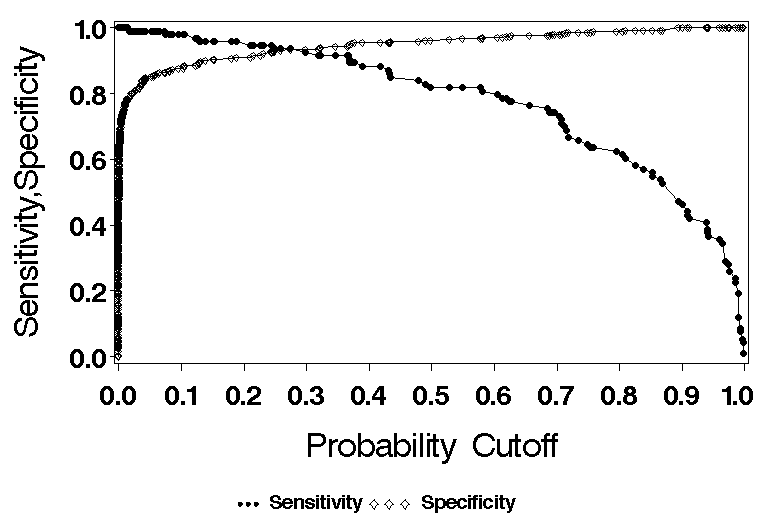
Figure 11.01: ROC curve for the final model.



Area under the ROC curve = 0.98

1. Plot sensitivity and specificity vs. possible cutpoints and select the “best” cutpoint *(2 points)*

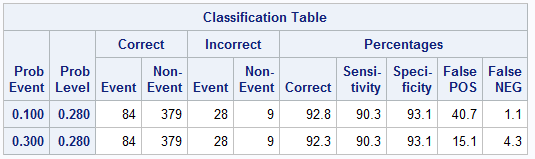
Figure 11.02: Sensitivity and Specificity curves plotted with best cut point indicate by a blue line



**Best cut point = 0.28**

1. Based on this cutpoint, calculate sensitivity and specificity. *(2 points)*

Table 11.01: Classification for final model with a cutpoint at 0.28 and split by the probability of the outcome at 10% and 30%.



Based on a cut point of 0.28:

* **Sensitivity = 90.3%**
* **Specificity = 93.1%**

1. Calculate the positive and negative predictive value assuming the prevalence of the outcome in the population of interest is 10%. *(2 points)*

Based on a prevalence of the outcome is 10%:

* **PPV = 100-FalsePOS** 🡪 **100-40.7 = 59.3%**
* **NPV = 100-FalseNEG** 🡪 **100-1.1 = 98.9%**

1. Draw conclusions keeping in mind the limitations of the above methods (show relevant SAS output). *(2 points)*

Based on the results of the ROC curve, to start, it can be seen that our Area under the ROC is 0.97, which indicates “Outstanding discrimination”. This may sounds nice, but in reality it is not. Based on that result, it is likely that there is Quasi-complete separation within our model. Therefore, the variables are most likely at the majority of data points are at the extreme ends and therefore, there is too much discrimination. All results indicated here should be interpreted with caution.

Based on these results, our model does a great job at predicting when someone will NOT buy a Kid creative magazine. Therefore, in the sense of our model. 98.9% of the time when we predict someone does not buy a Kid Creative magazine, we are correct. On the other hand, our model does a worse job at predicting when an individual does buy a Kid Creative magazine. 59.3% of the time, when we predict that someone will buy a kid creative magazine we are correct. It should be noted that our interpretations are based on the fact we are estimating that 10% of the population of interest are buying Kid Creative magazine.

SAS CODE

libname sdat 'C:\Users\ndyet\_000\Desktop\Class Folders\Spring 2016\ERHS 642\Data';

/\* data sdat.Kc\_46; set Kc\_46; \*/

**data** Kc\_46; set sdat.Kc\_46;

if Income <=**40000** then Incomea=**0**;

else if **40000**< Income then Incomea=**1**;

if Income <=**55000** then Incomeb=**0**;

else if **55000**< Income then Incomeb=**1**;

if Income <=**60000** then Incomec=**0**;

else if **60000**< Income then Incomec=**1**;

If Is\_married=**0** then Dual\_income=**.**;

ISFemale\_Income=Is\_Female\*Income;

White\_own=WHITE\*own;

**run**;

**proc** **print** data=KC\_46; **run**;

**proc** **freq** data=Kc\_46;

tables incomec\*BUy;

**run**;

**proc** **logistic** descending data=Kc\_46;

model BUY=Minors Is\_Female Is\_Married English;

**run**;

\*Power Analysis\*;

**proc** **power**;

logistic

vardist ("Minors") = binomial(**0.3**, **1**)

vardist ("Gender") = binomial(**0.05**,**1**)

vardist ("Marital Status") = binomial(**0.05**, **1**)

vardist ("English") = binomial(**0.05**,**1**)

testpredictor = "Minors"

covariates = ("Gender" "Marital Status" "English" )

responseprob = **0.15**

testoddsratio = **2.5**

covoddsratios = (**1.46** **5.54** **1.93**)

ntotal = **500**

power = **.**;

**run**;

\*categorical data frequency tables 2a;

**proc** **freq** data = Kc\_46;

tables (Is\_Female Is\_Married Has\_College Is\_Professional Is\_Retired Unemployed Dual\_Income Minors Own House White English Prev\_Child\_Mag Prev\_Parent\_Mag incomea)\*BUY/norow nopercent;

**run**;

**proc** **freq** data = Kc\_46;

tables Dual\_income\*Buy;

**run**;

**proc** **freq** data = Kc\_46;

tables BUY;

**run**;

\*Continuous variables descriptived 2b;

**proc** **sort** data=Kc\_46; by BUY; **run**;

**proc** **univariate** data=Kc\_46; by BUY;

var Income Residence\_Length;

**run**;

**proc** **univariate** data=Kc\_46;

var Income;

**run**;

**proc** **freq** data=Kc\_46;

tables Income\*BUY;

**run**;

\*Univariate Scale Assessment 2c;

**proc** **univariate** data=Kc\_46; var Income Residence\_Length; **run**;

\*Income;

**proc** **logistic** descending data=Kc\_46;

effect Incomes=spline(Income/knotmethod=list(**14000** **33000** **52500**) basis=tpf(noint) degree=**0**);

model BUY=Incomes; effectplot;

**run**;

**proc** **logistic** descending data=Kc\_46;

effect Incomes=spline(Income/knotmethod=list(**14000** **33000** **52500**) basis=tpf(noint) degree=**1**);

model BUY=Incomes; effectplot;

**run**;

**proc** **logistic** descending data=Kc\_46;

effect Incomes=spline(Income/knotmethod=list(**14000** **33000** **52500**) basis=tpf(noint) naturalcubic);

model BUY=Incomes; effectplot;

**run**;

\*Length of Residency;

**proc** **logistic** descending data=Kc\_46;

effect Residence\_Lengths=spline(Residence\_Length/knotmethod=list(**7.5** **16** **25**) basis=tpf(noint) degree=**0**);

model BUY=Residence\_Lengths; effectplot;

**run**;

**proc** **logistic** descending data=Kc\_46;

effect Residence\_Lengths=spline(Residence\_Length/knotmethod=list(**7.5** **16** **25**) basis=tpf(noint) degree=**1**);

model BUY=Residence\_Lengths; effectplot;

**run**;

**proc** **logistic** descending data=Kc\_46;

effect Residence\_Lengths=spline(Residence\_Length/knotmethod=list(**7.5** **16** **25**) basis=tpf(noint) naturalcubic);

model BUY=Residence\_Lengths; effectplot;

**run**;

**proc** **freq** data=Kc\_46;

tables Income Residence\_Length;

**run**;

\*Fp\_method INCOME;

\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*;

\* \*;

\* FP METHOD \*;

\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*;

\*\* Macro for fp assessment \*\*;

**%macro** fp1(dset,y,var,lb,p1);

%do %until(&p1=**7**);

%put \*\*\*\*\* &p1 \*\*\*\*\*;

ODS output FitStatistics = mfs;

data fpdat; set &dset; if &var>&lb; pc=&p1/**2**;

if pc ne **0** then F1=&var\*\*pc; else if pc = **0** then F1=log(&var);

run;

proc logistic descending data=fpdat;

model &y=F1; \*-------------------F1 represents the variable being tested for scale;

run;

data mfs; set mfs; if criterion='-2 Log L'; drop Criterion InterceptOnly; run;

proc append data=mfs base=tres; run;

proc datasets; delete fpdat mfs; run;

quit;

%let p1=%eval(&p1+1);

%end;

**%mend** fp1;

%***fp1***(Kc\_46,BUY,Income,**0**,-**4**); \*-----------Enter data set name, outcome variable name and name of variable being tested for scale;

**data** pvals; do p1=-**4** to **6**; output; end; **run**;

**data** pvals; set pvals; p1=p1/**2**; **run**;

**data** tres; merge pvals tres; if p1 in (-**1.5**, **1.5**, **2.5**) then delete; **run**;

**proc** **sort** data=tres; by InterceptAndCovariates; **run**;

**data** tres; set tres; if \_N\_=**1** or p1=**1**; **run**;

**%macro** fp2(dset,y,var,lb,p1,p2);

%do %until(&p1=**7**);

%do %until(&p2=**7**);

%put \*\*\*\*\* &p1 &p2 \*\*\*\*\*;

ODS output FitStatistics = mfs;

data fpdat; set &dset; if &var>&lb; pc1=&p1/**2**; pc2=&p2/**2**;

if pc1 ne **0** then F1=&var\*\*pc1; else if pc1 = **0** then F1=log(&var);

if pc1 ne pc2 then do; if pc2 ne **0** then F2=&var\*\*pc2;

else if pc2 = **0** then F2=log(&var); end;

if pc1=pc2 then F2=F1\*log(&var);

run;

proc logistic descending data=fpdat;

model &y=F1 F2; \*------------F1 and F2 represent the variable being tested for scale;

run;

data mfs; set mfs; if criterion='-2 Log L'; drop Criterion InterceptOnly; run;

proc append data=mfs base=tres2; run;

proc datasets; delete fpdat mfs; run;

quit;

%let p2=%eval(&p2+1);

%end;

%let p2=%eval(-4);

%let p1=%eval(&p1+1);

%end;

**%mend** fp2;

%***fp2***(Kc\_46,BUY,Income,**0**,-**4**,-**4**); \*-----------Enter data set name, outcome variable name and name of variable being tested for scale;

**data** pvals2; do p1=-**4** to **6**; do p2=-**4** to **6**; output;end; end; **run**;

**data** pvals2; set pvals2; p1=p1/**2**; p2=p2/**2**; **run**;

**data** tres2; merge pvals2 tres2;

if p1 in (-**1.5**, **1.5**, **2.5**) or p2 in (-**1.5**, **1.5**, **2.5**) then delete; **run**;

**proc** **sort** data=tres2; by InterceptAndCovariates; **run**;

**data** tres2; set tres2; if \_N\_=**1**; **run**;

**data** comb; set tres tres2; **run**;

**data** c1; set comb; if p1=**1** and p2=**.**; rename InterceptAndCovariates=Dev\_linear;

drop p1 p2; **run**;

**data** c2; set comb; if p1 ne **1** and p2=**.**; rename InterceptAndCovariates=Dev\_fp1;

rename p1=e\_fp1; drop p2; **run**;

**data** c3; set comb; if p2 ne **.**; rename InterceptAndCovariates=Dev\_fp2;

rename p1=e1\_fp2; rename p2=e2\_fp2; **run**;

**data** c;

merge c1 c2 c3;

diff\_lin\_fp1=Dev\_linear-Dev\_fp1;

diff\_lin\_fp2=Dev\_linear-Dev\_fp2;

diff\_fp1\_fp2=Dev\_fp1-Dev\_fp2;

p\_lin\_fp1=**1**-probchi(diff\_lin\_fp1,**1**);

p\_lin\_fp2=**1**-probchi(diff\_lin\_fp2,**3**);

p\_fp1\_fp2=**1**-probchi(diff\_fp1\_fp2,**2**);

**run**;

**proc** **print** noobs data=c;

var Dev\_linear e\_fp1 Dev\_fp1 e1\_fp2 e2\_fp2 Dev\_fp2 p\_lin\_fp1 p\_lin\_fp2 p\_fp1\_fp2;

format p\_lin\_fp1 p\_lin\_fp2 p\_fp1\_fp2 **6.4**;

**run**;

**proc** **datasets**; delete tres tres2 pvals pvals2 comb c c1 c2 c3; **run**; **quit**;

\* End macro for fp assessment \*;

\*Fp\_method Residence\_Length;

\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*;

\* \*;

\* FP METHOD \*;

\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*;

\*\* Macro for fp assessment \*\*;

**%macro** fp1(dset,y,var,lb,p1);

%do %until(&p1=**7**);

%put \*\*\*\*\* &p1 \*\*\*\*\*;

ODS output FitStatistics = mfs;

data fpdat; set &dset; if &var>&lb; pc=&p1/**2**;

if pc ne **0** then F1=&var\*\*pc; else if pc = **0** then F1=log(&var);

run;

proc logistic descending data=fpdat;

model &y=F1; \*-------------------F1 represents the variable being tested for scale;

run;

data mfs; set mfs; if criterion='-2 Log L'; drop Criterion InterceptOnly; run;

proc append data=mfs base=tres; run;

proc datasets; delete fpdat mfs; run;

quit;

%let p1=%eval(&p1+1);

%end;

**%mend** fp1;

%***fp1***(Kc\_46,BUY,Residence\_Length,**0**,-**4**); \*-----------Enter data set name, outcome variable name and name of variable being tested for scale;

**data** pvals; do p1=-**4** to **6**; output; end; **run**;

**data** pvals; set pvals; p1=p1/**2**; **run**;

**data** tres; merge pvals tres; if p1 in (-**1.5**, **1.5**, **2.5**) then delete; **run**;

**proc** **sort** data=tres; by InterceptAndCovariates; **run**;

**data** tres; set tres; if \_N\_=**1** or p1=**1**; **run**;

**%macro** fp2(dset,y,var,lb,p1,p2);

%do %until(&p1=**7**);

%do %until(&p2=**7**);

%put \*\*\*\*\* &p1 &p2 \*\*\*\*\*;

ODS output FitStatistics = mfs;

data fpdat; set &dset; if &var>&lb; pc1=&p1/**2**; pc2=&p2/**2**;

if pc1 ne **0** then F1=&var\*\*pc1; else if pc1 = **0** then F1=log(&var);

if pc1 ne pc2 then do; if pc2 ne **0** then F2=&var\*\*pc2;

else if pc2 = **0** then F2=log(&var); end;

if pc1=pc2 then F2=F1\*log(&var);

run;

proc logistic descending data=fpdat;

model &y=F1 F2; \*------------F1 and F2 represent the variable being tested for scale;

run;

data mfs; set mfs; if criterion='-2 Log L'; drop Criterion InterceptOnly; run;

proc append data=mfs base=tres2; run;

proc datasets; delete fpdat mfs; run;

quit;

%let p2=%eval(&p2+1);

%end;

%let p2=%eval(-4);

%let p1=%eval(&p1+1);

%end;

**%mend** fp2;

%***fp2***(Kc\_46,BUY,Residence\_Length,**0**,-**4**,-**4**); \*-----------Enter data set name, outcome variable name and name of variable being tested for scale;

**data** pvals2; do p1=-**4** to **6**; do p2=-**4** to **6**; output;end; end; **run**;

**data** pvals2; set pvals2; p1=p1/**2**; p2=p2/**2**; **run**;

**data** tres2; merge pvals2 tres2;

if p1 in (-**1.5**, **1.5**, **2.5**) or p2 in (-**1.5**, **1.5**, **2.5**) then delete; **run**;

**proc** **sort** data=tres2; by InterceptAndCovariates; **run**;

**data** tres2; set tres2; if \_N\_=**1**; **run**;

**data** comb; set tres tres2; **run**;

**data** c1; set comb; if p1=**1** and p2=**.**; rename InterceptAndCovariates=Dev\_linear;

drop p1 p2; **run**;

**data** c2; set comb; if p1 ne **1** and p2=**.**; rename InterceptAndCovariates=Dev\_fp1;

rename p1=e\_fp1; drop p2; **run**;

**data** c3; set comb; if p2 ne **.**; rename InterceptAndCovariates=Dev\_fp2;

rename p1=e1\_fp2; rename p2=e2\_fp2; **run**;

**data** c;

merge c1 c2 c3;

diff\_lin\_fp1=Dev\_linear-Dev\_fp1;

diff\_lin\_fp2=Dev\_linear-Dev\_fp2;

diff\_fp1\_fp2=Dev\_fp1-Dev\_fp2;

p\_lin\_fp1=**1**-probchi(diff\_lin\_fp1,**1**);

p\_lin\_fp2=**1**-probchi(diff\_lin\_fp2,**3**);

p\_fp1\_fp2=**1**-probchi(diff\_fp1\_fp2,**2**);

**run**;

**proc** **print** noobs data=c;

var Dev\_linear e\_fp1 Dev\_fp1 e1\_fp2 e2\_fp2 Dev\_fp2 p\_lin\_fp1 p\_lin\_fp2 p\_fp1\_fp2;

format p\_lin\_fp1 p\_lin\_fp2 p\_fp1\_fp2 **6.4**;

**run**;

**proc** **datasets**; delete tres tres2 pvals pvals2 comb c c1 c2 c3; **run**; **quit**;

\* End macro for fp assessment \*;

**proc** **logistic** descending data=Kc\_46; class Is\_Female/param=ref ref=first; model BUY=Is\_Female; **run**;

**proc** **logistic** descending data=Kc\_46; class Is\_Married/param=ref ref=first; model BUY=Is\_Married; **run**;

**proc** **logistic** descending data=Kc\_46; class Has\_College/param=ref ref=first; model BUY=Has\_College; **run**;

**proc** **logistic** descending data=Kc\_46; class Is\_Professional/param=ref ref=first; model BUY=Is\_Professional; **run**;

**proc** **logistic** descending data=Kc\_46; class Is\_Retired/param=ref ref=first; model BUY=Is\_Retired; **run**;

**proc** **logistic** descending data=Kc\_46; class Unemployed/param=ref ref=first; model BUY=Unemployed; **run**;

**proc** **logistic** descending data=Kc\_46; class Dual\_Income/param=ref ref=first; model BUY=Dual\_Income; **run**;

**proc** **logistic** descending data=Kc\_46; class Minors/param=ref ref=first; model BUY=Minors; **run**;

**proc** **logistic** descending data=Kc\_46; class Own/param=ref ref=first; model BUY=Own; **run**;

**proc** **logistic** descending data=Kc\_46; class House/param=ref ref=first; model BUY=House; **run**;

**proc** **logistic** descending data=Kc\_46; class White/param=ref ref=first; model BUY=White; **run**;

**proc** **logistic** descending data=Kc\_46; class English/param=ref ref=first; model BUY=English; **run**;

**proc** **logistic** descending data=Kc\_46; class Prev\_Child\_Mag/param=ref ref=first; model BUY=Prev\_Child\_Mag; **run**;

**proc** **logistic** descending data=Kc\_46; class Prev\_Parent\_Mag/param=ref ref=first; model BUY=Prev\_Parent\_Mag; **run**;

**proc** **logistic** descending data=Kc\_46; class Incomea/param=ref ref=first; model BUY=Incomea; **run**;

**proc** **logistic** descending data=Kc\_46; class Incomeb/param=ref ref=first; model BUY=Incomeb; **run**;

**proc** **logistic** descending data=Kc\_46; class Incomec/param=ref ref=first; model BUY=Incomec; **run**;

**proc** **logistic** descending data=Kc\_46; model BUY=Income/clodds=wald; units Income=**5000**; **run**;

**proc** **logistic** descending data=Kc\_46; model BUY=Residence\_Length/clodds=wald; units Residence\_Length=**1**; **run**;

\*\*\*\*Multivariate Scale Assessment 2e.\*\*\*;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married Has\_College Is\_Professional own House White English Prev\_Parent\_Mag/param=ref ref=first;

model BUY=Is\_Female Is\_Married Has\_College Is\_Professional own House White English Prev\_Parent\_Mag Income Residence\_Length/clodds=wald;

units Income=**5000** Residence\_Length=**1**;

**run**;

\*Is\_Professional removed;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married Has\_College own House White English Prev\_Parent\_Mag/param=ref ref=first;

model BUY=Is\_Female Is\_Married Has\_College own House White English Prev\_Parent\_Mag Income Residence\_Length/clodds=wald;

units Income=**5000** Residence\_Length=**1**;

**run**;

\*Length of Residency removed;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married Has\_College own House White English Prev\_Parent\_Mag/param=ref ref=first;

model BUY=Is\_Female Is\_Married Has\_College own House White English Prev\_Parent\_Mag Income/clodds=wald;

units Income=**5000**;

**run**;

\*Length of Residency removed;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married Has\_College own House White English/param=ref ref=first;

model BUY=Is\_Female Is\_Married Has\_College own House White English Income/clodds=wald;

units Income=**5000**;

**run**;

\*English removed;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married Has\_College own House White/param=ref ref=first;

model BUY=Is\_Female Is\_Married Has\_College own House White Income/clodds=wald;

units Income=**5000**;

**run**;

\*College Educated removed;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married own House White/param=ref ref=first;

model BUY=Is\_Female Is\_Married own House White Income/clodds=wald;

units Income=**5000**;

**run**;

\*Home ownership (Own) removed;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married House White/param=ref ref=first;

model BUY=Is\_Female Is\_Married House White Income/clodds=wald;

units Income=**5000**;

**run**;

\*Resident type (House) removed;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married White/param=ref ref=first;

model BUY=Is\_Female Is\_Married White Income/clodds=wald;

units Income=**5000**;

**run**;

\*Is Female removed;

**proc** **logistic** descending data=Kc\_46;

class Is\_Married White/param=ref ref=first;

model BUY=Is\_Married White Income/clodds=wald;

units Income=**5000**;

**run**;

\*Marital status removed;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female White/param=ref ref=first;

model BUY=Is\_Female White Income/clodds=wald;

units Income=**5000**;

**run**;

\*Race removed;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married/param=ref ref=first;

model BUY=Is\_Female Is\_Married Income/clodds=wald;

units Income=**5000**;

**run**;

\*Income removed;

\*\*\*\*\*\*INCOME effects model heavily\*\*\*\*\*;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married White/param=ref ref=first;

model BUY=Is\_Female Is\_Married White/clodds=wald;

**run**;

\*\*\*Adding Minors;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married White Minors/param=ref ref=first;

model BUY=Is\_Female Is\_Married White Minors Income/clodds=wald;

units Income=**5000**;

**run**;

\*Multivariate Fp\_method INCOME;

\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*;

\* \*;

\* Multivariate FP METHOD \*;

\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*;

\*\* Macro for fp assessment \*\*;

**%macro** fp1(dset,y,var,lb,p1);

%do %until(&p1=**7**);

%put \*\*\*\*\* &p1 \*\*\*\*\*;

ODS output FitStatistics = mfs;

data fpdat; set &dset; if &var>&lb; pc=&p1/**2**;

if pc ne **0** then F1=&var\*\*pc; else if pc = **0** then F1=log(&var);

run;

proc logistic descending data=fpdat;

model &y=F1 Is\_Female Is\_Married White Minors; \*-------------------F1 represents the variable being tested for scale;

run;

data mfs; set mfs; if criterion='-2 Log L'; drop Criterion InterceptOnly; run;

proc append data=mfs base=tres; run;

proc datasets; delete fpdat mfs; run;

quit;

%let p1=%eval(&p1+1);

%end;

**%mend** fp1;

%***fp1***(Kc\_46,BUY,Income,**0**,-**4**); \*-----------Enter data set name, outcome variable name and name of variable being tested for scale;

**data** pvals; do p1=-**4** to **6**; output; end; **run**;

**data** pvals; set pvals; p1=p1/**2**; **run**;

**data** tres; merge pvals tres; if p1 in (-**1.5**, **1.5**, **2.5**) then delete; **run**;

**proc** **sort** data=tres; by InterceptAndCovariates; **run**;

**data** tres; set tres; if \_N\_=**1** or p1=**1**; **run**;

**%macro** fp2(dset,y,var,lb,p1,p2);

%do %until(&p1=**7**);

%do %until(&p2=**7**);

%put \*\*\*\*\* &p1 &p2 \*\*\*\*\*;

ODS output FitStatistics = mfs;

data fpdat; set &dset; if &var>&lb; pc1=&p1/**2**; pc2=&p2/**2**;

if pc1 ne **0** then F1=&var\*\*pc1; else if pc1 = **0** then F1=log(&var);

if pc1 ne pc2 then do; if pc2 ne **0** then F2=&var\*\*pc2;

else if pc2 = **0** then F2=log(&var); end;

if pc1=pc2 then F2=F1\*log(&var);

run;

proc logistic descending data=fpdat;

model &y=F1 F2 Is\_Female Is\_Married White Minors; \*------------F1 and F2 represent the variable being tested for scale;

run;

data mfs; set mfs; if criterion='-2 Log L'; drop Criterion InterceptOnly; run;

proc append data=mfs base=tres2; run;

proc datasets; delete fpdat mfs; run;

quit;

%let p2=%eval(&p2+1);

%end;

%let p2=%eval(-4);

%let p1=%eval(&p1+1);

%end;

**%mend** fp2;

%***fp2***(Kc\_46,BUY,Income,**0**,-**4**,-**4**); \*-----------Enter data set name, outcome variable name and name of variable being tested for scale;

**data** pvals2; do p1=-**4** to **6**; do p2=-**4** to **6**; output;end; end; **run**;

**data** pvals2; set pvals2; p1=p1/**2**; p2=p2/**2**; **run**;

**data** tres2; merge pvals2 tres2;

if p1 in (-**1.5**, **1.5**, **2.5**) or p2 in (-**1.5**, **1.5**, **2.5**) then delete; **run**;

**proc** **sort** data=tres2; by InterceptAndCovariates; **run**;

**data** tres2; set tres2; if \_N\_=**1**; **run**;

**data** comb; set tres tres2; **run**;

**data** c1; set comb; if p1=**1** and p2=**.**; rename InterceptAndCovariates=Dev\_linear;

drop p1 p2; **run**;

**data** c2; set comb; if p1 ne **1** and p2=**.**; rename InterceptAndCovariates=Dev\_fp1;

rename p1=e\_fp1; drop p2; **run**;

**data** c3; set comb; if p2 ne **.**; rename InterceptAndCovariates=Dev\_fp2;

rename p1=e1\_fp2; rename p2=e2\_fp2; **run**;

**data** c;

merge c1 c2 c3;

diff\_lin\_fp1=Dev\_linear-Dev\_fp1;

diff\_lin\_fp2=Dev\_linear-Dev\_fp2;

diff\_fp1\_fp2=Dev\_fp1-Dev\_fp2;

p\_lin\_fp1=**1**-probchi(diff\_lin\_fp1,**1**);

p\_lin\_fp2=**1**-probchi(diff\_lin\_fp2,**3**);

p\_fp1\_fp2=**1**-probchi(diff\_fp1\_fp2,**2**);

**run**;

**proc** **print** noobs data=c;

var Dev\_linear e\_fp1 Dev\_fp1 e1\_fp2 e2\_fp2 Dev\_fp2 p\_lin\_fp1 p\_lin\_fp2 p\_fp1\_fp2;

format p\_lin\_fp1 p\_lin\_fp2 p\_fp1\_fp2 **6.4**;

**run**;

**proc** **datasets**; delete tres tres2 pvals pvals2 comb c c1 c2 c3; **run**; **quit**;

\* End macro for fp assessment \*;

\*Final Model ;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married White Minors/param=ref ref=first;

model BUY=Is\_Female Is\_Married White Minors Income/clodds=wald;

units Income=**5000**;

**run**;

\*\*\*\*\*Automated model selection, stepwise\*\*\*\*\*\*\*\*;

**proc** **logistic** descending data=Kc\_46; class Is\_Female Is\_Married Has\_College Is\_Professional Is\_Retired Dual\_Income Minors Own House White English Prev\_Child\_Mag Prev\_Parent\_Mag/param=ref ref=first;

model BUY=Is\_Female Is\_Married Has\_College Is\_Professional Is\_Retired Dual\_Income Minors Own House White English Prev\_Child\_Mag Prev\_Parent\_Mag Income Residence\_Length

/stepwise sle=**0.15** sls=**0.20** details; **run**;

**proc** **logistic** descending data=Kc\_46; class Is\_Female Minors House White Prev\_Child\_Mag Own/param=ref ref=first;

model BUY= Income Is\_Female Minors House White Prev\_Child\_Mag Own/clodds=wald;

units Income=**5000**;

**run**;

\*Final Model after stepwise selection;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married White Minors own house/param=ref ref=first;

model BUY=Is\_Female Is\_Married White Minors own house Income/clodds=wald;

units Income=**5000**;

**run**;

\*\*\*\*\*Interactions;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married White Minors own house/param=ref ref=first;

model BUY=Is\_Female Is\_Married White Minors own house Income

Is\_Female\*Is\_Married Is\_Female\*White Is\_Female\*Minors Is\_Female\*own Is\_Female\*house Is\_Female\*Income

Is\_Married\*White Is\_Married\*Minors Is\_Married\*own Is\_married\*house Is\_Married\*Income

White\*Minors White\*own White\*house White\*Income

Minors\*own Minors\*house Minors\*Income

own\*house own\*Income

House\*Income

/clodds=wald;

units Income=**5000**;

**run**;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married White Minors/param=ref ref=first;

model BUY=Is\_Female Is\_Married White Minors Income

Is\_Female\*Is\_Married Is\_Female\*White Is\_Female\*Minors Is\_Female\*Income

Is\_Married\*White Is\_Married\*Minors Is\_Married\*Income

White\*Minors White\*Income

Minors\*Income

/clodds=wald;

units Income=**5000**;

**run**;

**proc** **logistic** descending data=Kc\_46;

class Is\_Female Is\_Married White Minors own house/param=ref ref=first;

model BUY=Is\_Female Is\_Married White Minors own house Income Is\_Female\*Income White\*own/clodds=wald;

units Income=**5000**;

contrast 'White yes vs. otherwise at own=1' White **1** white\*own **1**/estimate=exp;

contrast 'White yes vs. otherwise at own=0' White **1** white\*own **0**/estimate=exp;

contrast 'Is\_Female 1 vs 0, Income=25,000'

Is\_Female **1** Income **0** Is\_Female\*Income **25000**/estimate=exp;

contrast 'Is\_Female 1 vs 0, Income=50,000'

Is\_Female **1** Income **0** Is\_Female\*Income **50000**/estimate=exp;

contrast 'Is\_Female 1 vs 0, Income=75,000'

Is\_Female **1** Income **0** Is\_Female\*Income **75000**/estimate=exp;

contrast 'Income+10,000 at Is\_Female=1' Income **10000** Is\_Female\*Income **10000**/estimate=exp;

contrast 'Income+10,000 at Is\_Female=0' Income **10000** Is\_Female\*Income **0**/estimate=exp;

**run**;

\*\*\*\*\*\*\*\*GOF tests\*\*\*\*\*\*\*;

**proc** **sort** data=Kc\_46; \*\*\*\* for a different data set change independent variable names in by statement \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

by Is\_Female Is\_Married White Minors own house Income;

**run**;

**proc** **logistic** descending data=Kc\_46;

model BUY=Is\_Female Is\_Married White Minors own house Income Is\_Female\*Income White\*own

/scale=n aggregate lackfit;

**run**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\* Osius-Rojek test \*\*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* If interaction terms are in the model, create the interaction terms ;

\* If categorical variables with more than 2 categories are in the model;

\* create design variables ;

\* Sort new data set by model covariates;

**proc** **sort** data=Kc\_46; \*\*\*\* for a different data set change independent variable names in by statement \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

by Is\_Female Is\_Married White Minors own house Income;

**run**;

\* For each covariate pattern, j, save m\_j= # with covariate pattern j and;

\* y\_j = # with outcome=1 in covariate pattern j ;

**proc** **means** n sum noprint data=Kc\_46; \*\*\*\* for a different data set change independent variable names in by statement \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\*\*\* and outcome variable name in var statement \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

by Is\_Female Is\_Married White Minors own house Income ISFemale\_Income White\_own;

var BUY; output out=jdat n=m\_j sum=y\_j; \*<--- indicate outcome;

**run**;

\*LOOK ABOVE! after "var" is where your outcome should be;

\* Run proc logistic for covariate patterns rather than individuals ;

\* outcome=y\_j / m\_j (not 0 or 1) , save fitted values ;

**proc** **logistic** noprint descending data=jdat; \*\*\*\* for a different data set change independent variable names in model statement. keep jdat as jdat. do NOT change this\*\*\*\*\*;

model y\_j/m\_j= Is\_Female Is\_Married White Minors own house Income ISFemale\_Income White\_own;

output out=pdat p=p\_j;

**run**;

\* Create v\_j, c\_j, the chi-square terms and the terms in the sum in A;

**data** pdat; set pdat;

v\_j=m\_j\*p\_j\*(**1**-p\_j); c\_j=(**1**-**2**\*p\_j)/v\_j; chisq\_j=(y\_j-m\_j\*p\_j)\*\***2**/v\_j;

m\_j\_inv=**1**/m\_j;

**run**;

\* Create and save chi-square & sum for A ;

\* Perform weighted linear regression, save SS ;

\* Calculate RSS,A,z & p-val for z ;

**proc** **means** sum noprint data=pdat;

var chisq\_j m\_j\_inv; output out=cdat sum=chisq m\_inv; **run**;

**proc** **reg** noprint data=pdat outest=ss; \*\*\*\* for a different data set change independent variable names in model statement \*\*\*\*\*;

model c\_j=Is\_Female Is\_Married White Minors own house Income ISFemale\_Income White\_own;

weight v\_j; **run**;

**data** zdat; merge cdat (keep=\_freq\_ chisq m\_inv) ss (keep=\_rmse\_);

rss=(\_freq\_-**9**-**1**)\*\_rmse\_\*\***2**; A=**2**\*(\_freq\_-m\_inv); \*\*\*\* for a different data set change 8 to number of variables in the model \*\*\*\*\*;

z=(chisq-(\_freq\_-**9**-**1**))/sqrt(A+rss); z=abs(z); \*\*\*\* for a different data set change 8 to number of variables in the model \*\*\*\*\*;

pval=(**1**-probnorm(z))\***2**;

**run**;

**proc** **print** noobs data=zdat; var pval; **run**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\*\*\*\* Stukel test of logistic regression model assumption \*\*\*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

ODS trace on;

ODS output GlobalTests=gt1;

**proc** **logistic** descending data=Kc\_46; \*\*\*\* for a different data set change outcome and independent variable names \*\*\*\*\*;

\*\*\*\* in model statement \*\*\*\*\*;

model BUY=Is\_Female Is\_Married White Minors own house Income ISFemale\_Income White\_own;

output out=pdat2 xbeta=g\_j p=p\_j;

**run**;

**data** pdat2;

set pdat2;

if p\_j>=**0.5** then ind1=**1**; else ind1=**0**;

if p\_j< **0.5** then ind2=**1**; else ind2=**0**;

z1\_j=**0.5**\*g\_j\*\***2**\*ind1;

z2\_j=-**0.5**\*g\_j\*\***2**\*ind2;

**run**;

ODS output GlobalTests=gt2;

**proc** **logistic** descending data=pdat2 ; \*\*\*\* for a different data set change outcome and independent variable names \*\*\*\*\*;

\*\*\*\* in model statement but keep z1\_j and z2\_j \*\*\*\*\*;

model BUY=Is\_Female Is\_Married White Minors own house Income ISFemale\_Income White\_own z1\_j z2\_j;

**run**;

**data** pval;

merge gt1(rename=(ChiSq=ChiSq1))

gt2(rename=(ChiSq=ChiSq2));

if \_N\_=**1**;

drop Test df ProbChisq;

lr=ChiSq2-ChiSq1;

pval=(**1**-probchi(lr,**2**));

**run**;

**proc** **print** noobs data=pval; var pval; **run**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*DIAGNOSTICS\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\*\*\*\* Diagnostics \*\*\*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**proc** **sort** data=Kc\_46; by Is\_Female Is\_Married White Minors own house Income; **run**;

**proc** **means** n sum noprint data=Kc\_46;

by Is\_Female Is\_Married White Minors own house Income;

var BUY; output out=jdat n=m\_j sum=y\_j;

**run**;

**proc** **logistic** descending data=jdat plots(only label)=(phat);

model y\_j/m\_j=Is\_Female Is\_Married White Minors own house Income Is\_Female\*Income White\*own;

output out=diag h=h difchisq=difchisq difdev=difdev c=db p=pihat;

**run**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\*\* Evaluate outliers \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**data** diag; set diag; j=\_N\_; **run**;

**proc** **print** data=diag noobs; where j in (**333**, **176**, **151**, **225**); \*<-- you change these numbers based on the predicted probability dignostics. You are looking for outliers;

var j Is\_Female Is\_Married White Minors own house Income y\_j m\_j pihat h difchisq difdev db; \*<-- Change to your variables do NOT include interactions keep everything else that is not a variable;

format pihat h difchisq difdev db **7.3**; \*<-- keep this the same, this is just making your table look nice.;

**run**;

\*You don't need to use the below macro if it confuses you... but it will make things way easier;

**%macro** outliers(j);

ODS output Logistic.ContrastEstimate=ors;

proc logistic descending data=diag; where j ne &j;

model y\_j/m\_j=Is\_Female Is\_Married White Minors own house Income Is\_Female\*Income White\*own;

contrast 'Married 1 vs 0' Is\_Married **1**/estimate=exp; \*<- insert one contrast per effect;

contrast 'Minors 1 vs 0' Minors **1**/estimate=exp; \*<- insert one contrast per effect;

contrast 'House 1 vs 0' own **1**/estimate=exp; \*<- insert one contrast per effect;

contrast 'White yes vs. otherwise at own=1' White **1** white\*own **1**/estimate=exp;

contrast 'White yes vs. otherwise at own=0' White **1** white\*own **0**/estimate=exp;

contrast 'Is\_Female 1 vs 0, Income=25,000'

Is\_Female **1** Income **0** Is\_Female\*Income **25000**/estimate=exp;

contrast 'Is\_Female 1 vs 0, Income=50,000'

Is\_Female **1** Income **0** Is\_Female\*Income **50000**/estimate=exp;

contrast 'Is\_Female 1 vs 0, Income=75,000'

Is\_Female **1** Income **0** Is\_Female\*Income **75000**/estimate=exp;

contrast 'Income+10,000 at Is\_Female=1' Income **10000** Is\_Female\*Income **10000**/estimate=exp;

contrast 'Income+10,000 at Is\_Female=0' Income **10000** Is\_Female\*Income **0**/estimate=exp;

run;

data ors&j; set ors;

drop type row stderr alpha lowerlimit upperlimit waldchisq;

rename estimate=OR&j probchisq=p&j; run;

proc print data=ors&j; run;

**%mend** outliers;

%***outliers***(**0**);

%***outliers***(**333**);

%***outliers***(**176**);

%***outliers***(**151**);

%***outliers***(**225**);

**data** ors; merge ors0 ors333 ors176 ors151 ors225; **run**;

**proc** **print** data=ors noobs;

var contrast or0 or333 or176 or151 or225 p0 p333 p176 p151 p225;

format ors0 ors333 ors176 ors151 ors225 **6.2**;

**run**;

**proc** **datasets**; delete ors; **run**; **quit**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Classification Tables\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**proc** **logistic** descending data=Kc\_46;

model Buy=Is\_Female Is\_Married White Minors own house Income Is\_Female\*Income White\*own / outroc=rocdat;

**run**;

**data** rocdat;

set rocdat;

spec=**1**-\_1mspec\_;

**run**;

axis1 label=(f=swiss h=**2.5** 'Probability Cutoff') minor=none;

axis2 label=(f=swiss h=**2.5** a=**90** 'Sensitivity,Specificity') minor=none;

goptions FTEXT=swissb HTEXT=**2.0** HSIZE=**8** in VSIZE=**6** in;

symbol1 v=dot i=join c=black h=**1**;

symbol2 v=diamond i=join c=black h=**1**;

footnote1 c=black f=special h=**1** 'J J J' f=swissb h=**1.5** ' Sensitivity'

c=black f=special h=**1** ' D D D' f=swissb h=**1.5** ' Specificity';

**proc** **gplot** data=rocdat;

plot (\_sensit\_ spec)\*\_prob\_

/overlay haxis=axis1 vaxis=axis2;

**run**; **quit**;

footnote;

**proc** **logistic** descending data=Kc\_46;

model Buy=Is\_Female Is\_Married White Minors own house Income Is\_Female\*Income White\*own

/ctable pprob=(**0.28**) pevent=**0.10** **0.30**;

**run**;