Test a Logistic Regression Model

(Regression Modeling in Practice Week 4 Assignment)

Expected Activities

- Test a logistic regression model.
- Write a blog entry that summarize in a few sentences
    1) what you found, making sure you discuss the results for the associations between all of your explanatory variables and your response variable. Make sure to include statistical results (odds ratios, p-values, and 95% confidence intervals for the odds ratios) in your summary.
    2) Report whether or not your results supported your hypothesis for the association between your primary explanatory variable and your response variable.
    3) Discuss whether or not there was evidence of confounding for the association between your primary explanatory and the response variable

    (Hint: adding additional explanatory variables to your model one at a time will make it easier to identify which of the variables are confounding variables).

Note

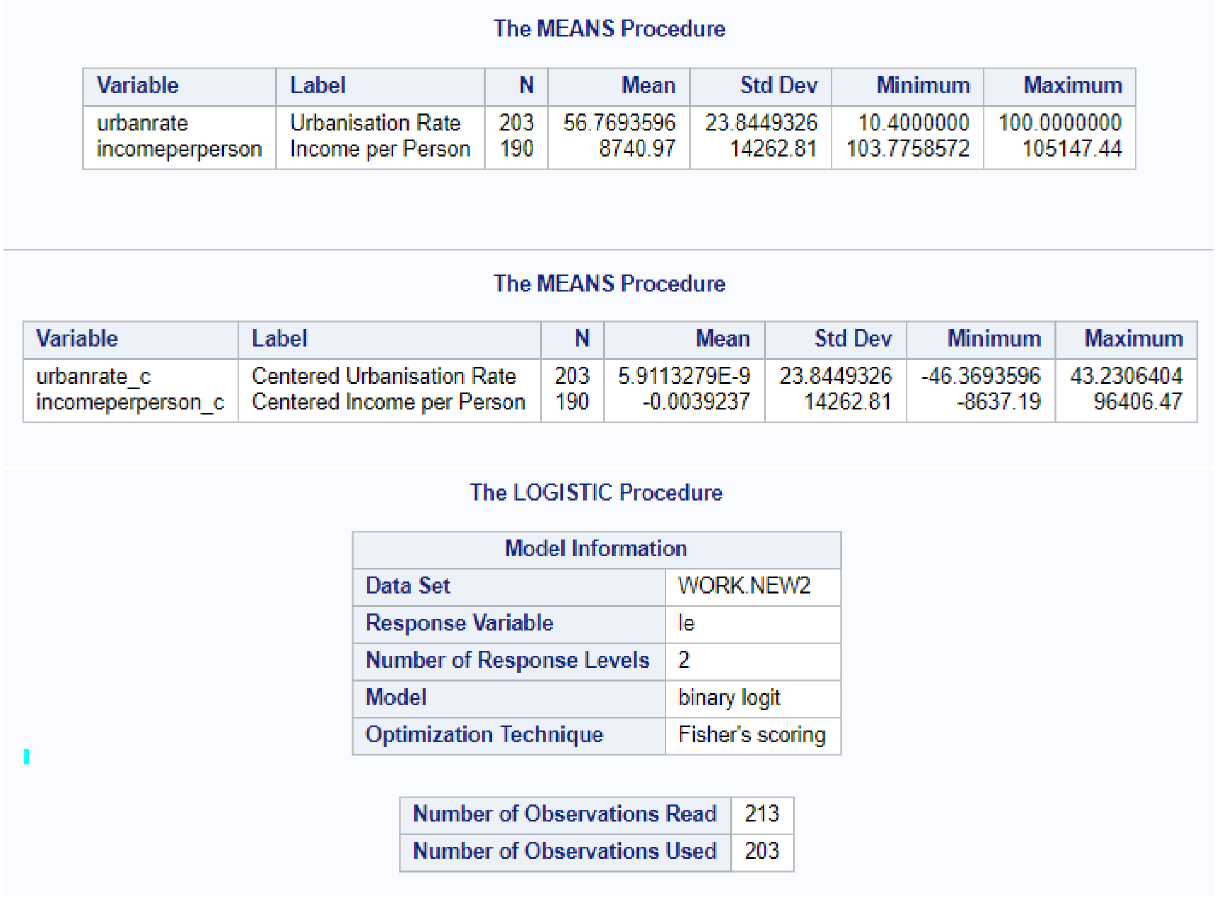
1. If your response variable is categorical with more than two categories, you will need to collapse it down to two categories, or subset your data to select observations from 2 categories.
2. If your response variable is quantitative, you will need to bin it into two categories.

SAS Program

```
LIBNAME mydata "/courses/d1406ae5ba27fe300 " ACCESS=readonly;

DATA new;
SET mydata.gapminder;
KEEP country urbanrate incomeperperson lifeexpectancy;
LABEL lifeexpectancy="Life Expectancy";
LABEL urbanrate="Urbanisation Rate";
LABEL incomeperperson="Income per Person";

/* Find the mean of explanatory variables */
PROC MEANS;
VAR urbanrate incomeperperson;

/* Coding responsive variable */
DATA new2;
SET new;

IF lifeexpectancy LE 69.7535236 THEN
le=0;
ELSE
le=1;
```

```
/* For quantitative explanatory variable, center it so that
the mean = 0 (or really close to 0) by subtracting the mean */
urbanrate_c=urbanrate - 56.7693596;
incomeperperson_c=incomeperperson - 8740.97;
LABEL urbanrate_c="Centered Urbanisation Rate";
LABEL incomeperperson_c="Centered Income per Person";

/* Calculate the mean to check centering */
PROC MEANS;
VAR urbanrate_c incomeperperson_c;

/* Logistic regression model */
PROC LOGISTIC DESCENDING;
MODEL le=urbanrate_c;

PROC LOGISTIC DESCENDING;
MODEL le=urbanrate_c incomeperperson_c;

RUN;
```

Output



The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
urbanrate incomeperperson	Urbanisation Rate Income per Person	203 190	56.7693596 8740.97	23.8449326 14262.81	10.4000000 103.7758572	100.0000000 105147.44

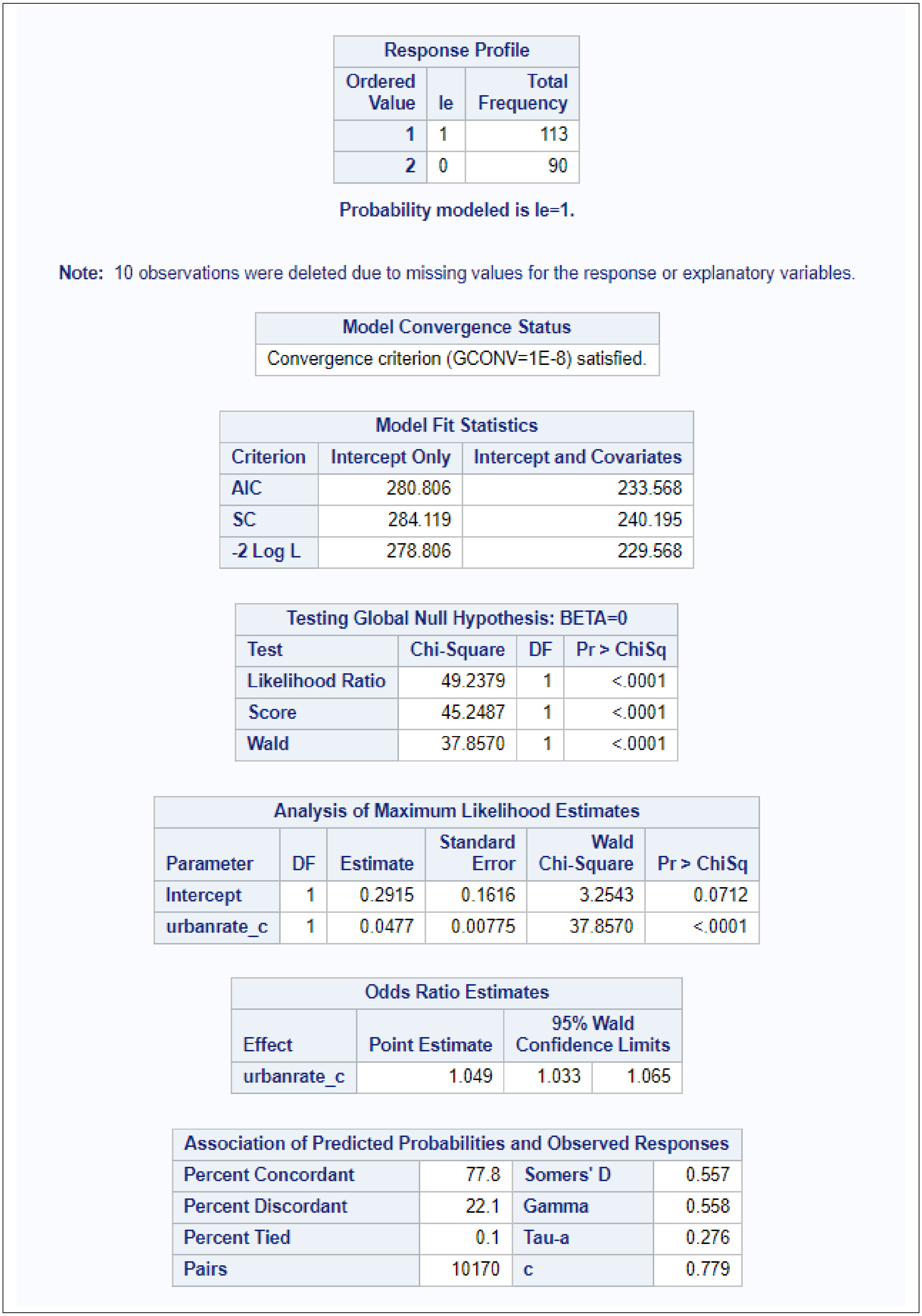
The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
urbanrate_c incomeperperson_c	Centered Urbanisation Rate Centered Income per Person	203 190	5.9113279E-9 -0.0039237	23.8449326 14262.81	-46.3693596 -8637.19	43.2306404 96406.47

The LOGISTIC Procedure

Model Information	
Data Set	WORK.NEW2
Response Variable	le
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	213
Number of Observations Used	203



Response Profile

Ordered Value	le	Total Frequency
1	1	113
2	0	90

Probability modeled is le=1.

Note: 10 observations were deleted due to missing values for the response or explanatory variables.

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	280.806	233.568
SC	284.119	240.195
-2 Log L	278.806	229.568

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	49.2379	1	<.0001
Score	45.2487	1	<.0001
Wald	37.8570	1	<.0001

Analysis of Maximum Likelihood Estimates

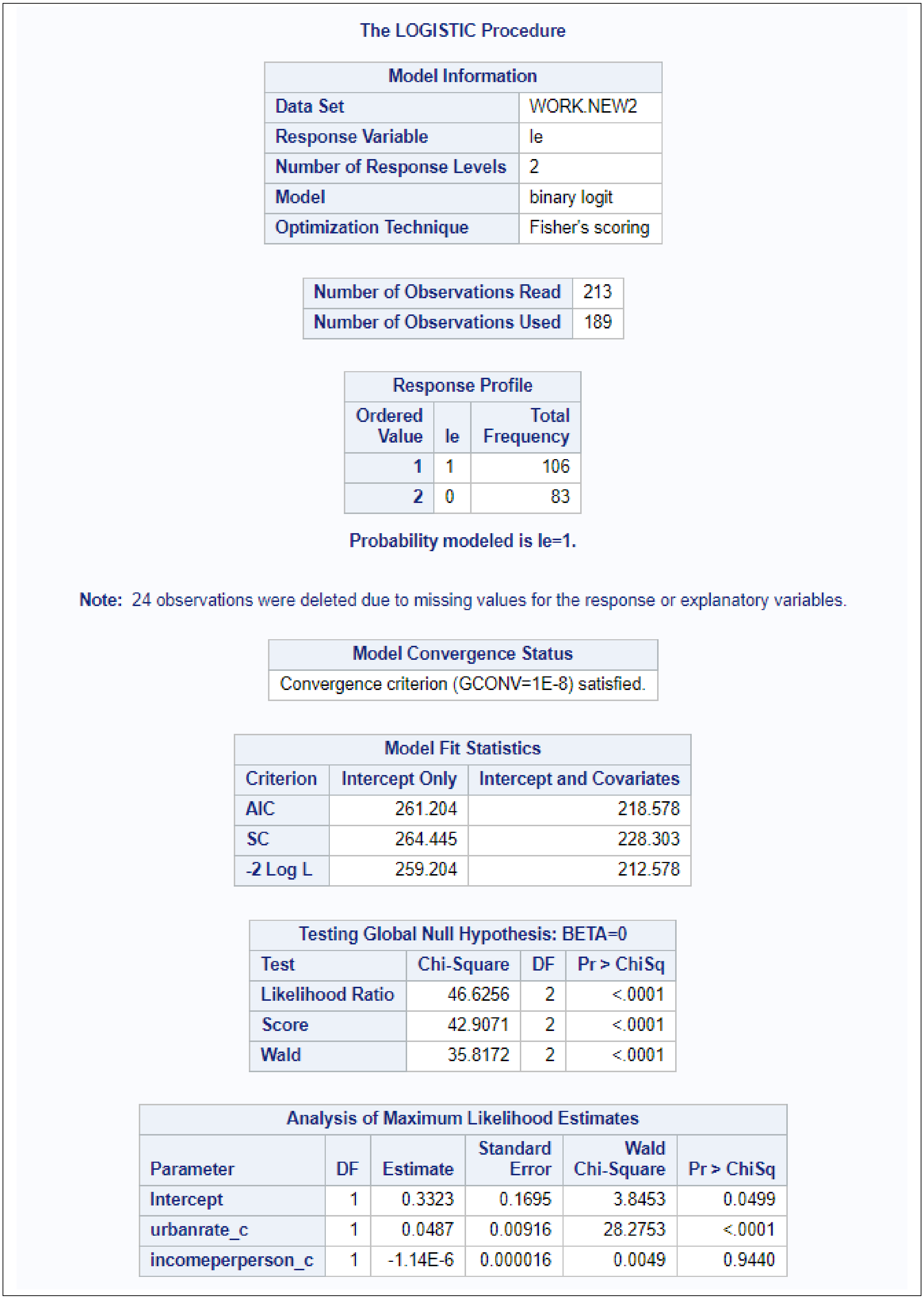
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.2915	0.1616	3.2543	0.0712
urbanrate_c	1	0.0477	0.00775	37.8570	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
urbanrate_c	1.049	1.033	1.065

Association of Predicted Probabilities and Observed Responses

Percent Concordant	77.8	Somers' D	0.557
Percent Discordant	22.1	Gamma	0.558
Percent Tied	0.1	Tau-a	0.276
Pairs	10170	c	0.779



The LOGISTIC Procedure

Model Information	
Data Set	WORK.NEW2
Response Variable	le
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	213
Number of Observations Used	189

Response Profile		
Ordered Value	le	Total Frequency
1	1	106
2	0	83

Probability modeled is le=1.

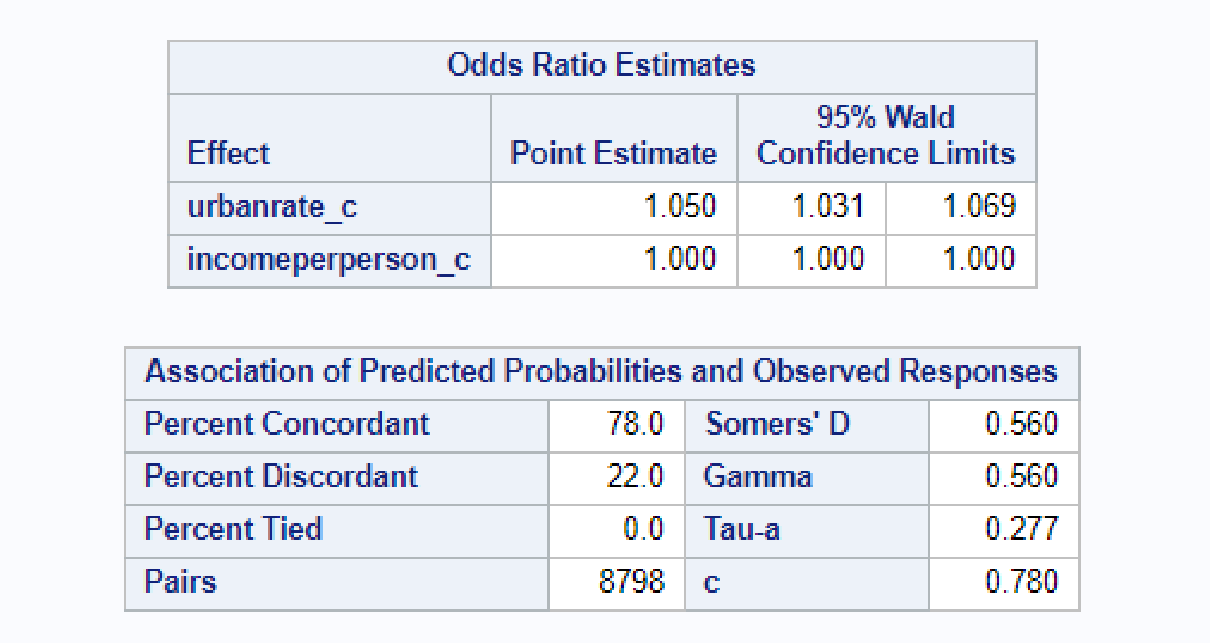
Note: 24 observations were deleted due to missing values for the response or explanatory variables.

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	261.204	218.578
SC	264.445	228.303
-2 Log L	259.204	212.578

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	46.6256	2	<.0001
Score	42.9071	2	<.0001
Wald	35.8172	2	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.3323	0.1695	3.8453	0.0499
urbanrate_c	1	0.0487	0.00916	28.2753	<.0001
incomeperperson_c	1	-1.14E-6	0.000016	0.0049	0.9440



Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
urbanrate_c	1.050	1.031	1.069
incomeperperson_c	1.000	1.000	1.000

Association of Predicted Probabilities and Observed Responses

Percent Concordant	78.0	Somers' D	0.560
Percent Discordant	22.0	Gamma	0.560
Percent Tied	0.0	Tau-a	0.277
Pairs	8798	c	0.780

Variables Used:

- Urbanisation Rate - explanatory variable
- Income per Person - explanatory variable
- Life Expectancy - response variable

All variables used in the analysis are quantitative.

➢ Explanatory variables were standardized for the Logistic procedures.

I used the centered Urbanisation Rate and centered Income per Person as the explanatory variables.

➢ Response variable (Life Expectancy) is binned into 2 categories.

In this logistic model I coded my response variable of Life Expectancy as 0 if the country has an average Life Expectancy below or equal to 69.7535236, and as 1 on the other hand.

Hypothesis:

There is a strong correlation between Life Expectancy and Urbanisation Rate.

Summary:

I have carried out a two stage analysis as part of this experiment. First, I ran logistic regression for the primary explanatory variable and response variable. After receiving positive results, I used the second explanatory in order to check whether it is significant or, on the contrary, confounding the relationship. The results from the two stages are as below:

- Stage 1

  The primary explanatory variable “Urbanisation Rate” has a significant relationship with the response variable “Life Expectancy” (p<0.0001). The null hypothesis may be rejected. The likelihood ratio in testing Null Hypothesis gives p<.0001.

  The explanatory variable (parameter estimate = 0.0477 p-value p<0.0001, odds ratio= 1.049) shows that countries with high Urbanisation rates are 1.049 times more likely to have average Life Expectancy more than 69.7535236.

  There is 95% confidence that the likelihood falls between 1.033 and 1.065.

- Stage 2

  After adding the second explanatory variable “Income per Person”, the correlation with “Life Expectancy” remains significant with p<0.0001 and 0.9440 for “Urbanisation Rate” and “Income per Person” respectively. Therefore, “Income per Person” does not confound the results.

  This time the odds ratio for Urbanisation Rate is 1.050. Countries with high Urbanisation rates are 1.050 times more likely to have high Life Expectancy. There is 95% confidence between 1.031 and 1.069.

  The odds ratio for average Income per Person is 1.000 and there is 95% confidence between 1.000 and 1.000.

The results support my original hypothesis of the significant and positive relationship between the Life Expectancy and the Urbanisation Rate. It appears that people have higher life expectancy with their lives in countries where urbanisation rate is high. There was no evidence of confounding for the association between my primary explanatory variable (Urbanisation Rate) and the response variable (Life Expectancy). After adding the second explanatory variable (Income per Person) the relationship remained statistically significant.