# GOLDMineR® : Stouffer's American Soldier Data

This note presents an analysis of a famous data example in GOLDMineR® and shows unique insights into the effects present in the data. This note also shows how you can use an odds framework to understand effects in the data.

In his 1972 article "A Modified Multiple Regression Approach to the Analysis of Dichotomous Variables," Leo Goodman presents a version of data presented by Stouffer et al. in their study "The American Soldier." Stouffer's study was published in 1949, and was based upon surveying American soldiers in World War II. The data should be understood in the context of the segregated American military and society of the time. The data have been analyzed many times, and Goodman's approach is both very fruitful for understanding what is going on in the data, as well as superior to other approaches used in the past. Goodman's analysis is replicable in SPSS Genlog or in GOLDMineR® .

Goodman presents a cross-classification of soldiers with respect to four dichotomized variables:

- Race: Black (b) or White (w).
- Region of origin: North (on) or South (os).
- Present camp location: North (pn) or South (ps).
- Preference as to camp location: North (rn) or South (rs).

Goodman fits a logit model in which Preference is a response variable and Race, Region, and Present camp location are predictors.

First, examine the two-way marginal tables relating each predictor to Preference. Here is the two-way marginal table relating Race to Preference for camp location.

| | | | PREFER | | |
|---|---|---|---|---|---|
| | | | rn | rs | Total |
| RACE | b | Count | 2027 | 2268 | 4295 |
| | | % within RACE | 47.2% | 52.8% | 100.0% |
| | w | Count | 2024 | 1717 | 3741 |
| | | % within RACE | 54.1% | 45.9% | 100.0% |
| Total | | Count | 4051 | 3985 | 8036 |
| | | % within RACE | 50.4% | 49.6% | 100.0% |

You can understand the data in terms of simple percentage differences. 52.8% of the Blacks prefer residing in the South relative to the North, while only 45.8% of the Whites prefer residing in the South relative to the North. Therefore, Blacks are about 7% more likely than Whites to prefer the South.

You can also understand the data in terms of odds. Blacks prefer the South to the North by odds of 2268 / 2027 or 1.119 to 1. Whites prefer the South to the North by odds of 1717 / 2024 or 0.848 to 1. Therefore, Blacks are 1.119 / 0.848 or 1.32 times as likely as Whites to prefer the South.

Here is the two-way marginal table relating Region of Origin to Preference for camp location.

| | | | PREFER | | |
|---|---|---|---|---|---|
| | | | rn | rs | Total |
| ORIGIN | on | Count | 3092 | 958 | 4050 |
| | | % within ORIGIN | 76.3% | 23.7% | 100.0% |
| | os | Count | 959 | 3027 | 3986 |
| | | % within ORIGIN | 24.1% | 75.9% | 100.0% |
| Total | | Count | 4051 | 3985 | 8036 |
| | | % within ORIGIN | 50.4% | 49.6% | 100.0% |

In percentage terms, 23.7% of those from the North preferred the South, while 75.9% of those from the South preferred the South, for a large percentage difference of 52.2%. In odds terms, those originating in the North prefer the South by odds of 958 / 3092 or 0.31 to 1, while those originating in the South prefer the South by odds of 3027 / 959 or 3.156 to 1. Therefore, those originating in the South are 3.156 / .31 or 10.186 times as likely as those originating in the North to prefer the South.

Here is the two-way marginal table relating Present camp location to Preference for camp location.

| | | | PREFER | | |
|---|---|---|---|---|---|
| | | | rn | rs | Total |
| PRESENT | pn | Count | 1829 | 644 | 2473 |
| | | % within PRESENT | 74.0% | 26.0% | 100.0% |
| | ps | Count | 2222 | 3341 | 5563 |
| | | % within PRESENT | 39.9% | 60.1% | 100.0% |
| Total | | Count | 4051 | 3985 | 8036 |
| | | % within PRESENT | 50.4% | 49.6% | 100.0% |

In percentage terms, 26% of those presently in the North prefer the South, while 60.1% of those presently in the South prefer the South, for a percentage difference of about 34%. In odds terms, those presently in the North prefer the South by odds of 644 / 1829 or 0.363 to 1, while those presently in the South prefer the South by odds of 3341 / 2222 or 1.5 to 1. Therefore, those presently in the South are 1.504 / 0.363 or 4.143 times as likely as those presently in the North to prefer the South.

To sum up so far: Analyzing the two-way marginal tables reveals variable importance order to be Origin, Present, and Race. However, the two-way tables above show the effect of each predictor as if it were the only one. Analogous to multiple regression, what is the effect of each predictor given the others?

One way to get a sense of that is to examine the four-way cross-classification of all variables. This is shown in the next figure.

| Race | Origin | Present | Prefer north | Prefer south | Odds |
|---|---|---|---|---|---|
| Black | North | North | 387 | 36 | 0.093 |
| Black | North | South | 876 | 250 | 0.285 |
| Black | South | North | 383 | 270 | 0.705 |
| Black | South | South | 381 | 1712 | 4.493 |
| White | North | North | 955 | 162 | 0.170 |
| White | North | South | 874 | 510 | 0.584 |
| White | South | North | 104 | 176 | 1.692 |
| White | South | South | 91 | 869 | 9.549 |

Reading the above table, for example, for a Black Northerner in a Northern camp, the odds are 36 to 387 (0.093 or about 1 to 10.75) that he will prefer a Southern camp or 387 to 36 (10.75 to 1) that he will prefer a Northern camp. Examining the table, you can see how the observed odds vary as the predictor variables vary.

Using GOLDMineR® , you can:

- Develop a model that describes quantitatively how the odds in the above table are affected by the predictor variables.
- Test whether the model fits the data.
- Measure how well the model fits the data.
- Assess the statistical significance of the model parameters.
- Estimate the main effects of the predictor variables and their effect on the odds.
- Obtain expected frequencies given the model.

The GOLDMineR® Define Model window appears in Figure 4 below.


Specify Prefer as the dependent variable, and specify Race, Origin, and Present camp location as predictors. The model fitted is a main effects model.

Figure 5 (below) shows the Association Summary.

| Association Summary | $L^2$ | df | p-value | $R^2$ | phi |
|---|---|---|---|---|---|
| Explained by Model | 3086.51 | 3 | 2.6e-669 | 0.345 | 0.7551 |
| Residual | 24.96 | 4 | 5.1e-5 | | |
| Total | 3111.47 | 7 | 6.5e-669 | | |

The Total $L^2$ (likelihood ratio chi-square) shown, 3111.47, is the same as that reported above in the test of joint independence. The main effects model $L^2$ is 3086.51, which is 99.2% of the total. The residual $L^2$ of 24.96 on 4 degrees of freedom is statistically significant but relatively small. The model R-squared is 0.345. Note that for dichotomous and ordered response variables, an R-square of 1 is not always mathematically attainable, so you should interpret this coefficient with that in mind. The phi coefficient of association is 0.7551. These numbers are both sizable under the circumstances.

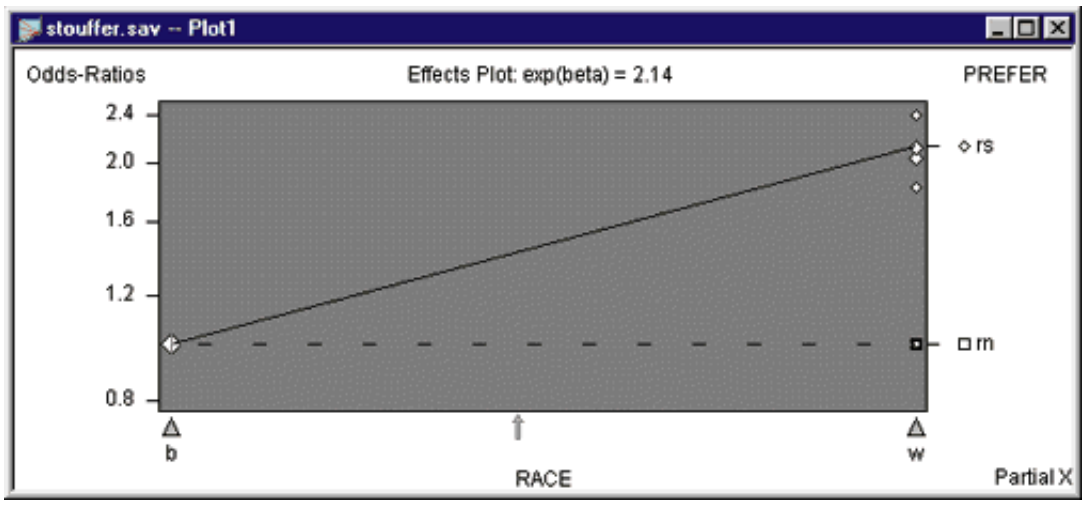
Here are Likelihood ratio tests for the individual effects in the model.

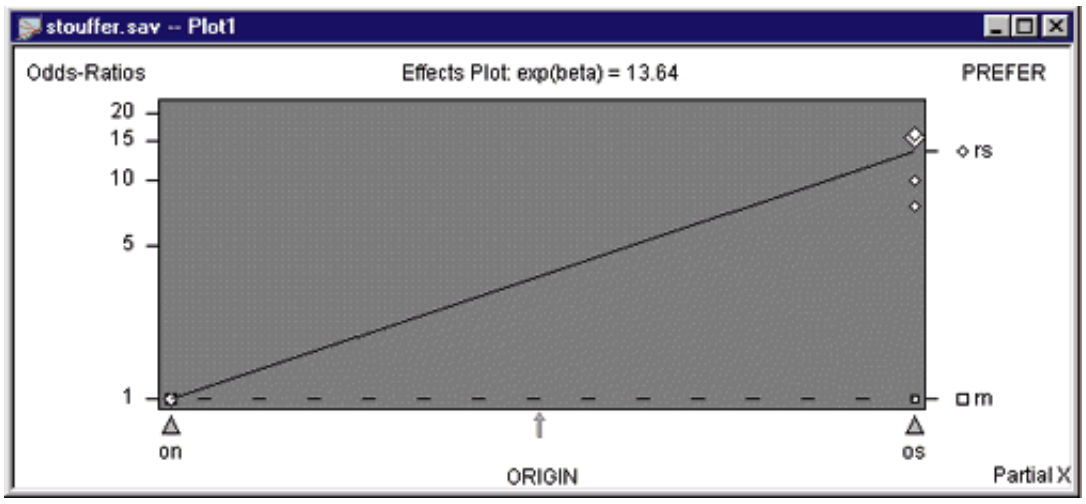
| | $L^2$(Y) | df | p-value | Beta | exp(Beta) |
|---|---|---|---|---|---|
| RACE (Fixed) | 161.39 | 1 | 5.6e-37 | 0.76 | 2.14 |
| ORIGIN (Fixed) | 2261.87 | 1 | 1.2e-493 | 2.61 | 13.64 |
| PRESENT (Fixed) | 670.05 | 1 | 9.8e-148 | 1.55 | 4.71 |

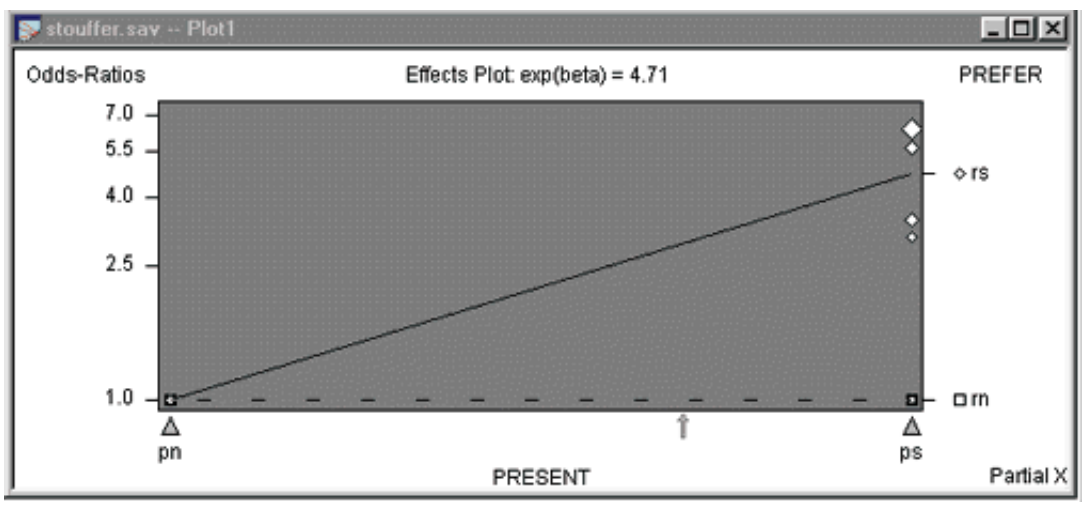
All terms are highly significant, with predictor importance order being Origin, Present, and Race. The exp(Beta) column shows coefficients that have an odds interpretation:

- Net of other terms in the model, Whites are 2.14 times as likely as Blacks to prefer the South to the North.
- Net of other terms in the model, those of Southern origin are 13.64 times as likely as those of Northern origin to prefer the South to the North.
- Net of other terms in the model, those presently in the South are 4.71 times as likely as those presently in the North to prefer the South to the North.

You can show these effects graphically in GOLDMineR® 's partial X plot. Here are the partial X plots for each of the three predictors.

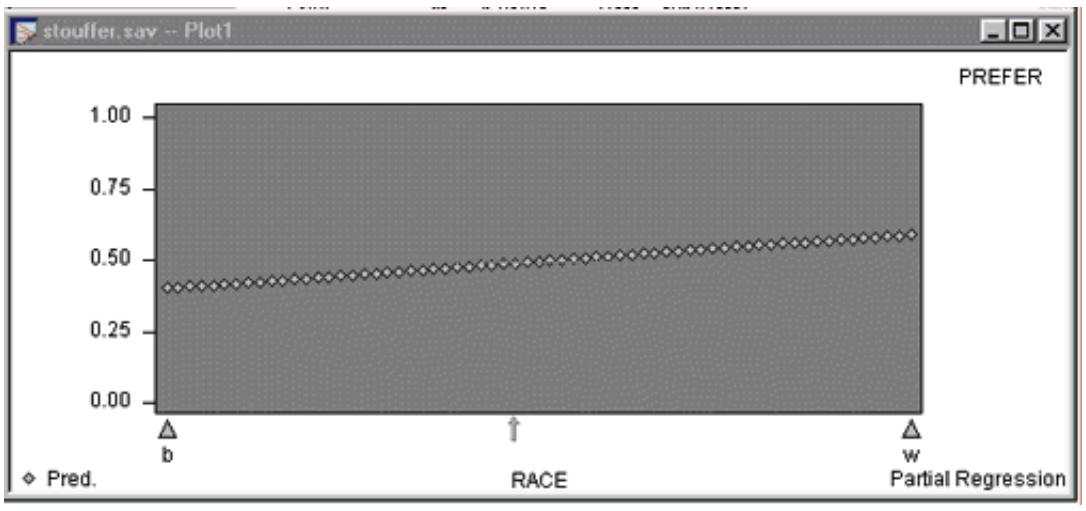


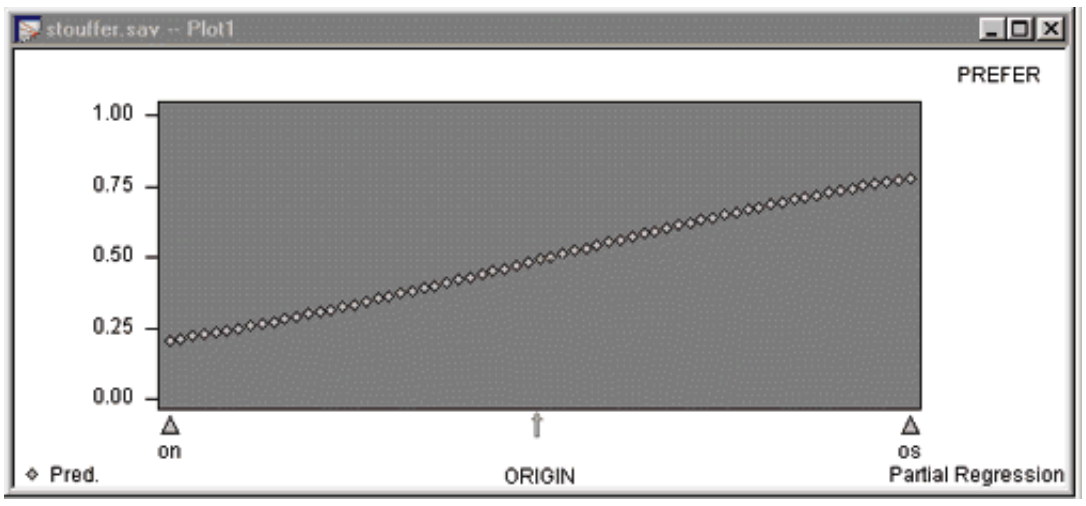


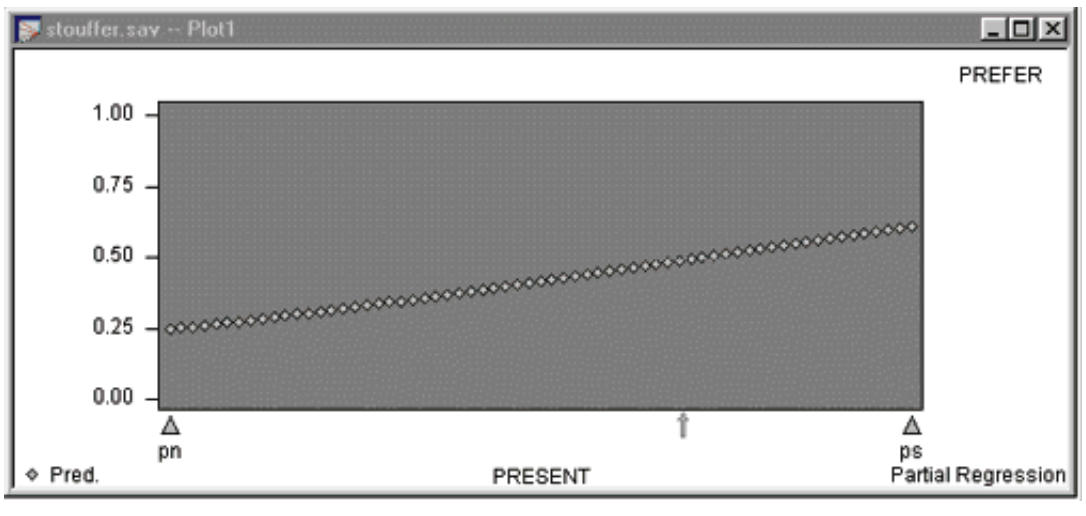




Consider the Race plot. The Partial X plot is so-named because it places the predictor categories for a given predictor on the horizontal axis. The baseline for Race is Black, so the odds comparison is White versus Black. The baseline for Prefer is "Prefer the North," so the odds comparison is "Prefer the South" versus "Prefer the north." The diagonal solid line is the expected odds ratio line, while the diamonds are the corresponding observed odds ratios observed in the table for preferring the South to the North. The expected odds, which is the slope of the line, is 2.14.

Another useful GOLDMineR® plot is the Partial regression plot. With the response variable coded 0,1, this plot shows the probability of a "1" response, here "Prefers the South," as a function of a given predictor's values conditional on values of the other predictors. One way to scale the plot is to use weighted average scoring for all variables. That is, with this scoring scheme, when viewing a particular partial regression plot, keep in mind that all other predictors are set to their mean value. This practice is analogous to what is sometimes done in logistic regression. To understand the impact of a given predictor, you plug in its low and high values while setting the other predictors to their means. Of course, you are free to set the other predictors to other values of interest.

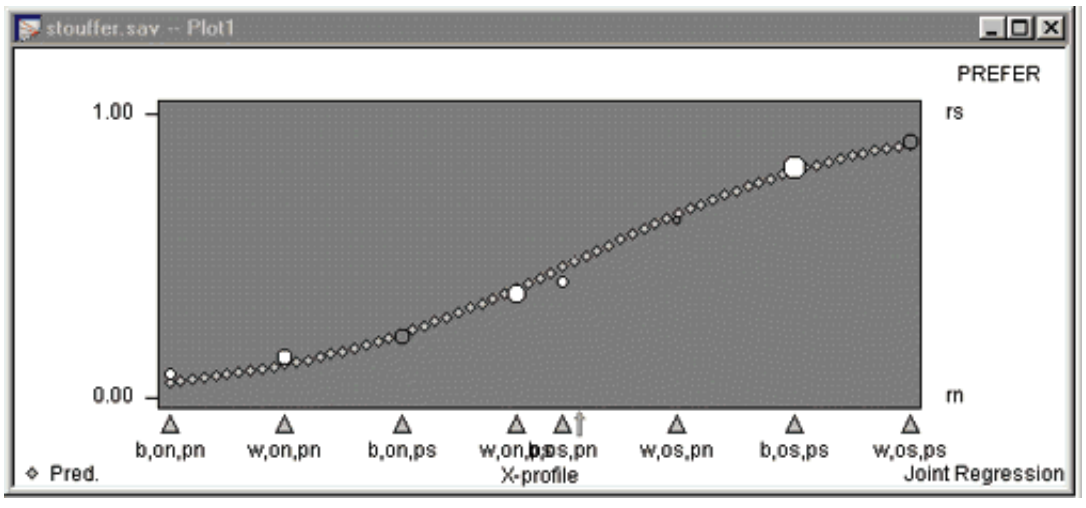



Note that the above three plots are on the same vertical scale.

A related GOLDMineR® plot that puts the multiple predictor information together in one plot is the joint regression plot. This plot shows categories of a "joint X" variable formed by crossing predictor categories on the horizontal axis.




This plot shows the probability of Preferring the South relative to the North on the vertical axis, with each joint-X category on the horizontal axis. The connected diamonds are the fitted model, while the circles are the probabilities for the observed categories of the joint-X variable. Note the extremes. At the left, Blacks originally from the North who are presently located in the North have the lowest probability of preferring the South to the North, while Whites originally from the South who are presently located in the South have the highest probability of preferring the South. The lack of fit indicated in the Association Summary above manifests itself in departures of the 8 observed points from the fitted curve. While the departures appear relatively slight, they are based on some sizable category frequencies. The circle size indicates the relative category size. Five of the eight points shown are whited out, indicating that they have statistically significant adjusted residuals. In order to see the value of the adjusted residual, click on the point in question. Or, you can open a Table Window in GOLDMineR® and view the adjusted residual values. Specify

```
Window

        New Table
```

Then specify

```
Table

        Adjusted residuals
```

Here is a portion of that table.

stouffer.sav -- Table1

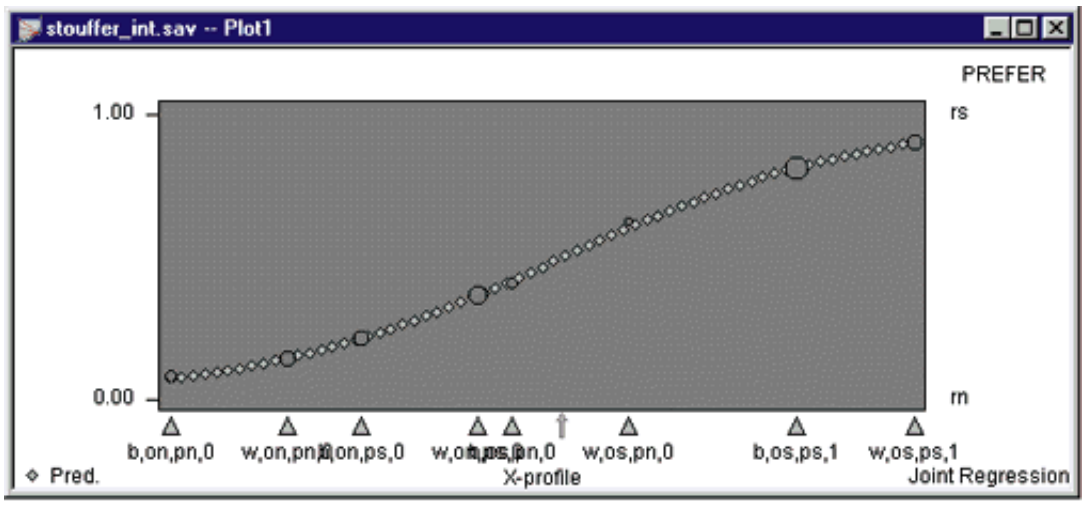
| (Joint X) | | | X-profile | | | |
|---|---|---|---|---|---|---|
| | | | b,on,pn | w,on,pn | b,on,ps | w,on,ps |
| PREFER | Y-scores | weights | 0.08 | 0.07 | 0.19 | 0.16 |
| rs | 1.00 | 0.50 | 36 | 162 | 250 | 510 |
| rn | 0.00 | 0.50 | 387 | 955 | 876 | 874 |
| | | Y-ref | 119.19 | 396.21 | 470.39 | 669.12 |
| | | adj.res | 2.39 | 3.63 | -1.07 | -3.08 |
| | average | score | 0.09 | 0.15 | 0.22 | 0.37 |

Note that the pattern of signs of the adjusted residuals is + + - - - - + +. This suggests that an added interaction term involving Origin and Present camp location might improve the fit.

Indeed, adding an interaction term to the model reduces the residual $L^2$ to 1.45 on 3 degrees of freedom, with a p value of 0.69.

| Association Summary | L² | df | p-value | R² | phi |
|---|---|---|---|---|---|
| Explained by Model | 3110.03 | 4 | 7.2e-673 | 0.349 | 0.7472 |
| Residual | 1.45 | 3 | 0.69 | | |
| Total | 3111.47 | 7 | 6.5e-669 | | |

Here is the joint regression plot with this revised model.




Note the better fit of the fitted curve to the observed points.

Here are the individual coefficients with the revised model.

| | L²(Y) | df | p-value | Beta | exp(Beta) |
|---|---|---|---|---|---|
| RACE (Fixed) | 151.20 | 1 | 9.5e-35 | 0.74 | 2.10 |
| ORIGIN (Fixed) | 468.74 | 1 | 6.0e-104 | 2.18 | 8.85 |
| PRESENT (Fixed) | 213.36 | 1 | 2.5e-48 | 1.22 | 3.39 |
| O_BY_P (Fixed) | 23.52 | 1 | 1.2e-6 | 0.60 | 1.81 |

Of course, the added interaction term complicates things a bit, but results in better fit.