When It's Not Random Chance: Creating Propensity Scores using SAS® EG

Josie Brunner, Austin Independent School District, Austin, Texas

Abstract

While randomized samples are ideal for hypothesis testing, they are not always possible, especially when evaluating programs where there is a self-selection process for treatment or control group. One quasi-experimental design approach is to use propensity scores to match treatment and control units to reduce selection bias in observable pre-treatment characteristics. This paper will focus on why and when propensity score analysis (PSA) should be included in a research design and will demonstrate how propensity scores can be created very simply using SAS EG 4.3.

Introduction

Ideally, researchers prefer randomized samples in their designs for causal inference because the estimated effect is assumed unbiased. In other words, the control and the treatment groups are assumed equivalent because of the subjects' random assignment into the groups. Resulting difference between the treatment group and the control group is attributed to the treatment effect.

$$\text{Treatment Effect} = \overline{Y}_{\text{Treatment}} - \overline{Y}_{\text{Control}}$$

However, randomized trials are not always possible, especially for studies that are retrospective, have ethical concerns when involving human subjects, or other logistical reasons. This is especially useful for education because selection for program intervention is usually driven by student characteristics and generally tries to be as inclusive as possible for those who qualify. Quasi-experimental designs are practical solutions to this dilemma. Methods for quasi-experimental designs try to estimate the pre-existing differences (i.e., selection bias) between the treatment and control groups to provide a more reliable estimate of the treatment effect. Propensity score analysis can be used for counterfactual models, in which investigation focuses on a particular cause for an outcome rather than all possible causes for an outcome (Holland, 1986).

Propensity Score Analysis

A propensity score (p-score) is the conditional probability for the unit's assignment into a condition based on a set of covariates (Rosenbaum & Rubin, 1983). The conditional probability in a randomized experiment, for example, equals .5 when there are two conditions and each unit has an equal chance of receiving the treatment (e.g., the fair coin toss). In non-randomized experiments, how conditions are assigned to units is not known, and the conditional probability must be estimated.

$$\text{P-score} = \Pr(X_{\text{Treatment}} \mid \{\text{set of covariates}\})$$

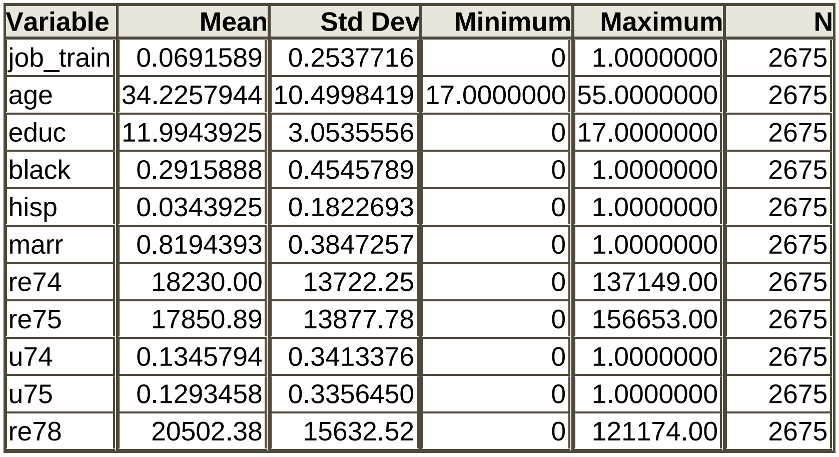
P-scores provide a means to summarize into a scalar value covariate relationships that determine treatment selection. The values can then be used to equate treatment and control groups in various ways, such as weighting, matching, analysis of covariance, or stratification. P-scores help reduce bias in non-randomized assignment so comparisons between treatment and control are between similar groups (i.e., “apples-to-apples”).

Selection bias is problematic In evaluations of programs open to any participant. Results are suspect when there are differences between those who participate in a program and those who do not, especially when those differences affect the estimate for the treatment.

Example: LaLonde’s (1986) Evaluation of the National Supported Work (NSW) Demonstration

The data for the SAS demonstration come from a subset of LaLonde’s (1986) NSW dataset of 2,675 men who were potential participants of a transitional, subsidized work experience program to help targeted groups of unemployed people. The data for the demonstration includes indicators for program participation: age, education, race/ethnicity (i.e., black or Hispanic), marital status, verified earnings in 1974 and 1975 and unemployment status in 1974 and 1975 (Figure 1). The measurable outcome was verified earnings in 1978. The evaluation question was: did people who participated in the job training program have higher incomes after the program than did those who did not participate in job training?

Figure 1. Summary of the National Supported Work (NSW) Demonstration Example Dataset



Variable	Mean	Std Dev	Minimum	Maximum	N
job_train	0.0691589	0.2537716	0	1.0000000	2675
age	34.2257944	10.4998419	17.0000000	55.0000000	2675
educ	11.9943925	3.0535556	0	17.0000000	2675
black	0.2915888	0.4545789	0	1.0000000	2675
hisp	0.0343925	0.1822693	0	1.0000000	2675
marr	0.8194393	0.3847257	0	1.0000000	2675
re74	18230.00	13722.25	0	137149.00	2675
re75	17850.89	13877.78	0	156653.00	2675
u74	0.1345794	0.3413376	0	1.0000000	2675
u75	0.1293458	0.3356450	0	1.0000000	2675
re78	20502.38	15632.52	0	121174.00	2675

Source. LaLonde (1986)

The goal of PSA is to create equivalent groups in which to test conditional effects. Figures 2 and 3 show box plots of the distribution of samples’ 1978 earnings by their participation in the program – Figure 2 without PSA and Figure 3 with PSA. There are many ways to use p-scores and techniques to match scores that are beyond the scope of this paper [see Coca-Perraillon (2006), Dehejia & Wahba (2002), Shadish, Cook, & Campbell (2002), Smith & Todd (2005)]. However, the box plots demonstrate how p-score matching altered the difference in the observed effect between the treatment and control groups. (For this subset, PSA did not change the final model’s outcome; “job_train” did not significantly contribute to 1978 earnings.)

Figure 2. Box Plot of 1978 Real Earnings, by Job Training (0 = did not participate, 1 = participated)

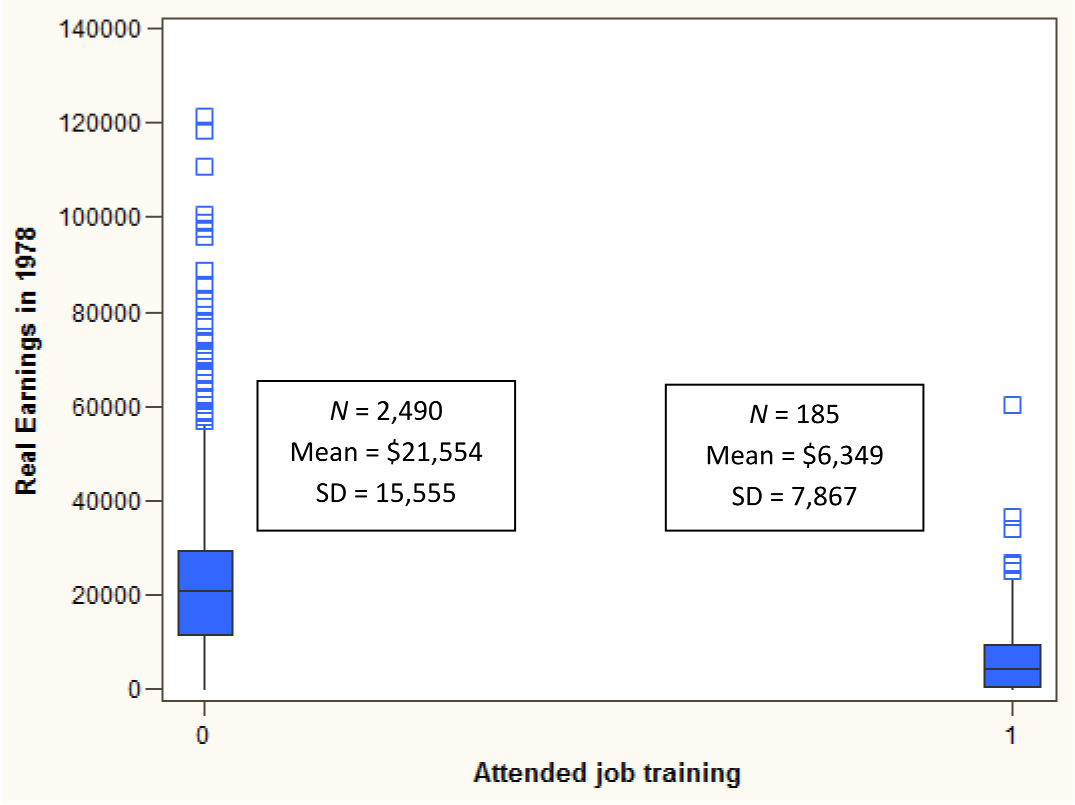
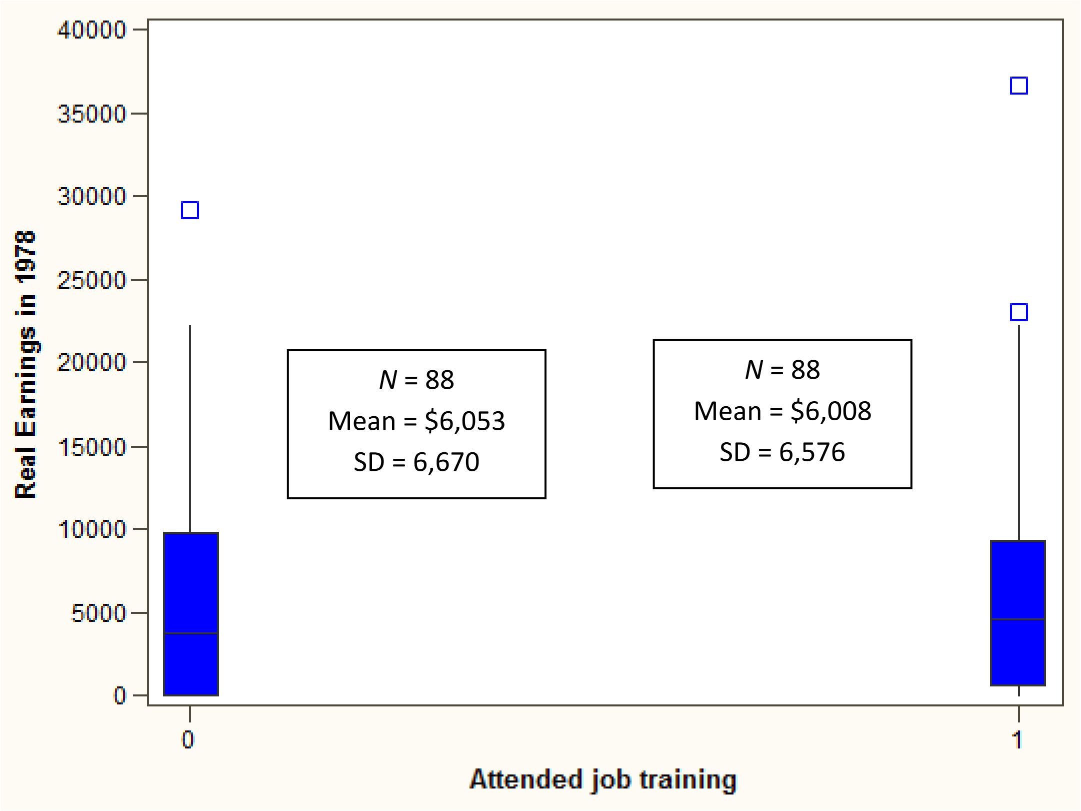


Figure 3. Box Plot of 1978 Real Earnings after Propensity Score Matching using Caliper Method without Replacement, by Job Training (0 = did not participate, 1 = participated)



Selection of covariates

Covariates used to create p-scores should be based on current theory or literature. Furthermore, since p-scores can only include observable indicators, consideration should be placed on the extent to which unmeasured bias might influence the model and the appropriateness of PSA for particular evaluation questions. For example, in early childhood education studies where pre-treatment data cannot be collected on students who did not participate in the program, PSA should not be used although program selection bias is assumed relevant.

Covariates used in the outcome model may also be used in calculating p-scores. P-scores are a way to reduce confounding factors related to selection, but those factors may still be related to the outcome as well. If the covariate is a potential confounder, it is generally included in the p-score model.

If a matching technique is applied to p-scores, then the number of covariates (especially poorly chosen ones) might affect the number of quality matches made. For PSA to effectively reduce selection bias, the treatment and control group must have "common support" in the covariates, i.e. overlap in characteristics by both groups. Figures 4a and 4b show the distributions of p-scores before and after matching has been applied.

Figure 4a. Distribution of P-scores Before Matching in NSW data

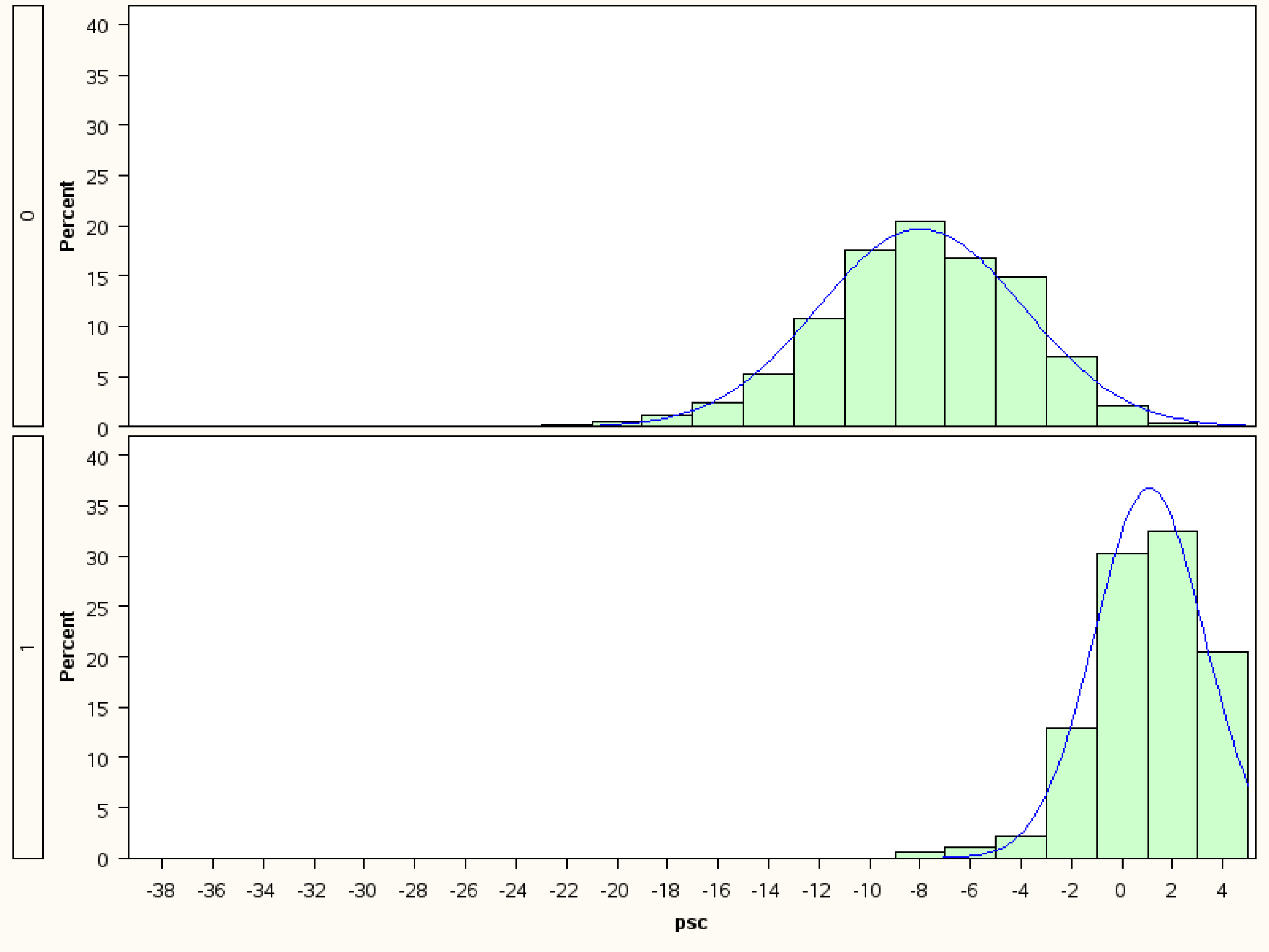
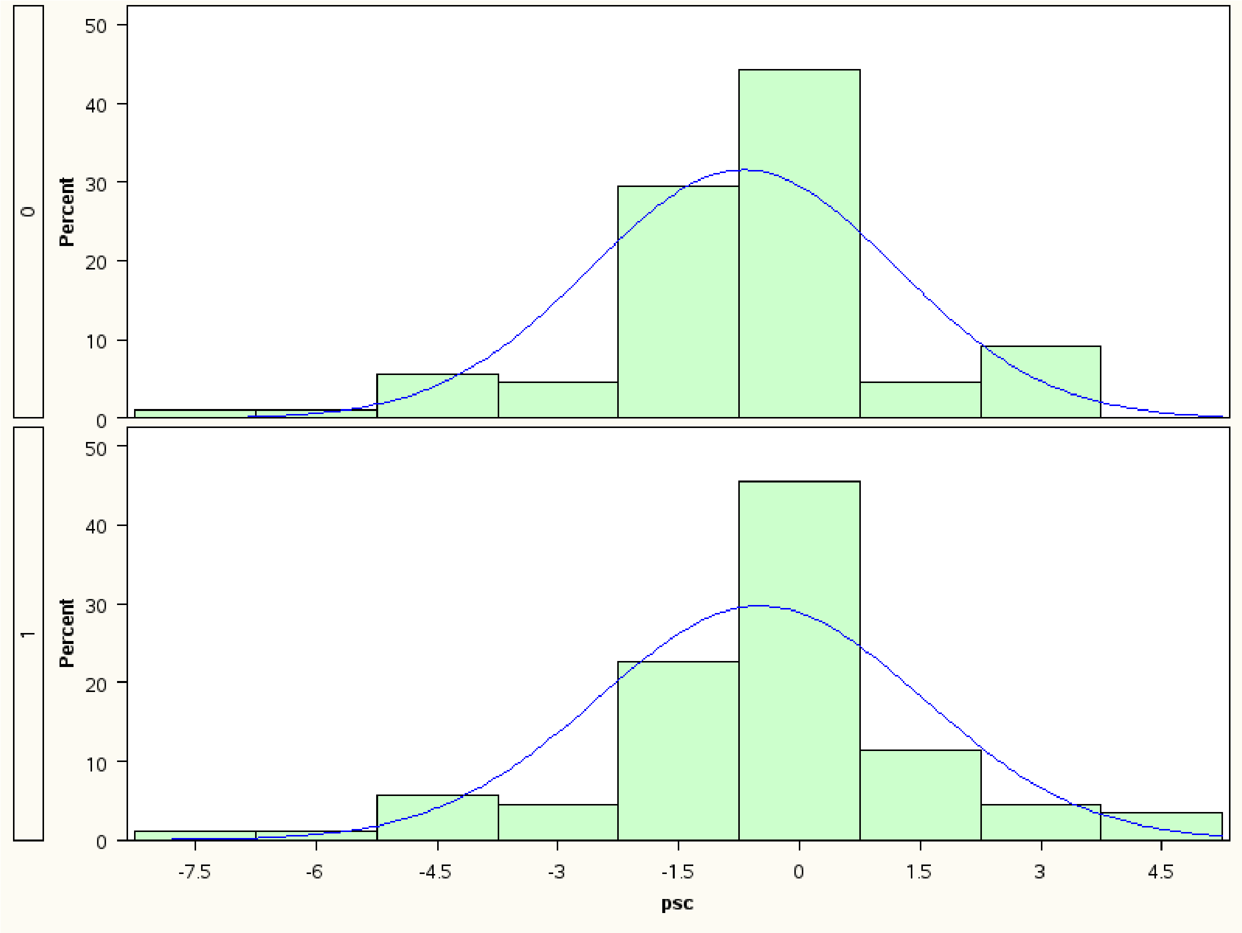


Figure 4b. Distribution of P-scores After Matching in NSW data

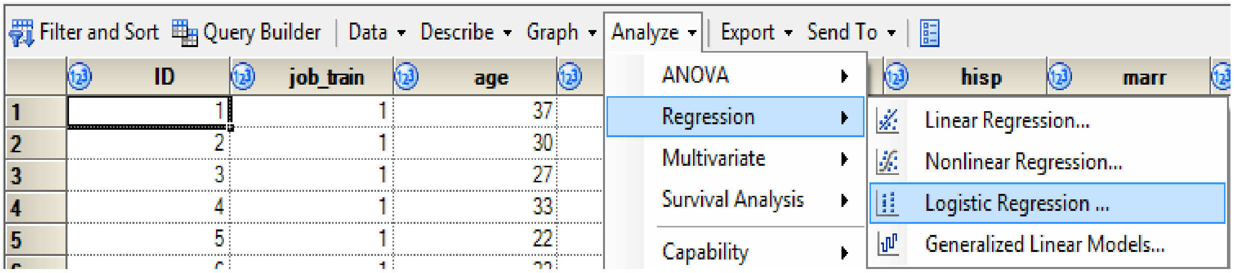


Logistic Regression to Create P-Scores

Again, p-scores are conditional probabilities, and logistic regression (Rosenbaum & Rubin; 1983) can be used to create estimates for a subject's probability into a set of binary conditions (i.e., treatment and control). Other methods, such as decision trees (Hastie, Tibshirani, & Friedman; 2001) or boosted regression (McCaffrey, Ridgeway & Morral; 2004) may also be used to estimate p-scores.

In SAS EG, **"Logistic Regression"** can be initiated under the category **"Regression"** in the **"Analyze"** menu options. This option allows users to perform logistic regression through point-and-click programming.

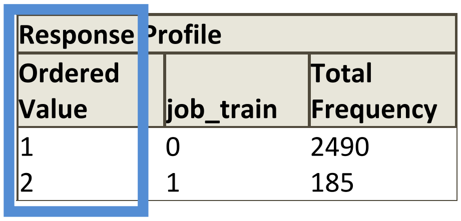
Figure 5. Analyze Menu



	ID	job_train	age	hisp	marr
1	1	1	37		
2	2	1	30		
3	3	1	27		
4	4	1	33		
5	5	1	22		

Dependent Variable

To use logistic regression, there must be only two conditions (groups) in which your subjects are categorized. This will be the dependent (i.e., outcome) variable in the p-score model. In this example, “job_train” is the indicator for program participation. (The outcome variable, real earnings in 1978, is not used.) Although “job_train” is coded as a numeric “1” or “0,” SAS EG automatically creates design variables so string data may be used (e.g., “Treatment” or “Control”). The response variable is based on the **ordered value** and not the actual value.

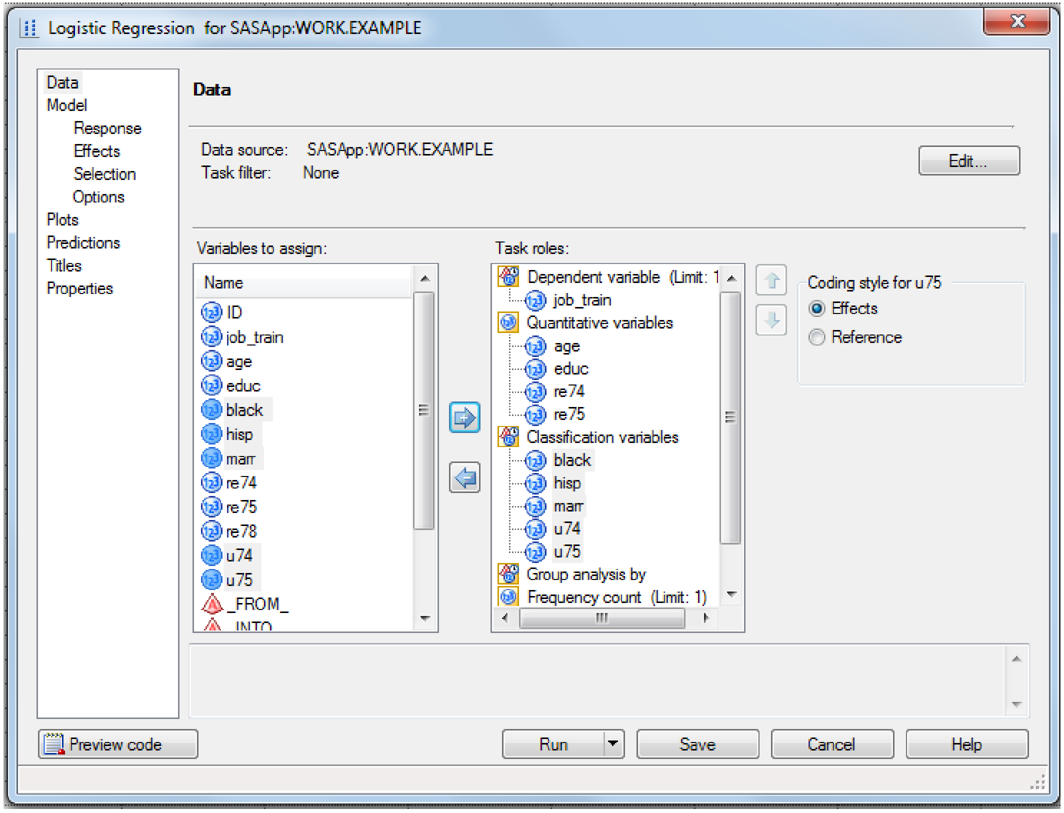


Response Profile		
Ordered Value	job_train	Total Frequency
1	0	2490
2	1	185

Adding Covariates

Under **Data** in the **Logistic Regression** window, the condition variable (i.e., job_train) can be added to **Dependent variable** by drag-and-drop method or by highlighting the variable and using the **+** arrow button. **Quantitative** (i.e., continuous) and **classification** (i.e., multinomial) variables can be added to the model as appropriate.

Figure 6. Logistic Regression>Data Window



Logistic Regression for SASApp:WORK.EXAMPLE

Data

Data source: SASApp:WORK.EXAMPLE
Task filter: None

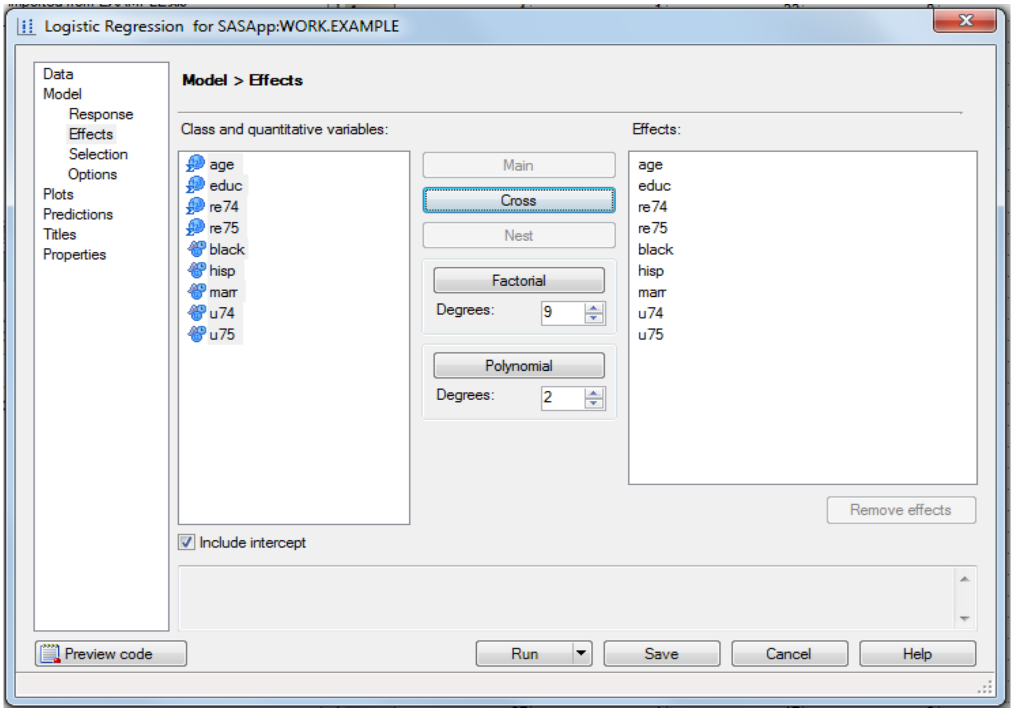
Variables to assign: ID, job_train, age, educ, black, hisp, marr, re74, re75, re78, u74, u75, _FROM_, _INTO_

Task roles:
- Dependent variable (Limit: 1): job_train
- Quantitative variables: age, educ, re74, re75
- Classification variables: black, hisp, marr, u74, u75
- Group analysis by
- Frequency count (Limit: 1)

Coding style for u75: ● Effects ○ Reference

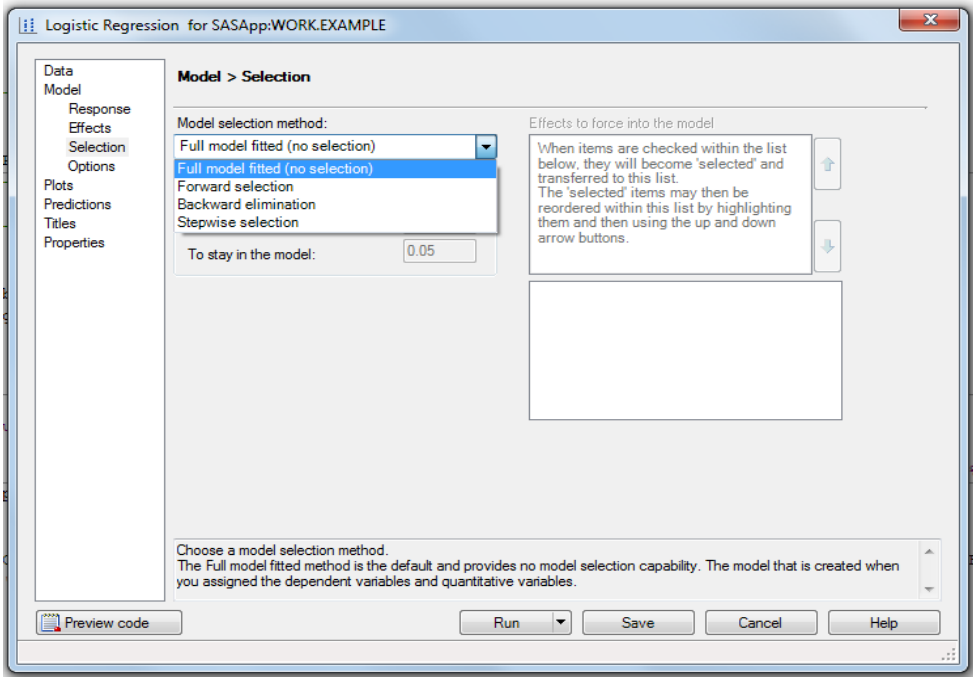
In the **Effects** window, move all covariates to right-hand **Effects** box. You may add interaction effects with the **Cross** button and polynomial terms using the **Polynomial** button to improve the p-scores created.

Figure 7. Logistic Regression>Effects Window



Under **Selection**, choose the preferred model method if all covariates do not need to be included in the model. **Full model fitted** is the default and will include all covariates in the **Effect** box.

Figure 8. Logistic Regression>Selection Window

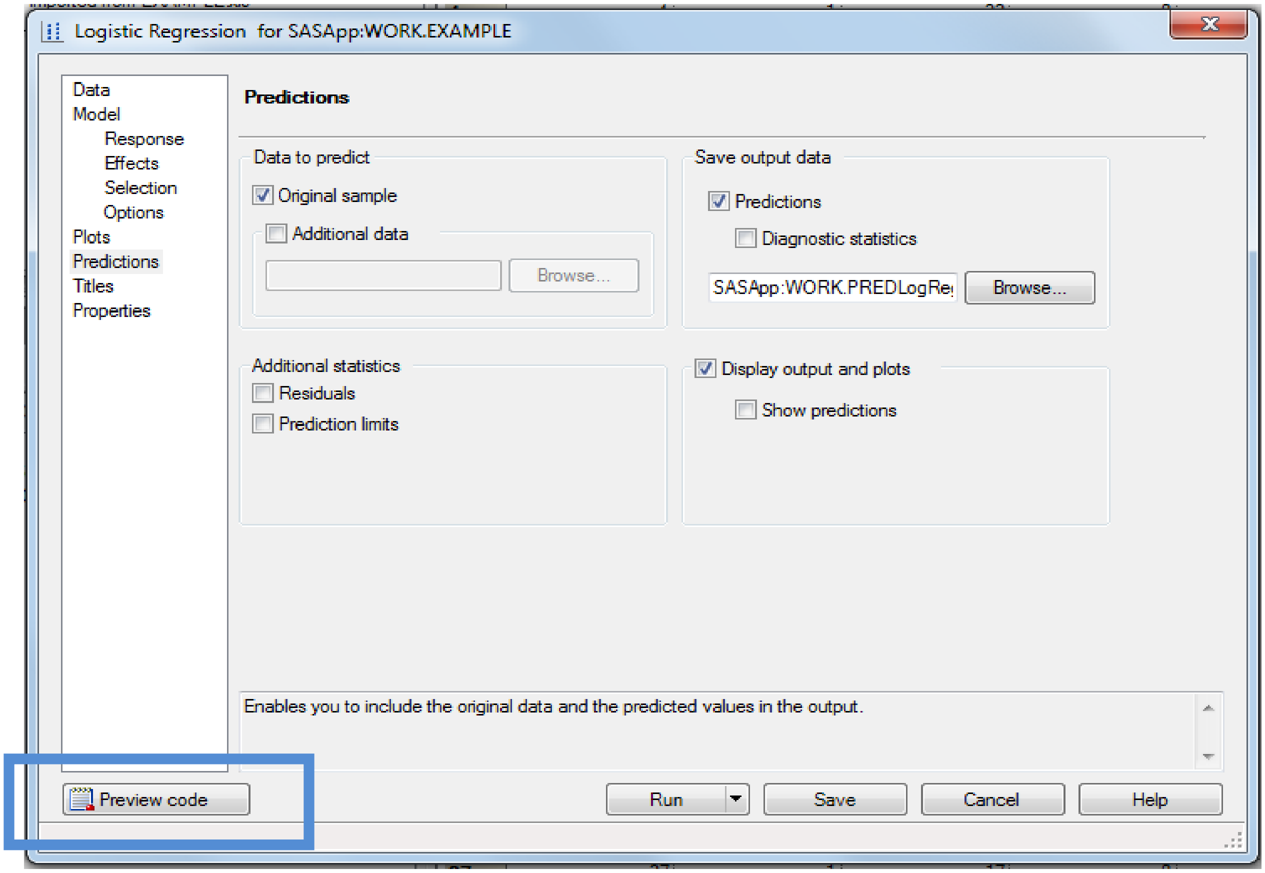


When items are checked within the list below, they will become 'selected' and transferred to this list. The 'selected' items may then be reordered within this list by highlighting them and then using the up and down arrow buttons.

Choose a model selection method.
The Full model fitted method is the default and provides no model selection capability. The model that is created when you assigned the dependent variables and quantitative variables.

To include the calculated conditional probabilities produced by the model for each observation to the original dataset, click **Original Sample** under **Predictions**. Click **Run** when finished.

Figure 9. Logistic Regression>Prediction Window



Predictions

Data to predict

☑ Original sample

☐ Additional data

Save output data

☑ Predictions

☐ Diagnostic statistics

SASApp:WORK.PREDLogRe

Additional statistics

☐ Residuals

☐ Prediction limits

☑ Display output and plots

☐ Show predictions

Enables you to include the original data and the predicted values in the output.

Preview code allows users to see the code generated by point-and-click interface. The partial code generated by SAS EG for this procedure was:

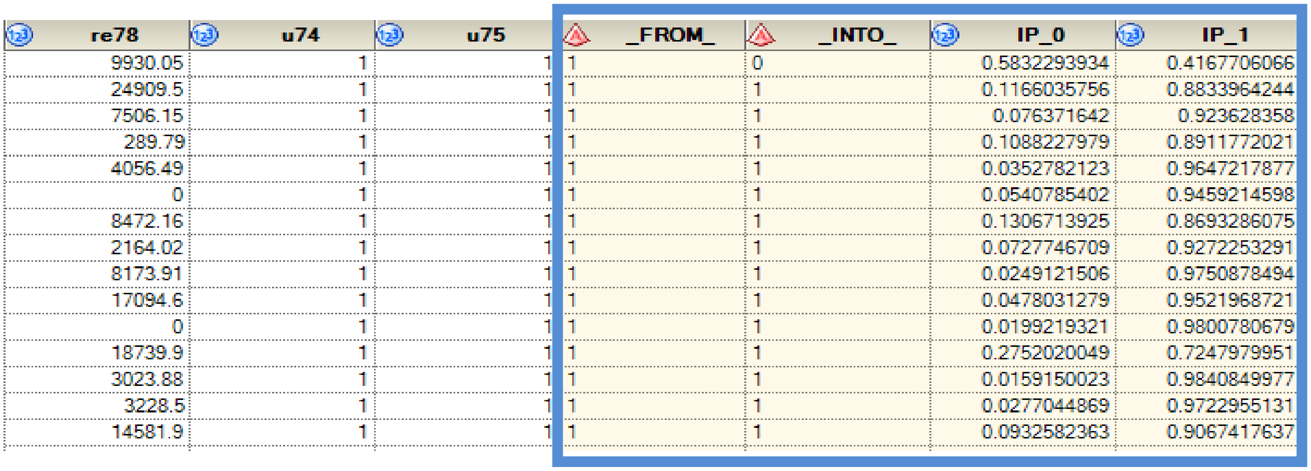
```
PROC LOGISTIC DATA=WORK.SORTTempTableSorted
           PLOTS(ONLY)=ALL;
    CLASS black  (PARAM=EFFECT) hisp (PARAM=EFFECT) marr (PARAM=EFFECT)
     u74 (PARAM=EFFECT) u75 (PARAM=EFFECT);
    MODEL job_train (Event = '0')=age educ re74 re75 black hisp marr
     u74 u75         /
           SELECTION=NONE
           LINK=LOGIT;

    OUTPUT OUT=WORK.PREDLogRegPredictions_EXAMPLE(LABEL="Logistic
     regression predictions and statistics for WORK.EXAMPLE")
           PREDPROBS=INDIVIDUAL;
RUN;
QUIT;
```

Code may be edited by first typing in the **Code** window. A pop-up box will ask if you want to create a copy of the code that can be modified. Click **Yes** if you want to make edits to the auto-generated code.

The fields **_FROM_**, **_INTO_, IP_<<Condition 1>>**, and **IP_<<Condition 2>>** are added to the original SAS dataset. In this example, **IP_0** is the condition probability of not being a participant in the job training program, and **IP_1** is the conditional probability of being a participant in the job training program. The conditional probabilities (i.e., p-scores) range from 0 to 1, where scores for **IP_1** below 0.5 are associated with not participating and above 0.5 are associated with participating. Furthermore, **IP_0 = 1 – IP_1**. **_From_** provides the observation's actual event, and **_INTO_** provides the predicted event.

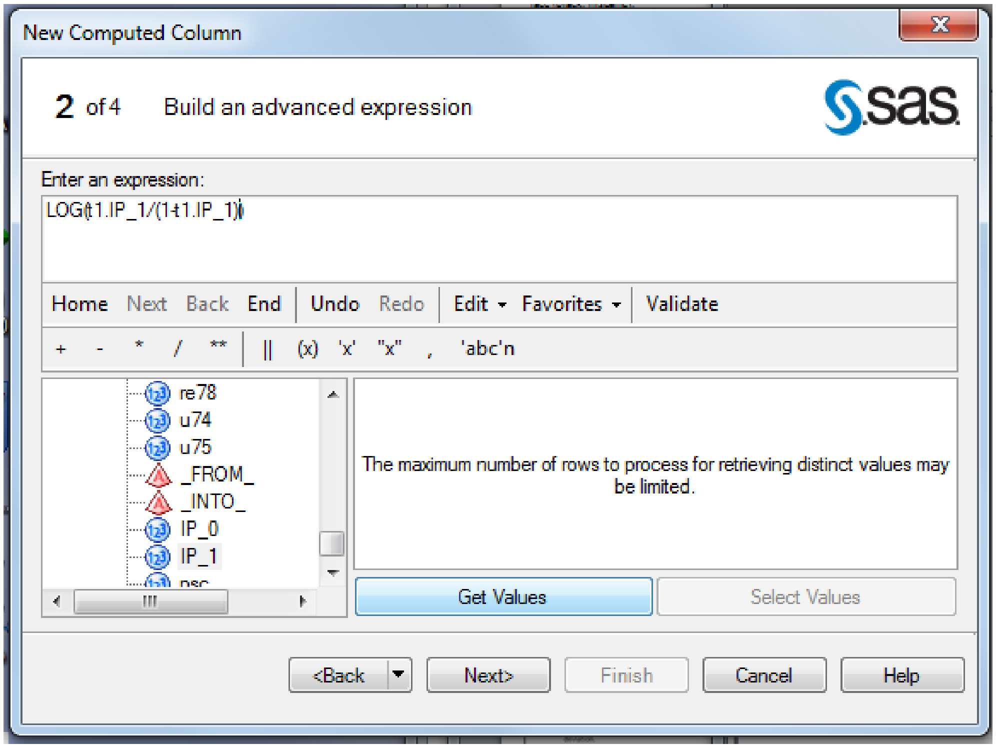
Figure 10. SAS EG View of Logistic Regression Predictions



re78	u74	u75	_FROM_	_INTO_	IP_0	IP_1
9930.05	1	1	1	0	0.5832293934	0.4167706066
24909.5	1	1	1	1	0.1166035756	0.8833964244
7506.15	1	1	1	1	0.076371642	0.923628358
289.79	1	1	1	1	0.1088227979	0.8911772021
4056.49	1	1	1	1	0.0352782123	0.9647217877
0	1	1	1	1	0.0540785402	0.9459214598
8472.16	1	1	1	1	0.1306713925	0.8693286075
2164.02	1	1	1	1	0.0727746709	0.9272253291
8173.91	1	1	1	1	0.0249121506	0.9750878494
17094.6	1	1	1	1	0.0478031279	0.9521968721
0	1	1	1	1	0.0199219321	0.9800780679
18739.9	1	1	1	1	0.2752020049	0.7247979951
3023.88	1	1	1	1	0.0159150023	0.9840849977
3228.5	1	1	1	1	0.0277044869	0.9722955131
14581.9	1	1	1	1	0.0932582363	0.9067417637

Rubin (2001) suggests p-scores be rescaled using a logit transformation (i.e., odds ratios). To recode, **IP_1** or **IP_0,** select **New** under **Computed Column** in the **Query Builder.** Choose **Advanced Expressions.** The formula for odds ratios is: **LOG(IP_1/(1-IP_1)).** In this example, the new variable name is **psc**.

Figure 11. Query>Computed Column>New>Advanced Expression Window



New Computed Column

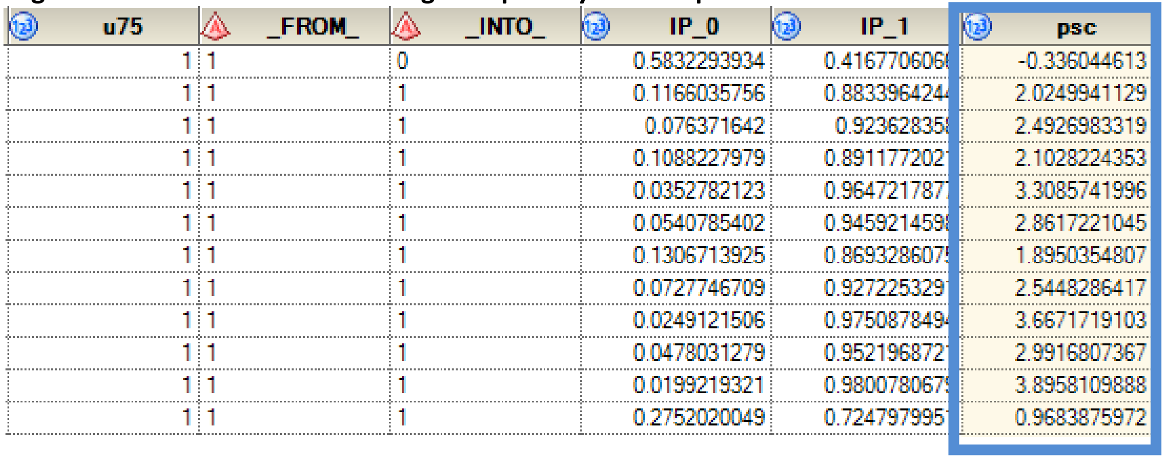
2 of 4 Build an advanced expression

Enter an expression:

```
LOG(t1.IP_1/(1-t1.IP_1))
```

The maximum number of rows to process for retrieving distinct values may be limited.

Figure 12. SAS EG View of Final Logit Propensity Scores "psc"



u75	_FROM_	_INTO_	IP_0	IP_1	psc
1	1	0	0.5832293934	0.416770606	-0.336044613
1	1	1	0.1166035756	0.883396424	2.0249941129
1	1	1	0.076371642	0.92362835	2.4926983319
1	1	1	0.1088227979	0.891177202	2.1028224353
1	1	1	0.0352782123	0.964721787	3.3085741996
1	1	1	0.0540785402	0.945921459	2.8617221045
1	1	1	0.1306713925	0.869328607	1.8950354807
1	1	1	0.0727746709	0.927225329	2.5448286417
1	1	1	0.0249121506	0.975087849	3.6671719103
1	1	1	0.0478031279	0.952196872	2.9916807367
1	1	1	0.0199219321	0.980078067	3.8958109888
1	1	1	0.2752020049	0.724797995	0.9683875972

Adequacy of Created Propensity Scores

A well-fitted model may not necessarily produce good enough p-scores to balance the distributions of covariates over the conditions (Shadish, Luellen, & Clark; 2006). In other words, while the logistic regression model may produce good predictors of subjects' probability for treatment or control group selection, there might not be enough overlap in the distributions of covariates between the treatment and control groups to create a balance.

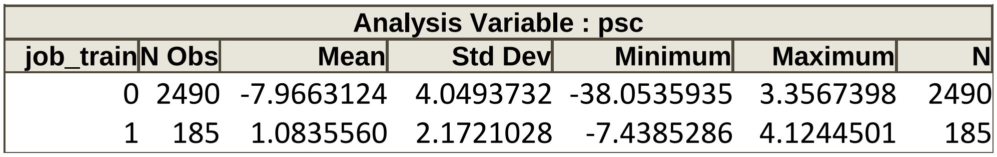
How the covariates are balanced will be determined by the method in which the p-scores are used to make statistical adjustments (e.g., weighting, stratification, matching, etc.). For this example, caliper matching (within 1 SD) without replacement was used (Cochran & Rubin, 1973). Example SAS code to perform the matching was from Coco-Perraillon (2006). The author found the code "with replacement" problematic because matches for treatment and control groups (despite being randomized) were based on first available approximate match, and some "good" control observations were never used, while others were over-represented in the final matched groups.

Since matching involves dropping, at times, a large number of control cases, methods to evaluate the quality of the propensity score (e.g., significance tests on covariate differences between the control and treatment group) may be related to reduction in power by the smaller sample size (Imai, King, & Stuart, 2008). Rubin (2001) provides three benchmarks to test the adequacy of the p-scores by examining the distribution quality of the scores:

1. The difference in the group means of the logit propensity scores should be less than half a standard deviation.

In SAS, under **Describe > Summary Statistics**, you may obtain the means and standard deviations for "psc" using the classification variable "job_train."

Figure 13. Before PSA adjustment: (1.08 – (-7.97))/2.17 = 4.17 SD



Analysis Variable : psc

job_train	N Obs	Mean	Std Dev	Minimum	Maximum	N
0	2490	-7.9663124	4.0493732	-38.0535935	3.3567398	2490
1	185	1.0835560	2.1721028	-7.4385286	4.1244501	185

Figure 14. After PSA adjustment: (-.49 – (-.69))/2.01 = .10 SD ✓

Analysis Variable : psc						
job_train	N Obs	Mean	Std Dev	Minimum	Maximum	N
0	88	-0.6928060	1.8962380	-7.4353981	3.3567398	88
1	88	-0.4863831	2.0138742	-7.4385286	3.9062371	88

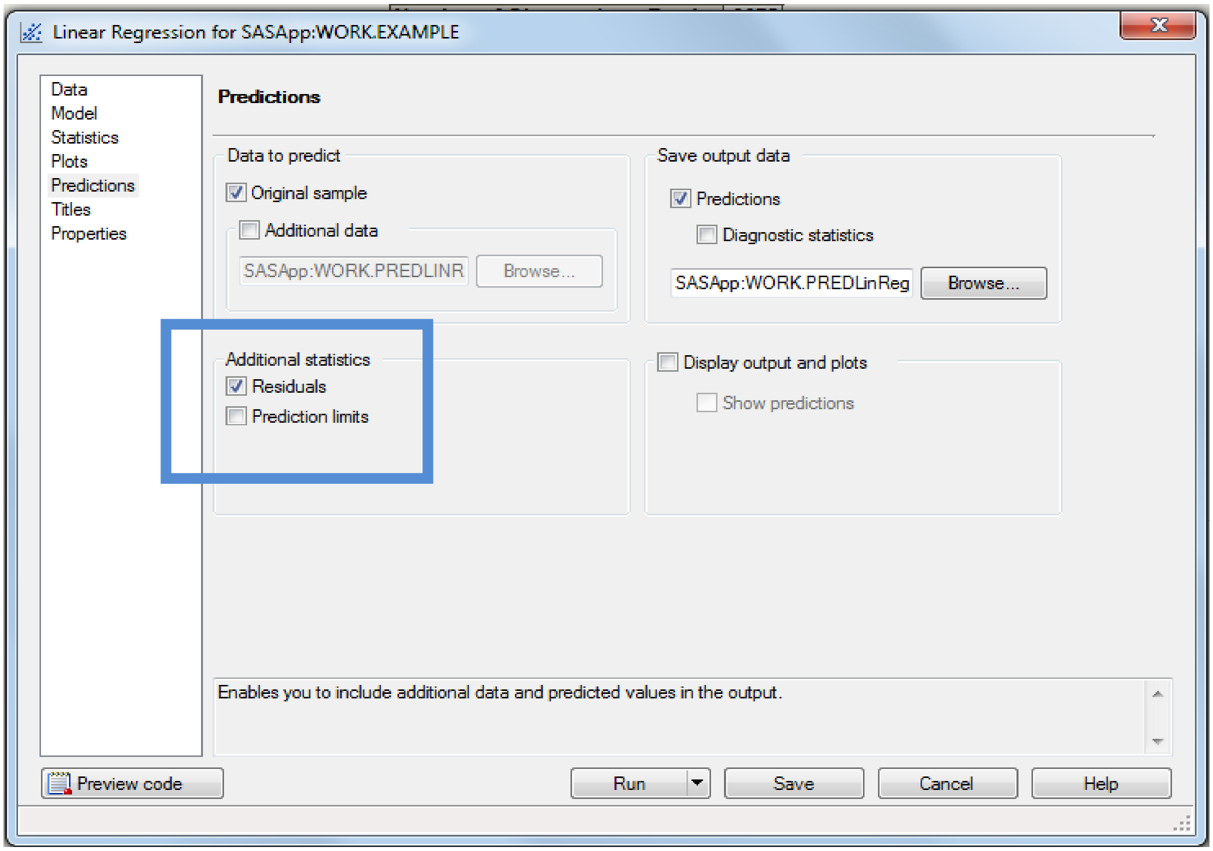
2. The ratio of the group variances of the logit propensity scores should be close to one.

Before PSA adjustment: 4.05/2.17 = 1.87
After PSA adjustment: 1.90/2.01 = 0.95 ✓

3. The ratio of the variance of the residuals of the covariates should be close to one (between 4/5 and 5/4).

To accomplish this in SAS, select **Regression** under **Analyze** and regress each covariate (individually) with p-scores. Save the residuals from the models by clicking **Residuals** in the **Predictions** window.

Figure 15. Analyze>Linear Regression>Predictions Window



Linear Regression for SASApp:WORK.EXAMPLE

Predictions

Data to predict
☑ Original sample
□ Additional data
SASApp:WORK.PREDLINR

Save output data
☑ Predictions
□ Diagnostic statistics
SASApp:WORK.PREDLinReg

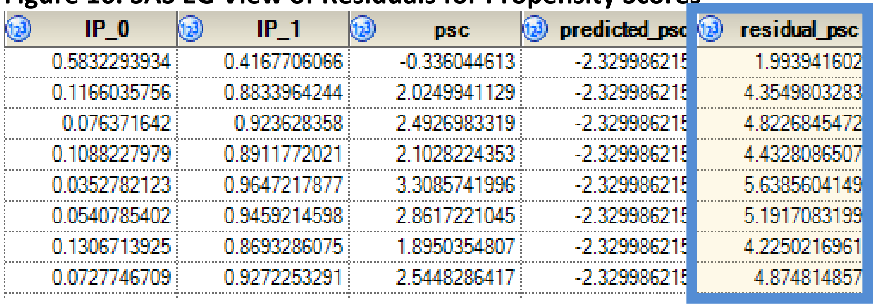
Additional statistics
☑ Residuals
□ Prediction limits

□ Display output and plots
□ Show predictions

Enables you to include additional data and predicted values in the output.

The SAS dataset will produce the field **residual_<<Propensity score name>>.** (**Predicted_psc** is the predicted value of the propensity score based on the covariate and can be ignored.)

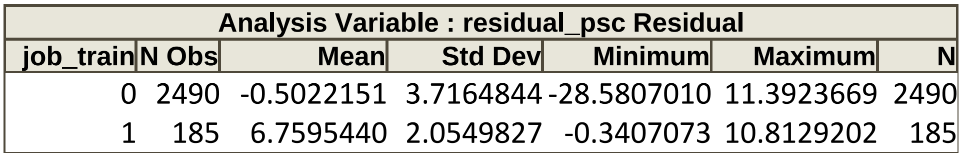
Figure 16. SAS EG View of Residuals for Propensity Scores



IP_0	IP_1	psc	predicted_psc	residual_psc
0.5832293934	0.4167706066	-0.336044613	-2.329986215	1.993941602
0.1166035756	0.8833964244	2.0249941129	-2.329986215	4.3549803283
0.076371642	0.923628358	2.4926983319	-2.329986215	4.8226845472
0.1088227979	0.8911772021	2.1028224353	-2.329986215	4.4328086507
0.0352782123	0.9647217877	3.3085741996	-2.329986215	5.6385604149
0.0540785402	0.9459214598	2.8617221045	-2.329986215	5.1917083199
0.1306713925	0.8693286075	1.8950354807	-2.329986215	4.2250216961
0.0727746709	0.9272253291	2.5448286417	-2.329986215	4.874814857

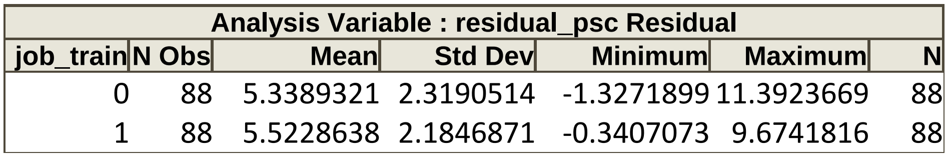
Use **Describe > Summary Statistics**, to obtain the standard deviations for “residual_psc” using the classification variable “job_train.” Figures 17 and 18 show the calculation for the variable “age.”

Figure 17. For age, before PSA adjustment: (3.71/2.05) = 1.81



Analysis Variable : residual_psc Residual						
job_train	N Obs	Mean	Std Dev	Minimum	Maximum	N
0	2490	-0.5022151	3.7164844	-28.5807010	11.3923669	2490
1	185	6.7595440	2.0549827	-0.3407073	10.8129202	185

Figure 18. For age, after PSA adjustment: (2.32/2.05) = 1.13 ✓



Analysis Variable : residual_psc Residual						
job_train	N Obs	Mean	Std Dev	Minimum	Maximum	N
0	88	5.3389321	2.3190514	-1.3271899	11.3923669	88
1	88	5.5228638	2.1846871	-0.3407073	9.6741816	88

Conclusion

PSA can be valuable when studying human subjects because random assignment is not always possible. Creating propensity scores can be done easily with SAS EG; however, the quality of the propensity scores will depend on how they are used (i.e., how well the covariates are balanced over the conditions), the quality of data, and whether unmeasured bias can be ignored. Rubin’s benchmarks should be used to evaluate the quality of the propensity scores, i.e., the degree of covariate overlap, but the quality will also be based on the method of statistical adjustment applied.

References

Cochran, W. & Rubin, D. (1973). Controlling Bias in Observational Studies. *Sankyha*. 35, 417-446.

Coco-Perraillon, M. (2006). Matching with propensity scores to reduce bias in observational studies. NESUG Conference.

Dehejia, R. H. & Wahba, S. (2002), Propensity score matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*. 84(1), 151-161. Data available at: http://www.nber.org/%7Erdehejia/nswdata.html

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association,* 81, 945-70.

Imai K., King, G., Stuart, E.A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of Royal Statistical Society*. 171(2), 481-502.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*. 76, 604-620.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies*. Psychological Methods.* 9(4), 403-425.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika,* 70, 41-55.

Rubin, D. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology.* 2, 169-188.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasiexperimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shadish, W. R., Luellen, J. K., & Clark, M. H. (2006). Propensity scores and quasi-experimentation: A testimony to the practical side of Lee Sechrest. In R. R.Bootzin (Ed.), *Strengthening research methodology: Psychological measurement and evaluation.* Washington, DC: American Psychological Association.

Smith, J. & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators. *Journal of Econometrics*. 125(1-2), 305-353.