Analyzing Household Income in Greater Louisville Area

Background

Greater Louisville Inc (GLI) conducted a random survey of persons 18 and older in the metro area to investigate household incomes. The **dataset** includes variables such as AGE, EDUC, HRS1, SPHRS1, EARNRS, CHILDS, HEAD, and INCOME.

Goals of Analysis

1. Determine if all variables in the dataset are significant in modeling household income.
2. Model the total number of hours worked per week using AGE, EDUC, and HEAD.
3. Assess the impact of an interaction term between EDUC and HRS1 on annual household income.
4. Model the likelihood of a household income exceeding $75,000 based on AGE, EDUC, and total hours worked (TOTAL).

Methodology

1. Significance of All Variables in Modeling Household Income

   Model Creation: To determine if all variables are significant in modeling household income, I start by creating a regression analysis model including all variables using SPSS. I set INCOME as the dependent variable, and AGE, EDUC, HRS1, SPHRS1, EARNRS, CHILDS, and HEAD as independent variables, then run the regression.

Coefficients[a] Full

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-63083.717	2247.415		-28.069	<.001	-67491.404	-58676.029
	Age	738.448	23.955	.446	30.827	<.001	691.467	785.428
	Educ	4625.592	116.473	.496	39.714	<.001	4397.162	4854.023
	HRS1	893.457	19.014	.693	46.990	.000	856.167	930.748
	SPHRS1	-144.417	20.028	-.122	-7.211	<.001	-183.696	-105.137
	Earners	-1841.138	562.711	-.063	-3.272	.001	-2944.742	-737.535
	Childs	166.073	233.944	.009	.710	.478	-292.744	624.891
	Head	104.542	713.219	.002	.147	.883	-1294.240	1503.325

a. Dependent Variable: Income

   Model Evaluation: Performing a general linear F-test comparing the full model to a reduced model excluding CHILDS and HEAD because the significance tests (p-values) for these variables show that they are not statistically significant in predicting household income, which justifies their exclusion in the reduced model.

Coefficients[a] Reduced

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-62843.776	2214.125		-28.383	<.001	-67186.170	-58501.382
	Age	744.353	22.467	.450	33.130	<.001	700.289	788.416
	Educ	4608.316	113.878	.494	40.467	<.001	4384.976	4831.656
	HRS1	893.781	18.988	.693	47.071	.000	856.542	931.021
	SPHRS1	-144.532	20.012	-.122	-7.222	<.001	-183.780	-105.285
	Earners	-1796.229	551.546	-.061	-3.257	.001	-2877.935	-714.524

a. Dependent Variable: Income

F test $\Rightarrow$ $\dfrac{\dfrac{1.017\times10^{9} - 1.019\times10^{9}}{5-7}}{\dfrac{2.755\times10^{11}}{7}} = 0$

Conclusion:

With the F-statistic being 0, the p-value is equal to 1. This means we fail to reject the null hypothesis, meaning there is no difference between the full and reduced models. This means we should continue to use the reduced model, which does not include the number of children or the sex of the head of the house.

2. Modeling Total Hours Worked

Model Creation:
I will start by creating a new variable called TOTAL which is the sum of HRS1 and SPHRS1, and then perform regression analysis using AGE, EDUC, and HEAD as independent variables.

Model Evaluation:
The output will provide coefficients, standard errors, t-values, p-values, and the R-squared value. These statistics help determine the significance and impact of each independent variable on the total number of hours worked per week.

Model Summary Table

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.473	0.223	0.222	8.649

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	76.790	3.995		19.223	<.001	68.955	84.624
	Age	-.787	.041	-.395	-19.344	<.001	-.867	-.707
	Educ	.618	.228	.055	2.710	.007	.171	1.064
	Head	14.091	1.367	.210	10.305	<.001	11.410	16.773

a. Dependent Variable: Total

Conclusion:
The R-squared value is 0.223, which indicates that 22.3% of the variation in the total hours is explained by this model. The model itself is significant as well. For every additional year added to age, there is an expected decrease of 0.787 in the total number of hours. For every additional year added to education, there is an expected increase of 0.618 total hours. When the respondent is labeled female, there is an expected increase of 14.091 total hours. The regression analysis reveals that age and the gender of the head of household are significant predictors of the total number of hours worked per week, while education is not.

3. Interaction Term in Modeling Annual Household Income

Model Creation:
Creating a model for annual household income using EDUC and HRS1 as independent variables and then assessing whether including an interaction term between these two variables improves the model.

Model Evaluation:
Running the regression in SPSS will provide the coefficients, t-values, and p-values. This model will help provide more insight into whether including an interaction has an impact or not.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-21677.249	2320.089		-9.343	<.001	-26227.459	-17127.038
	Educ	4501.359	154.192	.482	29.193	<.001	4198.955	4803.763
	HRS1	627.531	21.309	.487	29.449	<.001	585.738	669.323

a. Dependent Variable: Income

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-7344.599	4190.935		-1.752	.080	-15563.955	874.756
	Educ	3476.480	293.412	.372	11.848	<.001	2901.034	4051.926
	HRS1	221.002	101.420	.171	2.179	.029	22.094	419.910
	Inter	28.954	7.063	.344	4.099	<.001	15.101	42.806

a. Dependent Variable: Income

Conclusion:
The interaction term is shown to be significant, as the p-value is less than 0.001, indicating that the effect of education on household income varies with the number of hours worked per week. Including an interaction term between these two variables improved the model.

4. Modeling the Likelihood of Household Income Exceeding $75,000

Model Creation:
Using logistic regression, I aim to model the likelihood that a household's annual income exceeds $75,000 based on AGE, EDUC, and the total number of hours worked (TOTAL).

Model Evaluation:
The output is shown for the logistic model. Using the mean value for all 3 variables (48.83 years old, 14.156 years of education, and 56.54 total hours worked), I have a predicted probability of 27.58% for an income that is equal to or greater than $75,000.

The Pseudo R-squared is 0.2235 as $1 - \frac{D_k}{D_0} = 1 - \frac{1859.476}{(1859.476 + 535.705)} = 0.2235$

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	535.305	3	<.001
	Block	535.305	3	<.001
	Model	535.305	3	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1859.476[a]	.247	.343

Classification Table[a]

Observed		Predicted Greater75 0	Predicted Greater75 1	Percentage Correct
Step 1 Greater75	0	1124	144	88.6
	1	319	303	48.7
Overall Percentage				75.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Age	.049	.004	134.350	1	<.001	1.050	1.042	1.059
	Educ	.392	.024	268.630	1	<.001	1.480	1.412	1.551
	Total	.020	.002	101.999	1	<.001	1.020	1.016	1.024
	Constant	-10.019	.506	392.445	1	<.001	.000		

a. Variable(s) entered on step 1: Age, Educ, Total.

Conclusion:

This model had a percentage correct of 75.5%, which is much higher than the 67.1% of the simple expectation model. Also, age, education, and total hours worked per week are significant predictors of the likelihood of having a household income ≥ $75,000. This model also provides insights into the factors influencing the likelihood of achieving a higher household income and highlights the importance of education and work hours.

Final Thoughts

Our models provide valuable insights into household income dynamics in the Louisville metro area. The reduced model without CHILDS and HEAD is preferred for income analysis. Total hours worked can be effectively modeled using AGE, EDUC, and HEAD. The interaction term between EDUC and HRS1 is significant but does not drastically alter predictions. Lastly, logistic regression effectively models the likelihood of income exceeding $75,000.

This analysis provides a solid foundation for further investigations and decision-making for Greater Louisville Inc.