**Lab 3 – Multiple Regression Analysis with Qualitative Predictors**

**I. Data**

We will be using a common data set for all lab sessions throughout the quarter. The data set comes from the Census Bureau and uses data collected by the American Community Survey. The unit of analysis is metropolitan cities throughout the United States. There are 517 such observations in the data set. The variables in the data set include:

|  |  |
| --- | --- |
| GEO\_NAME | Metropolitan and Micropolitan Statistical Area |
| areatype | Metropolitan and Micropolitan |
| pop\_size | Total Estimated Population |
| division | US Division |
| region | Region |
| commute | Mean travel time to work |
| hh\_chil | Percent households with children <18 |
| edu\_hs | Percentage adults with High School Diploma |
| edu\_coll | Percentage adults with Bachelor's Degree |
| income | Median household income |
| income2 | Median household income (in thousands) |
| no\_car | Percentage of households without vehicle |
| female | Percent population - female |
| age | Median age |
| senior | Percent population 65 years or older |
| white | Percent Population - White |
| black | Percent Population - Black |
| indian | Percent Population - Indian |
| asian | Percent Population - Asian |
| hawaiian | Percent Population - Hawaiian |
| other | Percent Population - Other |
| hispanic | Percent Population - Hispanic |

The variables for *division* and *region* are defined by the map included on the last page. The variable *areatype* takes on two values – micro and metro – depending on the following definitions:

* A metro area contains a core urban area of 50,000 or more population
* A micro area contains an urban core of at least 10,000 (but less than 50,000) population

During the semester we will be answering the research question:

***What best predicts average commute times of US cities?***

There is a mix of qualitative and quantitative predictor variables that we will use to answer the question, using regression based approaches.

**II. Recode Variables Using Syntax**

In previous labs, we have looked at quantitative predictors of commute times – income, population size, percent population white. Qualitative predictors are also common in regression models. In this lab, we will consider the predictor *region* in a regression model predicting commute times.

**Dummy Coding**

Recall that our qualitative variable *region* is actually kept as a numeric variable in our data set with the following value labels:

1 = West

2 = Midwest

3 = Northeast

4 = South

5 = Other

If we input the *region* variable in a regression model, the results would be nonsense. We would not be able to interpret a slope coefficient. What would a “one unit change” in region mean? Since we are working with nominal level data, the intervals between numbers have no inherent meaning.

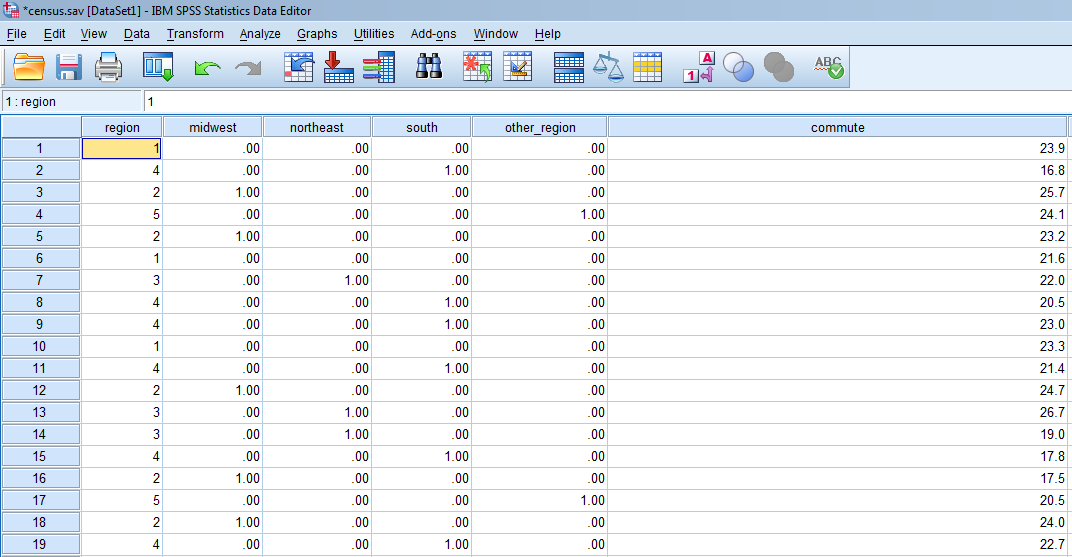
We have to code our data in a way where the intervals do mean something. One such way is dummy coding, where essentially we are creating separate dichotomous variables to indicate each region. You will need *k-1* dummy variables for each qualitative variable, where *k* is the number of levels or values in that qualitative variable. In this case, region takes on 5 different values so we need 4 different dummy variables.

The first thing we need is to define a reference group in our qualitative predictor. The reference group, in this case, is the region that all comparisons will be made with. The choice is arbitrary, but choose the group where most comparisons will be made.

We will use West as our reference group. Paste the following syntax to create dummy variables for region.

|  |
| --- |
| COMPUTE midwest = 0.  COMPUTE northeast = 0.  COMPUTE south = 0.  COMPUTE other\_region = 0.  EXE.  IF (region = 2) midwest = 1.  IF (region = 3) northeast = 1.  IF (region = 4) south = 1.  IF (region = 5) other\_region = 1.  EXE. |

This creates four dichotomous variables. A city in the West will have a zero for all four dummy variables. Otherwise, there will be a one for that region. Now the intervals are more meaningful, where a one unit change in *midwest* literally means whether a city is in the Midwest or not.

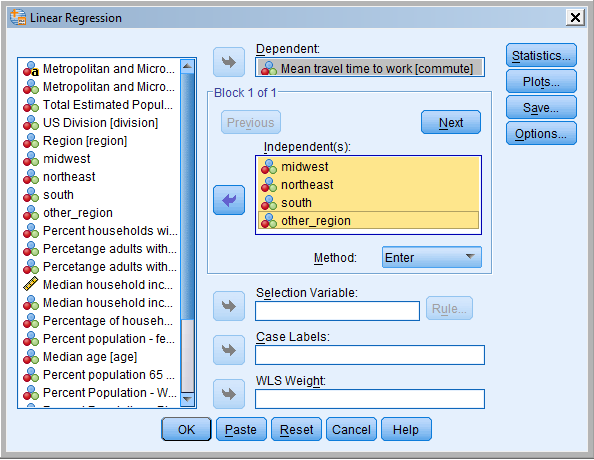


Also take a moment to find the mean commute times for all five regions.

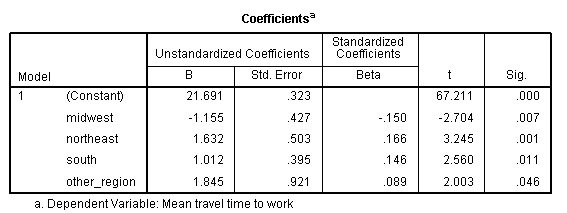
|  |  |
| --- | --- |
| Region | Mean commute time (in mins.) |
| West | 21.69 |
| Midwest | 20.54 |
| Northeast | 23.32 |
| South | 22.70 |
| Other | 23.54 |

**III. Regression with Dummy Variables (ANOVA)**

Now we can enter the four dummy variables into our model predicting commute times.



Here is the main output.



Notice that the y-intercept is 21.69, which is the mean of our reference group, in this case, the West region. All the slope coefficients represent the mean difference in commute times between that region and the West. The p-values represent the significance level in these mean differences. For example, the South has a mean commute time that is 1.012 minutes greater than the West. This difference is significantly different *t*(512) = 2.560, *p* < .05.

Notice that other region has the greatest standard error. Why is this?

You may want to recall how to conduct one-way ANOVA using SPSS.

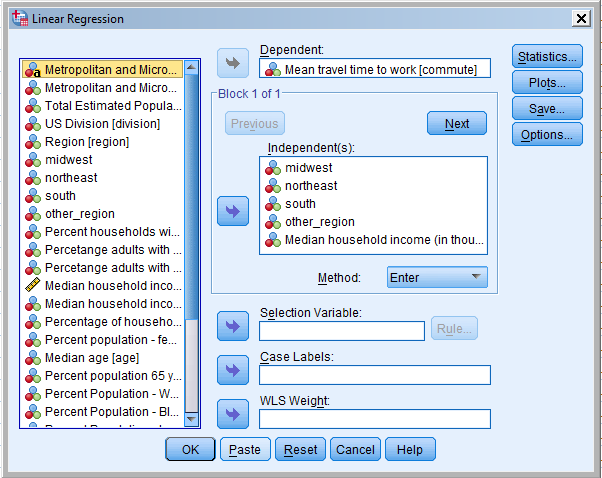
Analyze > Compare Means > One-Way ANOVA

1. **Regression with Qualitative and Quantitative Predictors (ANCOVA)**

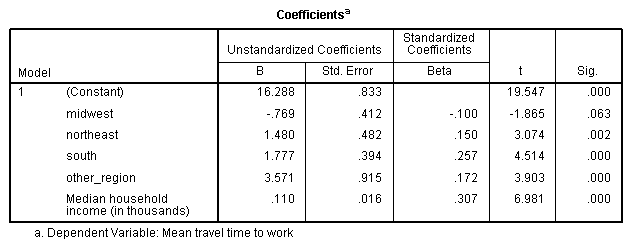
We have shown in previous analyses that income is also a significant predictor in commute times. Perhaps the differences in commute times are also due to difference in income across regions? We can account for any differences in income across region in our regression model. Let’s also note the mean median incomes across regions.

|  |  |  |
| --- | --- | --- |
| Region | Mean commute time (in mins.) | Mean median income (in thousands) |
| West | 21.69 | 49.32 |
| Midwest | 20.54 | 45.80 |
| Northeast | 23.32 | 50.71 |
| South | 22.70 | 42.34 |
| Other | 23.54 | 33.59 |

We can then enter median income in our regression model.



Here is the main output again for our regression model.



The coefficient for each of the dummy variables now compares differences in means adjusted for income. Said another way, the slope coefficients for each region is differences in region, controlling for differences in income.

The slope for income now represents the average slope across all five regions.

Notice that Midwest commute times are no longer significantly different than West commute times now that median income is accounted for. However, other regions remain significantly different than the West, even after controlling for income.

1. **Analysis of Covariance (ANCOVA)**

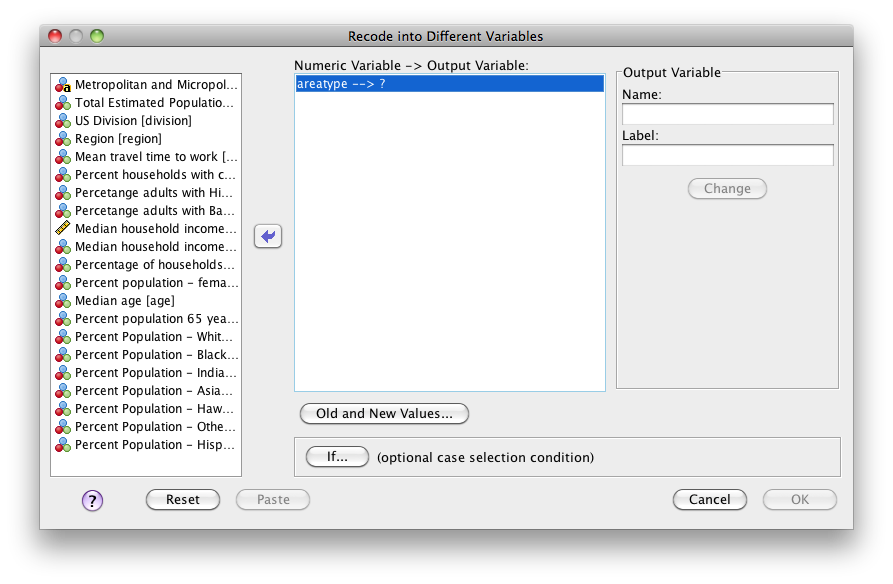
The analysis of covariance (ANCOVA) is a **general linear model** that combines the characteristics of analysis of variance (ANOVA) and regression. With ANOVA we were able to compare means of several groups. ANCOVA allows us to compare means of groups while statistically controlling for other continuous variables called covariates; that is, it allows us to **adjust the means** (of the outcome variable) to what they would be if all groups were equal on the covariates.

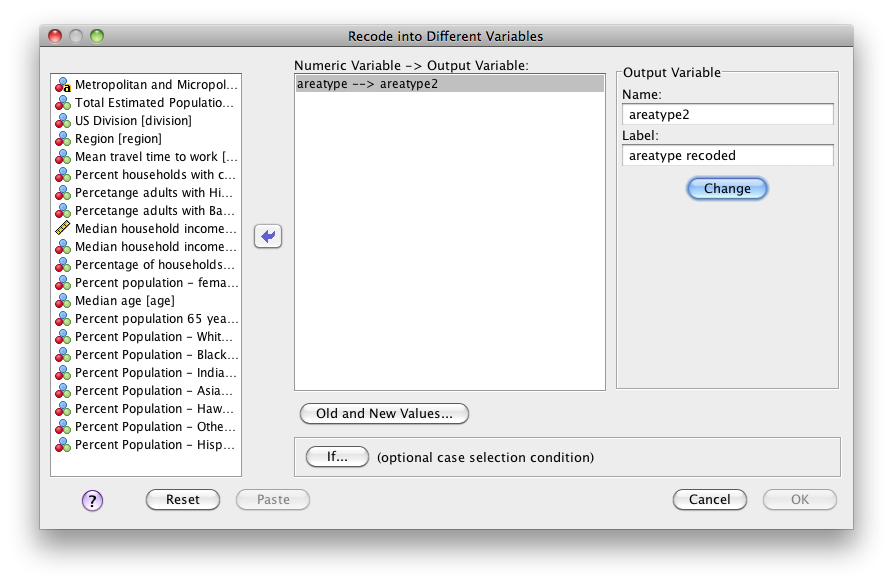
In previous analysis we have been showing variables that are significant predictors of commute times. We observed that the mean commute time was different across regions. In the following analysis we will explore if the commute time is different by **area type** (Micropolitan = 1, Metropolitan = 2) controlling for **income2**.

**Recode into Different Variables using Menu**

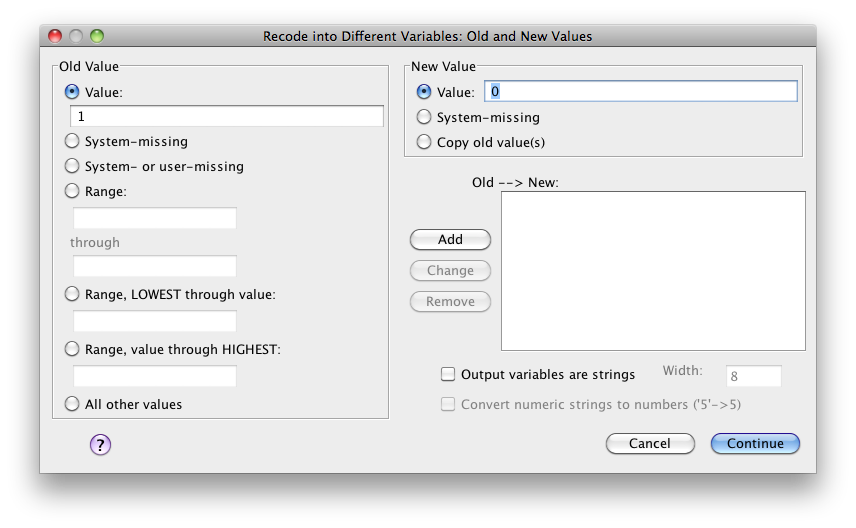
First, we can recode the values of Micropolitan and Metropolitan so that the results obtained are directly interpretable.

**Transform -> Recode into Different Variables**

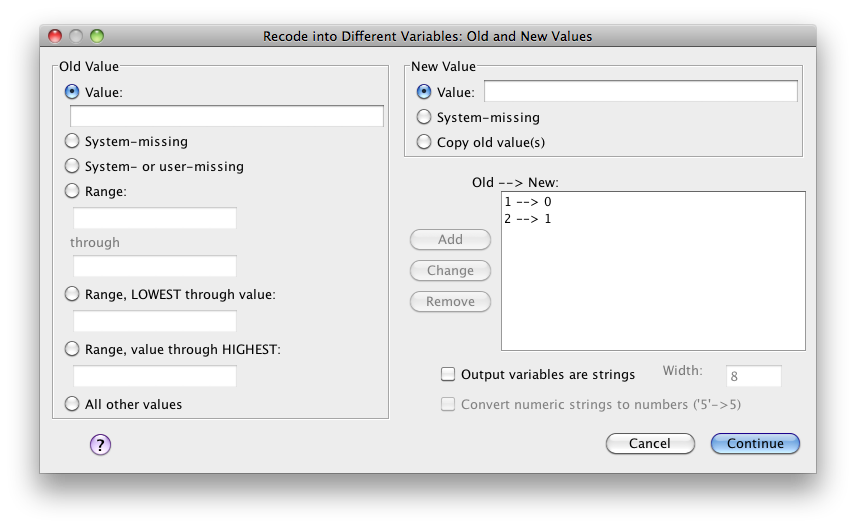


Select the variable **area type** and place it in the box *Numeric Variable - > Output Variable*. In the box Output Variable select the name (**areatype2**) and label for the new variable (areatype recoded). Then click the button Change. 

Go to **Old and New Values** and indicate which values in areatype do you want to change. Indicate the *Old Value* of 1 for Micropolitan and the *New Value* of 0 (zero).



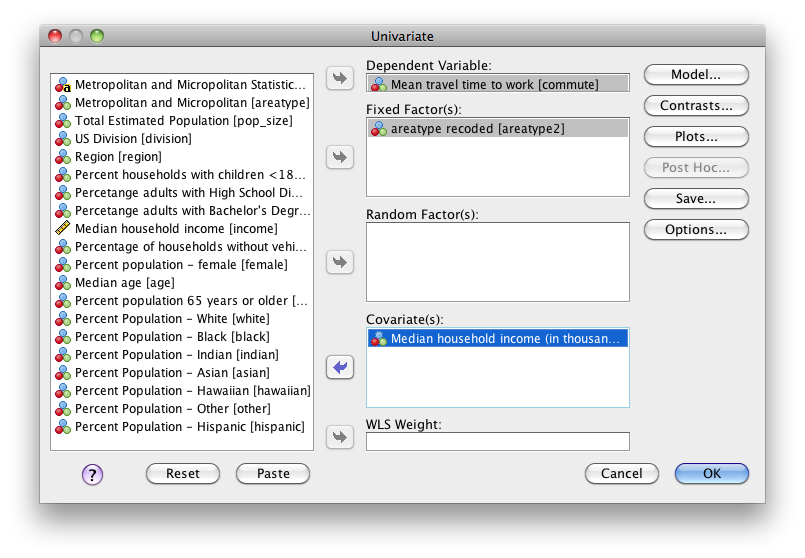
Then click on the button **Add** and repeat the process to indicate the *Old Value* of 2 for Metropolitan and the *New Value* of 1. Click Continue and then OK.



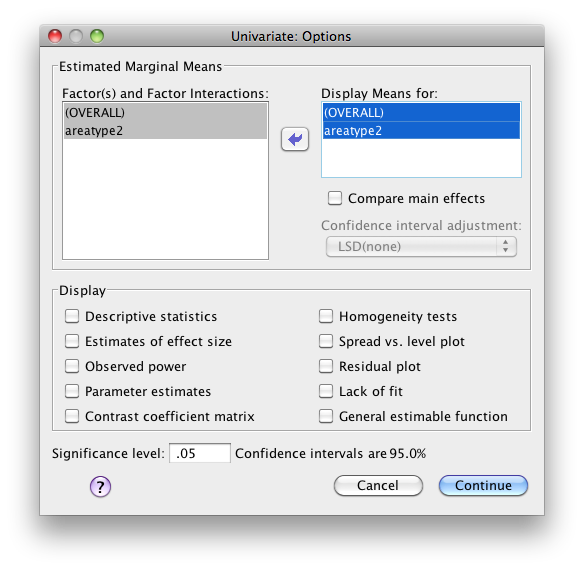
**ADJUSTED MEANS (or ESTIMATED MARGINAL MEANS)**

In order to obtain the adjusted means for each group go to:

**Analyze -> General Linear Models -> Univariate**



Select the variable commute as your **dependent variable**, the variable areatype2 as the **fixed factor** and the variable **income2** as your covariate. Then go to Options and place (OVERALL) and areatype2 in the box *Display Means for*.



|  |  |  |  |
| --- | --- | --- | --- |
| **1. Grand Mean** | | | |
| Dependent Variable: Mean travel time to work | | | |
| Mean | Std. Error | 95% Confidence Interval | |
| Lower Bound | Upper Bound |
| 22.211a | .162 | 21.893 | 22.529 |
| a. Covariates appearing in the model are evaluated at the following values: Median household income (in thousands) = 45.477. | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **2. areatype recoded** | | | | |
| Dependent Variable: Mean travel time to work | | | | |
| areatype recoded | Mean | Std. Error | 95% Confidence Interval | |
| Lower Bound | Upper Bound |
| .00 | 22.557a | .279 | 22.010 | 23.105 |
| 1.00 | 21.865a | .171 | 21.530 | 22.200 |
| a. Covariates appearing in the model are evaluated at the following values: Median household income (in thousands) = 45.477. | | | | |

If we compare the adjusted means to the means we can see that they are slightly different.

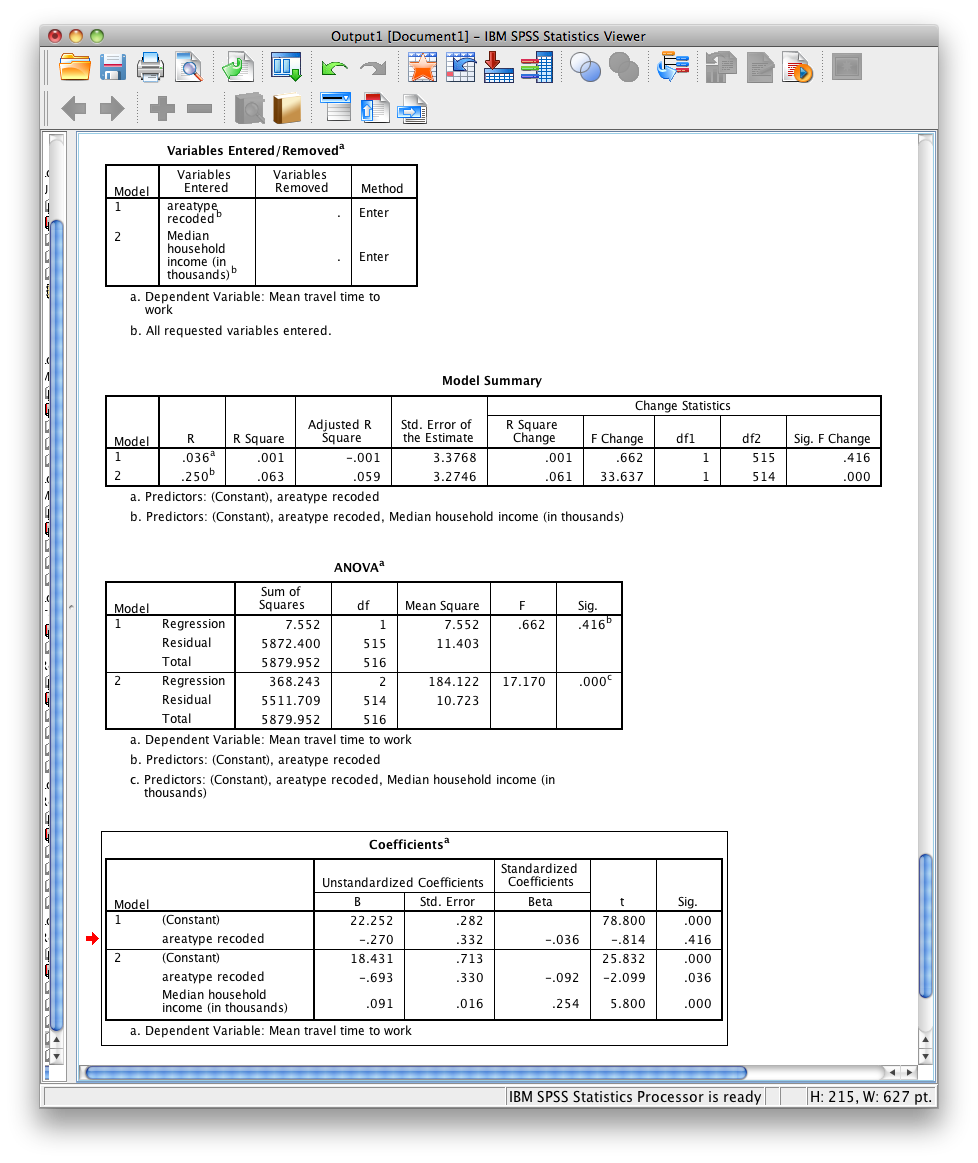
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Report** | | | | |
| Mean travel time to work | | | | |
| areatype recoded | Mean | N | Std. Deviation | Std. Error of Mean |
| .00 | 22.252 | 143 | 3.6492 | .3052 |
| 1.00 | 21.982 | 374 | 3.2671 | .1689 |
| Total | 22.056 | 517 | 3.3757 | .1485 |

**ANCOVA (run multiple regression models at the same time)**

To run the regression model that includes the categorical variable and the covariate go to: **Analysis -> Regression -> Linear**

Use commute as your dependent variable. In block 1 use areatype2 as your independent variable and in block 2 use income2. Ask for the R squared change statistics.

The test of the change in R squared for the categorical predictor is the test of difference between the adjusted means of the groups. The coefficient for the dummy variable areatype2 reflects the differences in adjusted means. The coefficient for the covariate income2 is the common slope (pooled regression slope).

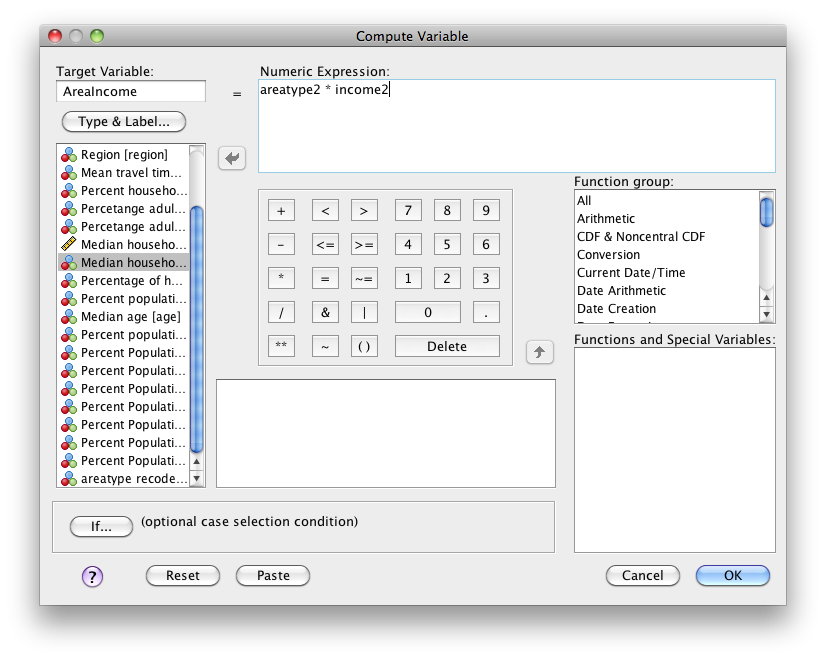


1. **Testing Homogeneity of Regression Slopes Assumption**

One of the assumptions of ANCOVA is that the slopes for each group are the same. To see if there is a significant interaction between the covariates and the dummy variables, the interaction terms need to be included in the model.

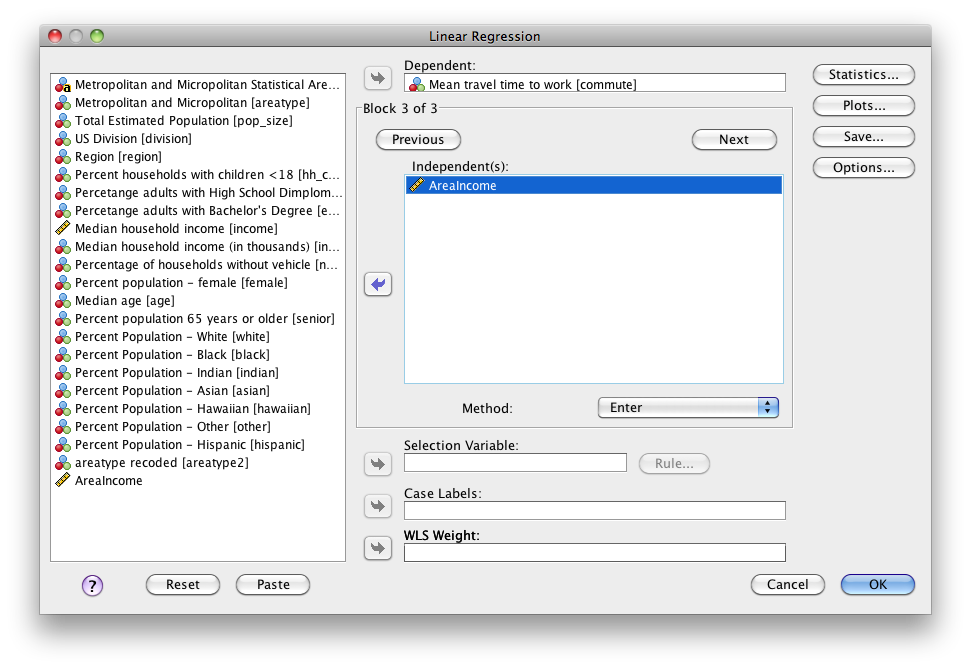
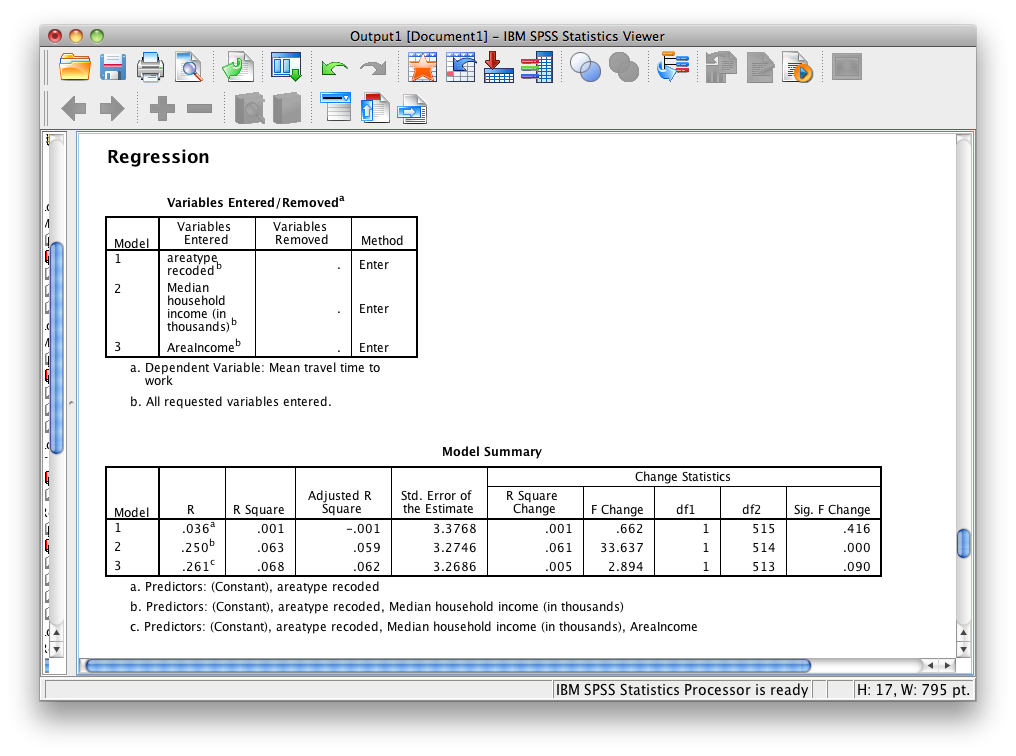
To create the interaction variable go to:

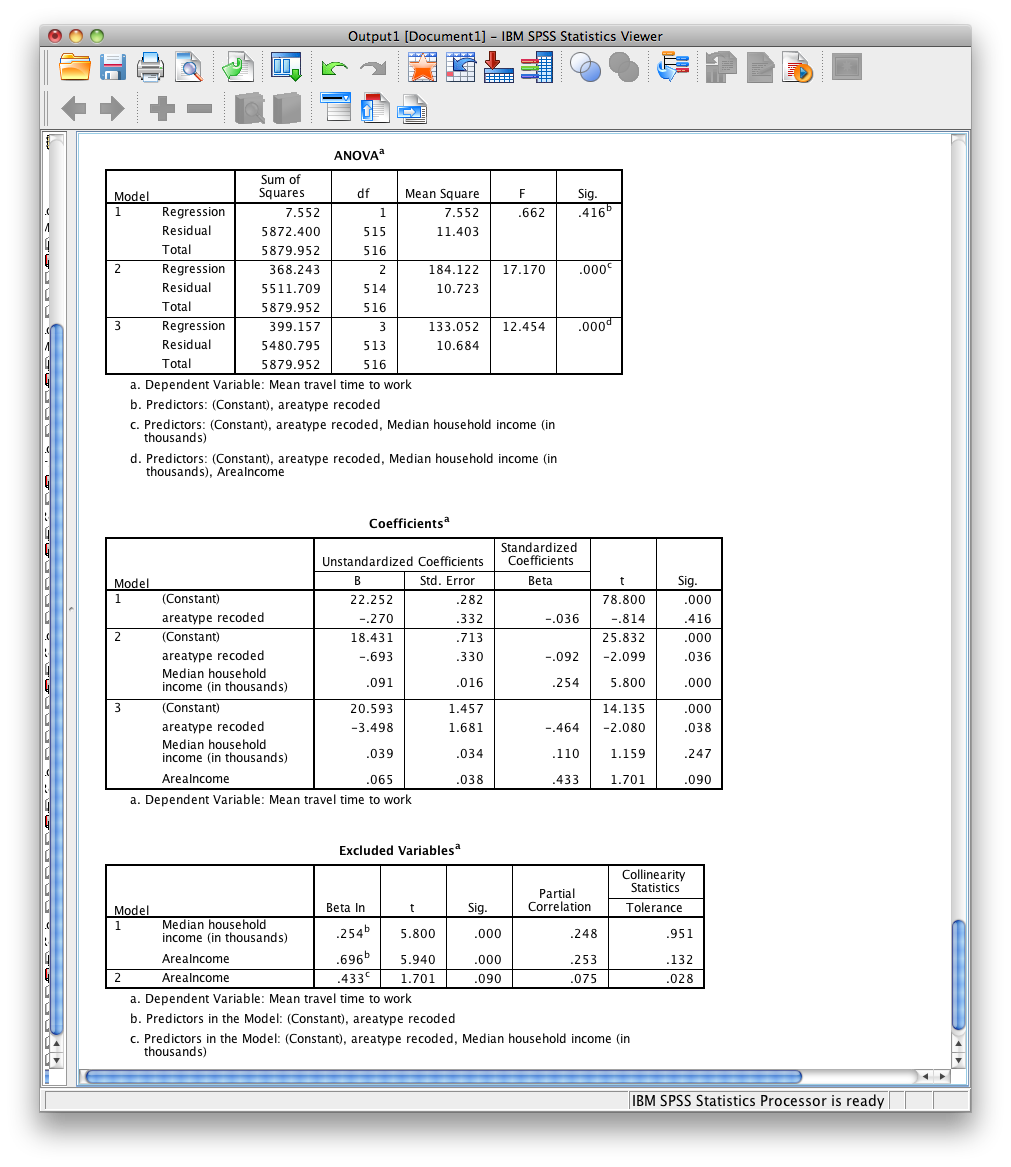
**Transform -> Compute Variable**



Name the variable **AreaIncome**. In the numeric expression multiply **areatype2** by **income2**.

Include the interaction variable in block 3 and rerun the analysis. See the following output and what is your conclusion about the assumption?





**Interactions**

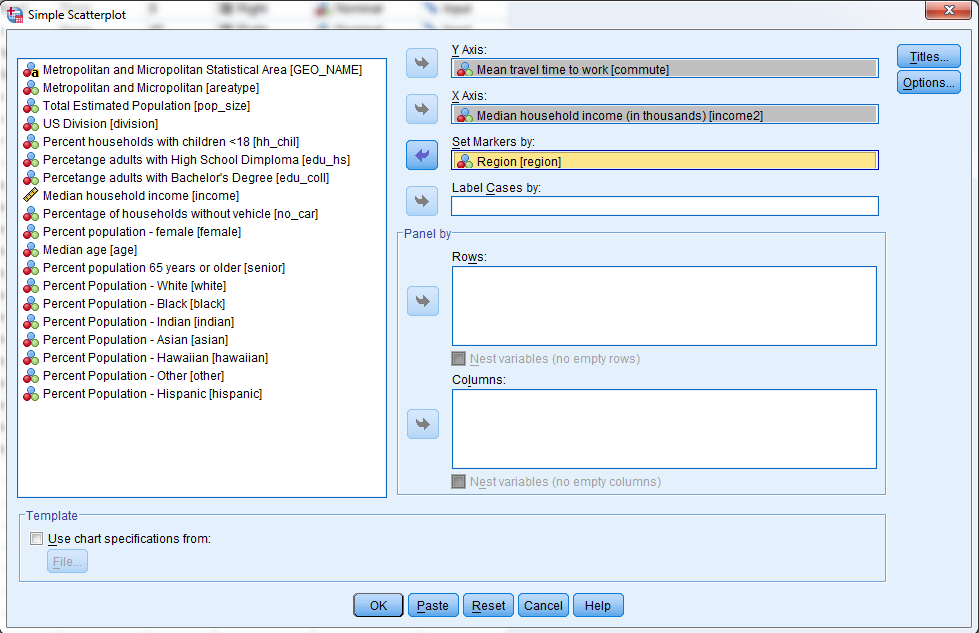
We looked at whether region of the US and median income predicted commute times. We also looked at whether there were significant differences in commute times across regions after controlling for differences in median income in each region. This ANCOVA assumed **homogeneous slopes**, that is, that the relationship between income and commute times was the same for each region. We will now look into how to test for this assumption by looking at interactions.

**Scatterplots**

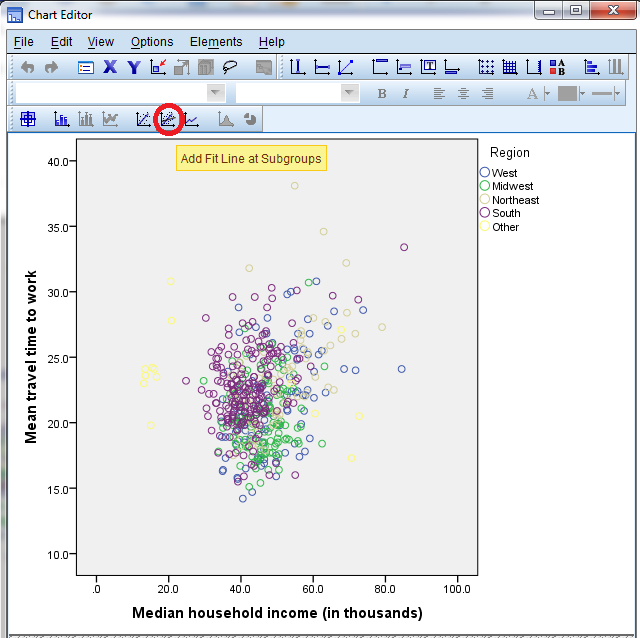
We can visually see whether slopes are homogeneous across groups.

**Graphs -> Legacy Dialogs -> Scatter/Dot**

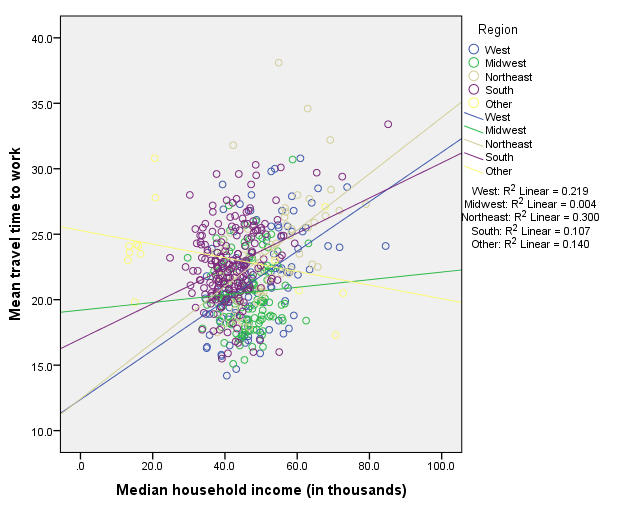
Choose “simple scatter”. Then assign median income to your x axis and commute times to your y axis. Finally, assign region to the window labeled “Set Markers by”.



Double click on the scatterplot that is produced and you can edit the plot. Choose to “Add Fit Line at Subgroups” to put regression lines at each group.



The scatterplot shows that the regression slopes for each region are not parallel, or even parallel enough. Some are negative even. When the regression is adjusting for differences in median income, it adjusts using a common slope across regions. However, is slopes are different across groups, then some adjustments are simply wrong. Consider the slope for “other region” below. There is a negative relationship between income and commute times. However, there will be a positive adjustment based upon the pooled slope between income and commute times. This will be the wrong adjustment.



**Testing Interactions**

You can formally test for homogeneous slopes by testing for interactions between your variables. Testing for an interaction tests whether the relationship between income and commute times is the same for each region.

Recall last week we made the following dummy codes for our qualitative predictors.

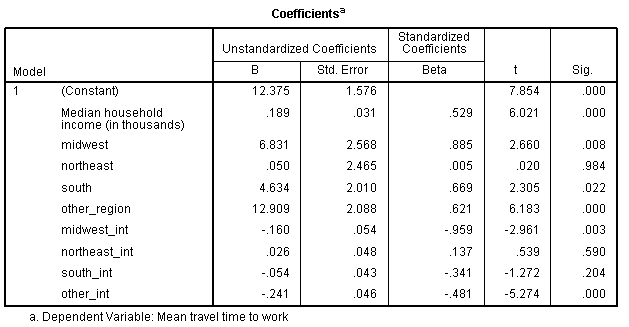
|  |
| --- |
| COMPUTE midwest = 0.  COMPUTE northeast = 0.  COMPUTE south = 0.  COMPUTE other\_region = 0.  EXE.  IF (region = 2) midwest = 1.  IF (region = 3) northeast = 1.  IF (region = 4) south = 1.  IF (region = 5) other\_region = 1.  EXE. |

We can then build on this code by adding interaction terms. Simple multiply each dummy variable with the other quantitative variable.

|  |
| --- |
| COMPUTE midwest\_int = midwest \* income2.  COMPUTE northeast\_int = northeast \* income2.  COMPUTE south\_int = south \* income2.  COMPUTE other\_int = other\_region \* income2.  EXE. |

This will create four additional interaction terms that can be entered into the regression model.

You can then enter the interaction terms into the regression model. You can either enter all the variables at once. Or, you can enter the variables into two blocks – the first block with all the variables and the second block with all the interaction terms – and then assess the *R*2 change to see whether the interaction terms add any significance to the model.



Here is our output. The slope for income (.189) and the intercept (12.375) represents the slope for income2 and the y-intercept in the reference group (in this case, the West). Each qualitative term (midwest, northeast, south, other\_region) represents the difference in y-intercept for each region compared with the West. Finally, each interaction term represents the difference in slope between income2 and commute across each region compared with the West. As we can see, the slopes for the Midwest and Other\_region is different than the West. Our assumption of homogeneous slopes has been violated, and the interaction terms significantly help predict commute times.

Exercise 1: Interpret each regression coefficient regardless of is significance.

Constant:

Median household income:

Midwest:

Northeast:

South:

Other-region:

Midwest\_int:

Northeast\_int:

South\_int:

Other-int:

Exercise 2: Write down a prediction equation for each region in terms of income predictor (You should end up with five prediction equations in this case)