MA684 HW from Class 6

Linear Regression with Categorical Predictors

1. From the hypothetical voter survey presented in class, focusing on factors associated with the political awareness score, based on a short current events quiz. The Table below presents results from three regression analyses performed on the same set of data. All three models predict the political awareness score from a voter’s age (in years), sex (1 for females, 2 for males), annual income (in $1,000), and political affiliation (democrat, republican, independent).

In each model, political affiliation is represented by two dummy variables, but the details around the dummy variables differ across the models.

In Model A, ‘repub’ is coded 1 for republicans and 0 for others, and indep is coded 1 for independents and 0 for others.

In Model B, ‘demo’ is coded 1 for democrats and 0 for others, and ‘repub’ is coded 1 for republicans and 0 for others.

In Model C, ‘repubPA’ is coded 1 for republicans, -1 for democrats, and 0 for others, and ‘indepPA’ is coded 1 for independents, -1 for democrats, and 0 for others.

All three models give an identical ANOVA table for the regression, with R2=0.220.

Results of multiple regression analyses predicting political awareness from age, sex, income, and political affiliation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Model A | | Model B | | Model C | |
|  | Slope | p-value | Slope | p-value | Slope | p-value |
| Intercept  Age  Sex  Income  Pol. Party  Demo  Repub  Indep  RepubPA  IndepPA | 1.93  0.06  -0.08  -0.01  ---  -0.34  2.31 | ---  <0.001  0.706  0.685  ---  0.169  <0.001 | 4.24  0.06  -0.08  -0.01  -2.31  -2.65  --- | ---  <0.001  0.706  0.685  <0.001  <0.001  --- | 2.58  0.06  -0.08  -0.01  -1.00  1.65 | ---  <0.001  0.706  0.685  <0.001  <0.001 |

1A. The slopes and p-values for age, sex, and income are the same for all three models, but the slopes associated with political party change across these three models. Give an interpretation for the slope of the ‘Repub’ variable (or, for model C, the ‘RepubPA’ variable) in each model, explaining why these slopes differ across these three models (and why Repub is significant in Model A but not in Model B).

1B. For each of the three models, calculate the predicted political awareness score for a 30 year old female democrat with a salary of 50 thousand dollars per year.

2. From the home sales data set that was presented in HW 4. Regression analysis was used to examine factors relating to the selling price of homes in a particular town. A real estate agent collected data on the last 25 houses that were sold in the town, including information on the selling price (in thousands of dollars), the size of the home (in 100 square feet), the number of bedrooms in the home, the age of the home (in years), and the neighborhood. We are interested in whether some neighborhoods have higher selling prices than others, controlling for the size and age of the home. So we ran two regression analysis:

Model 1: Selling Price = Age Size

The Anova table for the regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | DF | Sum of Squares | Mean Squares | F-Statistic |
| Model  Residual | 2  22 | 15,017  4,631 | 7,508.5  210.5 | 35.67 |
| Total | 24 | 19,648 |  |  |

Model R2 = 0.764

For the second model, we included 5 indicator variables representing the 6 neighborhoods in the town:

Model 2: Selling Price = Age Size Neighborhood (represented through 5 indicator variables)

The Anova table for the regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | DF | Sum of Squares | Mean Squares | F-Statistic |
| Model  Residual | 7  17 | 16,210  3,438 | 2,315.7  202.2 | 11.45 |
| Total | 24 | 19,648 |  |  |

Model R2 = 0.825

Do housing prices significantly differ across neighborhoods, after controlling for house size and the age of the house? Find the partial R2 for neighborhood, controlling for house size and age. Perform a multiple partial F-test (reporting the F-statistic, the degrees freedom, and the p-value) comparing the model with house age, size and neighborhood to the model with just house age and size.

3. (Hypothetical data based on an article from the American Journal of Public Health) A study examined the association between alcohol consumption and depression. A sample of 300 adults representing the general U.S. population was surveyed. Level of depression was measured through 7 questions that asked about the frequency of depressive symptoms, and was summarized as a ‘depressive symptoms score’ that can range from 0 to 28 with higher values indicating higher levels of depressive symptoms. Data are saved in the accompanying .csv file.

Variables in the data set are:

1) study id number, ranging from 1 to 300;

2) sex, coded 1 for males and 0 for females;

3) race-ethnicity, coded 1 for non-Hispanic Whites, 2 for non-Hispanic Blacks, 3 for Hispanics, 4 for Asians;

4) age, in years, ranging from 21 to 65;

5) years of education, ranging from 8 to 16 in this sample (a high school graduate has 12 years of education, and 16 years indicates a college graduate);

6) marital status, coded 1 for never married, 2 for currently married/living as married, and 3 for separated/divorced/widowed;

7) alcohol consumption, measured as the number of alcoholic drinks per week over the past month, ranging from 0 to 16 in the sample;

8) depressive symptom score, ranging from 1 to 16 in this sample;

9) categorized alcohol consumption, coded as 0 for abstainers, 1 for lighter drinkers (1 to 7 drinks per week), and 2 for heavier drinkers (8 or more drinks per week).

Our focus will be on the association between alcohol consumption and depressive symptoms (note that this is a cross-sectional survey, and does not provide information about the causation/direction of any association between alcohol use and depression – does alcohol use lead to depression, or do depressed people tend to drink more?).

3A. Run a multiple regression predicting depressive symptom score from age, sex, years of education, and alcohol consumption as the number of drinks per week. Present the results of this analysis in a table reporting the slopes, standard errors, and p-values.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Slope | SE | p-value |
| Intercept  Age  Sex (1-males 0-females)  Years Education  Drinks per week |  |  |  |

Model R2 = \_\_\_

How well does this regression predict depressive symptoms?

Based on this analysis, describe the association between alcohol consumption and depressive symptoms by interpreting the slope for alcohol consumption (drinks per week) in this model.

To look more closely at the association between alcohol consumption and depression, we want to categorize people as either 1) non-drinkers (0 drinks per week), 2) lighter drinkers (between 1 and 7 drinks per week, averaging 1 or fewer drinks per day), or 3) heavier drinkers (8 or more drinks per week, averaging more than 1 drink per day). We’ll do this analysis a couple of ways:

3B. Create two 0/1 dummy variables to indicate lighter drinkers and heavier drinkers (with abstainers serving as the reference group).

Using these dummy variables, run a multiple regression predicting depressive symptoms from age, sex, years of education, and the two dummy variables for alcohol consumption.

Present the results of this analysis in a table giving slopes, standard errors, and p-values (see table from 2A). Report the R2 for this regression model as well.

Discuss the association between alcohol consumption and depression, based on the slopes for the dummy variables representing alcohol consumption in this regression.

3C. To give an overall statement about the association between alcohol consumption and depression, based on this dummy variable model, calculated the increase in R2 due to the two dummy variables for alcohol consumption in this model. To do this, run two separate regressions (a Full model including age, sex, years of education, and the dummy variables versus a Reduced model with age, sex, years of education, but not the dummy variables) and find the Anova tables for these regressions:

Conduct a multiple-partial F-test to give a p-value for the association between alcohol consumption and depressive symptoms in this model (i.e., testing the two dummy variables controlling for the other variables in the model). Is the association between alcohol consumption and depressive symptoms significant in this model?

3D. As a second approach to this same dummy variable analysis, use the factor( ) command and the relevel( ) command within the lm( ) command to perform a regression using the categorized alcohol consumption variable (variable 9 in the data set, coded 0, 1, 2 for abstainers, lighter drinkers, and heavier drinkers). We want to use abstainers as the reference group.

Present the results of this analysis in a table giving slopes, standard errors, and p-values (see table from 2A). Report the R2 for this regression model as well.

How do the results of this analysis compare to the results from 3B?

3E. Using the Anova( ) command from the car package (Compliment to Applied Regression), find the Type II partial F-test for the categorized alcohol consumption variable in this model. How does this F-test compare to your result in 3C? (They should agree)

3F. (Not to be graded) Which analysis do you think is more appropriate – treating alcohol consumption as a measurement variable (drinks per day in 3A) or as a categorical variable (abstainers, light, heavy drinkers in 3D)? Explain.