MA684 HW from Class 6

Linear Regression with Categorical Predictors

Partial Correlation, Interaction Models

1. The attached ‘homeprices’ data set has information on the 25 homes that recently sold in a particular community (we looked at this data set in Class 4). We are interested in what relates to the price of a home, and we will focus on 3 variables – size of the house, number of bedrooms, and age of the house.

1a. (Work should be done using R) Find the correlation between selling price and each of these three variables (that is, find 3 bivariate correlation coefficients). Find these correlations, report significance, and offer an interpretation. Which of the 3 variables – size, number of bedrooms, age – is most strongly correlated with selling price?

1b. (Work should be done using R) We expect that bigger homes will sell at higher prices, and so house size and price are expected to be strongly correlated. So we would like to look at the correlation between number of bedrooms and selling price, and the correlation between age and selling price, controlling for the size of the house. Find these 2 partial correlations, report significance, and give an interpretation. I think these partial correlations are very different than the correlations reported in 1a – why?

1c. (Hand calculations) The purpose of this question is just to check the calculation of a partial correlation from bivariate correlations. Given the 3 correlations you found in 1a, calculate the partial correlation between selling price and age of the house (see the formula presented in class, and please carry 3 decimal places in your calculations). Confirm that the result of your hand calculations match the partial correlation from R in 1b.

1d. (Hand calculations) Below are the ANOVA tables from 2 regression models. The first (reduced model) predicts selling price from the size of the home. The second (full) model predicts selling price from the size of the home and the age of the home. (‘summary.aov’ gives Type I sequential tables, so in the Full model the SS for size matches the SS for size in the reduced model, and the SS for age is the increase in model SS when age is included with size):

Reduced Model: Price from size

> regoutR <- lm(price ~ size)

> summary(regoutR)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 46.0110 16.2186 2.837 0.00934 \*\*

size 3.2177 0.5845 5.505 1.34e-05 \*\*\*

---

Residual standard error: 19.2 on 23 degrees of freedom

Multiple R-squared: 0.5685, Adjusted R-squared: 0.5498

F-statistic: 30.31 on 1 and 23 DF, p-value: 1.344e-05

> summary.aov(regoutR)

Df Sum Sq Mean Sq F value Pr(>F)

size 1 11171 11171 30.31 1.34e-05 \*\*\*

Residuals 23 8478 369

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Full Model: Price from size and age

> regoutF <- lm(price ~ size + age)

> summary(regoutF)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 62.7164 12.8645 4.875 7.13e-05 \*\*\*

size 3.3218 0.4424 7.509 1.66e-07 \*\*\*

age -1.6879 0.3949 -4.275 0.000308 \*\*\*

---

Residual standard error: 14.51 on 22 degrees of freedom

Multiple R-squared: 0.7643, Adjusted R-squared: 0.7429

F-statistic: 35.67 on 2 and 22 DF, p-value: 1.247e-07

> summary.aov(regoutF)

Df Sum Sq Mean Sq F value Pr(>F)

size 1 11171 11171 53.07 2.69e-07 \*\*\*

age 1 3846 3846 18.27 0.000308 \*\*\*

Residuals 22 4631 211

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Given the results for these Anova tables, find the partial correlation coefficient between price and age, controlling for the size of the house. Confirm that the partial correlation you calculate matches the partial correlation from 1b. (Using Sums of Squares will give you absolute value of the partial correlation coefficient, but does not give the direction (sign) of the partial correlation. The slope for age in the full model gives the sign for the partial correlation).

1. (No compter work needed – from last week’s homework) Based on a study of maternal behavior during pregnancy, headed by Dr. Debbie Frank. We’re interested in the association between maternal alcohol use during pregnancy and pregnancy outcome. We’ll focus on birth weight, with higher birth weight generally reflecting a healthier pregnancy. A sample of 247 pregnant women were enrolled and followed during pregnancy. Alcohol use during pregnancy was coded as no use, lighter use, and heavier use. Control variables known to be associated with birth weight are gestational age, infant sex, and maternal pre-pregnancy weight.

Models A and B represent two different versions of a multiple regression predicting birth weight. In both models, two variables were created to represent the three category alcohol variable.

For Model A, the variable ‘lightalc’ was coded 1 for those with lighter alcohol use, 0 for others, and ‘heavyalc’ was coded 1 for those with heavy alcohol use, 0 for others.

In Model B, ‘lightalcP‘ was coded 1 for those with lighter alcohol use, 0 for those with heavier alcohol use, and -1 for those with no alcohol use, and ‘heavyalcP’ was coded 1 for those with heavy alcohol use, 0 for those with lighter alcohol use, and -1 for those with no alcohol use. The ‘child sex’ variable is coded 1 for males, 0 for females.

In these analyses, birth weight is measured in grams, with a mean (sd) of 3167 (484) grams in this sample.

The table below presents slopes and p-values from Models A and B:

Results of multiple regression analyses predicting Birthweight from gestational age, child sex, maternal pre-pregnancy weight, and alcohol consumption

|  |  |  |
| --- | --- | --- |
|  |  | |
|  | Slope | p-value |
| Model A  Intercept  Gestational age  Child Sex (Male)  Pre-preg Wt  LightAlc  HeavyAlc | -1559.1  107.6  223.5  2.5  -203.5  -301.2 | ---  <0.001  <0.001  <0.001  0.010  <0.001 |
| Model B  Intercept  Gestational age  Child Sex (Male)  Pre-preg Wt  LightAlcP  HeavyAlcP | -1727.3  107.6  223.5  2.5  -35.3  -132.9 | ---  <0.001  <0.001  <0.001  0.523  0.019 |

1A. Describe the association between alcohol use and Birthweight, focusing on the slopes and p-values from Model A.

1B. Describe the association between alcohol use and Birthweight, focusing on the slopes na p-values from Model B.

1C. Using both Model A and Model B, calculate the predicted birth weight for a female child, born at 40 weeks gestation, to a mom with pre-pregnancy weight of 120 lbs, who did not drink alcohol during pregnancy. How do the predicted birth weights compare between the two versions of the regression model?

The table below is the Anova table for both Models A and B (the Anova tables for the two models are identical). Both Model A and B have R2=0.26, p<0.001.

Anova for the Regression (for both Model A and Model B)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | df | Sum of Squares | Mean Square | F | p-value |
| Model  Error | 5  241 | 15042453  42558678 | 3008491  176592 | 17.04 | <0.001 |
| Total | 246 | 57601131 |  |  |  |

The table below is the Anova table for a model without the alcohol variables:

Anova for the Reduced Model (with gestational age, sex, and pre-pregnancy weight as predictors) R2=0.21

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | df | Sum of Squares | Mean Square | F | p-value |
| Model  Error | 3  243 | 11896782  45704349 | 3965594  188084 | 21.08 | <0.001 |
| Total | 246 | 57601131 |  |  |  |

1D. Since both Model A and Model B have the same Anova table for the regression, both models will give the same partial R-square for the two alcohol variables (controlling for gestational age, child sex, and pre-pregnancy weight) and the same multiple-partial F-test for the contribution of the two alcohol variables (controlling for gestational age, child sex, and pre-pregnancy weight). Report the partial R-square and the multiple partial F-test.

2. The purpose of this question is to work through a regression model with interaction terms. A (hypothetical) intervention was designed to increase awareness around environmental issues (recycling, grow local, energy conservation) in high school students. 100 students were enrolled, given a baseline assessment that measured their awareness of and attitudes toward environmental issues, and then randomized to either intervention or control. 3 months later, they were given a follow-up assessment of their awareness of and attitudes toward environmental issues. We are interested in whether the intervention increased awareness and improved attitudes.

(Hypothetical) data are saved in the attached .csv file. Variables in the file are: 1) a subject ID number; 2) an indicator for sex, coded 1 for males and 0 for females; 3) Intervention, coded 0 for those randomized to control and 1 for those randomized to intervention; 4) the baseline environmental awareness and attitudes score, where higher scores reflect better awareness and more environmental-friendly attitudes; and 5) the follow-up environmental awareness and attitudes score.

2a. As a first analysis, run a regression predicting the follow-up environmental awareness score from the baseline environmental awareness score, sex, and intervention. Did the intervention significantly increase the follow-up environmental awareness score? Estimate the intervention effect – how much did intervention increase the environmental awareness score.

2b. As a second analysis, run a regression predicting the follow-up environmental awareness score from the baseline environmental awareness score, sex, the intervention, and the interaction between the baseline score and intervention. With our focus on the intervention effect:

i) What does it mean if there is an interaction between the baseline score and intervention?.

ii) Is the interaction term significant in this regression? Report the t-value and p-value.

iii) Estimate the intervention effect – how much did intervention increase the environmental awareness score? (This is a bit of a trick question – I think the answer is ‘that depends’. To give a more specific answer, how much did intervention increase the environmental awareness score, for students with a baseline environmental awareness score of 20? How much did intervention increase the environmental awareness score for students with a baseline environmental score of 70?)

2c. We want to test whether the intervention was more effective for females than for males. To do this, run a regression predicting the follow-up environmental awareness score from the baseline environmental awareness score, sex, the intervention, and the interaction between sex and intervention. (We will ignore the interaction between the baseline score and intervention in this analysis – we could build a model with several different sets of interaction terms, but let’s not go there for this homework.) Is the intervention more effective for females than for males – report significance from this analysis.

3. The purpose of this question is to examine interaction with a more-than-2-category categorical independent variable. We will look at the Framingham Heart Study data set, with BMI categorized as normal weight, overweight, and obese (there were very few underweight subjects, and these have been combined with the ‘normal weight’ group).

We want to see if the effect of age differs by body weight – does age have a greater effect on systolic blood pressure for those with greater body weight?

As a preliminary analysis, I’ll first run a main-effects model (without interaction), and also get the Type II tests (to get a single p-value for the overall effect of body weight in this regression):

> regout <- lm(sysbp ~ age + sexmale + relevel(factor(bmicat3),'1'))

> summary(regout)

Call:

lm(formula = sysbp ~ age + sexmale + relevel(factor(bmicat3),

"1"))

Residuals:

Min 1Q Median 3Q Max

-37.193 -14.553 -3.296 13.089 72.943

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 83.7739 8.1616 10.264 < 2e-16 \*\*\*

age 0.9299 0.1601 5.808 2.54e-08 \*\*\*

sexmale -5.6905 2.8885 -1.970 0.0502 .

relevel(factor(bmicat3), "1")2 6.7379 3.0527 2.207 0.0285 \*

relevel(factor(bmicat3), "1")3 19.0984 4.2344 4.510 1.12e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.37 on 195 degrees of freedom

Multiple R-squared: 0.2753, Adjusted R-squared: 0.2605

F-statistic: 18.52 on 4 and 195 DF, p-value: 6.418e-13

> library("car", lib.loc="~/R/win-library/3.1")

Warning message:

package ‘car’ was built under R version 3.1.3

> Anova(regout)

Anova Table (Type II tests)

Response: sysbp

Sum Sq Df F value Pr(>F)

age 12650 1 33.732 2.536e-08 \*\*\*

sexmale 1455 1 3.881 0.05025 .

relevel(factor(bmicat3), "1") 7788 2 10.384 5.183e-05 \*\*\*

Residuals 73127 195

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>

Based on the Type II tests, does body weight significantly relate to systolic blood pressure, after controlling for age and sex? Interpret the slopes for the two BMI variables in this regression equation.

Does age significantly relate to systolic blood pressure? Interpret the slope for age.

To examine whether the effect of age differs for those with different BMI categories, the following interaction model was run:

> regout <- lm(sysbp ~ age + sexmale + relevel(factor(bmicat3),'1') + age\*relevel(factor(bmicat3),'1'))

> summary(regout)

Call:

lm(formula = sysbp ~ age + sexmale + relevel(factor(bmicat3),

"1") + age \* relevel(factor(bmicat3), "1"))

Residuals:

Min 1Q Median 3Q Max

-37.318 -14.572 -3.224 13.130 73.064

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 84.70262 12.96390 6.534 5.57e-10 \*\*\*

age 0.91006 0.26781 3.398 0.000824 \*\*\*

sexmale -5.64157 2.94815 -1.914 0.057152 .

relevel(factor(bmicat3), "1")2 5.43826 17.72644 0.307 0.759335

relevel(factor(bmicat3), "1")3 16.83926 24.54106 0.686 0.493431

age:relevel(factor(bmicat3), "1")2 0.02659 0.35347 0.075 0.940104

age:relevel(factor(bmicat3), "1")3 0.04558 0.48334 0.094 0.924970

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.46 on 193 degrees of freedom

Multiple R-squared: 0.2754, Adjusted R-squared: 0.2529

F-statistic: 12.22 on 6 and 193 DF, p-value: 1.216e-11

> Anova(regout)

Anova Table (Type II tests)

Response: sysbp

Sum Sq Df F value Pr(>F)

age 12650 1 33.3877 2.987e-08 \*\*\*

sexmale 1387 1 3.6618 0.05715 .

relevel(factor(bmicat3), "1") 7788 2 10.2783 5.732e-05 \*\*\*

age:relevel(factor(bmicat3), "1") 4 2 0.0052 0.99486

Residuals 73123 193

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Based on the Type II tests, does the effect of age differ by BMI category? Explain.

I think this interaction model gives different ‘slopes’ for age, for the 3 different BMI categories. Regardless of significance, give the slope for age for each of the three BMI categories.