Problem Set 4

Applied Stats/Quant Methods 1

Due: December 3, 2023

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub.
- This problem set is due before 23:59 on Sunday December 3, 2023. No late assignments will be accepted.

Question 1: Economics

In this question, use the `prestige` dataset in the `car` library. First, run the following commands:

```
install.packages(car)
library(car)
data(Prestige)
help(Prestige)
```

We would like to study whether individuals with higher levels of income have more prestigious jobs. Moreover, we would like to study whether professionals have more prestigious jobs than blue and white collar workers.

(a) Create a new variable `professional` by recoding the variable `type` so that professionals are coded as 1, and blue and white collar workers are coded as 0 (Hint: `ifelse`).

Answer.

To create a new variable in the dataset called 'professional', using the ifelse function, the following code can be used in R:

```
  Prestige$professional <- ifelse(Prestige$type == 'prof', 1, 0)

```

This code assigns a new variable the name professional in the Prestige dataset, using the subset function ($). It then defines the variable as a dummy, or binary variable, using the ifelse function to give every row a value of either one or zero in the new professional variable depending on if they have characters 'prof' in the already existing variable 'type'.

(b) Run a linear model with `prestige` as an outcome and `income`, `professional`, and the interaction of the two as predictors (Note: this is a continuous × dummy interaction.)

Answer.

To run a linear regression with those specific outcomes and predictors, the following code can be used in R:

```
model1 <- lm(prestige ~ income + professional + income * professional,
  data = Prestige)

```

The results of which can be viewed on the following table:

Table 1:

	Dependent variable:
	prestige
income	0.003^{***}
	(0.0005)
professional	37.781^{***}
	(4.248)
income:professional	-0.002^{***}
	(0.001)
Constant	21.142^{***}
	(2.804)
Observations	98
R^2	0.787
Adjusted R^2	0.780
Residual Std. Error	8.012 (df = 94)
F Statistic	115.878^{***} (df = 3; 94)
Note:	*p<0.1; **p<0.05; ***p<0.01

(c) Write the prediction equation based on the result.

Answer.

Given that basic the formula for the prediction equation is:

$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ...$

Where:

- μ_y = The predicted value of y (the outcome variable),
- β_0 = The intercept (The value of y, when x is equal to zero),
- β_1 = the coefficient of the first predictor variable (the change in y with a one-unit increase in of x1),
- x_1 = the value of the first predictor variable (dependent on what we are trying to predict)
- β_2 and x_2 follow the same logic as the first coefficient, and this continues for however many coefficients there are.

The prediction equation for the linear model 'model1' can be written as such:

μ_y **= 21.1422589 + 0.0031709***x_1 **+ 37.7812800***x_2 **- 0.0023257***x_1***x_2

(d) Interpret the coefficient for `income`.

Answer.

Income is a continuous predictor variable, so the coefficient for income in linear model model1 can be interpreted as such; for every one unit increase in income, prestige will increase by 0.003171 units on average.

(e) Interpret the coefficient for `professional`.

Answer.

Professional is a categorical variable, so the coefficient for professional in linear model model1 can be interpreted as such; similar to the coefficient for income, the coefficient for professional also represents how much prestige will increase with a one unit increase in professional, on average. Only, because professional is a binary categorical variable this means something different, the reference group (those that are not professional) is coded as 0 and the comparison group (those who are professional) is coded as 1. So, the coefficient represents the average difference between the reference and comparison groups.

(f) What is the effect of a $1,000 increase in income on prestige score for professional occupations? In other words, we are interested in the marginal effect of income when the variable `professional` takes the value of 1. Calculate the change in $\hat{y}$ associated with a $1,000 increase in income based on your answer for (c).

Answer.

To find the marginal effect of a $1000 increase in income on prestige when the value of professional is equal to 1, the predicted average value of $\hat{y}$ when income is equal to $0 and professional is equal to 1 must first be calculated, based on the assumptions in (c):

21.1422589 + 0.0031709*0 + 37.7812800*1 - 0.0023257*0*1

which returns an answer = 58.92354

The predicted average value of $\hat{y}$ when income is equal to $1000 and professional is equal to 1 must then be calculated, also based on the assumptions in (c):

21.1422589 + 0.0031709*1000 + 37.7812800*1 - 0.0023257*1000*1

which returns an answer = 59.76874

Then the difference between these two values must be calculated, which will give us the difference in the value of prestige based on a $1000 marginal increase in income when professional is set to a value of 1. This is the difference in prestige based on an increase of $0 to $1000, but, because the relationship is linear, all else being equal this increase in prestige will be the same for every increase of $1000 in income, and this can be checked using the same equation checking the difference between higher intervals of $1000.

59.76874 - 58.92354

Which returns an answer = 0.8452

(g) What is the effect of changing one's occupations from non-professional to professional when her income is \$6,000? We are interested in the marginal effect of professional jobs when the variable `income` takes the value of $6,000$. Calculate the change in $\hat{y}$ based on your answer for (c).

Answer.

To find the marginal effect of professional jobs when the income variable is equal to \$6000, first the predicted average value of $\hat{y}$ when the value of professional is set 0 and income variable is set to 6000 must be calculated, based on the assumptions made in (c).

21.1422589 + 0.0031709*6000 + 37.7812800*0 - 0.0023257*6000*0

Which returns the answer = 40.16766

Then the predicted average value of $\hat{y}$ when the value of professional is set 1 and income variable is set to 6000 must be calculated, based on the assumptions made in (c).

21.1422589 + 0.0031709*6000 + 37.7812800*1 - 0.0023257*6000*1

Which returns the answer = 63.99474

Then the difference between these two values must be calculated. Which will be the marginal effect of professional jobs on prestige when the income variable takes the value 6000.

63.99474 - 40.16766

Which returns the answer = 23.82708

Question 2: Political Science

Researchers are interested in learning the effect of all of those yard signs on voting preferences.[1] Working with a campaign in Fairfax County, Virginia, 131 precincts were randomly divided into a treatment and control group. In 30 precincts, signs were posted around the precinct that read, "For Sale: Terry McAuliffe. Don't Sellout Virgina on November 5."

Below is the result of a regression with two variables and a constant. The dependent variable is the proportion of the vote that went to McAuliff's opponent Ken Cuccinelli. The first variable indicates whether a precinct was randomly assigned to have the sign against McAuliffe posted. The second variable indicates a precinct that was adjacent to a precinct in the treatment group (since people in those precincts might be exposed to the signs).

Impact of lawn signs on vote share

Precinct assigned lawn signs (n=30)	0.042
	(0.016)
Precinct adjacent to lawn signs (n=76)	0.042
	(0.013)
Constant	0.302
	(0.011)

Notes: R^2=0.094, N=131

(a) Use the results from a linear regression to determine whether having these yard signs in a precinct affects vote share (e.g., conduct a hypothesis test with $\alpha = .05$).

Answer.

Null hypothesis, H_0 = the yard signs have no effect on votes. The coefficient for "Precinct assigned lawn signs" β_1 is not statistically different from zero.

$\beta_1 = 0$

Alternative hypothesis, H_a = the yard signs have a statistically significant effect on votes. The coefficient for "Precinct assigned lawn signs" β_1 is statistically different from zero.

$\beta_1 \neq 0$

[1] Donald P. Green, Jonathan S. Krasno, Alexander Coppock, Benjamin D. Farrer, Brandon Lenoir, Joshua N. Zingher. 2016. "The effects of lawn signs on vote outcomes: Results from four randomized field experiments." Electoral Studies 41: 143-150.

To perform this hypothesis test the test statistic and p-value must be calculated. The population standard deviation is unknown, so we will assume a t-distribution rather than a normal distribution for this test.

The formula for test statistic in this context is:

t = β_1 / SE(β_1)

For this hypothesis test it would be:

t = 0.042/0.016 **= 2.625**

Then the degrees of freedom need to be calculated. Which can be done with the following formula. N - K - 1, N being the number of observations and K the number of predictors.

30 - 1 - 1 = 28

Then using the degrees of freedom of 28 and α of 0.05, the critical t-value from a t-distribution table must be looked up. The hypothesis test is two-tailed because the test does not specify the direction of difference of β_1 from 0. So with this information and reference to a t-distribution table[2], a critical t-value of **2.048** is found.

Because the absolute value of the calculated test statistic for this hypothesis test, **2.625**, is greater than the critical t-value from the t-distribution table, **2.048**, the null hypothesis **H_0 can be rejected** and it can be concluded that variable "Precincts assigned lawn signs" has a statistically significant effect on vote share.

Then the p-value needs to be calculated. To calculate the p-value we can use the pnorm function in R, as seen in the following code:

```
    2*pnorm(-abs(2.625))

```

This code takes the test statistic that was previously calculated, and is multiplied by 2 because, again, this is a two tailed hypothesis test.

This code returns a p-value **= 0.008664897**

Because our calculated p-value of 0.008664897 is less than the threshold of significance, $\alpha = 0.05$ for this hypothesis test, the p-value provides further evidence to reject the H_0, the null hypothesis.

[2] https://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf

(b) Use the results to determine whether being next to precincts with these yard signs affects vote share (e.g., conduct a hypothesis test with $\alpha = .05$).

Answer.

Null hypothesis, H_0 = precincts being adjacent to precincts with yard signs has no effect on vote share. The coefficient for ”Precinct adjacent to lawn signs” β_2 is not statistically different from zero.

$\beta_2 = 0$

Alternative hypothesis, H_a = precincts being adjacent to precincts with yard signs has a statistically significant effect on votes. The coefficient for ”Precinct adjacent to lawn signs” β_2 is statistically different from zero.

$\beta_2 \neq 0$

To perform this hypothesis test the test statistic and p-value must be calculated. The population standard deviation is unknown, so we will assume a t-distribution rather than a normal distribution for this test.

The formula for test statistic in this context is:

t = β_1 / SE(β_1)

For this hypothesis test it would be:

t = 0.042/0.013 **= 3.230769**

Then the degrees of freedom need to be calculated. Which can be done with the following formula. N - K - 1, N being the number of observations and K the number of predictors.

76 - 1 - 1 = 74

Then using the degrees of freedom of 74 and α of 0.05, the critical t-value from a t-distribution table would usually be looked up. But, as most critical t-distribution tables do not go as high as 74 degrees of freedom exactly, in this case it is easier to use the following code in R to find the critical t-value:

```r
qt(p = 0.025, df = 74)

```

This code uses the qt function to calculate the critical t-value, taking into consideration that the hypothesis test is two-tailed by putting p = 0.025, when $\alpha = 0.05$. The code returns a value of -1.992543 which can be rounded to three decimals as t-values normally are and changed to positive as we are only concerned with the absolute value, which gives an answer = **1.993**.

Because the absolute value of the calculated test statistic for this hypothesis test **3.230769** is greater than the critical t-value from the t-distribution calculation made in R **1.993**, the null hypothesis **$\mathbf{H}_0$ can be rejected** and it can be concluded that variable ”Precincts adjacent to lawn signs” has a statistically significant effect on vote share.

Then the p-value needs to be calculated. To calculate the p-value we can use the pnorm function in R, as seen in the following code:

```
2*pnorm(-abs(3.230769))

```

This code takes the test statistic that was previously calculated, and is multiplied by 2 because a two sided hypothesis test is being performed.

This code returns a p-value = **0.001234577**

Because our calculated p-value of 0.001234577 is less than the threshold of significance, $\alpha = 0.05$ for this hypothesis test, the p-value provides further evidence to reject the H_0, the null hypothesis.

(c) Interpret the coefficient for the constant term substantively.

Answer.

The constant term in the case of this linear regression table represents the predicted value of the dependent variable, the baseline proportion of the vote that went to McAuliffe’s opponent (Ken Cuccinelli), when the two other coefficients, for the predictor variables, are equal to zero.

As there were 131 precincts randomly selected in the study, 30 of which had yard signs, and 71 of which were identified as being adjacent to precincts with yard signs, this leaves a remainder of 25 precincts which did not have yard signs, and were not adjacent to precincts with yard signs. The constant variable in this case may be the said to be the baseline average vote share that went to McAuliffe’s opponent (Ken Cuccinelli) in these precincts which had the lowest probability of being exposed to the yard signs, by neither having yard signs or being adjacent to precincts with yard signs.

(d) Evaluate the model fit for this regression. What does this tell us about the importance of yard signs versus other factors that are not modeled?

Answer.

The model has a very low R^2 of 0.094. An R^2 significantly lower than the maximum of 1.0 suggests that the model does not include all variables that are associated with the outcome. An R^2 lower than 0.4 is generally considered to show low correlation.

The formula for R^2 is:

$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$

So, the low R^2 in this case signifies that a large portion of the total variance is not explained by the model.

The distance of points in the regression plot to the regression line are called the standardized residuals, and in the case of this linear regression, the low R^2 indicates large standardized residuals, or a lot of spread of the data from the regression line.

All of this does not suggest that the yard signs have no effect on vote share, but that there other potential factors not considered in the model which are probably influencing vote share to a large extent. Because the R^2 is so low this suggests that most of the variance is not explained by the model, and is caused by other unaccounted for factors.

Bibliography:

- Kumar, A. (2023) ‘Linear Regression T-test: Formula, Example’, Analytics Yogi, 29 November. Available at: https://vitalflux.com/linear-regression-t-test-formula-example/ (Accessed: 30 November 2023).
- Kumar, A. (2022) ‘Linear regression hypothesis testing: Concepts, Examples’, Analytics Yogi, 18 April. Available at: https://vitalflux.com/linear-regression-hypothesis-testing-examples/ (Accessed: 2 December 2023).
- Grace-Martin, K. (2021) ‘Interpreting Regression Coefficients’, The Analysis Factor, 20 December. Available at: https://www.theanalysisfactor.com/interpreting-regression-coefficients/ (Accessed: 30 November 2023).
- Jeffery Ziegler’s lecture slides.
- Hannah Frank’s tutorial material.
- Frost, J. (2017) How To Interpret R-squared in Regression Analysis, Statistics By Jim. Available at: http://statisticsbyjim.com/regression/interpret-r-squared-regression/ (Accessed: 3 December 2023).

- Population Standard Deviation - an overview — ScienceDirect Topics (no date). Available at: https://www.sciencedirect.com/topics/mathematics/population-standard-deviation (Accessed: 3 December 2023).
- Kumar, A. (2022) ‘Linear regression hypothesis testing: Concepts, Examples’, Analytics Yogi, 18 April. Available at: https://vitalflux.com/linear-regression-hypothesis-testing-examples/ (Accessed: 3 December 2023).