Applied Stats - Problem Set 4

Yana Konshyna

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`).

After loading `Prestige` dataset into the working environment, I used summary() method to display summary statistics of each variable in the dataset.

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

summary(Prestige)
```

Output:

```
education          income          women           prestige          census          type
Min.   : 6.380   Min.   :  611   Min.   : 0.000   Min.   :14.80   Min.   :1113   bc  :44
1st Qu.: 8.445   1st Qu.: 4106   1st Qu.: 3.592   1st Qu.:35.23   1st Qu.:3120   prof:31
Median :10.540   Median : 5930   Median :13.600   Median :43.60   Median :5135   wc  :23
Mean   :10.738   Mean   : 6798   Mean   :28.979   Mean   :46.83   Mean   :5402   NA's: 4
3rd Qu.:12.648   3rd Qu.: 8187   3rd Qu.:52.203   3rd Qu.:59.27   3rd Qu.:8312
Max.   :15.970   Max.   :25879   Max.   :97.510   Max.   :87.20   Max.   :9517
```

Creating a new variable `professional` by using ifelse function to recoding the variable type. The professionals ("prof") are coded as 1, and blue and white collar workers ("bc", "wc") are coded as 0. Printing first 6 row of the table by using head() function.

```
# Converting categorical variable into factor
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)
head(Prestige)
```

Output:

```
                    education income women prestige census type   professional
gov.administrators      13.11  12351 11.16     68.8   1113 prof              1
general.managers        12.26  25879  4.02     69.1   1130 prof              1
accountants             12.77   9271 15.70     63.4   1171 prof              1
purchasing.officers     11.42   8865  9.11     56.8   1175 prof              1
chemists                14.62   8403 11.68     73.5   2111 prof              1
physicists              15.64  11030  5.13     77.6   2113 prof              1
```

(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 $\times$ dummy interaction.)

Executing the regression model in which the `prestige` variable is explained by the independent variables such as `income` and `professional`. The variable `income:professional` is the interaction of the two as predictors. Then I investigate the estimated coefficients of the model using summary().

```r
model <- lm(prestige ~ income + professional + income:professional, data
    = Prestige)
summary(model)
```

Output:

```
Residuals:    Min      1Q  Median      3Q     Max
          -14.852  -5.332  -1.272   4.658  29.932
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)          21.1422589  2.8044261   7.539 2.93e-11 ***
income                0.0031709  0.0004993   6.351 7.55e-09 ***
professionals        37.7812800  4.2482744   8.893 4.14e-14 ***
income:professionals -0.0023257  0.0005675  -4.098 8.83e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 8.012 on 94 degrees of freedom
(4 observations deleted due to missingness)
Multiple R-squared:  0.7872,Adjusted R-squared:  0.7804
F-statistic: 115.9 on 3 and 94 DF,  p-value: < 2.2e-16
```

Conclusion: All estimated coefficients are statistically differentiable from zero at the $\alpha = 0.05$ level because the p-value < 0.05.

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

The formula of prediction equation is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{income} + \hat{\beta}_2 \times \text{professional} + \hat{\beta}_3 \times \text{income} \times \text{professional}$$

Getting the coefficients for writing the prediction equation.

```r
coefficients_q1 <- model$coefficients
print(coefficients_q1)
```

Output:

```
(Intercept)              income        professional income:professional
21.142258854          0.003170909         37.781279955         -0.002325709
```

Writing the prediction equation.

```
#Printing the prediction equation
cat("Prediction Equation:\nprestige =", coefficients_q1[1], "+",
    coefficients_q1[2], "* income +",
     coefficients_q1[3], "* professionals", "+", "(", coefficients_q1[4],"
    )", "* income:professionals\n")
```

Output:

```
Prediction Equation:
prestige = 21.14226 + 0.003170909 * income + 37.78128 * professionals +
( -0.002325709 ) * income:professionals
```

(d) Interpret the coefficient for `income`.

There is a positive and statistically reliable relationship between the income and the prestige, such that an one unit increase in income, on average, is associated with the increase of 0.003170909 units in prestige score, under controlling for the effects of all other predictor variables in the model.

(e) Interpret the coefficient for `professional`.

There is a positive and statistically reliable relationship between the professional and prestige, such that in comparisson to non-professional, an one unit increase in professional, on average, is associated with the increase of 37.78128 units in prestige score, under controlling for the effects of all other predictor variables in the model.

(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).

Assigning \$0 to variable `income` and 1 to variable `professional`, then calculating $\hat{y}$ using the prediction equation.

```
income <- 0
professional <- 1
y_hat <- 21.14226 + 0.003170909 * income + 37.78128 * professional -
  0.002325709  * income * professional
print(y_hat)
```

Output:

```
58.92354
```

Assigning \$1000 to variable `income`, the variable `professional` doesn't change. Calculating $\hat{y}$*_new_income* using the prediction equation.

```
income <- 1000
y_hat_new_income <- 21.14226 + 0.003170909 * income + 37.78128 *
    professional -
  0.002325709  * income * professional
print(y_hat_new_income)
```

Output:

```
59.76874
```

Calculating marginal effect between $\hat{y}$*_new_income* and $\hat{y}$.

```
marginal_effect <- y_hat_new_income - y_hat
print(marginal_effect)
```

Output:

```
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).

Assigning to variable `income` the amount of \$6000, and 0 to variable `professional`, then calculating $\hat{y}\_non\_prof$ using the prediction equation.

```
income <- 6000
professional <- 0
y_hat_non_prof <- 21.14226 + 0.003170909 * income + 37.78128 * professional -
 0.002325709 * income*professional
print(y_hat_non_prof)
```

Output:

```
40.16771
```

Assigning 1 to variable `professional`, the variable `income` doesn't change. Calculating $\hat{y}\_prof$ using the prediction equation.

```
professional <- 1
y_hat_prof <- 21.14226 + 0.003170909 * income + 37.78128 * professional -
  0.002325709 * income*professional
print(y_hat_prof)
```

Output:

```
63.99474
```

Calculating marginal effect between $\hat{y}\_prof$ and $\hat{y}\_non\_prof$.

```
marginal_effect <- y_hat_prof - y_hat_non_prof
print(marginal_effect)
```

Output:

```
23.82703
```

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 $\alpha = .05$).

Null hypothesis: H_0: Having these yard signs in a precinct doesn't affect vote share.

Alternative hypothesis: H_A: Having these yard signs in a precinct affects vote share

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

1) Calculating test-statistic using formula $t = \frac{\hat{\beta}_1 - 0}{se_{\hat{\beta}_1}}$

[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.

```r
beta1 <- 0.042
se_beta1 <- 0.016
t1 <- (beta1 - 0)/(se_beta1)
print(t1)
```

Output:

```
2.625
```

2) Calculating degrees of freedom using formula $df = N - k$, where N - total number of observations, k - the number of parameters estimated in the model, included the intercept and the predictior variables.

```r
N <- 131
k <- 3
df = N-k
print(df)
```

Output:

```
128
```

3) Calculating P-value

```r
p_value <- 2* pt(t1, df, lower.tail = FALSE)
print(p_value)
```

Output:

```
0.00972002
```

Interpretation: The estimated coefficient is statistically differentiable from zero at the $\alpha = 0.05$ level because the p-value < 0.05 ($\approx$0.0097), so we can reject the null hypothesis that having these yard signs in a precinct doesn't affect vote share.

(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 $\alpha = .05$).

Null hypothesis: H_0: Being next to precincts with these yard signs doesn't affect vote share.

Alternative hypothesis: H_A: Being next to precincts with these yard signs affects vote share.

$$H_0 : \beta_2 = 0$$
$$H_A : \beta_2 \neq 0$$

1) Calculating test-statistic using formula $t = \frac{\hat{\beta}_2 - 0}{se_{\hat{\beta}_2}}$

```
beta2 <- 0.042
se_beta2 <- 0.013
t2 <- (beta2 - 0)/se_beta2
print(t2)
```

Output:

```
3.230769
```

2) Calculating degrees of freedom using formula $df = N - k$, where N - total number of observations, k - the number of parameters estimated in the model, included the intercept and the predictior variables.

```
N <- 131
k <- 3
df = N-k
print(df)
```

Output:

```
128
```

3) Calculating P-value

```
p_value <- 2*pt(t2, df, lower.tail = FALSE)
print(p_value)
```

Output:

```
0.00156946
```

Interpretation: The estimated coefficient is statistically differentiable from zero at the $\alpha = 0.05$ level because the p-value < 0.05 ($\approx$0.0016), so we can reject the null hypothesis that being next to precincts with these yard signs doesn't affect vote share.

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

In regression analysis, the constant term is the estimated y-intercept that represents the expected value of the dependent variable when all independent variables are equal to zero. Thus, constant term $\beta_0 = 0.302$ represents the estimated average proportion of the vote that went to McAuliff's opponent Ken Cuccinelli in the absence of lawn signs assigned and adjacent to.

(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?

The correlation r and its square describe the strength of association between y and the set of explanatory variables acting together as predictors in the model. R^2 falls between 0 and 1. The larger the value of R^2, the better the set of explanatory variables ($x_1, ..., x_p$) collectively predicts y (Agresti, 2018, section 11.2 Multiple Correlation and R^2). Our $R^2 = 0.094$, it is small, thus, we could suggest that the included variables do not provide the better explaination of the proportion of the vote that went to McAuliff's opponent Ken Cuccinelli. The other factors that are not modeled could be significant.