Problem Set 4

Ana Sofia Jesus 19327602

03/12/2023

Question 1

a)

Once I had loaded the prestige dataset from the car library, I created a new variable entitled "professional" by recoding the variable type so that professionals are coded as 1, and blue and white collar workers are coded as 0, using the following code:

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

b)

Next, I ran a linear model with occupation prestige as an outcome and income, professional, and the interaction of the two as predictors, using the following code:

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

This model produced the following output:

```
    Call:
lm(formula = prestige ~ income + professional + income:professional,
    data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-14.852  -5.332  -1.272   4.658  29.932

Coefficients:
                     Estimate Std. Error t value
(Intercept)         21.1422589  2.8044261   7.539
income               0.0031709  0.0004993   6.351
professional        37.7812800  4.2482744   8.893
income:professional -0.0023257  0.0005675  -4.098
                    Pr(>|t|)
(Intercept)         2.93e-11 ***
income              7.55e-09 ***
professional        4.14e-14 ***
income:professional 8.83e-05 ***
---
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
```

c)

The prediction equation is written as follows:

$$prestige = 21.1423 + 0.0032 \times income + 37.7813 \times professional - 0.0023 \times (income \times professional)$$

d)

The coefficient for income is 0.0032, which means that for each one-unit increase in income, the Prestige is expected to increase on average 0.0032 scale points, while holding other variables constant. The positive sign indicates that there is a positive relationship between income and occupation prestige with this model. It's important to note that this interpretation assumes a linear relationship between income and prestige, and it holds other variables in the model constant. In other words, within the same type of occupation, a higher income is expected to be associated to a higher prestige score.

e)

The coefficient for the "professional" dummy variable is 37.7813. If an individual has a type of occupation classified as professional, the prestige score is expected to be, on average, 37.7813 units higher compared to an individual who has a blue or white-collar occupation, while holding other variables constant. Hence, in this model, being a professional is associated with a substantial increase in the expected Prestige score compared to blue or white-collar occupations.

f)

To estimate the effect of a 1,000 dollars increase in income on the prestige score for professional occupations, I began by calculating the predicted prestige score for an income of 1000 using the prediction equation and then calculated the predicted prestige score for an income of 2000 using the prediction equation and estimated the difference in prestige scores between the two income levels.
This difference in predicted prestige scores gives me an estimate of the change in prestige associated with a 1,000 increase in income, assuming other variables in the model remain constant.

Equation where the income is 1000:

$$prestige = 21.1423 + 0.0032 \times 1000 + 37.7813 \times 1 - 0.0023 \times (1000 \times 1) = 59.7687389$$

Equation where income increases by 1000:

$$prestige = 21.1423 + 0.0032 \times 2000 + 37.7813 \times 1 - 0.0023 \times (2000 \times 1) = 60.6139389$$

Based on the above equations, the difference between them represents the change in y associated with a 1000 dollar increase in income: 60.6139389 - 59.7687389 = 0.8452

g)

To estimate the effect on prestige of changing one's occupations from non-professional (0) to professional (1) when income is set at 6,000 dollars, I calculated the difference between the change in y for non-professionals and then for professionals.
For non-professional (Professional = 0)

$$prestige = 21.1423 + 0.0032 \times 6000$$

For professional (Professional = 1)=

$$prestige = 21.1423 + 0.0032 \times 6000 + 37.7813 - 0.0023 \times 6000$$

Now, to calculate the difference between the two:

$$(21.1423+0.0032\times6,000+37.7813\times1-0.0023\times6,000\times1)-(21.1423+0.0032\times6,000+37.7813\times0-0.0023\times6,000\times0)$$

= 64,324 - 40,342 = 23.982

Therefore, the change in prestige associated with changing from a non-professional occupation to a professional occupation, when income is 6,000 dollars, is on average 23.98 scale points. In other words, for a person earning 6000 dollars, prestige increases on average 23.98 scale points if the job is categorized as professional.

Question 2

a)

In order to determine whether having yard signs in a precinct affects vote share, I am conducting a hypothesis test using the estimated coefficient (Precinct assigned lawn signs) from the regression model. In other words I am testing the partial effect of this predictor. This predictor is a binary variable, and I want to test whether the coefficient is significantly different from zero.

Formulating Hypotheses:

Null Hypothesis (H0): There is no discernible linear relationship between predictor variable of having yard signs and response variable vote share after controlling for the effects of all other predictor variables in the model.

Alternative Hypothesis (H1): There is a discernible linear relationship between predictor variable of having yard signs and response variable vote share after controlling for the effects of all other predictor variables in the model.

Given that the coefficient is b_1 and the standard error is $SE(b_1)$, the t-statistic for testing the significance of b_1 is calculated as:

$$t = \frac{b_1}{SE(b_1)}$$

$$t = \frac{0.042}{0.016} = 2.625$$

Next I determined the degrees of freedom: n-k-1 = 30-2-1 = 27 df

Based on the t-distribution table, I find the two-tailed critical value to be 2.052 at 27 degrees of freedom. The absolute value of my test statistic 2.625 is greater than the critical value 2.052. Hence, we are able to reject the null hypothesis that there is no discernible linear relationship between the predictor and the response variable in the model. In other words, there is evidence that having yard signs in a precinct affects vote share.

b)

In order to determine whether being next to precincts having yard signs affects vote share, I am conducting a hypothesis test using the estimated coefficient (Precinct adjacent to lawn signs) from the regression model. In other words I am testing the partial effect of this predictor. This predictor is a binary variable, and I want to test whether the coefficient is significantly different from zero.

Formulating Hypotheses:

Null Hypothesis (H0): There is no discernible linear relationship between predictor variable of being next to a precinct having yard signs and response variable vote share after controlling for the effects of all other predictor variables in the model.

Alternative Hypothesis (H1): There is a discernible linear relationship between predictor variable of being next to a precinct having yard signs and response variable vote share after controlling for the effects of all other predictor variables in the model.

Given that the coefficient is b_1 and the standard error is $SE(b_1)$, the t-statistic for testing the significance of b_1 is calculated as:

$$t = \frac{b_1}{SE(b_1)}$$

$$t = \frac{0.042}{0.013} = 3.230769$$

Next I determined the degrees of freedom: n-k-1= 76-2-1 = 73

Based on the t-distribution table, I find the two-tailed critical value to be 1.994 at 73 degrees of freedom. The value of my test statistic 3.231 is greater than the critical value 1.994. Hence, we are able to reject the null hypothesis that there is no discernible linear relationship between the predictor and the response variable in the model. In other words, there is evidence that being exposed to yard signs in precincts close by affects the vote share.

c)

The constant term, also known as the intercept, represents the estimated value of the output variable (Ken Cuccinelli's vote share) when all predictors in the model are set to zero. In this regression model, the constant term is 0.302. The constant term represents the estimated proportion of the vote that went to McAuliffe's opponent when there are no lawn signs in the precincts. In other words, it constitutes the baseline for the vote share of Ken Cuccinelli, since that for the precincts which did not have any yard signs and were not adjacent to any lawn signs the average vote share for Ken Cuccinelli was 0.302.

More importantly, the fact that this coefficient is positive suggests an increase in the vote share for Cuccinelli in precincts with yard signs supposedly favouring the opponent. If we compare the coefficient for the treatment variable to the constant term in this linear regression model we obtain the following result:

$$\left(\frac{0.042}{0.302}\right) \times 100 \approx 13.91$$

The coefficient for the treatment variable indicates that, on average, precincts with lawn signs had a roughly 13.91 percent higher vote share for Cuccinelli compared to the baseline, assuming a linear relationship between the presence of yard signs and the vote share for Cuccinelli.

d)

The R^2 value provides a measure of how well the predictors in a regression model explain the variability in the output. In this model the R^2 is 0.094, corresponding to the proportion of the variance in the vote share for Cuccinelli that is explained by the yard signs in the precint and adjacent precints. Only approximately 9.4 per cent of the variability in the vote share for Cuccinelli is explained by the predictors included in this regression model, the remaining 90.6 per cent of the variability is unexplained by the variables in the model and may be attributed to other factors or random variation.