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`).

First, I thoroughly review the data set in addition to what is suggested:

```
# Check top 5 observations and variable names
head(Prestige, 5)
# Check details about data set type, its dimensions, and
# the names and types for each variable
str(Prestige)
# Gives summary statistics for each variable
summary(Prestige)
```

Next, I check the levels of `type` to inform the ifelse expression for creating a new variable `professional`

```
levels(Prestige$type)
# "prof" = professional; "bc" = blue-collar; "wc" = white-collar
```

Lastly, I create a new dummy variable `professional`, and then coerce it as a factor to ensure the following model does not interpret the variable as a continuous numerical variable.

```
# Create new dummy variable "professional"
# where prof = 1 and remaining (bc and wc) = 0.
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)
# Coerce as factor to ensure model doe not
# interpret as a continuous numerical variable
Prestige$professional <- as.factor(Prestige$professional)
# Check new variable
Prestige$professional
```

(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.)

I run the linear model. Note that the `lm()` function omits observations which include at least one NA value among selected variables by default. In this case, I specify.

```
model_1 <- lm(prestige ~ income + professional + income:professional,
    Prestige, na.action = "na.omit")
```

Interpretations for **Income** and **Professional = 1** are discussed in following questions. I note the Interaction term represents the average diminished effect of Income on Prestige Score for Professional workers. In other words, the average effect of increasing income by \$1 on a worker's Prestige Score is, on average, less for professional workers than for white/blue-collar workers (see Figure 1).

The Constant term represents the average Prestige Score for a white/blue-collar worker with \$0 income (or the lowest possible value in Income).

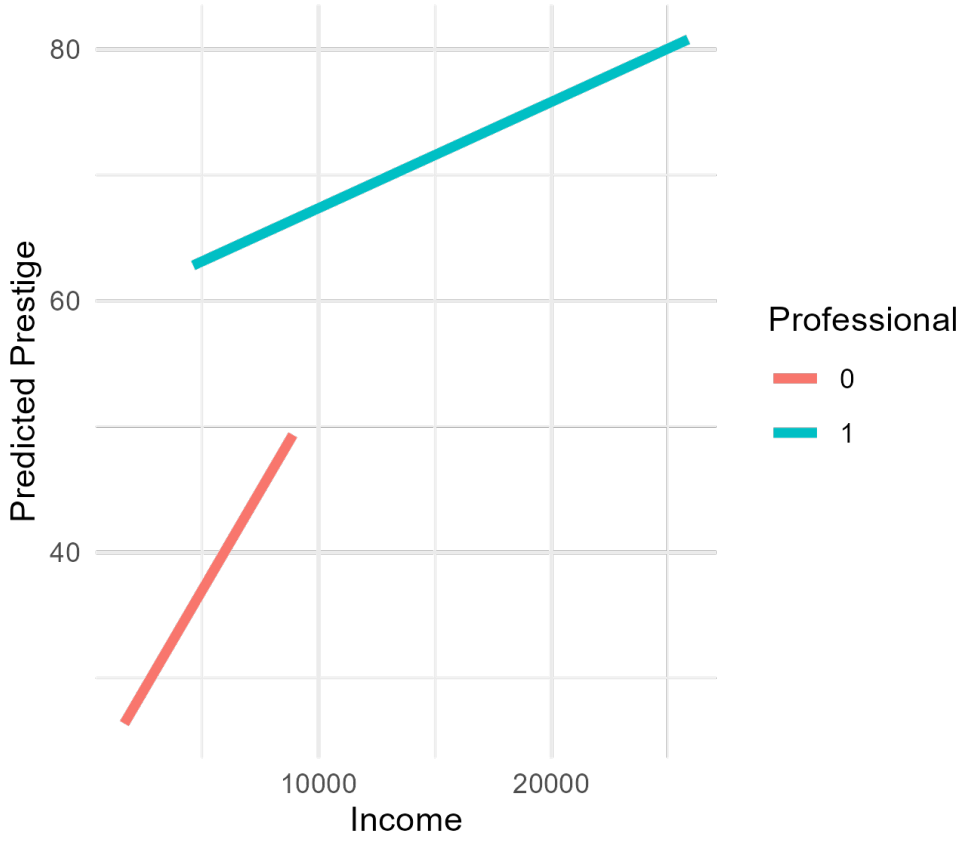
R^2 indicates the model predicts 78.7% of the variation in observed `Prestige Score` in the data set. `F Statistic` represents the hypothesis test that at least one of the

Table 1: Model 1 Regression Table

	Dependent variable:
	Prestige Score
Income ($1)	0.003^{***}
	(0.0005)
Professional = 1	37.781^{***}
	(4.248)
Interaction term: Income and Professional = 1	-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

partial effects (the coefficients) in the model is non-zero. In this case, the F-value is less than 0.05 and so gives evidence to reject the null hypothesis that all partial effects are equal to zero, and gives evidence for the alternative hypothesis that at least one of the partial effects is non-zero.

Figure 1: Marginal Effect of Income on Prestige



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

The following is a generic equation with two predictor variables X and D, an interaction term between the two, and stochastic error.

Note the stochastic error is assumed to cancel out given the assumptions for linear regression: $\varepsilon_i \overset{iid}{\sim} N(O, \sigma^2)$ (independence, normality, and constant variance of errors).

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 X_i D_i + \varepsilon_i$$

Below I input the prediction equation values:

$$\text{prestige} = 21.142 + 0.003 \times \text{income} + 37.781 \times \text{professional} + (-0.002 \times \text{income} \times \text{professional})$$

(d) Interpret the coefficient for `income`.

The coefficient for **income** represents the average increase in prestige per one additional dollar of income for white- and blue-collar workers (a.k.a. holding the professional variable constant).

Further, the p-value for the coefficient is less than the typcial $\alpha = 0.05$, therefore giving evidence to reject the null hypothesis that the partial effect of income on prestige for white/blue-collar workers is zero, and giving evidence for the alternative hypothesis that the partial effect of income on prestige for white/blue-collar workers is non-zero.

(e) Interpret the coefficient for `professional`.

The coefficient for **professional** represents the average increase in prestige for an individual who moves from white-/blue-collar employment to 'professional' employment, holding income constant.

Further, the p-value for the coefficient is less than the typcial $\alpha = 0.05$, therefore giving evidence to reject the null hypothesis that the partial effect of moving from white/blue-collar to professional employment is zero, holding income constant, and giving evidence for the alternative hypothesis that the partial effect is non-zero, holding income constant.

(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).

I am comparing the average increase in estimated prestige for a professional when their income increases from \$0 to \$1000.

$$\begin{aligned}
\text{prestige}_{professional+\$0} &= 21.142 + 0.003 \times 0 + 37.781 \times 1 + (-0.002 \times 0 \times 1) \\
&= 21.142 + 37.781 \times 1 \\
&= 58.92 \\
\text{prestige}_{professional+\$1000} &= 21.142 + 0.003 \times 1000 + 37.781 \times 1 + (-0.002 \times 1000 \times 1) \\
&= 59.77
\end{aligned}$$

```r
# Calculate average change in prestige associated with income increase
prof_zero_income <- sum(model_1$coefficients[1], model_1$coefficients[3]*1)
prof_zero_income
prof_1k_income <- sum(model_1$coefficients[1], model_1$coefficients[2]*1000, model_1$coefficients[3]*1, model_1$coefficients[4]*1000*1)
prof_1k_income
```

Subtracting the estimated prestige score of a professional with \$0 from that of a professional with \$1000 indicates the (average) marginal effect of increasing an average professional's income from \$0 to \$1000 is an increase of 0.85 prestige-score points.

```
# Calculate the marginal (average) effect of $1000
f_prestige <- prof_1k_income - prof_zero_income
f_prestige
```

(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).

$$\begin{aligned}
\text{prestige}_{non-professional+\$6000} &= 21.142 + 0.003 \times 6000 \\
&= 40.17 \\
\text{prestige}_{professional+\$6000} &= 21.142 + 0.003 \times 6000 + 37.781 \times 1 + (-0.002 \times 6000 \times 1) \\
&= 63.99
\end{aligned}$$

```
# Calculate average change in prestige associated with employment change
income_non_prof <- sum(model_1$coefficients[1], model_1$coefficients[2]*
    6000)
income_non_prof
income_prof <- sum(model_1$coefficients[1], model_1$coefficients[2]*6000,
     model_1$coefficients[3]*1, model_1$coefficients[4]*6000*1)
income_prof
```

Subtracting the estimated prestige score of a non-professional with \$6000 from that of a professional with \$6000 indicates the marginal (average) effect of moving from white/blue-collar employment to professional employment, when income is constant at \$6000, is associated with an average increase of 23.83 prestige-score points.

```
# Calculate the marginal (average) effect of $6000
g_prestige <- income_prof - income_non_prof
g_prestige
```

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 $\alpha = .05$).

I will conduct a t-test for the following hypothesis tests as the sample sizes for each group vary - treatment groups (n=30 and n=76) and control group (n=25) - two of which are below or at the margin of the typical cut-off point of n=30 for Z-scores.

$$H_0 : \hat{\beta}_1 = 0$$
$$H_a : \hat{\beta}_1 \neq 0$$

The test statistic is calculated by dividing the difference between the coefficient and null hypothesis by the standard error:

[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.

$$t = \frac{\hat{\beta} - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}}}$$

Below is the t-score for the coefficient representing the change in precinct voting preference correlated with assignment of lawn signs:

$$2.625 = \frac{0.042 - 0}{0.016}$$

Since we are testing the coefficient from a multiple regression model, I must calculate the residual degrees of freedom using $n-k-1$ to account for the estimated parameters, where n=sample size; k=number of predictor coefficients; -1= constant/intercept.

Therefore, we find 2.625 is greater than the critical value 1.703. This gives evidence to reject the null hypothesis that the effect of assigned lawn signs on voting preference is zero, and gives evidence for the alternative hypothesis that the effect is non-zero.

```r
# Calculate t-statistic
coefficient_a <- 0.042
se_a <- 0.016
n_a <- 30
null_hyp <- 0
t_score_a <- (coefficient_a - null_hyp) / se_a
t_score_a

# Calculate critical value and check
alpha <- 0.05
df_residual_a <- n_a - 2 - 1
critical_value_a <- qt(alpha, df = df_residual_a, lower.tail = FALSE)
critical_a <- ifelse(abs(t_score_a) > critical_value_a, "Greater than
    critical value", "Less than critical value")
critical_a
```

To be thorough, I also calculate the p-value to be 0.014, which is less than 0.05. This again gives evidence to reject the null hypothesis that the effect of assigned lawn signs on voting preference is zero, and gives evidence for the alternative hypothesis that the effect is non-zero.

```r
# Calculate p-value and check
p_value_a <- 2 * pt(abs(t_score_a), df = df_residual_a, lower.tail =
    FALSE)
null_hyp_a <- ifelse(p_value_a < alpha, "Less than 0.05", "Greater than
    0.05")
null_hyp_a
```

(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 $\alpha = .05$).

As explained above, while n=76 is relatively large in this case, I will conduct a t-test.

$$H_0 : \hat{\beta}_2 = 0$$
$$H_a : \hat{\beta}_2 \neq 0$$

Below is the t-score for the coefficient representing the change in precinct voting preference correlated with assignment of lawn signs to an adjacent precinct:

$$3.231 = \frac{0.042 - 0}{0.013}$$

Therefore, we find 3.231 is greater than the critical value 1.666. This gives evidence to reject the null hypothesis that the effect of assigned lawn signs to at least one adjacent precinct on voting preference is zero, and gives evidence for the alternative hypothesis that the effect is non-zero.

```r
# Calculate t-statistic
coefficient_b <- 0.042
se_b <- 0.013
n_b <- 76
t_score_b <- (coefficient_b - null_hyp) / se_b
t_score_b

# Calculate critical value and check
df_residual_b <- n_b - 2 - 1
critical_value_b <- qt(alpha, df = df_residual_b, lower.tail = FALSE)
critical_b <- ifelse(abs(t_score_b) > critical_value_b, "Greater than
    critical value", "Less than critical value")
critical_b
```

To be thorough, I also calculate the p-value to be 0.002, which is less than 0.05. This again gives evidence to reject the null hypothesis that the effect of assigned lawn signs to at least one adjacent on voting preference is zero, and gives evidence for the alternative hypothesis that the effect is non-zero.

```r
# Calculate p-value and check
p_value_b <- 2 * pt(abs(t_score_b), df = df_residual_b, lower.tail =
    FALSE)
null_hyp_b <- ifelse(p_value_b < alpha, "Less than 0.05", "Greater than
    0.05")
null_hyp_b
```

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

The constant represents the (estimated) average voting preference within a precinct when there is no lawn sign assigned within the given precinct or within an adjacent precinct.

In other words, the constant is the y-intercept for the line of best fit produced by the linear regression model, where the lawn sign-assignment and lawn sign-adjacent variables are zero (as this is their lowest possible value).

(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 model fit R^2, also known as the model's "explanatory power" or "predictive power", represents **how much total variation** in the outcome variable(s) covaries with the estimated line of best fit, which is based on the given predictor variable(s). It is calculated by dividing the variability *explained* by the model SSR by the *total variability* of the observed outcome data SST:

$$R^2 = \frac{SSR}{SST}$$

$R^2 = 1$ indicates 100% covariation, meaning the observed outcome(s) and prediction line have a "perfect" linear relationship without error. Meanwhile, $R^2 = 0$ indicates there is no covariation between the outcome(s) and prediction line. Figures at either extreme end (perfect collinearity or perfect absence of a relationship)are not typically expected without errors in coding or choice of variables.

In this case, $R^2 = 0.094$, indicating 9.4% of the variation in precinct voting preference is predicted by the regression model based on the predictor variables representing a) if a precinct has an assigned lawn sign, and b) if a precinct is adjacent to at least one which has an assigned lawn sign. In other words, 90.6% of the variation in voting preference is NOT explained by the model's prediction line.

Whether this is an acceptable 'fit' depends on the theoretical expectations and practical goals of the researcher: if the purpose is to explain the relationship between the predictor(s) and outcome(s), then achieving a high R^2 value isn't a priority. However, if the purpose is to maximise the *predictive power* of the model, then increasing the R^2 by changing predictor variables may be useful, so long as these inclusions are theoretically-informed. Note this can include adding and removing variables (see more below).

Note the sample sizes - precincts assigned (n=30) and precincts adjacent (n=76), and voting preference (n=25). Two out of three are below or at the margin of the standard arbitrary cutoff n=30 for assuming normality of sampling distributions. The control group does not meet this prerequisite for the validity of the regression model, and so

larger sample size for the control group may increase the model's validity alongside its R^2 while reducing coefficients' standard errors $\hat{\sigma}$'s. That said, the "assigned" group is on the margin of this prerequisite, and so the model may improve along these lines by increasing its n-size as well.

In addition, there is an explicit expectation in the given text that people in adjacent precincts to those assigned signs are likely to cross over and see them. Therefore, there is risk of collinearity (each variable measuring the same phenomenon - namely exposure to lawn signs) AND dependence (variation in one predictor is dependent on variation in the other i.e. assignment to one precinct necessarily 'creates' the adjacent precincts captured), each violating prerequisites for valid linear regression. Excluding one of these variables may increase the overall model fit alongside its overall theoretical validity of the model.