Problem Set 4

Applied Stats/Quant Methods 1

Due: November 26, 2021

Name: Gareth Moen

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 in `.pdf` form.
- This problem set is due before class on Friday November 26, 2021. No late assignments will be accepted.
- Total available points for this homework is 80.

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`.)

Creating a new empty "professional" column.

```
Prestige$professional <- NA
```

Filling the column with data

```
Prestige$professional <- as.factor(case_when(
Prestige$type == "prof" ~ "1",
Prestige$type == "wc" ~ "0",
Prestige$type == "bc" ~ "0",
))
```

View the new column

```
view(Prestige)
class(Prestige$professional)
```

Type is now 'factor'

(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.)

```
mod_1 <- lm(prestige ~ income + professional + income:professional,
data = Prestige)
```

View the results

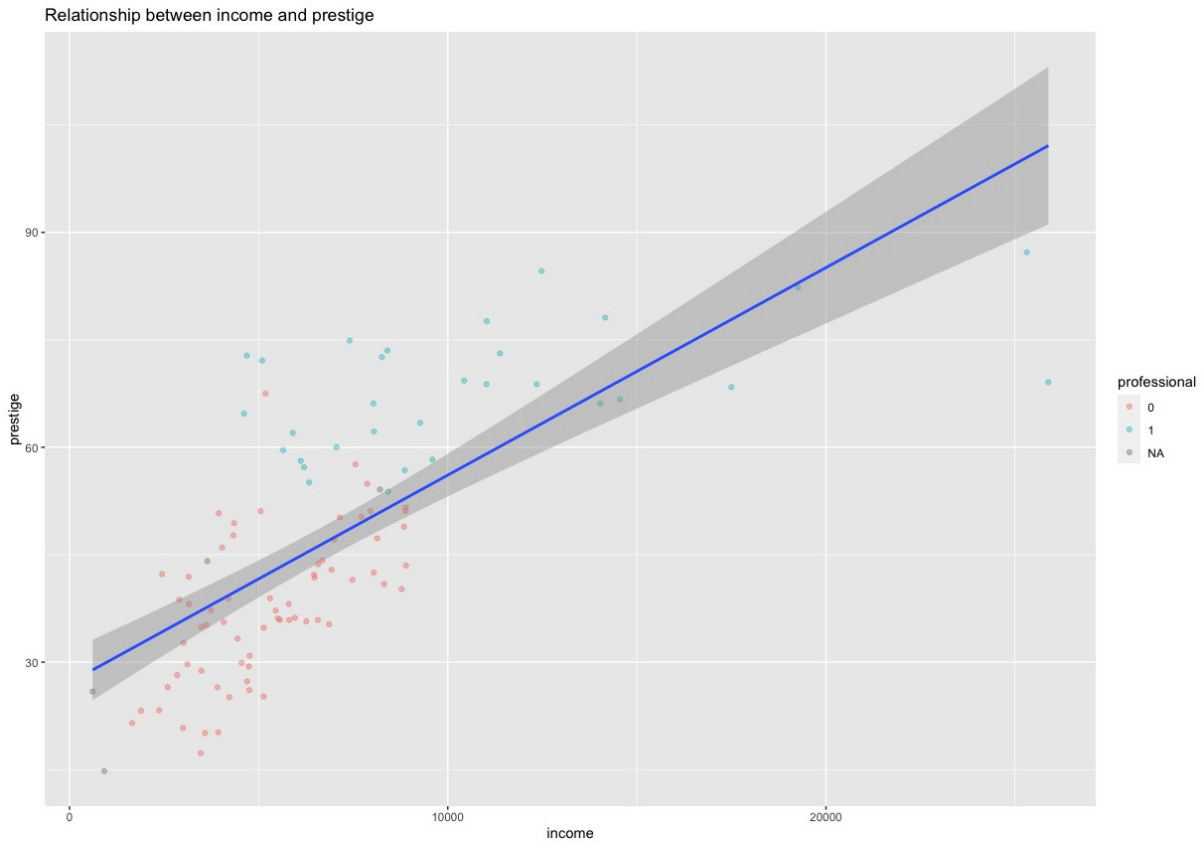
```
stargazer(mod_1, type = "text")
```

```
================================================
                        Dependent variable:
                    ----------------------------
                               prestige
------------------------------------------------
income                         0.003***
                               (0.0005)

professional1                  37.781***
                                (4.248)

income:professional1           -0.002***
                                (0.001)

Constant                       21.142***
                                (2.804)

------------------------------------------------
Observations                      98
R2                               0.787
Adjusted R2                      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
```

$R^2 = 0.787$ which means that the variables used add a high degree of predictive power to the model and also represent a 78.7% reduction in prediction error as compared to using $\bar{y}$.

A plot of prestige and income, colour coded for professional...

```
ggplot2::ggplot(aes(x = income, y = prestige), data = Prestige) +
geom_point(aes(colour = factor(professional)), alpha = 0.4) +
geom_smooth(method = "lm", formula = y ~ x) +
ggtitle("Relationship between income and prestige")
```



There is a strong relationship between being a professional and prestige

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

Prediction equation formula is...

$\hat{y}$ = a + b1x1 + b2x2 + b3x1x2

$\hat{y}$ = 21.142 + (0.003)x1 + (37.781)x2 + (-0.002)x1x2

(d) Interpret the coefficient for `income`.

There seems to be a significant (p = 7.55e-09) but weak correlation between income and prestige. A 1 unit increase in prestige corresponds to a $0.003 increase in income.

(e) Interpret the coefficient for `professional`.

There seems to be a significant (p = 4.14e-14) and strong correlation between professional and prestige, with a 1 unit increase in professional (a switch from not being a professional to being one) equaling a 37.781 unit increase in prestige.

(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).

From (c): $\hat{y} = 21.142 + (0.003)x1 + (37.781)x2 + (-0.002)x1x2$. An increase in \$1000 equals a $(0.003)(1000)$ increase in prestige = 3 units.

(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).

From (c) again: $\hat{y} = 21.142 + (0.003)x1 + (37.781)x2 + (-0.002)x1x2$

Now substituting \$6000 for income and 1 for professional $21.142 + (0.003)(6000) + (37.781)(1) + (-0.002)(6000)(1) = 21.142 + 18 + 37.781 - 12 = 39.142$ (non-professional prestige level at \$6000) + 25.781 (increase in prestige level of professional with \$6000 salary).

A switch from non-professional to professional represents a 25.781 unit increase in prestige at a salary level of \$6000.

```
# drop NA values in table, didn't work
tidyr::drop_na(Prestige, professional)
```

Plot of separate regression lines (includes NA values)

```
ggplot2::ggplot(aes(x = income, y = prestige), data = Prestige) +
geom_point(aes(colour = professional), alpha = 0.4) +
geom_smooth(method = "lm", aes(colour = professional)) +
ggtitle("Separate regression lines for professionals and non professionals")
```

Plot of separate regression lines while trying to eliminate NA-value line (which didn't work)

```
Prestige %>%
tidyr::drop_na(Prestige, professional) %>%
ggplot2::ggplot(Prestige, aes(x = income, y = prestige)) +
geom_point(aes(colour = professional), alpha = 0.4) +
geom_smooth(method = "lm", aes(colour = professional)) +
ggtitle("Separate regression lines for professionals and non professionals")
```

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 $\alpha = .05$).

Hypothesis test with $\alpha = 0.05$

Hypothesis test: $H_0 : \beta_i = 0$, with sample estimate $b_i for \beta_i$

$H_a : \beta_i \neq 0$

$t = b_i / se$

Our test statistic = $(0.042)/(0.016) = 2.625$

df = n - (k + 1) = 131 - (2 + 1) = 128

[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.

This gives us a two-tailed p-value of .00972 which tells us that we can reject the null, being in a precinct with signs has an effect.

(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 $\alpha = .05$).

Again we have a hypothesis test with $\alpha = 0.05$

Hypothesis test: $H_0 : \beta_i = 0$, with sample estimate $b_i for \beta_i$

$H_a : \beta_i \neq 0$

$t = b_i / se$

Our test statistic = $(0.042)/(0.013) = 3.23$

df = n - (k + 1) = 131 - (2 + 1) = 128

This gives us a two-tailed p-value of .001573 which tells us that we can reject the null and that being next to a precinct with signs also has an effect.

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

The constant term is .302 which is the predicted y value when both 'precinct assigned law signs' and 'precinct adjacent to lawn signs' equal 0.

(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?

$R^2 = 0.094$ which means that the variables used add a low degree of predictive power to the model as they only represent a 9.4% reduction in prediction error as compared to using $\bar{y}$. This tells us that there are other variables not in the model which are likely to have a much greater predictive power than those currently in the model.