# Submission for Problem Set 4

Applied Stats/Quant Methods 1

Duc Minh, VU
TCD StudentID: 22996761 / UCD StudentID: 19211157

## 1 Question 1

The ***prestige*** dataset will first be imported and the relevant libraries will be loaded in R for analysis.

```r
#install.packages("car")
library(car)
data(Prestige)
help(Prestige)
library(texreg)
```

(a) *Creating a new variable for type of work*
Using the ***ifelse*** function in R, the new variable ***professional*** has been created with value of 1 for professional ***type*** and value of 0 blue and white collar ***type***.

```r
summary(Prestige$type)
Prestige$professional <- as.factor(ifelse(Prestige$type == "prof", 1,
                                    ifelse(Prestige$type == "bc" | Prestige$type == "wc",0,NA)))
summary(Prestige$professional)
```

```
> summary(Prestige$type)
bc prof   wc NA's
44   31   23    4
> summary(Prestige$professional)
0    1 NA's
67   31    4
```

(b) *Linear model of **prestige** on **income**, **professional** and their interaction effects.*

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

```r
texreg(presti_lm,
       caption = "Regressing prestige on income, professional type and
    their interaction effect",
       float.pos = "H", digits = 4)
presti_lm.coff <- presti_lm$coefficients #Saving the coefficients for
    subsequent sections
```

```
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 Pr(>|t|)
(Intercept)          21.1422589  2.8044261   7.539 2.93e-11 ***
income                0.0031709  0.0004993   6.351 7.55e-09 ***
professional1        37.7812800  4.2482744   8.893 4.14e-14 ***
income:professional1 -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
```

| | Model 1 |
|---|---|
| (Intercept) | $21.1423^{***}$ |
| | (2.8044) |
| income | $0.0032^{***}$ |
| | (0.0005) |
| professional1 | $37.7813^{***}$ |
| | (4.2483) |
| income:professional1 | $-0.0023^{***}$ |
| | (0.0006) |
| $R^2$ | 0.7872 |
| Adj. $R^2$ | 0.7804 |
| Num. obs. | 98 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Regressing prestige on income, professional type and their interaction effect

(c) *Prediction equation*

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

$$\texttt{Prestige} = 21.1423 + 0.0032 \times \texttt{Income} + 37.7813 \times \texttt{Professional} - 0.0023 \times \texttt{Income} \times \texttt{Professional}$$

(d) *Interpreting* `Income` *coefficient*
The coefficient for `Income` is 0.0032 which reflects a positive relationship between `Income` and `Prestige` score. Holding other terms in the model constant, a one unit increase in `Income` is predicted to, on average, increase `Prestige` by 0.0032.

(e) *Interpreting* `Professional` *coefficient*
The coefficient for `Professional` is 37.7813. It should be noted that `Professional` is a dummy variable with `Non-Professional` (blue and white collars) as the baseline group, so the coefficient will be the 'offset' relative to the baseline group. Hence, holding `Income` constant, the prestige score is predicted to be, on average, 37.7813 higher for people in the `Professional` occupation group in comparison to `Non-Professional` occupation group of blue and white collars.

(f) Marginal effect of a $1,000 increase in income on prestige for professional occupations
1. `Prestige` score for a person with a professional occupation and with an `Income` of $1,000:

$$\texttt{Prestige} = 21.1423 + 0.0032 \times \texttt{Income} + 37.7813 \times \texttt{Professional} - 0.0023 \times \texttt{Income} \times \texttt{Professional}$$

$$\texttt{Prestige} = 21.1423 + 0.0032 \times 1,000 + 37.7813 \times 1 - 0.0023 \times 1 \times 1,000$$

$$\texttt{Prestige} = 59.8326$$

A person with a professional occupation and an income of $1,000 is predicted to have a prestige score of 59.8326.

```
presti.part.f_1 <- 21.1423 + 1000*0.0032 + 1*37.7813 + 1*1000*(-0.0023)
```

```
> presti.part.f_diff
[1] 0.9
```

2. $\mathtt{Prestige}$ score for a person with a professional occupation and with 0 $\mathtt{Income}$:

$$\mathtt{Prestige} = 21.1423 + 0.0032 \times \mathtt{Income} + 37.7813 \times \mathtt{Professional} - 0.0023 \times \mathtt{Income} \times \mathtt{Professional}$$

$$\mathtt{Prestige} = 21.1423 + 0.0032 \times 0 + 37.7813 \times 1 - 0.0023 \times 1 \times 0$$

$$\mathtt{Prestige} = 58.9236$$

A person with a professional occupation and no income is predicted to have a prestige score of 58.9236.

```
presti.part.f_2 <- 21.1423 + 0*0.0032 + 1*37.7813 + 1*0*(-0.0023)
```

```
> presti.part.f_2
[1] 58.9236
```

3. Hence, the marginal effect or difference of prestige point for professional going from $0 to $1,000 is 59.8326 - 58.9236 = 0.9.

```
presti.part.f_diff <- presti.part.f_1 - presti.part.f_2
```

```
> presti.part.f_diff
[1] 0.9
```

(g) Marginal effect of $\mathtt{Professional}$ jobs when the variable $\mathtt{Income}$ takes the value of $6,000

1. $\mathtt{Prestige}$ score for a person with a professional occupation and with an $\mathtt{Income}$ of $6,000:

$$\mathtt{Prestige} = 21.1423 + 0.0032 \times \mathtt{Income} + 37.7813 \times \mathtt{Professional} - 0.0023 \times \mathtt{Income} \times \mathtt{Professional}$$

$$\mathtt{Prestige} = 21.1423 + 0.0032 \times 6,000 + 37.7813 \times 1 - 0.0023 \times 1 \times 1,000$$

$$\mathtt{Prestige} = 64.3236$$

A person with a professional occupation and an income of $6,000 is predicted to have a prestige score of 64.3236.

```
presti.part.g.prof <- 21.1423 + 6000*0.0032 + 1*37.7813 + 1*6000*(-0.0023)
```

```
> presti.part.g.prof
[1] 64.3236
```

2. *Prestige* score for a person with a NON-professional occupation and with an *Income* of $6,000:

$$\mathit{Prestige} = 21.1423 + 0.0032 \times \mathit{Income} + 37.7813 \times \mathit{Professional} - 0.0023 \times \mathit{Income} \times \mathit{Professional}$$

$$\mathit{Prestige} = 21.1423 + 0.0032 \times 6,000 + 37.7813 \times 0 - 0.0023 \times 0 \times 1,000$$

$$\mathit{Prestige} = 40.3423$$

A person with a NON-professional occupation and an income of $6,000 is predicted to have a prestige score of 40.3423.

```
presti.part.g.non.prof <- 21.1423 + 6000*0.0032 + 0*37.7813 + 0*6000*
    (-0.0023)
```

```
> presti.part.g.non.prof
[1] 40.3423
```

3. And finally, for a person with an income of $6,000, the marginal effect or difference of prestige point in changing the occupation is 64.3236 - 40.3423 = 23.9813.

```
presti.part.g.diff <- presti.part.g.prof - presti.part.g.non.prof
```

```
> presti.part.g.diff
[1] 23.9813
```

# 2 Question 2

(a) Conducting hypothesis test for having yard signs in predicting vote share
Since we are concerned with whether having yard signs have any effect at all on the vote share, it will be a 2-sided hypothesis test. Therefore,

- The null hypothesis is: having yard sign has NO effect on the vote share:

$$\mathbf{H_0\text{:}}\ \beta_{\textbf{yard sign}} = \mathbf{0}$$

- The alternative hypothesis is: having yard sign has an effect on the vote share:

$$\mathbf{H_0\text{:}}\ \beta_{\textbf{yard sign}} \neq \mathbf{0}$$

- The test statistic is:

$$t = \frac{\beta_{\text{yard sign}} - \mu_0}{se} = \frac{0.042 - 0}{0.016} = 2.625$$

```r
test.stat.yard = 0.042/0.016
```

```
> test.stat.yard
[1] 2.625
```

• With 131 observations and 2 variables in the regression model, the degrees of freedom is: 131 - 2 - 1 = 128. The associated p-value for the above test statistic is:

```r
p.val.yard = 2*pt(test.stat.yard, df = 128, lower.tail = F)
```

```
> p.val.yard
[1] 0.00972002
```

which provides the value of 0.00972.
• With $\alpha = 0.05$ which is larger than our calculated p-value of 0.00972, there is statistical evidence for us to reject the null hypothesis $H_0$ and say that posting signs around the precint has a positive effect on the vote share.

(b) Conducting hypothesis test for being next to precincts with these yard signs in predicting vote share
Same as part (a), this will be a 2-sided test, since we are concerning with if there is any affect at all. Therefore:
• The null hypothesis is: being next to precincts with yard signs has NO effect on the vote share:

$$\mathbf{H_0\text{:}}\ \boldsymbol{\beta}_{\textbf{neighbour}} \mathbf{= 0}$$

• The alternative hypothesis is: being next to precincts with yard signs has an effect on the vote share:

$$\mathbf{H_0\text{:}}\ \boldsymbol{\beta}_{\textbf{neighbour}} \boldsymbol{\neq} \mathbf{0}$$

• The test statistic is:

$$t = \frac{\beta_{\text{neighbour}} - \mu_0}{se} = \frac{0.042 - 0}{0.013} = 3.231$$

```r
test.stat.nxt = 0.042/0.013
```

```
> test.stat.nxt
[1] 3.230769
```

• With a degree of freedom of 128, the associated p-value for the above test statistic is:

```r
p.val.yard = 2*pt(test.stat.yard, df = 128, lower.tail = F)
```

```
> p.val.nxt
[1] 0.00156946
```

which provides the value of 0.00157.

• With $\alpha = 0.05$ which is larger than our calculated p-value of 0.00972, there is statistical evidence for us to reject the null hypothesis $H_0$ and say that being in a precint next to a precint in the treatment group has a positive effect on the vote share.

(c) Interpreting the coefficent for the constant term

The two independent variables are all dummy variables, since both sign post and location are yes/no answers. Hence, the constant term is the basline level for the expected value of independent variables or when when the dummy varaibles are the baseline group. So the constant term is the average vote share that went to Cuccinelli, when a precint is NOT assigned with a lawn sign and NOT adjacent to a precint with a lawn sign, which is 0.302. It has a test statistic of 27.45 with a very small p-value which indicates that this constant is significance.

(d) Model fit

To evaluate the fit of a model, we use the $R^2$ value, which shows the proportion of variance in the dependent variable that can be explained by the independent variables. The $R^2$ value for the current model is 0.094 which is quite small, as only 9.4% of the variance in the vote share is explained by the location of yard sign and the location of the precint. Indeed, yard sign have a valuable contribution in explaining vote share with its strong statistical evidence. However, it only accounts for a small proportion of variation in vote share. There are still 90% of unexplained variances which shows that there other factors that can explain vote share.