Problem Set 4

Applied Stats/Quant Methods 1

Due: December 4, 2022

Question 1: Economics

Using the `prestige` dataset in the `car` library

```r
data(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) Created 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. [1]

```r
Prestige['professional'] <- ifelse(Prestige$type == 'prof', 1, 0)
```

(b) A linear model with `prestige` as an outcome and `income`, `professional`, and the interaction of the two as predictors was run.

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

The results are in Table 1. The *pvalue* for β_3 is 8.831093×10^{-05}, so we reject the hypothesis that the two variables do not interact, ie we conclude that a person's income level is not associated with the same level of prestige for professionals and non-professionals (see Figure 1).

[1] There were 4 nas in type - athletes, newsboys, babysitters, farmers, these were excluded from the model. See models in Table 2

Table 1: Prestige as a function of professional job and income

	Dependent variable:
	prestige
income	0.003171***
	(0.000499)
professional	37.781280***
	(4.248274)
income:professional	-0.002326^{***}
	(0.000567)
Constant	21.142260***
	(2.804426)
Observations	98
R^2	0.787154
Adjusted R^2	0.780361
Residual Std. Error	8.011644 (df = 94)
F Statistic	115.877800*** (df = 3; 94)
Note:	*p<0.1; **p<0.05; ***p<0.01

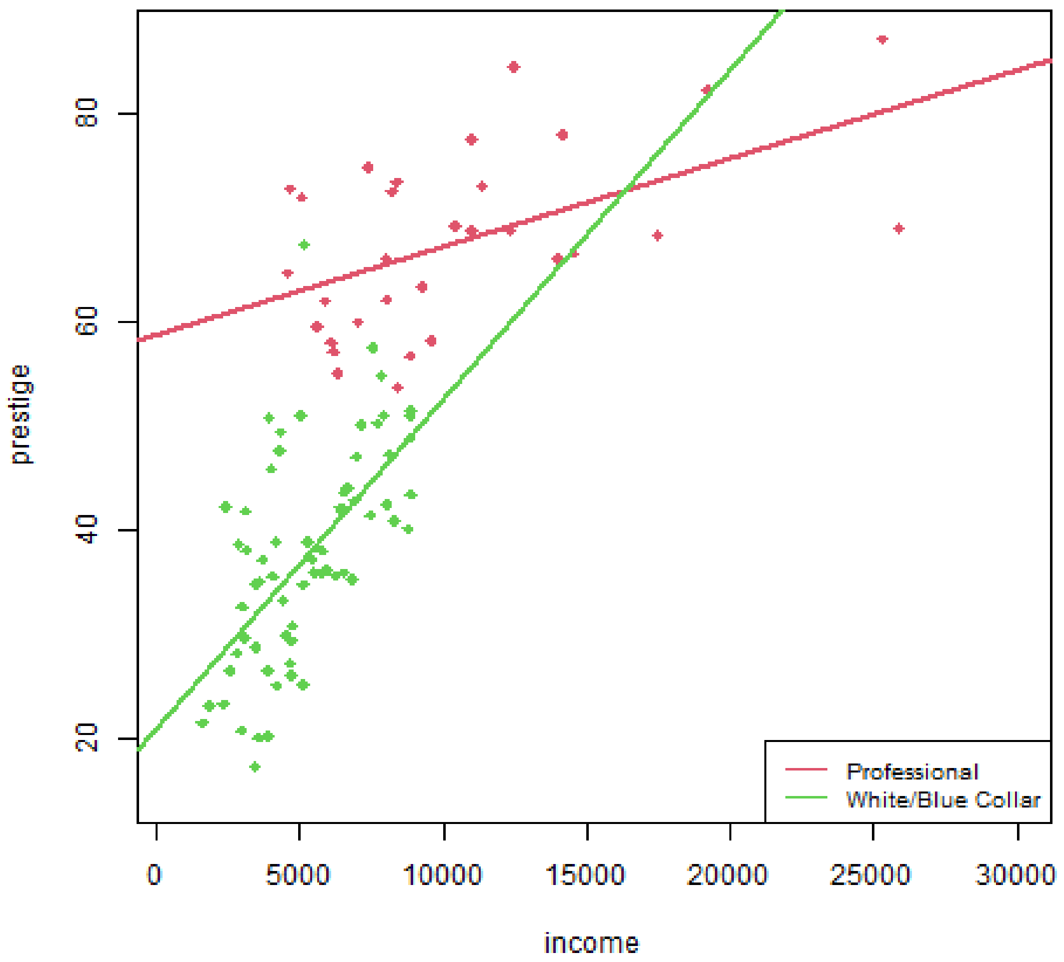

Figure 1: Prestige values for professionals/not as a function of income

(c) **Prediction Equation**:

if y = prestige, x = income, z = professional

$$\hat{y} = \beta_0 + \delta_1 z + \beta_1 x + \delta_2 xz + \epsilon \tag{1}$$

$$\beta_0 = 21.14226, \beta_1 = 0.003171, \delta_1 = 37.78128, \delta_2 = -0.002326$$

ie

$prestige = 21.14226 + 37.78128 \times professional + 0.003171 \times income + (-0.002326) \times professional \times income$

(d) The coefficient for `income` is $\beta_1 + \delta_2 z = 0.003171 - 0.002326 \times professional$

Assuming professional status is constant, a $1 rise in `income` results in a predicted rise in `prestige` of 0.003171 if in a blue collar or white collar job; `prestige` rises by 0.000845 per $ (ie 0.003171-0.002326) if in a professional job. The relationship of prestige with `income`, and the interaction between income and `professional`, are both statistically significant (at 0.1%). As per Figure 1, the `prestige` score for a person in a professional job starts at a higher value, for a 0 `income`, but the slope of the line predicting `prestige` is less steep, as increase in their `income` has less effect.

(e) The coefficient for `professional` is $\delta_1 + \delta_2 x = 37.78128 - 0.002326 \times income$

If `professional` switches to 1 (ie `type = prof`) then `prestige` increases by 37.78128 -0.002326× `income`, if `income` is held constant. If `professional` changes to 1 and `income` changes, the change in prestige is 37.78128 - 0.002326 × old income + 0.000845 × change in `income`. The relationships between `prestige` and `income` and the interaction between `income` and `professional` are both statistically reliable (at 0.1%). [2]

(f) If `professional` = 1, and $\Delta income = 1000$ then the equation becomes

$$\Delta\hat{y} = ((\beta_1 + \delta_2 \times 1) \times 1000$$

The marginal effect on `prestige` of an increase of $1,000 in income, for professional occupations: $= (0.003171 - 0.002326) \times 1000 = 0.845$

(g) From equation 1, if `income` x is 6000, then when `professional` z_0 was 0, `prestige` was:

$$\hat{y_0} = (\beta_0) + (\beta_1)x = 21.14226 + 0.003171 \times 6000 = 40.16826$$

[2] Hypothetically, a professional job would have more prestige for incomes below c$16,250 and less prestige for incomes above that level, as this is where the increase in the intersection value is wiped out by the decrease in the income coefficient. This that can't be supported by the current data, as the maximum income for any non-professional worker is $8,895 (fireworker).

when `professional` z_1 is 1, `prestige` is:

$$\hat{y}_1 = (\beta_0 + \delta_1 \times 1) + (\beta_1 + \delta_2 \times 1)x$$

$$= (21.14226 + 37.78128) + (0.003171 - 0.002326) * 6000 = 63.99354$$

so,

$$\Delta\hat{y} = \hat{y}_1 - \hat{y}_0 = 63.99354 - 40.16826 = 23.82528$$

(or, $\Delta\hat{y} = \delta_1 + \delta_2 x = 37.78128 - 0.002326 \times 6000$)

The marginal effect of professional jobs when the variable `income` takes the value of \$6,000 is an increase in `prestige` of 23.82528.

Question 2: Political Science

Researchers are interested in learning the effect of all of those yard signs on voting preferences.[3] 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

```r
n <- 131
est_coeffs <- 3
df <- n-est_coeffs
R2 <- 0.094
```

(a) To determine whether having these yard signs in a precinct affects vote share we conduct a hypothesis test with $\alpha = .05$. The null hypothesis is that having yard signs in a precinct have no effect on the voting in that precinct, ie $H_0 : \beta_1 = 0, H_{alt} : \beta_1 \neq 0$.

```r
beta1 <- 0.042
n1 <- 30
se1 <- 0.016
# get t-test value and pvalue
t_precinct <- beta1 / se1
pval_precinct <- 2*pt(abs(t_precinct), df , lower.tail = FALSE)
```

[3] 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.

$tvalue = 0.042/0.016 = 2.625$

$pvalue = Pr(tvalue = 2.625, df = 128) = 0.00972002$

Our $pvalue = 0.97\% < \alpha$, so we reject the null hypothesis and conclude that there is a positive and statistically significant relationship between the presence of signs in yards in a precinct and vote share. Specifically, having yard signs attacking his opponent in a precinct is associated with an increase in Ken Cuccinelli's vote share in that precinct of 0.042 (ie + 4.2%).

(b) To determine whether being next to precincts with these yard signs affects vote share we conduct a hypothesis test with $\alpha = .05$. The null hypothesis is that having yard signs in a precinct have no effect on the voting in the adjacent precincta, ie $H_0 : \beta_2 = 0, H_{alt} : \beta_2 \neq 0$.

```r
n2 <- 76
beta2 <- 0.042
se2 <- 0.013

t_adj <- beta2 / se2
pval_adj <- 2*pt(abs(t_adj), df, lower.tail = FALSE)
```

$tvalue = 0.042/0.013 = 3.230769$

$pvalue = 0.00156946$

Our $pvalue = 0.16\% < \alpha$, so we reject the null hypothesis and conclude that the presence of signs in yards in a precinct is supportive of the alternative, non-zero, hypothesis. A positive and statistically reliable relationship exists between the presence of yard signs in a precinct, and an increase in Cuccinelli's vote share in the adjacent precincts. The associated increase is 0.042 (4.2%).

(c) If not in a precinct with yard signs and not in an adjacent precinct, Ken Cuccinelli averages 30.2% of the vote. At the 95% confidence level, the vote for Cuccinelli in a precinct which is considered to be unaffected by signs is between 28.02% and 32.38%.

```r
beta0 <- 0.302
se0 <- 0.011

tscore <- qt(0.975, df) # get tscore for df 128,

CI0_L <- beta0 - tscore*se0
CI0_U <- beta0 + tscore*se0
```

t-statistic for two-sided test with $\alpha = 0.95\%$, 128 degrees of freedom = 1.978671

(d) If the model assumptions are valid, the presence of yard signs (even if they don't mention the candidate or their party) can predict an increase in votes of 4.2%[4]. That could be the difference between winning an losing a close election, so the signs would be

[4] CI β_1: 1.03% to 7.37%; CI β_2: 1.63% to 6.77%, at 95%

worth the cost, particularly as they increased vote share in 106 precincts, even though they were only placed in 30.

The model requires a linear relationship between the predictors and the independent variable (vote share), the independence of the predictors and equal variance of the residuals.

The model is assumed to have an effectively random/representative assignment of yard signs (it couldn't be actually random, or there would have been precincts with yard signs adjacent to each other). This assumption could fail if precincts are heterogeneous with regard to party affiliation, voter turnout, through traffic, size (affecting proportion of voters who would have seen the signs), etc. The model assumes predictors are independent of each other, ie that the vote share in the adjacent precincts is not being affected by some other characteristic based on proximity to the areas with yard signs, and that the change in vote share is not due to a confounding variable which affects the sign and adjacent precincts, but which is not being modelled.

However, the R^2 value is 0.094, ie only 9.4% of the variation in the dependent variable can be explained by the yard sign model. Most of the variation in vote share results from factors which are not modelled. If those missing variables were accounted for they could change both the values and the significance of the coefficients. Alternatively, the relationship between the dependent variable (vote share) and the predictors (yard signs) may not be linear, and an alternative model, with the same predictors, might be appropriate.

1 Appendix

1.0.1 Code

`PS04_ImeldaFinn.R`

Q1 - model variations depending on treatment of missing values

Went with conservative option of ignoring NAs rather than recoding. The difference to the coefficients and p-values wasn't significant.

```r
# class all nas as non-professional(based on education)
Prestige2 <- Prestige
Prestige2['professional'] <- ifelse(is.na(Prestige2$type) ,0,Prestige2$professional)

# class all nas as non-professional, apart from athletes(based on income)
Prestige3 <- Prestige2
Prestige3["athletes","professional"] <- 1

pres_nas <- lm(prestige ~ income + professional + income:professional,
                    data = Prestige2)
pres_athletes <- lm(prestige ~ income + professional + income:professional,
               data = Prestige3)
```

Table 2: Effect of na job types on model

	Dependent variable:		
	prestige		
	(1)	(2)	(3)
income	0.0032***	0.0033***	0.0032***
	(0.0005)	(0.0005)	(0.0005)
professional	37.7813***	38.1200***	37.1860***
	(4.2483)	(4.0798)	(4.0920)
income:professional	-0.0023^{***}	-0.0024^{***}	-0.0023^{***}
	(0.0006)	(0.0005)	(0.0005)
Constant	21.1423***	20.8035***	21.0533***
	(2.8044)	(2.5387)	(2.5748)
Observations	98	102	102
R^2	0.7872	0.7893	0.7863
Adjusted R^2	0.7804	0.7828	0.7797
Residual Std. Error	8.0116 (df = 94)	8.0181 (df = 98)	8.0747 (df = 98)
F Statistic	115.8778*** (df = 3; 94)	122.3380*** (df = 3; 98)	120.1705*** (df = 3; 98)
Note:			*p<0.1; **p<0.05; ***p<0.01

Q2

A full F-test was used to evaluate hypothesis that all the coefficients are 0; at 5% reject null hypothesis.

```
    F.test <-(R2/(k-1))/((1-R2)/(n-k))
    #F test statistic F = 6.640 with 2 and 129 degrees of freedom
     df1 <- k - 1
     df2 <- n-k

    F.pvalue <-df(F.test, df1, df2)
    # 0.001634304

```