Claire Ott
20302999
Applied Statistics I
Problem Set 4

Question 1

a.

```
# creating a new variable "professional" from "type" so that professionals = 1 and other (blue/white collar) = 0
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)
# checking to see if it worked
print(Prestige)
```

Example screenshot of output:

```
> print(Prestige)
                          education income women prestige census type professional
gov.administrators            13.11  12351 11.16     68.8   1113 prof            1
general.managers              12.26  25879  4.02     69.1   1130 prof            1
accountants                   12.77   9271 15.70     63.4   1171 prof            1
purchasing.officers           11.42   8865  9.11     56.8   1175 prof            1
chemists                      14.62   8403 11.68     73.5   2111 prof            1
physicists                    15.64  11030  5.13     77.6   2113 prof            1
biologists                    15.09   8258 25.65     72.6   2133 prof            1
architects                    15.44  14163  2.69     78.1   2141 prof            1
civil.engineers               14.52  11377  1.03     73.1   2143 prof            1
mining.engineers              14.64  11023  0.94     68.8   2153 prof            1
surveyors                     12.39   5902  1.91     62.0   2161 prof            1
draughtsmen                   12.30   7059  7.83     60.0   2163 prof            1
computer.programers           13.83   8425 15.33     53.8   2183 prof            1
economists                    14.44   8049 57.31     62.2   2311 prof            1
psychologists                 14.36   7405 48.28     74.9   2315 prof            1
social.workers                14.21   6336 54.77     55.1   2331 prof            1
lawyers                       15.77  19263  5.13     82.3   2343 prof            1
librarians                    14.15   6112 77.10     58.1   2351 prof            1
vocational.counsellors        15.22   9593 34.89     58.3   2391 prof            1
ministers                     14.50   4686  4.14     72.8   2511 prof            1
university.teachers           15.97  12480 19.59     84.6   2711 prof            1
primary.school.teachers       13.62   5648 83.78     59.6   2731 prof            1
secondary.school.teachers     15.08   8034 46.80     66.1   2733 prof            1
physicians                    15.96  25308 10.56     87.2   3111 prof            1
veterinarians                 15.94  14558  4.32     66.7   3115 prof            1
osteopaths.chiropractors      14.71  17498  6.91     68.4   3117 prof            1
nurses                        12.46   4614 96.12     64.7   3131 prof            1
nursing.aides                  9.45   3485 76.14     34.9   3135   bc            0
physio.therapsts              13.62   5092 82.66     72.1   3137 prof            1
```

b. Running a linear model where prestige is the outcome variable and income and professional, and the interaction between the two variables, are the predictors

```r
# running a linear model with interaction
interact_reg <- lm(prestige ~ income * professional, data=Prestige)
summary(interact_reg)
```

```
> summary(interact_reg)

Call:
lm(formula = prestige ~ 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 ***
professional        37.7812800  4.2482744   8.893 4.14e-14 ***
income:professional -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
```

c. Writing the prediction equation
   i. $\hat{y}$ = B0 + B1(income) + B2(professional) + B3(income x professional)
   ii. Prestige = 21.14266 + 0.003170909 x income + 37.78128 x professional - 0.002325709 x (income x professional)

d. Interpreting the coefficient for income
   i. The coefficient for income, which for this dataset is measured in (Canadian) dollars, measures the increase in prestige for every dollar earned.

e. Interpreting the coefficient for professional
   i. The coefficient for professional represents the effect of being a professional (professional = 1) versus a blue or white collar worker (professional = 0) when income is 0.

f. Calculating the change in $\hat{y}$ (prestige score) when there's a $1,000 increase in income and professional = 1
   i. Prestige = 21.14266 + 0.003170909 x income + 37.78128 - 0.002325709 x income
   ii. Solve for income = 1,000
   iii. Prestige = 21.14266 + 0.003170909 x 1,000 + 37.78128 - 0.002325709 x 1,000
   iv. Prestige = 59.76914
   v. Prestige score for a professional occupation will increase by approximately 59.8 when there is a $1,000 increase in income.

g. Calculating the effect of changing professions from non-professional to professional when income is $6,000
    i. To do this, I need to find the predicted values of ŷ (prestige) for both groups and then find the difference between them.
    ii. For non-professional (professional = 0)
        1. Prestige = 21.14266 + 0.003170909 x income
        2. Prestige = 21.14266 + 0.003170909 x 6,000
        3. Prestige = 40.168144
    iii. For professional (professional = 1)
        1. Prestige = 21.14266 + 0.003170909 x income + 37.78128 - 0.002325709 x income
        2. Prestige = 21.14266 + 0.003170909 x 6,000 + 37.78128 - 0.002325709 x 6,000
        3. Prestige = 63.99514
    iv. Calculating the change in ŷ
        1. $\Delta\hat{y} = \text{prestige}_{\text{professional}} - \text{prestige}_{\text{non-professional}}$
        2. $\Delta\hat{y} = 63.99514 - 40.168144$
        3. $\Delta\hat{y} = 23.826996$
    v. When income is $6,000, changing professions from non-professional to professional affects prestige, or ŷ, by 23.826996.

Question 2

a. Determining if the yard signs have an effect on the vote share:
    i. Coefficient for precincts with signs: $B_1 = 0.042$
    ii. SE (standard error) for that coefficient: SE = 0.016
    iii. H_0 = yard signs have no effect on vote share ($B_1 = 0$)
    iv. H_a= yard signs have an effect on vote share ($B_1 \neq 0$)
    v. Getting t-statistic
        1. T = coefficient / SE → 0.042/0.016 = 2.625 = t
    vi. Getting degrees of freedom
        1. df = N - k → 131 - 2 - 1 = 128
    vii. At a= 0.05, critical value for a 2-tailed test is approximately ± 1.98
        1. Since |t| = 2.625 > 1.98, we reject the null hypothesis
    viii. Therefore, we can conclude that having the yard signs in a precinct affects the vote share.
b. Determining whether being next to a precinct with the yard signs affects vote share:
    i. Coefficient for precincts adjacent to signs: $B_2 = 0.042$
    ii. SE (standard error) for that coefficient: SE = 0.013
    iii. H_0 = being adjacent to yard signs has no effect on vote share ($B_2 = 0$)
    iv. H_a= being adjacent to yard signs has an effect on vote share ($B_2 \neq 0$)
    v. Getting t-statistic
        1. T = coefficient / SE → 0.042/0.013 = 3.231 = t
    vi. Getting degrees of freedom
        1. df = N - k → 131 - 2 - 1 = 128
    vii. At a= 0.05, critical value for a 2-tailed test is approximately ± 1.98
        1. Since |t| = 3.231 > 1.98, we reject the null hypothesis
    viii. Therefore, we can conclude that being adjacent to a precinct with the yard signs affects the vote share.

c. Interpreting the coefficient for the constant term:
    i. The coefficient for the constant term, B_0, is 0.302. This constant represents the expected vote share for Ken Cuccinelli in precincts without yard signs and that are not adjacent to yard signs. This baseline corresponds to 30.2% of the vote share.
d. Evaluating the model fit:
    i. R^2 is 0.094. This means that the model only explains 9.4% of the variation in vote share, suggesting that there are other unmodeled factors, such as demographic and/or voter preferences maybe, that have a more significant role than the presence or proximity to yard signs in determining the vote share in different precincts. In other words, while the findings above suggest that the yard signs are statistically significant to the vote share, they are not the most influential factor.