STAR 513: MidTerm Exam

Spring 2025

Honor Pledge: I have worked independently on this exam. I have read the exam instructions. I have not given, received, or used any unauthorized assistance on this exam.
Yvette Uwineza

Instructions:

- This exam is due by Friday 3/14 at midnight.
- **Students are required to work independently on the exam.** Do NOT discuss the exam with anyone else (including other students).
- You may use the textbook, class notes, examples, HW solutions posted in the current Canvas course. You may use any other publicly available (print or online) statistics references or resources that you find helpful. Use of homework "helper" websites (ex: Chegg, NoteHall, etc) is NOT allowed. Use of chatbots (ex: ChatGPT) is NOT allowed.
- Knit frequently to avoid last minute problems. You may add or delete code chunks as needed. **It is the student's responsibility to check the knitted document (for correctness and completeness) before submitting.**
- For any questions that require calculations, you should provide R code for full credit.
- For some questions, there may be more than one possible answer, analysis or graph that could be used for full credit. **Choose one approach**, making a reasonable choice and justifying if needed.
- Given this is the final exam, you should present your best work. I will deduct points for things like printing full data to knitted document, unreadable tables, unclear, excess or unnecessary output, etc.
- Use $\alpha = 0.05$ and/or 95% confidence where needed.
- All questions are worth 4 points except where noted. Maximum score is 88.
- **I believe all students can do well on this exam. Please don't cheat!!!**

This exam will use Prestige Data available from Canvas as PrestigeData.csv. The data includes information for n = 95 occupations. The following variables are included:

- **prestige**: Pineo-Porter prestige score for occupation (no units).
- **education**: education in years.
- **income**: income in dollars ($)
- **type**: bc = blue collar, prof = professional, wc = white collar. Generally, blue-collar jobs involve manual labor, while white-collar jobs are more administrative or managerial.

Notes:

- **prestige** is the response variable (y) for all analyses.
- We will consider a few models with different predictors (x's).
- This data is from 1971. Hence the low education and income values.
- The education and income values are averages corresponding to a given occupation.
- When importing the data, it MAY be helpful to use code similar what is provided below in order to have the occupation names appear in some graphs. This is not required.

Summary Stats (Q1 - Q6)

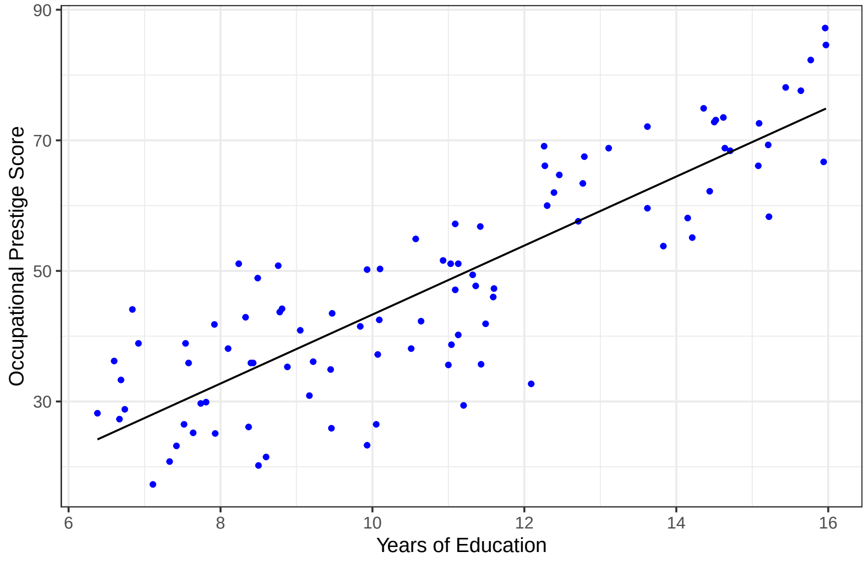
For this group of questions we will create and consider summary graphs and summary statistics.

Q1

Create a scatterplot of prestige vs **education** with fitted regression line overlaid. Your plot should include axis labels with units (where applicable). Briefly comment on one thing you **learn** from this plot (1-2 sentences).

Comment: Looking at the plot generated below, we see a strong positive relationship between years of education and occupational prestige score.

Prestige vs Education Plot

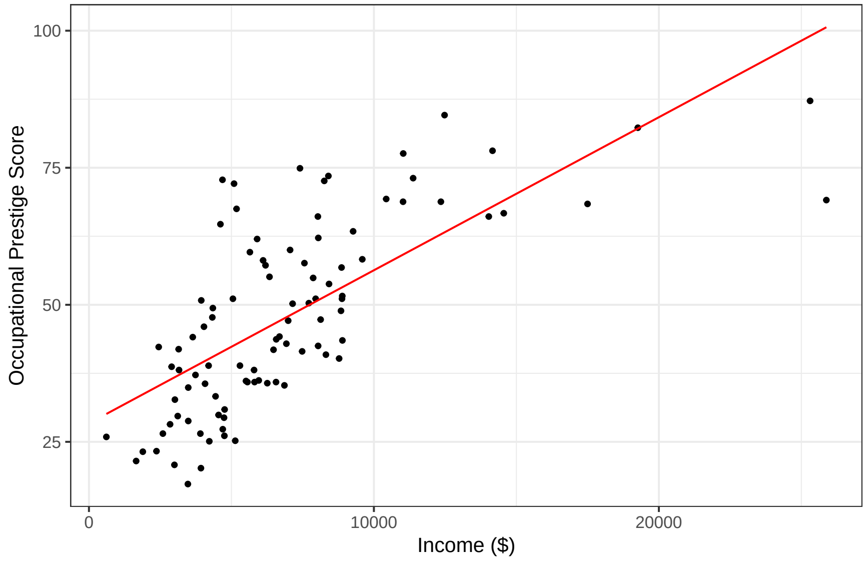


Q2

Create a scatterplot of prestige vs **income** with fitted regression line overlaid. Your plot should include axis labels with units (where applicable). Briefly comment on one **concern** you have after considering this plot (1-2 sentences).

Concern: One concern as I look at the graph, it is how the spread of prestige scores increases as income rises. The prestige scores vary widely as income levels are higher.

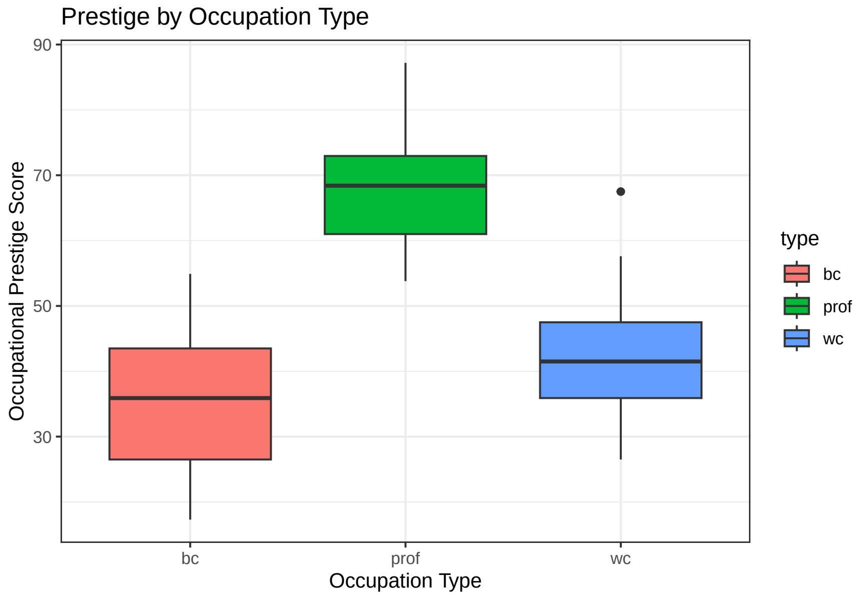
Prestige vs Income Plot



Q3

Create a boxplot of prestige vs **type**. Your plot should include axis labels with units (if applicable). Which type has the highest mean (or median) prestige?

Type Prof occupation has the highest mean prestige.



Q4 (2 pts)

Calculate pairwise (Pearson) correlations between all numeric variables (prestige, education, income). This should be presented as a 3 ×3 matrix.

```
##            prestige education    income
## prestige  1.0000000 0.8554382 0.7040783
## education 0.8554382 1.0000000 0.5706663
## income    0.7040783 0.5706663 1.0000000
```

Q5 (6 pts)

For each numeric variables (prestige, education, income), calculate the min, mean, max values. A table of results is preferred, but not required for full credit.

Variables	Min	Mean	Max
Prestige	17.30	47.77474	87.20
Education	6.38	10.84379	15.97
Income	611.00	6947.55789	25879.00

Q6

What occupation has the highest prestige? Briefly discuss if this is surprising or expected, based on the values of education, income and type (1-2 sentences).

```
##             prestige education income type
## physicians     87.2     15.96  25308 prof
```

Occupation: Physicians
Discussion: This is expected because physicians require so many years of education and in the end earn a high income. It also makes sense that they hold the highest occupational prestige.

Model 1 (Q7 - Q15)

For this group of questions we will focus on a model of prestige (y) versus education (x1) and income (x2).

Q7

Fit the model described above and show the coefficients table.

```
##
## Call:
## lm(formula = prestige ~ education + income, data = PrestigeData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.054  -4.936  -0.313   5.093  17.158
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.1056284  3.2717509  -1.866   0.0652 .
## education    4.1553895  0.3499237  11.875  < 2e-16 ***
## income       0.0012695  0.0002246   5.652 1.77e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.732 on 92 degrees of freedom
## Multiple R-squared:  0.8009, Adjusted R-squared:  0.7966
## F-statistic:   185 on 2 and 92 DF,  p-value: < 2.2e-16
```

Q8

Using this model, provide a detailed interpretation of the estimated coefficient corresponding to **education** in context of this study. Your interpretation should include appropriate units (where applicable) and the numeric value for the estimated coefficient.

A one year increase in education is associated with approximately 4.16 units increase in occupational prestige score, holding other factors constant.

Q9

Using this model, do we have evidence of a linear association between prestige and **education**? Provide a brief response in context including direction and appropriate statistical evidence (ex: test statistic and p-value or confidence interval).

Yes, we do have evidence of a positive linear association between prestige and education. This association is statistically significant, with a t value of 11.88 and a p value less than 0.001.

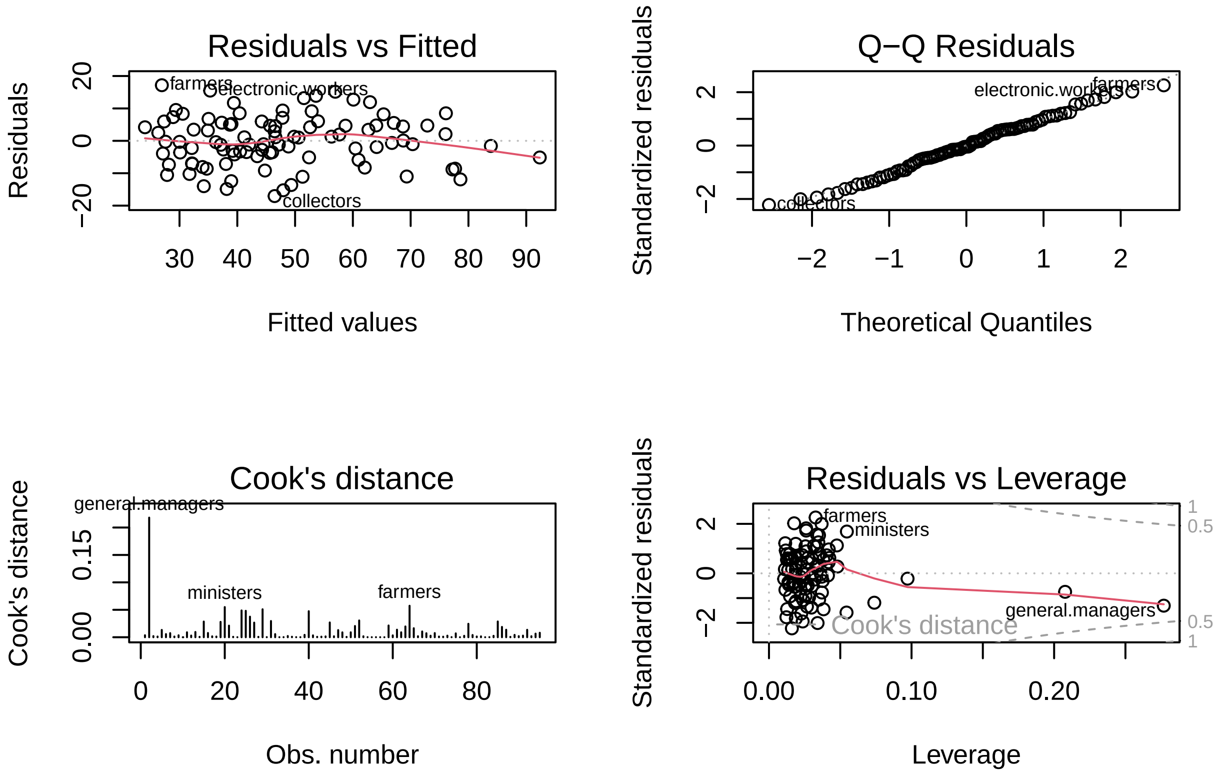
Q10 (2 pts)

What proportion of variation in prestige is explained by the linear regression including education and income? Give your answer to 3 decimal places.

The variation in prestige that is explained by the linear regression is 0.801. This means that 80.1% of the variation in prestige are explained by education and income.

Q11

Provide diagnostic plots: (1) Residuals vs Fitted, (2) QQplot of Residuals, (4) Cook's Distance and (5) Residuals vs Leverage.



Q12 (8 pts)

State the four assumptions of linear regression (in words) and briefly discuss whether each assumption is satisfied. Where applicable, your discussion should mention evidence from a specific plot (graph) from the previous question.

Assumption 1: Independence Since each observation (row) corresponds to a different occupation, this assumption is satisfied

Assumption 2: Equal Variance we see a football shape and a few outliers. So, we can say that there is no major concern and that the assumption is satisfied. We use the Residuals vs Fitted plot for this.

Assumption 3: Linearity or Model Fits, based on the residuals vs fitted plot, we can see some curvature in the red line, meaning that the linearity assumption is slightly but not entirely violated. So, it's not a major concern

Assumption 4: Normality of residuals: the Q-Q plot shows that the residuals are normal and they fall near the qqline. Hence this assumption is approximately satisfied.

Q13

There are TWO occupations with notably large (positive or negative) residuals. (These two occupations are nearly tied for the magnitude of their residuals.) Identify ONE of these occupations and provide the corresponding standardized residual, observed prestige (y) and fitted prestige ($\hat{y}$).

Below are the two occupations, we also see them with our Q-Q residuals plot above:

```
## Occupation: farmers
## Standardized Residual: 2.256
## Observed Prestige (y): 44.1
## Fitted Prestige (ŷ): 26.942
##
## Occupation: collectors
## Standardized Residual: -2.224
## Observed Prestige (y): 29.4
## Fitted Prestige (ŷ): 46.454
```

Q14

Identify the occupation with the highest **leverage**. Briefly discuss whether this observation is **influential**, referencing a specific diagnostic.

Occupation: General Managers
Discussion: Looking at the Residuals vs Leverage graph, general managers is at the extreme right, meaning it has the highest leverage. Looking at the cooks distance curve, we can see that general managers has the highest point, so we can say that it highly influential

---

Q15

Do you have concerns about collinearity for this model? Briefly discuss referencing a specific diagnostic.

Response

```
## education    income
##  1.482931  1.482931
```

Looking at our vif that are way below 10, we don't have any concern of high correlation. Each variable contributes independently to prestige prediction.

Model 2 (Q16 - Q22)

For this group of questions we will focus on a model of prestige (y) versus education (x1) and income (x2) and type (x3).

Q16

Fit the model described above and show the coefficients table.

```
##
## Call:
## lm(formula = prestige ~ education + income + type, data = PrestigeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0096  -4.8970   0.0093   5.1551  19.0443
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7108032  5.4528060   0.314    0.754
## education    3.4383155  0.6700592   5.131 1.64e-06 ***
## income       0.0010364  0.0002257   4.592 1.42e-05 ***
## typeprof     6.7675597  4.1000439   1.651    0.102
## typewc      -2.5998967  2.6554738  -0.979    0.330
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.336 on 90 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.8168
## F-statistic: 105.8 on 4 and 90 DF,  p-value: < 2.2e-16
```

Q17

A colleague looks at your coefficient table and asks (A) Type is a single predictor. Why are there 2 coefficients corresponding to type? (B) Why is there no coefficient corresponding to type bc? Briefly respond to each of these questions.

A: This is because type is categorical variable, so the other variable is used a reference, and we compare our results to it.

B: Given that bc is a reference category, the intercept would be the value of bc when prof and wc types are zero.

Q18

Using this model, provide a detailed interpretation of the estimated coefficient corresponding to **type prof** in context of this study. Include the numeric value for the estimated coefficient.

The estimated coefficient for type prof is 6.77. Holding other factors constant, professional occupations have an estimated prestige score that is 6.77 points higher on average compared to blue collar occupations.

Q19

Using this model, test for a difference between the mean Prestige comparing between levels of **type**. (In other words, test for a type effect using this model.) Provide an appropriate test statistic, p-value and conclusion in context.

Test statistic: F-value: 6.0895
p-value: 0.0033
Conclusion in context: Give our pvalue that is less than 0.05, we can conclude that there is strong evidence that occupation type has a statistically significant effect on prestige.

```
## Analysis of Variance Table
##
## Response: prestige
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## education  1 20214.0 20214.0 375.5567 < 2.2e-16 ***
## income     1  1909.6  1909.6  35.4780 4.921e-08 ***
## type       2   655.5   327.8   6.0895  0.003309 **
## Residuals 90  4844.2    53.8
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Q20

Using this model, calculate the predicted mean prestige for each type at the (overall) mean education and (overall) mean income. Also give the corresponding confidence intervals.
Notes: You calculated mean education and mean income in Q5. Some example starter code is provided but not required.

```
##   education   income type      fit      lwr      upr
## 1  10.84379 6947.558   bc 46.19582 42.22423 50.16742
## 2  10.84379 6947.558 prof 52.96338 48.06591 57.86085
## 3  10.84379 6947.558   wc 43.59593 40.41271 46.77914
```

The fit column shows our predicted mean prestige for each type

---

Q21

Using this model, provide emmeans and (unadjusted) pairwise comparisons for type.

```
##  type emmean   SE df lower.CL upper.CL
##  bc     46.2 2.00 90     42.2     50.2
##  prof   53.0 2.47 90     48.1     57.9
##  wc     43.6 1.60 90     40.4     46.8
##
## Confidence level used: 0.95
```

```
##  contrast  estimate   SE df t.ratio p.value
##  bc - prof    -6.77 4.10 90  -1.651  0.1023
##  bc - wc       2.60 2.66 90   0.979  0.3302
##  prof - wc     9.37 2.89 90   3.244  0.0017
```

Q22 (2 pts)

Consider your results from the previous question. Where did the emmeans (and confidence intervals) appear previously (besides Q21)? Where did the first two contrasts (estimates and tests) appear previously (besides Q21)?

---

emmeans appeared in Q20.
contrasts appeared in Q19

Appendix

```
#Retain this code chunk!!!
library(knitr)
library(knitr)
library(knitr)
library(tinytex)
library(broom)
library(emmeans)
library(tidyverse)
library(kableExtra)
library(ggplot2)
library(car)
library(emmeans)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
PrestigeData <- read.csv("PrestigeData.csv", row.names = 1)
#Q1

ggplot(PrestigeData, aes(x = education, y = prestige)) +
  geom_point(color = "blue", size=1) +
  geom_smooth(method = "lm", color = "black", se = FALSE, linewidth=0.5) +
  labs(title = "Prestige vs Education Plot",
       x = "Years of Education",
       y = "Occupational Prestige Score") +
  theme_bw()

#Q2

ggplot(PrestigeData, aes(x = income, y = prestige)) +
  geom_point(color = "black", size =1
             ) +
  geom_smooth(method = "lm", color = "red", se = FALSE, linewidth=0.5) +
  labs(title = "Prestige vs Income Plot",
       x = "Income ($)",
       y = "Occupational Prestige Score") +
  theme_bw()


#Q3


ggplot(PrestigeData, aes(x = type, y = prestige, fill = type)) +
  geom_boxplot() +
  labs(title = "Prestige by Occupation Type",
       x = "Occupation Type",
       y = "Occupational Prestige Score") +
  theme_bw()


#Q4


cor(PrestigeData%>%
                    select(prestige, education, income), method = "pearson")

#Q5
summary_stats <- PrestigeData %>%
  summarise(
    Min = c(min(prestige), min(education), min(income)),
    Mean = c(mean(prestige), mean(education), mean(income)),
    Max = c(max(prestige), max(education), max(income))
  ) %>%
  mutate(Variables = c("Prestige", "Education", "Income")) %>%
  relocate(Variables)

kable(summary_stats)

highest_prestige <- PrestigeData %>%
  filter(prestige == max(prestige))



highest_prestige <- PrestigeData %>%
  filter(prestige == max(prestige))
highest_prestige
#Q7

model1 <- lm(prestige ~ education + income, data = PrestigeData)

summary(model1)

#Q11

par(mfrow = c(2, 2))
plot(model1, which = c(1:2,4:5))


#Q13

standardized_residuals <- rstandard(model1)

top_residual_indices <- order(abs(standardized_residuals), decreasing = TRUE)[1:2]

observed_prestige <- PrestigeData$prestige[top_residual_indices]

fitted_prestige <- fitted(model1)[top_residual_indices]


for (i in 1:2) {
  cat("Occupation:", rownames(PrestigeData)[top_residual_indices[i]], "\n")
  cat("Standardized Residual:", round(standardized_residuals[top_residual_indices[i]], 3), "\n")
  cat("Observed Prestige (y):", round(observed_prestige[i], 3), "\n")
  cat("Fitted Prestige (ŷ):", round(fitted_prestige[i], 3), "\n\n")
}

#Q15

vif(model1)

#Q16

model2 <- lm(prestige ~ education + income + type, data = PrestigeData)

summary(model2)
```

```r
#Q19

anova(model2)

#Q20
# Compute predicted mean prestige and confidence intervals
new_data <- data.frame(
  education = mean(PrestigeData$education),
  income = mean(PrestigeData$income),
  type = factor(c("bc", "prof", "wc"))
)

pred <- predict(model2, newdata = new_data, interval = "confidence")


results <- cbind(new_data, pred)
results
#Q21

value <- emmeans(model2, ~ type)
value


pairs(value, adjust = "none")
```