Assignment 2

Introduction to Applied Statistics and Data Science

Sophie Ennis & Hannah Yoon

2025-02-17

# PART I

We continue our investigation into “the American Dream” – the idea that those from humble origins can become economically successful. A classical study on social mobility by Peter Blau and Otis Dudley Duncan introduced the “status attainment model,” which examines the transmission of social status for parents to children. They postulated that the SES of fathers influenced the education and first jobs of their sons, which then influenced the SES their sons attained later in life. We will explore part of this pathway using modern data, and unlike Blau and Duncan, we will consider the outcomes for women as well. We will use the cleaned 2018 General Social Survey dataset.

# QUESTION 1

### Transform Sex into a binary variable called Female coded so respondents are coded as 1 if female and 0 otherwise.

gss$Female <- ifelse (gss$Sex=="Female", 1, 0)

## Part A

### Create a correlation matrix with OccPres, DadPres, Female, and YearsEd. Use pairwise valid correlations in the sample.

correlation\_matrix <- subset (gss, select= c(OccPres, DadPres, Female, YearsEd))   
rcorr(as.matrix(correlation\_matrix))

## OccPres DadPres Female YearsEd  
## OccPres 1.00 0.19 0.00 0.48  
## DadPres 0.19 1.00 -0.03 0.26  
## Female 0.00 -0.03 1.00 0.01  
## YearsEd 0.48 0.26 0.01 1.00  
##   
## n  
## OccPres DadPres Female YearsEd  
## OccPres 2248 1778 2248 2246  
## DadPres 1778 1842 1842 1840  
## Female 2248 1842 2348 2345  
## YearsEd 2246 1840 2345 2345  
##   
## P  
## OccPres DadPres Female YearsEd  
## OccPres 0.0000 0.9540 0.0000   
## DadPres 0.0000 0.1641 0.0000   
## Female 0.9540 0.1641 0.5663   
## YearsEd 0.0000 0.0000 0.5663

## Part B

### Describe the bivariate relationships between these variables.

The strongest bivariate relationship present among these variables is between OccPres and YearsEd, with a correlation of 0.4799. The next strongest bivariate relationships are between DadPres and YearsEd, with a correlation of 0.2614, and between OccPres and DadPres, with a correlation of 0.1921. Following that are the bivariate relationships between DadPres and Female with a correlation of -0.032, and between YearsEd and Female with a correlation of 0.0118. Finally, the weakest bivariate relationship exists between OccPres and Female, with a correlation of -0.0012.

However, even if 0.4799 (OccPres & YearsEd) is the strongest relationship among the pairs of variables we have, given the general “rule” of

* | r | < 0.3 is a weak correlation
* 0.3 < | r | < 0.7 is a moderate correlation
* | r | > 0.7 is a strong correlation

We can say that OccPres and YearsEd (correlation of 0.4799) and DadPres and YearsEd (correlation of 0.2614) has a moderate correlation. All other pairs have a correlation less than 0.3, so we can assume that they have weak correlation (e.g., OccPres and Female, with a correlation of –0.0012 has a weak correlation).

## Part C

### Identify relationships that are statistically significant.

The traditional convention in the social sciences is to use a critical value of 0.05 (α = 0.05) (Jean, Lecture Day 6, Slide 15). Thus, we selected an alpha value of 0.05. Given this specified value, the relationships between (1) OccPres and DadPres, (2) OccPres and YearsEd, and (3) DadPres and YearsEd have p-values that are below the alpha value (as shown in Question 1, Part A, using rcorr) and are thus statistically significant.

# QUESTION 2

### Run a regression model with respondents’ occupational prestige being predicted by their fathers’ occupational prestige. This will be Model 1.

model1 <- lm(OccPres ~ DadPres, data=gss)  
model1

##   
## Call:  
## lm(formula = OccPres ~ DadPres, data = gss)  
##   
## Coefficients:  
## (Intercept) DadPres   
## 36.6668 0.2032

## Part A

### Interpret the intercept, slope, and r-squared for the model.

summary(model1)

##   
## Call:  
## lm(formula = OccPres ~ DadPres, data = gss)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.045 -10.543 -0.147 9.409 36.628   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 36.66684 1.13841 32.209 < 2e-16 \*\*\*  
## DadPres 0.20319 0.02462 8.253 2.98e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.43 on 1776 degrees of freedom  
## (570 observations deleted due to missingness)  
## Multiple R-squared: 0.03693, Adjusted R-squared: 0.03639   
## F-statistic: 68.11 on 1 and 1776 DF, p-value: 2.98e-16

Intercept: When DadPres is 0, OccPres is 36.667, so the expected value of occupational prestige of individuals in the population is 36.667 when their dad's occupational prestige is 0.

Slope: For each point that DadPres increases, we expect OccPres to increase by an average of 0.203 points.

R-squared: The r-squared is 0.037, so the model explains 3.7% of the variance in OccPres. This is low, so there's likely more variables that go into explaining OccPres than just DadPres.

## Part B

### Explain in plain language what the slope means for our research question. Explain why we would consider this a naïve estimate.

Model 1 has a weak positive slope, which indicates that as the father’s occupational prestige increases, we expect the children’s occupational to increase as well (DadPres seems to have an effect on OccPres). We would consider this a naive estimate because we’re only looking at the effect of one variable on the outcome and thus can’t make determinations about the effect of DadPres on OccPres without controlling for other variables that may moderate, mediate, or confound the relationship.

# QUESTION 3

### Add Female, Race, and Age to the same regression model you ran in Question 2. This will be Model 2.

model2 <- summ(lm(OccPres ~ DadPres + Female + Race + Age, data=gss), digits = 3)   
model2

## MODEL INFO:  
## Observations: 1774 (574 missing obs. deleted)  
## Dependent Variable: OccPres  
## Type: OLS linear regression   
##   
## MODEL FIT:  
## F(5,1768) = 19.236, p = 0.000  
## R² = 0.052  
## Adj. R² = 0.049   
##   
## Standard errors:OLS  
## ---------------------------------------------------  
## Est. S.E. t val. p  
## ----------------- -------- ------- -------- -------  
## (Intercept) 33.225 1.608 20.658 0.000  
## DadPres 0.205 0.025 8.291 0.000  
## Female -0.092 0.635 -0.145 0.885  
## RaceBlack -2.714 0.984 -2.757 0.006  
## RaceOther -1.001 1.026 -0.976 0.329  
## Age 0.076 0.018 4.186 0.000  
## ---------------------------------------------------

## Part A

### Interpret each of the new coefficients in your model.

1. The coefficient/slope for the Female variable is –0.092. This is the expected change in OccPres for a one-unit change in the Female variable. That is, the mean difference between males and females while holding all other variables constant.
2. The coefficient/slope for one of the race categories, Black, is –2.714. This is the mean difference in OccPress for between White and Black Respondents (while holding all other variables constant).
3. The coefficient/slope for one of the race categories, Other, is –1.001. This is the mean difference in OccPress for between White and Other Race Respondents (while holding all other variables constant).
4. The coefficient/slope for the Age variable is 0.076. This is the expected change in OccPres for each one unit change in Age (i.e., one year increase), holding constant all other variables.

## Part B

### Did your coefficient for DadPres change from Model 1? Explain what this means.

The coefficient for DadPres changed from 0.203 in Model 1 to 0.205 in Model 2. This means that controlling for other variables (specifically sex, race, and age) slightly increases the slope between DadPres and OccPres (i.e., strength of the effect that DadPres has on OccPres).

## 

## Part C

### Calculate the estimated occupational prestige for a 30-year old black male who had a father with occupational prestige of 40. Write down the equation you used to make this prediction.

The equation we used to make this prediction was:

ŷi = β0 + β1 DadPres i + β2 Female i + β3 Black i + β4 Other i + β5 Age i

Using the coefficients identified above and the fact that we have a case for a 30-year old black male with DadPres of 40 (i.e., Female = 0 because we have a male, Black = 1, Other = 0 because they are not in the other category, and Age = 30), we plug in these values into the equation:

ŷi = 33.225 + (0.205)\*(40) + (-0.092)\*(0) + (-2.714)\*(1) + (-1.001)\*(0) + (0.076)\*(30)  
ŷi

## [1] 40.991

ŷi = 40.991

Thus, we expect the estimated occupational prestige to be a score of 40.991 for a 30-year old Black male whose father had an occupational prestige of 40.

## Part D

### Calculate the estimated occupational prestige for a 40-year old white female who had a father with occupational prestige one standard deviation above the mean. Write down the equation you used to make this prediction (you do not need to show the calculations to get the mean and SD).

The equation we used to make this prediction was:

ŷi = β0 + β1 DadPres i + β2 Female i + β3 Black i + β4 Other i + β5 Age i

A father with occupational prestige one standard deviation (which was 12.971) above the mean (which was 44.393) is 57.364. Using this, the coefficients identified above, and that we have a 40-year old white female (i.e., Female = 1, Black = 0, Other = 0 because they are not in the Black & other categories, and Age = 40), we plug in these values into the equation:

ŷi = 33.225 + (0.205)\*(57.36428) + (-0.092)\*(1) + (-2.714)\*(0) + (-1.001)\*(0) + (0.076)\*(40)   
ŷi

## [1] 47.93268

ŷi = 47.933

Thus, we expect the estimated occupational prestige to be a score of 47.933 for a 40-year old white female who had a father with occupational prestige one standard deviation above the mean.

# QUESTION 4

### Add YearsEd to the regression model you ran in Question 3. This will be Model 3.

model3 <- summ(lm(OccPres ~ DadPres + Female + Race + Age + YearsEd, data=gss), digits = 3)   
model3

## MODEL INFO:  
## Observations: 1772 (576 missing obs. deleted)  
## Dependent Variable: OccPres  
## Type: OLS linear regression   
##   
## MODEL FIT:  
## F(6,1765) = 93.974, p = 0.000  
## R² = 0.242  
## Adj. R² = 0.240   
##   
## Standard errors:OLS  
## ---------------------------------------------------  
## Est. S.E. t val. p  
## ----------------- -------- ------- -------- -------  
## (Intercept) 9.240 1.835 5.036 0.000  
## DadPres 0.081 0.023 3.557 0.000  
## Female -0.253 0.568 -0.446 0.656  
## RaceBlack -2.190 0.883 -2.481 0.013  
## RaceOther 1.130 0.923 1.224 0.221  
## Age 0.089 0.016 5.444 0.000  
## YearsEd 2.056 0.098 21.065 0.000  
## ---------------------------------------------------

## Part A

### Interpret the new coefficient.

The coefficient/slope for the new variable, YearsEd, is 2.056. This is the expected change in OccPres for each one unit change in YearsEd (i.e., one year increase in years of school completed), holding constant all other variables.

## Part B

### Did any of your other coefficients change from Model 2? Explain what these changes mean.

In consideration of a reference group of White men, below are our interpretations of how the coefficients changed between Models 2 and 3:

1. The effect of DadPres is smaller in Model 3 when education is added to the model, changing from 0.205 in Model 2 to 0.081 in Model 3. This indicates that when accounting for educational attainment, one’s father’s occupational prestige has a smaller effect on their occupational prestige.
2. Race – Black consistently shows a negative effect in both models, but the effect is more profound in Model 2 where the coefficient is –2.714, as compared to Model 3, where the coefficient is –2.190. This suggests not only that Black individuals tend to have lower occupational prestige compared to the White reference group, but also that controlling for educational attainment accounts for some of these differences.
3. Race – Other flips from negative in Model 2 (-1.001 coefficient) to positive in Model 3 (1.130 coefficient), indicating that controlling for educational attainment creates a positive effect (increase) on other races’ occupational prestige scores in comparison with the White reference group, while this effect is negative when educational attainment is not controlled for.
4. Female remains consistently negative between the models but the value becomes more negative from –0.092 to –0.253 from Model 2 to Model 3, indicating that being female has a negative effect on (may lower) occupational prestige scores in comparison with the Male reference group, and that accounting for educational attainment lowers occupational prestige even more.
5. Age has a small positive effect in both models, increasing minimally from 0.076 in Model 2 to 0.089 in Model 3, indicating that controlling for educational attainment slightly increases the effect of age on occupational prestige.

Between Models 2 and 3, the p-values for each coefficient change, but the evaluation of them does not (i.e., those significant in Model 2 remains significant in Model 3). The coefficients that are statistically significant (below the alpha value of 0.05) are DadPres, Age, and Race-Black. Those *not* statistically significant are Female and Race – Other. We can interpret this as such: the relationship between DadPres & Age & Race – Black and OccPres, as presented by the coefficients, is unlikely to be due to chance, and there may be a meaningful association between DadPres & OccPres, Age & OccPres, and Race – Black & OccPres.

## Part C

### Discuss the potential causal pathways between fathers’ occupational prestige and that of their children. How are sex, race, and educational attainment different, in terms of how they might be related to our research question (e.g. related to the association between the prestige of fathers’ and children’s careers)?

We are looking at the core relationship between the prestige of fathers’ and children’s careers and how sex, race, and educational attainment affect this relationship.

Because the female coefficient was consistently negative across our models, this indicates that being female lowers children’s occupational prestige, affecting the relationship between the prestige of their career and their father’s. This could be due to gender differences in occupation (i.e., daughters do not go into the same field/job as their fathers) and differences in perceptions of occupations (i.e., feminized occupations may have lower occupational prestige). However, the female coefficient is not significant from our evaluation of the p-value, thus the effect of sex on the children’s career’s occupational prestige may not be profound.

The Race – Black variable has a consistently negative coefficient across our models, indicating that being in this racial category lowers children’s occupational prestige, affecting the relationship between the prestige of their career and their father’s. This could be due to racial inequality in obtaining and advancing within jobs (i.e., hiring/promotion discrimination, etc.). The p-value for this coefficient was statistically significant, thus Race – Black may have a profound effect on children’s occupational prestige.

The Race – Other variable has coefficients that “flip” across our models (negative to positive coefficient values from Model 2 to 3), indicating that the effect of being in this racial category on children’s occupational prestige depends on the consideration of the educational attainment variable. Since the “Other” category encompasses a variety of races, considering the educational attainment variable may affect the outcome variable. For instance, being Asian or Hispanic may influence children’s occupational prestige differently, and this might have a correlation with variations in educational attainment that resulted in the coefficient’s sign change.

The positive and significant coefficient for the YearsEd variable indicates that educational attainment has a profound effect on children’s occupational prestige. Their fathers’ occupational prestige may have an influence on their children’s career via an educational pathway (e.g., tuition support, legacies, etc.).

## Part D

### Compare the r-squared between Model 3 and Model 1. Explain what this difference means.

We found the r-squared for Model 1 from Question 2a, which was 0.037. The r-squared for Model 3 comes from Question 4 (earlier, when we run model3), which was 0.242. We can estimate how well our model fits the data, and this can be measured with the r2 (r-squared). r2 also tells us how much of the variance the data explains. Because the r-squared for Model 3 is 0.242, and the r-squared for Model 1 is 0.037, Model 3 explains 24.2% of the variance, while Model 1 only explains 3.69% of the variance. Since the r-squared for Model 3 is greater than Model 1 (0.242 > 0.037), this indicates that Model 3 fits the data better than Model 1.

# QUESTION 5

### Summarize in plain language what you learned about our research question.

Consistent with Peter Blau and Otis Dudley Duncan’s classical study on social mobility, our statistical analysis of the cleaned 2018 General Social Survey dataset indicates that the effect of father’s occupational prestige on their children’s occupational prestige remains significant, however there are other profound factors at play here. For instance, being female or being Black may lower occupational prestige when we compare these variables to our reference group of White men (hiring discrimination via sexism and racism may prevent social mobility). Increased educational attainment and age, on the other hand, increase occupational prestige (when an individual is more educated or older, they may gain the skills/degrees required to attain more prestigious jobs). In other words, even if one’s father’s occupational prestige is high, certain demographic characteristics like being female and black will lower one’s own occupational prestige, while other characteristics like educational attainment will increase it. Thus, we know from our analysis that when one would like to determine the effect of a predictor variable on an outcome variable, they need to account for other factors before they can conclude that there is a profound relationship.

# PART II

We are providing you data from a medical trial that included study of how biomarkers are predictive of blood glucose levels in diabetic patients. In 254 subjects, a panel of baseline biomarkers were measured and blood glucose readings were taken 6 hours later after a controlled diet. We will use the data on Canvas titled “glucose.rdata”. The outcome measure is change in glucose, and the predictor variables are measures of patient characteristics and biomarkers.

# QUESTION 6

### Use multiple regression to determine if any of the biomarkers studied are associated with change in blood glucose levels. For the purposes of this assignment, you should include all predictor variables in a single model (omit patient ID numbers).

glucose\_regression <- lm(delta\_glucose ~ male + race + ldl + hdl + trig + pnh1 + crot + ffactor + cvh1 + cvh2 + cvh5 + crp + afp + distalphl + cchannel + adsb + ark, data=glucose)   
glucose\_regression

##   
## Call:  
## lm(formula = delta\_glucose ~ male + race + ldl + hdl + trig +   
## pnh1 + crot + ffactor + cvh1 + cvh2 + cvh5 + crp + afp +   
## distalphl + cchannel + adsb + ark, data = glucose)  
##   
## Coefficients:  
## (Intercept) male raceb raceh racew ldl   
## 21.605976 -2.144267 -0.948026 -1.143877 -0.795120 -0.002033   
## hdl trig pnh1 crot ffactor cvh1   
## -0.009079 -0.007659 -0.006830 -0.069625 0.001811 -0.003620   
## cvh2 cvh5 crp afp distalphl cchannel   
## 0.001947 0.023146 -0.165532 0.043491 0.071188 0.002855   
## adsb ark   
## 0.013970 -0.023433

## PART A

### Identify any statistically significant predictors at conventional levels of confidence.

summary(glucose\_regression)

##   
## Call:  
## lm(formula = delta\_glucose ~ male + race + ldl + hdl + trig +   
## pnh1 + crot + ffactor + cvh1 + cvh2 + cvh5 + crp + afp +   
## distalphl + cchannel + adsb + ark, data = glucose)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.0386 -3.9141 -0.4959 3.4818 19.6045   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.605976 7.724246 2.797 0.00558 \*\*  
## male -2.144267 0.824065 -2.602 0.00986 \*\*  
## raceb -0.948026 1.873743 -0.506 0.61337   
## raceh -1.143877 1.701179 -0.672 0.50199   
## racew -0.795120 1.506991 -0.528 0.59826   
## ldl -0.002033 0.009256 -0.220 0.82636   
## hdl -0.009079 0.018938 -0.479 0.63209   
## trig -0.007659 0.010879 -0.704 0.48212   
## pnh1 -0.006830 0.013348 -0.512 0.60937   
## crot -0.069625 0.066240 -1.051 0.29430   
## ffactor 0.001811 0.013695 0.132 0.89491   
## cvh1 -0.003620 0.006765 -0.535 0.59311   
## cvh2 0.001947 0.008236 0.236 0.81336   
## cvh5 0.023146 0.008754 2.644 0.00875 \*\*  
## crp -0.165532 1.720991 -0.096 0.92346   
## afp 0.043491 0.110005 0.395 0.69294   
## distalphl 0.071188 0.025356 2.808 0.00541 \*\*  
## cchannel 0.002855 0.006190 0.461 0.64503   
## adsb 0.013970 0.011796 1.184 0.23748   
## ark -0.023433 0.018522 -1.265 0.20709   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.246 on 234 degrees of freedom  
## Multiple R-squared: 0.1086, Adjusted R-squared: 0.03617   
## F-statistic: 1.5 on 19 and 234 DF, p-value: 0.08638

Given that the conventional level of confidence is 0.05, the predictors that are statistically significant have p-values less than 0.05, which include male (p-value of 0.00986), cvh5 (p-value of 0.00875), and distalphl (p-value of 0.00541).

## PART B

### Interpret the r-squared and adjusted r-squared.

The r-squared is 0.109, so the model explains 10.9% of the variance in change in glucose (the outcome measure). R-squared represents the proportion of variance explained by the model – in multiple regression, it can be interpreted in the same way as with simple regression (Jean, Lecture Day 12, Slide 5). Essentially, the r-squared indicates how well the model fits the data, with values ranging from 0 to 1 (higher r-squared values signify a better fit). Since the r-squared value is closer to zero, we may interpret the model as not fitting the data very closely.

Since none of our predictors are completely uncorrelated by the outcome (e.g., none has a point estimate of β that is exactly 0), they each “explain” some of the variance in the outcome. Yet, this can be due to chance variation, and the adjusted r-squared takes this into account to give us a more conservative estimate of how well the predictors explain variance in the outcome in the population. (The adjusted r-squared adjusts for the number of predictors by adding a penalty for each variable added to the model.) Considering this, the adjusted r-squared is 0.036. Thus, this decrease from the r-squared of 0.109 to the adjusted r-squared of 0.036 indicates that the additional variables do not explain the variance in the outcome well.