Assignment 3 Solutions

Background

The data for this exercise were used in Ebonya Washington's paper: "Female Socialization: How Daughters Affect Their Legislator Fathers' Voting on Women's Issues." published in the American Economic Review in 2008. The paper asks whether having daughters influences the voting behavior of members of the US Congress. The hypothesis is that having (more) daughters makes legislators more likely to vote liberally (in terms of political alignment, and in contrast to conservatively) on issues concerning women.

For this exercise, we will focus on votes that took place in the 108th Congress, which held session in 2003/04. As a measure of a liberal voting record, we use scores assigned by the American Association of University Women (AAUW), a liberal group that concerns itself with issues of interest to women. For the 108th Congress, the AAUW selected 9 pieces of legislation in the areas of education, equality and reproductive rights. The AAUW then assigned a score to each member of Congress. The scores range from 0 to 100 and measure the percentage of times the legislator voted in favor of the position held by the AAUW.

The dataset `legislators.dta` contains the following characteristics for a random sample of 386 members of the 108th Congress:

- $ngirls$ number of daughters
- $totchi$ total number of children
- age age
- $female$ indicator for being female
- $repub$ indicator for being a Republican
- $moredef$ proportion of people in the legislator's district who are in favor of "more spending on defense"
- $aauw$ AAUW score

(For the purposes of this exercise, you can assume all members of the 108th Congress were either Democrats or Republicans and were either male or female.)

(a) Estimate and report results for the following regression models:

Load in the data set `legislators.dta`. Remember, you will first need to call the `haven` package to do so.

Generate a variable `ngirls2` $= \text{ngirls}^2$

Generate an interaction variable `repubngirls` $= \text{repub} * \text{ngirls}$

Generate an interaction variable `repubngirls2` $= \text{repub} * \text{ngirls2}$

Estimate the following three regression models, save the output as reg1, reg2, and reg3, and show the results of each using `summary()`:

$$aauw = \beta_0 + \beta_1 female + \beta_2 repub + \beta_3 ngirls + u \quad (1)$$
$$aauw = \beta_0 + \beta_1 female + \beta_2 repub + \beta_3 ngirls + \beta_4 ngirls2 + u \quad (2)$$
$$aauw = \beta_0 + \beta_1 female + \beta_2 repub + \beta_3 ngirls + \beta_4 ngirls2 + \beta_5 repubngirls + \beta_6 repubngirls2 + \beta_7 totchi + \beta_8 moredef + u \quad (3)$$

(Note: this method of generating interaction variables (multiplying them together) is appropriate when one of the interacted variables is a dummy variable, but may not be appropriate in all cases.)

In [1]:
```
# Add Code for part (a) here.
library(haven)
library(tidyverse)
options(warn=-1)

df <- read_dta("legislators.dta")
head(df)

#Create new variables
df <- mutate(df, ngirls2 = ngirls^2, repubngirl = repub*ngirls, repubngirls2 = repub*ngirls2)

#Estimate the three regressions
reg1 <- lm(aauw ~ female + repub + ngirls, data = df)
summary(reg1)

reg2 <- lm(aauw ~ female + repub + ngirls + ngirls2, data = df)
summary(reg2)

reg3 <- lm(aauw ~ female + repub + ngirls + ngirls2 + repubngirl + repubngirls2 + totchi + moredef, dat
a = df)
summary(reg3)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

✔ ggplot2 3.3.5     ✔ purrr   0.3.4
✔ tibble  3.1.3     ✔ dplyr   1.0.7
✔ tidyr   1.1.3     ✔ stringr 1.4.0
✔ readr   2.0.1     ✔ forcats 0.5.1

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

A tibble: 6 × 7

ngirls	totchi	repub	female	age	moredef	aauw
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	3	0	0	60	17.09234	75
1	1	1	0	37	31.40097	0
2	6	1	0	55	23.44828	0
2	2	0	0	45	16.47510	100
2	4	0	0	55	23.11688	100
2	5	1	0	55	31.40097	0

```
Call:
lm(formula = aauw ~ female + repub + ngirls, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-86.215  -6.668  -5.976  13.439  56.024 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  86.5608     1.6251  53.266  < 2e-16 ***
female       11.4167     2.8473   4.010 7.31e-05 ***
repub       -79.5468     1.7993 -44.210  < 2e-16 ***
ngirls       -0.3460     0.7894  -0.438    0.661    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.4 on 382 degrees of freedom
Multiple R-squared:  0.8449,	Adjusted R-squared:  0.8437 
F-statistic: 693.9 on 3 and 382 DF,  p-value: < 2.2e-16

Call:
lm(formula = aauw ~ female + repub + ngirls + ngirls2, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-86.382  -6.868  -6.233  13.618  55.767 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  85.9703     1.8061  47.600  < 2e-16 ***
female       11.6313     2.8632   4.062  5.9e-05 ***
repub       -79.5145     1.8009 -44.154  < 2e-16 ***
ngirls        0.6547     1.5490   0.423    0.673    
ngirls2      -0.2430     0.3235  -0.751    0.453    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.41 on 381 degrees of freedom
Multiple R-squared:  0.8452,	Adjusted R-squared:  0.8436 
F-statistic:   520 on 4 and 381 DF,  p-value: < 2.2e-16

Call:
lm(formula = aauw ~ female + repub + ngirls + ngirls2 + repubngirl + 
    repubngirls2 + totchi + moredef, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-85.436  -7.964  -1.367  11.292  54.591 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   95.5997     3.6777  25.995  < 2e-16 ***
female        11.6079     2.8334   4.097 5.13e-05 ***
repub        -79.4364     3.0424 -26.110  < 2e-16 ***
ngirls         0.4452     3.1682   0.141   0.8883    
ngirls2        0.5286     0.8568   0.617   0.5376    
repubngirl     2.1281     3.6217   0.588   0.5571    
repubngirls2  -0.7477     0.9302  -0.804   0.4220    
totchi        -2.0364     0.8066  -2.525   0.0120 *  
moredef       -0.3166     0.1247  -2.540   0.0115 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.2 on 377 degrees of freedom
Multiple R-squared:  0.8505,	Adjusted R-squared:  0.8474 
F-statistic: 268.2 on 8 and 377 DF,  p-value: < 2.2e-16
```

(b) Suggest which model is the best fit to the data. How did you determine this? (no more than 1 sentence is required)

The third model seems to be the best fit for the data, as the adjusted R^2 is the highest of the three models.

(c) Interpret the marginal effect at the mean of the number of daughters on AAUW score in each model.

(Hint: Calculate the total marginal effect, using the coefficients for all terms including the number of daughters. Calculating a marginal effect at the mean involves plugging in the mean number of daughters in the sample into any estimate of the total marginal effect of the number of daughters where this effect varies by the number of daughters.)

(Hint: Instead of typing in numbers from the regressions manually, you can call regression coefficients using `summary(reg1)$coefficients[#,1]` for coefficient number # starting with the intercept as number 1.)

(Hint: In model 3, the marginal effect will differ by particular subgroups. Interpret the different effects for each subgroup.)

In [2]:
```
mean_ng <- mean(df$ngirls, na.rm=T)

#Calculate the marginal effect for model 1:
summary(reg1)$coefficients[4,1]

#Calculate the marginal effect for model 2 at the mean:
summary(reg2)$coefficients[4,1]+ 2*summary(reg2)$coefficients[5,1]*mean_ng

#Calculate the marginal effect for model 2 at the mean for republicans/dems:
summary(reg3)$coefficients[4,1]+ 2*summary(reg3)$coefficients[5,1]*mean_ng #dems
summary(reg3)$coefficients[4,1]+ 2*summary(reg3)$coefficients[5,1]*mean_ng +
    summary(reg3)$coefficients[6,1]+ 2*summary(reg3)$coefficients[7,1]*mean_ng #repubs
```

-0.346022390451581

0.0643269405637884

1.72973734506802

2.04094391771639

Model 1: Each additional daughter of a legislator *reduces* their AAUW score by 0.346 points, holding the gender of the legislator and whether the legislator is a republican constant.

Model 2: Each additional daughter of a legislator changes their AAUW score by 0.6547 - 2 × 0.243 × $ngirls$ holding the gender of the legislator and the party affiliation of the legislator constant. Evaluating this at the mean number of daughters, this says that each additional daughter *increases* the AAUW score by 0.064 points.

Model 3: Holding constant the legislator's gender, party affiliation, total number of children, and the proportion of people in the legislator's district who are in favor of "more spending on defense", each additional daughter of a legislator changes their AAUW score by $.45 + 2 \times .53 \times\ ngirls + 2.12 \times repub - 2 \times .75 \times repub \times ngirls$. At the mean of $ngirls$, this corresponds to a marginal effect of 1.72 points for Democrats and 2.04 for Republicans.

(d) Test whether there is an effect of the number of daughters on AAUW scores using the second model. Be sure to describe carefully the null and alternative hypothesis.

(Hint: You can access the residuals from a regression you have saved as reg by calling `reg$residuals`, and you can access the r-squared by calling `summary(reg)$r.squared`.)

Using model 2, there are two terms in the regression that include the number of daughters. Testing whether this has an effect on AAUW scores this requires a joint hypothesis test:

$$H_0: \ \beta_3 = 0 \ \& \ \beta_4 = 0$$
$$H_1: \ \beta_3 \neq 0 \text{ and/or } \beta_4 \neq 0$$

Step 1: Estimate the restricted model.

Step 2: Calculate the sum of squared residuals from the restricted and unrestricted models.

Step 3: Apply the F-stat formula.

Alternative (easier):

Step 2 (alternative): Find the R^2 in the summary of the restricted and unrestricted models.

Step 3 (alternative): Apply the R^2 version of the F-stat formula.

In [3]:
```
#First calculate SSR_U
SSR_U<-sum(reg2$residuals^2)

reg_restricted <-lm(aauw ~ female + repub, data=df)
#summary(reg_restricted)

SSR_R<-sum(reg_restricted$residuals^2)

n<-nobs(reg2)
k<-4
q<-2

F<-((SSR_R-SSR_U)/q)/(SSR_U/(n-k-1))
F

###or
R2_U<-summary(reg2)$r.squared
R2_R<-summary(reg_restricted)$r.squared

F_2<-((R2_U-R2_R)/q)/((1-R2_U)/(n-k-1))
F_2

n-k-1
```

0.377972643431848

0.377972643431664

381

With 2 numerator degrees of freedom and 381 denominator degrees of freedom the critical value is ≈3 (for a 95% confidence level). Our F-statistic of 0.38 is lower than any reasonable critical value, so we fail to reject the null hypothesis.

(e) Using the third model, predict the AAUW score for male democrats who have 2 daughters and 1 son, and who have 25% of constituents who want more spending on defense, on average. Suggest a 95% CI for that predicted value.

(Hint: See part 3-A of Section Notes 8.)

In [4]:
```
#Recenter RHS variables around desired values
df$ngirls_aux <- df$ngirls-2
df$ngirls2_aux <- df$ngirls2-4
df$totchi_aux <- df$totchi-3
df$moredef_aux <- df$moredef-25

#Run auxilliary regression
reg3_aux <- lm(aauw ~  female + repub + ngirls_aux + ngirls2_aux + repubngirl + repubngirls2 + totchi_a
ux + moredef_aux, data = df)
summary(reg3_aux)
pred_val <- summary(reg3_aux)$coefficients[1,1]
ci_lower<- pred_val-1.96*summary(reg3_aux)$coefficients[1,2]
ci_upper<- pred_val+1.96*summary(reg3_aux)$coefficients[1,2]

print(paste0('Predicted value:', pred_val))
print(paste0('Confidence interval: [',ci_lower,' ',ci_upper, ']'))
```

```
Call:
lm(formula = aauw ~ female + repub + ngirls_aux + ngirls2_aux + 
    repubngirl + repubngirls2 + totchi_aux + moredef_aux, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-85.436  -7.964  -1.367  11.292  54.591 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   84.5800     2.0674  40.910  < 2e-16 ***
female        11.6079     2.8334   4.097 5.13e-05 ***
repub        -79.4364     3.0424 -26.110  < 2e-16 ***
ngirls_aux     0.4452     3.1682   0.141   0.8883    
ngirls2_aux    0.5286     0.8568   0.617   0.5376    
repubngirl     2.1281     3.6217   0.588   0.5571    
repubngirls2  -0.7477     0.9302  -0.804   0.4220    
totchi_aux    -2.0364     0.8066  -2.525   0.0120 *  
moredef_aux   -0.3166     0.1247  -2.540   0.0115 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.2 on 377 degrees of freedom
Multiple R-squared:  0.8505,	Adjusted R-squared:  0.8474 
F-statistic: 268.2 on 8 and 377 DF,  p-value: < 2.2e-16

[1] "Predicted value:84.5799623261095"
[1] "Confidence interval: [80.5277736498458 88.6321510023732]"
```

Here we are using the trick in the Section 8 Notes, where we first recenter our independent variables around the desired values (such that $\tilde{x}_j = 0$ when x_j equals the value we want to test). Then we re-run our regression model with these 'augmented' variables. $\hat{\beta}_0$ in this regression thus gives us the predicted AAUW score at our desired x values and comes with the correct standard error. We can use this estimated coefficient and its standard error to construct a 95% confidence interval in the usual way.

(f) Suppose a particular male Democrat has 2 daughters and 1 son in a state where 25% of constituents want more spending on defense. Generate a 95% CI for his *particular* AAUW score, continuing to use the third model.

(Hint: See part 3-B of Section Notes 8. Note that you can use most the values you already calculated in part (e) to answer this question.)

In [5]:
```
var_yhat <- (summary(reg3_aux)$coefficients[1,2])^2
sigma2_hat <- (summary(reg3_aux)$sigma)^2
ci_lower2 <- pred_val-1.96*sqrt(var_yhat+sigma2_hat)
ci_upper2 <- pred_val+1.96*sqrt(var_yhat+sigma2_hat)
print(paste0('Confidence interval: [',ci_lower2,' ',ci_upper2, ']'))
var_yhat
sigma2_hat
```

```
[1] "Confidence interval: [50.6348575966548 118.525067055564]"
```

4.27432139422112

295.671049048731

Note the difference from part (e) is that we are now predicting the value for one particular observation rather than *average* male Democrat with the specified values. Our point estimate will be the same, but we will need to calculate new standard errors to construct this confidence interval. Follow the formula in the Section Notes 8, 3-B to obtain that $\widehat{s.e.(\hat{u}^0)} = \sqrt{Var(\hat{y}) + \hat{\sigma}^2}$. Now we just have to plug this new standard error back into the confidence interval (our predicted value doesn't change)

$$CI^0 = [\hat{y} - 1.96se(\hat{u}^0), \hat{y} + 1.96se(\hat{u}^0)]$$

where $Var(\hat{y}) = 4.274$ and $\hat{\sigma}^2 = 295.67$

(g) Suppose you think Republicans and non-Republicans may have different gender patterns in voting with respect to the AAUW score. That is, republican men may vote differently than Republican women, who may vote differently than Democratic women who may vote differently than Democratic men. Write down an estimation equation you could use to test whether Republican women, Democratic women, and Democratic men each vote differently than Republican men. Specify what your null and alternative hypotheses would be.

To test this hypothesis, we could create variables for these categories of legislators and run the following regression:

$$aauw = \beta_0 + \beta_1 repubwoman + \beta_2 demwoman + \beta_3 demman + u$$

Note that we left out Republican men as the omitted category, meaning all effects will be interpreted relative to the Republican men. We can test the separate null hypotheses that $\beta_j = 0$ against the alternative hypothesis that $\beta_j \neq 0$ for $j = (1, 2, 3)$ using a t-test. (An F-test would be used to test the null that Republican men vote the same as Republican women and Democratic men and women as a whole, which is not what we are asking.)

One could also write

$$aauw = \beta_0 + \beta_1 dem + \beta_2 female + \beta_3 dem \times female + u$$

with a slightly different test and interpretation of β_3 in the next part.

(h) Implement your test. Interpret each coeffcient.

(Hint: To create dummy variables based on particular characteristics, it is easiest to first create the dummy variable and set it equal to 0 for all observations: `data$dummy<0`. Then, replace the values for that dummy with 1 for the observations that match the requirements you are looking for, as in `data[data$x1==0 & data$x2==1,]$dummy<-1`.)

(Hint: If you need to, you can include an interaction term in your regression using `:`. For example `lm(y~x1+x2+x1:x2,data=data)` includes an interaction between `x1` and `x2`. You will need to load the `car` package.)

In [6]:
```
df$repubwoman<-0
df[df$repub==1 & df$female==1,]$repubwoman<-1

df$demwoman<-0
df[df$repub==0 & df$female==1,]$demwoman<-1

df$demman<-0
df[df$repub==0 & df$female==0,]$demman<-1

#Direct way
reg4<-lm(aauw ~ repubwoman +demwoman +demman, data=df)
summary(reg4)

df <- mutate(df, dem= 1-repub)

#Interaction way
library(car)
reg4_alt<-lm(aauw ~ female+dem+female:dem, data=df)
summary(reg4_alt)
linearHypothesis(reg4_alt, 'female+dem+female:dem=0')
```

```
Call:
lm(formula = aauw ~ repubwoman + demwoman + demman, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-86.586  -6.246  -6.246  13.414  55.754 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    6.246      1.257    4.97 1.01e-06 ***
repubwoman    16.111      4.809    3.35 0.000889 ***
demwoman      89.168      3.462   25.76  < 2e-16 ***
demman        80.339      1.888   42.55  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.37 on 382 degrees of freedom
Multiple R-squared:  0.8455,	Adjusted R-squared:  0.8443 
F-statistic: 696.7 on 3 and 382 DF,  p-value: < 2.2e-16
```

```
Loading required package: carData


Attaching package: 'car'


The following object is masked from 'package:dplyr':

    recode


The following object is masked from 'package:purrr':

    some
```

```
Call:
lm(formula = aauw ~ female + dem + female:dem, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-86.586  -6.246  -6.246  13.414  55.754 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    6.246      1.257   4.970 1.01e-06 ***
female        16.111      4.809   3.350 0.000889 ***
dem           80.339      1.888  42.553  < 2e-16 ***
female:dem    -7.283      5.960  -1.222 0.222460    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.37 on 382 degrees of freedom
Multiple R-squared:  0.8455,	Adjusted R-squared:  0.8443 
F-statistic: 696.7 on 3 and 382 DF,  p-value: < 2.2e-16
```

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	383	315432.1	NA	NA	NA	NA
2	382	115250.6	1	200181.5	663.5051	1.55425e-85

$\hat{\beta}_1 = 16.11$ means that on average, Republican women have a 16.11 percentage point higher AAUW score than the omitted category, Republican men (who have a score of 6.24). $\hat{\beta}_2 = 89.168$ means that on average, Democratic women have an 89 percentage point higher AAUW score than Republican men and $\hat{\beta}_3 = 80.33$ means that on average, Democratic men have a 80.3 percentage point higher AAUW score than Republican men. All of these coefficients are significant at the 1\% level or lower, which means there is stastical evidence that each group votes differently from Republican men: we can reject the null that they vote the same with respect to the AAUW score.

If you ran the interacted model, then the coefficients for Republican women (female) and Democratic men (dem) would have the same interpretation as above, but for Democratic women, you would need to use a t-test to test the null that $\beta_1 + \beta_2 + \beta_3 = 0$ rather than just testing $\beta_3 = 0$. That is because the total effect of being a Democratric woman is the sum of those coefficients. Calculating the correct standard error for this t-test would be challenging. Using `linearHypothesis` or `lincom` will allow you to test this hypothesis. We show the output from using `linearHypothesis`, and observe that the F-stat of 663 is very large (and associated p-value is very small) so we can strongly reject the null hypothesis that the sum of those coefficients is 0. Notice also that the F-stat of 663 is the square of the t-stat we obtain in the first model for the coefficient on demwoman - this is due to the relationship between the F and t distributions.

(i) Adapt your regression to test whether Democratic women vote differently than Republican women with respect to the AAUW score. Write out the estimating equation and report your results.

In [7]:
```
options(warn=-1)
df$repubman<-0
df[df$repub==1 & df$female==0,]$repubman<-1

reg5<-lm(aauw~ repubman + demwoman +demman, data=df)
summary(reg5)

#Alternatively
linearHypothesis(reg4, 'demwoman = repubwoman')
```

```
Call:
lm(formula = aauw ~ repubman + demwoman + demman, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-86.586  -6.246  -6.246  13.414  55.754 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   22.357      4.642   4.816 2.11e-06 ***
repubman     -16.111      4.809  -3.350 0.000889 ***
demwoman      73.057      5.653  12.924  < 2e-16 ***
demman        64.228      4.851  13.239  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.37 on 382 degrees of freedom
Multiple R-squared:  0.8455,	Adjusted R-squared:  0.8443 
F-statistic: 696.7 on 3 and 382 DF,  p-value: < 2.2e-16
```

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	383	165644.4	NA	NA	NA	NA
2	382	115250.6	1	50393.8	167.0311	5.991823e-32

We rewrite the model to make Republican women the omitted category and estimate the equation:

$$aauw = \beta_0 + \beta_1 repubman + \beta_2 demwoman + \beta_3 demman + u$$

Here we see that: on average, Democratic women have an aauw score that is 73.05 points higher than that of Republican women and is highly significant ($t = 12.924$). We could also use the `linearHypothesis` function from the `car` package and test $\beta_1 = \beta_2$ from the model in part (i) directly. We would obtain an F-stat of 167, which is highly significant as well (and is in fact the square of the t-stat).

Downloading your Notebook

Download a PDF copy of your notebook by using **File > Download as HTML.** Alternatively, go to print preview and **Save as PDF** (but make sure your code does not get cut off at the side of the page!)