ECON 7710
Homework 2: Regression Fundamentals

Nate Moore

Robby Merrill

Gage Thompson

Will Noll

Gianluca Jones

General instructions

The `.Rmd` source for this document will be the template for your homework submission. You must submit your completed assignment as a single html document, uploaded to eLC by **11:59p on January 30** using the filename `sectiontime_teamnumber_hw1.html` (e.g. `935_1_hw1.html`).

Notes:

- Include the name of each teammate under `author` in the `yaml`.
- For questions requiring analytical solutions, you can type them in using markdown math code. Or, you can submit handwritten solutions, embedding them in the knitted document as clearly readable images.
- For questions requiring computation, some or all of the required code is included in associated chunks. Modify chunks where and how you are directed.
- For (almost) all questions about R Markdown, consult The Definitive Guide (https://bookdown.org/yihui/rmarkdown/).
- The `setup` chunk above indicates the packages required for this assignment.
- You will find a description of the variables in the referenced dataset through the Help tab in the Plot pane of RStudio.
- **Switch `eval` to `TRUE` in the global options command to execute code chunks**.

Part 1: Linear Regression

In this assignment, you are going to use data collected by Christopher Lemmon from a survey of Michigan State undergradutes in 1994. You have been tasked by the admissions office with assessing the relationship between ACT scores and college GPA.

a. Write down the initial bivariate linear regression you would estimate.

b. Estimate the parameters of this regression model. Write a sentence interpreting the coefficient estimates.

c. Create a scatter plot of the data with the best fit line. Include the regression parameters you estimated in (b).

d. Calculate the average GPA for each value of ACT scores we observe in the data set. Plot these values in a binned scatter plot. How does the information from this data plot compare to the scatter plot in part c?

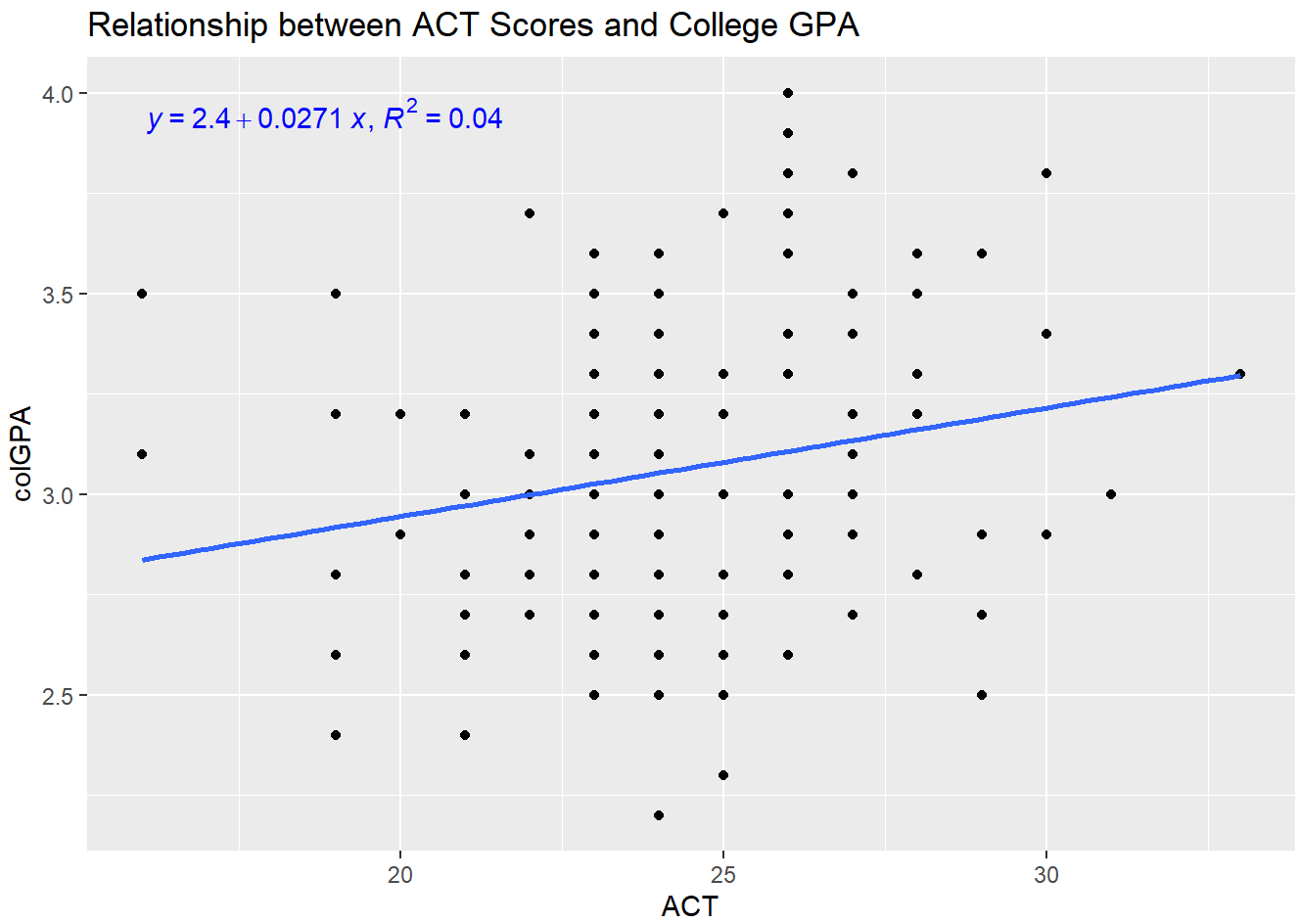
```
data(gpa1)

# (b) Estimate the baseline regression
educ_reg <- lm_robust(colGPA ~ ACT, data=gpa1)
summary(educ_reg)
```

```
##
## Call:
## lm_robust(formula = colGPA ~ ACT, data = gpa1)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  2.40298    0.26295   9.139 6.850e-16  1.88308  2.92288 139
## ACT          0.02706    0.01098   2.464 1.495e-02  0.00535  0.04878 139
##
## Multiple R-squared:  0.04275 ,   Adjusted R-squared:  0.03586
## F-statistic: 6.073 on 1 and 139 DF,  p-value: 0.01495
```

```
# (c) Create a scatter plot with the best fit line
ggplot(gpa1, aes(x = ACT, y = colGPA)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  stat_poly_eq(mapping = use_label(c("eq","R2")), color = "blue") +
  labs(title="Relationship between ACT Scores and College GPA")
```

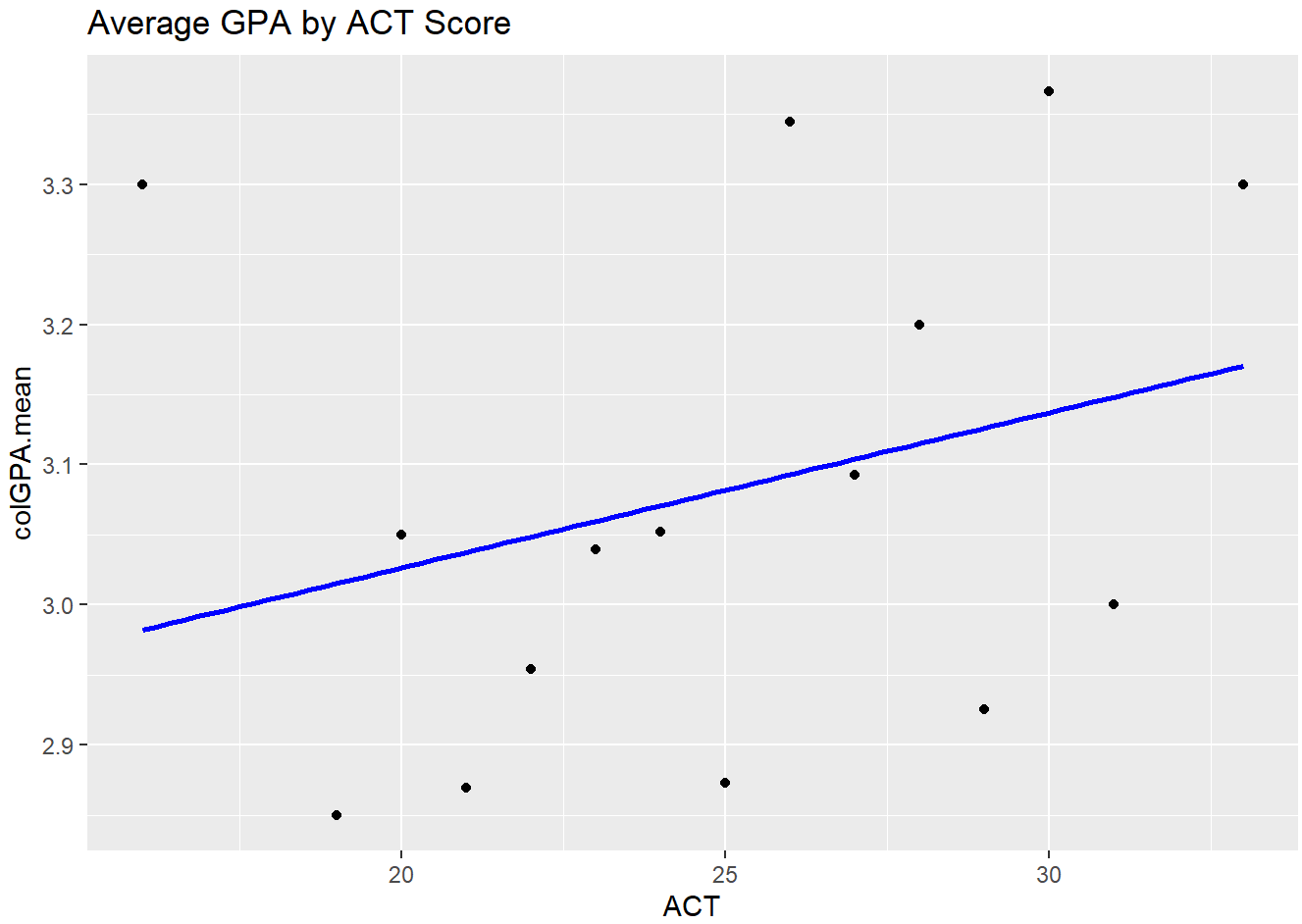
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# (d) Calculate average GPA for each value of ACT score
collapsed_data <- summaryBy(colGPA ~ ACT, data=gpa1, FUN=mean)

# Binned scatter plot
ggplot(collapsed_data, aes(x = ACT, y = colGPA.mean)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Average GPA by ACT Score")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Want to practice more R? Add the regression line to this last plot following the code from the lectures
```

**Answers:

a.

$$\text{GPA} = \beta_0 + \beta_1 \times \text{ACT} + \epsilon$$

b.

The estimated intercept is approximately 2.40298. This can be interpreted as the expected college GPA for a student with an ACT score of zero.The coefficient estimate of 0.0276 for ACT scores in the regression model indicates that each one-point increase in ACT score is associated with an average increase of 0.0276 in college GPA. This along with a p-value of 1.495e-02 suggests a significant relationship.

c.

See the plot above: “Relationship between ACT Scores and College GPA.”

d.

The first plot provides a more detailed view of the data, showing how GPA varies among students with the same ACT score. This could be useful for identifying outliers or understanding the distribution of GPA scores. The second plot simplifies the information and makes it easier to perceive the overall trend without the distraction of individual variances.

Part 2: Omitted Variables Bias

You are concerned about omitted variables bias from HS GPA.

a. Write down the omitted variables bias formula. What would need to be true for there not to be omitted variables bias here?

b. Create a scatter plot of ACT scores and HS GPA and calculate the correlation between ACT scores and HS GPA. What do you notice?

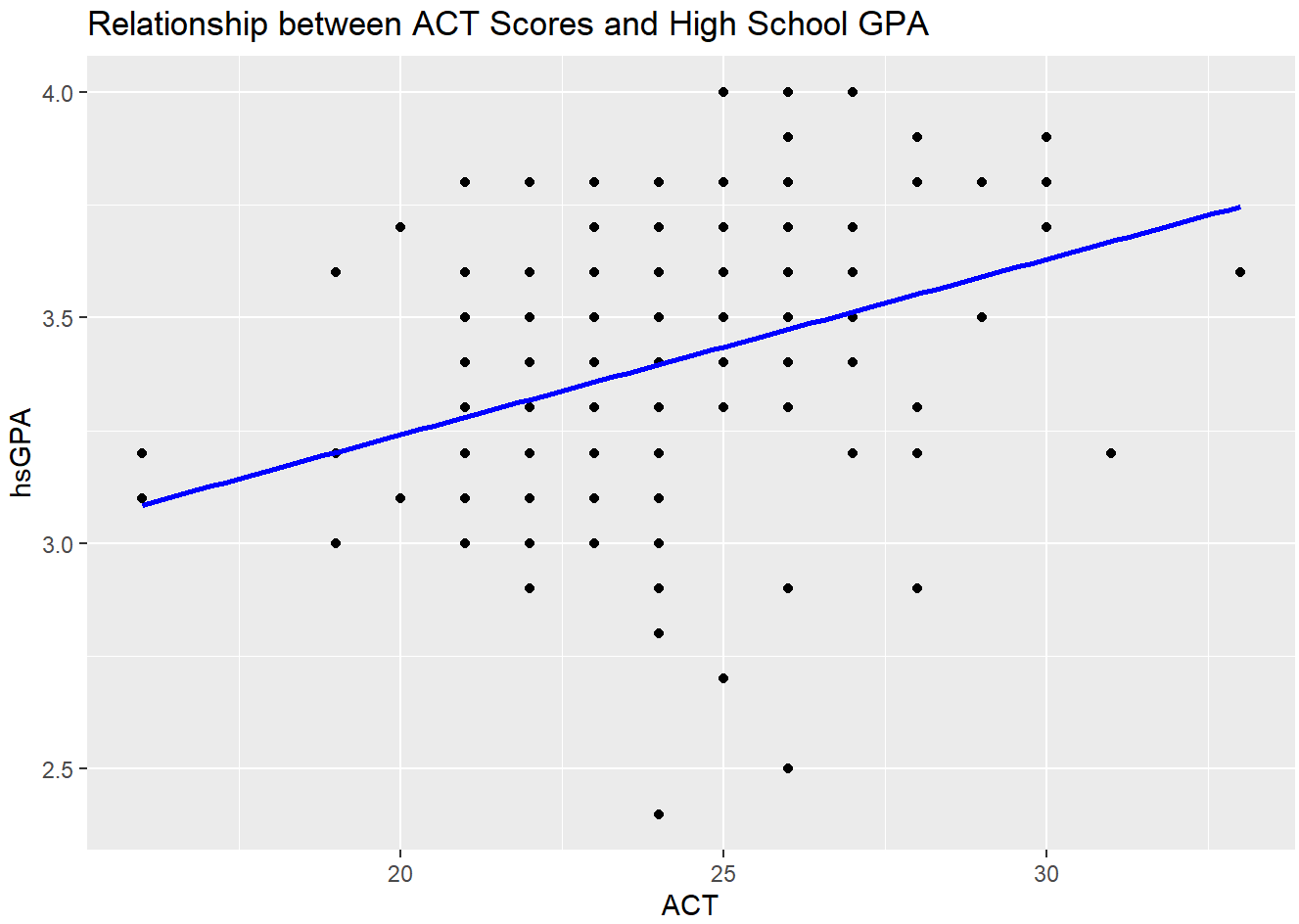
c. Write down the multivariate regression model you would estimate instead.

d. Estimate the parameters of this regression model. Write a sentence interpreting your results.

```
# (b)

ggplot(gpa1, aes(x = ACT, y = hsGPA)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title="Relationship between ACT Scores and High School GPA")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
cor(gpa1$ACT, gpa1$hsGPA)
```

```
## [1] 0.3458056
```

```
# (d)
educ_reg2 <- lm_robust(colGPA ~ ACT + hsGPA, gpa1)
summary(educ_reg2)
```

```
##
## Call:
## lm_robust(formula = colGPA ~ ACT + hsGPA, data = gpa1)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept) 1.286328    0.36197  3.5537 0.0005205  0.57061  2.00205 138
## ACT         0.009426    0.01090  0.8649 0.3885740 -0.01212  0.03097 138
## hsGPA       0.453456    0.09964  4.5509 0.0000116  0.25644  0.65048 138
##
## Multiple R-squared:  0.1764 ,    Adjusted R-squared:  0.1645
## F-statistic: 12.97 on 2 and 138 DF,  p-value: 6.874e-06
```

Answers

a. B2 has to be equal to 0 or there has to be no relationship between X1 and X2 for there to be no omitted variable bias.

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2\tilde{\delta}$$

b. There seems to be a postive moderate correlation between ACT scores and high school GPA. This means as ACT score increases high school GPA tends to increase as well.

c. $$\text{GPA} = \beta_0 + \ 0.009426 \times \text{ACT} + \ 0.453456 \times \text{HS GPA} + \epsilon$$

d. ACT is not a significant predictor of college GPA when high school GPA is accounted for. In this model high school GPA is a significant predictor of college GPA. The multiple r squared is .1764 which means the model only explains 17.64% of the variance, this is quite low which indicates that there are likely more variables that influence college GPA.

Part 3: Frisch-Waugh-Lovell

We now want to unpack what is going on in this multivariable regression using Frisch-Waugh-Lovell.

a. Write a few sentences explaining what the Frisch-Waugh-Lovell theorem states.
b. Regress ACT scores on HS GPA. Save the residuals. (No need to write anything here.)
c. Regress college GPA on HS GPA. Save the residuals. (No need to write anything here.)
d. Regress the college GPA residuals on ACT residuals. Verify that you got the same estimate as in 2c.
e. Now, we are going to create a residualized binned scatter plot. This plot allows us to visualize the relationship between two variables, conditional on other variables (“ceteris parabis”).

i. First, create deciles of the residualized ACT scores.

ii. Now, calculate the mean of both the residualized ACT scores and residualized college GPA for each decile of residualized ACT scores.
iii. Plot in a scatter plot. Discuss how this plot differs from the unresidualized scatter plot above. Pay attention to the scale of the y-axis.

Note: you have now through how to manually create a binned scatter (or binscatter) plot in R! Packages exist to do this, but collapsing the data yourself allows you more flexibility.

```
# (b) Run regression of ACT scores on HS GPA and save the residuals
reg1 <- lm_robust(ACT ~ hsGPA, gpa1)
ACT_resid <- (reg1$fitted.values - gpa1$ACT)

# (c) Run regression of College GPA on HS GPA and save the residuals
reg2 <- lm_robust(colGPA  ~ hsGPA, gpa1)
colGPA_resid <- (reg2$fitted.values - gpa1$colGPA)

#(d) Regress the residuals on each other
reg3 <- lm_robust(colGPA_resid ~ ACT_resid)
summary(reg3)
```

```
##
## Call:
## lm_robust(formula = colGPA_resid ~ ACT_resid)
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|) CI Lower CI Upper  DF
## (Intercept) 1.636e-16    0.02856 5.729e-15    1.000 -0.05647  0.05647 139
## ACT_resid   9.426e-03    0.01086 8.679e-01    0.387 -0.01205  0.03090 139
##
## Multiple R-squared:  0.005513 ,  Adjusted R-squared:  -0.001642
## F-statistic: 0.7532 on 1 and 139 DF,  p-value: 0.387
```

```
# (e) Create the residualized binned scatter plot

# Calculate deciles
act_deciles <- ntile(ACT_resid, 10)
resid_data <- data.frame(colGPA_resid, ACT_resid, act_deciles)
decile_data <- summaryBy(.  ~ act_deciles, data=resid_data)

# Add back the means (this is a scaling thing)
decile_data$colGPA_resid.mean <- decile_data$colGPA_resid.mean + mean(gpa1$colGPA)
decile_data$ACT_resid.mean <- decile_data$ACT_resid.mean + mean(gpa1$ACT)

# Plot the binned residuals
ggplot(decile_data, aes(ACT_resid.mean, colGPA_resid.mean)) + geom_point() +
  labs(title = "Residualized Binned Scatter Plot")
```


```
# Want more practice with plotting in R? Add the best fit line to this plot like we did in class
```

Answers

a. FWL decomposes a regression of Y on a set of variables, such as X1 and X2. FWL states that if we regress Y on all of the variables (i.e., X1 and X2) then we get the same coefficient estimates on X2 and the same residuals if we regress Y on all the variables in X1 and X2. It can be a three step process where you get the residuals of all the variables on X1 and X2. Then, regressing Y on X1, taking these residuals, and then regressing residuals from Y and back on the residuals from X2.

b. When running educ_reg2 from part 2c, we get a coefficient estimate for ACT of 0.009426. This is confirmed when we run reg3 for part 3d the coefficient estimate for ACT is again 0.009426, confirming the FWL.

$$\text{College GPA} = \beta_0 + \ 0.009426 \times \text{ACT} + \epsilon$$

c. This plot differs from the unresidualized scatter plot in that the data is clustered more closely together, and there are less outliers/extreme data points.