Problem Set 5

Due Date: February 13, 2026
Submission: https://canvas.northwestern.edu/courses/245562/assignments/1687750

Problem 1

1a. In your own words, explain: *1. The difference between residuals (e_i) and modeling errors (ϵ_i)*

Residuals are sample estimates; they are the difference between a model's prediction and the observed data. That is, they are $Y_i - \hat{Y}_i$. Modeling errors are population properties; they are the distribution of differences between the conditional expectation function and the distribution of data at the theoretical level. That is, they are $Y_i - E(Y_i|X_i)$.

2. Why the sum of residuals in OLS regression equals zero when an intercept is included

Because, once an intercept is included, the first-order condition forces the residuals to sum to zero. Otherwise, the residuals would not be orthogonal to the intercept, and the estimate would not be optimal.

3. How R^2 measures model fit and what it represents

R^2 is a measure of the variance of the residuals relative to the variance of the dependent variable. Specifically, it is 1 minus the variance of the residuals divided by the variance of the dependent variable. Because the residuals ordinarily will have a variance that is less than or equal to the unconditional variance of the dependent variable, R^2 typically ranges between 0 and 1. It represents the proportion of the variance in the dependent variable that has been partitioned into the fitted values of the regression, as opposed to the proportion that falls to the residuals.

1b. Using the Hibbs election data:

```
# Load data
library(rosdata)
data("hibbs")

# Fit the models
model_intercept <- lm(vote ~ 1, data = hibbs)  # Intercept only
model_econ <- lm(vote ~ growth, data = hibbs)   # With growth
```

Questions: 1. Verify that $\sum e_i = 0$ for both models.

```
sum(model_intercept$residuals)
```

```
## [1] 0
```

```
sum(model_econ$residuals)
```

```
## [1] -8.437695e-15
```

2. Calculate R^2 using the formula $R^2 = 1 - \frac{RSS}{TSS}$.

```
1 - sum(model_intercept$residuals^2)/sum((hibbs$vote- mean(hibbs$vote))^2)
```

```
## [1] 7.771561e-16
```

```
1 - sum(model_econ$residuals^2)/sum((hibbs$vote- mean(hibbs$vote))^2)
```

```
## [1] 0.5798462
```

3. Compare your calculation with the `summary()` output.

```r
summary(model_intercept)
```

```
##
## Call:
## lm(formula = vote ~ 1, data = hibbs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.455 -3.705 -1.300  3.440  9.735
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   52.055      1.402   37.12 3.54e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.609 on 15 degrees of freedom
```

```r
summary(model_econ)
```

```
##
## Call:
## lm(formula = vote ~ growth, data = hibbs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9929 -0.6674  0.2556  2.3225  5.3094
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.2476     1.6219  28.514 8.41e-14 ***
## growth        3.0605     0.6963   4.396  0.00061 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.763 on 14 degrees of freedom
## Multiple R-squared:  0.5798, Adjusted R-squared:  0.5498
## F-statistic: 19.32 on 1 and 14 DF,  p-value: 0.00061
```

1c. Limitations of R^2: *1. What happens to R^2 when you add more variables to a model, even irrelevant ones?*

It either goes up or stays the same. There is no way that the classic R^2 measure can ever decline when adding a new variable. Often even an irrelevant variable may marginally increase R^2 by capitalizing on chance.

2. Why might a high R^2 not indicate a good model?

Such a value might indicate that post-treatment variables have been included in addition to appropriate control variables, and that model construction proceeded in a post-hoc way intended to maximize a relatively unhelpful metric rather than on the basis of theoretical, causal, and substantive knowledge.

Problem 2

2a. *1. Write down the regression model in matrix form and explain each symbol.*

$$Y = \mathbf{X}\beta + \epsilon$$

Here, Y is an N by 1 vector of observations of the dependent variable. $\mathbf{X}$ is an N by $k+1$ matrix, where k is the number of independent variables. One column of $\mathbf{X}$ is a set of 1s to allow for the intercept. Each of the other columns contains the data for the independent variables. ϵ is the error term, which captures the difference between the model prediction and the actual value of the dependent variable for each of the N observations.

2b. Create a multicollinear scenario:

```
# Create multicollinear data
set.seed(123)
n <- 100
x1 <- rnorm(n)
x2 <- 0.95*x1 + rnorm(n, sd = 0.1)  # Highly correlated with x1
x3 <- rnorm(n)
y <- 2 + 1.5*x1 + 0.8*x3 + rnorm(n)

# Fit models
model_collinear <- lm(y ~ x1 + x2 + x3)
summary(model_collinear)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.49138 -0.65392  0.05664  0.67033  2.53210
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.9807     0.1073  18.452  < 2e-16 ***
## x1            1.0055     1.0410   0.966    0.337
## x2            0.4622     1.0946   0.422    0.674
## x3            0.7426     0.1122   6.617 2.09e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.052 on 96 degrees of freedom
## Multiple R-squared:  0.6503, Adjusted R-squared:  0.6394
## F-statistic: 59.52 on 3 and 96 DF,  p-value: < 2.2e-16
```

```
# Check variance inflation factors (VIF)
library(car)
```

```
## Loading required package: carData
```

```
vif(model_collinear)
```

```
##        x1        x2        x3
## 80.852638 80.779423  1.017576
```

Questions: *1. What happens to the standard errors when multicollinearity is present?*

Standard errors are inflated when multicollinearity is present.

2. How do the VIF values indicate multicollinearity?

When the VIF values are high, then we know that the variables in question are strongly interconnected in such a way that the uncertainty of estimation is being increased. A VIF greater than 10 typically indicates a quite problematic level of multicollinearity, so the levels we see in this example are extreme.

3. What are the practical implications for interpreting coefficients in the presence of multicollinearity?

The multicollinear coefficients are typically difficult to all estimate jointly, because their presence destabilizes the model. As such, it may be best to omit one or more to reduce overall excess variance. Because the variables are empirically close to redundant, it is likely that little will be lost in this dataset with such an omission.

Problem 3

Using the turnout data:

```
library(readr)
turnout <- read_csv("https://raw.githubusercontent.com/jnseawright/ps405/refs/heads/main/Data/turnout.cs
```

```
## Rows: 50 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): Year, Turnout, Temperature, GDP
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
turnout_clean <- na.omit(turnout)

# Calculate leverage (diagonal of hat matrix)
model <- lm(Turnout ~ GDP + Temperature, data = turnout_clean)

# Method 1: Using hatvalues()
leverage1 <- hatvalues(model)
leverage1
```

```
##          1          2          3          4          5          6          7
## 0.03624436 0.12541190 0.03550707 0.07860531 0.03497553 0.03715984 0.13289417
##          8          9         10         11         12         13         14
## 0.03978736 0.05114088 0.03609191 0.03497469 0.03474987 0.03579523 0.03816194
##         15         16         17         18         19         20         21
## 0.15598608 0.04780846 0.12429686 0.07938743 0.03955166 0.03395029 0.03394029
##         22         23         24         25         26         27         28
## 0.03783342 0.15512462 0.04704354 0.04059609 0.02907917 0.04231407 0.03839980
##         29         30         31         32         33         34         35
## 0.04384295 0.04684764 0.06587996 0.04801772 0.08548720 0.15357437 0.16794931
##         36         37         38
## 0.16075867 0.19348305 0.37734731
```

```
# Method 2: Calculate manually
xmat <- cbind(1,turnout_clean$GDP,turnout_clean$Temperature)

leverage2 <- diag(xmat%*%solve(t(xmat)%*%xmat, tol=1e-30)%*%t(xmat))
leverage2
```

```
##  [1] 0.03624436 0.12541190 0.03550707 0.07860531 0.03497553 0.03715984
##  [7] 0.13289417 0.03978736 0.05114088 0.03609191 0.03497469 0.03474987
## [13] 0.03579523 0.03816194 0.15598608 0.04780846 0.12429686 0.07938743
## [19] 0.03955166 0.03395029 0.03394029 0.03783342 0.15512462 0.04704354
## [25] 0.04059609 0.02907917 0.04231407 0.03839980 0.04384295 0.04684764
## [31] 0.06587996 0.04801772 0.08548720 0.15357437 0.16794931 0.16075867
## [37] 0.19348305 0.37734731
```

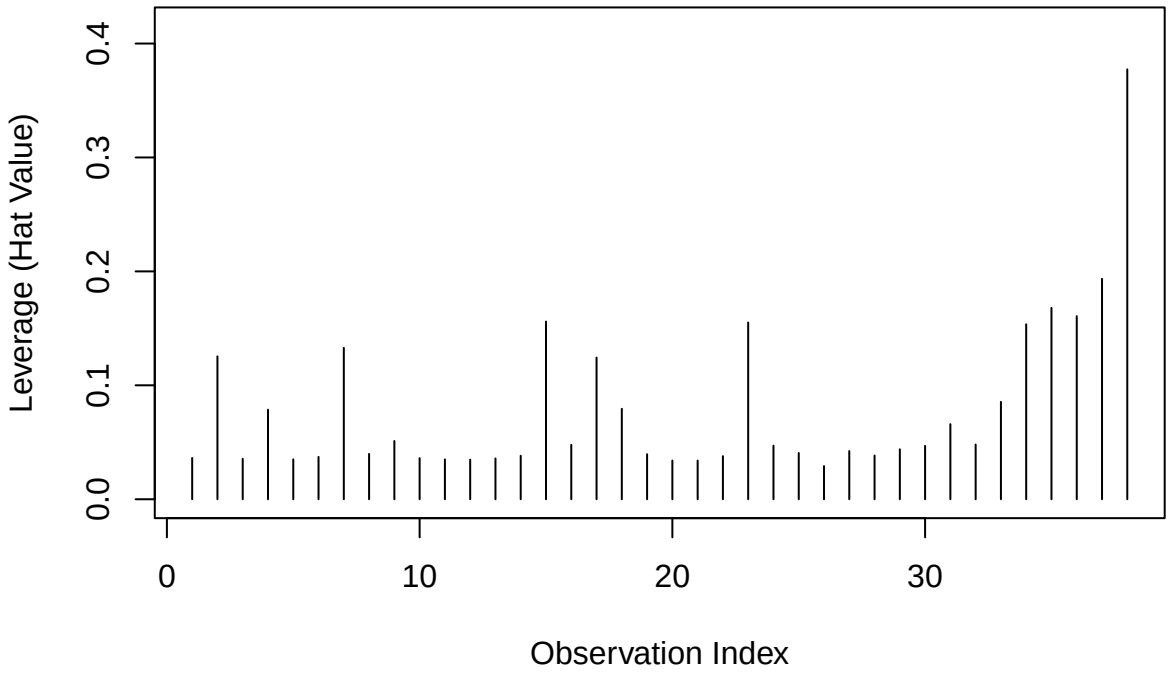
```r
# Compare

leverage1 - leverage2
```

```
##             1             2             3             4             5
##  9.589551e-15  6.264433e-14 -4.496403e-15  2.436940e-14 -1.040834e-15
##             6             7             8             9            10
##  9.089951e-16  6.783463e-14  9.013623e-15  1.669498e-14 -1.984524e-15
##            11            12            13            14            15
##  4.440892e-16  3.802514e-15  5.113965e-15 -3.622103e-15  9.547918e-14
##            16            17            18            19            20
##  1.153938e-14  6.449008e-14  2.925438e-14  5.800915e-15  6.779299e-15
##            21            22            23            24            25
##  7.854828e-15  6.362966e-15  9.764412e-14  7.480128e-15  1.540434e-15
##            26            27            28            29            30
## -4.832940e-15  7.278900e-15  4.704570e-15  7.258083e-15  9.534040e-15
##            31            32            33            34            35
##  1.915135e-14  9.561796e-15  2.023381e-14  4.296563e-14  5.062617e-14
##            36            37            38
##  1.099121e-14  1.387779e-16  2.237099e-14
```

```r
# Identify high leverage points

plot(leverage1,
     type = "h",  # vertical lines
     main = "Index Plot of Leverage Values",
     xlab = "Observation Index",
     ylab = "Leverage (Hat Value)",
     ylim = c(0, max(leverage1) * 1.1))
```



Questions: *1. What does leverage measure? Why do we care about high leverage points?*

Leverage measures how much the regression's overall coefficient estimates would change if a given point were deleted. High leverage points indicate points that both add an unusually high amount of information to the regression and have a distinctively high possibility of throwing the model off through measurement error, idiosyncratic factors, omitted variables, etc.

2. What is the average leverage value? What's the theoretical value?

```
mean(leverage1)
```

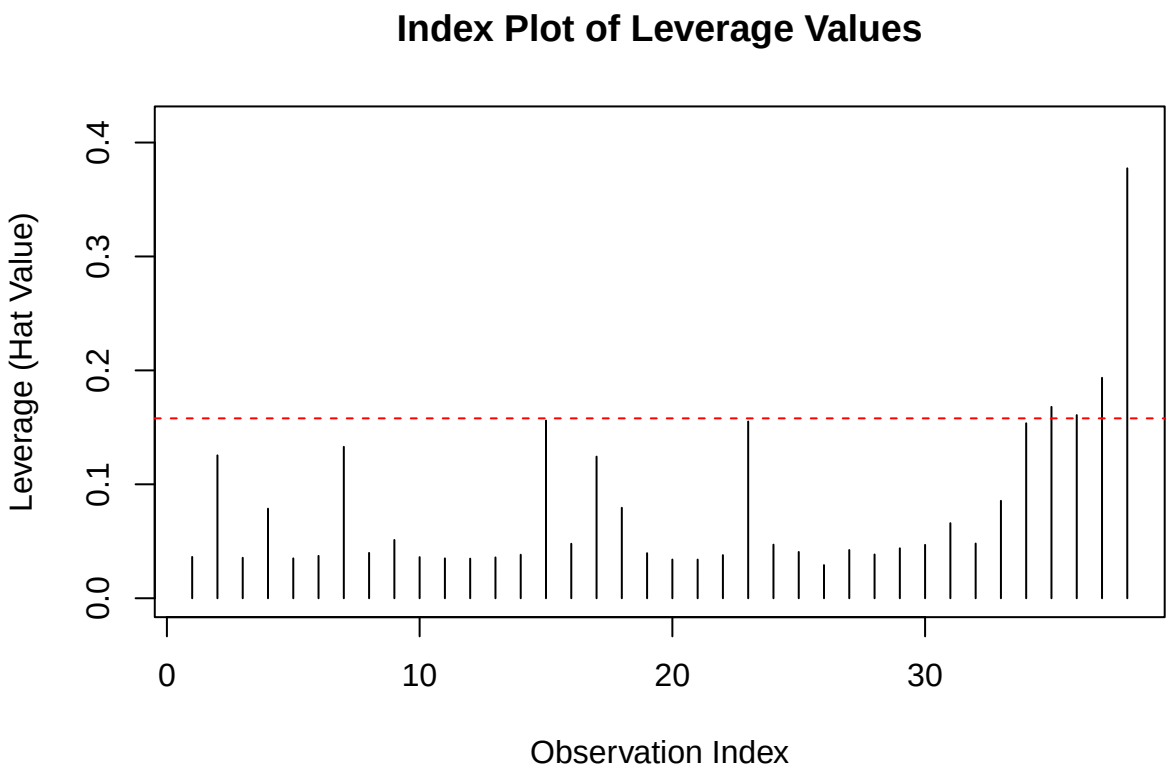
```
## [1] 0.07894737
```

The theoretical average is $\frac{k+1}{N} = \frac{3}{38} = 0.07894737$. Fantastic!

3. Create a plot of leverage vs. observation index. Add a horizontal line at $2p/n$ (where p = number of parameters).

```
# Create index plot
plot(leverage1,
     type = "h",  # vertical lines
     main = "Index Plot of Leverage Values",
     xlab = "Observation Index",
     ylab = "Leverage (Hat Value)",
     ylim = c(0, max(leverage1) * 1.1))

# Add horizontal reference lines
p <- length(coef(model))  # number of parameters
n <- nrow(turnout_clean)  # number of observations
abline(h = 2 * p / n, col = "red", lty = 2)  # common rule of thumb
```



Problem 4

4a. 1. Define DFBETA and Cook's Distance. What does each measure? 2. What's the difference between an outlier and an influential point?

4b. Using the turnout model:

```
# Calculate DFBETA
dfbeta(model)
# Plot DFBETA for Temperature coefficient
dfbetaPlots(model, terms = ~Temperature)
```

Questions: *1. Which observations are influential for the Temperature coefficient?*

Arguably none of them, but maybe the second one.

2. What happens to the Temperature coefficient if you remove the most influential observation?

```
# Simulation comparison
model_omit2 <- lm(Turnout ~ GDP + Temperature, data = turnout_clean, subset=c(1,3:38))

library(modelsummary)
influential_drop <- list("Full Model" = model,
               "Model Omitting Most Influential Observation" = model_omit2)
modelsummary(influential_drop, stars = TRUE, output = "markdown")
```

	Full Model	Model Omitting Most Influential Observation
(Intercept)	0.575	0.223
	(0.546)	(0.550)
GDP	-0.000	-0.000
	(0.000)	(0.000)
Temperature	0.001	0.008

	Full Model	Model Omitting Most Influential Observation
	(0.011)	(0.011)
Num.Obs.	38	37
R2	0.066	0.072
R2 Adj.	0.013	0.017
AIC	-68.6	-69.9
BIC	-62.1	-63.5
Log.Lik.	38.305	38.970
F	1.239	1.313
RMSE	0.09	0.08

- p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

The temperature coefficient gets quite a lot bigger but remains far from statistically significant; there is just a lot of uncertainty here.

3. Calculate Cook's Distance for all observations. Which observations have Cook's D > 0.5?

```
# Simulation comparison
cooks.distance(model)
```

```
##            1            2            3            4            5            6
## 5.932158e-02 1.848860e-01 3.703298e-02 9.937023e-02 2.377552e-02 4.370991e-02
##            7            8            9           10           11           12
## 7.567077e-02 1.172230e-03 1.738758e-03 1.648322e-03 1.317879e-05 2.598453e-02
##           13           14           15           16           17           18
## 2.690991e-02 4.465885e-03 7.017783e-02 6.031086e-03 7.971323e-03 3.041403e-02
##           19           20           21           22           23           24
## 1.748222e-02 9.359763e-05 6.944574e-04 5.160237e-04 1.185207e-03 2.354299e-04
##           25           26           27           28           29           30
## 4.342875e-03 3.993446e-03 7.704351e-03 5.840422e-03 1.814164e-02 3.834142e-03
##           31           32           33           34           35           36
## 2.890761e-02 1.051574e-02 6.527924e-04 5.307621e-04 4.493713e-03 6.324394e-05
##           37           38
## 7.343714e-02 1.113199e-01
```

```
cooks.distance(model)>0.5
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    14    15    16    17    18    19    20    21    22    23    24    25    26
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    27    28    29    30    31    32    33    34    35    36    37    38
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

No observations have an excessively large Cook's distance here, suggesting that overall leverage may not be too substantial for any observation.

4c. Case Study Analysis: *Identify the most influential observation in your model and analyze it:*

I am most interested in the coefficient for the GDP variable, so I will use DFBETA for that.

```
# Find most influential observation
dfbetaPlots(model, terms = ~GDP)
which.max(abs(dfbeta(model)[,2]))
# Analyze this observation
turnout_clean[38,]
```

```r
# We might suspect that the 2024 election was unusual because of the high polarization of the contest,
# multiple assassination attempts, court cases, and a late substitution of candidates for one party. Al
# have generated unusual patterns of engagement, aside from temperature and GDP!

# Run model without this observation

# Simulation comparison
# Compare coefficients
model_omit38 <- lm(Turnout ~ GDP + Temperature, data = turnout_clean, subset=c(1:37))

library(modelsummary)
influential_drop_2024 <- list("Full Model" = model,
               "Model Omitting 2024" = model_omit2)
modelsummary(influential_drop_2024, stars = TRUE, output = "markdown", fmt=fmt_significant(6))
```

Questions: *1. Why is this observation influential? Consider its leverage and residual.*

```r
hatvalues(model)[38]
summary(hatvalues(model))
model$residuals[38]
summary(model$residuals)
```

This observation has the single largest leverage score, combined with a relatively large (although not absolutely extreme) residual. This combination generates a high overall level of influence.

2. Should this observation be removed? What are the ethical and methodological considerations?

The observation has an outsized effect on the estimation, but also is potentially unusually informative. Deleting it could remove meaningful information about the relationships of interest, unless there are omitted variables or measurement error. At the same time, just deleting it could also mean losing the chance to do the extra work of uncovering sources of error that may affect other parts of the analysis. If this is a relationship that has important normative implications, putting in extra research attention to this observation may well be the better path.

3. What would you recommend to a researcher who found such an influential observation in their data?

Spend some time digging into the available information about this observation, treating it as a qualitative case study and an opportunity to learn more about the theory and social-science substance surrounding the model before making any statistical decisions. Then make decisions when informed by a deeper, expert understanding of primary and secondary evidence about this observation.