Answers to Problem Set 1

Nicholas Gonzalez

Work on with Tanner Bentley.

Problem #1

1.1

Regression residuals can be defined as the difference between the observed values of the dependent variable, or DV, and the predicted values of that variable.

Regression residuals can be used to evaluate how accurate a regression model fits a given data set.

$$e_i = y_i - \hat{y}_i$$

1.2

The population error can be defined as the difference between the estimate and the true value of a given population.

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

1.3

The population regression slope coefficient can be defined as the average change in y, the dependent variable, for everyone one unit increased in x, the independent variable. This can be symbolized has $\hat{\beta}_1$.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

1.4

The estimator of the regression slope coefficient is the average change in the dependent variable, or y, for one unit of movement in the independent variable, x. This can be symbolized has $\hat{\beta}$.

$$\hat{\beta_1} = \frac{Cov(X, Y)}{Var(X)}$$

Problem 2

2.1

The zero conditional mean ($\mathbb{E}[u|X] = 0 \quad \Rightarrow \quad \text{Cov}(X, u) = 0$) is the assumption that the average value of our error term, so u, given any specific value for the regressor x, would be zero.

If the zero conditional mean is unsatisfied, this means that there is some relationship with the regressor x and the error term u. This happens if their are important variables that correlate with x & y, but are not included in our model. This is omitted variable bias.

2.2.1

```
library(readr)
example <- read_csv("Problem_Set_1/example.csv")
```

```
Rows: 1000 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (5): x, y, u, y.alt, u.alt

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.
```

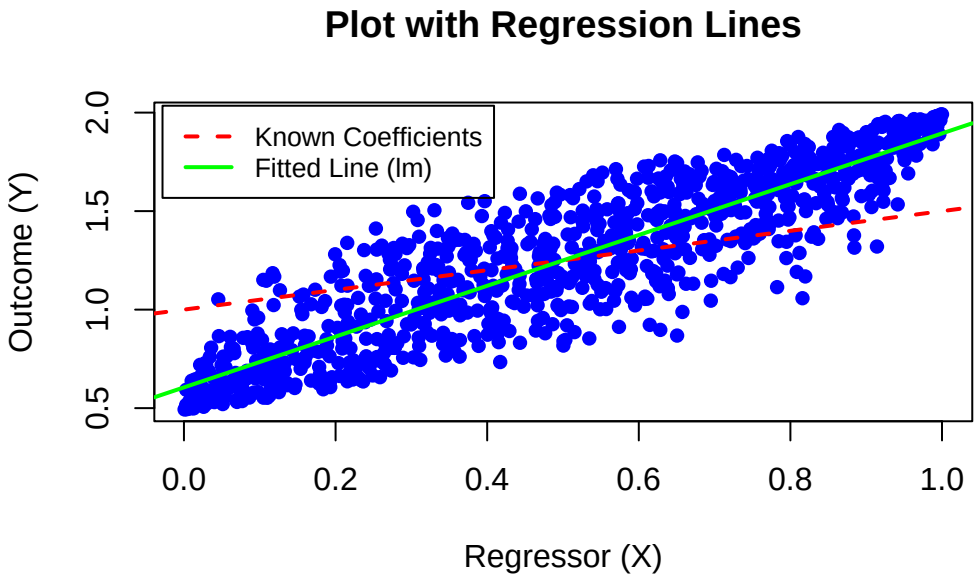
```
library(modelsummary)
```

	(1)
(Intercept)	0.605
	(0.011)
example$x	1.289
	(0.020)
Num.Obs.	1000
R2	0.812
R2 Adj.	0.812
AIC	−578.1
BIC	−563.3
Log.Lik.	292.036
RMSE	0.18

```
`modelsummary` 2.0.0 now uses `tinytable` as its default table-drawing
  backend. Learn more at: https://vincentarelbundock.github.io/tinytable/

Revert to `kableExtra` for one session:

  options(modelsummary_factory_default = 'kableExtra')
  options(modelsummary_factory_latex = 'kableExtra')
  options(modelsummary_factory_html = 'kableExtra')

Silence this message forever:

  config_modelsummary(startup_message = FALSE)
```

```r
plot(example$x, example$y, pch = 16, col = "blue",
     main = "Plot with Regression Lines",
     xlab = "Regressor (X)", ylab = "Outcome (Y)")

abline(a = 1, b = 0.5, col = "red", lwd = 2, lty = 2)

ex_model <- lm(example$y ~ example$x)

modelsummary(ex_model)
```

```r
abline(ex_model, col = "green", lwd = 2)

legend("topleft", inset = 0.01,
       legend = c("Known Coefficients", "Fitted Line (lm)"),
       col = c("red", "green"), lty = c(2, 1), lwd = 2,
       cex = 0.8)
```



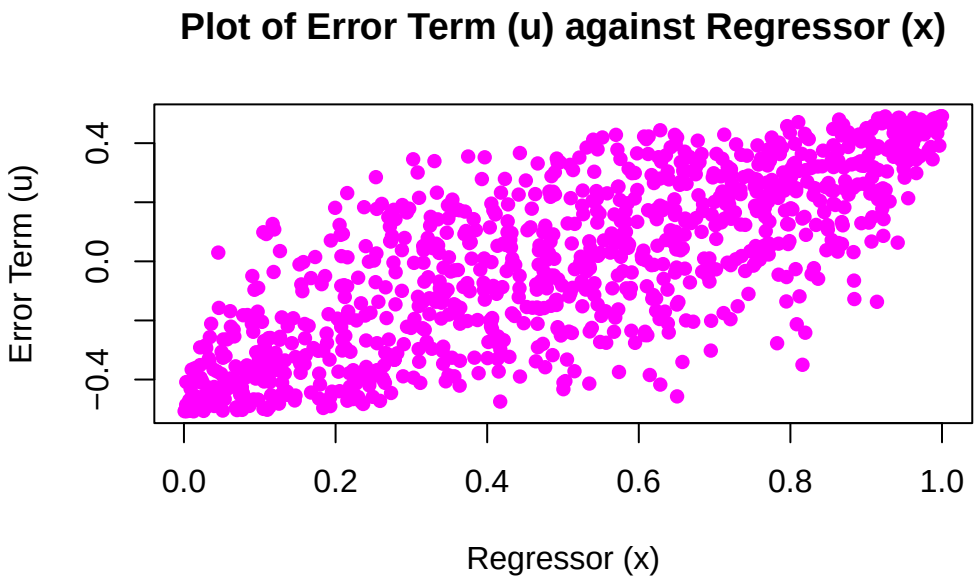
2.2.2

The line produced by lm() is more vertical, or has a steeper incline than the line created by the known-coefficients. The fitted line actually seems to tell a better story of the data because it lines up with the middle/center of the data more. The estimator generated via OLS is biased because there are clusters of data at the top(max) and bottom(min) of the graph.

2.3.1

The correlation (0.79) is significantly different from 0, suggesting that zero conditional mean assumption is violated within this data. When visually looking at the plot, we also see the assumption is violated because there is a discernible pattern where u increases with x. The plot also seems to have heteroskedastic features as there are some clusters at both the min and max, and spread is different across both axes. The fact that u increases with x also suggests non-constant variance.

```r
plot(example$x, example$u, pch = 16, col = "magenta",
     main = "Plot of Error Term (u) against Regressor (x)",
     xlab = "Regressor (x)", ylab = "Error Term (u)")
```




```r
correlation <- cor(example$x, example$u)
print(correlation)
```

```
[1] 0.7861619
```

2.3.2

The coefficients of these two pairs appear to be about half of one another. They also have the same standard errors.The relationship also has bias because the estimates are smaller, and the values are not centered. However, the small standard errors tell us there is not a ton of variance.

```r
ex_m2 <- lm(example$u ~ example$x)

m2_summary <- summary(ex_m2)

intercept <- m2_summary$coefficients[1, 1]
slope <- m2_summary$coefficients[2, 1]
intercept_se <- m2_summary$coefficients[1, 2]
slope_se <- m2_summary$coefficients[2, 2]
```

```r
# from part b

ex_summary <- summary(ex_model)


interceptB <- ex_summary$coefficients[1, 1]
slopeB <- ex_summary$coefficients[2, 1]
intercept_seB <- ex_summary$coefficients[1, 2]
slope_seB <- ex_summary$coefficients[2, 2]


cat("Intercept: ", intercept, "\n")
```

```
Intercept:  -0.3946463
```

```r
cat("Coefficient on X: ", slope, "\n")
```

```
Coefficient on X:  0.7890177
```

```r
cat("Standard error of intercept: ", intercept_se, "\n")
```

```
Standard error of intercept:  0.0113647
```

```r
cat("Standard error of coefficient on X: ", slope_se, "\n")
```

```
Standard error of coefficient on X:  0.01963417
```

```r
cat("Intercept: ", interceptB, "\n")
```

```
Intercept:  0.6053537
```

```r
cat("Coefficient on X: ", slopeB, "\n")
```

```
Coefficient on X:  1.289018
```

```
cat("Standard error of intercept: ", intercept_seB, "\n")
```

```
Standard error of intercept:  0.0113647
```

```
cat("Standard error of coefficient on X: ", slope_seB, "\n")
```

```
Standard error of coefficient on X:  0.01963417
```

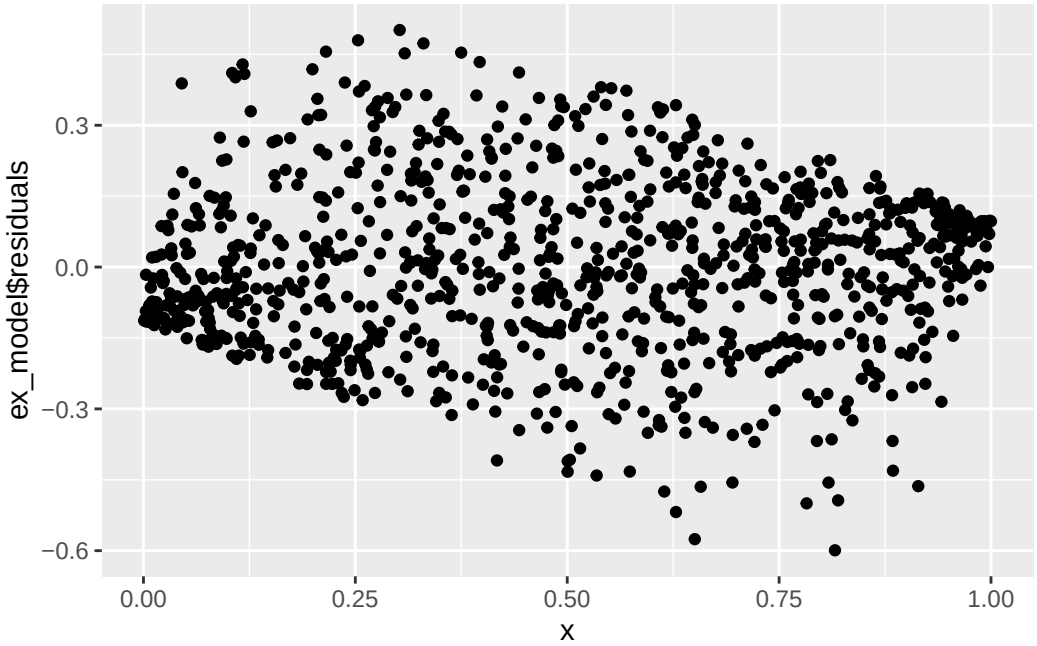
2.3.3

The failure of the zero conditional mean assumption explains why the OLS estimate for _1 in part (b) is biased, as X and u are correlated, because u increases with x. While a violation of the homoskedastic assumption does not bias the estimate, it could affect the reliability of the standard errors.

2.4

(i am hiding the code for the plot because it kept printing background items that made the PDF very long)

```
print(final_plot)
```



2.4.1

After generating the plot, we can tell that the data is now shifted. Therefore our same conclusions apply, however, this view of the data/graph, allows us to easier understand it is heteroskedastic.

2.4.2

I'd tell my friend that the residuals from a regression are always uncorrelated with the predictor because OLS minimizes the sum of squared residuals. It also forces the sample covariance between the residuals and the predictor to be zero. However, this does not mean the true errors (u) are uncorrelated with X. If the zero conditional mean assumption is violated, u and x could still be correlated, meaning there is bias in the OLS estimates. Residuals reflect the model's fit, not the underlying connection between u and x.

Problem 3

3.1

```
library(readr)
nc_precincts <- read_csv("Problem_Set_1/nc_precincts.csv")
```

```
Rows: 2582 Columns: 9
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): precinct_id
dbl (8): total_reg_2022, total_reg_2024, pct_GOP_2022, pct_GOP_2024, pct_tur...

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.
```

3.1.1

```
model_turnout <- lm(pct_turnout_2022 ~ pct_GOP_2022, data = nc_precincts)

summary(model_turnout)
```

```
Call:
lm(formula = pct_turnout_2022 ~ pct_GOP_2022, data = nc_precincts)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53955 -0.06093 -0.00150  0.05852  0.48798

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.412714   0.004342   95.05   <2e-16 ***
pct_GOP_2022 0.337009   0.012881   26.16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09622 on 2580 degrees of freedom
Multiple R-squared:  0.2097,    Adjusted R-squared:  0.2094
F-statistic: 684.5 on 1 and 2580 DF,  p-value: < 2.2e-16
```

```r
model_absentee <- lm(pct_absentee_vbm_2022 ~ pct_GOP_2022, data = nc_precincts)

summary(model_absentee)
```

```
Call:
lm(formula = pct_absentee_vbm_2022 ~ pct_GOP_2022, data = nc_precincts)

Residuals:
      Min        1Q    Median        3Q       Max
-0.054518 -0.015929 -0.003106  0.012628  0.110117

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.056798   0.001033   54.99   <2e-16 ***
pct_GOP_2022 -0.044659   0.003064  -14.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02289 on 2580 degrees of freedom
Multiple R-squared:  0.07608,   Adjusted R-squared:  0.07572
F-statistic: 212.4 on 1 and 2580 DF,  p-value: < 2.2e-16
```

```r
model_early <- lm(pct_early_2022 ~ pct_GOP_2022, data = nc_precincts)

summary(model_early)
```

```
Call:
lm(formula = pct_early_2022 ~ pct_GOP_2022, data = nc_precincts)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54160 -0.06164  0.01413  0.07631  0.26611

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.56855    0.00499  113.93   <2e-16 ***
pct_GOP_2022 -0.22009    0.01480  -14.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1106 on 2580 degrees of freedom
Multiple R-squared:  0.07891,   Adjusted R-squared:  0.07855
F-statistic:   221 on 1 and 2580 DF,  p-value: < 2.2e-16
```

```r
model_election <- lm(pct_election_day_2022 ~ pct_GOP_2022, data = nc_precincts)

summary(model_election)
```

```
Call:
lm(formula = pct_election_day_2022 ~ pct_GOP_2022, data = nc_precincts)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52194 -0.07737 -0.01339  0.06170  0.44563

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.363888   0.005162   70.49   <2e-16 ***
pct_GOP_2022 0.271854   0.015314   17.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	(1)	(2)	(3)	(4)
(Intercept)	0.413	0.057	0.569	0.364
	(0.004)	(0.001)	(0.005)	(0.005)
Coefficient on pct_GOP_2022	0.337	-0.045	-0.220	0.272
	(0.013)	(0.003)	(0.015)	(0.015)
Num.Obs.	2582	2582	2582	2582
R2	0.210	0.076	0.079	0.109
R2 Adj.	0.209	0.076	0.079	0.108
AIC	-4758.1	-12173.9	-4039.7	-3864.7
BIC	-4740.6	-12156.3	-4022.2	-3847.1
Log.Lik.	2382.060	6089.950	2022.869	1935.351
RMSE	0.10	0.02	0.11	0.11

```
Residual standard error: 0.1144 on 2580 degrees of freedom
Multiple R-squared:  0.1088,    Adjusted R-squared:  0.1085
F-statistic: 315.1 on 1 and 2580 DF,  p-value: < 2.2e-16
```

3.1.2

```
library(modelsummary)

model_turnout <- lm(pct_turnout_2022 ~ pct_GOP_2022, data = nc_precincts)
model_absentee <- lm(pct_absentee_vbm_2022 ~ pct_GOP_2022, data = nc_precincts)
model_early <- lm(pct_early_2022 ~ pct_GOP_2022, data = nc_precincts)
model_election <- lm(pct_election_day_2022 ~ pct_GOP_2022, data = nc_precincts)

model_table <- modelsummary(
  list(model_turnout, model_absentee, model_early, model_election),
  coef_rename = c("pct_GOP_2022" = "Coefficient on pct_GOP_2022"),
  output = "gt"
)

model_table
```

3.1.3

For the model with *turnout* as a variable, for every increase in republican voters in a precinct, the total election turnout increases by .0337. The positive sign of this model tells us that precincts that are republican dense tend to have a higher turnout.

For the model with *absentee voters*, the negative sign showcases that republican leaning precincts tend to have lower absentee voting than democratic ones. The magnitude suggests a consistent negative association.

The model looking at *early voting*, or voters, highlights a negative sign, meaning that precincts with higher republican support, early voting is found to be less common. The magnitude -.220 tells us that there is a negative relationship.

The fourth model looking at *election day voting* has a positive coefficient of 0.272. This suggest that as the amount of republicans in a precinct increases, the percent of people who vote on election day also increases. The magnitude tells us that this relationship is moderately positive.

3.1.4

Regarding *R^2*, for the social sciences, their is a rather robust, or strong *R^2* for our first model, but less so for the others. But in general none of our R^2 explains a great majority of the voters, as they are all >0.5, with the two of the 4 models being less than >.1. The model fails to capture other aspects of behavior that may influence these types of voting. In example, civic knowledge, income level, or even race.