Submission for Problem Set 3

Nicholas R. Gonzalez

Worked on with Tanner Bentley.

Problem 1

Problem 1 Part A

Problem 1, Part A, part i

```r
# this is the table for the regression asked for in Part i

library(modelsummary)
```

```
`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
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(haven)

exp_data <- read_csv("pset_3/exp_data.csv")
```

```
Rows: 5593 Columns: 14
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (14): leg_black, treat_out, responded, totalpop, medianhhincom, black_me...

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.
```

```r
respond_out <- lm(exp_data$responded ~ exp_data$treat_out, data = exp_data)

# modelsummary(respond_out)

# creating a table and cleaning it up

responded_model <- modelsummary(
  list("Model 1" = respond_out),
  title = "District Impact on Legislator Email Response",
  coef_map = c("exp_data$treat_out" = "Out-of-District Email"),
  stars = TRUE,
  fmt = 3,
  statistic = "conf.int")

print(responded_model)
```

```
\begin{table}
```

```
\centering
\begin{talltblr}[         %% tabularray outer open
caption={District Impact on Legislator Email Response},
note{}={+ p \num{< 0.1}, * p \num{< 0.05}, ** p \num{< 0.01}, *** p \num{< 0.001}},
]                     %% tabularray outer close
{                     %% tabularray inner open
colspec={Q[]Q[]},
column{2}={}{halign=c,},
column{1}={}{halign=l,},
hline{4}={1,2}{solid, black, 0.05em},
}                     %% tabularray inner close
\toprule
& Model 1 \\ \midrule %% TinyTableHeader
Out-of-District Email & \num{-0.266}***                \\
& [\num{-0.291}, \num{-0.241}] \\
Num.Obs.              & \num{5593}                     \\
R2                    & \num{0.073}                    \\
R2 Adj.               & \num{0.072}                    \\
AIC                   & \num{7568.7}                   \\
BIC                   & \num{7588.6}                   \\
Log.Lik.              & \num{-3781.346}                \\
RMSE                  & \num{0.48}                     \\
\bottomrule
\end{talltblr}
\end{table}
```

The model presented examines the likelihood of response when receiving an email from someone outside the district. Our coefficient, "Out-of-District Email" is -0.266 and is statistically significant at the 1% level ($p < 0.01$). So when holding other factors constant, receiving an email from outside the district is associated with a decrease of 0.266 units in the outcome variable, which is to receive a response, or for the (a) legislative person to email back. The model also shows an R^2 of 0.073. Therefore 7.3% of the variance in the outcome is explained by the predictors included in the model. This means that a fair amount of unobserved factors are also influencing the outcome. For this data, we are using a dummy variable, so the one unit increase would just be if the legislator is black or not, and if the legislator responded or not. But, in short, a one unit increase, so in the case of this model, receiving an email from outside the district, is associated with a small decrease in receiving an email back, or the outcome variable.

	(1)
(Intercept)	0.561
	(0.009)
treat_out	−0.275
	(0.013)
leg_black	−0.097
	(0.036)
treat_out × leg_black	0.128
	(0.052)
Num.Obs.	5593
R2	0.074
R2 Adj.	0.073
AIC	7564.8
BIC	7598.0
Log.Lik.	−3777.415
RMSE	0.48

Problem 1 Part B

Problem 1 Part B, part i

```
respond_black <- lm(responded ~ treat_out * leg_black, data = exp_data)

modelsummary(respond_black)
```

```
# creating a table and cleaning it up

black_leg_model <- modelsummary(
  list("Black Legislator's Response to Emails" = respond_black),
  title = "Black Legislator's Response to Emails",
  statistic = c("std.error", "conf.int"),
  stars = TRUE,
  fmt = 3,
  coef_rename = c(
    "treat_out" = "Out-of-District Email",
    "leg_black" = "Black Legislator"  )  # Clean up the coefficient names, including the inte
)


print(black_leg_model)
```

```
\begin{table}
\centering
\begin{talltblr}[         %% tabularray outer open
caption={Black Legislator's Response to Emails},
note{}={+ p \num{< 0.1}, * p \num{< 0.05}, ** p \num{< 0.01}, *** p \num{< 0.001}},
]                     %% tabularray outer close
{                     %% tabularray inner open
colspec={Q[]Q[]},
column{2}={}{halign=c,},
column{1}={}{halign=l,},
hline{14}={1,2}{solid, black, 0.05em},
}                     %% tabularray inner close
\toprule
& Black Legislator's Response to Emails \\ \midrule %% TinyTableHeader
(Intercept)                                & \num{0.561}***                \\
& (\num{0.009})                 \\
& [\num{0.543}, \num{0.580}]   \\
Out-of-District Email                      & \num{-0.275}***               \\
& (\num{0.013})                 \\
& [\num{-0.300}, \num{-0.249}] \\
Black Legislator                           & \num{-0.097}**                \\
& (\num{0.036})                 \\
& [\num{-0.167}, \num{-0.026}] \\
Out-of-District Email:Black Legislator & \num{0.128}*                  \\
& (\num{0.052})                 \\
& [\num{0.027}, \num{0.229}]   \\
Num.Obs.                                   & \num{5593}                    \\
R2                                         & \num{0.074}                   \\
R2 Adj.                                    & \num{0.073}                   \\
AIC                                        & \num{7564.8}                  \\
BIC                                        & \num{7598.0}                  \\
Log.Lik.                                   & \num{-3777.415}               \\
RMSE                                       & \num{0.48}                    \\
\bottomrule
\end{talltblr}
\end{table}
```

	(1)
(Intercept)	0.423
	(0.007)
poly(medianhhincom, 2)1	1.601
	(0.493)
poly(medianhhincom, 2)2	−1.280
	(0.493)
Num.Obs.	5593
R2	0.003
R2 Adj.	0.003
AIC	7974.7
BIC	8001.2
Log.Lik.	−3983.343
RMSE	0.49

Problem 1 Part B, part ii

The results of this model suggest that receiving an out of district email decreases the likelihood of a legislator (or their staff?) responding by 0.275 units. I however am not sure why this coefficient is different than the first model, when this part of the regression is the same. Nonetheless, the coefficient "leg_black" which references if the legislator was black or not, tells us that black legislators are .097 less likely to respond compared to non-Black legislators. Looking at the interaction, the effect of receiving an out of district email is greater for non-Black legislators, as opposed to black ones. Again, for this data, we are using a dummy variable, so the one unit increase would just be if the legislator is black or not, and if the legislator responded or not.

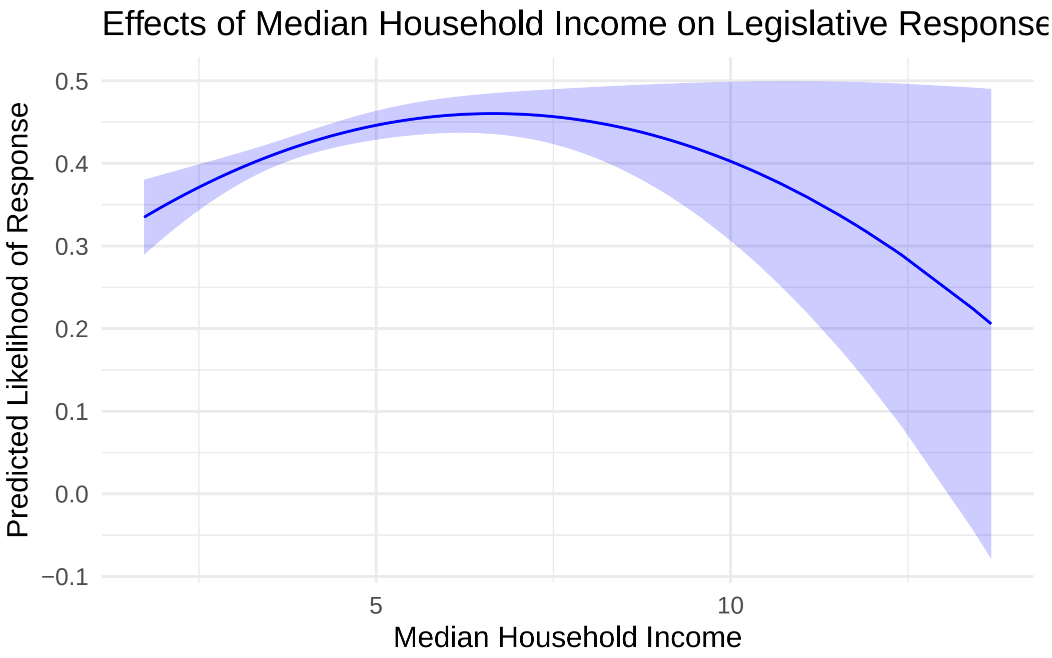
Problem 1 Part C

Problem 1 Part C, part i

```r
respond_income_poly <- lm(responded ~ poly(medianhhincom, 2), data = exp_data) # 2 for secon

modelsummary(respond_income_poly)
```

Problem 1 Part C, part ii

```r
library(ggeffects)

effect_plot <- ggpredict(respond_income_poly, terms = "medianhhincom [all]")

ggplot(effect_plot, aes(x = x, y = predicted)) +
  geom_line(color = "blue") +
  geom_ribbon(aes(ymin = conf.low, ymax = conf.high), fill = "blue", alpha = 0.2) +
  labs(
    title = "Effects of Median Household Income on Legislative Response",
    x = "Median Household Income",
    y = "Predicted Likelihood of Response"
  ) +
  theme_minimal()
```



Problem 1 Part C, part iii

When interpreting non-linear models, sometimes the coefficients can be difficult to understand, or evaluate. However, after graphing the polynomial model, we can see the shape of the plot suggest there is a strong non-linear relationship, where there is diminishing returns of having a higher household income, and a legislator responding. But as the median household income reaches its max, the there is actually less of a likelihood of getting a legislative response than the lowest median income. However, there is also a much larger confidence interval range, suggesting that the data might be inconclusive for higher house hold incomes. This is likely because there is less data on high income households emailing legislators. So there is likely much more variance going on. The very low R^2 even by social science standards at .003 is very telling howeverl

Problem 1 Part C, part iv

The linearity assumption in OLS refers to a model being linear in its parameters, not necessarily between predictors and the outcome. By using a polynomial model, I am modeling a non-linear relationship between median household income and the legislative response. Since the regression remains linear in the coefficients (because of no continous variables), I am not violating OLS's linearity assumption. Lastly, uusing a higher-order term captures curvature in the data while maintaining structure for OLS.

Problem 1 Part D

Problem 1 Part D, part i & ii

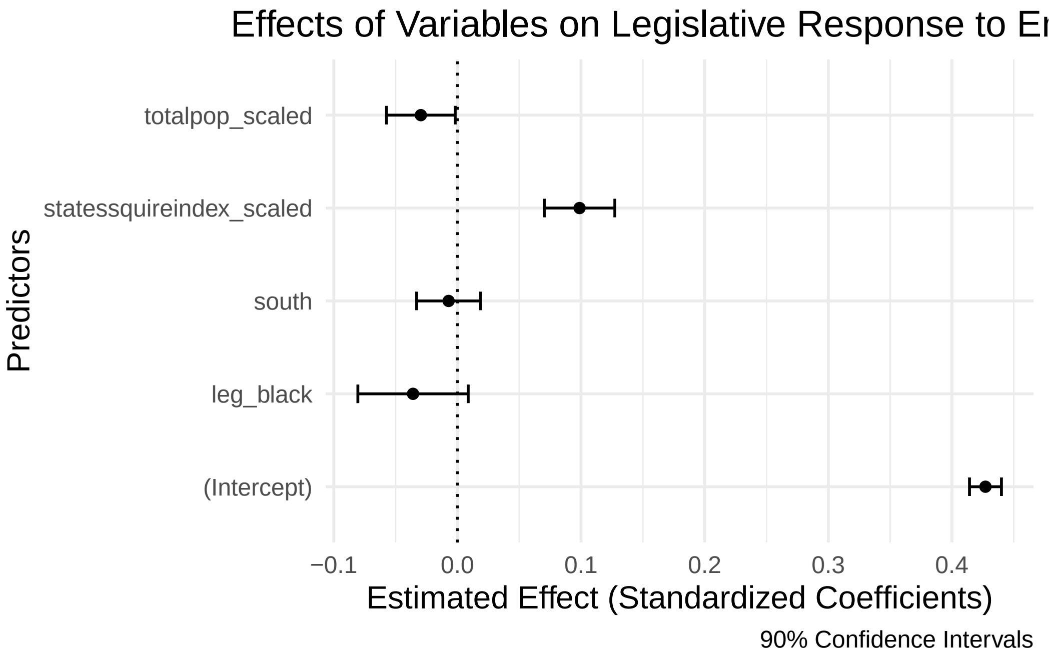
```r
# standardizing the data with scale() which is in base r

exp_data$statessquireindex_scaled <- scale(exp_data$statessquireindex) / 2

exp_data$totalpop_scaled <- scale(exp_data$totalpop) / 2

scaled_response_model <- lm(responded ~ leg_black + statessquireindex_scaled + south + total

summary(scaled_response_model)
```

```
Call:
lm(formula = responded ~ leg_black + statessquireindex_scaled + 
    south + totalpop_scaled, data = exp_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5539 -0.4192 -0.3800  0.5748  0.6700 

Coefficients:
```

```
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               0.427104   0.007860  54.341  < 2e-16 ***
leg_black                -0.035899   0.027125  -1.323    0.186    
statessquireindex_scaled  0.098789   0.017336   5.699 1.27e-08 ***
south                    -0.007125   0.015722  -0.453    0.650    
totalpop_scaled          -0.029576   0.016894  -1.751    0.080 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4923 on 5588 degrees of freedom
Multiple R-squared:  0.007831,  Adjusted R-squared:  0.00712 
F-statistic: 11.03 on 4 and 5588 DF,  p-value: 6.61e-09
```

Problem 1 Part D, part iii

```r
library(broom)
library(ggplot2)

model_tidy <- tidy(scaled_response_model, conf.int = TRUE, conf.level = 0.90)

#tidyverse plot

ggplot(model_tidy, aes(x = estimate, y = term, xmin = conf.low, xmax = conf.high)) +
  geom_point() +
  geom_errorbarh(height = 0.2) +
  geom_vline(xintercept = 0, linetype = "dotted") +
  labs(
    title = "Effects of Variables on Legislative Response to Emails",
    x = "Estimated Effect (Standardized Coefficients)",
    y = "Predictors",
    caption = "90% Confidence Intervals"
  ) +
  theme_minimal() +
  theme(
    axis.title.y = element_text(size = 12),
    axis.title.x = element_text(size = 12),
    plot.title = element_text(hjust = 0.5, size = 14)
  )
```



Problem 1 Part D, part iv

The Squire Index stands out as the most significant factor. This finding also theoretically makes sense. The data defines the Squire Index as a *"measure of professionalization of state legislatures that ranges from 0-100"*, which accounts for legislators' salaries, staff size, and their legislative session lengths. A more professionalized legislature likely means legislators have more resources and availability, making them more inclined to respond to emails. Or have staff who can respond to emails.