Problem Set 9

Due date: 4 December

Please upload your completed assignment to the ELMs course site (under the assignments menu). Remember to include an annotated script file for all work with R and show your math for all other problems (if applicable, or necessary). Please also upload your completed assignment to the Github repository that you have shared with us. *We should be able to run your script with no errors.*

Total points: 40

Question 1

Points: 10

Table 1 below reports the results from two regression models. In Model 1, in Table 1, Y is regressed on X_1 and, in Model 2, Y is regressed on both X_1 and X_2. Why might X_1 be statistically significant at conventional levels in Model 1 but statistically insignificant in Model 2? Be as specific as possible.

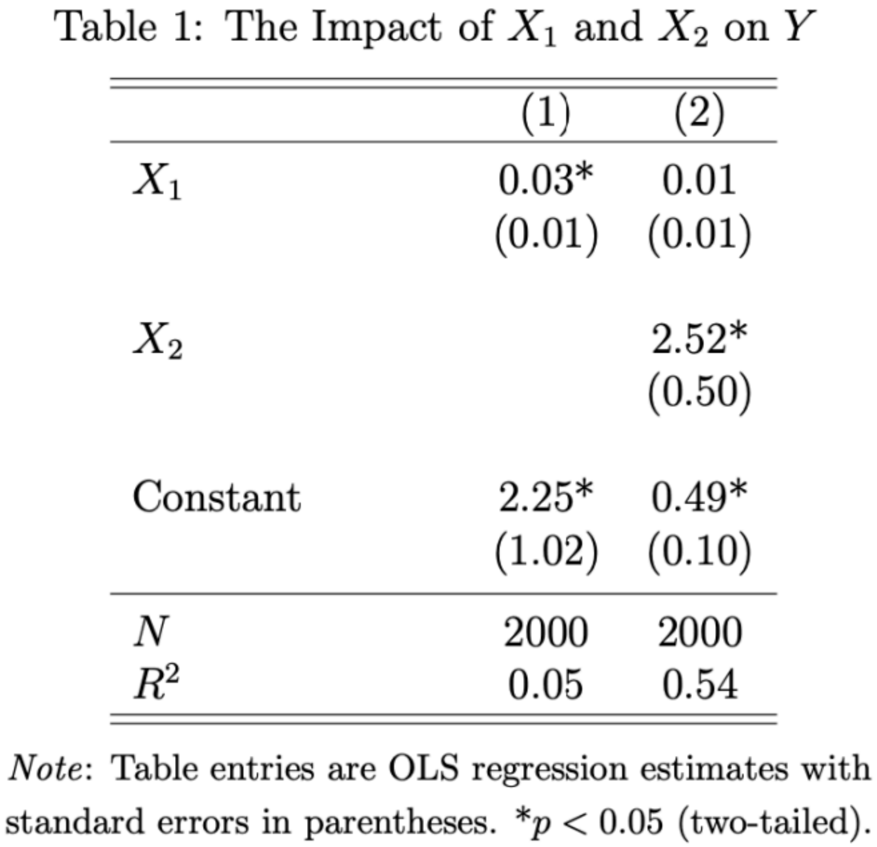


Table 1: The Impact of X_1 and X_2 on Y

	(1)	(2)
X_1	0.03*	0.01
	(0.01)	(0.01)
X_2		2.52*
		(0.50)
Constant	2.25*	0.49*
	(1.02)	(0.10)
N	2000	2000
R^2	0.05	0.54

Note: Table entries are OLS regression estimates with standard errors in parentheses. $^*p < 0.05$ (two-tailed).

Ans 1:

This usually indicates a situation where a variable is being suppressed by another variable. This can cause changes in statistical significance between the variables as well as causing changes in estimates.

There are a couple reasons why this can happen:

1. **Multicollinearity**: If X1 and X2 are highly correlated, they may provide redundant information when included together in the model. This can inflate the variance of the estimated regression coefficients, leading to higher standard errors and thus lower statistical significance.
2. **Mediation**: X2 could be a mediator variable that explains the relationship between X1 and Y. When X2 is not included in the model, the effect of X1 on Y might appear to be significant. However, once X2 is included and accounts for some or all of X1's effect on Y, the direct effect of X1X1 may no longer be significant.

3. **Specification Error**: Omitting a relevant variable (like X2 in Model 1) can lead to a misspecified model that falsely attributes the effect of X2 to X1. When X2 is included in Model 2, the estimates are corrected, and X1 may no longer be significant.
4. **Change in the Sample Size or Data Structure**: If the inclusion of X2 leads to a reduction in the sample size due to missing data, the reduced power could affect the significance of X1.
5. **Suppressor Effect**: X2 might be a suppressor variable, which means its inclusion in the model increases the predictive validity of X1, but due to the unique relationships among the variables, it might make X1 statistically insignificant.

In the given table, the increase in the R-squared value from 0.05 in Model 1 to 0.54 in Model 2 with the inclusion of X2 also indicates that X2 explains a substantial amount of the variation in Y that X1 does not, which could overshadow the effect of X1.

Question 2

Points: 10

Using the `censusAggregate` dataset (posted on ELMs) — which is survey data aggregated to the state level (1972-2000) — estimate a regression with `vote` as the dependent variable and the following independent variables: `nonSouth`, `edr`, and `pcthsg`. Report the results in a professionally formatted table and interpret the regression results.

Also, create a figure to display the predicted values (with confidence intervals) for the effect of `pcthsg` on the turnout rate. Lastly, is it meaningful to interpret the constant term on its own? Why, or why not?

> ℹ Note
>
> `vote` is the turnout rate in a state in a given year (i.e., the number of people who voted divided by the number eligible to vote).
> `nonSouth` is a dummy variable equal to `0` for Southern states and a `1` for non-Southern states.
> `pcthsg` is the percentage of the population in a state that graduated high school.
> `edr` is a dummy variable equal to `1` for states that used election-day registration and a `0` for states without election-day registration.

Ans 2:

```r
library(gtsummary)

data <- read.csv("Census.csv")
```

```r
model <- lm(vote ~ nonSouth + edr + pcthsg, data = data)
summary(model)
```

```

Call:
lm(formula = vote ~ nonSouth + edr + pcthsg, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.6259  -3.3554   0.0282   3.6670  15.9988 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 54.00180    2.61603  20.643  < 2e-16 ***
nonSouth     5.54555    0.79695   6.958 1.68e-11 ***
edr          5.79234    1.05645   5.483 7.99e-08 ***
pcthsg       0.10074    0.03615   2.787  0.00561 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.652 on 353 degrees of freedom
Multiple R-squared:  0.2813,	Adjusted R-squared:  0.2752 
F-statistic: 46.05 on 3 and 353 DF,  p-value: < 2.2e-16
```

```r
table <- gtsummary::tbl_regression(model)
table
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**Beta**	**95% CI**	**p-value**
nonSouth	5.5	4.0, 7.1	<0.001
edr	5.8	3.7, 7.9	<0.001
pcthsg	0.10	0.03, 0.17	0.006

Key Findings:

- **Model Formula**: `vote = 54.002 + 5.546*nonSouth + 5.792*edr + 0.101*pcthsg`

- **Key Findings**:
  - **nonSouth**: Being outside the South increases `vote` by about 5.55 points, statistically significant.
  - **edr**: Each unit increase in `edr` raises `vote` by approximately 5.79 points, statistically significant.
  - **pcthsg**: A one percentage point increase in `pcthsg` corresponds to a 0.10 point increase in `vote`, statistically significant.
- **Model Fit**:
  - **R-squared**: 28.13%, indicating the model explains about 28% of the variability in `vote`.
  - **Overall Significance**: The model is statistically significant (F-statistic p-value < 2.2e-16).

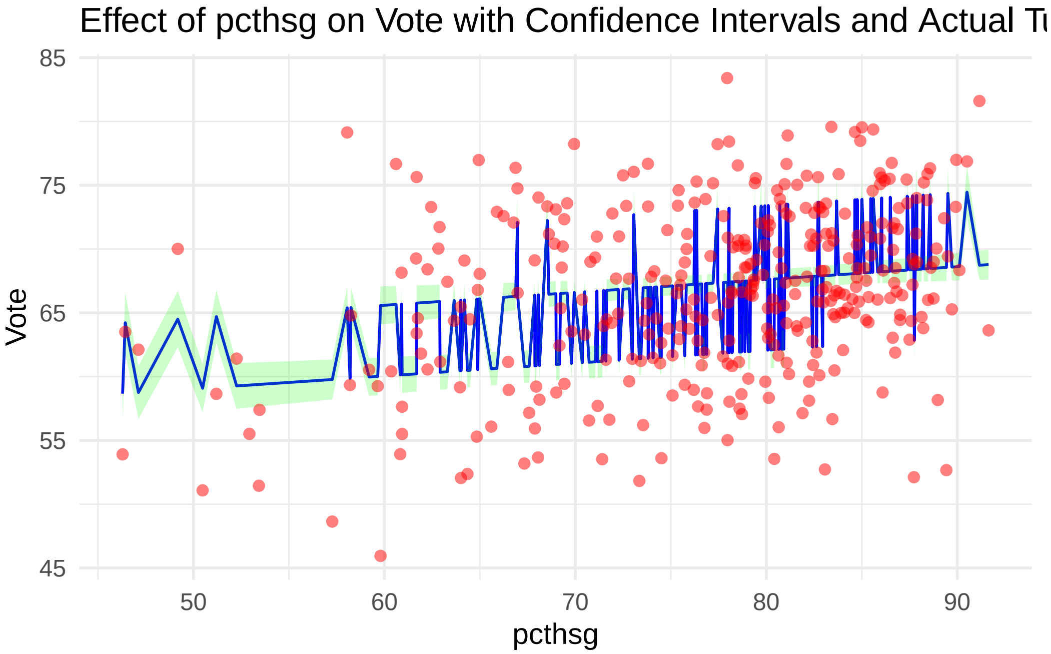
This model shows that `nonSouth`, `edr`, and `pcthsg` are significant predictors of `vote`, but it leaves a substantial portion of `vote`'s variability unexplained.

To display the effects of a single variable we can set the others to zero, usually not the best way as there are effects of those other variables. So I am electing to set the other values to their mean value and then driving prediction from that.

```r
# Set other predictors at their means
data$nonSouth_mean <- mean(data$nonSouth)
data$edr_mean <- mean(data$edr)

# Generate predictions
predictions <- predict(model, newdata = data, interval = "confidence")

# Combine with pcthsg for plotting
plot_data <- data.frame(pcthsg = data$pcthsg, predictions)
```

Plotting:

```r
library(ggplot2)

ggplot(plot_data, aes(x = pcthsg)) +
  geom_line(aes(y = fit), color = "blue") +  # Predicted values
  geom_ribbon(aes(ymin = lwr, ymax = upr), alpha = 0.2, fill = "green") +  # Confidence in
  geom_point(aes(y = data$vote), color = "red", alpha = 0.5) +  # Actual turnout rates
  labs(x = "pcthsg", y = "Vote", title = "Effect of pcthsg on Vote with Confidence Interva
  theme_minimal()
```



The chart denotes the effect of variable on the vote as well as seeing the actual values. As seen it does have significant predictive effect. We can see the confidence interval highlighted in green and the blue as the predicted value.

It would be beneficial to see the total predicted value as well, but the setting the other variables to their means is a good way to see the raw effects of those variables with the variability removed.

Intercept term interpretation:

The Intercept in a regression model is meaningful in contexts where it makes sense to have all predictor variables equal to zero. In this model, the intercept represents the predicted value of `vote` when `nonSouth`, `edr`, and `pcthsg` are all zero.

However, interpreting the constant term on its own might not always be meaningful, especially if the zero values of the predictors are outside their realistic range. For example, if having `nonSouth`, `edr`, and `pcthsg` all at zero is not a practical scenario, then the intercept does not represent a realistic situation. Therefore, its interpretation is more theoretical than practical.

In cases where the zero values of predictors are not meaningful or possible (e.g., negative values for inherently positive quantities), the intercept serves more as a statistical adjustment to ensure the model works well with other values of the predictors. In conclusion in this case, it does not make much sense to interpret it in this case as the variables we have especially things like NonSouth do not make sense being zero.

Question 3

Points: 5

Using the regression results from the previous question, evaluate the null hypothesis that the effects (i.e., regression coefficients) of `nonSouth` and `pcthsg` are jointly equal to zero. Can you reject the null hypothesis? Be sure to demonstrate how you reached a definitive conclusion.

Ans 3:

There are many ways to do this, I'm going to prove this using an F test.

Let my Null hypothesis be: Beta nonSouth = Beta Pcthsg = 0

Alternative: At least one of the betas is non-zero

Steps:

I already have a full model, I'm going to run a reduced model now

```
reduced_model <- lm(vote ~ edr, data = data)
```

Now we are running an Anova to see if the reduced model is significantly different that the full model that was previously run.

```
f_test_result <- anova(reduced_model, model)
f_test_result
```

```
Analysis of Variance Table

Model 1: vote ~ edr
Model 2: vote ~ nonSouth + edr + pcthsg
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    355 13919
2    353 11277  2    2641.8 41.347 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the very significant Pvalue here, we can reject the null hypothes. This means that adding `nonSouth` and `pcthsg` to the model significantly improves its fit, indicating that these variables have a significant joint effect on the dependent variable `vote`.

Question 4

Points: 15

Using one of the other datasets available in the `poliscidata` package pick one dependent variable and two or more independent variables. Run a regression of the dependent variable on the independent variables. In your answer, describe why you picked the variables you did, produce a professionally formatted results table, and describe your statistical and substantive findings.

Ans 4:

```r
library(poliscidata)
```

```
Registered S3 method overwritten by 'gdata':
  method         from
  reorder.factor gplots
```

```r
library(gtsummary)

# Load the 'world' dataset
data(world)

# Run the linear regression model
model <- lm(dem_score14 ~ gdp08 + literacy, data = world)

# Create a professionally formatted table using gtsummary
table <- tbl_regression(model)

# Print the table
table
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	Beta	95% CI	p-value
gdp08	0.00	0.00, 0.00	0.3
literacy	0.05	0.03, 0.07	<0.001

Why I picked the variables:

I wanted to see if there is a significant effect between GDP and literacy with the strength of democracy.

https://theconversation.com/are-poor-societies-stuck-with-dictators-57145. Shows that poverty is not a problem for democracy or literacy , but we can test it with data. I've picked 2008 just to test it for that time.

Dependent Variable:

- **Democracy Level**: A common measure of a country's democratic status, often quantified through various indexes.

Independent Variables:

- **GDP per Capita**: Economic prosperity can influence democratic structures.
- **Literacy Rate**: Higher literacy rates might correlate with stronger democratic institutions.

Hypothesis:

- **GDP per Capita**: Wealthier countries might have more resources to support democratic institutions.
- **Literacy Rate**: Education and literacy could foster political awareness and participation, reinforcing democracy.

Findings:

- **GDP in 2008 (`gdp08`)**: The coefficient is 0.00, with a 95% confidence interval ranging from 0.00 to 0.00, and a p-value of 0.3. This suggests that GDP in 2008 does not have a statistically significant impact on the democracy score. The effect size is also negligible.
- **Literacy Rate (`literacy`)**: The coefficient is 0.05, with a 95% confidence interval ranging from 0.03 to 0.07, and a p-value of less than 0.001. This indicates that higher literacy rates are significantly associated with higher democracy scores, with a moderate effect size.

Conclusion

While the GDP in 2008 doesn't seem to significantly influence a country's democracy score, the literacy rate appears to be a meaningful predictor, with higher literacy rates associated with higher democracy scores. This finding underscores the potential importance of education in supporting democratic structures.

If we really want to see the effects, we should run a full analysis with multiple variables with variable selection.