Problem Set 3

Applied Stats/Quant Methods 1

Due: November 11, 2024

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub.
- This problem set is due before 23:59 on Sunday November 11, 2024. No late assignments will be accepted.

In this problem set, you will run several regressions and create an add variable plot (see the lecture slides) in `R` using the `incumbents_subset.csv` dataset. Include all of your code.

Question 1

We are interested in knowing how the difference in campaign spending between incumbent and challenger affects the incumbent's vote share.

1. Run a regression where the outcome variable is `voteshare` and the explanatory variable is `difflog`.

```r
# Extract the dependent and independent variables
Y <- inc.sub$voteshare
anes <- anes[complete.cases(anes$caseid), ]
X <- inc.sub$difflog
n <- length(Y)

# Add intercept term
X_matrix <- cbind(1, X)

# Transpose the X matrix
```

```r
Xt <- t(X_matrix)

# Calculate X'X
XtX <- Xt %*% X_matrix

# Calculate the inverse of (X'X)
XtX_inv <- solve(XtX)

# Calculate X'Y
XtY <- Xt %*% Y

# Calculate the estimated beta coefficients
beta_hat <- XtX_inv %*% XtY

# Print the regression coefficients
```

So the regression coefficient is:

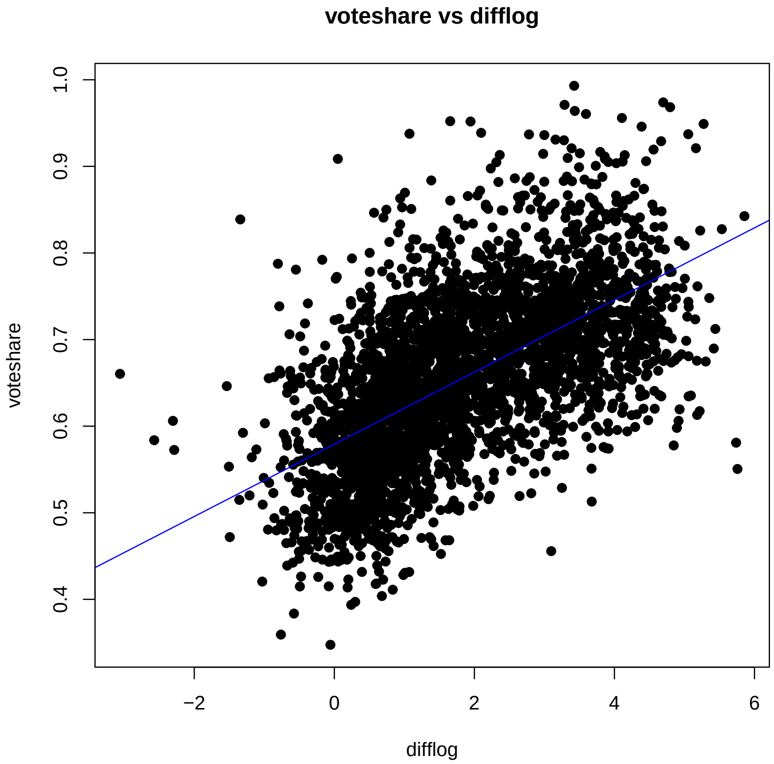
$$\hat{\beta}_0 = 0.57903071 \text{ (intercept)}$$
$$\hat{\beta}_1 = 0.04166632 \text{ (slope of difflog)}$$

2. Make a scatterplot of the two variables and add the regression line.

```r

# Plot the scatter plot
pdf("scatter_plot1.pdf")
plot(X, Y, main="voteshare vs difflog",
     xlab="difflog", ylab="voteshare", pch=19)

# Add the regression line
abline(a=beta_hat[1], b=beta_hat[2], col="blue")
```



3. Save the residuals of the model in a separate object.

```

# Calculate the predicted values
Y_hat <- X_matrix %*% beta_hat

# Calculate the residuals
residuals <- Y - Y_hat

# Save the residuals
residuals_model1 <- residuals

# View the first few residuals
```

4. Write the prediction equation.

Based on the calculated regression coefficients, the prediction equation is:

$$vote\hat{s}hare = \hat{\beta}_0 + \hat{\beta}_1 \times difflog$$

The specific values are:

$$vote\hat{s}hare = 0.57903071 + 0.04166632 \times difflog$$

Question 2

We are interested in knowing how the difference between incumbent and challenger's spending and the vote share of the presidential candidate of the incumbent's party are related.

1. Run a regression where the outcome variable is `presvote` and the explanatory variable is `difflog`.

```r
# Extract the dependent and independent variables
Y <- inc.sub$presvote
X <- inc.sub$difflog
n <- length(Y)

# Add intercept term
X_matrix <- cbind(1, X)

# Transpose the X matrix
Xt <- t(X_matrix)

# Calculate X'X
XtX <- Xt %*% X_matrix

# Calculate the inverse of (X'X)
XtX_inv <- solve(XtX)

# Calculate X'Y
XtY <- Xt %*% Y

# Calculate the estimated beta coefficients
beta_hat <- XtX_inv %*% XtY

# Print the regression coefficients
print(beta_hat)
```

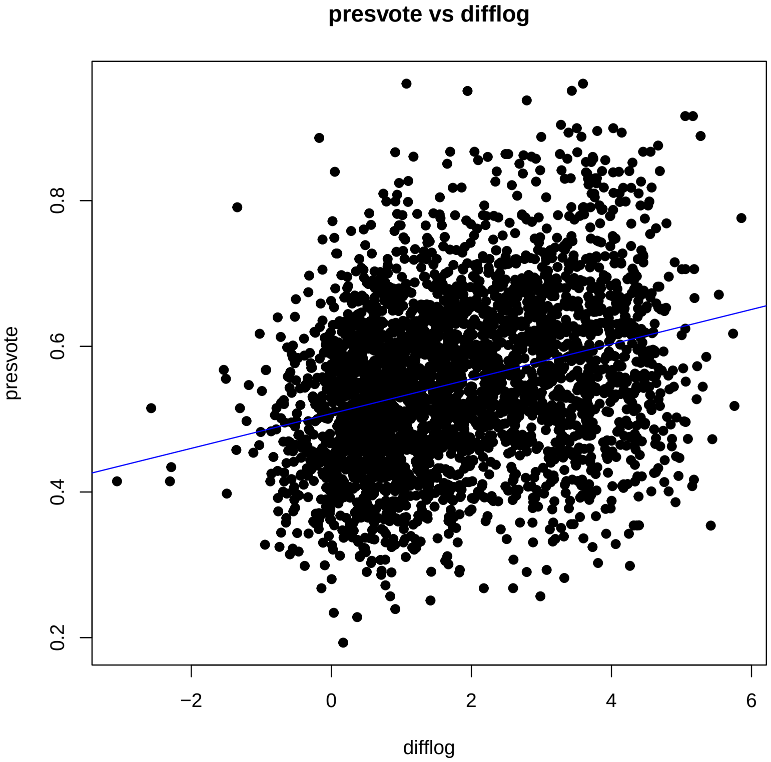
So the regression coefficient is:

$$\hat{\beta}_0 = 0.50758333 \text{ (intercept)}$$
$$\hat{\beta}_1 = 0.02383723 \text{ (slope of difflog)}$$

2. Make a scatterplot of the two variables and add the regression line.

```r
# Plot the scatter plot
pdf("scatter_plot2.pdf")
plot(X, Y, main="presvote vs difflog",
      xlab="difflog", ylab="presvote", pch=19)

# Add the regression line
abline(a=beta_hat[1], b=beta_hat[2], col="blue")
dev.off()
```



3. Save the residuals of the model in a separate object.

```
# Calculate the predicted values
Y_hat <- X_matrix %*% beta_hat

# Calculate the residuals
residuals <- Y - Y_hat

# Save the residuals
residuals_model2 <- residuals

# View the first few residuals
head(residuals_model2)
```

4. Write the prediction equation.

Based on the calculated regression coefficients, the prediction equation is:

$$\hat{voteshare} = \hat{\beta}_0 + \hat{\beta}_1 \times difflog$$

The specific values are:

$$\hat{voteshare} = 0.50758333 + 0.02383723 \times difflog$$

Question 3

We are interested in knowing how the vote share of the presidential candidate of the incumbent's party is associated with the incumbent's electoral success.

1. Run a regression where the outcome variable is `voteshare` and the explanatory variable is `presvote`.

```r
# Extract the dependent and independent variables
Y <- inc.sub$voteshare
X <- inc.sub$presvote
n <- length(Y)

# Add intercept term
X_matrix <- cbind(1, X)

# Transpose the X matrix
Xt <- t(X_matrix)

# Calculate X'X
XtX <- Xt %*% X_matrix

# Calculate the inverse of (X'X)
XtX_inv <- solve(XtX)

# Calculate X'Y
XtY <- Xt %*% Y

# Calculate the estimated beta coefficients
beta_hat <- XtX_inv %*% XtY

# Print the regression coefficients
print(beta_hat)
```

So the regression coefficient is:

$$\hat{\beta}_0 = 0.4413299 \text{ (intercept)}$$

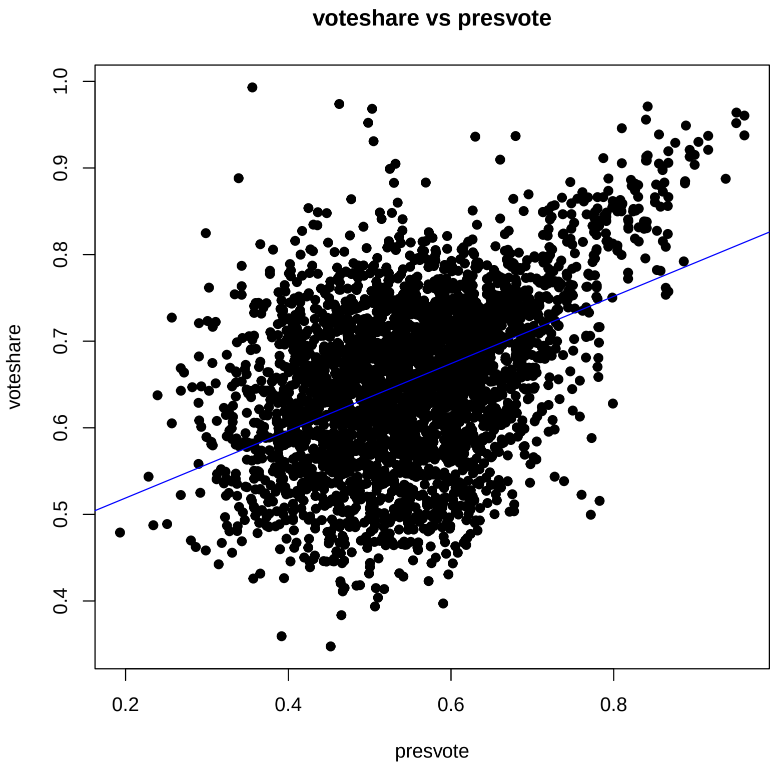
$$\hat{\beta}_1 = 0.3880184 \text{ (slope of presvote)}$$

2. Make a scatterplot of the two variables and add the regression line.

```r
pdf("scatter_plot3.pdf")
plot(X, Y, main="voteshare vs presvote",
     xlab="presvote", ylab="voteshare", pch=19)

```

```
# Add the regression line
abline(a=beta_hat[1], b=beta_hat[2], col="blue")
dev.off()
```



3. Write the prediction equation.

Based on the calculated regression coefficients, the prediction equation is:

$$\hat{voteshare} = \hat{\beta}_0 + \hat{\beta}_1 \times difflog$$

The specific values are:

$$\hat{voteshare} = 0.4413299 + 0.3880184 \times difflog$$

Question 4

The residuals from part (a) tell us how much of the variation in `voteshare` is *not* explained by the difference in spending between incumbent and challenger. The residuals in part (b) tell us how much of the variation in `presvote` is *not* explained by the difference in spending between incumbent and challenger in the district.

1. Run a regression where the outcome variable is the residuals from Question 1 and the explanatory variable is the residuals from Question 2.

```r
# Ensure that residuals_model1 and residuals_model2 have the same length
length(residuals_model1)
length(residuals_model2)

# Assign the dependent and independent variables
Y_resid <- residuals_model1
X_resid <- residuals_model2
n <- length(Y_resid)

# Add intercept term
X_matrix_resid <- cbind(1, X_resid)

# Transpose the X matrix
Xt_resid <- t(X_matrix_resid)

# Calculate X'X
XtX_resid <- Xt_resid %*% X_matrix_resid

# Calculate the inverse of (X'X)
XtX_inv_resid <- solve(XtX_resid)

# Calculate X'Y
XtY_resid <- Xt_resid %*% Y_resid

# Calculate the estimated beta coefficients
beta_hat_resid <- XtX_inv_resid %*% XtY_resid

# Print the regression coefficients
print(beta_hat_resid)
```

So the regression coefficient is:

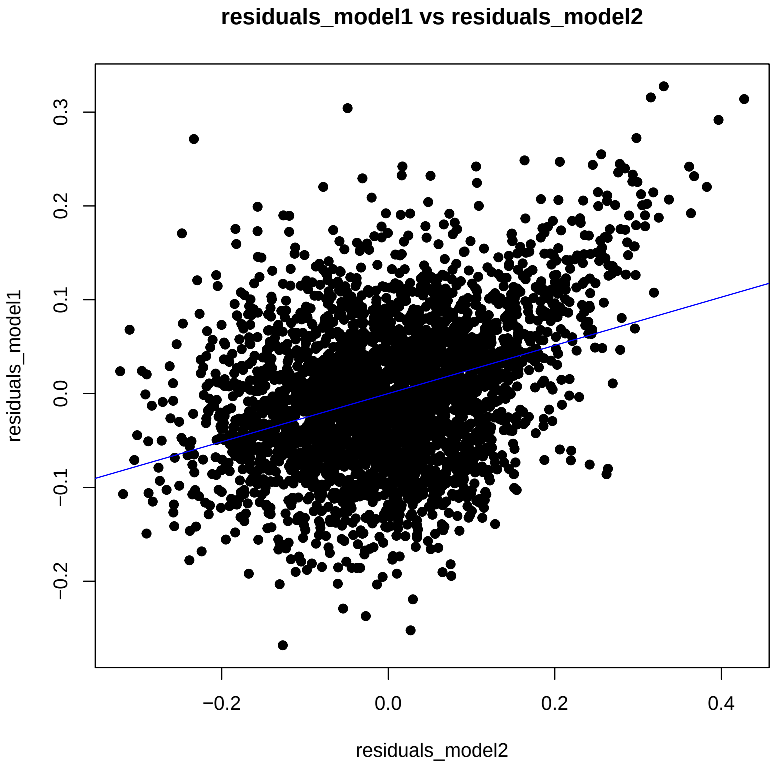
$$\hat{\beta}_0 = 5.326175 \times 10^{-17} \text{ (intercept)}$$

$$\hat{\beta}_1 = 0.3880184 \text{ (slope of presvote)}$$

2. Make a scatterplot of the two residuals and add the regression line.

```
# Plot the scatter plot
pdf("scatter_plot4.pdf")
plot(X_resid, Y_resid, main="residuals_model1 vs residuals_model2",
     xlab="residuals_model2", ylab="residuals_model1", pch=19)

# Add the regression line
abline(a=beta_hat_resid[1], b=beta_hat_resid[2], col="blue")
dev.off()
```



3. Write the prediction equation.

   Based on the calculated regression coefficients, the prediction equation is:

$$\hat{e}_{model1} = \hat{\beta}_0 + \hat{\beta}_1 \times e_{model2}$$

   With specific values:

$$\hat{e}_{model1} = 5.326175 \times 10^{-17} + 0.2568770 \times e_{model2}$$

   After simplification:

$$\hat{e}_{model1} \approx 0 + 0.2568770 \times e_{model2}$$

   Therefore:

$$\hat{e}_{model1} = 0.2568770 \times e_{model2}$$

Question 5

What if the incumbent's vote share is affected by both the president's popularity and the difference in spending between incumbent and challenger?

1. Run a regression where the outcome variable is the incumbent's `voteshare` and the explanatory variables are `difflog` and `presvote`.

```r
# Assign the dependent and independent variables
Y <- inc.sub$voteshare
X1 <- inc.sub$difflog
X2 <- inc.sub$presvote
n <- length(Y)

# Add intercept term
X_matrix <- cbind(1, X1, X2)

# Transpose the X matrix
Xt <- t(X_matrix)

# Calculate X'X
XtX <- Xt %*% X_matrix

# Calculate the inverse of (X'X)
XtX_inv <- solve(XtX)

# Calculate X'Y
XtY <- Xt %*% Y

# Calculate the estimated beta coefficients
beta_hat <- XtX_inv %*% XtY

# Print the regression coefficients
print(beta_hat)
```

So the regression coefficient is:

$$\hat{\beta}_0 = 0.44864422 \text{ (intercept)}$$
$$\hat{\beta}_1 = 0.03554309 \text{ (slope of difflog)}$$
$$\hat{\beta}_2 = 0.25687701 \text{ (slope of presvote)}$$

2. Write the prediction equation.

Based on the calculated regression coefficients, the prediction equation is:

$$vote\hat{s}hare = \hat{\beta}_0 + \hat{\beta}_1 \times difflog + \hat{\beta}_2 \times presvote$$

With specific values:

$$\hat{voteshare} = 0.44864422 + 0.03554309 \times difflog + 0.25687701 \times presvote$$

3. What is it in this output that is identical to the output in Question 4? Why do you think this is the case?

Here's the English translation and LaTeX code:
In regression model of Question 5, which includes two independent variables difflog and presvote, we obtained the regression coefficient for presvote:

$$\hat{\beta}_2 = 0.25687701$$

This is identical to the regression coefficient of presvote (0.256877) in Question 4.
Explanation:
According to the properties of multiple linear regression, when we include multiple independent variables in a regression model, each variable's regression coefficient reflects its independent effect on the dependent variable after controlling for other independent variables. In Question 4, we effectively regressed residuals_model1 (voteshare residuals after controlling for difflog) on residuals_model2 (presvote residuals after controlling for difflog), where the regression coefficient reflected the effect of presvote on voteshare after controlling for difflog.
Therefore, in Question 5, when we simultaneously regress voteshare on difflog and presvote, the regression coefficient of presvote (0.25687701) is exactly the same as its effect coefficient on voteshare in Question 4. This is because in multiple regression models, presvote's regression coefficient represents its independent effect on voteshare after controlling for difflog, consistent with the results from the residual regression in Question 4.