Problem Set 2

Applied Stats II

Due: February 18, 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 in `.pdf` form.
- This problem set is due before 23:59 on Sunday February 18, 2024. No late assignments will be accepted.
- Total available points for this homework is 80.

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're interested in what types of international environmental agreements or policies people support (Bechtel and Scheve 2013). So, we asked 8,500 individuals whether they support a given policy, and for each participant, we vary the (1) number of countries that participate in the international agreement and (2) sanctions for not following the agreement.

Load in the data labeled `climateSupport.RData` on GitHub, which contains an observational study of 8,500 observations.

- Response variable:
  - `choice`: 1 if the individual agreed with the policy; 0 if the individual did not support the policy
- Explanatory variables:
  - `countries`: Number of participating countries [20 of 192; 80 of 192; 160 of 192]
  - `sanctions`: Sanctions for missing emission reduction targets [None, 5%, 15%, and 20% of the monthly household costs given 2% GDP growth]

Please answer the following questions:

1. Remember, we are interested in predicting the likelihood of an individual supporting a policy based on the number of countries participating and the possible sanctions for non-compliance.
   (a) Fit an additive model. Provide the summary output, the global null hypothesis, and p-value. Please describe the results and provide a conclusion.
   (b) How many iterations did it take to find the maximum likelihood estimates?
2. If any of the explanatory variables are significant in this model, then:
   (a) For the policy in which nearly all countries participate [160 of 192], how does increasing sanctions from 5% to 15% change the odds that an individual will support the policy? (Interpretation of a coefficient)
   (b) For the policy in which very few countries participate [20 of 192], how does increasing sanctions from 5% to 15% change the odds that an individual will support the policy? (Interpretation of a coefficient)
   (c) What is the estimated probability that an individual will support a policy if there are 80 of 192 countries participating with no sanctions?
   (d) Would the answers to 2a and 2b potentially change if we included the interaction term in this model? Why?
      - Perform a test to see if including an interaction is appropriate.

1. Remember, we are interested in predicting the likelihood of an individual supporting a policy based on the number of countries participating and the possible sanctions for non-compliance.

(a). Fit an additive model. Provide the summary output, the global null hypothesis, and p-value. Please describe the results and provide a conclusion.

```r
# check the details of the dataframe
summary(climateSupport)

# check if the type of variable is factor
var_types_0 <- sapply(climateSupport, str)
# countries and sanctions are both Ord.factor
# prepare the to be converted variables
convert <- c("countries", "sanctions")

# use for loop to convert variables
climateSupport$countries <- factor(climateSupport$countries,
                           levels = c("20 of 192", "80 of 192", "160 of 192"),
                           ordered = FALSE)
climateSupport$sanctions <- factor(climateSupport$sanctions,
                           levels = c("None", "5%", "15%", "20%"),
                           ordered = FALSE)

# check the type of variables
var_types_1 <- sapply(climateSupport, str)

# because the response variable is binary, so choose logistic regression here
q1mod <- glm(choice ~ ., # Y and Xs
            data = climateSupport, # select dataset
            family = "binomial") # select method as binomial
# summary the model
summary(q1mod)
```

Table 1: Outcome variable is `choice` and the explanatory variables are `countries` and `sanctions`

	Dependent variable:
	choice
countries80 of 192	0.336^{***}
	(0.054)
countries160 of 192	0.648^{***}
	(0.054)
sanctions5%	0.192^{***}
	(0.062)
sanctions15%	-0.133^{**}
	(0.062)
sanctions20%	-0.304^{***}
	(0.062)
Constant	-0.273^{***}
	(0.054)
Observations	8,500
Log Likelihood	−5,784.130
Akaike Inf. Crit.	11,580.260
Note:	*p<0.1; **p<0.05; ***p<0.01

There is a **positive** and statistically reliable relationship between the `choice` and the levels 80 of 192, 160 of 192 in `countries`, and the level 5% in `sanctions`. There is a **negative** and statistically reliable relationship between the `choice` and the level 15%, 20% in `sanctions`.
Below is the description under certain level of sanctions situations:

- For a certain level of sanctions, few countries participate [20 of 192] **decreases** the **log odds** of individual support 0.273 (the constant).
- For a certain level of sanctions, some countries participate [80 of 192] **increases** the **log odds** of individual support 0.336.
- For a certain level of sanctions, most countries participate [160 of 192] **increases** the **log odds** of individual support 0.648.

Below is the description under certain level of participate countries situations

- For a certain level of participate countries, none sanction **decreases** the **log odds** of individual support 0.273 (the constant).
- For a certain level of participate countries, 5% sanction **increases** the **log odds** of individual support 0.192.
- For a certain level of participate countries, 15% sanction **decreases** the **log odds** of individual support 0.133.
- For a certain level of participate countries, 20% sanction **decreases** the **log odds** of individual support 0.304.

The global null hypothesis is:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_p = 0$$
$$H_1\text{: at least one slope is not equal to } 0$$

```
# the global hypothesis test
# create the null mod for hypothesis test
nullMod <- glm(choice ~ 1, # 1 = fit an intercept only
               data = climateSupport, # select dataset
               family = "binomial") # select method as binomial

# use chi square method to make global test
anova(nullMod, q1mod, test = "Chisq")
# and we can also try this way, they are equal!
anova(nullMod, q1mod, test = "LRT")
```

And we get the outputs:

```
Analysis of Deviance Table
Model 1: choice ~ 1
Model 2: choice ~ countries + sanctions
Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      8499      11783
2      8494      11568  5   215.15 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the p-value (< 2.2e - 16) is below the $\alpha = 0.05$ threshold, so we would say that we find sufficient evidence to reject the null hypothesis. So, we know at least one slope is not equal to 0.

(b). How many iterations did it take to find the maximum likelihood estimates?

In our logistic regression model summary, we will find an outputs like:

```
# summary the model
summary(q1mod)
```

```
Number of Fisher Scoring iterations: 4
```

So, it takes 4 iterations to find the maximum likelihood estimate.

2.If any of the explanatory variables are significant in this model, then:

In Question 1, we know the formula of this module is:

$$logit(p_{choice}) = -0.273 + 0.336contries[80/192] + 0.648countries[160/192] + 0.192sanctions[5\%] - 0.133sanctions[15\%] - 0.304sanctions[20\%]$$

(a).For the policy in which nearly all countries participate [160 of 192], how does increasing sanctions from 5% to 15% change the odds that an individual will support the policy? (Interpretation of a coefficient)

We know the countries level is [160 of 192], then we can write two formulas to present the 5% sanctions situation and 15% sanctions situation:

$$logit(p_{choice}) = -0.273 + 0.648 * 1 + 0.192 * 1 ....(1)$$
$$logit(p_{choice}) = -0.273 + 0.648 * 1 - 0.133 * 1 ....(2)$$
$$(2) - (1) = \Delta logit(p_{choice}) = 0.242 - 0.567 = -0.325$$
$$odd \quad ratios = e^{(-0.325)} \approx 0.723$$

So, we can know: If a policy is nearly all countries participate [160 of 192], increased sanctions from 5% to 15% will increase odds of individual agree with the policy by a multiplicative factor of 0.723.

(b).For the policy in which very few countries participate [20 of 192], how does increasing sanctions from 5% to 15% change the odds that an individual will support the policy? (Interpretation of a coefficient)

We know the countries level is [20 of 192], then we can write two formulas to present the 5% sanctions situation and 15% sanctions situation:

$$logit(p_{choice}) = -0.273 + 0.192 * 1 ....(1)$$
$$logit(p_{choice}) = -0.273 - 0.133 * 1 ....(2)$$
$$(2) - (1) = \Delta logit(p_{choice}) = -0.406 - (-0.081) = -0.325$$
$$odd \quad ratios = e^{(-0.325)} \approx 0.723$$

So, we can know: If a policy is very few countries participate [20 of 192], increased sanctions from 5% to 15% will increase odds of individual agree with the policy by a multiplicative factor of 0.723.

(c).What is the estimated probability that an individual will support a policy if there are 80 of 192 countries participating with no sanctions?

We know the countries level is [80 of 192], and the sanctions level is none, then we can write the formula:

$$logit(p_{choice}) = -0.273 + 0.336 = 0.063$$

And we know the formula is:

$$P = \frac{1}{1+e^{-logit(P)}} \text{ -> } P = \frac{1}{1+e^{-0.063}} \text{ -> } P \approx 0.516$$

And we can use predict code in `R` to check again:

```
# check the predict value
q2c_predict <- predict(q1mod, # set model
                       newdata = data.frame(countries = "80 of 192",
                                            sanctions = "None"), # set dataframe
                       type = "response") # get dependent variable value
print(q2c_predict) # print the results
```

```
1
0.5159191
```

So, we can conclude that if there are 80 out of 192 countries participating with no sanctions, the estimated probability will be approximately 0.516 on average.

(d).Would the answers to 2a and 2b potentially change if we included the interaction term in this model? Why?

The answers will not change. Because we didn't find enough evidence that including an interactive effect of `countries` and `sanctions` is a significant predictor for odds of deciding in individual policy support choice. Below is the process:

```
# add interaction into the regression model
q2mod <- glm(choice ~ countries * sanctions, # Y and Xs
            data = climateSupport, # select dataset
            family = "binomial") # select method as binomial
summary(q2mod)
```

Table 2: Outcome variable is `choice` and explanatory variables are `countries`, `sanctions` and interaction

	Dependent variable:
	choice
countries80 of 192	0.376***
	(0.106)
countries160 of 192	0.613***
	(0.108)
sanctions5%	0.122
	(0.105)
sanctions15%	−0.097
	(0.108)
sanctions20%	−0.253**
	(0.108)
countries80 of 192:sanctions5%	0.095
	(0.152)
countries160 of 192:sanctions5%	0.130
	(0.151)
countries80 of 192:sanctions15%	−0.052
	(0.152)
countries160 of 192:sanctions15%	−0.052
	(0.153)
countries80 of 192:sanctions20%	−0.197
	(0.151)
countries160 of 192:sanctions20%	0.057
	(0.154)
Constant	−0.275***
	(0.075)
Observations	8,500
Log Likelihood	−5,780.983
Akaike Inf. Crit.	11,585.970
Note:	*p<0.1; **p<0.05; ***p<0.01

The global null hypothesis is:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_p = 0$$
$$H_1\text{: at least one slope is not equal to 0}$$

```
# use chi square method to make global test
anova(nullMod, q2mod, test = "Chisq")
# and we can also try this way, they are equal!
anova(nullMod, q2mod, test = "LRT")
```

And we get the outputs:

```
Analysis of Deviance Table

Model 1: choice ~ 1Model
2: choice ~ countries * sanctions
Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      8499      11783
2      8488      11562 11   221.44 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the p-value ($<$ 2.2e - 16) is below the $\alpha = 0.05$ threshold, so we would say that we find sufficient evidence to reject the null hypothesis. So, we know at least one slope is not equal to 0.

And we can make a significant test for different slopes.

$$H_0 : \beta_{\#of countries|sanctions} = \beta_{\#of countries|sanctions}$$
$$H_1 : \text{Effect of countries participate is different by sanctions levels.}$$

```r
# make significant test for different slopes
anova(q1mod, q2mod, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: choice ~ countries + sanctions
Model 2: choice ~ countries * sanctions
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      8494      11568
2      8488      11562  6   6.2928   0.3912
```

We can see that the p-value (0.3912) is upper the $\alpha = 0.05$ threshold, so, there is not evidence that including an interactive effect of `countries` and `sanctions` is a significant predictor for odds of deciding in individual policy support choice.

And we can also visualize these two models scatter to see details:

```r
# Make a data frame
predicted_data <- data.frame(
  choice = climateSupport$choice,
  q1mod_hat = q1mod$fitted.values,
  q2mod_hat = q2mod$fitted.values
)

# Reorder and plot for q1mod_hat
ordered_data <- arrange(predicted_data, q1mod_hat)
ordered_data <- mutate(ordered_data, rank = row_number())

q1_plot <- ggplot(ordered_data, aes(rank, q1mod_hat)) +
  geom_point(aes(colour = choice), alpha = 0.5) +
  theme_bw() +
  scale_y_continuous(limits = c(0, 1)) +
  labs(title = "Q1 Model - without interaction")

# Reorder and plot for q2mod_hat
ordered_data <- arrange(predicted_data, q2mod_hat)
ordered_data <- mutate(ordered_data, rank = row_number())
```

```r

q2_plot <- ggplot(ordered_data, aes(rank, q2mod_hat)) +
  geom_point(aes(colour = choice), alpha = 0.5) +
  theme_bw() +
  scale_y_continuous(limits = c(0, 1)) +
  labs(title = "Q2 Model - with interaction")

# Save plots to a PDF
pdf("q2_plot1.pdf")
# Plot in two columns
gridExtra::grid.arrange(q1_plot, q2_plot, nrow = 2)
dev.off()
```

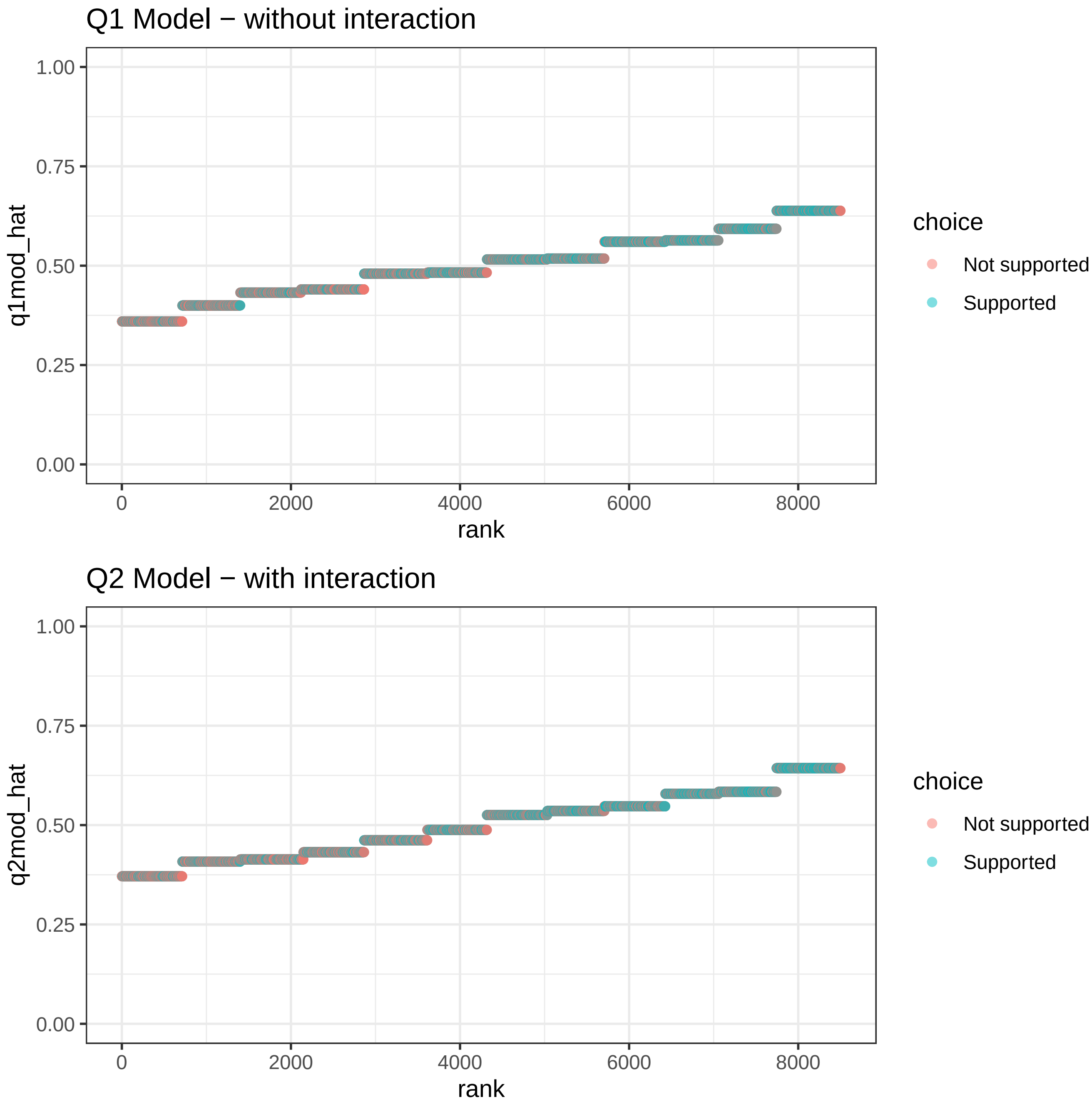

Figure 1: Scatter - Model 1 and Model 2