# Module 6 Assessment

## Instructions

Please submit the markdown file **and compiled pdf** to canvas before April 30 at 11:59pm. For this assignment, you can discuss with classmates, but please at least attempt to go through it individually first so that you can see what you understand or don't understand. Ultimately, the final product you turn in should be your own work. So you can discuss questions with classmates, but your answers should be written in your own words.

## Groundhog Day

Legend has it that Punxsutawney Phil, a groundhog from Punxsutawney, Pennsylvania is capable of predicting the severity of the weather. On Groundhog day each year (Feb 2), Phil rises from his burrow and if he sees his shadow, it means that it will be a long winter. If Phil doesn't see his own shadow, it means that there will be a early spring. Phil has appeared on The Oprah Winfrey Show and was immortalized by the 1993 movie "Groundhog Day" starring Bill Murray. But has Phil been fooling us this whole time, or is he the real deal? Let's take a look. We will consider Phil's predictions from 1900-2022 and for the purposes of this assignment, if the temperature in March is higher than average, we will consider that as an early spring, and if the temperature in March is lower than average we will consider that as a long winter.

```
# load data
phil_data <- read.csv("https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lectureData/groundh
names(phil_data)
```

```
## [1] "Year"        "mar_avg"     "jan_avg"     "feb_avg"     "Prediction"
## [6] "Actual_warm"
```

In the data set, we have the following variables * jan_avg : The average temperature in Pennsylvania in January * feb_avg : The average temperature in Pennsylvania in February * mar_avg : The average temperature in Pennsylvania in January * Prediction : The outcome predicted by Punxsutawney Phil. "Winter" for long winter, "Spring" for early spring * Actual_warm : The dependent variable of interest. It is a 1 if the temperature in March is higher than average, 0 if the temperature in March is lower than average

### Question 1 (1 pt)

Fit a logistic regression model where the outcome is whether or not the temperature in March was higher than average (Actual_warm), and the only covariate we consider is Phil's prediction (Prediction).

#### Answer to Question 1

### Question 2 (1 pt)

Give an interpretation for the estimated coefficient for Phil's prediction in the model above.

#### Answer to Question 2

## Question 3 (1 pt)

Given that Phil predicts an early spring, what are the **odds** that the temperature in March will be above average?

**Answer to Question 3**

## Question 4 (1 pt)

Given that Phil predicted an early spring, what is the **probability** that the temperature in March will be above average?

**Answer to Question 4**

## Question 5 (2 pts)

Does Phil seem to be helpful in predicting the weather in March? Why or why not?

**Answer to Question 5**

# NYC Bike Data

In the following data we have recorded the total number of bicycles which crossed the Manhattan Bridge in New York City each day during 2018[1]. We also have included information about the day of the week (week day or weekend), and information about the weather that day[2].

```
bike_data_train <- read.csv("https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lectureData/

dim(bike_data_train)
names(bike_data_train)
```

The variables in the data set include:

- date: the date of the observation
- bike_counts : the dependant variable of interest which is the total number of bikes which crossed the bridge that day
- weekEnd: Weekend or Weekday
- Wind_avg : average wind on that day
- Precip : precipitation in inches
- Snowfall : snowfall in inches
- Snodepth : amount of snow on the ground
- Temp_avg : the average temperature throughout the day
- Temp_max : the maximum temperature for the day
- Temp_min : the minimum temperature for the day
- Wind_fast2m : the fastest wind speed which was sustained for at least 2 minutes
- Wind_fast5s : the fastest wind speed which was sustained for at least 5 seconds
- FOG : whether there was fog
- Thunder : whether there was Thunder
- IcePellets : whether there was Ice Pellets
- Smoke : whether there was smoke

## Question 6 (1 pt)

If we are interested in seeing how the weather affects the number of bicycles which cross the Manhattan bridge. It seems link a Poisson regression might be appropriate for this data since it is count data. What is the link function used in Poisson regression? What is the relationship between the mean and variance in a Poisson distribution?

### Answer to Question 6

## Question 7 (1 pt)

Fit a Poisson regression model where the outcome of interest is the number of bikes crossing the Manhattan bridge and the covariates of interest are weekend, precipitation, the average temperature, the minimum temperature, snowfall, and fog.

### Answer to Question 7

## Question 8 (1 pt)

How should you interpret the estimated coefficient for **minimum** temperature in the model above?

[1]The data below is a cleaned version of data from the NYC open data https://data.cityofnewyork.us/Transportation/Bicycle-Counts/uczf-rk3c

[2]The weather data was recorded at the JFK airport weather station and is available from NOAA at https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND

**Answer to Question 8**

### Question 9 (1 pt)

We can test an individual coefficient using the output of `summary`. But as we discussed in class, we can also create confidence intervals and tests using the $\chi^2$ test which comes from using the likelihood. Test whether the average temperature is statistically significant using the $\chi^2$ test.

**Answer to Question 9**

### Question 10 (5 pts)

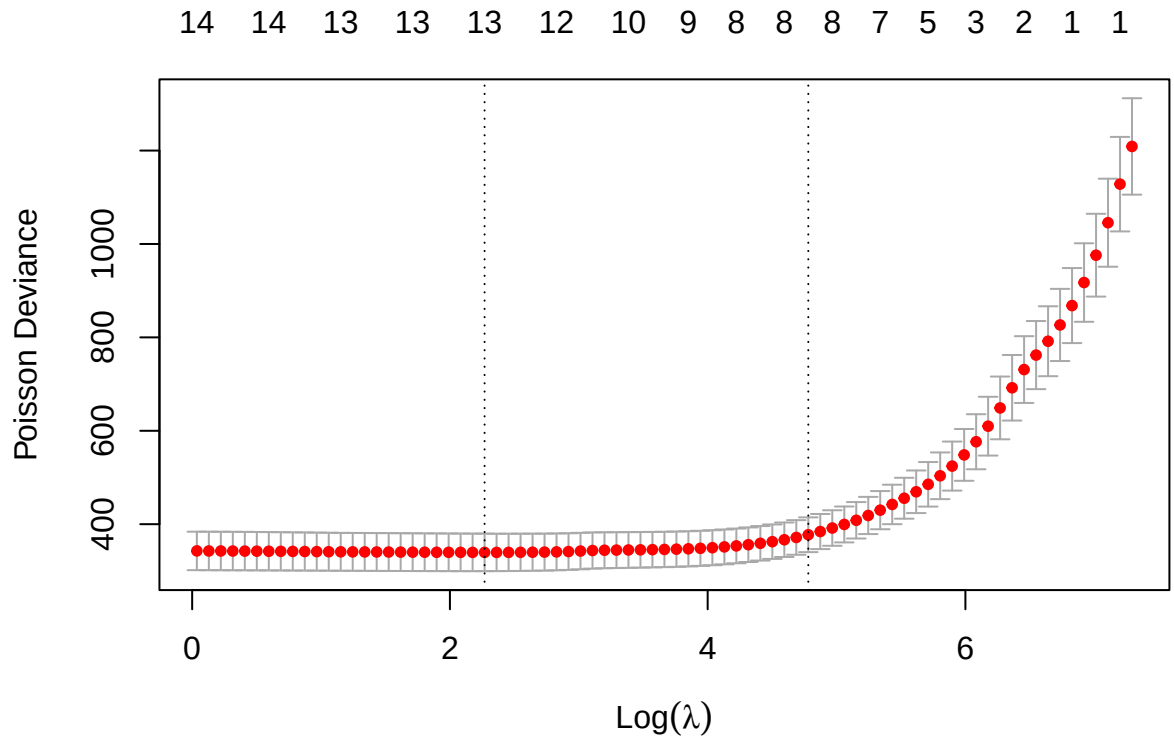
Do the assumptions for Poisson regression seem satisfied for this data? Why or why not? You can use plots or other code to justify your answers if needed.

**Answer to Question 10**

## High-dimensional Regression

Even though we have more observations than covariates here, we can still use the **Lasso** procedure to select a model. In particular, the following code selects the penalty parameter $\lambda$ value through a cross validation procedure. The plot shows the Deviance (a measure of fit calculated using the likelihood) and the horizontal axis shows various values of the penalty parameter. When the deviance is larger, this indicates that the coefficients do not fit the data as well. The estimated coefficients for the selected model are printed below.

```
# fit the lasso (alph = 1 indicates lasso) with a poisson family
lasso_mod <- glmnet::cv.glmnet(y = bike_data_train$bike_counts,
                               x = as.matrix(bike_data_train[, -c(1,2)]),
                               alpha = 1, family = "poisson")
plot(lasso_mod)
```

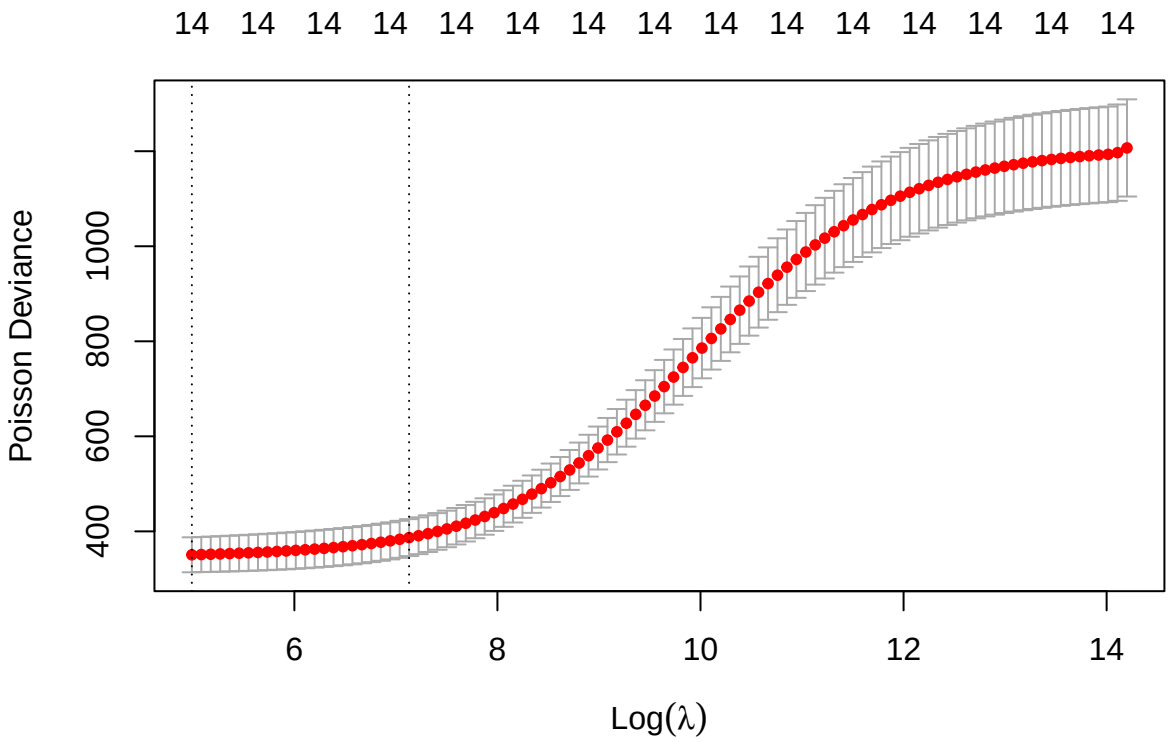


```r
# We can see the coefficients selected by the largest lambda value with a
# CV error within 1 standard error of the lowest CV error
coef(lasso_mod, s = lasso_mod$lambda.1se)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept)  7.2670015741
## weekEnd     -0.3322201103
## Wind_avg    -0.0085806791
## Precip      -0.3821864754
## Snowfall    -0.0233748083
## Snowdepth   -0.0487821396
## Temp_avg     .
## Temp_max     0.0188952637
## Temp_min     .
## Wind_fast2m  .
## Wind_fast5s -0.0003883544
## FOG         -0.0927611094
## Thunder      .
## IcePellets   .
## Smoke        .
```

The following code uses **ridge regression** to estimate coefficients for the Poisson regression and uses a $\lambda$ value which is selected by cross validation. The plot shows the Deviance (a measure of fit calculated using the likelihood) and the horizontal axis shows various values of the penalty parameter. When the deviance is larger, this indicates that the coefficients do not fit the data as well. The estimated coefficients for the selected model are printed below.

```r
# fit the ridge regression (alpha = 0 indicates ridge) with a poisson family
ridge_mod <- glmnet::cv.glmnet(y = bike_data_train$bike_counts,
                               x = as.matrix(bike_data_train[, -c(1,2)]),
                               alpha = 0, family = "poisson")
plot(ridge_mod)
```




```
# We can see the coefficients selected by the largest lambda value with a
# CV error within 1 standard error of the lowest CV error
coef(ridge_mod, s = ridge_mod$lambda.1se)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept)  7.5972355275
## weekEnd     -0.3072004266
## Wind_avg    -0.0101221904
## Precip      -0.3114437911
## Snowfall    -0.0752549513
## Snowdepth   -0.0605829220
## Temp_avg     0.0056313115
## Temp_max     0.0076688222
## Temp_min     0.0037712861
## Wind_fast2m -0.0007320694
## Wind_fast5s -0.0030456146
## FOG         -0.1407990105
## Thunder      0.0350538815
## IcePellets  -0.2346937856
## Smoke        0.1506612684
```

## Question 11 (1 pts)

There is not one right answer, but which model would you prefer to use? Explain why.

**Answer to Question 11**

## Question 12 (2 pts)

Using the data from 2019, we compare the predictive accuracy of the coefficients estimated by the lasso and the ridge to the predictive accuracy of the coefficients estimated without any penalty (which we calculate below for you). We see that the lasso and ridge both have better predictions for new data. Explain why this is might happen using one of the fundamental statistical trade-offs we have discussed in class.

```
# Test data from 2019
bike_data_test <- read.csv("https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lectureData/bi

# regular glm without any penalization
unpenalized_model <- glm(bike_counts ~ ., data = bike_data_train[, -1], family = "poisson")

# predictive accuracy for 2019 when using all covariates but no model seletion or penalization
## use type = "response" to get predictions in bikes, instead of log(bikes)
mean((bike_data_test$bike_counts - predict(unpenalized_model,
                                           newx = as.matrix(bike_data_test[,-c(1,2)]),
                                           type = "response"))^2)
```

```
## [1] 3251833
```

```
## Mean squared error for predictions using lasso
mean((bike_data_test$bike_counts - predict(lasso_mod, s=lasso_mod$lambda.1se,
                                           newx = as.matrix(bike_data_test[,-c(1,2)]),
                                           type = "response"))^2)
```

```
## [1] 1032445
```

```
## Mean squared error for predictions using ridge regression
mean((bike_data_test$bike_counts - predict(ridge_mod, s=ridge_mod$lambda.1se,
                                           newx = as.matrix(bike_data_test[,-c(1,2)]),
                                           type = "response"))^2)
```

```
## [1] 1058384
```

**Answer to Question 12**