Logistic Regression

In logistic regression, outcome is dichotomous:

- Yes or No
- Win vs. Lose
- Pass vs. Fail

The outcome is specified as a 1 or a 0:
The outcome of interest (Yes, Win, Pass) is specified as the 1
The predictors must still be ratio/interval/ or dichotomous

Logistic Regression

Coefficients are interpreted as the log-odds increase in observing the outcome, if the predictor variable increases by 1 unit, holding all other variables constant

- Example, if B_1 = 0.7
- Then if the predictor variable X_1 went from 1 to 2, the odds of the outcome will go up by $e^{0.7}$ = 2.00.
- Therefore, if the predictor variable for this person is 1 unit higher than another person, and all is the same, then, that person should have double the odds of being in the "yes" outcome
- $Y = -2.3 + .7X_1$
    - If X_1 is 4, then odds of "yes" outcome = $e^{-2.3 + .7*4}$ = 1.64
    - If X_1 is 5, then odds of "yes" outcome = $e^{-2.3 + .75}$ = 3.32

Penalized Regression

Imagine collecting 100 predictor variables to try to predict a person's stress

- Social Support
- Time of day
- Color shirt etc.

The regression equation will look like the following

- y = .20 + .75*Social Support + .23*Time of Day - .02*Color shirt

Notice how even bad predictors still have a coefficient value

Probably those values are not the real value of 0

For the really good predictors, they probably don't a coefficient that high either.

Penalized Regression

- In ordinary multiple regression, we never penalized the coefficients
- Whatever the line of best fit was, we accepted
- Regression tends to overfit - the line is a little “too” close to all the points
- We might want to make them coefficients a little smaller to compensate for overfitting
- Penalized regression = making the regression coefficients smaller in a systematic way

Ridge Regression

- Ridge regression shrinks the regression coefficients, so that variables, with minor contribution to the outcome, have their coefficients close to zero.
- Shrinking the coefficients is achieved by penalizing the regression model by focusing on finding the line of best fit especially on predictors with big coefficients (good predictors)
- Ridge Regression changes the formula for calculating the line of best fit

Ridge Regression

In Ridge regression, we add a "cost" to the calculation of what the line of best fit should be

- In ordinary least squares, the line of best fit, was one that minimized the squared differences between the outcome and predicted values

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

- In Ridge regression, we add an additional variable to the cost function, which includes
  - The sum of the squared coefficients (if the coefficient is larger, it needs to justify its contribution)
  - Lambda: how much we want to emphasize the squared coefficient penalty

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Ridge Regression

The amount of the penalty can be fine-tuned using a constant called **lambda (λ)**

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

λ ranges from 0 to Infinity

When λ=0, the penalty term has no effect, and ridge regression becomes the same as ordinary multiple regression

As λ increases to infinity, the impact of the penalty grows, and the regression coefficients will get closer zero.

Benefits of Ridge Regression

- In ridge regression, **the data need to be standardized** - need coefficients to be comparable to each other
  - Make sure to standardize each column before perform Ridge Regression
  - Otherwise the coefficient of a variable measured on a small scale (feet) will be bigger than the coefficient of a variable measured on a larger scale (inches), even if they are equally important at predicting the outcome
- **Works better than ordinary multiple regression when you have many predictors**
  - Ordinary Multiple Regression needed to have fewer predictor to prevent overfitting (Through Dimension Reduction or Manually removing them)
  - Ridge allows you to throw all the predictors at the model
  - Can have more predictors that you have people measured
- **Works well with many highly correlated predictors**

Limitations of Ridge Regression

- **Need to remember to standardize the predictors**

- **Need to choose the right value for lambda**

- **It will include all the predictors in the final model**
  - Even useless ones will be in the final model
  - Ideally want to thrown out predictors that don’t help, so
    - you don’t have to measure them in the future
    - you have a simpler model to describe

- **Ridge regression shrinks the coefficients towards zero, but it will not set any of them exactly to zero**
  - Lasso regression overcomes this drawback.

Lasso Regression

➢ **Lasso regression (Least Absolute Shrinkage and Selection Operator)**

- Lasso regression also shrinks the regression coefficients, so that variables, with minor contribution to the outcome, have their coefficients close to zero.

- Lasso regression and Ridge regression are similar, but it changes the formula for the penalty

➢ Lasso regression uses the **L1-norm**, which means that the sum of the absolute value of the coefficients are used, instead of the sum of the squared coefficients

Lasso Regression

In Lasso regression, we add a "cost" to the calculation of what the line of best fit should be-

- In ordinary least squares, the line of best fit, was one that minimized the squared differences between the outcome and predicted values

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

In Lasso regression, we add an additional variable to the cost, which includes

- The sum of the absolute value of the coefficients (if the coefficient is larger, it needs to justify its contribution)
- Lambda: how much we want to emphasize the absolute value coefficient penalty

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Benefits of Lasso Regression

- **Works better than ordinary multiple regression when you have many predictors**
    - Lasso allows you to throw all the predictors at the model
- **Not as good as Ridge regression, when more predictors than people**
- **Allows Predictor Coefficients to be 0**
    - Allows you to perform "Feature selection" (remove non-essential predictors)
    - Produces simpler models (fewer predictor variables)
    - Produces more interpretable models (fewer variables to describe)

Limitations of Lasso Regression

- **Need to choose the right value for lambda**

- **Need to remember to standardize the predictors**

- **Not ideal for many highly correlated variables**
    - Ridge is preferred in that case
    - When there are highly correlated predictors lasso tends to just pick one predictor out of the group.

Elasticnet Regression

- **ElasticNet Regression**: A combination of Ridge and Lasso regression
  - Uses both L1 and L2 norms to penalize the coefficients
  - Can set the amount/ratio of L1 vs L2 normalization using **α**
- Alpha (**α**) represents the ratio of Ridge to Lasso normalization that we want
  - **α** = 1: Pure Lasso
  - **α** .9: Mostly Lasso
  - **α** = .5: Equally Lasso and Ridge
  - **α** = .1 Mostly Ridge
  - **α** = 0: Pure Ridge

$$P = \lambda \left\{ (1-\alpha) \|\beta\|_2^2 / \; 2 + \alpha \|\beta\|_1 \right\}$$

Benefits of Elasticnet Regression

- **Potentially the same benefits as Lasso and Ridge**
  - **Balances generating a model with two predictors and too many predictors**
  - L1-penalty helps generating a sparse model
  - L2-part overcomes a strict selection
- **Does not require you to know whether Lasso or Ridge is the best model**
  - By changing alpha, and cross-validating, you can see which form works better

Limitations of Elasticnet Regression

- **Need to remember to standardize the predictors**
- **Need to find the best values for lambda and alpha**