

Lecture 19: Multiple Logistic Regression

STAT 432, Spring 2021

Multiple Logistic Regression

- ▶ Multiple logistic regression is a method to model a binary response variable, $Y \in \{0, 1\}$, using predictor variables x_1, x_2, \dots, x_p .
- ▶ Specifically, the method models $p(\mathbf{x}) = Pr(Y = 1|\mathbf{x})$, the probability $Y = 1$ given predictors $\mathbf{x} = (x_1 \ x_2 \ \dots \ x_p)'$.

Multiple Logistic Regression

Two ways to express multiple logistic regression model:

Probability form:

$$p(\mathbf{x}) = \Pr(Y = 1|\mathbf{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}} = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \dots - \beta_p x_p}}$$

Logit form:

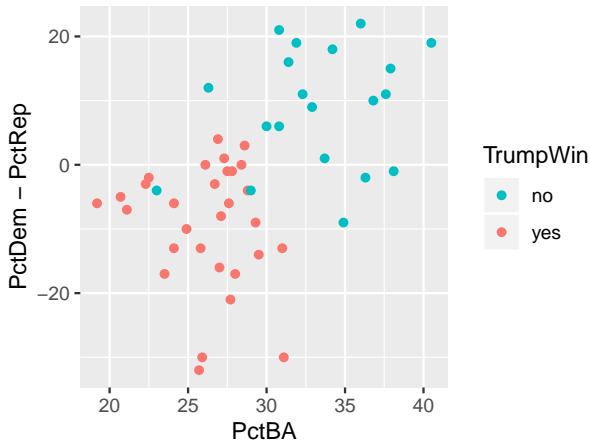
$$\log \left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})} \right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

Example: 2016 US Presidential Election

- ▶ Again, we use the data set called `Election16` from the `Stat2Data` library. The data contain results from the 2016 presidential election and demographic information from all 50 states.
- ▶ The binary response variable is `TrumpWin`, whether Trump won the state (1=yes, 0=no).
- ▶ The predictors are
 - ▶ `HS`: Percent of high school graduates in the state
 - ▶ `BA`: Percent of college graduates in the state
 - ▶ `Adv`: Percent with advanced degrees in the state
 - ▶ `Dem.Rep`: Percent Democratic - Percent Republican
 - ▶ `Income`: Per capita income in the state

Example

We'll start by considering a multiple logistic regression model for `TrumpWin`, using two predictors from the data set: `BA` and `Dem.Rep`.



Example

We'll start by considering a multiple logistic regression model for TrumpWin, using two predictors from the data set: BA and Dem.Rep.

```
> glm2 <- glm(TrumpWin ~ BA + Dem.Rep, data=Election16,  
              family=binomial)
```

```
> summary(glm2)
```

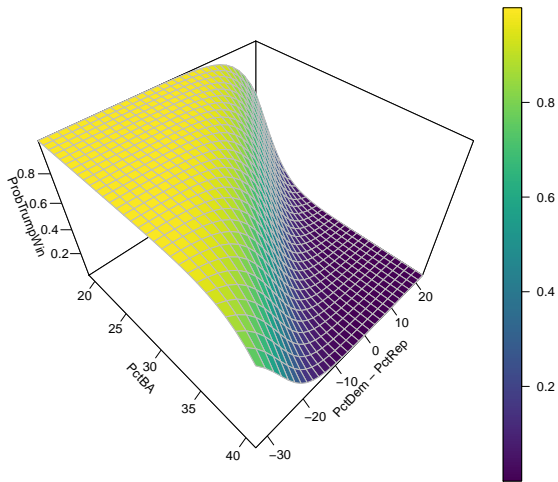
Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	15.34796	6.13130	2.503	0.0123 *
BA	-0.51792	0.21181	-2.445	0.0145 *
Dem.Rep	-0.21406	0.08778	-2.439	0.0147 *

```
> confint(glm2)
```

	2.5 %	97.5 %
(Intercept)	5.7651914	31.36586108
BA	-1.0790415	-0.18972992
Dem.Rep	-0.4357941	-0.07794399

Example



The code used to create the last plot:

```
> library(plot3D)
> library(viridis)
> glm2 <- glm(TrumpWin ~ BA + Dem.Rep, data=Election16, family=binomial)
> x1vals <- seq(19, 41, len=30)
> x2vals <- seq(-33, 23, len=30)
> grd <- expand.grid(BA = x1vals, Dem.Rep = x2vals)
> preds <- predict(glm2, grd, type="response")
> persp3D(x1vals, x2vals, matrix(preds, 30, 30), col = viridis(200),
  theta=45, phi= 45, ticktype="detailed", expand=0.7, border="grey",
  xlab = "PctBA", ylab = "PctDem - PctRep", zlab = "ProbTrumpWin")
```


Example

The equation for the fitted logistic regression model in probability form is given by:

$$\hat{p}(x_1, x_2) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2}} = \frac{e^{15.348 - 0.518x_1 - 0.214x_2}}{1 + e^{15.348 - 0.518x_1 - 0.214x_2}}$$

In California, the % with a BA is 31.4, and Dem.Rep, the % Democrat minus the % Republican, is 16. So the estimate for the probability that Trump won is

$$\hat{p}(31.4, 16) = \frac{e^{15.348 - 0.518(31.4) - 0.214(16)}}{1 + e^{15.348 - 0.518(31.4) - 0.214(16)}} = \frac{e^{-4.34}}{1 + e^{-4.34}} = 0.0129$$

Example

To make the predictions in R use the `predict()` function:

```
> new_x <- data.frame(BA = 31.4, Dem.Rep = 16)

# prediction for logit
> predict(glm2, newdata=new_x)
      1
-4.339819

# prediction for probability
> predict(glm2, newdata=new_x, type="response")
      1
0.01287107
```

Variable Selection

- ▶ The AIC can also be used to select variables for a logistic regression model.
- ▶ The *smaller* the value for the AIC the better the logistic regression model.
- ▶ If we fit several logistic regression models, using the same data set, we would select the model with the smallest AIC value.

Example

Consider all 8 possible logistic regression models for TrumpWin, using BA, Dem.Rep, and Income as potential predictors.

Predictor Variables	AIC
Null model (intercept only)	69.30
BA	38.43
Dem.Rep	37.60
Income	49.92
BA, Dem.Rep	27.08
BA, Income	40.12
Dem.Rep, Income	26.79
BA, Dem.Rep, Income	27.62

Here is the code used for the previous table:

```
> glm0 <- glm(TrumpWin ~ 1, data=Election16, family=binomial)
> glm1_1 <- glm(TrumpWin ~ BA, data=Election16, family=binomial)
> glm1_2 <- glm(TrumpWin ~ Dem.Rep, data=Election16, family=binomial)
> glm1_3 <- glm(TrumpWin ~ Income, data=Election16, family=binomial)
> glm2_1 <- glm(TrumpWin ~ BA + Dem.Rep, data=Election16, family=binomial)
> glm2_2 <- glm(TrumpWin ~ BA + Income, data=Election16, family=binomial)
> glm2_3 <- glm(TrumpWin ~ Dem.Rep + Income, data=Election16, family=binomial)
> glm3 <- glm(TrumpWin ~ Income + BA + Dem.Rep, data=Election16, family=binomial)

> AIC(glm0, glm1_1, glm1_2, glm1_3, glm2_1, glm2_2, glm2_3, glm3)
      df      AIC
glm0    1 69.30117
glm1_1  2 38.43266
glm1_2  2 37.59612
glm1_3  2 49.92250
glm2_1  3 27.08020
glm2_2  3 40.12465
glm2_3  3 26.79370
glm3    4 27.62363
```

We can also use the `step()` function to implement backwards stepwise selection for logistic regression using the AIC.

```
> glm5 <- glm(TrumpWin ~ HS + BA + Adv + Dem.Rep + Income,  
              data=Election16, family = binomial)
```

```
> glm_sel <- step(glm5)  
> summary(glm_sel)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	1.458e+01	5.308e+00	2.747	0.00601	**
Dem.Rep	-3.099e-01	1.135e-01	-2.731	0.00632	**
Income	-2.677e-04	9.866e-05	-2.713	0.00667	**

```
> AIC(glm5, glm_sel)  
      df      AIC  
glm5    6 31.55431  
glm_sel  3 26.79370
```