# Logistic Regression

Jeffrey Woo

MSDS, University of Virginia

# Logistic Regression Notation & Terminology

- $\pi$: **probability** of “success” (of belonging in a certain class).
- $1 - \pi$: probability of “failure” (of belonging in the other class).
- $\frac{\pi}{1-\pi}$: **odds** of “success”. Odds is not the same as probability (if probability is close to 0 then odds is approximately the same as the probability). For rare events, the terms can be interchangeable.
- $\log\left(\frac{\pi}{1-\pi}\right)$: log-odds of “success”.

From my experience, a lot of people tend to use “odds” and “probability” interchangeably. In my view, this should only be done for rare events.

# Logistic Regression Equation

Typically written as

$$\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{1}$$

Alternate way of writing:

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} \tag{2}$$

# Interpreting Coefficient

Assume we only have one predictor which is quantitative. There are a few interpretations of the coefficient of the predictor, $\beta_1$.

- For a one-unit increase in the predictor, the **log odds** changes by $\beta_1$.
- For a one-unit increase in the predictor, the **odds are multiplied** by a factor of $\exp(\beta_1)$.
- For a one-unit increase in the predictor, the **odds ratio** is $\exp(\beta_1)$.

These are all equivalent interpretations. If $\beta_1$ is positive, $\exp(\beta_1)$ is greater than 1, so multiply by a factor greater than 1, which means the odds increase. Increase in odds is also an increase in probability

# Titanic Example

$y$: survived (1) or not (0), predictor: fare paid (adjusted for inflation)

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.92671    0.10866  -8.528  < 2e-16 ***
Fare         0.01613    0.00252   6.399 1.56e-10 ***
```

- Coefficient estimates are $\hat{\beta}_0 = -0.92671$, and $\hat{\beta}_1 = 0.01613$.
- So $\pi =$probability of survival for given $X$ is estimated by

$$\hat{\pi} = \frac{e^{-0.92671+0.0.01613x}}{1+e^{-0.92671+0.0.01613x}}$$

- The predictor is statistically significant.

# Titanic Example

In this example,

- The predicted log odds of a passenger surviving increases by 0.01613 for a one-unit increase in paid fare.
- The predicted odds of a passenger surviving is multiplied by $\mathbf{exp\,(0.01613) = 1.016261}$ for a one-unit increase in paid fare.
- The predicted odds ratio of a passenger surviving for a one-unit increase in paid fare 1.016261.

# Interpreting Coefficient

If the coefficient $\beta_1$ is for an indicator variable, the interpretation is easier. Assume we have a binary variable, coded 1 for males and 0 for females, and we are predicting if the person survives the Titanic.

- The difference in log odds of survival between males and females is $\beta_1$.
- The odds of survival for males is $\exp(\beta_1)$ times the odds of survival for females.
- The odds ratio of survival for males and females is $\exp(\beta_1)$.

These are all equivalent interpretations.

# Hypothesis Tests in Logistic Regression

- Remove single predictor: Wald Test based on the $Z$ distribution (analogous to the $t$ test in MLR)
- Is model useful: $\Delta G^2 =$ null deviance - residual deviance based on the $\chi^2_{p-1}$ distribution (analogous to the ANOVA $F$ test in MLR)
- Remove a subset of terms: $\Delta G^2 =$ residual deviance of reduced model - residual deviance of full model based on the $\chi^2_{\# \text{ of terms to drop}}$ distribution (analogous to partial $F$ test)

# Goodness of Fit Tests

Three Goodness of Fit (GOF) tests in logisitic regression. In other words is the log odds a linear combination of coefficients and predictors:

1. Deviance GOF test
2. Pearson GOF test
3. Hosmer-Lemeshow test

Deviance and Pearson require **grouped** data (usually found in designed experiments); cannot have (a number of) observations that have unique combination of predictors. Difficult to implement in observational studies.

# Residuals in Logistic Regression

Due to the binary nature of the response variable in logistic regression, examining residuals is not very helpful (due to discrete nature of response, the values of the residuals are typically discrete). Again, require data to be grouped to reliably use the plots.