STAT 302: Logistic Regression

Sheridan Grant

University of Washington Statistics Department

slgstats@uw.edu

May 11, 2020

Review: Prediction vs. Model Fit

Model Fit

- Compare predictions to truth
- on data used to train model

BAD measures of fit (no *df* adjustment):

- $SSE \equiv \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- Root-Mean Squared Error
  $RMSE \equiv \sqrt{SSE/n}$
- Bad because nonsense variables can artificially decrease

GOOD measure of fit (*df* adjustment):

- $df = n - \#$Parameters
- Residual Standard Error
  $RSE \equiv \sqrt{SSE/df}$
- Good because nonsense can't improve fit

Review: Prediction vs. Model Fit

Model Fit

- Compare predictions to truth
- on data used to train model

BAD measures of fit (no *df* adjustment):

- $SSE \equiv \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- Root-Mean Squared Error $RMSE \equiv \sqrt{SSE/n}$
- Bad because nonsense variables can artificially decrease

GOOD measure of fit (*df* adjustment):

- $df = n - \#\text{Parameters}$
- Residual Standard Error $RSE \equiv \sqrt{SSE/df}$
- Good because nonsense can't improve fit

Predictive Power

- Compare predictions to truth
- **NOT** on data used to train model

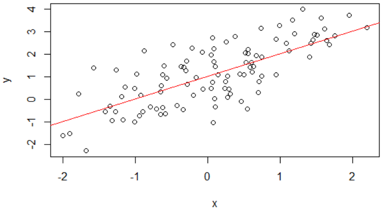
Measures of predictive power

- Root-Mean Squared Error $RMSE \equiv \sqrt{SSE/n}$
- Mean Absolute Error $MAE \equiv \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$
- Many others

Linear vs. Logistic Regression

Linear Regression

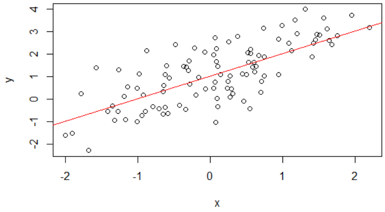
- Continuous response
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta} X_i$



Linear vs. Logistic Regression

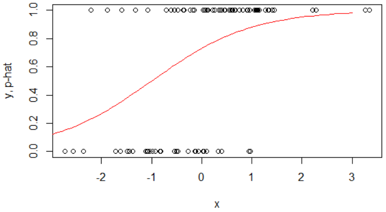
Linear Regression

- Continuous response
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta} X_i$



Logistic Regression

- Binary response
- $\hat{\mathrm{P}}(Y_i = 1) = f(\hat{\beta}_0 + \hat{\beta} X_i)$



More on Logistic Regression

- *Generalized* linear model
- Must transform $p_i = \mathrm{P}(Y_i = 1)$ so that a linear model makes sense
- Trick is to find 1-1 transformation $f : p \in [0, 1] \to f(p) \in \mathbb{R}$

More on Logistic Regression

- *Generalized* linear model
- Must transform $p_i = \mathrm{P}(Y_i = 1)$ so that a linear model makes sense
- Trick is to find 1-1 transformation $f : p \in [0, 1] \to f(p) \in \mathbb{R}$

1. $p \in [0, 1]$

More on Logistic Regression

- *Generalized* linear model
- Must transform $p_i = \mathrm{P}(Y_i = 1)$ so that a linear model makes sense
- Trick is to find 1-1 transformation $f : p \in [0, 1] \to f(p) \in \mathbb{R}$

1. $p \in [0, 1]$
2. $\frac{p}{1-p} \in [0, \infty)$

More on Logistic Regression

- *Generalized* linear model
- Must transform $p_i = \mathrm{P}(Y_i = 1)$ so that a linear model makes sense
- Trick is to find 1-1 transformation $f : p \in [0, 1] \to f(p) \in \mathbb{R}$

1. $p \in [0, 1]$
2. $\frac{p}{1-p} \in [0, \infty)$
3. $\mathrm{logit}(p) \equiv \log\left(\frac{p}{1-p}\right) \in \mathbb{R}$

More on Logistic Regression

- *Generalized* linear model
- Must transform $p_i = \mathrm{P}(Y_i = 1)$ so that a linear model makes sense
- Trick is to find 1-1 transformation $f : p \in [0, 1] \rightarrow f(p) \in \mathbb{R}$

1. $p \in [0, 1]$
2. $\frac{p}{1-p} \in [0, \infty)$
3. $\text{logit}(p) \equiv \log\left(\frac{p}{1-p}\right) \in \mathbb{R}$
4. $p_i = \text{expit}(\beta_0 + \beta X_i + \epsilon_i)$
5. $\hat{p}_i = \text{expit}(\hat{\beta}_0 + \hat{\beta} X_i)$ $(\text{expit}(x) = \exp(x)/[\exp(x) + 1])$

Finally, get $\hat{Y}_i$ from $\hat{p}_i$. For example, $\hat{Y}_i = \mathbf{1}[\hat{p}_i > 0.5]$. Other suggestions?