Business Analytics

Session 2b. Classification and Logistic Regression

Renyu (Philip) Zhang

New York University Shanghai

Spring 2019

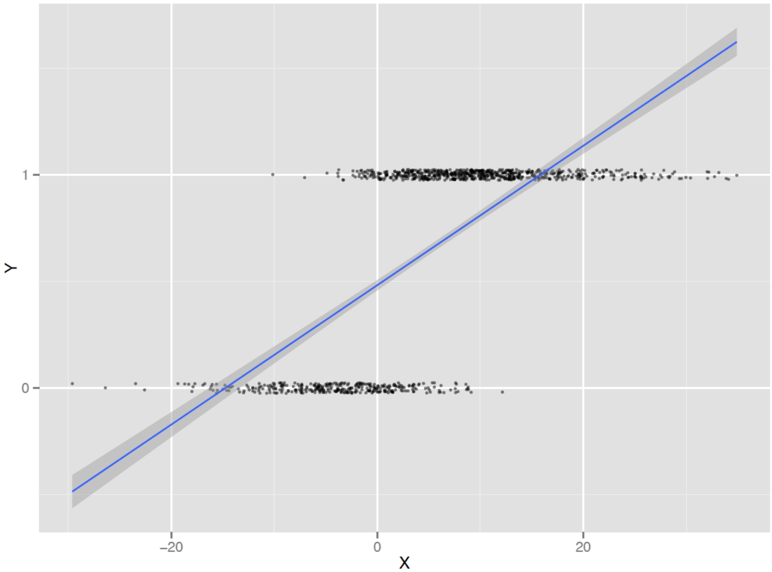
Classification in a Nutshell

Regression Analysis with Binary Outcomes

- **Goal**: Find the relationship between the outcome Y and the covariates $X = (X_1, X_2, \cdots, X_p)$, where $Y \in \{0, 1\}$.
  - Examples: Will the user watch the video? Will the user click the ad? Is this patient with cancer?
  - Y is called a *categorical variable*.

- Why linear regression does not work well in this case? Let's see a picture.

Linear Regression with Binary Outcomes



Goal of (Binary) Classification

We attempt to build a classifier $\hat{f}(\cdot)$ that predect the outcome according to decision rules of the form:

$$\hat{Y}_i = \begin{cases} 1, & \text{if } X_i \in \mathcal{X} \\ 0, & \text{otherwise} \end{cases} \quad \text{for all } i = 1, 2, \cdots, n$$

We want the classifier to make as few mistakes as possible:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\hat{Y}_i = Y_i\} \text{ is minimized, where } \mathbf{1}\{\hat{Y}_i \neq Y_i\} = \begin{cases} 1, & \text{if } \hat{Y}_i \neq Y_i \\ 0, & \text{otherwise.} \end{cases}$$

Logistic Regression

- Predict the *probability* of $Y = 1$ using X.
  - Turns a classification problem into a regression problem.

- Input: Sample data $\mathcal{D} = \{Y_i \in \{0,1\}, X_{ij} : 1 \leq i \leq n, 1 \leq j \leq p\}$.

- Output: A fitted model $\hat{f}(\cdot)$, such that

$$\mathbb{P}(Y_i = 1 | X_i) \approx \hat{f}(X_i), \text{ where } X_i = (X_{i1}, X_{i2}, \cdots, X_{ip}).$$

Healthcare Quality Assessment Using Logistic Regression (Analytics Edge, Chapter 1.2)

Assessing Healthcare Qualities

- Healthcare quality assessment is essential for medical interventions.
  - Good quality care educates patients and controls costs.
- No single set of guidelines for defining the quality of healthcare.
- Rely on healthcare experts to assess the quality.
- Expert physicians evaluate the healthcare quality by examining patients' records.
  - Time-consuming, inefficient, and non-scalable.
- Question: How to identify poor healthcare quality using analytics?

Data

- Claims data for 131 diabetes patients randomly sampled from a large health insurance claims database.
  - Basic claims information like age, cost, etc.
  - Expert review data about diagnoses, treatments, and prescrptions.
  - Expert assessment about healthcare quality (poor or high quality).

Data

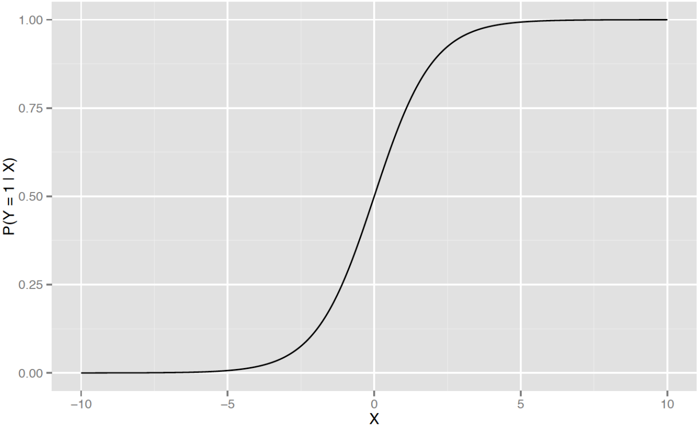
- Claims data for 131 diabetes patients randomly sampled from a large health insurance claims database.
  - Basic claims information like age, cost, etc.
  - Expert review data about diagnoses, treatments, and prescrptions.
  - Expert assessment about healthcare quality (poor or high quality).
- Outcome (Y_i)
  - **Quality of care**: 1 for low-quality care, 0 for high-quality care
- Covariates (X_i)
  - Diabetes treatment
  - Patient demographics
  - Healthcare utilization
  - Providers
  - Claims
  - Prescriptions

Logistic Regression to Predict Quality of Care

- Logistic regression assumes the following model:

$$\mathbb{P}(Y_i = 1|X_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})} = 1 - \mathbb{P}(Y_i = 0|X_i)$$

- A plot of one covariate, with $\beta_0 = 0$ and $\beta_1 = 1$



Understanding the Logistic Curve

- The logistic curve is increasing.
  - Larger values of covariates with positive coefficients will tend to increase the probability that $Y = 1$.

- Positive (resp. negative) values of $\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$ are predictive of $Y_i = 1$ (resp. $Y_i = 0$).

Understanding the Logistic Curve

- The logistic curve is increasing.
  - Larger values of covariates with positive coefficients will tend to increase the probability that $Y = 1$.

- Positive (resp. negative) values of $\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$ are predictive of $Y_i = 1$ (resp. $Y_i = 0$).

- Odds:

$$\text{Odds} = \frac{\mathbb{P}(Y = 1|X)}{\mathbb{P}(Y = 0|X)} \in (0, +\infty)$$

- In logistic regression:

$$\log(\text{Odds}) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \in (-\infty, +\infty)$$

Finding Logistic Regression Coefficients

- Model:

$$\mathbb{P}(Y_i = 1|X_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})} = 1 - \mathbb{P}(Y_i = 0|X_i)$$

- Find the coefficients $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)$ that maximize the chance of seeing the data $\mathcal{D} = \{Y_i \in \{0,1\}, X_{ij} : 1 \le i \le n, 1 \le j \le p\}$
  - Maximum likelihood estimation.

$$\hat{\beta} = \mathrm{argmax}_{\beta} \sum_{i=1}^{n} \left(Y_i \log \left(\frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})} \right) + (1 - Y_i) \log \left(\frac{1}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})} \right) \right)$$

Finding Logistic Regression Coefficients

- Model:

$$\mathbb{P}(Y_i = 1|X_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})} = 1 - \mathbb{P}(Y_i = 0|X_i)$$

- Find the coefficients $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)$ that maximize the chance of seeing the data $\mathcal{D} = \{Y_i \in \{0,1\}, X_{ij} : 1 \leq i \leq n, 1 \leq j \leq p\}$
  - Maximum likelihood estimation.

$$\hat{\beta} = \text{argmax}_{\beta} \sum_{i=1}^{n} \left(Y_i \log \left(\frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})} \right) + (1 - Y_i) \log \left(\frac{1}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})} \right) \right)$$

- Demo: Fitting the model in *R*.
- How do we interpret the fitted coefficients $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$?
  - $\hat{\beta}_j$ is the change in the log odds for Y associated with unit change in X_j.

Classification with Logistic Regression

- We want to make a binary prediction (classification).
  - Did this patient receive poor care or good care?
- Use a threshold value t:
  - If $\mathbb{P}(Y = 1|X) \geq t$, predict poor quality.
  - If $\mathbb{P}(Y = 1|X) < t$, predict good quality.
- Confusion Matrix

	Predicted=0	Predicted=1
Actual=0	True Negative (TN)	False Positive (FP)
Actual=1	False Negative(FN)	True Positive (TP)

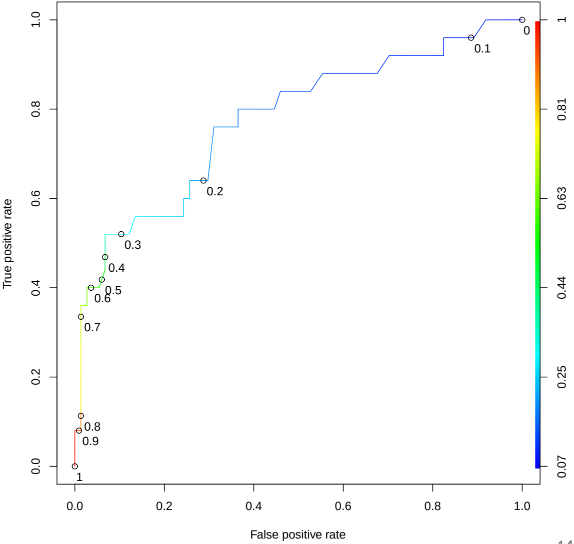
$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN}$$

- Trade-off between false-negatives and false-positives.
  - If t is large, false-positive $\downarrow$ and false-negative $\uparrow$.
  - If t is small, false-positive $\uparrow$ and false-negative $\downarrow$.
  - If $t = 0.5$, the model predicts the more likely outcome.

Receiver Operator Characteristic (ROC) Curve

- True positive rate (Sensitivity) on $y-$axis.
- False positive rate (1-specificity) on $x-$axis.
- Best t value trades off the cost of failing to detect positives and the cost of raising false alarms.



Interpreting Logistic Regression Model

- Multicollinearity
  - Do the coefficients make sense?
  - Check covariate correlations.
- Measures of accuracy:

	Predicted=0	Predicted=1
Actual=0	True Negative (TN)	False Positive (FP)
Actual=1	False Negative(FN)	True Positive (TP)

N=number of observations

Overall Accuracy=(TN+TP)/N

Overall error rate=(FP+FN)/N

Sensitivity=TP/(TP+FN)

False Negative Error Rate=FN/(TP+FN)

Specificity=TN/(TN+FP)

False Positive Error Rate=FP/(TN+FP)

Area Under the ROC Curve (AUC)

- Another measure of model accuracy: Area Under the ROC Curve (AUC).
  - Given a random positive and a random negative outcome, proportion of the time you predict which is which is correct.
  - Less affected by sample (im)balance than accuracy.

- AUC=1 means perfect prediction, AUC=0.5 means pure guessing. In our model, AUC=0.775.

Homework

- Submit the members and name of your group by 10:00PM, February 24, Sunday.
- Join the class WeChat group (if you haven't done it yet).
- Finish the homework assignment (NO need to submit it).
- Read "*The Analytics Edge*", Chapters 21.2.