STOR 320 Modeling VII

Lecture 30

Yao Li

Department of Statistics and Operations Research

UNC Chapel Hill

Introduction

- Now We Consider
    - Categorical Response Variables
    - Numerical/Categorical Explanatory Variables
- Focus is on Classification
- Read Chapter 4 in ISLR

Introduction

- Basic Case: Binary Response
  - Variable Has Two Possible Outcomes
  - Typically, Yes or No Responses to a Question
  - Example
    - Y = Who Will Win the 2024 Presidential Election?
    - Y = Did You Pass Your STOR 320 Class?
    - Y = What Factors Influence the Admission into Graduate School?

Scenario

- Question: Are Students Who Get Good Grades Likely to be Admitted to Graduate School?
  - Y = Would the Student be Admitted to a Graduate School?
  - X = College GPA
- Why is Linear Regression Inappropriate?

$$P(Admission|X) = \beta_0 + \beta_1 X$$

Problem Setting

- Bernouilli Random Variable

$$Y = \begin{cases} 1 & if\ Yes \\ 0 & if\ No \end{cases}$$

$$p = E(Y) = P(Y = 1)$$

- Sample n Students

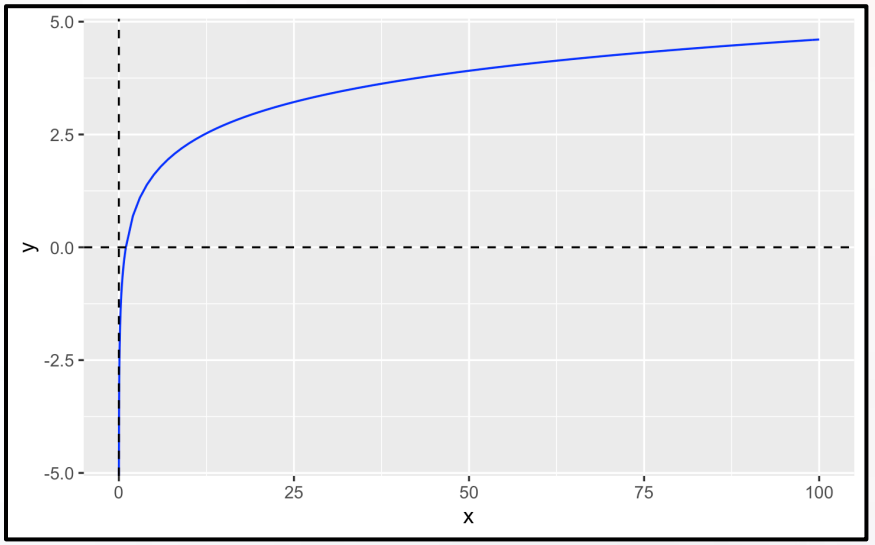
$$Y' = \sum Y_i \sim Binomial(n, p)$$

$$\hat{p} = \frac{\sum y_i}{n}$$

Estimated Probability that a Student Would be Admitted to a Graduate School

- Analyze the Effect of X on p: $p = E(Y|X) \neq \beta_0 + \beta_1 X$

Logit Link



- Modeling the Mean
  - Logit Link Function

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Odds of Admission

  - Understanding Odds
    - Odds of Admission = 1
    - Odds of Admission < 1
    - Odds of Admission > 1

Model Construction

- Solving for $\frac{p}{1-p}$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 X}$$

→ Odds of Admission Given the Student's GPA

- Solving for p

$$p = e^{\beta_0+\beta_1 X} - pe^{\beta_0+\beta_1 X}$$

$$p(1 + e^{\beta_0+\beta_1 X}) = e^{\beta_0+\beta_1 X}$$

$$p = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

→ Probability of Admission Given the Student's GPA

Logistic Regression for Classification

- Recall: $Y = \begin{cases} 1 & if\ Yes \\ 0 & if\ No \end{cases}$

- After Getting Data, We Estimate
  - $\hat{\beta}_0$
  - $\hat{\beta}_1$
  - $\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$ ➡ Estimated Probability of Admission Given the Student's GPA

- Two Scenarios
  - $\hat{p} < 0.5$ ➡ $\hat{Y} = 0$
  - $\hat{p} > 0.5$ ➡ $\hat{Y} = 1$

Evaluating the LR Model

- Two Methods
  - Leave Out Data Intentionally
  - Use Cross-Validation

- Positives and Negatives
  - True Positive = Predicted an Admission and the Student Got Admitted
  - False Positive=Predicted an Admission and the Student Didn't Get Admitted
  - False Negative = Predicted a Student Wouldn't be Admitted and They Did Get Admitted
  - True Negative = Predicted a Student Wouldn't be Admitted and They Didn't Get Admitted

Confusion Matrix

- Confusion Matrix

	Predicted	
Actual	*Will be Admitted*	*Won't be Admitted*
Admission	n_{11}	n_{12}
Isn't Admitted	n_{21}	n_{22}

  - Sensitivity:
    $$n_{11}/(n_{11} + n_{12})$$
  - Specificity:
    $$n_{22}/(n_{21} + n_{22})$$
  - False Positive Rate:
    $$n_{21}/(n_{21} + n_{22})$$
  - False Negative Rate:
    $$n_{12}/(n_{11} + n_{12})$$

Titanic: Data

- Titanic Survival Data

```
> library(titanic)
```

  - Response Variable

$$Y = \begin{cases} 1 & if\ \ Survived \\ 0 & if\ \ Did\ Not\ Survive \end{cases}$$

  - Explanatory Variables
    - Passenger Class
    - Sex
    - Age
    - Siblings/Spouses Aboard
    - Parents/Children Aboard
    - Passenger Fare
    - Port of Embarkation

Titanic: Data

- Titanic Survival Data (Continued)
  - Selecting Variables of Interest

```
> TRAIN=titanic_train[,c(2,3,5,6,7,8,10,12)]
> TEST=titanic_test[,c(2,4,5,6,7,9,11)])
```

  - Glimpse of Data

```
glimpse(TRAIN)

## Observations: 891
## Variables: 8
## $ Survived <int> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1,...
## $ Pclass   <int> 3, 1, 3, 1, 3, 3, 1, 3, 3
## $ Sex      <chr> "male", "female", "female
## $ Age      <dbl> 22, 38, 26, 35, 35, NA, 5
## $ SibSp    <int> 1, 1, 0, 1, 0, 0, 0, 3, 0
## $ Parch    <int> 0, 0, 0, 0, 0, 0, 0, 1, 2
## $ Fare     <dbl> 7.2500, 71.2833, 7.9250,
## $ Embarked <chr> "S", "C", "S", "S", "S",
```

```
glimpse(TEST)
```

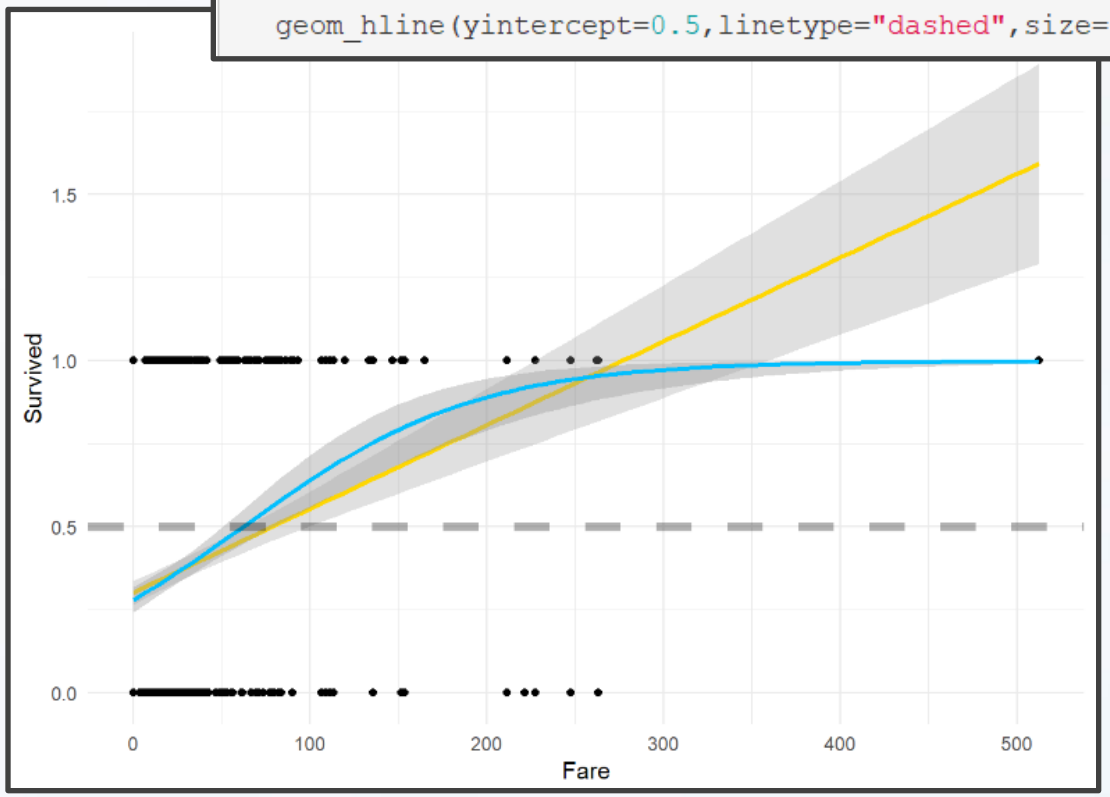
Problem?

```
## Observations: 418
## Variables: 7
## $ Pclass   <int> 3, 3, 2, 3, 3, 3, 3, 2, 3, 3, 3, 1, 1, 2, 1, 2, 2, 3,...
## $ Sex      <chr> "male", "female", "male", "male", "female", "male", "...
## $ Age      <dbl> 34.5, 47.0, 62.0, 27.0, 22.0, 14.0, 30.0, 26.0, 18.0,...
## $ SibSp    <int> 0, 1, 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 1, 1, 1, 1, 0, 0,...
## $ Parch    <int> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ Fare     <dbl> 7.8292, 7.0000, 9.6875, 8.6625, 12.2875, 9.2250, 7.62...
## $ Embarked <chr> "Q", "S", "Q", "S", "S", "S", "Q", "S", "C", "S", "S"...
```

Visualization: Survival vs. Fare

- Visualizing the Data

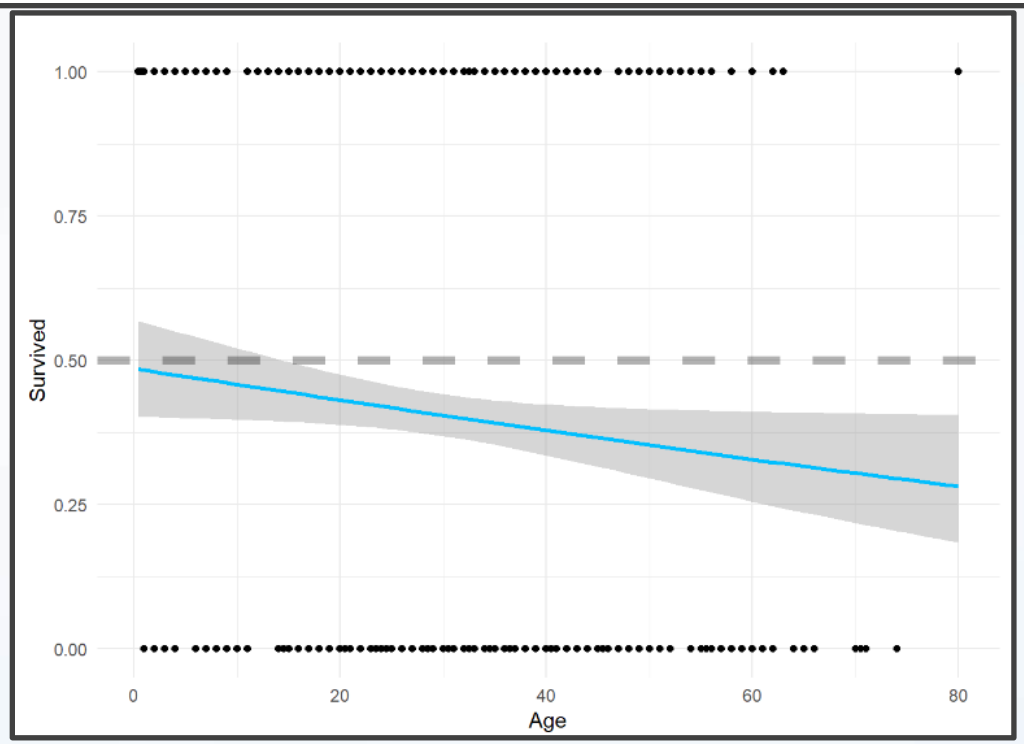
```
ggplot(TRAIN) + geom_point(aes(x=Fare,y=Survived)) + theme_minimal() +
  geom_smooth(aes(x=Fare,y=Survived),method="lm",alpha=0.3,color="gold") +
  geom_smooth(aes(x=Fare,y=Survived),method="glm",
              method.args=list(family="binomial"),color="deepskyblue1") +
  geom_hline(yintercept=0.5,linetype="dashed",size=2,alpha=0.3)
```



Visualization: Survival vs. Age

- Visualizing the Data (Continued)

```
ggplot(TRAIN) + geom_point(aes(x=Age,y=Survived)) + theme_minimal() +
  geom_smooth(aes(x=Age,y=Survived),method="glm",
              method.args=list(family="binomial"),color="deepskyblue1") +
  geom_hline(yintercept=0.5,linetype="dashed",size=2,alpha=0.3)
```



Visualization: Survival vs. Sex

- Visualizing the Data (Continued)

```
TRAIN %>%
  mutate(Sex=factor(Sex)) %>%
  group_by(Sex) %>%
  summarize(Prop.Survived=mean(Survived)) %>%
  ggplot() +
  geom_bar(aes(x=Sex,y=Prop.Survived),
           stat="Identity",fill="deepskyblue1") +
  theme_minimal() +
  theme(text=element_text(size=20))
```