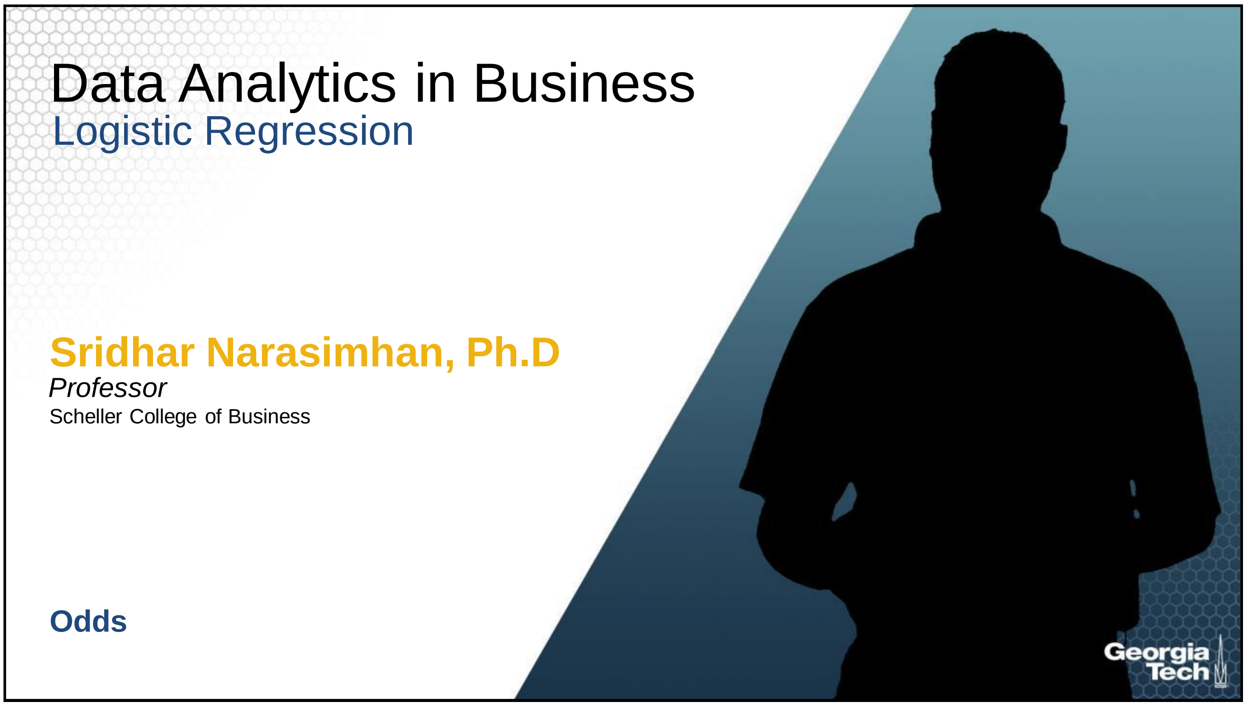
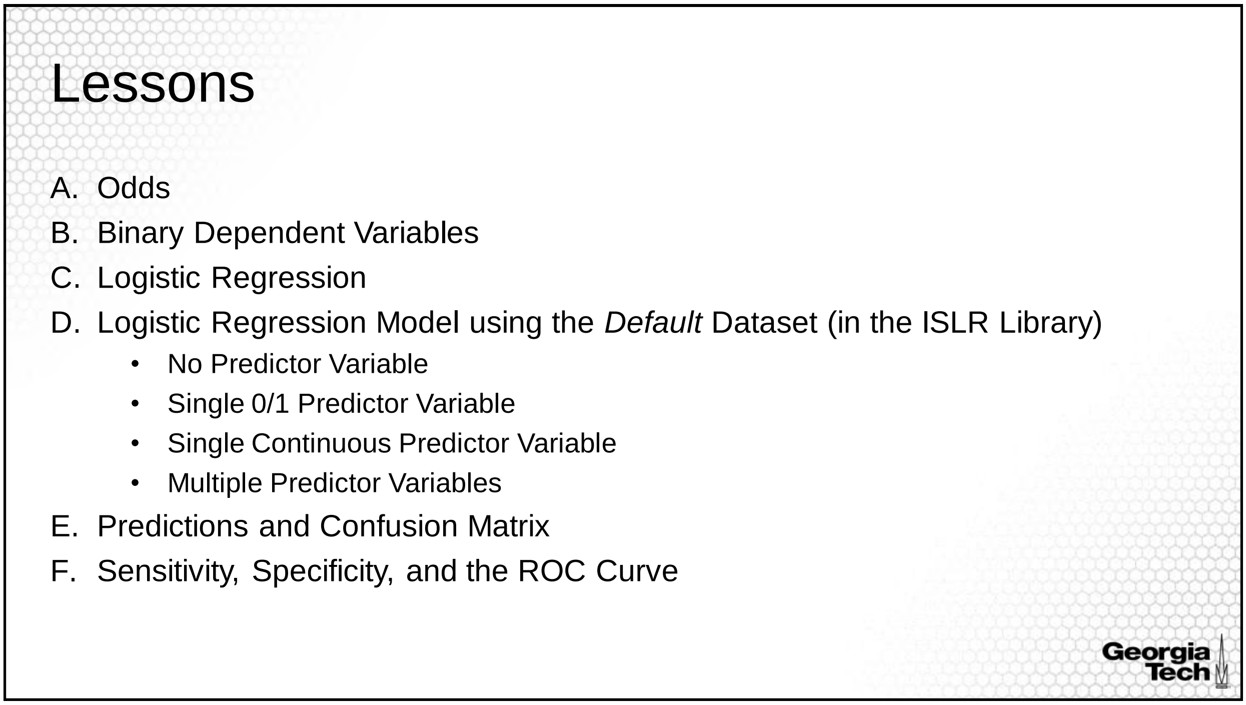


Data Analytics in Business

Logistic Regression

Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business

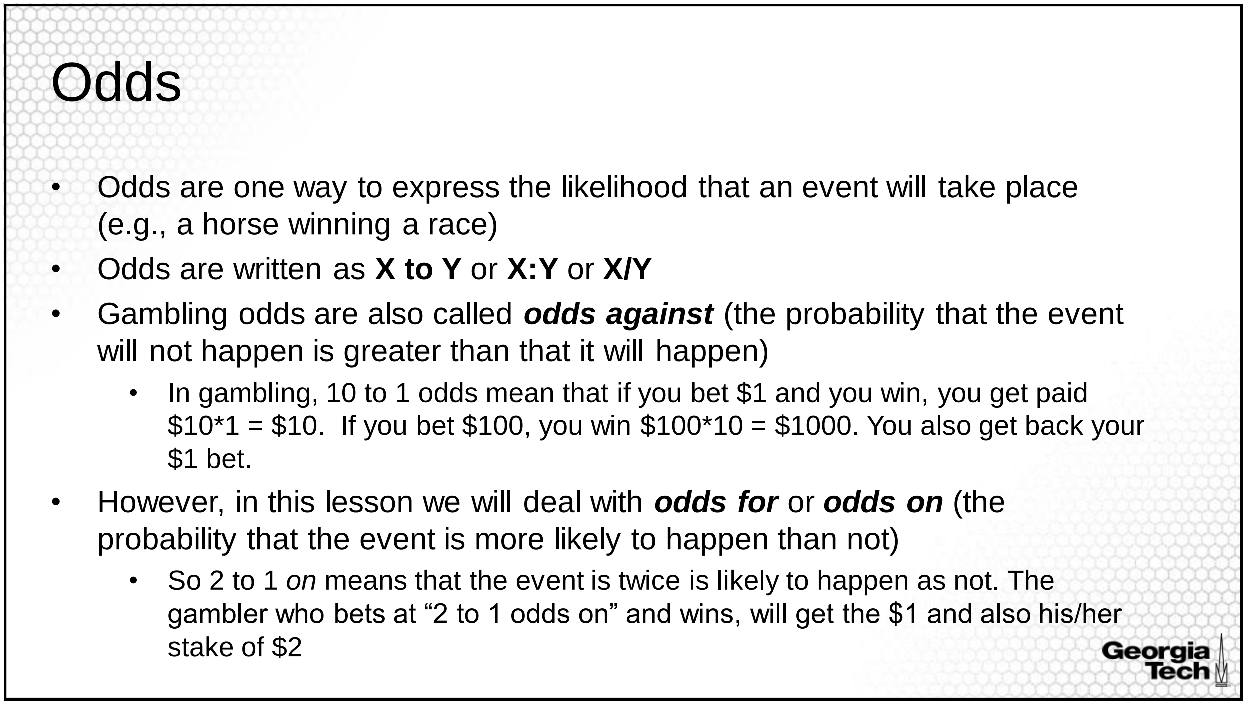
Odds



Lessons

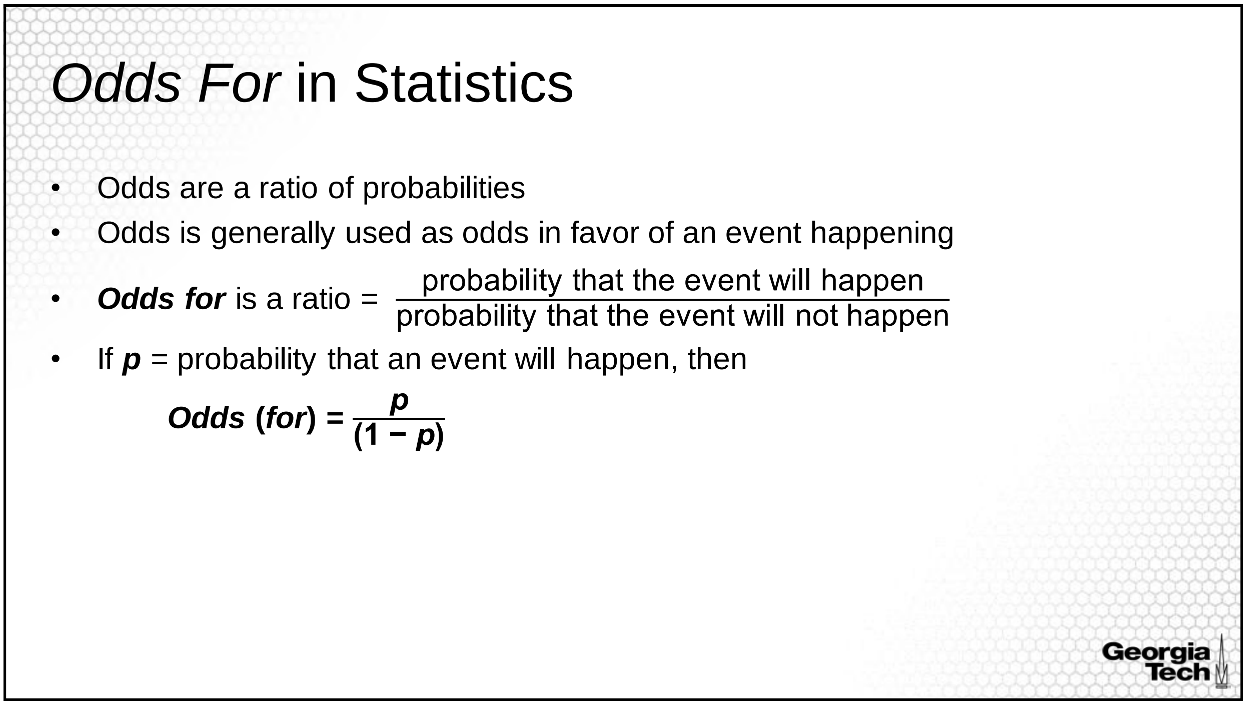
A. Odds
B. Binary Dependent Variables
C. Logistic Regression
D. Logistic Regression Model using the *Default* Dataset (in the ISLR Library)
- No Predictor Variable
- Single 0/1 Predictor Variable
- Single Continuous Predictor Variable
- Multiple Predictor Variables

E. Predictions and Confusion Matrix
F. Sensitivity, Specificity, and the ROC Curve



Odds

- Odds are one way to express the likelihood that an event will take place (e.g., a horse winning a race)
- Odds are written as **X to Y** or **X:Y** or **X/Y**
- Gambling odds are also called ***odds against*** (the probability that the event will not happen is greater than that it will happen)
  - In gambling, 10 to 1 odds mean that if you bet $1 and you win, you get paid $10*1 = $10. If you bet $100, you win $100*10 = $1000. You also get back your $1 bet.
- However, in this lesson we will deal with ***odds for*** or ***odds on*** (the probability that the event is more likely to happen than not)
  - So 2 to 1 *on* means that the event is twice is likely to happen as not. The gambler who bets at “2 to 1 odds on” and wins, will get the $1 and also his/her stake of $2



Odds For in Statistics

- Odds are a ratio of probabilities
- Odds is generally used as odds in favor of an event happening
- ***Odds for*** is a ratio = $\frac{\text{probability that the event will happen}}{\text{probability that the event will not happen}}$
- If ***p*** = probability that an event will happen, then

$$\textbf{\textit{Odds}}\ \mathbf{(}\textbf{\textit{for}}\mathbf{)} = \frac{\boldsymbol{p}}{\mathbf{(1-}\boldsymbol{p}\mathbf{)}}$$

Odds For in Statistics

- Using the previous example, if *Odds for* is 2:1, then

$$Odds\ (for) = \frac{2}{1} = \frac{p}{(1-p)}\text{ , thus}$$

$$2(1 - p) = p$$

$$2 - 2p = p$$

$$2 = 3p$$

$$p = \frac{2}{3}\text{ or 0.6667 or 66.67\%}$$

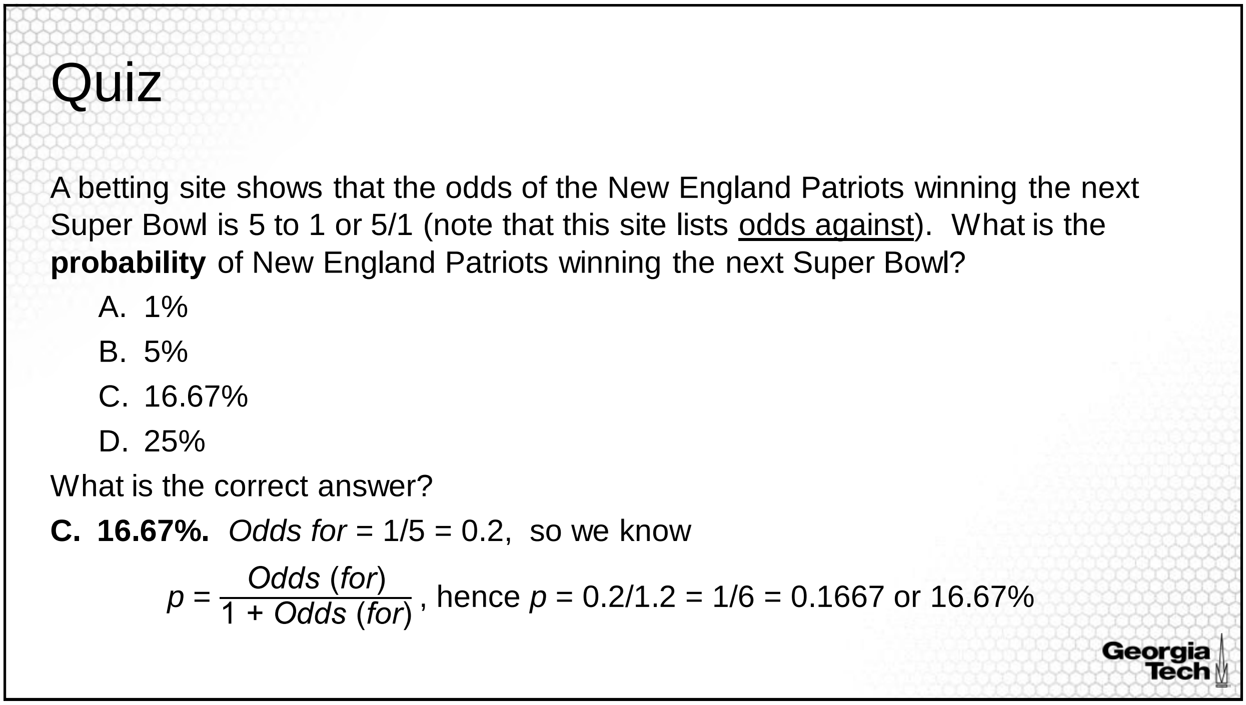
Odds and Probability

- Knowing *Odds* (*for*) we can get the value of *p* by using the equation:

$$p = \frac{Odds\ (for)}{1 + Odds\ (for)}$$

- Some examples of *Odds* and their respective *p* values are shown in the table

Odds	***Odds* (*for*)**	***p***	**(1 – *p*)**
2:1	2/1 = 2	0.67	0.33
3:2	3/2 = 1.5	0.60	0.40
3:1	3/1 = 3	0.75	0.25
4:1	4/1 = 4	0.80	0.20
9:1	9/1 = 9	0.90	0.10
10:1	10/1 = 10	0.91	0.09
1:2	1/2 = 0.5	0.33	0.67
1:3	1/3 = 0.33	0.25	0.75
2:3	2/3 =0.67	0.40	0.60
1:4	1/4 = 0.25	0.20	0.80



Quiz

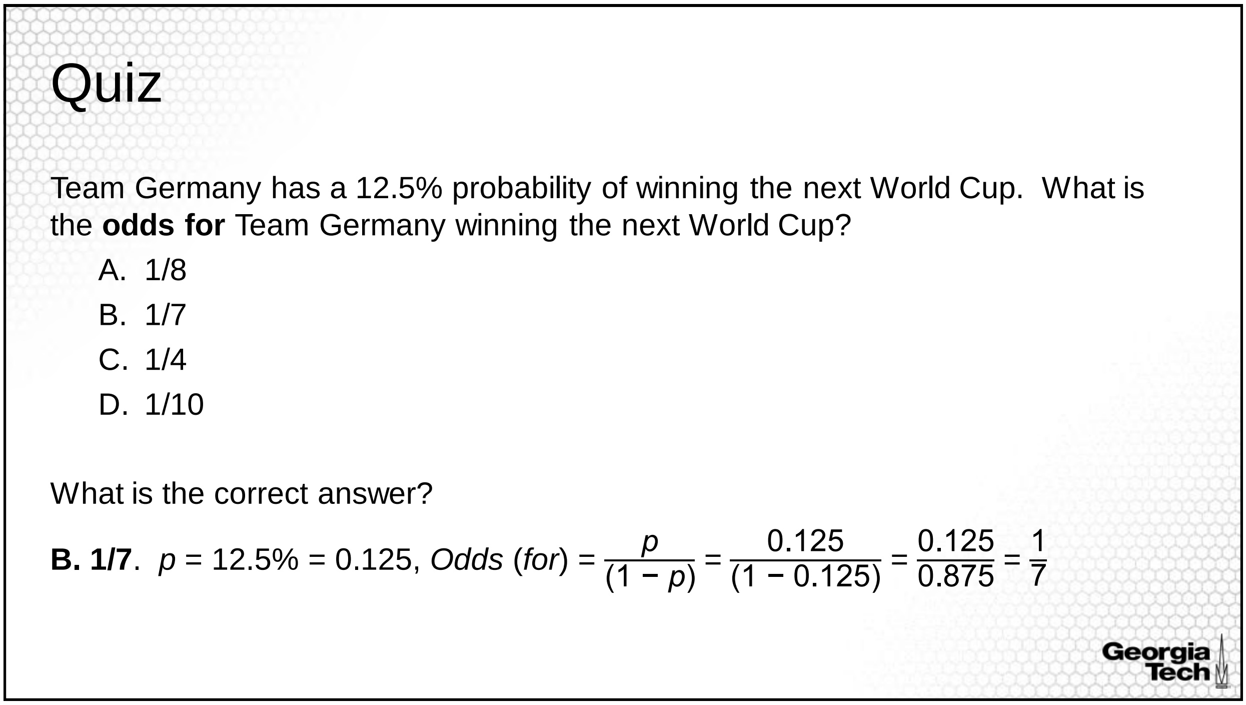
A betting site shows that the odds of the New England Patriots winning the next Super Bowl is 5 to 1 or 5/1 (note that this site lists odds against). What is the **probability** of New England Patriots winning the next Super Bowl?

A. 1%
B. 5%
C. 16.67%
D. 25%

What is the correct answer?

C. 16.67%. *Odds for* = 1/5 = 0.2, so we know

$$p = \frac{Odds\ (for)}{1 + Odds\ (for)}\text{, hence } p = 0.2/1.2 = 1/6 = 0.1667 \text{ or } 16.67\%$$



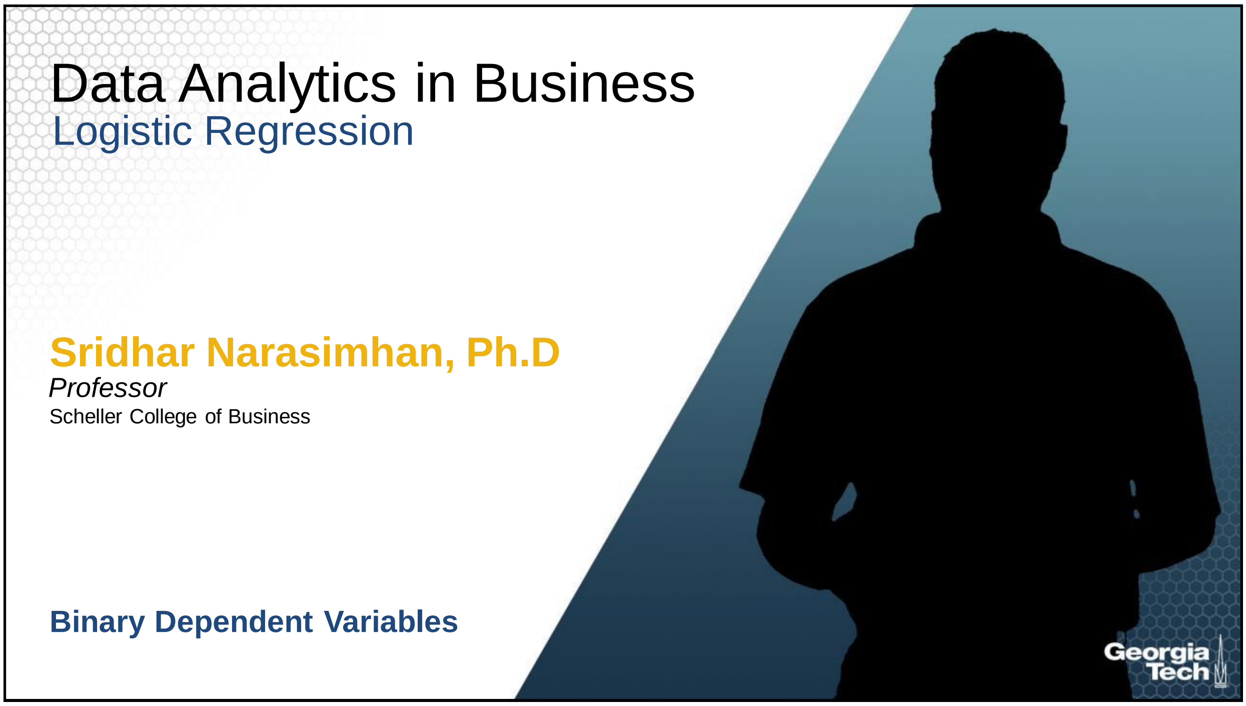
Quiz

Team Germany has a 12.5% probability of winning the next World Cup. What is the **odds for** Team Germany winning the next World Cup?

A. 1/8
B. 1/7
C. 1/4
D. 1/10

What is the correct answer?

B. 1/7. p = 12.5% = 0.125, $Odds\ (for) = \frac{p}{(1-p)} = \frac{0.125}{(1-0.125)} = \frac{0.125}{0.875} = \frac{1}{7}$



Data Analytics in Business

Logistic Regression

Sridhar Narasimhan, Ph.D

Professor

Scheller College of Business

Binary Dependent Variables

Relationship Between *Odds for* and *p*

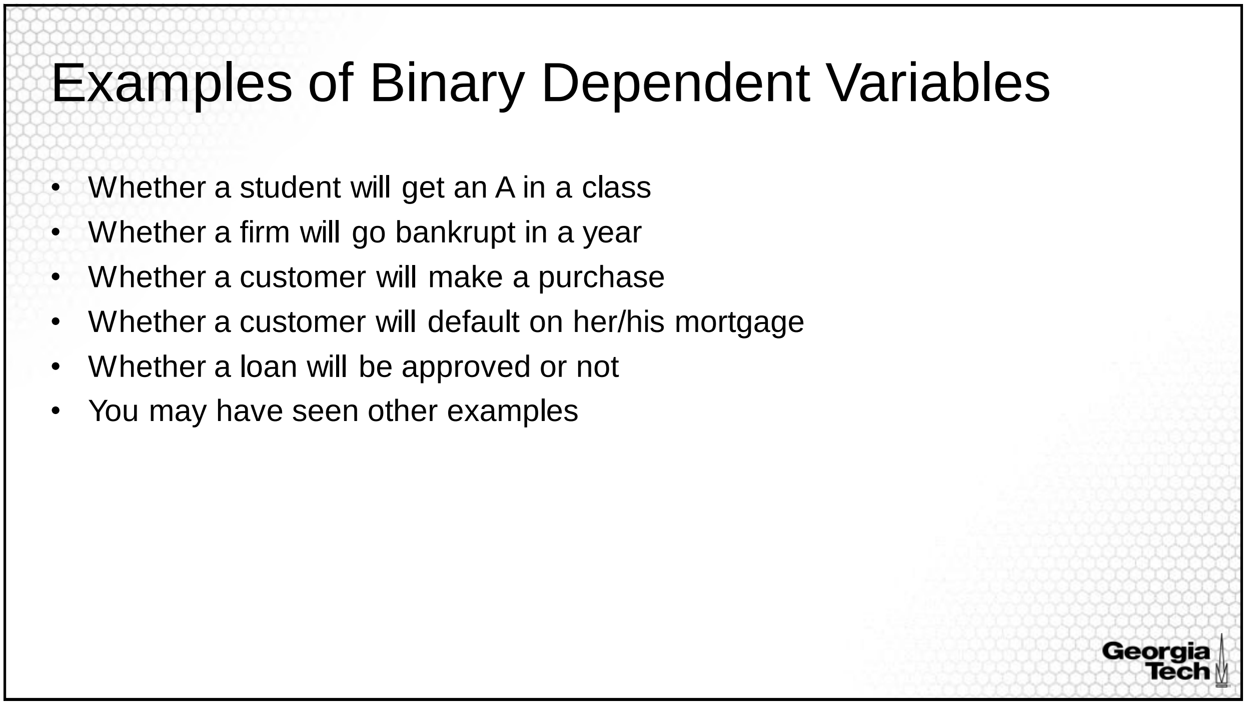
- Relationship between *Odds for* and *p* (the probability that an event will happen)
- Knowing *p*, we can get *Odds* (*for*) by using the equation:

$$Odds\ (for) = \frac{p}{(1 - p)}$$

- Similarly, knowing *Odds* (*for*) we can get the value of *p* by using the equation:

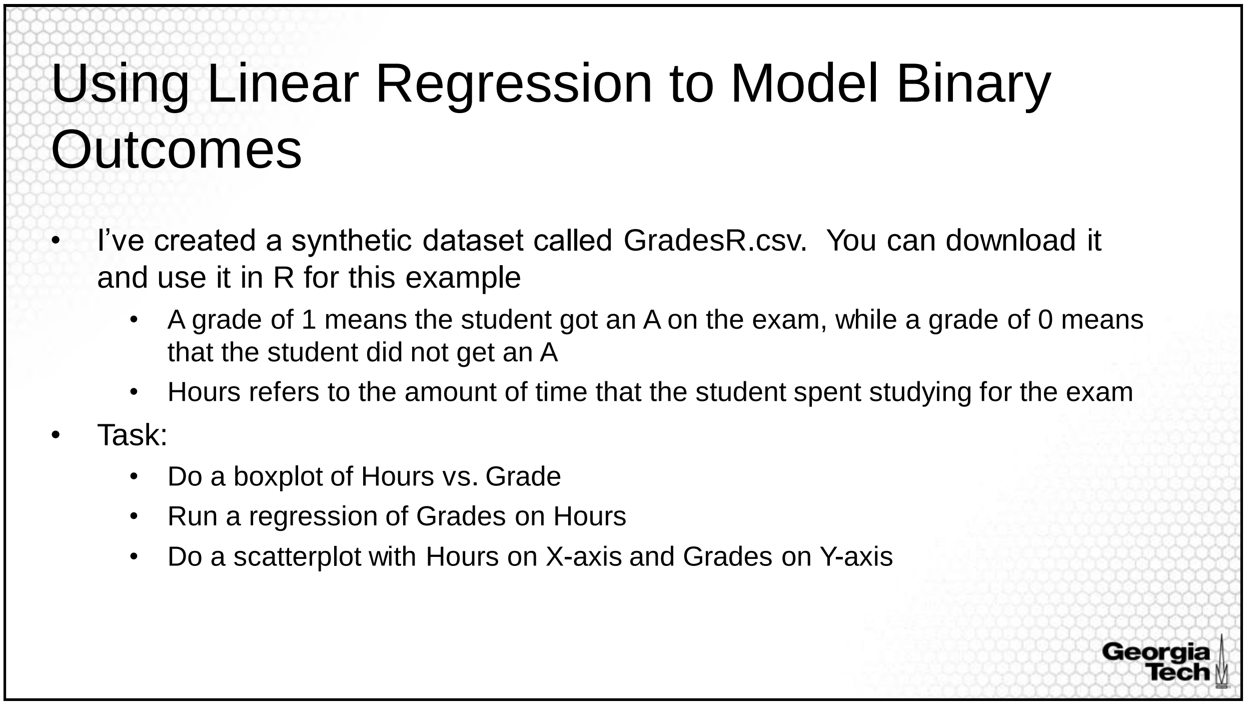
$$p = \frac{Odds\ (for)}{1 + Odds\ (for)}$$

Georgia Tech



Examples of Binary Dependent Variables

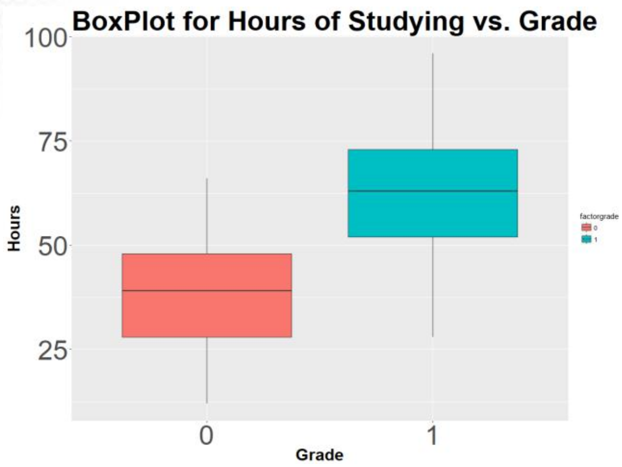
- Whether a student will get an A in a class
- Whether a firm will go bankrupt in a year
- Whether a customer will make a purchase
- Whether a customer will default on her/his mortgage
- Whether a loan will be approved or not
- You may have seen other examples



Using Linear Regression to Model Binary Outcomes

- I've created a synthetic dataset called GradesR.csv. You can download it and use it in R for this example
  - A grade of 1 means the student got an A on the exam, while a grade of 0 means that the student did not get an A
  - Hours refers to the amount of time that the student spent studying for the exam
- Task:
  - Do a boxplot of Hours vs. Grade
  - Run a regression of Grades on Hours
  - Do a scatterplot with Hours on X-axis and Grades on Y-axis

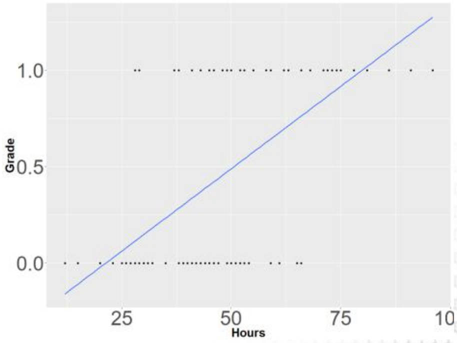
Boxplot of Hours for Each Grade



Using Linear Regression to Model Binary Outcomes

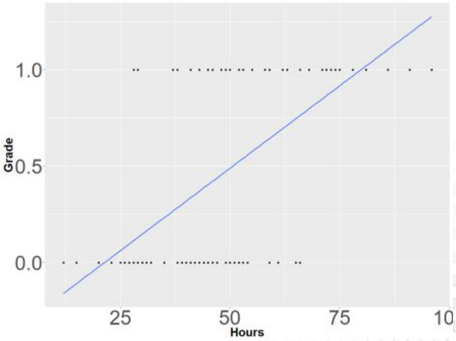
- Is this a good model? How good is the fit?

```
lm(formula = Grade ~ Hours, data = logit_grade)

Residuals:
    Min        1Q       Median      3Q        Max
-0.7630   -0.3060   -0.0284    0.2883    0.8862

Coefficients:
                  Estimate     Std. Error    t value      Pr(>|t|)
(Intercept)    -0.364520     0.111263     -3.276      0.00146 **
Hours            0.017084     0.002084      8.197      9.61e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error:  0.3889 on 98 degrees of freedom
Multiple R-squared:   0.4068,    Adjusted R-squared:  0.4007
F-statistic:  67.2 on 1 and 98 DF,    p-value: 9.606e-13
```



Using Linear Regression to Model Binary Outcomes

- Questions:
  - Are all the predicted values either = 0 or 1?
  - Are some of the predicted values below 0?
  - Are some of the predicted values above 1?
- All the predicted values lie on the regression line!
- How does one predict a grade of 1 or 0 using these predicted values on the regression line? What should we do?



Data Analytics in Business

Logistic Regression

Sridhar Narasimhan, Ph.D

Professor

Scheller College of Business

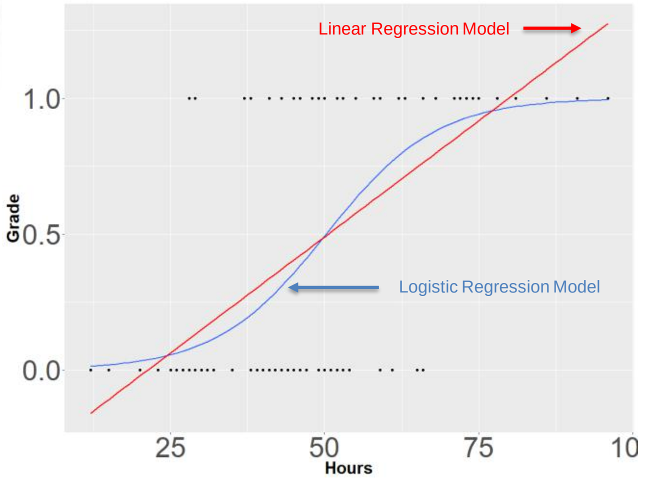
Logistic Regression

Logistic Regression

Logistic regression is similar to linear regression, but with two main differences:

1. Y (outcome or response) is categorical
   - Yes/No
   - Approve/Reject
   - Responded/Did Not Respond
   - Pass/Fail
2. Result is expressed as a ***probability*** of being in a group; this implies that the predicted value is always between 0 and 1

Comparing the Logistic and Linear Regression Models


Linear Regression Model

Logistic Regression Model

Logistic Regression

- We use the ***logistic function*** in logistic regression, which gives us the probability of being in a group
- Let *p(x)* = Prob(y = 1|x), or Probability that y=1 given a value of x
- We define the logistic function as:

  $p(x) = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}$,

  i.e., $p(x)$ = exp(b_0 + b_1x)/[1 + exp(b_0 + b_1x)]
- *p(x)* has the property that it is always between 0 and 1 for all values of x

Logistic Regression

- Let *p* = *p(x)* to simplify the notation:
  - p = exp(b_0 + b_1x)/[1 + exp(b_0 + b_1x]
  - 1 – p = 1 $-\frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}} = \frac{1}{1+e^{b_0+b_1x}}$ = 1/[1 + exp(b_0 + b_1x)]
  - Therefore, p/(1 – p) = exp(b_0 + b_1x)
- Taking natural logs on both sides, we get:
  - log(p/(1 – p)) = b_0 + b_1x
- As we have previously defined, *Odds* (*for*) = $\frac{p}{(1-p)}$, therefore
  - log(p/(1 – p)) is the log of odds for, or “logit,” and
  - The logit model is: **logit(*p*) = log(*p*/(1 – *p*)) = b_0 + b_1x**
- All other components of the regression model are the same

Why Transform from Probability to Log Odds?

- Why do we do the transformation from probability to log odds?
  - Because it is usually difficult to model a variable which has restricted range, such as probability
  - The transformation is an attempt to get around the restricted range problem because it maps probability ranging between 0 and 1 to log odds ranging from negative infinity to positive infinity
  - Another reason is that, among all of the infinite choices of transformation, the log of odds is one of the easiest to understand and interpret
- The transformation is called logit and **logit(*p*) = log(*p*/(1 – *p*)) =** $\boldsymbol{b_0 + b_1 x}$

Georgia Tech

Interpreting the Logistic Regression Model

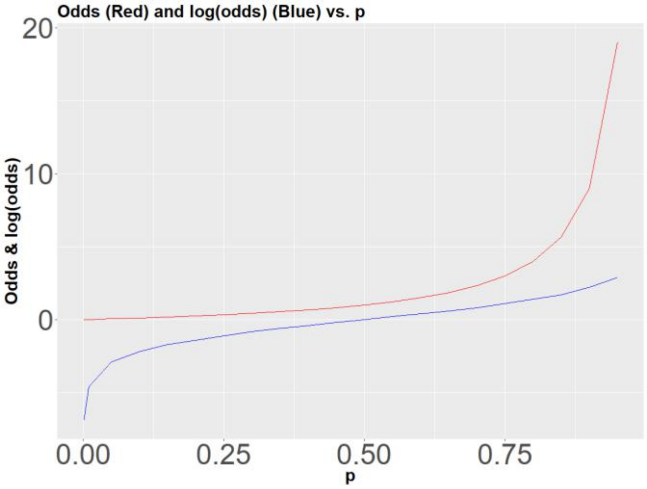
- **logit(*p*) =** log(p/(1 – p)) = $b_0 + b_1 x$ means that as *x* increases by 1 unit, the natural log of the odds increases by b_1
- This is the same as the odds increasing by a factor of exp(b_1), which is roughly 100*b_1 percent
- Note that the exact odds change is $(e^{b_1}$ - 1)*100 %

Georgia Tech

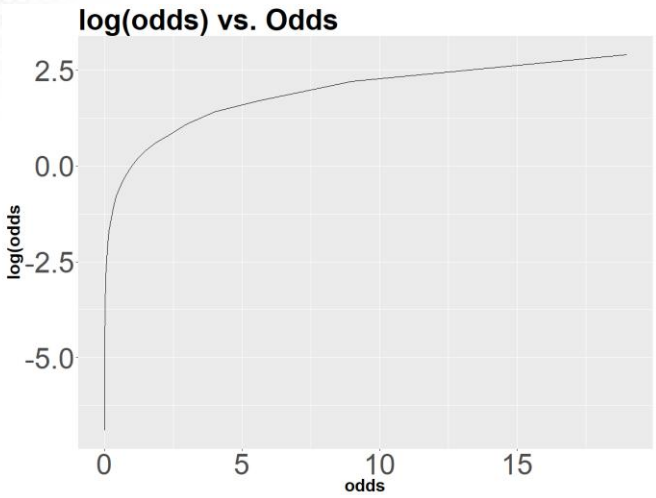
Logistic Regression

p	odds = *p*/(1-*p*)	ln(odds)
0.001	0.001	-6.9
0.01	0.010	-4.6
0.05	0.053	-2.9
0.1	0.111	-2.2
0.15	0.176	-1.7
0.2	0.250	-1.4
0.25	0.333	-1.1
0.3	0.429	-0.8
0.35	0.538	-0.6
0.4	0.667	-0.4
0.45	0.818	-0.2
0.5	1.000	0.0
0.55	1.222	0.2
0.6	1.500	0.4
0.65	1.857	0.6
0.7	2.333	0.8
0.75	3.000	1.1
0.8	4.000	1.4
0.85	5.667	1.7
0.9	9.000	2.2
0.95	19.000	2.9
0.999	999.000	6.9
0.9999	9999.000	9.2

Probability, Odds, and Log of Odds



Log(odds) vs. Odds



Quiz

The logistic function, *p(x),* returns values

A. Between -10, + 10
B. Between -1, 0
C. Between (-infinity, + infinity)
D. Between 0, 1

What is the correct answer?

D. Between 0, 1

Quiz

log(p/(1 – p)) = $b_0 + b_1x$ means that as x increases by 1 unit,

A. the natural log of the odds increases by b_1
B. the odds increase by a factor of exp(b_1)
C. the odds increase by roughly 100*b_1 percent
D. All of the Above
E. None of the Above

What is the correct answer?

D. All of the Above (i.e., A, B, and C)

Data Analytics in Business

Logistic Regression

Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business

Logistic Regression Model
Using the *Default* Dataset

The *Default* Dataset in ISLR Package

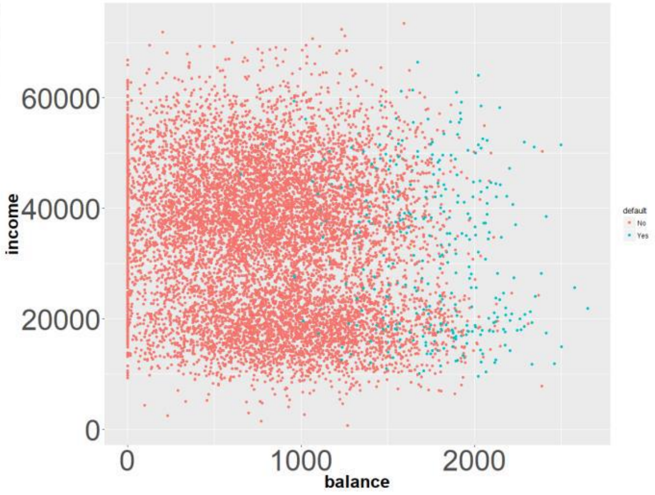
The *Default* Dataset:

```
str(ISLR::Default)
'data.frame':  10000 obs. of 4 variables:
 $ default: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ student: Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 2 1 1 ...
 $ balance: num  730 817 1074 529 786 ...
 $ income : num  44362 12106 31767 35704 38463 ...
```

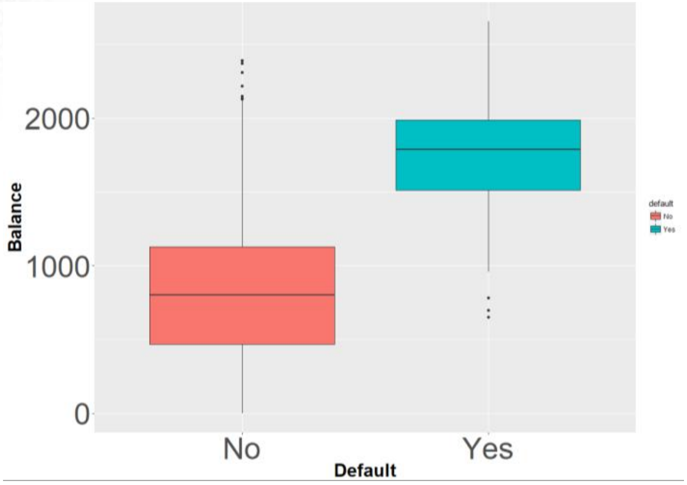
Create a new dataframe df <- ISLR::Default and add two dummy variables

$$\text{dft} = \begin{cases} 1, & if\ default = "Yes" \\ 0, & otherwise \end{cases} \qquad \text{stdt} = \begin{cases} 1, & if\ student = "Yes" \\ 0, & otherwise \end{cases}$$

Income vs. Balance (*defaulters* in Blue)



BoxPlot for Balance vs. Default Status



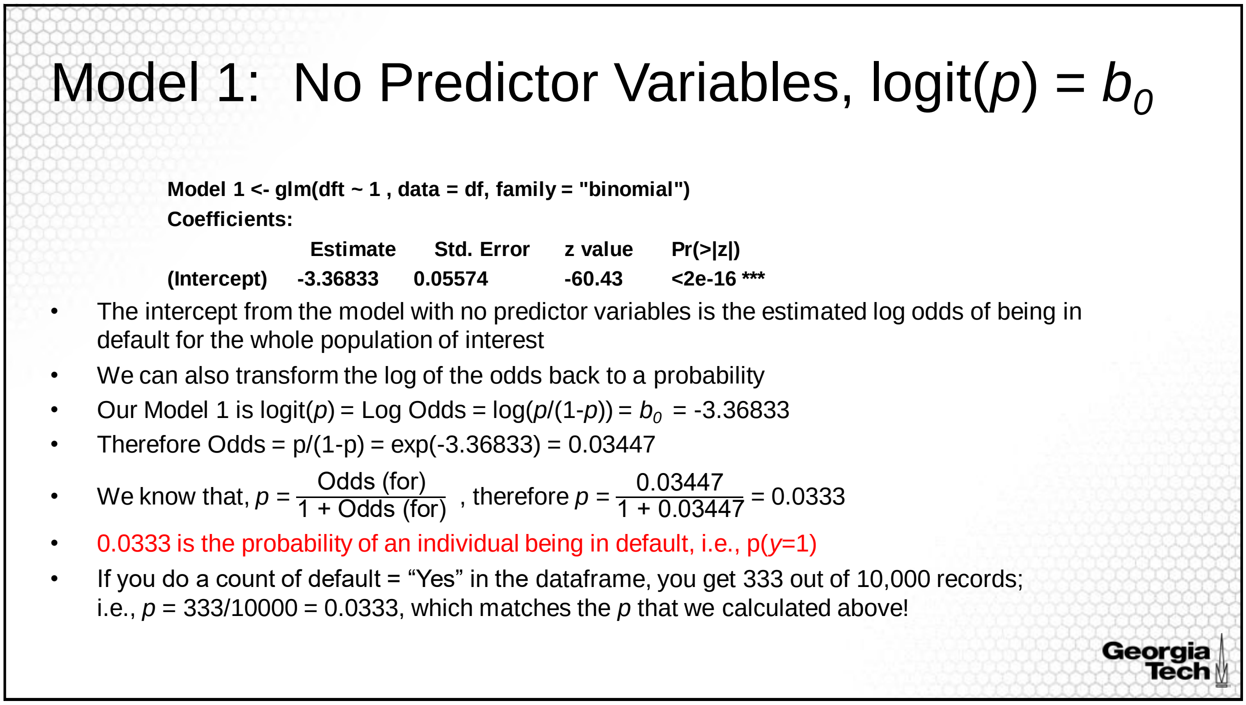
Logistic Regression Models

We will look at the following:

- Model 1 (No Predictor Variables): logit(p) = b_0
- Model 2 (Single 0/1 Predictor Variable): logit(p) = b_0 + b_1*$stdt$
- Model 3 (Single Continuous Predictor Variable): logit(p) = b_0 + b_1*$balance$
- Model 4 (Multiple Predictors): logit(p) = b_0 + b_1*$balance$ + b_2*income + b_3*$stdt$

Where $p = \frac{Odds\ (for)}{1 + Odds\ (for)}$, probability that default = “Yes,” and

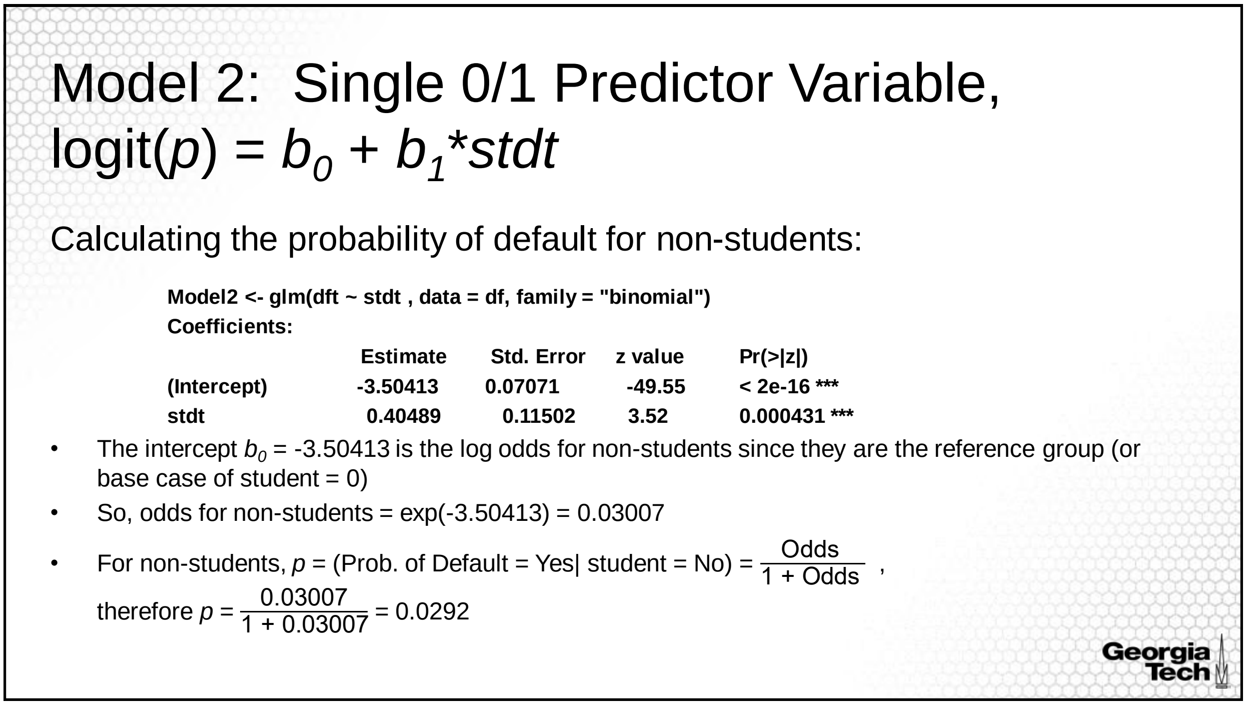
logit(p) = log(p/(1 – p))



Model 1: No Predictor Variables, logit(p) = b_0

```
Model 1 <- glm(dft ~ 1 , data = df, family = "binomial")
Coefficients:
               Estimate    Std. Error    z value    Pr(>|z|)
(Intercept)   -3.36833    0.05574        -60.43     <2e-16 ***
```

- The intercept from the model with no predictor variables is the estimated log odds of being in default for the whole population of interest
- We can also transform the log of the odds back to a probability
- Our Model 1 is logit(p) = Log Odds = log(p/(1-p)) = b_0 = -3.36833
- Therefore Odds = p/(1-p) = exp(-3.36833) = 0.03447
- We know that, $p = \frac{\text{Odds (for)}}{1 + \text{Odds (for)}}$, therefore $p = \frac{0.03447}{1 + 0.03447} = 0.0333$
- 0.0333 is the probability of an individual being in default, i.e., p(y=1)
- If you do a count of default = "Yes" in the dataframe, you get 333 out of 10,000 records; i.e., p = 333/10000 = 0.0333, which matches the p that we calculated above!



Model 2: Single 0/1 Predictor Variable, logit(p) = b_0 + b_1**stdt*

Calculating the probability of default for non-students:

```
Model2 <- glm(dft ~ stdt , data = df, family = "binomial")
Coefficients:
                      Estimate     Std. Error    z value      Pr(>|z|)
(Intercept)          -3.50413     0.07071        -49.55       < 2e-16 ***
stdt                  0.40489      0.11502        3.52        0.000431 ***
```

- The intercept b_0 = -3.50413 is the log odds for non-students since they are the reference group (or base case of student = 0)
- So, odds for non-students = exp(-3.50413) = 0.03007
- For non-students, p = (Prob. of Default = Yes| student = No) = $\frac{\text{Odds}}{1 + \text{Odds}}$,

  therefore $p = \frac{0.03007}{1 + 0.03007} = 0.0292$

Model 2: Single 0/1 Predictor Variable, logit(p) = $b_0 + b_1{*}stdt$

Calculating the probability of default for students:

```
Model2 <- glm(dft ~ stdt , data = df, family = "binomial")
Coefficients:
                  Estimate    Std. Error    z value      Pr(>|z|)
(Intercept)      -3.50413     0.07071       -49.55       < 2e-16 ***
stdt              0.40489      0.11502       3.52        0.000431 ***
```

- The coefficient for student (b_1 = 0.40489) is the amount that we have to add to b_0 to get the log odds for students = -3.50413 + 0.40489 = -3.09924
- So, odds for students = exp(-3.09924) = 0.04508
- For students, p = (Prob. of Default = Yes| student = Yes) $= \frac{\text{Odds}}{1 + \text{Odds}}$,

  therefore $p = \frac{0.04508}{1 + 0.04508} = 0.0431$
- Students have higher default probability than non-students!

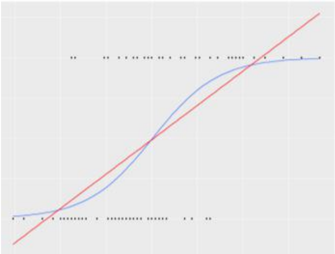
Model 3: Single Continuous Predictor Variable, logit(p) = $b_0 + b_1{*}balance$

```
Model3 <- glm(dft ~ balance, data = df, family = "binomial")
Coefficients:
               Estimate      Std. Error    z value     Pr(>|z|)
(Intercept)    -1.065e+01    3.612e-01     -29.49      <2e-16 ***
balance        5.499e-03     2.204e-04      24.95      <2e-16 ***
```

- Note that b_1 = 0.0055.
- An increase in the balance is associated with increasing the log odds of default, hence the odds, and hence the probability of default.
- Adding one unit (i.e., $1) to the balance increases the log odds of default by 0.0055.

Making Predictions
(*p*(*x*) = probability of default) Using Model 3

- $p(x) = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}$, b_0 = -10.65 and b_1 = 0.0055
- If *x* = \$1,000, then *p*(*x*) = $\frac{e^{-10.65\ +0.0055*1000}}{1+e^{-10.65\ +0.0055*1000}}$ = 0.00576, or less than 1%
- If *x* = \$1,500, then *p*(*x*) = $\frac{e^{-10.65\ +0.0055*1500}}{1+e^{-10.65\ +0.0055*1500}}$ = 0.08317, or 8.3%
- If *x* = \$2,000, then *p*(*x*) = $\frac{e^{-10.65\ +0.0055*2000}}{1+e^{-10.65\ +0.0055*2000}}$ = 0.5866, or 58.7%



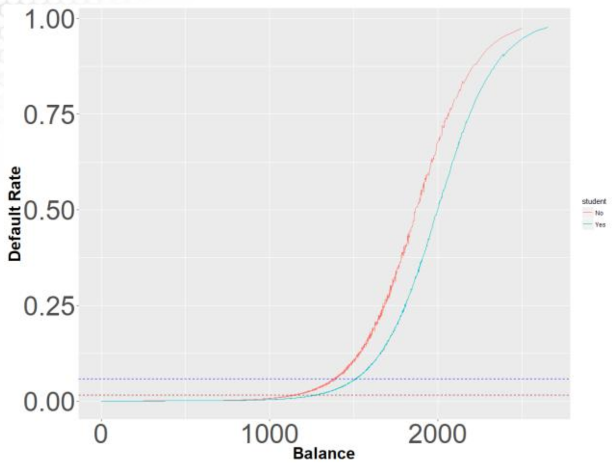
For the logit model, increasing *x* by \$500 has a nonlinear effect on p(*x*)

Model 4: Multiple Predictors,
logit(*p*) = b_0 + b_1**balance* + b_2*income + b_3**stdt*

```
Model4 <- glm(dft ~ balance + income + stdt, data = df, family = "binomial")
Coefficients:
                Estimate        Std. Error     z value      Pr(>|z|)
(Intercept)     -1.087e+01      4.923e-01      -22.080      < 2e-16 ***
balance          5.737e-03      2.319e-04       24.738      < 2e-16 ***
income           3.033e-06      8.203e-06        0.370      0.71152
stdt            -6.468e-01      2.363e-01       -2.738      0.00619 **
```

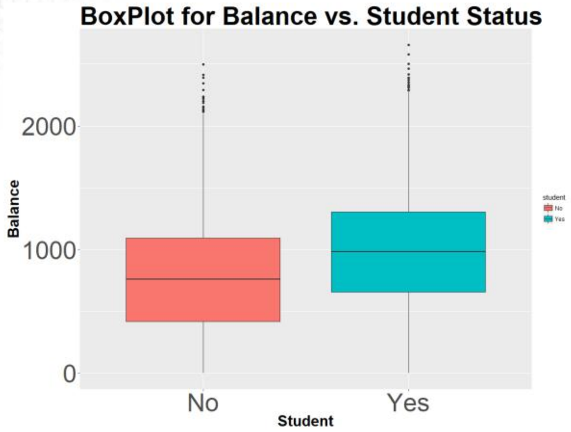
- Note that b_0 = -10.87, b_1 = 0.0057, b_2 = 0.000003, b_3 = -0.65
- An increase in the balance is associated with increasing the log odds of default, hence the odds, and hence the probability of default
- Adding one unit (i.e., one \$) to the balance increases the log odds of default by 0.0057
- The coefficient of student = -0.65, implies that at a fixed value of balance and income, students are less likely to default than non-students! This is different from the result that you saw in Model 2

Default Rates for Students and Non-Students



- At the same credit card balance, an individual student will have a lower probability of default than a non-student (logistics curves)
- On the whole, students carry more credit card balances, thus, overall, students default at a higher rate than non-students (horizontal lines)
- This phenomenon is called ***confounding***

Students (On Average) Have a Bigger Balance Than Non-Students



- There is correlation between Student and Balance

Quiz

For this logistic model logit(*p*) = b_0, where Y=1 is the default case, b_0 means that the intercept from the model with no predictor variables is the estimated

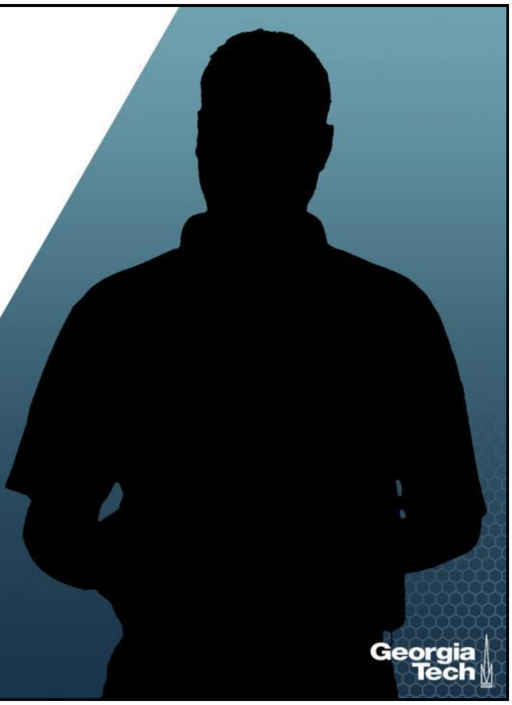
A. odds of being in default for the whole population of interest
B. log odds of being in default for the whole population of interest
C. probability of being in default for the whole population of interest
D. log odds of not being in default for the whole population of interest

What is the correct answer?

B. log odds of being in default for the whole population of interest

Data Analytics in Business

Logistic Regression



Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business

Predictions and Confusion Matrix

Making Predictions on the Fitted Data

- Let's use **Model 4: logit(*p*) = b_0 + b_1*balance + b_2*income + b_3*stdt**
- We can make predictions on the fitted data, and add them to the dataframe *df*, that we defined earlier

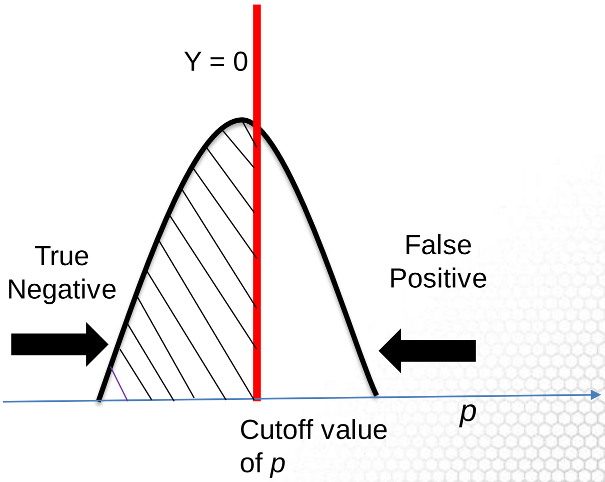
```
df <-  df %>%
        mutate(pred_prob_model4 = predict(Model4, newdata = ., type = "response")) %>%
        mutate(pred_outcome_model4 = ifelse(pred_prob_model4 >= 0.5,1,0))
# we are using 0.5 as cutoff for predicting Y=1.
View(df)
```

- We now have:

$$\text{pred\_outcome\_model4} = \begin{cases} 1, & \text{if pred\_prob\_model4 is equal or greater than 0.5 (50\%);} \\ 0, & otherwise \end{cases}$$

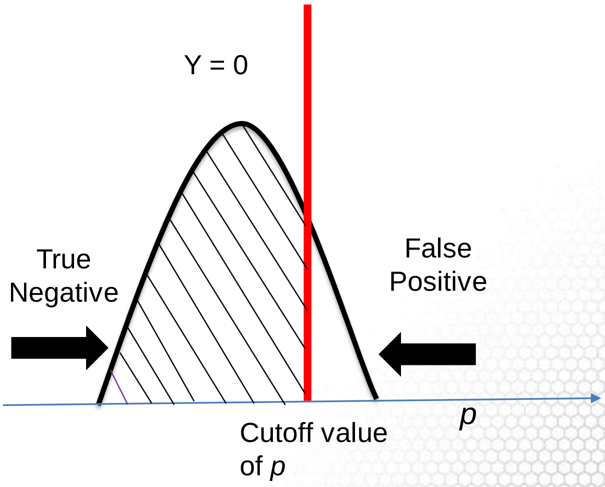
True Negative and False Positive

- Consider all the **Y = 0** observations
- For each of these observations, we use the Logit model to make a prediction (using the X values)
- If the predicted value is 0, we get a **true negative**
- If the predicted value is 1, we get a **false positive**
- If we increase the cutoff value of *p*, the true negatives will increase and the false positives will decrease



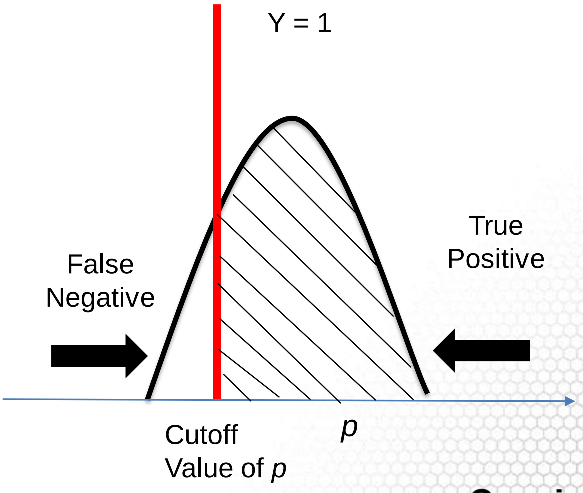
True Negative and False Positive

- Consider all the **Y = 0** observations
- For each of these observations, we use the Logit model to make a prediction (using the X values)
- If the predicted value is 0, we get a **true negative**
- If the predicted value is 1, we get a **false positive**
- If we increase the cutoff value of *p*, the true negatives will increase and the false positives will decrease



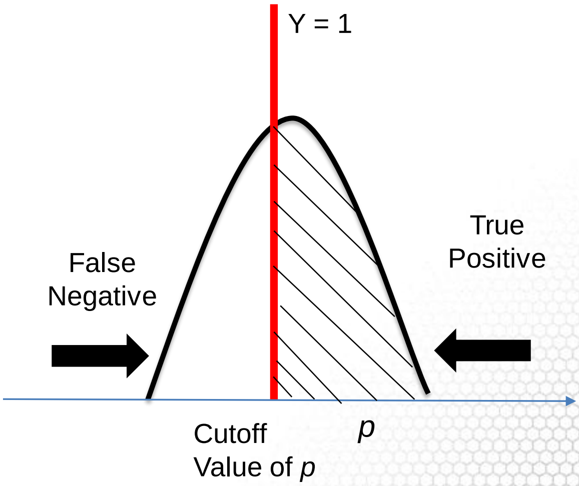
True Positive and False Negative

- Consider all the **Y = 1** data observations
- For each observation, we use the Logit model to make a prediction (using the X values)
- If the predicted value = 1, we get a **true positive**
- If the predicted value = 0, we get a **false negative**
- If we increase the cutoff, the false negatives will increase and the true positives will decrease



True Positive and False Negative

- Consider all the **Y = 1** data observations
- For each observation, we use the Logit model to make a prediction (using the X values)
- If the predicted value = 1, we get a **true positive**
- If the predicted value = 0, we get a **false negative**
- If we increase the cutoff, the false negatives will increase and the true positives will decrease



Confusion Matrix

		Predicted Value (pred_outcome_model4)		
		0	1	Total
True Value	0			9667
	1			333
	Total			10000

- A confusion matrix is where you record the performance of a classifier
- In this lesson, we want to gauge how our logit models perform
- In the ISLR: *Default* dataset, we have 333 cases with default = “Yes” (or Y =1) and 9667 records with default = “No” (or Y = 0). We have the row totals for this confusion matrix
- We need to record the predicted value (i.e., $\hat{Y}$) after we fit our logit model on this dataset

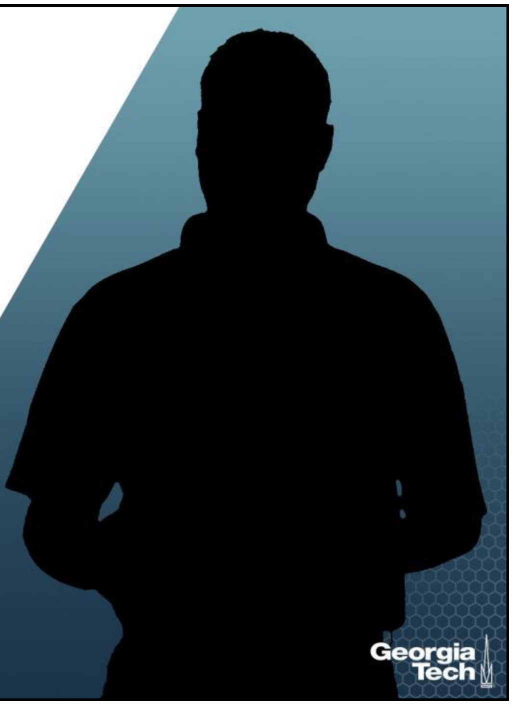
Confusion Matrix

		Predicted Value (pred_outcome_model4)		
		0	1	Total
True Value	0	9627 (true negative)	40 (false positive)	9667
	1	228 (false negative)	105 (true positive)	333
	Total			10000

- **xtabs(~dft + pred_outcome_model4, data = df)**
- We have fitted Model 4, and have used estimated predicted p = 0.5 as a cutoff for setting $\hat{Y}$ to 1. Otherwise, we set $\hat{Y}$ = 0
  - What is the count of True Negatives? 9627
  - What is the count of True Positives? 105
  - What is the count of False Positives? 40
  - What is the count of False Negatives? 228

Data Analytics in Business

Logistic Regression

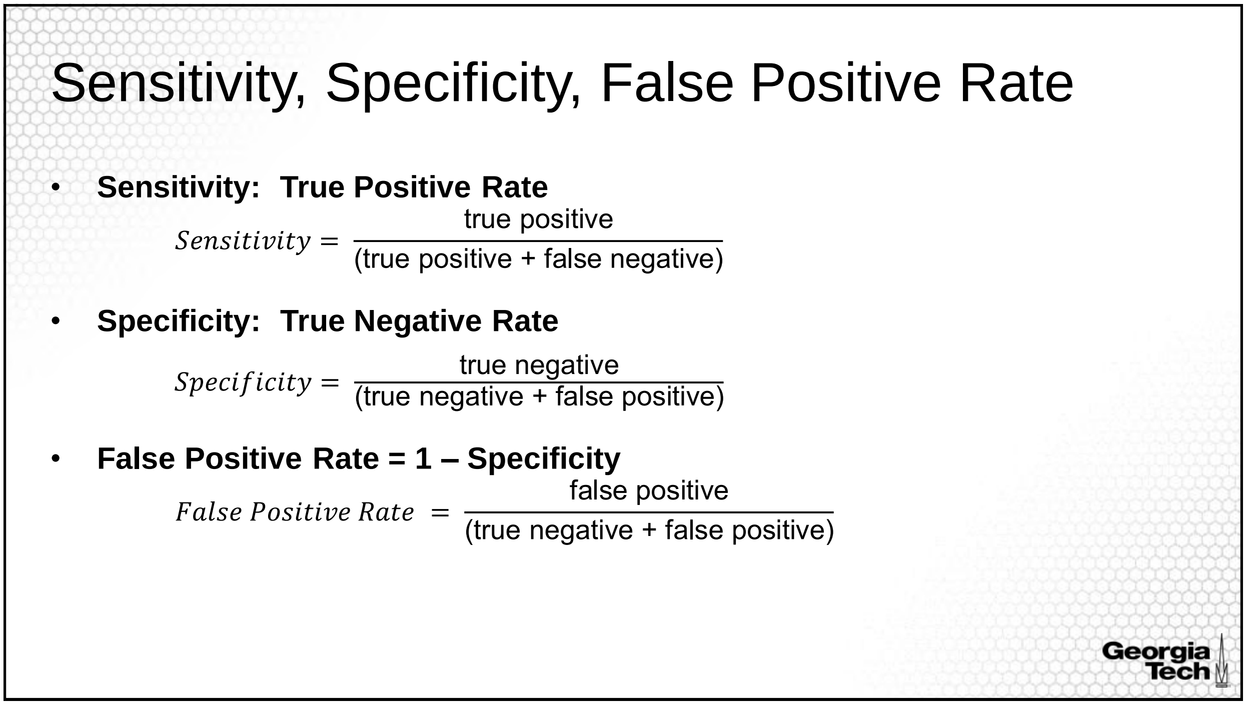


Sridhar Narasimhan, Ph.D

Professor

Scheller College of Business

Sensitivity, Specificity, and the ROC Curve



Sensitivity, Specificity, False Positive Rate

- **Sensitivity: True Positive Rate**

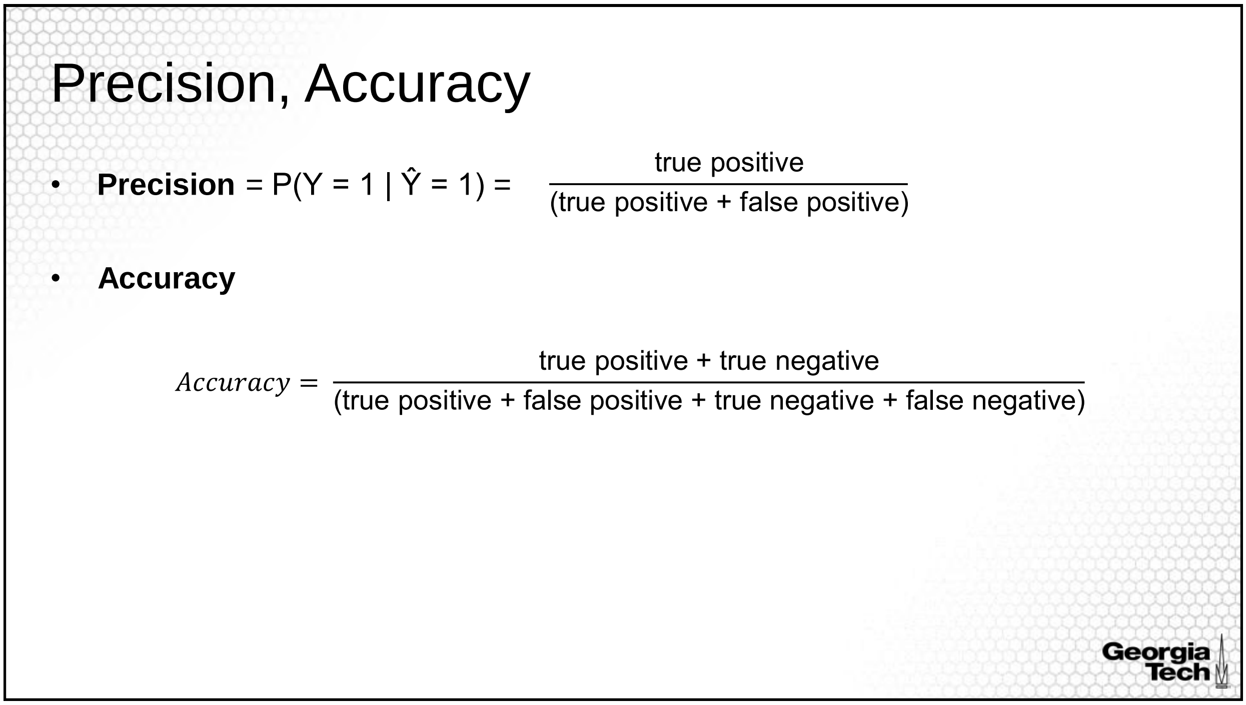
$$Sensitivity = \frac{\text{true positive}}{\text{(true positive + false negative)}}$$

- **Specificity: True Negative Rate**

$$Specificity = \frac{\text{true negative}}{\text{(true negative + false positive)}}$$

- **False Positive Rate = 1 – Specificity**

$$False\ Positive\ Rate\ = \frac{\text{false positive}}{\text{(true negative + false positive)}}$$

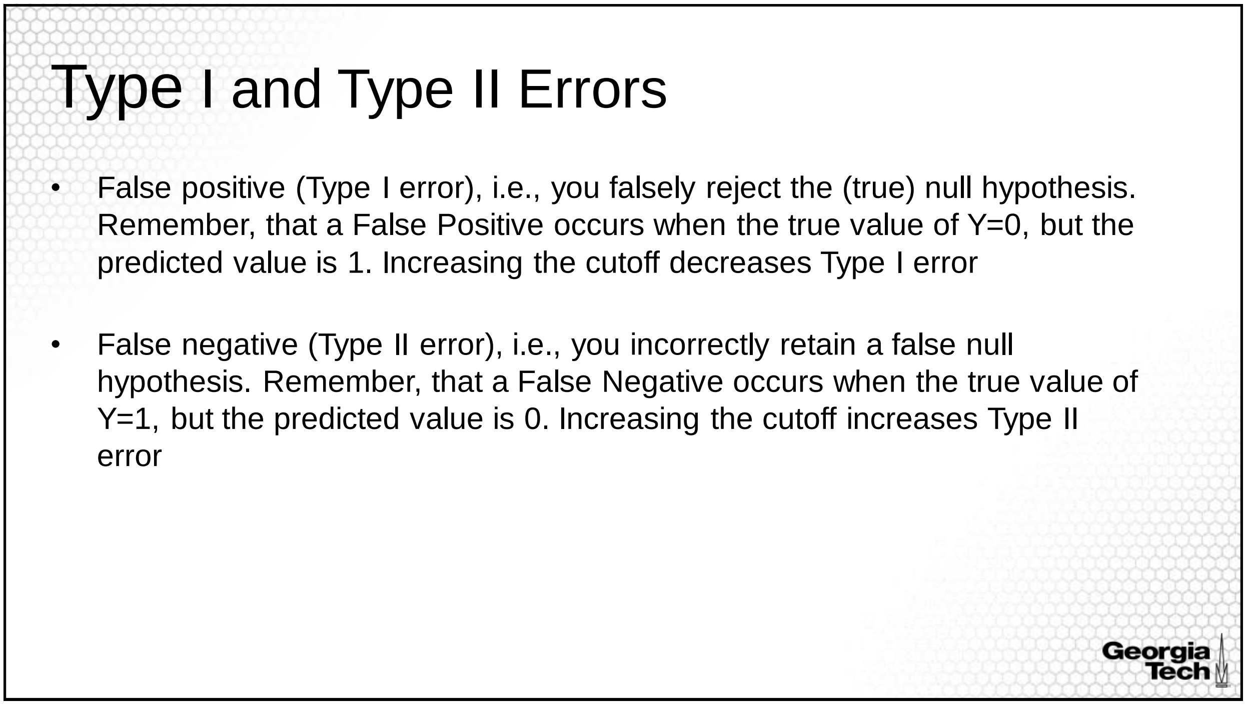


Precision, Accuracy

- **Precision** = P(Y = 1 | Ŷ = 1) = $\frac{\text{true positive}}{\text{(true positive + false positive)}}$

- **Accuracy**

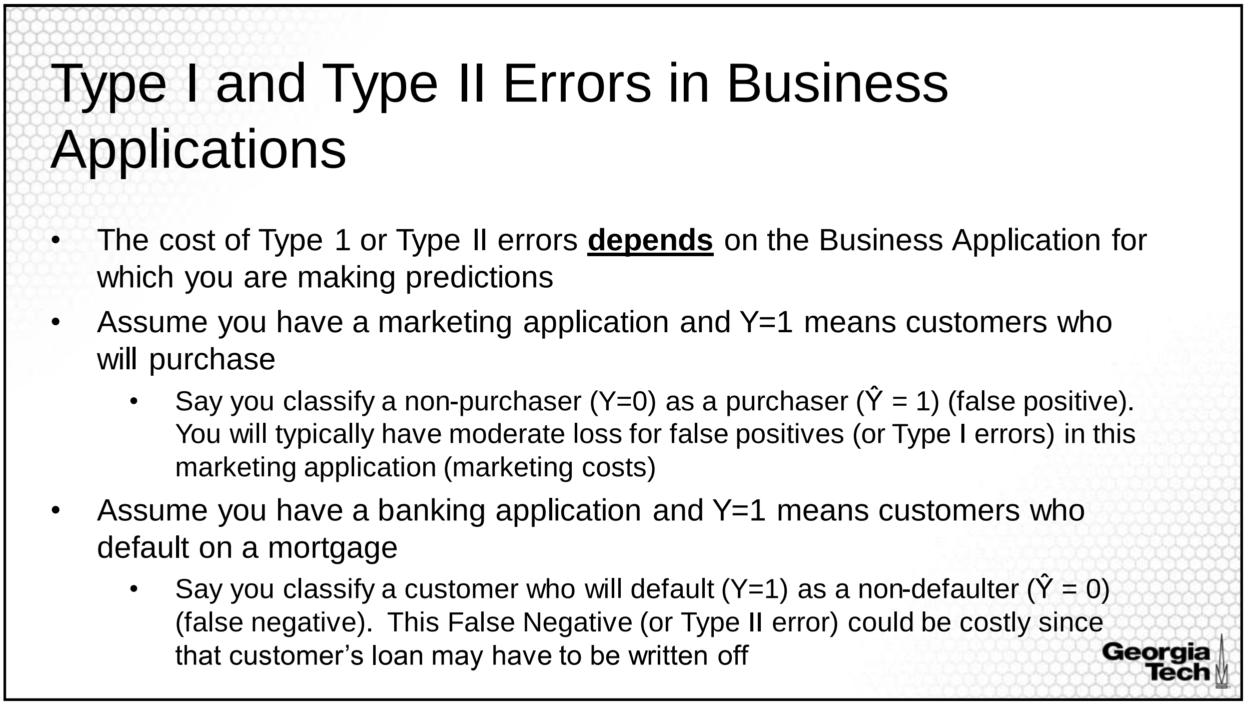
$$Accuracy = \frac{\text{true positive + true negative}}{\text{(true positive + false positive + true negative + false negative)}}$$



Type I and Type II Errors

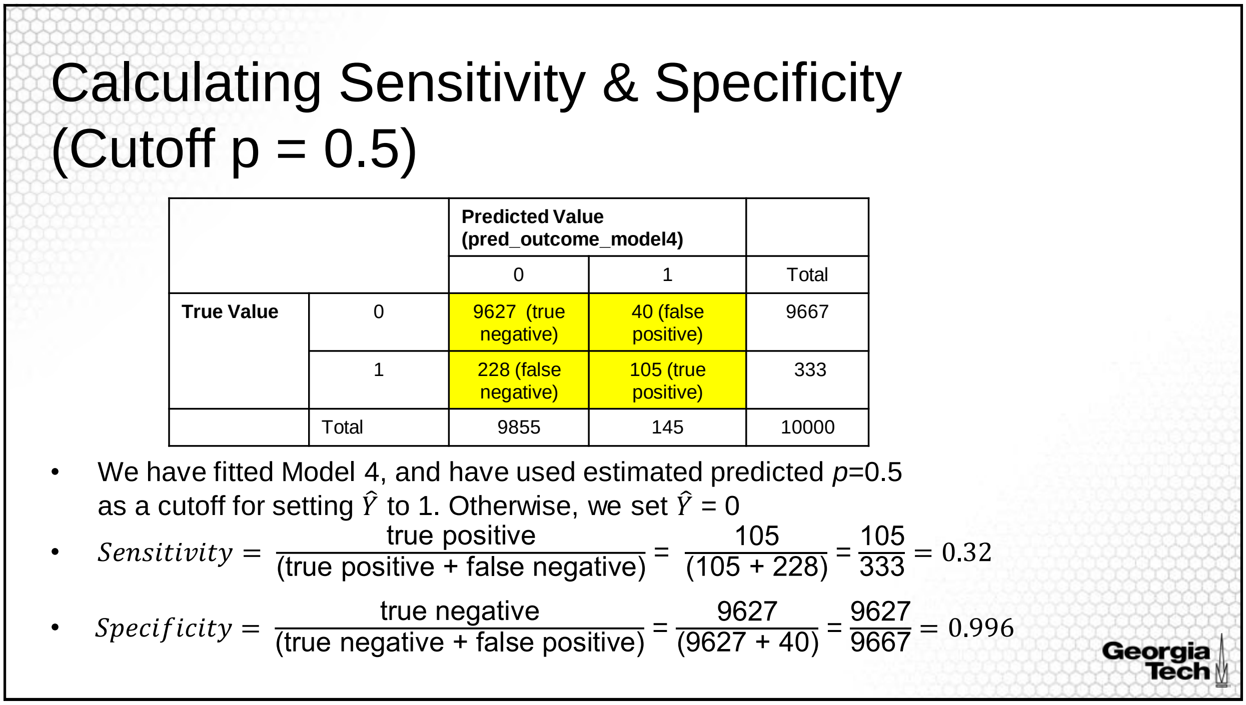
- False positive (Type I error), i.e., you falsely reject the (true) null hypothesis. Remember, that a False Positive occurs when the true value of Y=0, but the predicted value is 1. Increasing the cutoff decreases Type I error

- False negative (Type II error), i.e., you incorrectly retain a false null hypothesis. Remember, that a False Negative occurs when the true value of Y=1, but the predicted value is 0. Increasing the cutoff increases Type II error



Type I and Type II Errors in Business Applications

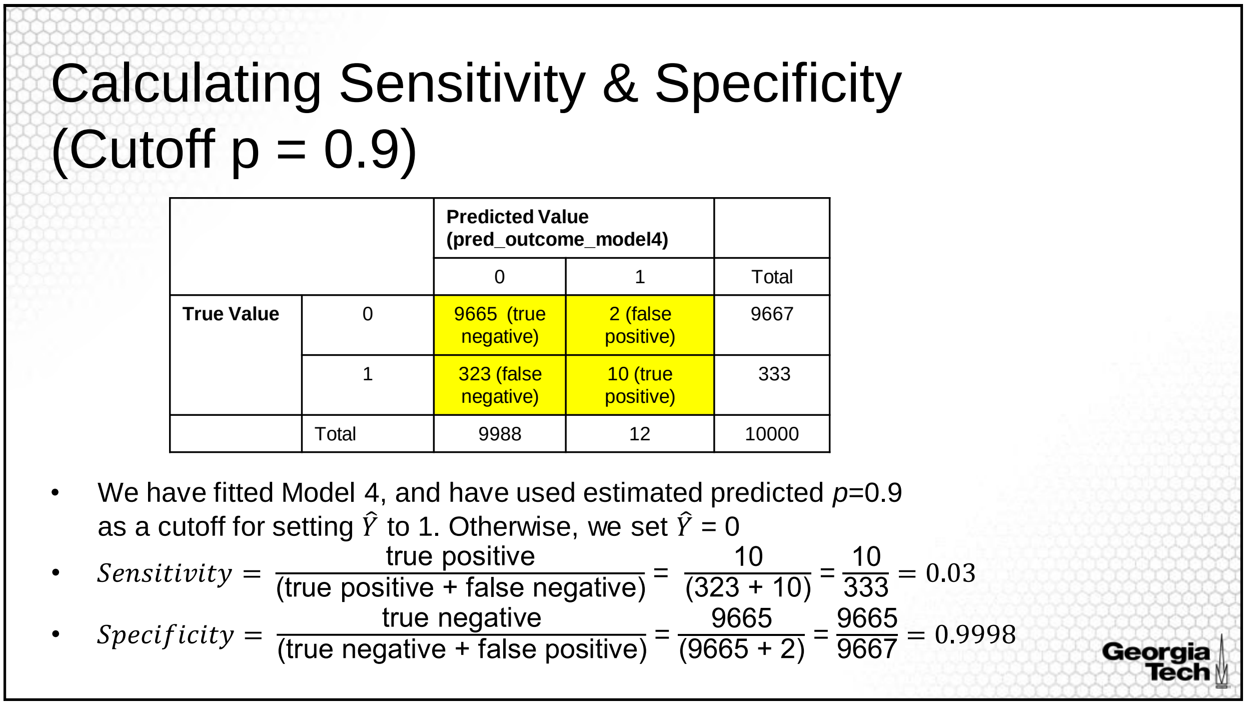
- The cost of Type 1 or Type II errors **depends** on the Business Application for which you are making predictions
- Assume you have a marketing application and Y=1 means customers who will purchase
  - Say you classify a non-purchaser (Y=0) as a purchaser ($\hat{Y}$ = 1) (false positive). You will typically have moderate loss for false positives (or Type I errors) in this marketing application (marketing costs)
- Assume you have a banking application and Y=1 means customers who default on a mortgage
  - Say you classify a customer who will default (Y=1) as a non-defaulter ($\hat{Y}$ = 0) (false negative). This False Negative (or Type II error) could be costly since that customer's loan may have to be written off



Calculating Sensitivity & Specificity (Cutoff p = 0.5)

		Predicted Value (pred_outcome_model4)		
		0	1	Total
True Value	0	9627 (true negative)	40 (false positive)	9667
	1	228 (false negative)	105 (true positive)	333
	Total	9855	145	10000

- We have fitted Model 4, and have used estimated predicted *p*=0.5 as a cutoff for setting $\hat{Y}$ to 1. Otherwise, we set $\hat{Y}$ = 0
- $Sensitivity = \frac{\text{true positive}}{\text{(true positive + false negative)}} = \frac{105}{(105 + 228)} = \frac{105}{333} = 0.32$
- $Specificity = \frac{\text{true negative}}{\text{(true negative + false positive)}} = \frac{9627}{(9627 + 40)} = \frac{9627}{9667} = 0.996$



Calculating Sensitivity & Specificity (Cutoff p = 0.9)

		Predicted Value (pred_outcome_model4)		
		0	1	Total
True Value	0	9665 (true negative)	2 (false positive)	9667
	1	323 (false negative)	10 (true positive)	333
	Total	9988	12	10000

- We have fitted Model 4, and have used estimated predicted *p*=0.9 as a cutoff for setting $\hat{Y}$ to 1. Otherwise, we set $\hat{Y}$ = 0
- $Sensitivity = \frac{\text{true positive}}{\text{(true positive + false negative)}} = \frac{10}{(323 + 10)} = \frac{10}{333} = 0.03$
- $Specificity = \frac{\text{true negative}}{\text{(true negative + false positive)}} = \frac{9665}{(9665 + 2)} = \frac{9665}{9667} = 0.9998$

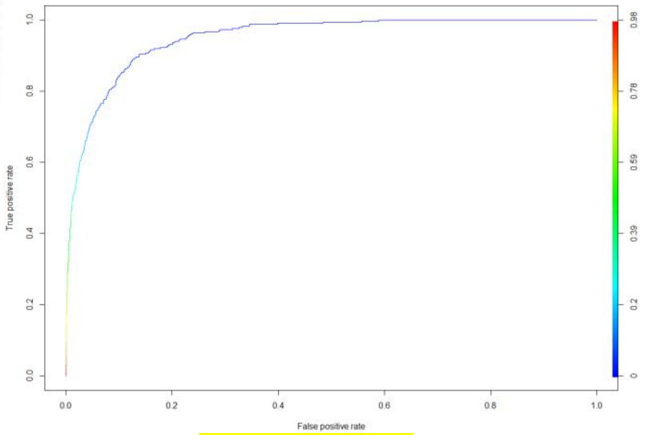
Increasing the Cutoff Value of *p*

- If we increase the cutoff, true negatives will increase and false positives will decrease
- If we increase the cutoff, false negatives will increase and true positives will decrease
- $Sensitivity = \dfrac{\text{true positive}}{\text{(true positive + false negative)}}$
- $Specificity = \dfrac{\text{true negative}}{\text{(true negative + false positive)}}$
- Therefore, as we increased the cutoff for *p* from 0.5 to 0.9
  - Sensitivity decreased from 0.32 to 0.03
  - Specificity increased from 0.996 to 0.9998

Logit Prediction and ROC Curve in R

```
if (!require(ggExtra))install.packages("ggExtra")
library("ggExtra")
if (!require(ROCR)) install.packages("ROCR")
library("ROCR")

#ROC Curve
pred <- prediction(df$pred_prob_model4,df$dft)  # create a prediction object in R
class(pred)
perf <- performance(pred, "tpr", "fpr")  # tpr and fpr are true and false positive rates
plot(perf, colorize=T)

# calculate Area Under the Curve for this Logit Model
auc.perf <-  performance(pred, measure = "auc")
auc.perf@y.values
```

Receiver Operating Characteristic (ROC Curve) for Model 4



TPR = Sensitivity

FPR = 1 - Specificity

- The ROC curve is used to show the diagnostic ability of a binary classifier as the cutoff value is varied
- Area under the curve = 0.95. We want this value to be greater than 0.5

Quiz

		Predicted Value (pred_outcome_model4)		
		0	1	Total
True Value	0	9627 (true negative)	40 (false positive)	9667
	1	228 (false negative)	105 (true positive)	333
	Total	9855	145	10000

True or False: Sensitivity is 105/333

Answer: TRUE, because sensitivity = $\dfrac{\text{true positive}}{\text{(true positive + false negative)}}$

Quiz

		Predicted Value (pred_outcome_model4)		
		0	1	Total
True Value	0	9627 (true negative)	40 (false positive)	9667
	1	228 (false negative)	105 (true positive)	333
	Total	9855	145	10000

True or False: Specificity is 40/9667.

Answer: FALSE, because specificity = $\frac{\text{true negative}}{\text{(true negative + false positive)}}$

Recap of the Module

A. Odds
B. Binary Dependent Variables
C. Logistic Regression
D. Logistic Regression Model using the *Default* Dataset (in the ISLR Library)
- No Predictor Variable
- Single 0/1 Predictor Variable
- Single Continuous Predictor Variable
- Multiple Predictor Variables

E. Predictions and Confusion Matrix
F. Sensitivity, Specificity, and the ROC Curve