Image removed due to copyright restrictions

Modeling the Expert

An Introduction to Logistic Regression

15.071 – The Analytics Edge

Ask the Experts!

- Critical decisions are often made by people with expert knowledge
- Healthcare Quality Assessment
  - Good quality care educates patients and controls costs
  - Need to assess quality for proper medical interventions
  - No single set of guidelines for defining quality of healthcare
  - Health professionals are experts in quality of care assessment

Experts are Human

- Experts are limited by memory and time
- Healthcare Quality Assessment
  - Expert physicians can evaluate quality by examining a patient's records
  - This process is time consuming and inefficient
  - Physicians cannot assess quality for millions of patients

Replicating Expert Assessment

- Can we develop analytical tools that replicate expert assessment on a large scale?
- Learn from expert human judgment
  - Develop a model, interpret results, and adjust the model
- Make predictions/evaluations on a large scale
- Healthcare Quality Assessment
  - Let's identify poor healthcare quality using analytics

Claims Data

Medical Claims

Diagnosis, Procedures,
Doctor/Hospital, Cost

Pharmacy Claims

Drug, Quantity, Doctor,
Medication Cost

- Electronically available
- Standardized
- Not 100% accurate
- Under-reporting is common
- Claims for hospital visits can be vague

Creating the Dataset – Claims Samples

Claims Sample

- Large health insurance claims database
- Randomly selected 131 diabetes patients
- Ages range from 35 to 55
- Costs $10,000 – $20,000
- September 1, 2003 – August 31, 2005

Creating the Dataset – Expert Review

Claims Sample

Expert Review

- Expert physician reviewed claims and wrote descriptive notes:

"Ongoing use of narcotics"

"Only on Avandia, not a good first choice drug"

"Had regular visits, mammogram, and immunizations"

"Was given home testing supplies"

Creating the Dataset – Expert Assessment

Claims Sample

Expert Review

Expert Assessment

- Rated quality on a two-point scale (poor/good)

"I'd say **care was poor** – poorly treated diabetes"

"No eye care, but overall I'd say **high quality**"

Creating the Dataset – Variable Extraction

- Claims Sample
- Expert Review
- Expert Assessment
- Variable Extraction

- Dependent Variable
  - **Quality of care**
- Independent Variables
  - ongoing use of **narcotics**
  - **only on Avandia**, not a good first choice drug
  - Had **regular visits, mammogram, and immunizations**
  - Was given **home testing supplies**

Creating the Dataset – Variable Extraction

- Claims Sample
- Expert Review
- Expert Assessment
- Variable Extraction

- Dependent Variable
  - **Quality of care**
- Independent Variables
  - Diabetes treatment
  - Patient demographics
  - Healthcare utilization
  - Providers
  - Claims
  - Prescriptions

Predicting Quality of Care

- The dependent variable is modeled as a binary variable
  - 1 if low-quality care, 0 if high-quality care
- This is a *categorical variable*
  - A small number of possible outcomes
- Linear regression would predict a continuous outcome
- How can we extend the idea of linear regression to situations where the outcome variable is categorical?
  - Only want to predict 1 or 0
  - Could round outcome to 0 or 1
  - But we can do better with logistic regression

Logistic Regression

- Predicts the probability of poor care
  - Denote dependent variable "PoorCare" by y
  - $P(y = 1)$
- Then $P(y = 0) = 1 - P(y = 1)$
- Independent variables $x_1, x_2, \ldots, x_k$
- Uses the Logistic Response Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}$$

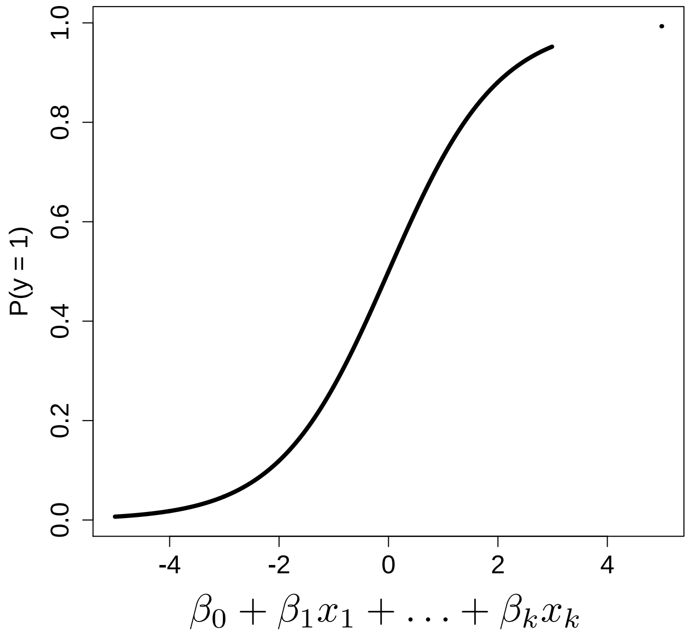
  - Nonlinear transformation of linear regression equation to produce number between 0 and 1

Poor Care = 1
Good Care = 0

Understanding the Logistic Function

$$P(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)}}$$

- Positive values are predictive of class 1
- Negative values are predictive of class 0



Understanding the Logistic Function

$$P(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)}}$$

- The coefficients are selected to
  - Predict a high probability for the poor care cases
  - Predict a low probability for the good care cases

Understanding the Logistic Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}$$

- We can instead talk about Odds (like in gambling)

$$\text{Odds} = \frac{P(y = 1)}{P(y = 0)}$$

- Odds > 1 if y = 1 is more likely
- Odds < 1 if y = 0 is more likely

The Logit

- It turns out that

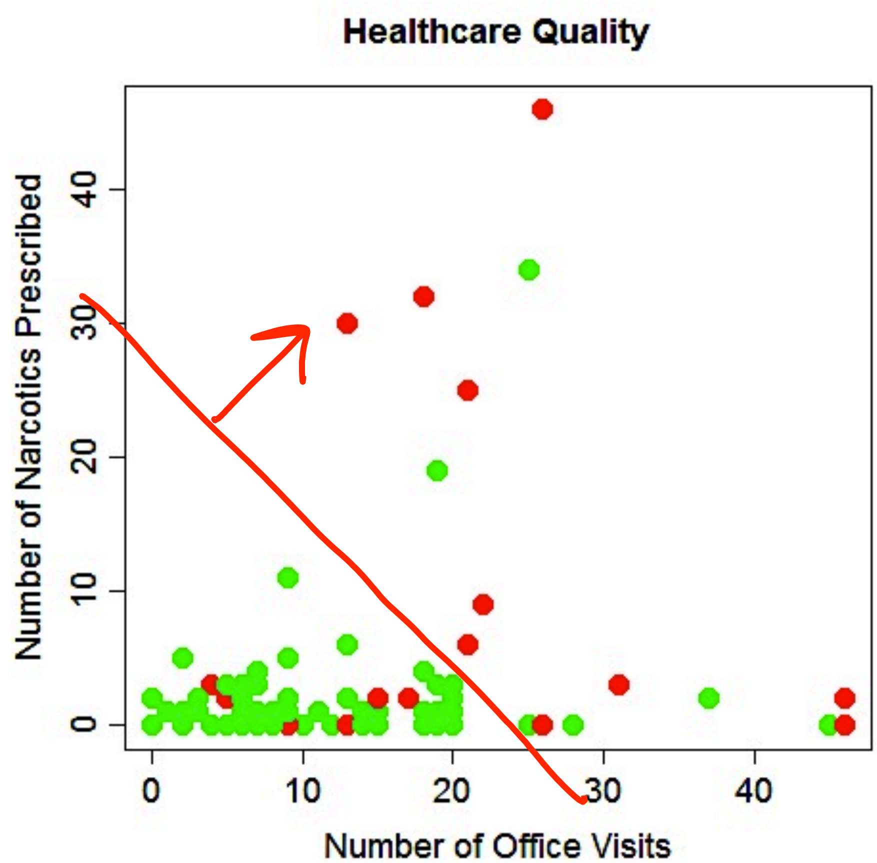
$$\text{Odds} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}$$

$$log(\text{Odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- This is called the "Logit" and looks like linear regression
- The bigger the Logit is, the bigger $P(y = 1)$

Model for Healthcare Quality

- Plot of the independent variables
  - Number of Office Visits
  - Number of Narcotics Prescribed
- Red are poor care
- Green are good care



Threshold Value

- The outcome of a logistic regression model is a probability
- Often, we want to make a binary prediction
  - Did this patient receive poor care or good care?
- We can do this using a *threshold value* t
- If $P(\text{PoorCare} = 1) \geq t$, predict poor quality
- If $P(\text{PoorCare} = 1) < t$, predict good quality
- What value should we pick for t?

Threshold Value

- Often selected based on which errors are "better"
- If t is **large**, predict poor care rarely (when P(y=1) is large)
  - More errors where we say good care, but it is actually poor care
  - Detects patients who are receiving the worst care
- If t is **small**, predict good care rarely (when P(y=1) is small)
  - More errors where we say poor care, but it is actually good care
  - Detects all patients who might be receiving poor care
- With no preference between the errors, select t = 0.5
  - Predicts the more likely outcome

Selecting a Threshold Value

Compare actual outcomes to predicted outcomes using a *confusion matrix (classification matrix)*

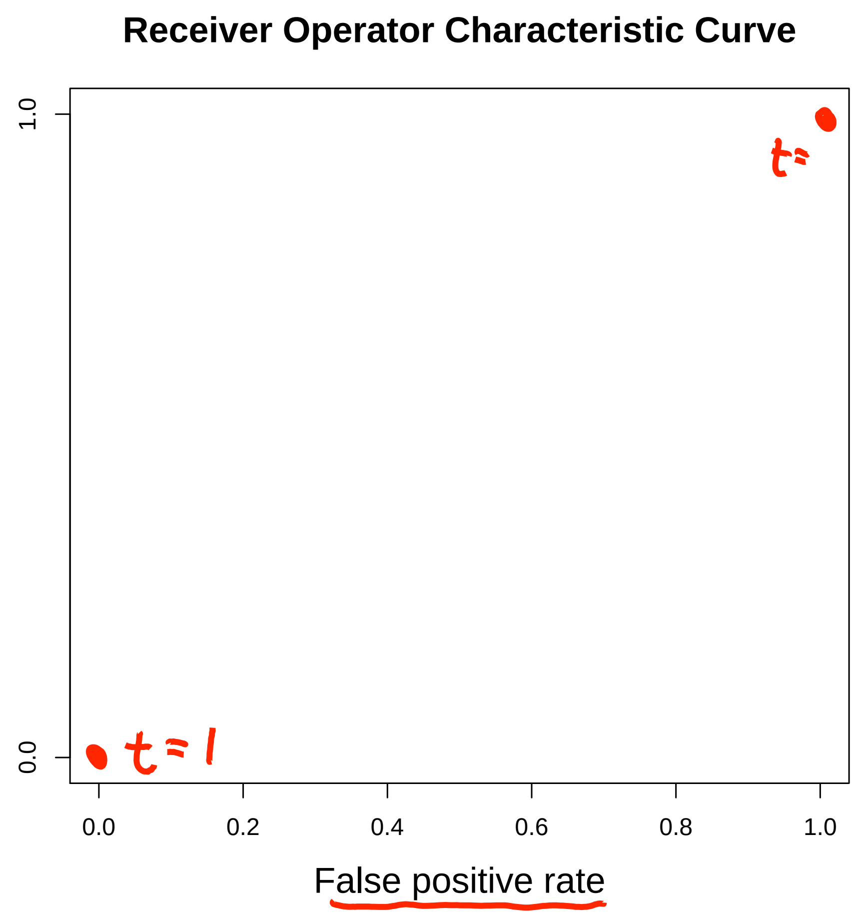
	Predicted = 0	**Predicted = 1**
Actual = 0	True Negatives (TN)	False Positives (FP)
Actual = 1	False Negatives (FN)	True Positives (TP)

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

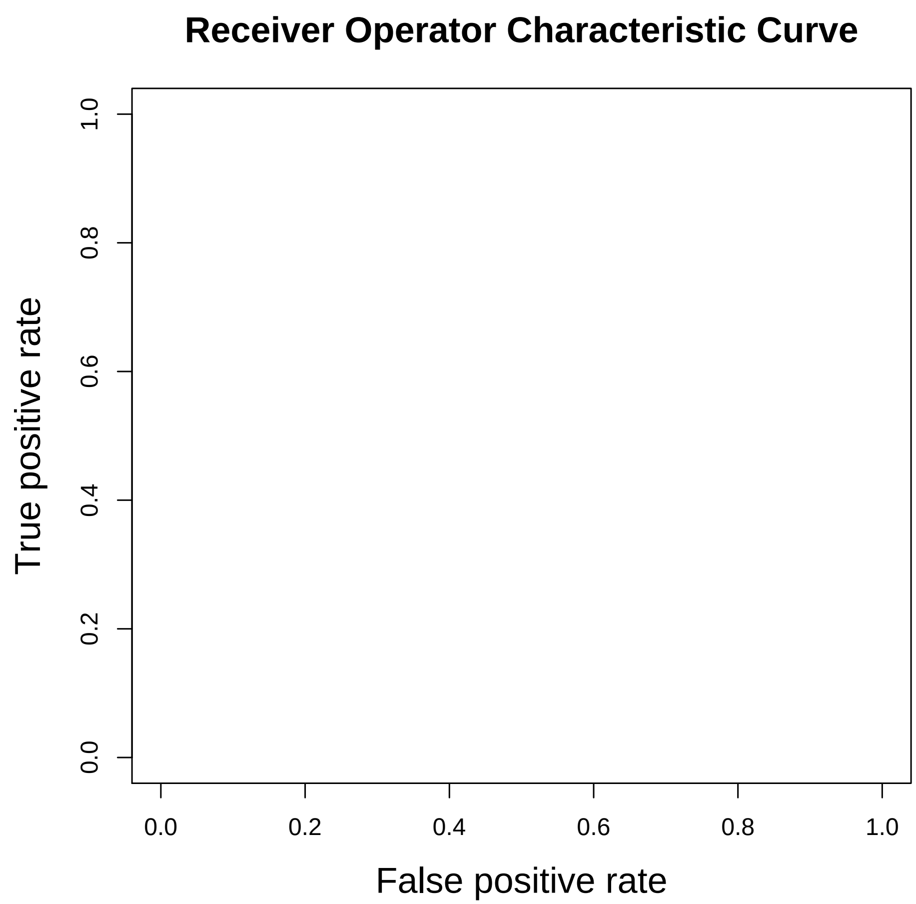
Receiver Operator Characteristic (ROC) Curve

- True positive rate (sensitivity) on y-axis
  - Propor care caught
- False posit (1-specific
  - Propor care labeled as poor care



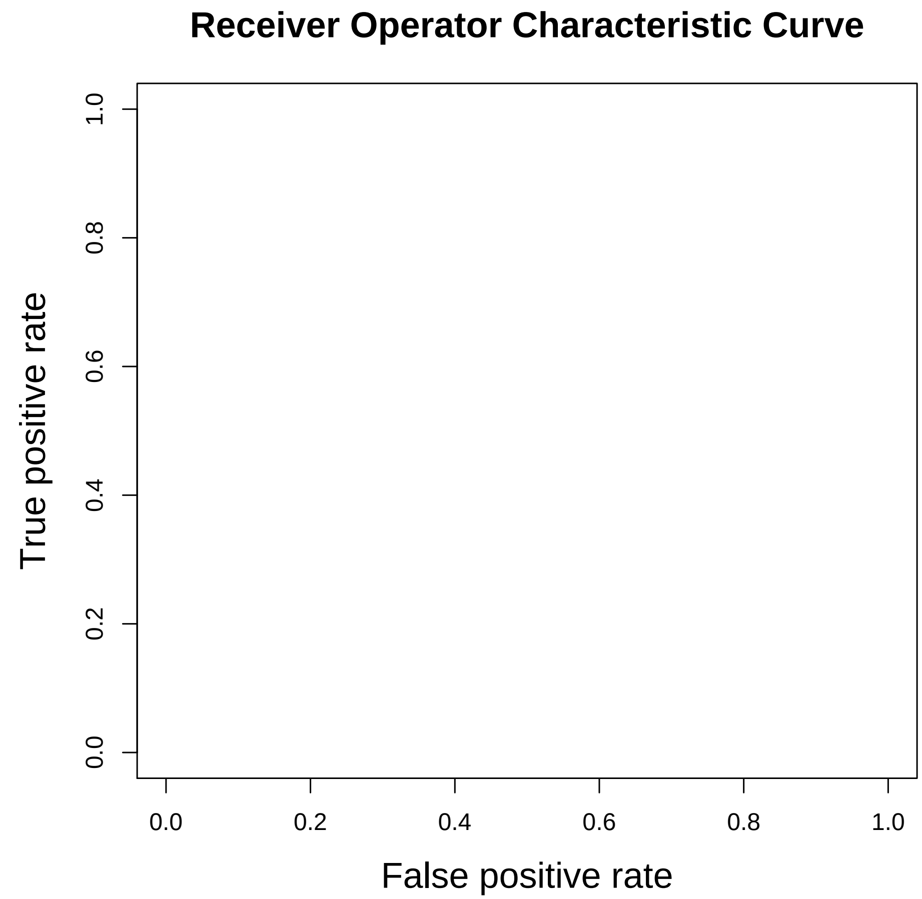
Selecting a Threshold using ROC

- Captures all thresholds simultaneously
- **High threshold**
  - High specificity
  - Low sensitivity
- **Low Threshold**
  - Low specificity
  - High sensitivity



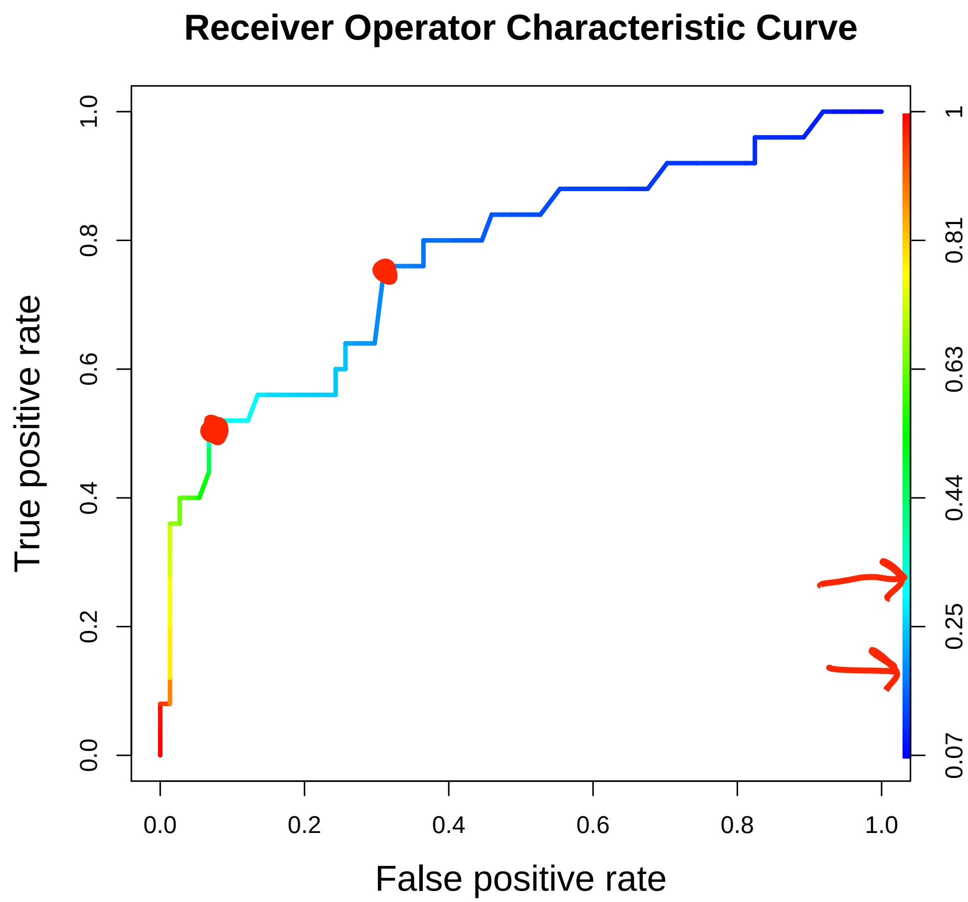
Selecting a Threshold using ROC

- Choose **best threshold** for **best trade off**
  - **cost of failing to detect positives**
  - **costs of raising false alarms**



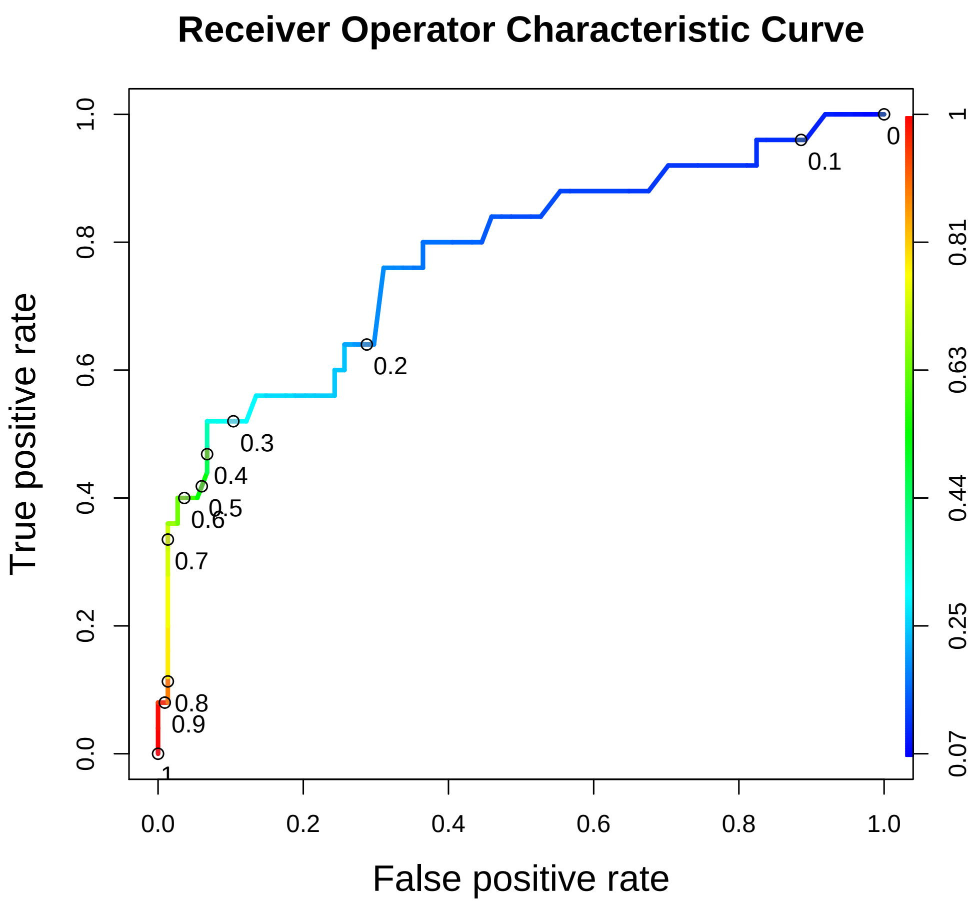
Selecting a Threshold using ROC

- Choose **best threshold** for **best trade off**
  - **cost of failing to detect positives**
  - **costs of raising false alarms**



Selecting a Threshold using ROC

- Choose **best threshold** for **best trade off**
  - **cost of failing to detect positives**
  - **costs of raising false alarms**



Interpreting the Model

- Multicollinearity could be a problem
  - Do the coefficients make sense?
  - Check correlations

- Measures of accuracy

Compute Outcome Measures

Confusion Matrix:

	Predicted Class = 0	Predicted Class = 1
Actual Class = 0	True Negatives (TN)	False Positives (FP)
Actual Class = 1	False Negatives (FN)	True Positives (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)

Making Predictions

- Just like in linear regression, we want to make predictions on a test set to compute out-of-sample metrics

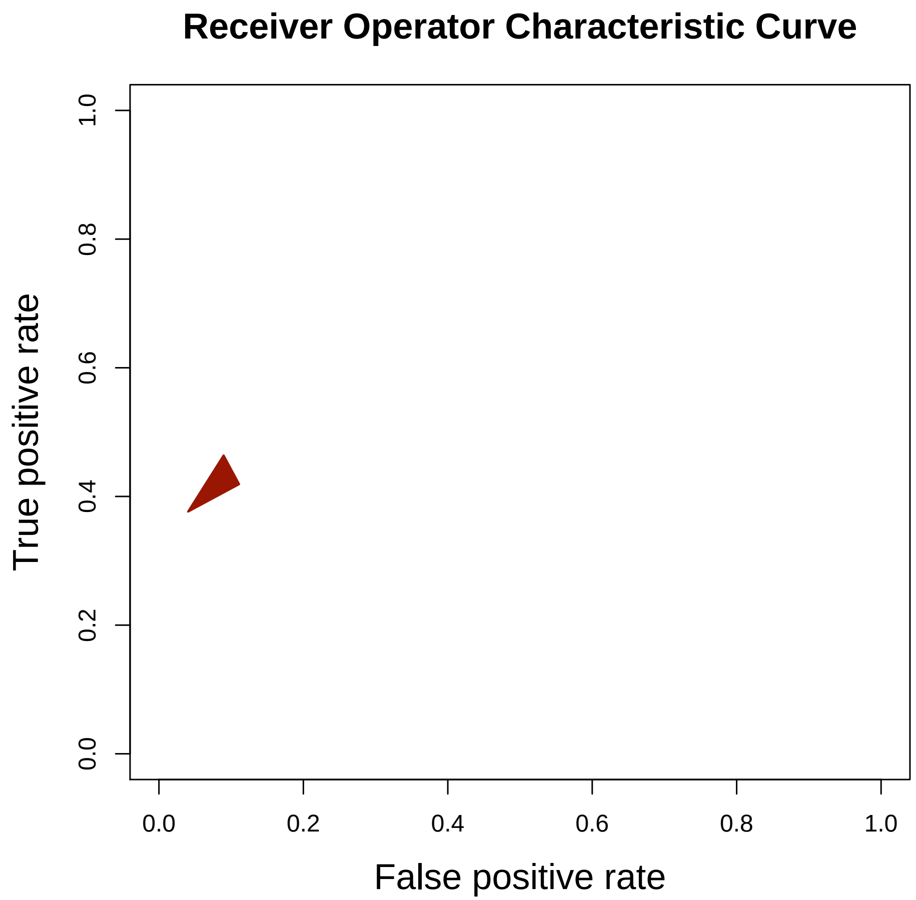
```
> predictTest = predict(QualityLog,
type="response", newdata=qualityTest)
```

- This makes predictions for probabilities
- If we use a threshold value of 0.3, we get the following confusion matrix

	Predicted Good Care	Predicted Poor Care
Actually Good Care	19	5
Actually Poor Care	2	6

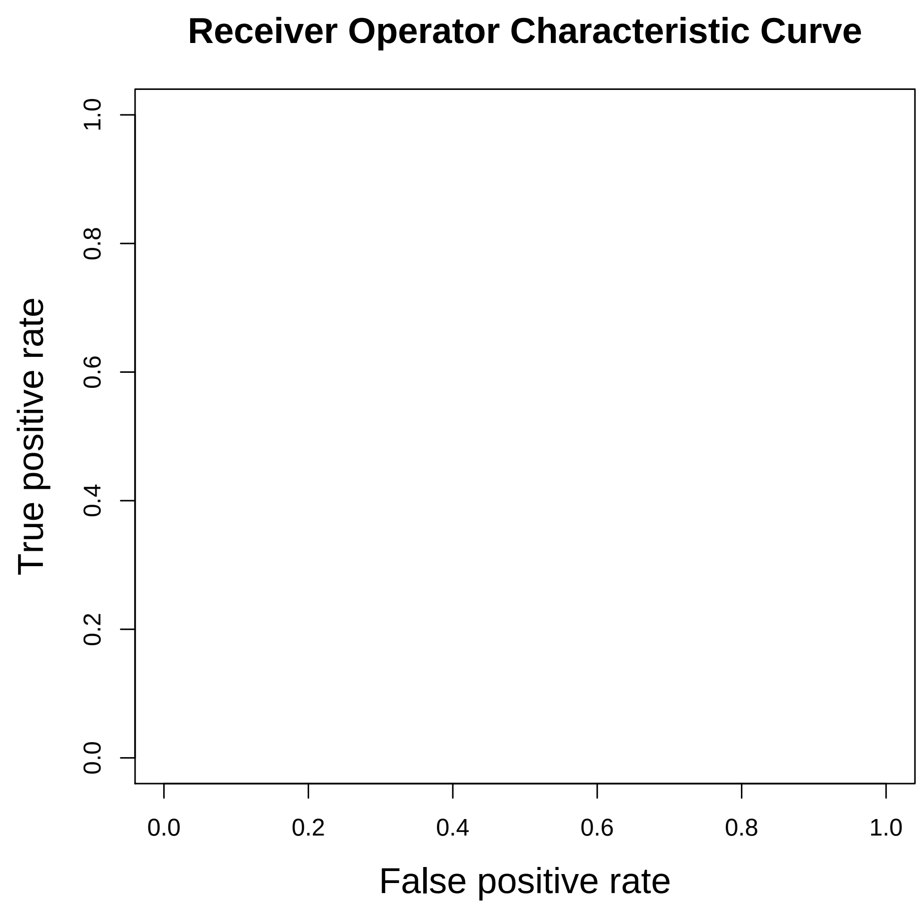
Area Under the ROC Curve (AUC)

- Just take the area under the curve
- Interpretation
  - Given a random positive and negative, proportion of the time you guess which is which correctly
- Less affected by sample balance than accuracy



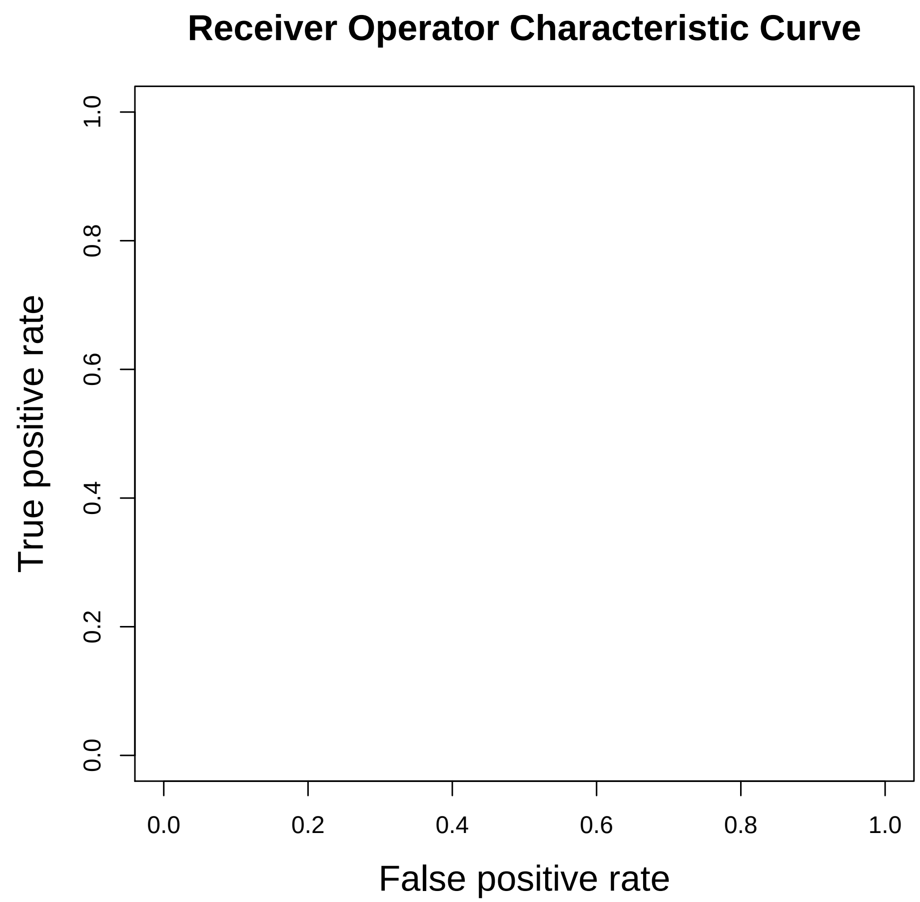
Area Under the ROC Curve (AUC)

- What is a good AUC?
  - Maximum of 1 (perfect prediction)



Area Under the ROC Curve (AUC)

- What is a good AUC?
  - Maximum of 1 (perfect prediction)
  - Minimum of 0.5 (just guessing)



Conclusions

- An expert-trained model can accurately identify diabetics receiving low-quality care
  - Out-of-sample accuracy of 78%
  - Identifies most patients receiving poor care
- In practice, the probabilities returned by the logistic regression model can be used to prioritize patients for intervention
- Electronic medical records could be used in the future

The Competitive Edge of Models

- While humans can accurately analyze small amounts of information, models allow larger scalability
- Models do not replace expert judgment
  - Experts can improve and refine the model
- Models can integrate assessments of many experts into one final unbiased and unemotional prediction