Logistic Regression

Module 7

DATA 5600
Introduction to Regression and Machine Learning for Analytics
Marc Dotson

Module Overview

- Visualization methods
- Logistic regression formulation
- Logistic regression interpretations
- Logistic regression model performance

CHD Data Set

What are the risk factors associated with CHD?

Variable	Description
chd	Developed coronary heart disease (CHD): yes or no
age	Age in years
height	Height in inches
weight	Weight in pounds
sbp	Systolic blood pressure in mmHg (millimeters of mercury)
dbp	Diastolic blood pressure in mmHg (millimeters of mercury)
chol	Cholesterol in mg/dL (milligrams of cholesterol per deciliter of blood)
cigs	Number of cigarettes smoked a day

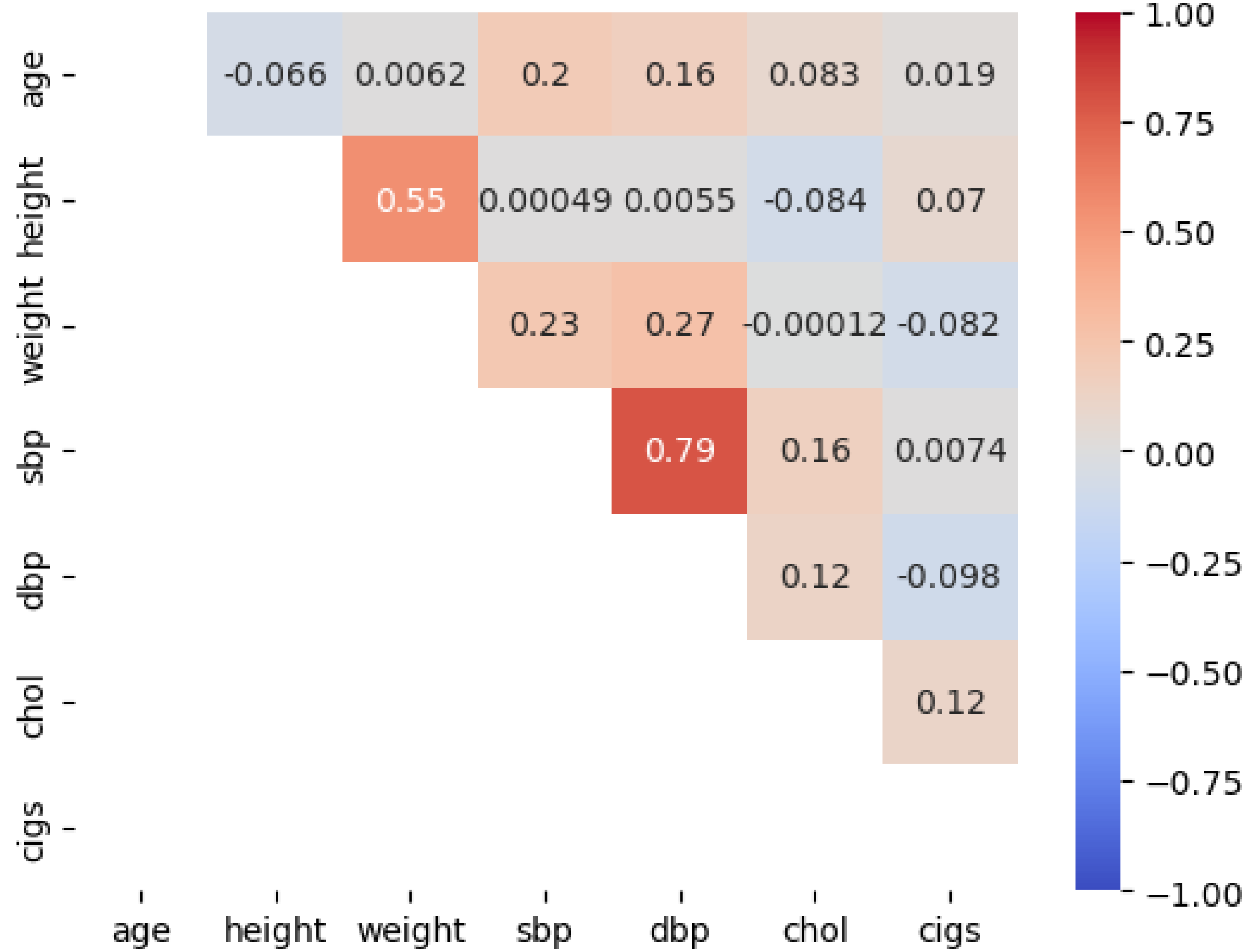
757 subjects (randomly selected) aged 39 to 59 years old and free of heart disease as determined by electrocardiogram at an initial screening. At baseline the variables in the following table were collected. Follow-up continued for 8.5 years with repeat examinations to determine if patients developed CHD. The goal is to determine risk factors (ways of healthy living) to avoid CHD.

What is the response variable? Is it continuous or categorical?

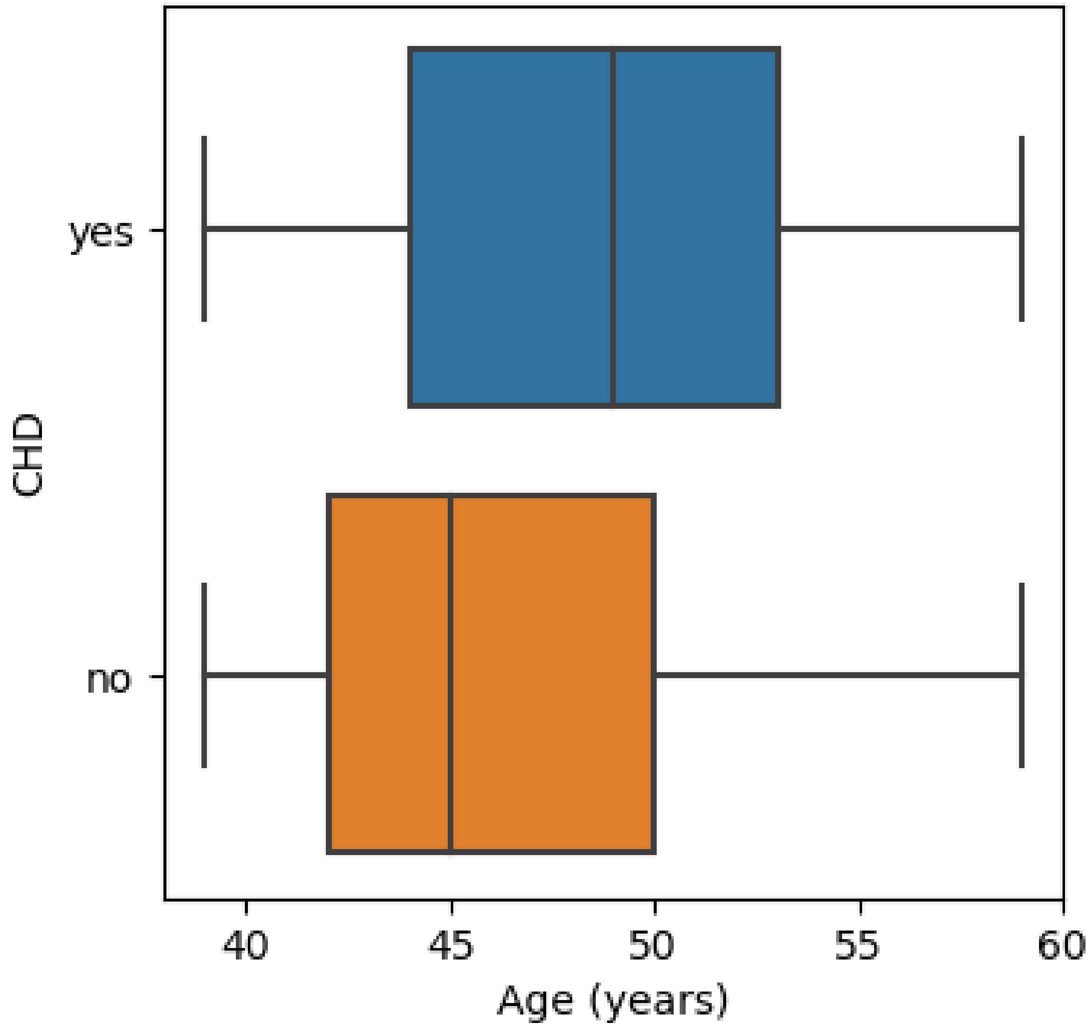
CHD Data Set: Baseline Prediction for the Response

- Of the 757 patients in the data set, 500 of them did not develop CHD, while 257 of them did develop CHD.
- If we had no other information than this, and if a patient was presented to you and you had to classify whether or not they would develop CHD, which would you choose?
- Based on the data, approximately how accurate would you be?
- This is our “baseline” accuracy rate that we hope to beat with the predictors we have in the data set and our model.

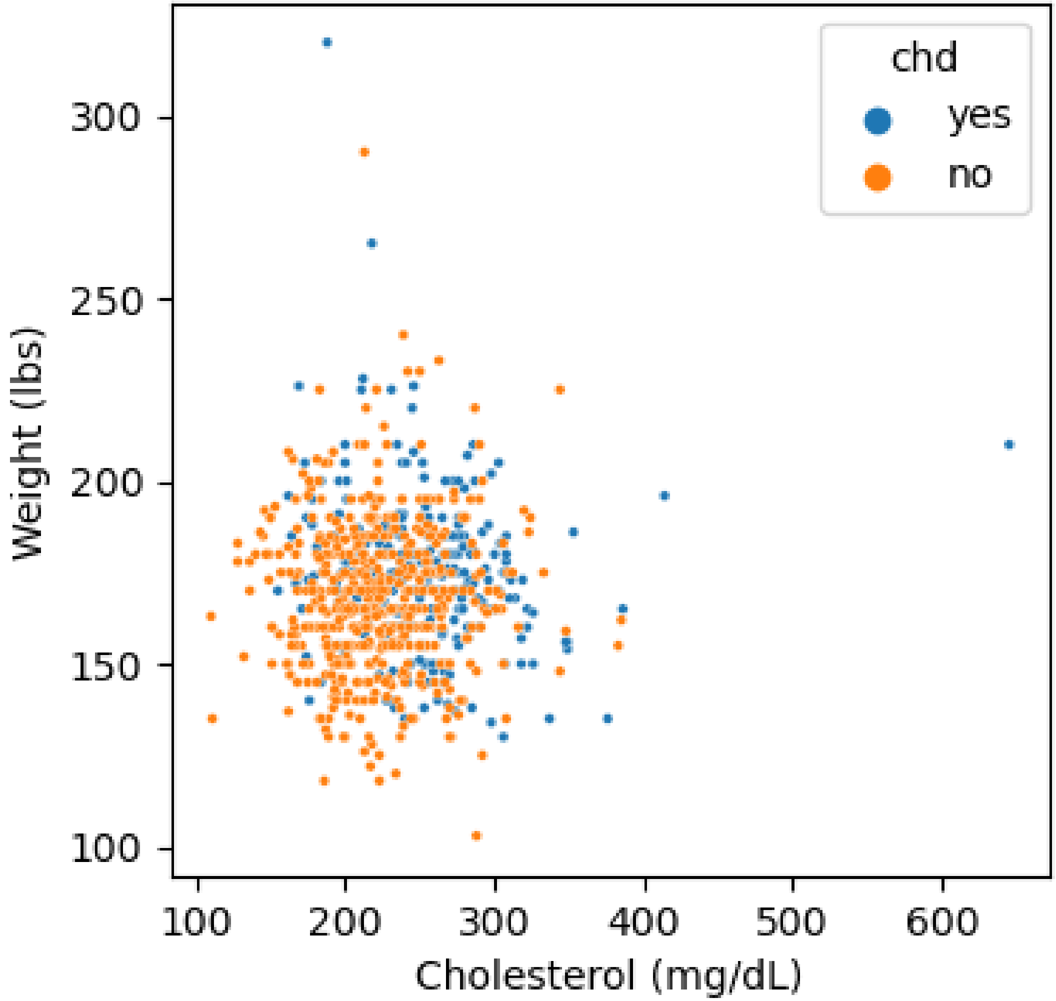
CHD Data Set: Correlation Matrix



CHD Data Set: Side-By-Side Boxplots



CHD Data Set: Color-Coded Scatterplot



CHD Data Set: Cross-Tabulation (Contingency Table)

- Only works well for discrete data
- Works really well for factors with few levels
  - Example:

Age	CHD No	CHD Yes	Sum
<= 50	386	156	542
> 50	114	101	215

Age	CHD No	CHD Yes	Sum
39	32	19	51
40	50	12	62
41	38	12	50
42	33	5	38
43	43	13	56
44	38	13	51
45	26	12	38
46	34	11	45
47	23	8	31
48	24	19	43
49	27	21	48
50	18	11	29
51	20	17	37
52	13	15	28
53	17	12	29
54	20	11	31
55	14	10	24
56	11	12	23
57	10	7	17
58	7	6	13
59	2	11	13

CHD Data Set: Test/Train

- Logistic Regression is often used for prediction, so it is important to create a model that will perform well on new/future data (i.e., one that **does not overfit**)

- We will randomly split our data into a training and testing set
  - Build the logistic regression model on the training set
  - Use that model to make predictions on the testing set
  - Report model performance on *testing* set

- We will do an 80/20 split
  - 80% of data in the training set, 20% of data in the testing set
  - This is a common split to use, but it certainly depends on sample size

Full data set:
757 rows

Training data set:
605 rows
(0.80*757=605.6)

Testing data set:
152 rows
(0.20*757=151.4)

CHD Data Set: Variable Selection

- Variable selection was performed on the training data set, and the following variables were chosen to be included in the model:
  - age
  - weight
  - sbp
  - chol
  - cigs
- Note that I would have gone with a different simpler (smaller) model in “real life,” but I want to keep it larger for illustration.

Code!

Logistic Regression Formulation

Can We Use Linear Regression?

- Our response is categorical, so can we just use indicator variables and set

$$y_i = \begin{cases} 1 & \text{if CHD} \\ 0 & \text{otherwise} \end{cases}$$

then use regular least squares multiple linear regression?

Can We Use Linear Regression?

- Our response is categorical, so can we just use indicator variables and set

$$y_i = \begin{cases} 1 & \text{if CHD} \\ 0 & \text{otherwise} \end{cases}$$

then use regular least squares multiple linear regression?

- NO! Because
  - Predictions will be outside the range of {0, 1}
  - Linear assumption might be violated
  - Errors certainly will not be normally distributed
  - Equal variance (homoscedasticity) is also likely to be violated

- We need an entirely new regression framework!

Can We Use Linear Regression?

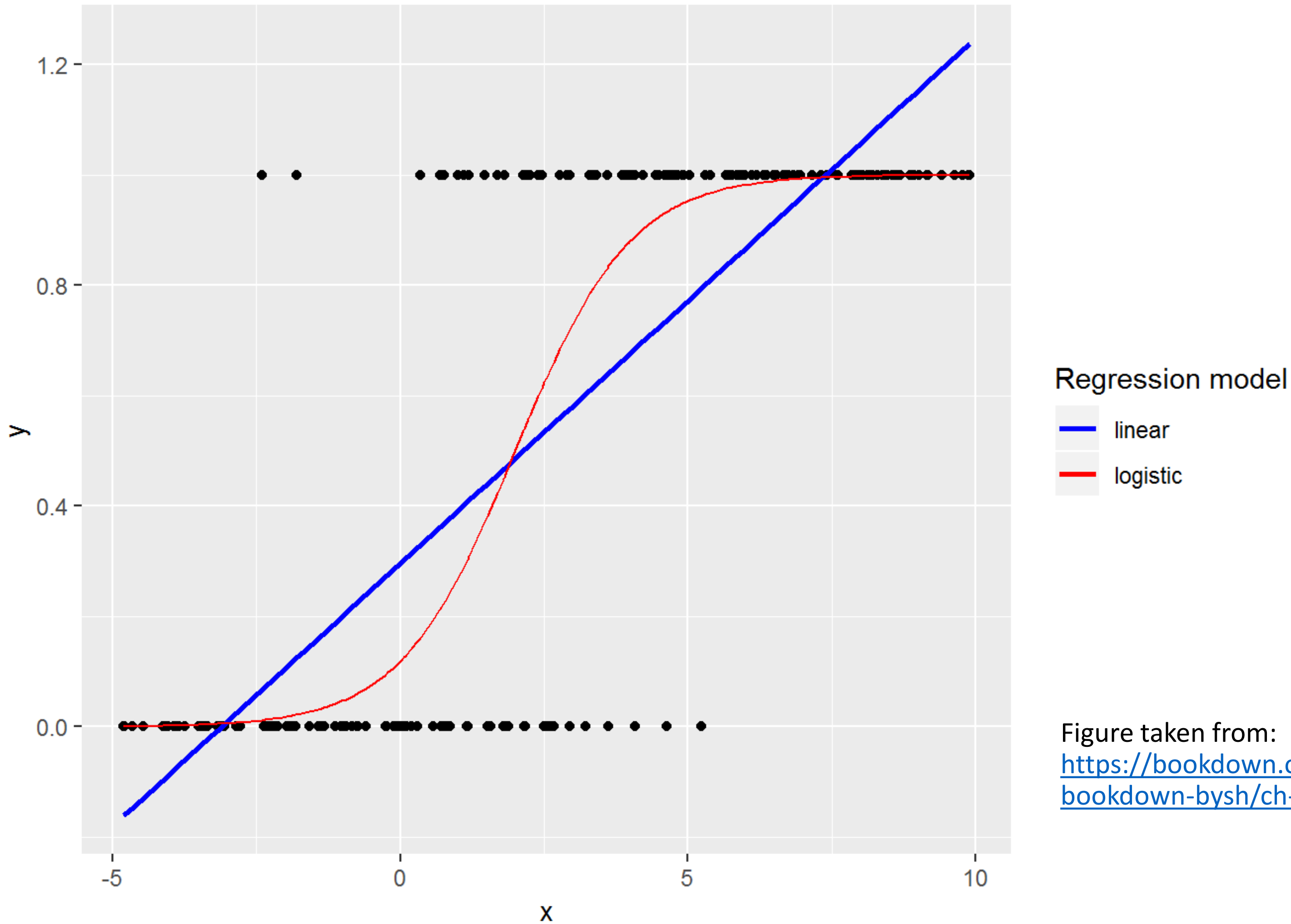


Figure taken from:
https://bookdown.org/roback/bookdown-bysh/ch-logreg.html

Logistic Regression

- What is an appropriate distribution for when $Y_i \in \{0,1\}$? (Reason why normality and homoscedasticity assumptions are violated)

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

$$P(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$P(Y_i = 1) = \quad \pi_i$$

$$P(Y_i = 0) = 1 - \pi_i$$

$$E(Y_i) = \pi_i$$

$$\text{Var}(Y_i) = \pi_i(1-\pi_i)$$

Logistic Regression

- If our response follows a Bernoulli distribution then,

$$E(Y_i) = \pi_i = \text{Prob}(Y_i = 1)$$

- So, can we just set,

$$\text{Prob}(Y_i = 1) = \pi_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ik} \ ?$$

Logistic Regression

- A "Generalized Linear Model" for a binary response using the most common link function: the logit function

Odds

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ik}$$

Logit Function

$$\Rightarrow \pi_i = \frac{\exp\{\beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ik}\}}{1 + \exp\{\beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ik}\}} \in (0,1)$$

Logistic Function

where $\pi_i = \text{Prob}(Y_i = 1 | x_{i1}, \dots, x_{ip-1})$

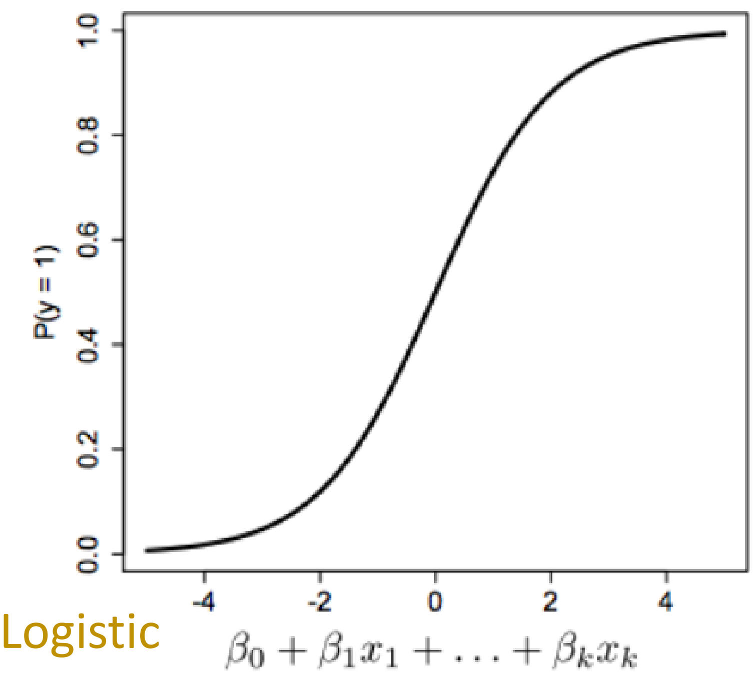


Figure taken from: https://medium.com/analytics-vidhya/a-guide-to-machine-learning-in-r-for-beginners-part-5-4c00f2366b90

Logistic Regression: CHD Data Set

- Theoretical/general logistic regression model (not the "fitted" model):

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{weight}_i + \beta_3 \text{sbp}_i + \beta_4 \text{chol}_i + \beta_5 \text{cigs}_i$$

where $\pi_i = \text{Prob}(\text{chd}_i = 1 | \text{age}_i, \text{weight}_i, \text{sbp}_i, \text{chol}_i, \text{cigs}_i)$

Logistic Regression: Estimation of the Coefficients

Formulation

$$y_i \overset{\text{ind}}{\sim} \text{Bern}(\pi_i)$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ik}$$

- How do we estimate the β_ks?

There are no residuals in the traditional sense anymore, so we can't use OLS which minimizes the distance of the line to the points

- We use maximum likelihood instead of ordinary least squares (requires a larger sample size since ML doesn't have an analytical/closed-form solution like OLS)
- In this class, we will let Python do it for us

Code!

Logistic Regression Interpretations

Logistic Regression: Odds

Interpret

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ik} \quad \text{where } \pi_i = \text{Prob}\left(Y_i = 1 | x_{i1}, \dots, x_{ip-1}\right)$$

- $\pi_i = \text{Prob}\left(Y_i = 1 | x_{i1}, \dots, x_{ip-1}\right)$
- $1 - \pi_i = \text{Prob}\left(Y_i = 0 | x_{i1}, \dots, x_{ip-1}\right)$
- Odds (the probability of an event occurring divided by the probability of the event not occurring):

$$\text{Odds of } Y_i = 1 | x_{i1}, \dots, x_{ip-1} = \frac{\text{Prob}\left(Y_i = 1 \middle| x_{i1}, \dots, x_{ip-1}\right)}{\text{Prob}\left(Y_i = 0 \middle| x_{i1}, \dots, x_{ip-1}\right)} = \frac{\pi_i}{1-\pi_i}$$

Natural log – base e

- Log Odds that $Y_i = 1 | x_{i1}, \dots, x_{ip-1} = \log\left(\frac{\pi_i}{1-\pi_i}\right)$

Do not confuse probability with odds!!!

Probability: In a deck of 52 cards, there are 13 spades. Probability of drawing a spades is 13/52 = 0.25 = 25%.

Odds: Probability of drawing a spade is 0.25. The probability of not drawing a spade is 1 - 0.25=0.75. So, the odds is 0.25/0.75 or 1:3 (or 1/3 pronounced 1 to 3 odds). [13:39 = 1:3s]

Logistic Regression: Odds Ratio

- Odds Ratio: (the ratio of two odds) the odds of the event in one group (ex: $x_{\mathrm{i}1} = 1$) divided by the odds in another group (ex: $x_{\mathrm{i}1} = 0$).
- Odds Ratio for $x_{\mathrm{i}1} = 1$, holding all other variables constant:

Comparing x=1 to x=0 (think one unit increase…)

$$\mathrm{OR}_i = \frac{\mathrm{Prob}\left(Y_i = 1 \middle| x_{i1} = 1, \dots, x_{ip-1}\right) / \mathrm{Prob}\left(Y_i = 0 \middle| x_{i1} = 1, \dots, x_{ip-1}\right)}{\mathrm{Prob}\left(Y_i = 1 \middle| x_{i1} = 0, \dots, x_{ip-1}\right) / \mathrm{Prob}\left(Y_i = 0 \middle| x_{i1} = 0, \dots, x_{ip-1}\right)}$$

- Log Odds Ratio for $x_{\mathrm{i}1} = 1$, holding all other variables constant:

$$\log(\mathrm{OR}_i)$$

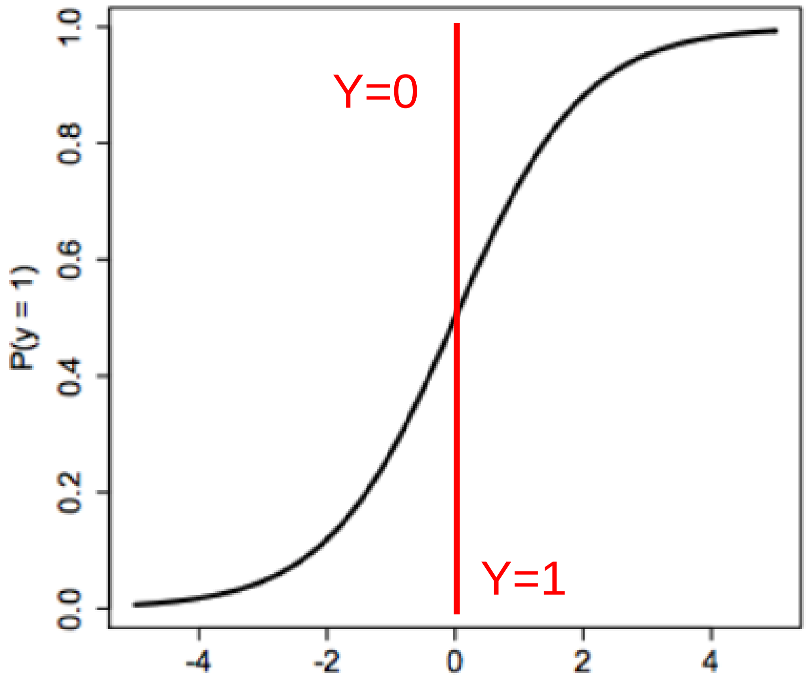
Logistic Regression: Coefficient Interpretations

Interpret

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ik} \quad \text{where } \pi_i = \text{Prob}(Y_i = 1 | x_{i1}, \dots, x_{ip-1})$$

- If $Y_i = 1$ is more likely, then $\log(OR_i) > 0$ and $\text{OR}_i > 1$

- If $Y_i = 0$ is more likely, then $\log(OR_i) < 0$ and $\text{OR}_i < 1$



Log(OR) $\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

0.02 0.1 1 7.4 54.6

OR $\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$

Logistic Regression: Coefficient Interpretations

$$y_i \overset{\text{ind}}{\sim} \text{Bern}(\pi_i)$$

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \sum_{k=1}^{p-1} \hat{\beta}_k x_{ik}$$

- How do we interpret $\exp\left(\hat{\beta}_k\right)$?
  - **Exponentiated coefficients are odds ratios.**
  - *Odds ratio* for X_k
    (odds of $Y_i = 1$ when $X_k + 1$ vs. odds of $Y_i = 1$ when X_k)

$$\frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k (x_{ik} + 1) + \cdots + \hat{\beta}_{p-1} x_{ip-1}\right)}{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} + \cdots + \hat{\beta}_{p-1} x_{ip-1}\right)} = \exp(\hat{\beta}_k)$$

Logistic Regression: Coefficient Interpretations

$$y_i \overset{\text{ind}}{\sim} \text{Bern}(\pi_i)$$

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}$$

- How do we interpret $\exp(\hat{\beta}_1)$?
  - As x_{i1} increases by one unit (ex: from 0 to 1), $\log\left(\widehat{\text{odds}}_i\right) = \hat{\beta}_0 + \hat{\beta}_1 \quad \Rightarrow \quad \widehat{\text{odds}}_i = \exp\left(\hat{\beta}_0\right)\exp\left(\hat{\beta}_1\right)$.
  - So, an increase of one unit in X_1 ***multiplies the odds*** (in favor of $Y_i = 1$) by a factor of $\exp\left(\hat{\beta}_1\right)$.
  - The odds ratio reflects the multiplicative change in odds (of Y_i) as a variable increases by 1 unit, holding all other variables constant.

Logistic Regression: Coefficient Interpretations

$$y_i \overset{\text{ind}}{\sim} \text{Bern}(\pi_i)$$

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \sum_{k=1}^{p-1} \hat{\beta}_k x_{ik}$$

- How do we interpret $\hat{\beta}_k$?

Based on $\log(OR)$

- Holding all else constant, for every one unit increase in X_k, the **log odds** (of $Y = 1$) increase by $\hat{\beta}_k$.
- Just interpret the sign: If $\hat{\beta}_k > 0$, then the **log odds** (of $Y = 1$) increase as X_k increases, holding all else constant. If $\hat{\beta}_k < 0$, then the **log odds** (of $Y = 1$) decrease as X_k increases, holding all else constant.

Based on OR

- Holding all else constant, as X_k increases by one, the **odds** (of $Y = 1$) is $\exp\{\hat{\beta}_k\}$ *times* more likely.
- Holding all else constant, as X_k increases by one, the **odds** (of $Y = 1$) increase by $100 \times (\exp\{\hat{\beta}_k\} - 1)\%$.

Logistic Regression: Coefficient Interpretations

$$y_i \overset{\text{ind}}{\sim} \text{Bern}(\pi_i)$$

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \sum_{k=1}^{p-1} \hat{\beta}_k x_{ik}$$

- $\hat{\beta}_{\text{age}} = 0.0596$. How do we interpret this number?

Based on $\log(OR)$

- Holding all else constant, for every additional year in age, the log odds of developing CHD increase by 0.0596.
- Since $0.0596 > 0$, then the log odds of developing CHD increase as age increases, holding all else constant.

Based on OR

- Holding all else constant, for every additional year in age, the odds a patient develops CHD is $\exp\{0.0596\} = 1.0614$ times more likely.
- Holding all else constant, for every additional year in age, the odds a patient develops CHD increase by $100 \times (\exp\{0.0596\} - 1)\% = 6.14\%$.

Logistic Regression: Negative Coefficient Interpretations

Interpret

$$\frac{\pi_i}{1-\pi_i} = \frac{\exp\{-0.067\}}{1}$$

$$y_i \overset{\text{ind}}{\sim} \text{Bern}(\pi_i)$$

$$\frac{1-\pi_i}{\pi_i} = \frac{1}{\exp\{-0.067\}}$$

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \sum_{k=1}^{p-1} \hat{\beta}_k x_{ik}$$

- *If* $\hat{\beta}_{\text{age}} = -0.067$, how would we interpret this number? (Hint: flip the odds)

Based on $\log(OR)$

- Since $-0.067 < 0$, we know the log odds of developing CHD *decrease* as age increases, holding all else constant.

Based on OR

- Holding all else constant, for every additional year in age, the odds a patient does **NOT** develop CHD is $1/\exp\{-0.067\} = 1.0693$ times more likely.
- Holding all else constant, for every additional year in age, the odds a patient does **NOT** develop CHD increase by $100 \times ((1/\exp\{-0.067\}) - 1)\% = 6.93\%$.

Logistic Regression: Confidence Intervals – CHD Data Set

Interpret

$$y_i \overset{\text{ind}}{\sim} \text{Bern}(\pi_i) \qquad \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \sum_{k=1}^{p-1} \hat{\beta}_k x_{ik}$$

- Confidence intervals are calculated the same way as before
- 95% CI for $\hat{\beta}_{\text{age}} = (0.026, 0.093)$. How do we interpret this interval?

Logistic Regression Model Predicted Probability

Logistic Regression: Prediction

$$y_i \overset{\text{ind}}{\sim} \text{Bern}(\pi_i) \qquad \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \sum_{k=1}^{p-1} \hat{\beta}_k x_{ik} \qquad \hat{\pi}_i = \frac{\exp\{\hat{\beta}_0 + \sum_{k=1}^{p-1} x_{ik}\hat{\beta}_k\}}{1 + \exp\{\hat{\beta}_0 + \sum_{k=1}^{p-1} x_{ik}\hat{\beta}_k\}}$$

- We can use our model for prediction, as well.
- Often, it is nice to compute a predicted probability (of developing CHD) for a specific observation.
- If we want our interpretations based on probabilities, rather than log odds or odds, then we can convert the log odds to probabilities:

$$\pi_i = \frac{\exp\{\log \text{odds}_i\}}{1 + \exp\{\log \text{odds}_i\}}$$

Logistic Regression: CHD Data Set

- Fitted model:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -12.71 + 0.06 \times \text{age}_i + 0.02 \times \text{weight}_i + 0.02 \times \text{sbp}_i + 0.01 \times \text{chol}_i + 0.02 \times \text{cigs}_i$$

$$\text{where } \hat{\pi}_i = \text{Prob}(\text{chd}_i = 1 | \text{age}_i, \text{weight}_i, \text{sbp}_i, \text{chol}_i, \text{cigs}_i)$$

Logistic Regression: Prediction – CHD Data Set

Predict

- The log odds that a patient has CHD if the patient has the following characteristics: age=50, weight=182, sbp=136, chol=253, and cigs=20 is

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -12.71 + 0.06 \times \text{age}_i + 0.02 \times \text{weight}_i + 0.02 \times \text{sbp}_i$$
$$+ 0.01 \times \text{chol}_i + 0.02 \times \text{cigs}_i$$
$$= -12.71 + 0.06(50) + 0.02(182) + 0.02(136) + 0.01(253) + 0.02(20)$$

$= 0.21$ (keeping all decimals in Python)

- So, the predicted probability that a patient with these characteristics has CHD is:

$$\hat{\pi}_i = \frac{\exp\{0.21\}}{1+\exp\{0.21\}} = 0.55$$

Logistic Regression Model Assumptions

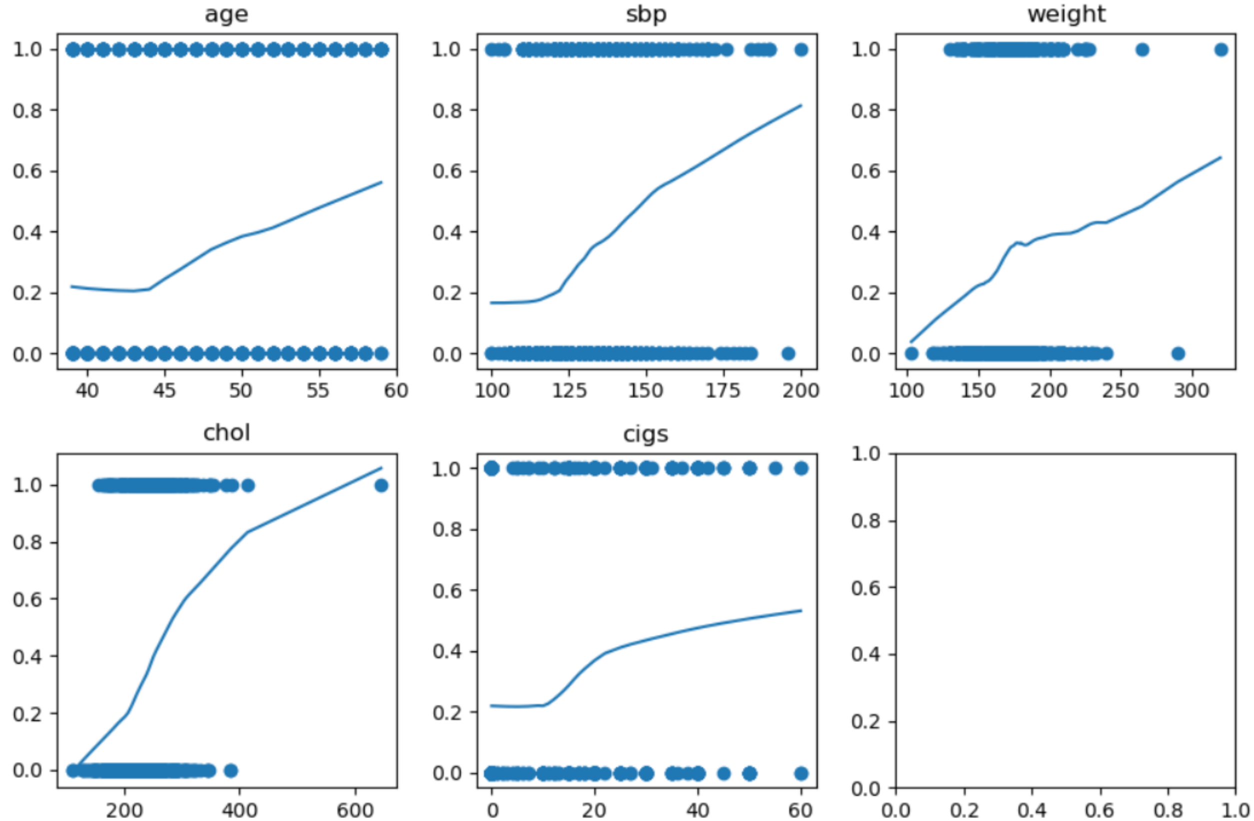
Logistic Regression: Model Assumptions

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ik} \quad \text{where } \pi_i = \text{Prob}(Y_i = 1 | x_{i1}, \dots, x_{ip-1})$$

1. The Xs vs log odds are linear ($\log\left(\frac{\pi_i}{1-\pi_i}\right)$ is a linear function of the Xs)
2. The ~~residuals~~ observations are independent
3. ~~The residuals are normally distributed and centered at zero~~
4. ~~The residuals have equal (constant) variance σ^2 across all values of the Xs (homoscedastic)~~
5. The model describes all observations (i.e., there are no influential points)
6. Additional predictor variables are not required
7. No Multicollinearity (can perform variable selection the same way as for linear regression)

Logistic Regression: Model Assumptions

1. The Xs vs log odds are linear ($\log\left(\frac{\pi_i}{1-\pi_i}\right)$ is a linear function of the Xs)
    - (monotone in probability) Check this with a scatterplot with a smoother



Logistic Regression: Model Assumptions

2. The observations are independent
   - Assess the same as before

3. The model describes all observations (i.e., there are no influential points)
   - Assess the same as before

4. Additional predictor variables are not required
   - Assess the same as before

5. No Multicollinearity
   - VIFs are based on R^2, but we don't have a way to calculate R^2 in logistic regression (no sums of squares)
   - What do we do?

Many diagnostics for multicollinearity can be obtained by using an OLS regression model. "Because the concern is with the relationship among the independent variables, the functional form of the model for the dependent variable is irrelevant to the estimation of collinearity"

Logistic Regression Model Performance

Logistic Regression: Model Performance

- How can we describe how well a logistic model performs?

  - We no longer can use the sum of squares metrics we used before (because we no longer assume normality and we don't have residuals)

  - So, we cannot use metrics like the MSE, RMSE, R^2, F statistic, etc.

Logistic Regression: Model Performance

- Instead, we use:
  - Deviance/Likelihood Ratio Test (generalization of the residual sum of squares, RSS)
  - Pseudo R^2
  - Confusion Matrix & Associated Metrics
  - ROC Curve & AUC Value

Logistic Regression: Model Output

Performance

```
Optimization terminated successfully.
         Current function value: 0.554881
         Iterations 6
```

Logit Regression Results

Dep. Variable:	chd	**No. Observations:**	605
Model:	Logit	**Df Residuals:**	599
Method:	MLE	**Df Model:**	5
Date:	Mon, 13 Nov 2023	**Pseudo R-squ.:**	0.1446
Time:	09:45:57	**Log-Likelihood:**	-335.70
converged:	True	**LL-Null:**	-392.47
Covariance Type:	nonrobust	**LLR p-value:**	7.312e-23

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-12.7071	1.426	-8.914	0.000	-15.501	-9.913
age	0.0596	0.017	3.498	0.000	0.026	0.093
weight	0.0175	0.005	3.889	0.000	0.009	0.026
sbp	0.0204	0.006	3.332	0.001	0.008	0.032
chol	0.0138	0.002	5.931	0.000	0.009	0.018
cigs	0.0245	0.006	3.970	0.000	0.012	0.037

Logistic Regression: Deviance

- To test the overall performance of the model, we use the model χ^2 test (df = number of covariates in model), which is analogous to the model F-test in linear regression):

$$\underbrace{\text{Model } \chi^2}_{\text{Deviance/Likelihood Ratio Test Statistic}} = \underbrace{-2\log\mathcal{L}_{\text{intercept}}}_{\text{-2 times the log likelihood of a model only including the intercept called the null deviance}} - \underbrace{\left(-2\log\mathcal{L}_{\text{int\&covariates}}\right)}_{\text{-2 times the log likelihood of a model including the intercept and all predictors called the residual deviance}}$$

$$=$$

$$=$$

Interpretation:

Note: the likelihood for logistic regression is: $\mathcal{L} = \prod_{i=1}^{n} P(Y_i = y_i | x_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$

Logistic Regression: Deviance

- To test the overall performance of the model, we use the model χ^2 test (df = number of covariates in model), which is analogous to the model F-test in linear regression):

Take the difference in the two deviances shown in the R output

Equivalent to us classifying everyone as not having CHD and being accurate ~66% of the time

$$\underbrace{\text{Model } \chi^2}_{\text{Deviance/Likelihood Ratio Test Statistic}} = \underbrace{-2\log\mathcal{L}_{\text{intercept}}}_{\text{Total variation}} - \underbrace{\left(-2\log\mathcal{L}_{\text{int\&covariates}}\right)}_{\text{Variation not explained by model}}$$

Deviance/Likelihood Ratio Test Statistic

-2 times the log likelihood of a model only including the intercept **called the null deviance**

-2 times the log likelihood of a model including the intercept and all predictors **called the residual deviance**

LL-Null: **Log-Likelihood:**

$$= -2(-392.47) - (-2)(-335.70)$$

$$= 784.94 - 671.40 = 113.54 \ (p\text{-value} \approx 0)$$

Interpretation:

Note: the likelihood for logistic regression is: $\mathcal{L} = \prod_{i=1}^{n} P(Y_i = y_i | x_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$

Logistic Regression: Pseudo R^2

- Pseudo $R^2 = 1 - \frac{\text{residual deviance}}{\text{null deviance}}$ **Log-Likelihood:** **LL-Null:**

$$= 1 - \frac{-335.70}{-392.47} = 1 - 0.8554 = 0.1446 \text{ for CHD data}$$

- Interpretation: percent of variation in $\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)$ explained by the model.

- Note: in practice, the upper bound typically is not 1 (low Pseudo R^2 values are normal even if your model does well at classifying)

Logistic Regression: Classification

Performance

$$y_i \overset{\text{ind}}{\sim} \text{Bern}(\pi_i) \qquad \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \sum_{k=1}^{p-1} x_{ik}\hat{\beta}_k \qquad \hat{\pi}_i = \frac{\exp\{\hat{\beta}_0 + \sum_{k=1}^{p-1} x_{ik}\hat{\beta}_k\}}{1+\exp\{\hat{\beta}_0 + \sum_{k=1}^{p-1} x_{ik}\hat{\beta}_k\}}$$

logit function

logistic function(derived from logit)

- Many times we want to classify, so we set

$$\hat{y} = \begin{cases} 1 & \text{if } \hat{\pi} > c \\ 0 & \text{if } \hat{\pi} \leq c \end{cases}$$

where c is a cutoff probability.

Can use 0.5 as the cutoff, but if you have imbalance (more “No”s than “Yes”s, then can adjust cutoff probability accordingly

Logistic Regression ON TEST DATA SET Performance

		Predicted	
		Yes	No
Truth	Yes	19	25
	No	13	95

- Using a cutoff value, we can produce a **confusion matrix:**
  - **True Positives**: Predicted “Yes” and Truth “Yes” (19)
  - **True Negatives**: Predicted “No” and Truth “No” (95)
  - **False Positives**: Predicted “Yes” and Truth “No” (13, type I error)
  - **False Negatives**: Predicted “No” and Truth “Yes” (25, type II error)
  - **Sensitivity/Recall**: Percent of correctly predicted “Yes”s among all “Yes”s $\left(\frac{19}{19+25} = 0.43\right)$
  - **Specificity**: Percent of correctly predicted “No”s among all “No”s $\left(\frac{95}{95+13} = 0.88\right)$
  - **Positive Predictive Value/Precision**: Percent of correctly predicted “Yes”s $\left(\frac{19}{19+13} = 0.59\right)$
  - **Negative Predictive Value**: Percent of correctly predicted “No”s $\left(\frac{95}{95+25} = 0.79\right)$
  - **Percent Correctly Classified/Accuracy:** Percent of correctly predicted “Yes”s and “No”s $\left(\frac{19+95}{152}\right) = 0.75$ (compare with 66% without using any variables)

Logistic Regression: Classification

- A large value for c results in _______________ patients classified as developing CHD, which results in a _______________ specificity and a _______________ sensitivity.

- A small value for c results in _______________ patients classified as developing CHD, which results in a _______________ specificity and a _______________ sensitivity.

Logistic Regression

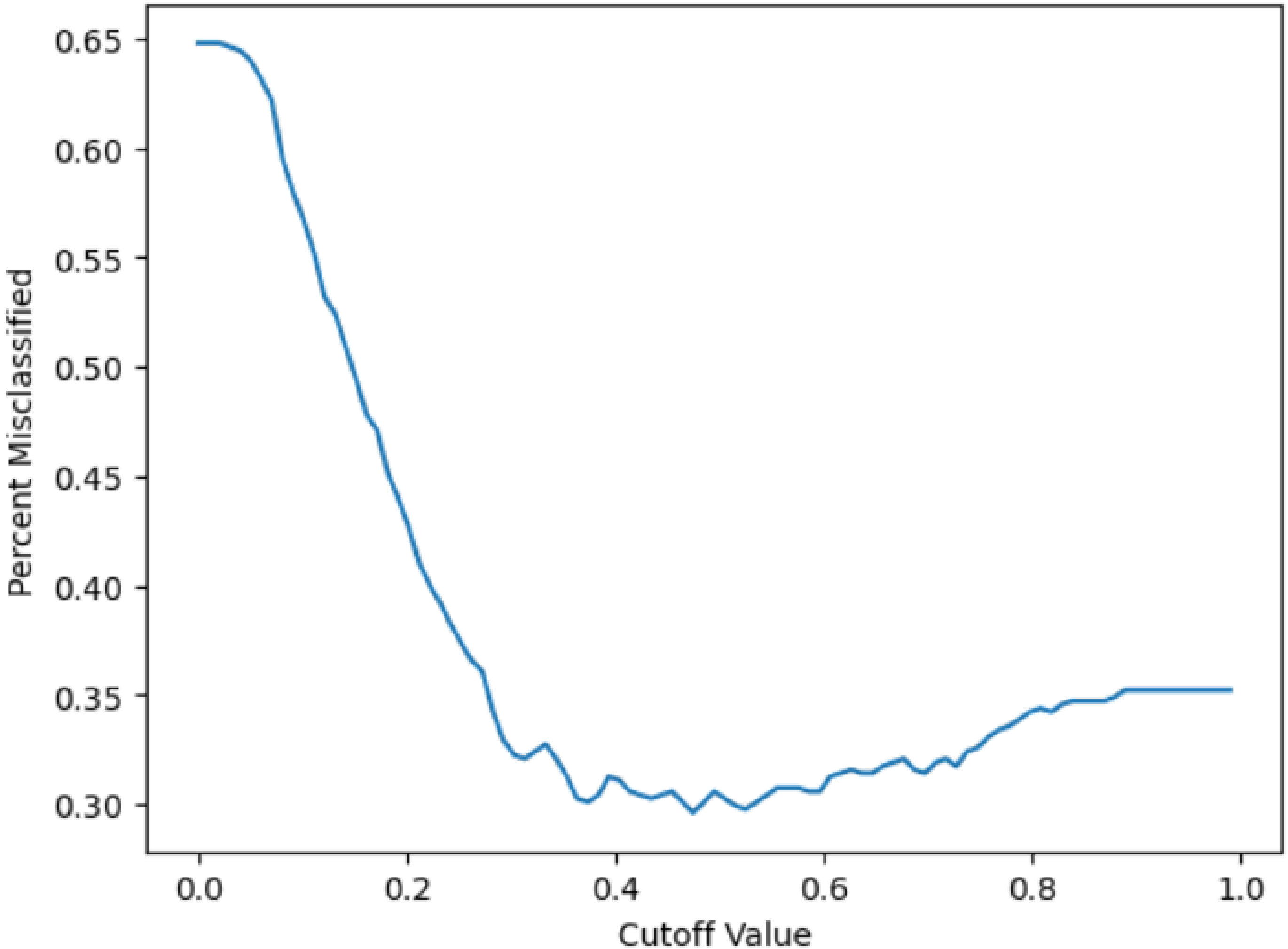
- So, how do we choose the cutoff value c?

1. $c = 0.5$

2. Choose c to minimize the misclassification rate

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) = \text{Percent Misclassified}$$

CHD Example


Percent Misclassified
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.0
0.2
0.4
0.6
0.8
1.0
Cutoff Value

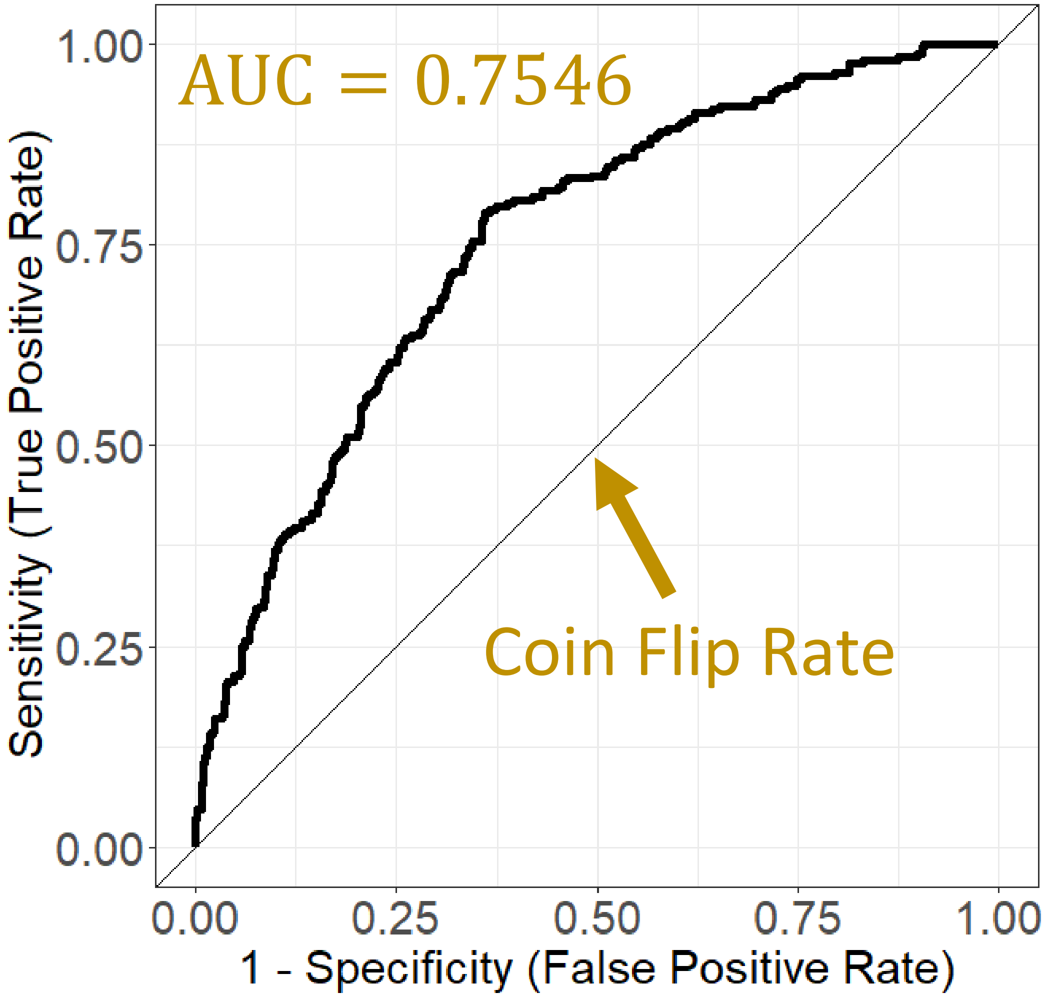
Logistic Regression: ROC and AUC

- Classification (and the confusion matrix) is built on a cutoff. We can see how well our model does across all cutoff values using the ROC curve.

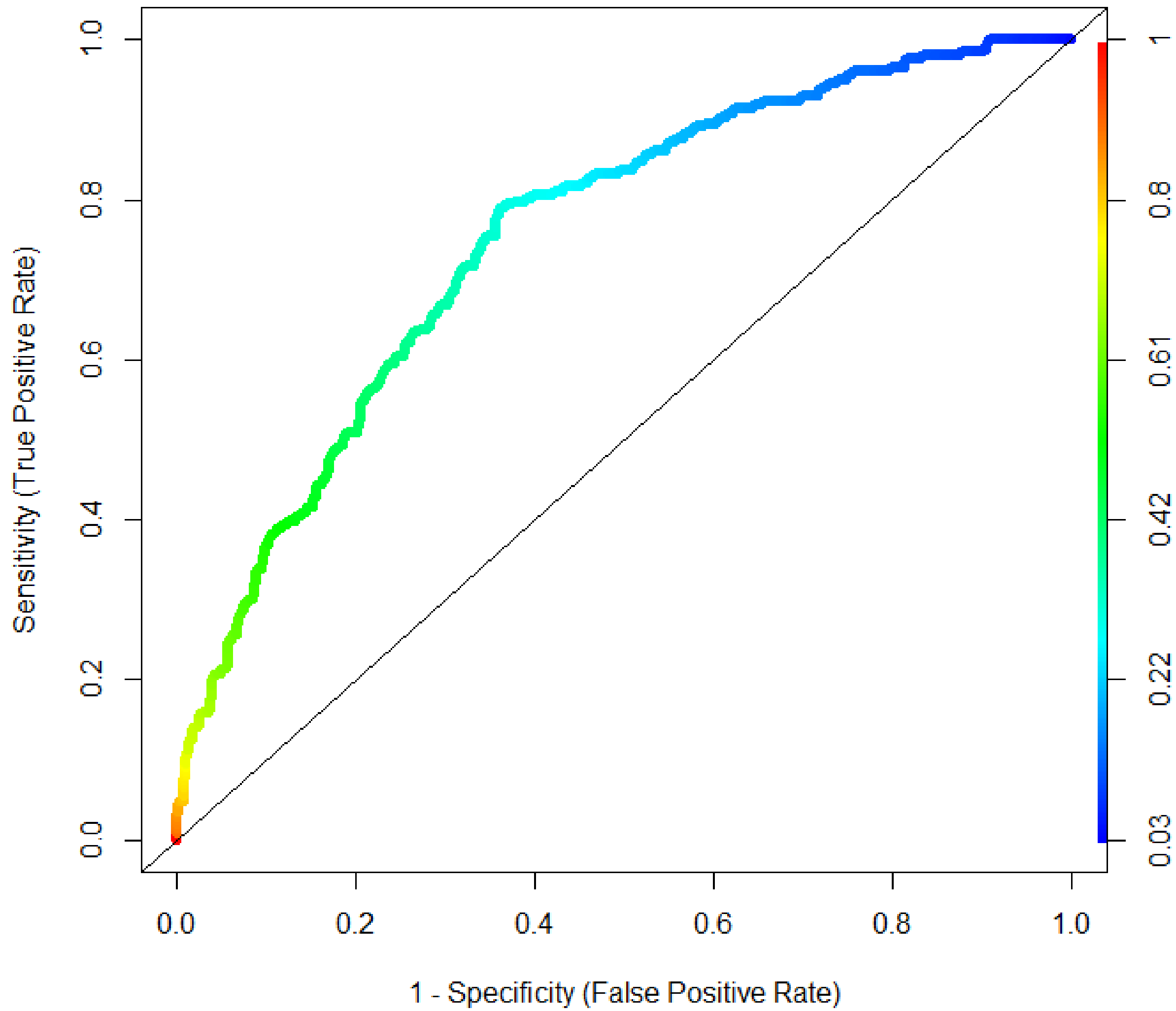
- ROC (Receiver Operating Characteristic) Curves: For many cutoff values, compare the true positive rate (sensitivity) to the false positive rate (1 – specificity)

- We can summarize an ROC curve by the area under the curve (AUC)
  - AUC is the rate of successful classification
    - We want $\mathrm{AUC} >> 0.50$ (we want our model to do better than guessing)

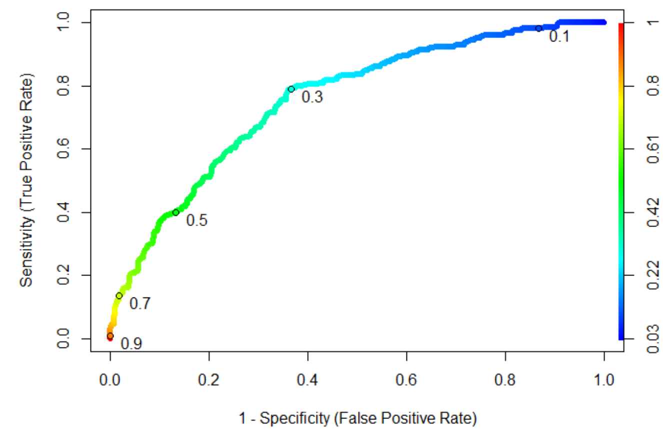
Logistic Regression: ROC and AUC



Logistic Regression: ROC Colored by Cutoff Value



Logistic Regression: ROC Colored by Cutoff Value



- $c = 0.90$

Sensitivity (Percent of true positives) $\frac{1}{1+256} = 0.004$

Specificity (Percent of true negatives) $\frac{500}{500+0} = 1$

1-Specificity (Percent of false positives) $1 - 1 = 0$

		Predicted	
		Yes	No
Truth	Yes	1	256
	No	0	500

- $c = 0.50$

Sensitivity (Percent of true positives) $\frac{102}{102+155} = 0.40$

Specificity (Percent of true negatives) $\frac{434}{434+66} = 0.87$

1-Specificity (Percent of false positives) $1 - 0.87 = 0.13$

		Predicted	
		Yes	No
Truth	Yes	102	155
	No	66	434

- $c = 0.10$

Sensitivity (Percent of true positives) $\frac{252}{252+5} = 0.98$

Specificity (Percent of true negatives) $\frac{67}{67+433} = 0.13$

1-Specificity (Percent of false positives) $1 - 0.13 = 0.87$

		Predicted	
		Yes	No
Truth	Yes	252	5
	No	433	67

Recall:

$$\hat{y} = \begin{cases} 1 & \text{if } \hat{\pi} > c \\ 0 & \text{if } \hat{\pi} \le c \end{cases}$$