Logistic regression

Model Predictions & Assumptions

Prof. Maria Tackett

03.30.20

Click for PDF of slides

Risk of coronary heart disease

This data is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The goal is to predict whether a patient has a 10-year risk of future coronary heart disease.

Response:

`TenYearCHD`:

- 0 = Patient doesn't have 10-year risk of future coronary heart disease
- 1 = Patient has 10-year risk of future coronary heart disease

Predictor:

- **`age`**: Age at exam time.
- **`currentSmoker`**: 0 = nonsmoker; 1 = smoker
- **`totChol`**: total cholesterol (mg/dL)

Logistic Regression Model

- Suppose $P(y_i = 1|x_i) = \pi_i$ and $P(y_i = 0|x_i) = 1 - \pi_i$
- The **logistic regression model** is

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

- $\log\left(\frac{\pi_i}{1-\pi_i}\right)$ is called the **logit** function

Modeling risk of coronary heart disease

```
risk_m <- glm(TenYearCHD ~ ageCent + currentSmoker + totCholCent,
              data = heart_data, family = binomial)
tidy(risk_m, conf.int = TRUE) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.111	0.077	-27.519	0.000	-2.264	-1.963
ageCent	0.081	0.006	13.477	0.000	0.070	0.093
currentSmoker1	0.447	0.099	4.537	0.000	0.255	0.641
totCholCent	0.003	0.001	2.339	0.019	0.000	0.005

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -2.111 + 0.081\text{ageCent} + 0.447\text{currentSmoker} + 0.003\text{totChol}$$

Part 1: Prediction

Using the model for prediction

- We are often interested in predicting whether a given observation will have a "yes" response
- To do so
  - Use the logistic regression model to calculate the predicted log-odds that an observation has a "yes" response
  - Then, use the log-odds to calculate the predicted probability of a "yes" response
  - Then, use the predicted probabilities to classify the observation as having a "yes" or "no" response

Calculating the predicted probability

$$\hat{\pi}_i = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}}$$

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

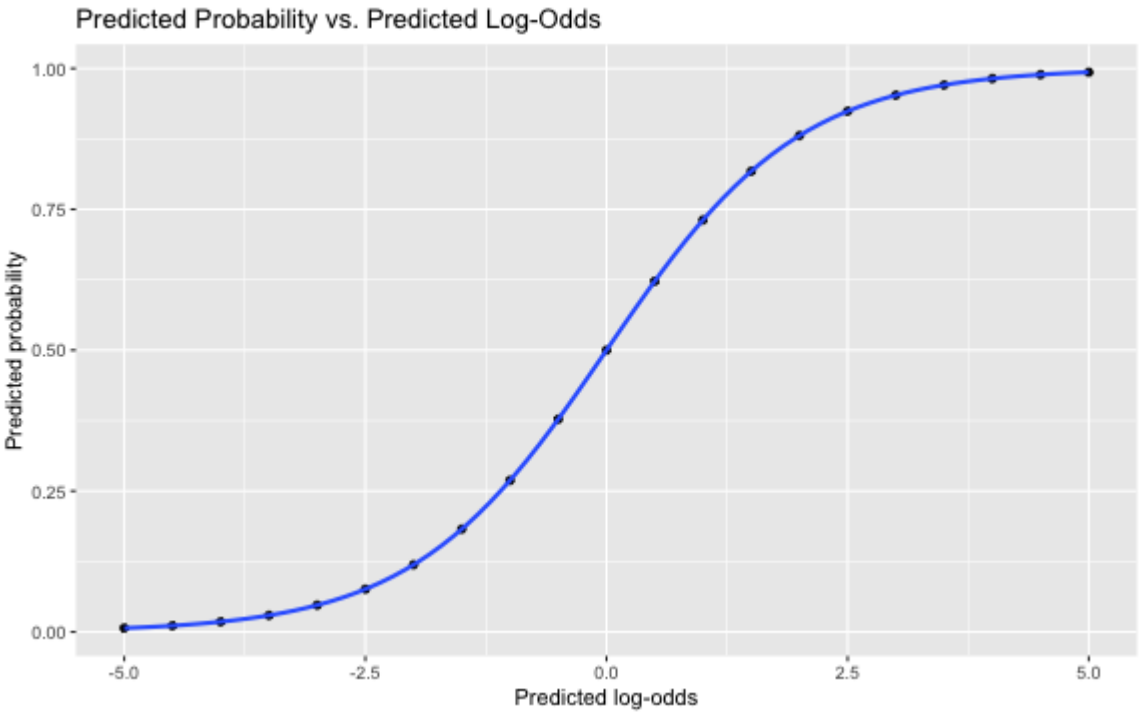
$$\Rightarrow \exp\left\{\log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right)\right\} = \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}$$

$$\Rightarrow \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} = \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}$$

$$\Rightarrow \hat{\pi}_i = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}}$$

$\hat{\pi}$ vs. $\widehat{\text{log-odds}}$

$$\hat{\pi}_i = \frac{\exp(hat\beta_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)} = \frac{\exp(\widehat{\text{log-odds}})}{1 + \exp(\widehat{\text{log-odds}})}$$



Predicted response for a patient

- Suppose a patient comes in who is 60 years old, does not currently smoke, and has a total cholesterol of 263 mg/dL.
- Predicted log-odds that this person is high risk for coronary heart disease in the next 10 years:

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -2.111 + 0.081\text{ageCent} + 0.447\text{currentSmoker} + 0.003\text{totC}$$

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -2.111 + 0.081 \times (60 - 49.552) + 0.447 \times 0 + 0.003 \times (263 - 236.848) \approx -1.186$$

- The probability this patient is high risk for coronary heart disease in the next 10 years:

$$\hat{\pi}_i = \frac{\exp\{-1.186\}}{1+\exp\{-1.186\}} = 0.234$$

Predictions in R

```
x0 <- data_frame(ageCent = (60 - 49.552),
                 totCholCent = (263 - 236.848),
currentSmoker = as.factor(0))
```

- Predicted log-odds

```
predict(risk_m, x0)
```

```
##         1
## -1.192775
```

- Predicted probabilities

```
predict(risk_m, x0, type = "response")
```

```
##         1
## 0.2327631
```

Is this patient high risk?

```
predict(risk_m, x0, type = "response")
```

```
##         1
## 0.2327631
```

The probability the patient is at risk for coronary heart disease is 0.233.

Based on this probability, would you consider this patient as being high risk for getting coronary heart disease in the next 10 years? Why or why not?

Confusion Matrix

- We can use the predicted probability to predict the outcome for a given observation
    - In other words, we can classify the observations into two groups: "yes" and "no"
- **How**: Establish a threshold such that $y = 1$ if predicted probability is greater than the threshold $(y = 0 \text{ otherwise})$
- To assess the accuracy of our predictions, we can make a table of the observed (actual) response versus the predicted response.
    - This table is the confusion matrix
- We can use this table to calculate the proportion of observations that were misclassifed. This is the misclassification rate.

Confusion Matrix

Suppose we use 0.3 as the threshold to classify observations

```
threshold <- 0.3
risk_m_aug <- augment(risk_m, type.predict = "response",
                      type.residuals = "response")
```

```
risk_m_aug %>%
  mutate(risk_predict = if_else(.fitted > threshold, "Yes", "No")) %>%
  group_by(TenYearCHD, risk_predict) %>%
  summarise(n = n()) %>%
  kable(format="markdown")
```

TenYearCHD	risk_predict	n
0	No	2899
0	Yes	202
1	No	457
1	Yes	100

Confusion matrix

TenYearCHD	risk_predict	n
0	No	2899
0	Yes	202
1	No	457
1	Yes	100

What proportion of observations were misclassified?

What is the disadvantage of relying on the confusion matrix to assess the accuracy of the model?

Confusion matrix: 2 X 2 table

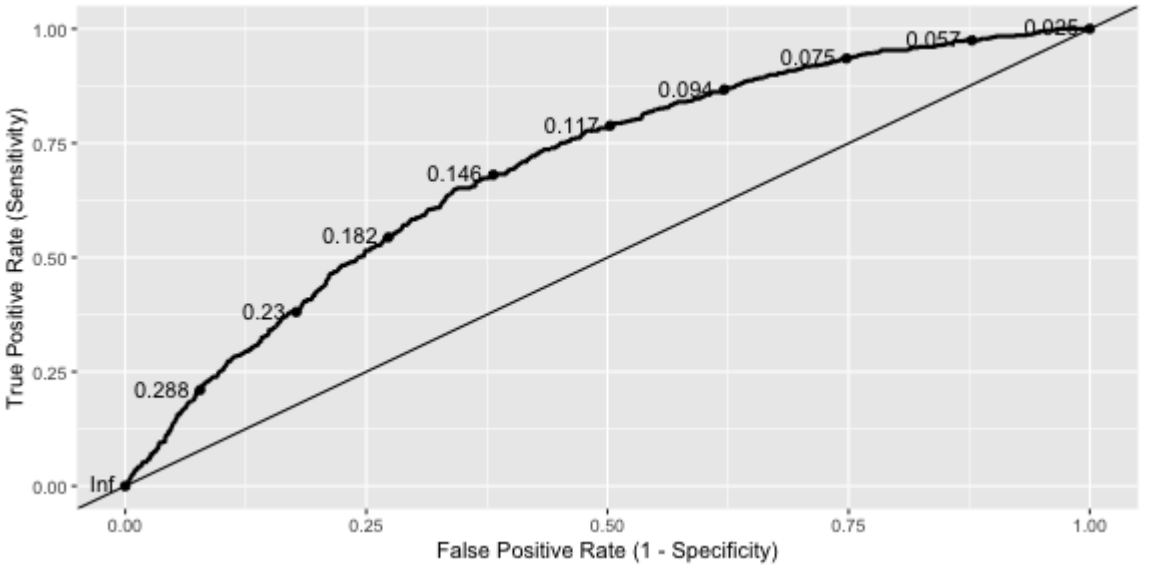
In practice, you often see the confusion matrix presented as a 2×2 table as shown below:

```
risk_m_aug %>%
  mutate(risk_predict = if_else(.fitted > threshold, "Yes", "No")) %>%
  group_by(TenYearCHD, risk_predict) %>%
  summarise(n = n()) %>%
  spread(risk_predict, n) %>%
  kable(format="markdown")
```

TenYearCHD	No	Yes
0	2899	202
1	457	100

Receiver Operating Characteristic (ROC) curve

```
library(plotROC) #extension of ggplot2
(roc_curve <- ggplot(risk_m_aug,
                     aes(d = as.numeric(TenYearCHD) - 1,
                         m = .fitted)) +
  geom_roc(n.cuts = 10, labelround = 3) +
  geom_abline(intercept = 0) +
  labs(x = "False Positive Rate (1 - Specificity)",
       y = "True Positive Rate (Sensitivity)") )
```



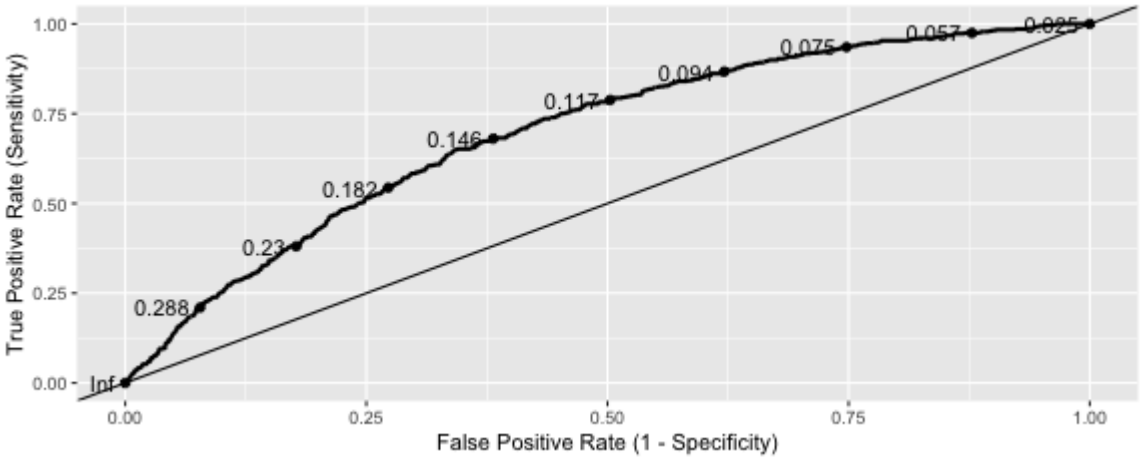
Sensitivity & Specificity

- **Sensitivity:** Proportion of observations with $y = 1$ that have predicted probability above a specified threshold
  - Called **true positive rate** (y-axis)
- **Specificity:** Proportion of observations with $y = 0$ that have predicted probability below a specified threshold
  - (1 - specificity) called **false positive rate** (x-axis)
- What we want:
  - High sensitivity
  - Low values of 1-specificity

Area under curve (AUC)

We can use the area under the curve (AUC) as one way to assess how well the logistic model fits the data

- $AUC = 0.5$ very bad fit (no better than a coin flip)
- AUC close to 1: good fit




```
calc_auc(roc_curve)$AUC
```

```
## [1] 0.6972743
```

Which threshold would you choose?

A doctor plans to use the results from your model to help select patients for a new heart disease prevention program. She asks you which threshold would be best to select patients for this program. Based on the ROC curve from the previous slide, which threshold would you recommend to the doctor? Why?

Part 2: Checking Assumptions

Assumptions for logistic regression

We want to check the following assumptions for the logistic regression model:

- **Linearity**: Is there a linear relationship between the log-odds and the predictor variables?

- **Randomness**: Was the sample randomly selected? Or can we reasonably treat it as random?

- **Independence**: There is no obvious relationship between observations

Linearity: binned residual plots

- It is not useful to plot the raw residuals, so we will examine binned residual plots

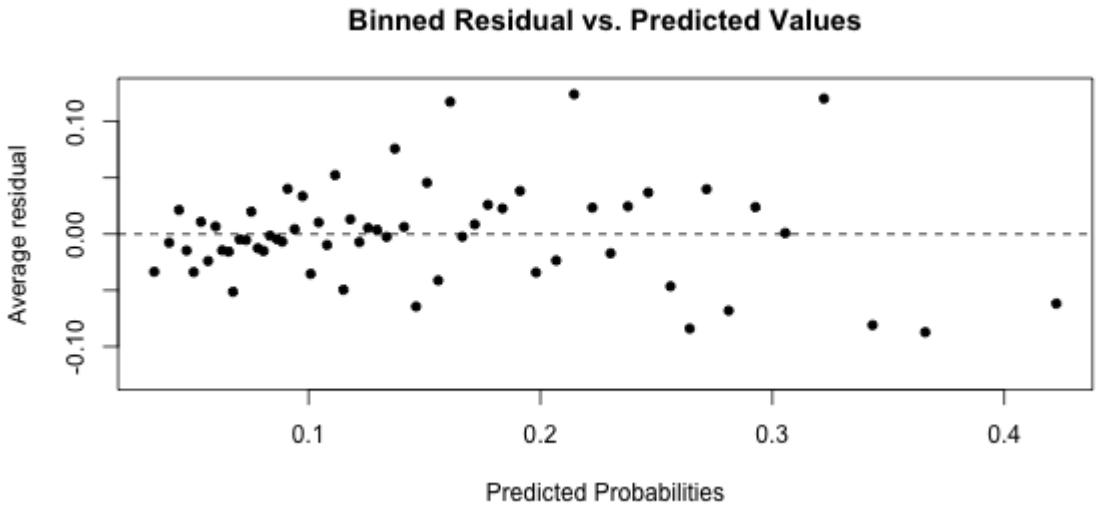
When examining binned residuals

- Plot should have no discernible pattern or trend
    - Nonlinear trend may be indication that squared term or log transformation of predictor variable required
- If bins have average residuals with large magnitude
    - Look at averages of other predictor variables across bins
    - Interaction may be required if large magnitude residuals correspond to certain combinations of predictor variables

Binned plot vs. predicted values

- Use the **`binnedplot`** function in the **arm** package.
  - *Tip: Don't load the **arm** package to avoid conflicts with tidyverse*

```
arm::binnedplot(x = risk_m_aug$.fitted, y = risk_m_aug$.resid,
                xlab = "Predicted Probabilities",
                main = "Binned Residual vs. Predicted Values",
                col.int = FALSE)
```



Making binned residual plot

- Calculate raw residuals $(y_i - \hat{\pi}_i)$
- Order observations either by the values of the predicted probabilities (or by numeric predictor variable)
- Use the ordered data to create g bins of approximately equal size. Default value: $g = \sqrt{n}$
- Calculate average residual value in each bin
- Plot average residuals vs. average predicted probability (or average predictor value)

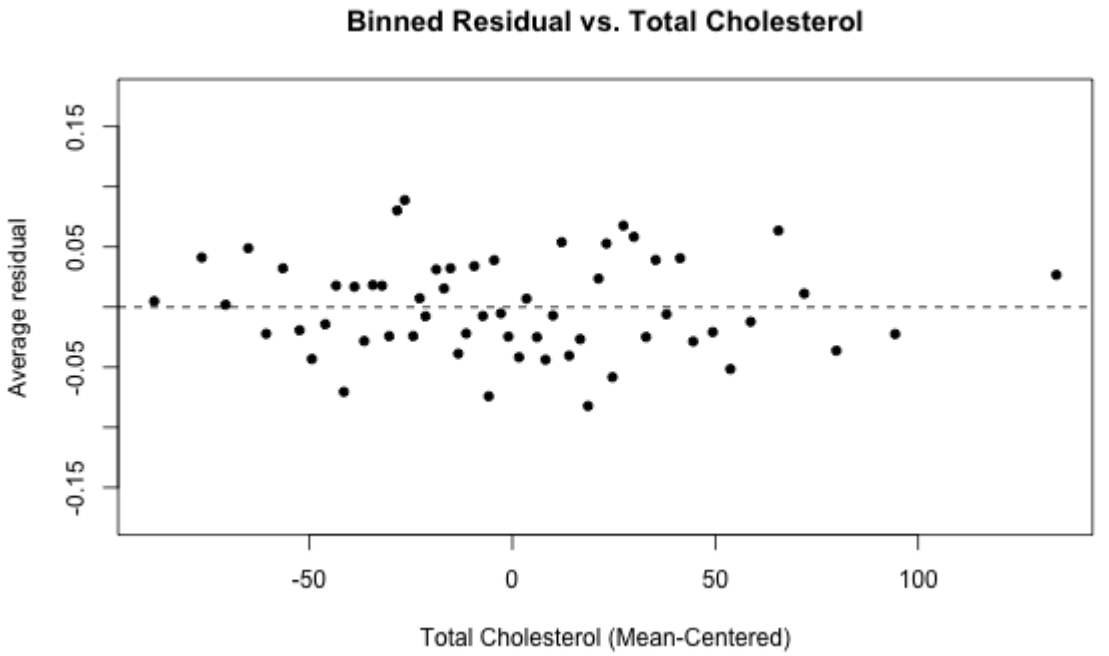
Residuals vs. Age

Make binned plot with predictor on x axis

```
arm::binnedplot(x = risk_m_aug$ageCent,
                y = risk_m_aug$.resid,
                col.int = FALSE,
                xlab = "Age (Mean-Centered)",
                main = "Binned Residual vs. Age")
```

Residuals vs. `totChol`

```
arm::binnedplot(x = risk_m_aug$totCholCent,
                y = risk_m_aug$.resid,
                col.int = FALSE,
                xlab = "Total Cholesterol (Mean-Centered)",
                main = "Binned Residual vs. Total Cholesterol")
```



Residuals vs. categorical predictors

- Calculate average residual for each level of the predictor
  - Are all means close to 0? If not, there is a problem with model fit.

```
risk_m_aug %>%
  group_by(currentSmoker) %>%
  summarise(mean_resid = mean(.resid))
```

```
## # A tibble: 2 x 2
##   currentSmoker mean_resid
##   <fct>              <dbl>
## 1 0              -2.95e-14
## 2 1              -2.42e-14
```

Randomness

Assess randomness based on a description of the data collection

- Was the sample randomly selected?
- If the sample was not randomly selected, is there reason to believe the observations in the sample differ systematically from the population of interest?

What do you conclude about the randomness assumption for our dataset?

Independence

Assess independence based on a description of the data collection

- Is there an obvious relationship between observations?
  - This assumption is most often violated when data was collected over time or there is a spatial relationship between observations?

What do you conclude about the independence assumption for our dataset?