Logistic regression

Model Predictions

Dr. Maria Tackett

10.28.19

Click for PDF of slides

Announcements

- Lab 07 **due Tue, Oct 29 at 11:59p**
- Reading 09 for Wednesday
- Project Proposal **due Wed, Oct 30 at 11:59p**

Packages

```r
library(tidyverse)
library(knitr)
library(broom)
library(pROC) #ROC curves
```

Review

- y: binary response
  - 1: yes
  - 0: no
- $Mean(y) = p$
- $Var(y) = p(1-p)$
- **Odds of "yes":** $\omega = \frac{p}{1-p}$

Logistic Regression Model

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

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

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

Interpreting the intercept: β_0

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

- Intercept: When $x = 0$, odds of y are $\exp\{\beta_0\}$

Interpreting slope coefficient β_1

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

If x is a quantitative predictor

- As x_i increases by 1 unit, we expect the odds of y to multiply by a factor of $\exp\{\beta_1\}$

If x is a categorical predictor

- The odds of y for group k are expected to be $\exp\{\beta_1\}$ times the odds of y for the baseline group.

Inference for coefficients

- The standard error is the estimated standard deviation of the sampling distribution of $\hat{\beta}_1$
- We can calculate the C confidence interval based on the large-sample Normal approximations
- CI for $\boldsymbol{\beta}_1$:

$$\hat{\beta}_1 \pm z^* SE(\hat{\beta}_1)$$

CI for $\exp\{\boldsymbol{\beta}_1\}$:

$$\exp\{\hat{\beta}_1 \pm z^* SE(\hat{\beta}_1)\}$$

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)

Modeling risk of coronary heart disease

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

1. Interpret age in terms of the odds of being at risk for coronary heart disease.
2. Interpret `currentSmoker1` in terms of the odds of being at risk for coronary heart disease.

Predictions & Model Fit

Predictions

- We are often interested in predicting if a given observation will have a "yes" response

- To do so, we will use the logistic regression model to predict the probability of a "yes" response for the given observation. If we have one predictor variable, then...

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

- We will use the predicted probabilities to classify the observation as having a "yes" or "no" response

Is the patient at risk for coronary heart disease?

- 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 at risk for coronary heart disease:

$$\begin{aligned}\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = & -2.111 + 0.081 \times (60 - 49.552) - 0.447 \times 0 \\ & + 0.002 \times (263 - 236.848) \approx -1.212\end{aligned}$$

- The probability this passenger thinks reclining the seat is rude:

$$\hat{p}_i = \frac{\exp\{-1.212\}}{1+\exp\{-1.212\}} = 0.229$$

Predictions in R

```r
x0 <- data_frame(ageCent = (60 - 49.552), totCholCent = (263 - 236
currentSmoker = as.factor(0))
```

- Predicted log-odds

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

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

- Predicted probabilities

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

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

Is the patient at risk for coronary heart disease?

```r
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 the patient at risk for coronary heart disease? Why or why not?

Confusion Matrix

- We can use the estimated probabilities to predict outcomes
- *Ex.*: Establish a threshold such that $y = 1$ if predicted probability is greater than the threshold ($y=0$ otherwise)
- Determine how many observations were classified correctly and incorrectly and put the results in a 2×2 table
  - This table is the **confusion matrix**
- If the proportion of misclassifications is high, then we might conclude the model doesn't fit the data well

Confusion Matrix

Suppose we use 0.3 as the threshold to classify responses

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

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

risk_predict	0	1
No	2899	457
Yes	202	100

Confusion matrix

risk_predict	0	1
No	2899	457
Yes	202	100

What proportion of observations were misclassified?

Sensitivity & Specificity

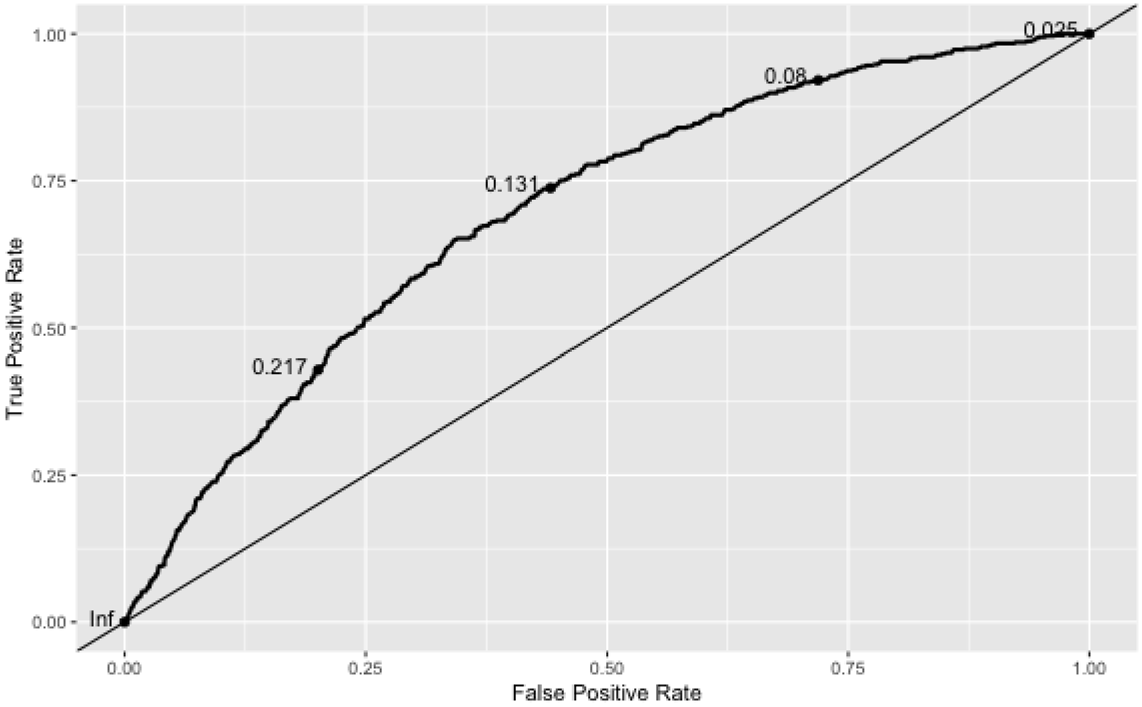
- **Sensitivity:** Proportion of observations with $y = 1$ that have predicted probability above a specified threshold
  - Called true positive rate
- **Specificity:** Proportion of observations with $y = 0$ that have predicted probability below a specified threshold
  - (1 - specificity) called false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

ROC Curve

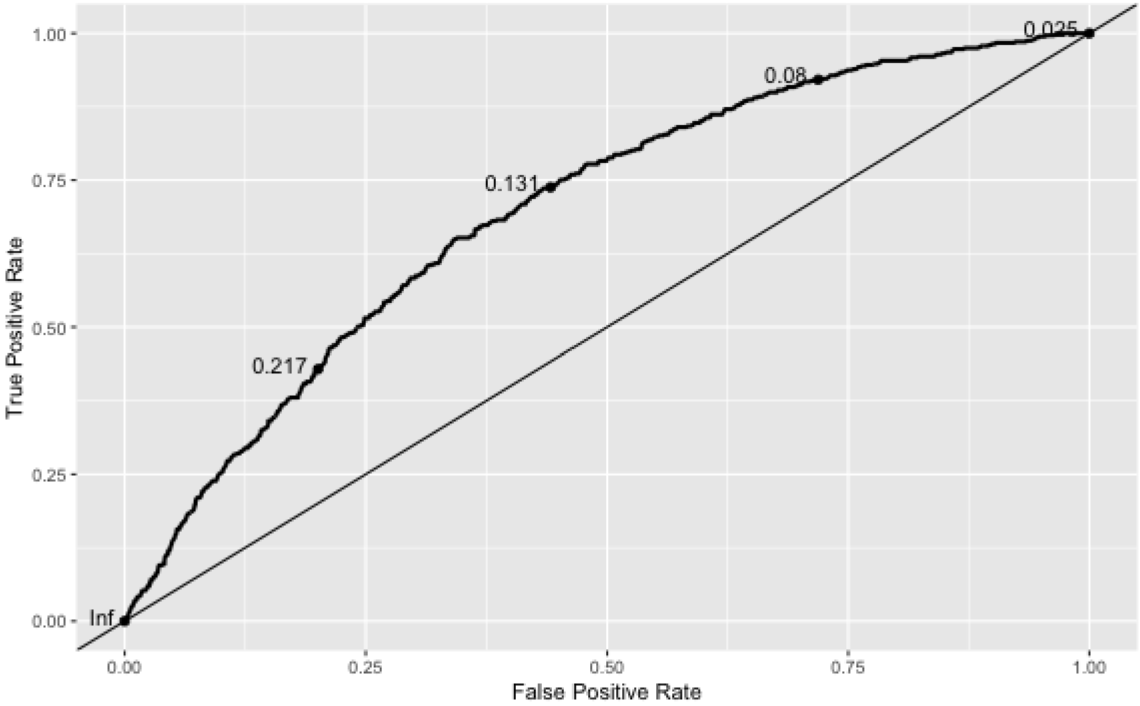
- Receive Operating Characteristic (ROC) curve :
  - *x-axis*: $1 - \text{specificity}$
  - *y-axis*: Sensitivity
- Evaluated with a lot of different values for the threshold
- Logistic model fits well if the area under the curve (AUC) is close to 1
- ROC in R
  - Use the **`roc`** function in the `pROC` to calculate AUC
  - Use **`geom_roc`** layer in ggplot to plot the ROC curve

Visualize ROC curve

```
library(plotROC) #extension of ggplot2
roc_curve <- ggplot(risk_m_aug, aes(d = as.numeric(TenYearCHD) -1,
  geom_roc(n.cuts = 5, labelround = 3) +
  geom_abline(intercept = 0) +
  labs(x = "False Positive Rate", y = "True Positive Rate")
roc_curve
```



Area under curve



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

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

Application Exercise

Copy the **Logistic Regression** project on RStudio Cloud