Logistic regression

Model fit & Exploratory data analysis

Dr. Maria Tackett

10.30.19

Click for PDF of slides

Announcements

- Reading 10 for Monday
- Project Proposal **due TODAY at 11:59p**
- Electronic Undergraduate Research Conference

Packages

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

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 $P(Y_i = 1|X_i) = p_i$ 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

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

Logistic Regression Model

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

- We can calculate p_i by solving the logit equation:

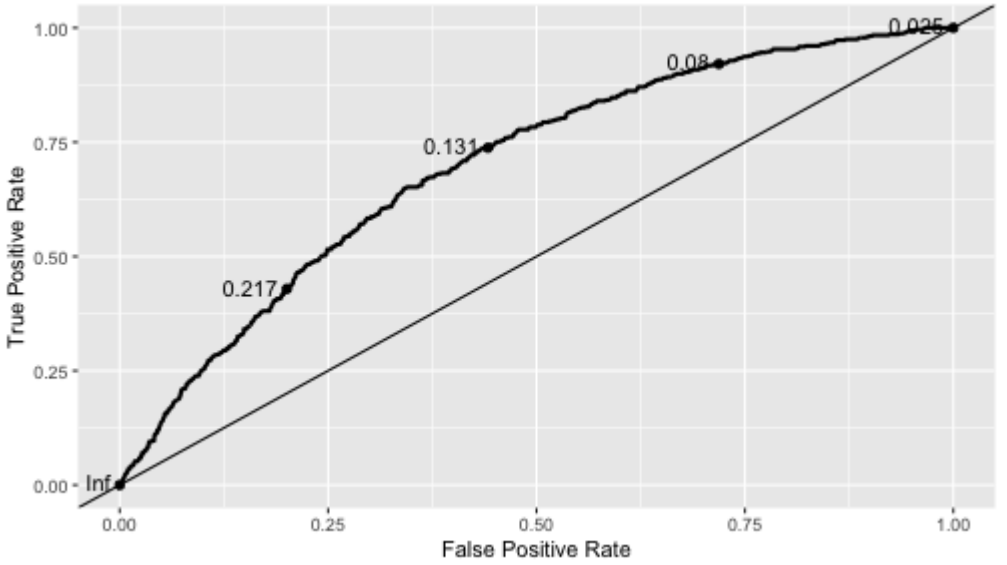
$$p_i = \frac{\exp\{\beta_0 + \beta_1 X_i\}}{1 + \exp\{\beta_0 + \beta_1 X_i\}}$$

ROC Curve

- Receiver Operating Characteristic (ROC) curve:
  - *X-axis*: $1 -$ 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

ROC Curve

.small[

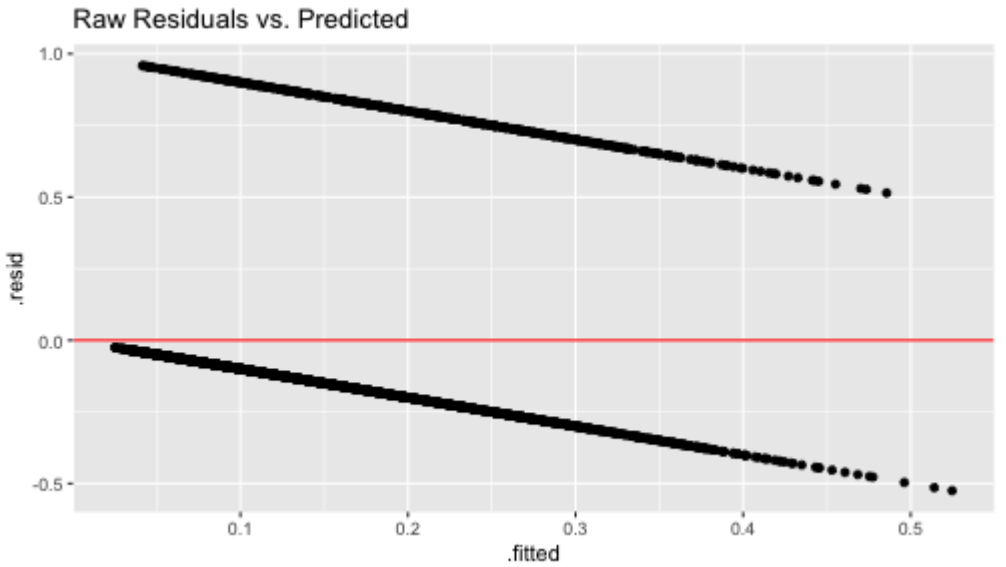



```
## [1] 0.6972743
```

Not useful: Raw residuals vs. predicted

- Include **`type.residuals = "response"`** in the `augment` function to get the raw residuals.

$$e_i = Y_i - \hat{p}_i$$



Binned Residuals

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

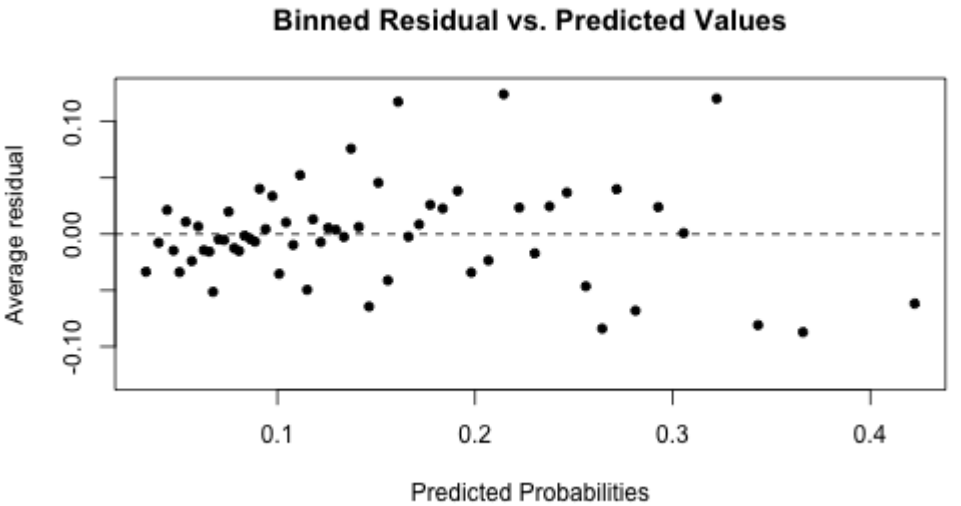
When examining binned residuals

- Look for patterns
- 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
- 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)

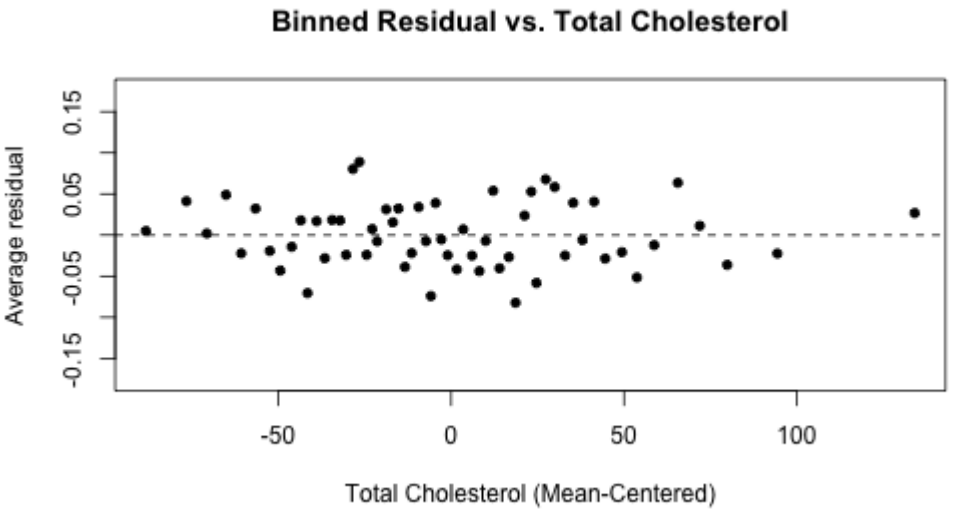
Residuals vs. Age

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

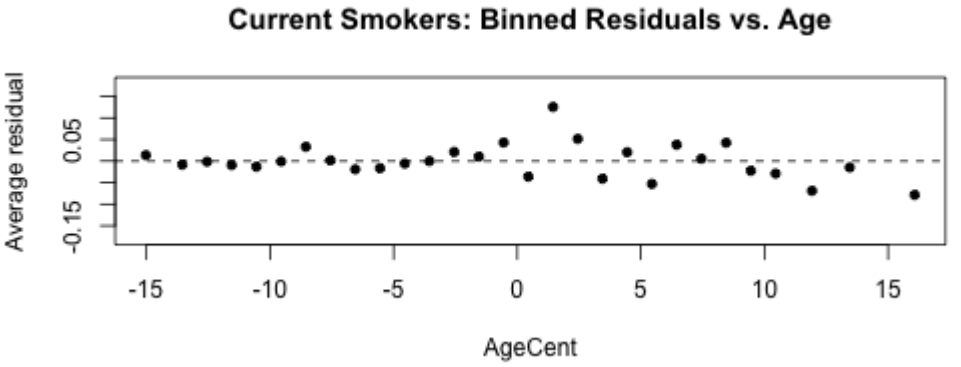
- Calculate average residual for each level of the predictor
  - Are all means close to 0? If not, there is a problem with model fit.

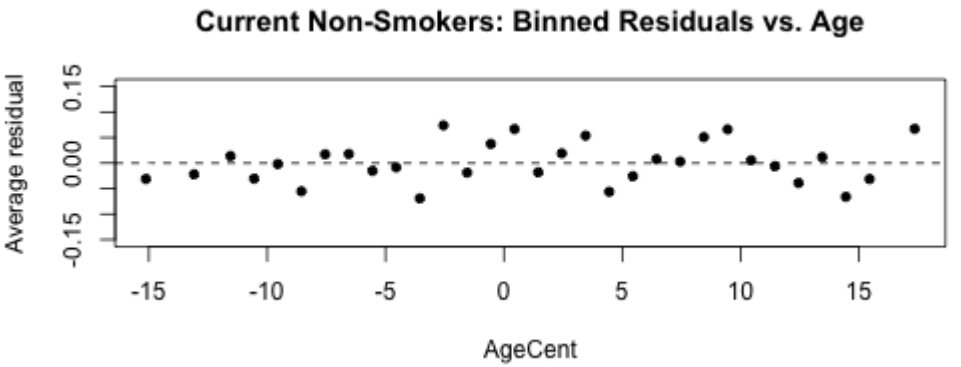
```
risk_m_aug %>%
  group_by(currentSmoker) %>%
  summarise(mean_resid = mean(.resid))
```

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

Residuals

Let's look at the binned residuals versus `AgeCent` separately for those who currently smoke and those who do not



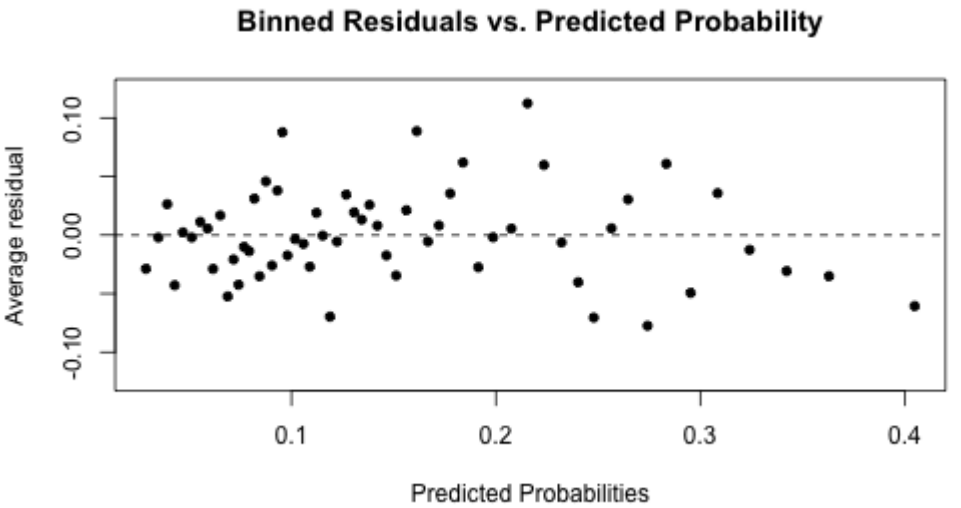


Model with interaction term

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.159	0.088	-24.498	0.000	-2.337	-1.992
ageCent	0.090	0.009	9.988	0.000	0.072	0.107
totCholCent	0.002	0.001	2.293	0.022	0.000	0.005
currentSmoker1	0.507	0.111	4.570	0.000	0.292	0.727
ageCent:currentSmoker1	-0.015	0.012	-1.241	0.215	-0.039	0.009

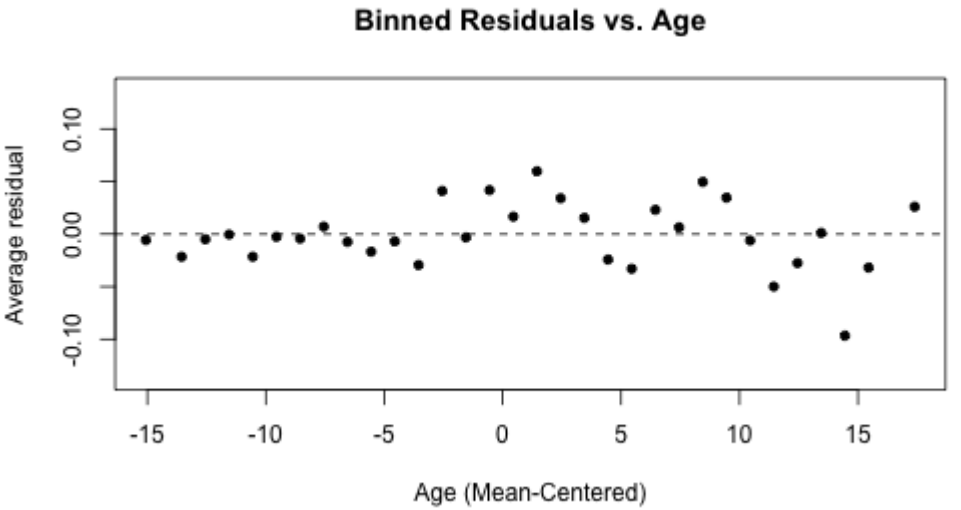
1. What is the effect of age on the odds of being at risk of heart disease for smokers?
2. What is the effect of age on the odds of being at risk of heart disease for non-smokers?
3. Is the effect of age on being at risk of heart disease significantly different for the two groups?

Binned residuals



Residuals vs. quantitative predictor

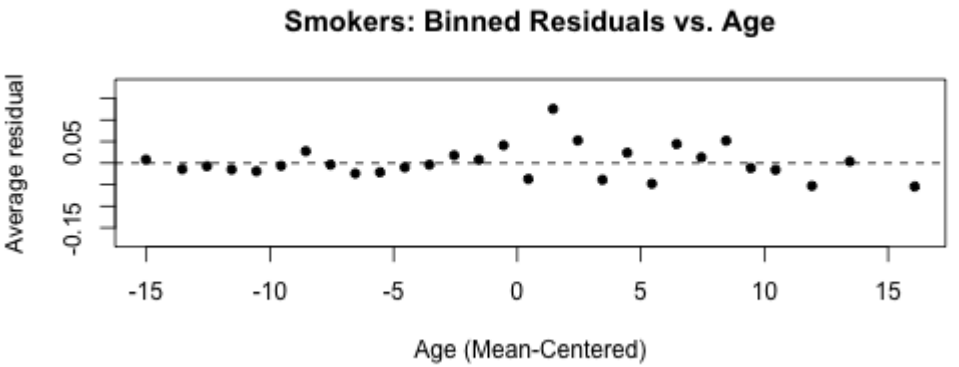
```
arm::binnedplot(x=risk_m_int_aug$ageCent,y=risk_m_int_aug$.resid,›
```

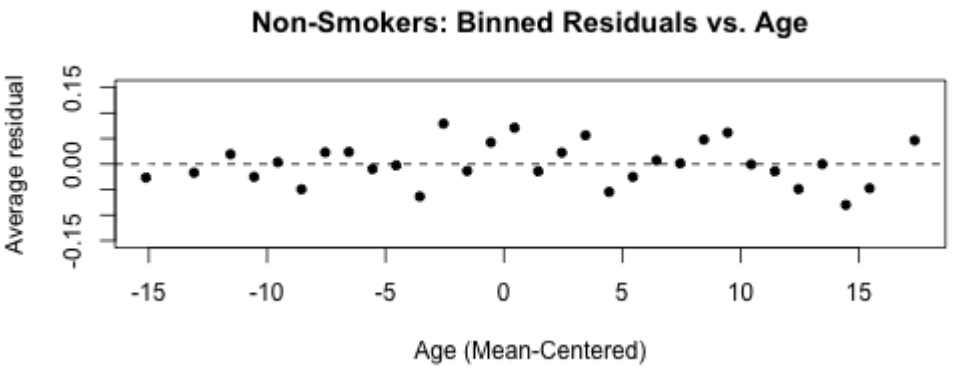


Residuals vs. categorical predictor

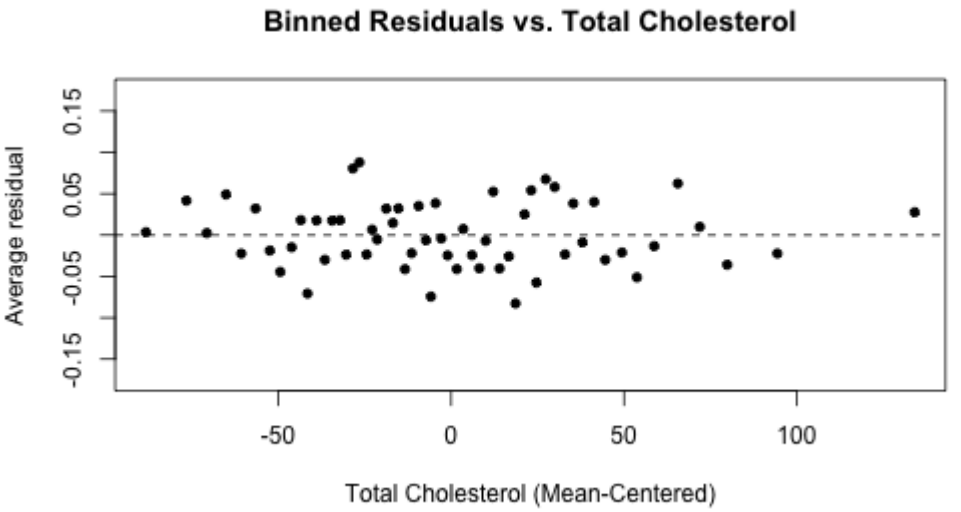
```
## # A tibble: 2 x 2
##   currentSmoker mean_resid
##   <fct>              <dbl>
## 1 0              -2.99e-12
## 2 1              -1.02e-14
```

Binned Residuals vs. Age: Smokers vs. Non-Smokers





Binned Residuals vs. total cholesterol



Exploratory Data Analysis

Exploratory Data Analysis

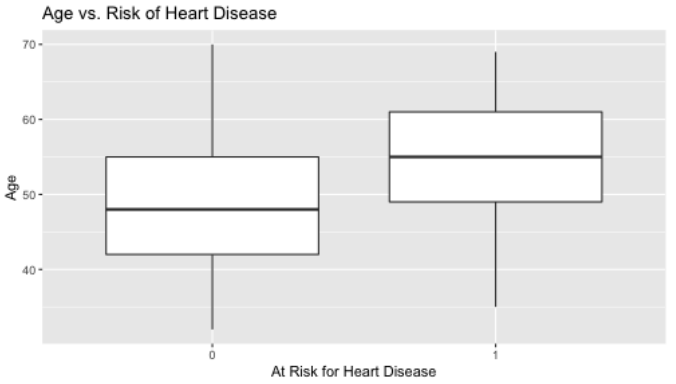
Categorical predictors:

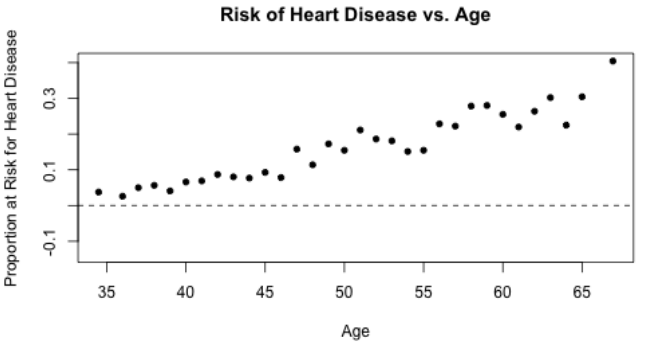
- Examine the percentage of $y = 1$ for each level (category)
- You can visualize using a stacked bar chart

Quantitative predictors:

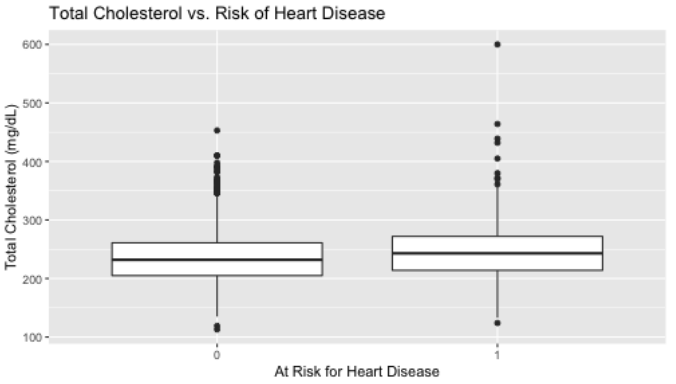
- Use side-by-side boxplots to examine the distribution of the predictor for each level of the response
- Use binned plots to examine how the probability of $y = 1$ changes as the predictor increases

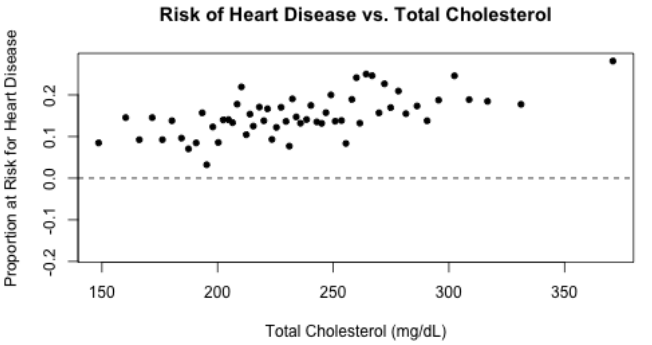
EDA: `TenYearCHD` vs. Age





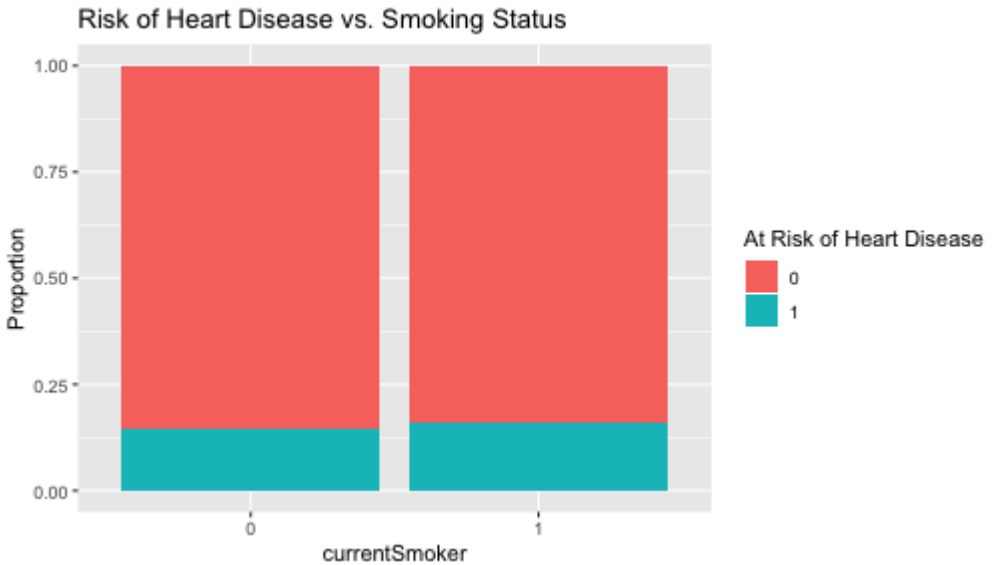
EDA: `TenYearCHD` vs. `totChol`





TenYearCHD vs. currentSmoker

```
ggplot(data = heart_data, aes(x = currentSmoker, fill = as.factor(
  geom_bar(position = "fill") +
  labs(y = "Proportion",
       fill = "At Risk of Heart Disease",
       title = "Risk of Heart Disease vs. Smoking Status")
```



Drop-in-deviance test

Comparing Nested Models

- Suppose there are two models:
  - Model 1 includes predictors $x_1, \ldots, x_q$
  - Model 2 includes predictors $x_1, \ldots, x_q, x_{q+1}, \ldots, x_p$
- We want to test the hypotheses

$$H_0 : \beta_{q+1} = \cdots = \beta_p = 0$$
$$H_a : \text{ at least 1 } \beta_j \text{ is not} 0$$

- We used a *Nested F Test* to compare two nested models in linear regression
- We will use the **drop-in-deviance test** in logistic regression

Deviance residual

- The **deviance residual** is the a measure of how much the observed data differs from what is measured using the likelihood ratio
- The deviance residual for the i^{th} observation is

$$d_i = \text{sign}(Y_i - \hat{p}_i)\sqrt{2\left[Y_i \log\left(\frac{Y_i}{p_i}\right) + (1 - Y_i)\log\left(\frac{1 - Y_i}{1 - \hat{p}_i}\right)\right]}$$

where $\text{sign}(Y_i - \hat{p}_i)$ is positive when $Y_1 = 1$ and negative when $Y_i = 0$.

Drop-in-Deviance Test

- The **deviance statistic** for Model k is $D_k = \sum_{i=1}^{n} d_i^2$
- To test the hypotheses

$$H_0 : \beta_{q+1} = \cdots = \beta_p = 0$$
$$H_a : \text{ at least 1 } \beta_j \text{ is not0}$$

the drop-in-deviance statistic is $D_1 - D_2$

- When the sample size is large, the drop-in-deviance statistic has an approximately Chi-squared distribution with degrees of freedom equal to the difference in number of predictor variables in Model 1 and Model 2

Should we add `Education` to the model?

- Suppose
  - Model 1 includes `AgeCent`, `currentSmoker`, `totCholCent`
  - Model 2 includes `AgeCent`, `currentSmoker`, `totCholCent`, `education` (categorical)

```
model1 <- glm(TenYearCHD ~ ageCent + currentSmoker + totChol,
              data = heart_data, family = binomial)
model2 <- glm(TenYearCHD ~ ageCent + currentSmoker + totChol +
                as.factor(education),
              data = heart_data, family = binomial)
```

```
# Deviances
(dev_model1 <- glance(model1)$deviance)
```

```
## [1] 2894.989
```

```
(dev_model2 <- glance(model2)$deviance)
```

```
## [1] 2887.206
```

Should we add `education` to the model?

```
# Drop-in-deviance test statistic
(test_stat <- dev_model1 - dev_model2)
```

```
## [1] 7.783615
```

```
# p-value
1 - pchisq(test_stat, 3) #3 = number of new model terms in model2
```

```
## [1] 0.05070196
```

Should we add `Education` to the model?

- We can use the **anova** function to conduct this test
  - Add **`test = "Chisq"`** to conduct the drop-in-deviance test

```
anova(model1, model2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: TenYearCHD ~ ageCent + currentSmoker + totChol
## Model 2: TenYearCHD ~ ageCent + currentSmoker + totChol + as.factor(educ
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3654     2895.0
## 2      3651     2887.2  3   7.7836   0.0507 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appex: Framingham Study

- We will analyze data from a cardiovascular study on residents in Framingham, MA
- **Goal**: Predict whether or not a participant has a 10-year risk of future coronary heart disease
- Original data contains information from 4,000+ participants. We will use 500 for this analysis.