27.4: Extending Regression to Binary Outcomes.

Let's say that we have a blood test (which is often referred to as a *biomarker*) and we want to know whether it predicts who is going to have a heart attack within the next year. We will generate a synthetic dataset for a population that is at very high risk for a heart attack in the next year.

```
# sample size
npatients=1000

# probability of heart attack
0.5p_heartattack =

# true relation to biomarker
0.6true_effect <-

# assume biomarker is normally distributed
disease_df <- tibble(biomarker=rnorm(npatients))

# generate another variable that reflects risk for
# heart attack, which is related to the biomarker
disease_df <- disease_df %>%
  mutate(risk = biomarker*true_effect + rnorm(npatients))

# create another variable that shows who has a
# heart attack, based on the risk variable
disease_df <- disease_df %>%
  mutate(
    heartattack = risk > quantile(disease_df$risk,
                                  1-p_heartattack))

glimpse(disease_df)
```

```
## Observations: 1,000
## Variables: 3
## $ biomarker   <dbl> 1.15, 0.68, 1.21, -0.72, -1.00, -0.12…
## $ risk        <dbl> 1.054, -0.529, 0.675, -0.474, -1.398,…
## $ heartattack <lgl> TRUE, FALSE, TRUE, FALSE, FALSE, TRUE…
```

Now we would like to build a model that allows us to predict who will have a heart attack from these data. However, you may have noticed that the heartattack variable is a binary variable; because linear regression assumes that the residuals from the model will be normally distributed, and the binary nature of the data will violate this, we instead need to use a different kind of model, known as a *logistic regression* model, which is built to deal with binary outcomes. We can fit this model using the `glm()` function:

```
glm_result <- glm(heartattack ~ biomarker, data=disease_df,
                  family=binomial())
summary(glm_result)
```

```
##
## Call:
## glm(formula = heartattack ~ biomarker, family = binomial(), data = disease_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1301  -1.0150   0.0305   1.0049   2.1319
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.00412    0.06948   -0.06     0.95
## biomarker    0.99637    0.08342   11.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1386.3  on 999  degrees of freedom
## Residual deviance: 1201.4  on 998  degrees of freedom
## AIC: 1205
##
## Number of Fisher Scoring iterations: 3
```

This looks very similar to the output from the `lm()` function, and it shows us that there is a significant relationship between the biomarker and heart attacks. The model provides us with a predicted probability that each individual will have a heart attack; if this is greater than 0.5, then that means that the model predicts that the individual is more likely than not to have a heart attack.
We can start by simply comparing those predictions to the actual outcomes.

```
# add predictions to data frame
disease_df <- disease_df %>%
0.5  mutate(prediction = glm_result$fitted.values>,
         heartattack = heartattack)

# create table comparing predicted to actual outcomes
CrossTable(disease_df$prediction,
           disease_df$heartattack,
           prop.t=FALSE,
           prop.r=FALSE,
           prop.chisq=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  1000
##
##
##                      | disease_df$heartattack
## disease_df$prediction |     FALSE |      TRUE | Row Total |
## ----------------------|-----------|-----------|-----------|
##                 FALSE |       332 |       157 |       489 |
##                       |     0.664 |     0.314 |           |
## ----------------------|-----------|-----------|-----------|
##                  TRUE |       168 |       343 |       511 |
##                       |     0.336 |     0.686 |           |
## ----------------------|-----------|-----------|-----------|
##          Column Total |       500 |       500 |      1000 |
##                       |     0.500 |     0.500 |           |
## ----------------------|-----------|-----------|-----------|
##
##
```

This shows us that of the 500 people who had heart attacks, the model corrected predicted a heart attack for 343 of them. It also predicted heart attacks for 168 people who didn't have them, and it failed to predict a heart attack for 157 people who had them. This highlights the distinction that we mentioned before between statistical and practical significance; even though the biomarker shows a highly significant relationship to heart attacks, it's ability to predict them is still relatively poor. As we will see below, it gets even worse when we try to generalize this to a new group of people.