Logistic Regression

Linear regression is designed for a *quantitative* response variable; in the model equation

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

the random noise term ϵ is usually assumed to be at least approximately Gaussian.

When the response Y is the *indicator* of success versus failure in some experiment with just those two outcomes, that model is inappropriate.

Example: Semiconductor manufacturing

A silicon wafer is cut into many dice, and each die is classified as acceptable or defective. The probability of being defective is found to vary with the radial distance from the center of the wafer.

Response:

$$Y = \begin{cases} 1 & \text{if the die is defective} \\ 0 & \text{if the die is acceptable} \end{cases}$$

Predictor:

$$x = \text{radial distance from center}$$

The predictor x determines the probability of success:

$$P(Y = 1) = \beta(x) \quad \text{for some function } \beta(x),$$

and

$$\begin{aligned} P(Y = 0) &= 1 - P(Y = 1) \\ &= 1 - \beta(x). \end{aligned}$$

Then

$$E(Y) = P(Y = 1) = \beta(x),$$

and we could write

$$Y = \beta(x) + \epsilon$$

with

$$E(\epsilon) = 0.$$

If in addition

$$\beta(x) = \beta_0 + \beta_1 x,$$

then we have

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

but ϵ is *not* Gaussian and does *not* have constant variance.

We could use least squares to fit the model anyway; however, the model itself is inappropriate, because for some x it gives "probabilities" that are either negative or greater than 1.

The issue is that we are modeling $P(Y = 1)$, which must lie between 0 and 1.

We could instead model the *odds ratio*

$$\frac{P(Y = 1)}{P(Y = 0)} = \frac{\beta(x)}{1 - \beta(x)},$$

which can take any positive value, or its logarithm

$$\log \frac{\beta(x)}{1 - \beta(x)},$$

which can take any value, either positive or negative.

Logistic regression

In the logistic regression model, we assume that

$$\log \frac{\beta(x)}{1-\beta(x)} = \beta_0 + \beta_1 x.$$

Equivalently, if we solve for $\beta(x)$:

$$\beta(x) = P(Y=1) = \frac{\exp(\beta_0 + \beta_1 x)}{1+\exp(\beta_0 + \beta_1 x)}.$$

In R:

```
b0 <- 0; b1 <- 1;
curve(exp(b0 + b1 * x) / (1 + exp(b0 + b1 * x)), -5, 5)
```

Example: Space shuttle O-rings

In January, 1986, Space Shuttle *Challenger* was destroyed when an O-ring seal in its right solid rocket booster failed.

In 24 prior launches, O-rings had been damaged in 7 launches at various temperatures:

```
oRing <- read.csv("Data/o-ring.csv");
plot(oRing, xlim = c(30, 85));
abline(v = 31, col = "red") # Launch temperature
```

We can fit the logistic regression model using the R function `glm()`, which handles this and several other models; because the response Y is a Bernoulli random variable, which is a special case of the binomial random variable, we use `family = binomial`:

```
oRingGlm <- glm(Failure ~ Temperature, oRing, family = binomial);
summary(oRingGlm)
```

The output shows that $\hat{\beta}_0 = 10.87535$, and $\hat{\beta}_1 = -0.17132$; to test $H_0 : \beta_1 = 0$, use the z-statistic, and note that the associated P-value is 0.0400.

That is, the risk of failure has a moderately significant dependence on temperature, with lower temperatures increasing the risk.

Estimated probability of failure:

```
curve(predict(oRingGlm, data.frame(Temperature = x),
              type = "response"),
      from = 30, to = 85, add = TRUE)
```

Adding confidence intervals is more work, but necessary:

```
x <- seq(from = 30, to = 85, length = 100);
oRingPred <- predict(oRingGlm, data.frame(Temperature = x),
                     se.fit = TRUE);
y <- oRingPred$fit + oRingPred$se.fit %o% qnorm(c(0.025, 0.975));
matlines(x, exp(y) / (1 + exp(y)) , lty = 2, col = "blue")
```

The logistic regression model can include more than one predictor:

$$\log \frac{\beta(x)}{1-\beta(x)} = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \cdots + \beta_k x_k.$$

Challenger again

A different data set includes information about a pressure:

```
challenger <- read.csv("Data/challenger.csv");
challenger$Y <- challenger$distress_ct > 0;
summary(glm(Y ~ temperature + pressure, challenger,
            family = binomial))
```

Poisson Regression

This second data set includes the *number* of O-rings that were damaged, with values 0, 1, and 2.

We might want to model that count as a Poisson random variable, again with expected value as a function of temperature:

$$E(Y) = \beta(x).$$

In this case, the only constraint on $\beta(x)$ is that it should be positive, and the usual model is

$$\log[\beta(x)] = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \cdots + \beta_k x_k.$$

The same function `glm()` can be used, with `family = poisson`:

```
challengerGlm <- glm(distress_ct ~ temperature, challenger,
                     family = poisson);
summary(challengerGlm)
```

This Poisson regression model offers another way to estimate the probability of any O-ring failures (using only temperature, as pressure is not significant):

$$\begin{aligned} P(Y \geq 1) &= 1 - P(Y = 0) \\ &= 1 - e^{-\beta(x)} \\ &= 1 - \exp(-\exp(\beta_0 + \beta_1 x)) \end{aligned}$$

```
plot(Y ~ temperature, challenger, xlim = c(30, 85));
curve(1 - exp(-predict(challengerGlm,
                       data.frame(temperature = x),
                       type = "response")),
      add = TRUE)
# 95% confidence intervals:
challengerPred <- predict(challengerGlm,
                          data.frame(temperature = x),
                          se.fit = TRUE);
y <- challengerPred$fit +
            challengerPred$se.fit %o% qnorm(c(0.025, 0.975));
matlines(x, 1 - exp(-exp(y)) , lty = 2, col = "blue")
```

Example: Logistic regression in a Designed Experiment

Cut roses are susceptible to wilting caused by a fungus, but development of the fungus can be inhibited by treatment with ethylene.

Different cultivars have varying susceptibility to the fungus; some are also damaged by the ethylene treatment.

Designed experiment

Response: $Y = 1$ if the rose's quality is unacceptable, $Y = 0$ if acceptable;

Factors:
- Cultivar, with 4 levels;
- Treatment, with 2 levels (treated or not treated).

Replication: 10 replicates.

Model:

$$\log\left[\frac{P(Y_{i,j,k}=1)}{1-P(Y_{i,j,k}=1)}\right] = \mu + \tau_i + \beta_j + (\tau\beta)_{i,j}$$

When the interactions $(\tau\beta)_{i,j}$ are significant, the cultivars have varying response to the ethylene treatment.

The `predict` method can be used to assess the best strategy for each cultivar, and which cultivar will suffer the least damage.