Statistics for Health Economics

Andrea Tancredi

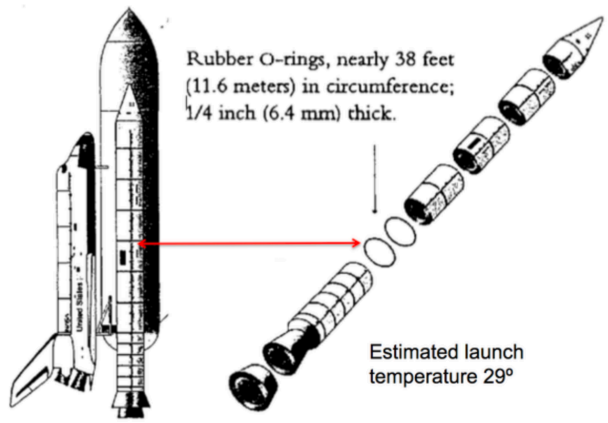
Sapienza University of Rome

Lectures 19 and 20: Regression 3

Logistic regression model

- A binary response Y_i takes values 1 and 0 with probabilities π_i and $1 - \pi_i$, denoting a dichotomous outcome such as success/failure, won/lost, or well/ill.
- Such data are common in applications. The simplest relation for $E(Y_i) = \pi_i$ is $\pi_i = \beta_0 + \sum_{j=1}^{p} X_{ji}\beta_j$. This is unsuitable for general use because π_i may then lie outside the unit interval.
- It is usually better to force $0 < \pi_i < 1$ by taking it be a nonlinear monotone function of $\beta_0 + \sum_{j=1}^{p} X_{ji}\beta_j$
- When $Y_i \sim Binomial(N_i, \pi_i)$ with N_i known similar considerations apply since Y_i represent the number of successes in N_i independent trials with success probability π_i

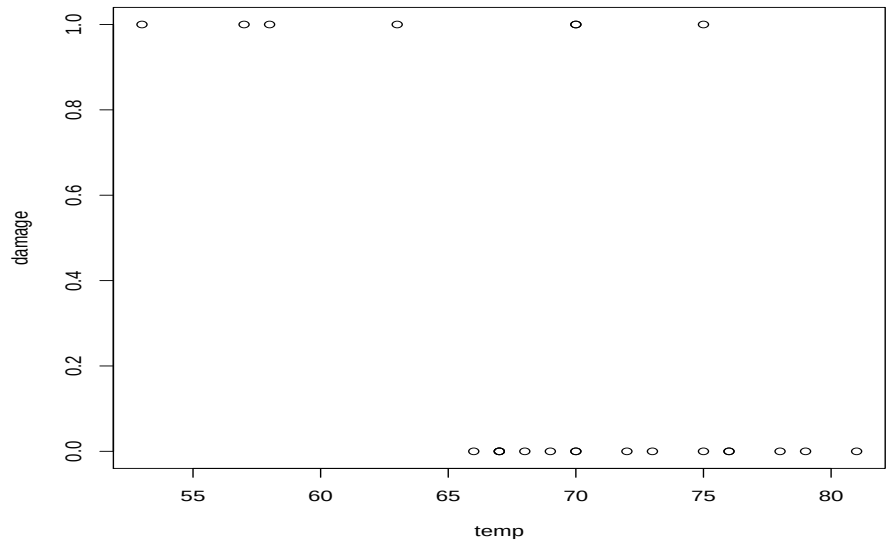
- **Shuttle data** On January 28, 1986 America was shocked by the destruction of the space shuttle Challenger, and the death of its seven crew members
- Up until 1986 the space shuttle was lifted into space by a pair of booster rockets, that were comprised of four sections stacked vertically on top of each other. The joints between the sections were sealed by O- rings.



- On January 28, 1986 the temperature at launch time was so cold that the O- rings became brittle and failed to seal the joints, allowing hot exhaust gas to come into contact with unburned fuel.
- One of the issues was whether NASA could or should have forseen that cold weather might diminish performance of the O-rings.

- The plot of Y = presence of damage against X = temperature for the launches prior to the Challenger accident suggests that colder launches are more likely to have damaged O-rings.

```
temp<-c(66,70,69,68,67,72,73,70,57,63,70,78,67,53,67,75,70,81,76,
79,75,76,58)
damage<-c(0,1,0,0,0,0,0,0,1,1,1,0,0,1,0,0,0,0,0,0,1,0,1)
plot(y=damage,x=temp)
```



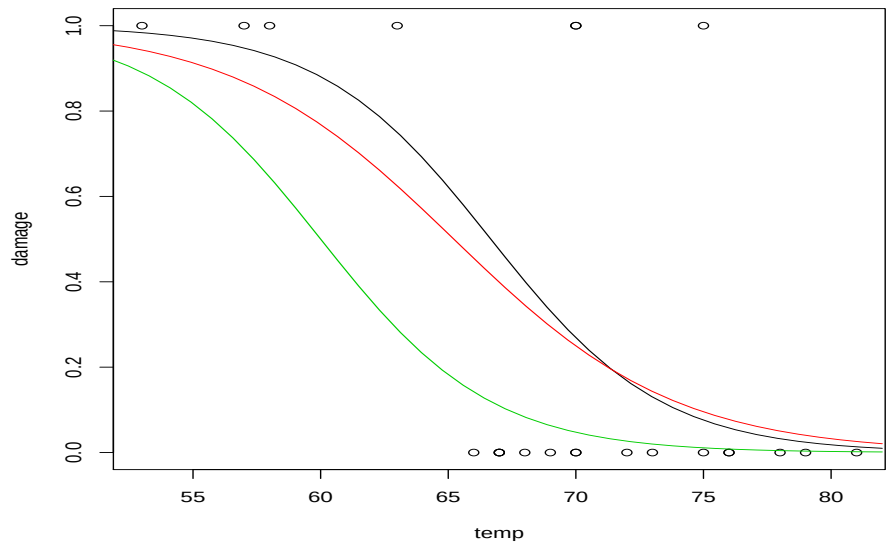
- We want model for damage probability as a function of temperature

- The most commonly adopted model in such situations is

$$E[Y|X] = P(Y = 1|X) = \pi(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- For example by taking the following values for β_0 and β_1 (20,-0.3), (15,-0.23) (18,-0.3) we have the following functions for π

```
b0=c(20, 15, 18); b1=c ( -.3, -.23, -.3 )
x=seq ( 50, 82, length=40 )
plot(y=damage,x=temp)
for ( i in 1:3 ) lines (x,
          exp(b0[i] + b1[i]*x)/(1 + exp(b0[i] + b1[i]*x) ), col=i )
```



- The previous model is known as logistic regression model. Let the $i'th$ observation have covariate x_i and probability of success $\theta_i = E[Y_i|x_i]$. Define

$$\phi_i = \log\left(\frac{\theta_i}{1-\theta_i}\right)$$

ϕ_i is called the logit of θ_i. The inverse transformation is

$$\theta_i = \frac{e^{\phi_i}}{1+e^{\phi_i}}$$

The logistic regression model assumes

$$\phi_i = \beta_0 + \beta_1 x_i$$

and more generally

$$\phi_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

- This model is a generalized linear model or glm because it is a linear model for ϕ, a transformation of $E(Y|x)$ rather than for $E(Y|x)$
- The quantity $\beta_0 + \beta_1 x$ is called thel inear predictor.
- If $\beta_1 > 0$, then as $x \to \infty$ $\theta \to 1$ and as $x \to -\infty$ $\theta \to 0$. If $\beta_1 < 0$ the situation is reversed.
- β_0 is like an intercept; it controls how far to the left or right the curve is. β_1 is like a slope; it controls how quickly the curve moves between its two asymptotes
- Logistic regression and, indeed, all generalized linear models differ from linear regression in two ways: the regression function is nonlinear and the distribution of $Y|x$ is not Normal. Hence the likelihood function is different

- When we have only one covariate the likelihood function is

$$
\begin{aligned}
p(y_1, \ldots, y_n | x_1, \ldots, x_n, \beta_0, \beta_1) &= \prod_{i=1}^{n} p(Y_i | x_i, \beta_0, \beta_1) = \prod_{i=1}^{n} \theta_i^{y_i} (1 - \theta_i)^{1 - y_i} \\
&= \prod_{i=1:y_i=1} \theta_i \prod_{i=1:y_i=0} (1 - \theta_i) \\
&= \prod_{i=1:y_i=1} \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \prod_{i=1:y_i=0} \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}}
\end{aligned}
$$

- R will maximise this likelihood via numerical techniques, even with more covariates (independent variables) x

- The command to fit the logistic regression model is glm

```
m=glm(damage~temp,family=binomial)
summary(m)$coeff

##                Estimate Std. Error   z value   Pr(>|z|)
## (Intercept) 15.0429016  7.3786301  2.038712 0.04147878
## temp        -0.2321627  0.1082364 -2.144959 0.03195610
```

- Note that the estimate of β_1 is negative, hence with low value of temperature the damage probability increases. Morever the test for rejecting $H_0 : \beta_1 = 0$ has a small p-value (0.03), then we can reject the hypothesis of independence with respect to the temperature.
- Finally note that the estimated O-rings damage probability at 37 F degrees (when the shuttle exploded) is

```
exp(sum(coef(m)*c(1,37)))/(1+exp(sum(coef(m)*c(1,37))))

## [1] 0.9984265
```

- Note that when x_i also is a binary (0,1) variable we have that

$$\theta_0 = P(Y_i = 1|x_i = 0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

$$1 - \theta_0 = P(Y_i = 0|x_i = 0) = \frac{1}{1 + e^{\beta_0}}$$

$$\frac{\theta_0}{(1 - \theta_0)} = e^{\beta_0}$$

$$\log \frac{\theta_0}{(1 - \theta_0)} = \beta_0$$

- Note that when x_i also is a binary (0,1) variable we have that

$$\theta_1 = P(Y_i = 1|x_i = 1) = \frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}}$$

$$1 - \theta_1 = P(Y_i = 0|x_i = 1) = \frac{1}{1 + e^{\beta_0+\beta_1}}$$

$$\frac{\theta_1}{(1 - \theta_1)} = e^{\beta_0+\beta_1}$$

$$\log \frac{\theta_1}{(1 - \theta_1)} = \beta_0 + \beta_1$$

$$\beta_1 = \log \frac{\theta_1}{(1 - \theta_1)} - log \frac{\theta_0}{(1 - \theta_0)} = log\psi$$

where

$$\psi = \left(\frac{\theta_1}{(1 - \theta_1)}\right) : \left(\frac{\theta_0}{(1 - \theta_0)}\right)$$

```
HDR=read.csv("HDR_sample2.csv",sep=";",dec=",")
diabetes=matrix(,nrow=nrow(HDR),ncol=6)
for (i in 1:6) {
diabetes[,i]=substr(as.character(HDR[,(25+i-1)]),1,3)=='250'}
diabetes01=as.numeric(apply(diabetes,FUN=any,MAR=1))
HDR=cbind(HDR,diabetes=diabetes01)
rm(diabetes)
attach(HDR)
table(diabetes,gender)
```

```
##          gender
## diabetes Females Males
##        0   26985 16608
##        1     634   743
```

```r
m=glm(diabetes~gender,family=binomial)
summary(m)$coef
```

```
##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -3.7509875 0.04017805 -93.35913 0.000000e+00
## genderMales  0.6440437 0.05495801  11.71883 1.020675e-31
```

```r
exp(coef(m)[1])
```

```
## (Intercept)
##  0.02349453
```

```r
exp(coef(m)[2])
```

```
## genderMales
##    1.904165
```

```r
m=glm(diabetes~gender+age,family=binomial)
summary(m)$coef
```

```
##                Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept) -4.5583099 0.069209212 -65.862762 0.000000e+00
## genderMales  0.5309629 0.055889677   9.500196 2.094955e-21
## age          0.0177332 0.001125329  15.758234 6.029425e-56
```

```r
exp(coef(m)[1])
```

```
## (Intercept)
##  0.01047976
```

```r
exp(coef(m)[2])
```

```
## genderMales
##    1.700569
```

```r
m=glm(diabetes~gender+age_cl,family=binomial)
summary(m)$coef
```

```
##                Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) -3.5977816 0.06389735 -56.305645 0.000000e+00
## genderMales  0.2817446 0.05753151   4.897223 9.720067e-07
## age_cl25-44 -1.0847161 0.11022122  -9.841263 7.476622e-23
## age_cl45-64 -0.2192634 0.08492182  -2.581944 9.824551e-03
## age_cl65-74  0.1909121 0.08963068   2.129986 3.317277e-02
## age_cl75+    0.9960778 0.07332754  13.583953 4.986124e-42
```

```r
exp(coef(m)[1])
```

```
## (Intercept)
##  0.02738441
```

```r
exp(coef(m)[2])
```

```
## genderMales
##     1.32544
```