Logistic Regression

Description of Data

A survey of workers in the US cotton industry examined whether or not an individual suffered from a specific lung disease. The value of the following five variables were also recorder:

- race:ethnic group of worker (1=white, 2=other)
- sex (1=male, 2=female)
- smoking status (1=smoker, 2=non–smoker)
- employment length (1= less than 10 years, 2=10–22 years, 3= over 20 years)
- dust: amount of dust in the workspace (1=high, 2=mediioum 3=low)

The data are available in the web page of the class (`logistic1.txt`).

The problem for these data is to assess the significance of explanatory variables. That is, which of these variables are predictive and whether or not a worker suffers from the lung disease. Since the response variable is binary we will be using the logistic regression model for the analysis.

Logistic Regression

Instead of using a linear model for the dependence of probability of suffering from the lung disease on the explanatory variables, we use the logistic transformation which is defined as

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p. \tag{1}$$

Here p is the probability of suffering from the lung disease and the number of explanatory variables equals to p. The regression parameters are estimated by the method of maximum likelihood where the response variable is assumed to have the binomial distribution. Notice that (1 can be rewritten as

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)} \tag{2}$$

Analysis Using S-Plus

The data are in the form of counts of disease cases and counts of workers without the disease with the corresponding values of the explanatory variables. They are available as a data frame in S-Plus called `logreg`. Estimates of the parameters in the logistic regression model can be obtained by the following:

```
> attach(logreg)
> glm( cbind(Yes, No)~dust+race+sex+smoking+Empleng, family=binomial) -> out1
> out1
Call:  glm(formula = cbind(Yes, No) ~ dust + race + sex + smoking +Empleng,
           family = binomial)

Coefficients:
(Intercept)         dust         race          sex      smoking      Empleng
    -0.4852      -1.3751       0.2463      -0.2590      -0.6292       0.3856

Degrees of Freedom: 64 Total (i.e. Null);  59 Residual
Null Deviance:      322.5
Residual Deviance: 69.51
```

These are the parameter estimates together with the null deviance, the residual deviance. To get a complete summary of the estimates together with their standard errors use

```
> summary(out1)

Call:
glm(formula = cbind(Yes, No) ~ dust + race + sex + smoking +
    Empleng, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4126  -0.7573  -0.2421   0.3688   1.9804

Coefficients:
            Estimate Std. Error     t     p-value
(Intercept)  -0.4852     0.6059  -0.801  0.42326
dust         -1.3751     0.1155 -11.903  < 2e-16
race          0.2463     0.2061   1.195  0.23198
sex          -0.2590     0.2116  -1.224  0.22086
smoking      -0.6292     0.1930  -3.259  0.00112
Empleng       0.3856     0.1069   3.607  0.00031

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 322.527  on 64  degrees of freedom
```

```
Residual deviance:  69.509  on 59  degrees of freedom

Number of Fisher Scoring iterations: 4
```

The above results show that variables `dust`, `smoking`, `Empleng`, are the most significant in the prognosis of lung disease. We can fit a model with these regressors and then compare it with the full model:

```
> glm( cbind(Yes, No)~dust+smoking+Empleng, family=binomial) -> out2
> anova(out2,out1)
Analysis of Deviance Table

Model 1: cbind(Yes, No) ~ dust + smoking + Empleng
Model 2: cbind(Yes, No) ~ dust + race + sex + smoking + Empleng
  Resid. Df Resid. Dev Df Deviance
1        61     72.562
2        59     69.509  2    3.053
> 1-pchisq(3.053, df=2)
[1] 0.2172949
> summary(out2)

Call:
glm(formula = cbind(Yes, No) ~ dust + smoking + Empleng, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3421  -0.7700  -0.2518   0.4001   2.0523

Coefficients:
            Estimate Std. Error      t   p-value
(Intercept)  -0.1418     0.3412  -0.416 0.677742
dust         -1.4657     0.1058 -13.859  < 2e-16
smoking      -0.6778     0.1887  -3.592 0.000328
Empleng       0.3331     0.0886   3.760 0.000170

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 322.527  on 64  degrees of freedom
Residual deviance:  72.562  on 61  degrees of freedom


Number of Fisher Scoring iterations: 4
```

Object `out2` contains the results of the reduced model whose results are summarized above. Accordingly, the estimated logistic model is

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.1418 - 1.1657 \times \text{dust} - 0.6778 \times \text{smoking} + 0.3331 \times \text{Empleng}$$

and it is possible to calculate the predicted value of the probability of suffering from the lung disease for any combination of values of the three explanatory variables. For instance, if a worker has `dust=1`, `smoking=1` and `Empleng=3`, the equation above yields $\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.286$, so that $\hat{p} = 0.2165$.

To get the Pearson and deviance residuals, use

```
> residuals(out2, type="d")
 [1] -1.01112780 -2.18492321  0.03337703  1.05930412 -2.38265770  0.78599964 -1.15218086
 [8] -1.35054640  1.20244805 -0.64235459 -1.41172893  0.67924262 -1.49239389 -1.07494660
[15] -1.01392725  0.24210289  0.23378143  0.67100821 -0.74619694  0.10394677  1.33677030
[22] -0.58548730  0.39372083  2.05225729  1.49151960 -0.94215661 -0.76417711  0.72544147
[29] -0.66877675 -0.83273916  0.00000000 -0.43343338  0.88110306  0.00000000 -0.59817211
[36] -0.25108030  1.04689402  0.83623897 -0.78766678  0.09325768  0.00000000 -0.27363872
[43]  0.00000000 -1.17367866  0.64751569  0.00000000 -0.42856685 -0.20685143  1.71825626
[50] -3.34213290  1.57828278  0.41914800 -0.35176829 -1.14748789 -0.69863437 -1.25301162
[57]  0.24410437 -0.69863437  0.00000000 -0.24191169 -0.61128692 -1.57719376  1.20318632
[64]  0.53824349 -0.25255313 -0.58560121 -0.72477262 -1.36106299 -0.35333090  0.00000000
[71] -0.35716405 -0.21149430
> residuals(out2, type="pear")
 [1] -0.94535295 -1.55751682  0.03350811  1.09147989 -1.69847134  0.85775982 -0.84251658
 [8] -1.16707986  1.39123857 -0.60853457 -1.25426495  0.73131825 -1.07391312 -0.76324965
[15] -0.71764295  0.24581171  0.24358032  0.77245841 -0.53695656  0.10577991  1.66583158
[22] -0.54375653  0.40972225  2.62599096  1.60218010 -0.84268248 -0.69200923  0.74984267
[29] -0.47823390 -0.59038560  0.00000000 -0.40853032  1.00017959  0.00000000 -0.42774540
[36] -0.17800796  1.19196118  1.00011214 -0.55771007  0.09454214  0.00000000 -0.19375080
[43]  0.00000000 -0.83470245  0.74057667  0.00000000 -0.30479024 -0.14646184  1.77849577
[50] -2.65712373  1.72113934  0.42524859 -0.25263518 -0.81437328 -0.52573924 -1.14642169
[57]  0.25005055 -0.52573924  0.00000000 -0.17168496 -0.59074676 -1.12419548  1.39228504
[64]  0.56308501 -0.18001535 -0.41485553 -0.52978676 -1.23313725 -0.34015947  0.00000000
[71] -0.25458014 -0.14982821
> plot(residuals(out2, type="d"), xlab="Index", ylab="Deviance Residuals")
> abline(h=0)   # get Figure 1
> plot(residuals(out2, type="pear"), xlab="Index", ylab="Pearson Residuals")
> abline(h=0)   # get Figure 2
```

The plots of the deviance and Pearson residuals are shown in Figures 1 and 2. Both plots show that observation number 50 is rather problematic. The negative value of the residual indicates that the predicted value of suffering from the lung disease for this observation is larger than the observed value. Looking at the data, observation 50 has `dust=2`, `smoking=1` and `Empleng=3` so that $\hat{p} = 0.059$. The data estimated probability is $1/141 = 0.007$.

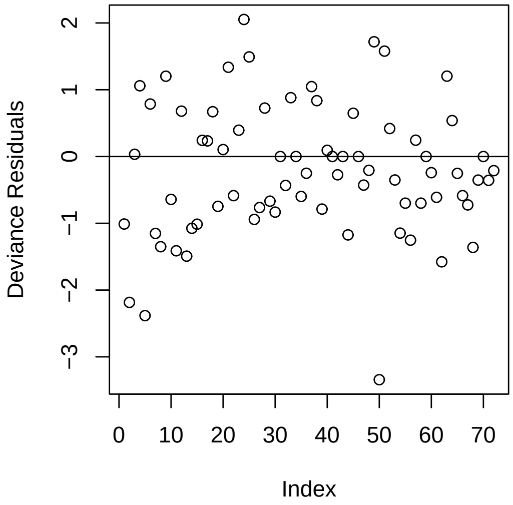

Figure 1: Deviance Residuals.

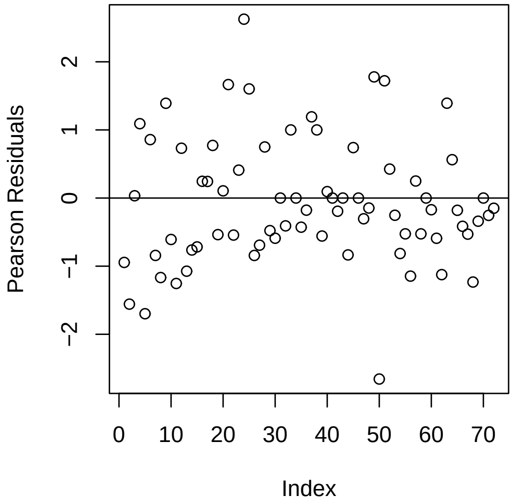

Figure 2: Pearson Residuals.

Fitting a Probit Model

The same analysis can be carried over using a probit model

$$p = \Phi(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p),$$

where Φ stands for the cdf of the standard normal distribution.

```
> glm( cbind(Yes, No)~dust+smoking+Empleng, family=binomial(link=probit)) -> out3
> out3

Call:  glm(formula = cbind(Yes, No) ~ dust + smoking + Empleng,
           family = binomial(link = probit))
Coefficients:
(Intercept)         dust      smoking      Empleng
    -0.4044      -0.6268      -0.2840       0.1406

Degrees of Freedom: 64 Total (i.e. Null);  61 Residual
Null Deviance:      322.5
Residual Deviance: 84.59
> summary(out3)

Call:
glm(formula = cbind(Yes, No) ~ dust + smoking + Empleng, family = binomial(link = probit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5085  -0.7912  -0.2626   0.2894   2.5516

Coefficients:
            Estimate Std. Error   t      p-value
(Intercept) -0.40440    0.15877  -2.547 0.010864
dust        -0.62684    0.04633 -13.531  < 2e-16
smoking     -0.28396    0.08214  -3.457 0.000546
Empleng      0.14064    0.04056   3.468 0.000525


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 322.527  on 64  degrees of freedom
Residual deviance:  84.587  on 61  degrees of freedom

Number of Fisher Scoring iterations: 4
```