Generalized Linear Models

Objectives

- GLM and LINK
- Binomial Cases
- Model Selection
- Count Data
- Survival Analysis

GLM: Variance and Link Families

Generalized linear models are fit using the `glm( )` function. The form of the `glm` function is shown below. As long as you want the default link, you only have to specify the family name. If you want an alternative link, then you must write a link argument.

```
glm(formula, family=familytype(link="linkfunction"), data= )
```

family	default `link` function	other choices
binomial	`family=binomial(link = "logit")`	probit or cloglog
gaussian	`family=gaussian(link = "identity")`	
Gamma	`family=Gamma(link = "inverse")`	identity or log
inverse.gaussian	`inverse.gaussian(link = "1/mu^2")`	
poisson	`poisson(link = "log")`	identity or sqrt
quasibinomial	`quasibinomial(link = "logit")`	probit or cloglog
quasipoisson	`quasipoisson(link = "log")`	identity or sqrt

- `quasibinomial` and `quasipoisson` are used in case of *overdispersion.*
  - Binomial with the (default) "logit" link is the *logistic* regression.

GLM begins with LINK function

Suppose, for example, the response variable Y_i $(i = 1, 2, \ldots, n)$ has a binomial distrbution with two parameters (n_i, p_i), i.e.,

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

Assume that the Y_i's are independent. Assume also that the trials that compose the response Y_i's are all subject to the same predictor variables, i.e., $(x_{i1}, x_{i2}, \ldots, x_{iq})$. In order to describe the relationship between $(x_{i1}, x_{i2}, \ldots, x_{iq})$ and p_i, we construct:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}$$

This idea can be extended to models for other types of response and is one of the defining features of the GLM.

Note that setting $\eta_i = p_i$ is not appropriate because $0 \leq p_i \leq 1$, (i.e., p_i does not go from $-\infty$ to $+\infty$). Instead we use a **link function** $g(\cdot)$ such that $\eta_i = g(p_i)$. For a binomial probability p_i, there are three common choices:

- Logit: $\eta = \log(p/(1-p))$.
- Probit: $\eta = \Phi^{-1}(p)$, where Φ is the normal cdf.
- Complementary log-log: $\eta = \log\{-\log(1-p)\}$.

In all three cases, η now goes from $-\infty$ to $+\infty$.

For example, when "logit" link is used, we are fitting the following model:

$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}$$

Some algebra:

$$\begin{aligned}
\frac{p_i}{1-p_i} &= e^{\eta_i} = \exp\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}\right) \\
p_i &= (1-p_i) \cdot e^{\eta_i} \\
(1+e^{\eta_i})\, p_i &= e^{\eta_i} \\
\therefore \quad p_i &= \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{1}{1+e^{-\eta_i}}
\end{aligned}$$

Notice that $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}$ (but without ε_i term). Some students might ask: "why do we not simply fit a linear regression of $\log\left(\frac{p}{1-p}\right)$ against the explanatory x-variable?" GLM has three great advantages:

- GLM allows for the non-constant binomial variance.
- GLM deals with the fact that logits for p's near 0 or 1 are infinite.
- GLM allows for differences in the sample sizes by weighted regression.

For another example, when "complementary log-log" link is used, we are fitting the following model:

$$\log\{-\log(1-p_i)\} = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}$$

More algebra:

$$\begin{aligned}
-\log(1-p_i) &= e^{\eta_i} = \exp\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}\right) \\
\log(1-p_i) &= -e^{\eta_i} \\
1-p_i &= e^{-e^{\eta_i}} \\
\therefore p_i &= 1 - e^{-e^{\eta_i}}
\end{aligned}$$

OK, here are some basic facts you need to know about "logistic" modeling. First, what's called a logistic model is about log of "odds ratio" vs. $x_1, x_2, \ldots, x_k$, i.e.,

$$\text{logit}(p) = \text{logistic}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Equivalently, this can also be written as

$$p = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}}$$

Next, suppose we consider two individual cases A and B that have the same values for all the variables except for a dichotomous jth variable as shown below:

case	x_1	x_2	$\cdots$	x_j	$\cdots$	x_k
A	x_{1a}	x_{2a}	$\cdots$	1	$\cdots$	x_{ka}
B	x_{1a}	x_{2a}	$\cdots$	0	$\cdots$	x_{ka}

That is, two individual cases A and B are the same except that A has $x_j = 1$ and B has $x_j = 0$. According to the logit of the probability of success for indivudals A and B, denoted by $\mathit{logit}(p_a)$ and $\mathit{logit}(p_b)$ are given by:

$$\begin{aligned}\text{logit}(p_a) &= \beta_0 + \beta_1 x_{1a} + \cdots + \beta_j(1) + \cdots + \beta_k x_{ka}\\ \text{logit}(p_b) &= \beta_0 + \beta_1 x_{1a} + \cdots + \beta_j(0) + \cdots + \beta_k x_{ka}\end{aligned}$$

This leads to

$$\text{logit}(p_a) - \text{logit}(p_b) = \beta_j$$

Or equivalently,

$$\frac{p_a/(1-p_a)}{p_b/(1-p_b)} = e^{\beta_j}$$

That is, in words, the odds in favor of success for subject A divided by the odds in favor of success for subject B becomes e^{β_j}. This can also be interpreted as the **odds ratio** relating exposure to the jth dichotomous variable for two hypothetical individuals, one of whom is exposed to the variable (individual A) and the other is not exposed for the jth variable (individual B), where the individuals are the same for all other variables considered in the model.

In case of a continuous variable, the argument goes something like:

$$\begin{aligned}\text{logit}(p_a) &= \beta_0 + \beta_1 x_{1a} + \cdots + \beta_j(x_j + \Delta) + \cdots + \beta_k x_{ka}\\ \text{logit}(p_b) &= \beta_0 + \beta_1 x_{1a} + \cdots + \beta_j(x_j) \qquad + \cdots + \beta_k x_{ka}\end{aligned}$$

This leads to

$$\text{logit}(p_a) - \text{logit}(p_b) = \beta_j \Delta$$

Or equivalently,

$$\frac{p_a/(1-p_a)}{p_b/(1-p_b)} = e^{\beta_j \Delta}$$

Thus, the odds in favor of success for subject A divided by the odds in favor of success for subject B can be calculated by $e^{\beta_j \Delta}$ when A has Δ much more of x_j variable than B.

Ex 1. GLM with LOGIT link, a.k.a Logistic Model

This sample data are from an experiment measuring death rates for insects, with 30 insects at each of five treatment levels. Variables are: `dead` = number dead, `alive` = number alive, and `conc` = concentration of insecticide.

```
> library(faraway)
> dim(bliss)
[1] 5 3
> head(bliss)
  dead alive conc
1    2    28    0
2    8    22    1
3   15    15    2
4   23     7    3
5   27     3    4
> attach(bliss)
> model1 <- glm(cbind(dead,alive)~conc, family=binomial, data=bliss)
> model2 <- glm(cbind(dead,alive)~conc, family=binomial(link=probit), data=bliss)
> model3 <- glm(cbind(dead,alive)~conc, family=binomial(link=cloglog), data=bliss)
> fitted(model1)
         1          2          3          4          5
0.08917177 0.23832314 0.50000000 0.76167686 0.91082823
> fitted(model2)
         1          2          3          4          5
0.08424186 0.24487335 0.49827210 0.75239612 0.91441122
> fitted(model3)
        1         2         3         4         5
0.1272700 0.2496909 0.4545910 0.7217655 0.9327715
> x = conc
> y = (dead/(dead+alive))
> plot(x,y)
> 1/(1+exp(-(-2.324+1.162*1))) #To verify R calculation of "logit" at x=1
[1] 0.2383041
> pred1 <- ilogit(model1$coef[1]+model1$coef[2]*x)
> pred2 <- pnorm(model2$coef[1]+model2$coef[2]*x)
> pred3 <- 1-exp(-exp((model3$coef[1]+model3$coef[2]*x)))
> plot(x,y)
> lines(x,pred1,lty=2,col=2)
> lines(x,pred2,lty=3,col=3)
> lines(x,pred3,lty=4,col=4)
> text(locator(1),c("logit"),col=2)
```

```
> text(locator(1),c("probit"),col=3)
> text(locator(1),c("c-log-log"),col=4)
> summary(model3)
> summary(model2)
> summary(model1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3238     0.4179  -5.561 2.69e-08 ***
conc          1.1619     0.1814   6.405 1.51e-10 ***
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 64.76327  on 4  degrees of freedom
Residual deviance:  0.37875  on 3  degrees of freedom
AIC: 20.854

Number of Fisher Scoring iterations: 4
> confint(model1)
                 2.5 %    97.5 %
(Intercept) -3.2060617 -1.557314
conc         0.8301789  1.546129
> exp(confint(model1)) # 95% CI for exponentiated coefficients
                 2.5 %    97.5 %
(Intercept) 0.04051586 0.2107012
conc        2.29372916 4.6932687
> predict(model1, type="response") # predicted values
         1          2          3          4          5
0.08917177 0.23832314 0.50000000 0.76167686 0.91082823
> residuals(model1, type="deviance") # residuals
          1           2           3           4           5
-0.45101510  0.35969607  0.00000000  0.06430235 -0.20449347
```

(See Fig. 1 below.) Notice that the residual deviance is small (compared to the degrees of freedom) indicating a very good fit and x-variable is very significant.

Ex 2. Another example for more calculation of "logistic" model

This dataset is about sex ratios in insects (the proportion of all individuals that are males). In the species in question, it has been observed that the sex ratio is highly variable, and an experiment was set up to see whether population density was involved in determining the fraction of males.

```
> data1 <- read.table("U:\\STAT510\\sexratio.txt",header=T)
> data1
  density females males
1       1       1     0
2       4       3     1
3      10       7     3
4      22      18     4
5      55      22    33
6     121      41    80
7     210      52   158
8     444      79   365
> attach(data1)
> model1 <- glm(cbind(males,females)~density,family=binomial)
```

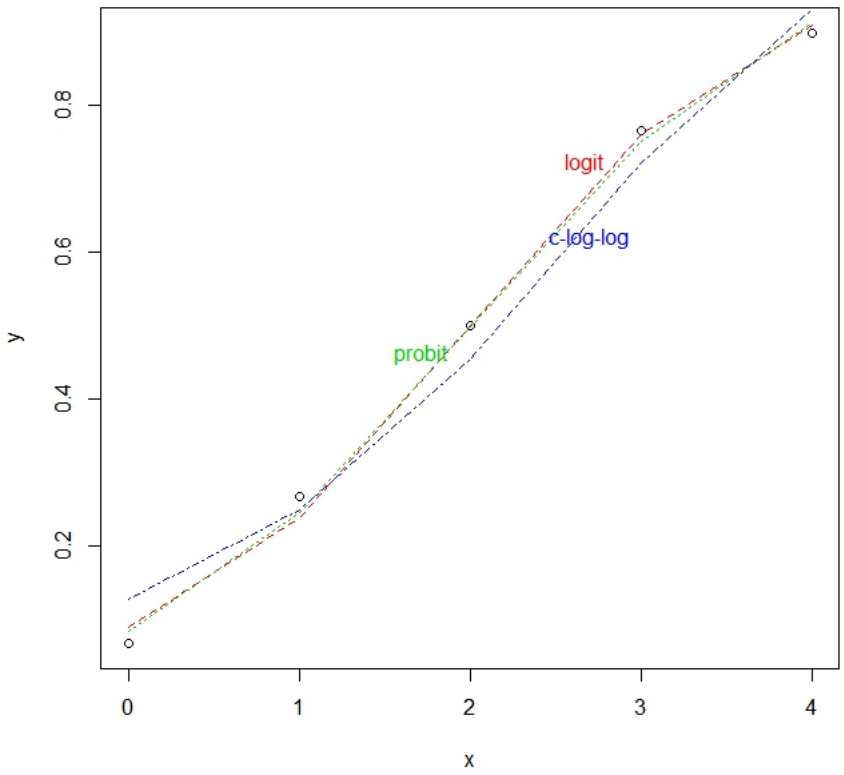

Figure 1: `logit`, `probit`, and `cloglog`

```
> summary(model1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.0807368  0.1550376   0.521    0.603
density     0.0035101  0.0005116   6.862 6.81e-12 ***
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 71.159  on 7  degrees of freedom
Residual deviance: 22.091  on 6  degrees of freedom
AIC: 54.618

Number of Fisher Scoring iterations: 4
> # model1 has "overdispersion" problem.
> p = males/(males+females)
> par(mfrow=c(1,2))
> plot(p~density,pch=16,col=2,ylab="Proportion of males")
> plot(p~log(density),pch=16,col=2,ylab="Proportion of males")
> # This plot shows p is more linear with log(x) than raw x.
> model2 <- glm(cbind(males,females)~log(density),family=binomial)
> summary(model2)

Coefficients:
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.65927    0.48758  -5.454 4.92e-08 ***
log(density)  0.69410    0.09056   7.665 1.80e-14 ***
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 71.1593  on 7  degrees of freedom
Residual deviance:  5.6739  on 6  degrees of freedom
AIC: 38.201

Number of Fisher Scoring iterations: 4

> par(mfrow=c(2,2))
> plot(model2)
> exp(4.5)
[1] 90.01713
> predict(model2,list(density=exp(4.5)))
        1
0.4641847
> predict(model2,list(density=exp(4.5)),type="response")
        1
0.6140064
> newx <- seq(0,6,0.01)
> pred1 <- predict(model2,list(density=exp(newx)),type="response")
> plot(p~log(density),pch=16,ylab="Proportion of males")
> lines(newx,pred1,lty=2,col=2)
```

(See Fig. 2 as well) The second model (i.e., `model2` with log(`density`)) fits much better than model1. In a GLM model like this, it is assumed that the residual deviance is the same as the residual degrees of freedom. If the residual deviance is larger than the residual degrees of freedom, it's called **overdispersion**. It means that there is extra, unexplained variation, over and above the binomial variance assumed by the model specification. In `model1`, there is substantial overdispersion (residual deviance =22.091 on 6 d.f.). In `model2`, there is no evidence of overdispersion (residual deviance = 5.67 on 6 d.f.). So we take `model2` as the final one.

Also, the estimated final model is $\hat{y} = -2.65927 + 0.6941 \log(x)$. So for example, the predicted y at $\log(x)$=4.5 (i.e., $x = e^{4.5}$=90.01713) would be $\hat{y} = -2.65927 + (0.6941 \times 4.5) = 0.4618$.

This means $\hat{p} = \dfrac{1}{1 + e^{-0.4618}} = 0.61400532$. Notice that you need to write `type="response"` in `predict` command above to obtain $\hat{p}$.

You can go through usual model diagnostics to ensure the health of the final model. It turns out there is no peculiar pattern in the residuals against the fitted values, and the normal plot of the residuals is reasonably linear. The 4^{th} observation is highly influential (it has a large Cook's D value), but the model is still significant with this point omitted.

We conclude that the proportion of animals that are males increases significantly with increasing density, and that the logistic model is linearized by logarithmic transformation of the explanatory variable (population density). We finish this example by overlaying the fitted line on the scatterplot (see Fig. 3).

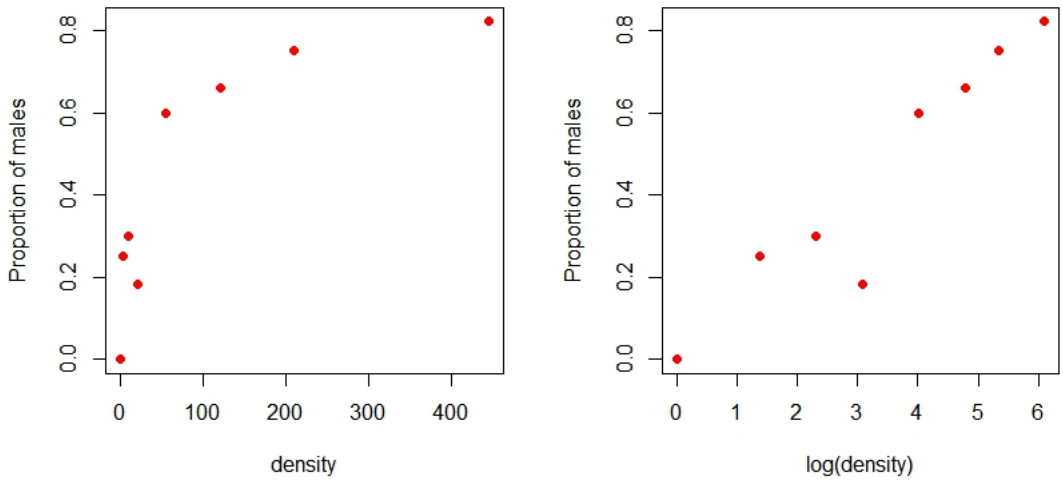

Figure 2: Proportion of males is more linear with $\log(x)$, than raw x.

Model Selection

Ex 3. Binary Response

In this example, the response variable is called **incidence**; a value of 1 means that an island was occupied by a particular species of bird, and 0 means that the bird did not breed there. The explanatory variables are the **area** of the island (km^2) and the **isolation** of the island (distance from the mainland, km).

```
> data1 <- read.table("U:\\STAT510\\isolation.txt",header=T)
> dim(data1)
[1] 50  3
> head(data1)
  incidence  area isolation
1         1 7.928     3.317
2         0 1.925     7.554
3         1 2.045     5.883
4         0 4.781     5.932
5         0 1.536     5.308
6         1 7.369     4.934
> attach(data1)
```

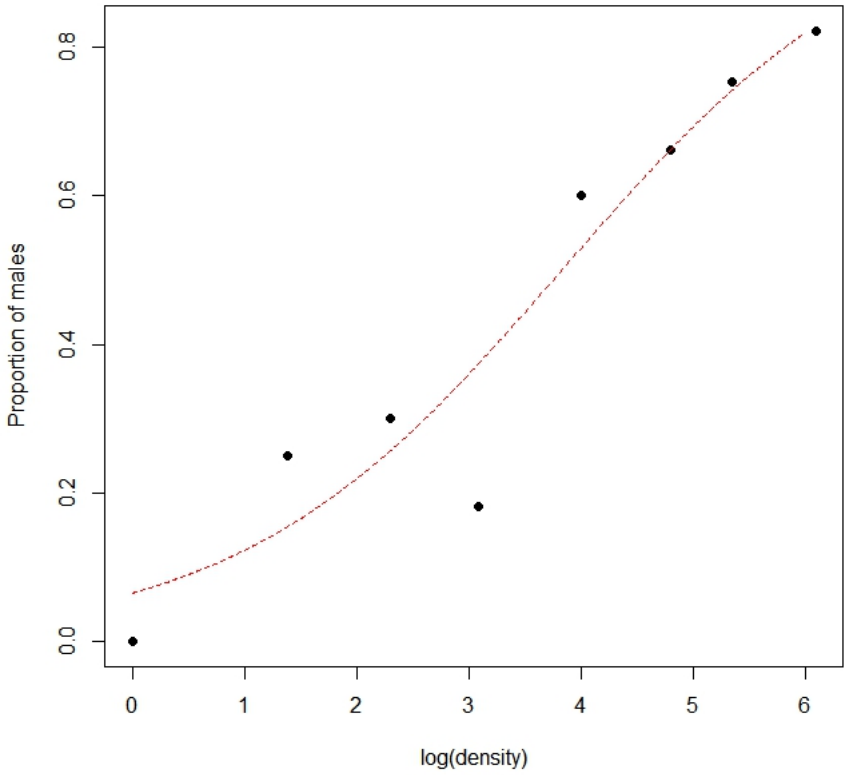

Figure 3: Observed proportion of males and estimated logistic model

```
> model1 <- glm(incidence~area*isolation,family=binomial)
> summary(model1)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.0313     7.1747   0.562    0.574
area            1.3807     2.1373   0.646    0.518
isolation      -0.9422     1.1689  -0.806    0.420
area:isolation -0.1291     0.3389  -0.381    0.703

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.029  on 49  degrees of freedom
Residual deviance: 28.252  on 46  degrees of freedom
AIC: 36.252

Number of Fisher Scoring iterations: 7

> model2 <- glm(incidence~area+isolation,family=binomial)
> summary(model2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.6417     2.9218   2.273  0.02302 *
```

```
area          0.5807     0.2478   2.344  0.01909 *
isolation    -1.3719     0.4769  -2.877  0.00401 **
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.029  on 49  degrees of freedom
Residual deviance: 28.402  on 47  degrees of freedom
AIC: 34.402

Number of Fisher Scoring iterations: 6

> anova(model2,model1,test="Chi")
Analysis of Deviance Table

Model 1: incidence ~ area + isolation
Model 2: incidence ~ area * isolation
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        47     28.402
2        46     28.252  1  0.15043   0.6981
```

The simpler model (`model2`) is not significantly worse, so we accept this, and inspect the parameter estimates and standard errors.

```
> summary(model2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.6417     2.9218   2.273  0.02302 *
area          0.5807     0.2478   2.344  0.01909 *
isolation    -1.3719     0.4769  -2.877  0.00401 **
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.029  on 49  degrees of freedom
Residual deviance: 28.402  on 47  degrees of freedom
AIC: 34.402

Number of Fisher Scoring iterations: 6

> # The following are done to check y vs. each variable, separately.
> model_a <- glm(incidence~area, family=binomial)
> model_i <- glm(incidence~isolation, family=binomial)
>
> par(mfrow=c(1,2))
> newx1 <- seq(0,10,0.01)
> newy1 <- predict(model_a,list(area=newx1),type="response")
> plot(area, incidence); lines(newx1,newy1,col=2,lwd=2)
>
> newx2 <- seq(2,10,0.01)
> newy2 <- predict(model_i,list(isolation=newx2),type="response")
> plot(isolation, incidence); lines(newx2,newy2,col=2,lwd=2)
```

Area has a significant positive effect (larger islands are more likely to be occupied), but isolation has a very strong negative effect (isolated islands are much less likely to be occupied). This is the minimal adequate model. We also plotted the fitted model through the scatterplot of the data. It is much easier to do this for each variable separately, as shown in Fig. 4.

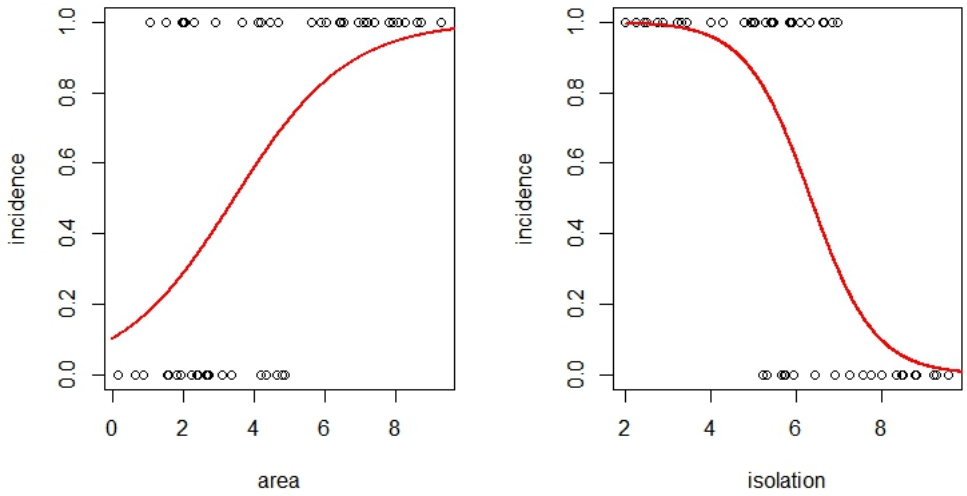

Figure 4: Plots of separately fitted models superimposed on the data plots

Count Data

Ex 4. Count Data

In the sample dataset `gala` from the `faraway` library, there are 30 Galapagos islands and 7 variables:

- `Species` = the number of plant species found on the island
- `Endemics` = the number of endemic species
- `Area` = the area of the island (km^2)
- `Elevation` = the highest elevation of the island (m)
- `Nearest` = the distance from the nearest island (km)
- `Scruz` = the distance from Santa Cruz island (km)
- `Adjacent` = the area of the adjacent island (square km)

We don't want to include `Endemics` for this analysis, so we drop it at the beginning.

```
> library(faraway)
> dim(gala)
[1] 30  7
> head(gala)
             Species Endemics  Area Elevation Nearest Scruz Adjacent
Baltra            58       23 25.09       346     0.6   0.6     1.84
Bartolome         31       21  1.24       109     0.6  26.3   572.33
Caldwell           3        3  0.21       114     2.8  58.7     0.78
Champion          25        9  0.10        46     1.9  47.4     0.18
Coamano            2        1  0.05        77     1.9   1.9   903.82
Daphne.Major      18       11  0.34       119     8.0   8.0     1.84
> gala <- gala[,-2]
> head(gala)
             Species  Area Elevation Nearest Scruz Adjacent
```

```
Baltra            58 25.09       346     0.6   0.6     1.84
Bartolome         31  1.24       109     0.6  26.3   572.33
Caldwell           3  0.21       114     2.8  58.7     0.78
Champion          25  0.10        46     1.9  47.4     0.18
Coamano            2  0.05        77     1.9   1.9   903.82
Daphne.Major      18  0.34       119     8.0   8.0     1.84
> model1 <- glm(Species~.,family=poisson, data=gala)
> summary(model1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent    -6.630e-04  2.933e-05 -22.608  < 2e-16 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68

Number of Fisher Scoring iterations: 5
```

The residual deviance of 716.85 with 24 df means a serious *overdispersion*, and it means an ill-fitting model if the Poisson is correct for the response. Sometimes such an overdispersion is due to outliers, but the normal plot of the residuals showed no serious outliers. We also note that the proportion of deviance explained by this model, $1 - \frac{716.85}{3510.73} = 0.7957$, which is almost the same as the R^2 value of the plain `lm` model. So this could be that this model needs some improvement.

Sometimes it can be improved by specifying negative binomial (instead of Poisson), for example. If the underlying mechanism is not known, we can introduce a dispersion parameter ϕ such that $\text{Var}(Y) = \phi E(Y) = \phi\mu$. For the regular Poisson case, $\phi = 1$; in case of overdispersion, $\phi > 1$; and in case of underdispersion, $\phi < 1$. The dispersion parameter can be estimated by:

$$\hat{\phi} = \frac{\chi^2}{n-p} = \frac{\Sigma\left(y_i - \hat{\mu}_i\right)^2 / \hat{\mu}_i}{n-p}$$

```
> (phi = sum(residuals(model1,type="pearson")^2/model1$df.res))
[1] 31.74914
> summary(model1,dispersion=phi)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.1548079  0.2915897  10.819  < 2e-16 ***
Area        -0.0005799  0.0001480  -3.918 8.95e-05 ***
Elevation    0.0035406  0.0004925   7.189 6.53e-13 ***
Nearest      0.0088256  0.0102621   0.860    0.390
Scruz       -0.0057094  0.0035251  -1.620    0.105
Adjacent    -0.0006630  0.0001653  -4.012 6.01e-05 ***
```

```
---
(Dispersion parameter for poisson family taken to be 31.74914)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68

Number of Fisher Scoring iterations: 5
```

Notice that the inclusion of the dispersion estimate has no effect on the regression parameter estimates, but the p-values of the variables are quite different. The p-values shown here are more similar to those of the plain `lm` printout.

When comparing Poisson models with overdispersion, an F-test rather tahna χ^2 test must be used. So, test the significance of each of the predictors relative to the full model in the following way.

```
> drop1(model1,test="F")
Single term deletions

Model:
Species ~ Area + Elevation + Nearest + Scruz + Adjacent
          Df Deviance     AIC F value    Pr(>F)
<none>         716.85  889.68
Area       1  1204.35 1375.18 16.3217 0.0004762 ***
Elevation  1  2389.57 2560.40 56.0028 1.007e-07 ***
Nearest    1   739.41  910.24  0.7555 0.3933572
Scruz      1   813.62  984.45  3.2400 0.0844448 .
Adjacent   1  1341.45 1512.29 20.9119 0.0001230 ***
---
Warning message:
In drop1.glm(model1, test = "F") : F test assumes 'quasipoisson' family
> model2 <- glm(Species~.,family=quasipoisson, data=gala)
> summary(model2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.1548079  0.2915901  10.819 1.03e-10 ***
Area        -0.0005799  0.0001480  -3.918 0.000649 ***
Elevation    0.0035406  0.0004925   7.189 1.98e-07 ***
Nearest      0.0088256  0.0102622   0.860 0.398292
Scruz       -0.0057094  0.0035251  -1.620 0.118380
Adjacent    -0.0006630  0.0001653  -4.012 0.000511 ***
---
(Dispersion parameter for quasipoisson family taken to be 31.74921)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Ex 5. Comparing two models

In the following sample dataset, The response variable is `n`, the count for each contingency. The explanatory variables are all categorical: `sun` is a two-level factor (Sun and Shade), `height` is a two-level factor (High and Low), `perch` is a two-level factor (Broad and Narrow), `time` is a three-level factor (Afternoon, Mid.day and Morning), and there are two lizard `species` both belonging to the genus Anolis (A. grahamii and A. opalinus). In this example, we will just show how to compare two particular models:

```
> data1 <- read.table("U:\\STAT510\\lizards.txt",header=T)
> dim(data1)
[1] 48  6
> head(data1)
   n   sun height  perch    time  species
1 20 Shade   High  Broad Morning opalinus
2 13 Shade    Low  Broad Morning opalinus
3  8 Shade   High Narrow Morning opalinus
4  6 Shade    Low Narrow Morning opalinus
5 34   Sun   High  Broad Morning opalinus
6 31   Sun    Low  Broad Morning opalinus
> data1 <- read.table("U:\\STAT510\\lizards.txt",header=T)
> dim(data1)
[1] 48  6
> head(data1)
   n   sun height  perch    time  species
1 20 Shade   High  Broad Morning opalinus
2 13 Shade    Low  Broad Morning opalinus
3  8 Shade   High Narrow Morning opalinus
4  6 Shade    Low Narrow Morning opalinus
5 34   Sun   High  Broad Morning opalinus
6 31   Sun    Low  Broad Morning opalinus
> attach(data1)
> model1 <- glm(n~sun+height+perch+time*species,family=poisson)
> model2 <- update(model1, ~. -time:species,family=poisson)
> anova(model2,model1,test="Chi") #Use F-test in case of quasipoisson
Analysis of Deviance Table

Model 1: n ~ sun + height + perch + time + species
Model 2: n ~ sun + height + perch + time * species
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        41     152.61
2        39     146.25  2   6.3676  0.04143 *
---
```

At $\alpha = 0.05$, we conclude that the interaction term (between time & species) is significant.

Survival Analysis

Survival analysis (also called reliability analysis) covers a set of techniques for modeling the time to an event. Data may be **right censored** – the event may not have occured by the end of the study or we may have incomplete information on an observation but know that up to a certain time the event had not occured (e.g., the participant dropped out of study in week 10 but was alive at that time).

While generalized linear models are typically analyzed using the `glm( )` function, survival analyis is typically carried out using functions from the `survival` package. The `survival` package can handle one and two sample problems, parametric accelerated failure models, and the Cox proportional hazards model.

Data are typically entered in the format *start time*, *stop time*, and *status* (e.g., 1=event occured, 0=event did not occur). Alternatively, the data may be in the format *time to event* and *status* (1=event occured, 0=event did not occur). A status=0 indicates that the observation is right cencored. Data are bundled into a `Surv` object via the `Surv( )` function prior to further analyses.

- `survfit( )` is used to estimate a survival distribution for one or more groups.
- `survdiff( )` tests for differences in survival distributions between two or more groups.
- `coxph( )` models the hazard function on a set of predictor variables.

Ex 6. Survival in patients with advanced lung cancer from the North Central Cancer Treatment Group. Performance scores rate how well the patient can perform usual daily activities.

variable	description
`inst`	Institution code
`time`	Survival time in days
`status`	censoring status 1=censored, 2=dead
`sex`	Male=1, Female=2
`ph.ecog`	ECOG performance score (0=good, 5=dead)
`ph.karno`	Karnofsky performance score (bad=0, good=100) rated by physician
`meal.cal`	Calories consumed at meals
`wt.loss`	Weight loss in last six months

```
> library(survival)
> dim(lung)
> head(lung)
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1    3  306      2  74   1       1       90       100     1175      NA
2    3  455      2  68   1       0       90        90     1225      15
3    3 1010      1  56   1       0       90        90       NA      15
4    5  210      2  57   1       1       90        60     1150      11
5    1  883      2  60   1       0      100        90       NA       0
6   12 1022      1  74   1       1       50        80      513       0
> # Mayo Clinic Lung Cancer Data
> # Learn about the dataset
> help(lung)
> # Create a Surv object
> survobj <- with(lung, Surv(time,status))
>
> # Plot survival distribution of the total sample
> # Kaplan-Meier estimator
> model1 <- survfit(survobj~1, data=lung)
> # summary(model1)
```

```
> plot(model1, xlab="Survival Time in Days",
+   ylab="% Surviving", yscale=100,
+    main="Survival Distribution (Overall)")
> # Compare the survival distributions of men and women
> model2 <- survfit(survobj~sex,data=lung)
>
> # Plot the survival distributions by sex
> plot(model2, xlab="Survival Time in Days",
+   ylab="% Surviving", yscale=100, col=c("red","blue"),
+   main="Survival Distributions by Gender")
>   legend("topright", title="Gender", c("Male", "Female"),
+   fill=c("red", "blue"))
> # Test for difference between male and female
> # Survival curves (logrank test)
> survdiff(survobj~sex, data=lung)
Call:
survdiff(formula = survobj ~ sex, data = lung)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=1 138      112     91.6      4.55      10.3
sex=2  90       53     73.4      5.68      10.3

 Chisq= 10.3  on 1 degrees of freedom, p= 0.00131
>
> # Predict male survival from age and medical scores
> MaleMod <- coxph(survobj~age+ph.ecog+ph.karno+pat.karno,
+   data=lung, subset=sex==1)
>
> # Display results
> MaleMod
Call:
coxph(formula = survobj ~ age + ph.ecog + ph.karno + pat.karno,
    data = lung, subset = sex == 1)

               coef exp(coef) se(coef)     z      p
age         0.02247   1.02272  0.01222  1.84 0.0659
ph.ecog     0.66545   1.94537  0.22571  2.95 0.0032
ph.karno    0.02555   1.02588  0.01178  2.17 0.0300
pat.karno -0.01106   0.98900  0.00889 -1.24 0.2136

Likelihood ratio test=17.9  on 4 df, p=0.00131
n= 134, number of events= 108
   (4 observations deleted due to missingness)
>
> # Evaluate the proportional hazards assumption
> cox.zph(MaleMod)
               rho   chisq     p
age        0.00534 0.00363 0.952
ph.ecog    0.02851 0.09155 0.762
ph.karno   0.16922 2.43462 0.119
pat.karno 0.02988 0.12793 0.721
GLOBAL          NA 5.62951 0.229
```

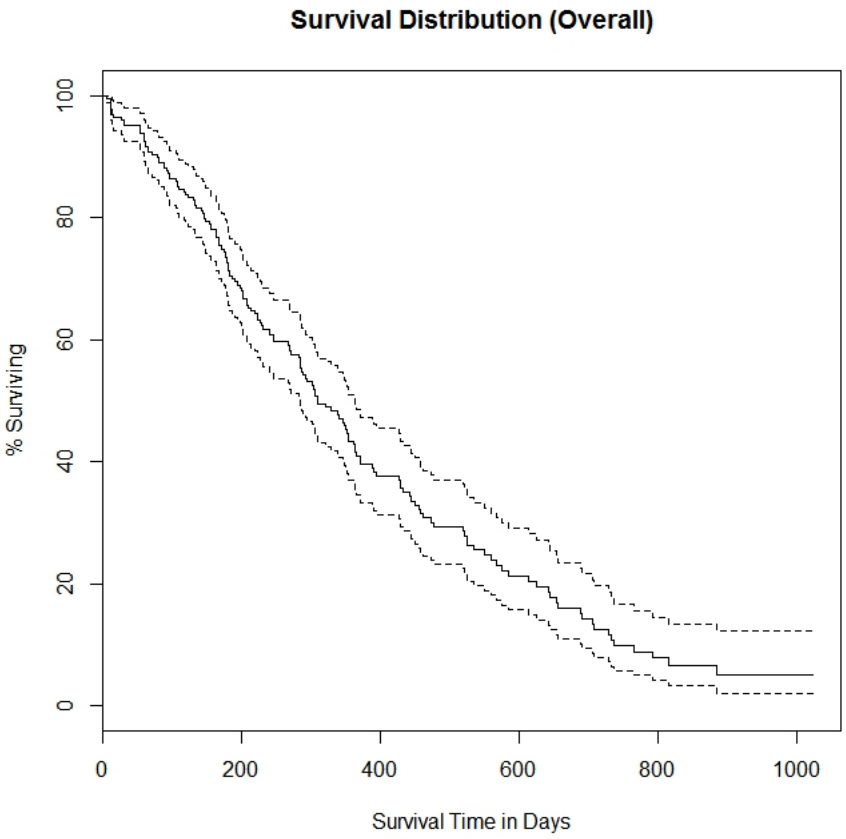

Figure 5: Survival curve with confidence interval

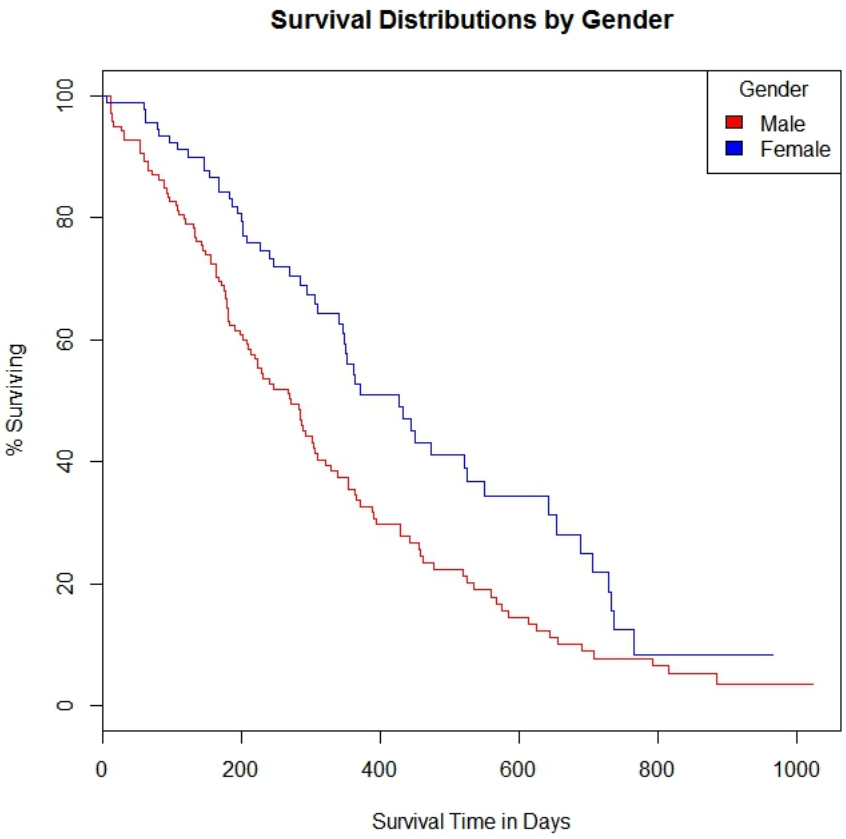

Figure 6: Survival curve for men & women