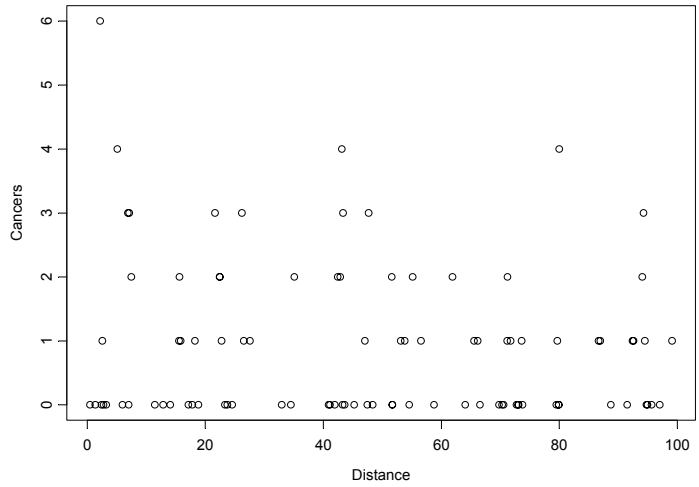
Zobecněné lineární modely

Pokud naše data neodpovídají předpokladům metod uvedených v předchozích cvičeních a ani nejsme schopni je rozumně transformovat a ani nám nepomohou neparametrické metody je řešením použití některé z metod ze skupiny zobecněných lineárních modelů (GLIM). Tyto modely mohou a také ve skutečnosti pracují s daty, kde se např. variance v datech mění s průměrem, dále můžeme mít data o četnostech s nulovými hodnotami, binární závislé proměnné. Detaily a úvod do GLIMů jsou uvedeny např. v sylabech od T. Herbena. To že pro analyzovaná data použiji model s daným rozdělením neznamená, že moje data z toto rozdělení mají, stačí jen pokud jím mohu moje data aproximovat.

Zobecněné lineární modely mají dva důležité parametry: **transformační funkci** (link function), která převádí hodnoty prediktoru na smysluplné hodnoty závislé proměnné; **typ rozložení** dané tak, aby postihlo vztah mezi rozptylem a očekávanou hodnotou y (binomické, Poissonovo, Gamma, exponenciální). To jak daný GLM model odpovídá studovaným datům se vyjadřuje pomocí tzv. **deviance**.

V S+ je analýza GLMů obdobná jako jakákoli jiná analýza (regrese, ANOVA). Jen pro výpočet použijeme funkci `glm` a dále musíme specifikovat typ rozložení (stejně jako když jsme počítali frekvenční tabulku či logistickou regresi). Důležité je u testovaného modelu brát v potaz, že získané p hodnoty nemusí být přesné a také sledovat, zda model nepřekročil tzv. „overdispersion“. Pokud spočteme minimální adekvátní model, pak by reziduální škálovaná deviance měla zhruba odpovídat reziduálním stupňům volnosti. Pokud je výrazně vyšší, pak je buď špatně definovaný model, či pravděpodobnost p není konstantní. Řešení je možné, a použijeme pak místo testovací statistiky Chí-kvadrát pro srovnání modelů, F statistiku. Dále overdispersion ovlivňuje odhady chyb, kdy ve skutečnosti jsou větší a signifikance parametrů může být přeceněna.

Vliv vzdálenosti od zdroje na počty událostí
V následujícím příkladu chceme zjistit, zda má na počet pacientů s rakovinou prostaty v dané nemocnici vliv jak je daleko od jaderné elektrárny. To je analogie např. počtu potomků ve vzdálenosti od mateřské rostliny, počtu odchycených jedinců... V našem případě je tedy vzdálenost od elektrárny kontinuální vysvětlující proměnná a vysvětlujeme počet případů, kdy počty jsou vždy celočíselné. Dále je zde při velké vzdálenosti od zdroje také velký počet nulových hodnot a nejvyšší hodnoty máme v blízkosti zdroje. Pro tyto typy dat používáme Poissonovské rozdělení chyb stejně jako v příkladu s kontingenční tabulkou. V datovém souboru (clusters; Crawley) máme dvě proměnné: počet případů a vzdálenost.



Po vizuální kontrole dat pro výpočet voláme v S+ jednoduše: `model<-glm(Cancers~Distance, poisson)`.

```
Coefficients:
                   Value Std. Error    t value
(Intercept)  0.186933539 0.18777172  0.9955362
   Distance -0.006139087 0.00365356 -1.6803029

(Dispersion Parameter for Poisson family taken to be 1 )

    Null Deviance: 149.4839 on 93 degrees of freedom

Residual Deviance: 146.6431 on 92 degrees of freedom
```

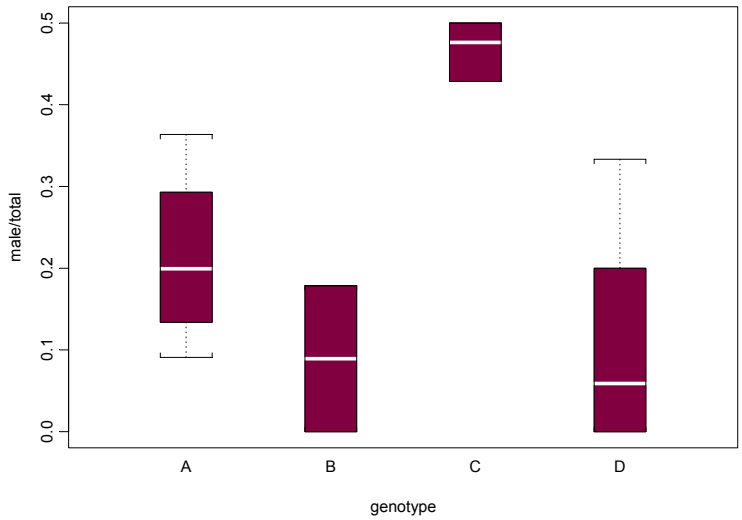
Z hodnoty koeficientu u proměnné Distance vidíme, že závislost je negativní, avšak pokud zkontrolujeme hodnotu overdisperze, pak musíme říci, že poměr mezi reziduální deviancí a stupni volnosti je výrazně vyšší než 1 (testovat tedy budeme pomocí F statistiky, normálně bychom použili Chí-kvadrát). Pokračujeme tedy v testu zda distance je významná pomocí zjednodušení modelu (fce update) a poté srovnání pomocí funkce anova s testovací statistikou F: `anova (model, model.1, test="F").`

```
> anova(model, model.1, test="F")
Analysis of Deviance Table

Response: Cancers

     Terms Resid. Df Resid. Dev Test Df  Deviance  F Value     Pr(F)
1 Distance        92   146.6431
2        1        93   149.4839      -1 -2.840826 1.841963 0.1780414
```

F hodnota a pravděpodobnost, která ji odpovídá, neukazují na výrazné rozdíly mezi modely, takže jsme neprokázali, že vzdálenost jakkoli ovlivňuje počet výskytů rakoviny prostaty.

Poměrová data

V dalším příkladu nás zajímá, zda se liší různé genotypy nějakého druhu hmyzu v tom, jaký je podíl pohlaví jejich potomků. Použijeme data „Sexratio“ (Crawley). Naimportujeme datový soubor, připojíme (attach) a následně si data vizualizujeme. Znázorníme si např. poměry samčích potomků pro jednotlivé genotypy; voláme tedy: `plot(genotype,male/total).`



Z grafu vidíme, že tam zřejmě nějaké rozdíly budou a také, že velmi kolísá variabilita v rámci genotypů. Pro analýzu tohoto typu dat (poměry mezi dvěma hladinami) si musíme vytvořit pomocnou závislou proměnnou, která obsahuje jak hodnoty pro počty samců tak i samic. Použijeme funkci cbind, ta spojí v jeden objekt několik jiných objektů. V našem případě vytvoříme proměnnou y, která se skládá ze dvou „sloupců“, každý s počty pro dané pohlaví. Voláme tedy `y<-cbind(male, total-male).`

Proměnná y vypadá takto:

```
> y
   male
 1    3 14
 2    8 14
 3    1 10
 4    2  7
...
```

Vytvoříme tedy model `g.typy<-glm(y~genotype, binomial).`

Vypíšeme si model: `summary (g.typy)`

```
Residual Deviance: 14.5101 on 10 degrees of freedom
```

Nejdříve zkontrolujeme poměr mezi hodnotou reziduální deviance a počtem odpovídajících stupňů volnosti (overdispersion). V tomto případě se jedná o 14.51 ku 10, což znamená, že bychom raději měli použít pro testování F statistiku.

Pokud odebereme z modelu vliv genotypů, zůstane nám tzv. nulový model. S+ nám umožňuje testovat náš model oproti nulovému jednoduše. V minulém příkladě jsme jej vytvořili tak, že jsme odebrali jedinou nezávislou proměnnou. Pokud zadáme do funkce anova jméno jen

jednoho modelu, testuje S+ zadaný model oproti nulovému modelu. Takže voláme: `anova(g.typy, test="F")`.

```
> anova(g.typy, test ="F")
Analysis of Deviance Table

Binomial model

Response: y

Terms added sequentially (first to last)
         Df Deviance Resid. Df Resid. Dev  F Value      Pr(F)
    NULL                    13   32.24692
genotype  3 17.73682         10   14.51010 4.668026 0.02741896
```

Výsledek ukazuje průkazný vliv efektu genotypů na poměr mezi pohlavím potomků.
Pokud bychom chtěli spočíst poměry pohlaví pro jednotlivé hladiny genotypů, voláme: `fitted (g.typy)`. Takže pro genotyp A je poměr 0.237, atd.

```
> fitted (g.typy)
         1         2         3         4         5         6
0.2372881 0.2372881 0.2372881 0.2372881 0.1612903 0.1612903
         7         8         9        10        11        12
0.4693878 0.4693878 0.4693878 0.1311475 0.1311475 0.1311475
        13        14
0.1311475 0.1311475
```