SMIII lectures

Week 9

Poisson regression

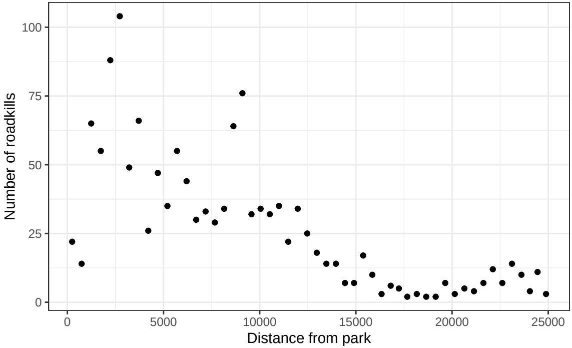
Poisson data

- Independent observations
- Constant rate of success λ (time based intervals)
- The mean rate λ is equal to the variance λ

Example - roadkills

```
## Rows: 52 Columns: 23
## -- Column specification --------------------------------
## Delimiter: "\t"
## dbl (23): Sector, X, Y, BufoCalamita, TOT.N, S.RICH, OPE
##
## i Use `spec()` to retrieve the full column specificatio
## i Specify the column types or set `show_col_types = FALS
```

distance_park	roadkills
250.214	22
741.179	14
1240.080	65
1739.885	55
2232.130	88
2724.089	104



Number of roadkills
100
75
50
25
0
0
5000
10000
15000
20000
25000
Distance from park

Poisson Regression Models

Consider data

$$(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)$$

For count data consider a Poisson model

$$Y_i \sim Po(\lambda_i) \text{ independently, for } i = 1, 2, \ldots, n.$$

The regression problem is to relate the Poisson mean, λ_i, to the predictor $\boldsymbol{x}_i$.

That is, we seek a suitable model

$$M : \lambda_i = f(\boldsymbol{x}_i).$$

Log-linear Models

To ensure that the Poisson mean λ_i is positive, we define

$$\eta_i = \log(\lambda_i).$$

Consider log linear regression models

$$M : \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

In matrix notation,

$$\boldsymbol{\eta} = X\boldsymbol{\beta}.$$

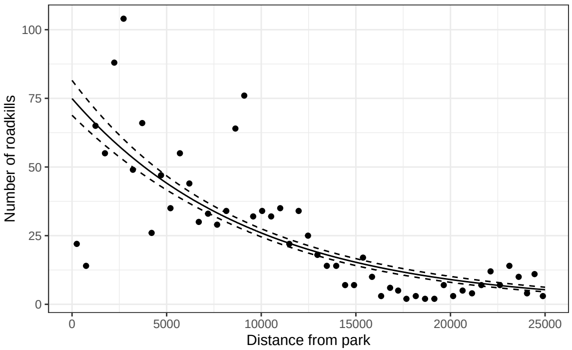
Generalised linear models

Regression Model	Response	Link Function
Linear	$Y_i \sim N(\mu_i, \sigma^2)$	$\eta_i = \mu_i$
Logistic	$Y_i \sim B(n_i, \pi_i)$	$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$
Poisson	$Y_i \sim Po(\lambda_i)$	$\eta_i = \log(\lambda_i)$

Fitting in R

```
roadkill_glm <- glm(roadkills ~ distance_park, data = roadkill, family = poisson())
summary(roadkill_glm)
```

```
##
## Call:
## glm(formula = roadkills ~ distance_park, family = poisson(),
##     data = roadkill)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.1100  -1.6950  -0.4708   1.4206   7.3337
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    4.316e+00  4.322e-02   99.87   <2e-16 ***
## distance_park -1.059e-04  4.387e-06  -24.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1071.4  on 51  degrees of freedom
## Residual deviance:  390.9  on 50  degrees of freedom
## AIC: 634.29
##
## Number of Fisher Scoring iterations: 4
```



Number of roadkills
100
75
50
25
0
0
5000
10000
15000
20000
25000
Distance from park

Interpretation of coefficients

$$\log(Y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Consider increasing x_i by 1

$$\begin{aligned} \log(Y_i') &= \hat{\beta}_0 + \hat{\beta}_1(x_i + 1) \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_1 \\ &= \log(Y_i) + \hat{\beta}_1 \end{aligned}$$

$$\Rightarrow Y_i' = e^{\hat{\beta}_1} Y_i$$

Example - roadkill

```r
b <- coef(roadkill_glm)
exp(b[2])
```

```
## distance_park
##     0.9998942
```

So increasing distance by 1 km, decreases by 1%.

Poisson rate regression

Poisson rate regression

$$
\begin{aligned}
\log(r_i) = \log\left(\frac{\lambda_i}{t_i}\right) &= \beta_0 + \beta_1 x_i \\
\Rightarrow \log(\lambda_i) = \log(t_i) + \beta_0 + \beta_1 x_i
\end{aligned}
$$

Example - insurance

An insurance company recorded the number of policies held and the number of accident claims in 64 categories defined by

- `District`: 1-4, where 4 represents major cities;
- `Engine`: Engine capacity of the car,
  - < 1 litre,
  - $1 - 1.5$ litre,
  - $1.5 - 2$ litre,
  - > 2 litre;
- `Age`:
  - < 25,
  - $25 - 29$,
  - $30 - 35$,
  - > 35.

```
## Rows: 64 Columns: 5
## -- Column specification -----------------------------
## Delimiter: ","
## chr (3): District, Engine, Age
## dbl (2): Policies, Claims
##
## i Use `spec()` to retrieve the full column specificatio
## i Specify the column types or set `show_col_types = FALS
```

District	Engine	Age	Policies	Claims
D1	<1l	<25	197	38
D1	<1l	25-29	264	35
D1	<1l	30-35	246	20
D1	<1l	35+	1680	156
D1	1-1.5l	<25	284	63
D1	1-1.5l	25-29	536	84

Fitting an offset in R

```
M1 <- glm(Claims ~ District + Engine + Age,
          family=poisson,
          offset=log(Policies),
          data=insurance
)
```

```r
summary(M1)
```

```
##
## Call:
## glm(formula = Claims ~ District + Engine + Age, family = poisson,
##     data = insurance, offset = log(Policies))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.46558  -0.50802  -0.03198   0.55555   1.94026
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.82174    0.07679 -23.724  < 2e-16 ***
## DistrictD2   0.02587    0.04302   0.601 0.547597
## DistrictD3   0.03852    0.05051   0.763 0.445657
## DistrictD4   0.23421    0.06167   3.798 0.000146 ***
## Engine1-1.5l  0.16134    0.05053   3.193 0.001409 **
## Engine1.5-2l  0.39281    0.05500   7.142 9.18e-13 ***
## Engine2+l     0.56341    0.07232   7.791 6.65e-15 ***
## Age25-29    -0.19101    0.08286  -2.305 0.021149 *
## Age30-35    -0.34495    0.08137  -4.239 2.24e-05 ***
## Age35+      -0.53667    0.06996  -7.672 1.70e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 236.26  on 63  degrees of freedom
## Residual deviance:  51.42  on 54  degrees of freedom
## AIC: 388.74
##
## Number of Fisher Scoring iterations: 4
```