# Using R for Generalized Linear Models

Nathaniel MacNell

EPID 799B

Many popular models used in epidemiology fall in the family of *generalized linear models*. These models are based on the prototypical linear model

$$Y = X\beta \ + \epsilon$$

Where $Y$ is a vector of response values and $X$ is a vector of covariates including our exposure(s). $\epsilon$ is a vector of residuals representing the difference from the modeled and actual responses; values of $\epsilon$ are assumed to be independent from one another. R uses **model formula** objects to specify the values of Y and X. Model formula objects are built using the  operator. For example, in the births dataset:

```
> formula <- births$GEST ~ births$CIGDIR + births$WIC
> formula

births$GEST ~ births$CIGDIR + births$WIC
```

This doesn't look too exciting at the moment, but we can put the model into the glm function to fit our model. Remember that we can look into more model diagnostics using **names()** on the model object and inspecting the parts with **$**.

```
> model <- glm(births$GEST ~ births$CIGDUR + births$WIC)
> summary(model)

Call:
glm(formula = births$GEST ~ births$CIGDUR + births$WIC)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-21.597   -0.597    0.403    1.403   60.785

Coefficients:
               Estimate Std. Error  t value Pr(>|t|)
(Intercept)    38.59681    0.01213 3182.915  < 2e-16 ***
births$CIGDURU  1.55066    0.17455    8.884  < 2e-16 ***
```

```
births$CIGDURY -0.23293    0.02811   -8.285  < 2e-16 ***
births$WICU    -0.38180    0.13651   -2.797  0.00516 **
births$WICY    -0.03267    0.01736   -1.883  0.05977 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 8.927083)

    Null deviance: 1095342  on 122512  degrees of freedom
Residual deviance: 1093639  on 122508  degrees of freedom
AIC: 615876

Number of Fisher Scoring iterations: 2
```

The **glm()** function by default builds a linear model. We can specify other models using a different link function and outcome distribution (much like SAS). This is specified by adding a *family* argument to **glm()** using the general syntax

$$family = OutcomeDistribution(\text{”}link\text{”})$$

Here are some of the common choices for link and distributions. For the models that use transformations, **coef()** and **confint()** can be used on a model object to get the mean estimates and confidence limits. R returns on the scale of the underlying linear regression, so for regressions with log or logit links you'll need to exponentiate the results using **exp()** to get the measure of interest.

| Name | Coefficients | family= |
|---|---|---|
| Linear | Risk Difference | gaussian(”identity”) |
| Log-binomial | log(Risk Ratio) | binomial(”log”) |
| Logistic | log(Odds Ratio) | binomial(”logit”) |
| Poisson | log(Rate Ratio) | poisson(”log”) |
| Quasi-poisson | log(Rate Ratio) | quasipoisson(”log”) |

R has packages for other distributions that use a similar syntax. Another helpful argument to **glm()** is *weights* which can be used to specify the weights for each observation.

## Activity

1. Calculate the risk difference and 95% confidence interval for Cesarian Section (ROUT value 4) between hospital (DELTYPE value 1) and non-hospital births.

2. Calculate the odds ratio and 95% confidence interval for low birthweight (BWGRP 04 and lower) between adequate (KOTEL 3 or 4) and inadequate prenatal care.