

## Overview: Statistical Modelling (Generalized Linear Models)

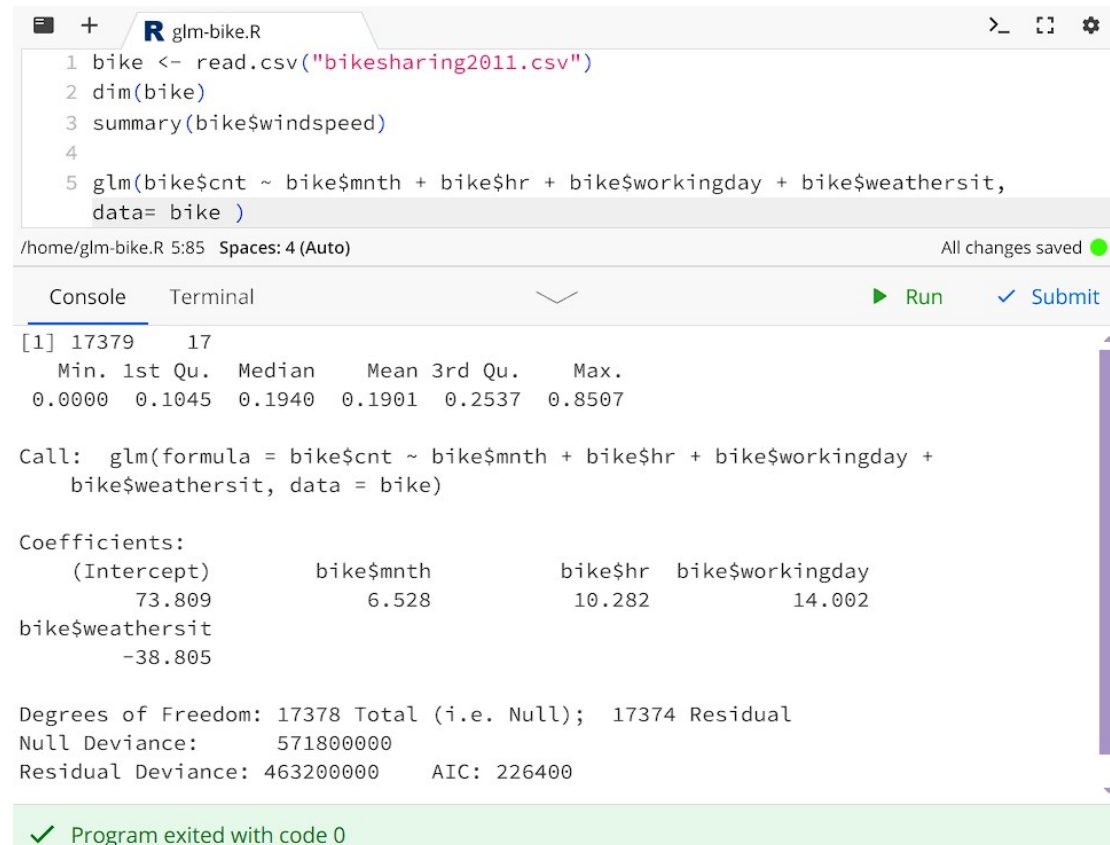
In Week 4, we built linear and logistic regression models for response variables which are continuous and binary, respectively. In some problems, the response variables are not continuous and neither are they categorical. For example, we may be looking at modelling the number of traffic accidents occurring at a particular section of a highway in accident analysis and prevention, daily number of unique visitors to an e-commerce website or number of claims for a certain insurance policy in risk analysis. In such cases, linear and logistic regression models are not appropriate due to the inherent characteristics of the response variable. In this week, we explore a family of regression models, known as generalized linear models for modelling response variables which come from family of distributions known as the exponential family. The binomial, Poisson and normal distributions are some special cases of the exponential family.

### 5.1 Introduction to Generalized Linear Models

In Week 4, we built linear and logistic regression models for response variables which are continuous and binary, respectively. In some problems, the response variables are not continuous and neither are they categorical. For example, in the `bikesharing2011.csv` data set (UCI Machine Learning Repository, <https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset>), the response variable of interest is `cnt` which is the number of hourly users of a bike sharing program in 2011. In this case, the response variable takes on non-negative integers, also known as counts. The potential covariates in the model are `mnth` (month of the year), `hr` (hour of the day, from 0 to 23), `workingday` (an indicator variable that equals 1 if it is neither a weekend nor a holiday), and `weathersit` (a qualitative variable that takes on one of four possible values: misty or cloudy, light rain or light snow, heavy rain, heavy snow).

## Engage Activity: Explore the Data Set

In the previous topic, we have discussed the use of linear regression models in statistical modelling of quantitative response variables. Perform a linear regression analysis on the **bikesharing2011.csv** data set. You may also conduct a descriptive analysis of the variables of interest to better understand the data set.



```
1 bike <- read.csv("bikesharing2011.csv")
2 dim(bike)
3 summary(bike$windspeed)
4
5 glm(bike$cnt ~ bike$mnth + bike$hr + bike$workingday + bike$weathersit,
  data= bike )
```

/home/glm-bike.R 5:85 Spaces: 4 (Auto) All changes saved

Console Terminal

[1] 17379 17

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0000	0.1045	0.1940	0.1901	0.2537	0.8507

Call: glm(formula = bike\$cnt ~ bike\$mnth + bike\$hr + bike\$workingday + bike\$weathersit, data = bike)

Coefficients:

(Intercept)	bike\$mnth	bike\$hr	bike\$workingday	bike\$weathersit
73.809	6.528	10.282	14.002	-38.805

Degrees of Freedom: 17378 Total (i.e. Null); 17374 Residual  
Null Deviance: 571800000  
Residual Deviance: 463200000 AIC: 226400

✓ Program exited with code 0

### 5.1.1 Poisson Regression

In Week 1, we discussed the Poisson distribution which is used for modelling counts.

Recall that the Poisson distribution with mean  $\lambda > 0$  has probability formula:

$$Pr(Y = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where  $k=0,1,2,3,\dots$  is the number of counts and  $e$  is the exponential number. The Poisson distribution has a unique characteristic in which its mean is equal to its variance, a phenomenon known as equidispersion. As an application example, the Poisson distribution can be used to model  $YY$  which denotes the number of users of the bike sharing program during a particular hour of the day, for a fixed weather conditions and during a particular month of the year. If we let  $\lambda=5\lambda=5$ , then the

probability distribution of  $Y$  can be computed using R, <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Poisson.html>:

```
► Run R [ ]
1 k <- 0:10
2 dpois(k,5)

[1] 0.006737947 0.033689735 0.084224337 0.140373896 0.175467370 0.175467370
[7] 0.146222808 0.104444863 0.065278039 0.036265577 0.018132789

✓ Program exited with code 0
```

However, in reality the mean number of users  $Y$  may vary depending on the hour of the day, the month of the year, weather conditions and so on. Therefore, instead of modelling  $Y$  as a Poisson distribution with a fixed mean, we would like to allow the mean  $\lambda=E(Y)$  to vary as a function of the covariates:

$$\log(\lambda(X_1, X_2, \dots, X_p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_p$$

Or

$$\lambda(X_1, X_2, \dots, X_p) = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_p}$$

The  $\beta_0, \dots, \beta_p$  are values to be estimated using maximum likelihood estimation from the data set, which in some ways is similar to what we do in logistic regression modelling. The use of log function is to ensure that  $\lambda(X_1, X_2, \dots, X_p)$  takes on nonnegative values since the counts are nonnegative.

To fit a generalized linear model in R, we use the `glm()` function.

```
R [ ]
1 dat <- read.csv("AmesHouseNormal.csv")
2 glm.fit <- glm(GarageCars ~ GarageArea + LandContour, data=dat, family=poisson())
3 summary(glm.fit)
```

The above GLM fit indicates that based on this Poisson regression model, `GarageArea` is a significant variable in explaining number of `GarageCars`.

## 5.1.2 Generalized Linear Model

In the three regression models that we have discussed so far in this course, a set of predictors (also known as covariates)  $X_1, X_2, \dots, X_p$  is used to predict a response variable  $Y$ . In using these regression models, we assume that conditional on the covariates  $X_1, X_2, \dots, X_p$ ,  $Y$  belongs to a certain family of distributions. For linear regression, we typically assume that  $Y$  follows a normal distribution. For logistic regression, we assume that  $Y$  follows Bernoulli distribution. Finally, for Poisson regression, we assume that  $Y$  follows a Poisson distribution. The normal, Bernoulli and Poisson distributions are members of a wider class of distributions, known as the exponential family. Other well-known members of the exponential family are the exponential distribution, the Gamma distribution and the negative binomial distribution.

In general, we can perform a regression by modelling the response  $Y$  as coming from a particular member of the exponential family, and then transforming the mean of  $Y$  so that the transformed mean is a linear function of the predictors. This approach is known as a generalized linear model (GLM) which express the regression models as:

$$\eta(E(Y|X_1, X_2, \dots, X_p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_p$$

There are three components to a GLM:

- random component - specifies the probability distribution of the response variable  $Y$
- systematic component - specifies the covariates or explanatory variables  $(X_1, X_2, \dots, X_p)$  in the model
- link function - specifies the link between the random and the systematic components

The  $\eta$  on the left hand side of the GLM equation above is the link function which applies a transformation to  $E(Y|X_1, X_2, \dots, X_p)$  so that the transformed mean is a linear function of the predictors. The link functions for linear, logistic and Poisson regression are  $\eta(\mu) = \mu$ ,  $\eta(\mu) = \log(\mu/(1 - \mu))$  and  $\eta(\mu) = \log(\mu)$ , respectively.

Watch the following video for an explanation on generalized linear models.

### Statistical Learning: 4.8 Generalized Linear Models

<https://www.youtube.com/watch?v=n8Nj64FyjSo>

Source: (Stanford Online, 2022)

The `glm()` function in R works for all common GLM models by specifying the `family` parameter.

#### Question 1

Generalized linear models work by transforming the mean of the response variable (from the exponential family) as a linear function of the predictor variables.

**ANS:** True

#### Question 2

What are the three components in a generalized linear model? Select all that applies.

**ANS:**

- Random component
- Systematic component
- Link function

## 5.2 Measuring Goodness-of-Fit in Generalized Linear Models

### Analysis of Deviance

The simplest model that can be fitted to a data set consisting of  $n$  observations or data points is the null model which has only one parameter (e.g.  $\beta_0$  in our linear regression model) representing a common  $\mu$ . The null model attributes all the variation in our response variable to the random component. At the other extreme is the full model which has  $n$  parameters and thus attributes all the variation in our response variable to the systematic component. It is obvious that the null model is too simple and the full model is uninformative since it merely repeats the data information.

Using the full model as the baseline, the discrepancy for an intermediate model with pp parameters can be measured using the deviance. For the interested reader, a more mathematical exposition of the deviance is given in the following video.

### GLM - 8 - Deviance

<https://www.youtube.com/watch?v=JUrKnBbDDzU>

Source: (*Meerkat Statistics*, 2020)

## Residuals

Residuals can be used to explore the goodness-of-fit of a model and may also indicate the presence of influential values in the data set which require further investigation.

The Pearson residual, Anscombe residual and deviance residual can also be used as measure of goodness of fit.

In R `glm()` output, the null deviance, residual deviance and AIC are reported. A low null deviance implies that the data can be modeled well without any independent variables hence it is not necessary to use include many independent variables in the model. On the other hand, a low residual deviance implies that the model you have trained is appropriate. Ideally, we would like the residual deviance to be close to its degrees of freedom.

The Akaike information criterion (AIC) is defined as  $AIC = 2p - 2\ln(\hat{L})$ , where p is the number of model parameters and  $\hat{L}$  is the maximum of the likelihood function. A model with a low AIC is characterized by low complexity (minimizes p) and a good fit (maximizes  $\hat{L}$ ).

## Application Activity: Poisson Regression

**Time:** 30 minutes

**Purpose:** The purpose of this activity is to fit a generalized linear model to data set using R.

**Task:** Let  $Y$  denote the number of users of the bike sharing program in the `bikesharing2011.csv` data set.

1. Fit a Poisson regression model for the response variable.
2. Compare the results with the linear regression model fitted in 5.1.
3. Discuss the goodness-of-fit of the GLM.

+

R glm-bike.R

>\_ [] ⚙

```

1 bike <- read.csv("bikesharing2011.csv")
2
3 glm.fit <- glm(bike$cnt ~ bike$mnth + bike$hr + bike$workingday +
  bike$weathersit, data= bike, family=poisson())
4
5 glm.fit
6 summary(glm.fit)

```

Console Terminal

▶ Run ✓ Submit

Call: glm(formula = bike\$cnt ~ bike\$mnth + bike\$hr + bike\$workingday + bike\$weathersit, family = poisson(), data = bike)

Coefficients:

(Intercept)	bike\$mnth	bike\$hr	bike\$workingday
4.54409	0.03510	0.05573	0.07392
bike\$weathersit			
-0.22060			

Degrees of Freedom: 17378 Total (i.e. Null); 17374 Residual  
Null Deviance: 2892000  
Residual Deviance: 2309000 AIC: 2420000

Call:  
glm(formula = bike\$cnt ~ bike\$mnth + bike\$hr + bike\$workingday + bike\$weathersit, family = poisson(), data = bike)

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	4.544e+00	2.289e-03	1985.55	<2e-16 ***
bike\$mnth	3.510e-02	1.618e-04	216.99	<2e-16 ***
bike\$hr	5.573e-02	8.335e-05	668.60	<2e-16 ***
bike\$workingday	7.392e-02	1.199e-03	61.64	<2e-16 ***
bike\$weathersit	-2.206e-01	9.447e-04	-233.52	<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: 2891591 on 17378 degrees of freedom  
Residual deviance: 2308972 on 17374 degrees of freedom  
AIC: 2419883

Number of Fisher Scoring iterations: 5

✓ Program exited with code 0