Binomial regression in R

In this lesson, we’ll analyze real data using binomial regression.

The data that we'll explore come from the University of California Irvine (UCI) Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+#). The original work was published in "Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models". Luis M. Candanedo, Véronique Feldheim. *Energy and Buildings*. Volume 112, 15 January 2016, Pages 28-39.

The variables in the dataset include:

1. date: time year-month-day hour:minute:second
2. Temperature: in Celsius
3. Relative Humidity: as a percentage
4. Light: measured in Lux
5. CO2: in ppm
6. Humidity Ratio: Derived quantity from temperature and relative humidity, in kgwater-vapor/kg-air
7. Occupancy: 0 for not occupied, 1 for occupied status

Our goal will be to predict occupancy using these variables as predictors. Note that the dataset also includes dates and a humidity ratio, which we will ignore for simplicity. To load the data, we'll use the `RCurl` package.

In [1]:

```
library(RCurl) #a package that includes the function getURL(), which allows fo
r reading data from github.
library(ggplot2)

url = getURL("https://raw.githubusercontent.com/LuisM78/Occupancy-detection-da
ta/master/datatest.txt")
occ = read.csv(text = url)
head(occ[,c(2,3,4,5,7)])
summary(occ[,c(2,3,4,5,7)])
```

A data.frame: 6 × 5

	Temperature	Humidity	Light	CO2	Occupancy
	<dbl>	<dbl>	<dbl>	<dbl>	<int>
140	23.7000	26.272	585.2000	749.2000	1
141	23.7180	26.290	578.4000	760.4000	1
142	23.7300	26.230	572.6667	769.6667	1
143	23.7225	26.125	493.7500	774.7500	1
144	23.7540	26.200	488.6000	779.0000	1
145	23.7600	26.260	568.6667	790.0000	1

```
  Temperature       Humidity          Light              CO2        
 Min.   :20.20   Min.   :22.10   Min.   :   0.0   Min.   : 427.5  
 1st Qu.:20.65   1st Qu.:23.26   1st Qu.:   0.0   1st Qu.: 466.0  
 Median :20.89   Median :25.00   Median :   0.0   Median : 580.5  
 Mean   :21.43   Mean   :25.35   Mean   : 193.2   Mean   : 717.9  
 3rd Qu.:22.36   3rd Qu.:26.86   3rd Qu.: 442.5   3rd Qu.: 956.3  
 Max.   :24.41   Max.   :31.47   Max.   :1697.2   Max.   :1402.2  
   Occupancy     
 Min.   :0.0000  
 1st Qu.:0.0000  
 Median :0.0000  
 Mean   :0.3647  
 3rd Qu.:1.0000  
 Max.   :1.0000  
```

In [2]:

```
#hist(occ$Light)
```

Now let's fit a binomial regression model. To do this, we'll have to use the `glm()` function; `lm()` does not have the flexibility to work with GLMs. The first argument in `glm()` is the same as `lm()`: `response predictor1 + predictor2+. . .`. For binomial regression the response can "be specified as a factor (when the first level denotes failure and all others success) or as a two-column matrix with the columns giving the numbers of successes and failures." (From R help file) So, if you have a response where the total number of trials is greater than 1, the second method might be helpful.

Since this function works for GLMs broadly, we'll have to specify that we want binomial regression in particular. We can do this in a few different ways.

```
In [3]: glmod = glm(Occupancy ~ Temperature + Humidity + Light + CO2, data = occ, fami
        ly = "binomial")
        summary(glmod)
        exp(-29.32)
        exp(0.022)
```

```
Call:
glm(formula = Occupancy ~ Temperature + Humidity + Light + CO2, 
    family = "binomial", data = occ)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.4969  -0.0624  -0.0179   0.1038   2.6544  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -29.316563  11.038232  -2.656  0.00791 ** 
Temperature  -0.333612   0.318492  -1.047  0.29488    
Humidity      1.353727   0.298368   4.537  5.7e-06 ***
Light         0.021921   0.001586  13.819  < 2e-16 ***
CO2          -0.006839   0.003257  -2.099  0.03578 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3496.96  on 2664  degrees of freedom
Residual deviance:  375.66  on 2660  degrees of freedom
AIC: 385.66

Number of Fisher Scoring iterations: 9
```

1.84708036048018e-13

1.02224378447044

The model output provides us with a lot of information. Some of this information we are in the position to interpret now, and some we will learn how to interpret soon.

1. First, notice that the output provides the code used to generate the model, under `call`.
2. Second, the output contains the "deviance residuals". We will define these in the next lesson, and so will ignore them for now, but, we can think of them as similar to the residuals in standard linear regression.
3. Third, we see a coefficients table, similar to the one provided by standard linear regression and the `lm()` function. Since we haven't covered inference for GLMs yet, we will focus our attention on the estimate column.
   A. As discussed in a previous lesson, these estimates were calculated using maximum likelihood estimation. Recall that the (log) liklihood function was nonlinear, and so we have to rely on an iterative algorithm to converge to the MLE. The algorithm used in `glm()` is iteratively reweighted least squares (IWLS).
   B. Let's now interpret these values. From a previous lesson, we know that

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 = \log\left(\underbrace{\frac{p}{1-p}}_{\text{odds}}\right)$$

First, note that the intercept term is ≈ -29.32. So, assuming the model is correct, the average log odds of an office being occupied when `temperature = Humidity = Light = CO2 = 0` is ≈ -29.32. Exponentiating, the average odds are $\approx 1.84 \times 10^{-13}$, basically, zero. That seems to make some sense. If the temperature were very low, the lights were off, and no CO2 measured, it would be very unlikely that a person was occupying the office.

Second, let's interpret the light coefficient. Assuming our model is correct, a one-lux increase in light, with all other predictors held constant would result in a ≈ 0.022 increase in the log-odds, on average. Exponentiating, a one-lux increase in light, with all other predictors held constant would result in an average multiplicative increase in odds of ≈ 1.02; a very slight increase. Note that the increase is multiplicitive.

Here's an estimate of the odds of an office being occupied:

$$e^{\hat{\eta}} = e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \widehat{\beta}_4 x_4} = \underbrace{\frac{\hat{p}}{1-\hat{p}}}_{\text{odds}}$$

If we increase

$$\begin{aligned} & e^{\hat{\eta}_{+1}} = e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 (x_3 + 1) + \widehat{\beta}_4 x_4} \\ \implies & e^{\hat{\eta}_{+1}} = exp\{\widehat{\beta}_3\} exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \widehat{\beta}_4 x_4\} \\ \implies & e^{\hat{\eta}_{+1}} = exp\{\widehat{\beta}_3\}\hat{\eta} \end{aligned}$$

So, we see that $exp\{\widehat{\beta}_3\} \approx 1.02$ is an average multiplicitive increase in the odds of an office being occupied, adjusting for temperature, humidity, and CO2.

In the next lesson, we'll learn a bit more about the binomial regression model, which will allow us to interpret more of the output from R table.

In [ ]: