Week ON15

Elmer V Villanueva

01 June 2020

SET YOUR WORKING DIRECTORY!

```
setwd("C:/Users/1/Dropbox/00 - Working Folder/Teaching/DPH112/2019-2020/Week ON15")
```

Announcements

- This is final last week of sessions.
- The final paper is due on 17 June
- All the assessments are to be completed INDIVIDUALLY. No collusion is allowed. This means that you cannot discuss this assignment with other students, nor can you share your approach or code.
- The running count for students forwarding errors is as follows:

Student	Items Identified
Yijia Jiang	6
Jing Wang	4
Yuxuan Wu	2
Xinwen Hu	1
Yuxin Zhang	1

Reading

Read and understand Vittinghoff et al., Chapter 8.1 and 8.2.

Review of Previous Learning

In the previous weeks, we extended the multiple linear regression model

$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$

which is appropriate for continuous outcomes, to apply to outcomes that are binary, resulting in the logistic regression model. We did this by using a link function or a transformation technique that preserves the right-hand side of the equation. The logistic regression model uses the logit link function so that the resulting model is

$$logit(Y) = ln\ Odds(Y) = ln\left(\frac{Pr(Y)}{1 - Pr(Y)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Today, we will end this module by extending the regression model further.

Poisson regression

Poisson regression is useful when the outcome is a count (ie., a discrete or integer variable), with large-count outcomes being "rare"" events. For example, we might wish to model the number of deaths due to airline accidents or the number of COVID-19 infections. The idea of a rare event is rather unspecified. For example, we can model the number of death due to heart attacks using a Poisson regression, even if the number of deaths is large and is hardly rare. For example, Poisson regression has been used to model bacterial cell counts on a Petri dish. Now, bacterial cell counts can run to the billions. How can that be thought of as rare?

The reason for this is that the number itself is not the issue, but the count must pertain to different units of time or space. In the case of bacteria on a Petri dish, we can divide the surface area of the dish into smaller and smaller units so that we reach a unit of area in which only one bacterium is present. Thus, we don't actually model the total number of bacterium, but we model the number *per area*. This is an important concept.

Let us use another example. We might want to model the number of deaths in Shanghai in 2018. Data show that there were about 125,700 deaths in the municipality registered in that year. Does this violate the "rare event" requirement of the Poisson model? Not at all. This is because we can divide time into smaller and smaller units so that we reach a point in which only one death occurs per unit time and it is rare for large number of deaths to occur within that time. Thus, by varying the denominator, we can apply the Poisson regression model even to large numbers of counts.

The Poisson regression model uses the the natural logarithm as the link function. That is, the Poisson regression model is

$ln(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

The test and inferences on the Poisson model are carried out in a similar way as previously demonstrated.

Example 1

Table 1 shows the number of new drugs (`D`) that were released in the US market from 1992 to 2005 for 16 major diseases. Also included are the number of people affected by that condition (`P`) per 100,000 and the amount of money spent on research in those disease categories (`M`) in millions of US dollars in 1994.

Table. Data on new drugs.

Disease	D	P	M
Ischaemic heart disease	6	8976	198.4
Lung cancer	3	874	80.2
HIV/AIDS	21	1303	1049.6
Alcohol abuse	2	18092	222.6

Disease	D	P	M
Cerebrovascular disease	2	9467	108.5
Chronic obstructive pulmonary disease	1	4271	48.9
Depression	7	12785	149.5
Diabetes	13	37850	278.4
Osteoarthritis	5	12345	151.3
Drug abuse	1	4000	442.1
Dementia	9	8931	344.1
Asthma	3	15919	41.8
Colorectal cancer	2	1926	70.6
Prostate cancer	4	2020	40.1
Breast cancer	9	2262	159.5
Bipolar disorder	2	2418	35.0

We are interested in exploring the relationship between the number of drugs released to the market as a function of the prevalenc of the disease and the amount of funding spent. That is, we want to regress `D` on `P` and `M`.

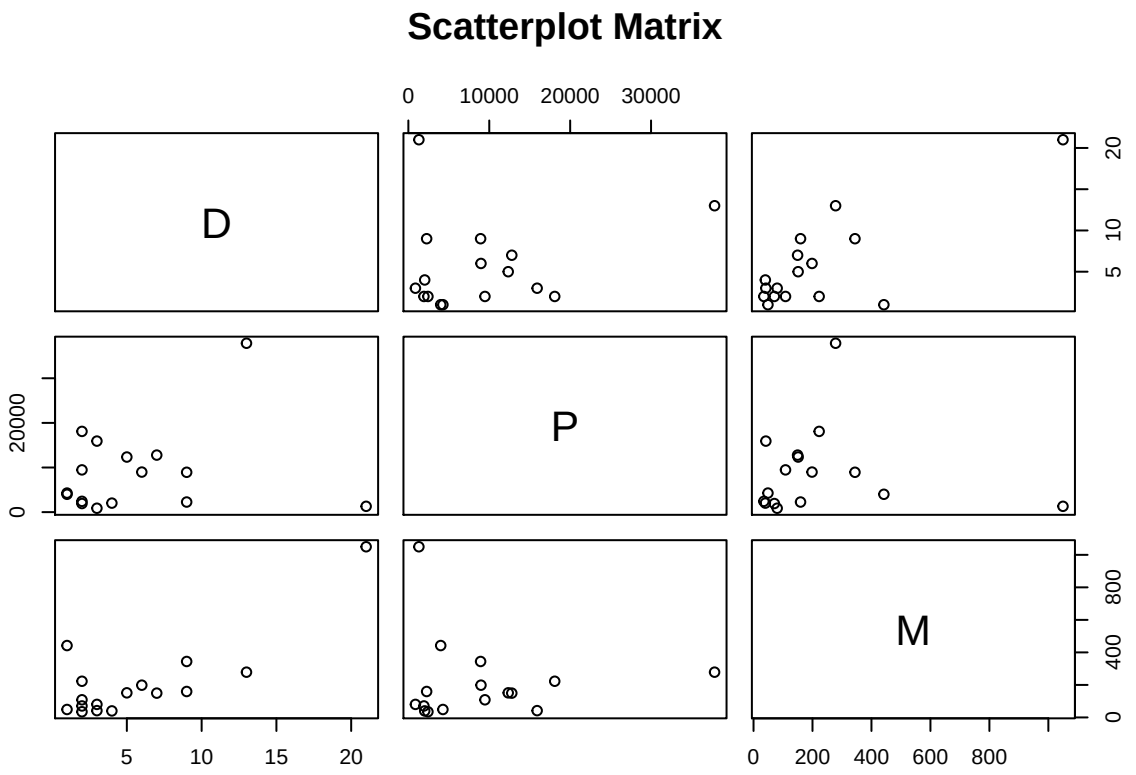
```
DISEASE <- c("Ischaemic heart disease", "Lung cancer", "HIV/AIDS", "Alcohol abuse",
             "Cerebrovascular disease", "Chronic obstructive pulmonary disease",
             "Depression", "Diabetes", "Osteoarthritis", "Drug abuse",
             "Dementia", "Asthma", "Colorectal cancer", "Prostate cancer",
             "Breast cancer", "Bipolar disorder")
D <- c(6, 3, 21, 2, 2, 1, 7, 13, 5, 1, 9, 3, 2, 4, 9, 2)
P <- c(8976, 874, 1303, 18092, 9467, 4271, 12785,
       37850, 12345, 4000, 8931, 15919, 1926, 2020,
       2262, 2418)
M <- c(198.4, 80.2, 1049.6, 222.6, 108.5, 48.9,
       149.5, 278.4, 151.3, 442.1, 344.1, 41.8,
       70.6, 40.1, 159.5, 35.0)
DRUGS <- data.frame(DISEASE, D, P, M)
str(DRUGS)
```

```
## 'data.frame':    16 obs. of  4 variables:
##  $ DISEASE: Factor w/ 16 levels "Alcohol abuse",..: 13 14 12 1 5 6 9 10 15 11 ...
##  $ D      : num  6 3 21 2 2 1 7 13 5 1 ...
##  $ P      : num  8976 874 1303 18092 9467 ...
##  $ M      : num  198.4 80.2 1049.6 222.6 108.5 ...
```

```
head(DRUGS)
```

```
##                                  DISEASE  D     P      M
## 1               Ischaemic heart disease  6  8976  198.4
## 2                           Lung cancer  3   874   80.2
## 3                              HIV/AIDS 21  1303 1049.6
## 4                         Alcohol abuse  2 18092  222.6
## 5               Cerebrovascular disease  2  9467  108.5
## 6 Chronic obstructive pulmonary disease  1  4271   48.9
```

```r
pairs(~D + P + M, data = DRUGS,
      main = "Scatterplot Matrix")
```



The Poisson regression is run using the `glm()` function, this time specifying the `family = poisson` option.

```r
DRUGS.LM <- glm(D ~ P + M, family = poisson, data = DRUGS)
summary(DRUGS.LM)
```

```
##
## Call:
## glm(formula = D ~ P + M, family = poisson, data = DRUGS)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6883  -0.6883   0.1103   0.6594   2.4389
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 8.778e-01  2.074e-01   4.233 2.31e-05 ***
## P           2.700e-05  9.508e-06   2.840  0.00451 **
## M           1.998e-03  3.008e-04   6.642 3.10e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 62.961  on 15  degrees of freedom
```

```
## Residual deviance: 24.077  on 13  degrees of freedom
## AIC: 82.14
##
## Number of Fisher Scoring iterations: 5
```

Interpretation of the model estimates

The results from Example 1 show that both the prevalence and the funding are statistically significantly related to the number of drugs brought to market. In particular,

- For every 100,000 people affected by a disease, the natural logarithm of the number of drugs brought to market increases by 0.000027, after holding the amount of funding constant.
- For every million US dollars spend on research funding, the natural logarithm of the number of drugs brought to market increases by 0.0020, after adjusting for the disease prevalence.

Now, this interpretation, while technically correct, if very difficult for a lay person to understand. This is the case, also, for logistic regression. So, as in the case of logistic regression, we need to back-transform the coefficients so that the results are understandable. We can do this easily by exponentiating the coefficients. The resulting values represent the **RATE OF CHANGE** in the outcome.

```
exp(coef(DRUGS.LM))
```

```
## (Intercept)           P           M
##    2.405498    1.000027    1.002000
```

- For every 100,000 people affected by the disease, the rate of drugs released to the market increases by 0.000027 or 0.0027% after controlling for research funding.
- For every million dollars in funding received, the rate of drugs released to the market increases by 0.0020 or 0.20% after controlling for prevalence.

I would adjust these interpretations so that the numbers are more useful.

- For every 100 million people affected by the disease, the rate of drugs released to market increases by 2.7%, after adjusting for research funding.
- For every 10 million dollars in funding received, the rate of drugs released to the market increases by 2.0%, after controlling for prevalence.

Diagnostic testing

As with linear and logistic regression, linearity, collinearity measured by variance inflation factors, leverage and Cook's distance need to be assessed. Goodness of fit tests, however, are quite important in the case of Poisson regression. This is because, as we learned in DPH101, the Poisson distribution on which this regression metehod is based as a strict requirement: that $\mu(Y) = \sigma^2(Y)$. That is, the mean and the variance must be equal to each other. Thus, we need to test this.

In the output for the model, this is given by the line showing the "Residual Deviance":

```
summary(DRUGS.LM)
```

```
##
## Call:
## glm(formula = D ~ P + M, family = poisson, data = DRUGS)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6883  -0.6883   0.1103   0.6594   2.4389
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) 8.778e-01  2.074e-01   4.233 2.31e-05 ***
## P           2.700e-05  9.508e-06   2.840  0.00451 **
## M           1.998e-03  3.008e-04   6.642 3.10e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 62.961  on 15  degrees of freedom
## Residual deviance: 24.077  on 13  degrees of freedom
## AIC: 82.14
##
## Number of Fisher Scoring iterations: 5
```

The rule of thumb is that the residual variance must be about equal to the degrees of freedom. That is, you need to compare 24.077 versus 13. In the above example, it is clear that the two values are not the same. In this case, we say that the data are *overdispersed* and that the Poisson model may not be appropriate.

Solutions to overdispersed models is beyond the scope of this module.

THE END