Regression and Linear Models (37252)

Lecture 10 – Logistic Regression I

Lecturer: Joanna Wang
Notes adopted from Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

Acknowledgement

These notes are based, in part, on earlier versions prepared by Dr Ed Lidums and Prof. James Brown.

10. Lecture outline

Topics:

- binary response variable
- modelling p using OLS
- modelling a transform of p as a linear function
- link functions
- modelling p using logistic regression
- simple logistic regression
  - continuous predictor
  - categorical predictor
  - confidence intervals
  - hypothesis tests
- examples
  - continuous predictor
  - categorical predictor

Parts of this discussion have been motivated by Chapter 1 of Hosmer and Lemeshow (2000).

10. Binary response variable

Up until now we have been constructing regression models with continuous, numerical response variables.

However, in many situations we are interested in modelling a **binary categorical variable** – a variable taking states such as "yes/no", "success/failure", "on/off", "accept/reject" etc.

Examples include regression models of:

- outcomes from medical treatment with explanatory variables such as dosage, age, sex, etc.
- results from applications for jobs with explanatory variables such as degree, experience etc.
- normal birth weight with explanatory variables gestational age etc.

We will see that the properties of binary response variables require a different approach to OLS/GLS regression, the logistic regression.

10. Binary response variable

Of course, regression is a numerical procedure and we cannot model categorical variables directly.

The solution is to adopt the approach we took for categorical predictors.

Instead of using the binary categorical variable directly, we substitute a numerical variable, defined as taking only two values (one for each of the binary states).

For use in logistic regression, it is necessary to define such **dummy variables** with the rule

$$Y = \begin{cases} 1, & \text{“success” with probability } p \\ 0, & \text{“failure” with probability } 1-p \end{cases}.$$

So Y is a **Bernoulli RV** with probability of success p.

Our aim is to model p, which we assume to be a function of the explanatory variables $x_1, \ldots, x_m$.

10. Modelling p using OLS

An early, and crude, approach is to assume a population model

$$Y = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \epsilon \tag{1}$$

from which the sample data $(X_{i,1}, \ldots, X_{i,m}, Y_i)$, $i \in \{1, \ldots, n\}$, are observations described by

$$Y_i = \beta_0 + \sum_{j=1}^{m} \beta_j X_{i,j} + \epsilon_i. \tag{2}$$

Assuming $\mathbb{E}[\epsilon] = 0$, a model can be fitted to the conditional expectation

$$\begin{aligned} \mathbb{E}[Y|x_1, \ldots, x_m] &= \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \mathbb{E}[\epsilon] \\ &= \beta_0 + \sum_{j=1}^{m} \beta_j x_j. \end{aligned}$$

10. Modelling p using OLS

If this model is consistent, the interpretation

$$p(x_1, \ldots, x_m) := \text{Prob}(Y = 1 | x_1, \ldots, x_m) = \mathbb{E}[Y | x_1, \ldots, x_m]$$

applies, given the sample response data $Y_i \in \{0, 1\}$.

So the population model (1) can also be stated as

$$Y = p(x_1, \ldots, x_m) + \epsilon. \tag{3}$$

Making the usual assumption

$$\epsilon_i \sim \text{N}(0, \sigma^2) \tag{4}$$

and independent, the least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ lead to the fitted model

$$\hat{p}(x_1, \ldots, x_m) = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_j.$$

10. Modelling a transform of p as a linear function

The statistical model for the sample data (2) means that assumption (4) can be restated as

$$Y_i \sim \mathrm{N}\left(\beta_0 + \sum_{j=1}^{m} \beta_j X_{i,j}, \sigma^2\right). \tag{5}$$

This assumption means the OLS model has two major problems:

- the response variable $Y \in \mathbb{R}$, not $Y \in \{0, 1\}$ as a Bernoulli RV should be.
- the variance of the response $\mathrm{var}(Y) = \sigma^2$, not $\mathrm{var}(Y) = p(1-p)$ as it should be for a Bernoulli RV.

A different approach is needed.

10. Modelling a transform of p as a linear function

Instead of modelling p with OLS as

$$p(x_1, \ldots, x_m) = \beta_0 + \sum_{j=1}^{m} \beta_j x_j$$

we model a function g of p as

$$g\big(p(x_1, \ldots, x_m)\big) = \beta_0 + \sum_{j=1}^{m} \beta_j x_j =: \eta(x_1, \ldots, x_m) \tag{6}$$

such that p can be recovered from the inverse of g as

$$p(x_1, \ldots, x_m) = g^{-1}\big((\eta(x_1, \ldots, x_m)\big). \tag{7}$$

So the population model (3) can be re-stated as

$$\begin{aligned} Y &= p(x_1, \ldots, x_m) + \epsilon \\ &= g^{-1}\big(\eta(x_1, \ldots, x_m)\big) + \epsilon. \end{aligned} \tag{8}$$

10. Modelling a transform of p as a linear function

In terms of the sample data $(X_{i,1}, \ldots, X_{i,m}, Y_i)$, $i \in \{1, \ldots, n\}$, we have the statistical model

$$\begin{aligned} Y_i &= p_i + \epsilon_i \\ &= g^{-1}(\eta_i) + \epsilon_i \end{aligned} \tag{9}$$

where

$$g(p_i) = \beta_0 + \sum_{j=1}^{m} \beta_j X_{i,j} =: \eta_i. \tag{10}$$

It remains to specify the errors. For this is it useful to revisit the population model

$$Y = p(x_1, \ldots, x_m) + \epsilon,$$

which is (8) re-stated.

Given that $Y \in \{0, 1\}$ the RV $\epsilon \in \{-p, 1 - p\}$ and so the errors ϵ_i must be assumed to be from a distribution consistent with this property.

Clearly, the assumption $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ does not meet this condition.

10. Modelling a transform of p as a linear function

The details of the error assumption and the estimation procedure applied to (9)-(10) we omit, but it involves a technique called **maximum likelihood estimation (MLE)**.

The term ''logistic regression'' comes from the representation (10), but the method of least squares is not the estimation techniques.

In any event, the estimated parameters $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ result in the fitted model for g

$$g\big(\hat{p}(x_1, \ldots, x_m)\big) = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_j =: \hat{\eta}(x_1, \ldots, x_m) \tag{11}$$

from which the fitted probabilities can be recovered from the inverse as

$$\hat{p}(x_1, \ldots, x_m) = g^{-1}\big(\hat{\eta}(x_1, \ldots, x_m)\big). \tag{12}$$

10. Link functions

The relationships (7) and (12) require the function g to satisfy certain asymptotic properties.

These properties are those of **cumulative distribution functions (CDFs)**, namely

$$\lim_{\eta \to -\infty} g^{-1}(\eta) = 0,$$

$$\lim_{\eta \to \infty} g^{-1}(\eta) = 1$$

and

$$g^{-1}(\eta) \in (0, 1)$$

for $\eta \in \mathbb{R}$.

Such functions are called **link functions**.

Link functions

Various link functions g are used including **probit**

$$g(p) = \text{probit}(p) := \Phi^{-1}(p) \quad (\Phi \text{ standard normal CDF}),$$

complementary log-log

$$g(p) = \text{cloglog}(p) := \ln\big(-\ln(1-p)\big),$$

log-log

$$g(p) = \text{loglog}(p) := -\ln\big(-\ln(p)\big)$$

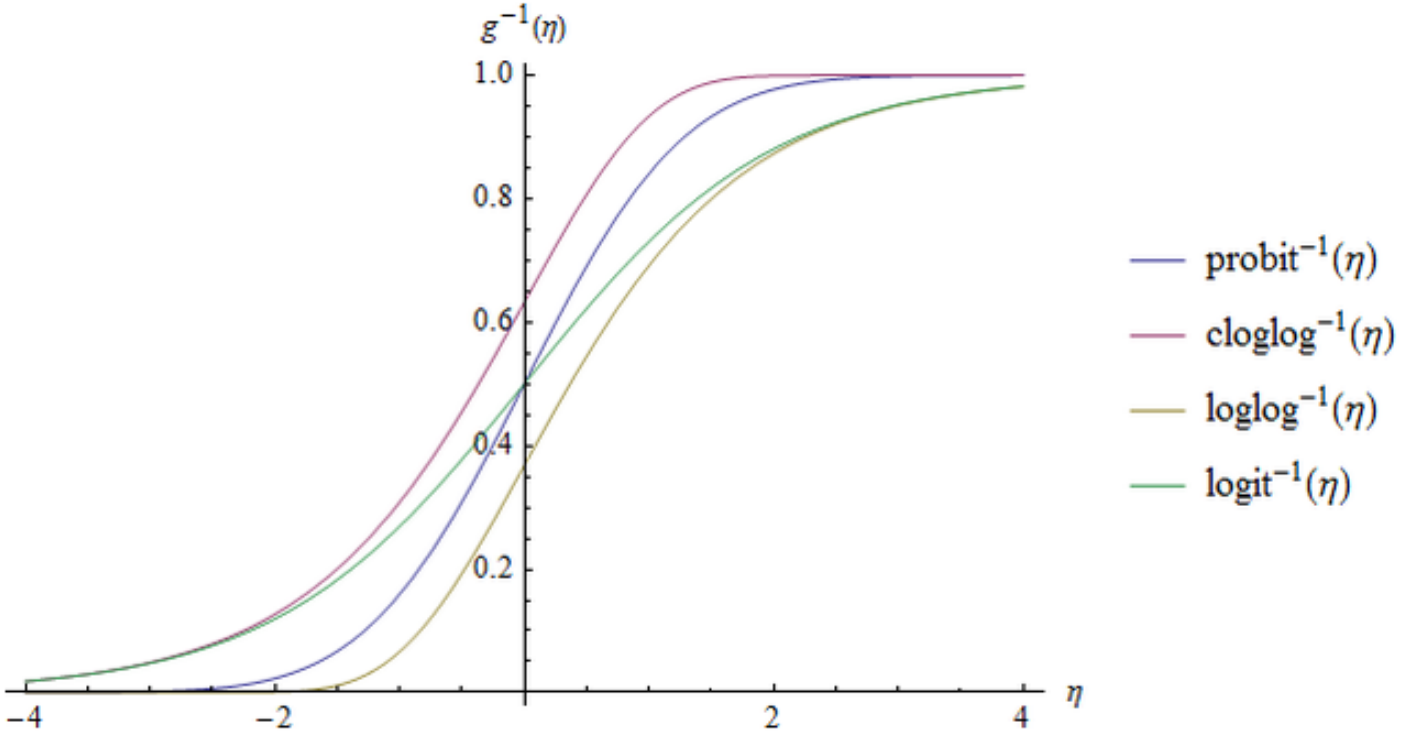
and **logit**

$$g(p) = \text{logit}(p) := \ln\left(\frac{p}{1-p}\right).$$

10. Link functions

The plot below shows that the inverse of these example link functions do behave as CDFs.

Plot of example inverse link functions g^{-1}



10. Modelling p using logistic regression

The link function that we will use is the logit function

$$g(p) = \text{logit}(p) \equiv \ln\left(\frac{p}{1-p}\right) \tag{13}$$

which we model with the **logistic regression equation**

$$\text{logit}\big(p(x_1, ..., x_m)\big) = \beta_0 + \sum_{j=1}^{m} \beta_j x_j := \eta(x_1, ..., x_m) \tag{14}$$

So this is a model of the **log-odds**

$$\ln\left(\frac{p}{1-p}\right) \tag{15}$$

from which the **odds**

$$\frac{p}{1-p} \tag{16}$$

or the variable of interest, p, can be obtained as

$$p(x_1, ..., x_m) = \text{logit}^{-1}\big(\eta(x_1, ..., x_m)\big) = \frac{1}{1 + e^{-\beta_0 - \sum_{j=1}^{m} \beta_j x_j}}. \tag{17}$$

10. Modelling p using logistic regression

The choice as to which outcome we define as "success" may seem arbitrary, but the log-odds for "failure" is the inverse of the log-odds for success.

Observe that is we take the negative of (14) we have

$$-\beta_0 - \sum_{j=1}^{m} \beta_j x_j = -\text{logit}(p) = -\ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{1-p}{p}\right)$$

where we recognise $1 - p$ as the probability of "failure", and so the last equation is the logistic regression model of the log-odds of "failure".

The choice of outcome labelled "success" will be the one most convenient for addressing the modelling question(s) we are trying to answer.

10. Simple logistic regression

Let's look more closely at the case of **simple logistic regression**, i.e. when $m = 1$. The model has the form

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \tag{18}$$

which, solving for p, results in the **logistic equation**

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-\beta_0-\beta_1 x}}. \tag{19}$$

Taking $x = 0$ gives the **baseline** or **reference** probability

$$p(0) = \frac{1}{1+e^{-\beta_0}} \tag{20}$$

which can also be interpreted as the solution of the **null** model ($m = 0$).

Baseline probabilities do not always make sense, as we will see in our example.

10. Simple logistic regression – continuous predictor

Let the independent variable x be continuous.

The value x when $p = 1/2$ is called the **median effective level**.

The median effective level is the value of x where there is a 50% chance of "success".

From (18) this is readily seen to be

$$x = -\frac{\beta_0}{\beta_1} \tag{21}$$

as

$$p\Big(-\frac{\beta_0}{\beta_1}\Big) = \frac{1}{1+e^{-\beta_0+\beta_1\frac{\beta_0}{\beta_1}}} = \frac{1}{1+e^{-\beta_0+\beta_0}} = \frac{1}{2}.$$

It should also be clear from (19) that

$$\lim_{\eta\to-\infty} p(\eta) = 0 \quad \text{and} \quad \lim_{\eta\to\infty} p(\eta) = 1.$$

10. Simple logistic regression – continuous predictor

On the **logit scale** we interpret β_1 as the increase in log-odds for a unit increase in x. To see this note that

$$\beta_0 + \beta_1(x+1) = \beta_0 + \beta_1 x + \beta_1 = \ln\left(\frac{p(x)}{1-p(x)}\right) + \beta_1.$$

On the **odds scale**, we interpret e^{β_1} as the multiple of the odds, i.e. the **odds ratio**, for a unit increase in x. To see this consider

$$e^{\beta_0+\beta_1(x+1)} = e^{\beta_0+\beta_1 x} e^{\beta_1} = \frac{p(x)}{1-p(x)} e^{\beta_1}.$$

Of course, $e^{\beta_1} - 1$ is the proportional change in the odds for a unit increase in x, positive when $\beta_1 > 0$ and negative when $\beta_1 < 0$.

So the effects of the independent variable are **linear** on the logit scale and **multiplicative** on the odds scale.

10. Simple logistic regression – continuous predictor

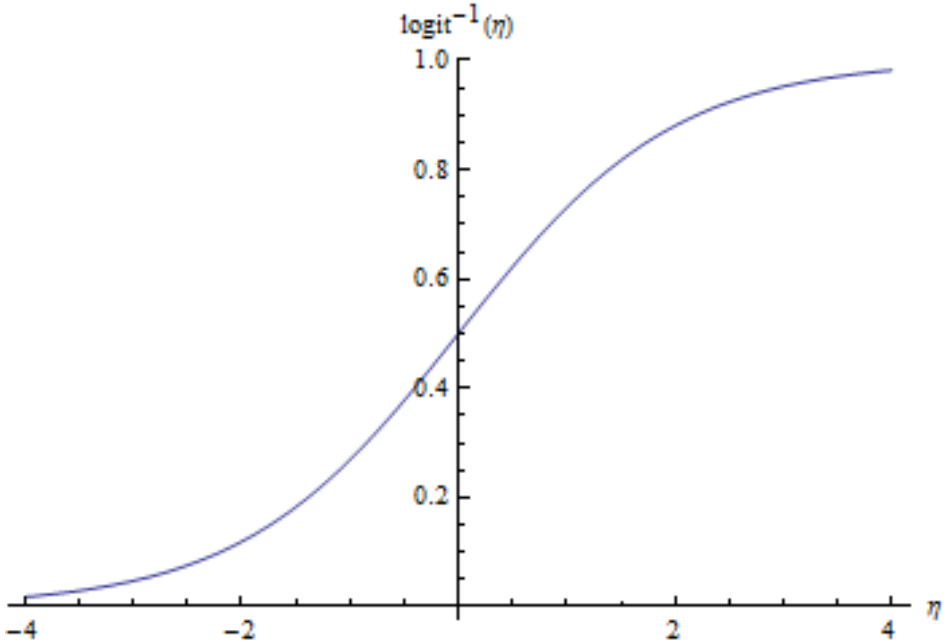
We can also consider the **instantaneous** or **marginal** change in p. Note that

$$p'(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \frac{1}{1+e^{\beta_0+\beta_1 x}} = p(x)\big(1-p(x)\big)$$

which is maximised when $p = 1/2$, that is the median effective level

$$\eta(x) = \beta_0 + \beta_1 x = 0 \quad \text{or} \quad x = -\frac{\beta_0}{\beta_1}$$

(see plot below).



10. Simple logistic regression – categorical predictor

Now let the independent variable x be a dummy variable for a categorical variable taking two states; i.e. we define x as

$$x = \begin{cases} 0, & \text{state A} \\ 1, & \text{state B} \end{cases} .$$

On the **logit scale**, when $x = 0$ we have the log-odds of state A

$$\ln\left(\frac{p(0)}{1-p(0)}\right) = \beta_0$$

and when $x = 1$ we have the log-odds for state B

$$\ln\left(\frac{p(1)}{1-p(1)}\right) = \beta_0 + \beta_1.$$

10. Simple logistic regression – categorical predictor

On the **odds scale**, when $x = 0$ we have the odds of success for state A

$$\frac{p(0)}{1-p(0)} = e^{\beta_0}$$

and when $x = 1$ we have the odds of success for state B

$$\frac{p(1)}{1-p(1)} = e^{\beta_0+\beta_1}.$$

The **odds-ratio** for state B over state A are

$$\frac{\frac{p(1)}{1-p(1)}}{\frac{p(0)}{1-p(0)}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

So the odds of success for state B are e^{β_1} times the odds of success for state A.

10. Simple logistic regression – confidence intervals

Let $S_{\hat{\beta}_1}$ be the standard error (sample standard deviation) of the parameter estimate $\hat{\beta}_1$.

Although not normal, we can use the approximation $\hat{\beta}_1 \sim \mathsf{N}(\beta_1, S^2_{\hat{\beta}_1})$.

This approximation allows the calculation of approximate confidence intervals on the true value of β_1. For instance, a $100(1-\alpha)\%$ two-sided confidence interval is given by

$$\hat{\beta}_1 - z_{1-\alpha/2} S_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + z_{1-\alpha/2} S_{\hat{\beta}_1} \tag{22}$$

where for $Z \sim \mathsf{N}(0,1)$ we have

$$\mathsf{Prob}(|Z| > z_{1-\alpha/2}) = \alpha.$$

10. Simple logistic regression – hypothesis tests

The approximation $\hat{\beta}_1 \sim \mathrm{N}(\beta_1, S^2_{\hat{\beta}_1})$ can also be used for hypothesis tests on the true value of β_1.

The following hypotheses are those tested and reported by R.

$$H_0\text{: } \beta_1 = 0$$
$$H_A\text{: } \beta_1 \neq 0.$$

We could use the test statistic

$$z_* = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

comparing it against a standard normal RV.

10. Examples - continuous predictor

Although the response variable is log-odds, we don't actually calculate this when building logistic regression models in R (or in other software packages).

Instead we pass the sample data $(X_{i,1}, \ldots, X_{i,m}, Y_i)$, $i \in \{1, \ldots, n\}$, into R and request a binary logistic regression.

Consider the data set *BirthWeightExample.csv*, available on Canvas, consisting of the dummy response variable

$$bwght = \begin{cases} 1 & \text{normal birth weight (“success”)} \\ 0 & \text{low birth weight (“failure”)} \end{cases}$$

and predictor *gage* (gestational age).

Define $p(gage) = \text{Prob}(bwght = 1|gage)$.

10. Examples - continuous predictor

The model we construct is

$$\text{logit}(\hat{p}) = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 \times gage$$

or

$$\hat{p}(gage) = \frac{1}{1 + e^{-\hat{\beta}_0 - \hat{\beta}_1 \times gage}}.$$

10. Examples - continuous predictor

```
> dat <- read.csv("~/2022_37252/Lecture_notes/Week11/Lecture/BirthWeigh
> mod1 <- glm(bwght ~ gage, family = "binomial", data = dat)
> summary(mod1)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6084  -0.3858   0.2324   0.4402   1.9120

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -48.9085    20.3382  -2.405   0.0162 *
gage          1.3127     0.5409   2.427   0.0152 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.975  on 23  degrees of freedom
Residual deviance: 16.298  on 22  degrees of freedom
AIC: 20.298
```

10. Examples - continuous predictor

The estimates $\hat{\beta}_0, \hat{\beta}_1$ give the fitted probability model

$$\hat{p}(gage) = \frac{1}{1 + e^{48.909 - 1.3131 \times gage}}.$$

We also see from the R output that the p-values for both $\hat{\beta}_0, \hat{\beta}_1$ are less than our preferred significance level 0.05, so we can reject the null hypotheses that $\beta_0 = 0$ and $\beta_1 = 0$ and conclude that $\beta_0 \neq 0$ and $\beta_1 \neq 0$.

Two-sided 95% confidence intervals for β_1 are also reported.

```
> confint.default(mod1)
                  2.5 %    97.5 %
(Intercept) -88.7707100 -9.046307
gage          0.2525393  2.372790
```

We can use the fitted model to predict the probability of normal birth weight; e.g., when $gage = 39$

$$\hat{p}(39) = \frac{1}{1 + e^{48.909 - 1.3131 \times 39}} \approx 0.91.$$

10. Examples - continuous predictor

The **baseline probability** (20) is

$$\hat{p}(0) = \frac{1}{1 + e^{-\hat{\beta}_0}} = \frac{1}{1 + e^{48.909}} \approx 0.0$$

which is the probability of being born with normal birth weight after a zero week gestation!

(One could argue that it is nonsensical to take $gage = 0$ in the model or alternatively that the model is doing its job by assigning such a small probability.)

The **median effective level** in (21) works out as

$$gage = -\frac{\hat{\beta}_0}{\hat{\beta}_1} = \frac{48.909}{1.313} \approx 37.2498$$

so $\hat{p}(37.2498) \approx 0.5$. This also the value of *gage* where the instantaneous change of $\hat{p}$ is maximised.

10. Examples - categorical predictor

Now consider the data set *GPvisitExample.csv*, available on Canvas, consisting of the dummy response variable

$$GP = \begin{cases} 1 & \text{frequent (“success”)} \\ 0 & \text{infrequent (“failure”)} \end{cases}$$

and dummy predictor *Gender*

$$Gender = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases} .$$

$GP = 1$ (frequent GP visits) is considered a “success” in the sense that it defines the state for which odds are modelled.

So define $p(Gender) = \text{Prob}(GP = 1|Gender)$.

10. Examples - categorical predictor

We can test for association between the categorical variables with a Chi-square independence test.

```
> library('gmodels')
> datGP <- read.csv("~/2022_37252/Lecture_notes/Week11/Lecture/GPvisitE
> CrossTable(datGP$Gender, datGP$GP, expected = T, chisq = T)
Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
```

10. Examples - categorical predictor

Total Observations in Table: 6223

	datGP$GP		
datGP$Gender	0	1	Row Total
0	2146	764	2910
	1854.581	1055.419	
	45.792	80.466	
	0.737	0.263	0.468
	0.541	0.339	
	0.345	0.123	
1	1820	1493	3313
	2111.419	1201.581	
	40.222	70.678	
	0.549	0.451	0.532
	0.459	0.661	
	0.292	0.240	
Column Total	3966	2257	6223
	0.637	0.363	

10. Examples - categorical predictor

We can test for association between the categorical variables with a Chi-square independence test.

```
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  237.1566     d.f. =  1     p =  1.639466e-53

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  236.3435     d.f. =  1     p =  2.466067e-53
```

With a test statistic of 237.157 the null hypothesis of independence is rejected.

So there is a statistically significant association to model.

10. Examples - categorical predictor

The R output for parameter estimates is below.

```
> mod2 <- glm(GP ~ Gender, family = "binomial", data = datGP)
> summary(mod2)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0945  -1.0945  -0.7804   1.2626   1.6355

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.03279    0.04213  -24.52   <2e-16 ***
Gender       0.83474    0.05472   15.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8151.5  on 6222  degrees of freedom
Residual deviance: 7911.0  on 6221  degrees of freedom
AIC: 7915
```

10. Examples - categorical predictor

Extracting these estimates $\hat{\beta}_0, \hat{\beta}_1$ gives the fitted probability model

$$\hat{p}(\mathit{Gender}) = \frac{1}{1 + e^{1.033 - 0.835 \times \mathit{Gender}}}.$$

The p values for both $\hat{\beta}_0, \hat{\beta}_1$ are less than our preferred significance level 0.05, so we can reject the null hypotheses that $\beta_0 = 0$ and $\beta_1 = 0$ and conclude that $\beta_0 \neq 0$ and $\beta_1 \neq 0$.

Two-sided 95% confidence intervals for β_1 are also reported.

```
> confint.default(mod2)
                 2.5 %     97.5 %
(Intercept) -1.1153651 -0.9502212
Gender       0.7274977  0.9419906
```

10. Examples - categorical predictor

As a dummy variable for gender, the model only makes sense when *Gender* equals zero or one.

Setting $Gender = 0$ in the fitted odds model

$$\frac{\hat{p}(Gender)}{1 - \hat{p}(Gender)} = e^{-1.033 + 0.835 \times Gender}$$

gives the predicted odds of "success" (frequent GP visits) for males

$$\frac{\hat{p}(0)}{1 - \hat{p}(0)} = e^{-1.033} \approx 0.3559$$

while setting $Gender = 1$ gives the predicted odds of "success" (frequent GP visits) for females

$$\frac{\hat{p}(1)}{1 - \hat{p}(1)} = e^{-1.033 + 0.835} \approx 0.8204.$$

Of course, the same results could be obtained from the cross-tab data. However, this model can be extended to included continuous variables, something a cross-tab approach can't handle.

References I

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd edition.