Topic 3

Weighted Least Squares and Logistic Regression

Peter Tea

Department of Mathematics and Statistics
University of Ottawa

Contents

1 Weighted least squares

1.1 Review of Ordinary Least Squares

Ordinary least squares (OLS) is often used in regression analysis to find estimates of parameters by minimizing the sums of squared residuals. The residual sums of squares (RSS) is given by:

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2$$
$$i = 1, 2, ..., n$$

Where y_i denotes the ith response variable, x_i denotes the ith predictor variable and β denotes the regressor parameters.

However, depending on the given data, it may not be appropriate to estimate parameters with the OLS method. The ordinary least squares method assumes that the general linear model has the form:

$$y = x\beta + \epsilon \tag{1}$$

where

- y is the (n x 1) vector of observed values for the response variables
- X is the (n x p) matrix of predictors (including the intercept)
- β is a (p x 1) matrix of regression parameters
- ϵ is the (n x 1) vector of errors

In this model, it is assumed that the error term (ϵ) is constant for all values of the response variable (i.e. homoscedastic errors) and that these errors are uncorrelated. As such, the variance-covariance matrix can be written as:

$$Cov(\epsilon) = \sigma^2 I_n$$

Where I_n is a (n x n) identity matrix. Thus, the variance-covariance matrix can also be written as:

$$Cov(\epsilon) = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

1.2 WLS: How it works

In many cases, assumptions of uncorrelated and homoscedastic errors, defined for (1), may not be valid. Thus, alternative methods should be considered in order to compute improved parameter estimates.

One alternative is weighted least squares (WLS). This method redistributes the influence of data points on the estimation of the parameters. This approach is consulted when there are *heteroscedastic* errors in the model (non-constant variances). Through the weighted least squares approach, observations with greater variances will have smaller weights, and conversely observations with smaller variances will have greater weights. With this method, the expression of the variance-covariance matrix is replaced in favour of a more general expression:

$$Cov(\epsilon) = \sigma^2 W^{-1}$$

where W is a positive definite matrix (*Fahrmeir 2013*).

W is defined as the diagonal matrix of weights as follows:

$$W = diag(w_1, w_2, ..., w_n)$$

or:

$$W = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{bmatrix}, W^{-1} = \begin{bmatrix} \frac{1}{w_1} & 0 & \dots & 0 \\ 0 & \frac{1}{w_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{w_n} \end{bmatrix}$$

Thus, the heteroscedastic variances can be written as $Var(\epsilon_i) = \dfrac{\sigma^2}{w_i}$

For the WLS approach, *transformation* of the response variable (Y), the regressor matrix (x) and the errors (ϵ) are performed such that the transformed variables follow the linear model with homoscedastic errors denoted in (1).

- Transformed errors: $\epsilon_i^* = \sqrt{w_i}\epsilon_i$

  Transformation of errors induces constant variance where $Var(\epsilon_i^*) = Var(\sqrt{w_i}\epsilon_i) = \sigma^2$

- Transformed response variable: $y_i^* = \sqrt{w_i}y_i$
- Transformed predictors: $x_{ik}^* = \sqrt{w_i}x_{ik}$

$i = 1, 2, ..., n$, and k is the number of regressors in the model.

Thus, the linear model (1) can be rewritten as:

$$W^{\frac{1}{2}}y = W^{\frac{1}{2}}x\beta + W^{\frac{1}{2}}\epsilon \tag{2}$$

Where

$$W^{\frac{1}{2}} = diag(\sqrt{w_1}, \sqrt{w_2}, ..., \sqrt{w_n})$$

1.3 Parameter Estimation: Part I

Regressor coefficient estimates are obtained by minimizing the weighted sums of squares (WSS):

$$WSS(\beta) = \sum_{i=1}^{n} \hat{e}_{wi}^2$$

$$WSS(\beta) = \sum_{i=1}^{n} w_i(y_i - x_i^{'}\beta)^2$$

$$i = 1, 2, ..., n$$

where the residuals of a weighted least squares model is defined as:

$$\hat{e}_{wi} = \sqrt{w_i}(y_i - x_i^{'}\beta)$$

Note that in matrix notation, the WLS function is written as

$$WSS(\beta) = (y - X\beta)'W(y - X\beta)$$

Example

For a simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

the weighted least squares estimates can be obtained as follows:

1. We define the weighted sums of squares:

$$WSS(\beta_0, \beta 1) = \sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_1 x_i)^2$$

2. We then differentiate $WSS(\beta_0, \beta 1)$ with respect to β_0 and $\beta 1$, and equate them to zero:

(i) Let $Q = \sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_1 x_i)^2$

(ii) $\dfrac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n} w_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$

(iii) $0 = -2\sum_{i=1}^{n} w_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$

(iv) $\hat{\beta}_0 \sum_{i=1}^{n} w_i + \hat{\beta}_1 \sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} w_i y_i$

The same procedure (with respect to β_1) is done to obtain:

$\hat{\beta}_0 \sum_{i=1}^{n} w_i x_i + \hat{\beta}_1 \sum_{i=1}^{n} w_i x_i^2 = \sum_{i=1}^{n} w_i x_i y_i$

3. Finally, we solve these equations for the weighted least squares estimate of β_0 and β_1. For example, it can be shown that the weighted estimator for β_1 is

$$\hat{\beta}_{1w} = \frac{\sum_{i=1}^{n} w_i x_i y_i}{\sum_{i=1}^{n} w_i x_i^2}$$

1.4 Parameter Estimation: Part II

Originally, with ordinary least squares, the estimator for β was:

$$\hat{\beta} = (X'X)^{-1}X'y$$

In the case where $W \neq I_n$, the ordinary least squares estimates (OLSE) are still unbiased,

$$E(\hat{\beta}) = \beta$$

however these estimates are no longer the best linear unbiased estimator of β. This is because, in this case, the OLSE has higher variability

$$Var(\hat{\beta}) = \sigma^2(X'X)^{-1}X'W^{-1}X(X'X)^{-1}$$

Now, for weighted least squares, it can be shown that these estimates are obtained by:

$$\hat{\beta}_w = (X'WX)^{-1}X'Wy$$

where $\hat{\beta}_w$ denotes the weighted least squares estimator.

With the derived WLS estimator, the linear model can be written as:

$$\hat{Y} = X\hat{\beta}_w = X(X'WX)^{-1}X'Wy = H_w y$$

where the weighted hat matrix, H_w is:

$$H_w = X(X'WX)^{-1}X'W$$

The diagonal elements of the hat matrix, h_{ii}, is called the leverage and represents a measure of the extremity of the ith predictor value relative to the mean of the predictor variable. The leverage can be written as

$$h_{ii} = [X(X'WX)^{-1}X'W]_{ii}$$

As h_{ii} approaches 1, the ith predicted response, $\hat{y}_i$ gets closer to the ith observed response y_i

It should be noted that if $w_i = 0$, then the matrix W will no longer have full rank. Therefore, W is no longer invertible (i.e. W is singular) and W^{-1} no longer exists. The WLS estimator, $\hat{\beta}_w$, may not exist since the matrix X'WX may no longer be invertible.

Tests and confidence intervals used for ordinary least squares approach can be easily adapted to weighted least squares. The work then lies upon obtaining the weight matrix, W. Depending on different scenarios, the matrix can be obtained by many different approaches.

1.5 Weight matrix for Replicate data

In general, smaller weights are assigned to the less precise measurements and greater weights are assigned to the more precise measurements when estimating the parameters of the model. If the data contains replicate data where observations can be grouped together (i.e. aggregate data), then the weight matrix can be easily obtained.

Let n_i denote the number of replicates for the predictor x_i and let j represent the number of unique predictor variables. Then, W is estimated as follows:

$$\hat{W} = diag(n_1, n_2, ..., n_j)$$

The variance of each response can then be written as:

$$Var(Y_i|x_i) = \frac{\sigma^2}{w_i} = \frac{\sigma^2}{n_i}$$

The larger the number of replicates (n_j) for a predictor variable, the greater its weight. Intuitively, the more replicate observations we have for a predictor, the more confidence there is in its variance estimation. Thus if we are more confident in the accuracy of a measured data point, we may wish to give such points more weight (more influence) on the regression model. Fitting regression models to regions where the background noise is small is more sensible and more reliable compared to fitting models where the background noise is large.

Example:
In a dose-response study, the drug Rosuvastin was tested on its effect on reducing cholesterol levels. In this study, only 6 unique doses of the drug was used. The response variable (% change in LDL cholesterol) was measured and the data was reported in response mean by dose, $\bar{Y_j}$, where $j = 1, 2, ..., 6$. Sample sizes, n_j, varied by doses. Assuming equal variances of individual patients, the variance for each dose is denoted by:

$$V(\bar{Y_j}) = \frac{\sigma^2}{n_j}$$

The weights are determined by the samples sizes, n_j.

1.6 Other approaches for estimating W

Another approach to obtain the weight matrix is to first perform ordinary least squares (W $= I_n$), and then use the the residuals of this model to provide an estimate, $\hat{W}$, of W. Thus, the weights are estimated as:

$$\hat{w}_i = \frac{1}{\sigma_i^2}$$

Another scenario involves the variance of the errors being proportional to some predictor x_i.

$$Var(\epsilon_i) \propto x_i$$

For example, if the variance of Y_i increases as the value of x_i increases, then the estimated weights

$$\hat{w}_i = \frac{1}{x_i}$$

may be a solution to account for these heteroscedastic errors.

1.7 An example in R

The following dataset explores the relationship between exercise level and the concentration of plasma cholesterol. The variable *exercise* is categorical and self reported. We would like to test whether the *plasma cholesterol concentration* differ for any of levels of *exercise*.

```
> data <-read.csv("heartdata.csv", header = TRUE, sep = ",")
> new.data <- na.omit(data) #Delete the missing values
> class(new.data$exercise)
[1] "integer"

> new.data$exercise <- as.factor(new.data$exercise)

> model.plasma <- lm(plasma.ch~ exercise, data = new.data)
```

```
#Set up the One-Way ANOVA

> summary(model.plasma)

Call:
lm(formula = plasma.ch ~ exercise, data = new.data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9320 -0.7364 -0.1085  0.6315  6.7415

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.12854    0.01231 416.575  < 2e-16 ***
exercise1    0.15193    0.03947   3.849 0.000119 ***
exercise2   -0.09217    0.02117  -4.354 1.34e-05 ***
exercise3   -0.14875    0.02033  -7.316 2.66e-13 ***
exercise4   -0.20657    0.02169  -9.523  < 2e-16 ***
exercise8    0.08776    0.15252   0.575 0.565019
exercise9   -0.32454    0.26650  -1.218 0.223321
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.031 on 18802 degrees of freedom
Multiple R-squared:  0.007984,Adjusted R-squared:  0.007667
F-statistic: 25.22 on 6 and 18802 DF,  p-value: < 2.2e-16
```

The regular ANOVA indicates that there is significance given the small p-value. However, there may be issues with some of the underlying assumptions:

```
> library("car")
> ncvTest(model.plasma)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 15.51721    Df = 1     p = 8.175733e-05
```

In the formal test of homogeneity of variances, the null hypothesis is rejected, meaning that we may be worried about the assumption of homoscedastic errors. One approach to proceed with is the weighted least squares approach. We will divide the weights for every factor level present in the categorical variable *exercise*. Let $j = 1, 2, ..., 7$ denote the 7 groups. Thus, $W_i = \frac{1}{\sigma_i^2}$.

```
###Weighted Least Squares
```

```
> standard.deviation <- tapply(new.data$plasma.ch, new.data$exercise, sd)
#Calculate the standard deviation for each group level.

> standard.deviation
#Check if it worked

       0         1         2         3         4         8         9
1.0394139 1.0894371 1.0536737 1.0101979 0.9969116 1.2562225 0.9084823

> new.data$Std.Dev = with(new.data, ifelse(exercise=="0", standard.deviation[1],

                  ifelse(exercise=="1", standard.deviation[2],

                    ifelse(exercise=="2", standard.deviation[3],

                      ifelse(exercise=="3", standard.deviation[4],

                        ifelse(exercise=="4", standard.deviation[5],

                          ifelse(exercise=="8", standard.deviation[6],

                            standard.deviation[7])))))))

#This code just adds the standard deviation(previously calculated) to the dataframe.

> model.weight <- lm(plasma.ch ~ exercise, data = new.data, weights=1/Std.Dev^2)
#Run new model, using reciprocal of sd^2 as the weight.

> summary(model.weight)

Call:
lm(formula = plasma.ch ~ exercise, data = new.data, weights = 1/Std.Dev^2)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-2.9411 -0.7125 -0.1044  0.6171  6.4858

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.12854    0.01241 413.227  < 2e-16 ***
exercise1    0.15193    0.04152   3.659 0.000254 ***
exercise2   -0.09217    0.02153  -4.280 1.88e-05 ***
```

```
exercise3   -0.14875    0.02013  -7.388 1.55e-13 ***
exercise4   -0.20657    0.02127  -9.714  < 2e-16 ***
exercise8    0.08776    0.18564   0.473 0.636388
exercise9   -0.32454    0.23490  -1.382 0.167099
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1 on 18802 degrees of freedom
Multiple R-squared:  0.007939,Adjusted R-squared:  0.007622
F-statistic: 25.08 on 6 and 18802 DF,  p-value: < 2.2e-16


> ncvTest(model.weight)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 2.650173e-06    Df = 1     p = 0.9987011
```

From the fitted model using the weighted approach, the formal homogeneity of variance test was not significant, meaning we have fixed the issue of heteroscedastic errors. However, the results for this model closely resemble the results obtained using ordinary least squares.

2 Logistic Regression

2.1 Categorical Response Data

Logistic regression is a model used when an analyst is interested in modelling the relationship between predictor variables and *categorical* response variables. For example, if we are interested in the outcome of heart disease (where the levels are only "diseased" and "non-diseased) related to some predictor variables, then a logistic model may be used.

More importantly, this model is used to predict the probability of occurrence of a categorical response from some predictor variables. For example, logistic regression allows an analyst to quantify how the presence of a risk factor affects the probability of contracting a disease. The predictor variables may be categorical, continuous or a mix of both types of variables.

As an extension to logistic regression, if there are 3 or more possible response outcomes then multinomial logistic regression is used. If these outcomes have a natural ordering, then ordinal logistic regression is used.

Since the response variables are categorical, typical regression models (eg. linear regression) cannot be used to predict the categorical responses. The reasons are as follows:

1. The response variables are not normally distributed (they follow Binomial distribution).

2. Using a linear model to predict responses, while making the assumption of constant variance, will not be appropriate.

   If
   $$Y_i = x_i'\beta + \epsilon_i$$
   where:

   - $Y_i = \begin{cases} 1 \text{ if ith subject has disease} \\ 0 \text{ otherwise} \end{cases}$
     $i = 1, 2, ...n$
   - x_i is a vector of predictors
   - β is a vector of regression coefficients

   With probabilities:
   $$Pr(Y_i = 1) = \pi_i$$
   $$Pr(Y_i = 0) = 1 - \pi_i$$

   Then the mean and variance of Y_i can be written as:
   $$E(Y_i) = \pi_i$$
   $$Var(Y_i) = (1 - \pi_i)(\pi_i)$$

   Therefore, any predictor variable which affects the mean of Y_i will affect its variance as well.

3. The response variables can only take values inside the range of (0,1). Using typical regression techniques (which are unbounded in the sense that they may produce values outside the specified range) may produce nonsensical predictions of the binary response variable.

2.2 Logit Transformation

In order to express π_i as a linear function of its predictor variables,
$$E(y_i) = \pi_i = x'\beta$$
where $0 \leq E(y_i) = \pi_i \leq 1$,
a transformation of π_i is required to remove the restriction of its range.

1. Firstly, we express π_i in terms of *odds*:
   $$odds = \frac{\pi_i}{1 - \pi_i}$$
   $$i = 1, 2, ..., n$$

2. Secondly, we take the natural logarithm of the odds. This is often times referred to as the *logit*.

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right)$$

It can be noted that logistic regression is a *generalized linear model* with a binomial distributed response variable and a logit *link function*. Generalized linear models allow response variables, which are not normally distributed, to be related to linear predictors using a link function. A link function, $g(\mu)$, is a function of the expected value of the response variable Y which is able to relate Y to some predictor variables through a linear equation.

- Let $\mu = E(Y)$ be the expected value of the response variable Y
- Let g$(\cdot)$ be the link function.

  g$(\cdot)$ relates μ to linear predictors of the form:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

  where k is the number of regressors.

In logistic regression, the link function is the logit:

$$g(\mu) = logit(\pi_i)$$

The purpose of the logit transformation is to map the expected value of the response, π_i, from the range (0, 1) to the entire real line *(Rodriguez 2007)*. This indicates that the logit (the link function of π_i) can now take on values from the range $(-\infty, +\infty)$. Thus, the responses can now be treated as a continuous variable.

In utilizing the logit of π_i, we can now model the expected value of our response variable with a linear expression of its predictors.

$$logit(\pi_i) = x_i^{'}\beta$$
$$i = 1, 2, ..., n$$

2.3 Interpretation of the model coefficients

The interpretation for the regression coefficients remain the same as the interpretation for any linear model, however the coefficients affect the logit rather than the expected values of the response Y. There is also an added interpretation if we consider the regression coefficients in terms of odds:

Consider the logit which can be written as:

$$exp\left\{log\left(\frac{\pi_i}{1-\pi_i}\right)\right\} = exp\{x_i^{'}\beta\}$$

which can then be simplified to:

$$\frac{\pi_i}{1-\pi_i} = exp(x_i^{'}\beta)$$

Now consider a logistic model with only one predictor:

$$\hat{g}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The fitted value at $x_i + 1$ is then written as:

$$\hat{g}(x_i + 1) = \hat{\beta}_0 + \hat{\beta}_1(x_i + 1)$$

The difference of the model at $x_i + 1$ and x_i is:

$$\hat{g}(x_i + 1) - \hat{g}(x_i) = \hat{\beta}_1$$

Recall that $\hat{g}(x_i)$ represents the log odds $\left(log\left(\frac{\pi_i}{1-\pi_i}\right)\right)$ at x_i. Therefore, the regressor coefficient, $\hat{\beta}_1$, represents the change in log odds between $x_i + 1$ and x_i.

$$\hat{\beta}_1 = log\left(\frac{\pi_i(x_i+1)}{1-\pi_i(x_i+1)}\right) - log\left(\frac{\pi_i(x_i)}{1-\pi_i(x_i)}\right) = log\left(\frac{\text{odds for } x_i + 1}{\text{odds for } x_i}\right)$$

Finally, with exponentiation we get the following result:

$$exp\{\beta_1\} = \left(\frac{\text{odds for } x_i + 1}{\text{odds for } x_i}\right)$$

The odds ratio (OR) is the ratio of the odds for $x_i + 1$ and x_i, and is interpreted as the increase in the probability of success associated with a one unit change in the value of the predictor variable.

The odds ratio is estimated as:

$$\hat{O}_R = exp\{\hat{\beta}_1\}$$

If the ith predictor is continuous, then for one unit increase in the ith predictor the odds increase by a multiplicative factor of $exp(\beta)$. If the ith predictor is a dichotomous categorical variable, then $exp(\beta)$ represents the odds ratio of these two factor levels (i.e. the difference in odds between these two levels).

2.4 Parameter estimation

Since the variances of a Binomial distributed response variable is not constant, least squares methods to estimate model parameters will not be appropriate. Instead, maximum likelihood estimation (MLE) will be used to generate these parameter estimates. Through maximum likelihood, values for the unknown parameters are obtained (iteratively) such that the probability of obtaining the observed set of data is maximized *(Hosmer 2013)*.

Since $Y_i \sim Bernouilli(\pi_i)$ with

$$Y_i = \begin{cases} 1 \text{ if event occurs} \\ 0 \text{ otherwise} \end{cases}$$

and

$$Pr(Y_i = 1) = \pi_i$$

The log likelihood function (LF) can be written as:

$$LF = \sum_{i=1}^{n} [y_i log(\pi_i) + (1 - y_i) log(1 - \pi_i)]$$

Through maximum likelihood estimation, it can be shown that the probability, π_i can be estimated by $\hat{\pi}_i$:

$$\hat{\pi}_i = \frac{1}{1 + e^{-(x_i' \hat{\beta})}}$$

Example in R:
The following dataset comes from a survey given to secondary school students. We would like to model the odds that a student has enrolled in extra tutoring in a math class (denoted by the variable "paid") given some predictors in the dataset. We first go through variable selection to determine a set of plausible regressor variables to include in the model.

```
> data <-read.csv("student-mat.csv", header = TRUE, sep = ",")

> model <-glm(paid ~ school + sex + age + address + famsize + Pstatus + Medu + Fedu +
studytime + traveltime + failures + schoolsup + famsup + activities + nursery + higher
+ internet + romantic,
 family = binomial(link = "logit"), data = data)
  #Run the logistic regression model with ALL variables

> model2 = step(model, direction = "backward", trace=FALSE)
#Perform a backwards elimination from the "saturated" model.

> formula(model2)
paid ~ sex + traveltime + failures + famsup + nursery + higher +
    internet

> library(car)
> vif(model2)
  #VIF >10 indicates high correlation with the other regressors.
              GVIF Df GVIF^(1/(2*Df))
sex        1.029729  1        1.014755
traveltime 1.073688  3        1.011920
failures   1.060479  3        1.009835
famsup     1.030063  1        1.014920
nursery    1.017480  1        1.008702
higher     1.021259  1        1.010574
internet   1.026811  1        1.013317
```

Since all VIF's appear to be lower than 10, there is no concerning issues of multicollinearity in the model. We move on to interpreting the model's parameter estimates:

```
> exp(confint(model2))
                   2.5 %      97.5 %
(Intercept) 0.0008827905   0.1340635
sex1        0.4498215193   1.0981443
traveltime2 0.9208299558   2.5602845
traveltime3 0.1557649977   1.3102712
traveltime4 0.0525182958   2.0275772
failures1   0.1845678875   0.8018716
failures2   0.1754544347   1.5807459
failures3   0.0382221109   1.2225864
famsup1     2.1225820643   5.4138351
nursery1    0.8750530379   2.6936652
higher1     1.8117954967 193.8674558
internet1   1.2495075050   4.5260885
```

It turns out that some of the regressors are not significant at the level of 5% (since the 95% confidence interval contains the value 1). We will discuss parameter estimation in the next chapter. However, we can still interpret some of the regressors which are significant.

For example, the variable **famsup1** (family's support of education, yes = 1 and no = 0) has an odds ratio confidence interval of (2.1225820643 , 5.4138351). The interpretation is that the odds of a student enrolled in extra tutoring lessons is at least 2.122 times greater when the student's family supports education compared to the odds when the student's family does not support education.
The interpretation of the other significant regressors are as follows:

- failures1: The odds that a student is enrolled in extra tutoring lessons is at least 18.45% smaller for students that have already failed one class compared to a student who has failed 0 classes.
- higher1: The odds that a student is enrolled in extra tutoring lessons is at least 81% greater if the student plans to pursue higher education compared to students who do not plan to pursue higher education.
- internet1: The odds that a student is enrolled in extra tutoring lessons is at least 25% greater if the student has access to internet at home, compared to a student who does not have access to the internet at home.

Had any of the variables in the model been continuous, then the interpretation of the exponentiated coefficient would have been an increase in odds of success given a one unit increase in that predictor variable.

2.5 Goodness of Fit

Goodness of Fit:
Testing for the significance of the regression coefficients is done by comparing 2 models: one null model without the variable in question, and another model that includes the variable in question. A regressor is significant if its inclusion in the model is able to increase the fit of the model to the observed data.

The degree to which these parameter estimates model the observed data can be investigated by examining the *deviance* of the model which represents the difference in the observed response to its predicted value. The deviance is analogous to the residual sums of squares (SSE) in linear regression. The deviance statistic is based on the log likelihood function and is computed as:

$$D = -2log(\text{likelihood of the fitted model})$$

and can written as:

$$D = -2\sum_{i=1}^{n}\left[y_i log\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)log\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right]$$

- y_i denotes the observed value
- $\hat{\pi}_i$ denotes the MLE of the fitted ith observation.

The smaller the deviance, the greater the fit of the model.

To assess the fit of the model, we compare the deviance with and without the regressor in question in the fitted model. This is given by the G (Goodness of fit) statistic.

$$G = -2log\left(\frac{\text{likelihood without the regressor}}{\text{likelihood with the regressor}}\right)$$

The G statistic represents the change in deviance when comparing the two models. This is referred as the likelihood ratio test.

The likelihood ratio test can be used to compare nested models when we consider whether the fit of the model is actually enhanced with some of its regressors. The null hypothesis that we would like to test is:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_j = 0$$

Under the null hypothesis the G statisic follows a chi square distribution with ν degrees of freedom:

$$G \sim \chi^2_\nu$$

An example in R:
Continuing with the previous example, the summary of the model is as follows:

```
> summary(model2)

Call:
glm(formula = paid ~ sex + traveltime + failures + famsup + nursery +
    higher + internet, family = binomial(link = "logit"), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7355  -1.0197  -0.3466   0.9818   2.3061

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.9819     1.1779  -3.380 0.000724 ***
sex1         -0.3516     0.2274  -1.546 0.122008
traveltime2   0.4244     0.2603   1.630 0.103057
traveltime3  -0.7419     0.5352  -1.386 0.165721
traveltime4  -0.9425     0.8835  -1.067 0.286063
failures1    -0.9320     0.3722  -2.504 0.012283 *
failures2    -0.6202     0.5529  -1.122 0.261964
```

```
failures3    -1.3057     0.8422  -1.550 0.121061
famsup1       1.2140     0.2385   5.091 3.56e-07 ***
nursery1      0.4233     0.2860   1.480 0.138839
higher1       2.3240     1.0802   2.151 0.031439 *
internet1     0.8516     0.3269   2.605 0.009183 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 544.83  on 394  degrees of freedom
Residual deviance: 465.97  on 383  degrees of freedom
AIC: 489.97

Number of Fisher Scoring iterations: 5
```

From the R output, we see that only some of the regressors are significant at the level of $\alpha = 0.05$. The null deviance is 544.83 and represents the fit of the null model (i.e. a model without regressors). With the fitted model, the deviance is reduced to 465.97.

It should be noted that since the variable "traveltime" has more than 2 levels, R creates design variables to accommodate the polychotomous variable. R performs a type of reference coding where each traveltime variable is relative to the first level. For example, the regressor coefficient travel3 represents the difference in log odds ratio (i.e. logit) between level 1 (<15 minutes commute) and level 3 (30 minutes commute).

The overall fit of this logistic regression model can be further analysed with a goodness of fit test:

```
> null <-glm(paid ~ 1, family = binomial(link = "logit"), data = data)
> anova(model2, null,
+       test="Chisq")
Analysis of Deviance Table

Model 1: paid ~ sex + traveltime + failures + famsup + nursery + higher +
    internet
Model 2: paid ~ 1
  Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
1       383     465.97
2       394     544.83 -11  -78.856 2.455e-12 ***
---
```

Given the low p-value (2.455e-12), we reject the null hypothesis that the fitted model is similar to the null model. Thus, our fitted model is indeed significant.

2.6 Wald Tests

Hypothesis testing on logit models can also be done using *Wald* tests which assumes an asymptotic normal distribution under the null hypothesis. This test is mostly used when there is a large sample size. We test the null hypothesis of

$$H_0 : \beta_j = 0$$

for some j regressor coefficient.
The Wald statistic is calculated as:

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim N(0, 1)$$

where:

- W is the Wald statistic
- $\hat{\beta}$ is the MLE of the regression coefficient
- $SE(\hat{\beta}_j)$ is the standard error of the regression coefficient

The confidence interval for the regression coefficient can be found as follows:

$$\hat{\beta} \pm Z_{1-\frac{\alpha}{2}}(SE(\hat{\beta}_j))$$

where:

- $Z_{1-\frac{\alpha}{2}}$ is the normal critical value

Exponentiating the endpoints of these intervals provide confidence intervals for the odds ratios.

$$exp\{\hat{\beta} \pm Z_{1-\frac{\alpha}{2}}(SE(\hat{\beta}_j)\}$$

The Wald test for each regressor coefficient can be evaluated as follows:

```
> library(aod)

> wald.test(b = coef(model2), Sigma = vcov(model2), Terms =6)
Wald test:
----------

Chi-squared test:
X2 = 6.3, df = 1, P(> X2) = 0.012
```

The output provides the wald statistic and its corresponding p-value. Again, we see that the regressor "failures1" is significant. The same p-value is obtained as the one previously seen from the summary output.

In the previous R output, the 95% confidence intervals were given for all regression estimates. Some did not include the value 1, and thus were significant. The regressors that did include the value 1 were not significant.

3 Sources

Agresti, Alan. (2007). An Introduction to Categorical Data Analysis. John Wiley & Sons, Inc.

Breheny, Patrick. (2013). Weighted least squares. Advanced Regression, University of Kentucky.

Eng, Ken, Chen, Yin-Yu, and Kiang, J.E. (2009), Users guide to the weighted-multiple-linear-regression program (WREG version 1.0): U.S. Geological Survey Techniques and Methods, book 4, chap. A8, 21 p. URL: http://pubs.usgs.gov/tm/tm4a8.)

Fahrmeir, Ludwig, Kneib, Thomas, Lang, Stefan, Marx, Brian. Regression Models, Methods and Applications. (2013). Springer.

Hosmer, David W., Stanley Lemeshow, and Rodney Sturdivant.(2013). Applied Logistic Regression. John Wiley & Sons, Inc.

Rodriguez, G. (2007). Lecture Notes on Generalized Linear Models. URL: http://data.princeton.edu/w