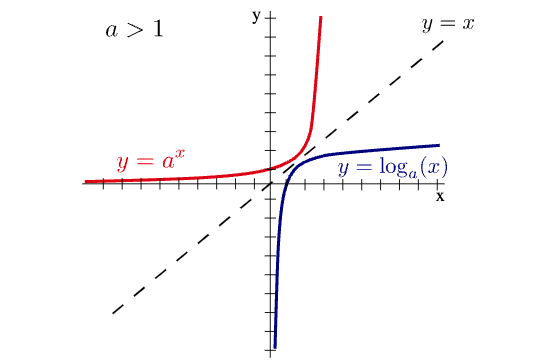
**Exponential and logarithmic functions**

A logarithmic function reverses the action of the corresponding exponential function:

**If and only if:**

**Graphically, we can represent both functions like this:**



**Where: a>0 and different from 0.**

**Theory**

The idea of logistic regression is similar to that of multiple regression. The main difference is that the dependent variable is qualitative and only takes two values. For example, a bank might be interested in knowing if a customer will be approved for an insurance policy depending on some variables. The dependent variable in this case takes only two forms: either the customer is approved or rejected from the program. The main purpose of logistic regression is to develop a model that predicts the probability that a customer will be approved based on some initial values for the predictors.

With only one predictor, the logistic regression equation takes the following form:

if there are several predictors:

With:

Ln(odds) =

odds:

odds =

why can’t we use linear regression?

Because the assumption of linearity is violated when the predicted variable is categorical. One way around this issue is to transform the data using logarithmic transformation.

How are the coefficients found?

Maximum likelihood estimation is used in an iterative process to find best values for the predictors that can estimate the scores for the dependent variable.

**Model assessment**

**Log-likelihood**

As in multiple regression, we can use the observed and fitted values to assess the fit of the model. The measure we use is the log-likelihood which is analogous to the residual sum of squares. It answers the question of how much unexplained variance remains after the model has been fitted to the data. The measure sums the probabilities for the predicted and actual outcomes:



It follows that small values are better.

**Deviance (-2LL)**

The deviance is convenient because it follows a chi-square distribution. It also can serve as a measure to compare fitness between models. Smaller values are better.

Deviance= -2 \* log-likelihood

In multiple regression, we compare the best model with the “baseline model”, the mean of scores. In logistic regression, since the mean is insignificant, we would use the outcome category that has more frequency as baseline model-that is, the model when only the constant is included. We can use this simple principle to compute the improvement in model fit after adding predictors or the likelihood ratio:

**R and R2**

We can calculate correlation coefficients for each predictor. They represent partial correlations between each predictor and the outcome. Their interpretation is done in similar ways to the simple correlation coefficient.

Z2 is the Wald statistic for the predictor. Because this formula has the Wald statistic, which can be inaccurate, it should be treated with some caution.

Concerning R2, an easy measure is presented by Hosmer & Lemeshow (1989) :

It provides a gauge of the improvement of fit in the model.

**Information criterion**

These measures (AIC and BIC) are used to solve problems that R2 brings: with more predictors R2 tend to get bigger. Thus, these measures penalize models that have more predictors. Less values are better.

**The z-statistic**

The contribution of each predictor can be assessed with the z-statistic. It is calculated exactly in the same ways as the t-statistic for multiple regression. It follows a normal distribution.

**The odd-ratio**

The most important interpretation is the odd-ratio. It is the exponential value of the coefficient. It is an indicator of the change in odds of the outcome resulting from one unit change in the predictor.

The odds are the ratio of the probability of an event occurring divided by the probability of the event not occurring.

Example:

What is the odds ratio for this sample?



Let’s consider that purchase behavior is a factor with two levels: 0 for not bought, 1 for bought

Let’s consider gender as a factor with these levels: 0 for male, 1 for female

R provides us with the following output:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.2528 0.8018 1.562 0.118

XFemale **-2.8622** 1.3575 -2.108 0.035 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 20.728 on 14 degrees of freedom

Residual deviance: 14.941 on 13 degrees of freedom

AIC: 18.941

How to calculate the Odds ratio manually?

prob\_buy\_0 = 1/(1+exp (-(1.2528-2.8622\*0)))

inverse1 = 1- prob\_buy\_male

odds1 = prob\_buy\_male/inverse1

odds1 = 3.5

# with one unit change in the predictor

prob\_buy\_1 = 1/(1+exp(-(1.2528-2.8622\*1)))

inverse2 = 1 - prob\_buy\_1

odds2 = prob\_buy\_1/inverse2

Odds2 = 0.2

Odds ratio = odds2/odds1 = **0.057**

What does it mean?

If we calculate the odds ratio for purchasing, going from male to female, we get 0.057 which is equal to exp (-2.8622). **It means that as we go from male to female (0 to 1, or one unit change in the predictor), the probability of purchasing behavior (outcome) drops by 0.057 (bought to not bought) because it is less than 1.**

**Assumptions**

As in linear regression, some assumptions must be checked for logistic regression.

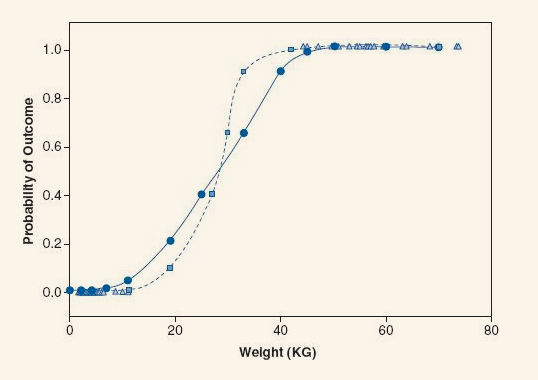
**Linearity:** there is a linear relationship between the logit of the outcome and the continuous predictors.

How to test: interaction between the predictor and its log transformation is significant.

**Independence of errors**: cases should not be related

**Multicollinearity**: we may use VIF and tolerance

Sometimes, logistic regression can have problems of its own. This happens if a solution isn’t plausible and so R usually provides large standard errors. This may be due to problem related to the ratio of cases to variables: incomplete information from the predictors (some categories are empty), complete separation (predictors perfectly predict the outcome, no overlap between values of the predictor between the two categories of the outcome).



**Casewise analysis**

* Predicted probabilities:

Cured Intervention Duration predicted\_prob

1 Not Cured No Treatment 7 0.4285714

2 Not Cured No Treatment 7 0.4285714

3 Not Cured No Treatment 6 0.4285714

4 Cured No Treatment 8 0.4285714

5 Cured Intervention 7 0.7192982

6 Cured No Treatment 6 0.4285714

As can be seen, the values provided give us strong evidence that having the intervention enhances the chance of getting better.

* Residuals

stand\_resid student\_resid dfbeta.(Intercept) dfbeta.InterventionIntervention leverage

1 -1.0675117 -1.0643627 -3.886912e-02 0.03886912 0.01785714

2 -1.0675117 -1.0643627 -3.886912e-02 0.03886912 0.01785714

3 -1.0675117 -1.0643627 -3.886912e-02 0.03886912 0.01785714

4 1.3135473 1.3110447 4.782751e-02 -0.04782751 0.01785714

5 0.8189783 0.8160435 -3.039582e-17 0.03225994 0.01754386

6 1.3135473 1.3110447 4.782751e-02 -0.04782751 0.01785714

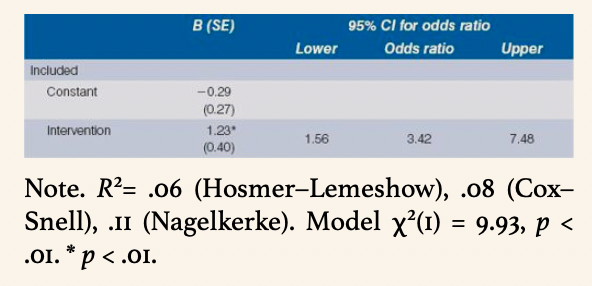
Std. Residuals: only 5% should lie outside ± 1.96. values ±3 are cause for concern.

Leverage: 0 is no influence and 1 is complete influence. Expected leverage is (k+1)/n, in this case it is 0.018.

Dfbeta: values should be less than 1.

The diagnostics in our example seem to be ok.

**How to report logistic regression**



**How to report multinomial logistic regression**

