



The Odds Ratio

by Katrien Declercq

Academic

The concept of the odds ratio will be explained using the same example as in the paper on logistic regression. Suppose a company that sells products via a catalog wants to identify those customers to whom advertising efforts should be directed. It has been decided that customers who spend 100 dollars or more are the target group. Based on orders received over the last six months, the company wants to characterize this group of customers. For each order, the data available are the purchase price (coded as 1 if purchase \$100 or more, 0 if purchase under \$100), the age of the customer in years, the gender of the customer and the annual income (Low, Middle, High). To help the company to answer their question, an analysis has to be done to identify what factors (age, gender, income) determine whether a customer purchases for \$100 or more (i.e. shows the response 1). For this explanation, we will focus on the influence of the gender on the amount customers spend.

Data are available on 431 customers (dataset in newsletter of last month) and a table of gender (F(emale) or M(ale)) by purchase ($\geq 100\$$ or $< 100\$$) shows that out of the 240 women in the trial, 101 bought for 100\$ or more (i.e.42%) and out of the 191 men, 61 spent 100\$ or more (i.e. 32%). So, apparently, the chance to spend 100\$ or more seems to be higher for women than for men (surprisingly enough!). The odds ratio is a measure that can be used to describe this association. For both groups separately, we can calculate the ratio of the probability to spend 100\$ or more to the probability to spend less than 100\$ and this value is called the odds. For the females in the example, the odds of spending 100\$ or more is calculated as $0.42/0.58$ which is 0.73. For the males in the example, the odds of spending 100\$ or more is $0.32/0.68$, which is 0.47. So, the odds of spending 100\$ or more is higher in women than in man. By taking the ratio of these odds, we can calculate the odds ratio (OR) of spending 100\$ or more for women relative to men as $0.73/0.47$ which is 1.55.

The odds ratio shows the strength of association between a predictor and the response of interest. It can vary from 0 to infinity. If the odds ratio is one, there is no association. In our example this would mean that the ratio of the probability to purchase for 100\$ or more to the probability to purchase for less than 100\$ is the same for women and men. However, we found an OR of 1.55, indicating that women are more likely to spend 100\$ or more than men. If the OR would have been less than one, then the men would have been more likely to spend 100\$ or more than women.

The OR and logistic regression are very closely related. The logistic regression model with only gender as explanatory variable can be written as $\text{logit}(p) = \alpha + \beta * X$, where $X = 1$ for females and 0 for males (if we use reference cell coding as we did in last months newsletter) and p is the

probability to spend 100\$ or more. This means that for women, the $\text{logit}(p) = \alpha + \beta$ and for men it is α . Now, since the logit is $\log(p/(1-p))$, it is the log of the odds. So, the log of the OR for purchasing 100\$ or more of females compared to males is the log of the odds for females - the log of the odds for males, so the logit for females - the logit for males, so it is equal to β and an estimate for the OR itself can be found as $\exp(\beta)$ which is $\exp(0.4373) = 1.55$.

So, the parameter estimates of a logistic regression can be interpreted easily in terms of odds ratios. The advantage of this measure of association is that it is independent of the way in which the data were collected. If the explanatory variable has more than two levels, the parameter estimates can be interpreted by calculating more odds ratios (comparing the groups two by two). If more explanatory variables are present in a model, the odds ratios for one predictor may be calculated keeping all other predictors at a fixed level. Of course, this only makes sense if there are no interactions going on. If that were the case, the interactions should be interpreted instead of the main effects. For a continuous explanatory variable, the odds ratio corresponds to a unit increase in the explanatory variable.

Odds ratios can be calculated with PROC FREQ by specifying the CMH option in the TABLES statement. The odds ratio value is then listed beside "Case-control" in the section labeled "Estimates of the Relative Risk (Row1/Row2)." The confidence intervals are labeled "Confidence bounds" and are 95% confidence intervals by default (can be changed using the ALPHA= option in the TABLES statement). When performing a logistic regression with PROC LOGISTIC, the "Odds Ratio Estimates" table contains the odds ratio estimates and the corresponding 95% Wald confidence intervals. Also with PROC GENMOD, the logistic regression can be performed and the odds ratio can be found using the ESTIMATE statement with the EXP option. Also a 95% confidence interval for the OR is calculated.

```
PROC SORT DATA=b_sales;
  BY descending purchase gender;
RUN;

PROC FREQ DATA=b_sales order=data;
  TABLES gender*purchase/cmh;
RUN;

PROC LOGISTIC DATA=b_sales descending;
  CLASS gender (param=ref ref='M');
  MODEL purchase = gender;
  TITLE 'LOGISTIC REGRESSION MODEL: Purchase = Gender';
RUN;

PROC GENMOD DATA=b_sales descending;
  CLASS gender;
  MODEL purchase = gender / dist=binomial link=logit;
  ESTIMATE 'Female vs. Male' gender 1 -1 / e exp;
RUN;
```