---

This page shows an example of logistic regression with footnotes explaining the output. The data were collected on 200 high school students, with measurements on various tests, including science, math, reading and social studies. The response variable is high writing test score (**honcomp**), where a writing score greater than or equal to 60 is considered high, and less than 60 considered low; from which we explore its relationship with gender (**female**), reading test score (**read**), and science test score (**science**). The dataset used in this page can be downloaded from SAS Web Books Regression with SAS (https://stats.idre.ucla.edu/sas/webbooks/reg/).

```
proc logistic data= logit descending;
 model honcomp = female read science;
run;
```

```
The LOGISTIC Procedure

              Model Information
Data Set                      WORK.LOGIT
Response Variable             honcomp
Number of Response Levels     2
Model                         binary logit
Optimization Technique        Fisher's scoring

Number of Observations Read         200
Number of Observations Used         200

          Response Profile
 Ordered                      Total
   Value      honcomp     Frequency
       1            1            53
       2            0           147
Probability modeled is honcomp=1.
```

```
         Model Fit Statistics
                              Intercept
              Intercept             and
Criterion          Only      Covariates
AIC             233.289         168.236
SC              236.587         181.430
-2 Log L        231.289         160.236

        Testing Global Null Hypothesis: BETA=0
Test                 Chi-Square       DF     Pr > ChiSq
Likelihood Ratio        71.0525        3          ChiSq
Intercept     1    -12.7772      1.9759       41.8176
```

## Model Information

```
Model[d]                            binary logit
Optimization Technique[e]           Fisher's scoring

Number of Observations Read[f]            200
Number of Observations Used[f]            200

          Response Profile
 Ordered                      Total
   Value[g]     honcomp[g]    Frequency[h]
       1            1             53
       2            0            147
Probability modeled is honcomp=1.[i]
```

a. **Data Set** – This the data set used in this procedure.

b. **Response Variable** – This is the response variable in the logistic regression.

c. **Number of Response Levels** – This is the number of levels our response variable has.

d. **Model** – This is the type of regression model that was fit to our data. The term logit and logistic are exchangeable.

e. **Optimization Technique** – This refers to the iterative method of estimating the regression parameters. In SAS, the default is method is Fisher's scoring method, whereas in Stata, it is the Newton-Raphson algorithm. Both techniques yield the same estimate for the regression coefficient; however, the standard errors differ between the two methods. For further discussion, see Regression Models for missing values for any variables in the equation. By default, SAS does a listwise deletion of incomplete cases.

g. **Ordered Value** and **honcomp** – **Ordered value** refers to how SAS orders/models the levels of the dependent variable. When we specified the **descending** option in the **procedure** statement, SAS treats the levels of **honcomp** in a descending order (high to low), such that when the logit regression coefficients are estimated, a positive coefficient corresponds to a positive relationship for high write status, and a negative coefficient has a negative relationship with high write status. Special attention needs to be placed on the ordered value since it can lead to erroneous interpretation. By default SAS models the 0s, whereas most other statistics packages model the 1s. The **descending** option is necessary so that SAS models the 1's.

h. **Total Frequency** – This is the frequency distribution of the response variable. Our response variable has 53 observations with a high write score and 147 with a low write score.

i. **Probability modeled is honcomp=1** – This is a note informing which level of the response variable we are modeling. See superscript g for further discussion of the **descending** option and its influence on which level of the response variable is being modeled.

---

## Model Fit Statistics

| Criterion[k] | Intercept Only[l] | Intercept and Covariates[m] |
|---|---|---|
| AIC | 233.289 | 168.236 |
| SC | 236.587 | 181.430 |
| -2 Log L | 231.289 | 160.236 |

Testing Global Null Hypothesis: BETA=0

| Test[n] | Chi-Square[o] | DF[o] | Pr > ChiSq[o] |
|---|---|---|---|
| Likelihood Ratio | 71.0525 | 3 | |

j. **Model Convergence Status** – This describes whether the maximum-likelihood algorithm has converged or not, and what kind of convergence criterion is used to assess convergence. The default criterion is the relative gradient convergence criterion **(GCONV)**, and the default precision is $10^{-8}$.

k. **Criterion** – Underneath are various measurements used to assess the model fit. The first two, Akaike Information Criterion (**AIC**) and Schwarz Criterion (**SC**) are deviants of negative two times the Log-Likelihood (**-2 Log L**). **AIC** and **SC** penalize the log-likelihood by the number of predictors in the model.

**AIC –** This is the Akaike Information Criterion. It is calculated as AIC = -2 Log L + 2(($k$-1) + $s$), where $k$ is the number of levels of the dependent variable and $s$ is the number of predictors in the model. **AIC** is used for the comparison of nonnested models on the same sample. Ultimately, the model with the smallest **AIC** is considered the best, although the **AIC** value itself is not meaningful.

**SC –** This is the Schwarz Criterion. It is defined as – 2 Log L + (($k$-1) + $s$)*log(Σ $f_i$), where $f_i$'s are the frequency values of the $i^{th}$

itself is not meaningful.

l. **Intercept Only** – This column refers to the respective **criterion** statistics with no predictors in the model, i.e., just the response variable.

m. **Intercept and Covariates –** This column corresponds to the respective **criterion** statistics for the fitted model. A fitted model includes all independent variables and the intercept. We can compare the values in this column with the criteria corresponding **Intercept Only** value to assess model fit/significance.

n. **Test** – These are three asymptotically equivalent Chi-Square tests. They test against the null hypothesis that at least one of the predictors' regression coefficient is not equal to zero in the model. The difference between them are where on the log-likelihood function they are evaluated. For further discussion, see Categorical Data Analysis, Second Edition (http://www.stat.ufl.edu/~aa/cda/cda.html), by Alan Agresti (pages 11-13).

**Likelihood Ratio –** This is the Likelihood Ratio (LR) Chi-Square test that at least one of the predictors' regression coefficient is not equal to zero in the model. The LR Chi-Square statistic can be calculated by -2 Log L(null model) – 2 Log L(fitted model) = 231.289-160.236 = 71.05, where L(null model) refers to the **Intercept Only** model and L(fitted model) refers to the **Intercept and Covariates** model.

**Score** – This is the Score Chi-Square Test that at least one of the predictors' regression coefficient is not equal to zero in the model.

**Wald** – This is the Wald Chi-Square Test that at least one of the predictors' regression coefficient is not equal to zero in the model.

o. **Chi-Square, DF** and **Pr** > **ChiSq –** These are the **Chi-Square** test statistic, Degrees of Freedom (**DF**) and associated p-value (**PR>ChiSq**) corresponding to the specific **test** that all of the predictors are simultaneously equal to zero. We are testing the probability (**PR>ChiSq**) of observing a **Chi-Square** statistic as extreme as, or more so, than the observed one under the null hypothesis; the null hypothesis is that

## Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates

| Parameter[p] | DF[q] | Estimate[r] | Standard Error[s] | Wald Chi-Square[t] | Pr > ChiSq[t] |
|---|---|---|---|---|---|
| Intercept | 1 | -12.7772 | 1.9759 | 41.8176 | <.0001 |
| female | 1 | 1.4825 | 0.4474 | 10.9799 | 0.0009 |
| read | 1 | 0.1035 | 0.0258 | 16.1467 | <.0001 |
| science | 1 | 0.0948 | 0.0305 | 9.6883 | 0.0019 |

Odds Ratio Estimates

| Effect[u] | Point Estimate[v] | 95% Wald Confidence Limits[w] | |
|---|---|---|---|
| female | 4.404 | 1.832 | 10.584 |
| read | 1.109 | 1.054 | 1.167 |
| science | 1.099 | 1.036 | 1.167 |

in the model.

r. **Estimate** – These are the binary logit regression estimates for the **Parameters** in the model. The logistic regression model models the log odds of a positive response (probability modeled is honcomp=1) as a linear combination the predictor variables. This is written as **log[ p / (1-p) ] = b0 + b1*female + b2*read + b3 *science**,

where p is the probability that **honcomp** is 1. For our model, we have, **log[ p / (1-p) ] = -12.78 + 1.48*female + 0.10*read + 0.09*science**.

We can interpret the parameter estimates as follows: for a one unit change in the predictor variable, the difference in log-odds for a positive outcome is expected to change by the respective coefficient, given the other variables in the model are held constant.

**Intercept** – This is the logistic regression estimate when all variables in the model are evaluated at zero. For males (the variable **female** evaluated at zero) with zero **read** and **science** test scores, the log-odds for high write score is -12.777. Note that evaluating **read** and **science** at zero is out of the range of plausible test scores. If the test scores were mean-centered, the intercept would have a natural interpretation: the expected log-odds for high write score for males with an average **read** and **science** test score.

**female** – This is the estimated logistic regression coefficient comparing females to males, given the other variables are held constant in the model. The difference in log-odds is expected to be 1.4825 units higher for females compared to males, while holding the other variables constant in the model.

**read** – This is the estimate logistic regression coefficient for a one unit change in **read** score, given the other variables in the model are held constant. If a student were to increase her **read** score by one point, her difference in log-odds for high write score is expected to increase by 0.10 unit, given the other variables in the model are held constant.

**science** – This is the estimate logistic regression coefficient for a one unit change in **science** score, given the other variables in the model are held constant. If a student were to increase her **science** score by one point, the difference in log-odds for high write score is

t. **Chi-Square** and **Pr** > **ChiSq** – These are the test statistics and p-values, respectively, testing the null hypothesis that an individual predictor's regression coefficient is zero, given the other predictor variables are in the model. The **Chi-Square** test statistic is the squared ratio of the **Estimate** to the **Standard Error** of the respective predictor. The **Chi-Square** value follows a central Chi-Square distribution with degrees of freedom given by **DF**, which is used to test against the alternative hypothesis that the **Estimate** is not equal to zero. The probability that a particular **Chi-Square** test statistic is as extreme as, or more so, than what has been observed under the null hypothesis is defined by **Pr>ChiSq**.

u. **Effect** – Underneath are the predictor variables that are interpreted in terms of odds ratios.

v. **Point Estimate** – Underneath are the odds ratio corresponding to **Effect**. The odds ratio is obtained by exponentiating the **Estimate**, exp[**Estimate**]. The difference in the log of two odds is equal to the log of the ratio of these two odds. The log of the ratio of two odds is the log odds ratio. Hence, the interpretation of **Estimate**–the coefficient was interpreted as the difference in log-odds–could also be done in terms of log-odds ratio. When the **Estimate** is exponentiated, the log-odds ratio becomes the odds ratio. We can interpret the odds ratio as follows: for a one unit change in the predictor variable, the odds ratio for a positive outcome is expected to change by the respective coefficient, given the other variables in the model are held constant.

w. **95% Wald Confidence Limits** – This is the Wald Confidence Interval (CI) of an individual odds ratio, given the other predictors are in the model. For a given predictor variable with a level of 95% confidence, we'd say that we are 95% confident that upon repeated trials, 95% of the CI's would include the "true" population odds ratio. The CI is equivalent to the **Chi-Square** test statistic: if the CI includes one, we'd fail to reject the null hypothesis that a particular regression coefficient equals zero and the odds ratio equals one, given the other predictors are in the model. An advantage of a CI is that it is illustrative; it provides information on where the "true" parameter may lie and the precision of the point estimate for the odds ratio.

| | | | |
|---|---|---|---|
| Percent Discordant[y] | 14.2 | Gamma | 0.715 |
| Percent Tied[z] | 0.2 | Tau-a[dd] | 0.279 |
| Pairs[aa] | 7791 | c[ee] | 0.857 |

x. **Percent Concordant** – A pair of observations with different observed responses is said to be concordant if the observation with the lower ordered response value (honcomp = 0) has a lower predicted mean score than the observation with the higher ordered response value (honcomp = 1). See **Pairs**, superscript aa, for what defines a pair.

y. **Percent Discordant** – If the observation with the lower ordered response value has a higher predicted mean score than the observation with the higher ordered response value, then the pair is discordant.

z. **Percent Tied** – If a pair of observations with different responses is neither concordant nor discordant, it is a tie.

aa. **Pairs** – This is the total number of distinct pairs in which one case has an observed outcome different from the other member of the pair. In the Response Profile table in the Model Information section above, we see that there are 53 observations with honcomp=1 and 147 observations with honcomp=0. Thus the total number of pairs with different outcomes is 53*147=7791.

bb. **Somers' D** – Somer's D is used to determine the strength and direction of relation between pairs of variables. Its values range from -1.0 (all pairs disagree) to 1.0 (all pairs agree). It is defined as $(n_c-n_d)/t$ where $n_c$ is the number of pairs that are concordant, $n_d$ the number of pairs that are discordant, and t is the number of total number of pairs with different responses. In our example, it equals the difference between the percent concordant and the percent discordant divided by 100: (85.6-14.2)/100 = 0.714.

cc. **Gamma** – The Goodman-Kruskal Gamma method does not penalize for ties on either variable. Its values range from -1.0 (no association) to 1.0 (perfect association). Because it does not penalize for ties, its value will generally be greater than the values for Somer's D.

ee. **c** – **c** is equivalent to the well known measure ROC. **c** ranges from 0.5 to 1, where 0.5 corresponds to the model randomly predicting the response, and a 1 corresponds to the model perfectly discriminating the response.

---

Click here to report an error on this page or leave a comment

How to cite this page (https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-cite-web-pages-and-programs-from-the-ucla-statistical-consulting-group/)

© 2021 UC REGENTS (http://www.ucla.edu/terms-of-use/) HOME (/) CONTACT (/contact)