Credit Scoring via Logistic Regression[☆]

Ali Al-Aradi

Department of Statistical Sciences, University of Toronto, Toronto, Canada

Abstract

The goal of credit scoring models is to predict the creditworthiness of a customer and determine whether they will be able to meet a given financial obligation or default on it. Such models allow a financial institution to minimize the risk of loss by setting decision rules regarding which customers receive loan and credit card approvals. Logistic regression can be used to predict default events and model the influence of different variables on a consumer's creditworthiness. In this paper we use a logistic regression model to predict the creditworthiness of bank customers using predictors related to their personal status and financial history. Model adequacy and robustness checks are performed to ensure that the model is being properly fitted and interpreted.

1. Introduction

Logistic regression is one of the most important models for categorical response data. It is an example of a generalized linear model whose main use is to estimate the probability that a binary response occurs based on a number of predictor variables. Logistic regression is used in a wide variety of applications including biomedical studies, social science research, marketing as well as financial applications. One example of the latter is the use of binary logistic regression models for credit-scoring, that is: modeling the probability that a customer is creditworthy (i.e. able to meet a financial obligation in a timely manner) using a number of predictors. These predictors can include the size of the loan as well as other personal information such as the customer's annual income, occupation, other outstanding debts, their past default behavior and their credit history.

In this paper, we use a data set that includes 20 covariates for 1000 observations (loan applicants) to build a model for creditworthiness. The model allows us to identify the variables

[☆]This work is based on a problem posed in Chapter 6 of Agresti (2003) modified by Prof. Nancy Reid for the Winter 2014 session of the Methods of Applied Statistics II course at the University of Toronto.

most strongly associated with a customer's credit score. The conclusions are then presented in the form of a report to the bank manager which would help them assess loan applications based on the applicant's profile to decide whether to proceed with loan approval or not.

2. Data and Preprocessing

The data set used is the German Credit dataset obtained from the UCI machine-learning data archive and includes 20 covariates (7 numerical, 13 categorical) and 1000 observations. Each observation represents an individual customer with the response indicating their actual classification (1 = "Good" or 2 = "Bad") and the covariates indicating various attributes related to the customer's personal or financial information. For the purpose of this paper we will focus on the predictors listed in Table 1 below.

Variable	**Possible Values**
Checking account status	Less than 0 DM (Deutsche Mark) between 0 DM and 200 DM More than 200 DM/salary assignments for at least 1 year
Credit duration	Numerical value in months
Credit history	no credits taken/all credits paid back duly all credits at this bank paid back duly existing credits paid back duly till now delay in paying off in the past critical account/other credits existing (not at this bank)
Intended use	Car (new) Car (used) Furniture/equipment Radio/television Domestic appliances Repairs Education Vacation Retraining Business Other
Marital status and gender	Divorced/separated male Divorced/separated/married male Single male Married/widowed male Single female

Table 1: List of Predictor Variables

First, some minor data preprocessing is done to make the analysis simpler. The required variables are extracted and the covariate values are renamed for ease of interpretation where possible (e.g. instances of "A91" are replaced by "divorcedMale" to indicate the gender and marital status of the consumer). Also, to ensure that responses are in the form of binary data, the "bad" credit quality responses are changed from 2 to 0 so that success (good credit) is indicated by a value of 1, and the odds we consider are those of being creditworthy, i.e. not defaulting on the loan.

3. Binary Logistic Model

We fit a binary logistic model to the data, using the logit link function. That is, the classification of the i^{th} customer as good or bad is modeled using a Bernoulli random variable:

$$Y_i = \begin{cases} 1 & \text{if the customer is creditworthy} \\ 0 & \text{otherwise} \end{cases}$$

with conditional probabilities $\mathbb{P}(Y_i = 1|\mathbf{x}_i) = \pi_i$ and $\mathbb{P}(Y_i = 0|\mathbf{x}_i) = 1 - \pi_i$ where $\mathbf{x}_i$ is a vector of covariates associated with this customer. The conditional expectation is then given by:

$$\mathbb{E}[Y_i|\mathbf{x}_i] = \pi_i$$

and this is associated to a linear predictor via the logit function, i.e.

$$\text{logit } \pi_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i'\boldsymbol{\beta} = \eta_i$$

where $\boldsymbol{\beta}$ is a vector of parameters that needs to be estimated. The estimation is performed by iterative weighted least squares (IWLS) which is described in more detail in Davison (2003). Note that the conditional joint probability of $Y_1, ..., Y_n$ (assuming conditional independence) is:

$$\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \exp\left[\sum_{i=1}^{n} y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} \log(1-\pi_i)\right]$$

which implies that this probability distribution is a member of the exponential family.

The choice of link function is motivated mostly by ease of interpretation of model parameters. Additionally, alternative models were fitted using the probit and complementary log-log link functions and the resulting conclusions are similar. It should also be noted that, while aggregation to binomial data is possible, the binomial denominators of the aggregated data remain too small for confidence in chi-square asymptotics[1]. This makes it difficult to use usual model adequacy checking procedures, such as looking at residuals and deviance.

[1] See Section 10.3.2. of Davison (2003).

A possible solution is to further group the covariate data prior to aggregation to achieve a smaller range of covariate patterns, and in turn decrease the instances of small binomial denominators. Since this process involves some loss of information as the covariate patterns become less granular, the use of aggregated data is deferred to a later section and is used mainly as a robustness check for the results obtained by the binary model.

4. Model Adequacy

The first test we perform to check the suitability of this model/link is a test of non-additivity, where we compute the fitted linear predictor $\hat{\eta}$, then estimate a second model with $\hat{\eta}^2$ added to the original list of explanatory variables and, finally, test the significance of the deviance reduction. The deviance of the extended model is lower by 1.62. So, the test statistic for the non-additivity test is 1.62, and this is compared against a χ^2_1 distribution for a p-value of 0.203. This suggests that there is weak evidence against the model/link.

Since we are working with binary data, the usual model checking procedures such as using the Pearson chi-square statistic or the deviance likelihood ratio test are not informative. So, instead of using these tests or looking at the usual residual plots, we will employ the Hosmer-Lemeshow Test.[2] The idea is to group observations into g categories (usually taken to be 10) based on fitted probabilities, computing the Pearson chi-squared statistic for the resulting $g \times 2$ contigency table, and using this as a measure of fit by comparing the test statistic to a χ^2_{g-2} distribution. For this model, the test statistic is 4.861 which gives a p-value of 0.772, suggesting that there is little evidence against the model fit.

One matter of concern with the Hosmer-Lemeshow test is that it has been shown to be sensitive to the choice of grouping parameter g.[3] To address this we run the test using values of g from 3 to 100. The smallest p-value we obtain is 0.123, which supports our original conclusion regarding model adequacy. Other concerns regarding this test (that are not addressed in this report) include its ineffectiveness at detecting small nonlinear terms in the predictor, interaction effects (both mild and extreme), and incorrect but symmetric link functions.[4]

Next, we consider the effect of outliers on the model. For this we plot the Cook statistics for each observation to identify outliers in the dataset. We find from Figure 1 that consumers #106, #204 and #736 appear to be outliers. So, we fit a second model with these cases excluded to study their impact on the estimation and conclusions. We find that removal of these cases does not lead to noticeable changes in estimated parameters, parameter sig-

[2] See section 7.5 of Dobson and Barnett (2008).
[3] See Hosmer et al. (1997).
[4] Ibid.

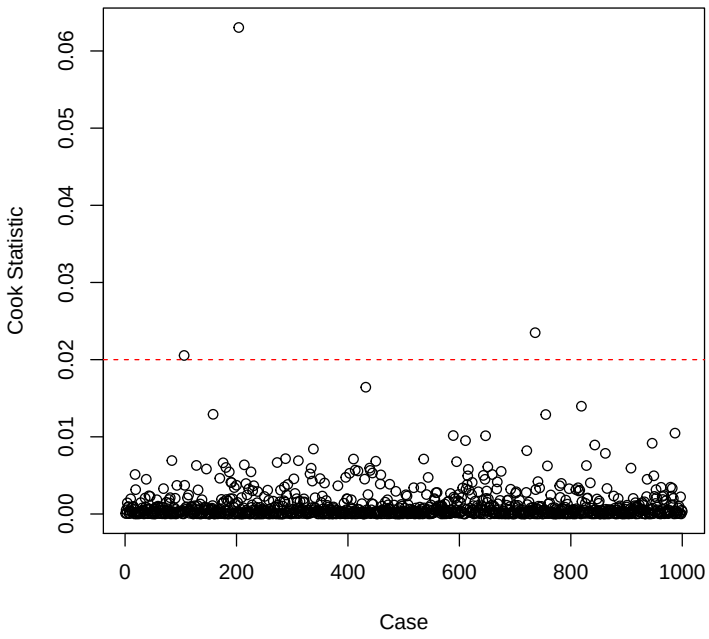

Figure 1: Cook statistics for each observation

nificance, goodness-of-fit test results, multiple comparisons or any other procedure and its associated conclusions. One exception is the estimated beta associated with the retraining purpose indicator variable. The estimate of this parameter changes from 1.825 with the outliers included to 15.213 when they are removed. However, in this case one of the removed observations (#204, the most noticeable outlier) is the only consumer that indicated retraining as their purpose and had bad credit, i.e. the remaining consumers that indicated their purpose to be retraining had good credit. This is what causes the large change in the associated coefficient, but this is not a concern for the model's overall fit or adequacy. At this point we are not concerned with overdispersion, as we are working with binary data where overdispersion is undetectable.

5. Results and Interpretation

The linear predictor of the model we are fitting is given by:

$$\begin{aligned}\eta = \alpha &+ \beta_{\text{0-200}} C_{\text{0-200}} + \beta_{\text{moreThan200}} C_{\text{moreThan200}} + \beta_{\text{none}} C_{\text{none}} \\ &+ \beta_{\text{atBankPaid}} H_{\text{atBankPaid}} + \beta_{\text{existingPaid}} H_{\text{existingPaid}} + \beta_{\text{pastDelay}} H_{\text{pastDelay}} + \beta_{\text{critical}} H_{\text{critical}} \\ &+ \beta_{\text{divOrMarriedF}} G_{\text{divOrMarriedF}} + \beta_{\text{singleM}} G_{\text{singleM}} + \beta_{\text{marriedM}} G_{\text{marriedM}} \\ &+ \beta_{\text{usedCar}} P_{\text{usedCar}} + \beta_{\text{other}} P_{\text{other}} + \beta_{\text{radioTV}} P_{\text{radioTV}} + \beta_{\text{appliance}} P_{\text{appliance}} + \beta_{\text{repairs}} P_{\text{repairs}} \\ &+ \beta_{\text{education}} P_{\text{education}} + \beta_{\text{retrain}} P_{\text{retrain}} + \beta_{\text{business}} P_{\text{business}}\end{aligned}$$

where the indicators C_x, H_x, G_x, P_x correspond to checking account status, credit history, gender/status or purpose x, respectively. Since we are using a logistic regression model with logit link function, the coefficients for each variable can be interpreted in terms of multiplicative factors for the odds of a consumer's creditworthiness, relative to the reference category (reference categories are given in Table 2 below). In particular, the creditworthiness of a consumer in category x change by a factor of e^{β_x} relative to the reference category after controlling for all other variables. A positive (resp. negative) coefficient indicates greater (resp. smaller) odds of having good credit compared to the reference category. Differences between coefficients of the same variable type can be interpreted in the same manner; as differences in odds of creditworthiness. In particular, the odds of creditworthiness for two consumers in two categories x and y vary by a factor of $e^{\beta_x - \beta_y}$, after controlling for all other variables.

Variable	**Reference Category**
Checking account status	Less than 0 DM
Credit history	No credits taken/all credits paid
Gender/marital status	Divorced/separated male
Purpose	New car

Table 2: List of Reference Categories

To find which variables explain creditworthiness and in what way, we begin by testing the significance of each group of variables. We do this by running a likelihood ratio test, comparing the full model deviance to a reduced model deviance in which a single group of variables is removed, and comparing this test statistic to a χ^2_k where k is equal to the number of removed parameters. The findings are summarized in the Table 3:

We find that all the predictor variables are significant at the 95% significance level, so they are all important in determining the consumer's creditworthiness. To find the ways in which different categories relate to one another in terms of odds of creditworthiness, we compute simultaneous 95% confidence intervals for all contrasts of each group of variables. Table 4 shows the confidence intervals for significant differences.

Removed Variable	df	Deviance	AIC	LRT Stat	p-value
		980.197	1022.197		
checking	3	1068.997	1104.997	88.801	3.964×10^{-19}
duration	1	1020.499	1060.499	40.303	2.175×10^{-10}
history	4	1010.195	1044.195	29.998	4.899×10^{-6}
purpose	9	1013.427	1037.427	33.23	1.218×10^{-4}
genderStatus	3	989.076	1025.076	8.88	0.031

Table 3: Tests of Significance for Predictors

From this table we can conclude the following:

- The odds of creditworthiness for a single male are greater than a divorced or married female by a factor of $e^{\beta_{\text{singleM}} - \beta_{\text{divOrMarriedF}}} = 1.636$. There are no other significant differences between the various gender/status categories, and there is insufficient information to make a general statement on gender/status effect on creditworthiness.

- Odds of a consumer's creditworhtiness increase with an increase in the amount in their checking account. In particular, relative to consumers with less than 0 DM in their checking account, odds of creditworthiness for customers with checking accounts with 0-200 DM, more than 200 DM and those with no checking account are greater by factors of $e^{\beta_{\text{0-200}}} = 1.662$, $e^{\beta_{\text{moreThan200}}} = 3.015$, and $e^{\beta_{\text{none}}} = 6.25$ respectively. Moreover, the odds for consumers with no checking account are greater than those in the 0-200DM group by a factor of $e^{\beta_{\text{none}} - \beta_{\text{0-200}}} = 3.76$. This is partially expected, as customers with more money in their checking accounts would be less likely to default, but it is surprising to see that customers with no checking account being more creditworthy.

- With respect to purpose, the largest difference in odds is between consumers whose purpose was education and those whose purpose was the purchase of a used car; the latter's odds were greater by a factor of $e^{\beta_{\text{usedCar}} - \beta_{\text{education}}} = 5.497$. Additionally, customers whose purpose was the purchase of a used car or a radio/TV had greater odds of creditworthiness than those buying a new car by factors of $e^{\beta_{\text{usedCar}}} = 4.034$ and $e^{\beta_{\text{radioTV}}} = 2.232$, respectively. Once again there is not enough information to make a general statement on the relation between purpose and creditworthiness.

- A one-month increase in duration drops the expected odds of creditworthiness by a factor of $e^{\beta_{\text{duration}}} = 0.958$. This makes sense as a longer loan has a greater chance of defaulting than a shorter one, after controlling for other variables.

- Consumers with "critical" credit history show large increases in expected odds of creditworthiness. The odds for these customers relative to customers with fully paid credits, fully paid credits at this bank and those with existing credits paid back are greater by

Linear Hypothesis = 0	Lower Bound	Estimate	Upper Bound
Credit History			
β_{critical}	0.549	1.662	2.774
$\beta_{\text{critical}} - \beta_{\text{atBankPaid}}$	0.549	1.546	2.542
$\beta_{\text{critical}} - \beta_{\text{existingPaid}}$	0.108	0.647	1.187
Gender and Marital Status			
$\beta_{\text{singleM}} - \beta_{\text{divOrMarriedF}}$	0.035	0.492	0.949
Checking Account Status			
$\beta_{\text{0-200}}$	0.014	0.508	1.003
$\beta_{\text{moreThan200}}$	0.229	1.104	1.979
β_{none}	1.292	1.833	2.373
$\beta_{\text{none}} - \beta_{\text{0-200}}$	0.777	1.324	1.872
Purpose			
β_{usedCar}	0.367	1.395	2.422
β_{radioTV}	0.113	0.803	1.493
$\beta_{\text{education}} - \beta_{\text{usedCar}}$	-3.083	-1.704	-0.326
Duration			
β_{duration}	-0.056	-0.042	-0.029

Table 4: 95% confidence intervals for significant coefficient differences

factors of $e^{\beta_{\text{critical}}} = 5.267$, $e^{\beta_{\text{critical}}-\beta_{\text{atBankPaid}}} = 4.691$, $e^{\beta_{\text{critical}}-\beta_{\text{existingPaid}}} = 1.91$, respectively. This result is counterintuitive as it suggests that consumers with worse credit history are less likely to default. This might be due to a form of "bias" associated with the way loans are issued - the bank may be more stringent when it comes to loaning a consumer with bad credit history, whereas consumers with good credit history do not face the same kind of scrutiny and may end up being issued a loan they eventually cannot repay. An alternative explanation is that there may be a data issue in which the categories were incorrectly labeled. It would be best to be cautious with the interpretation of this result.

6. Robustness Checks

In order to check the robustness of the conclusions made above, an alternative model is fit using aggregated data, to see if the conclusions of these models agree with those of the original model. For the aggregated data, we aggregate the responses to achieve binomal variables. To do this in such a way that the binomial denominators are not too small, we must group some of the predictor variables to reduce the number of covariate patterns. To this end, we have done the following:

- History is defined as "good" (all paid/at bank paid/existing paid) and "bad" (past delay/critical).
- Durations are grouped into 2-year bins (i.e. 0-2, 2-4, or 4-6 years).
- Marital status ("divorced/married" or "single") are considered seperately from gender.
- Purpose predictors are grouped into "car" (for new and used cars), "home" (for furniture, appliances, radio/TV and repairs) and "other" (for the remaining purposes).

We find that with this grouping strategy, 54% of covariate patterns have binomial denominators of 5 or more, so we can have more confidence in χ^2 asymptotics. The model appears to be adequate, since the p-value for the non-additivity test is 0.606, and the deviance is 138.251 on 122 degrees of freedom, so the p-value for the χ^2 goodness-of-fit test is 0.149. Furthermore, there does not appear to be any indication of overdispersion based on the reisudal deviance. Finally, the diagnostic plots given below indicate reasonable proximity to normality for the residuals and constant variance. Note that one covariate pattern was omitted from the fitted model as it was deemed to be an outlier.

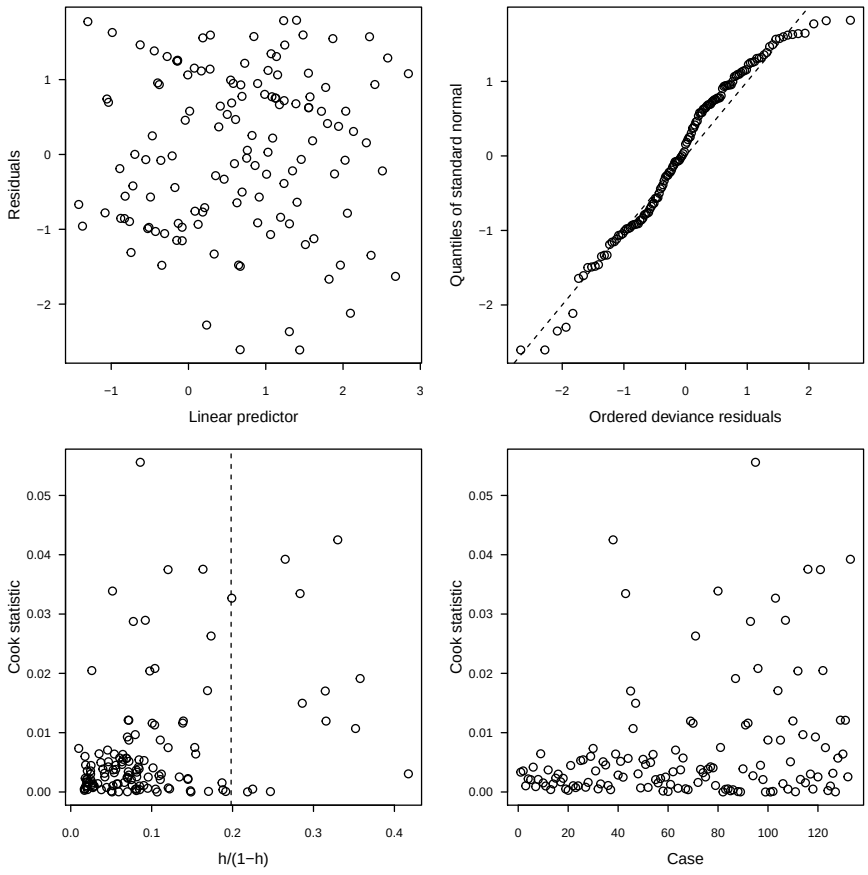

Figure 2: Diagnostic plots for binomial logistic regression model

If we test the significance of each of the variables individually for this model with likelihood ratio tests (results in the table below), we find that the results agree with those of the original model. The exceptions are that when gender and status are considered separately, they are not significant variables, and purpose is also insignificant at 95% significance level. So, there is no difference on average in the odds of creditworthiness for males and females, for single and married/divorced consumers, and for customers whose purpose is "car", "home" or "other". This is not surprising as the first model concluded that there was a difference between only one pair of gender/status groups: single males and divorced/married females, and differences between just three pairs of purposes (used car vs. new car, radio/TV vs. new car, education vs. used car).

Removed Variable	**df**	**Deviance**	**AIC**	**LRT Stat**	**p-value**
none		138.251	338.085		
checking	3	245.256	439.09	107.005	4.839×10^{-23}
duration	2	166.252	362.085	28	8.314×10^{-7}
history	1	148.904	346.738	10.653	0.001
purpose	2	143.329	339.162	5.077	0.079
gender	1	139.623	337.457	1.372	0.241
status	1	138.335	336.168	0.083	0.773

Table 5: Tests of Significance for Predictors - Aggregated Data

When considering the significant differences (pairwise comparisons of means using the Tukey approach) between the other categories we find:

- The conclusions for consumers with varying checking account status match those of the first model both directionally, and in terms of magnitude.
- Differences between duration groups show a decrease in creditworthiness odds as duration increases, with the largest difference being between the 4-6 year and 0-2 year groups (decrease by a factor of 0.276)
- Being in the "bad" credit history group increases the odds of creditworthiness by 1.722

This verifies the robustness of the conclusions of the binary logistic regression model.

7. Conclusions

The following report summarizes the findings of a statistical analysis conducted to determine the credit-worthiness of consumers based on the consumer's checking account, the duration of the credit, the consumer's credit history, the intended use for the credit, and the consumer's gender and marital status.

The main findings of the analysis are as follows:

- Odds of a consumer's creditworhtiness increase with an increase in the size of their checking account. In particular, relative to consumers with less than 0 DM in their checking account, odds of creditworthiness for customers with checking accounts with 0-200 DM, more than 200 DM and those with no checking account are greater by factors of 1.662, 3.015, and 6.25, respectively. Moreover, the odds for consumers with no checking account are greater than those in the 0-200DM group by a factor of 3.76. This is partially expected, as customers with larger checking accounts would be less likely to default, but it is surprising to see that customers with no checking account being more creditworthy.

- Broadly speaking, when considering gender and marital status separately there are no differences among males and females or among divorced/married and single consumers. There is, however, a difference between divorced/married females and single males: the odds of creditworthiness of the latter group are 1.636 times greater.

- Broadly speaking, when the purpose of the credit is grouped into "car," "home," and "other" we find no difference in odds of creditworthiness. If we take a more granular view of purpose, we find that the largest difference in odds is between consumers whose purpose was education and those whose purpose was the purchase of a used car; the latter's odds are 5.497 times greater. Additionally, customers whose purpose was the purchase of a used car or a radio/TV had greater odds of creditworthiness than those buying a new car by factors of 4.034 and 2.232, respectively.

- Increased duration decreases the odds of creditworthiness. In particular, a one-month increase in duration drops the expected odds of creditworthiness by a factor of 0.958. This makes sense as a longer loan has a greater chance of defaulting than a shorter one, after controlling for other variables.

- A somewhat counterintuitive result is that consumers with "critical" credit history show large increases in expected odds of creditworthiness. The odds for these customers relative to customers with fully paid credits, fully paid credits at this bank and those with existing credits paid back are greater by factors of 5.267, 4.691, 1.91, respectively. The result suggests that consumers with worse credit history are less likely to default.

This might be due to a form of "bias" associated with the way loans are issued - the bank may be more stringent when it comes to loaning a consumer with bad credit history, whereas consumers with good credit history do not face the same kind of scrutiny and may end up being issued a loan they eventually cannot repay. An alternative explanation is that there may be a data issue in which the categories were incorrectly labeled. It would be best to be cautious with the interpretation of this result.

8. References

Agresti, A. (2003). *Categorical data analysis*. John Wiley & Sons.

Davison, A. C. (2003). *Statistical models*. Cambridge University Press.

Dobson, A. J. and A. Barnett (2008). *An introduction to generalized linear models*. CRC press.

Hosmer, D. W., T. Hosmer, S. Le Cessie, S. Lemeshow, et al. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine 16*(9), 965–980.