- 27th Feb, 2018

- Data Science Bootcamp

Logistic Regression and Moving Beyond Linearity

- Anindya Moitra
- Shovon Sengupta

Shuvayon Dey

Logistic Regression

Modeling discrete response variables

- In a very large number of problems in cognitive science and related fields, the response variable is categorical, often binary (yes/no; acceptable/not acceptable; phenomenon takes place/does not take place)
- Potentially explanatory factors (independent variables) are categorical, numerical or both.

Examples: binomial/ multinomial responses

- How many people answer YES to question X in the survey?
- Subjects answer YES, MAYBE, NO
- Subject decides to buy one of 4 different products.
- Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 0 (did not vote) or 1(did vote)

Why use logistic regression?

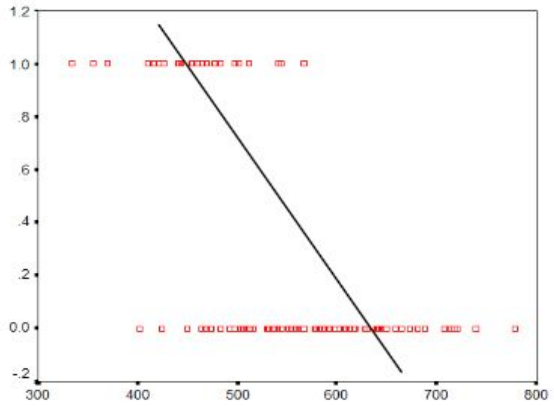
- In the OLS regression:
  $Y = \gamma + \phi X + e$; where Y = (0, 1)
- The error terms are heteroskedastic
- e is not normally distributed because Y takes on only two values
- The predicted probabilities can be greater than 1 or less than 0

How Logistic solve these problems?

- $\ln[p/(1-p)] = a + bx + e$; p is the probability of the event Y occurs, P(Y=1)
- $p/(1-p)$ is the odds ratio; $\ln[p/(1-p)]$ is the log odds ratio or "logit"
- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.

Why Bother With This Logit Function?

If we used Y as the outcome variable and tried to fit a line, it wouldn't be a very good representation of the relationship. The following graph shows an attempt to fit a line between one X variable and a binary outcome Y.



Probability vs Odds Ratio

- The probability that an event will occur is the **fraction of times you expect to see that event in many trials**. Probabilities always range between 0 and 1. If the probability of an event is 0.80 (80%), then the probability that the event will not occur is 1-0.80 = 0.20, or 20%.
- The odds of an event represent the ratio of the (probability that the event will occur) / (probability that the event will not occur). This could be expressed as follows: **Odds of event = Y / (1-Y)**

  If the probability of the event occurring = 0.80, then the odds are 0.80 / (1-0.80) = 0.80/0.20 = 4 (i.e., 4 to 1).

Examples:

- If a race horse runs 100 races and wins 25 times and loses the other 75 times, the probability of winning is 25/100 = 0.25 or 25%, but the odds of the horse winning are 25/75 = 0.333 or 1 win to 3 loses.
- If the horse runs 100 races and wins 5 and loses the other 95 times, the probability of winning is 0.05 or 5%, and the odds of the horse winning are 5/95 = 0.0526.

Something that never happens will have odds of 0:1 in favor, and something that always happens will have odds of 1:0 in favor (0:1 against). Odds of 1:1 are "fifty-fifty", equally like to occur or not; this corresponds to 50% probability.

Logistic Regression - Illustration

- The simple and multiple linear regression methods are used to model the relationship between a quantitative response variable and one or more quantitative explanatory variables.
- However Logistic Regression performs a regression analysis when the Response Variable is dichotomous in nature.

Two Outcomes (conventionally)

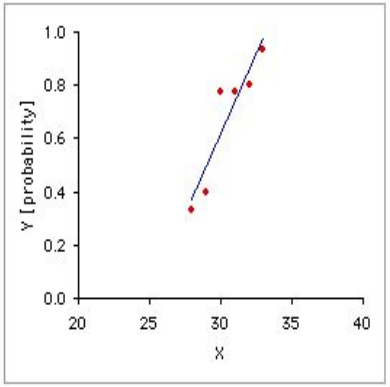
Success Probability = **'p'**

Failure Probability = **'1-p'**

'odds' of **Success**

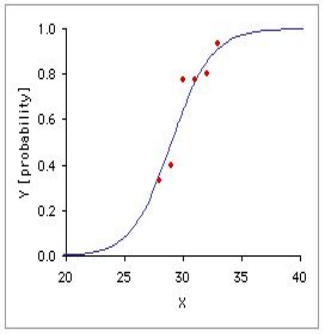
Independent variable (quantitative)

The equation of the logistic regression is : $\log [p/(1-p)] = \beta 0 + \beta 1 * X$



Linear Regression

- Entirely quantitative model that eventually aims at calculating the β0 and β1, estimating the equation.
- Then the X values are fed in from the data which yields the Log 'odds'.
- The 'odds' are calculated from the log odds using the logarithmic table.



Since logistic function is monotonic, increase in log odds => increase in odds

The logistic transformation is to convert the function to linearity, since the probability function is as shown adjacently.

Maximum Likelihood Estimation (MLE)

- MLE is a statistical method for estimating the coefficients of a model.
- The likelihood function (L) measures the probability of observing the particular set of dependent variable values (p_1, p_2, ..., p_n) that occur in the sample:
  L = Prob (p_1* p_2* * * p_n)
- The higher the L, the higher the probability of observing the p_s in the sample.

- Let the log of odds is represented by,

$$\log \frac{P(\mathbf{x})}{1-P(\mathbf{x})} = \sum_{j=0}^{K} b_j x_j$$

- Adding the intercept term (x_0=1) we have,

$$\frac{P(\mathbf{x})}{1-P(\mathbf{x})} = \exp(\sum_{j=0}^{K} b_j x_j)$$

$$= \prod_{j=0}^{K} \exp(b_j x_j)$$

- Thus the likelihood can be represented as,

$$\mathrm{L}(X|P) = \prod_{i=1,y_i=1}^{N} P(\mathbf{x}_i) \prod_{i=1,y_i=0}^{N} (1-P(\mathbf{x}_i))$$

Maximum Likelihood Estimation (MLE)

- Taking log on both sides of (3), we have the log likelihood function as,

$$\mathcal{L}(X|P) = \sum_{i=1, y_i=1}^{N} \log P(\mathbf{x}_i) + \sum_{i=0, y_i=0}^{N} \log(1 - P(\mathbf{x}_i))$$

- Our aim is to minimize -2* $\mathcal{L}(X|P)$ (known as the deviance of the model, analogous to residual sum of squares (RSS) to linear regression).

- Taking gradient w.r.t the parameters **b,** we have where, $P_i' = P_i'(x_i)$

$$\nabla_{\mathbf{b}}\mathcal{L} = \sum_{\substack{i=0 \\ y_i=1}}^{N} \frac{P_i'}{P_i}\mathbf{x}_i - \sum_{\substack{i=0 \\ y_i=0}}^{N} \frac{P_i'}{1-P_i}\mathbf{x}_i$$

- Solving for (5), we have

$$\begin{aligned}
\nabla_{\mathbf{b}}\mathcal{L} &= \sum_{\substack{i=i \\ y_i=1}}^{N} \frac{P_i(1-P_i)}{P_i}\mathbf{x}_i - \sum_{\substack{i=1 \\ y_i=0}}^{N} \frac{P_i(1-P_i)}{1-P_i}\mathbf{x}_i \\
&= \sum_{\substack{i=1 \\ y_i=1}}^{N} (1-P_i)\mathbf{x}_i - \sum_{\substack{i=1 \\ y_i=0}}^{N} P_i\mathbf{x}_i \\
&= \sum_{i=1}^{N} [y_i(1-P_i) - (1-y_i)P_i]\,\mathbf{x}_i
\end{aligned}$$

Note: **$P_i'(x_i) = P_i(x_i)\,(1-Pi(x_i))$**

- Thus using the Newton – Raphson method, we iteratively try to solve for the equation

$$\sum_{i=1}^{N} y_i\mathbf{x}_i - P_i\mathbf{x}_i = \mathbf{0}$$

Sample output

```
Call:
glm(formula = newy ~ marital + default1 + housing + loan + contact +
    poutcome + age + balance + campaign + pdays + previous, family = binomial(link = "logit"),
    data = trainingData)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.63309  -1.03353   0.05055   1.05357   2.93939

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.880e-01  2.050e-01   2.380   0.0173 *
maritalmarried   -3.310e-01  8.258e-02  -4.009 6.11e-05 ***
maritalsingle     1.959e-01  9.501e-02   2.062   0.0392 *
default1yes      -1.874e-01  2.191e-01  -0.855   0.3925
housingyes       -5.985e-01  5.437e-02 -11.008  < 2e-16 ***
loanyes          -4.607e-01  7.879e-02  -5.847 5.00e-09 ***
contacttelephone -2.133e-01  1.030e-01  -2.071   0.0384 *
contactunknown   -1.058e+00  7.101e-02 -14.898  < 2e-16 ***
poutcomeother     1.468e-01  1.327e-01   1.107   0.2684
poutcomesuccess   2.198e+00  1.607e-01  13.679  < 2e-16 ***
poutcomeunknown   6.837e-02  1.428e-01   0.479   0.6320
age               2.602e-03  2.563e-03   1.015   0.3100
balance           1.373e-04  3.114e-05   4.409 1.04e-05 ***
campaign         -9.591e-02  1.158e-02  -8.279  < 2e-16 ***
pdays             9.425e-05  4.589e-04   0.205   0.8373
previous          3.615e-02  1.645e-02   2.197   0.0280 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10264.1  on 7403  degrees of freedom
Residual deviance:  8786.1  on 7388  degrees of freedom
AIC: 8818.1

Number of Fisher Scoring iterations: 5
```

Null Deviance and Residual Deviance

- **Null Deviance = 2(LL(Saturated Model) - LL(Null Model)) on df = df_Sat - df_Null**
- **Residual Deviance = 2(LL(Saturated Model) - LL(Proposed Model)) df = df_Sat - df_Proposed**

- The **Saturated Model** is a model that assumes **each data point has its own parameters** (which means you have n parameters to estimate.)

- The **Null Model** assumes the exact "opposite", in that is **assumes one parameter for all of the data points**.

- The **Proposed Model** assumes you can explain your data points with **p parameters + an intercept term**, so you have p+1 parameters.

In case of logistic regression **R^2 doesn't lead to us any information**, instead heuristically we can use **pseudo-R^2** (*also known as McFadden's pseudo-R squared*) to test goodness of fit.

Output interpretation

- **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 Y =1) as a linear combination the predictor variables. This is written as

  $\log[p / (1-p)] = b_0 + b_1*x_1 + b_2*x_2 + b_3 *x_3 + \ldots.$, where p is the probability that Y is 1.

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.

- **Standard Error** - These are the standard errors of the individual regression coefficients. They are used in the 95% Confidence Limits.

- **Z value and Pr (> |z|) -** 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.

- **Odds ratio** - The odds ratio for your coefficient is the increase in odds when you increase X by one unit ***When the odds ratio is greater than 1, it describes a positive relationship and an odds ratio less than 1 implies a negative relationship.***

- **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(Σf_i), where f_i's are the frequency values of the i^{th} observation, and k and s were defined previously. Like AIC, SC penalizes for the number of predictors in the model and the smallest SC is most desirable and the value itself is not meaningful

Assessing Model Fit

- **Percent Concordant** - A pair of observations with different observed responses is said to be concordant if the observation with the lower ordered response value (Y= 0) has a lower predicted mean score than the observation with the higher ordered response value (Y= 1).

- **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.

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

- **Pairs** - This is the total number of distinct pairs with one case having a positive response (Y= 1) and the other having a negative response (Y= 0).

- **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 (nc-nd)/t where nc is the number of pairs that are concordant, nd 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.

- **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.

- **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.

- **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), 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.

Hosmer - Lemeshow Goodness of Fit

The Hosmer-Lemeshow test is used to determine the goodness of fit of the logistic regression model. Essentially it is a chi-square goodness of fit test for grouped data, usually where the data is divided into 10 equal subgroups..

Let the predicted probabilities be,

$$\hat{\pi} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + .. + \hat{\beta}_p X_p)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + .. + \hat{\beta}_p X_p)}$$

- The observations in the sample are then split into g groups according to their predicted probabilities. Suppose (as is commonly done) that g=10.
- Since this is a chi-square goodness of fit test, we need to calculate the HL statistic

$$\sum_{k=0}^{1} \sum_{l=1}^{g} \frac{(o_{kl} - e_{kl})^2}{e_{kl}}$$

where o_{kl} denotes the number of observed Y=0 observations in the l^{th} group, o_{1l} denotes the number of observed Y=1observations in the l^{th} group, e_{0l} and $e_{1l}$and similarly denote the expected number of zeros.

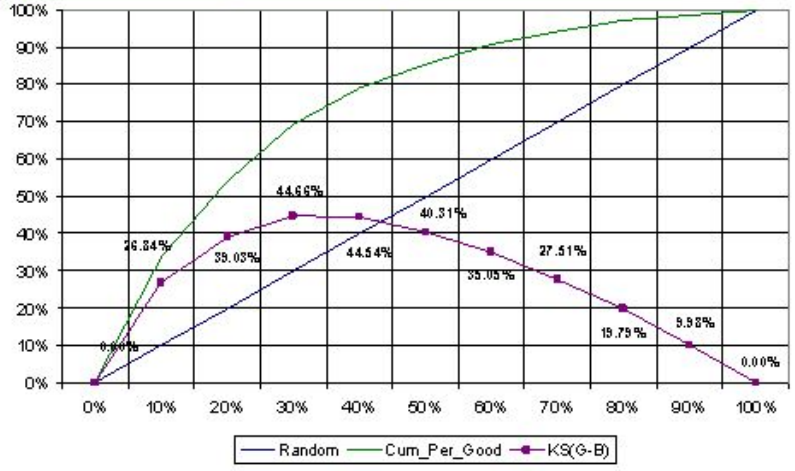
Small values with large p-values indicate a good fit to the data while large values with p-values below 0.05 indicate a poor fit. The null hypothesis holds that the model fits the data.

Hosmer
Lemeshow

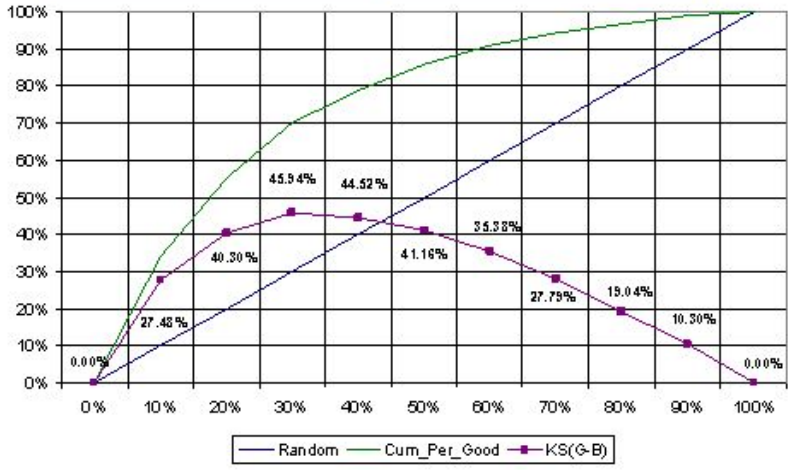
Kolmogorov-Smirnov (KS)

Kolmogorov-Smirnov (KS) measures the maximum vertical separation/deviation between the cumulative distributions of good and bads. The weakness of this method is that the separation measured only at one point (which may not be around the expected cut-off point) and not on the entire scoring range. If the intended model cut-off is at the upper or lower range of scores , this method may not give provide a very good indication of model comparison. In such cases, it might be better to compare the deviation at the intended cut-off since that is where the maximum separation is most required.

Model Development Sample



Model Validation Sample



KS_Calc

Confusion Matrix:

Predicted class	Actual Class		Marginal
	Yes	No	
Yes	True Positives (TP)	False Positive (FP)	P
No	False Negative (FN)	True Negatives (TN)	N
Marginal	P'	N'	Total

If these results are from a population-based study, prevalence can be calculated as follows:

- ***Prevalence of Disease= P'/ Total***

Sensitivity is the probability that a test will indicate 'Y=1' among those with the Y=1:

- ***Sensitivity/ Recall: TP / P'***

Specificity is the fraction of those "Y=0" who will have a negative test result:

- ***Specificity: TN / N'***

Sensitivity and specificity are characteristics of the test. The population does not affect the results.

- ***Accuracy = (TP + TN) / Total; Error rate: 1 – Accuracy***

Precision and F measure

A modeler and a partner have a different question: what is the chance that a person with a positive test truly default? If the subject is in the first row in the table above, what is the probability of being in cell A as compared to cell B? A modeler calculates across the row as follows:

- ***Positive Predictive Value/ Precision: TP/(TP + FP) × 100***
- ***Negative Predictive Value: TN/(FP + TN) × 100***

Perfect score for Recall is 1.0
Positive and negative predictive values are influenced by the prevalence of target in the population that is being tested. If we test in a high prevalence setting, it is more likely that persons who test positive will truly default than if the test is performed in a population with low prevalence..

- ***F* measure (*F*-score)**: harmonic mean of precision and recall, $F = \frac{2 * precision * recall}{precision + recall}$

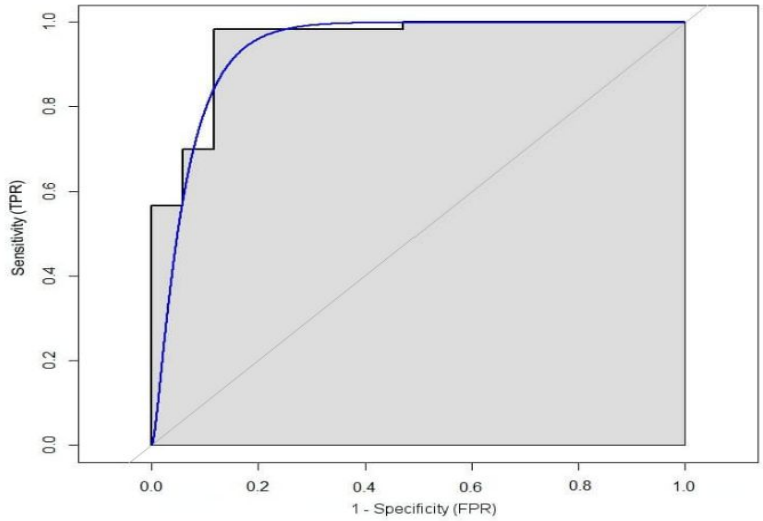
Predicted class	Actual Class		
	Yes	No	Total
Yes	**90**	**210**	300
No	**140**	**9560**	9700
Total	230	9770	10000

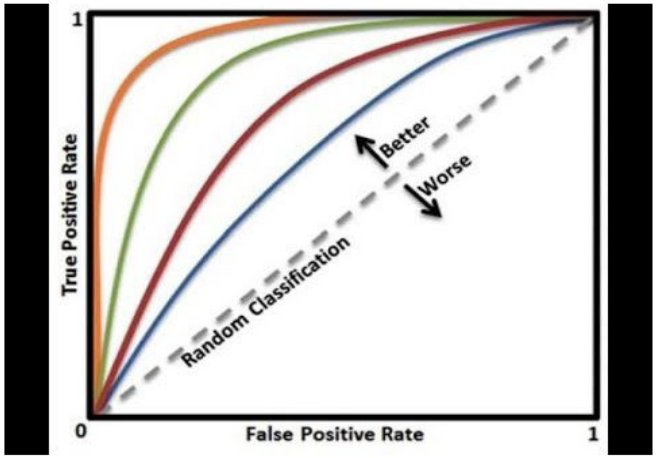
- *Sensitivity/ Recall = 90/230 = 0.39*
- *Specificity = 9560/9770 = 0.98*
- *PPV/Precision = 90/300 = 0.30*
- *NPV = 0.98*
- *Accuracy = (9560+90)/10000 = 0.96*

ROC

- **ROC –** *A Receiving Operating Characteristic (ROC) curve is a useful tool that allows us to examine the trade-off between these two metrics by plotting the TP against the FP for a variety of different classification thresholds. Each point on the line in Figure represents a different threshold for classification, ranging from all probabilities classified as failures in the bottom left-hand corner (i.e., 0% TP and FP) and all probabilities classified as successes in the top right-hand corner (i.e., 100% TP and FP).*

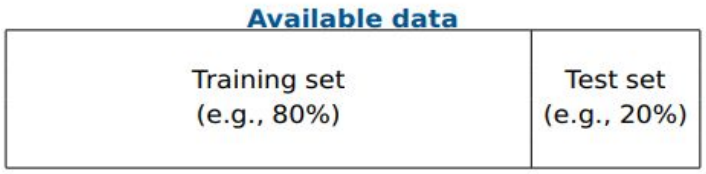
- This is the most powerful non-parametric two sample test and measure is equivalent to the area under the Receiver Operating Characteristics (ROC) curve, Gini co-efficient, and the Wilcoxon-Mann-Whitney test.
- It measures classifier performance across all score ranges and is a better measure of overall scorecard strength.
- The C-Statistics measures the area under the Sensitivity vs (1-Specificity) curve for the entire score range. The random line denotes C statistics = 0.5. Therefore for a good model, the c-statistics should be above 0.5. As per industry norms, a c-statistics above 0.7 is considered as a good model.





Validation Methods - Holdout & Cross-Validation

- **Holdout method**
  - Given data is randomly partitioned into two independent sets
    - Training set (e.g., 4/5) for model construction
    - Test set (e.g., 1/5) for accuracy estimation
  - Random sampling: a variation of holdout
    - Repeat holdout k times, accuracy = avg. of the accuracies obtained



- **Cross-validation** (*k*-fold, where k = 10 is most popular)
  - Randomly partition the data into *k mutually exclusive* subsets, each approximately equal size
  - At i^{th} iteration, use D_i as test set and others as training set

This is a 7-fold cross validation.

Moving Beyond Linearity

Generalized Additive Models

- GAMs are simply a class of statistical Models in which the usual Linear relationship between the Response and Predictors are replaced by several Non linear smooth functions to model and capture the Non linearities in the data.
- These are also a flexible and smooth technique which helps us to fit Linear Models which can be either linearly or non linearly dependent on several Predictors X_i to capture Non linear relationships between Response and Predictors.
- GAMs are just a Generalized version of Linear Models in which the Predictors X_i depend Linearly or Non linearly on some Smooth Non Linear functions like Splines , Polynomials or Step functions etc.

$$f(x) = y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \ldots . f_p(x_{ip}) + \epsilon_i$$

where the functions f_1, f_2, f_3 ,.... f_p are different Non Linear Functions on variables X_p .

Logistic Regression using GAM

We can also fit a Logistic Regression Model using GAMs for predicting the Probabilities of the Binary Response values. We will use the identity I() function to convert the Response to a Binary variable.

GAM code

Polynomial Regression

In multiple linear regression, we have outlined ways to check assumptions of linearity by looking for curvature in various plots:

- For instance, we look at the scatterplot of the residuals versus the fitted values.
- We also look at a scatterplot of the residuals versus each predictor.

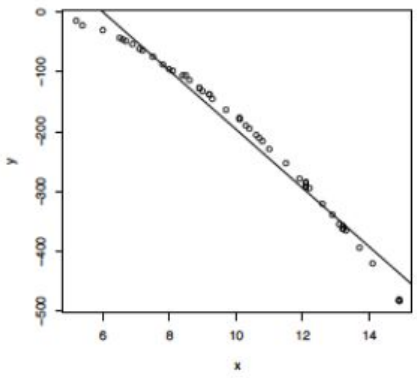
Sometimes, a plot of the residuals versus a predictor may suggest there is a nonlinear relationship. One way to try to account for such a relationship is through a polynomial regression model. Such a model for a single predictor, X, is:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_h X^h + \epsilon,$$

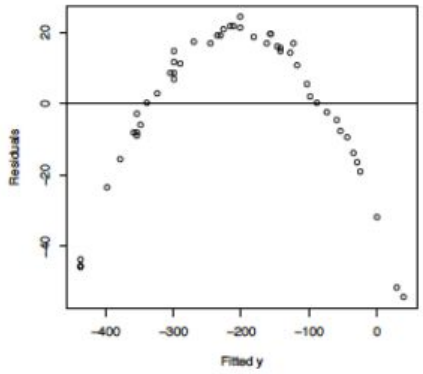
where h is called the degree of the polynomial.

Although this model allows for a nonlinear relationship between Y and X, polynomial regression is still considered linear regression since it is linear in the regression coefficients

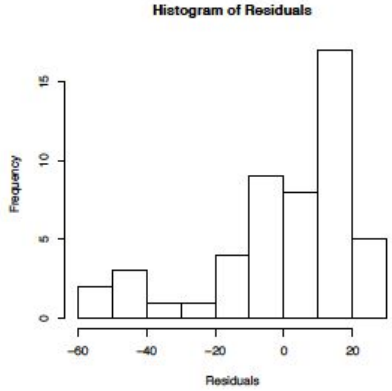
Polynomial Regression



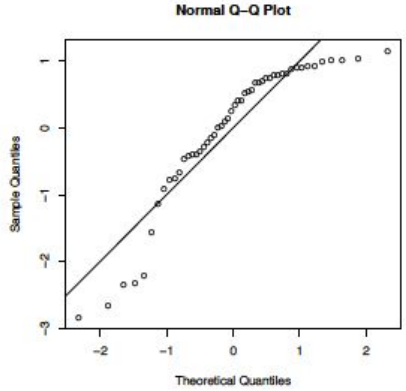
(a)



(b)



(c)



(d)

(a) Scatterplot of the quadratic data with the OLS line. (b) Residual plot for the OLS fit.
(c) Histogram of the residuals. (d) NPP for the Studentized residuals.

- Residuals versus predictor plots how there is obvious curvature and it does not show uniform randomness as we have seen before.
- The histogram appears heavily left-skewed and does not show the ideal bell-shape for normality.
- The NPP seems to deviate from a straight line and curves down at the extreme percentiles.

These plots alone suggest that there is something wrong with the model being used and indicate that a higher-order model may be needed.

Polynomial regression

Code_Polyreg

Poisson Regression

Suppose we are interested in knowing:

- What is the expected number of credit cards a person may have, given his/her income?, or
- What is the sample rate of possession of credit cards?

In Poisson regression Response/outcome variable Y is a count. But we can also have Y/t, the rate (or incidence) as the response variable, where t is an interval representing time, space or some other grouping.

GLM Model for Counts with its assumptions:

$$g(\mu)=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k$$

Random component: Response Y has a Poisson distribution that is yi~Poisson(μi) for i=1,...,N where the expected count of yi is E(Y)=μ.
Systematic component: Any set of X = (X_1, X_2, … X_k) are explanatory variables.

Link:

Identity link: $\mu=\beta_0+\beta_1 x_1$
Sometimes the identity link function is used in Poisson regression. This model is the same as that used in ordinary regression except that the random component is the Poisson distribution.
Issue: can yield $\mu < 0$

Poisson Regression

Natural log link: $\log(\mu)=\beta_0+\beta_1 x_1$

The Poisson regression model for counts is sometimes referred to as a "Poisson loglinear model". We will focus on this one and a rate model for incidences.

For simplicity, with a single explanatory variable, we write: **$\log(\mu)=\alpha+\beta x$**. This is equivalent to: **$\mu=\exp(\alpha+\beta x)=\exp(\alpha)\exp(\beta x)$**

Interpretation of Parameter Estimates:
$\exp(\alpha)$ = effect on the mean of Y, that is μ, when X = 0
$\exp(\beta)$ = with every unit increase in X, the predictor variable has multiplicative effect of $\exp(\beta)$ on the mean of Y, that is μ

- If $\beta = 0$, then $\exp(\beta) = 1$, and the expected count, $\mu = E(y) = \exp(\alpha)$, and Y and X are not related.
- If $\beta > 0$, then $\exp(\beta) > 1$, and the expected count $\mu = E(y)$ is $\exp(\beta)$ times larger than when X = 0
- If $\beta < 0$, then $\exp(\beta) < 1$, and the expected count $\mu = E(y)$ is $\exp(\beta)$ times smaller than when X = 0

Poisson Regression

Parameter Estimation: Same as used in case of Logistic Regression, i.e. MLE and Newton-Raphson.

Inference:

- Confidence Intervals and Hypothesis tests for parameters
  - Wald statistics and asymptotic standard error (ASE)
  - Likelihood ratio tests
  - Score tests
- Distribution of probability estimates

Model Fit:

- Pearson chi-square statistic, χ^2
- Deviance, G^2
- Likelihood ratio test, and statistic, ΔG^2

Residual analysis: Pearson, deviance, adjusted residuals, etc...

tudents Data Se

Poisson Code

25

Types of Logistic Regression models

Method	Description	Example	SAS Method
Binary logistic regression model	Used to model a binary (two-level) response	yes or no, Default or non - default	• PROCLOGISTIC • PROC GENMOD
Ordinal (ordered) logistic regression model	Used to model an ordered response — for example, low, medium, or high.	low, medium, or high.	• PROC LOGISTIC - UNEQUALSLOPES/ EQUALSLOPES option in the MODEL statement • PROC GENMOD - DIST=MULT option is specified in the MODEL statement.
Nominal (unordered) logistic regression model	Used to model a multilevel response with no ordering	Eye color with levels brown, green, and blue	• PROC LOGISTIC - REPEATED
Discrete choice models	Used to model a response that is the choice of individuals	Transportation modes (car, bus, train, plane)	• PROC BCHOISE (Bayesian choice modelling)
Logistic model for longitudinal (or repeated measures) data	These models are for a response that is observed more than once on each subject (or item), either at multiple times or under multiple conditions. The response can be binary, ordinal, or nominal.	Incidence of pulmonary disease among family members may be correlated because of hereditary factors	• PROC GENMOD – REPEATED statement

26

GLM in a nutshell

Three elements make up the generalized linear model:

- A probability distribution from the exponential family (as outlined above).
- A linear predictor η = Xβ . The linear predictor gives you information about the model's independent variables.
- The link function relates the linear predictor to the expected value. The following table shows some examples of link functions for various types of models.

Distribution	Support of distribution	Typical uses	Link name	Link function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$
Gamma				
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-2}$
Poisson	integer: $[0, +\infty)$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$
Bernoulli	integer: $[0, 1]$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$
Binomial	integer: $[0, N]$	count of # of "yes" occurrences out of N yes/no occurrences		
Categorical	integer: $[0, K)$	outcome of single K-way occurrence		
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1			
Multinomial	K-vector of integer: $[0, N]$	count of occurrences of different types (1 .. K) out of N total K-way occurrences		

Applying Logistic Regression using Regularization.
See: https://datascienceplus.com/logistic-regression-regularized-with-optimization/

27

Appendix

Further Reads:

- http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/viewer.htm#statug_bchoice_details01.htm
- http://www2.sas.com/proceedings/sugi22/STATS/PAPER278.PDF
- https://www.analyticsvidhya.com/blog/2016/02/multinomial-ordinal-logistic-regression/
- http://support.sas.com/kb/22/871.html
- https://people.maths.bris.ac.uk/~sw15190/mgcv/tampere/mgcv.pdf
- https://datascienceplus.com/logistic-regression-regularized-with-optimization/
- https://stats.idre.ucla.edu/sas/output/proc-logistic/


GD Code


GD2 Code


Hosmer Lemeshow


Bank Data Code