Multiple Factor Analysis: Logistic Regression

I. Logistic Regression

Recall: Ordinary Linear Regression

The **ordinary linear regression** is of the form

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta X_{k,i} + \varepsilon_i$$

Assumptions:

1. $\text{Corr}\left(X_{p,i}, X_{q,i}\right) \neq \pm 1$ for $p \neq q$
2. $\mathbb{E}(\varepsilon_i | \boldsymbol{X_i}) = 0$
3. $\text{Var}(\epsilon_i) = \sigma^2 < \infty$
4. $\text{Corr}\left(\varepsilon_i, \varepsilon_j\right) = 0$ for $i \neq j$
5. $\varepsilon_i \sim N(0, \sigma^2)$

Distribution of $\boldsymbol{Y}$: $\boldsymbol{Y_i \sim Be(\pi_i)}$

- Y_i has a probability density function of

$$f(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$Y_i = \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & \text{with probability } 1-\pi_i \end{cases}$$

- The mean and variance of Y_i is given by

$$0 \leq \mathbb{E}(Y_i) = \pi_i \leq 1$$

$$\text{Var}(Y_i) = \pi_i(1-\pi_i)$$

- Each of the $Y_i{}'s$ are independent of each other

Remark; In our context, π_i refers to the conditional probability that account i will become "bad" during its payment term to the bank.

Definition: Link Function

The **link function** of the dependent variable Y is the transformation on Y that provides the linear relationship between the linear predictor, $\boldsymbol{X^T\beta}$ and $\mathbb{E}(Y)$. That is, $g(\cdot)$ is a link function for the dependent variable Y if

$$g\left(\mathbb{E}(Y)\right) = \boldsymbol{X^T\beta} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Remark: $\boldsymbol{X} = [1, X_1, X_2, \dots, X_k]^T$ *and* $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_k]^T$

For the ordinary linear regression, $g\left(\mathbb{E}(Y)\right) = Y$

Problem: We can't use our ordinary linear regression as our model, since $Y_i \in \{0,1\}$ (either "good" or "bad" account), while the Y_i in the ordinary linear regression can range from $-\infty$ to ∞.

Definition: Logit Link Function

The **logit link function** gives the log-odds that $Y = 1$. This means that

$$\text{logit}\left(\mathbb{E}(Y)\right) = \ln\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{X^T\beta}$$

The probability π can then be computed as

$$\pi = \frac{1}{1 + \exp(-\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta})}$$

Remark: The logistic function is

$$f(x) = \frac{1}{1 + e^{-x}}$$

Model Parameters

We will be using the method of maximum likelihood to obtain the regression parameters $\beta_0, \beta_1, \dots, \beta_k$

Consider n independent Bernoulli observations (*referring to n accounts*), $Y_1, Y_2, \dots, Y_n$. The probability distribution function of Y_i is given by

$$f(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}$$

Hence, we aim to choose $\boldsymbol{\beta}$ that maximizes the likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}$$

Or, the log-likelihood function

$$\ln(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i \ln(\pi_i) + (1 - Y_i)\ln(1 - \pi_i))$$

Recall:

$$\pi = \frac{1}{1 + \exp(-\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta})}$$

We can use partial derivatives to obtain a system of equations

$$\begin{aligned}
l(\boldsymbol{\beta}) &= \ln(\boldsymbol{\beta}) \\
&= \sum_{i=1}^{n} (Y_i \ln(\pi_i) + (1 - Y_i)\ln(1 - \pi_i)) \\
&= \sum_{i=1}^{n} (Y_i \ln(\pi_i) + (-Y_i)\ln(1 - \pi_i) + \ln(1 - \pi_i)) \\
&= \sum_{i=1}^{n} \left(Y_i \ln\left(\frac{\pi_i}{1 - \pi_i}\right) + \ln(1 - \pi_i)\right) \\
&= \sum_{i=1}^{n} \left(Y_i(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta}) + \ln\left(1 - \frac{1}{1 + \exp(-\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta})}\right)\right) \\
&= \sum_{i=1}^{n} \left(Y_i(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta}) + \ln\left(\frac{1}{1 + \exp(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta})}\right)\right) \\
&= \sum_{i=1}^{n} \left(Y_i(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta}) - \ln(1 + \exp(\boldsymbol{X}^{T}\boldsymbol{\beta})\right)
\end{aligned}$$

We then have

$$\begin{aligned}\frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n}\left(Y_i\left(X_{j,i}\right) - \frac{X_{j,i}\exp\left(\boldsymbol{X^T\beta}\right)}{1+\exp(\boldsymbol{X^T\beta})}\right)\\ &= \sum_{i=1}^{n} X_{j,i}\left(Y_i - \frac{\exp\left(\boldsymbol{X^T\beta}\right)}{1+\exp(\boldsymbol{X^T\beta})}\right)\\ &= \sum_{i=1}^{n} X_{j,i}\left(Y_i - \frac{1}{1+\exp(-\boldsymbol{X^T\beta})}\right)\\ &= \sum_{i=1}^{n} X_{j,i}(Y_i - \pi_i)\end{aligned}$$

We then equate the partial derivatives to zero to get critical points.

What should the $X_i's$ be?

Generally, in a scorecard format, various factors are used to determine whether or not an account will go bad (eg: age, income, education, etc). However, these factors as is are in units. To avoid the inconsistencies from the units of variables, they can be standardized by using the WOE per grouped attribute in place for actual values. Recall that WOE is given by

$$\text{WOE} = \ln\left[\frac{P(c|Good)}{P(c|Bad)}\right]$$

Hence, our logistic regression model is given by

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 WOE_{1,i} + \beta_2 WOE_{2,i} + \cdots + \beta_k WOE_{k,i}$$

In this model, if π_i refers to the bad probability of account i and if a higher WOE can indicate a lower probability of default, then the $\beta_i < 0$ for $i = 1,2,3 \dots, k$.

II. Building the Logistic Regression Model

We will be using a development sample to build the logistic regression model, while the validation sample is to evaluate the performance of the model. Common splits are 70%-30% and 80%-20%.

A. Multicollinearity in Credit Scorecards

- A lot of independent variables in credit scorecards.
- Issue: we might include redundant/include variables i.e, some variables can be written as a linear combination of others.
- The individual effect of a single variable (if it also depends on other variables in the model) cannot be properly extracted.

A1. Checking for collinearity

Pairwise Correlation

Compute for the pairwise correlations $\rho_{p,q}$ for each variable pair X_p and X_q. The correlation is given by

$$\rho_{p,q} = \frac{\mathbb{E}\left[\left(X_p - \mathbb{E}[X_p]\right)\left(X_q - \mathbb{E}[X_q]\right)\right]}{\sigma_p\sigma_q}$$

With the estimator

$$r_{p,q} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_{p,i}X_{q,i} - \overline{X_p X_q}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{p,1} - \overline{X_p}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{q,1} - \overline{X_q}\right)^2}}$$

As the **sample correlation coefficient.**

Test for Multicollinearity

Assumption: X_p and X_q have a bivariate normal distribution. We test:

$$H_0\colon \rho_{p,q} = 0 \text{ vs } H_a\colon \rho_{p,q} \neq 0$$

With test statistic

$$T_{n-2} = \frac{\sqrt{n-2}\, r_{p,q}}{\sqrt{1 - r_{p,q}^2}}$$

Which has a Student t distribution with $n - 2$ degrees of freedom.

Remark: This has to be done for each pair of variables, which is difficult if k (the number of variables) is very large.

Variance Inflation Factor (VIF)

Problem with multicollinearity: parameter estimates will have large variances and covariances.

It describes the speed at which variances and covariances are increasing. The VIF is obtained by regressing an independent variable with the other independent variable and taking its adjusted R-square R^2, the VIF for each predictor is computed as

$$\text{VIF}_j = \frac{1}{1 - \widehat{R_k}^2}$$

If $\text{VIF}_j > 5.0$, this indicates a severe multicollinearity problem. These variables should be excluded from the model. If there is a group of independent variables with high VIFs, then include only one of those variables in the final model.

B. Variable Selection Procedures

1. **Significance of the variables.** For each parameter estimate $\widehat{\beta_j}$, we want to test whether the estimate is statistically significant. We have the following test.

$$H_0\colon \beta_j = 0 \quad \text{vs} \quad H_a\colon \beta_j \neq 0$$

If the $p -$ value given is less than a given level of significance, α, then we reject the null hypothesis and therefore the parameter estimate is significant.

2. **Akaike Information Criterion:** The Akaike Information Criterion is a penalized log-likelihood criterion, which means that it measures goodness-of-fit but penalizes the complexity of the model. The AIC is given by

$$\text{AIC} = -2l + 2k$$

Where l is the log-likelihood function $\ln(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}(Y_i \ln(\widehat{\pi_i}) + (1 - Y_i)\ln(1 - \widehat{\pi_i}))$, and k is the number of parameters.

A better model is the one with the lower AIC.

3. **Information Value:** Even before obtaining the parameters, some variables can be eliminated by taking note of the information value.

The selection of variables performed in R mainly makes use of variable significance or the AIC. Here are three methods:

1. **Forward Selection**
2. **Backward Selection**
3. **Stepwise Selection**

Forward Selection: The process starts with the null model, which is the model consisting of only an intercept term. Then the software determines which of the set of independent variables is the most statistically significant, given a significance level α. This variable is then added to the model, and the model parameters are refitted. The process continues until there are no more variables that are statistically significant.

On the basis of AIC, the model adds the variable such that when the model is refitted, the smallest AIC is obtained. This process is repeated until the it is better to stop adding variables than to add more, in terms of the AIC (i.e. if the current model has the lowest AIC compared to when one more variable is added to the model).

Backward Selection: In this procedure, all variables are initially included in the model. Then one-by-one, the software removes the variables that are least statistically significant, or if the removal results to a lower AIC. This process is repeated until all variables are statistically significant, or if the minimal AIC has been attained.

Stepwise Selection: Stepwise selection incorporates both forward and backward selection processes by removing or adding variables to the model until the "best" model has been constructed. The process ends when the incremental predictive power of adding the next variable is negligible, or if the optimal AIC has already been achieved with the current model.

While these are statistical methods to obtain the best model, ultimately, a big influence on the selection of variables will be on expert judgment. Some variables are deemed by senior management as crucial to the decision of credit-worthiness.

With this, we should be able to get the parameter estimates $\widehat{\boldsymbol{\beta}}$ and thus our model. The estimates of the probabilities π_i of default for account i can be obtained from an equation we had earlier

$$\widehat{\pi_\imath} = \frac{1}{1 + \exp\left(-\boldsymbol{X}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}\right)}$$

Which can be used to determine probabilities for new data sets.

C. Model Validation

C1. Goodness of Fit: **Hosmer-Lemeshow Test**

We test the hypotheses:

H_0: The model is a good fit vs H_a: The model is not a good fit

Here are the steps

The observations are first ranked in increasing order of $\widehat{\pi_\iota}$. Then two possible schemes can be carried out to categorize the data into deciles:

- The first group will contain the observations in the bottom 10%, the second group will contain the observations in the second 10%, and so on. Cut offs are determined on the probabilities; that is, observations with an estimated probability of less than the first cut off will fall into the first group, and so on.

	Probability of Default
1	1.48%
2	7.49%
3	21.38%
4	24.97%
5	31.09%
6	56.52%
7	56.57%
8	60.00%
9	63.05%
10	64.36%
11	65.33%
12	66.81%
13	68.98%
14	71.51%
15	86.79%
16	87.07%
17	89.85%
18	91.41%
19	93.55%
20	95.31%

The H-L statistic can be computed through

$$\chi^2_{HL} = \sum_{i=1}^{10} \frac{\left(O_j - N_j\bar{\pi}_j\right)^2}{N_j\bar{\pi}_j\left(1-\bar{\pi}_j\right)}$$

where N_j is the number of observations in the j^{th} decile, O_j is the number of bad accounts in the j^{th} decile, $\bar{\pi}_j$ is the average probability among bad accounts in the j^{th} decile.

Here, $\chi^2_{HL} \sim \chi(8)$. Hence, the null hypothesis is rejected if $\chi^2_{HL} > \chi_\alpha(8)$ or if the $p -$ value is less than α.

C3. Association Measures

Generally, if account i is "good" while account j is "bad", then we should expect that $\widehat{\pi_\iota} < \widehat{\pi_J}$. That is, a good model must associate higher default probabilities to bad accounts than good accounts. One thing that can be used to measure this is **Somer's D.** It has the following steps:

1. Consider all the possible $T = n_G \times n_B$ pairings of a good and bad account and the computed $\widehat{\pi_\iota}$ for each account.
2. Classify each pair as
   a. **Concordant** if $\widehat{\pi_B} > \widehat{\pi_G}$ (they agree on the rating)
   b. **Discordant** if $\widehat{\pi_B} < \widehat{\pi_G}$ (they disagree)
   c. **Tied** if $\widehat{\pi_B} = \widehat{\pi_G}$

The **Somer's D** statistic is given by

$$SD = \frac{(n_C - n_D)}{T}$$

Where n_C and n_D are the number of concordant and discordant pairs respectively.

Exercise: Give an example of a concordant pair and discordant pair from the given:

	Probability of Default
1	1.48%
2	7.49%
3	21.38%
4	24.97%
5	31.09%
6	56.52%
7	56.57%
8	60.00%
9	63.05%
10	64.36%
11	65.33%
12	66.81%
13	68.98%
14	71.51%
15	86.79%
16	87.07%
17	89.85%
18	91.41%
19	93.55%
20	95.31%

A concordant pair is given by (1,20) while a discordant pair is given by (3,4)

C4. Contingency Table

The prediction of whether or not an account is a good or a bad account depends on a cut off probability $z \in [0,1]$. More specifically,

$$\hat{Y}_i = \begin{cases} 1 & \text{if } \hat{\pi}_i > z \\ 0 & \text{if } \hat{\pi}_i \leq z \end{cases}$$

Remark: $\hat{Y}_i$ is a prediction, which can sometimes be different from the actual value Y_i.

A good model is a model for which most predictions are equal to the actual values. As a basis for evaluation, we can construct a **confusion matrix/contingency table,** which is summarized below:

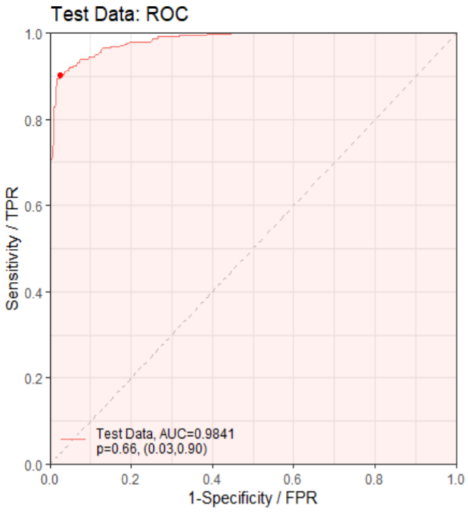
	Predicted 0	Predicted 1
Actual 0	**True Negative (TN)**	**False Positive (FP)**
Actual 1	**False Negative (FN)**	**True Positive (TP)**

Let n_B and n_G be the number of actual good and actual bad accounts respectively. The following ratios can be obtained from the four quantities mentioned earlier.

1. $TPR = \frac{TP}{n_B}$ (sensitivity; probability of detection)
2. $FPR = \frac{FP}{n_G}$ (fall-out; probability of false alarm)
3. $TNR = \frac{TN}{n_G}$ (specificity)
4. $FNR = \frac{FN}{n_B}$ (miss rate)
5. $Acc = \frac{TP+TN}{n_B+n_G}$ (accuracy)
6. $Err = \frac{FP+FN}{n_B+n_G}$ (error rate)

C5. Receiver Operating Characteristic (ROC) curve

The ROC curve is used to evaluate the goodness-of-fit of a model. It is also used to compare predictive powers of several logistic models. The ROC curve is a plot of the true positive rate $TPR(z)$ against the false positive rate $FPR(z)$ across different cut-offs $z \in [0,1]$



The predictive power of the model can be obtained by the area under the curve **AUC** or the **C −statistic**. A perfect model is both perfectly specific and perfectly sensitive, and so the area under the ROC curve of the perfect model is 1.0. An area of 0.5 under the ROC indicates absence of any predictive power. *An area of 0.5 indicates that the ROC curve is just the equation* $y = x$.

Another interpretation of the AUC is that it is the probability that a randomly selected good account has a higher score than that of a randomly selected bad account.

C6. Empirical Distribution and Cut-off Probabilities

Before we determine the optimal probabilities, we need to first to define the **empirical cumulative distribution** of the good and bad accounts. Let B and G be the set of bad accounts and good accounts respectively, with n_B and n_G observations each.

The **empirical cumulative distribution of bad accounts** is given by

$$F_B(z) = \frac{1}{n_B} \sum_{i \in B} \mathbf{1}(\widehat{\pi_\iota} \le z)$$

This describes the proportion of bad accounts that are predicted to be good if the cut off probability is $z \in [0,1]$. On the other hand, the **empirical cumulative distribution of good accounts** is given by

$$F_G(z) = \frac{1}{n_G} \sum_{i \in B} \mathbf{1}(\widehat{\pi_\iota} \le z)$$

Which describes the proportion of good accounts that are predicted to be good if the cut-off probability is $z \in [0,1]$.

We want to choose the cut-off z such that the empirical cdfs of the good and bad accounts are as different from each other as possible. That is, our cut-off probability is the value z that maximizes $|F_B(z) - F_G(z)|$.

Now, the maximum value, given by $KS = \sup_z |F_B(z) - F_G(z)|$ is actually the **Kolmogorov-Smirnov (KS) statistic,** which is used to determine whether the distribution of the "bad" accounts is significantly different from the distribution "good" accounts. We test:

H_0: The two distirbutions are the same vs H_a: The two distributions are not the same

We reject H_0 at the significance level of α if

$$KS \geq k(\alpha)\sqrt{\frac{1}{n_B} + \frac{1}{n_G}}$$

With $k(\alpha) = \sqrt{-\frac{1}{2}\ln\left(\frac{\alpha}{2}\right)}$

C7. The Cumulative Accuracy Profile

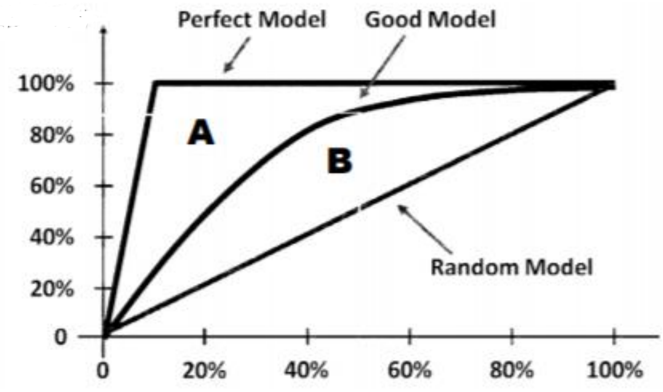
The CAP visually describes the descriptive power of a model. The y-axis is the 1-ECDF of the bad accounts and the x-axis is the 1-ECDF of all accounts, i.e the CAP is the graph of the points $\left(1 - F(z), 1 - F_B(z)\right)$ as z ranges from [0,1]

The empirical distribution of all accounts is given by

$$F(z) = \frac{1}{n_B + n_G}\sum_{i \in B \cup G} \mathbf{1}(\hat{\pi}_\iota \leq z)$$

This means we are plotting

$$1 - F(z) = \frac{1}{n_B + n_G}\sum_{i \in B \cup G} \mathbf{1}(\hat{\pi}_\iota > z) \quad \text{vs} \quad 1 - F_B(z) = \frac{1}{n_B}\sum_{i \in B} \mathbf{1}(\hat{\pi}_\iota > z)$$



A perfect model is such that there is a cut off probability that splits all accounts with the bad accounts. The accuracy ratio is given by the formula

$$AR = \frac{B}{A + B}$$

D. Final Scorecard

Goal: Let us transform the probabilities of default π_i into something easier to understand: **Scores.** Scores are generally a linear transformation

$$\text{Score} = \text{Offset} + \text{Factor} \times \ln(\text{Odds})$$

However, most scorecards specify a certain **good:bad odds** at a certain score and the points required to double the odds p_2.

$$\text{Score} + p_2 = \text{Offset} + \text{Factor} \times \ln(2 \times \text{odds})$$

It follows that

$$\text{Factor} = \frac{p_{do}}{\ln 2}$$

$$\text{Offset} = \text{Score} - \text{Factor} \times \ln(\text{odds})$$

```
#---------------------------------------------------------------------------------
# Final Scorecard
#---------------------------------------------------------------------------------
scorecard(bins,
          reg_backward,
          points0 = 600, # points at the specified odds
          odds0 = 1/50, # this is the bad/good odds
          pdo = 20, # points needed to double the odds
          basepoints_eq0 = FALSE)
```

We know that

$$\begin{aligned}
\text{Score} &= \text{Offset} + \text{Factor} \times \ln(\text{Odds}) \\
&= \text{Offset} + \text{Factor} \times \ln\left(\frac{1-\pi_i}{\pi_i}\right) \\
&= \text{Offset} + \text{Factor} \times \left[-\ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] \\
&= \text{Offset} + \text{Factor} \times \left[-(\beta_0 + \beta_1 WOE_{1,i} + \beta_2 WOE_{2,i} + \cdots + \beta_k WOE_{k,i})\right] \\
&= \text{Offset} - \text{Factor} \times \beta_0 - \text{Factor} \times \beta_1 WOE_{1,i} - \text{Factor} \times \beta_2 WOE_{2,i} \ldots - \text{Factor} \times \beta_k WOE_{k,i}
\end{aligned}$$