Co-financed by the Connecting Europe Facility of the European Union

Model Selection

Trustworthy AI - Lecturer: Emanuela Raffinetti; Python instructor: Alex Gramegna
E-mail: emanuela.raffinetti@unipv.it; alex.gramegna01@universitadipavia.it

Interpreting the multiple linear regression model - I

- We have studied R^2 as a measure of goodness of fit for the linear model.
- An important aim of the analysis, in a multiple regression setting, is to understand not only the absolute contribution of the fitted plan to the explanation of the variability of Y, as expressed by R^2, but also the determination of the partial contribution of each explanatory variable.
- To such end we now examine in detail the variance decomposition identity for a multiple regression model with p regressors. It can be demonstrated that:

$$Var(Y) = \sum_{j=1}^{p} \beta_j Cov(X_j, Y) + Var(\varepsilon)$$

Interpreting the multiple linear regression model - II

- We know that, in the simple linear regression model, $\beta = \frac{Cov(X,Y)}{Var(X)}$
- If in the multiple linear regression model we had $\beta_j = \frac{Cov(X_j,Y)}{Var(X_j)}$

  then it would follow

  $$Var(Y) = \sum_{j=1}^{p} Var(Y)\rho^2_{X_j,Y} + Var(\varepsilon)$$

  and therefore

  $$Var(\hat{Y}) = Var(\hat{Y}_1) + Var(\hat{Y}_2) + ... + Var(\hat{Y}_p)$$

  so that

  $$R^2 = \sum_{j=1}^{p} \rho^2_{Y,X_j}$$

  that is, the variance of Y explained by the fitting plane would be equal to the sum of the variance of Y explained by each of the fitting lines, built separately for each of the explanatory variables.

Interpreting the multiple linear regression model - III

In general, we have instead the following recursive relationship:

$$R^2 = \sum_{j=1}^{p} \rho^2_{Y,X_j|X_{i<j}} (1 - R^2_{Y,X_1,\ldots,X_{j-1}})$$

where $R^2_{Y,X_1,\ldots,X_{j-1}}$ indicates the coefficient of multiple correlation between Y and the fitted plane determined by the explanatory variables $X_1, \ldots, X_{j-1}$, while $\rho^2_{Y,X_j|X_{i<j}}$ indicates the partial correlation coefficient between Y and X_j, conditional on the "previous" variables $X_1, \ldots, X_j - 1$.

This means that adding a new variable X_j reduces the quote of variance not explained by the regression of Y on the previously added variables of a fraction equal to the square of the partial correlation coefficient between itself and the response variable, conditional on the variables already present.

Interpreting the multiple linear regression model - IV

Summarising:

- the contribution of a single explanatory variable, say X_j, to the fitting plane is additive, and therefore R^2 always increases as the number of variables increases.
- However, the increase is not necessarily equal to ρ^2_{Y,X_j}. This occurs only in the uncorrelated case.
- In general, it can be lesser or greater, according to the degree of correlation of the response variable with the regressors already present, and of the latter with X_j.

Selecting the set of regressors: F-test - I

- Recall the total variance decomposition:

$$MSTO = \frac{SSTO}{N-1},\ MSE = \frac{SSE}{N-p-1},\ MSR = \frac{SSR}{p}$$

- We can represent it with the so-called ANOVA table:

Source	Formula	DF
SSTO	$\sum_{i=1}^{N}(y_i - \bar{y})^2$	$N-1$
SSE	$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$	$N-p-1$
SSR	$\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$	p

- The f-statistic is calculated as follows:

$$f = \frac{MSR}{MSE}$$

Selecting the set of regressors: F-test - II

- Under the normal linear model assumptions, the f-statistic is distributed as an F variable with parameters p, $N-p-1$:

$$f \sim F(p, N-p-1)$$

- If $f > F_{1-\alpha}(p, N-p-1)$, where α is the significance level, then we reject the null hypothesis that all βs are equal to zero.
- Correspondingly, we reject the null hypotheses when the associated p-value is small.

Nested Linear Models - I

- We were actually comparing the model with all our available regressors with the model only including the intercept.
- More in general, we can use the F-test to compare two *nested* models, like:

$$y_i = \alpha + \beta_1 x_{1i} + \varepsilon_i \quad (1)$$
$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \quad (2)$$

- (1) is the *reduced* model and is nested within (2) which is the *full* model.

Nested Linear Models - II

- Including additional explanatory variables always increases the explained variance, but it should be verified whether the improvement is "large enough".

- Parsimonious models (models with a small number of predictors) are preferred as long as their goodness of fit is "similar" to that of more complex models.

F-test for model selection - I

- When we compare two models, the f-statistic can be calculated as follows:

$$f = \frac{(SSE_R - SSE_F)/k}{SSE_F/(N - p - k - 1)}$$

where k is the number of additional explanatory variables in the full model (F), while p is the number of explanatory variables in the reduced model (R).

- And we have that

$$f \sim F(k, N - p - k - 1)$$

F-test for model selection - II

- In our example with models (1) and (2), we have $p = 1$, $k = 2$. Then we calculate

$$f = \frac{(SSE_1 - SSE_2)/2}{SSE_2/(N-4)}$$

and we reject the null hypothesis that $\beta_2 = \beta_3 = 0$ if

$$f > F_{1-\alpha}(2, N-4)$$

- Correspondingly, we reject the null hypotheses when the associated p-value is small.

F-test for model selection - III

When comparing two models, one with p regressors, one with $p+1$, the f-statistic can also be expressed in terms of partial correlation coefficients as

$$f = \frac{\rho^2_{Y,X_{p+1}|X_1,\ldots,X_p}/1}{(1-R^2_{Y,X_1,\ldots,X_p})/(N-p-2)}$$

from which we get

$$f = \frac{Var(\hat{Y}_{p+1}) - Var(\hat{Y}_p)}{(Var(Y) - Var(\hat{Y}_p))/(N-p-2)}$$

Therefore, the f-statistic can be interpreted as the ratio between the additional variance explained by the $p+1$-th variable and the mean residual variance
In other terms, it expresses the relative importance of the $p+1$-th variable.

T-test - I

- Remind that, when $k = 1$ (only one regressor added), performing an f-test is equivalent to performing a t-test:

$$f = t^2$$

$$t = \frac{\hat{\beta}_1}{\sigma(\hat{\beta}_1)}$$

- Through the t-test we verify the null hypothesis that a single coefficient is equal to zero.

- $\sigma(\hat{\beta}_1)$ is the standard error of $\hat{\beta}_1$.

T-test - II

```
Call:
lm(formula = btc_coinbase ~ ., data = data2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.052330 -0.003256  0.000023  0.002987  0.052098

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0001059  0.0003123   0.339  0.73454
btc_kraken   0.0210321  0.0123977   1.696  0.09025 .
btc_bitstamp 0.0384385  0.0359272   1.070  0.28503
btc_itbit    0.0130343  0.0256007   0.509  0.61082
btc_bitfinex 0.2297741  0.0315236   7.289 8.47e-13 ***
btc_hitbtc   0.0821093  0.0184755   4.444 1.03e-05 ***
btc_gemini   0.5981632  0.0308680  19.378  < 2e-16 ***
btc_bittrex  0.0056419  0.0145595   0.388  0.69850
usdyuan     -0.1045943  0.2066436  -0.506  0.61291
usdeur       0.2060414  0.0986501   2.089  0.03710 *
gold         0.0712161  0.0575053   1.238  0.21597
oil         -0.0595675  0.0192726  -3.091  0.00208 **
sp500       -0.0952889  0.0569865  -1.672  0.09495 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure: R output of a linear regression model

T-test - III

- For a significance level α, we reject the null hypothesis that $\beta_1 = 0$ if

$$|t| > T_{1-\alpha/2}$$

where T follows a Student's t distribution with $(N - p - 1)$ degrees of freedom.

- Correspondingly, we reject the null hypotheses when the associated p-value is small.

Stepwise regression

- The procedure that allows to choose the set of predictive variables in a model is called **stepwise regression**. It selects the final model through subsequent hypothesis tests, in each of which two alternative models will be compared.
- Stepwise algorithms can be:
  - **Forward**: they start with the model with no predictors
  - **Backward**: they start with the model with all the specified predictors.
- In each step, a variable is considered for addition to or subtraction from the initial set of explanatory variables, based on some model selection criterion.

Evaluating performance: error measures - I

- A way of comparing models (nested or not) in terms of their performance is using measures based on the distance between observed and fitted values.
- Among them, the most commonly used are the mean squared error (MSE), the root mean squared error (RMSE), the mean absolute error (MAE) and the mean absolute percentage error (MAPE).
- We prefer models with lower values of the error measures.

Evaluating performance: error measures - II

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$ → **Mean Squared Error**

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$ → **Root Mean Squared Error**

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|(\hat{y}_i - y_i)|$$ → **Mean Absolute Error**

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{(\hat{y}_i - y_i)}{y_i}\right|$$ → **Mean Absolute Percentage Error**

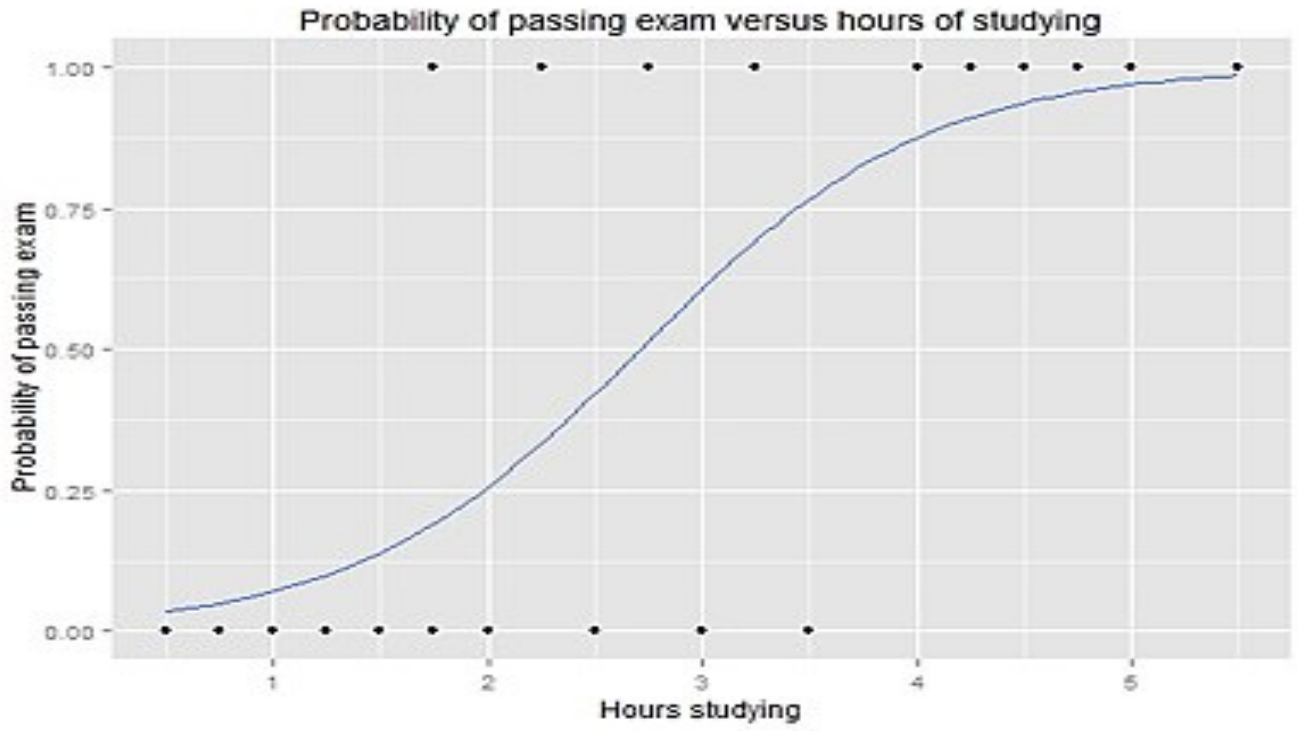
Interpreting the logistic regression model - I

From an interpretational viewpoint, the logit function implies that the dependence of probability π that $Y = 1$ on the explanatory variables is described by a sigmoid or S-shaped curve.
Precisely, inverting the expression that defines the logit function it is obtained that:

$$\pi_i = \frac{\exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})}{1 + \exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})}$$

Interpreting the logistic regression model - II

When only one regressor is considered, it is easy to represent the relationship between the latter and the response:



Interpreting the logistic regression model - III

The parameter β determines the rate of growth or increase of the curve, in particular the sign of β indicates if the curve increases or decreases while the magnitude determines the velocity with which the curve does so:

- When $\beta > 0$, $\pi(x)$ increases as x increases
- When $\beta < 0$, $\pi(x)$ decreases as x increases
- For $\beta \to 0$, the curve tends to become a horizontal straight line. In particular, when $\beta = 0$, Y results independent of X.

Interpreting the logistic regression model - IV

To properly interpret β, recall the linear relationship found for the log-odds:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

Positive log-odds is evidence in favour of $Y = 1$ while negative log-odds favours $Y = 0$.
The previous expression establishes that the logit increases of β units in correspondence to a unit increase in x.

Interpreting the logistic regression model - V

It is also useful to consider the exponential of the previous expression, moving to the odds:

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\alpha + \beta x) = \exp(\alpha)\exp(\beta)^x$$

This exponential relationship provides an useful interpretation of the parameter:

- the odds increase in a multiplicative way, of $\exp(\beta)$, in correspondence to every unitary increase of x;
- in other terms, the odds at level $x+1$ are equal the odds at level x multiplied by $\exp(\beta)$;
- note that, when $\beta = 0$, we obtain $exp(\beta) = 1$ and, therefore, the odds do not depend on X.

Interpreting the logistic regression model - VI

To extend the interpretation of β to the case in which we consider p regressors, we should recall that, while the probability of success is a logistic function and therefore not linear in the explanatory variables, the logarithm of the odds is a linear function of the explanatory variables:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Thus,

- β_j is the increase in the logit in correspondence to a unit increase in x_j *ceteris paribus*;
- in other terms, the odds at level $x_j + 1$ are equal the odds at level x_j multiplied by $\exp(\beta_j)$ as long as the levels of the other x variables remain the same.

Testing the significance of logistic regression coefficients

We have already considered testing the null hypothesis $\beta_j = 0$, against the alternative $\beta_j \neq 0$. If the sample size is sufficiently large, the Wald's statistics:

$$Z = \frac{\hat{\beta}_j}{\sigma\left(\hat{\beta}_j\right)}$$

where $\sigma\left(\hat{\beta}_j\right)$ indicates the standard error of the estimator, is approximately distributed as a standardised normal.
Therefore, to decide whether to accept or reject the null hypothesis we can use the following rejection region:

$$R = \{|Z| > z_{1-\alpha/2}\}$$

with $z_{1-\alpha/2}$ the $1-\alpha/2$ 100% percentile of a standard normal distribution, and reject when the associated p-value is small.

Testing the significance of the logistic regression model

The overall significance of a logistic regression model can be evaluated by taking the difference in **deviance** between the considered model and the saturated one (the model having as many parameters as observations, thus leading to a perfect fit), obtaining the following statistic:

$$G^2(M) = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right]$$

where the $\hat{\pi}_i$ terms are the fitted probabilities of success based on model M.

The value of G^2 should be compared with critical values of a $\chi^2(N-p)$ distribution. "Small" values of G^2 indicate a good fit: the model is close to the more complex "saturated" model.
Note that G^2 can be interpreted as a distance function, expressed in terms of entropy differences between the fitted model and the saturated model.

Comparing logistic regression models

The deviance statistic can also be used to compare two nested logistic regression models, model M_A, with q parameters, and model M_B, with p parameters, with $q < p$:

$$D = G^2(M_A) - G^2(M_B) =$$
$$2\sum_{i=1}^{N}\left[y_i \log\left(\frac{n_i\hat{\pi}_i^B}{n_i\hat{\pi}_i^A}\right) + (n_i - y_i)\log\left(\frac{n_i - n_i\hat{\pi}_i^B}{n_i - n_i\hat{\pi}_i^A}\right)\right]$$

where $\hat{\pi}_i^A$ and $\hat{\pi}_i^B$ indicate the success probabilities fitted, respectively, on the basis of models M_A and M_B
and

$$D \sim \chi^2_{p-q}$$

Small values of indicate that model A is not much worse than the more complex model B. Correspondingly, the p-value is large and we do not reject the simpler model (which is in H0).

Evaluating performance: confusion matrix - I

- When the dependent variable is binary, to assess the accuracy of a model we consider its ability to classify the observations.
- The confusion matrix, containing information about actual and predicted classifications, is typically used.

	Predicted: 0	Predicted: 1
Actual: 0	TN (True negative)	FP (False positive)
Actual: 1	FN (False negative)	TP (True positive)

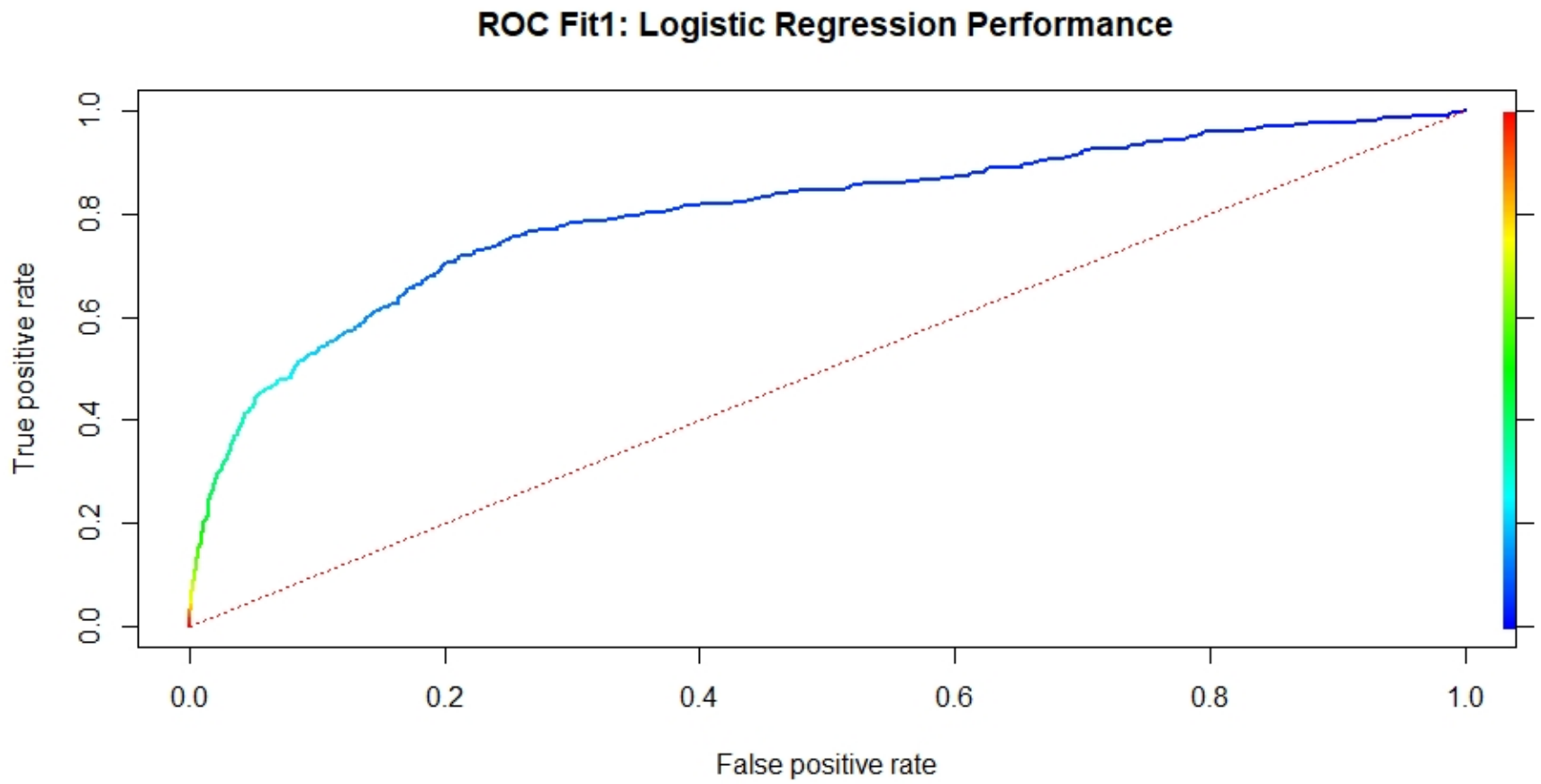
Evaluating performance: confusion matrix - II

- Several measures can be obtained from the confusion matrix.
  - Accuracy $= \frac{TP + TN}{TP + FP + TN + FN}$
  - Sensitivity or Recall $= \frac{TP}{TP + FN}$
  - Specificity $= \frac{TN}{TN + FP}$
  - Precision $= \frac{TP}{TP + FP}$
- These measures depend on the chosen cut-off.
- We need a measure which takes all possible cut-offs into account and summarises predictive accuracy into a single statistics.

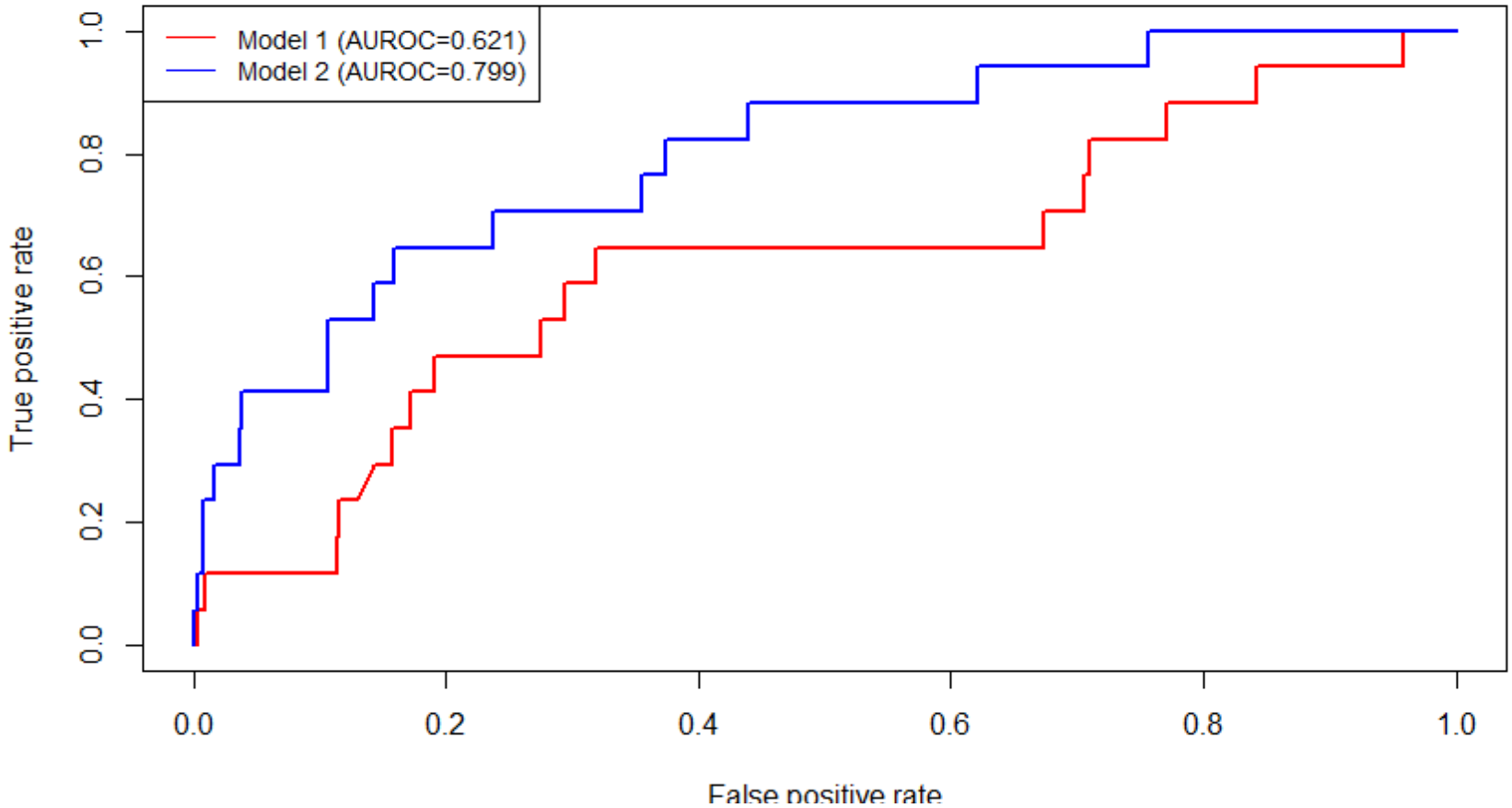
Evaluating performance: the ROC curve and the AUROC - I

- A widely used predictive accuracy measure for binary responses is the Area Under the Receiver Operating Characteristics curve (AUROC).
- The ROC curve displays the relationship between the false positive rate (1- specificity, on the x-axis) and the true positive rate (sensitivity, on the y-axis), across a series of predetermined cut-off points.
- The bisector line corresponds to a random prediction model with no ability to classify the observations. In this case the AUROC is equal to 0.5.
- The ideal curve coincides with the y-axis between 0 and 1, and the AUROC, in this case, is equal to 1.

Evaluating performance: the ROC curve and the AUROC - II



Evaluating performance: the ROC curve and the AUROC - III



Reference

- Kutner M.H., Nachstsheim C.J., Neter J.: Applied Linear Regression Models, 4° edition, McGraw Hill Irwin (2004), available at: https://www.academia.edu/32804953/Applied_Linear_Regression_Models_4th_edition