

MASM22/FMSN30/FMSN30F: Linear and Logistic Regression, 7.5 hp

FMSN40: ... with Data Gathering, 9 hp

Lecture 9, spring 2026

Goodness-of-fit

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

4/5-26

Goodness of fit – Classification

Confusion matrix

Example

Cohen's kappa

McNemar test

ROC-curve

ROC-curve

Optimal threshold

Area Under the Curve

Goodness of fit — Classification

Sometimes we want to use our model to classify future objects as "success" or "failure", depending on the probabilities given by their x -values.

The easiest is to classify using the most probable category:

$$\hat{Y}_i = \begin{cases} \text{failure} & \text{if } \hat{p}_i \leq 0.5, \\ \text{success} & \text{if } \hat{p}_i > 0.5 \end{cases}$$

Note: there are situations where we will want a different threshold than 0.5.

- ▶ We want to examine the proportion of the observations that are correctly classified.

Confusion matrix

Predicted (\hat{Y}_i)	True (Y_i)		Total
	Failure ($Y_i = 0$)	Success ($Y_i = 1$)	
Failure ($\hat{p}_i \leq 0,5$)	true negative	false negative	TN + FN
Success ($\hat{p}_i > 0,5$)	false positive	true positive	FP + TP
Total	TN + FP	FN + TP	n

Statistics

- **Sensitivity.** The proportion of the true successes that have been correctly classified as successes: **true positive rate** or

$$\text{recall} = \Pr(\hat{Y}_i = 1 \mid Y_i = 1) = \frac{\text{TP}}{\text{FN} + \text{TP}}.$$

- **Specificity.** The proportion of the true failures that have been correctly classified as failures: **true negative rate**

$$= 1 - \text{false positive rate} = \Pr(\hat{Y}_i = 0 \mid Y_i = 0) = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

Statistics (cont.)

- ▶ **Accuracy.** The overall proportion that have been correctly classified $= \frac{TP + TN}{n}$.
- ▶ **Positive predictive value (PPV).** The proportion of the predicted successes that are true successes: **precision**
 $= \Pr(Y_i = 1 \mid \hat{Y}_i = 1) = \frac{TP}{FP + TP}$
- ▶ **Negative predictive value (NPV).** The proportion of the predicted failures that are true failures
 $= \Pr(Y_i = 0 \mid \hat{Y}_i = 0) = \frac{TN}{FN + TN}$

We would prefer all of these to be large. However, there will be a tradeoff between them.

Example: The Oslo model

The largest model: $I(\text{cars}/1000) * \text{windspeed} + \text{tempdiff}$:

Predicted	True		Total	
	Low(0)	High(1)		
Low(0)	374 (TN)	90 (FN)	464	NPV = $374/464 = 80.6\%$
High(1)	12 (FP)	24 (TP)	36	PPV = $24/36 = 66.7\%$
Total	386	114	500	Accuracy = $(374 + 24)/500 = 79.6\%$
	Specificity = $374/386 = 96.6\%$	Sensitivity = $24/114 = 21.1\%$		

The sensitivity is quite bad.

Additional statistics

From `confusionMatrix()` in the `caret` R-package we also get:

- ▶ Prevalence = $\bar{Y} = \frac{FN + TP}{n} = \frac{114}{500} = 22.8\%$.
- ▶ Detection rate = $\frac{TP}{n} = \frac{24}{500} = 4.8\%$.
- ▶ Detection prevalence = $\frac{FP + TP}{n} = \frac{36}{500} = 7.2\%$.
- ▶ Balanced accuracy = $\frac{1}{2}(\text{Sensitivity} + \text{Specificity}) =$
 $= \frac{0.211 + 0.966}{2} = 59.0\%$.

Additional statistics, cont'd.

- ▶ A confidence interval for the accuracy (based on an exact test for a Binomial distribution, cf. profile likelihood intervals).
- ▶ No Information Rate = $\frac{386}{500} = 77.2\%$. The proportion correctly classified when assigning everything to the most prevalent class.
- ▶ A test for whether the accuracy is significantly higher than the No Information Rate. In this case, it's not.
- ▶ Kappa = $\kappa = 0.236$. Measure of how much closer to 100% the accuracy is, compared to random assignment. In this case, less than 1/4 of the distance.
- ▶ McNemar's test of whether the proportions classified as negative and positive are the same as the observed proportions. In this case, they are not. There are significantly more false negatives (90) than false positives (12).

Cohen's κ (kappa) coefficient

Measures how well the observed and predicted outcomes agree, compared to what would be expected by chance.

$$\kappa = \frac{\text{Accuracy} - p_c}{1 - p_c} = \frac{0.796 - 0.722}{1 - 0.722} = 0.2364$$

where p_c is the expected accuracy if the true (Y_i) and predicted (\hat{Y}_i) classifications are independent:

$$\begin{aligned} p_c &= \underbrace{\frac{\text{TN} + \text{FN}}{n}}_{\text{predicted failure}} \cdot \underbrace{\frac{\text{TN} + \text{FP}}{n}}_{\text{true failure}} + \underbrace{\frac{\text{FP} + \text{TP}}{n}}_{\text{predicted success}} \cdot \underbrace{\frac{\text{FN} + \text{TP}}{n}}_{\text{true success}} = \\ &= \frac{464}{500} \cdot \frac{386}{500} + \frac{36}{500} \cdot \frac{114}{500} = 72.2\% \end{aligned}$$

$\kappa = 1$ perfect (Accuracy = 100%);

$\kappa = 0$ same as random; $\kappa < 0$ worse than random.

McNemar's test for paired binary data

Test for whether the predicted proportions of positives and negatives are the same as the observed proportions, i.e.,

$$\underbrace{\frac{TN + FN}{n}}_{\text{predicted failure}} = \underbrace{\frac{TN + FP}{n}}_{\text{true failure}} \quad \text{and} \quad \underbrace{\frac{TP + FP}{n}}_{\text{predicted success}} = \underbrace{\frac{TP + FN}{n}}_{\text{true success}}.$$

This is equivalent to $FN = FP$, i.e., the number of false positives is the same as the number of false negatives.

The **McNemar test** rejects this if

$$\chi^2 = \frac{(FP - FN)^2}{FP + FN} > \chi_{\alpha}^2(1) \quad (\text{FP and FN must be large enough})$$

Note: `confusionMatrix()` uses the continuity corrected version $\chi^2 = \frac{(|90-12|-1)^2}{90+12} = 58.1 > \chi_{0.05}^2(1) = 3.84$, as an approximation of an exact test.

Warnings

- ▶ There is no easy way to "punish" the addition of more x -variables.
- ▶ Larger models generally have higher values when predicting the same data that was used when estimating the model, due to over-fitting.
- ▶ If the main purpose of the model is to classify future observations, we should validate on a separate data set, not involved when fitting the model.
Note: The caret package has a large number of functions for this but we won't use them in this course.
- ▶ On the other hand, if the model cannot even predict its own data, it is not a very good model.

Goodness of fit – Classification

Confusion matrix

Example

Cohen's kappa

McNemar test

ROC-curve

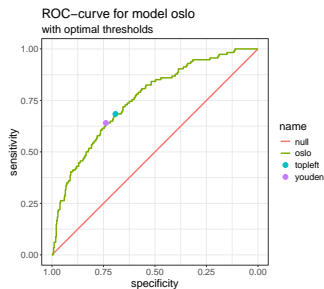
ROC-curve

Optimal threshold

Area Under the Curve

ROC-curve

- ▶ By changing the threshold value from 0.5 to something else we can change the specificity and sensitivity. We can make either of them as large as we like, but at the cost of the other becoming small.
- ▶ It may often be important to have both a large specificity *and* a large sensitivity,
- ▶ We could try all possible threshold values and calculate the specificity and sensitivity for each one.
- ▶ This ability of the model to separate two categories is illustrated in the **ROC-curve** (Receiver Operating Characteristics)



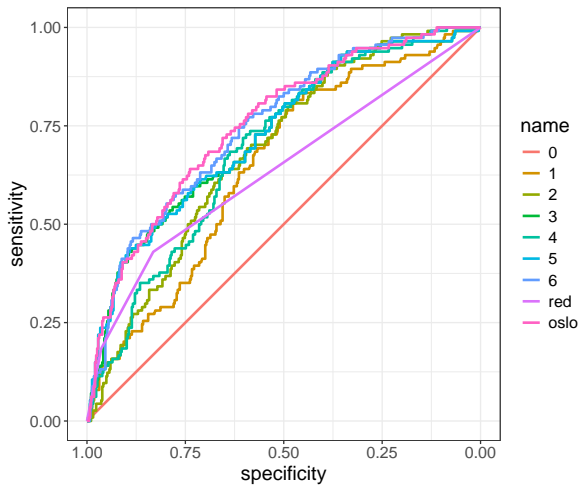
- ▶ Note the reversed scale on the x -axis!
- ▶ Using the optimal threshold $\hat{p}_i > 0.233$ gives a sensitivity of 68.4% and a specificity of 69.2%.

Optimal threshold

Use the ROC-curve to choose an optimal threshold value:

- ▶ "youden": the point farthest from the diagonal, maximizing the balanced accuracy $\frac{1}{2}(\text{Sensitivity} + \text{Specificity})$
- ▶ "closest.topleft": the point closest to the top left, ideal, corner, minimizing $(1 - \text{Sensitivity})^2 + (1 - \text{Specificity})^2$

ROC-curves for all the models



Area Under the Curve (AUC)

- ▶ The area under the ROC-curve (**AUC**) measures how close we are to the ideal curve. Ideal area = 1. Null model (toss a coin) = 0.5. Areas below 0.5 are worse than "toss a coin".
- ▶ The AUC is the probability that, if you take a random pair of observations, where one is a success and the other a failure, the success has a higher predicted probability of being a success than the failure does. The AUC thus gives the probability that the model correctly ranks such pairs of observations.
- ▶ There are several techniques for calculating confidence intervals for AUC and testing whether two ROC-curves have the same AUC.

Model	AUC	95 % C.I.	P-value
0:null	50.0	(50.0, 50.0)	< 0.001
1:cars	64.8	(59,3, 70.3)	< 0.001
2:cars+wind	68.6	(63.5, 73.7)	0.002
red:tempdiff	63.8	(58.8, 68.9)	< 0.001
3:cars+zerodiff	72.8	(67.5, 78.1)	0.03
4:cars*wind	69.5	(64.3, 74,6)	0.003
5:cars+tempdiff	73.0	(67.7, 78.2)	0.046
6:cars*wind + zerodiff	74.9	(69.9, 79.8)	0.12
oslo:cars*wind+tempdiff	75.5	(70.5, 80.4)	-

- ▶ The largest model has the largest AUC, as expected.
- ▶ Only model 6 is not significantly different from "oslo".
- ▶ On the other hand, the differences between model 3, 5, 6, and oslo, are small.
- ▶ General conclusion: the number of cars and some type of temperature difference should be in the model.