MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: . . . with Data Gathering, 9 hp

Lecture 9, spring 2024

Goodness-of-fit

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

2/5-24

Goodness of fit – Classification
- Confusion matrix
- Example
- Cohens kappa
- McNemar test

ROC-curve
- ROC
- AUC

Goodness-of-fit test
- Hosmer-Lemeshow

Goodness of fit — Classification

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

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

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

	True (Y_i)		
Predicted ($\hat{Y}_i$)	Failure ($Y_i = 0$)	Success ($Y_i = 1$)	Total
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 **recall** $= Pr(\hat{Y}_i = 1 \mid Y_i = 1) = \dfrac{\text{TP}}{\text{FN} + \text{TP}}$.
- **Specificity**. The proportion of the true failures that have been correctly classified as failures: **true negative rate** $= 1$ – **false positive rate** $= Pr(\hat{Y}_i = 0 \mid Y_i = 0) = \dfrac{\text{TN}}{\text{TN} + \text{FP}}$.

Statistics (cont.)

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

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(cars/1000)*windspeed + tempdiff`:

	True			
Predicted	Low(0)	High(1)	Total	
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 Specificity = $374/386 = 96.6\,\%$	114 Sensitivity = $24/114 = 21.1\,\%$	500	Accuracy = $(374 + 24)/500 = 79.6\,\%$

The sensitivity is quite bad.

Additional statistics

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

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

Additional statistics, cont'd.

- A confidence interval for the accuracy (based on an exact test for a Binomial distribution, cf. partial 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 independant:

$$p_c = \underbrace{\frac{\mathrm{TN} + \mathrm{FN}}{n}}_{\text{predicted failure}} \cdot \underbrace{\frac{\mathrm{TN} + \mathrm{FP}}{n}}_{\text{true failure}} + \underbrace{\frac{\mathrm{FP} + \mathrm{TP}}{n}}_{\text{predicted success}} \cdot \underbrace{\frac{\mathrm{FN} + \mathrm{TP}}{n}}_{\text{true success}} =$$

$$= \frac{464}{500} \cdot \frac{386}{500} + \frac{36}{500} \cdot \frac{114}{500} = 72.2\,\%$$

$\kappa = 1$ 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.,

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

This is equivalent to $\mathrm{FN} = \mathrm{FP}$, i.e., the number of false positives is the same as the number of false negatives.
The **McNemar test** rejects this if

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

Note: `confusionMatrix()` uses the continuity corrected version $\chi^2 = \frac{(|90-12|-1)^2}{90+12} = 58.1 > \chi^2_{0.05}(1) = 3.84$, 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
- Cohens kappa
- McNemar test

ROC-curve

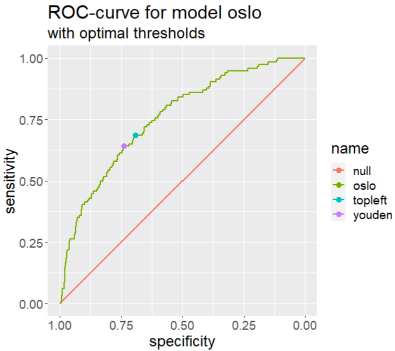
- ROC
- AUC

Goodness-of-fit test

- Hosmer-Lemeshow

ROC-curve

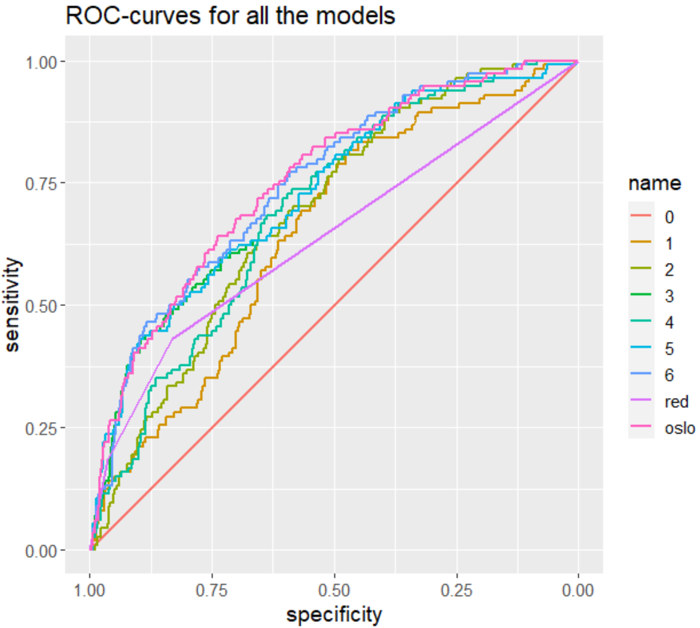
- By changing the threshold value from 0.5 to something else we can change the specificity and sensitivity. We can make either of them as large as we like, but at the cost of the other becoming small.
- It may often be important to have both a large specificity *and* a large sensitivity,
- We could try 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 %.

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
sensitivity
specificity
1.00
0.75
0.50
0.25
0.00
1.00
0.75
0.50
0.25
0.00
name
0
1
2
3
4
5
6
red
oslo

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.

Goodness of fit – Classification

- Confusion matrix
- Example
- Cohens kappa
- McNemar test

ROC-curve

- ROC
- AUC

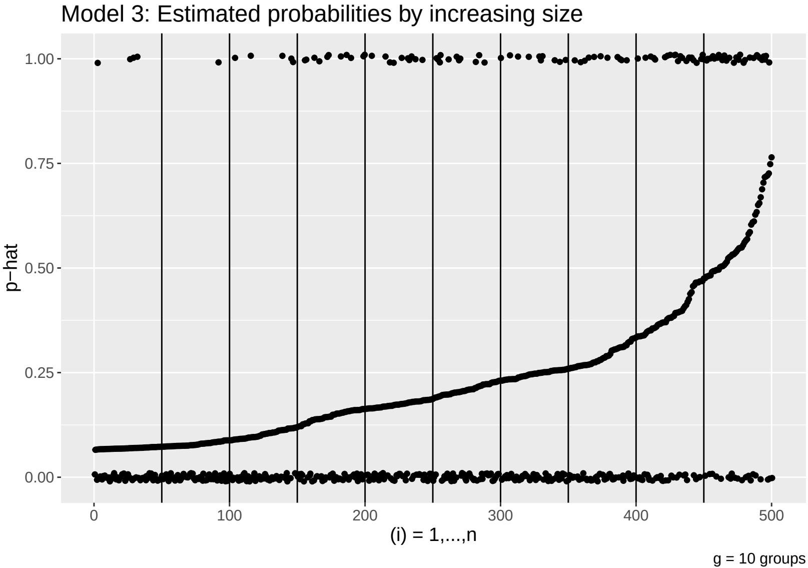
Goodness-of-fit test

- Hosmer-Lemeshow

- The sensitivity and specificity only look at the overall ability of the model to correctly predict the outcome.
- We want the model to be equally good at predicting the number of successes and failures, for all probabilities, large or small.

Hosmer-Lemeshow goodness-of-fit test

- Estimate the probabilities of success, $\hat{p}_i$, and sort them in increasing order, $\hat{p}_{(1)}, \ldots, \hat{p}_{(n)}$.
- Divide them into g groups with $n_g = n/g$ observations each: the n_g smallest, $\hat{p}_{(1)}, \ldots, \hat{p}_{(n_g)}$, in group 1, the n_g next smallest, $\hat{p}_{(n_g+1)}, \ldots, \hat{p}_{(2n_g)}$, in group 2, etc.
- We should choose $g > p + 1$ to allow enough flexibility to find discrepancies.
- At the same time we should have n_g large enough that the expected number of successes and failures are both at least 5, approximately, in all groups.


Model 3: Estimated probabilities by increasing size
p−hat
1.00
0.75
0.50
0.25
0.00
0
100
200
300
400
500
(i) = 1,...,n
g = 10 groups

Hosmer-Lemeshow cont'd.

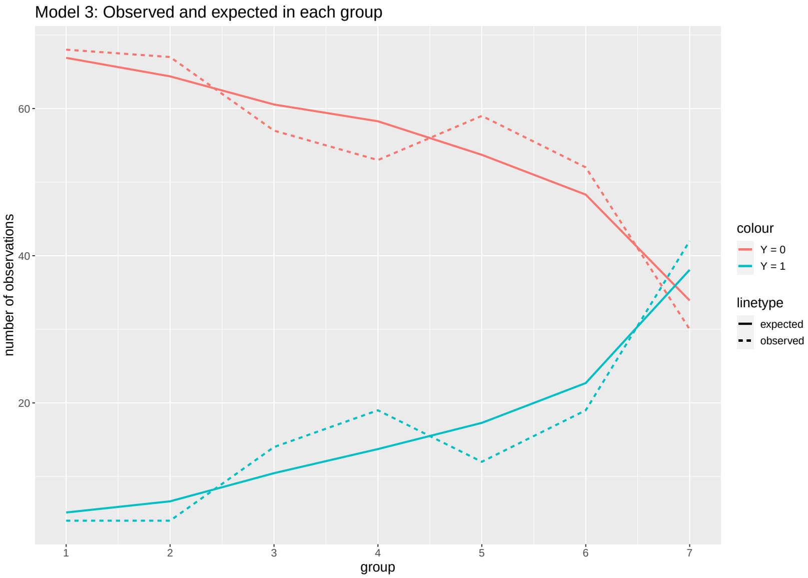
- The expected number of successes, E_{1k}, and failures, E_{0k}, in group k, for $k = 1, \ldots, g$, is then the sum of the probabilities

$$E_{1k} = \sum_{i=(k-1)n_g+1}^{kn_g} \hat{p}_{(i)}, \qquad E_{0k} = n_g - E_{1k}$$

- We want to compare them with the observed number of successes, O_{1k}, and failures, O_{0k}, in each group:

$$O_{1k} = \sum_{i=(k-1)n_g+1}^{kn_g} Y_{(i)}, \qquad O_{0k} = n_g - O_{1k}$$

  where $Y_{(i)}$ is the Y_i-value corresponding to $\hat{p}_{(i)}$.
- If the model is correct, the differences should be small.


Model 3: Observed and expected in each group
number of observations
60
40
20
1
2
3
4
5
6
7
group
colour
Y = 0
Y = 1
linetype
expected
observed

If these conditions are satisfied, and H_0: "the model gives correct probabilities" is correct, the weighted sum of squared differences, χ^2_{HL}, is approximately χ^2-distributed:

$$\chi^2_{\mathsf{HL}} = \sum_{j=0}^{1} \sum_{k=1}^{g} \frac{(O_{jk} - E_{jk})^2}{E_{jk}} \sim \chi^2(g-2)$$

and we should reject H_0 at significance level α if $\chi^2_{\mathsf{HL}} > \chi^2_{\alpha}(g-2)$.

- Rejecting H_0 means that the model is unable to predict the number of outcomes correctly.
- Not rejecting H_0 does not mean that the model is correct! Just that we could not prove that it is wrong.

Model 3: with $g = 7$ we get $\chi^2_{\mathsf{HL}} = 9.2 < \chi^2_{0.05}(5) = 11.1$ (P-value $= 0.10 > 0.05$) so we can not reject H_0.

Warning: different number of groups, g, can give different conclusions! Try several values for g and follow the majority conclusion.

(*) χ^2-test motivation

- The O_{jk} can be seen as random observations from dependent Binomial distributions.
- The Binomial distributions can, for large n and small p, with $np(1-p) \approx np$, be approximated by Poisson distributions: $O_{jk} \sim Po(E_{jk})$ with $E(O_{jk}) = V(O_{jk}) = E_{jk}$.
- If E_{jk} is large enough the Poisson distributions can be approximated by Normal distributions: $O_{jk} \sim N(E_{jk},\, E_{jk})$.
- Standardization gives $\dfrac{O_{jk} - E_{jk}}{\sqrt{E_{jk}}} \sim N(0,\, 1)$ and $\dfrac{(O_{jk} - E_{jk})^2}{E_{jk}} \sim \chi^2(1)$.
- The sum of these $2g$ dependent χ^2-variables is then also χ^2-distributed but we loose some of the degrees of freedom due to the dependance between them.