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

Lecture 9, spring 2023

Goodness-of-fit

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

3/5-23

Goodness of fit

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 might want a different threshold.

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

This can be done using the **Confusion matrix**:

	Predicted ($\hat{Y}_i$)		
True (Y_i)	Failure ($\hat{p}_i \leq 0,5$)	Success ($\hat{p}_i > 0.5$)	Total
Failure ($Y_i = 0$)	true negative	false positive	TN + FP
Success ($Y_i = 1$)	false negative	true positive	FN + TP
Total	TN + FN	FP + TP	n

- **Sensitivity** is the proportion of the true successes that have been correctly classified as successes (true positive rate or recall) $= Pr(\hat{Y}_i = 1 \mid Y_i = 1) = \frac{TP}{FN + TP}$.
- **Specificity** is the proportion of the true failures that have been correctly classified as failures (true negative rate $= 1 -$ false positive rate)
  $= Pr(\hat{Y}_i = 0 \mid Y_i = 0) = \frac{TN}{TN + FP}$.
- **Accuracy** is the overall proportion that have been correctly classified $= \frac{TP + TN}{n}$.
- **Precision** is the proportion of the predicted successes that are true successes $Pr(Y_i = 1 \mid \hat{Y}_i = 1) = \frac{TP}{FP + TP}$

All of these should be large.

Example: PM_{10}

- The largest model: `I(cars/1000)*windspeed + tempdiff`:

	Predicted			Correctly	
True	0	1	total	classified	
0	374	12	386	96.9 %	specificity
1	90	24	114	21.1 %	sensitivity
total	464	36	500		

Accuracy: $(374 + 24)/500 = 79.6\,\%$. Precision: $24/36 = 66.7\,\%$.

- The best (BIC) model: `I(cars/1000) + zerodiff`:

	Predicted			Correctly	
True	0	1	total	classified	
0	369	17	386	95.6 %	specificity
1	92	22	114	19.3 %	sensitivity
total	461	39	500		

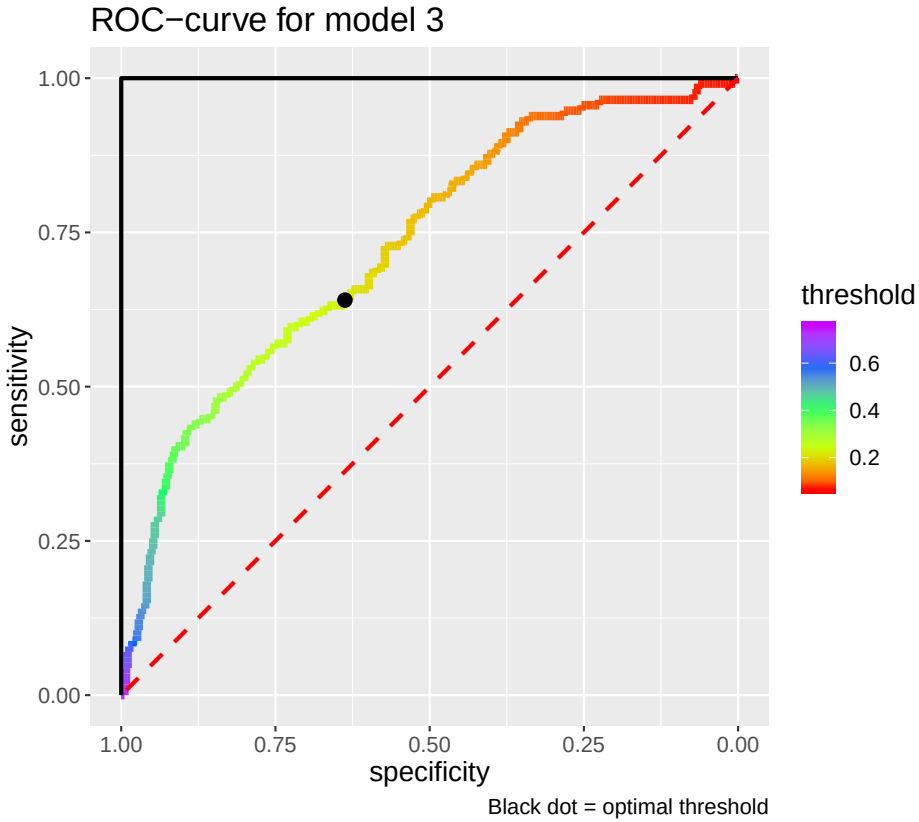
Accuracy: $(369 + 22)/500 = 78.2\,\%$. Precision: $22/39 = 56.4\,\%$

Warnings

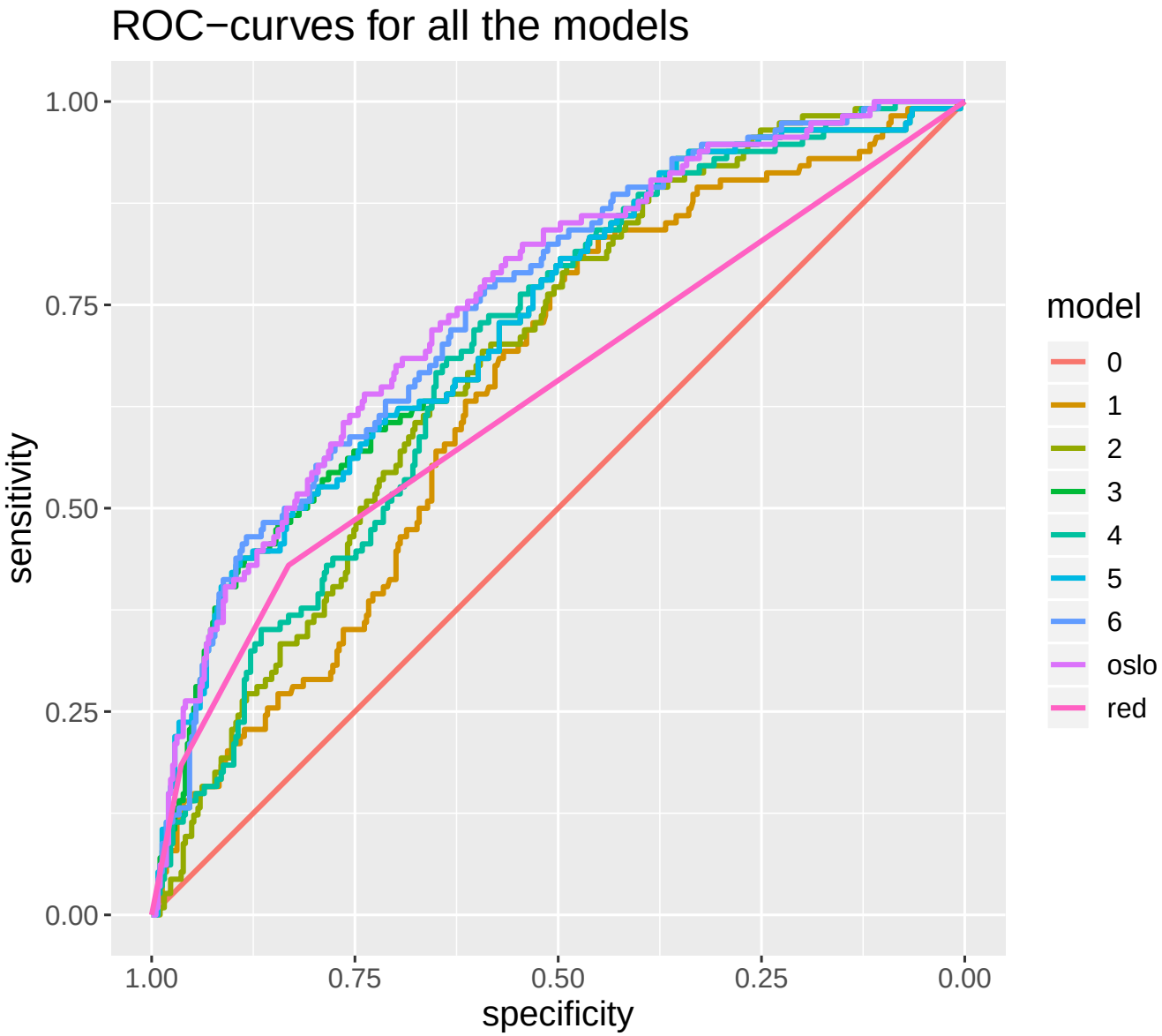
- There is no easy way to "punish" the addition of more x-variables.
- Larger models generally have higher values when predicting the same data that was used when estimating the model, due to over-fitting.
- If the main purpose of the model is to classify future observations, we should validate on a separate data set, not involved when fitting the model.
- On the other hand, if the model cannot even predict its own data, it is not a very good model.

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.
- Choose the threshold value that makes both specificity and sensitivity as large as possible, at the same time.
- 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! (or plot 1 - specificity)
- The solid black is the ideal ROC-curve.
  The dashed red is from the null model.
- Using the threshold $\hat{p}_i > 0.2215$ gives a sensitivity of 64.0 % and a specificity of 63.7 %.


ROC−curves for all the models
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
1.00
0.75
0.50
0.25
0.00
model
0
1
2
3
4
5
6
oslo
red

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.

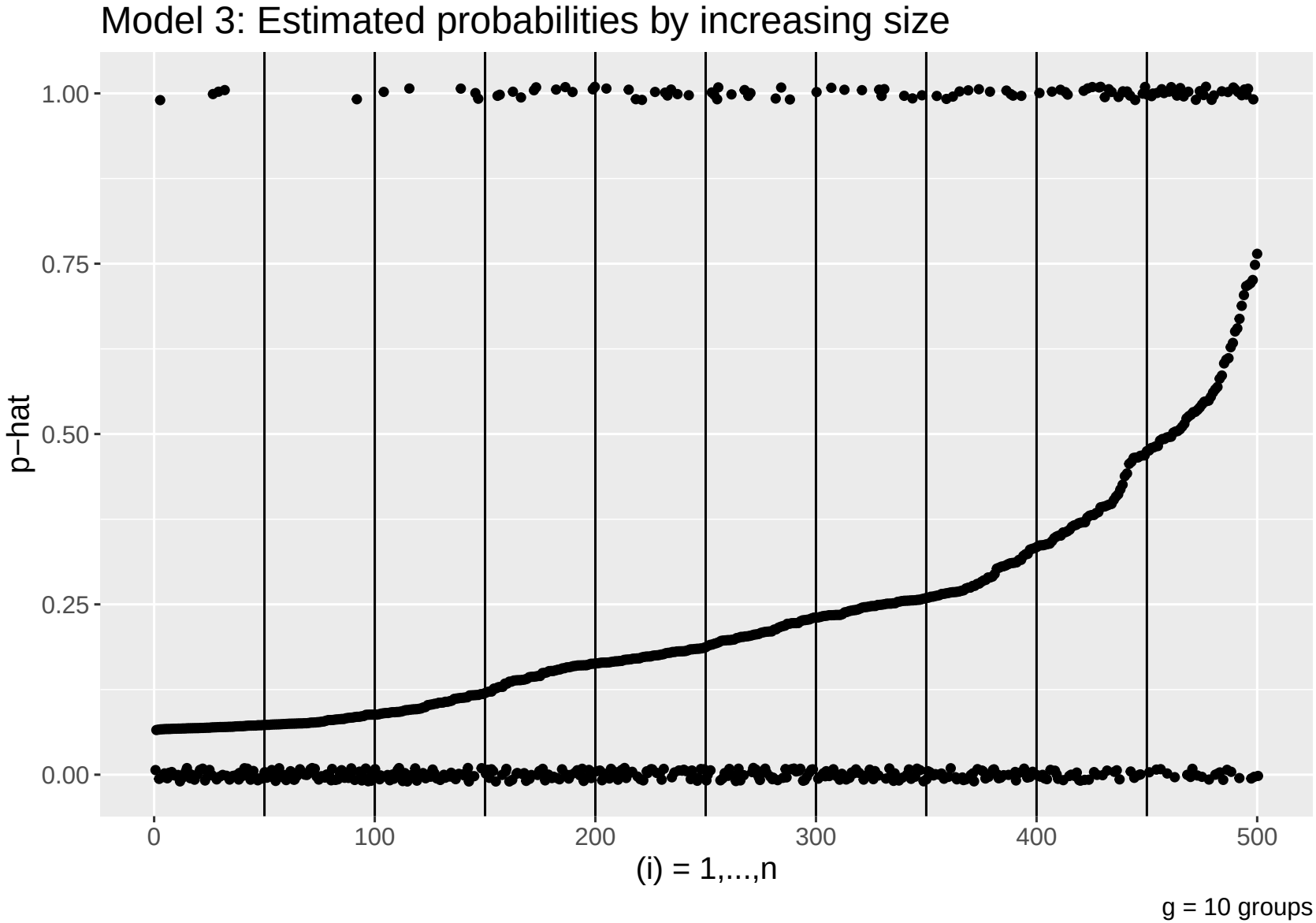
Model	AUC	95 % C.I.	P-value
0:null	50.0	(50.0, 50.0)	< 0.001
1:cars	64.8	(59,3, 70.3)	< 0.001
2:cars+wind	68.6	(63.5, 73.7)	0.002
red:tempdiff	63.8	(58.8, 68.9)	< 0.001
3:cars+zerodiff	72.8	(67.5, 78.1)	0.03
4:cars*wind	69.5	(64.3, 74,6)	0.003
5:cars+tempdiff	73.0	(67.7, 78.2)	0.046
6:cars*wind + zerodiff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

- 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
1.00
0.75
0.50
0.25
0.00
p−hat
0
100
200
300
400
500
(i) = 1,...,n
g = 10 groups

Hosmer-Lemeshow cont'd.

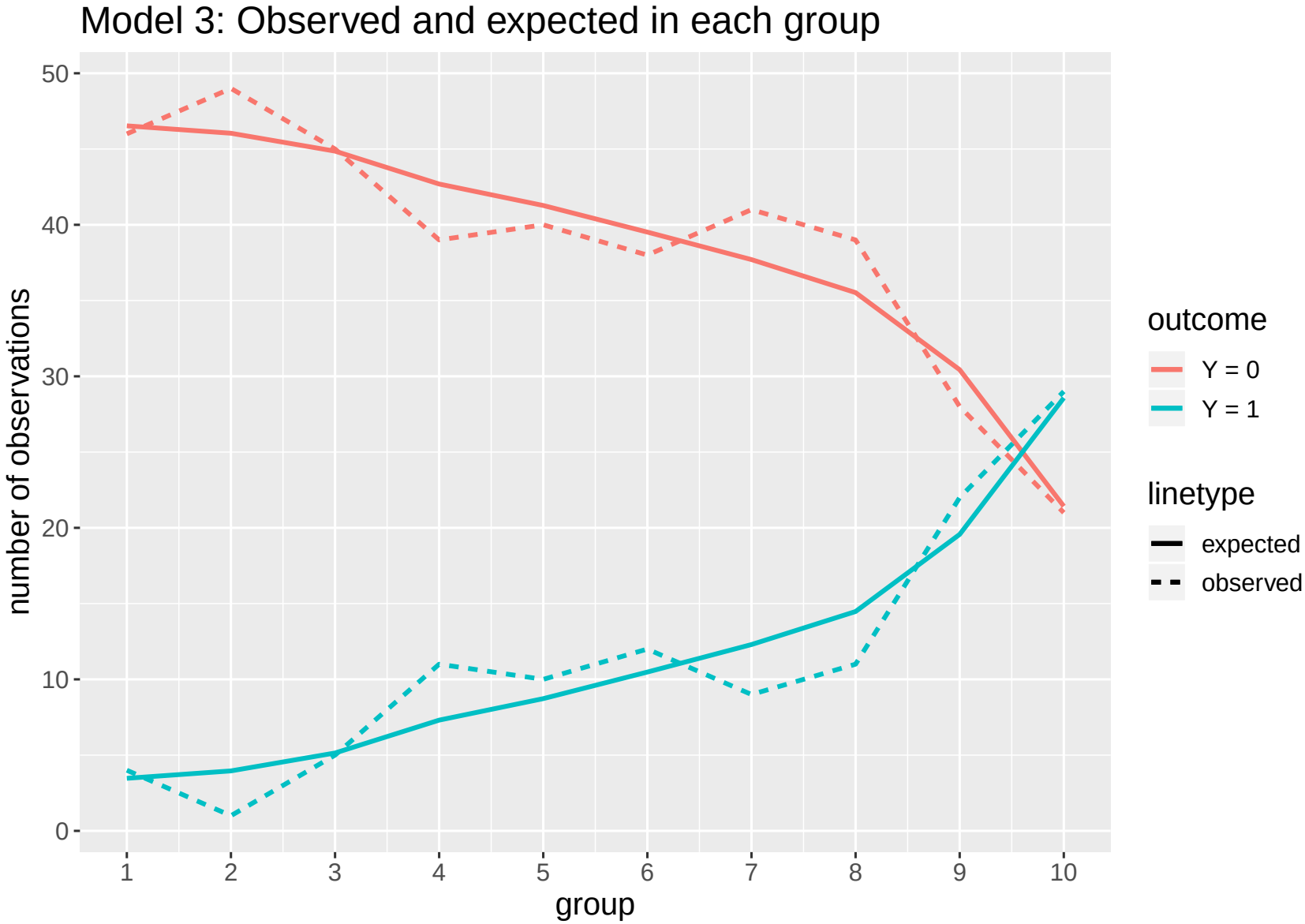
- The expected number of successes, E_{1k}, and failures, E_{0k}, in group k, for $k = 1, \ldots, g$, is then the sum of the probabilities

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

- 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
group
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
outcome
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:

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

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

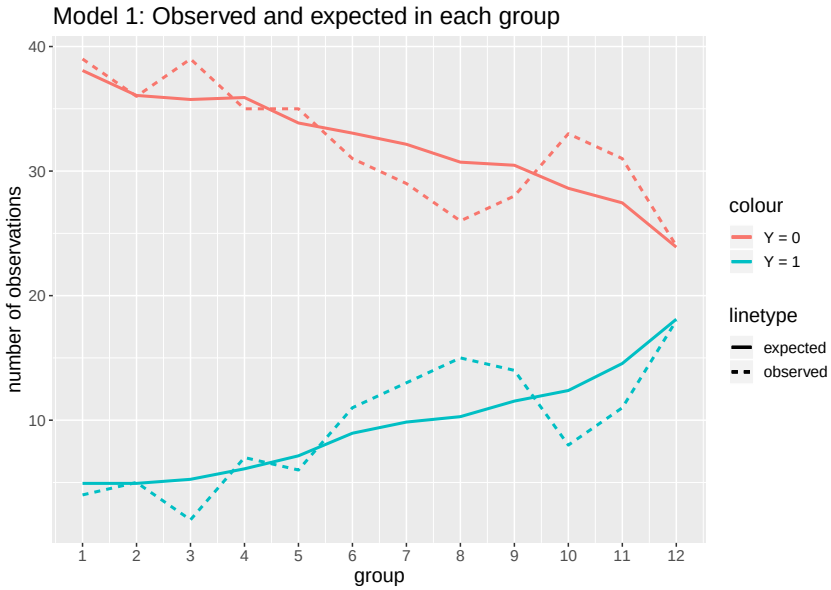
- Rejecting H_0 means that the model is unable to predict the number of outcomes correctly.
- Not rejecting H_0 does not mean that the model is correct! Just that we could not prove that it is wrong.

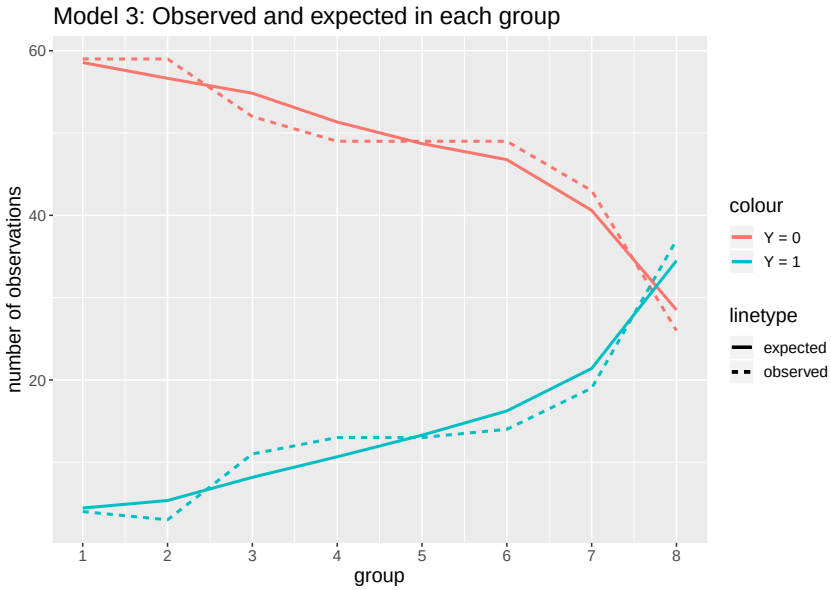
Model 3: with $g = 10$ we get $\chi^2_{\text{HL}} = 8.04 < \chi^2_{0.05}(8) = 15.5$ so we can not reject H_0.

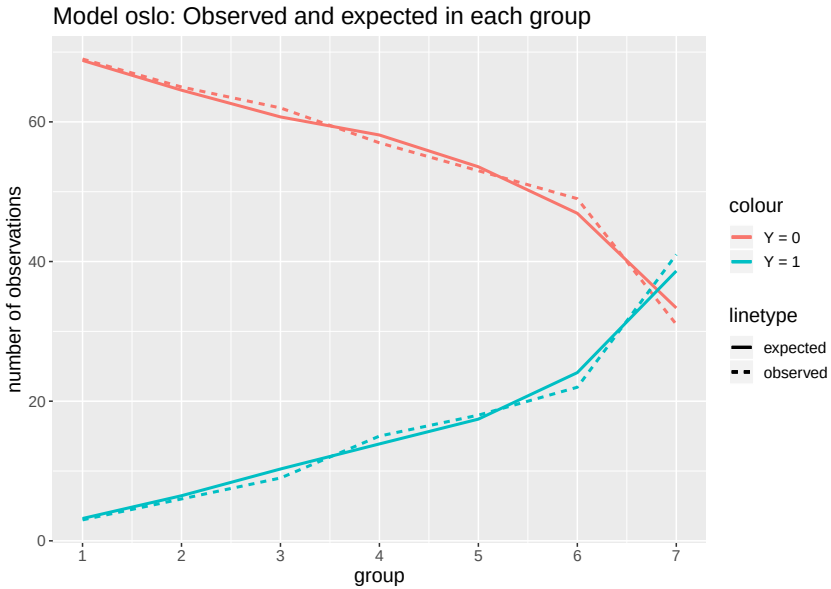
Warning: different number of groups, g, can give different conclusions! Try several values for g and follow the majority conclusion.

(*) 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 $\frac{O_{jk} - E_{jk}}{\sqrt{E_{jk}}} \sim N(0, 1)$ and $\frac{(O_{jk} - E_{jk})^2}{E_{jk}} \sim \chi^2(1)$.
- 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.







Model	p-value
1:cars	0.29
3:cars+zerodiff	0.66
oslo:cars*wind+tempdiff	0.97

The largest model gives the best predictions, but they are all OK.