Course 42588 – week 11

Data analytics – classification models

Today's program

- Logistic regression
  - Motivation and the model
  - Example I and predictions
  - Interpretation and effects - odds ratios, relative risk and elasticities
  - Maximum likelihood estimation
  - The LR test and example II
  - Overall fit and ROC curves
  - Comparison with other models

- Work on Project 3

The course plan

Week	Date	Subject/Lecture	Literature	Exercises	Teachers
1	31/1	Introduction + questions and data	AoS chap. 3	Form groups + week 1 exercise	Stefan
2	7/2	Basics on data and variables	AoS chap. 1-2 (+ OM 1)	Project 1 – start	Stefan/Guest from Genmab
3	14/2	Surveys + data types + experimental data	Paper 1 (+ OM 2-5)	Project 1 – work	Sonja / Stefan
4	21/2	Governance + causality	Paper 2 + AoS chap. 4 (+ OM 6)	Project 1 – deadline	Hjalmar / Stefan
5	28/2	More on data, e.g. real-time data, online data	Paper 3 (+ OM 7-10)	Discuss data for project 2	Guido/ Stefan
6	6/3	Visualisation	Chap. 1,5,6,7,10,23, 24,29 in Wilke + (AM 1-2)	Integrated exercises + work on project 2	Mads
7	13/3	Spatial data	Chap. 1,14 in Gimonds	Week 7 exercises + work on project 2	Mads / Guest from Niras
8	20/3	Imputation/weighting/presentation proj. 2	Paper 4	First deadline of project 2 + Week 8 exercises	Mads
9	3/4	Data analytics I	ISL ch. 3 + paper 5	Work on project 3	Stefan
10	10/4	Data analytics II	ISL ch. 6-6.1 + 7-7.4	Work on project 3	Stefan
11	17/4	Data analytics III	ISL ch. 4-4.3	Work on project 3	Stefan
12	24/4	Data analytics IV	Train 2.1-2.3 + 3.1, 3.6, 3.8-9	Work on project 3	Stefan
13	1/5	Summary and perspective	Paper 6	Project 3 – deadline	Stefan

Logistic regression - motivation

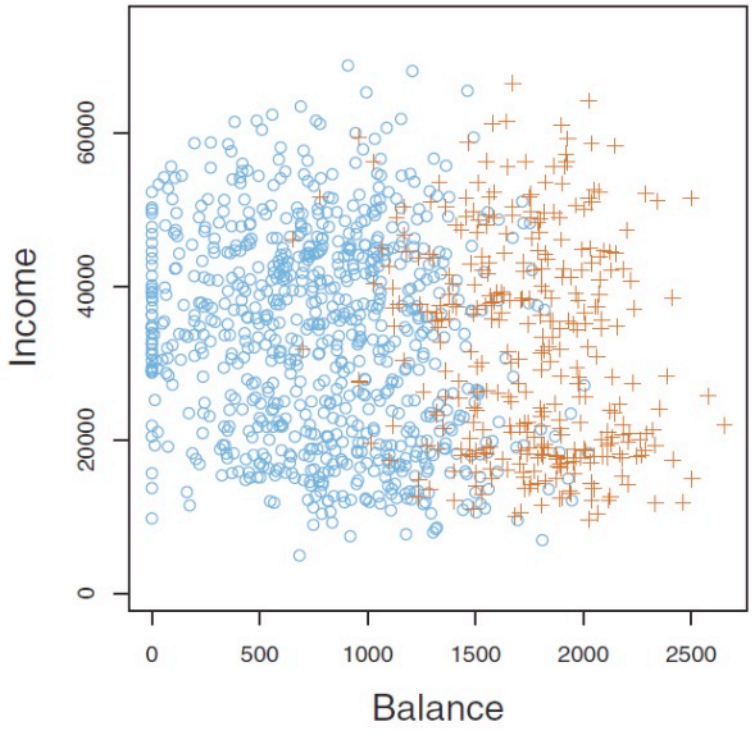
- Logistic regression is used to model binary random variables, i.e. variables with two possible outcomes.

- Examples (discuss a 1 minute with your neighbour)
  - a person's choice between car and public transport for commuting.
  - a household's decision between having or not having a car.
  - whether an individual has been involved in a car accident or not.

Logistic regression - motivation

- Another example is from the book chapter 4 looking at factors that influence whether individuals default on their credit cards, i.e. Default = {0,1} is a binary qualitative variable.

- Overall 3% default so the plot only shows a fraction of the individuals who did not default.



Logistic regression – the model

- We have something in the real world that can be represented by a random variable, Y, that has two outcomes, i.e. they can be represented by 0 and 1.

- In a logistic regression (model), we model

$$P(Y = 0) \text{ and } P(Y = 1)$$

- But since $P(Y = 1) = 1 - P(Y = 0)$, we only need to model one of them.

- A restriction for a model of $P(Y = 1)$ is that it should only give results between 0 and 1, and at the same time, it should be able to give any number between 0 and 1.

Logistic regression – the model

- If we use the notation $\pi(x) = P(Y = 1|X = x) = P(Y = 1|x)$ then the logistic regression is defined by

$$\pi(x) = \frac{\exp(\beta' x)}{1+\exp(\beta' x)} \text{ or } \ln\left(\frac{\pi}{1-\pi}\right) = \beta' x \qquad (1)$$

  where β are the parametres that we would like to estimate while the $x's$ are the explanatory variables.

- The function $\ln\left(\frac{x}{1-x}\right)$ is known as the logit transformation. Logistic regression is said to use the logit as link function, i.e. the link function is used to get from $\pi(x)$ to $\beta' x$, and for logistic regression the link function is equal to the logit transformation.

Logistic regression – the model

- It is important to remember the expression

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta' x \qquad (1)$$

  where β are the parametres that we would like to estimate while the $x's$ are the explanatory variables.

- The right hand side is similar to a linear regression and we can use the same explanatory variables:
  - Quantitative continuous variables, e.g. income, travel time
  - Discrete quantitative variables, e.g. age groups
  - Qualitative variables, e.g. gender or education, but these have to be recoded as dummies, if they have more than two levels

- NB. When you interpret the parametres it is important to remember that their effect on probabilities are non-linear because of the logit transformation in (1).

Question – dependent variables

- What types of variables can enter as a dependent variable in a linear and logistic regression?

A. $Y\ continuous\ (lin)\ and\ Y\ binary\ (log)$
B. $Y\ continuous\ (lin)\ and\ Y\ continuous\ (log)$
C. $Y\ binary\ (lin)\ and\ Y\ continuous\ (log)$
D. $Y\ all\ types\ of\ variables\ in\ both\ models$

Answer: A

Logistic regression – example I

- The model is $\ln\left(\frac{\pi}{1-\pi}\right) = \beta' x$, where $\pi = P(y = 1)$, and where β are the parametres that we would like to estimate and x are the explanatory variables.

- Accident data: We have a data set with 3643 observations of which 723 have been involved in an accident and the remaining 2920 have not. In the data, we have the variables: gender, age group and blood alcohol concentration (BAC). What types are the variables? Nominally, ordinally, interval, or ratio scaled?

- To investigate how alcohol affects the risk of accident, we apply the model:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_{age24} x_{age18-24} + \beta_{age34} x_{age25-34} + \beta_{age49} x_{age35-49} + \beta_{fem} x_{fem} + \beta_{bac} x_{bac}$$

where x_{fem} is a dummy for female, $x_{age18-24}$, $x_{age25-34}$, $x_{age35-49}$ are dummies for age groups and x_{bac} is the alcohol percentage.

Reference codning of a discrete variable

- The variable age group has four levels:
  - 18-24
  - 25-34
  - 35-49
  - 50+

- Since the groups do not cover similar age groups, it does not make sense to use a single parameter to find the effects of the age groups.

- Instead of a linear effect, we can choose a reference group, e.g. 50+, and then introduce dummies for the remaining groups, i.e. three dummy variables
  - $x_{age18-24}$, the variable is 1 for the 18-24 year old and 0 for the rest
  - $x_{age25-34}$, the variable is 1 for the 25-34 year old and 0 for the rest
  - $x_{age35-49}$, the variable is 1 for the 35-49 year old and 0 for the rest

Logistic regression – example I

- The model is $\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_{age24}x_{age18-24} + \beta_{age34}x_{age25-34} + \beta_{age49}x_{age35-49} + \beta_{fem}x_{fem} + \beta_{bac}x_{bac}$

- When both discrete variables are reference coded, the result is:

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.4793	0.1055	551.8127	<.0001
bac		1	2.9943	0.3027	97.8260	<.0001
agegroup	18-24	1	2.1685	0.1357	255.3857	<.0001
agegroup	25-34	1	1.0536	0.1381	58.2232	<.0001
agegroup	35-49	1	0.4192	0.1305	10.3273	0.0013
gender1	F	1	0.3435	0.0963	12.7347	0.0004

- How should we interpret this (discuss with your neighbour for a minute)?
- We see that all parametres are significant.
  - The probability for accidents decreases with age everything else equal
  - The probability rises with the alcohol percentage everything else equal
  - The probability is higher for females everything else equal

Logistic regression – the model

- Predictions using the model can be done using the expression

$$P(y = 1|x) = \pi(x) = \frac{\exp(\beta' x)}{1+\exp(\beta' x)},$$

  which is only a function of the x's once the β's are estimated.

- We can therefore calculate $P(y = 1|x)$ for the values in the data.

- If we believe that a variable changes, e.g. if income is increased by 10%, then we can calculate the probabilities for a data set where the income is 10% higher.

- To compare the effect of the change, we can compare the average value of $P(y = 1|x)$ in the sample before and after the change.

Logistic regression – example I

- The estimated model is:

$$\ln\left(\frac{\pi}{1-\pi}\right) = -2.48 + 2.17 * x_{age18-24} + 1.05 * x_{age25-34} + 0.42 * x_{age35-49} + 0.34 * x_{fem} + 2.99 * x_{bac}$$

- To calculate $\mathrm{E(Y|x)} = P(Y = 1|x)$, we use the corresponding formula:

$$\pi = \frac{\exp(\beta' x)}{1 + \exp(\beta' x)}$$

- For a sober 22 year old male the probability becomes:

$$P(Y = 1|x) = \frac{\exp(-2{,}48 + 2{,}17 * 1)}{1 + \exp(-2{,}48 + 2{,}17 * 1)} = 0{,}42$$

- What is the probability for a 22 year old male with x = 0.52

How to judge parameters (variable influence) I

- Looking at our example we have

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_{age24}x_{age18-24} + \beta_{age34}x_{age25-34} + \beta_{age49}x_{age35-49} + \beta_{fem}x_{fem} + \beta_{bac}x_{bac}$$

- There are various ways to judge the influence of variables
  - Direct interpretation of the β's is NOT a good idea
  - You could consider marginal effects. For the above logistic regression, we have

$$P(y = 1|x) = \pi(x) = \frac{\exp(\beta' x)}{1+\exp(\beta' x)} \text{ and } \frac{\partial \pi}{\partial x_k} = (1-\pi)\pi\beta_k$$

  - An alternative is of course elasticities

$$E = \frac{\partial \pi}{\partial x_k}\frac{x_k}{\pi} = (1-\pi)x_k\beta_k$$

How to judge parameters (variable influence) II

- Another common way is to look at odds ratios. Suppose we calculate the log odds for a male and female, respectively, with the same remaining values, and take the difference.

$$\ln\left(\frac{\pi_f}{1-\pi_f}\right) - \ln\left(\frac{\pi_m}{1-\pi_m}\right) = \beta_{fem} \text{ i.e. odds ratio is equal to } \frac{\frac{\pi_f}{1-\pi_f}}{\frac{\pi_m}{1-\pi_m}} = \exp(\beta_{fem})$$

- Note that you should only use odds ratio (OR) if you remember that they are ratios of the odds and NOT relative risks (RR), i.e. ratios of the risks.

- Suppose an example where observations are either treatment or control. In the control group there is a 0.50 risk of death while in the treatment group there is a 0.25 risk of death.

- Here the relative risk of dying with treatment versus without is 0.25/0.50 = 0.50. On the other hand the odds ratio can be found to be 0.33.

Question – parameter interpretation

- Suppose a group of patients get a treatment (X=1) and another do not (X=0). They either die (Y=1) or get cured (Y=0). Suppose that if P(Y=1|X=0)=0.4 and P(Y=1|X=1) = 0.2. What can we conclude about the odds ratio (OR)= $(\frac{\pi_1}{1-\pi_1}/\frac{\pi_0}{1-\pi_0})$ and the relative risk (RR) when $\pi_i = P(Y = 1|X = i)$?

A. *The RR is 0.5 and the OR is 0.25*

B. *The RR is 0.25 and the OR is 0.5*

C. *The RR is* 0.25 *and the OR is* 0.375

D. *The RR is* 0.5 *and the OR is* 0.375

Answer: D

Break

Question – estimation

- What function is minimised in the least squares method that we use to estimate linear regression $y = \beta_0 + \beta_1 x + u$?

A. $min \sum_{i=1}^{N} (Y_i - E(Y|X_i))$

B. $min \sum_{i=1}^{N} |Y_i - E(Y|X_i)|$

C. $min \sum_{i=1}^{N} (Y_i - E(Y|X_i))^2$

D. $min \sum_{i=1}^{N} (Y_i - E(Y|X_i))^3$

Answer: C

Maximum likelihood estimation

- Examples of estimators
  - The average is an estimator of the mean
  - Ordinary least squares (OLS) for linear regression

- Maximum likelihood estimation is a general method that can be used if you know the probability of all outcomes, i.e. the probability function $P(Y = y|x)$, e.g. $\mathrm{P(Y = 1|x)} = \frac{\exp(\beta' x)}{1+\exp(\beta' x)}$ for logistic regression.

- For a specific data set, we can perceive the function $f(y|\beta, x)$ as a function of β. For an observation in a logistic regression, we have

$$f(y_n|\beta, x_n) = \begin{cases} \mathrm{P}(y_n = 0|\beta, x_n), & if\ y_n = 0 \\ \mathrm{P}(y_n = 1|\beta, x_n), & if\ y_n = 1 \end{cases}$$

- We denote $\mathrm{L}(\beta|y, x) = \prod_n f(y_n|\beta, x_n)$ as the likelihood function.

Maximum likelihood estimation

- Normally, we work with the natural logarithm of the likelihood function, i.e. the loglikelihood function (LL):

$$\mathrm{LL} = \ln\big(\mathrm{L}(\beta|x)\big).$$

- Maximum likelihood estimation (MLE) uses

$$\hat{\beta} = argmax_{\beta}\mathrm{LL}(\beta)$$

  where

$$\mathrm{LL}(\beta) = \sum_{n} \ln(P(y_n|x_n,\beta))$$

- MLE is consistent (but biased) and asymptotically normal.

- Take a minute to discuss with your neighbour what consistent and asymptotically normal means.

Logistic regression - estimation

- In logistic regression, we have

$$\mathrm{P}(\mathrm{y}=1)=\frac{\exp(\beta' x)}{1+\exp(\beta' x)},$$

and

$$\mathrm{P}(\mathrm{y}=0)=1-\mathrm{P}(\mathrm{y}=1)$$

where β are the parametres that we would like to estimate, and the x's are the explanatory variables.

- Because we have these expressions for the probabilities, we can use maximum likelihood estimation. Note that for a linear-in-parameters logistic regression, the LL is globally concave. Why is this important?

- Two important expressions related to MLE are:
  - $\mathrm{LL}(0)$, which is the value of the LL function when all parametres are 0.
  - $\mathrm{LL}(\hat{\beta})$, which is the maximum value of the LL function.

Logistic regression - estimation

- Using the maximum (log)likelihood value, $\mathrm{LL}(\hat{\beta})$, we can compare two models estimated on the same data:

$$LR = -2 * \left(LL(\hat{\beta}_1) - LL(\hat{\beta}_2)\right) \sim \chi_f^2$$

- This test is known as a likelihood ratio (LR) test. The test size is χ_f^2 distributed with f degrees of freedom, where f is the difference in the number of parametres between the two models.

- The test is only valid if one of the models is a special case of the other. This is the case if you have a model and assume some of the parametres equal to zero or some of the parametres equal to each other, which gives you a smaller model. We always have f>0.

- You may end up by mistake switching the values from model 1 and 2 in the formula. The result should always be a positive number so if you get a negative you have somehow switched the order of the models.

Exercise 2 on the week info page (10 min)

- Open the data accidents_logistic. Estimate the model from example 1, i.e. a logistic regression that explains the probability for a car accident using the explanatory variables: age, gender and blood alcohol content.

- Also estimate the model without the female dummy and check that the LR test statistic between this model and the full model corresponds to the Wald Chi Square statistic for the female dummy in the full model.

- Calculate the difference in the average probability of a car accident if all individuals in the sample have a BAC of 0 compared to 0.5.

Logistic regression – example II

- Mode choice: We have a data set with 430 individuals, who had the choice between train (205) and car (225) as mode. We specify a model for $\pi = P(bil) = P(Y = 0)$:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_{t\_diff}x_{time} + \beta_{c\_diff}x_{price} + \beta_{w\_diff}x_{walk} + \beta_{tr\_diff}x_{transf}$$

where x_{time} is travel time car – travel time train (min),

x_{walk} is the difference in walk time for the mode (min),

x_{price} is price car – price train (Euro), and

x_{transf} is the difference in transfers

Logistic regression – example II

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.3288	0.3722	0.7804	0.3770
t_diff	1	-0.0621	0.0177	12.2940	0.0005
w_diff	1	-0.0214	0.0141	2.3029	0.1291
c_diff	1	-0.9440	0.4784	3.8939	0.0485

- We see that time is very significant $p<0.01$. Price is also significant $p<0.05$, while walk time is not significant $p=0.13$.
- The signs on all parametres are negative as expected.
- -2 times the loglikelihood is 578.2.

Logistic regression – example II

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.1956	0.3806	0.2643	0.6072
t_diff		1	-0.0662	0.0180	13.5565	0.0002
w_diff		1	-0.0316	0.0152	4.3271	0.0375
c_diff		1	-1.0711	0.4851	4.8747	0.0273
tr_diff	0	1	-0.2045	0.1090	3.5228	0.0605

- We see that time is very significant $p<0.01$. Walk and price are also significant $p<0.05$, while number of changes is not significant $p=0.06$.
- The signs on all parametres are negative as expected.
- -2 times the loglikelihood is 574.6.

Logistic regression – LR test example

- If we calculate

$$LR = -2 * \left(LL(\hat{\beta}_1) - LL(\hat{\beta}_2)\right) = -2 * LL(\hat{\beta}_1) + 2 * LL(\hat{\beta}_2) = 578.2 - 574.6 = 3.6$$

- The LR test of these two models follow a χ_1^2 distribution with 1 degree of freedom. So we can look up in a table that $\chi_1^2(3.84) = 0.95$, i.e. 95 % of the probability mass is to the left of 3.84. Since $3.6 < 3.84$, we cannot reject the hypothesis that the simple model explains data as well as the larger model, but it is a borderline call as we are very close to the critical value

- When you are close as here, you should find other arguments than the LR test to decide on which model to prefer.

Break

Question – parameter interpretation

- How can we interpret β_{fem} in a logistic regression of the probability of an accident:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_{fem} x_{fem},$$

where $\beta_{fem} = -0.2$?

A. *If you are a woman then the prob. is 0.2 higher*

B. *If you are a woman then the prob. is 0.2 lower*

C. *If you are a woman then the prob is lower*

D. *If you are a woman then the prob. is higher*

Answer: C

Overall model evaluation

- For linear regression models, we use (adjusted) R^2 as a measure for how much of the variation in data that can be explained by the model.

- There are similar generalised measures for the logistic regression.

- One of these expression for the models overall explanation of the data is (McFaddens) ρ^2:

$$\rho^2 = 1 - \frac{LL(\hat{\beta})}{LL(0)} \text{ and } \bar{\rho}^2 = 1 - \frac{LL(\hat{\beta}) - K}{LL(0)},$$

  where K is the number of parametres in the model. This measure is similar to AIC and BIC that you can also calculate for logistic regression estimated by MLE.

- Other methods exist based on the confusion matrix, e.g. accuracy or F_1-scores, the area under the curve known as the ROC curve.

Logistic regression – ROC curves

- We can place all observations in one of the four boxes if we use a threshold, T. Denote $\mathrm{P}(y_i = 1) > T$ as positive and $\mathrm{P}(y_i = 1) < T$ as negative:

True negative, i.e. **P(Y=1)<T and Y=0** **(Specificity)**	**False negative, i.e.** **P(Y=1)<T and Y=1** **(1-Sensitivity)**
False positive, i.e. **P(Y=1) > T and Y=0** **(1-Specificity)**	**True positive, i.e.** **P(Y=1) > T and Y=1** **(Sensitivity)**

- The share of true negative (Y=0) observations is denoted as specificity.
- The share of true positive (Y=1) observations is denoted as sensitivity.

Example – default classification

- Using a threshold of 0.5, i.e. P(default|X=x)=P(Y=1|X=x)>0.1, we get the following confusion matrix for the book example

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

- So only 23 out of 9667 non-defaulters were incorrectly labelled. However, 252/333= 75.7% defaulters were incorrectly labelled, i.e. the specificity is high (1-23/9667)=0.998 while the sensitivity is low (1-252/333)=0.243. The total error rate is 0.0275 (training rate).

- If we use the threshold 0.2 instead, we get
  - Specificity = 1-235/9667 = 0.976
  - Sensitivity = 1-138/333 = 0.586

		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

- Total error rate increases slightly to 0.0373.

Logistic regression – ROC kurver

- An ROC curve is drawn by plotting sensitivity against 1-specificity.

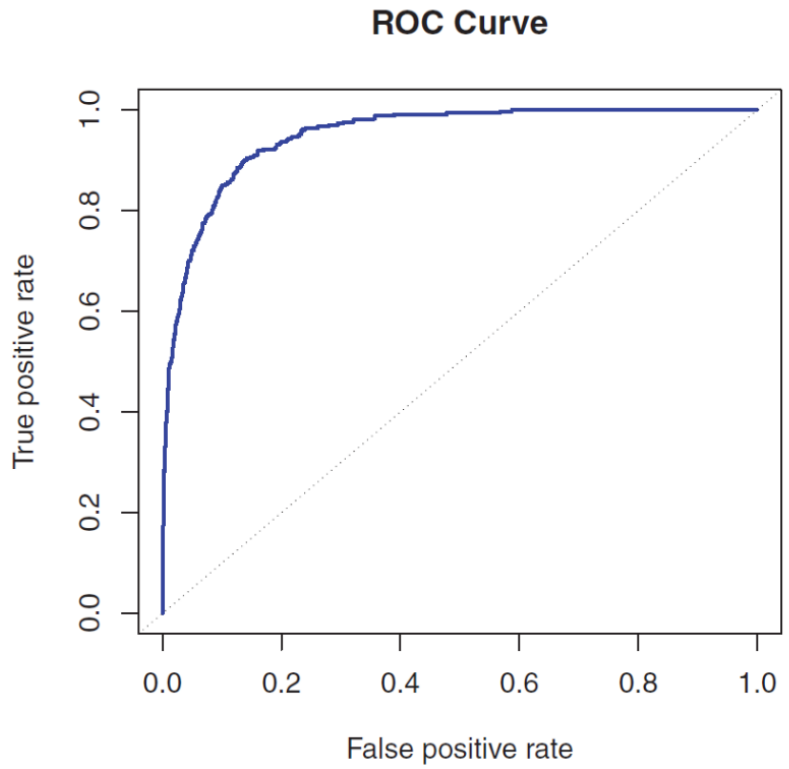
- In the limit $T = 1$ then 1-specificity = 0 because all (Y=0) observations become true negative. Likewise all (Y=1) observations become false negative, so sensitivity is also equal to 0.

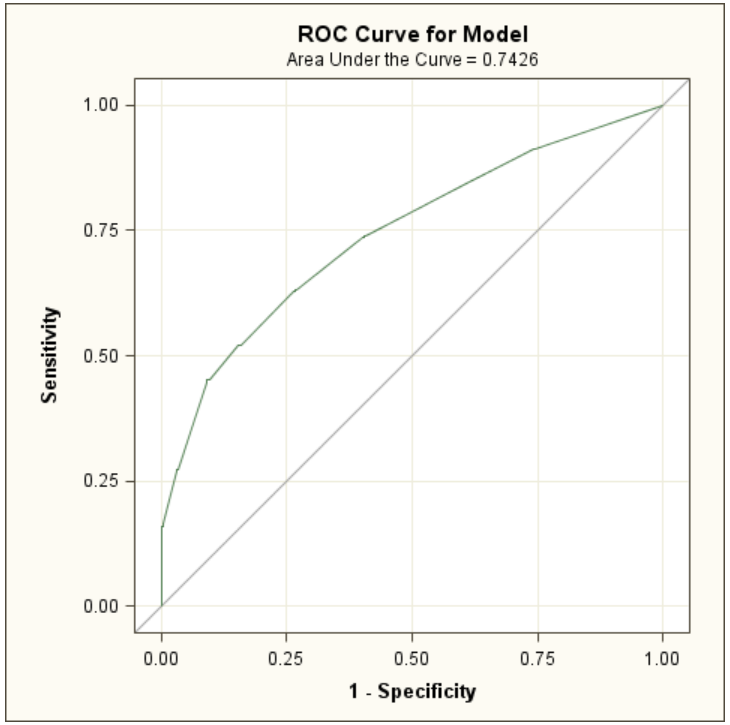
Similarly for the limit $T = 0$, we get the point (1,1).

- The model is evaluated by summing the area under the curve.

Two examples of ROC curves

- In the Default example to the left the ROC area looks high while for the accident example on the right, it is 0.74.





How to predict an outcome?

- In the above confusion matrices, we used the standard way to predict a likely outcome for each observation, i.e. for a given threshold T

$$\hat{y}_n = \begin{cases} 0 & ,if\ \mathrm{P}(y_n = 1|x_n) < T \\ 1 & ,if\ \mathrm{P}(y_n = 1|x_n) > T \end{cases}$$

- This is known as the Bayes classifier and can be shown to give the lowest total error rate if the P estimates are correct.

- However as the example showed, it might not be the most useful approach that gives the lowest total error rate. Another approach is to make predictions that follow the distribution captured by the model, i.e. draw a random number $s \sim U[0,1]$ and calculate

$$\hat{y}_n = \begin{cases} 0 & ,if\ \mathrm{P}(y_n = 1|x_n) < s \\ 1 & ,if\ \mathrm{P}(y_n = 1|x_n) > s \end{cases}$$

Question – binary variable

- What is the average for a binary variable that can take on the values 0 and 1?

A. $P(y = 0)$

B. $P(y = 1)$

C. $P(y = 0) + P(y = 1)$

D. $0.5 * (P(y = 0) + P(y = 1))$

Answer: B

Logistic regression – gen. lin. models

- We use logistic regression if we need to model a binary variable, Y. A logistic regression is a model of

$$P(Y = 1)$$

- For a binary random variable, we have the result

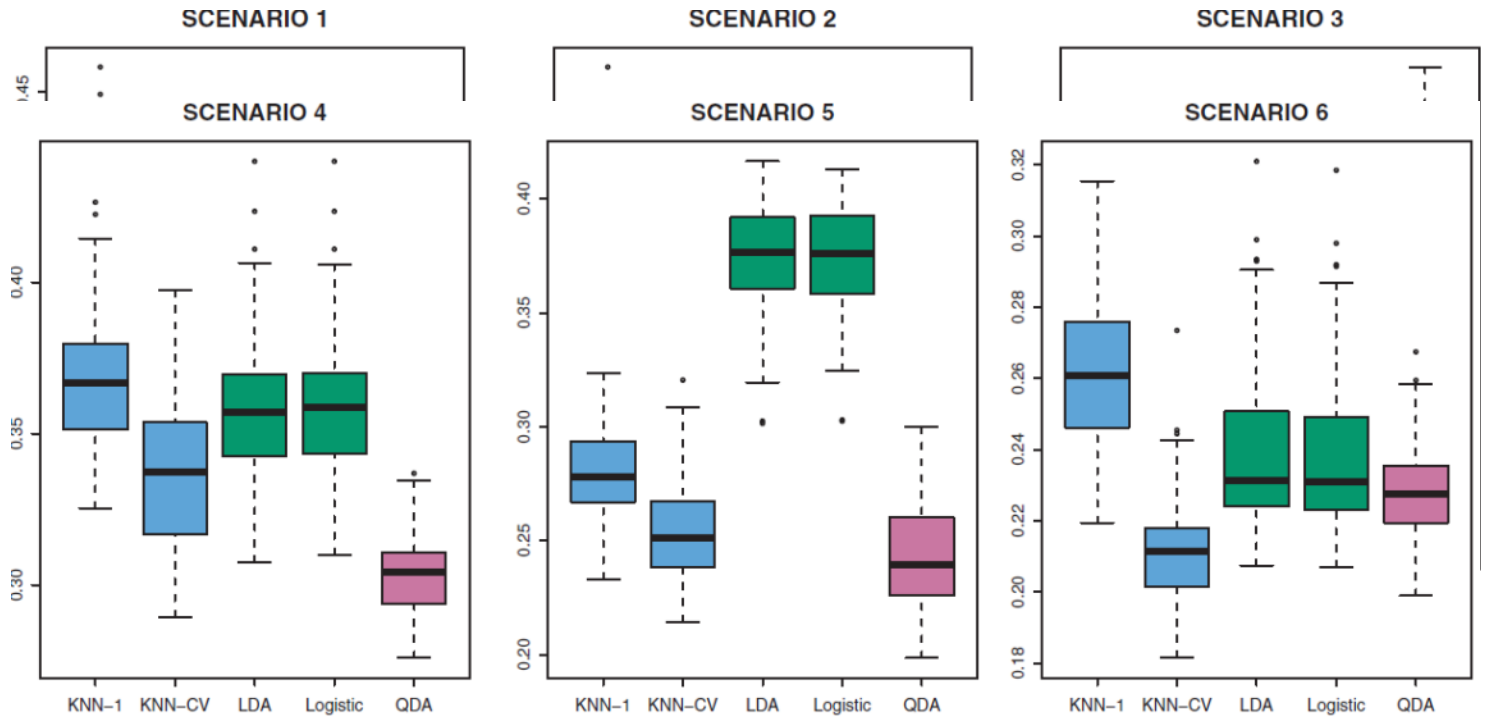
$$\mathrm{E}(\mathrm{Y}) = 1 * P(Y = 1) + 0 * P(Y = 0) = P(Y = 1).$$

- So when we model $P(Y = 1)$, we actually model $\mathrm{E}(\mathrm{Y})$, hence $\mathrm{Y} = \mathrm{P}(\mathrm{Y} = 1) + \varepsilon$.

- This makes logistic regression similar to linear regression where we also model the expectation. The difference is that Y is Bernouilli distributed, that $\mathrm{E}(Y)$ only can take on values between 0 and 1, and that $\mathrm{E}(\mathrm{Y})$ and the residual are correlated.

Comparison to other methods (4.5 in book)

- Even though we have not covered all methods, it is interesting to see how the methods compare on various test data sets. Below are shown test error rates for five models across data generated



Project 3

- Most groups have started working on Project 3 - part 1, i.e. the two exercises on linear regression. Part 2 on will be uploaded today.

- All groups should continue the work on Project 3 today.

Feedback

- Final questions

    1. What was the most interesting you learned during the lecture?

    2. What is your most important unanswered question based on the lecture?

- Group 7 (Benedicte, Mads, Rasmus, Rasmus, Benjamin) should send their feedback to Stefan. Everyone else are very welcome to give feedback as well!

For next time

- Read for this week
  - Introduction to statistical learning chap 4-4.3 (2.2.3 is recap)
  - Confusion matrices and ROC curves, pp. 148-152
  - (Empirical comparison of methods, pp. 161-164)

- To prepare for lecture 11, you should read
  - Train chap. 2.1-2.3 + 3.1, 3.6, 3.8-9

- Course evaluations have opened. Remember to evaluate all your course.

- Today is the deadline for exam date wishes 28th vs. 29th!

- Work on Project 3 (deadline 7/5).