Logistic Regression: the model, estimation, interpretation and diagnostics

QFRM Course

Dr Svetlana Borovkova

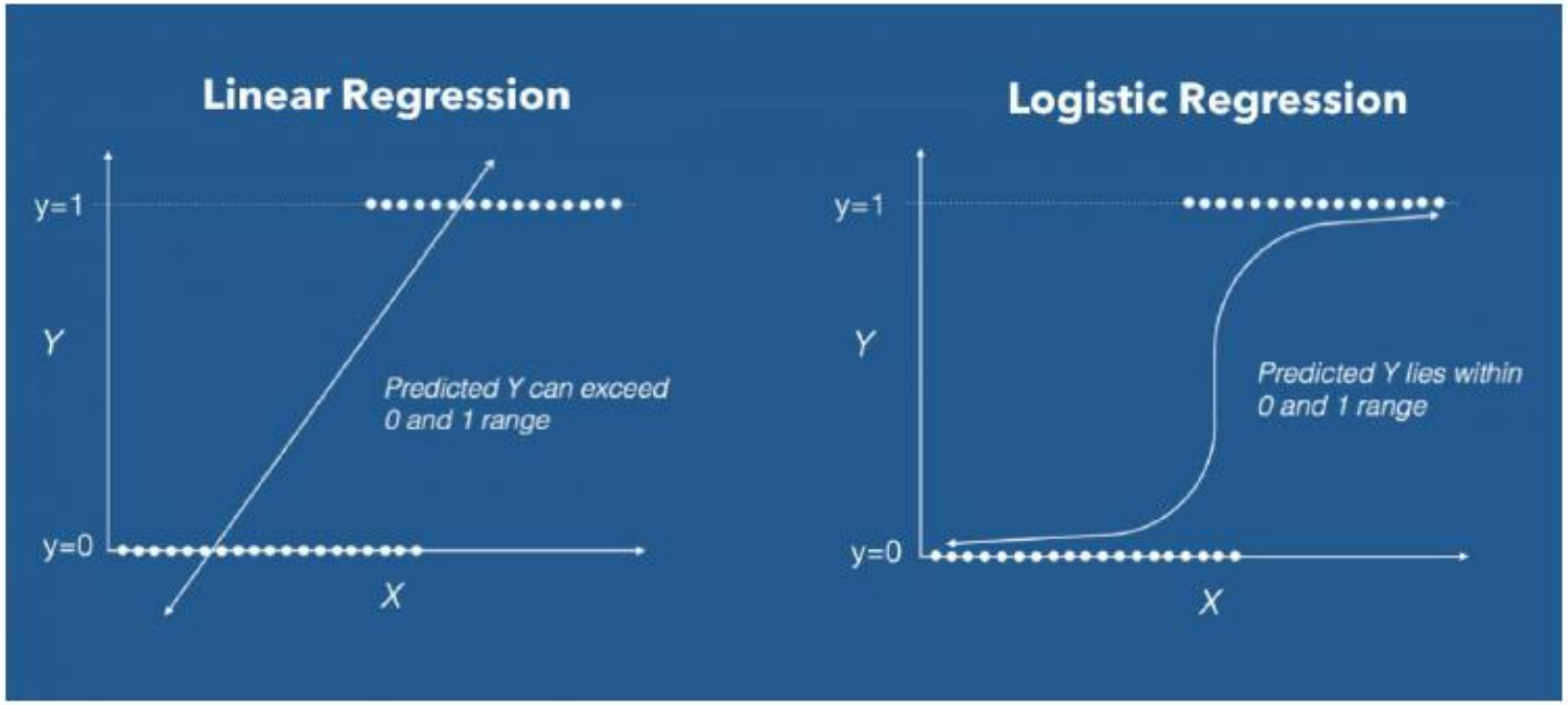
Logistic regression

Generalized Linear Model (GLM) where the response variable is either zero or one

Aims to forecast the probability that the response is 1, given the set of regressors (explanatory variables)

Main concept: not probability but **odds** (often used in gambling)

Why linear regression cannot be used?



Odds

Odds are another way of quantifying the probability of an event, commonly used in gambling (and logistic regression).

Odds

For some event E,

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

Similarly, if we are told the odds of E are x to y then

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

which implies

$$P(E) = x/(x+y), \quad P(E^c) = y/(x+y)$$

Why linear regression cannot be used?

- P must be between 0 and 1, and linear functions are unbounded.
- Changing P by the same amount requires a bigger change in X when P is already large (or small) than when P is close to ½
- Next idea: let log P(X) be a linear function of X, but logarithms are unbounded in only one direction, and linear functions are not at all.
- the easiest modification of log P which has an unbounded range is the logistic (or logit) transformation, log (P/1–*P)*. We can make this a linear function of X without fear of nonsensical results. (Of course the results could still happen to be wrong, but they're not guaranteed to be wrong.)
- This last alternative is **logistic regression**.

Probability vs odds

Measure	Min	Max	Name
$P(Y=1)$	0	1	"probability"
$\frac{P(Y=1)}{1-P(Y=1)}$	0	∞	"odds"
$\log\left[\frac{P(Y=1)}{1-P(Y=1)}\right]$	$-\infty$	∞	"log-odds" or "logit"

Logit function

The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and ∞.

Inverse logit (logistic) function

$$g^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$

The inverse logit function takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1.

This formulation also has some use when it comes to interpreting the model as logit can be interpreted as the log odds of a success, more on this later.

Two alternative formulae for logistic regression

$$logit(p_X) = \log\left(\frac{p_X}{1-p_X}\right) = \beta_0 + \beta_1 X$$

$$p_X = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Two alternative formulae for logistic regression: multiple regressors

$$Z_i = ln\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta_1 x_1 + .. + \beta_n x_n$$

Taking exponent on both sides of the equation gives:

$$P_i = E(y = 1|x_i) = \frac{e^z}{1+e^z} = \frac{e^{\alpha+\beta_i x_i}}{1+e^{\alpha+\beta_i x_i}}$$

Probability of default

$$p = \frac{\exp(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n)}{1 + \exp(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n)}$$

- p is the probability of default
- x_i is the explanatory factor i
- β_i is the regression coefficient of the explanatory factor i
- n is the number of explanatory variables

Classification and estimation

- We predict Y=1 if P>=0.5 and Y=0 if P<0.5. This means guessing 1 if $\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ is positive and 0 if it is negative. This means that logistic regression is a **linear classifier** and $\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ is a hyperplane separating predicted 1s from 0s.

- Estimation of parameters is done via maximum likelihood, noticing that, given the vector of observed features **x**, the outcome (default/no default) is a Bernoilli random variable with probability of 1 equal to P(1|**x**). This immediately gives us expression for likelihood which needs to be maximized numerically:

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i)^{1-y_i}$$

Estimation of parameters

- ML also gives us standard errors of estimates and we can determine p-values (and test null hypothesis of significant effects) by e.g., Wald test statistics:

Wald test statistic

$$W = \frac{\hat{\beta}_1}{\widehat{\mathrm{SE}}(\hat{\beta}_1)}$$

p–value: use of approximate normal distribution of $\hat{\beta}_1$ and standard error.

Example: Nodal metastases vs. phosphatase

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.9919	0.6033	1.64	0.1001
$\log_2$(phosph)	2.4198	0.8778	2.76	0.0058

$$W = \frac{\hat{\beta}_1}{\widehat{\mathrm{SE}}(\hat{\beta}_1)} = \frac{2.42}{0.9} = 2.8$$

Odds Ratio: important for interpretation

		Outcome status +	Outcome status −
Exposure status	+	a	b
	−	c	d

$$\text{OR} = \frac{\text{Odds of being a case given exposed}}{\text{Odds of being a case given unexposed}}$$

$$= \frac{\frac{a}{a+b} / \frac{b}{a+b}}{\frac{c}{c+d} / \frac{d}{c+d}} = \frac{a/c}{b/d} = \frac{ad}{bc}.$$

- Odds Ratios (OR) can be useful for comparisons.
- Suppose we have a trial to see if an intervention T reduces mortality, compared to a placebo, in patients with high cholesterol. The odds ratio is

$$OR = \frac{\text{odds(death|intervention T)}}{\text{odds(death|placebo)}}$$

- The OR describes the benefits of intervention T:
  - OR<1: the intervention is better than the placebo since odds(death|intervention T) $<$ odds(death|placebo)
  - OR$=1$: there is no difference between the intervention and the placebo
  - OR>1: the intervention is worse than the placebo since odds(death|intervention T) $>$ odds(death|placebo)

Interpretation of logistic regression coefficients

$$\log\left(\frac{p_X}{1-p_X}\right) = \beta_0 + \beta_1 X$$

- β_0 is the log of the odds of success at zero values for all covariates.
- $\frac{e^{\beta_0}}{1+e^{\beta_0}}$ is the probability of success at zero values for all covariates

Interpretation cont'd

Slope β_1 is the increase in the log odds ratio associated with a one-unit increase in X:

$$\beta_1 = (\beta_0 + \beta_1(X+1)) - (\beta_0 + \beta_1 X)$$

$$= \log\left(\frac{p_{X+1}}{1+p_{X+1}}\right) - \log\left(\frac{p_X}{1-p_X}\right) = \log\left\{\frac{\left(\frac{p_{X+1}}{1-p_{X+1}}\right)}{\left(\frac{p_X}{1-p_X}\right)}\right\}$$

and e^{β_1}=OR!.

- If $\beta_1 = 0$, there is no association between changes in X and changes in success probability (OR= 1).
- If $\beta_1 > 0$, there is a positive association between X and p (OR> 1).
- If $\beta_1 < 0$, there is a negative association between X and p (OR< 1).

Interpretation cont'd

- OR>1: positive relationship: as X increases, the probability of Y increases; exposure $(X = 1)$ associated with higher odds of outcome.
- OR<1: negative relationship: as X increases, probability of Y decreases; exposure $(X = 1)$ associated with lower odds of outcome.
- OR$=1$: no association; exposure $(X = 1)$ does not affect odds of outcome.

In logistic regression, we test null hypotheses of the form $H_0 : \beta_1 = 0$ which corresponds to OR$=1$.

Interpretation in our example:

Example: OR when phosphatase changes by a factor of 2:

$$\text{OR} = \exp(\beta_1) = 11.2$$

OR for a change by a factor of 1.5:

$$1.5 = 2^{0.585} \longrightarrow \text{OR} = 11.2^{0.585} = 4.1$$

What if more than one regressor?

- Rule of thumb: at least 20 observations for event per regressor, and 20 observations of non-event per regressor

	Estimate	Std. Error	z value	Pr(>|z|)	OR
$\log_2$(phosph)	2.4198	0.8778	2.76	0.0058	11.2
Age	-0.0448	0.0468	-0.96	0.3379	1.0
X-ray	2.1466	0.6984	3.07	0.0021	8.6
Size	1.6094	0.6325	2.54	0.0109	5.0
Grade	1.1389	0.5972	1.91	0.0565	3.1

New model

	Estimate	Std. Error	z value	Pr(>|z|)	OR
(Intercept)	-0.5418	0.8298	-0.65	0.5138	
$\log_2$(phosph)	2.3645	1.0267	2.30	0.0213	10.6
X-ray	1.9704	0.8207	2.40	0.0163	7.2
Size	1.6175	0.7534	2.15	0.0318	5.0

Interpretation:

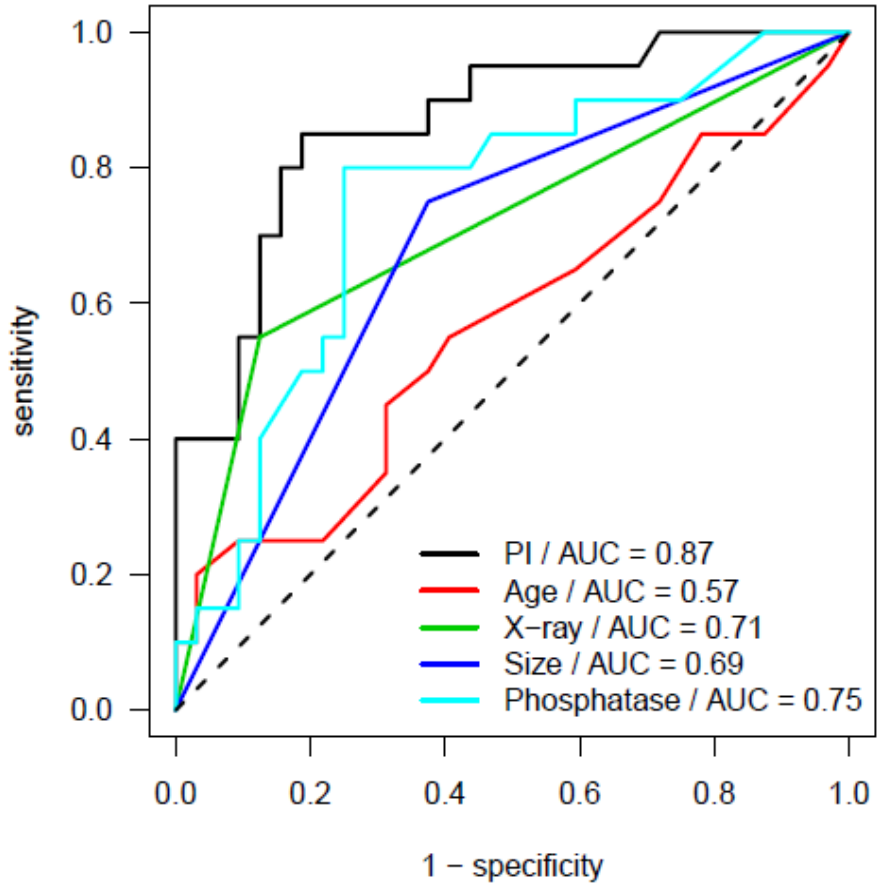
$\hat{\beta}_i$ Influence of x_i when remaining variables are fixed

p–values Does x_i, given the fixed remaining variables, yield additional information about $P(y = 1)$? Significant variables are called "independent risk factors".

Goodness of prediction

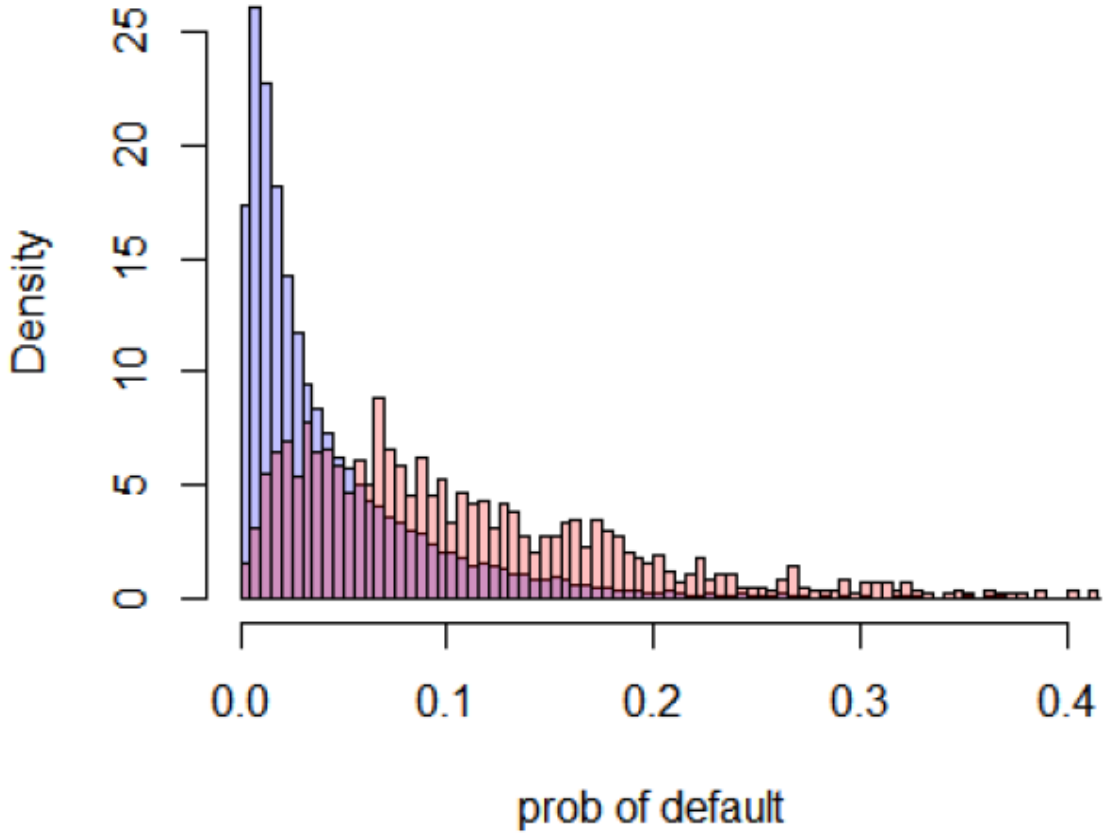
Example: Nodal metastases with prostate cancer

- ROC (receiver operating characteristic) curve:

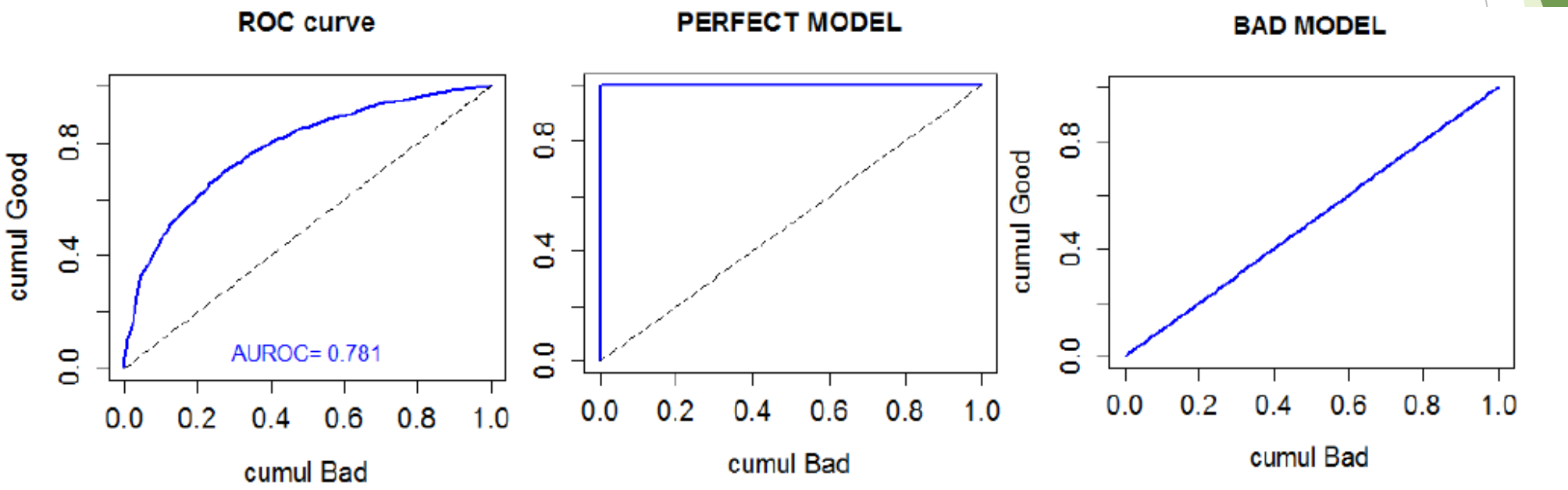


Red: defaulted, blue: not defaulted

- So the problem is choosing correct cut-off value of PoD



ROC: includes in one graph the performance of the model for all cut-off PoD values



Model selection

(i) Compute univariate model for each x variable, eliminate e. g. those with $p > 0.2$

(ii) Build a multiple model with the remaining variables; eliminate clearly non–significant variables

Big issue: class imbalance

To understand that lets assume you have a dataset where 95% of the Y values belong to non-defaulted class and 5% belong to defaulted class.

Had I just blindly predicted all the data points as non-defaulted, I would achieve an accuracy percentage of 95%. Which sounds pretty high. But obviously that is flawed. What matters is how well you predict the defaulted class.

So that requires the defaulted and non-defaulted classes are balanced AND on top of that I need more refined accuracy measures and model evaluation metrics to improve my prediction model. This is **confusion matrix** discussed last time.

Remedies:

- Down Sampling
- Up Sampling
- Hybrid Sampling using SMOTE and ROSE

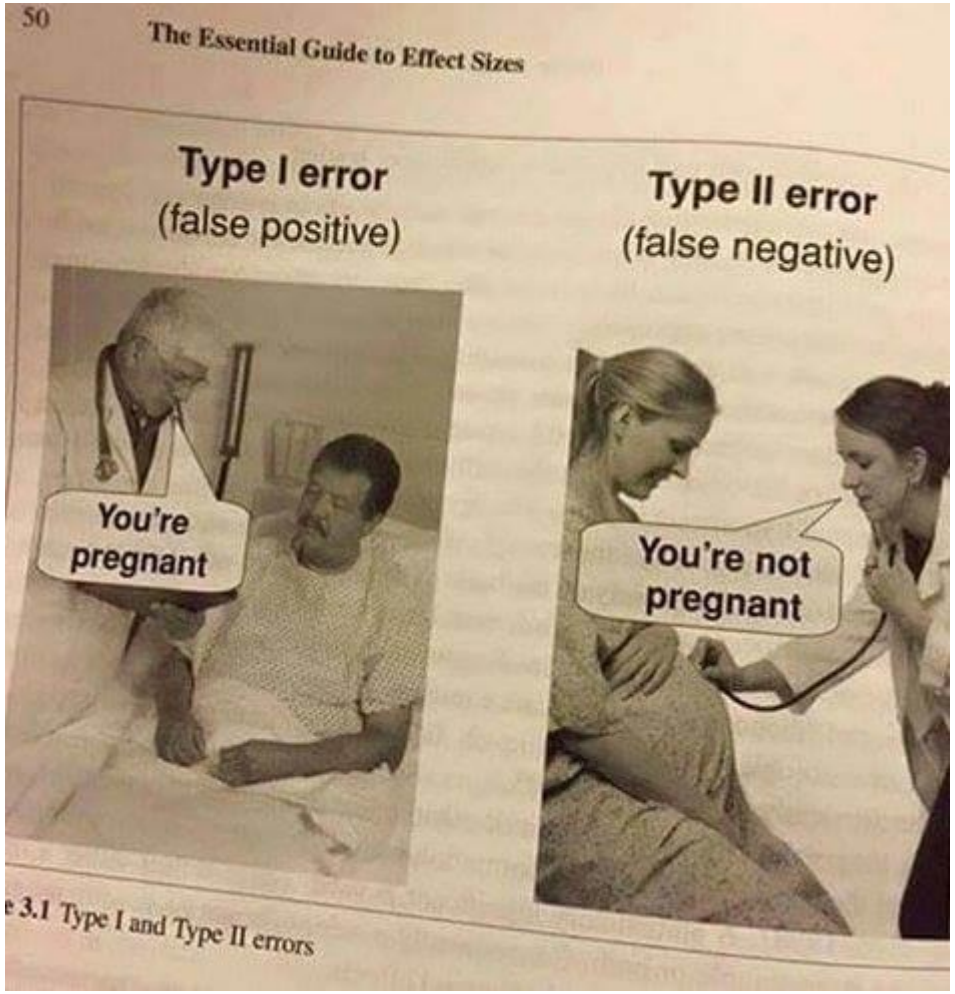
Recall: confusion matrix

	Predicted Bad	Predicted Good
Observed Bad	357	178
Observed Good	3171	7014

- True Positive (= 7014/(7014+3171)) (also called **Recall**)
- True Negative (=357/(357+178))
- **Accuracy**: (True Positive + True Negative) / Total Population

- False positive: Type I error
- False negative: Type II error

- **Precision**: true predicted positive / total predicted positive (=7014/(7014+178))


50
The Essential Guide to Effect Sizes
Type I error
(false positive)
Type II error
(false negative)
You're pregnant
You're not pregnant
e 3.1 Type I and Type II errors

Read this for good explanation of performance measures (confusion matrix etc)

- https://www.machinelearningplus.com/machine-learning/evaluation-metrics-classification-models-r/

Diagnostic compared to linear regression

- R2 is controversial and rarely used

- Residual analysis is different: need different kinds of residuals, such as
  - **Deviance residuals**
  - **Pearson residuals**
  - Schonefeld residuals

- Once these are properly calculated, they can be visually examined and conclusions drawn about possible outliers, heteroscedasticity, nonlinearity etc.

- See documents on Canvas for good reference on these residuals