Biostatistics
Week XII

Ege Ülgen, M.D.

23 December 2021

ACIBADEM
MEHMET ALİ AYDINLAR ÜNİVERSİTESİ

Generalized Linear Models

- A generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for the response variable to have **an error distribution other than the normal distribution**
- The GLM generalizes linear regression by allowing the linear model to be related to the response variable via **a link function**

Logistic Regression

- Logistic regression is a specialized form of regression used when the dependent variable is **binary outcome**
  - Having a binary outcome (dependent variable) violates the assumption of linearity in linear regression

- The goal of logistic regression is to find the best fitting model to describe the relationship between the binary outcome and a set of independent variables
  - e.g., predicting whether the treatment will be successful or not, the presence/absence of a disease, etc.

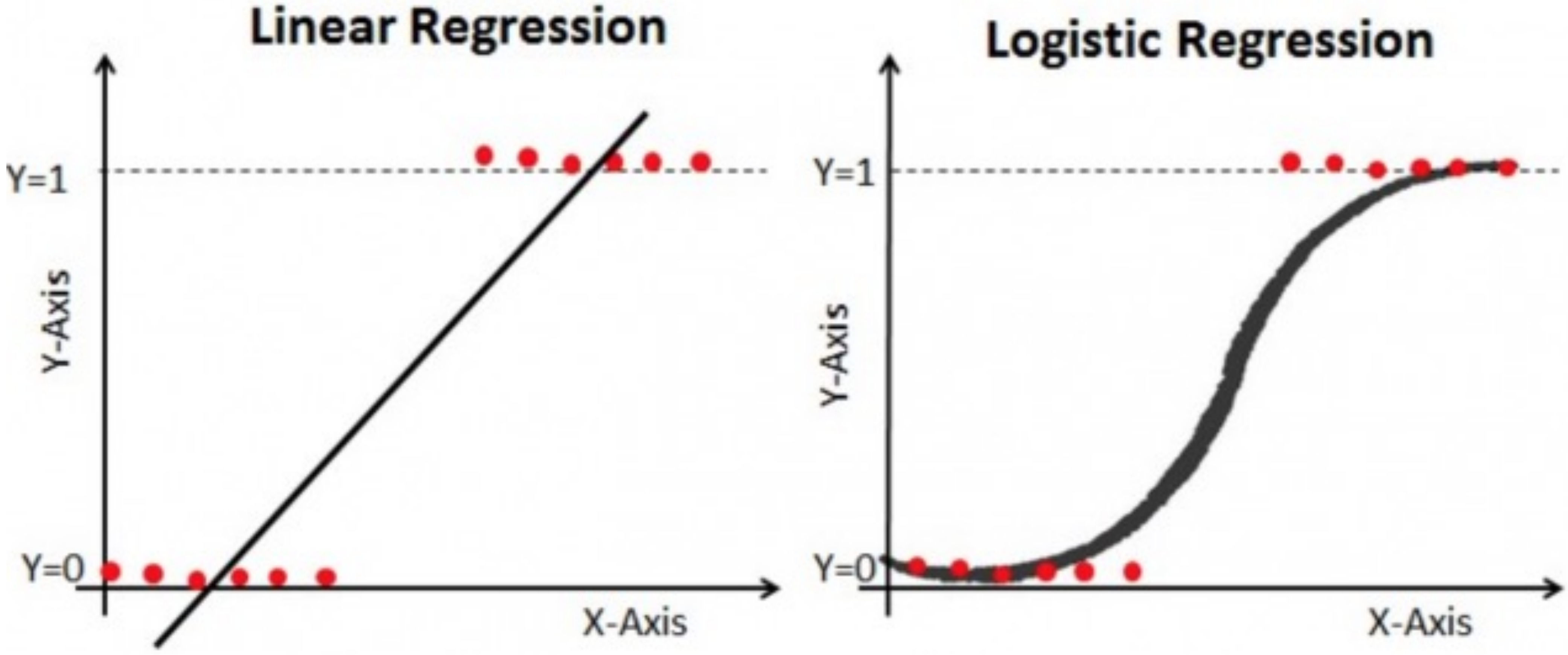
Logistic Regression

- Logistic regression generates the coefficients of the following formula to predict a **logit transformation** of the probability of presence of the outcome:

$$logit(P(Y=1)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- The coefficients are estimated via Maximum Likelihood Estimation (MLE)
- *logit* is in fact the log of odds:

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$



Science O-OD. Logistic regression with python [Internet]. Medium. 2019 [cited 2021 Oct 9]. Available from: https://medium.com/@ODSC/logistic-regression-with-python-ede39f8573c7

Logistic Regression – Example

- Identification of risk factors for lymph node metastases with prostate cancer
- n = 52 patients
- y = nodal metastases (0 = none, 1 = metastases)
- x = phosphatase, age , X-ray result, tumor size, tumor grade
  - The first two variables are continuous, the rest are binary

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffff9de0876f/slides-logistic.pdf

Lymph node metastases – Univariate Models

	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

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffff9de0876f/slides-logistic.pdf

Lymph node metastases – Final 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

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffff9de0876f/slides-logistic.pdf

Interpretation

- With 95% confidence, it could be said that a patient with $\log_2$(phosphatase) = 0, negative X-ray result, size = 0 was equally-likely in terms of having nodal metastases ($p = 0.5138$)
- With 95% confidence, it could be said that $\log_2$(phosphatase) and having nodal metastases are associated ($p = 0.0213$)
  - A one unit increase in $\log_2$(phosphatase) was associated with approximately 963.87% increase in the odds of having nodal metastases
  - $(\exp(2.3645) – 1) * 100 = 963.87$
- …

Poisson Regression

- Linear regression was for continuous outcome, whereas logistic regression for binary outcome
- For **count** outcome, Poisson regression can be used

Poisson Regression - Example

- For 59 epilepsy patients the following data were collected:
  - **treatment:** the **treatment group**, a factor with levels placebo and Progabide
  - **base**: the **number of seizures** collected during 8-week period **before** the trial started
  - **age**: the **age of the patient**
  - **seizure rate:** the **number of seizures** occurred during the 2-week period **after** the trial was started

Poisson Regression – Example (cont.)

- First 10 patients:

treatment	base	age	seizure.rate	subject
placebo	11	31	5	1
placebo	11	30	3	2
placebo	6	25	2	3
placebo	8	36	4	4
placebo	66	22	7	5
placebo	27	29	5	6
placebo	12	31	6	7
placebo	52	42	40	8
placebo	23	37	5	9
placebo	10	28	14	10

Poisson Regression – Example (cont.)

- A Poisson regression with treatment group, previous seizures and age are related to the mean number of of seizure for patient i, λ_i, is given by:

$$log(\lambda_i) = \beta_0 + \beta_1 * I(treatment = Progabide) + \beta_2 * (base - 6) + \beta_3(age - 18)$$

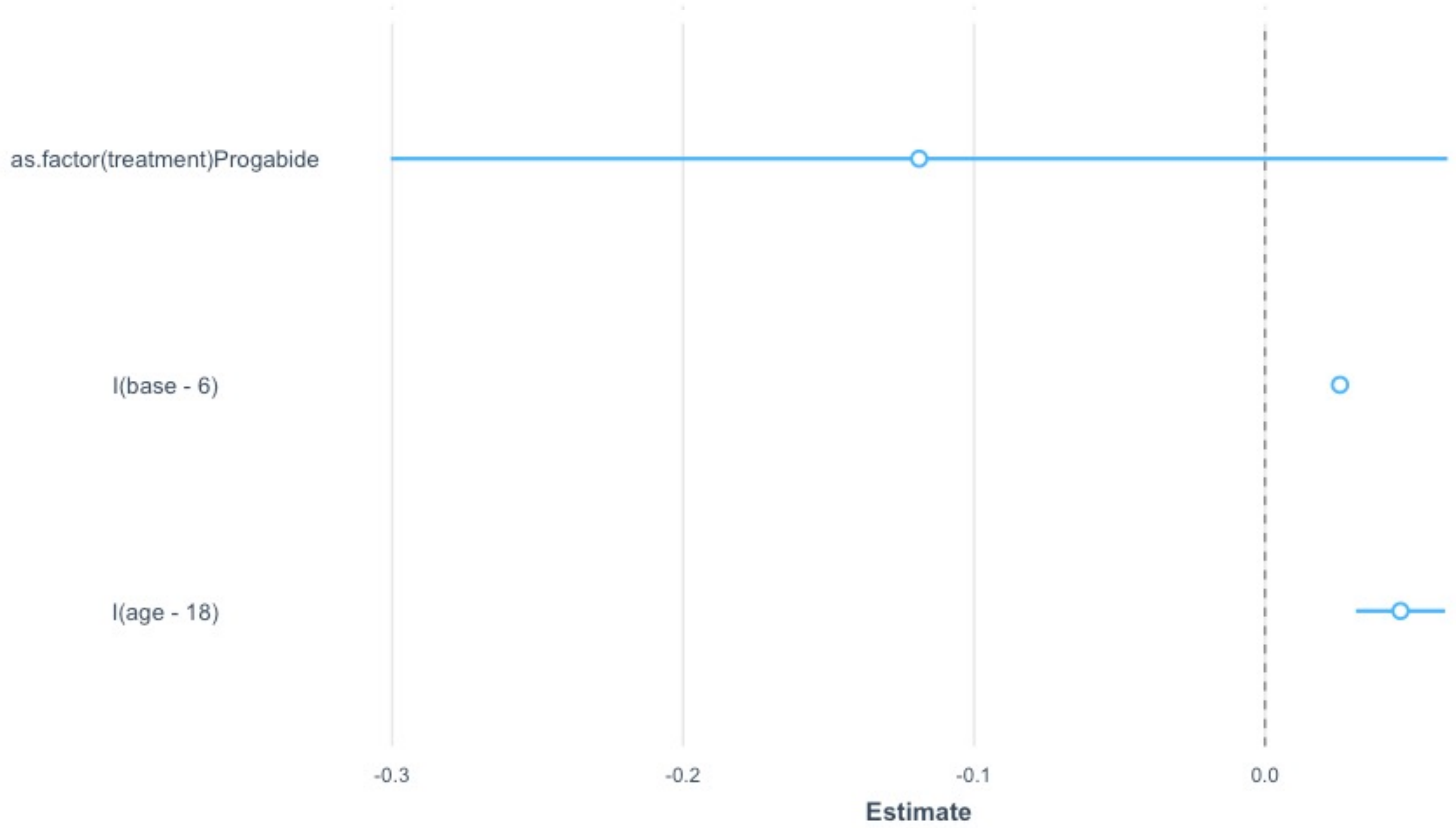
Poisson Regression – Example (cont.)

$$log(\lambda_i) = \beta_0 + \beta_1 * I(treatment = Progabide) + \beta_2 * (base - 6) + \beta_3(age - 18)$$

	Estimate	Std. Error	z value	p
(Intercept)	0.75	0.14	5.33	<0.001
treament = Progabide	-0.12	0.09	-1.28	0.20
base	0.03	0.00	26.37	<0.001
age	0.05	0.01	5.95	<0.001

Poisson Regression – Example (cont.)

$$log(\lambda_i) = \beta_0 + \beta_1 * I(treatment = Progabide) + \beta_2 * (base - 6) + \beta_3(age - 18)$$



Poisson Regression – Example (cont.)

- A patient in placebo group, with 6 previous seizures, and aged 18 had approximately 2 seizures on average in the first two weeks after the trial was started
    - exp(0.75)
- With 95% confidence, it could be said that there was no difference between placebo and progabide (p-value = 0.199)
    - Negative estimate for β_1 indicates lowered mean number of seizures for progabide, but the difference from placebo was not significant

Poisson Regression – Example (cont.)

- With 95% confidence, it could be said that previous number of seizures occurred in the 8-week interval prior to the study start and mean seizure rate was significantly associated (p-value < 0.001)
- One unit increase in previous seizure is associated with approximately 2.6% increase in the mean number of seizures in the first two weeks of the trial
  - (exp(0.03) - 1) * 100

Poisson Regression – Example (cont.)

- With 95% confidence, it could be said that age sand mean seizure rate was significantly associated (p-value < 0.001)
- One unit increase in age is associated with approximately 4.8% increase in the mean number of seizures in the first two weeks of the trial
    - (exp(0.05) - 1) * 100

Other Regression Models

- Multinomial Logistic Regression
  - generalizes logistic regression to multiclass problems, i.e., with more than two possible discrete outcomes
- Ordinal Regression
  - used for predicting an ordinal variable
- Polynomial Regression
  - the relationship between the independent variable(s) and the dependent variable y is modelled as an n^{th} degree polynomial

…

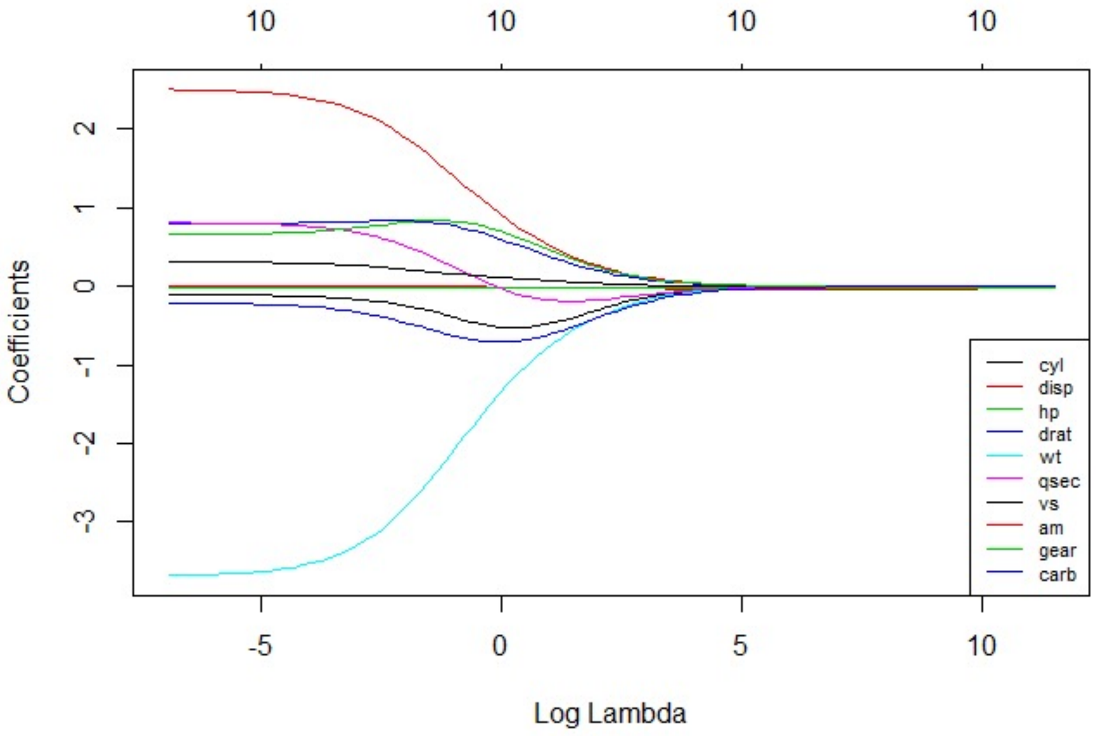
Ridge Regularization

$$\sum_{i=1}^{M} (y_i - \hat{y_i})^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} w_j^2$$

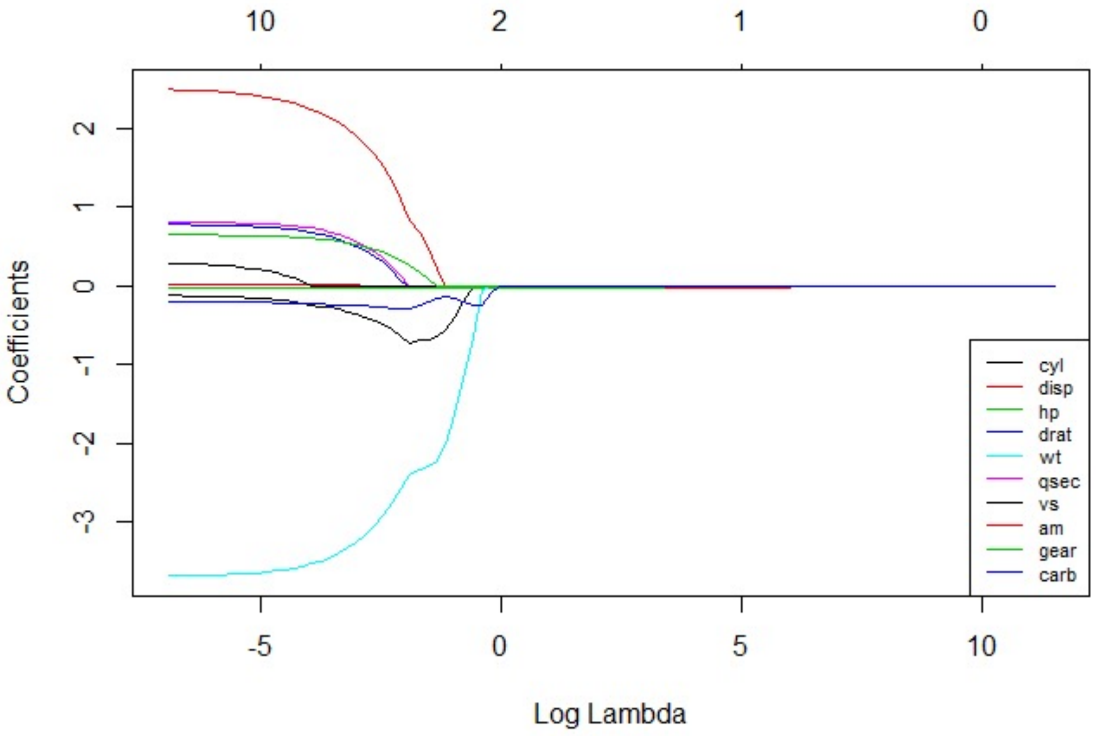
Lasso Regularization

$$\sum_{i=1}^{M} (y_i - \hat{y_i})^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} |w_j|$$

Ridge

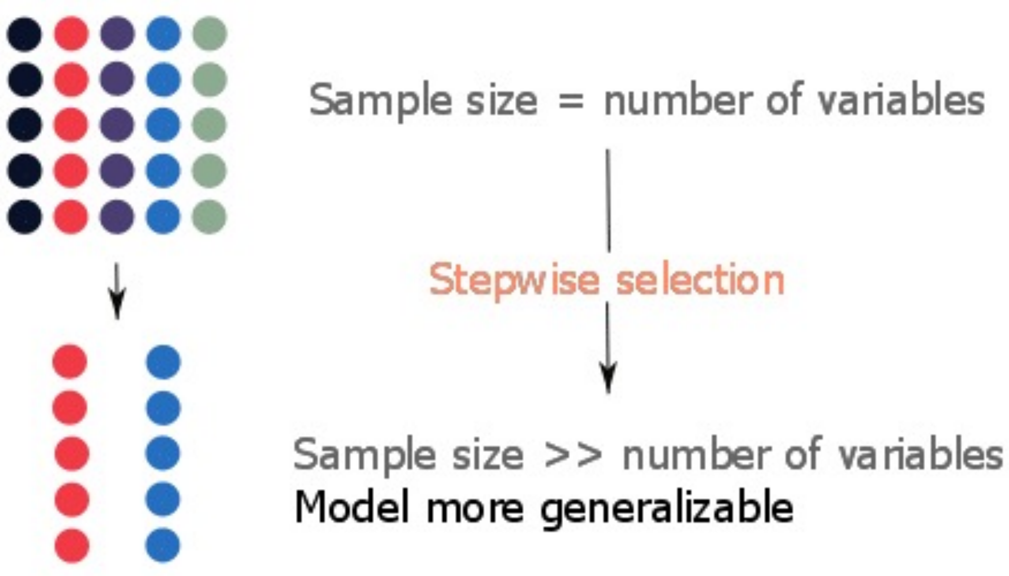


Lasso



https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net

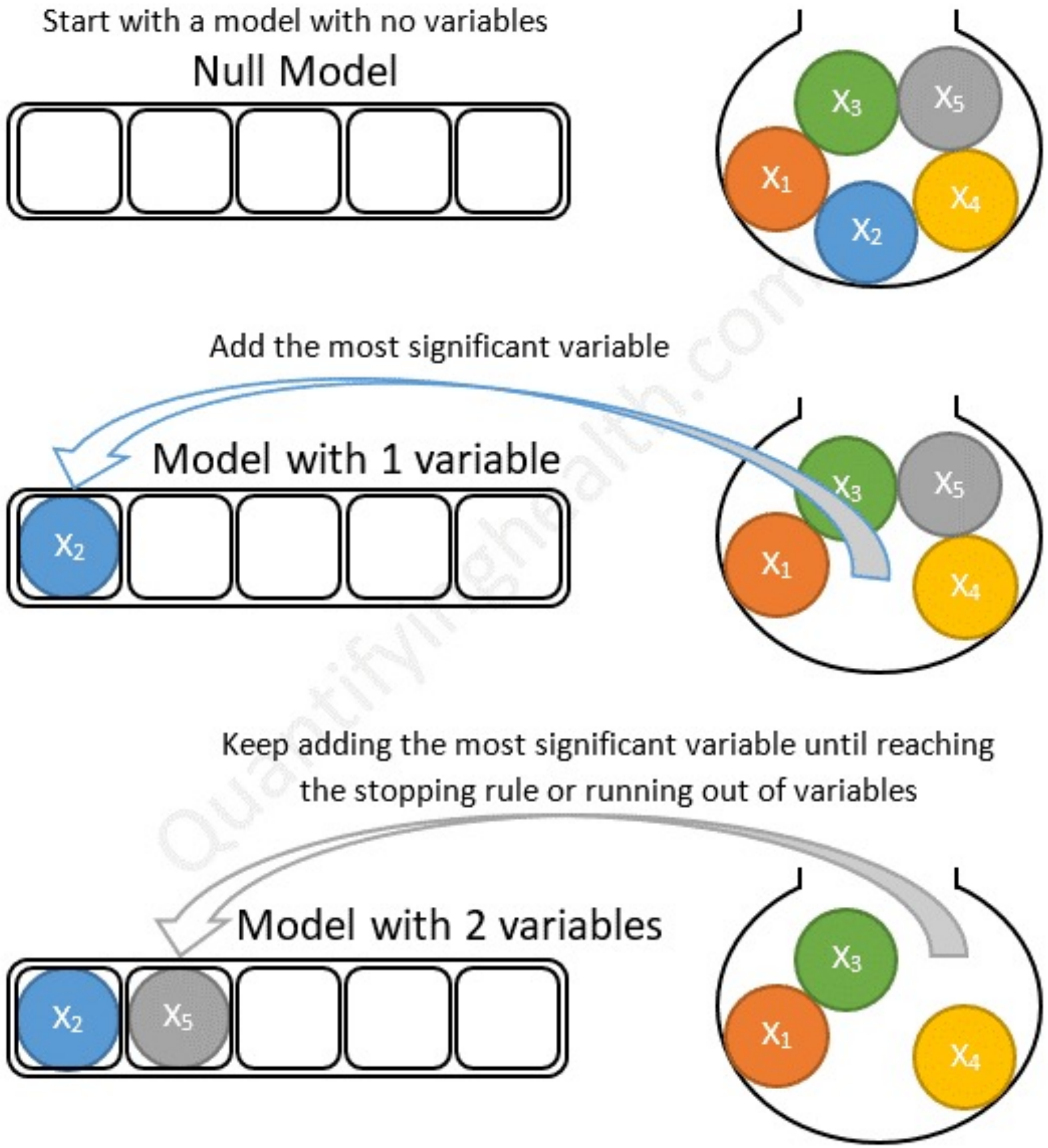
Stepwise Regression



- When the sample size is not much larger than the number of predictors, the regression model will perform poorly in terms of out-of-sample accuracy
- Reducing the number of predictors in the model by using stepwise regression will improve out-of-sample accuracy
  - Forward stepwise selection
  - Backward stepwise selection

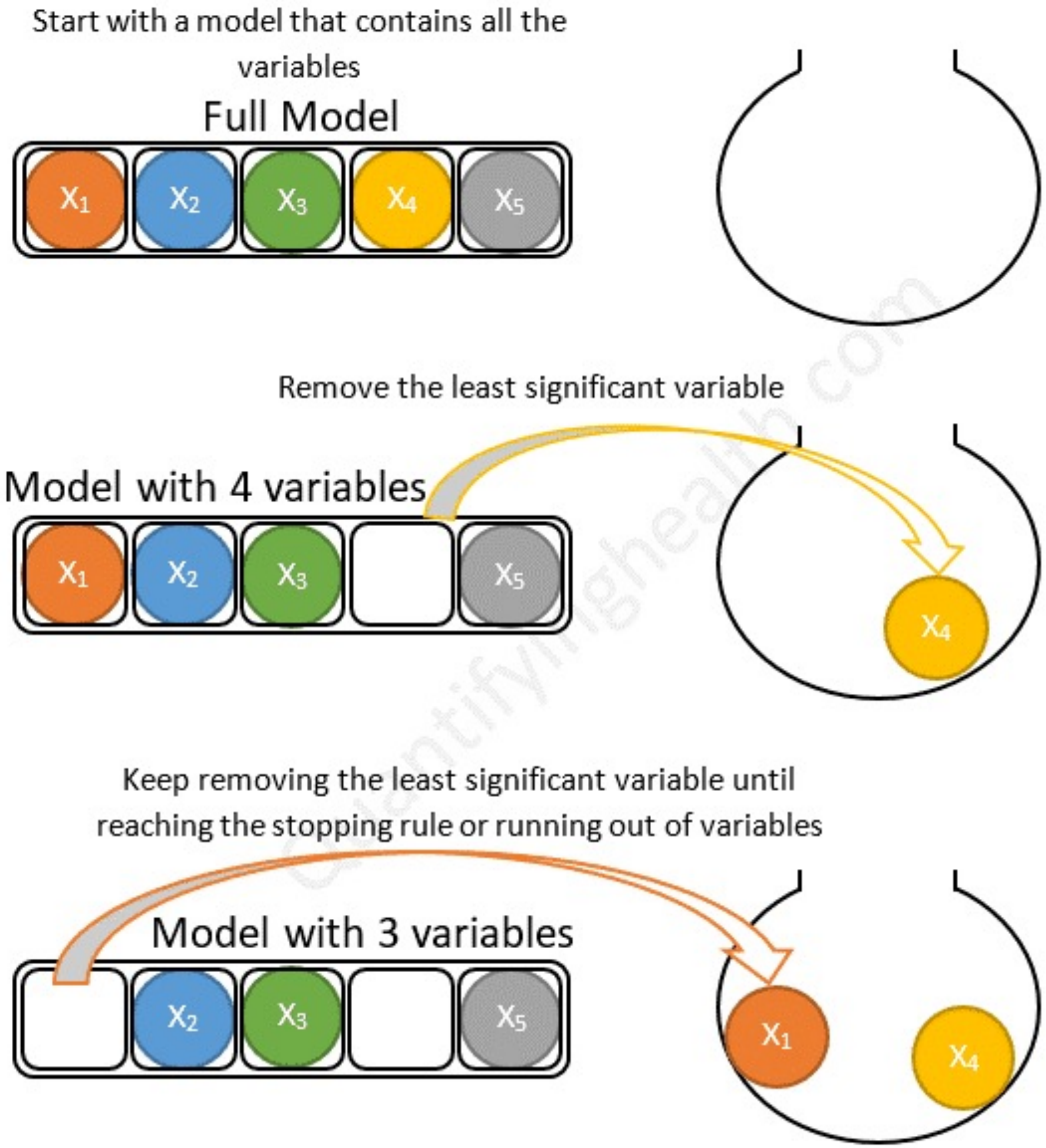
Choueiry G. Understand forward and backward stepwise regression – quantifying health [Internet]. [cited 2021 Nov 2]. Available from: https://quantifyinghealth.com/stepwise-selection/

Forward stepwise selection example with 5 variables:



Choueiry G. Understand forward and backward stepwise regression – quantifying health [Internet]. [cited 2021 Nov 2]. Available from: https://quantifyinghealth.com/stepwise-selection/

Backward stepwise selection example with 5 variables:



oueiry G. Understand forward and backward stepwise regression – quantifying health Internet]. [cited 2021 Nov 2]. Available from: https://quantifyinghealth.com/stepwise-selection/

Brief Summary

Dependent Variable	Link function	Regression Model
Continuous	Y	Linear Regression
Binary	logit(Y)	Logistic Regression
Count	log(Y)	Poisson Regression (Log-linear model)