Logistic Regression

1. Ask

Can we predict whether the patient has 10-year risk of future coronary heart disease (CHD)?

Is dependent variable Y
(binary: having a desease 1 or not 0)
depends on independent variables
(factors X_1,X_2,X_3,X_4,X_5)?
And if so, what is the relationship,
can we use it to predict Y?

2. Prepare

Data: “Heart desease”

from Turing provided Spreadsheet

1. Data Preparation

- Ineffective variables removed
- Missing values updated
- Outliers
- Data Randomisation
- Data Splitting: 80% for training / 20% for testing

3. Process

1. Data Preparation in detail

- **Ineffective variables removed**

“education” variable (values 1234)
because there is no explanation about it in the description

- **NA replaced with avg, mode, median**

with average: cigsPerDay (because continious: 1,2,3,4,5,6, ...)
with mode: BPmeds (because nominal: 0 and 1)
with median: totChol, BMI, heartRATE, glucose (because of outliers)

It’s best to use the **mean** to describe the center of a dataset when the distribution is mostly symmetrical and there are **no outliers**.

It is best to use the **median** when the distribution is either skewed or there **are outliers** present.

- **Outliers decision**

IQR method (small ranges)
Box Plot (glucose case)
Medical parameters (specific)

Generally speaking, an extreme value called an outlier,
is one that is **distant** from most of the other observations.

- **Randomisation of data**

A process of rearranging or **shuffling the order** of data elements in a dataset in a random manner.
When building a logistic regression model, it's common practice to split the available data into training and testing sets. Randomization ensures that each observation has an equal chance of being included in either the training or testing set, helping **to avoid bias** in model evaluation.

Fraction of classes:
rare vs dominant **1/7 or 0,15**
in both 80% and 20% data sets.

- **Splited data into 80% / 20%**

The purpose of this split is **to assess** how well the model generalizes to unseen data, which is crucial for evaluating its performance and ensuring it doesn't overfit the training data.

This split strikes a balance between having sufficient data for training and ensuring a reasonable amount of data for evaluation.

These ratios are not strict rules and can be adjusted based on factors such as the size of the dataset, the complexity of the model, and the specific requirements of the problem at hand. The key is to strike a **balance** between having enough data for training to capture patterns in the data and having enough data for testing to evaluate the model's generalization performance accurately.

4. Analyze

2.Model Building:

- Tool: Jamovi
- 80% of data for training
- Assigned data types
  Nominal / Continious
- Selected variables (p < 0.05)
  statistically significant
- Selected Cut-Off to 0.1

Binomial Logistic Regression

Model Fit Measures

Model	Deviance	AIC	R^2_{McF}
1	2582	2596	0.108

Model Coefficients - TenYearCHD

Predictor	Estimate	SE	Z	p	Odds ratio
Intercept	−8.50092	0.43707	−19.45	<.001	2.03e−4
age	0.06056	0.00660	9.18	<.001	1.06
cigsPerDay Full	0.02420	0.00430	5.63	<.001	1.02
sysBP	0.01803	0.00225	8.02	<.001	1.02
male:					
1 – 0	0.40851	0.10909	3.74	<.001	1.50
prevalentStroke:					
1 – 0	1.12316	0.52918	2.12	0.034	3.07
glucose Full	0.00939	0.00203	4.61	<.001	1.01

Note. Estimates represent the log odds of "TenYearCHD = 1" vs. "TenYearCHD = 0"

Classification Table – ...

Observed	Predicted 0	Predicted 1	% Correct
0	1280	1593	44.6
1	96	421	81.4

Note. The cut-off value is set to 0.1

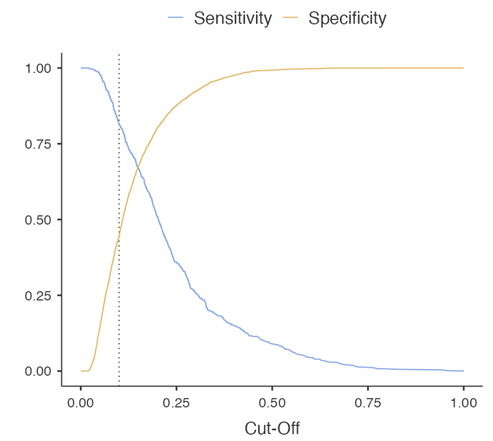
Predictive Measures

Accuracy	Specificity	Sensitivity	AUC
0.502	0.446	0.814	0.725

Note. The cut-off value is set to 0.1

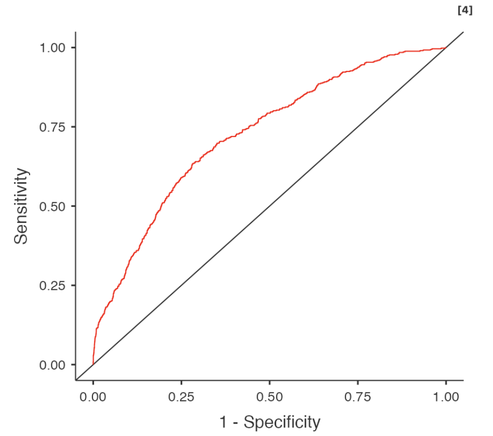
Prediction

Cut-Off Plot



ROC Curve

[4]



4. Analyze

2. Model building in detail

Jamovi (Model Builder)

- **assigned data types: Nominal / Continious**

Nominal data concept consists of **categories** or labels without any order.
Continious data concept consists of **numerical values** that can be measured and take any value within a range.

- **Variables selected according to p value < 0.05**
  statistically significant

In model building, researchers often use p-values to decide which independent variables to include in the final model. Variables with p-values below a chosen significance level (e.g., 0.05) are considered statistically significant and are more likely to be included in the model, as **they suggest a meaningful relationship with the outcome**.

A significance level of 0.05 is quite stringent. It means that the probability of observing the data if the null hypothesis were true (i.e., no true relationship between the independent variable and the outcome) must be very low (less than 0.05) for us to reject the null hypothesis. By setting such a strict criterion, we ensure that **only variables with a very strong relationship with the outcome are included in the model.**

age of the patient (Continuous)
cigsPerDay on average (Continuous)
sysBP systolic blood pressure (Continuous)
glucose level (Continuous)
male male or female (Nominal)

- **Setting Cut-Off**

The default **decision point**, or Cut-Off, is typically 0.50 or 50%.
If the probability is above 0.5, the classification is “1”; otherwise it is “0.”

Setting the cutoff at 0.1 to **prioritize high sensitivity** can be acceptable, especially in medical scenarios where **detecting positive cases is crucial** and the consequences of missing them (false negatives) outweigh the costs of false positives.

Lower cutoff is often appropriate if the goal is to identify members of a rare class.

Selecting the best model

Classification Table – ...

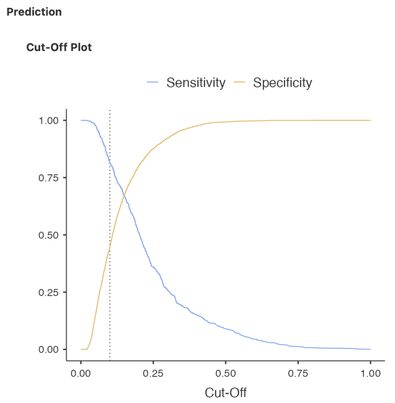
Observed	Predicted 0	Predicted 1	% Correct
0	1280	1593	44.6
1	96	421	81.4

Note. The cut-off value is set to 0.1

Predictive Measures

Accuracy	Specificity	Sensitivity	AUC
0.502	0.446	0.814	0.725

Note. The cut-off value is set to 0.1



Classification Table – ...

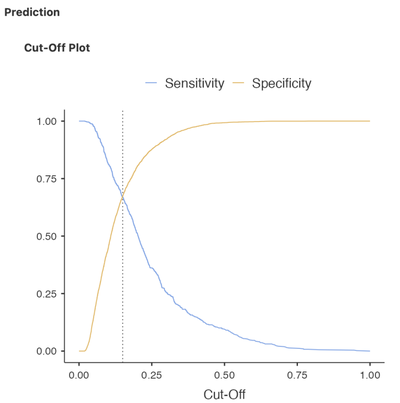
Observed	Predicted 0	Predicted 1	% Correct
0	1937	936	67.4
1	171	346	66.9

Note. The cut-off value is set to 0.15

Predictive Measures

Accuracy	Specificity	Sensitivity	AUC
0.673	0.674	0.669	0.726

Note. The cut-off value is set to 0.15



4. Analyze

3. Model testing

- Tool: Exel Online
- Test with 20% of data
- Model evaluation by comparison of metrics

Testing results with 20% of data

		Predicted 0	Predicted 1		% Correct
Actual	0	280	441	721	38,83%
Actual	1	14	113	127	88,98%
	Total	294	554	848	

Accuracy	46,34%
Specificity / TN Rate	38,83%
Sentitivity / Recall	88,98%
AUC	≈70,82%
Precission	20,40%
TP Rate	88,98%
FP Rate	61,17%

SUCCES

		Predicted 0	Predicted 1	
Actual	0	True Negative	False Positive	TotalActualNegatives
Actual	1	False Negative	True Positive	TotalActualPositives
	Total	TotalNegatives	TotalPositives	Total Sample Size/Population

Training results with 80% of data

Classification Table – ...

Observed	Predicted 0	Predicted 1	% Correct
0	1280	1593	44.6
1	96	421	81.4

Note. The cut-off value is set to 0.1

Predictive Measures

Accuracy	Specificity	Sensitivity	AUC
0.502	0.446	0.814	0.725

Note. The cut-off value is set to 0.1

In both the training and testing datasets, sensitivity is relatively high (>80%), indicating that the **model is effective at identifying positive cases** which is very important for predicting CHD.

However, specificity is a bit lower in the testing dataset, suggesting that **the model may incorrectly classify some non-CHD individuals as having CHD**.

While the AUC values for both the training and testing datasets are moderate (72.5% and 70.82% respectively), they indicate that **the model performs not perfect but better than random chance**.

In both datasets, the accuracy is relatively low (around 50%), indicating that the model's performance is not much better than random guessing. **The difference between sensitivity and specificity may contribute to this low accuracy, as the model may prioritize one at the expense of the other.**

5. Share

4. Metrics interpretation

In this case, the sensitivity (81.4%) is relatively high, indicating the model's ability to correctly identify individuals with CHD.
However, the specificity (44.6%) is lower, suggesting that the model may incorrectly classify some non-CHD individuals as having CHD.

Room for Improvement: While a 72.5% AUC is moderate to good, there may still be room for improvement in the model's performance. Optimizing the model's parameters, feature selection, or incorporating additional data could potentially enhance its predictive ability and overall performance.

Classification Table – ...

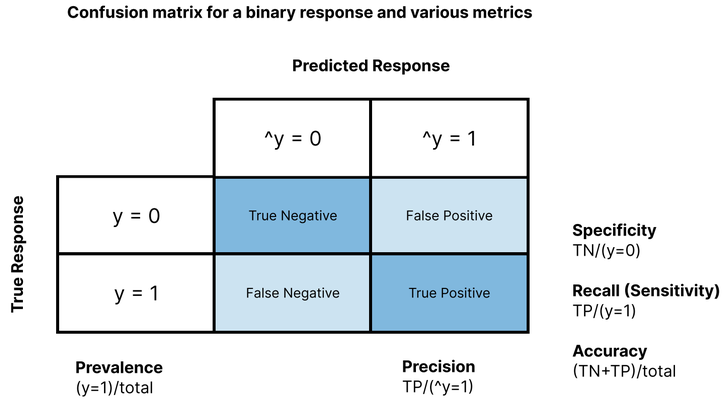
	Predicted		
Observed	0	1	% Correct
0	1280	1593	44.6
1	96	421	81.4

Note. The cut-off value is set to 0.1

Predictive Measures

Accuracy	Specificity	Sensitivity	AUC
0.502	0.446	0.814	0.725

Note. The cut-off value is set to 0.1



Accuracy

The percent (or proportion) of **cases classified correctly**.

50.2%
Model's ability to predict correctly is not high.

Specificity

The percent (or proportion) of all 0s that are **correctly classified as 0s**.

44.6%
Model's ability to predict a negative outcome is also not high.

Sensitivity/Recall

The percent (or proportion) of all 1s that are **correctly classified as 1s**.

81.4%
Model strength to predict a positive outcome - the proportion of the 1s that it correctly identifies is quire high.

In the context of CHD prediction, high sensitivity is crucial because it ensures that the model can correctly identify as many CHD patients as possible, minimizing false negatives. **Missing a CHD diagnosis could have serious consequences, so maximizing sensitivity is often a priority.**

AUC

Area under the ROC curve.

The **larger the value of AUC, the more effective** the classifier.

An AUC of **1 indicates a perfect classifier**: it gets all the 1s correctly classified, and it doesn't misclassify any 0s as 1s.
A completely **ineffective classifier** - the diagonal line - will have an **AUC of 0.5**.
The model has an AUC of about 0.69, corresponding to a **relatively weak classifier**.

72.5%
Moderate to good performance in CHD prediction. Significantly outperforms random guessing.

Rare Class Problem

In many cases, there is an imbalance in the classes to be predicted,
with one class much more dominant than the other.
The rare class is usually the class of more interest and is typically designated 1,
in contrast to the more dominant 0s.

Imbalance in the data in Graded task case:
Total data rows 4238
“having desease” 644 rare class of more interest 1s
“not having desease” 3594 dominant class 0s

In the typical scenario of healthcare data,
the 1s are the more important case (having desease),
in the sense that misclassifying them as 0s (not having) is worse than
misclassifying 0s as 1s.
For example, correctly identifying a desease may save a patient.
On the other hand, correctly identifying no desease merely saves a patient.

In such cases, **unless the classes are easily separable,**
the most accurate classification model may be one
that simply classifies everything as a 0.
For example, if only 0.1% of patients end up having,
a model that predicts that each patient without desease will be 99.9% accurate.
However, it will be useless.
Instead, we would be happy with a model that is
less accurate overall but is good at picking out those who has desease,
even if it misclassifies some nondeseasers along the way.

5. Share

Binomial Logistic Regression

Model Fit Measures

Model	Deviance	AIC	R^2_{McF}
1	2582	2596	0.108

Model Coefficients - TenYearCHD

Predictor	Estimate	SE	Z	p	Odds ratio
Intercept	–8.50092	0.43707	–19.45	<.001	2.03e–4
age	0.06056	0.00660	9.18	<.001	1.06
cigsPerDay Full	0.02420	0.00430	5.63	<.001	1.02
sysBP	0.01803	0.00225	8.02	<.001	1.02
male:					
1 – 0	0.40851	0.10909	3.74	<.001	1.50
prevalentStroke:					
1 – 0	1.12316	0.52918	2.12	0.034	3.07
glucose Full	0.00939	0.00203	4.61	<.001	1.01

Note. Estimates represent the log odds of "TenYearCHD = 1" vs. "TenYearCHD = 0"

Odds Ratio

The odds ratio in logistic regression is a fundamental concept for understanding the **relationship between independent variables (predictors) and binary outcomes**, providing **insight** into **how changes in predictor variables influence the odds of the outcome occurring**.

e.g. The smoking group has 2% (1.02 – 1 = 0.02) more odds of having CHD than the non-smoking group.

The standard error is a measure of uncertainty of the logistic regression coefficient. It is useful for calculating the p-value and the confidence interval for the corresponding coefficient.

When fitting a logistic regression model, the coefficients associated with each independent variable represent the **log of the odds ratio**.

Specifically, if we denote **β as the coefficient** for a particular independent variable, then the odds ratio (OR) corresponding to that variable is given by:

$$\mathbf{OR = e^{\beta}}$$

This means that for a one-unit increase in the independent variable, the odds of the outcome (e.g., CHD or non-CHD) occurring will be multiplied by the odds ratio.

For **example**, we have a logistic regression model predicting the probability of a patient having CHD based on their age. If the coefficient for age is 0.06, then the odds ratio associated with age is e0.06, which is approximately 1.06. This means that for every one-unit increase in age, the odds of having the disease increase by a factor of 1.06, or approximately 6%.

The **interpretation** of the odds ratio depends on whether the value is greater than 1, equal to 1, or less than 1:

- If the odds ratio is greater than 1, it indicates that as the independent variable increases, the odds of the outcome increase.
- If the odds ratio is equal to 1, it suggests that the independent variable does not have an effect on the odds of the outcome.
- If the odds ratio is less than 1, it suggests that as the independent variable increases, the odds of the outcome decrease.

6. Act

Actionable recommendations

- **Prediction of having future CHD or not**
  is associated with statistically significant factors such as:
  age(X_1),
  cigaretsPerDay(X_2),
  systolicBloodPressure(X_3),
  gender(X_4),
  prelevantStroke(if had previously had a stroke)(X_5),
  glucose level(X_6).

- **The relationship is equation:**
  $Y = -8.5 + 0.06X1 + 0.02X2 + 0.02X3 + 0.4X4 + 1.12X5 + 0.01X6$

- **We can use it to predict whether the patient has**
  10-year risk of future coronary heart disease:
  with 81.4% Sensitivity, 72.5% AUC, 44.6% Specificity, 50.2% Accuracy.
  Moderate to good performance in CHD prediction.
  Significantly outperforms random guessing.

The model could be improved with:

- More Data in general
- More Balanced Data
- Better identification and Removal of Outliers (specific medical knowledge required)

Recommendations for patients:

- reducing or quitting smoking can substantially decrease the risk of developing CHD.
- lower glucose levels are typically associated with a lower risk of CHD.

Thank You