

Question 1

1 pts

Which one of the following statements is FALSE?

- ☐ The "dredge" function can be used to fit all possible subset models using BIC as the model selection criterion.
- ☐ Cross-validation can be used to get a realistic estimate of prediction error when the data set is not large enough to be divided into a training set and a test set.
- ☐ Two common purposes for fitting regression models are for explanation or prediction.
- ☐ For large sample sizes, AIC and AICc will give almost the same values.
- ☒ Using the same data to fit a predictive model and to assess the prediction error rate tends to result in an estimate that is too high.

E

Comments:

A: "dredge" function generates a model selection table of models with combinations (subsets) of fixed effect terms in the global model, with optional model inclusion rules.

D: When the number of observations is large the Akaike Information Criterion (AIC) and the small-sample corrected Akaike Information Criterion (AICc) become extremely similar because AICc converges to AIC. Therefore we gain (or lose) almost nothing by switching between the two criteria.

E: Using same data to fit predictive model and access prediction error rate will result in estimate of prediction error rate lower.

Question 2

1 pts

Suppose that W has a chi-squared distribution with 24 degrees of freedom.

We wish to estimate the probability $\Pr(W > 26)$ using simulation. What R command will achieve this? Assume that N is a very large number.

- ☐ `var(rchisq(N, 26) > 24)`
- ☐ `mean(rchisq(N, 26) > 24)`
- ☒ `mean(rchisq(N, 24) > 26)`
- ☐ `var(rchisq(N, 24) > 26)`
- ☐ `mean(rchisq(N, 24) < 26)`
- ☐ `mean(rchisq(N, 26) < 24)`

C

Comments:

`rchisq(n, df)` generates random deviates.

What model formula corresponds to the model that should be used to estimate all of the direct causal effects on variable C?

- ☒ C~B+D
- ☐ C~A
- ☐ C~A+E
- ☐ C~A+B+D+E
- ☐ C~A+B+D
- ☐ C~E

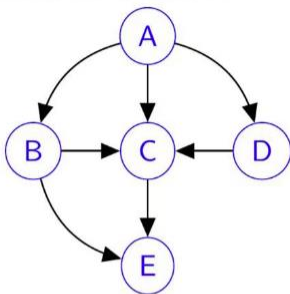
A

Comments:

A is confounder and should be controlled.

Colliders should be left uncontrolled and confounders that should be controlled.

Consider the following causal diagram.



Which model formula corresponds to the model that should be used to estimate the total causal effect that A has on E?

- ☐ E~B+C
- ☒ E~A+B+C+D
- ☐ E~A+B+C
- ☐ E~A+D
- ☐ E~A

B

Comments:

B is mediator in this case and the total effect can be broken into two parts: the direct and indirect effect. The direct effect is the effect of exposure on the outcome absent the mediator. The indirect pathway is the effect of exposure on the outcome that works through the mediator.

What are the estimated values of sensitivity and specificity?

- ☐ sensitivity = .687 and specificity = .131
- ☐ sensitivity = .817 and specificity = .764
- ☐ sensitivity = .764 and specificity = .817
- ☐ sensitivity = .313 and specificity = .131
- ☒ sensitivity = .687 and specificity = .869

E

Comments:

Sensitivity (true positive rate) = the probability of a positive test, conditioned on truly being positive.

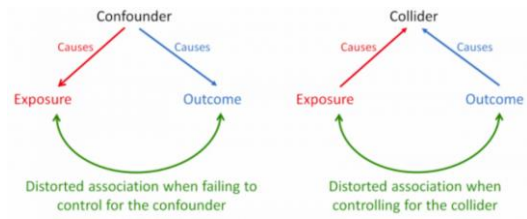
Specificity (true negative rate) = the probability of a negative test, conditioned on truly being negative.

Which of the following statements is correct with respect to the direct effect of B on C?

- ☐ To estimate this direct causal effect both A and E should be included as regressors (in addition to B).
- ☐ Both A and E are colliders.
- ☒ A is a confounder and E is a collider.
- ☐ A is a collider and E is a confounder..
- ☐ Both A and E are confounders.
- ☐ To estimate this direct causal effect, neither A or E should be included as regressors (in addition to B).

C

Comments: Understand the difference between confounder and collider.



Question 7

1 pts

Which of the following statements about predicting a binary response is FALSE.

- ☒ Specificity is the probability of a false positive.
- ☐ The area under a ROC curve is always less than or equal to 1.
- ☐ Sensitivity is the probability of a true positive.
- ☐ Using a random process to predict a binary response will give an area under the ROC curve of 0.5.
- ☐ An ROC curve shows the impact of changing the threshold on sensitivity and specificity for a predictive logistic model.

A

Comments:

Specificity is true negative rate (TNR)

Question 8

1 pts

Which one of the following statements concerning diagnostic procedures is false?

- ☐ A Standardized Residuals vs Leverage Plot is useful for identifying influential points.
- ☒ Variance inflation factors are useful for detecting non-constant variance.
- ☐ A gam plot is useful for identifying a suitable transformation of a regressor.
- ☐ A correlogram is useful for detecting serial correlation.
- ☐ A Box-Cox plot is useful for identifying a suitable transformation of the response.

B

Comments:

Variance inflation factor measures how much the variance of an independent variable is influenced by its interaction/correlation with the other independent variables.

To detect non-constant variance, usually we make scatter plots of the residuals versus the explanatory variables and versus the predicted values from the model, which is a comparison of the amount of random variation in different parts of the data.

Question 9		1 pts		
A classic British study on smoking and mortality investigated the effect of smoking on coronary death rates (measured as deaths per 100,000 person-years). Assume the data have been arranged in a data frame with variables <code>person.years</code>, <code>deaths</code>, <code>age.group</code> and <code>smokers</code>.				
	<code>person.years</code>	<code>deaths</code>	<code>smokers</code>	<code>age.group</code>
1	18793	2	No	35-44
2	52407	32	Yes	35-44
3	10673	12	No	45-54
4	43248	104	Yes	45-54
5	5710	28	No	55-64
6	28612	206	Yes	55-64
7	2585	28	No	65-74
8	12663	186	Yes	65-74
9	1462	31	No	75-84
10	5317	102	Yes	75-84
Given that we wish to model death rate per 100,000 person years, which of the following is TRUE?				
☐ We should use <code>deaths</code> as the response and <code>person.years</code> as an explanatory variable.				
☐ We should include an offset of <code>log(person.years/100000)</code> .				
☒ We should divide <code>person.years</code> by <code>deaths</code> and use this as the response.				
☐ We should take <code>person.years</code> as the response and <code>deaths</code> as an explanatory variable.				
☐ We should include an offset of <code>log(person.years)</code> .				

C

Comments:

Based on the question, since the death rate is measured by 100k person-years, thus the `person.years` should be divided by `deaths` and then the death rate can be calculated as this ratio $\times 100,000$. For choices with offsets mentioned, either the ratio is not correct or it doesn't worth being transformed given the magnitude is not significant.

Question 10

1 pts

Within a causal diagram consider the following possible pathways from variable A to variable C:

1. $A \leftarrow B \rightarrow C$
2. $A \rightarrow B \rightarrow C$
3. $A \rightarrow B \leftarrow C$

Now consider fitting a regression model where the variable B is included as an explanatory variable. What would the status of the pathway be for each of these scenarios?

- ☐ The pathway would be closed for 1 and it would be open for 2 and 3.
- ☐ The pathway would be closed for 2 and 3 and it would be open for 1.
- ☐ The pathway would be closed for 2 and it would be open for 1 and 3.
- ☐ The pathway would be open for all three scenarios.
- ☐ The pathway would be closed for all three scenarios.
- ☐ The pathway would be closed for 1 and 3 and it would be open for 2.
- ☒ The pathway would be closed for 3 and it would be open for 1 and 2.
- ☐ The pathway would be closed for 1 and 2 and it would be open for 3.

G

Comments:

(Assume it is to explain effects on variable C).

If the causal pathways represent interest associations, they should typically remain open in a statistical analysis. Therefore, 1 and 2 should be open. For 3, since no association can be transmitted along it, it should be closed.