Department of Statistics

Quaid-i-Azam University, Islamabad

Linear Models Dr.Zahid Asghar

Note: Attempt all the questions. Question Number 1 has 15 marks.

Q#1

In the expression Sales ≈ f(TV, Radio, Newspaper), "Sales" is the:

Response

Training Data

Independent Variable Feature

While doing a homework assignment, you fit a Linear Model to your data set. You are thinking about changing the Linear Model to a Quadratic one. Which of the following is most likely true:

Using the Quadratic Model will decrease your Irreducible Error.

Using the Quadratic Model will decrease the Bias of your model.

Using the Quadratic Model will decrease the Variance of your model

Using the Quadratic Model will decrease your Reducible Error

A hypercube with side length 1 in d dimensions is defined to be the set of points (x1, x2, ..., xd) such that0≤*xj*≤1 for all j = 1, 2, ..., d. Define the boundary region of the hypercube to be the set of all points such that there exists a j for which 0≤*xj*≤.05 or .95≤*xj*≤1 (i.e., it is the set of all points that have at least one dimension in the most extreme 10% of possible values). What proportion of the volume of a hypercube of dimension 50 is in the boundary region? Please give your answer as a value between 0 and 1 with 3 significant digits. If you think the answer is 50.52%, you should say 0.505: (2 points)

True or False: A fitted model with more predictors will necessarily have a lower Training Set Error than a model with fewer predictors.

Why is linear regression important to understand? Select all that apply:

Top of Form

a) The linear model is often true

Linear regression is very extensible and can be used to capture nonlinear effects

Simple methods can outperform more complex ones if the data are noisy

Understanding simpler methods sheds light on more complex ones

Which of the following are true statements? Select all that apply:

Top of Form

A 95% confidence interval is a random interval that contains the true parameter 95% of the time

If I perform a linear regression and get confidence interval from 0.4 to 0.5, then there is a 95% probability that the true parameter is between 0.4 and 0.5.

The true parameter (unknown to me) is 0.5. If I sample data and construct a 95% confidence interval, the interval will contain 0.5 95% of the time.

Which of the following indicates a fairly strong relationship between X and Y?

R2=0.9

The p-value for the null hypothesis β1=0 is 0.0001

The t-statistic for the null hypothesis β1=0 is 30

What is the difference between lm(y ~ x\*z) and lm(y ~ I(x\*z)), when x and z are both numeric variables?

The first one includes an interaction term between x and z, whereas the second uses the product of x and z as a predictor in the model.

The second one includes an interaction term between x and z, whereas the first uses the product of x and z as a predictor in the model.

The first includes only an interaction term for x and z, while the second includes both interaction effects and main effects.

The second includes only an interaction term for x and z, while the first includes both interaction effects and main effects.

9. What is the difference between lm(y ~ x\*z) and lm(y ~ I(x\*z)), when x and z are both numeric variables?

a) The first one includes an interaction term between x and z, whereas the second uses the product of x and z as a predictor in the model.

b) The second one includes an interaction term between x and z, whereas the first uses the product of x and z as a predictor in the model.

c) The first includes only an interaction term for x and z, while the second includes both interaction effects and main effects.

d) The second includes only an interaction term for x and z, while the first includes both interaction effects and main effects.

10. Suppose we collect data for a group of students in a statistics class with variables X1= hours studied, X2= undergrad GPA, and Y= receive an A. We fit a logistic regression and produce estimated coefficients β^o=−6, β^1=0.05,β^2=1.Estimate the probability that a student who studies for 40h and has an undergrad GPA of 3.5 gets an A in the class:

Which of the following is NOT a linear function in x:

f(x)=a+b2x

The discriminant function from LDA

δk(x)=xμkσ2−μ2k2σ2+log(πk)

logit(P(y=1|x)) where P(y=1|x) is as in logistic regression

P(y=1|x) from logistic regression