PRINCIPLES OF STATISTICAL MODELLING
MINIQUIZ 3

NOVEMBER 15, 2016
JACOBS UNIVERSITY BREMEN

DUE IN:: Tuesday, 15.11.2016 at 23:59,
HOW:: electronically in pdf-format via submission to `www.turnitin.com`
Class id:: 13900910
enrollment password:: 20SigmaCV16
Please register for the class on turnitin ahead of time.
GROUP WORK:: You are encouraged to work in teams of two – but no larger. Only one solution is accepted and graded per group. Please include the names of all group members on the assignment.
FORMAT:: Please do the required analyses and provide answers in complete sentences. Just report those statistics that are relevant; do not copy complete software output. Integrate requested figures or tables into your document and give a brief verbal comment/caption on them.

House Prices in King County, WA

Economic theory tells us that house prices are based on a variety of features. The file `kc_house_data.csv`) contains a data set with house sale prices for homes in King County, WA that were sold between May 2014 and May 2015. King County has its seat in Seattle and it is the most populous county in Washington, and the 13th-most populous in the United States.

The data set comprises 19 different features. The general goal is to predict house prices (`price`) using all the available predictors, except the case identifiying information (i.e. `id, date`).

Variables: Description of variables:
price : price of home in USD
bedrooms: number of bedrooms in home
bathrooms: number of bathrooms in home
sqft_living: living aerea (in sq.ft)
sqft_lot: lot size of the house (in sq.ft)
floors: number of floors
waterfront: Waterfront dummy variable (= 1 if home is at Waterfront; 0 otherwise)

view: Scenic view dummy variable (= 1 if home has a scenic view; 0 otherwise)
condition: condition of home
grade: Classification by construction quality which refers to the types of materials used and the quality of workmanship, higher grade = higher quality
sqft_basement: size of the basement
sqft_above: sqft_above = sqft_living - sqft_basement
yr_built: year in which house was built
yr_renovated: year in which house was renovated for the last time, '0' indicating that no major renovation took place
zipcode: ZIP code
lat: geographic latitude of location
long: geographic longitude of location
sqft_living15: the average house square footage of the 15 closest houses
sqft_lot15: the average lot square footage of the 15 closest houses

(1) (10 points) First compute summary statistics for house prices. The summary must include the following statistics: minimum, maximum, mean, median, standard deviation. Report these summary statistics.

(2) Split the data set into training and evaluation set by using an 80/20 split.
   (a) (1 point) How many cases are in your training data set?
   (b) (1 point) How many cases are in your evaluation data set?
   (c) (3 points) Compute and report summaries as defined above for house prices in both your training and evaluation data set.
   (d) (20 points) Fit a multiple regression model to predict the response using all of the predictors except `id` and `date`. Describe your results. How good is the model? For which predictors can we not reject the null hypothesis $H_0 : \beta_j = 0$?
   (e) (5 points) Starting with the null model as lower boundary and having the above model as upper bound, use the automatic forward/backward selection method to derive the "best" model. Report the significant predictors and the AIC score of this model.
   (f) (10 points) Draw scatterplots of the residuals in the "best" model against each of the predictors. Add the partial regression line (i.e. the line defined by $\hat{y} = \hat{\beta}_0 + \hat{\beta}_j x_j$) and some local smoothing line to see whether there are non-linear effects to consider.
   (g) (20 points) Select three predictors for which non-linear effects seem to be existent. For each of them separately, add polynomials up to degree 5 to the "best" model. Decide on the order of the three predictors in which they should be added. Run 5-fold cross validation to evaluate which higher order polynomial of the three predictors should be included in your final model. Plot the mean squared errors of your cross-valdiation results against complexity of the model. Use MSE and the one-standard-error rule to decide which higher order polynomials to include.

CONCEPTUAL QUESTIONS

(3) (15 points) Suppose we have a data set with five predictors, $X_1 = GPA, X_2 = IQ, X_3 =$ Gender (1 for Female and 0 for Male), $X_4 =$Interaction between GPA and IQ, and $X_5 =$Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = 10$.

(a) Which answer is correct, and why?

(i) For a fixed value of IQ and GPA, males earn more on average than females.

(ii) For a fixed value of IQ and GPA, females earn more on average than males.

(iii) For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

(iv) For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

(4) (15 points) We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

(a) What is the probability that the first bootstrap observation is not the jth observation from the original sample? Justify your answer.

(b) What is the probability that the second bootstrap observation is not the jth observation from the original sample?

(c) Argue that the probability that the jth observation is not in the bootstrap sample is $(1 - 1/n)^n$.

(d) When $n = 5$, what is the probability that the jth observation is in the bootstrap sample?

(e) When $n = 100$, what is the probability that the jth observation is in the bootstrap sample?

(f) When $n = 10000$, what is the probability that the jth observation is in the bootstrap sample?

(g) Create a plot that displays, for each integer value of n from 1 to 100,000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

(h) Now investigate numerically the probability that a bootstrap sample of size $n = 1000$ contains the jth observation. Let $j = 7$. Repeatedly create bootstrap samples (a total of 20 000), and each time record whether or not the seventh observation is contained in the bootstrap sample. What ratio of the 20 000 bootstrap samples contains the seventh observation?