Exam 1

Subrata Paul

9/23/2020

Problem 1 (5 points)

Write down a linear regression model with assumptions.

Problem 2 (5 points)

In fitting a simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$, it was found that observation Y_i fell directly on the fitted regression line. If this case were deleted, would the least square regression line fitted on the remaining $n - 1$ cases be changed? [Hint: try to use the function that we minimize in the least square procedure.]

Problem 3(a) (10 points)

In this problem, you will simulate data with 5000 observations. About 50% of them are male. Use a binomial distribution to choose the number of males randomly (Hint: You are flipping a fair coin and counting the number of heads). Call the variable `Gender`. Diastolic blood pressure of male and female follows a normal distribution with mean $\mu_{\text{male}} = 82$, $\mu_{\text{female}} = 80$ mmHg and standard deviation $\sigma_{\text{male}} = \sigma_{\text{female}} = 10.5$. Total cholesterol in blood follows a normal distribution with a mean of 5.69 and variance 1.31. Glucose follows a normal distribution with a mean 5.12 and a standard deviation of 1.24. Gender, cholesterol, and glucose are predictor variables. The response variable is BMI. The error term, ϵ $N(0, 9)$ accounts for the randomness and effect of other factors that affect BMI. The mean BMI while all the predictors are zero is 23. Simulate BMI so that the effect sizes (regression coefficients) of Gender (Male), blood pressure, cholesterol, and blood glucose are 0.01, 0.07, 0.1, and -0.1, respectively. Run a multiple linear regression and discuss if the regression model could identify the simulated relationship. You should also discuss if you find an estimated coefficient for a variable that is much different than the parameter used in the simulation.

Problem 3(b) (2 points)

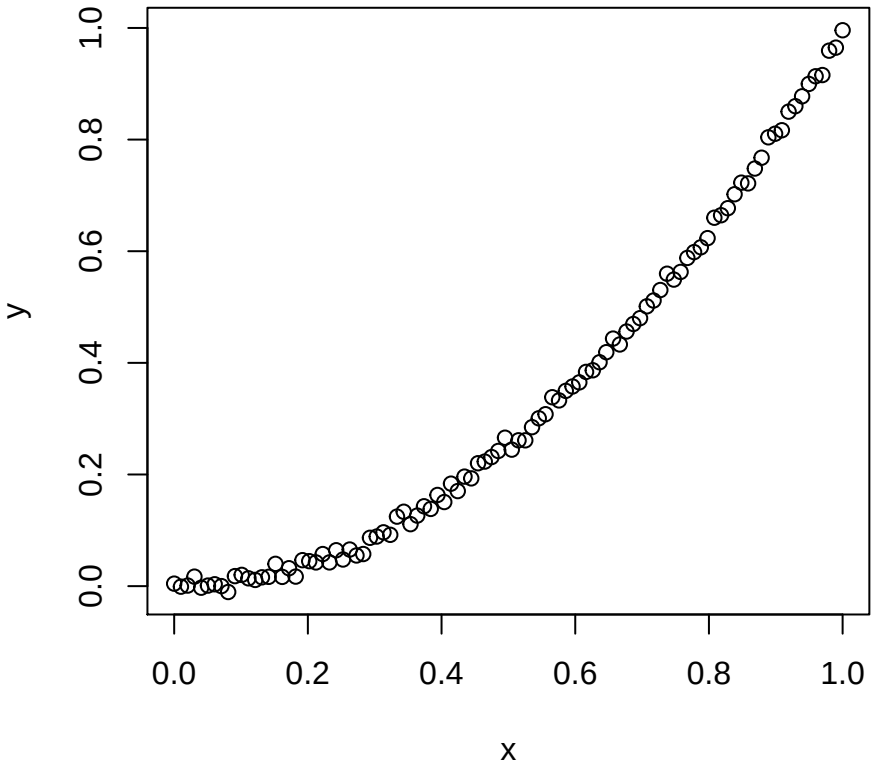
Run the same analysis as problem 3(a) multiple times. Do you get the same or different estimates? Why?

Problem 4 (10 points)

a. What does it mean for a regression model to be a linear model? (Specifically, explain what linear model means in the context of a regression model.)

b. Consider a setting where there are four observations ($n = 4$) and two predictors ($p = 3$). Construct a 4×3 design matrix X that would lead to an unidentifiable model but where no two columns are identical.

Consider the figure below for parts (c) and (d) of this question.



(c) Is the relationship between x and y linear? Why?

(d) Explain how the relationship between y and x can be approximated reasonably well by a linear model.

Problem 5 (8 points)

Consider the model

$$\log(\text{ppgdp}) = \beta_0 + \beta_1 \text{fertility} + \beta_2 \log(\text{pctUrban}) + \epsilon$$

You can find the description of the data and the variables using `?alr4::UN11`. Fit the model, print the summary of the model and, interpret the coefficient of `pctUrban`.

You will not get full credit for using generic terms or variable names like fertility or pctUrban. Clearly indicate what these variables are measuring/representing.

Problem 6 (10 points)

Assume that the observations for the response variable are correlated i.e. $\text{cov}(y_i, y_j) \neq 0$. So the variance-covariance matrix $Var(\epsilon) \neq \sigma^2 I$, where σ is a constant and I is the identity matrix. Instead assume that $Var(\epsilon) = \sigma^2 I + \gamma^2 K$, where K is not a diagonal matrix. How does this phenomena effects the estimates $\hat{\beta}$. (More specifically is $E[\hat{\beta}]$ and $Var(\hat{\beta})$ in this case and how they vary from that under usual linear regression model assumption?)

Problem 7 (10 points)

Consider a simple linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

with usual notations and assumptions.

a. How does the parameter β_0 and β_1 chaanges if we center the predictor variable X (i.e. substract $\overline{X}$ from X).

b. How do the parameters changes if we scale the predictor variable X (i.e. divide X by its standard deviation?)

c. If X and Y are uncorrelated what can be said about β_0 and β_1?

Problem 8 (10 points)

Download the data `simu_exam1.txt` from the canvas. Fit a multiple linear regression model with Y as the response variable and $x1, x2, x3$ as predictors (just one model with three predictors). Perform model diagnostic for structure. If there are issues, suggest a model that is more appropriate for the data. Give the coefficients of the final model and interpret them.

Problem 9 (20 points)

Select data from http://archive.ics.uci.edu/ml/datasets.php. On the left sidebar select `Regression`, `Numerical`, and `Multivariate`. You can choose any data from the list that has `Default Task = Regression` and the number of instances more than 500. Do not select data that has `Time Series` in the `Default Task` column. You will describe the data, run an appropriate multiple linear regression model, perform diagnostic for model structure, and transform variable if appropriate, and at the end interpret your result.

Problem 10 (MATH 5387 only)

Consider the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j + \cdots + \beta_{p-1} X_{p-1}$$

Show that the OLS linear fit to the data in an added variable plot for predictor x_j will have slope β_j and intercept 0.