Homework 2

36-401, Modern Regression, Fall 2015

Due at the start of class, 17 September 2015

AGENDA: EDA; Simple Linear Regression Models; Simulation
Note: The problem set is scored out of 100 points; the problems add up to 90 points; the remaining ten points will be graded according to a writing rubric, given at the end of the assignment.

1. (20) Consider a regression model, obeying equation (1.1) from the textbook with $\beta_0 = 20, \beta_1 = 5, \sigma^2 = 3$.

   (a) (5) Write the equations of the regression model, including all the assumptions.

   (b) (5) A new observations Y is made at $X = 4$. Can we know the exact probability that Y is greater than 46? If yes, what is that probability? If not, why not?

   (c) (5) Further studies suggest that a normal error regression model, the textbook's (1.24), is appropriate. What are the distributions of Y at $X = 3$ and $X = 5$? Could you plot the regression model in the fashion of Figure 1.6?

   (d) (5) Under the assumptions of part (c), can you know the exact probability that Y is greater than 46 when $X = 4$? If yes, what is that probability? If not, why not?

2. (50) The data set `chicago`, in the package `gamair`, contains data on the relationship between air pollution and the death rate in Chicago from 1 January 1987 to 31 December 2000. The response variable of interest is `death`, the total number of non-accidental deaths each day. The other variables in the data set are `time`, recorded in days before or after 31 December 1993, and five possible predictor variables:

   - `pm10median`: the median density over the city of large pollutant particles
   - `pm25median`: the median density of smaller pollutant particles
   - `o3median`: the median concentration of ozone (O_3) in the air
   - `so2median`: the median concentration of sulfur dioxide (SO_2) in the air

- `tmpd`: the mean daily temperature.

We will use this problem to practice EDA and setting up regression models.

(a) (5) Run `summary` on each variable. (You will need to load the data first, and probably install the `gamair` package before that.)

(b) (15) *Examining the variables*

i. (2) Is temperature given in degrees Fahrenheit or degrees Celsius? How can you tell?

ii. (2) We will ignore the `pm25median` variable in the rest of this problem set. Why is this reasonable?

iii. (2) Report the mean, variance and median of each variable.

iv. (9) Create a histogram (labeled and titled) for each variable with an appropriate numbers of bins. (*Hint:* use `par(mfrow=c(2,3))` to put all plots on the same graph page.) Describe each distribution in words — location of the central tendency, skew, shape, symmetry, tails, etc. Note any potential outliers.

(c) (15) Plot each predictor variable (x-axis) against the response variable (y-axis) in a labeled, titled scatterplot. Use different plotting symbols for each graph. (You should have four graphs in all, and can use `par(mfrow=c(2,2))` again.)

i. (5) Describe each bivariate relationship between a predictor and the response. Is there one to speak of? If so, does it look linear, nonlinear, positive, negative ...?

ii. (5) Are there any outliers in each of the plot? If there are, which days are outliers? (For full credit, use `time` to give calendar dates, not day numbers.) Do the different plots share outlier days?

iii. (5) For each predictor variable, does a linear regression of the number of deaths on the predictor seem appropriate? (That is, for which variables do the modeling assumptions seem to hold?) Which predictor variable do you think would be the most appropriate? Why?

(d) (15) We will take a closer look at the relationship between `death` and `tmpd`. Someone proposes that the relationship follows a normal error linear regression model with $\epsilon \sim N(0, 14.2^2)$ and the true regression function $\mathbb{E}[Y|X = x] = 130 - 0.28x$.

i. (3) Write the theoretical regression model in context; include all assumptions.

ii. (3) Explain the interpretations of the proposed β_0, β_1 coefficients in this context.

iii. (3) Add the proposed function to the scatterplot of `death` and `tmpd`. Do you agree with the normal error regression model assumption? Why / why not?

iv. (3) Find the predicted change in the number of deaths from a 2 C° degree warming over the course of a whole year (not per day).

v. (3) Can we claim the relationship between temperature and deaths is causal? Explain.

3. (20) *Simulating from a Regression model*

   (a) (5) Generate 50 equally spaced values of X from 2 to 30. Simulate values of Y from a regression model with $\beta_0 = 21, \beta_1 = -3$, and $\epsilon \sim N(0, 25)$ (*Hint: Use* `rnorm()`).

   (b) (5) Plot the pairs of $(X_i,\, Y_i)$ values as black solid squares and add the underlying regression line (use `lwd=2, col="red"`).

   (c) (5) Calculate the sample mean $\bar{X}$ and the corresponding predicted value of $\mathbb{E}\left[Y|X = \bar{X}\right]$. Add this point as a red filled circle to the plot. (*Hint:* `points()`).

   (d) (5) Generate a second set of responses from a model with the same underlying function, but $\epsilon \sim N(0, 100)$. Make a scatter plot. Compare the two graphs and describe what are similar and what are different.

RUBRIC (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is either properly integrated with a tool like R Markdown or knitr, or included as a separate R file. In the former case, both the knitted and the source file are included. In the latter case, the code is clearly divided into sections referring to particular problems. In either case, the code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output. For full credit, all code runs, and the Markdown file knits (if applicable).