UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Basic Exam - Applied Statistics
January 2023

Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level. The number of points for each part of each question is listed inline with the part.

Question:	1	2	3	4	5	Total
Points:	22	20	20	20	18	100
Score:						

1. A company is spends X_{1i} dollars on web-based advertising and X_{2i} dollars on print advertising in year i. The company is interested in studying the effectiveness of advertising in increasing annual sales, Y_i.

   (a) (4 points) You fit the linear regression model: $Y_i\ X_{1i} + X_{2i}$. This model depends on the usual assumptions of linear regression. Which two of these assumptions are you most concerned about? Briefly, describe why or what you are concerned about.

   (b) (4 points) For each of the assumptions you identified in the previous part, describe how you would check that assumption. For each, either sketch a diagnostic or give numerical values that would signal the assumption failed.

   (c) (3 points) For one of the assumptions you have been discussing, describe how you would change the model if that assumption proved false.

   (d) (3 points) Describe one way you might use a lagged variable (Y_{i-1},X_{1i-1}, or X_{2i-1}) to address a question of interest with these data.

   Suppose now that you have these data for J different companies in the same industry, such that Y_{ij} is the annual sales of the j^{th} company in the i^{th} year. You wish to model the impact of advertising strategy on sales.

   (e) (4 points) One concern you have about these data is that the companies are of different sizes. A bigger company might spend more on advertising and have more sales, but you don't want to just measure size. Briefly describe one approach to model these companies of different sizes in the same model.

   (f) (4 points) Write the form of a mixed effect model you could use to address this question. Be sure to specify the random effect(s) used and be clear which observations share common random effects.

2. You must construct a sampler for a complex distribution that samples from the following process. Draw a random number uniformly between 0 and 1. If it exceeds a threshold, a, then keep it. If not, you must re-draw a uniform number and keep the result.

   (a) (5 points) Write `python` or `R` function that returns a sample from the process and takes as input the parameter a.

   The python function `random.randrange(start, stop)` returns a uniform random number between start and stop. The R function `runif(1, min, max)` returns a single uniform random number between min and max.

   (b) (10 points) Derive the probability density function (pdf) of the distribution for the process and sketch a plot of the histogram. Prove that the function is a valid pdf.

   (c) (5 points) Suppose one person (Player I) draws a sample from the distribution with threshold a and a second person (Player II) draws a sample from the distribution with threshold b. Player II wins the game if their draw is greater than Player I and vice versa. You can ignore draws. Write a program that simulates 10,000 plays of the game (Player I and Player II drawing one sample) and produces the empirical chance that Player II wins.

3. Suppose that a researcher comes to you once a day for a year ($n = 365$ days) with an offer of a hypothesis to test. You can choose to test the hypothesis or not test the hypothesis, but once you have made your decision, you cannot go back and recall a passed test. You are able to rank the hypotheses from best to worst (based on your own assessment of the value of the question underlying the hypothesis) The order the researcher presents the hypotheses is random. You may only choose one hypothesis to test for the year. Your goal is to select the best hypothesis for testing.

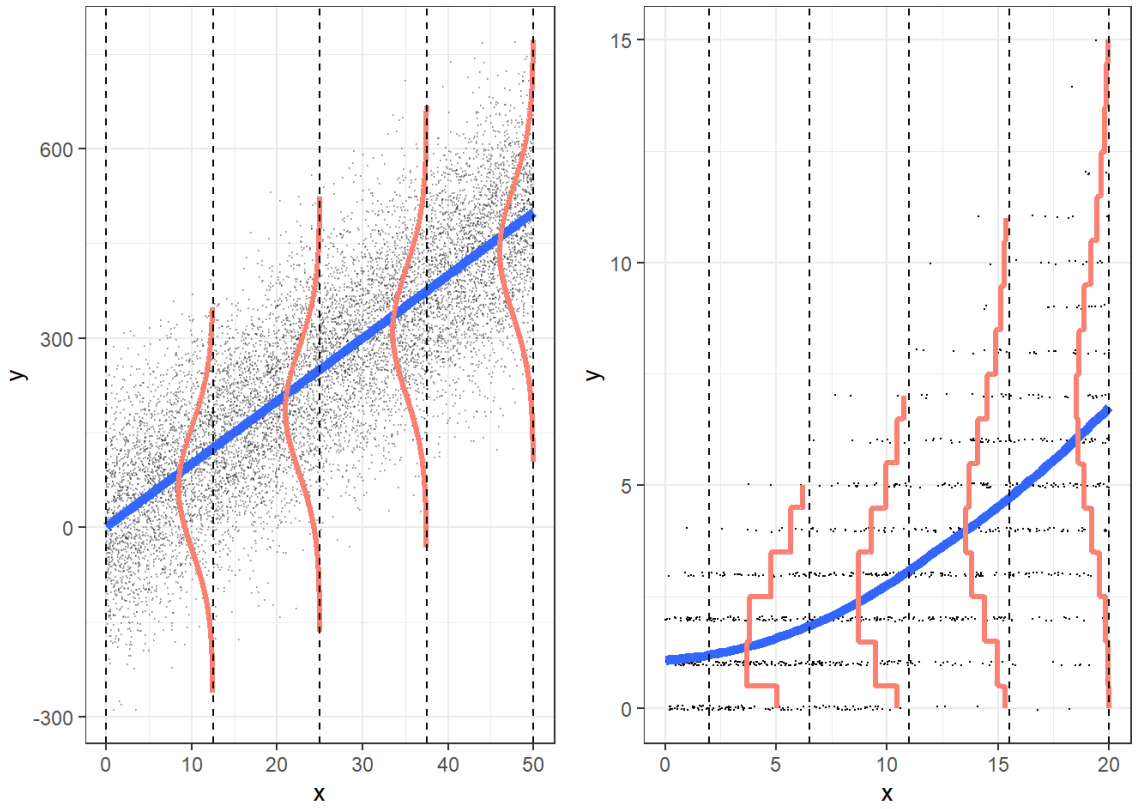
   The list of values of the hypotheses is stored in a list constructed as

   `h = list(numpy.random.permutation(365))` which is a randomly permuted list containing the integers from 0 to 364. The "best" hypothesis has value 364.

   (a) (10 points) **Simulation** Suppose that your rule for selecting a hypothesis to test is that you reject r hypotheses and test the hypothesis with the best ranking relative to the r that have already been observed. Write a `python` function to implement this decision rule. The function should return the value in h that corresponds to the hypothesis to be tested (equivalently, the day to accept the hypothesis proposal from the researcher). The function should take as inputs the list h and the threshold r. Remember, you can only look at one value of h at a time and it must be in the sequence given in h.

   (b) (10 points) **Optimize** r Given the function you wrote in the previous part, write a program that scans across r from 100 to 199. For each value of r the program should simulate 1000 runs for a fixed value of r and compute the average value of the accepted hypothesis. The result should be a list of length 100 where each element is the average value. Use `h = list(numpy.random.permutation(365))` to randomly permute the list for each of the simulation runs. The function `numpy.mean(a)` computes the average for an array-like a.

4. This graphic, from Roback and Legler, 2021, compares linear and Poisson regression.



(a) (5 points) Please specify which side is linear regression and which is Poisson regression and why.

(b) (5 points) Which part of each plot represents each of the following:

- The mean model
- The conditional distribution of Y
- The data

And briefly describe how each of these features differs between linear and Poisson regression.

These plots consider a single continuous regressor, X. Suppose there is an additional binary regressor, Z.

(c) (5 points) Sketch a version of the linear regression plot that shows additive effects of both X and Z on Y. Make the labels clear enough to read, and specify the full mean model including coefficients.

(d) (5 points) Sketch a version of the linear regression plot that shows additive effects of X and Z on Y, as well as an interaction effect of X and Z. Make the labels clear enough to read, and specify the full mean model including coefficients.

5. The Kentucky Derby is a major annual horse race. Consider the time of the winning horse each year since 1896 (called `speed`). We will consider the pattern of winning times over years (called `year`, and re-coded as $yearnew = year - 1896$, $yearnew2 = yearnew^2$). The last model also considers a binary variable `fast` which indicates whether the track conditions were fast that day. Consider the below model fits and plots:

```
model1 <- lm(speed ~ year, data = derby.df)

##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  2.05347   4.543754  0.4519 6.521e-01
## year         0.02613   0.002322 11.2515 1.717e-20

##  R squared =  0.5134
##  Residual standard error =  0.9032

model2 <- lm(speed ~ yearnew, data = derby.df)

##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept) 51.58839   0.162549  317.37 2.475e-177
## yearnew      0.02613   0.002322   11.25  1.717e-20

##  R squared =  0.5134
##  Residual standard error =  0.9032

model2q <- lm(speed ~ yearnew + yearnew2, data = derby.df)

##               Estimate Std. Error t value   Pr(>|t|)
## (Intercept) 50.5874566  2.082e-01 243.010 2.615e-162
## yearnew      0.0761728  7.950e-03   9.581  1.839e-16
## yearnew2    -0.0004136  6.359e-05  -6.505  1.921e-09

##  R squared =  0.641
##  Residual standard error =  0.779

model5 <- lm(speed ~ yearnew + fast + yearnew:fast,
             data=derby.df)

##              Estimate Std. Error t value   Pr(>|t|)
## (Intercept)  50.52863   0.205072 246.394 6.989e-162
## yearnew       0.03075   0.003471   8.859  9.839e-15
## fast          1.83352   0.262175   6.994  1.730e-10
## yearnew:fast -0.01149   0.004117  -2.791  6.128e-03

##  R squared =  0.7068
##  Residual standard error =  0.7071
```

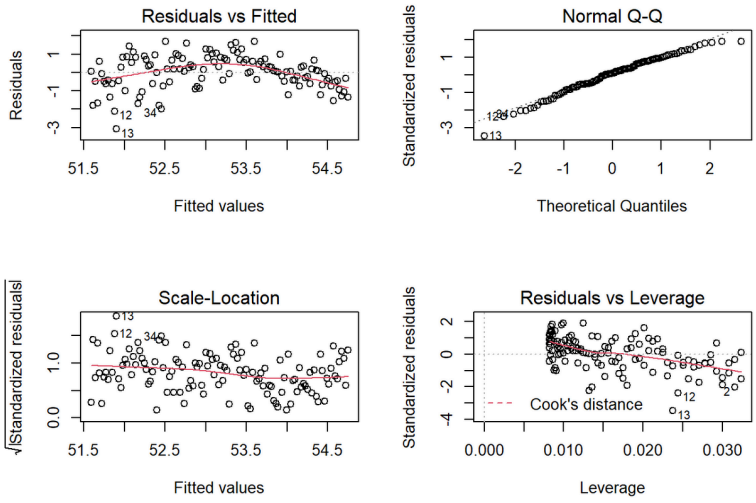


Figure 1.6: Residual plots for Model 2.

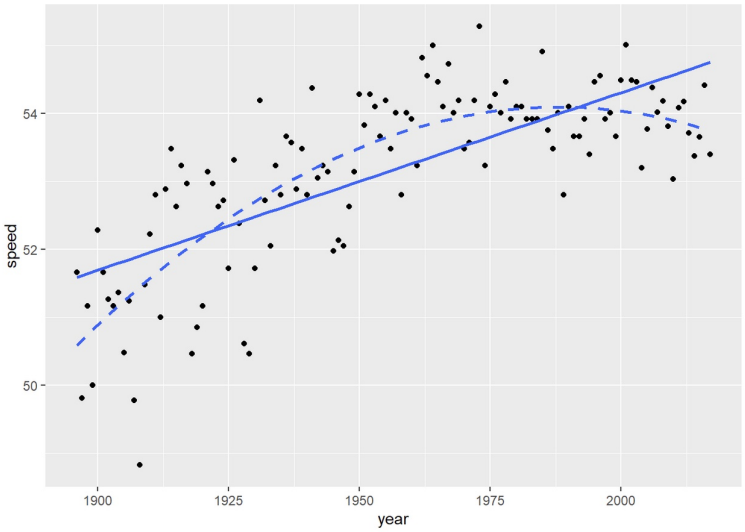


Figure 1.7: Linear (solid) vs. quadratic (dashed) fit.

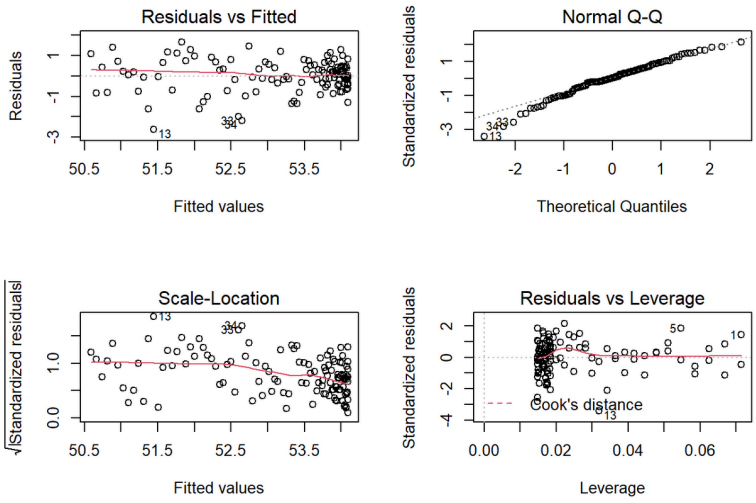


Figure 1.8: Residual plots for Model 2Q.

(a) (3 points) What changed between Model 1 and Model 2. What stayed the same? Which of these models would you prefer to use? Why?

(b) (5 points) What changed between Model 2 and Model 2q? Which of these models would you prefer, and why? Be sure to reference features you see in the plots.

(c) (5 points) For Model 5, sketch the model fit. Write down the mean model for each of the levels of `fast`.

(d) (5 points) Based on the information provided, do you know which model you would prefer as your final model? If yes, briefly explain which model and why. If not, briefly explain how you would gather more information and choose your final model.