18.650 PROBLEM SET 3, DUE WEDNESDAY, SEPT. 30, 2015

in class, or by 4 P.M. at E18-412

This PS will include various regressions done on the data set, a "data frame" in R, "kyderb40-15", which gives the winning times, in seconds, in the Kentucky Derby (horse race) from 1940 through 2015. The data frame is on the course website. Make a folder (subdirectory) in which you will do R work and on which R is available (as in Athena, or in case you've downloaded R from some CRAN website onto your own laptop). Read in the data frame by a command: whatevername = read.table("kyderb40-15"). There are 76 rows for 1940 through 2015. The row numbers are 1 through 76, and there are just two columns of data, for years and seconds. For any given data frame say "oldframe" with N rows and $1 \leq j \leq k \leq N$ you can extract a subframe having just rows j through k by newframe = oldframe[j:k,]. Or to extract a column, say the ith column, one would do for example vec = oldframe[,i]. in the current case you might call column 1 "years" and column 2 "times" or "seconds".

To do y-on-x simple linear regression in R for a vector y on a vector x with the same number of components, one first creates a "regression object" by someobj = lm(y∼ x) where "lm" abbreviates "linear model". Then if you give the command "summary(someobj)" you will see output such as

```
  ............
  Call: lm(formula = y∼ x)
Residuals:
Min  1Q   Median 3Q   Max
-8   -6   -2       4   12

Coefficients:
             Estimate Std. Error  t value  Pr(> |t|)
(Intercept)  -22.0000   5.5498    -3.964   0.00415   **
x             11.0000   0.8944    12.298   1.78e-06  ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Date: 18.650,

```
Residual standard error: 8.124 on 8 degrees of freedom
Multiple R-squared: 0.9498, Adjusted R-squared: 0.9435
F-statistic: 151.3 on 1 and 8 DF, p-value: 1.778e-06
........
```

How to interpret this? A regression line $y = a+bx$ was fitted to some data. The “Intercept” a, the value given by the model when $x = 0$, is estimated as -22, and the slope b is estunated as 11. The hypothesis that $a = 0$ is rejected, with a p-value of .00416, so the intercept is significantly different from 0. The hypothesis that the slope $b = 0$ is (strongly) rejected, with a p-value of $1.78\cdot 10^{-6}$. The p-value for each coefficient is the probability in the next to last column for that coefficient. The last column is *** if the p-value is less than 0.001, or ** if it is not but is less than 0.01, or * if it is not but is less than 0.05. If none of these occurs, in other words the last column is “.” or blank, we would not reject the hypothesis that the given coefficient is 0 and would provisionally accept it.

1. (20 points) A letter to *Science* magazine, issue of Aug. 8, 2014, said that “Kentucky Derby winning times between 1950 and 2012” show “no significant increase in speed” (i.e. decrease in winning times). The issue is, selective breeding of race horses is intended to increase their speeds, but did this succeed over the given time span? Select the data from 1950 through 2015 from the provided data set (which begins with 1940) and check the statement.
(a) What is the p-value for the slope (coefficient of “years”) in the regression of winning times (seconds) on years in the 1950-2015 range? Is the slope significantly different from 0?
(b) What is the p-value for the intercept? Is the intercept significantly different from 0?
(c) Based on parts (a) and (b), what model, for prediction of future winning times, could we provisionally accept?

2. Now consider the entire data set from 1940 through 2015. Answer parts (a) and (b) of the preceding problem for this whole data set. Also do a quadratic regression: if winning times (seconds) are being regressed on years, the R command for creating the regression object would be something of the form

```
qobj = lm(seconds ~ years + I(years^2))
```

(I(·) is the identity function, but R requires it to be written in this case.) Is the coefficient of the quadratic (squared) term significantly different from 0? If so, what does that tell us about the simple linear regression model in this case?

3-4. Give answers to problem 1(a) and (b) for the years (i) 1940-1949, (ii) 1940-1959, (iii) 1940-1969.
(iv) Among the following possible choices (u), (x), (y), which do you think is preferable, for fitting the data 1940–2015 and predicting future times? All models will include some random errors. For the other two models which you do not choose, explain why not.
(u) A model of constant times (no slope);
(x) A linear model with non-zero intercept and slope;
(y) A model with a "change point", using a model of form (u) or (v) (which?) up to some year, and a different model of one of those forms (which?) after that year.

5. There is a data set "quaddata" on the course website. It is a 2 by 10 matrix of numbers. As described in the first paragraph of this pset, read it into R, calling it qdd (a "data frame"). Find the corresponding 2 by 10 matrix by mqd = as.matrix(qdd). Define a vector x with 10 components as the first row, y as the second row. Form the quadratic regression object

qobj = lm(y∼ x + I(x^2))

where I(·) is the identity function but R requires it to be written. Which of the coefficients: (a) the intercept, (b) the coefficient b of x, and (c) the coefficient, say c, of x^2, are significantly different from 0 at the $\alpha = 0.05$ level?
(d) Consider the simple linear regression model

$$Y_j = a + bx_j + \varepsilon_j$$

where ε_j are i.i.d. $N(0, \sigma^2)$ for some unknown $\sigma > 0$. Do the data in "quaddat" fit with that model? Why, or why not?