Solutions to 18.443 PS3, due Wednesday Sept. 29, 2015

1. To get years 1950 through 2015, one should take rows 11 through 76 of the given data set from 1940 through 2015.
(a) (7 points) The p-value for the slope is 0.709 (rmd draft), or 0.453 (TA). Either way, this is not (is far from being) less than 0.05, so the slope is not significantly different from 0, confirming the finding in the letter to *Science*.
(b) (7 points) The p-value of the intercept is 2.2e-11 (rmd draft) or 1.4e-10 (TA), eithern way extremely small, so the intercept is (highly) significantly different from 0. (A model saying the race could be run instantly or in negative time is not reasonable.)
(c) (6 points) By part (a), one would omit the slope from the regression and model the data as a constant plus random error i.i.d. $N(0, \sigma^2)$. The constant is not necessarily close to the estimated intercept in the original regression, which is affected by the estimated slope.

2. (a′) (6 points) The p-value for the slope now is $0.00034 < .01 < .05$ (rmd draft) or 0.0115 ¡ 0.05 (TA), so the slope is significantly different from 0, starting in 1940. Its estimated value is negative, indicating that in 1940-1949 times were relatively slow.
(b′) (6 points) The p-value of the intercept is less than 2e-16, again tiny, certainly less than 0.05, so the intercept is again significantly different from 0.
(c′) (8 points) The quadratic regression gives p-values less than .001 for all three coefficients, specifically including the coefficient of $(\text{years})^2$, with p-value 0.00034. This means we can reject the simple linear regression hypothesis, which is the special case of the quadratic regression in which that coefficient is 0. (This does not mean by the way that the quadratic regression itself is a correct model.)

3-4 (40 points total) (i) (8 points) For 1940–49, neither coefficient is significantly different from 0. (Not surprising, as $n = 10$ is a rather small data set.)
(ii) (8 points) For 1940–59, the intercept is significantly different from 0 (p-value = .0115) but the slope is not (p-value 0.0796) at the .05 level.
(iii) (8 points) For 1940–1969, the slope is (highly) significantly different from 0 (p-value $1.41 \cdot 10^{-6}$) and so is the intercept (p-value 0.000294). (This is about same result, both significant, as for the entire span of years 1940–2015 in Problem 2, so apparently the first 30 years is a long enough sample

to show that.)
(iv) (16 points, divided 5+5+6 for the three alternatives as follows):
(u) (5 points) Does not fit the data. The model constant plus random error was rejected in Problem 2 (a'), where the slope was found significantly different from 0.
Answer (x) to part (iv)) (5 points): Does not fit the data. The simple linear regression model was rejected in problem 2 (c').
(answer (y) to part (iv)) (6 points) Such a piecewise model does fit the data, but there are different possible choices. Putting the change point between 1949 and 1950, a constant plus random error model from 1940–49 fits those data (part (i)) and one with a different constant from 1950–2014 fits the data for those years (Problem 1) as in both cases, fitted slopes are not significantly different from 0. Suitable values for the separate constants in the respective periods, minimizing sum of squared errors, would just be the sample means over the periods, not required. But, estimating the constant value as the output intercept in a regression would be incorrect, as that is based on a fit with a non-zero slope, warned against in an email (−2 points for that).

Another possible correct answer is to take a simple linear regression model from 1940 to 1969, as in part (iii), then do a new regression on the data 1970 through 2015 and find that its slope is not significantly different from 0, so to use a constant plus random error from 1970 on. But again, the constant should not be estimated by the regression intercept (−2), which again would be based on a fit with a non-zero slope.

There is an R software package for "change point models" (how best to choose the change point?) but that's a field in itself, seemingly too complicated for a beginning course.
5. The coefficients of x and x^2 are both significantly different from 0, although the intercept is not. The conclusion about the coefficient of x^2 shows that the simple linear regression model does not fit the data.