2.4: Simultaneous Inference and Important Considerations

Dr. Bean - Stat 5100

(Hypothetical) Suppose you are searching for a relationship between a person's genetics and their likelihood to contract SARS-COV-2. You conduct individual t-tests between 1000 prominent genes (expressed vs non-expressed) and SARS-COV-2 infection/non-infection rates and find that 45 genes of them share a significant link with the likelihood of infection. Based on these results, what would you conclude about the 45 genes?

Nothing. At least not from this test. We would have expected around 50 genes to have "significant" results just due to random chance. Without a multiple hypothesis adjustment, these results suggest that there is nothing different in the genetic expression of those who were infected and those who were not infected.

Regression Through the Origin

Sometimes we wish to force the regression line to go through the origin (i.e. the point (0,0)), making the theoretical linear model become

$$Y_i = \beta_1 X_{i,1} + \epsilon_i$$

When might regression through the origin be a good idea?

- When the point (0,0) makes sense in the context of the data.
- When our sample size is small (avoiding an estimate of β_0 saves us one degree of freedom).
- If BOTH of the above conditions are not met, don't bother with regression through the origin.

Cautions for regression through the origin:

- $\sum_i e_i$ not necessarily equal to 0 (residuals might be unbalanced)
- R^2 can be negative, giving it a nonsensical interpretation

The following is a table of SAS output from a linear regression model fit with 16 observations. Observations that normally appear in the table but have been removed are denoted by a period. Please fill in all missing values

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5297.51250	.	.	<.0001
Error	.	.	.		
Corrected Total	15	5443.93750			

Root MSE	.	R-Square	.
Dependent Mean	225.56250	Adj R-Sq	0.9712
Coeff Var	1.43376		

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5297.51250	5297.51250	506.51	<.0001
Error	14	146.42500	10.45893		
Corrected Total	15	5443.93750			

Root MSE	3.23403	R-Square	0.9731
Dependent Mean	225.56250	Adj R-Sq	0.9712
Coeff Var	1.43376		

Extrapolation

Extrapolation in linear regression is the use of your linear model to make a prediction for Y that falls outside the range of observed response variables. When there are multiple X variables, this can include a situation where each individual X variable is technically within the range of observed observations, but the *combination* of X variable inputs falls outside the range of the observed data.

If I have created a linear model for prediction, isn't extrapolation the point?

Slight extrapolations are usually not bad and even desirable. However, we do not know if the observed relationships between variables remains the same in unobserved territory. Abuse of this assumption may cause us to come to unreasonable conclusions.

Please check out the following story available at the following link:

`https://www.nature.com/articles/431525a`

After reading this story:

Does this analysis provide convincing evidence that female Olympic sprinters will overtake male sprinters in 2156? Discuss why or why not.

No. The extrapolation required for this conclusion lies far beyond the range of observed data. We can have no confidence that the trend observed over the past 100 years will continue for the next 150 years.

What other interesting conclusions/observations might be made from these regression models fit to sprinter times?

- The percent variance in run times explained by the line is notably higher for the men than it is for the women.
- The range of available men's running times is notably longer than the availability of women's running times.
- (Possible): Is there anything notable about the times when the run times were notably lower than the prediction from the regression line?

What, if anything, could be done to improve this analysis?

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

The 2004 Olympic women's 100-metre sprint champion, Yuliya Nesterenko, is assured of fame and fortune. But we show here that — if current trends continue — it is the winner of the event in the 2156 Olympics whose name will be etched in sporting history forever, because this may be the first occasion on which the race is won in a faster time than the men's event.

The Athens Olympic Games could be viewed as another giant experiment in human athletic achievement. Are women narrowing the gap with men, or falling further behind? Some argue that the gains made by women in running events between the 1930s and the 1980s are decreasing as the women's achievements plateau[1]. Others contend that there is no evidence that athletes, male or female, are reaching the limits of their potential[1,2].

In a limited test, we plot the winning times of the men's and women's Olympic finals over the past 100 years (ref. 3; for data set, see supplementary information) against the compe-

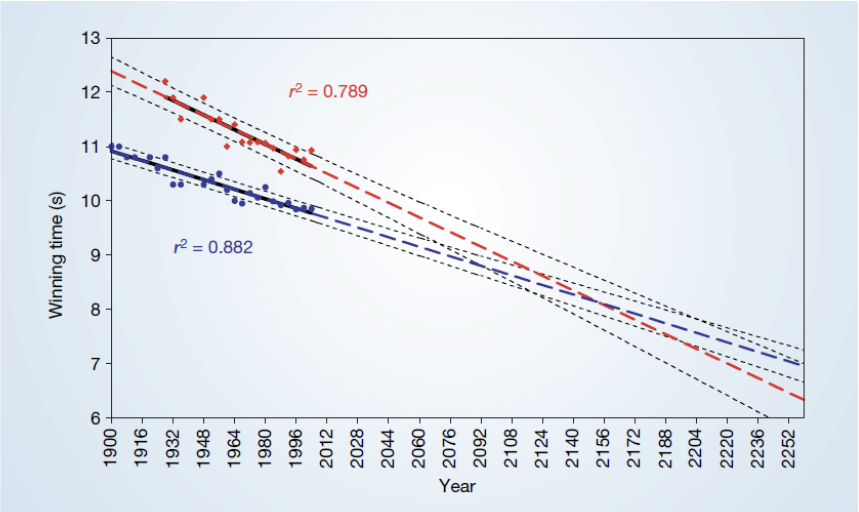

Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.