2.4: Simultaneous Inference and Important Considerations

Dr. Bean - Stat 5100

Simultaneous inference is when we want to conduct multiple tests of significance at the same time.

1 Why Simultaneous Inference?

In handout 2.3, we conducted inference for parameters one at a time. We need to change our approach when looking at multiple parameters simultaneously.

(Groups) How and why do we need to change our approach when conducting simultaneous inference?
(check out **this comic** for help).

If we conduct several tests at the same level of significance, the probability of getting one false positive result (a type I) error becomes much higher than α.

As a result, we need to adjust the level of significance to account for a "multiplicity" of testing.

2 Bonferroni Adjustment

Multiplicity:

- Let A_j = event that an individual $(1-\alpha)100\%$ CI does not contain the true value of β_j.
- $P(A_0) = P(A_1) = \alpha \rightarrow$ Type I Error
  - $P(\text{NOT} A_j)$ = probability that an interval contains the true value of β_j.
- **Bonferroni Inequality:** $P(\text{NOT} A_0 \text{ AND NOT } A_1) \geq 1 - P(A_0) - P(A_1)$

This means that if we conduct g tests at a confidence level of $(1-\frac{\alpha}{g})$, then we are guaranteed that overall level of confidence for all intervals *considered jointly* will be at least $(1-\alpha)$, we call this the **Bonferroni adjustment**.

- **Bonferroni Advantage:** Can be literally applied in *any* situation requires a multiplicity adjustment, including simultaneous intervals for $\hat{Y}$ at multiple X_h levels.
- **Bonferroni Disadvantage:** Can be overly conservative, producing inefficient (unnecessarily wide) intervals.

Comparison of Simultaneous Intervals for $\hat{Y}$

- Confidence intervals (mean response)
  - Bonferroni

$$\hat{Y} \pm t_{n-p}(1-\frac{\alpha}{2g}) * s\{\hat{Y}_h\}$$

- Working-Hotelling (WH)

$$\hat{Y} \pm W * s\{\hat{Y}_h\} \qquad \left(W = \sqrt{pF_{p,n-p}(1-\alpha)}\right)$$

Notice that the W-statistic does not consider g

* WH provides a "confidence band" for the entire regression line (all possible X_h levels).
* This means the WH interval at any individual X_h will be wider than the t-based confidence interval, but the WH intervals will eventually be narrower than Bonferroni confidence intervals if enough X_h are considered.

- Prediction intervals (new response)
  - Bonferroni

$$\hat{Y} \pm t_{n-p}(1 - \frac{\alpha}{2g}) * s\{\hat{Y}_{h(new)}\}$$

  - Scheffe (chef-eh)

$$\hat{Y} \pm S * s\{\hat{Y}_{h(new)}\} \qquad \left(S = \sqrt{gF_{g,n-p}(1-\alpha)}\right)$$

Rule of Thumb: Always pick the most efficient interval that guarantees your intended type I error (α).

Table 1: Summary of Methods for Simultaneous Intervals

Simultaneous Interval on:	**Methods**
β's	Bonferroni
Population means of Y at multiple X_h	Bonferroni or Working-Hotelling
Predictions for Y at multiple X_h	Bonferroni of Scheffe

3 Inverse Prediction

Problem: What is the value of X_h necessary to achieve a specific value of $\hat{Y}$.

Solution: solve for X.

$$\begin{aligned} \hat{Y} &= b_0 + b_1 X_h \\ b_1 X_h &= \hat{Y} - b_0 \\ X_h &= \frac{\hat{Y} - b_0}{b_1} \end{aligned}$$

Problem: Use Y to predict values of X.

Solution: DO NOT solve for X.

Why?

- The least squares slope estimate of regression model that predicts Y using X: $\rho\frac{SD\{Y\}}{SD\{X\}}$.
- The least squares slope estimate of regression model that predicts X using Y: $\rho\frac{SD\{X\}}{SD\{Y\}}$.
- Notice that the slopes are NOT inverses of each other.

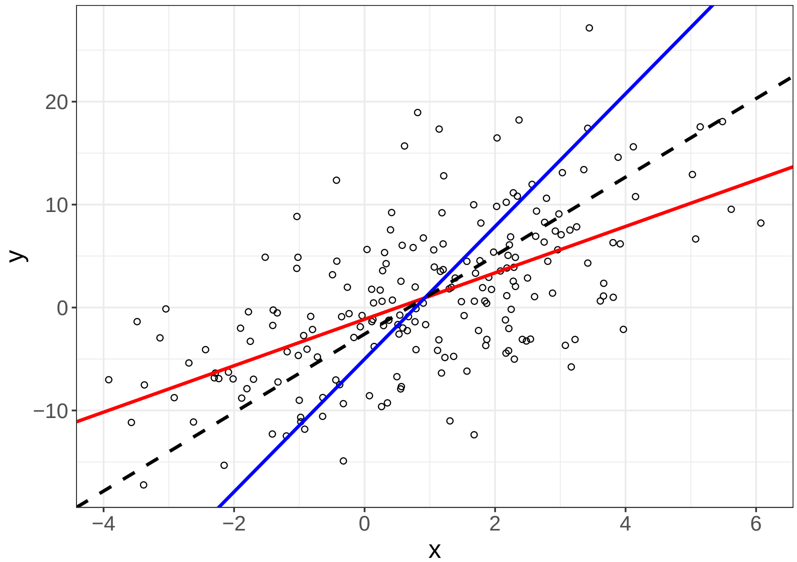

Figure 1: Scatterplot of points along with the regression line that uses X to predict Y (red), the regression line that uses Y to predict X (blue), the SD line (black).

4 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

5 Cautions for Linear Regression

- Remedial measures may not fix violations of assumptions
  - May need to abandon OLS regression altogether

- Interpretation: Sometimes the X vs Y relationship may look counterintuitive
  - May be the result of omitted predictors
- R^2 can be abused
  - Higher $R^2 \to$ not always better model
  - Lower $R^2 \to$ does not mean there is no linear relationship
- Extrapolation
  - Linear model is only good for prediction/inference at or near the range of observed data points.

(Individual) If I have created a linear model for prediction, would I WANT to extrapolate?

Slight extrapolations are usually not bad and even desirable. However, the farther you get away from the range of your observed data the more inappropriate your predictions will become.

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-

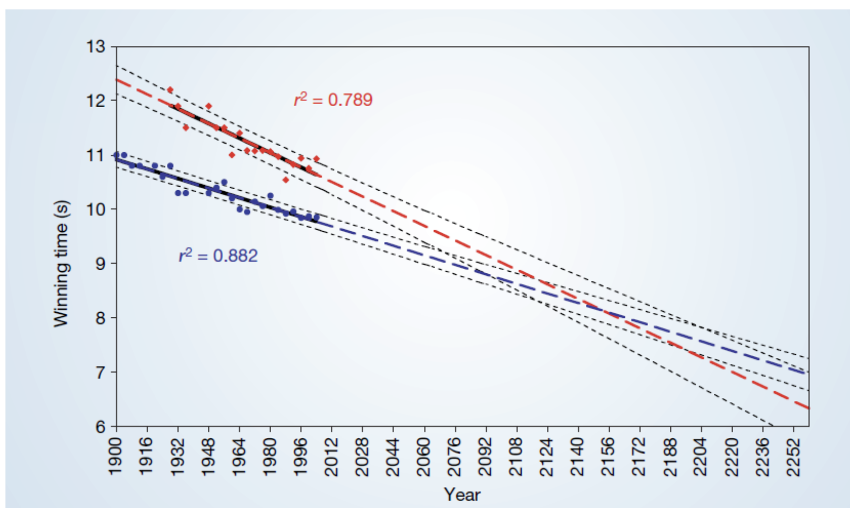

Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) 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.