P8130: Biostatistical Methods I

Lecture 14: Inferences in Simple Linear Regression

Cody Chiuzan, PhD
Department of Biostatistics
Mailman School of Public Health (MSPH)

Outline

- Lectures 12 & 13 introduced the simple linear regression model
- Least Squares (LS) and Maximum Likelihood (ML) estimation methods
- Today will discuss:
  - Inferences in SLR
  - Correlation and coefficient of determination

Inferences about parameters

- The two basic tools for statistical inference are:
  - **Confidence intervals (CI):** provide a range of values for the TRUE parameter of interest
  - **Hypothesis testing**: how probable are the data gathered under a null hypothesis about the data generating distribution
  - **P-values**? Often (ab)used.

    Compare the data vs the null (no effect) rather that testing whether the data are consistent with your hypothesis

    How 'unlikely' that we would observe such an extreme test statistic (e.g., t^*) in the direction of the alternative, given that the null hypothesis is true

Inference about parameters

- To make inferences we need to understand the sampling distribution of the model estimators:

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2})$$

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}))$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}} \sim t_{n-2}$$

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSE(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2})}} \sim t_{n-2}$$

Inference on β_1 (true regression slope)

- A *(1-α)* 100% confidence interval for the true slope is given by:

$$\widehat{\beta_1} \pm t_{n-2,1-\alpha/2} \cdot se(\widehat{\beta_1}), \text{ where } se(\widehat{\beta_1}) = \sqrt{MSE / \sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- Hypothesis test: $H_0: \beta_1 = \beta_{10}$ *vs* $H_1: \beta_1 \neq \beta_{10}$ (β_{10} is usually 0).

- Test statistic t^* has the following distribution:

$$\frac{\widehat{\beta_1} - \beta_{10}}{se(\widehat{\beta_1})} \sim t_{n-2,1-\alpha/2}$$

Inferences – regression output

Parameter	Estimate	Standard error	t^*	p-value
Intercept β_0	b_0	$\text{se}(b_0)$	$t_0^* = \frac{b_0}{se(b_0)}$	$P(\lvert T\rvert > \lvert t_0^*\rvert)$
Slope β_1	b_1	$\text{se}(b_1)$	$t_1^* = \frac{b_1}{se(b_1)}$	$P(\lvert T\rvert > \lvert t_1^*\rvert)$

The p-values in the table correspond to the two hypotheses tests:

$$H_0\colon\ \beta_1 = 0\ and\ H_0\colon\ \beta_0 = 0$$

P-value = Probability(observing something as or more extreme than test statistic$|H_0$)
Conclusions: reject or fail to reject H_0 (NEVER accept H_1)

Inference about the regression line

- Let X_h be any predictor. We want to estimate the mean of all outcomes in the population that have covariate X_h:

$$E(Y_h) = \beta_0 + \beta_1 X_h$$

- The point estimate is given by: $\widehat{Y_h} = \widehat{\beta_0} + \widehat{\beta_1} X_h$
- The *(1-α)* 100% confidence interval (CI) for $\beta_0 + \beta_1 X_h$ is given by:

$$\widehat{\beta_0} + \widehat{\beta_1} X_h \pm t_{n-2,1-\alpha/2} \cdot se(\widehat{\beta_0} + \widehat{\beta_1} X_h),$$

$$se(\widehat{\beta_0} + \widehat{\beta_1} X_h) = \sqrt{MSE \left\{ \frac{1}{n} + [(X_h - \bar{X})^2 / \sum_{i=1}^{n} (X_i - \bar{X})^2] \right\}}$$

Prediction intervals

- Previous slide showed the CI for the unknown *mean* at X_h
- What if we want an interval for an actual (single), NEW value Y_h?

$$Y_h = \beta_0 + \beta_1 X_h + \varepsilon_h$$

- We are making inferences about an individual value, not the mean of all *Xs*
- It should be wider than the interval for the mean of Y|X, i.e., CI

Prediction interval

- The *(1-α)* 100% *prediction* interval for one, new value Y_h is given by:

$$\widehat{\beta_0} + \widehat{\beta_1} X_h \pm t_{n-2,1-\alpha/2} \cdot se(\widehat{\beta_0} + \widehat{\beta_1} X_h),$$

$$se(\widehat{\beta_0} + \widehat{\beta_1} X_h) = \sqrt{MSE \left\{ \frac{1}{n} + [(X_h - \bar{X})^2 / \sum_{i=1}^{n} (X_i - \bar{X})^2] + 1 \right\}}$$

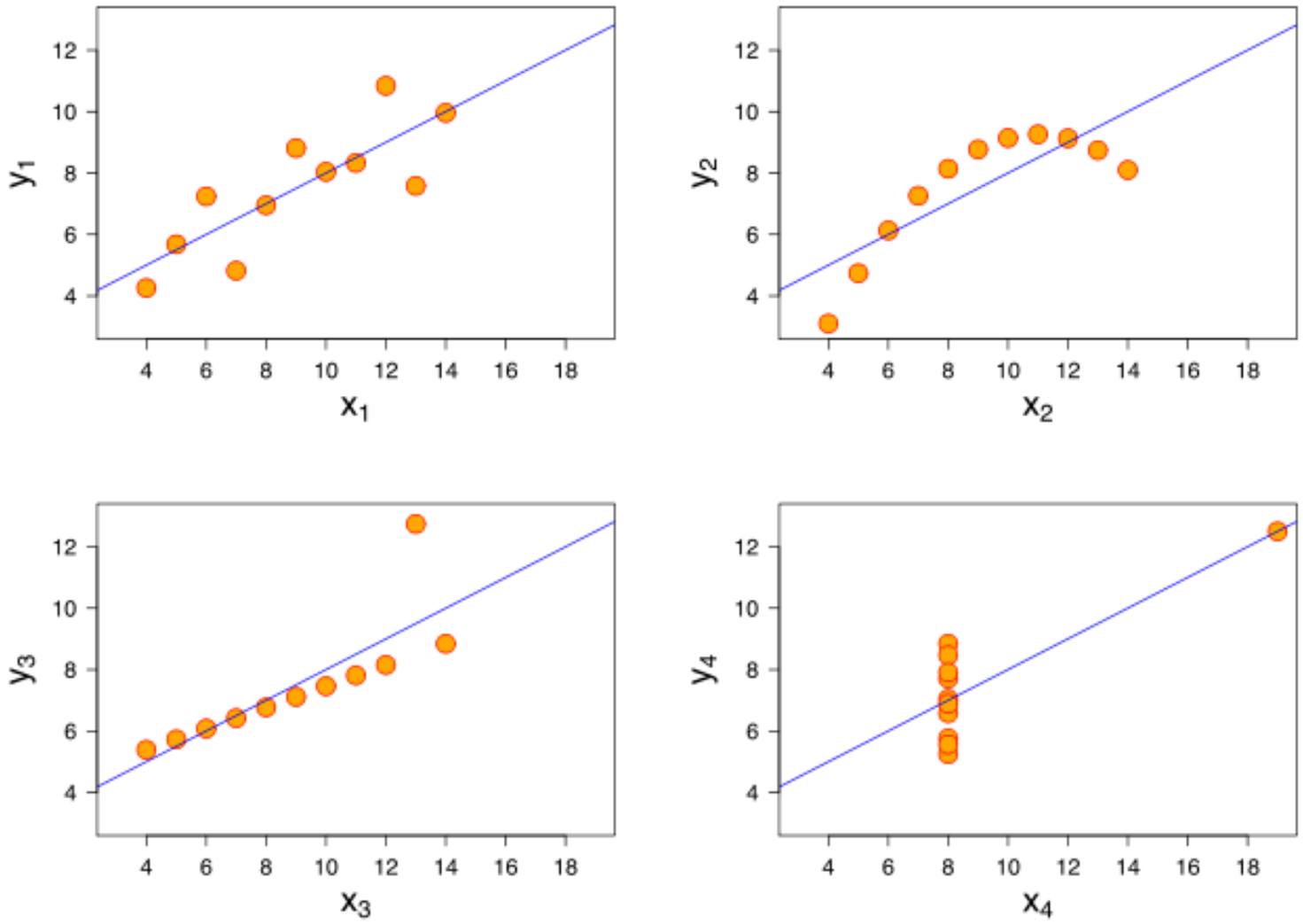
- Where is +1 coming from?

Slope vs Correlation

- Slope: measures if there is a linear trend (relationship) b/w X and Y
- Correlation: provides a measure of the strength of this linear relationship
  - Person correlation
- Statistical significance is the same

  The p-value for testing that correlation is 0 is the SAME as the p-value for testing that slope is 0 (why?)

Examples: (Same) Slope vs Correlation



Correlation

- The estimated correlation between X and Y is given by:

$$\hat{\rho}_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = r$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

Correlation

- The correlation coefficient r is a unitless measure: $-1 \leq r \leq 1$
- When $r = 0$, there is no linear relationship between X and Y
- Estimated correlation is a function of the estimated slope:

That's why we have the same p-value for both hypothesis tests!

Correlation vs Coefficient of Determination

- Coefficient of determination is denoted by r^2 or R^2
- Used in regression as a measure of goodness of fit
- It is in fact a proportion which tells us how much variation in Y is explained by the variation in X

$$0 \leq r^2 \leq 1$$

- If $r^2 = 1$, then all points fall perfectly on the regression line.
- If $r^2 = 0$, then the estimated regression line is perfectly horizontal ☺.

Coefficient of Determination

- Remember 'Sum of squares error': $SSE = \sum_{i-1}^{n} (Y_i - \hat{Y}_i)^2 = \hat{\varepsilon}_i^2$
- 'Sum of squares regression': $SSR = \sum_{i-1}^{n} (\hat{Y}_i - \bar{Y})^2$
- 'Sum of squares total': $SSTO = \sum_{i-1}^{n} (Y_i - \bar{Y})^2$

SSTO = SSR + SSE

- *Coefficient of determination* is given by: $r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$
- For simple linear regression only: $r^2 = \hat{\rho}^2$

Readings

Kutner *et al*., Applied Linear Statistical Models

- Chapter 2, Sections: 2.1 – 2.5 and 2.9

Next class:

- Multiple Linear Regression (Continuous and Categorical Predictors)