14.2 Confidence and Prediction Intervals

MA 116

June 2025

Review: 14.1

1. Population and sample with paired data points (x_i, y_i)
2. Such a sample gives a least-square regression equation $\hat{y} = b_1 x + b_0$, in which b_1 and b_0 are statistics.
3. The corresponding population parameters are β_1 and β_0
4. $y = \beta_1 x + \beta_0$ is the true linear relation
5. For a a number that the explanary variable x can reasonably assume, $mu_{y|a}$ is the population mean of response variable y when $x = a$.
6. To conduct inference on the least squares regression line (e.g. do hypothesis test about β_1), we require two conditions to be met:
   - Condition 1. It is reasonable to say x and y are linearly relate. Equivalently, There are fixed numbers β_1 and β_0 such that $\mu_{y|x} = \beta_1 x + \beta_0$ is approximately true for any value of x. If the residual plot has no pattern, then we say Condition 1 is met.

Review: 14.1

1.
   - Condition 2. For any fixed value of x, the response variable y is approximately normally distributed with mean $\mu_{y|x} = \beta_1 x + \beta_0$ and standard deviation σ, a fixed value that does not depend on value of x. (Need to know or be able to approximate this value.)
   -
     - To check Condition 2, we introduce the **least-squares regression model**

       $$y_i = \mu_{y|x_i} - \epsilon_i = \beta_1 x_i + \beta_0 - \epsilon_i,$$

       where ϵ_i is a random error term with mean 0 and standard deviation $\sigma_{\epsilon_i} = \sigma$
     - If ϵ_i is normally distributed with mean 0 and standard deviation σ that we can approximate, then Consition 2 is met. Indeed, we are able to use $s_e = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}}$ to approximate σ. In this class, we make an additional assumption that ϵ_i is normally distributed. Then Condition 2 follows directly from this additional assumption.
2. Hypothesis test about β_1: Using new variable $t = \frac{b_1 - \beta_1}{s_{b_1}}$.

Predicted value

Given a sample we can obtain a least squares regression line $\hat{y} = b_1 x + b_0$. At any numerical value of x, the corresponding numerical value $\hat{y} = b_1 x + b_0$ is called the predicted value of the response varaible for a given value of x with respect to this particular least squares regression line.

Two interpretations about the predicted value

1. It estimates the mean of all response variable values at this x value
2. It estimates the predicted y value for a randomly selected individual from the population at this x value.

Since the predicted value $\hat{y}$ has these two different interpretations, there are two different interval estimators that both centered at $\hat{y}$. The interval estimator that estimates $\mu_{y|x}$ is called a **confidence intervals for a mean response**. The interval estimator that estimates the y value from a random selected data point (x, y) from the population is called a **prediction interval for an individual response.**

Despite the fact that both use the predicted value of the response $\hat{y} = b_1 x + b_0$ as a point estimator, prediction interval for an individual response often needs to have a larger margin of error compared to confidence intervals for a mean response.

Suppose a family doctor is interested in examining the relationship between a patient's age and total cholesterol level (in mg/dL). In this setting, our explanatory variable is x=age, and response variable is y=cholesterol level (in mg/dL). Population=all American people. $\mu_{y|32}$ is the population mean of cholesterol level (in mg/dL) of all Americans of age 32. Suppose this doctor calculate an interval estimator $[198.8, 224, 4]$ to estimate $\mu_{y|32}$ from a sample. However, given a randomly selected Americans of age 32, can we say their cholesterol level is between 198.8 and 224.4 mg/dL? NO! This is way too restricted.

For the second scenario, we use **prediction interval for an individual response**. ***Calculating intervals will not be tested, but concepts/definitions may be tested.

Confidence intervals for a mean response

Assume the sample size is large enough and Condition 1 and 2 are met. Fix a numerical value x

1. Population parameter to be estimated: $\mu_{y|x}$ where x is a numerical value
2. Point estimator (center of our interval) $\hat{y} = b_1 x + b_0$, a numerical value calculated from a least squares regression line obtained from a sample
3. Confidence level $(1-\alpha)100\%$
4. Margin of error $E = t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_i (x_i - \overline{x})^2}}$
   - In this formula, x is a fixed numerical value.
   - $\sum_i (x_i - \overline{x})^2$ is just the sum of square of this least squares regression line

Prediction intervals for an individual response

Assume the sample size is large enough and Condition 1 and 2 are met. Fix a numerical value x

1. Quantity to be estimated: The y value of a randomly selected (x, y) from population.
2. Point estimator (center of our interval) $\hat{y} = b_1 x + b_0$, a numerical value calculated from a least squares regression line obtained from a sample
3. Confidence level $(1 - \alpha)100\%$
4. Margin of error $E = t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}$
   - In this formula, x is a fixed numerical value.
   - $\sum_i (x_i - \overline{x})^2$ is just the sum of square of this least squares regression line
   - This formula gives a larger E than the previous margin of error formula for $\mu_{y|x}$.

Review: Hypothesis test about β_1

We want to answer the following question: Do the sample data provide sufficient evidence to conclude that a linear relation exists between the two variables? If there is no linear relation between the response and explanatory variables, the slope of the true regression line will be zero. Do you know why? A slope of zero means that information about the explanatory variable, x, does not change our estimate of the value of the response variable, y.

1. the null hypothesis $H_0 : \beta_1 = 0$ means there is no linear relation between the explanatory and response variables.
2. the alternative hypothesis $H_1 : \beta_1 \neq 0$ means there is linear relation between the explanatory and response variables.
3. the alternative hypothesis $H_1 : \beta_1 > 0$ means there is linear relation between the explanatory and response variables that is positively associated.
4. the alternative hypothesis $H_1 : \beta_1 > 0$ means there is linear relation between the explanatory and response variables that is negatively associated.

Example of hypothesis test about β_1

Suppose we want to test if temperature (explanary variable) is in linear relation to garlic leaves growing speed (response variable).

x_i	y_i
50	1
60	3
70	3
80	3

The table shows a sample we get.

1. Step I. Calculate $\overline{x}$, $\overline{y}$, s_x and s_y.
2. Step II. Calculate r.

Example of hypothesis test about β_1

Suppose we want to test if temperature (explanary variable) is in linear relation to garlic leaves growing speed (response variable), to a significance level $\alpha = 0.05$. Before we determine H_0 and H_1, we first calculate all the values that might be used in our hypothesis test.

x_i	y_i
50	1
60	3
70	3
80	3

The table shows a sample we get.

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775.$

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775.$
3. Calculate

$$b_1 = r\frac{s_y}{s_x}$$

and

$$b_0 = \overline{y} - b_1\overline{x}.$$

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775.$
3. $$b_1 = r\frac{s_y}{s_x} = 0.06$$

   and

   $$b_0 = \overline{y} - b_1\overline{x} = -1.4.$$

   Therefore, the least squares regression line we obtained from this sample is

   $$\hat{y} = 0.06x - 1.4.$$

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775$.
3. Step III. $\hat{y} = 0.06x - 1.4$.
4. Step IV. Calculate the residues and draw a residual plot for this least squares regression line.

x_i	y_i	$\hat{y}_i = 0.06x_i - 1.4$	residual
50	1		
60	3		
70	3		
80	3		

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775$.
3. Step III. $\hat{y} = 0.06x - 1.4$.
4. Step IV. Calculate the residues and draw a residual plot for this least squares regression line.

x_i	y_i	$\hat{y}_i = 0.06x_i - 1.4$	residual
50	1	1.6	-0.6
60	3	2.2	0.8
70	3	2.8	0.2
80	3	3.4	-0.4

Since the sample data points seem to scatter around the zero line in the residual plot without a pattern, we say that a linear relationship between x and y is reasonable.

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775$.
3. Step III. $\hat{y} = 0.06x - 1.4$.
4. Step IV.

x_i	y_i	$\hat{y}_i = 0.06x_i - 1.4$	residual
50	1	1.6	-0.6
60	3	2.2	0.8
70	3	2.8	0.2
80	3	3.4	-0.4

5. Step V. Calculate

$$s_e = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_i \text{residual}^2}{n-2}}$$

and

$$s_{b_1} = \frac{s_e}{(\sqrt{n-1})s_x}$$

.

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775$.
3. Step III. $\hat{y} = 0.06x - 1.4$.
4. Step IV.

x_i	y_i	$\hat{y}_i = 0.06x_i - 1.4$	residual
50	1	1.6	-0.6
60	3	2.2	0.8
70	3	2.8	0.2
80	3	3.4	-0.4

5. Step V.

$$s_e = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_i \text{residual}^2}{n-2}} = \sqrt{\frac{1.2}{2}} = 0.77,$$

$$s_{b_1} = \frac{s_e}{(\sqrt{n-1})s_x} = 0.034.$$

1. Step I. $\overline{x} = 65$, $\overline{y} = 2.5$, $s_x = 12.9$ and $s_y = 1$.
2. Step II. $r = \dfrac{22.5 - 2.5 + 2.5 + 7.5}{12.9 \cdot 1 \cdot 3} = \dfrac{30}{38.73} = 0.775$.
3. Step III. $\hat{y} = 0.06x - 1.4$.
4. Step IV.

x_i	y_i	$\hat{y}_i = 0.06x_i - 1.4$	residual
50	1	1.6	-0.6
60	3	2.2	0.8
70	3	2.8	0.2
80	3	3.4	-0.4

5. Step V. $s_e = \sqrt{\dfrac{\sum_i (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\dfrac{\sum_i \text{residual}^2}{n-2}} = \sqrt{\dfrac{1.2}{2}} = 0.77$,
   $s_{b_1} = \dfrac{s_e}{(\sqrt{n-1})s_x} = 0.034$.
6. Step VI. Hypothesis test. H_0, $H_1 : \beta_1 \neq 0$, test type. Critical value(s). Calculate test statistic. Conclusion.