Inference on the least-squares regression model

MA 116

June 2025

Items marked with *** are important concepts that may be tested on quizzes or exams.

In Chapter 14 we start by asking how to relate concepts in Chapter 4 (a sample with pair data points (x_i, y_i) and a least squares regression line $\hat{y} = b_1 x + b_0$) to our big picture (population vs. sample, sampling distribution, hypothesis test, interval estimator).

In the least-square regression equation $\hat{y} = b_1 x + b_0$, we note that b_1 and b_0 are statistics. The statistics b_1 and b_0 are sample statistics for the population parameters β_1 and β_0.

The true linear relation between the explanatory variable x and the response variable y is given by $y = \beta_1 x + \beta_0$.

Inference about β_1 and β_0?

Because b_1 and b_0 are statistics, their value vary from sample to sample, so a sampling distribution is associated with each. We use this sampling distribution to perform inference on β_1 and β_0.

inference example

We might want to **test** whether there is a linear relation between x and y. One method is to do a hypothesis test with $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$. If the result is statistically significant, we say that there is sufficient evidence to conclude that there is a linear relation between x and y.

However, to investigate the distribution of b_1 or to do any other inference, we require two conditions on our populationto be met.

Definition.

Suppose 32 is a possible x value, i.e. a value that the variable x can assume. Then, we let $\mu_{y|32}$ be the mean value of the response variable y given that the value of the explanatory variable x is 32.

Idea: 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). Let's think our population to be all American people. Then people at the age 32 can verywell have different cholesterol level (in mg/dL). $\mu_{y|32}$ is the population mean of cholesterol level (in mg/dL) of all Americans of age 32.

Condition 1: Linearity is a reasonable approximation.

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.

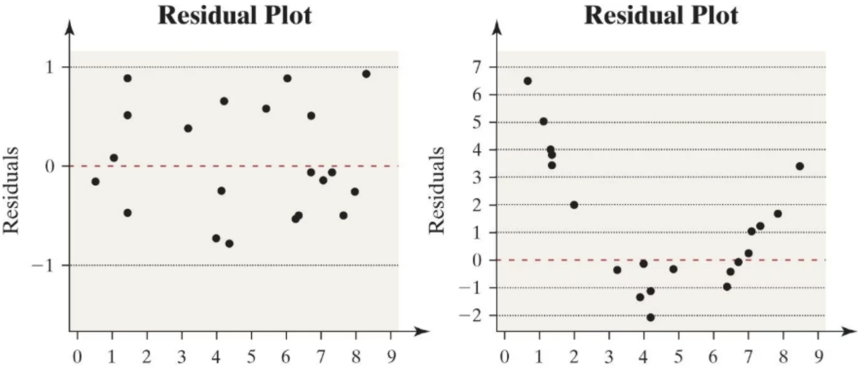
Given a population, how to check if this condition is met? **Steps:**

1. Obtain a random sample of a reasonable size n
2. Calculate the least squares regression line of the sample $\hat{y} = b_1 x + b_0$
3. Calculate the residuals of the sample $y_i - \hat{y}_i$
4. Plot the residual against the explanatory variable x

If a plot of the residuals against the explanatory variable shows a discernible pattern, such as a curvy, then the explanatory and response variable may not be linearly related.

***

Figure 20



(a) Linear Model Appropriate

(b) Linear Model Not Appropriate: Patterned Residuals

Condition 2. Normality

***

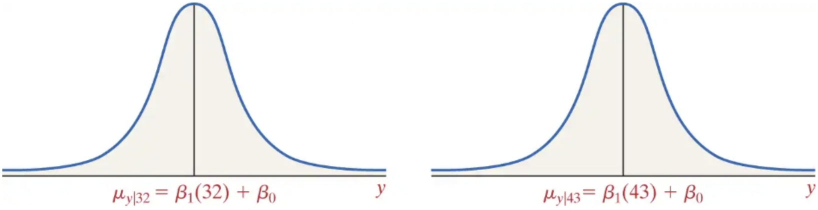
For any fixed value of x, the response variable y is approximately normally distributed with mean $\mu_{y|x} = \beta_1 x + \beta_0$ and standard deviation σ, a fixed value that does not depend on value of x.

Idea. Fix a value of x let's consider the probability density of getting a particular value of y. This condition is requiring that the probability density function of y with a fixed value of x is approximately a normal curve.

How do we verify if this condition is met? (†)

Idea of the normality condition

Figure 4


$\mu_{y|32} = \beta_1(32) + \beta_0$
y
$\mu_{y|43} = \beta_1(43) + \beta_0$
y

As the value of x changes from 32 to 43, the distribution of y with respect to a fixed value of x remains normal but shifts horizontally–the peak of its curve shifts from $\mu_{y|32} = 32\beta_1 + \beta_0$ to $\mu_{y|43} = 43\beta_1 + \beta_0$. But the shape of the normal curve remains unchanged.

Concepts

A linear relation is of the form $y = \beta_1 x + \beta_0$ for x, y variables and β_1, β_0 values.

In question solving, we use the equation $\mu_{y|x} = \beta_1 x + \beta_0$ for $\mu_{y|x}$, β_1, β_0 population parameters. If we've checked that the residual plot of a random sample is good, we then assume that **linerality is a reasonable approximation** and assume this equation holds. $***$

The difference between the observed (y_i, from a data point (x_i, y_i) in our sample) and predicted value ($\mu_{y|x_i}$) of the response variable is an error term, ϵ_i.

Least-squares regression model

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

where

1. β_1 and β_0 are parameters to be estimated based on sample data
2. ϵ_i is a random error term with mean 0 and standard deviation $\sigma_{\epsilon_i} = \sigma$, the error terms are independent (reason: x_i is a fixed value so $\beta_1 x_i + \beta_0$ is also a fixed value (despite the fact that β_1 and β_0 are to be estimated), so we can calculate the mean and standard deviation of ϵ_i from those of y for a fixed value x_i)
3. $i = 1, \ldots, n$ where n is the sample size

***

Standard error of the estimate s_e

Recall the second condition requires that

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.

To try to verify this, we realize that this condition is equivalent to say that ϵ_i is normally distributed with mean 0 and standard deviation σ, independent of i. We then define what's called a **standard error of the estimate** s_e to estimate σ_{ϵ_i} hence σ. s_e is an unbiased estimator of σ.

Given a sample of size n we first obtain the least squares regression line $\hat{y} = b_1 x + b_0$. Then

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

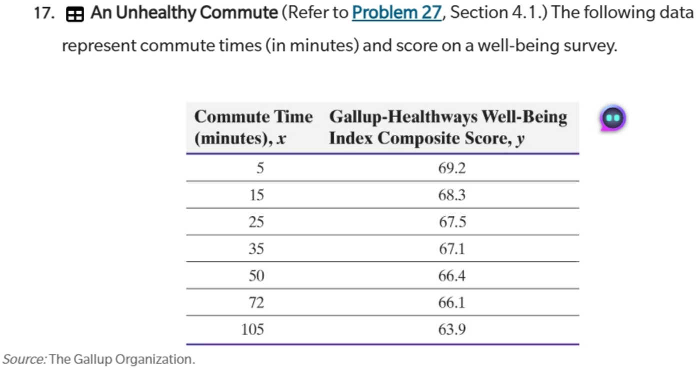
***

Given a sample of size n we first obtain the least squares regression line $\hat{y} = b_1 x + b_0$. Then

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

The reason we divide by $n-2$ is that in a least squares regression line we have estimated two population parameters β_0 and β_1, that is, we lose 2 degrees of freedom.

Example. Compute the standard error of the estimate



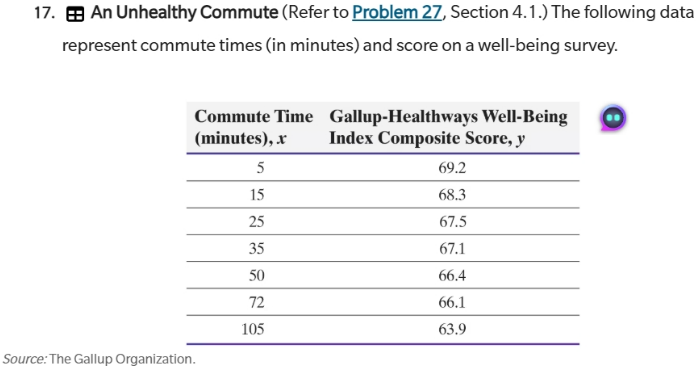
17. **An Unhealthy Commute** (Refer to Problem 27, Section 4.1.) The following data represent commute times (in minutes) and score on a well-being survey.

Commute Time (minutes), x	Gallup-Healthways Well-Being Index Composite Score, y
5	69.2
15	68.3
25	67.5
35	67.1
50	66.4
72	66.1
105	63.9

Source: The Gallup Organization.

1. Compute the least-squares regression line.

Example. Compute the standard error of the estimate



17. An Unhealthy Commute (Refer to Problem 27, Section 4.1.) The following data represent commute times (in minutes) and score on a well-being survey.

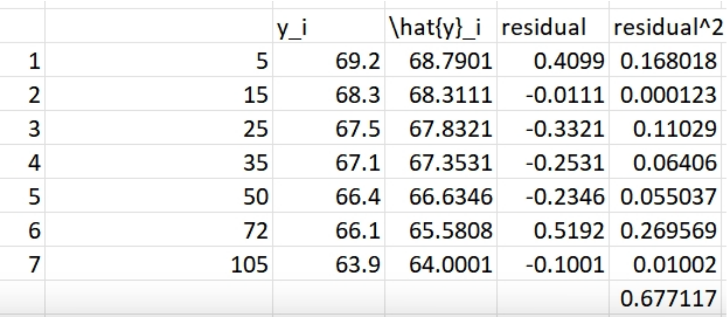
Commute Time (minutes), x	Gallup-Healthways Well-Being Index Composite Score, y
5	69.2
15	68.3
25	67.5
35	67.1
50	66.4
72	66.1
105	63.9

Source: The Gallup Organization.

1. The least-squares regression line of this sample is $\hat{y} = -0.0479x + 69.0296$.
2. Compute the predicted values for each observation in the data set, i.e. compute $\hat{y}_i$ for each i.
3. Compute the residuals $y_i - \hat{y}_i$ for each i.
4. Compute $\sum_i$ resduals^2 of this sample.

Example. Compute the standard error of the estimate

1. The least-squares regression line of this sample is $\hat{y} = -0.0479x + 69.0296$.



		y_i	\hat{y}_i	residual	residual^2
1	5	69.2	68.7901	0.4099	0.168018
2	15	68.3	68.3111	-0.0111	0.000123
3	25	67.5	67.8321	-0.3321	0.11029
4	35	67.1	67.3531	-0.2531	0.06406
5	50	66.4	66.6346	-0.2346	0.055037
6	72	66.1	65.5808	0.5192	0.269569
7	105	63.9	64.0001	-0.1001	0.01002
					0.677117

2.

3. Compute the standard error of the estimate s_e of this sample.

*** Students are expected to solve questions like this by hand with a smaller sample size

Recall that we want to verify if Condition 2 on normality is met. (†)

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.

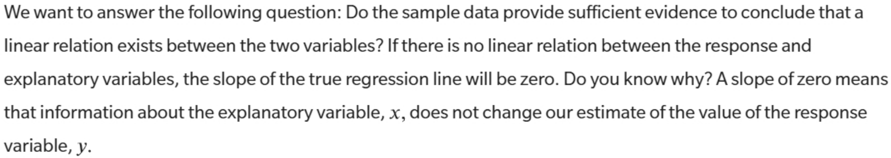
We translated this condition into the following condition about ϵ_i.

For any x value x_i, $\epsilon_i = y_i - \beta_1 x_i - \beta_0$ is normally distributed with a mean at 0 and a standard deviation $\sigma_{\epsilon_i} = \sigma$.

We claimed that $s_e = \sqrt{\dfrac{\sum_i (y_i - \hat{y}_i)^2}{n-2}}$ is an estimator of σ_{ϵ_i} hence σ. Since s_e does not depend on a choice of fixed value of x, we conclude that σ_{ϵ_i} hence σ does not depend on a choice of fixed value of x.

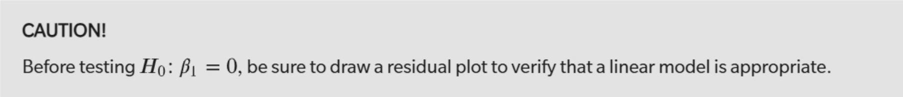
In this class, we make an additional assumption that ϵ_i is normally distributed. Then Condition 2 follows directly from this additional assumption.

Have both Condition 1 and 2 met, we may now conduct inference 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.



CAUTION!

Before testing $H_0\colon \beta_1 = 0$, be sure to draw a residual plot to verify that a linear model is appropriate.

We need to varify that Condition 1 is met, i.e. a linear model is appropriate.

In the following part of this section, we assume that a linear model is appropriate and that ϵ_i are normally distributed. I.e. both Condition 1 and 2 are met. Then we may conduct inference on β_1.

Obtaining different random samples of the same size n gives us different least squares regression lines $\hat{y} = b_1 x + b_0$ (i.e. different values of b_1 and b_0), so it makes sense to view b_1 as a random variable and talk about its sampling distribution. We claim that the standard deviation of the variable b_1 is given by

$$\frac{\sigma}{\sqrt{\sum_i (x_i - \overline{x})}}.$$

Since we don't know the population parameter σ, we use

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

as an estimator of σ.***

It follows that by replacing σ by s_e we defined a sample statistics

$$s_{b_1} = \frac{s_e}{\sqrt{\sum_i (x_i - \overline{x})}} = \frac{s_e}{(\sqrt{n-1})s_x}.$$

Note for a particular sample, $s_x = \sqrt{\frac{\sum (x_i - \overline{x})}{n-1}}$ and $s_e = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}}$ are both numerical values, so s_{b_1} is also a numerical value, which justifies that s_{b_1} is a sample statistic.***

Sampling distribution

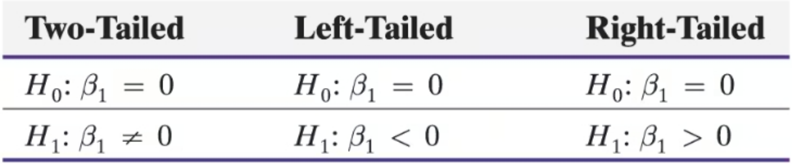
Under all our previous assumptions, we have the following result of sampling distribution.

View b_1 as a variable, we construct a new variable

$$t = \frac{b_1 - \beta_1}{s_{b_1}}.$$

This new variable t follows Student's t-distribution with $n-2$ degrees of freedom.

Hypothesis testing



Two-Tailed	Left-Tailed	Right-Tailed
$H_0\colon \beta_1 = 0$	$H_0\colon \beta_1 = 0$	$H_0\colon \beta_1 = 0$
$H_1\colon \beta_1 \neq 0$	$H_1\colon \beta_1 < 0$	$H_1\colon \beta_1 > 0$

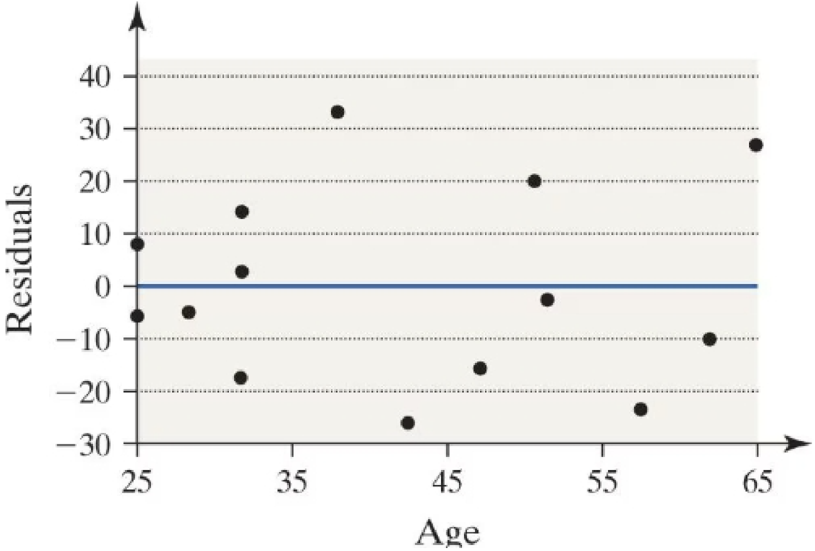
1. Look at a residual plot and determine if linerality is a reasonable approximation
2. Determine H_0 and H_1 and test type, then **assume that $\mathbf{H_0}$ is true**
3. Calculate the critical value(s) and critical region based on α and test type
4. Compute the test statistics t_0
5. Determine whether t_0 lies in the critical region, if so, reject H_0
6. State the conclusion: There is/is not sufficient evidence to conclude that a linear relation exists between [explanatory variable] and [response variable].

Compute the test statistics t_0: Recall that in the setting of doing hypothesis testing, we obtain a sample and try to find evidence from sample data analysis to support our H_1. Having this sample, we should compute s_{b_1} and b_1 to get numerical values.
Under the assumption that H_0 is true, we have $\beta_1 = 0$. Then the formula to calculate the test statistics is

$$t_0 = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1}{s_{b_1}}.$$

***

Example



Example 5, p.708 (classical approach)

Confidence interval about β_1

If both Condition 1 and 2 are met, we may say that a $(1-\alpha)100\%$ confidence interval for the slope of the true regression line, β_1, is given by

$$b_1 \pm t_{\alpha/2} \cdot s_{b_1},$$

with $t_{\alpha/2}$ wrt $n-2$ degrees of freedom.***

With this formula, we may conduct hypothesis text about β_1 using confidence interval approach.

1. Look at a residual plot and determine if linerality is a reasonable approximation
2. Determine H_0 and H_1 and test type
3. Calculate the confidence interval
4. If $\beta_1 = 0$ is not in the interval, reject H_0, otherwise do not reject H_0.
5. State the conclusion: There is/is not sufficient evidence to conclude that a linear relation exists between [explanatory variable] and [response variable].

Example 6, p.711.