EE 381 Probability & Statistic with Applications to Computing (Fall 2020)

Lecture 12
Hypothesis Testing & Regression

Duc H. Tran, PhD

Hypothesis Testing and Regression

- Statistical Decisions
- Test of Hypothesis and Significance
- Curve Fitting
- Regression and Correlation

Statistical Decisions

Definition: making decisions about populations on the basis of sample information.

Example:

We wish to decide on the basis of sample data whether:

- A new serum is really effective in curing a disease
- One educational procedure is better than another
- Or whether a given coin is loaded.

Statistical Hypotheses. Null Hypotheses

- Statistical hypotheses: statements about the probability distributions of the populations.
- In simple terms, it means we make assumptions or guesses about the populations involved in attempting to reach decisions.
- Two types of statistical hypotheses:
  - ✓null hypotheses
  - ✓alternative hypothesis.

Statistical Hypotheses. Null Hypotheses

Example:

- ✓We want to decide whether a given coin is loaded, we formulate the hypothesis that the coin is fair, i.e., p=0.5, where p is the probability of heads.
- ✓We want to decide whether one procedure is better than another, we formulate the hypothesis that there is *no difference* between the procedures.

- These hypotheses are called null hypotheses, denoted by H_0

- Hypotheses that differs from null hypotheses are called alternative hypothesis, denoted by H_1.

Example:

- ✓If the null hypothesis is p=0.5, possible alternative hypotheses are p=0.7, or p>0.5.

Tests of Hypotheses and Significance

- Procedure that enable us to decide whether to accept or reject hypotheses or to determine whether observed samples differ significantly from expected results are called *tests of hypotheses*, *tests of significance*, or *decision rules.*

Example: If 20 tosses of a coin yield 16 heads, we might reject the hypotheses that the coin is fair, although we might be wrong.

- If we reject a hypothesis when it happens to be true, a *Type I error* has been made.
- On the other hand, if we accept a hypothesis when it should be rejected, a *Type II error* has been made.

Note: It is not simple to minimize errors of decision. The only way to reduce both types of error is to increase the sample size, which may or may not be possible.

Level of Significance

- In testing a given hypothesis, the maximum probability with which we would be willing to risk a *type I error* is called the *level of significance* of the test.
- This probability is often specified before any samples are drawn so that it will not influence our decision.

Example: if a 0.05 or 5% level significance is chosen in designing a test of a hypothesis, then there are about 5% that we would reject the hypothesis when it should be accepted.

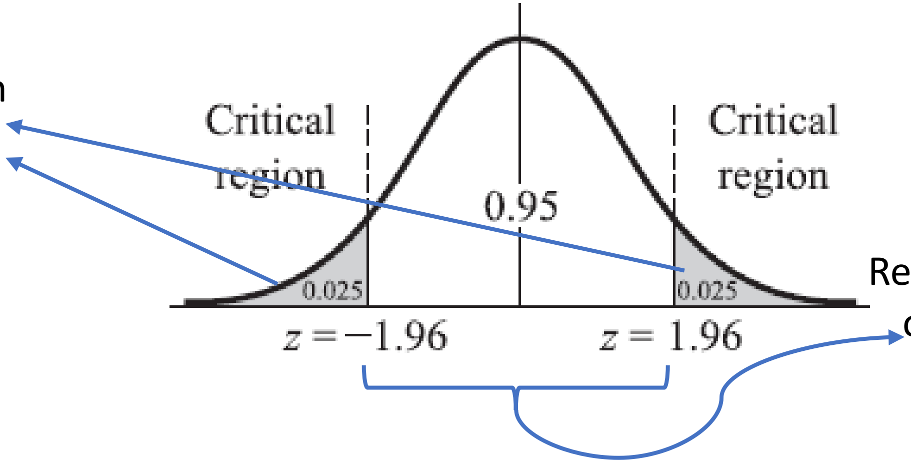
In other words, we say that the hypothesis has been rejected at a 0.05 level of significance, which mean that we could be wrong with probability 0.05.

- In practice, a level of significance of 0.05 or 0.01 is customary.

Tests involving the Normal Distribution

Suppose that under a given hypothesis, the sampling distribution of a statistic S is a normal distribution with mean μ_S and standard deviation σ_S.

Also, suppose we decide to reject the hypothesis if S is either too small or too large.



Region of rejection of the hypothesis

Region of acceptance of the hypothesis

We can be 95% confident that, if the hypothesis is true, the z score of an actual sample statistic S will lie in (-1.96, 1.96).

The shaded area 0.05 is the level of significance of the test. It represents the probability of our being wrong in rejecting the hypothesis (Type I error)

Tests of Hypotheses and Significance Example

Example:

a) Find the probability of getting between 40 and 60 heads inclusive in 100 tosses of a fair coin.

$\mu = np = 50 \ \ \sigma = \sqrt{npq} = 5$. Both greater than 5, we can use normal approximation. On continuous scale, between 40 and 60 is the same as 39.5 and 60.5.

Required probability = area btw -2.1 and 2.1 = 0.9642

b) To test the hypothesis that a coin is fair, the following decision rules are adopted: *(1) Accept the hypothesis if the number of heads in a single sample of 100 tosses is btw 40 and 60 inclusive, (2) reject the hypothesis otherwise*. Find the probability of rejecting the hypothesis when it is actually correct.

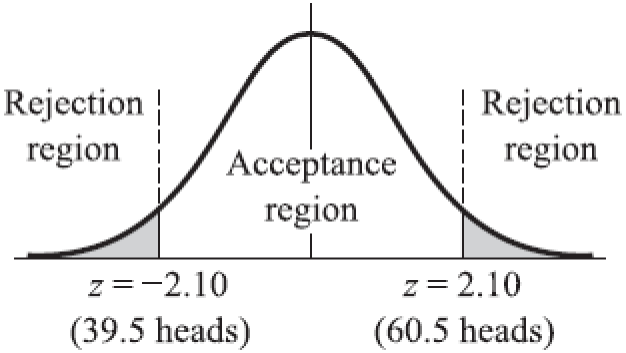
The probability of not getting btw 40 and 60 heads inclusive if the coin is fair equal 1-0.9642=0.0358. Then the probability of rejecting the hypothesis when it is correct equals 0.0358.

Tests of Hypotheses and Significance Example

Example (ctn.):

c) Interpret graphically the decision rule and the result of (b).

The decision rule is illustrated in the below figure, which shows the probability distribution of heads in 100 tosses of a fair coin.



If a single sample of 100 tosses yields a z score btw -2.1 and 2.1, we accept the hypothesis; otherwise, we reject it and decide that the coin is not fair.

Probability of making *type I error*: 0.0358. This probability is represented by the total shaded area and is called the *level of significance*.

Tests of Hypotheses and Significance Example

Example (ctn.):

d) What conclusion would you draw if the sample of 100 tosses yielded 53 heads? 60 heads?

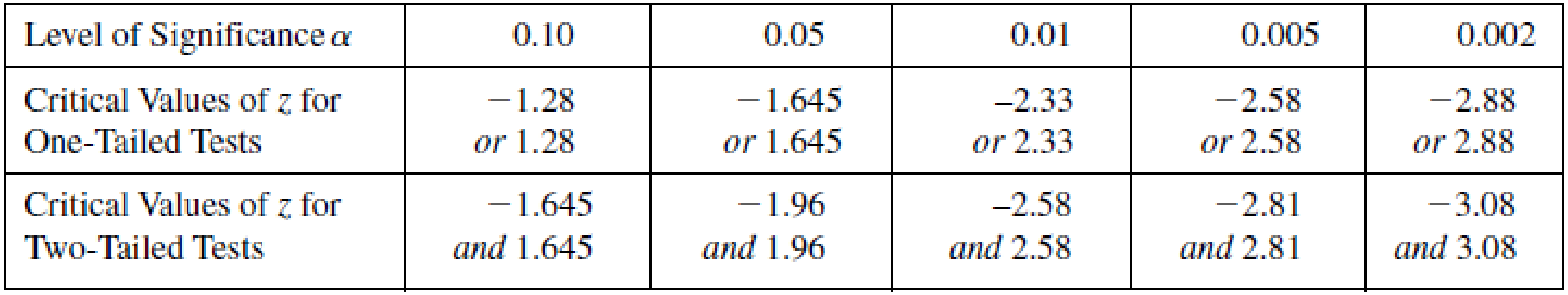
According to the decision rule, we would have to accept the hypothesis that the coin is fair in both cases. We might argue that if only one more head had been obtained, we would have rejected the hypothesis. This is what we must face when any sharp line of division is used in making decisions.

e) Could we be wrong in our conclusion to (d)?

Yes. We could accept the hypothesis when it actually should be rejected, as would be the case, for example, when the probability of heads is really 0.7 instead of 0.5.

One-tailed and Two-tailed Tests

- In previous example, we consider the extreme values of the statistic *S* on both side of the mean (both tails of the distribution) → called *two-tailed tests* or *two-sided tests*.
- If we consider extreme values only to one side of the mean (one tail of the distribution) → called *one-tailed tests* or *one-sided tests*. In this case, the *region of rejection of the hypothesis* is a region to one side of the distribution.



Level of Significance α	0.10	0.05	0.01	0.005	0.002
Critical Values of z for One-Tailed Tests	−1.28 *or* 1.28	−1.645 *or* 1.645	–2.33 *or* 2.33	−2.58 *or* 2.58	−2.88 *or* 2.88
Critical Values of z for Two-Tailed Tests	−1.645 *and* 1.645	−1.96 *and* 1.96	–2.58 *and* 2.58	−2.81 *and* 2.81	−3.08 *and* 3.08

Critical values of z for both one-tailed and two-tailed tests at various levels of significance

P value

- P-value is used as another way to determine if we can reject/accept the null hypothesis.
- P-value is the probability for the null hypothesis to be true.
- Checking condition:
  - ✓If p-value < α (the level of significance), we reject the null hypothesis.
  - ✓If p-value $\geq \alpha$ (the level of significance), we accept the null hypothesis.

P value

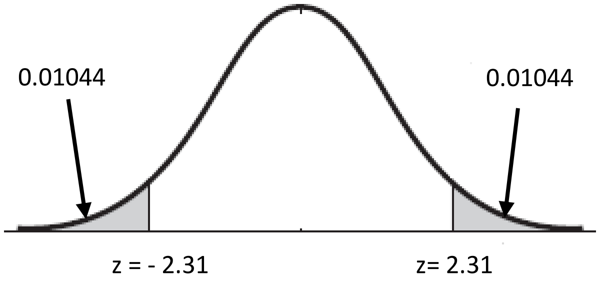
- In most of the tests we will consider, the null hypothesis H_0 will be an assertion that a population parameter has a specific value, and the alternative hypothesis H_1 will be one of the following assertions:
  - i. The parameter is greater than the stated value (right-tailed test).
  - ii. The parameter is less than the stated value (left-tailed test).
  - iii. The parameter is either greater than or less than the stated value (two-tailed test).

P value

Example: The average weight of all resident in town XYZ is 168 lbs. A nutritionist believes the true mean to be different. She measured the weight of 36 individuals and found the mean to be 169.5 lbs with a standard deviation of 3.9.

a) State the null and alternative hypotheses.

b) At a 95% confidence level, is there enough evidence to discard the null hypothesis?



Answers:

a) $H_0 = 168$ and $H_1 \neq 168$.

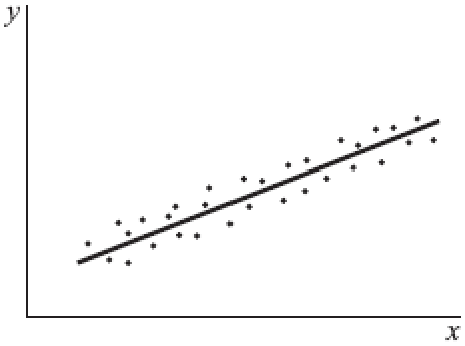
b) $n = 36, \bar{X} = 169.5, s = 3.9, C = 0.95, \alpha = 0.05$

$$z_c = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{169.5 - 168}{3.9/\sqrt{36}} = 2.31$$

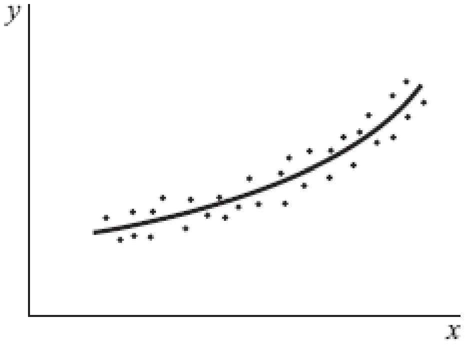
P-value = 0.02088 which is less than $\alpha = 0.05$, so we reject the null hypothesis that the average mean weight of all resident is 168 lbs.

Curve Fitting

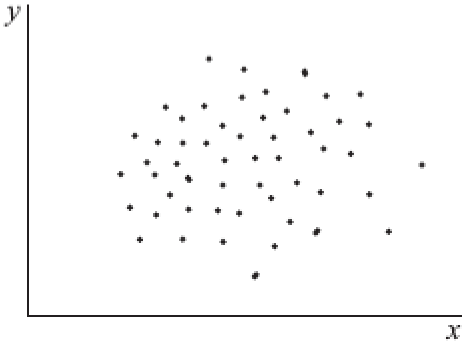
- Definition: expressing the relationship exists between 2 (or more) variables in mathematical form by determining an equation connecting the variables.
- Steps of curve fitting process:
  - Showing corresponding values of the variables.
  - Plotting the points on a rectangular coordinate system (scatter diagram)
  - Visualizing a smooth curve approximating the data. (approximating curve)



Linear relationship



Nonlinear relationship



No relationship exists

Curve Fitting

- In general, for a linear relationship, we could use a straight line:

$$y = a + bx$$

- For a nonlinear relationship, we could use *parabola* or *quadratic curve*:

$$y = a + bx + cx^2$$

- Sometimes it helps to plot scatter diagrams in terms of transformed variables.

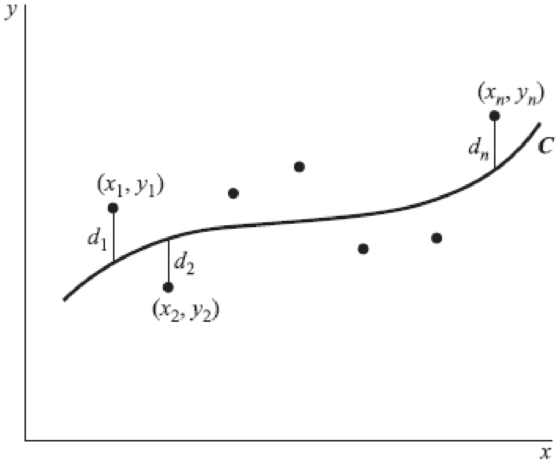
Example: If log y versus x leads to a straight line, we can try $\log y = a + bx$ as an equation for the approximating curve.

Regression

- One of the main purposes of curve fitting is to estimate one of the variables (the *dependent variable*) from the other (the *independent variable*).
- The process of estimation is often referred to as *regression*.
- If *y* is to be estimated from *x* by means of some equation, we call the equation a *regression equation of y on x*, and the corresponding curve a *regression curve of y on x*.

The Method of Least Squares

- This method is use as a measure of the goodness of fit of an approximating curve to a set of data.



- $d_1, d_2, \dots, d_n$ is called deviation, or error, or residual and may be positive, negative, or zero.
- A curve having the property that:

$$d_1^2 + d_2^2 + \cdots + d_n^2 = a\ minimum$$

is called a best-fitting curve.

A curve having this property is said to fit the data in the *least-squares sense*, and is called *least-square regression curve*.

The Least-Squares Line

- The least-squares line approximating the set of points $(x_1, y_1), \ldots (x_n, y_n)$ has the equation:

$$y = a + bx$$

Where the constant a and b are determined by solving simultaneously the equations:

$$\sum y = an + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Which are called the *normal equations* for the least-squares line.

The Least-Squares Line

- The values of a and b will be:

$$a = \frac{(\sum y)\,(\sum x^2) - (\sum x)(\sum xy)}{n\sum x^2 - (\sum x)^2} \qquad b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

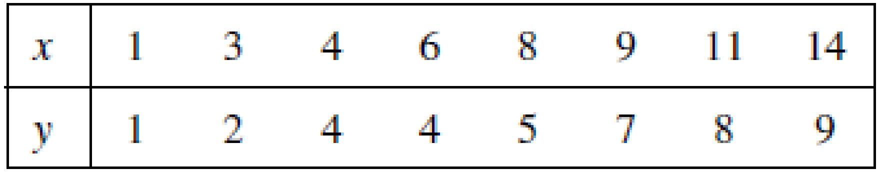
- The value for b can also be written as:

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

since $\bar{x} = (\sum x)$/n

The Least-Squares Line

Example: Find the least square line to the following data, using *x* as independent variable.



x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

The equation of the line is *y*=*a*+*bx*. Using normal equations to find *a* and *b*.

$$\sum y = an + b\sum x$$

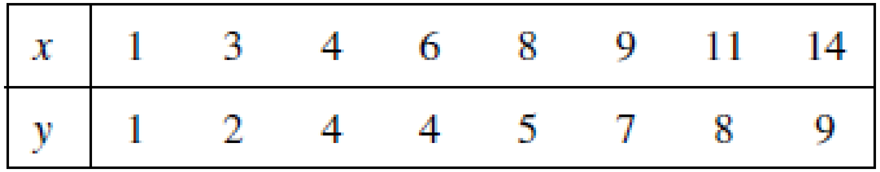
$$\sum xy = a\sum x + b\sum x^2$$

$$8a + 56b = 40\ and\ 56a + 524b = 364$$

We got $a = 6/11$ and b= $7/11$.

In-class Exercise

Find the least square line to the following data, using *x* as dependent variable.



x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

The Least-Squares Line

If we divide both side of the normal equation $\sum y = an + b\sum x$ by n, we will get:

$$\bar{y} \; = a + b\bar{x}$$

This yields: a $= \bar{y} - b\bar{x}$.

Then if we replace this value of a to the original least-squares line $y = a + bx$, we will get:

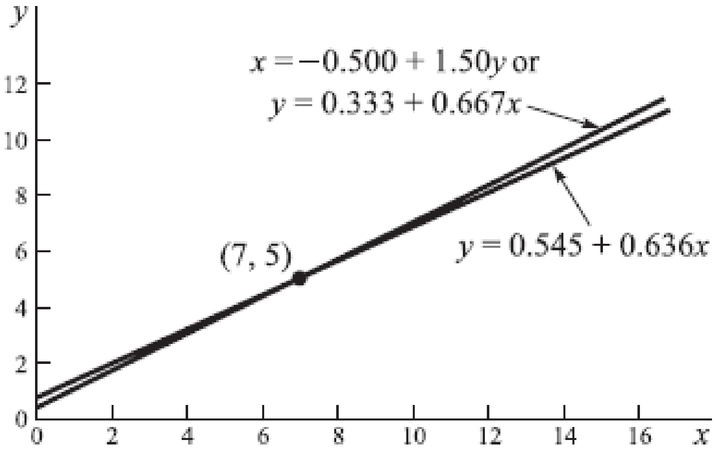
$$y - \bar{y} = b(x - \bar{x}) \quad or \qquad y - \bar{y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}(x - \bar{x})$$

The above equation shows that the least-squares line passes through the point ($\bar{x}$, $\bar{y}$), which is called the *centroid* or *center of gravity* of the data.

The Least-Squares Line

Example: The centroid of previous examples.

$$\bar{x} = \frac{\sum x}{n} = \frac{56}{8} = 7, \qquad \bar{y} = \frac{\sum y}{n} = \frac{40}{8} = 5$$



The Least-Squares Parabola

The least-squares parabola that fits a set of sample points is given by:

$$y = a + bx + cx^2$$

Where a, b, c are determined from the normal equations:

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

Multiple Regression

We use multiple regression when working on 3 or more variables.

For example, if there is a linear relationship between a dependent variable *z* and two independent variables *x* and *y*, then we would seek an equation connecting the variables that has the form:

$$z = a + bx + cy$$

This is called regression equation of *z* on *x* and *y*. It represents a plane in three-dimensional rectangular coordinate system (regression plane).

Where a, b, c are determined from the normal equations:

$$\sum z = na + b\sum x + c\sum y$$

$$\sum xz = a\sum x + b\sum x^2 + c\sum xy$$

$$\sum yz = a\sum y + b\sum xy + c\sum y^2$$

Standard Error of Estimate

If we denote the estimated value of *y* for a given value of *x*, as obtained from the regression curve of *y* on *x*, then a measure of the scatter about the regression curve is supplied by the quantity:

$$s_{y,x} = \sqrt{\frac{\sum(y - y_{est})^2}{n}}$$

Which is called the *standard error of estimate of y on x*.

Note: $\sum(y - y_{est})^2 = \sum d^2$

The Linear Correlation Coefficient

The correlation coefficient, *r*, measures *how well* the least-squares regression line fits the sample data.

It can be computed from either of the results:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

or

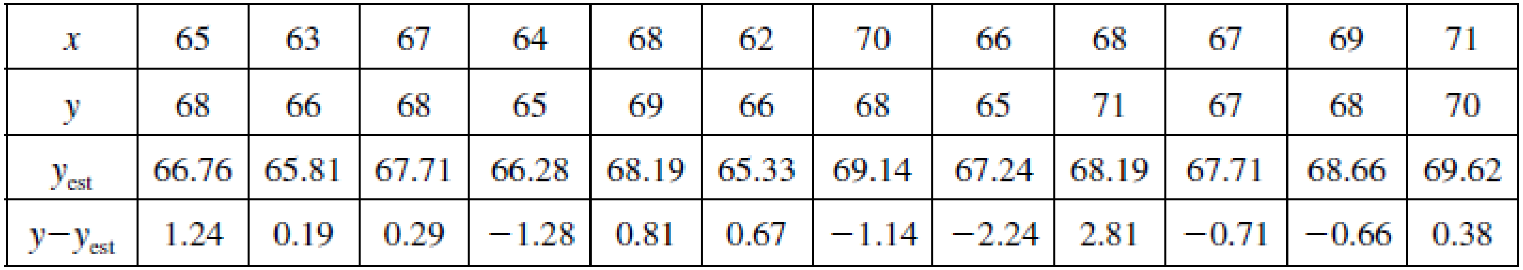
$$r^2 = \frac{\sum(y_{est}-\bar{y})^2}{\sum(y-\bar{y})^2}$$

Formula equivalent to those above, which are often used in practice, are:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

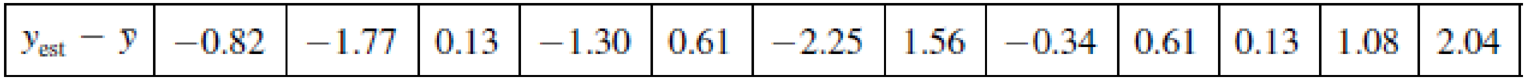
The Linear Correlation Coefficient

Example: Given the following information, find the correlation coefficient.



x	65	63	67	64	68	62	70	66	68	67	69	71
y	68	66	68	65	69	66	68	65	71	67	68	70
y_{est}	66.76	65.81	67.71	66.28	68.19	65.33	69.14	67.24	68.19	67.71	68.66	69.62
$y - y_{\text{est}}$	1.24	0.19	0.29	−1.28	0.81	0.67	−1.14	−2.24	2.81	−0.71	−0.66	0.38

We need to find $y_{est} - \bar{y}$ and y $- \bar{y}$ where $\bar{y} = 67.58$



$y_{\text{est}} - \bar{y}$	−0.82	−1.77	0.13	−1.30	0.61	−2.25	1.56	−0.34	0.61	0.13	1.08	2.04

$$r^2 = \frac{\sum(y_{est} - \bar{y})^2}{\sum(y - \bar{y})^2} = 0.4938$$

So $r = \pm 0.7027$. Since y_{est} increases as *x* increases, the correlation is positive, and $r = 0.7027$.

Reference

Notes, equations, and figures in the lecture are based on or taken from materials in the course textbook:

“Probability and Statistics”, by Spiegel, Schiller and Srinivasan, ISBN 987-007-179557-9 (McGraw-Hill/Schaun’s)