Chapter 3: Simple Linear Regression, Part 1

Dr.Ting Wei

Section 3.1:
Relationship between variables

Relationship between random variables.

- Two random variables can be independent or dependent.
- If they are dependent, the relationship between them can be strong or weak.
- Question: how can we find a numerical measure of the strength of a relationship between two random variables?

Question

Consider the following two experiments:

- Experiment 1: measure X and Y where
  $X =$ weight of a sample of water,
  $Y =$ volume of the same sample of water.

- Experiment 2: measure X and Y where
  $X =$ body weight of a human,
  $Y =$ same human height.

- Question: in which experiment do you think the relationship between X and Y be strongest?

Measuring strength of relationship.

- Main idea: the covariance and correlation are two measures that quantify the difference in strength of a relationship between two random variables.

Covariance

- Let X, Y be random variables with means $E(X) = \mu_X$ and $E(Y) = \mu_Y$. The covariance between them, often denoted σ_{XY}, is defined as

$$\sigma_{XY} = \text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

- Note: $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$.

The correlation coefficient

- Problem with the covariance: it is measured in units of X times units of Y. We want to make this scale-free. We do this by introducing the coefficient of correlation.

- The coefficient of correlation is defined by

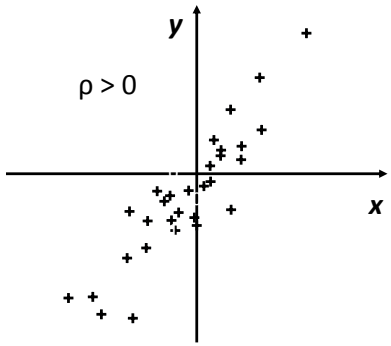
$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

and

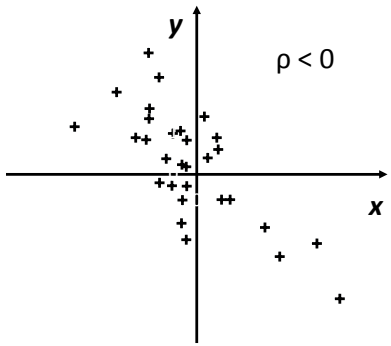
$$-1 \leqslant \rho \leqslant 1$$

The sign of the correlation coefficient

- The sign of ρ is the sign of $\sigma_{XY} = \text{cov}(X, Y) = E[(X-\mu_X)(Y-\mu_Y)]$.



$\rho > 0$: we tend to have more positive values of $(X-\mu_X)(Y-\mu_Y)$: X and Y tends to increase together.



$\rho < 0$: we tend to have more negative values of $(X-\mu_X)(Y-\mu_Y)$: as X increases, Y tends to decrease and vice-versa.

Estimating the correlation coefficient

- We make n pairs of observations

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

of random variables X and Y.

- We want an estimate $\hat{\rho}$ of ρ from these observations.
- By definition, $\rho = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$, therefore

$$\hat{\rho} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}$$

Estimating the correlation coefficient, cont.

- The estimates of σ_X, σ_Y, σ_{XY} are:

$$\hat{\sigma}_X^2 = s_X^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2 \ ,$$

$$\hat{\sigma}_Y^2 = s_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2 \ ,$$

$$\hat{\sigma}_{XY} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \ .$$

- Therefore

$$\hat{\rho} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- This is called the sample correlation coefficient between X and Y.

Some useful notation and formulas

- We denote

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n},$$
$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n},$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}.$$

- Therefore, in this notation

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Example: Pulse Data

- The following data gives results from 11 people in three columns: first pulse rate (at rest), second pulse rate (after exercise) and smoking indicator (1=smokes regularly, 2=does not smoke regularly).

First pulse rate	Second pulse rate	Smoking indicator
96	140	2
62	100	2
78	104	1
82	100	2
100	115	1
68	112	2
96	116	2
78	118	2
88	110	1
62	98	1
80	128	2

Example: Pulse Data, cont.

- Let X = first pulse rate, Y = second pulse rate

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$
$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$
$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$
$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$\sum_{i=1}^{n} x_i = 890$; $\sum_{i=1}^{n} x_i^2 = 73780$; $S_{xx} = 1770.909$.
$\sum_{i=1}^{n} y_i = 1241$; $\sum_{i=1}^{n} y_i^2 = 141653$; $S_{yy} = 1645.636$.
$\sum_{i=1}^{n} x_i y_i = 101404$; $S_{xy} = 995.818$.

Therefore

$$\hat{\rho} = \frac{995.818}{\sqrt{1770.909 \times 1645.636}} = 0.583$$

Correlation as a measure of linear relationship

- Theorem: *The sample correlation coefficient satisfies the inequality* $|\hat{\rho}| \leq 1$, *with equality if and only if the points* $\{(x_i, y_i)\}$ *lie on a straight line.*

- Therefore correlation is a measure of linear association

$\hat{\rho} = 0$	no linear association.
$\hat{\rho} = 1$	perfect positive linear association; x and y increase together.
$\hat{\rho} = -1$	perfect negative linear association; y decreases as x increases.

The meaning of zero correlation

- Let $X \sim N(\mu, \sigma^2)$ and let $Y = (X - \mu)^2$. Then $E(Y) = \text{var}(X) = \sigma^2$.

- Since $E(X-\mu)=E[(X - \mu)^3]=0$ we have

$$\text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] =$$
$$= E\left[(X - \mu)\left((X - \mu)^2 - \sigma^2\right)\right] = 0$$

And thus, $\rho = 0$.

Inferences about the population correlation

- The sample correlation coefficient $\hat{\rho}$ is an estimate of the population correlation coefficient ρ.
- Hence, if $\rho = 0$ is true, we expect $\hat{\rho} \approx 0$.
- We want to test the hypotheses $H_0 : \rho = 0$, and compute a confidence interval for ρ.

Testing the hypothesis $\rho = 0$

- We want to test the hypothesis: $H_0 : \rho = 0$, $H_1 : \rho \neq 0$.
- We can use the t-statistics

$$\mathrm{T} = \hat{\rho}\sqrt{\frac{n-2}{1-\hat{\rho}^2}}$$

- Under the null hypothesis this has a t distribution with $n-2$ degrees of freedom.

Example: Pulse data

- We computed the sample correlation coefficient for the Pulse data and found it to be $\hat{\rho} = 0.583$.
- Using the formula $\mathrm{T} = \hat{\rho}\sqrt{\frac{n-2}{1-\hat{\rho}^2}}$ with $\hat{\rho} = 0.583$ and $n = 11$, we obtain

$$\mathrm{T} = 0.5833\sqrt{\frac{9}{1 - 0.5833^2}} = 2.1545 \ .$$

- From the table of the t-distribution with 9 df, we see that critical region is $(-\infty, -2.262) \cup (2.262, +\infty)$. Hence the hypothesis $\rho = 0$ is acceptable.

Confidence intervals for ρ

- It can be shown that the 95% confidence interval for ρ is given by (l_1, l_2) where

$$l_1 = \frac{e^{-\frac{2\times 1.96}{\sqrt{n-3}}} \cdot \frac{1+\hat{\rho}}{1-\hat{\rho}} - 1}{e^{-\frac{2\times 1.96}{\sqrt{n-3}}} \cdot \frac{1+\hat{\rho}}{1-\hat{\rho}} + 1}, \qquad l_2 = \frac{e^{\frac{2\times 1.96}{\sqrt{n-3}}} \cdot \frac{1+\hat{\rho}}{1-\hat{\rho}} - 1}{e^{\frac{2\times 1.96}{\sqrt{n-3}}} \cdot \frac{1+\hat{\rho}}{1-\hat{\rho}} + 1}$$

- Thus if we repeat the experiment many times, the CI covers the true value of ρ with 95% probability.

Pulse data

- For the pulse data we calculated $\hat{\rho} = 0.5833$; also $n = 11$.
- Hence 95% confidence interval is

$$(l_1,\, l_2) = \left(\frac{\left(\frac{1+0.5833}{1-0.5833}\right) \exp\left(\frac{-2\times 1.96}{\sqrt{8}}\right) - 1}{\left(\frac{1+0.5833}{1-0.5833}\right) \exp\left(\frac{-2\times 1.96}{\sqrt{8}}\right) + 1}, \right.$$
$$\left. \frac{\left(\frac{1+0.5833}{1-0.5833}\right) \exp\left(\frac{2\times 1.96}{\sqrt{8}}\right) - 1}{\left(\frac{1+0.5833}{1-0.5833}\right) \exp\left(\frac{2\times 1.96}{\sqrt{8}}\right) + 1} \right)$$
$$= (-0.026,\ 0.877)\ .$$

- Note: the interval contains 0, as expected.

Summary

- Coefficient of correlation $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$.
- Estimation: ρ is estimated by the sample correlation coefficient $\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$.
- Theorem: $|\hat{\rho}| \leq 1$ and $|\hat{\rho}| = 1$ iff $\{(x_i, y_i)\}$ lie on a straight line. Thus $\hat{\rho}$ (and ρ) is a measure of linear association.
- We showed in an example that zero covariance between two random variables does not imply their independence.
- We discussed the hypothesis test $\rho = 0$ using the t-statistics, and we discussed how to compute the corresponding confidence interval.
- Example: Pulse data.

Section 3.2: Least Squares Estimation

Introduction to regression: response versus explanatory variables

- Regression analysis is used for explaining or modelling the relationship between a single variable Y, called the response variable, and one or more variables $X_1, \ldots, X_k$, called predictors or explanatory variables.
- In regression analysis we treat the X variables as fixed constants. We are interested in the effect of the predictors on the response.
- If $k = 1$ we have a simple regression model.

Correlation versus simple regression

- Regression is different from correlation.
- Correlation describes the joint distribution between random variables X and Y. The correlation between X and Y is the same as the correlation between Y and X.
- In simple regression, one of the variables (Y) is the response, which we predict from the X. So regression describes the conditional distribution of Y given X.

Introduction to regression: objectives

- Possible objectives of regression analysis include:
  - Prediction of future observations.
  - Assessment of the effect of, or relationship between, explanatory variables and the response.
  - General description of data structure.

Linear models

- Suppose we want to model the response Y in terms of three predictors X_1, X_2, X_3. One general form of the model could be

$$Y = f(X_1, X_2, X_3) + \varepsilon$$

where f is an unknown function and ε is an error term.

- In a linear model , the function f has the following form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where β_i, $i = 0, 1, 2, 3$, are unknown parameters.

The simple linear regression model

- The simple linear regression model has the form

$$Y = a + bX + \varepsilon$$

where ε is a random variable with a specified probability distribution with mean 0.

- It follows that $E(Y) = a + bX$.

- We call $a + bX$ the systematic (or deterministic) part of the model, while ε is the random component.

Assumptions in the simple linear regression model

- In the simple linear regression model $Y = a + bX + \varepsilon$ we assume

  a) ε has a normal distribution with $E(\varepsilon) = 0$

  b) $\text{var}(\varepsilon) = \sigma^2$ is independent of X.

- If n independent observations are made on this model, we have n independent random variables

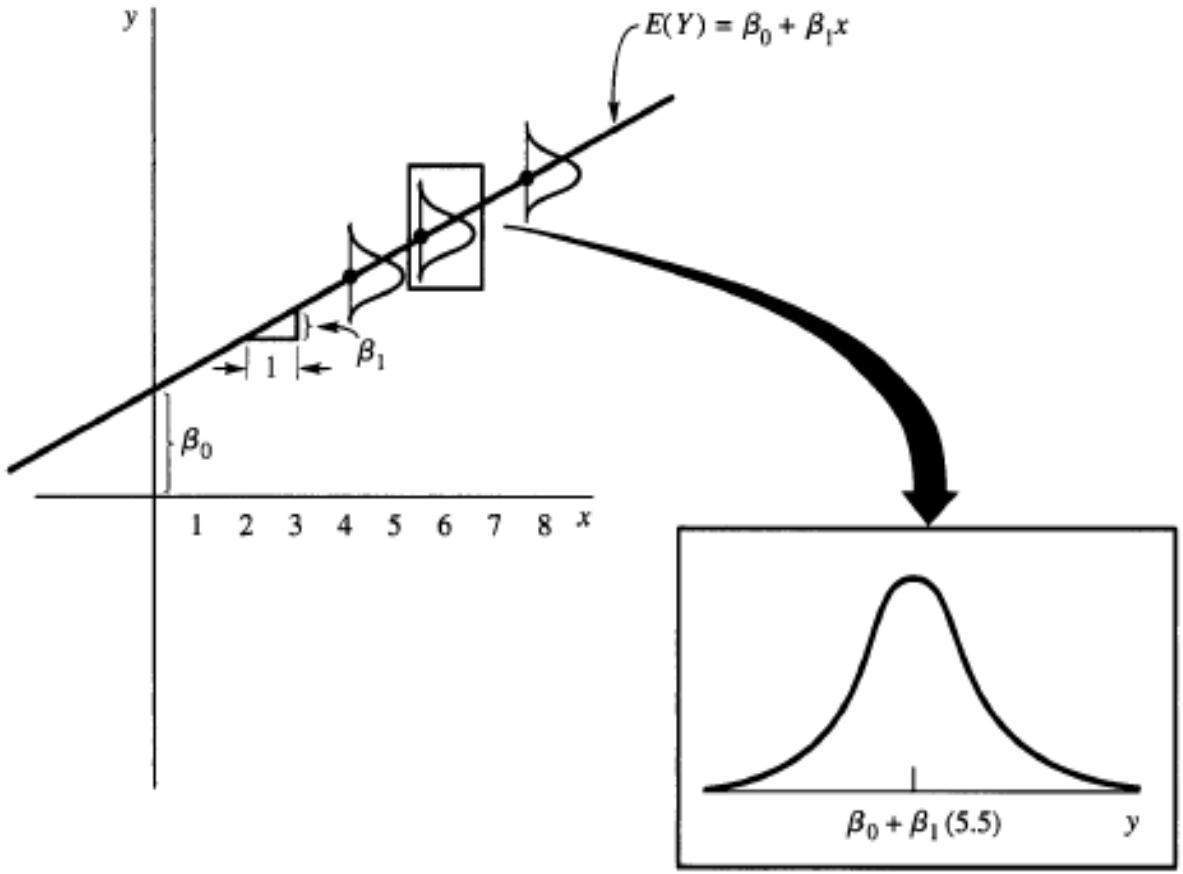
$$Y_i = a + bX_i + \varepsilon_i \qquad i = 1, \ldots, n$$

  where ε_i are independent random variables with normal distribution $N(0, \sigma^2)$

- Equivalently: $Y_i \sim N(a + bX_i, \sigma^2)$.

Graphical representation

For each X there is a population of possible values of Y, which is normally distributed with mean $a + bX$ and variance σ^2.



Estimation of parameters: main aim

- Our main aim is to estimate the model parameters a and b. That is, to produce the line $E(Y) = \hat{a} + \hat{b}X$ which best fits the data points.

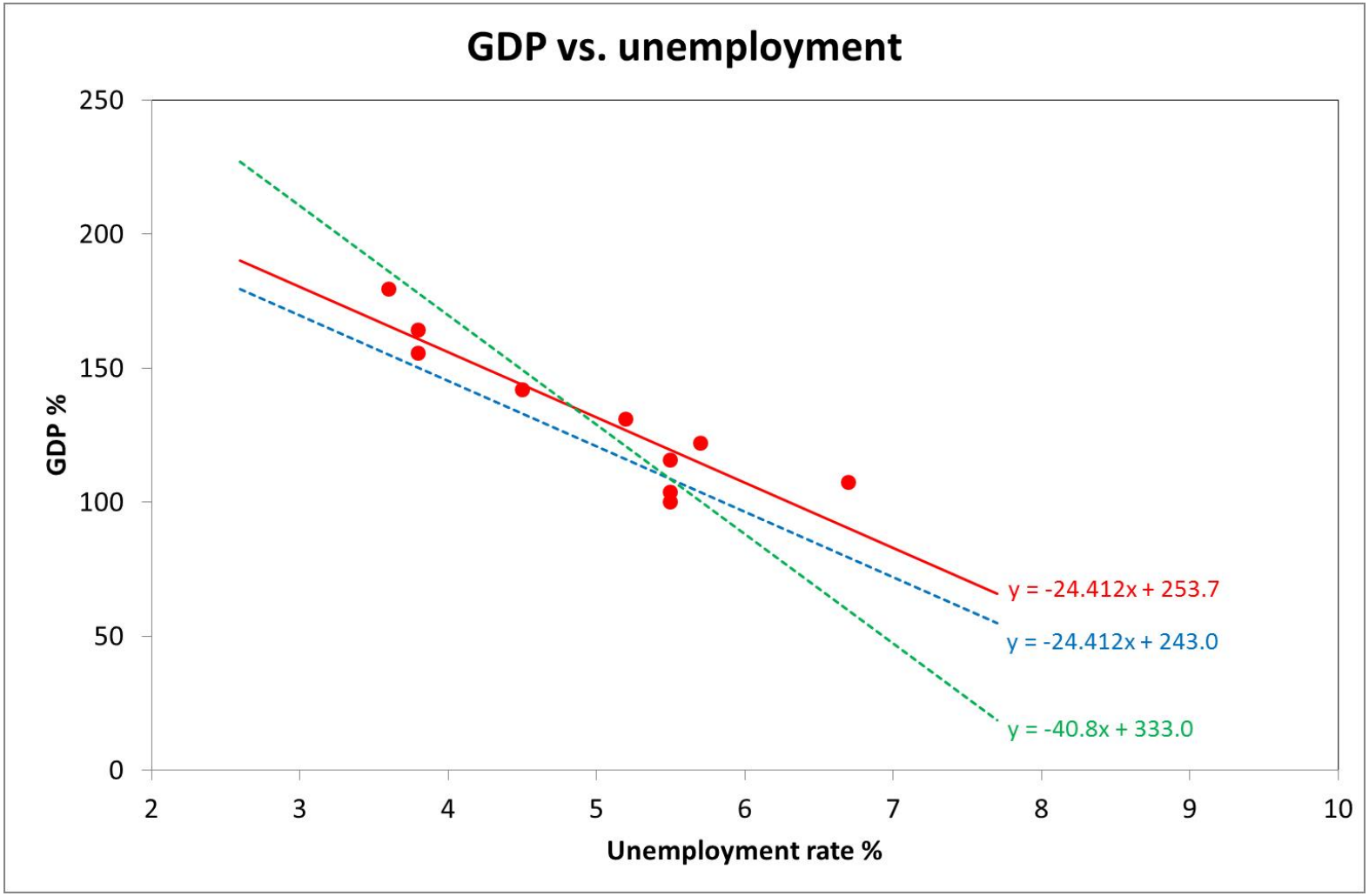
- Then $\hat{y}_i = \hat{a} + \hat{b}x_i$ is called the i^{th} fitted value. It is the predictor of y_i given the value x_i.

Example: Unemployment rate and Gross Domestic Product (GDP)

- The following table shows the behaviour of the Gross Domestic Product (GDP) in USA, for the decade 1960-1969, as a function of the unemployment rate percentage. GDP is expressed in percent of the GDP value of the initial year 1960.

Year	Unempl. rate % (x)	GDP % (y)
1	5.5	100.0
2	5.5	103.9
3	6.7	107.5
4	5.5	115.6
5	5.7	121.9
6	5.2	130.9
7	4.5	141.9
8	3.8	155.6
9	3.8	164.4
10	3.6	179.6

Question



Question: Which of the three lines best fit the data, and why?

MLE for simple linear regression

- Recall that in the simple linear regression model $Y_i \sim N(a + bX_i, \sigma^2)$.

- Likelihood $= \mathcal{L}(a, b, \sigma^2 | \text{data}) =$

$$= \prod_{i=1}^{n} f(y_i | a, b, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{-(y_i - a - bx_i)^2}{2\sigma^2} \right\}$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \right\} .$$

- To maximize the likelihood we need to minimize

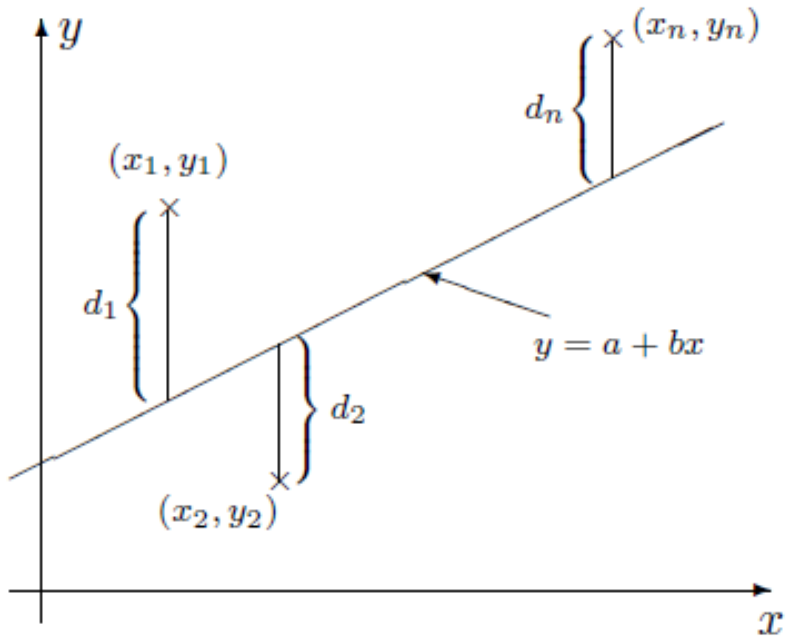
$$SS = \sum_{i=1}^{n} (y_i - a - bx_i)^2 .$$

- Thus the maximum likelihood estimates are those values which minimise the sum of squared distances between the observed and fitted values.

Least square estimation

- The estimates obtained by minimizing SS are also called the least square estimates.

- The line $\hat{a} + \hat{b}x$ is the one that minimizes the sum of squares of the vertical distances between the observed points and the line



Minimizing *SS*

- $SS = \sum_{i=1}^{n} (y_i - a - bx_i)^2$ is a function of the two variables a and b.
- To minimize it, we need to equate to zero the partial derivatives

$$\frac{\partial SS}{\partial a} = \sum_{i=1}^{n} 2(-1)(y_i - a - bx_i) \quad \text{and}$$

$$\frac{\partial SS}{\partial b} = \sum_{i=1}^{n} 2(-x_i)(y_i - a - bx_i) \; .$$

Some calculations

- $\frac{\partial SS}{\partial a} = \sum_{i=1}^{n} 2(-1)(y_i - a - bx_i) = 0$ implies

$$\sum_{i=1}^{n} y_i = n\hat{a} + \hat{b}\sum_{i=1}^{n} x_i \quad \text{or} \quad \bar{y} = \hat{a} + \hat{b}\bar{x} \ .$$

- $\frac{\partial SS}{\partial b} = \sum_{i=1}^{n} 2(-x_i)(y_i - a - bx_i) = 0$ implies

$$\sum_{i=1}^{n} x_i y_i = \hat{a}\sum_{i=1}^{n} x_i + \hat{b}\sum_{i=1}^{n} x_i^2 \ .$$

- These equations are called the likelihood equations.

Some calculations, cont.

- From previous slide

$$\bar{y} = \hat{a} + \hat{b}\bar{x} \ , \qquad \sum_{i=1}^{n} x_i y_i = \hat{a}\sum_{i=1}^{n} x_i + \hat{b}\sum_{i=1}^{n} x_i^2 \ .$$

- Therefore

$$\sum_{i=1}^{n} x_i y_i = (\bar{y} - \hat{b}\bar{x})\sum_{i=1}^{n} x_i + \hat{b}\sum_{i=1}^{n} x_i^2 = n\bar{y}\bar{x} - \hat{b}\bar{x}^2 n + \hat{b}\sum_{i=1}^{n} x_i^2$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \hat{b}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = \hat{b}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Equivalently

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \hat{b}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Some calculations, cont.

- From previous slide:

$$\bar{y} = \hat{a} + \hat{b}\bar{x} \ , \qquad \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \hat{b}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ .$$

- In conclusion, the MLE of a and b are

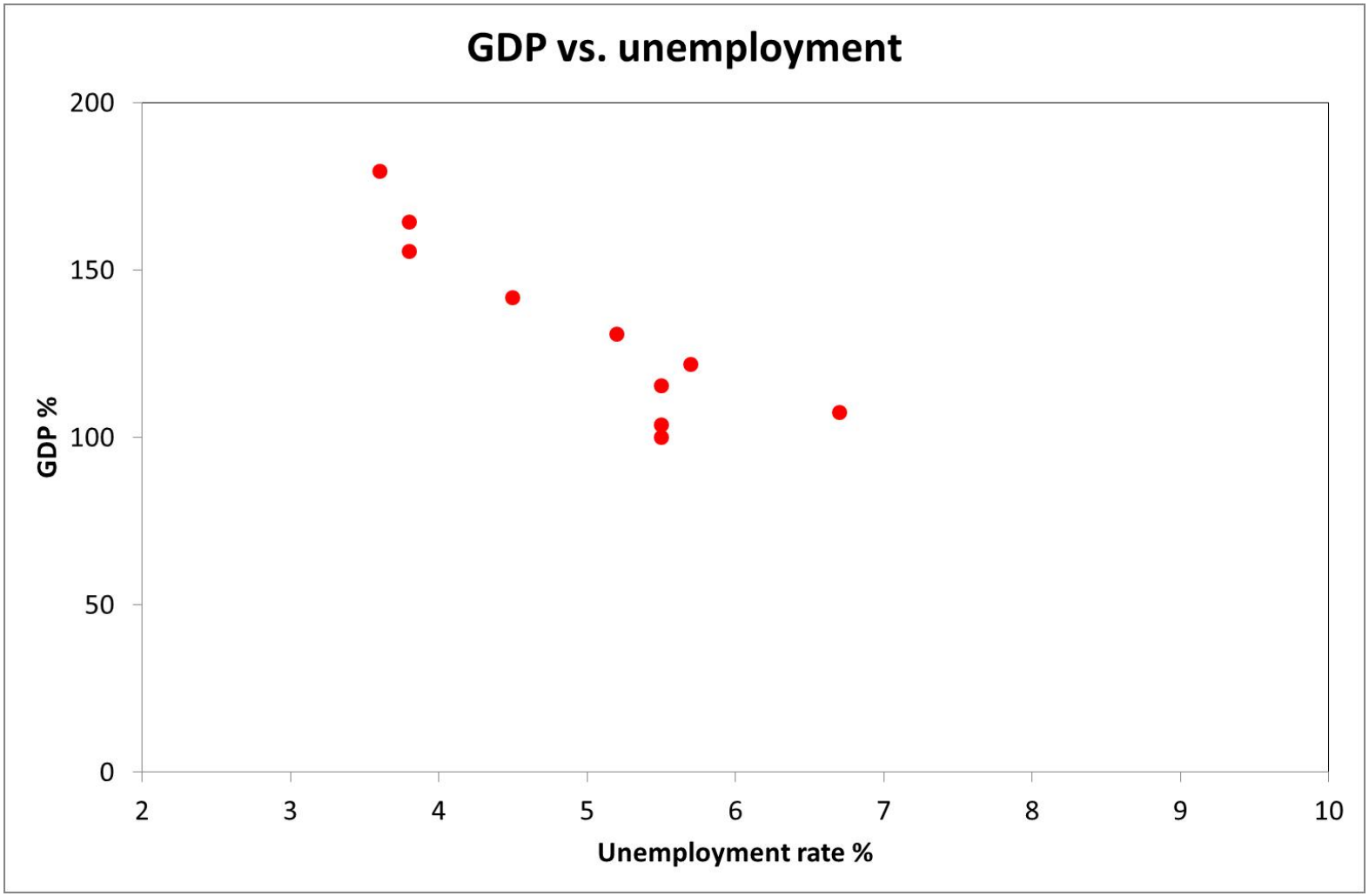
$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \ .$$

Residuals

- The i^{th} residual is the difference between the i^{th} observed and fitted values

  $r_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i$.

Example: Unemployment and GDP, cont.



Plot of the GDP against the unemployment rate rate

Example: Unemployment and GDP, cont.

- We can calculate as done in lecture 10:

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} = -225.954$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 9.256$$

$$\sum_i x_i = 49.8\,, \quad \bar{x} = 4.98\,, \quad \sum_i y_i = 1321.3\,, \quad \bar{y} = 132.13$$

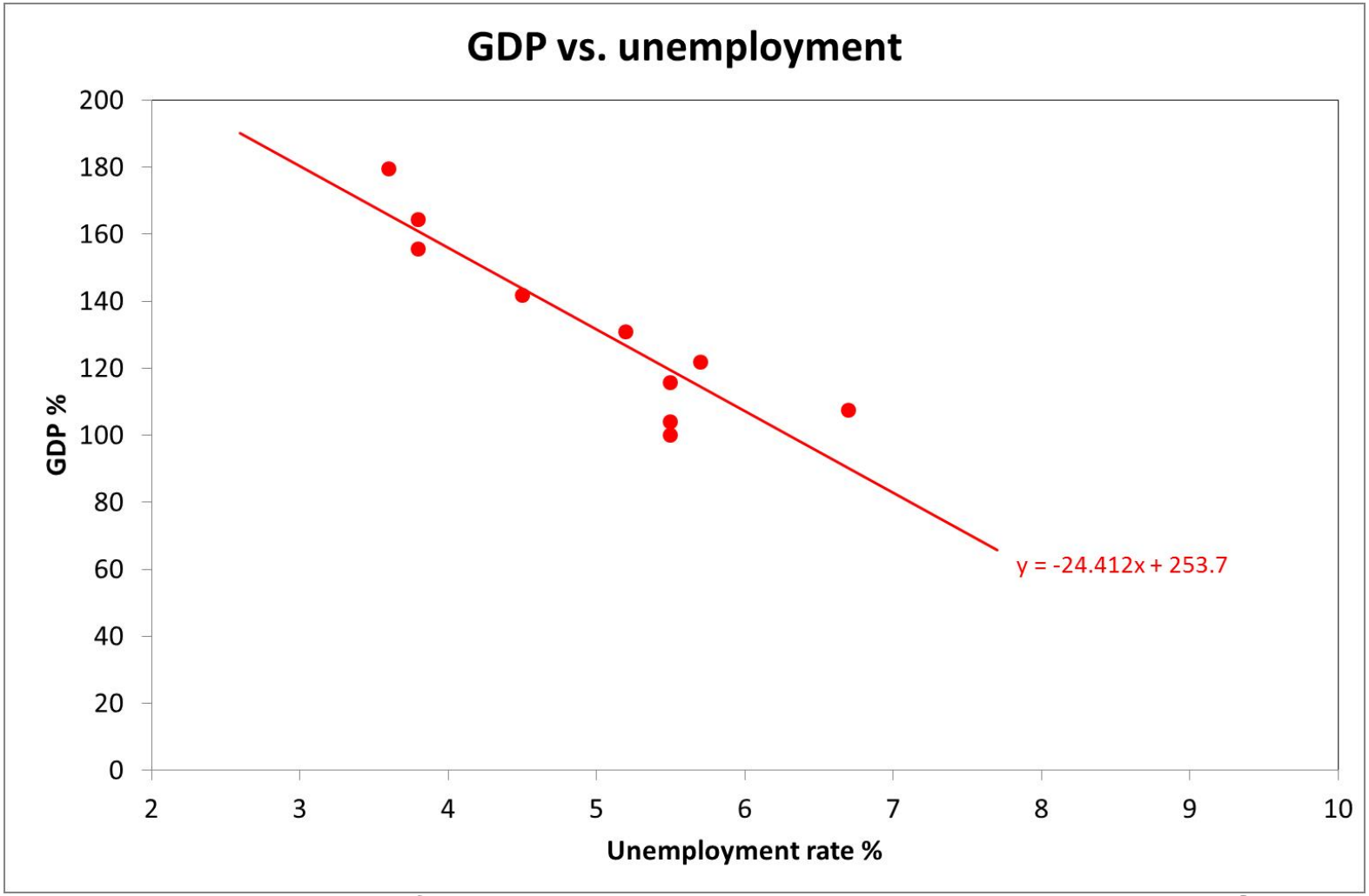
- Therefore

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = -24.4116 \quad \hat{a} = \bar{y} - \hat{b}\bar{x} = 253.7$$

- The regression line is

$$y = \hat{a} + \hat{b}x = 253.7 - 24.4116x\ .$$

Example: Unemployment and GDP, cont.



Regression line of the unemployment rate against the GDP percentage.

Example: Unemployment and GDP, cont.

- Calculating and comparing RSS for the three lines.

$y = -24.412x + 253.7$ Equation A (regression)

$y = -24.412x + 243.0$ Equation B

$y = -40.8x + 333.0$ Equation C

Data		Fitted values			Square of residuals		
UN.%	GDP %	Eq. A	Eq. B	Eq. C	Eq. A	Eq. B	Eq. C
5.5	100.0	119.436	108.736	108.6	377.766	76.321	73.960
5.5	103.9	119.436	108.736	108.6	241.374	23.389	22.090
6.7	107.5	90.142	79.442	59.64	301.290	787.236	2290.58
5.5	115.6	119.436	108.736	108.6	14.716	47.112	49.000
5.7	121.9	114.554	103.854	100.44	53.965	325.662	460.532
5.2	130.9	126.760	116.060	120.84	17.142	220.235	101.204
4.5	141.9	143.848	133.148	149.4	3.794	76.601	56.250
3.8	155.6	160.936	150.236	177.96	28.472	28.773	499.970
3.8	164.4	160.936	150.236	177.96	12.000	200.621	183.874
3.6	179.6	165.818	155.118	186.12	189.937	599.357	42.510
				RSS	1240.46	2385.31	3779.97

Summary

- The simple linear regression model is used for modeling the relationship between a response variable Y and an explanatory variable X.
- The model has the form $Y = aX + b + \varepsilon$ where ε is a random variable with mean 0 and constant variance σ^2.
- If n independent observations are made on this model, then $Y_i \sim N(a + bX_i, \sigma^2)$.
- The MLE of the parameters a and b correspond to the least square estimators and are given by:

$$\hat{b} = \frac{S_{xy}}{S_{xx}}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \ .$$

- The regression line $y = \hat{a}x + \hat{b}$ is the best line fitting the data points. The value $\hat{y}_i = \hat{a}x_i + \hat{b}$ is the i^{th} fitted value while $y_i - \hat{y}_i$ is the i^{th} residual.

Section 3.3: Point estimation of sigma

Why estimating σ^2?

Reasons to find an estimate of σ^2 include:

a) We want to obtain an indication of the variability of the probability distribution of Y.

b) As we will see later on, a variety of inferences concerning the regression function and the prediction of Y require an estimate of σ^2.

Residuals

- The i^{th} residual is the difference between the i^{th} observed and fitted values

$$r_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i$$

An estimate of σ^2 for the simple regression model

- The residual sum of squares:

$$RSS = \sum_{i=1}^{n} r_i^2$$

- The estimate of the variance is:

$$\hat{\sigma}^2 = \frac{RSS}{(n-2)}$$

Properties of $\hat{\sigma}^2$

- Recall that a point estimator $\hat{\theta}$ of a parameter θ is called unbiased if $E(\hat{\theta}) = \theta$. Otherwise, it is called biased.
- It can be shown that $\hat{\sigma}^2 = \dfrac{RSS}{(n-2)}$ is unbiased.
- One can prove that the maximum likelihood estimator of σ^2 is $\dfrac{RSS}{n}$ and this is a biased estimator.
- Hence $\hat{\sigma}^2$ is a bias-corrected version of the MLE.

Properties of RSS

- Recall that in the simple linear regression model, $Y_i = a + bX_i + \varepsilon_i$ where ε_i are normal independent random variables with mean 0 and variance σ^2.
- Hence $\frac{\varepsilon_i}{\sigma}$ has the standard normal distribution.
- It follows that $\sum_{i=1}^{n} \left(\frac{\varepsilon_i}{\sigma}\right)^2 \sim \chi^2_n$ since it is sum of n independent standard normal distributions.
- What about $\frac{RSS}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{r_i}{\sigma}\right)^2$?

Properties of RSS, cont.

- In Section 3.2 we showed that, equating to zero the partial derivatives of SS we found

$$\frac{\partial SS}{\partial a} = \sum_{i=1}^{n} 2(-1)(y_i - a - bx_i) = 0$$

$$\frac{\partial SS}{\partial b} = \sum_{i=1}^{n} 2(-x_i)(y_i - a - bx_i) = 0$$

- Since $r_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i$ this is same as

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} x_i r_i = 0$$

Properties of RSS, cont.

- Hence the r_i are not independent random variables as they satisfy two linear restrictions.
- It can be shown that

$$\frac{RSS}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{r_i}{\sigma}\right)^2 \sim \chi^2_{n-2}$$

Example: Unemployment rate and Gross Domestic Product (GDP)

- The following table shows the behaviour of the Gross Domestic Product (GDP) in USA, for the decade 1960-1969, as a function of the unemployment rate percentage. GDP is expressed in percent of the GDP value of the initial year 1960.

Year	Unempl. rate % (x)	GDP % (y)
1	5.5	100
2	5.5	103.9
3	6.7	107.5
4	5.5	115.6
5	5.7	121.9
6	5.2	130.9
7	4.5	141.9
8	3.8	155.6
9	3.8	164.4
10	3.6	179.6

- In lecture 11 we calculated:
  $\bar{x} = 4.98 \quad S_{xx} = 9.256, \quad S_{yy} = 6756.361, \quad S_{xy} = -225.954.$
  $\hat{a} = 253.7, \qquad \hat{b} = -24.4116\ .$

Example: Unemployment and GDP, cont.

- From previous slide the regression line equation is: $y = 253.7 - 24.4116x$.

- Hence we can calculate the residuals:

Year	Unempl. rate % (x)	GDP % (y)	Residuals (r_i)
1	5.5	100	-19.436
2	5.5	103.9	-15.536
3	6.7	107.5	17.358
4	5.5	115.6	-3.836
5	5.7	121.9	7.346
6	5.2	130.9	4.140
7	4.5	141.9	-1.948
8	3.8	155.6	-5.336
9	3.8	164.4	3.464
10	3.6	179.6	13.782

- Hence $RSS = \sum_i r_i^2 = 1240.457$,
$\hat{\sigma}^2 = RSS/(n-2) = 1240.457/8 = 155.057$

Computing $\hat{\sigma}^2$: alternative formulas

- Since $\hat{\sigma}^2 = \dfrac{RSS}{(n-2)}$, we need to compute RSS in order to compute $\hat{\sigma}^2$.
- Computing RSS directly is tedious.
- There is an alternative formula for RSS, as follows:

$$RSS = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

RSS and sample correlation

- From previous slide: $RSS = S_{yy} - \dfrac{S_{xy}^2}{S_{xx}}$
- From Section 3.1: $\hat{\rho} = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$, that is

$$\hat{\rho}^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

- Therefore

$$RSS = S_{yy} - \frac{\hat{\rho}^2 S_{xx} S_{yy}}{S_{xx}} = S_{yy}(1 - \hat{\rho}^2)$$

- Equivalently

$$\hat{\rho}^2 = \frac{S_{yy} - RSS}{S_{yy}}$$

An interpretation

- From previous slide: $\hat{\rho}^2 = \dfrac{(S_{yy} - RSS)}{S_{yy}}$ (*)

- $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total variation in y about its mean. RSS is the residual variation in y after fitting the regression line (that is, not explained by x).
  $S_{yy} - RSS$ is the amount of variation in y explained by x.

- Therefore the formula (*) says that $\hat{\rho}^2$ is the proportion of variation in the response y explained by the explanatory variable x.

An interpretation, cont.

- From previous slide: $\hat{\rho}^2 = \dfrac{(S_{yy} - RSS)}{S_{yy}}$
- Some special cases:

$\hat{\rho} = \pm 1$	$RSS = 0$	all variation in y explained as a function of x
$\hat{\rho} = 0$	$RSS = S_{yy}$	no variation in y explained as a function of x
$\hat{\rho} = 0.8$	64%	of variation in y explained as a function of x

where 'as a function of x' means the simple linear regression line.

Example: Unemployment and GDP, cont.

- In slide 11 we calculated:

$$S_{xx} = 9.256, \quad S_{yy} = 6756.361, \quad S_{xy} = -225.954.$$

- Therefore we compute

$$RSS = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 6756.361 - \frac{(-225.954)^2}{9.256} = 1240.457$$

$$\hat{\rho}^2 = \frac{S_{yy} - RSS}{S_{yy}} = \frac{6756.361 - 1240.457}{6756.361} = 0.8164$$

- Hence 81.64% of the variation in y is explained by x.

Summary

- We discussed the unbiased estimator $\hat{\sigma}^2 = \frac{RSS}{(n-2)}$ of the parameter σ, where RSS is the residual sum of squares.
- We discussed the distribution of $\frac{RSS}{\sigma^2}$.
- We showed that the sample correlation coefficient and RSS are related by $\hat{\rho}^2 = \frac{S_{yy} - RSS}{S_{yy}}$.
- We deduced that $\hat{\rho}^2$ is the proportion of variation in the response y explained by the explanatory variable x.
- We illustrated this in the GDP example.

Section 3.4: Sampling distributions

The idea of sampling distributions

- Think of repeating the experiment many times. Each time you have new observations, hence new estimates $\hat{a}, \hat{b}$.
- We want to analyze this variability in the estimates. That is we want to find $E(\hat{b})$, $var(\hat{b})$, $E(\hat{a})$, $var(\hat{a})$ and then the distributions of $\hat{b}, \hat{a}$, which are called sampling distributions.

The estimate $\hat{b}$

- From Section 3.2:

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- On the other hand, $\sum_{i=1}^{n}(x_i - \bar{x})\bar{y} = \bar{y}(\sum_{i=1}^{n} x_i - n\bar{x}) = 0$.

- Therefore

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Discussion

- Consider the expression of the previous slide

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Which part of this expression is random, which part is not?
- Is $\hat{b}$ a linear combination of random variables?

Answer to discussion questions

- From previous slide:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Only y_i's are random.

- Since x_i and $\bar{x}$ are constant, this has the form

$$\hat{b} = \sum_{i=1}^{n} c_i y_i \qquad c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

a linear combination of independent random variables $y_1, \ldots, y_n$.

Reminder

- The expectation of a linear combination of random variables is given by $E(\sum_{i=1}^{n} c_i y_i) = \sum_{i=1}^{n} c_i E(y_i)$.
- The variance of a linear combination of *independent* random variables is

$$\text{var}\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2 \,\text{var}(y_i)$$

Expectation of $\hat{b}$

- In our case $c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $\quad E(y_i) = a + bx_i$ therefore

$$
\begin{aligned}
E(\hat{b}) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(y_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(a + bx_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})bx_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{since } \sum_{i=1}^{n}(x_i - \bar{x})a = 0 \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} b \qquad \text{since } \sum_{i=1}^{n}(x_i - \bar{x})\bar{x} = 0 \\
&= b
\end{aligned}
$$

- Hence $\hat{b}$ is an unbiased estimator of b.

Variance of $\hat{b}$

- In our case $c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $\quad \text{var}(y_i) = \sigma^2$.
- Therefore

$$\begin{aligned}
\text{var}(\hat{b}) &= \text{var}\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \\
&= \frac{1}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} \sum_{i=1}^{n}(x_i - \bar{x})^2 \,\text{var}(y_i) \\
&= \frac{\sigma^2}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

Distribution of $\hat{b}$

- We calculated $E(\hat{b}) = b$ and $\text{var}(\hat{b}) = \dfrac{\sigma^2}{S_{xx}}$.
- Since y_i are normally distributed, we conclude that

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{S_{xx}}\right)$$

Expectation of $\hat{a}$

- Recall from Section 3.2:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

- Therefore

$$\begin{aligned} E(\hat{a}) &= E(\bar{y} - \hat{b}\bar{x}) = E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) - E(\hat{b})\bar{x} \\ &= \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) - b\bar{x} \\ &= a + b\bar{x} - b\bar{x} \\ &= a \end{aligned}$$

- Hence $\hat{a}$ is an unbiased estimator of a.

Distribution of $\hat{a}$

- One can calculate that

$$\text{var}(\hat{a}) = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

- If y_i are normally distributed, we conclude

$$\hat{a} \sim N\left(a, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right)$$

Example: Unemployment rate and Gross Domestic Product (GDP)

- The following table shows the behaviour of the Gross Domestic Product (GDP) in USA, for the decade 1960-1969, as a function of the unemployment rate percentage. GDP is expressed in percent of the GDP value of the initial year 1960.

Year	Unempl. rate % (x)	GDP % (y)
1	5.5	100
2	5.5	103.9
3	6.7	107.5
4	5.5	115.6
5	5.7	121.9
6	5.2	130.9
7	4.5	141.9
8	3.8	155.6
9	3.8	164.4
10	3.6	179.6

- In Section 3.2 we calculated:
  $\bar{x} = 4.98 \quad S_{xx} = 9.256, \quad S_{yy} = 6756.361, \quad S_{xy} = -225.954$
  $\hat{a} = 253.7, \quad \hat{b} = -24.4116, \quad RSS = 1240.457$

Example: Unemployment and GDP

- We want to calculate the standard errors for $\hat{a}$ and $\hat{b}$ using the formulas:

$$\text{var}(\hat{a}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \qquad \text{var}(\hat{b}) = \frac{\sigma^2}{S_{xx}}$$

σ^2 is unknown, so we use the estimate from Section 3.3

$$\hat{\sigma}^2 = \frac{RSS}{(n-2)}$$

- For unemployment and GDP example, we previously calculated

$$\hat{\sigma} = \sqrt{\frac{RSS}{(n-2)}} = \sqrt{\frac{1240.457}{8}} = 12.452$$

Example: Unemployment and GDP, cont.

- Therefore

$$\text{s.e.}(\hat{a}) = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} = 12.452\sqrt{\left(\frac{1}{10} + \frac{4.98^2}{9.256}\right)} = 20.759$$

$$\text{s.e.}(\hat{b}) = \hat{\sigma}\sqrt{\left(\frac{1}{S_{xx}}\right)} = 12.452\sqrt{\left(\frac{1}{9.256}\right)} = 4.093\ .$$

Distribution of residuals

- The i^{th} residual is $r_i = y_i - \hat{a} - \hat{b}x_i$, therefore

$$E(r_i) = E(y_i) - E(\hat{a}) - E(\hat{b})x_i = a + bx_i - a - bx_i = 0$$

- It can be shown that

$$\text{var}(r_i) = \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\sigma^2$$

Standardized residuals

- The standardized residual is defined as the residual divided by its standard error

$$\frac{r_i}{\hat{\sigma}\sqrt{\left(1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{S_{xx}}\right)}}$$

- Standardized residuals immediately tell us how many estimated standard deviations any point is away from the fitted regression model.

- For large samples these *standardized* residuals will approximately have a standardized Normal distribution and hence values larger than 2 or smaller than -2 are considered to be *unusual*.

Summary

- The estimators $\hat{a}$ and $\hat{b}$ are unbiased with variances

$$\text{var}(\hat{a}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \qquad \text{var}(\hat{b}) = \frac{\sigma^2}{S_{xx}}$$

  If we assume normal errors, $\hat{a}$ and $\hat{b}$ are normally distributed.

- The residual r_i is normally distributed with mean 0 and variance

$$\text{var}(r_i) = \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \sigma^2$$

- The standardized residuals are

$$\frac{r_i}{\hat{\sigma}\sqrt{\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right)}}$$

  and have a standard normal distribution.

Section 3.5: Confidence intervals and hypothesis tests

Confidence intervals: recap

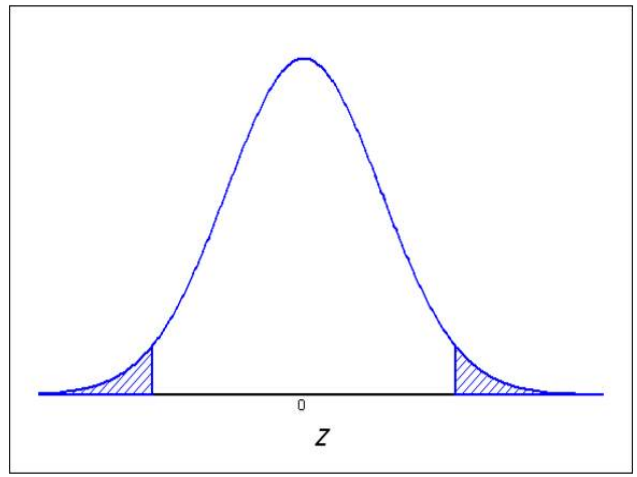
- Fact: Let $\hat{\theta}$ be a statistic that is normally distributed with mean θ and standard error $\sigma_{\hat{\theta}}$. Then

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim N(0, 1)$$

by central limit theorem.

Confidence intervals: recap, cont.

- Let $Z_{0.025}$ and $-Z_{0.025}$ be as follows



  where $P(-Z_{0.025} \leq Z \leq Z_{0.025}) = 0.95$.

- The 95% confidence interval for θ is

$$(\hat{\theta} - 1.96\sigma_{\hat{\theta}}\,,\, \hat{\theta} + 1.96\sigma_{\hat{\theta}})$$

Discussion

- We showed in Section 3.4 that, under the normal error assumption

$$\hat{a} \sim N\left(a, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right)$$

- What is the corresponding standard normal distribution?
- What is the corresponding 95% C.I?

Confidence intervals for a

- We showed in Section 3.4 that, under the normal error assumption

$$\hat{a} \sim N\left(a, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right) \ .$$

- Hence a 95% confidence interval for a is

$$\left(\hat{a} - 1.96\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \ , \ \hat{a} + 1.96\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

- However, σ is usually unknown.

Confidence intervals for a, using t-distribution

- We are going to replace σ with its estimator $\hat{\sigma}$.
- One obtains

$$\frac{\hat{a} - a}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}$$

where t_{n-2} is the t-distribution with $n-2$ degrees of freedom.

Confidence intervals for a, using t-distribution, cont.

- Hence a 95% confidence interval for a is

$$\left(\hat{a} - t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \; , \; \hat{a} + t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \right)$$

- That is, replacing σ^2 by the unbiased estimate $\hat{\sigma}^2$ merely changes the normal percentage point to the corresponding t percentage point.

Example: Pulse Data

- The following data gives results from 11 people in three columns: first pulse rate (at rest), second pulse rate (after exercise) and smoking indicator (1=smokes regularly, 2=does not smoke regularly). Let X = first pulse rate, Y = second pulse rate,

First pulse rate	Second pulse rate	Smoking indicator
96	140	2
62	100	2
78	104	1
82	100	2
100	115	1
68	112	2
96	116	2
78	118	2
88	110	1
62	98	1
80	128	2

We have $\bar{x} = 80.909$, $\bar{y} = 112.828$, $S_{xx} = 1770.909$, $S_{yy} = 1645.636$, $S_{xy} = 995.818$.

Example: Pulse data

- In this example, $n = 11$ and we calculate $\hat{a} = 67.323$, $\hat{\sigma} = 10.983$, $\bar{x} = 80.909$, $S_{xx} = 1770.909$.
- From the t-distribution table, $t_{0.025,9} = 2.262$.
- Hence a 95% C.I. for a is

$$\left(\hat{a} - t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \ , \ \hat{a} + t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \right)$$

$$= 67.32 \pm 2.262 \times 10.98\sqrt{\frac{1}{11} + \frac{80.91^2}{1770.91}} = (18.97, 115.67) \ .$$

Hypothesis test for a

- We want to test the hypothesis $H_0 : a = a_0$ against $H_1 : a \neq a_0$.
- If H_0 is true, then

$$t = \frac{\hat{a} - a_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$$

  is an observation from a t_{n-2} distribution.
- We choose significance level 5%. The corresponding critical region for this two-sided test is

$$(-\infty, -t_{0.025,n-2}) \cup (t_{0.025,n-2}, +\infty)$$

Example, cont.

- We want to test the hypothesis $a = 0$, that is, the regression line goes through the origin.
- As before, $\hat{a} = 67.323$. $\hat{\sigma} = 10.983$, $\bar{x} = 80.909$, $n = 11$, $S_{xx} = 1770.909$.
- The t-statistic is

$$t = \frac{\hat{a} - a_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} = \frac{67.323 - 0}{10.983\sqrt{\frac{1}{11} + \frac{80.909^2}{1770.909}}} = 3.15$$

- With a significance level of 5%, the critical region is $(-\infty, -2.262) \cup (2.262, +\infty)$ which contains 3.15. Hence we reject the null hypothesis.

Inference for b

- Inference for the slope of regression line is very often very useful in practice.
- When testing the hypothesis $H_0 : b = 0$ we want to assess if the explanatory variable X has really an influence on the response Y, or else is not needed.
- If $H_0 : b = 0$ is rejected, we say the explanatory variable is statistically significant to the response Y.

Confidence intervals for b

- We showed in Section 3.4 that, under the normal errors assumption,

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{S_{xx}}\right)$$

- Hence a 95% C.I. for b has limits

$$\left(\hat{b} - 1.96\frac{\sigma}{\sqrt{S_{xx}}}\ ,\ \hat{b} + 1.96\frac{\sigma}{\sqrt{S_{xx}}}\right)$$

- As in the case of $\hat{a}$, there is the problem that σ is in general unknown.

Confidence intervals for b, cont.

- We substitute σ with its estimate $\hat{\sigma}$.
- Similarly to the case for a, this changes the limits of the C.I. from the normal percentage point $Z_{0.025} = 1.96$ to the t-percentage point $t_{0.025,n-2}$.
- Hence a 95% C.I. for b is

$$\left(\hat{b} - t_{0.025,n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \ , \ \hat{b} + t_{0.025,n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \right)$$

Hypothesis test for b

- We want to test the hypothesis $H_0 : b = b_0$ against $H_1 : b \neq b_0$.
- If H_0 is true, then

$$t = \frac{\hat{b} - b_0}{\hat{\sigma}/\sqrt{S_{xx}}}$$

is an observation from a t_{n-2} distribution.

- We choose significance level 5%. The corresponding critical region for this two-sided test is

$$(-\infty, -t_{0.025,n-2}) \cup (t_{0.025,n-2}, +\infty)$$

Example: Pulse Data

- For Pulse Data we have:

  $n = 11$, $\hat{b} = 0.5623$, $\hat{\sigma} = 10.983$, $S_{xx} = 1770.909$.

- A 95% C.I. for b has limits

$$\left(\hat{b} - t_{0.025,n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \ , \ \hat{b} + t_{0.025,n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \right)$$

$$= 0.5623 \pm 2.262 \times 10.983 \sqrt{\frac{1}{1770.909}}$$

$$= (-0.03, 1.15)$$

- Note: this interval includes 0.

Example: Pulse Data, cont.

- We want to test the hypothesis $b = 0$.
- As before, $\hat{b} = 0.5623$, $\hat{\sigma} = 10.983$, $S_{xx} = 1770.909$.

  We calculate the t-statistic

$$t = \frac{\hat{b} - b_0}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{\hat{b}\sqrt{S_{xx}}}{\hat{\sigma}} = \frac{0.5623\sqrt{1770.909}}{10.983} = 2.15$$

- With a significance level of 5%, the critical region is $(-\infty, -2.262) \cup (2.262, +\infty)$ which doesn't contain 2.15. Hence we accept the null hypothesis $b = 0$.
- We conclude that the explanatory variable X has no influence on the response Y.

Confidence intervals for σ^2

- From Section 3.3,

$$\hat{\sigma}^2 = \frac{RSS}{(n-2)}, \qquad \frac{RSS}{\sigma^2} \sim \chi^2_{n-2}$$

- We have

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

- Hence,

$$P\left(\chi^2_{0.975,n-2} < \frac{(n-2)\hat{\sigma}^2}{\sigma^2} < \chi^2_{0.025,n-2}\right) = 0.95$$

Confidence intervals for σ^2, cont.

- The 95% C.I. for σ^2 is

$$\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{0.025,n-2}}, \frac{(n-2)\hat{\sigma}^2}{\chi^2_{0.975,n-2}} \right)$$

- The corresponding C.I. for σ is

$$\left(\hat{\sigma}\sqrt{\frac{n-2}{\chi^2_{0.025,n-2}}}, \hat{\sigma}\sqrt{\frac{n-2}{\chi^2_{0.975,n-2}}} \right)$$

Example: Pulse data

- In this case, $n = 11$, $RSS = (n-2)\hat{\sigma}^2 = 1085.67$, $\chi^2_{0.025,9} = 19.02$, $\chi^2_{0.975,9} = 2.7$.
- Hence a 95% C.I. for σ^2 is given by

$$\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{0.025,n-2}}, \frac{(n-2)\hat{\sigma}^2}{\chi^2_{0.975,n-2}} \right) = \left(\frac{1085.67}{19.02}, \frac{1085.67}{2.7} \right) = (57.04, 401.87)$$

- The corresponding C.I. for σ is

$$(\sqrt{57.04}, \sqrt{401.87}) = (7.55, 20.05)$$

Summary

- We established the limits of 95% C.I. for a, b and σ^2 to be

$$\left(\hat{a} - t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \ , \ \hat{a} + t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \right)$$

$$\left(\hat{b} - t_{0.025,n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}} \ , \ \hat{b} + t_{0.025,n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}} \right)$$

$$\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{0.025,n-2}}, \frac{(n-2)\hat{\sigma}^2}{\chi^2_{0.975,n-2}} \right)$$

- The corresponding hypothesis tests require the critical value from the t-distribution.
- Of particular interest is the hypothesis test for $H_0 : b = 0$ which assesses the need for the explanatory variable.

Section 3.6:
Confidence and prediction intervals

The need of confidence intervals for predictions

- Given a value x_0 of the predictor, the predicted response is $\hat{y}_0 = \hat{a} + \hat{b}x_0$.
- We want to assess the uncertainty in this prediction, as often needed in decision making.
- If the prediction has a wide C.I., we need to allow for outcomes far from the point estimate.

Different types of prediction

- We consider two different types of prediction:
  1. Predictions of future mean response.
  2. Prediction of future observation.

Discussion

Suppose we build a simple linear regression model where Y is the selling price of houses in a given area and X is the number of bedrooms. We can make two kinds of predictions for houses with 4 bedrooms:

1. We want to know what would a house with 4 bedrooms sell for on average.

2. Suppose that a specific house with 4 bedrooms comes on the market, we want to predict its price.

Question: for which type of prediction of the selling price do you think the error is greater?

Answer to discussion question

Suppose we build a simple linear regression model where Y is the selling price of homes in a given area and X is the number of bedrooms. We can make two kinds of predictions for a given x_0:

1. We want to know the average price of houses with characteristic x_0. This average selling price is $a + bx_0$ since $E(\varepsilon) = 0$, and this is predicted by $\hat{a} + \hat{b}x_0$, so only the variances of $\hat{a}$ and $\hat{b}$ need to be taken into account.

2. Suppose that a specific house comes on the market with characteristic x_0. Its selling price is $a + bx_0 + \varepsilon$.
The predicted price of this house is still $\hat{a} + \hat{b}x_0$, but in assessing the variance of this prediction, we must include the variance of ε.

Confidence intervals for expectations

- When $x = x_0$, then $E(y_0) = a + bx_0$, which has unbiased estimator

$$\hat{y}_0 = \hat{a} + \hat{b}x_0$$

- Recalling that $\hat{a} = \bar{y} - \hat{b}\bar{x}$, we obtain

$$\hat{y}_0 = \bar{y} - \hat{b}\bar{x} + \hat{b}x_0 = \bar{y} + \hat{b}(x_0 - \bar{x})$$

- We now use this expression to calculate the variance of $\hat{y}_0$.

Confidence intervals for expectation, cont.

- From previous slide:

$$\hat{y}_0 = \bar{y} + \hat{b}(x_0 - \bar{x})$$

- Therefore

$$\mathrm{var}(\hat{y}_0) = \mathrm{var}(\bar{y}) + 2(x_0 - \bar{x})\,\mathrm{cov}(\bar{y}, \hat{b}) + (x_0 - \bar{x})^2\,\mathrm{var}(\hat{b})$$

Confidence intervals for expectation, cont.

- Recall $\text{var}(\hat{b}) = \frac{\sigma^2}{S_{xx}}$ and $\text{var}(\bar{y}) = \frac{\sigma^2}{n}$, and

$$\begin{aligned}\text{cov}(\bar{y}, \hat{b}) &= \text{cov}\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} c_i y_i\right) = \frac{1}{n}\sum_{i=1}^{n} c_i \,\text{var}(y_i) \\ &= \frac{\sigma^2}{n}\sum_{i=1}^{n} c_i = 0\end{aligned}$$

- Therefore,

$$\text{var}(\hat{y}_0) = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

- Question: When does $\text{var}(\hat{y}_0)$ obtain the minimum?

Remark

- We calculated

$$\text{var}(\hat{y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

- We see that $\text{var}(\hat{y}_0)$ is minimal at $x_0 = \bar{x}$ and increases as x_0 moves away from $\bar{x}$.
- Thus the regression line is best estimated at the average value of the predictor and poorly estimated for extreme values of the predictor.

Confidence intervals for expectation, cont.

- We established that $\hat{y}_0 = \hat{a} + \hat{b}x_0$ is normally distributed with mean $a + bx_0$ and variance $\text{var}(\hat{y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$.

- It follows that

$$\frac{\hat{y}_0 - a - bx_0}{\sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}}$$

has a ..... distribution.

Fill in the gaps

Confidence intervals for expectation, cont.

- We substitute σ^2 with its estimate $\hat{\sigma}^2$. It can be shown that

$$\frac{\hat{y}_0 - a - bx_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}}$$

has a ..... distribution.

- A 95% C.I. for $a + bx_0$ has therefore limits

$$\left(\hat{a} + \hat{b}x_0 - t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}, \right.$$
$$\left. \hat{a} + \hat{b}x_0 + t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

Fill in the gaps

Example: Acid Rain

- Pure water has a pH of 7.0. Normal rain is slightly acidic because carbon dioxide (CO2) dissolves into it forming weak carbonic acid, with pH of approximately 5.5.
- A laboratory is equipped to make pH measurements.
- The following slide shows the results of measurements of pH made by a laboratory on 18 samples of water. Each sample is prepared by a chemical engineer and have a *known* level of pH.
- We want to compare the "True" values (x) with the laboratory measurements (y).

Example: Acid Rain, cont.

- The following table contains the data of the experiment

Sample	True pH	Measur. pH	Sample	True pH	Measur. pH
1	7.50	7.37	10	5.74	5.52
2	6.20	6.36	11	4.94	4.71
3	6.20	6.12	12	7.10	7.12
4	7.10	6.96	13	6.55	6.43
5	6.20	6.25	14	7.20	6.76
6	5.95	5.93	15	5.34	5.11
7	6.55	6.38	16	9.50	8.50
8	6.55	6.34	17	8.00	8.50
9	5.34	4.93	18	10.0	9.50

- For these data we calculate:
  $\hat{a} = 0.3595, \quad \hat{b} = 0.921, \quad \hat{\sigma} = 0.29848, \quad \bar{x} = 6.776,$
  $\bar{y} = 6.599, \quad S_{xx} = 30.626, \quad S_{yy} = 27.400, \quad S_{xy} = 28.205,$
  $RSS = 1.4254, \quad n = 18.$

Example: Acid Rain, cont.

- From previous slide
  $\hat{a} = 0.3595, \quad \hat{b} = 0.921, \quad \hat{\sigma} = 0.29848, \quad \bar{x} = 6.776,$
  $S_{xx} = 30.626, \; n = 18.$

- Suppose we want the 95% C.I. at $x_0 = 5$. Since $t_{0.025,16} = 2.12$, we have

$$\hat{a} + \hat{b}x_0 \pm t_{0.025,16}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$= 0.3595 + 0.921 \times 5 \pm 2.12 \times 0.29848\sqrt{\frac{1}{18} + \frac{(5 - 6.776)^2}{30.626}}$$

$$= (4.713, 5.216)$$

Prediction intervals for new observation

- If y_0 is the actual value, then

$$y_0 = a + bx_0 + \varepsilon_0 \ ,$$

where ε_0 is some unknown error and $\varepsilon_0 \sim N(0, \sigma^2)$.

- A point estimate of y_0 is $\hat{y}_0 = \hat{a} + \hat{b}x_0$ and

$$\hat{y}_0 - y_0 = \hat{y}_0 - a - bx_0 - \varepsilon_0$$

- It follows that $E(\hat{y}_0 - y_0) = 0$.

Prediction intervals for new observation, cont.

- It can be shown that

$$\text{var}(\hat{y}_0 - y_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

- In conclusion, $\hat{y}_0 - y_0$ is normally distributed with mean 0 and variance $\sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$.

- It follows that

$$\frac{\hat{y}_0 - y_0}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim N(0, 1)$$

- However, σ is in general unknown.

Prediction intervals for new observation, cont.

- We substitute the unknown σ with its estimate $\hat{\sigma}$ and find

$$\frac{\hat{y}_0 - y_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

- The 95% prediction interval for y_0 has limits

$$\left(\hat{a} + \hat{b}x_0 - t_{0.025,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}, \right.$$
$$\left. \hat{a} + \hat{b}x_0 + t_{0.025,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

Example: Acid Rain

- We want to calculate the 95% prediction interval for Y at the value $x = 5$ for the Acid Rain.

- We previously calculated $\hat{a} = 0.3595, \quad \hat{b} = 0.921,$ $\hat{\sigma} = 0.29848, \quad \bar{x} = 6.776, \quad S_{xx} = 30.626,\ n = 18,$ $t_{0.025,16} = 2.12.$

- Therefore

$$\hat{a} + \hat{b}x_0 \pm t_{0.025,16}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$= 0.3595 + 0.921 \times 5 \pm 2.12 \times 0.29848\sqrt{1 + \frac{1}{18} + \frac{(5 - 6.776)^2}{30.626}}$$

$$= (4.283, 5.646)$$

Confidence interval versus prediction interval

- We showed that the C.I. for the mean response when $x = x_0$ is

$$\hat{a} + \hat{b}x_0 \pm t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

  while the prediction interval for a new observation at $x = x_0$ is

$$\hat{a} + \hat{b}x_0 \pm t_{0.025,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

- Which interval is wider?

Summary

- We gave the notions of CI for mean response and prediction interval for new observations, and illustrated their differences.
- We gave the corresponding formulas

$$\hat{a} + \hat{b}x_0 \pm t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

$$\hat{a} + \hat{b}x_0 \pm t_{0.025,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

- We illustrated this on the Acid Rain example.