Introduction to Probability and Statistics Fourteenth Edition

Chapter 12
Linear Regression and Correlation

CH11 VS CH12?

- In Chapter 11, we used ANOVA to investigate the effect of various factor-level combinations (treatments) on a response.
  - Our objective was to see whether the treatment means were different.
- In Chapters 12 and 13, we investigate a response y which is affected by various independent variables, x_i.

Introduction

- Our objective is to use the information provided by the x_i to predict the value of y.
- We plot the value of x in the x-axis, and the value of y in the y-axis.
  - x, factor, explanatory variable, independent variable, input
  - y, response variable, dependent variable, output

Sum of squares: Definitions and simplications

$$S_{xx} := \sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \sum x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2$$

$$S_{yy} := \sum (y_i - \bar{y})^2 = \sum (y_i^2 - 2\bar{y}y_i + \bar{y}^2) = \sum y_i^2 - 2\bar{y}(n\bar{y}) + n\bar{y}^2 = \sum y_i^2 - n\bar{y}^2$$

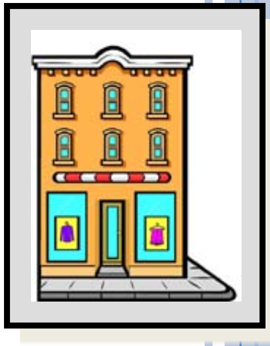
$$S_{xy} := \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i y_i - \bar{x}y_i - \bar{y}x_i + \bar{x}\bar{y}) = \sum x_i y_i - \bar{x}(n\bar{y}) - \bar{y}n\bar{x} + n\bar{x}\bar{y} = \sum x_i y_i - n\bar{x}\bar{y}\,.$$

12.1 Simple linear regression

EXAMPLE

- Let y be a student's college achievement, measured by his/her GPA. This might be a function of several variables:
  - x_1 = rank in high school class
  - x_2 = high school's overall rating
  - x_3 = high school GPA
  - x_4 = SAT scores
- We want to predict y using knowledge of x_1, x_2, x_3 and x_4.

Example



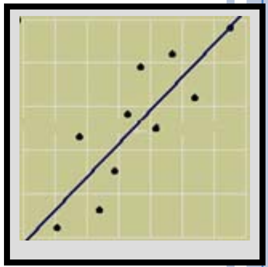
- Let y be the monthly sales revenue for a company. This might be a function of several variables:
  - x_1 = advertising expenditure
  - x_2 = time of year
  - x_3 = state of economy
  - x_4 = size of inventory
- We want to predict y using knowledge of x_1, x_2, x_3 and x_4.

Some Questions

- Which of the independent variables are useful and which are not?
- How could we create a prediction equation to allow us to predict y using knowledge of x_1, x_2, x_3 etc?
- How good is this prediction?

We start with the simplest case, in which the response *y* is a function of a single independent variable, *x*.

A Simple Linear Model



- In Chapter 3, we used the equation of a line to describe the relationship between y and x for a **sample** of n pairs, (x, y).
- If we want to describe the relationship between y and x for the **whole population**, there are two models we can choose

•**Deterministic (Economic) Model:** $y = \alpha + \beta x$

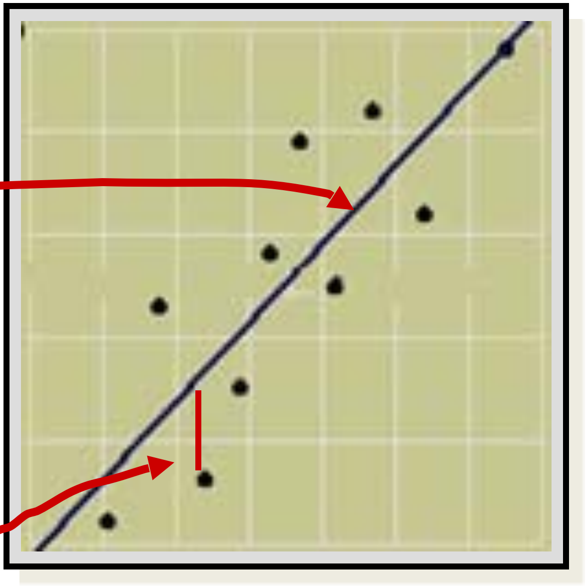
•**Probabilistic (Econometric) Model:**

  •**y = deterministic model + random error**

  •$\mathbf{y} = \alpha + \beta \mathbf{x} + \varepsilon$

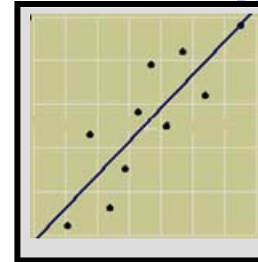
A Simple Linear Model

- Since the bivariate measurements that we observe do not generally fall **exactly** on a straight line, we choose to use:
- **Probabilistic Model:**
  - $y = \alpha + \beta x + \varepsilon$
  - $E(y|x) = \alpha + \beta x$

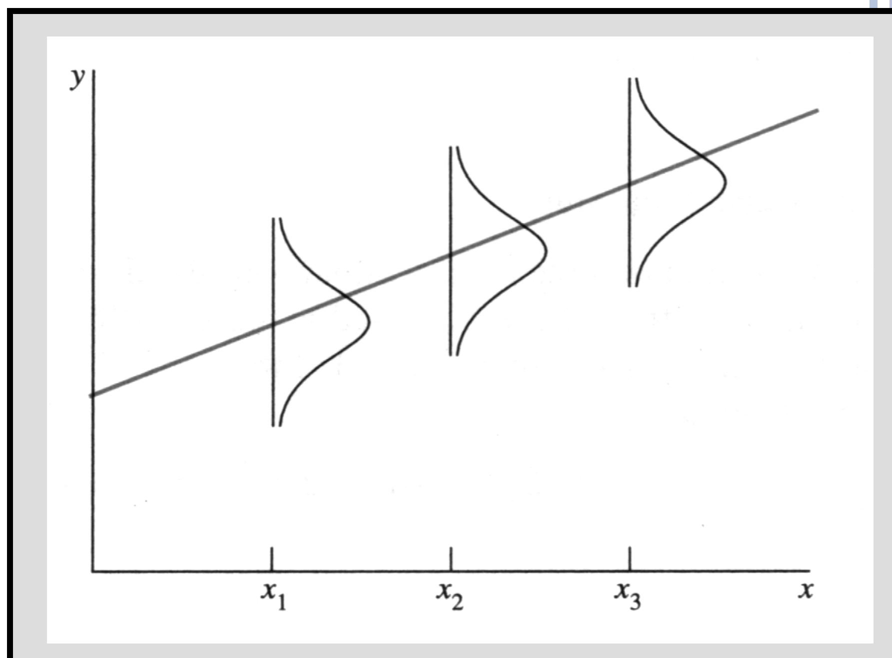


Points deviate from the **line of means** by an amount ε where ε has a normal distribution with mean 0 and variance σ^2.

The Random Error



- The line of means, $E(y|x) = \alpha + \beta x$, describes average value of y for any fixed value of x.
- The population of measurements is generated as y deviates from the population line by ε. We estimate α and β using sample information.



$E(y) := E(y|x) = \alpha + \beta x$

The line of means

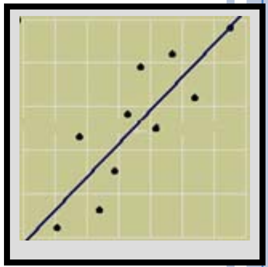
$E(y) := E(y|x) = \alpha + \beta x$,

is also called:

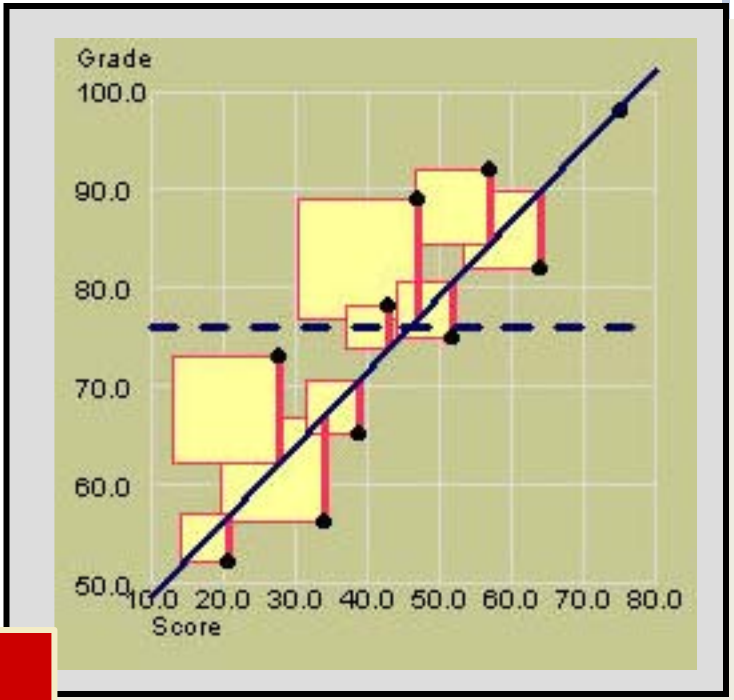
- the conditional mean given x:
- the conditional expectation given x,
- the simple regression function
- the regression function
- Interpretations of α and β
  - α: when $x = 0$, on avereage, y is α . 當x=0, y平均而言是 α
  - β: when x increases one unit, on average, y increases β unit. 當 x增加一個單位，平均而言，y增加β個單位。

MY APPLET

The Method of Least Squares



- The equation of the best-fitting line is calculated using a set of n pairs (x_i, y_i).
  - We choose our estimates *a* and *b* to estimate α and β so that the vertical distances of the points from the line, are minimized.



$$\text{Bestfitting line}: \hat{y} = a + bx$$

Choose a and b to minimize

$$\text{SSE} = \sum (y - \hat{y})^2 = \sum (y - a - bx)^2$$

APPENDIX: PROOF OF LSE

- Least squares estimators a and b minimize $\text{SSE}(a,b)$, where
$$\text{SSE}(a,b) = \sum (y-\hat{y})^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 .$$
- Results: The least squares estimators are:
$$b = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2} = \frac{S_{xy}}{S_{xx}},$$
$$a = \bar{y} - b\bar{x} .$$
- The *estimated* or *fitted* regression line is: $\hat{y}_i = a + bx_i$.

Example

The table shows the math achievement test scores for a random sample of n = 10 college freshmen, along with their final calculus grades.

Student	1	2	3	4	5	6	7	8	9	10
Math test, x	39	43	21	64	57	47	28	75	34	52
Calculus grade, y	65	78	52	82	92	89	73	98	56	75

Use your calculator to find the sums and sums of squares.

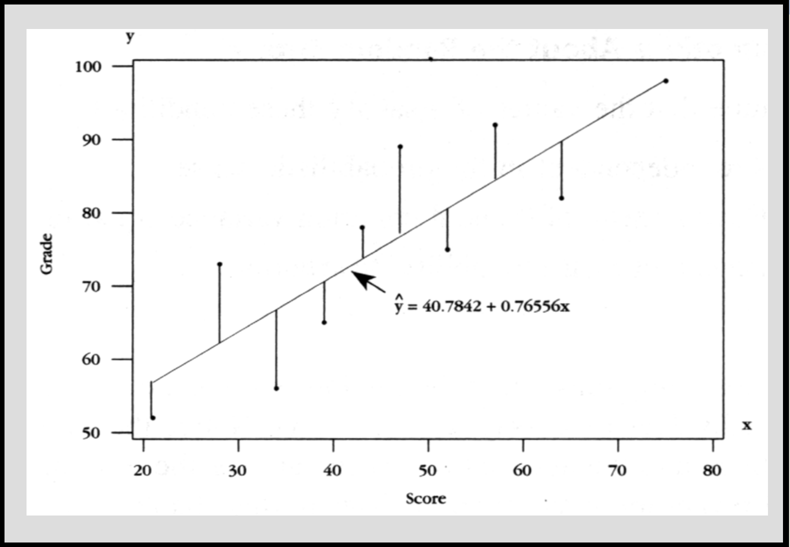
$$\sum x = 460 \qquad \sum y = 760$$

$$\sum x^2 = 23634 \quad \sum y^2 = 59816$$

$$\sum xy = 36854$$

$$\bar{x} = 46 \qquad \bar{y} = 76$$

Example



$$S_{xx} = 23634 - \frac{(460)^2}{10} = 2474$$

$$S_{yy} = 59816 - \frac{(760)^2}{10} = 2056$$

$$S_{xy} = 36854 - \frac{(460)(760)}{10} = 1894$$

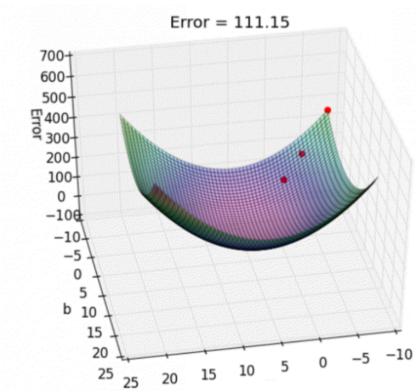
$$b = \frac{1894}{2474} = .76556 \quad \text{and} \quad a = 76 - .76556(46) = 40.78$$

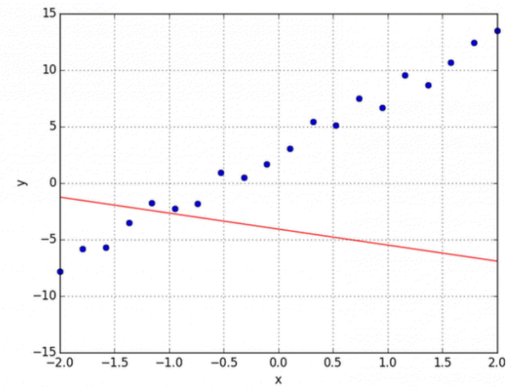
$$\text{Bestfitting line:} \quad \hat{y} = 40.78 + .77x$$

Demo.

$$\mathrm{SSE}(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 .$$

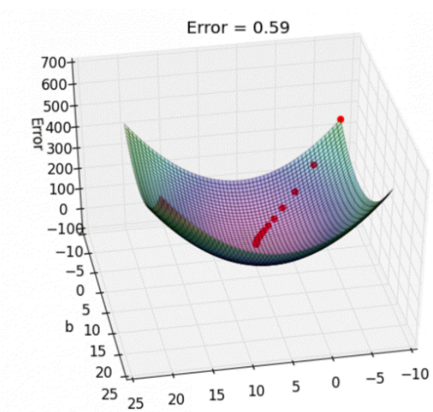
a = -1.41, b = -4.10

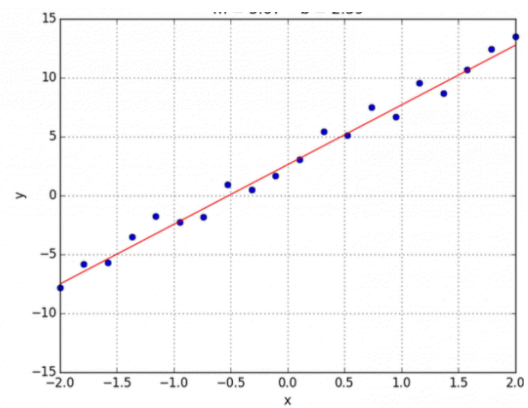




a

a = 5.07, b = 2.59





a

Source: https://medium.com/@savannahar68/getting-started-with-regression-a39aca03b75f

APPENDIX A: LSE formulas derivations

- Recall that $\mathrm{SSE}(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2$. First-order-Condition:

$$\frac{\partial SSE(a,b)}{\partial a} = -2\sum (y_i - a - bx_i) = 0,$$
$$\frac{\partial SSE(a,b)}{\partial b} = -2\sum x_i(y_i - a - bx_i) = 0.$$

- Simple math gives

$$\sum y_i = na + (\sum x_i)b, \qquad (1)$$
$$\sum x_i y_i = (\sum x_i)a + (\sum x_i^2)b. \qquad (2)$$

- Multipling (1) by $(\sum x_i)$ and multiply (2) by n, we have

$$(\sum x_i)(\sum y_i) = n(\sum x_i)a + (\sum x_i)^2 b, \qquad (3)$$
$$n(\sum x_i y_i) = n(\sum x_i)a + n(\sum x_i^2)b. \qquad (4)$$

- Thus, (4)-(3) gives

$$b = \frac{n(\sum x_i y_i) - (\sum x_i \sum y_i)}{n(\sum x_i)^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}.$$

- Plug b back to (1), we have $a = \bar{y} - \bar{x}b$.

12.2 An Analysis of variance for lienar regression

The ANOVA Table

Total df = n -1

Regression df = 1

Error df = n –1 – 1 = n - 2

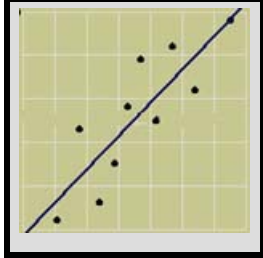
Mean Squares

MSR = SSR/(1)

MSE = SSE/(n-2)

Source	df	SS	MS	F
Regression	1	SSR	MSR=SSR/(1)	F=MSR/MSE
Error	n - 2	SSE	MSE=SSE/(n-2)	
Total	n -1	Total SS		

The Analysis of Variance



- The total variation in the experiment is measured by the **total sum of squares**:

$$\text{Total SS} = S_{yy} = \sum (y - \bar{y})^2$$

The **Total SS** is divided into two parts:

✓**SSR** (sum of squares for regression): measures the variation explained by using *x* in the model.

✓**SSE** (sum of squares for error): measures the leftover variation not explained by *x*.

TOTAL SS

- Total SS $= \sum (y_i - \bar{y})^2 = S_{yy}$.
- Note that $\sum_{i=1}^{n} y_i = n\bar{y}$.
- Thus,

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

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

How to calculate ANOVA table?

1. Total SS $= \sum (y - \bar{y})^2 = S_{yy}.$

2. SSR

$$\text{SSR}:= \sum (\hat{y}_i - \bar{y})^2 = \sum ((\boxed{a} + bx_i) - \bar{y})^2 = \sum ((\boxed{\bar{y} - b\bar{x}}) + bx_i - \bar{y})^2$$

$$= \sum (b(x_i - \bar{x}))^2 = b^2 S_{xx} = \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}.$$

3. SSE = Total SS − SSR

The Analysis of Variance

We calculate

$$SSR = \frac{(S_{xy})^2}{S_{xx}} = \frac{1894^2}{2474}$$

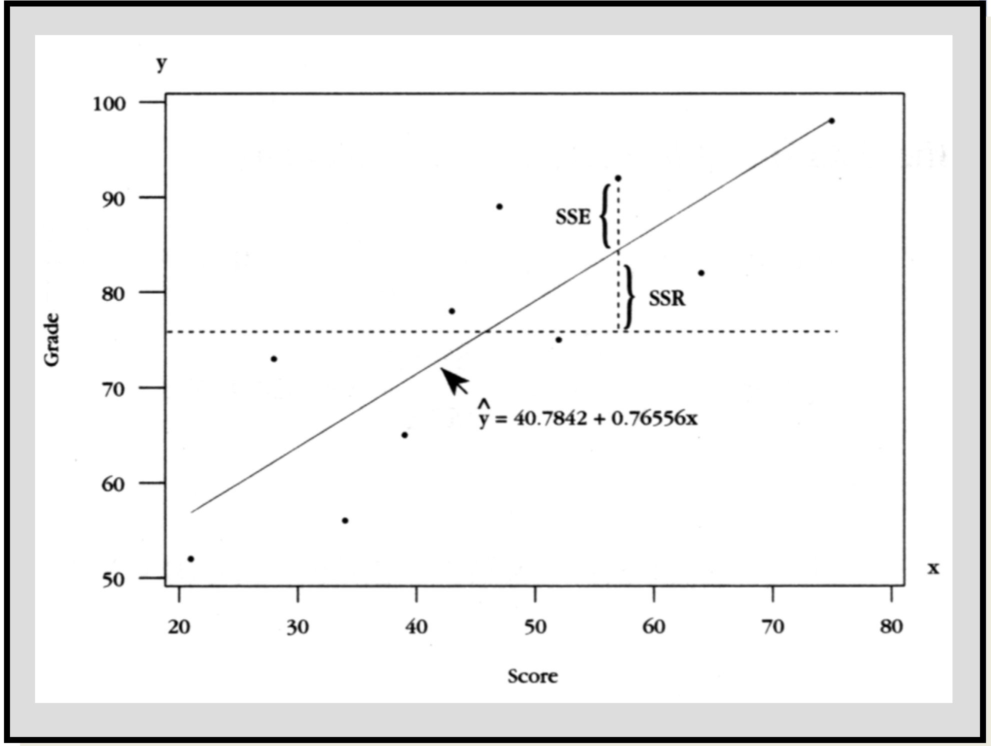
$$= 1449.9741$$

$$SSE = \text{Total SS} - SSR$$

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

$$= 2056 - 1449.9741$$

$$= 606.0259$$



The Calculus Problem

$$SSR = \frac{(S_{xy})^2}{S_{xx}} = \frac{1894^2}{2474} = 1449.9741$$

$$SSE = \text{Total SS} - SSR = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

$$= 2056 - 1449.9741 = 606.0259$$

Source	df	SS	MS	F
Regression	1	1449.9741	1449.9741	19.14
Error	8	606.0259	75.7532	
Total	9	2056.0000		

Calculation of ch 12.1-12.2

- A: Calculate$(\sum x_i)$, $(\sum y_i)$, $(\sum x_i^2)$, $(\sum y_i^2)$,$(\sum x_i y_i)$, $\bar{x}$, and $\bar{y}$
- B: Calculate
  - $S_{xx} := \sum (x-\bar{x})^2 = \sum x^2 - (\sum x)^2/n$.
  - $S_{yy} := \sum (y-\bar{y})^2 = \sum y^2 - (\sum y)^2/n.$
  - $S_{xy} := \sum (x-\bar{x})(y-\bar{y}) = \sum xy - (\sum x)(\sum y)/n.$
- C: Calculate
  - $b = S_{xy}/S_{xx}$
  - $a = \bar{y} - b\bar{x}$
  - Best fitting line: $y = a + bx$
- D: Complete the ANOVA table:
  - (1) totalSS $= S_{yy}$
  - (2) SSR $= S_{xy}^2/S_{xx}$,
  - (3) SSE $=$ totalSS $-$ SSR.

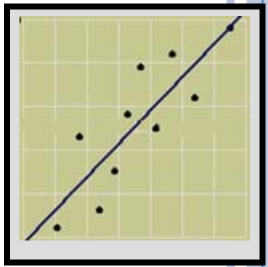
Next, Hypothesis Testing!

12.3 Testing the usefulness of the linear regression model

THREE EQUIVELENT APPROAHCES

1. H_0: model is not useful in explaining the response variable vs H_1: model is useful in explaining the response variable
   - Using ANOVA, F test, right-tailed test
2. $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$
   - $t_{STAT} = \dfrac{b}{\sqrt{MSE/S_{xx}}} \sim t_{df=(n-2)}$, t test, two-sided test
3. $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$
   - $t_{STAT} = r\sqrt{\dfrac{n-2}{(1-r^2)}} \sim t_{(n-2)}$, t test, two-sided test

Testing the Usefulness of the Model



- The first question to ask is whether the independent variable *x* is of any use in predicting *y*.
- If it is not, then the value of *y* does not change, regardless of the value of *x*. This implies that the slope of the line, β, is zero.

$$H_0 : \beta = 0 \quad \text{versus} \; H_a : \beta \neq 0$$

Key ingredients

$$a \sim N\left(\alpha, \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}\right),$$

- $$b \sim N\left(\beta, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right).$$

- See appendix for derivations in p.44 - p.49.
- With the above information, we can do statistical inference:
  - Confidence interval
  - Hypothesis test

t-TEST

1. H_0: $\beta = 0$ vs $H_1 : \beta \neq 0$
2. Set α
3. $t_{STAT} = \dfrac{b - 0}{\sqrt{MSE/S_{xx}}} \sim t_{(n-2)}$.
4. Calulate t^*
5. Find the rejection region or p-value. (two-tail test).
6. Conclude.

Summary

- The $100(1-\alpha)\%$ CI for β is

$$b \pm t_{(n-2);\alpha/2}\sqrt{\frac{\text{MSE}}{S_{xx}}}.$$

- To ask whether x is useful in predicting y, we set $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$. The test statistic is

$$t_{STAT} = \frac{b-0}{\sqrt{\frac{\text{MSE}}{S_{xx}}}} \sim t_{(n-2)}.$$

The Calculus Problem

- Is there a significant relationship between the calculus grades and the test scores at the 5% level of significance?

MY APPLET

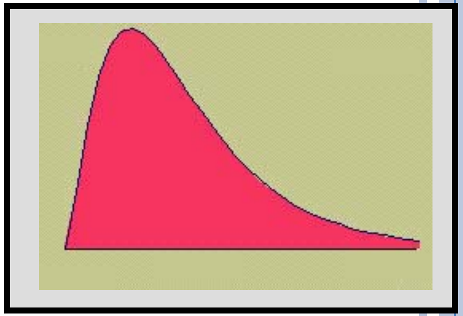
$$H_0 : \beta = 0 \ \text{versus} \ H_a : \beta \neq 0$$

$$t = \frac{b - 0}{\sqrt{\text{MSE}/S_{xx}}} = \frac{.7656 - 0}{\sqrt{75.7532/2474}} = 4.38$$

Reject H_0 when $|t| > 2.306$. Since $t = 4.38$ falls into the rejection region, H_0 is rejected .

There is a significant linear relationship between the calculus grades and the test scores for the population of college freshmen.

THE F TEST



- You can test the overall usefulness of the model using an F test. If the model is useful, MSR will be large compared to the unexplained variation, MSE.

To test H_0 : model is useful in predicting y

$$\text{Test Statistic: } F = \frac{MSR}{MSE}$$

Reject H_0 if $F > F_\alpha$ with 1 and $n - 2$ df.

THE F TEST

You can test the overall usefulness of the model using an F test. If the model is useful, MSR will be large compared to the unexplained variation MSE.

1. H_0: The model is not useful in predicting y vs H_1: The model is useful in predicting y
2. Set up α
3. Test statistic: $F_{STAT} = \frac{MSR}{MSE} \sim F_{1,(n-2)}$.
4. Calculate realized statistic F^*.
5. Find rejection region of p-value.

This is a right-tailed test!

6. Conclude

This test is exactly equivalent to the t-test, with $t^2 = F$.

Minitab Output

Least squares regression line

To test $H_0 : \beta = 0$

Regression Analysis: y versus x

The regression equation is y = 40.8 + 0.766 x

Predictor	Coef	SE Coef	T	P
Constant	40.784	8.507	4.79	0.001
x	0.7656	0.1750	4.38	0.002

S = 8.70363   R-Sq = 70.5%   R-Sq(adj) = 66.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1450.0	1450.0	19.14	0.002
Residual Error	8	606.0	75.8		
Total	9	2056.0			

$\sqrt{MSE}$

Regression coefficients, *a* and *b*

$t^2 = F$

WHY $t_{STAT}^2 = F_{STAT}$?

- The t test and F test are indeed the same! This is because:

$$t_{STAT}^2 = \left(\frac{b}{\sqrt{MSE/S_{xx}}}\right)^2 = \frac{b^2}{MSE/S_{xx}} = \frac{b^2 S_{xx}}{MSE}$$

$$= \frac{S_{xy}^2}{S_{xx}^2}\frac{S_{xx}}{MSE}$$

$$= \frac{S_{xy}^2}{S_{xx}MSE} = \frac{SSR}{MSE} = \frac{SSR/1}{MSE} = \frac{MSR}{MSE} = F_{STAT}$$

MEASURING THE STRENGTH OF THE RELATIONSHIP

- If the independent variable x is useful in predicting y, you will want to know how well the model fits.
- The strength of the relationship between x and y can be measured using:
- Correlation coefficient: $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$.
- Coefficient of determination: $R^2 = \dfrac{\text{SSR}}{\text{Total SS}}$

(pronounced "R squared").

R^2

- Because Total = SS+SSE, R^2 measures the proportion of the total variation in the response that can be explained by using the explanatory variable x in the model.
- The percent reduction the total variation by using the regression equation rather than just using the sample mean $\bar{y}$ to estimate y.
- For the calculation problem, $R^2 = 0.705$ or 70.5%. The model is working well!

WHY $R^2 = r^2$?

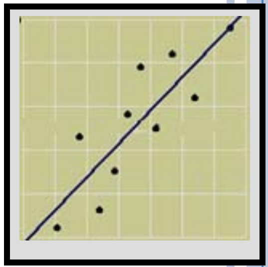
- This is because

$$r^2 = \left(\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\right)^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{S_{xy}^2}{S_{xx}}\frac{1}{S_{yy}} = \frac{\text{SSR}}{\text{Total SS}} = R^2$$

-

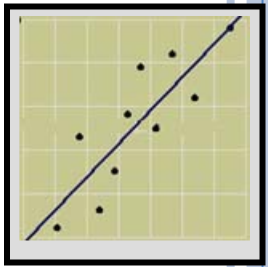
Recall that SSR $= \dfrac{S_{xy}^2}{S_{xx}}$**!**

Interpreting a Significant Regression

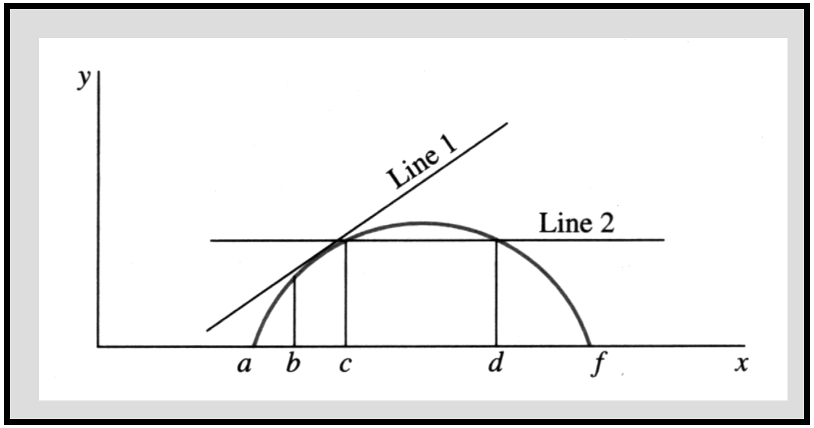


- Even if you do not reject the null hypothesis that the slope of the line equals 0, it does not necessarily mean that *y* and *x* are unrelated.
- Type II error—falsely declaring that the slope is 0 and that *x* and *y* are unrelated.
- It may happen that *y* and *x* are perfectly related in a nonlinear way.

Some Cautions



- You may have fit the wrong model.



- **Extrapolation**—predicting values of *y* outside the range of the fitted data.
- **Causality**—Do not conclude that *x* causes *y*. There may be an unknown variable at work!

Appendix B: A useful formula

- Linear combinations of indepedenent normal distributions remain a normal distribution. Specifically, if $X_i \sim N(\mu_i, \sigma^2)$and X_i are independent, then
- $\sum a_i X_i \sim N\left(\sum a_i \mu_i, \sigma^2 (\sum a_i^2) \right).$

Appendix: why $b \sim N(\beta, \frac{\sigma^2}{\sum (x_i - \bar{x})^2})$?

- Because $\sum (x_i - \bar{x}) = 0$, we rewrite b as

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

$$= \sum \frac{(x_i - \bar{x})y_i}{\sum (x_i - \bar{x})^2}$$

- $$= \sum \left(\frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2} \right) y_i$$

$$= \sum w_i y_i,$$

- where $w_i = \dfrac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$.

- b is a linear combination of normal distribution, and hence a normal distribution.

- To find the expectation and variance of b, note the following identities:

- $\sum w_i = 0;$
- $\sum w_i x_i = \sum \frac{(x_i - \bar{x})x_i}{\sum (x_i - \bar{x})^2} = \sum \frac{(x_i - \bar{x})x_i - (x_i - \bar{x})\bar{x}}{\sum (x_i - \bar{x})^2} = \sum \frac{(x_i - \bar{x})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = 1;$
- $\sum w_i^2 = \sum \left(\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right)^2 = \frac{1}{\sum (x_i - \bar{x})^2}.$

- Rewrite

$$b = \sum w_i(\alpha + \beta x_i + e_i)$$

$$= \alpha \sum w_i + \beta \sum w_i x_i + \sum w_i e_i$$

- $$= \beta + \sum w_i e_i \,.$$

- Therefore,

$$E(b) = \beta$$

- $$var(b) = \sum w_i^2 \sigma^2 = \sigma^2 \frac{1}{\sum (x_i - \bar{x})^2} \,.$$

-

APPENDIX: WHY $a \sim N(\alpha, \sigma^2 \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2})$?

- Now, we write

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

$$= \sum (\frac{1}{n} - \bar{x} w_i) y_i$$

$$= \sum (\frac{1}{n} - \bar{x} w_i)(\alpha + \beta x_i + e_i)$$

- $$= (\alpha - \alpha \bar{x} \sum w_i) + \left(\beta \bar{x} - \bar{x}\beta \sum x_i w_i\right) + \sum (\frac{1}{n} - \bar{x} w_i) e_i$$

$$= \alpha + \sum (\frac{1}{n} - \bar{x} w_i) e_i \, .$$

Hence, a is a normal distribution.

- To find the expectation and variance of a, it is easy to see

$$E(a) = \alpha + \sum [(\frac{1}{n} - \bar{x} w_i)0] = \alpha$$

$$var(a) = \sum (\frac{1}{n} - \bar{x} w_i)^2 \sigma^2$$

$$= \sigma^2 \left(\sum \frac{1}{n^2} - 2 \sum \frac{\bar{x}}{n} w_i + \bar{x}^2 \sum w_i^2 \right)$$

$$= \sigma^2 \left(\frac{1}{n} + \bar{x}^2 \frac{1}{\sum (x_i - \bar{x})^2} \right)$$

- $$= \sigma^2 \frac{\sum (x_i - \bar{x})^2 + n\bar{x}^2}{n \sum (x_i - \bar{x})^2}$$

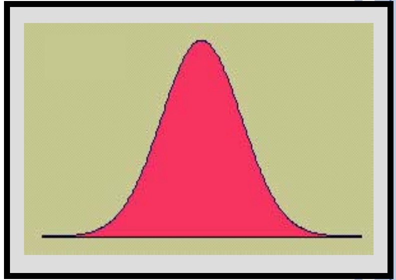
$$= \sigma^2 \frac{(\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2) + n\bar{x}^2}{n \sum (x_i - \bar{x})^2} = \sigma^2 \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}.$$

-

12.4

Diagnostic Tools for checking the regression assumptions

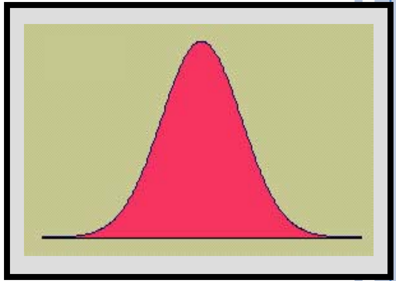
Checking the Regression Assumptions



1. The relationship between x and y is linear, given by $y = \alpha + \beta x + \varepsilon$. ε：error 誤差
2. The random error terms ε are independent and, for any value of x, have a normal distribution with mean 0 and variance σ^2.

- Remember that the results of a regression analysis are only valid when the necessary assumptions have been satisfied.

Diagnostic Tools



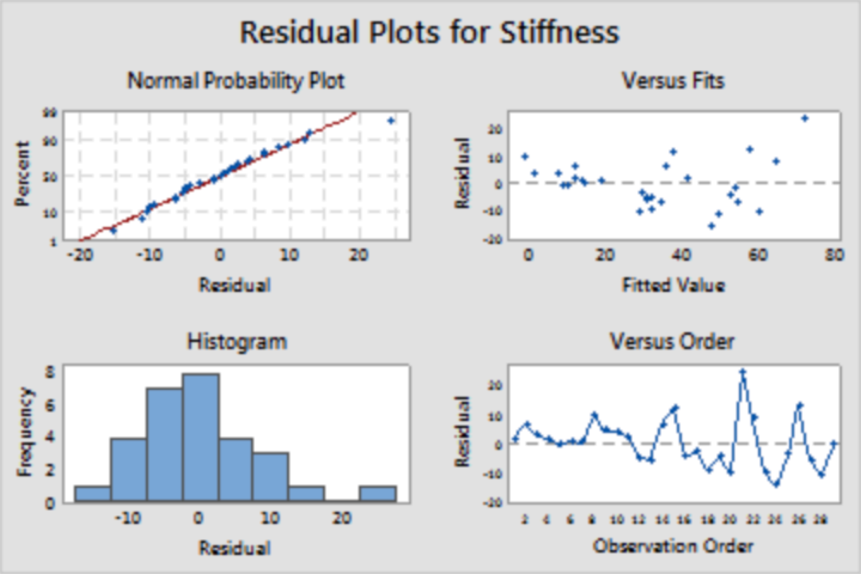
- We use the same diagnostic tools used in Chapter 11 to check the normality assumption and the assumption of equal variances.

1. **Normal probability plot** of residuals
2. Plot of **residuals versus fit** or **residuals versus variables**

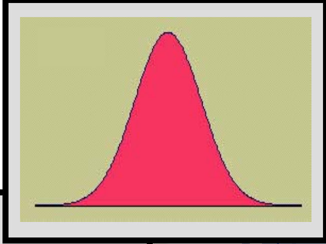
RESIDUALS

- Residual = 殘差, estimated error:
  $r_i := y_i - \hat{y}_i = y_i - a - bx_i\,.$
- If all assumptions have been met, these residuals should be $N(0,\sigma^2)$.
- Informal way
  1. Residuals plots (identical and independent)
     - r_i vs y_i
     - r_i vs i
     - r_i vs x_i
  - Normal distribution
     - Normal probability plot
     - histogram

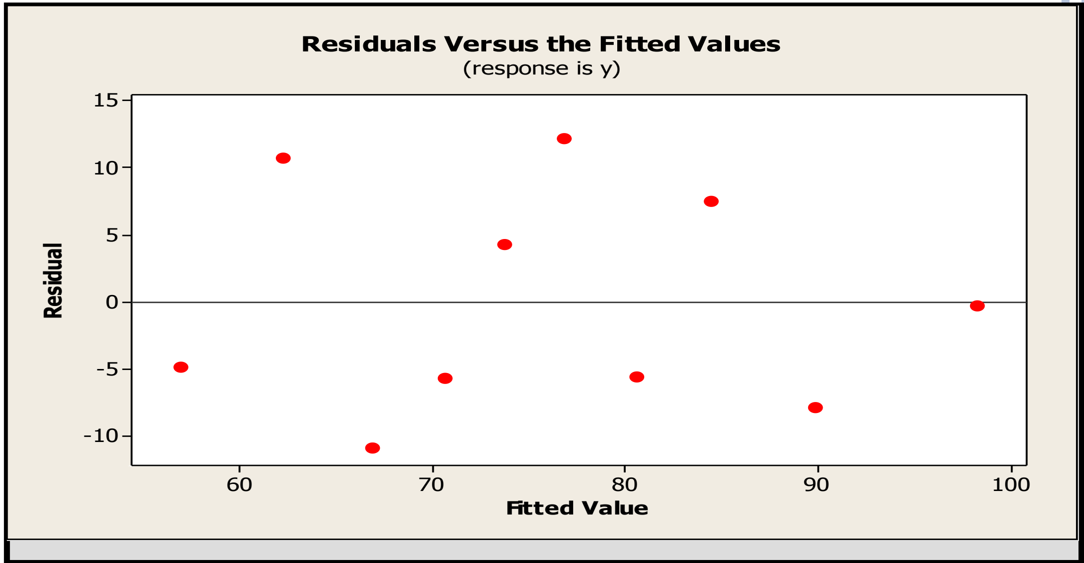
The **residual error** is the “leftover” variation in each data point after the variation explained by the regression model has been removed.



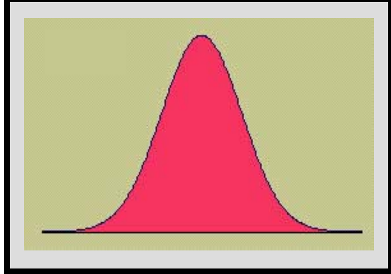
Residuals versus Fits



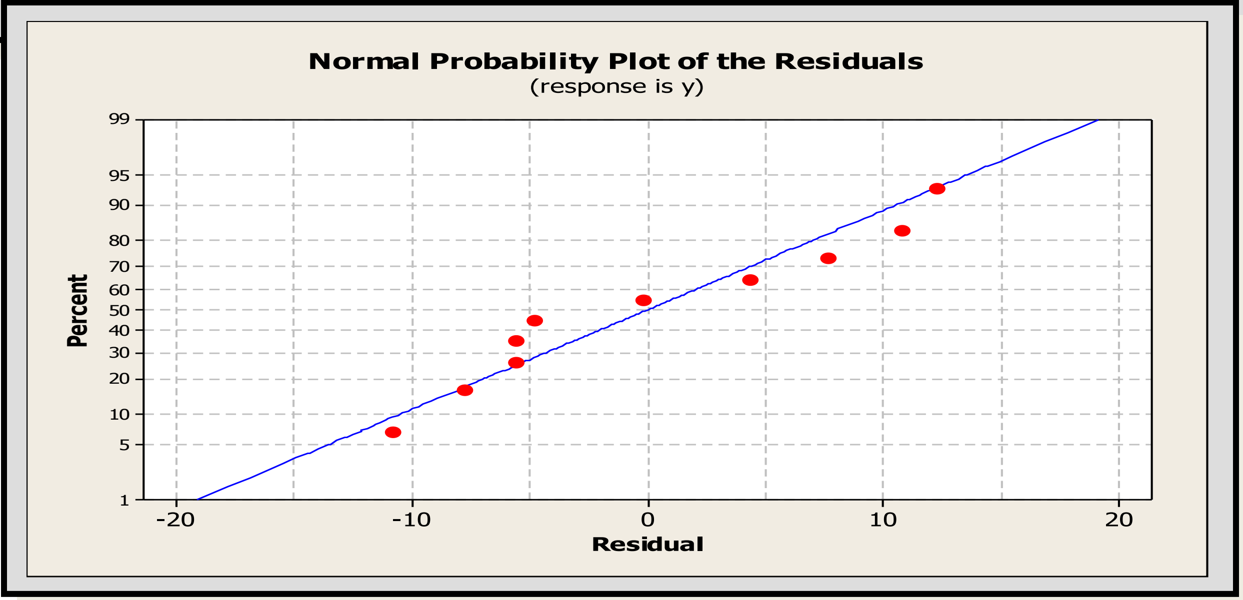
- ✓ If the equal variance assumption is valid, the plot should appear as a random scatter around the zero center line.
- ✓ If not, you will see a pattern in the residuals.



NORMAL PROBABILITY PLOT



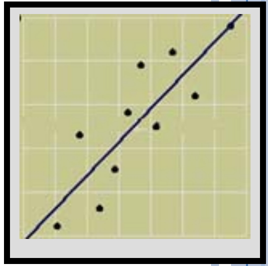
- ✓ If the normality assumption is valid, the plot should resemble a straight line, sloping upward to the right.
- ✓ If not, you will often see the pattern fail in the tails of the graph.



12.5

Estimation and prediction using the fitted line

Estimation and Prediction

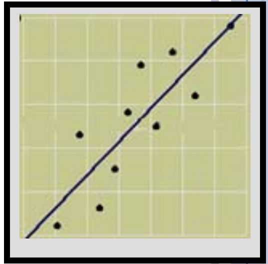


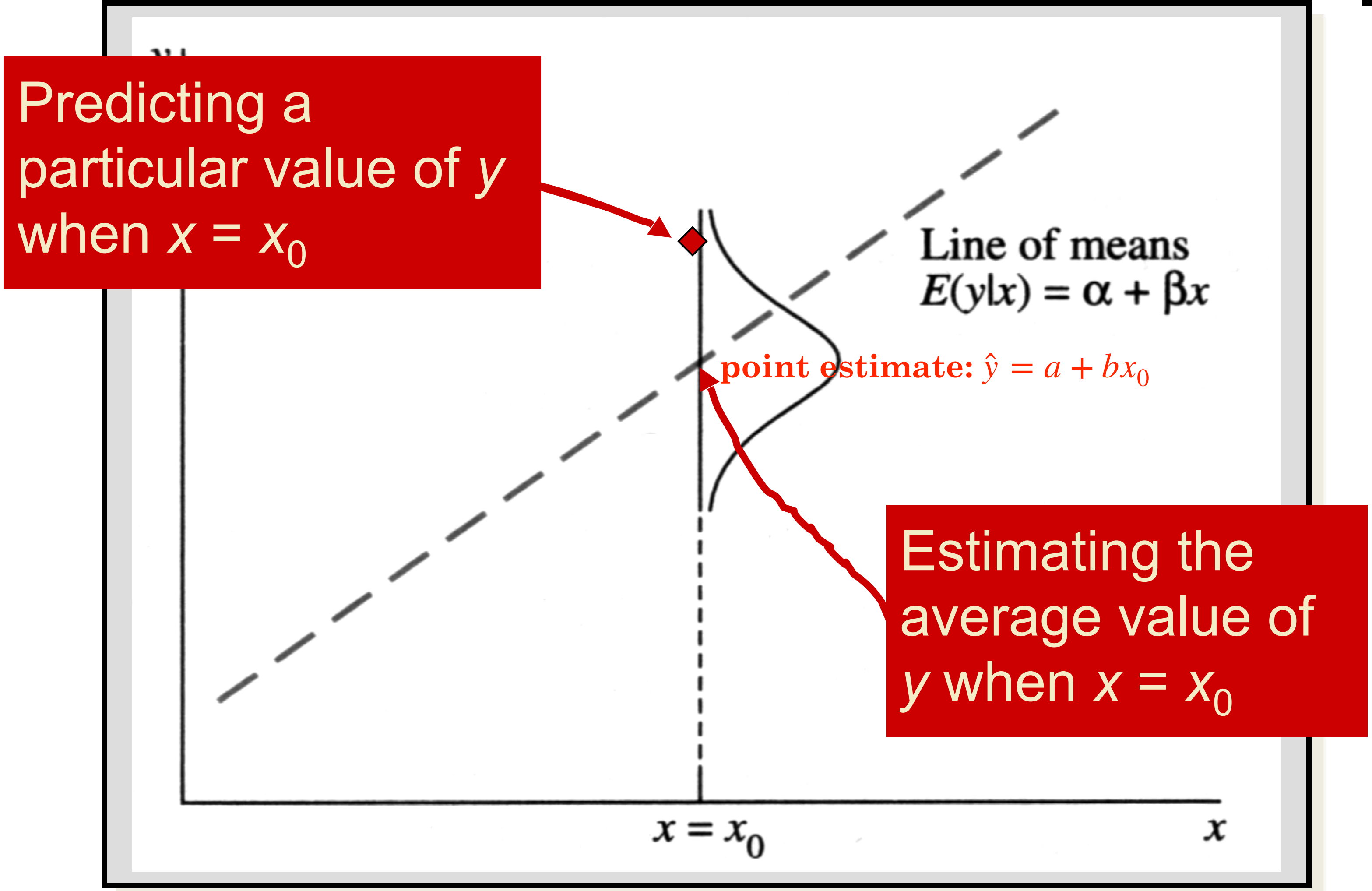
- Once you have
  - ✓ (12.4) used the diagnostic plots to check for violation of the regression assumptions.
  - ✓ (12.2) determined that the regression line is useful $(\beta \neq 0)$
- You are ready to use the regression line to

✓ *Estimate* the average value of *y* for a given value of *x:* $E(y \mid x)$

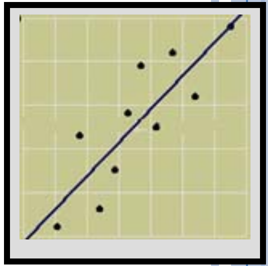
✓ *Predict* a particular value of *y* for a given value of *x:* y_{new}

Estimation and Prediction

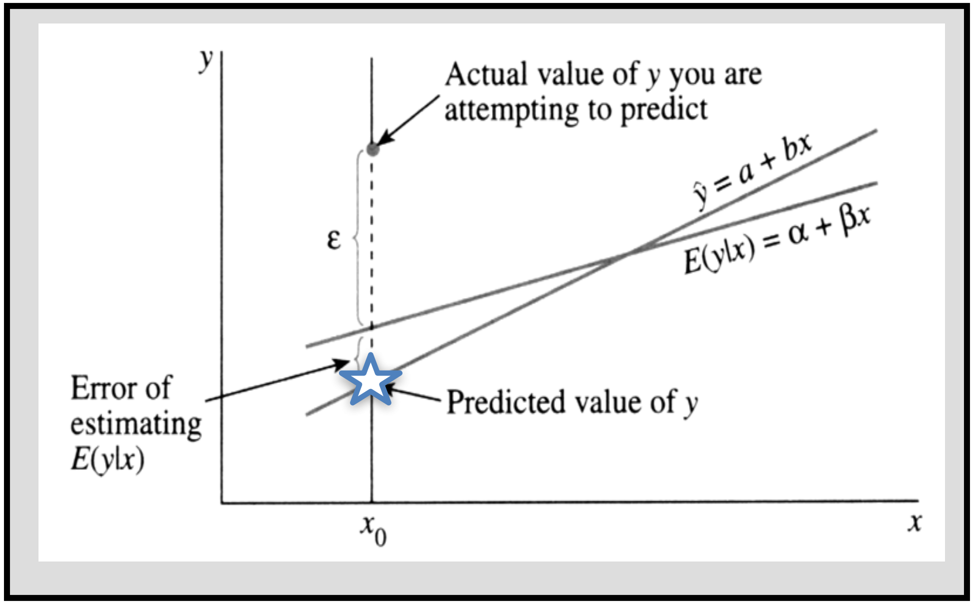




ESTIMATION AND PREDICTION



- The best estimate of either E(*y*) or *y* for a given value $x = x_0$ is $\hat{y} = a + bx_0$
- Particular values of *y* are more difficult to predict, requiring a wider range of values in the prediction interval.



How do we obtain CI

- Target: $E(y|x_0) = \alpha + \beta x_0$

- We use $\hat{y} = a + bx_0$ to estimate $E(y|x_0)$

- Appendix has shown that $(a + bx_0) \sim N(\alpha + \beta x_0, \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}))$

- Therefore, $\dfrac{(a + bx_0) - (\alpha + \beta x_0)}{\sqrt{MSE(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}} \sim t_{(n-2)}$

- To obtain the CI, we start with

- $$1 - \alpha = P\left(-t_{(n-2);\alpha/2} < \frac{(a + bx_0) - (\alpha + \beta x_0)}{\sqrt{MSE(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}} < t_{(n-2);\alpha/2} \right)$$

How do we obtain prediction interval?

✓ *Predict* a particular value of *y* for a given value of *x*: y_{new}

- Target: $y_{new} = \alpha + \beta x_0 + \varepsilon_{new}$
- We use $\hat{y}_{new} = a + bx_0 + 0 = a + bx_0$ to predict y_{new}
- Vedio shows the forecast error:

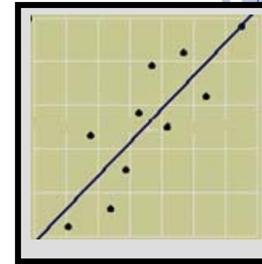
$$\epsilon = y_{new} - \hat{y}_{new} = (\alpha + \beta x_0 + \varepsilon_{new}) - ((a + bx_0)) \sim N(0, \sigma^2(1 + \frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx}))$$

- Therefore, $\dfrac{(y_{new}) - (\hat{y}_{new})}{\sqrt{MSE(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}} \sim t_{(n-2)}$
- To obtain the PI, again, we start with

$$1 - \alpha = P\left(-t_{(n-2);\alpha/2} < \frac{y_{new} - (\hat{y}_{new})}{\sqrt{MSE(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}} < t_{(n-2);\alpha/2}\right)$$

- $$= P\left(-t_{(n-2);\alpha/2}\sqrt{MSE(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})} < y_{new} - (\hat{y}_{new}) < t_{(n-2);\alpha/2}\sqrt{MSE(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}\right)$$

Estimation and Prediction



Confidence interval (CI) for the *average value* of y given $x = x_0$

$$\hat{y} \pm t_{(n-2),\alpha/2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Prediction interval (PI) for a *particular value* of y given $x = x_0$

$$\hat{y} \pm t_{(n-2),\alpha/2}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

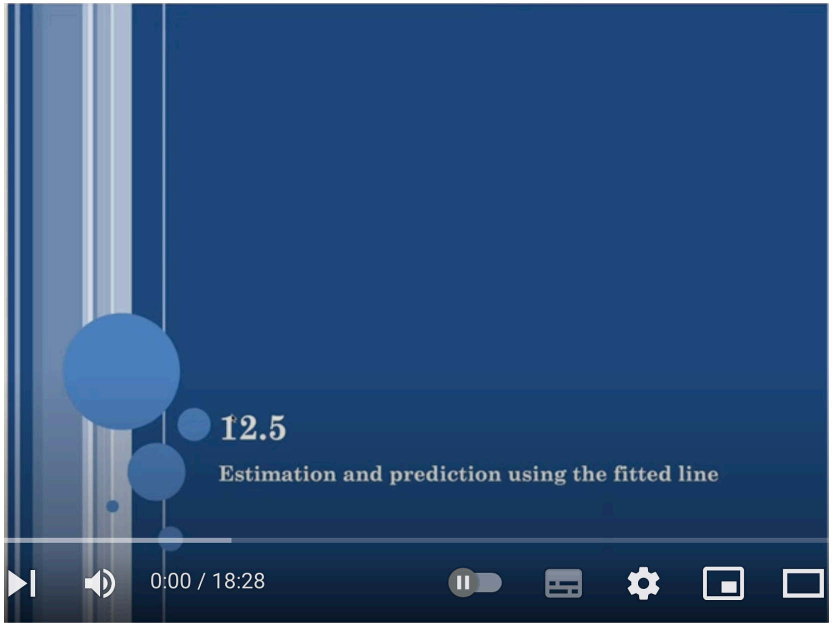
$$MSE = \frac{SSE}{n-2}$$

Procedures in finding the best fitting regression line

1. Calculate $(\sum x_i)$, $(\sum y_i)$, $(\sum x_i^2)$, $(\sum y_i^2)$, $(\sum x_i y_i)$, $\bar{x}$, and $\bar{y}$
2. Calculate
   - $S_{xx} = \sum(x-\bar{x})^2 = \sum x^2 - (\sum x)^2/n.$
   - $S_{yy} = \sum(y-\bar{y})^2 = \sum y^2 - (\sum y)^2/n.$
   - $S_{xy} = \sum(x-\bar{x})(y-\bar{y}) = \sum xy - (\sum x)(\sum y)/n.$
3. Calculate
   - $b = S_{xy}/S_{xx}$
   - $a = \bar{y} - b\bar{x}$

   Best fitting line: $\hat{y} = a + bx$ with a and b in Step 3.
4. Complete the ANOVA table:

   $$\text{totalSS} = S_{yy},\ \text{SSR} = S_{xy}^2/S_{xx},\ \text{SSE} = \text{totalSS} - \text{SSR}.$$



https://youtu.be/onFoHVDoWhA?si=9X8NB8GdE0sZFUaA

https://youtu.be/y1LXtX_G1mk?si=I-fs5XB_qR7MwU0i

THE CALCULUS PROBLEM

- Estimate the average calculus grade for students whose achievement score is 50 with a 95% confidence interval.

$$\text{Calculate } \hat{y} = 40.78424 + .76556(50) = 79.06$$

$$\hat{y} \pm 2.306\sqrt{75.7532\left(\frac{1}{10} + \frac{(50-46)^2}{2474}\right)}$$

$$79.06 \pm 6.55 \quad \text{or } 72.51 \text{ to } 85.61.$$

THE CALCULUS PROBLEM

- Predict the calculus grade for **a particular student** whose achievement score is 50 with a 95% prediction interval.

$$\text{Calculate } \hat{y} = 40.78424 + .76556(50) = 79.06$$

$$\hat{y} \pm 2.306\sqrt{75.7532\left(1 + \frac{1}{10} + \frac{(50-46)^2}{2474}\right)}$$

$$79.06 \pm 21.11 \quad \text{or } 57.95 \text{ to } 100.17.$$

Notice how much wider this interval is!

MINITAB OUTPUT

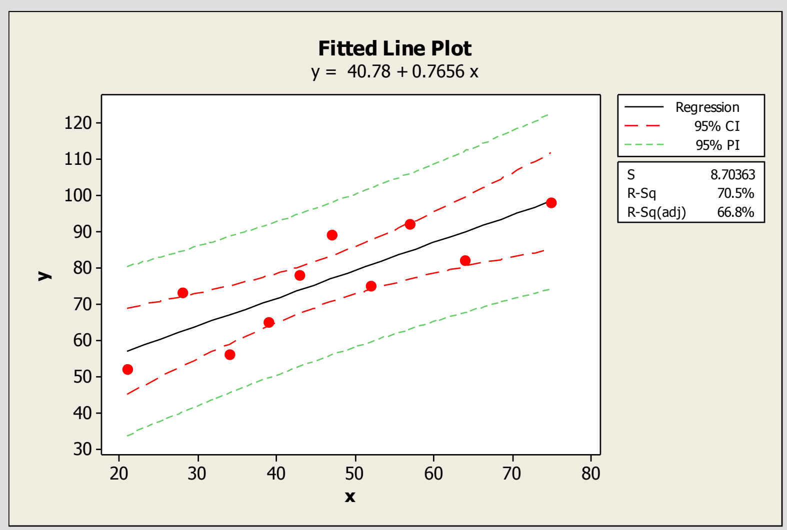
Confidence and

```
Predicted Values for New Observations
New Obs    Fit       SE Fit      95.0% CI            95.0% PI
1         79.06       2.84     (72.51, 85.61)   (57.95,100.17)

Values of Predictors for New Observations
New Obs             x
1                 50.0
```

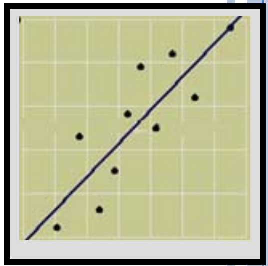
✓Green prediction bands are always wider than red confidence bands.

✓Both intervals are narrowest when x = x-bar.



12.6 Correlation Analysis

CORRELATION ANALYSIS



- The strength of the relationship between *x* and *y* is measured using the **coefficient of correlation**:

$$\text{Correlation coefficient}: r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

- Recall from Chapter 3 that

(1) $-1 \le r \le 1$ (2) r and b have the same sign

(3) $r \approx 0$ means no linear relationship

(4) $r \approx 1$ or -1 means a strong (+) or (-) linear relationship

Example

The table shows the heights and weights of $n = 10$ randomly selected college football players.

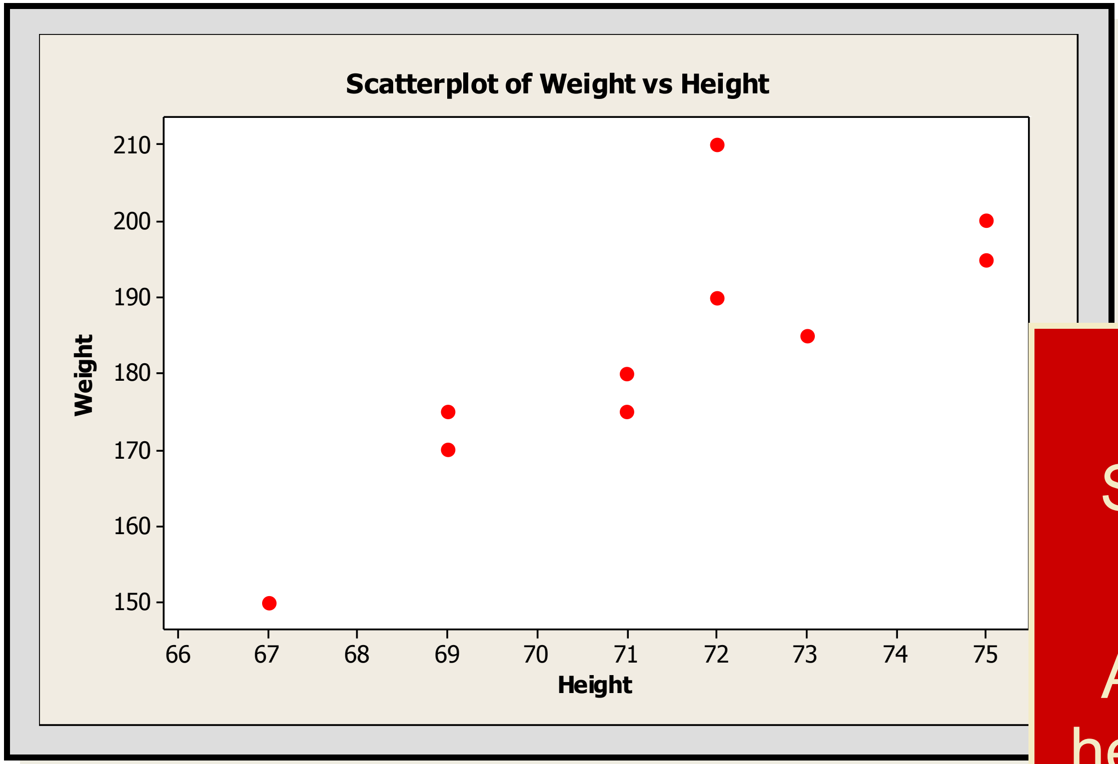
Player	1	2	3	4	5	6	7	8	9	10
Height, x	73	71	75	72	72	75	67	69	71	69
Weight, y	185	175	200	210	190	195	150	170	180	175

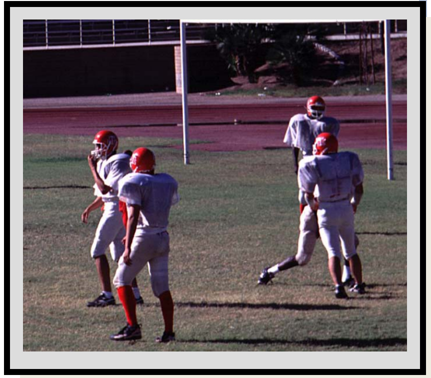
Use your calculator to find the sums and sums of squares.

$$S_{xy} = 328 \quad S_{xx} = 60.4 \quad S_{yy} = 2610$$

$$r = \frac{328}{\sqrt{(60.4)(2610)}} = .8261$$

Football Players





r = .8261

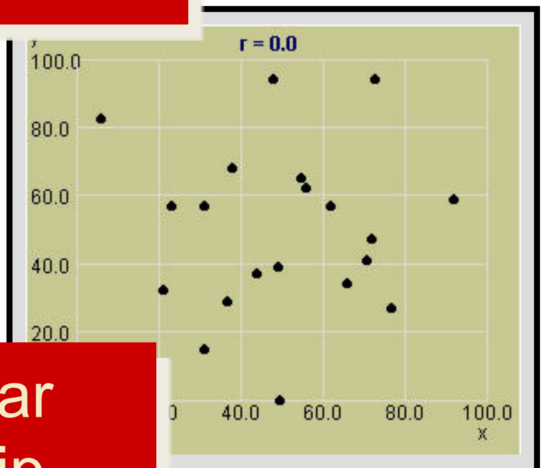
Strong positive correlation

As the player's height increases, so does his weight.

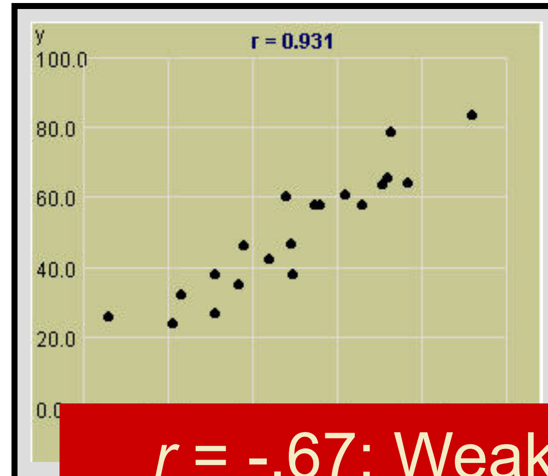
Some Correlation Patterns

- Use the **Exploring Correlation** applet to explore some correlation patterns:

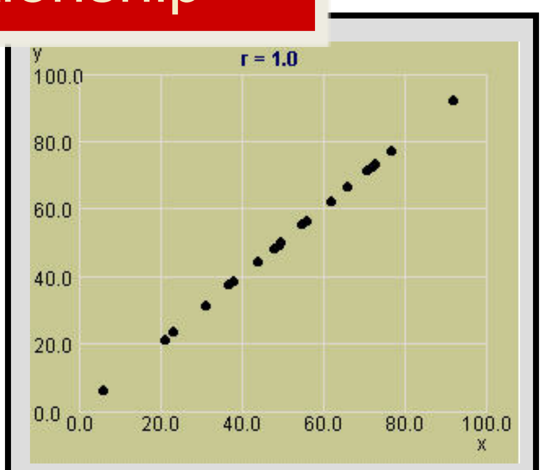
r = 0; No (linear) correlation



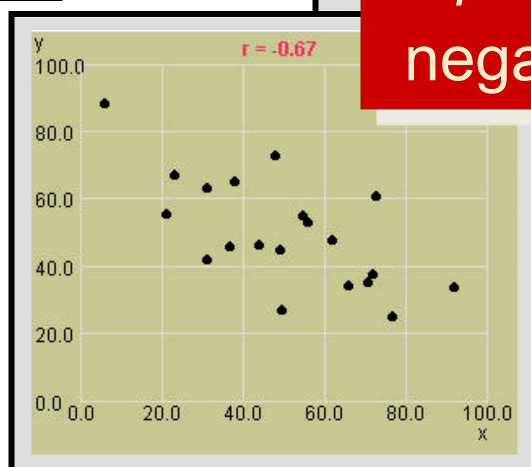
r = .931; Strong positive correlation



r = 1; Linear relationship



r = -.67; Weaker negative correlation



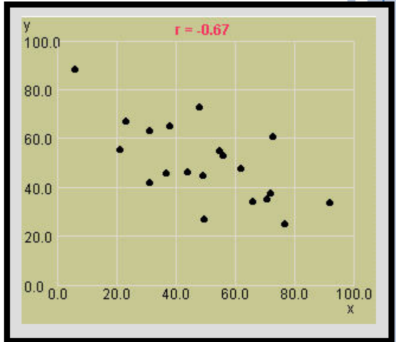
MY APPLET

ρ V.S. r

- Population quantity:
- $$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sqrt{E(X-\mu_X)^2 E(Y-\mu_Y)^2}}$$

- Sample estimate for ρ
- $$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2}}.$$
- $$t_{STAT} = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{(n-2)}$$

INFERENCE USING *R*



- The population **coefficient of correlation** is called ρ ("rho"). We can test for a significant correlation between *x* and *y*

$$\text{To test } H_0: \rho = 0 \text{ versus } H_a: \rho \neq 0$$

$$\text{Test Statistic: } t = r\sqrt{\frac{n-2}{1-r^2}}$$

$$\text{Reject } H_0 \text{ if } t > t_{\alpha/2} \text{ or } t < -t_{\alpha/2} \text{ with } n-2 \; df.$$

This test is exactly equivalent to the *t*-test for the slope β=0.

Equal to the t-test for $H_0 : \beta = 0$?

$$\left(r\sqrt{\frac{n-2}{1-r^2}}\right)^2 = r^2\frac{n-2}{1-r^2} = R^2\frac{n-2}{1-R^2}$$

$$= \frac{(n-2)\text{SSR}/\text{SSTotal}}{\text{SSE}/\text{SSTotal}}$$

$$= \frac{\text{SSR}}{\text{SSE}/(n-2)} = \frac{\text{SSR}/1}{\text{SSE}/(n-2)}$$

$$= \frac{\text{MSR}}{\text{MSE}} = F_{STAT} = (t_{STAT})^2 .$$

- Recall
  - $r^2 = R^2$
  - $\text{SSE} + \text{SSR} = \text{SSTotal}$
  - $F_{STAT} = (t_{STAT})^2$

$r = .8261$

EXAMPLE

Is there a significant positive correlation between weight and height in the population of all college football players?

$$H_0 : \rho = 0$$
$$H_a : \rho > 0$$

Test Statistic: $t = r\sqrt{\dfrac{n-2}{1-r^2}}$

$$= .8261\sqrt{\frac{8}{1-.8261^2}} = 4.15$$

Use the *t*-table with *n*-2 = 8 df to bound the *p*-value as *p*-value < .005. There is a significant positive correlation.

MY APPLET

Key Concepts

I. A Linear Probabilistic Model

1. When the data exhibit a linear relationship, the appropriate model is $y = \alpha + \beta x + \varepsilon$.
2. The random error ε has a normal distribution with mean 0 and variance σ^2.

II. Method of Least Squares

1. Estimates a and b, for α and β, are chosen to minimize SSE, the sum of the squared deviations about the regression line, $\hat{y} = a + bx$.
2. The least squares estimates are $b = S_{xy}/S_{xx}$ and $a = \bar{y} - b\bar{x}$.

Key Concepts

III. Analysis of Variance

1. Total SS = SSR + SSE, where Total SS = S_{yy} and SSR = $(S_{xy})^2 / S_{xx}$.
2. The best estimate of σ^2 is MSE = SSE / $(n - 2)$.

IV. Testing, Estimation, and Prediction

1. A test for the significance of the linear regression— $H_0 : \beta = 0$ can be implemented using one of two test statistics:

$$t = \frac{b}{\sqrt{\text{MSE} / S_{xx}}} \quad \text{or} \quad F = \frac{\text{MSR}}{\text{MSE}}$$

Key Concepts

2. The strength of the relationship between x and y can be measured using

$$R^2 = \frac{\text{SSR}}{\text{Total SS}}$$

which gets closer to 1 as the relationship gets stronger.

3. Use **residual plots** to check for nonnormality, inequality of variances, and an incorrectly fit model.

4. **Confidence intervals** can be constructed to estimate the intercept α and slope β of the regression line and to estimate the average value of y, $E(y)$, for a given value of x.

5. **Prediction intervals** can be constructed to predict a particular observation, y, for a given value of x. For a given x, prediction intervals are always wider than confidence intervals.

KEY CONCEPTS

V. Correlation Analysis

1. Use the correlation coefficient to measure the relationship between x and y when both variables are random:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

2. The sign of r indicates the direction of the relationship; r near 0 indicates no linear relationship, and r near 1 or –1 indicates a strong linear relationship.
3. A test of the significance of the correlation coefficient is identical to the test of the slope β.