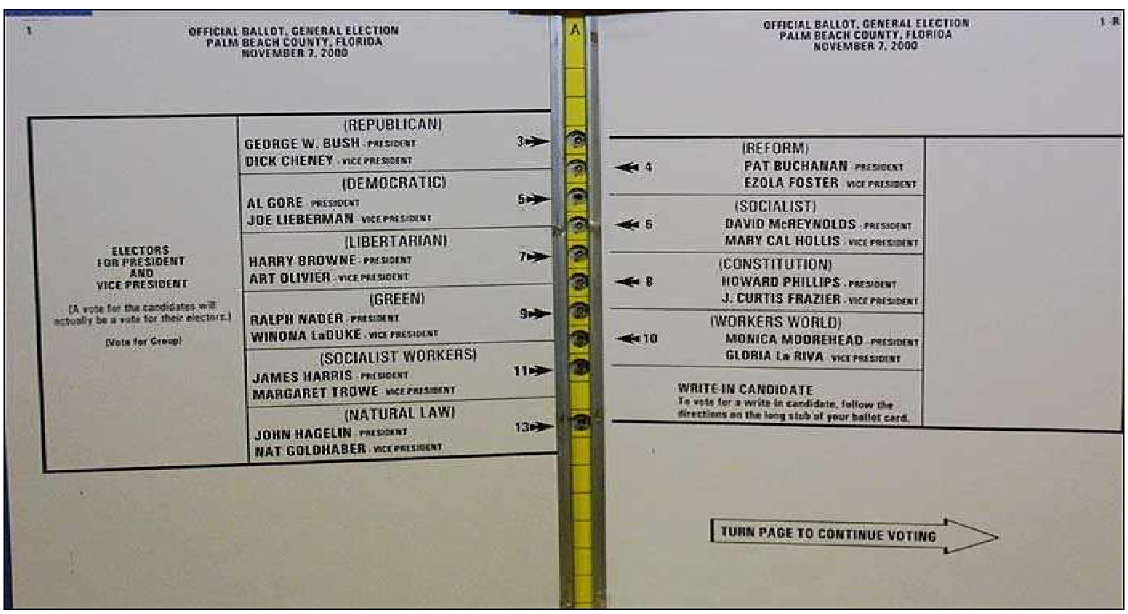
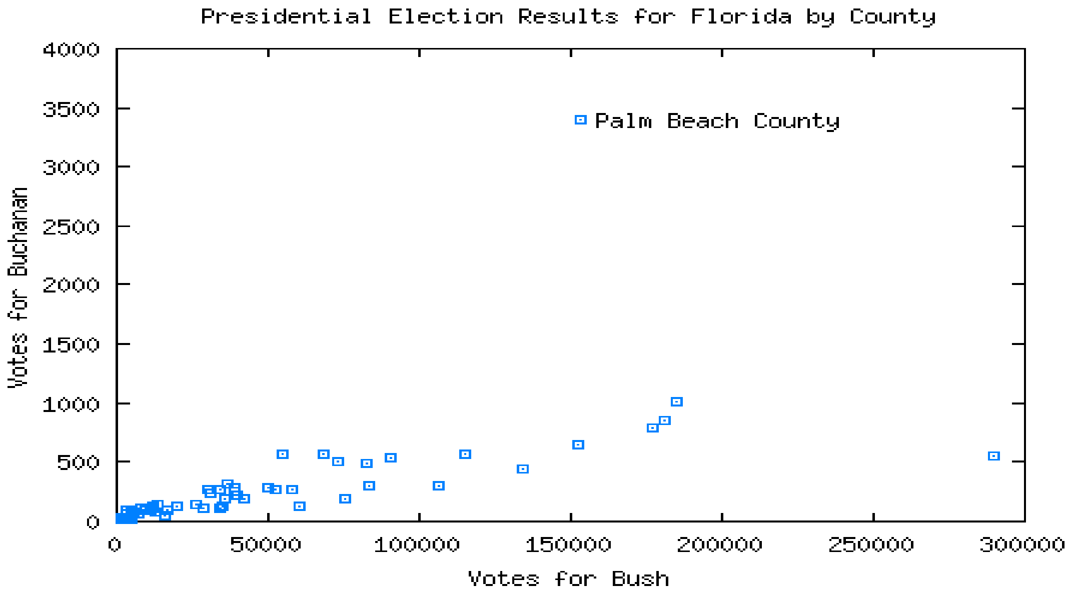
12 Simple Linear Regression and Correlation

Introduction:

A **scatterplot** graphically illustrates the relationship between two quantitative variables.

Example: In the Presidential Election of 2000, George W. Bush earned 537 votes more than Al Gore in Florida, granting the Presidency to Bush. However, the Palm Beach County, Florida, the "butterfly ballot" possibly caused some individuals to mistakenly vote for Pat Buchanan rather than Gore.





Examining the relationship between variables is called **regression analysis**.

Typically for **two** variables,

x is the **explanatory** variable.

y is the **response** variable.

Example: For the following pairs of variables, which is the explanatory variable, and which is the response variable?

(a) number of years of education and income

(b) blood pressure (systolic) and weight

(c) height of sons and height of fathers

(d) score on midterm exam and score on final exam

(e) score on SAT and final GPA in college

(f) deficit spending and interest rates

(g) temperature and ozone in atmosphere

Two purposes of regression analysis:

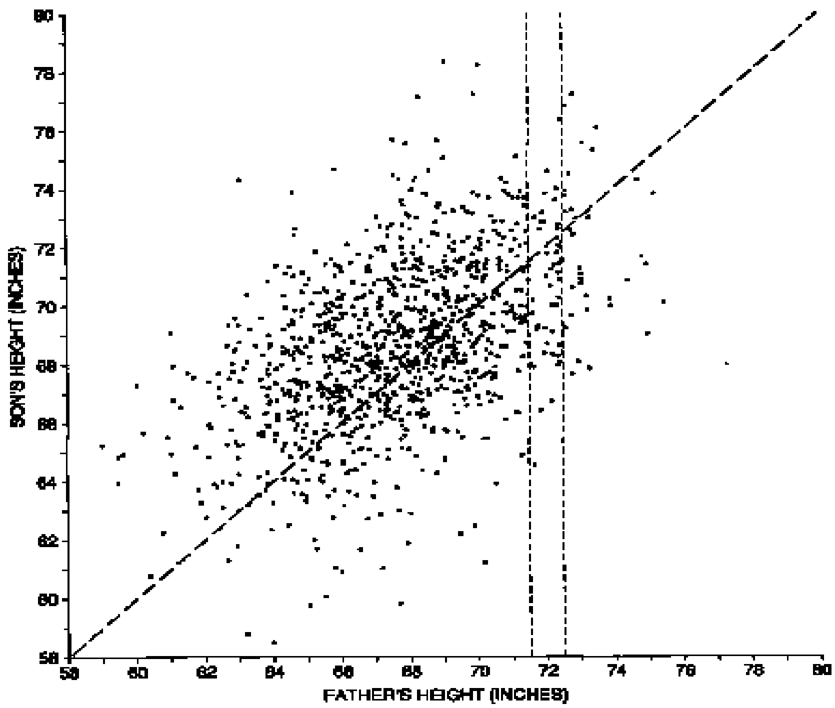
1. explain

2. predict

12.1 The Simple Linear Regression Model

Examining the **linear** relationship between **two** variables is called **simple linear regression**.

Example: Heights of 1078 fathers and sons, England, around year 1900.



□

For **simple linear regression**, the (X_i, Y_i) data pairs are **linearly** related, and the Y_i observations given X_i are *usually* assumed **independent**, for $i = 1, \ldots, n$.

Simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where the ε_i are **independent** $N(0, \sigma)$, such that ε_i is **independent** of x_i, for $i = 1, \ldots, n$.

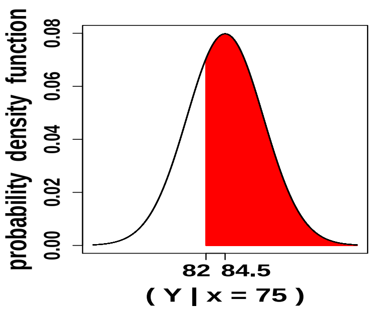
- Determine $E(Y_i|x_i)$.

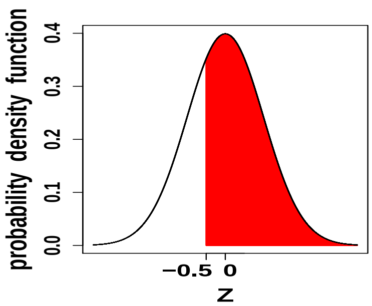
- Determine $V(Y_i|x_i)$.

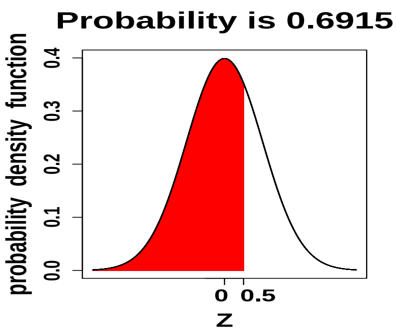
$\square$

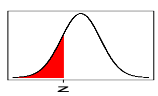
Example: Let x be the student's score on exam **#1**, and y be the student's score on exam **#2**. Suppose that the **true** regression model is $y = 2 + 1.1x + \varepsilon$, where $\varepsilon \sim N(0, \sigma = 5)$ such that ε is **independent** of x.

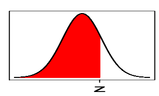
If a student scored a **75** on exam **#1**, determine the probability that this same student will score at least an **82** on exam **#2**.











Standard normal table, pp. 722–723

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

□

12.2 Estimating Model Parameters

Simple linear regression model:

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

where the ε are **independent** $N(0, \sigma)$, such that ε is **independent** of x.

The parameters $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are **unknown** and will be estimated by the method of **least squares**.

The **least squares** estimates of β_0 and β_1 are the values of b_0 and b_1 which **minimize**

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2.$$

The **least squares** estimators produce the **estimated regression line** or **least squares line**: $\boldsymbol{y = \hat{\beta}_0 + \hat{\beta}_1 x}$

Example: Show how the **least squares** estimators $\boldsymbol{\hat{\beta}_1}$ and $\boldsymbol{\hat{\beta}_2}$ may be determined.

□

Using short-cut formulas,

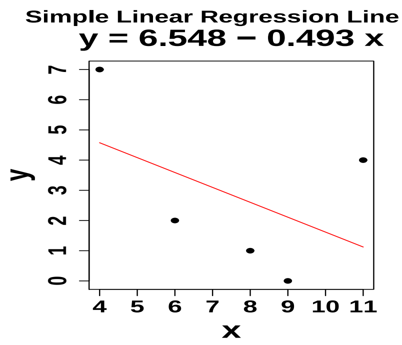
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \, \bar{x}$.

It can be shown that $\boldsymbol{\hat{\beta}_0}$ and $\boldsymbol{\hat{\beta}_1}$ are **unbiased** for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$, respectively.

Example: Consider the following (x, y) data.

x	y
4	7
6	2
9	0
8	1
11	4



(a) Determine the **least squares line**.

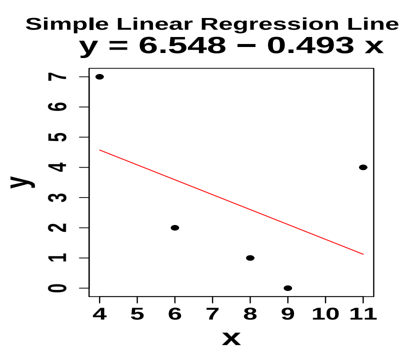
(b) **Predict** a new value of y if $x = 9$.

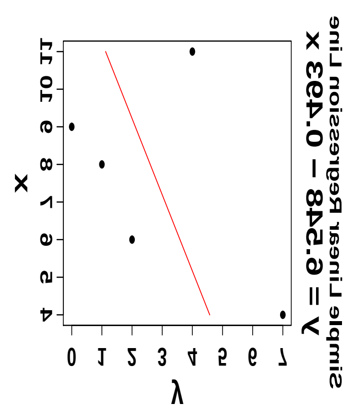
(c) **Estimate** the **mean** value of y if $x = 9$.

(d) **Predict** a new value of y if $x = 15$.

□

Remark: The least squares line of y on x differs from the least squares line of x on y.





For a given value of x, the **fitted** or **predicted** value of y is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1\, x$.

The **residuals** are the **vertical** deviations, $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$.

Why are these **residuals** valuable?

Example: *Revisit.* Using the following (x, y) data, estimate σ in the **linear regression model**,

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

where the ε are **independent** $N(0, \sigma)$, such that ε is **independent** of x.

x	y
4	7
6	2
9	0
8	1
11	4

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}^2}{n-2}$$

Short-cut formula:

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-2}\left[\sum_{i-1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i\right]$$

You need NOT memorize this formula.

□

12.3 Inferences About the Slope Parameter, $\boldsymbol{\beta_1}$

Simple linear regression model:

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

where the ε are **independent** $N(0, \sigma)$, such that ε is **independent** of x.

Consider the hypothesis test,

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

What are we effectively testing?

What is a reasonable point estimator of β_1?

The **standard error** of $\hat{\beta}_1$ is

$$s_{\hat{\beta}_1} = s/\sqrt{S_{xx}},$$

where

$$S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2, \quad \text{and}$$

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-2} [\sum_{i-1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i]$$

You need NOT memorize this formula for $s_{\hat{\beta}_1}$.

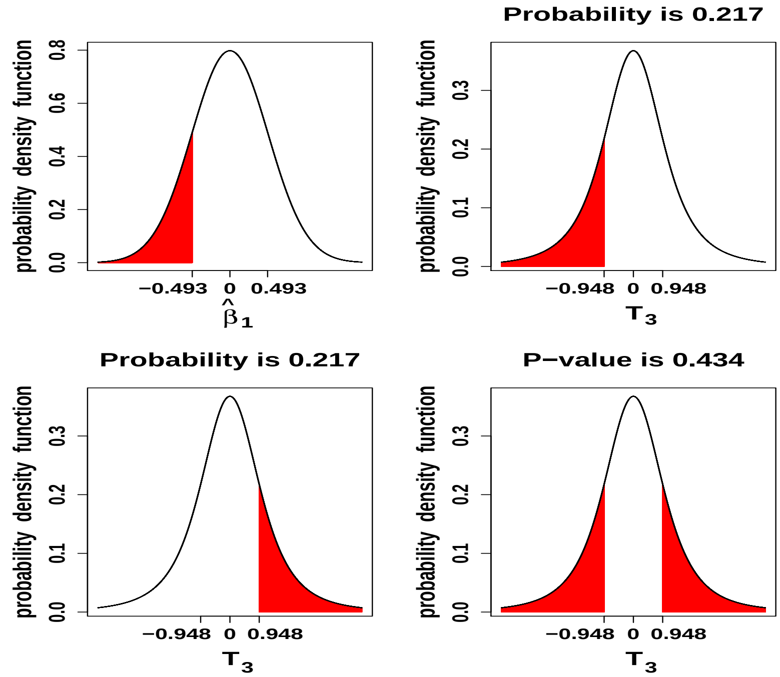
Example: *Revisit.* Using the following (x, y) data, test for nonzero slope in the **simple linear regression** model at level $\alpha = 0.1$.

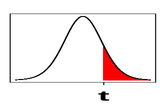
x	y
4	7
6	2
9	0
8	1
11	4

(a) State the null and alternative hypotheses.

(b) Determine the value of the **standardized test statistic**.

(c) Determine the P-value.



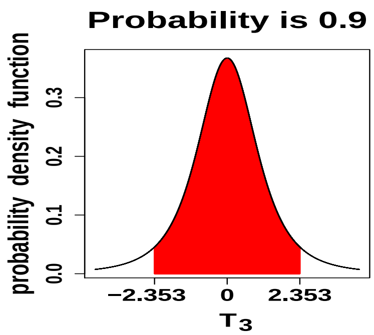


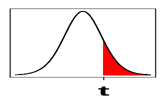
t-table, pp. 728–729

$t \setminus \nu$	1	2	3	4			
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.8	.285	.254	.241	.234			
0.9	.267	.232	.217	.210			
1.0	.250	.211	.196	.187			
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(d) State the conclusion in statistical terms and in regular English.

(e) Construct a **90%** confidence interval on the slope parameter.





t-table, p. 725

	α						
ν	**.10**	**.05**	**.025**	**.01**	**.005**	**.001**	**.0005**
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

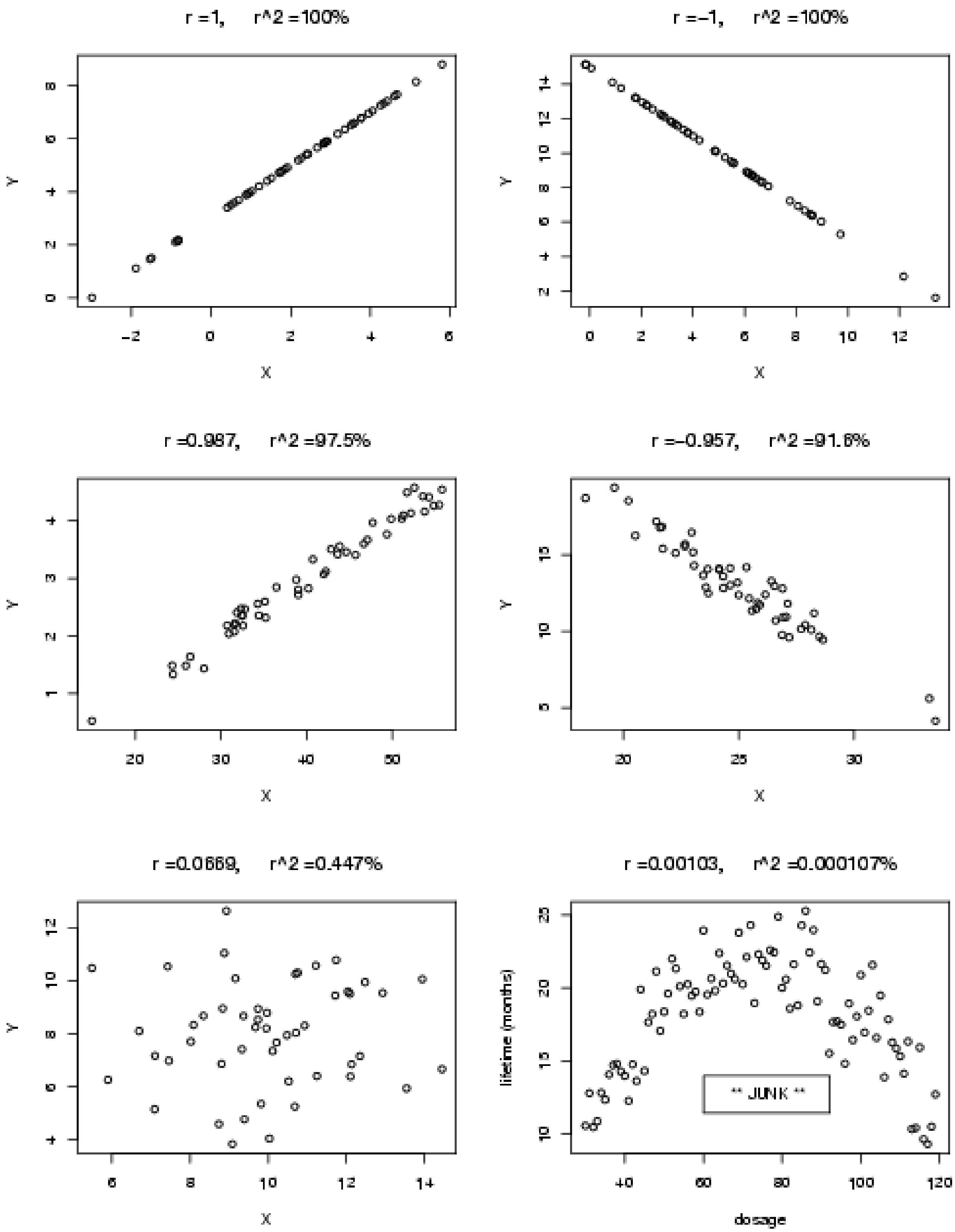
□

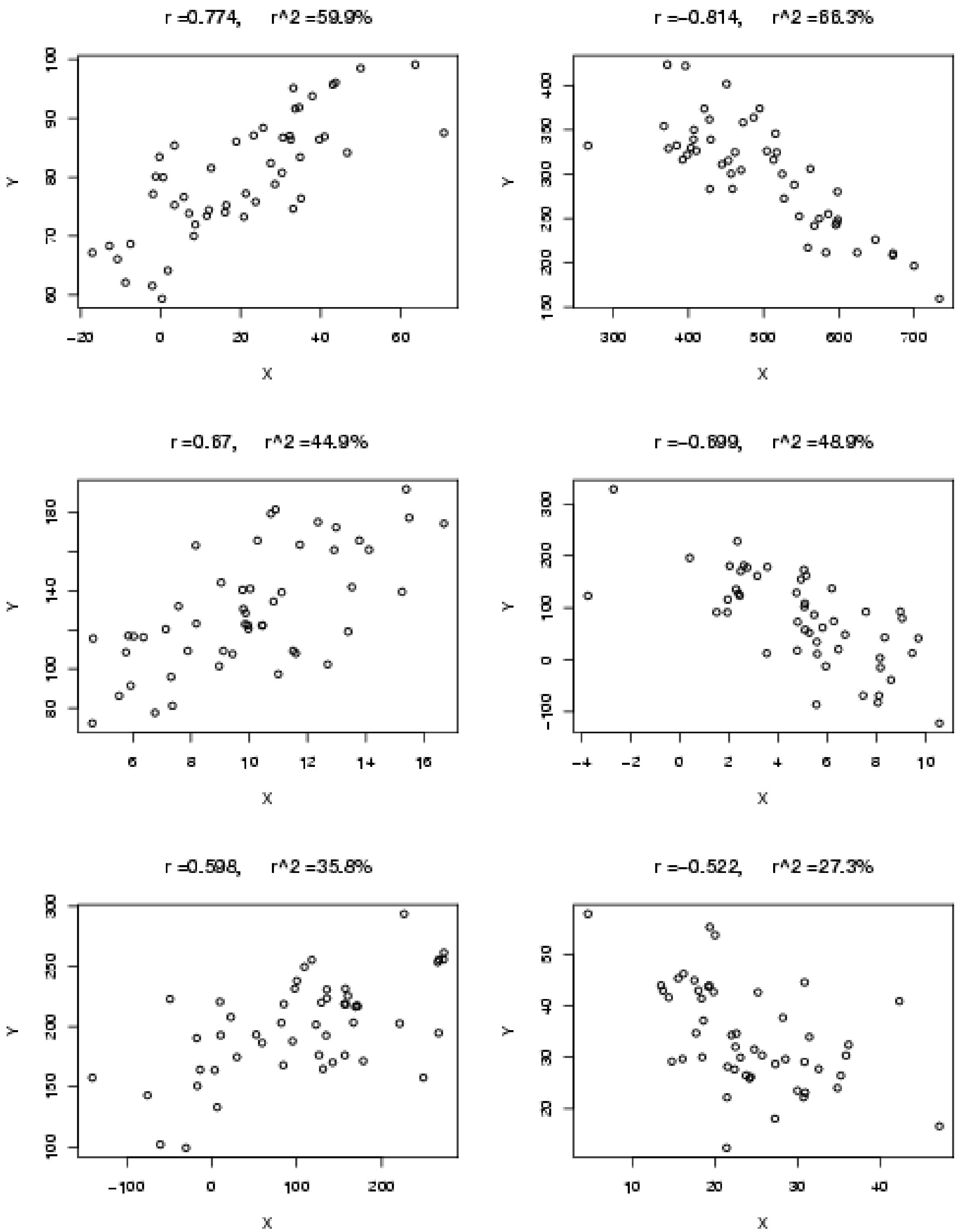
12.5 Correlation

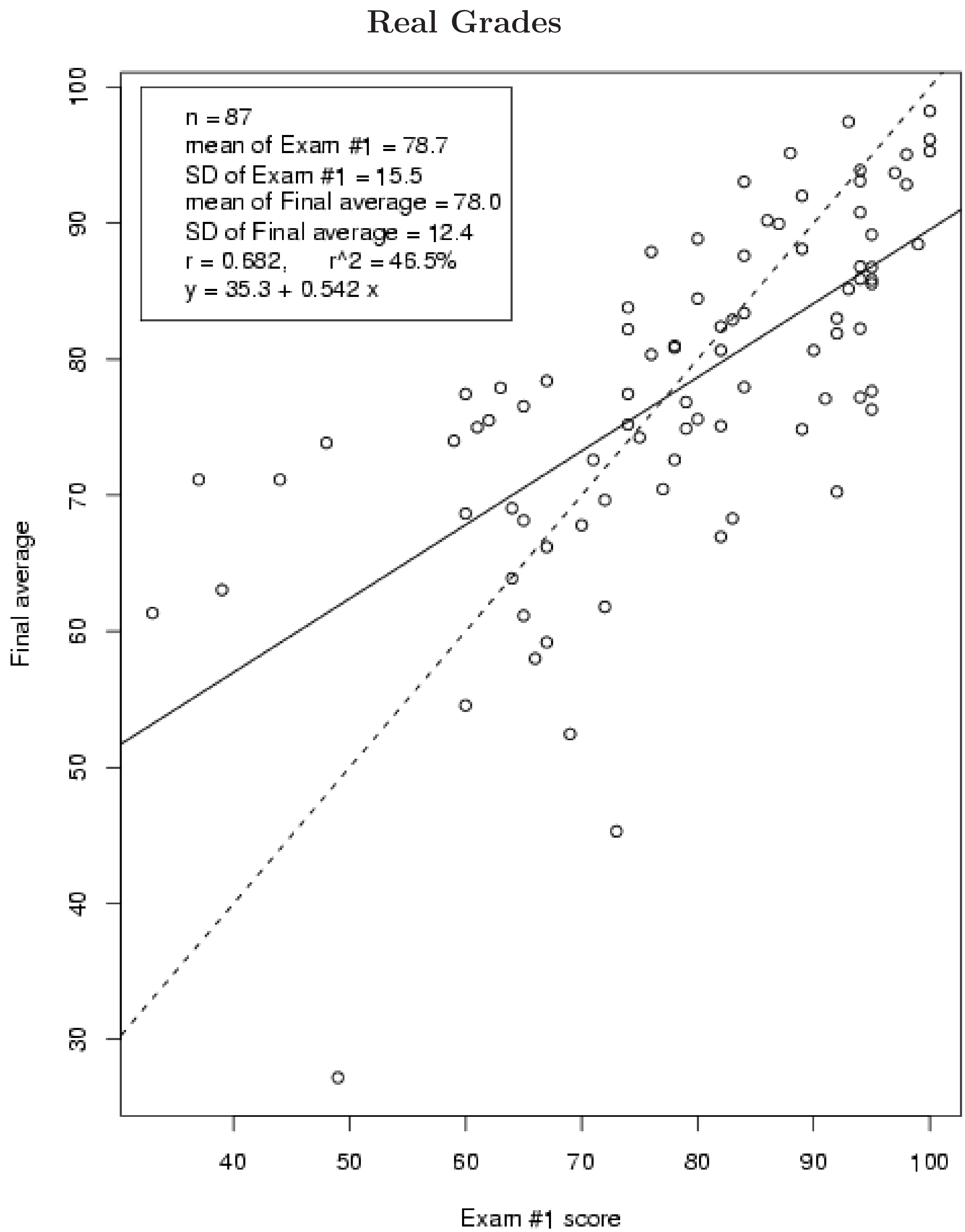
Recall from chapter 5, the **population** correlation coefficient between X and Y is $\rho_{x,y} = \rho = \text{cov}(X,Y)/[\sigma_x \ \sigma_y]$, and is a measure of the **linear** association between X and Y.

Based on data, we can estimate $\boldsymbol{\rho}$ using the (Pearson) **sample** correlation coefficient

$$\hat{\rho} = r = \frac{\frac{1}{n-1} \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x \ s_y}.$$


r =1, r^2 =100%
r =−1, r^2 =100%
r =0.987, r^2 =97.5%
r =−0.957, r^2 =91.6%
r =0.0669, r^2 =0.447%
r =0.00103, r^2 =0.000107%
** JUNK **
lifetime (months)
dosage
X
Y


r =0.774, r^2 =59.9%
r =-0.814, r^2 =66.3%
r =0.67, r^2 =44.9%
r =-0.699, r^2 =48.9%
r =0.598, r^2 =35.8%
r =-0.522, r^2 =27.3%
X
Y



Remarks regarding r (similar to those regarding ρ in chapter

5):

(a) Is r random or fixed?

(b) What are the units on r?

(c) What are the possible values of r?

(d) $r = 1$ implies what type of correlation?

(e) $r = -1$ implies what type of correlation?

(f) Is selection of x and y relevant when calculating r?

(g) r makes sense for linear associations only.

(h) A linear transformation on the data does not affect $|r|$.

(i) As the number of (x, y) data pairs becomes huge, r "converges" to the **population** correlation, ρ.

(j) Correlation does not imply causation.

Example: Consider the two variables "weight of **older** brother at age 5" and "weight of **younger** brother at age 5."

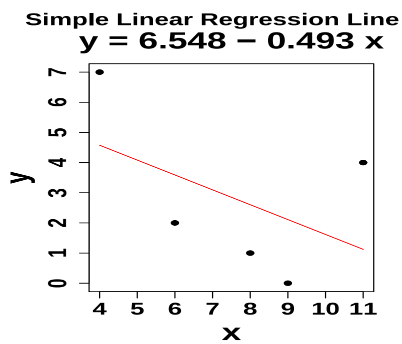
A **lurking variable** is a third variable which confuses the relationship between the two variables of interest.

(k) r^2 is called the **coefficient of determination** (introduced in section 12.2) and measures the proportion of variability in y that can be explained by x due to the **linear** relationship between x and y.

What are the possible values of r^2?

Example: Revisit the following (x, y) data. Compute r and r^2.

x	y
4	7
6	2
9	0
8	1
11	4



□

Example: Under the *lofty* assumption that the final score is based upon 5 equally weighted INDEPENDENT exams with a common variance, then r^2 (at least for the entire data set of 87 students) should be about what number?

□

Additional formulas:

$\hat{\beta}_1 = r s_y / s_x$

$$s_{\hat{\beta}_1} = \frac{s_y}{s_x}\sqrt{\frac{1-r^2}{n-2}}$$

You need NOT memorize this formula for $s_{\hat{\beta}_1}$.

Example: *New data set.*

Consider the **simple linear regression model:**

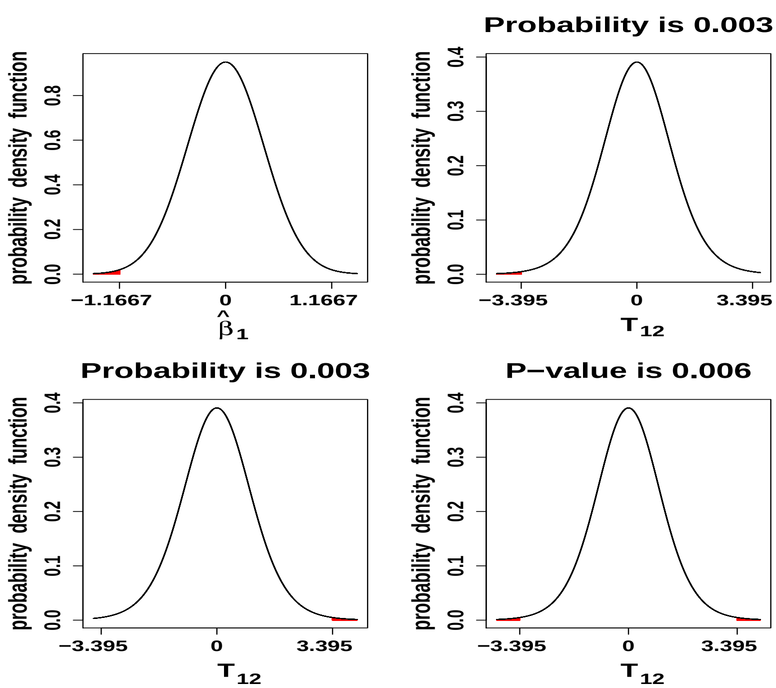
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

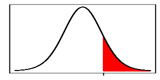
where the ε_i are **independent** $N(0, \sigma)$, such that ε_i is **independent** of x_i, for $i = 1, \ldots, n$.

Suppose $n = 14$, $\bar{x} = 21$, $s_x = 3$, $\bar{y} = 38$, $s_y = 5$, and $r = -0.7$.

(a) Determine the equation of the **fitted regression line**.

(b) Test for a nonzero slope parameter at level 0.1.





t-table, pp. 728–729

$t \setminus \nu$	9	10	11	12	13	14	15
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.3	.005	.004	.004	.003	.003	.003	.002
3.4	.004	.003	.003	.003	.002	.002	.002
3.5	.003	.003	.002	.002	.002	.002	.002
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

□