1 Statistical Models and Conditional Expectation

"essentially, all models are wrong, but some are useful"

George E.P. Box

Mathematical model

A **mathematical model** is a description of a system using mathematical concepts and language.

Statistical model

A **statistical model** embodies a set of assumptions concerning the generation of the observed data, and similar data from a larger population.

Relations between variables

A **functional relation** between two variables is expressed by a mathematical formula. If x is the independent variable and y is the dependent variable, then a function relation is of the form:

$$y = f(x).$$

A **statistical relation**, unlike a functional relation, is not a perfect one. In general, the observations for a statistical relation do not fall directly on the curve of relationship. This is commonly expressed as a functional relation coupled with a random error ϵ. If x is the independent variable and Y is the dependent variable, then a statistical relation *often* takes the form:

$$Y = f(x) + \epsilon.$$

A statistical relation is also commonly expressed in terms of **conditional expectation**. That is, for random variables Y and X,

$$Y = E[Y|X = x] + \epsilon.$$

1.1 Conditional Expectation

Goal: The goal of this section is to motivate conditional expectation:

$$E[Y|X = x]$$

Consider an example from probability theory.

Example 1

Consider rolling two fair six sided dice (D_1 & D_2) and recording the sum of the faces and the maximum of the faces. Define two random variables $Y = D_1 + D_2$ and $X = \max\{D_1, D_2\}$. The joint probability distribution $P(X = x, Y = y)$ of these two random variables is:

$X \diagdown Y$	2	3	4	5	6	7	8	9	10	11	12	$P(X = x)$
1	1/36	0	0	0	0	0	0	0	0	0	0	1/36
2	0	2/36	1/36	0	0	0	0	0	0	0	0	3/36
3	0	0	2/36	2/36	1/36	0	0	0	0	0	0	5/36
4	0	0	0	2/36	2/36	2/36	1/36	0	0	0	0	7/36
5	0	0	0	0	2/36	2/36	2/36	2/36	1/36	0	0	9/36
6	0	0	0	0	0	2/36	2/36	2/36	2/36	2/36	1/36	11/36
$P(Y = y)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	1

DEFINITION 1.1 The **conditional probability mass function** of $Y|X = x$ is defined by

$$p(y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)},$$

where $P(X = x, Y = y)$ is the joint distribution of X and Y and $P(X = x)$ is the marginal distribution of X. Note: $P(X = x) > 0$ for all x.

Example 1 continued

DEFINITION 1.2 The conditional expectation of $Y|X = x$ is defined by

$$E[Y|X = x] = \sum_y y * p(y|X = x).$$

Note: We can also define conditional variance Var$[Y|X = x]$ analogously.

Note:

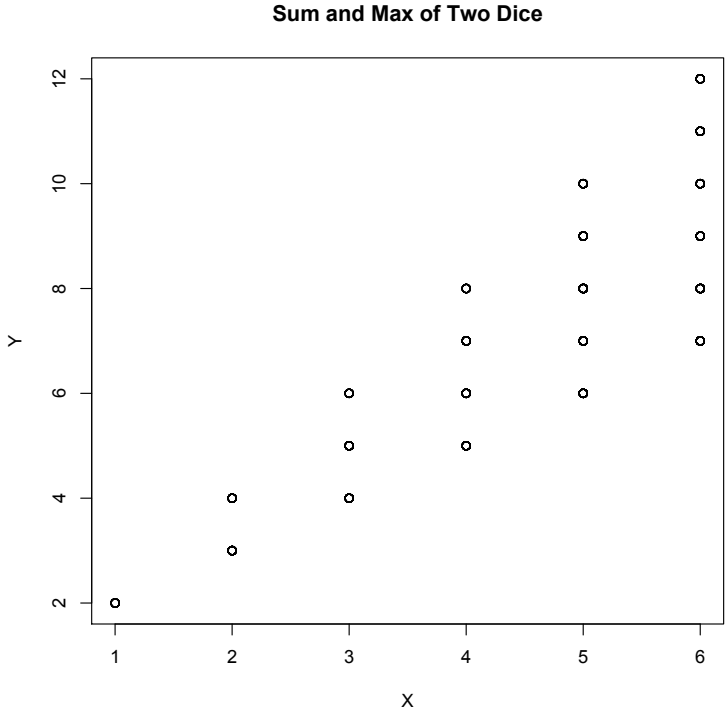
Note:

Example 1 continued

Find the conditional expectation of $Y|X = x$. Note: There will be six values corresponding to $X = 1, 2, 3, 4, 5, 6$.

$$E[Y|X = 1] = 2 * \frac{1/36}{1/36} + 3 * \frac{0}{1/36} + 4 * \frac{0}{1/36} + \cdots + 12 * \frac{0}{1/36} = 2$$
$$E[Y|X = 2] = 2 * \frac{0}{3/36} + 3 * \frac{2/36}{3/36} + 4 * \frac{1/36}{3/36} + \cdots + 12 * \frac{0}{3/36} = \frac{10}{3}$$
$$E[Y|X = 3] = 2 * \frac{0}{5/36} + 3 * \frac{0}{5/36} + 4 * \frac{2/36}{5/36} + \cdots + 12 * \frac{0}{5/36} = \frac{24}{5}$$

x	1	2	3	4	5	6
$E[Y\|X = x]$	2	10/3	24/5	44/7	70/9	102/11



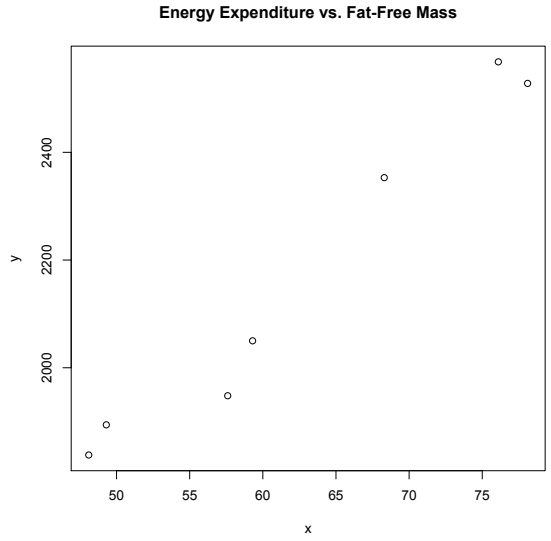
Criticize this motivating example:

Example 2

Realistic Example:

To investigate the dependence of energy expenditure on body build, researches used underwater weighing techniques to determine the fat-free body mass for each of seven men. They also measured the total 24-hour energy expenditure for each man during conditions of quiet sedentary activity. The results are shown in the table.

Subject	1	2	3	4	5	6	7
x	49.3	59.3	68.3	48.1	57.61	78.1	76.1
y	1,894	2,050	2,353	1,838	1,948	2,528	2,568



Questions:

i. How do we compute the conditional expectation $E[Y|X = x]$?

ii. What type of functional form will $E[Y|X = x]$ take on?

iii. Are the variables X and Y discrete or continuous?

iv. What probability distributions govern the behavior of X and Y?

v. Should X be thought of as fixed? (non-random)

vi. Is our model correct? How off are we?

Computing $E[Y|X = x]$

- Suppose X and Y are jointly normally distributed.
- We use the bivariate normal distribution to model this relationship.

Continuous random variables

DEFINITION 1.3 Let X and Y be two continuous random variables. The **conditional probability density function** of $Y|X = x$ is defined by

$$f(y|X = x) = \frac{f(x, y)}{f_X(x)},$$

where $f(x, y)$ is the joint density of X and Y and $f_X(x)$ is the marginal density of X. Note: $f_X(x) > 0$ for all x.

DEFINITION 1.4 Let X and Y be two continuous random variables and let $f(y|X = x)$ be the conditional density function of $Y|X = x$. The conditional expectation of $Y|X = x$ is defined by

$$E[Y|X = x] = \int y * f(y|X = x)dy.$$

Note:

1.2 Bivariate Normal Distribution

DEFINITION 1.5 The **bivariate normal distribution** of random vector (X, Y) has probability density function defined by

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_1\sigma_2} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\},$$

with

$$-\infty < x < \infty \qquad -\infty < y < \infty.$$

Note:

Note: The Bivariate Normal distribution is a special case of a multivariate normal.

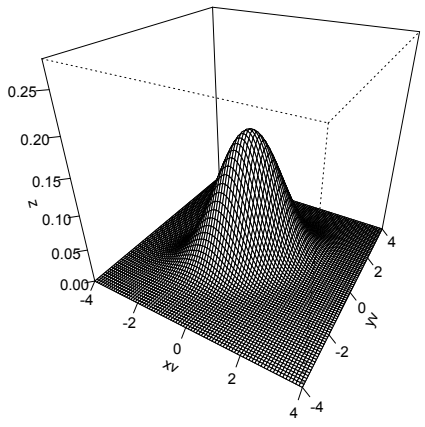
PROPOSITION 1.1 *If (X, Y) is a random vector from the bivariate normal distribution, then the conditional expectation and variance of Y given $X = x$ are*

$$E[Y|X = x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = \beta_0 + \beta_1 x \tag{1.1}$$

and

$$Var[Y|X = x] = \sigma_Y^2(1 - \rho^2). \tag{1.2}$$

Figure 1: Bivariate Normal with parameters $\mu_1 = 0,\ \mu_2 = 0,\ \sigma_1^2 = 1,\ \sigma_2^2 = 1,\ \rho = .25$



Summary:

1. $E[Y|X = x]$ is a linear function (assuming X, Y are bivariate normal).

2. The goal is to estimate the parameters β_0 and β_1.

3.

2 Simple Linear Regression

2.1 The Simple Linear Regression Model

Notation

Y= dependent variable, response variable

x= independent variable, covariate, predictor

n paired observations $(x_1, Y_1), (x_2, Y_2), \ldots (x_n, Y_n)$.

Model statement

Parameters: β_0, β_1, and σ^2

Model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n, \tag{2.1}$$

with

$$\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Note:

Famous picture

2.2 Why Isn't the Independent Variable Random?

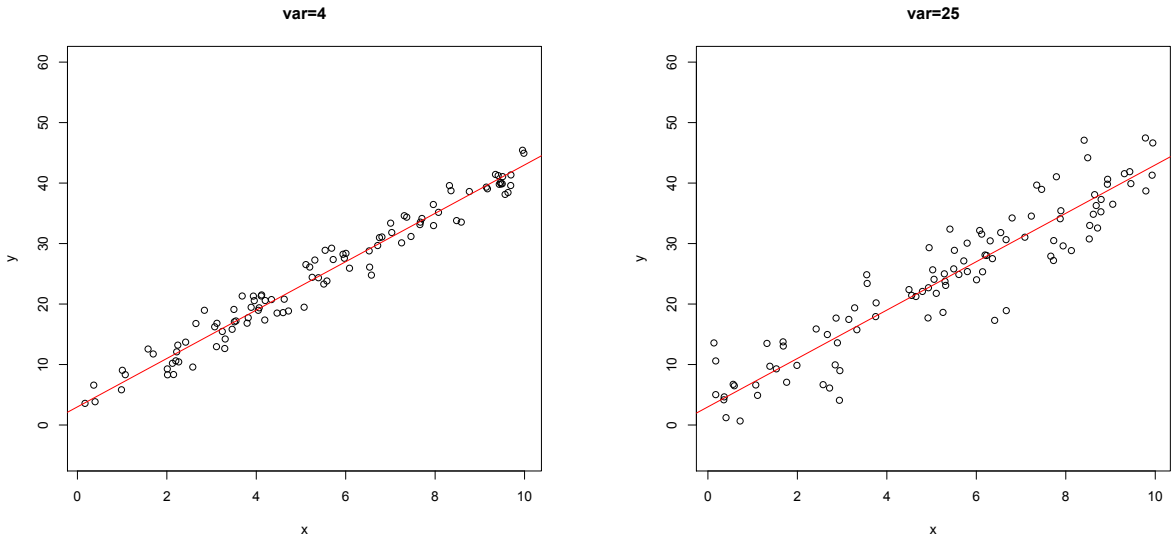
- To investigate "*Why Isn't the Independent Variable Random?"*, consider the following exercise.
- **Setup:**
  - Suppose that the independent variable is a random variable, i.e. $X \sim$ Distribution with pdf $f_X(x)$.
  - Assume a normal distribution on the errors, i.e., $\epsilon \sim N(0, \sigma^2)$.
  - Define the response as the random variable $Y = \beta_0 + \beta_1 X + \epsilon$.
  - Assume X and ϵ are independent (**Key assumption**).

Conclusion:

- Define the linear model by
$$Y = \beta_0 + \beta_1 X + \epsilon,$$
where X is a random variable that is independent of $\epsilon \sim N(0, \sigma^2)$. Then
$$E[Y|X = x] = \beta_0 + \beta_1 x.$$

- Equivalently, if X and ϵ are independent in the famous simple linear regression model, then the conditional expectation $E[Y|X = x]$ is a linear function of x.

- When do we have to worry about dependence between X and ϵ?



Example 3

Suppose the true model relating x and Y is: $Y = 3 + 4x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$. Simulate two data sets using this model. Assume $\sigma^2 = 4$ and $\sigma^2 = 25$. Below is the R code for the first case $\sigma^2 = 4$.

```
x <- sample(1:2000/200,100,replace=TRUE)
error <- rnorm(100,mean=0,sd=2)
y <- 3+4*x+error
plot(x,y,ylim=c(-5,60))
abline(a=3,b=4,col=2)
```

2.3 Least squares method

Notation for a data set consisting of n ordered pairs

Observation Number	Response Variable Y	Predictor X
1	y_1	x_1
2	y_2	x_2
$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n

Deviations

- $y_i - \bar{y} =$ the **deviation** of each case y_i from the sample mean of the response variable $\bar{y}$.
- $x_i - \bar{x} =$ the **deviation** of each case x_i from the sample mean of the predictor variable $\bar{x}$.
- $(x_i - \bar{x})(y_i - \bar{y}) =$ the product of the **deviations**.

Sums of squares

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

Other notation

Denote *some* line by: $\hat{y} = b_0 + b_1 x$

Let the line at a point (x_i, y_i) be denoted by: $\hat{y}_i = b_0 + b_1 x_i$

Define the residual of the i^{th} observation by: $e_i = y_i - \hat{y}_i$.

Objective function

$$Q(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$$

Derivation of the line of best fit

PROPOSITION 2.1 *Let*

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

Then Q is minimized when

$$\hat{\beta}_1 = b_1 = \frac{S_{xy}}{S_{xx}},$$

and

$$\hat{\beta}_0 = b_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

The derivation follows:

$$\begin{aligned}
\frac{\partial Q}{\partial b_0} &= \frac{\partial}{\partial b_0} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n} \frac{\partial}{\partial b_0} (y_i - b_0 - b_1 x_i)^2 \\
&= 2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)(-1) = 2\Big[- \sum_{i=1}^{n} y_i + b_0 n + b_1 \sum_{i=1}^{n} x_i \Big] \\
\frac{\partial Q}{\partial b_1} &= \frac{\partial}{\partial b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n} \frac{\partial}{\partial b_1} (y_i - b_0 - b_1 x_i)^2 \\
&= 2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)(-x_i) = 2\Big[- \sum_{i=1}^{n} y_i x_i + b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 \Big]
\end{aligned}$$

Setting the partial derivatives equal to zero yields the normal equations:

$$n b_0 + \Big(\sum_{i=1}^{n} x_i \Big) b_1 = \sum_{i=1}^{n} y_i \tag{2.2}$$

$$\Big(\sum_{i=1}^{n} x_i \Big) b_0 + \Big(\sum_{i=1}^{n} x_i^2 \Big) b_1 = \sum_{i=1}^{n} y_i x_i \tag{2.3}$$

Solving for b_0 in equation (2.2) and substituting the expression into (2.3) gives

$$\Big(\sum_{i=1}^{n} x_i \Big) \Big(\frac{1}{n} \sum_{i=1}^{n} y_i - \frac{1}{n} \Big(\sum_{i=1}^{n} x_i \Big) b_1 \Big) + \Big(\sum_{i=1}^{n} x_i^2 \Big) b_1 = \sum_{i=1}^{n} y_i x_i.$$

Then solving for b_1 we get

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\Big(\sum_{i=1}^n x_i\Big)\Big(\sum_{i=1}^n y_i\Big)}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^n x_i\Big)^2},$$

and (2.2) gives,

$$\hat{\beta}_0 = b_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Q must be a minimum at the point $(\hat{\beta}_0, \hat{\beta}_1)$ since $Q(b_0, b_1) \geq 0$ for all real b_0 and b_1. □

Proposition 2.1 implies that the line of best fit is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Note: The valid prediction range (VPR) is any x contained in the interval $[min\{x_i\}, max\{x_i\}]$. **Extrapolation** occurs when a response value is predicted using an x value that is outside the VPR.

Recall Example 2

To investigate the dependence of energy expenditure on body build, researches used underwater weighing techniques to determine the fat-free body mass for each of seven men. They also measured the total 24-hour energy expenditure for each man during conditions of quiet sedentary activity. The results are shown in the table.

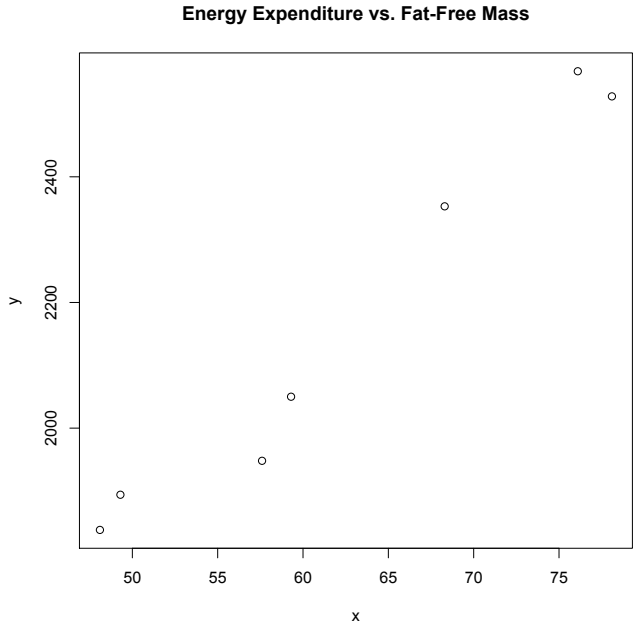
Subject	1	2	3	4	5	6	7
x	49.3	59.3	68.3	48.1	57.61	78.1	76.1
y	1,894	2,050	2,353	1,838	1,948	2,528	2,568

The `lm` function is the most widely used function in STAT2103. The `R` code follows:

```
x <- c(49.3,59.3,68.3,48.1,57.6,78.1,76.1)
y <- c(1894,2050,2353,1838,1948,2528,2568)
model <- lm(y~x)
model
```

The R code for the scatter plot with the line of best fit follows:

```
plot(x,y,main="Energy Expenditure vs.  Fat-Free Mass")
abline(model,col=2)
```



PROPOSITION 2.2 *The line of best fit crosses the point* $(\bar{x}, \bar{y})$

PROOF:

□

Interpretation of the slope β_1 (or $\hat{\beta}_1$)

"For every 1 unit increase in x, the average of the response variable increases (or decreases) by $\hat{\beta}_1$ units."

Example 2 continued

Estimating σ^2

DEFINITION 2.1 The **mean square error** denoted MSE is defined by

$$MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

Example 2 continued

```
n <- length(x)
y_hat <- fitted(model)
SSE <- sum((y-y_hat)*(y-y_hat))
SSE/(n-2)
```

Coefficient of determination

The **coefficient of determination** denoted r^2 is the proportion of variation in the response variable y explained by the model (or explained by covariate x). The computational formula is given by

$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{SSE}{S_{yy}}$$

Note:

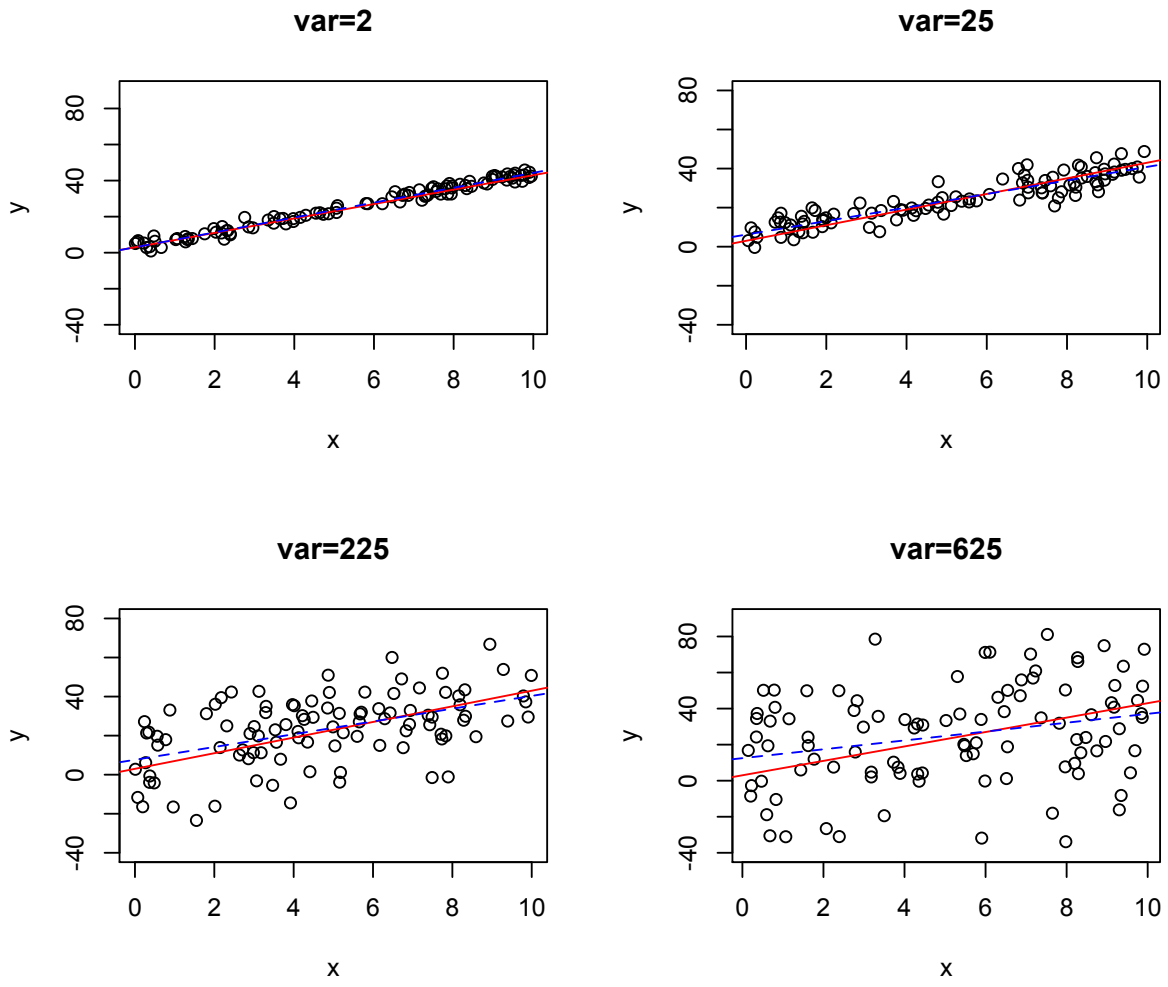
Example 2 continued

```
y_hat <- fitted(model)
SSE <- sum((y-y_hat)*(y-y_hat))
SYY <- sum((y-mean(y))*(y-mean(y)))
1-SSE/SYY
```

Example 2 continued

Suppose the true model relating x and Y is: $Y = 3+4x+\epsilon, \quad \epsilon \sim N(0, \sigma^2)$. Simulate four data sets using this model. Assume $\sigma^2 = 4, \sigma^2 = 25, \sigma^2 = 225$ and $\sigma^2 = 625$. For each simulated data set, find the coefficient of determination. For *simple linear regression*, one R method for finding the coefficient of determination is:

```
cor(x,y)^2
```



Properties of the slope and variance estimators

PROPOSITION 2.3

i. $\hat{\beta}_1$ *is an unbiased estimator of* β_1. $E[\hat{\beta}_1] = \beta_1$

ii. $\hat{\beta}_0$ *is an unbiased estimator of* β_0. $E[\hat{\beta}_0] = \beta_0$

iii. MSE *is an unbiased estimator of* σ^2. $E[MSE] = \sigma^2$.

Note:

PROOF:

DEFINITION 2.2 The **residual** denoted e_i is the difference between the observed value y_i and its corresponding fitted value $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i$$

The fitted value is given by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimators of β_0 and β_1.

Distinction between e_i and ϵ_i

Properties of fitted regression line

1. $$\sum_{i=1}^{n} e_i = 0$$
2. $$\sum_{i=1} y_i = \sum_{i=1} \hat{y}_i$$
3. $$\sum_{i=1}^{n} x_i e_i = 0$$
4. $$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$
5. The line of best fit always crosses the point $(\bar{x}, \bar{y})$.

PROOF:

2.4 Probability Distributions of Estimators and Residuals

Linearity of the normal distribution

THEOREM 2.1 *Let $Y_1, Y_2, \ldots, Y_n$ be an indexed set of independent normal random variables. Then for real numbers $a_1, a_2, \ldots, a_n$, the random variable $W = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$ is normally distributed with mean*

$$E[W] = E[a_1Y_1 + a_2Y_2 + \cdots + a_nY_n] = a_1E[Y_1] + a_2E[Y_2] + \cdots + a_nE[Y_n],$$

and variance

$$\text{Var}[W] = \text{Var}[a_1Y_1 + a_2Y_2 + \cdots + a_nY_n] = a_1^2\text{Var}[Y_1] + a_2^2\text{Var}[Y_2] + \cdots + a_n^2\text{Var}[Y_n].$$

Expressing least squares estimators as a linear combination of the response values Y_i

THEOREM 2.2 *Under the conditions of regression model* (2.1)*, least squares estimator* $\hat{\beta}_1$ *is normally distributed with mean* β_1 *and variance* σ^2/S_{xx}.

THEOREM 2.3 *Under the conditions of regression model* (2.1)*, the least squares estimators* $\hat{\beta}_0$ *and* $\hat{\beta}_1$ *are unbiased and have minimum variance among all other unbiased linear estimators.*

Expressing fitted values $\hat{Y}_i$ as a linear combination of the response values Y_i

Properties of hat values h_{ij}

THEOREM 2.4 *Define the* **hat value** h_{ij} *by*

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}},$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Then

1. $h_{ij} = h_{ji}$
2. $\sum_{j=1}^{n} h_{ij} = 1$
3. $\sum_{j=1}^{n} h_{ij} x_j = x_i$
4. $\sum_{j=1}^{n} {h_{ij}}^2 = h_{ii}$
5. $\sum_{i=1}^{n} h_{ii} = 2$

PROOF: **Homework**

THEOREM 2.5 *Under the conditions of regression model* (2.1)*, the fitted response value* $\hat{Y}_i$ *is normally distributed with respective mean and variance*

$$E[\hat{Y}_i] = \beta_0 + \beta_1 x_i$$

and

$$\text{Var}[\hat{Y}_i] = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right).$$

THEOREM 2.6 *Under the conditions of regression model* (2.1)*, the residuals* e_i *are normally distributed with respective mean and variance*

$$E[e_i] = 0$$

and

$$\text{Var}[e_i] = \sigma^2(1 - h_{ii}),$$

where

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

Distribution of MSE

THEOREM 2.7 *Under the conditions of regression model* (2.1)*, the random variable*

$$W = \frac{(n-2)MSE}{\sigma^2}$$

is distributed χ^2 *with degrees of freedom* $n-2$*.*

Relationship between the slope and intercept

THEOREM 2.8 *Let $\hat{\beta}_1$ and $\hat{\beta}_0$ be the least squares estimators of β_1 and β_0. Then*

$$\mathit{Cov}(\hat{\beta}_1, \hat{\beta}_0) = -\bar{x}\mathrm{Var}[\hat{\beta}_1].$$

PROOF:

2.5 Maximum Likelihood Estimation of the Simple Linear Regression Model

Consider a random sample $X_1, X_2, \ldots, X_n$ each having common probability density function (or probability mass function) $f(x_i|\theta)$ where θ is a generic parameter of that distribution. θ could also be a vector of parameters. The joint probability density function (or joint probability mass function) is

$$f(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta).$$

Define the likelihood function as

$$\mathcal{L}(\theta; x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta). \tag{2.4}$$

It is often convenient working with the log-likelihood function

$$\log(\mathcal{L}(\theta; x_1, x_2, \ldots, x_n)) = \log(f(x_1, x_2, \ldots, x_n|\theta)).$$

Note:

DEFINITION 2.3 The **maximum likelihood estimate** $\hat{\theta}$ is the value of θ that maximizes the likelihood function (2.4), so that

$$\mathcal{L}(\hat{\theta}; x_1, x_2, \ldots, x_n) \geq \mathcal{L}(\theta; x_1, x_2, \ldots, x_n).$$

Example 4

Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution each having common probability density function $f(x_i|\mu) = \frac{1}{\mu} \exp\big(-\frac{1}{\mu}x_i\big),\ x_i \geq 0$. Find the maximum likelihood estimator of μ.

Maximim likelihood estimators of β_0, β_1, and σ^2

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample satisfying the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Notice from the above model: $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

The probability density function of a single Y_i is

$$f(y_i|\beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_i))^2 \right\}.$$

Consequently, the maximum likelihood function is:

$$\begin{aligned}
\mathcal{L}(\beta_0, \beta_1, \sigma^2; y_1, y_2, \ldots, y_n) &= f(y_1|\beta_0, \beta_1, \sigma^2) \times f(y_2|\beta_0, \beta_1, \sigma^2) \times \cdots \times f(y_n|\beta_0, \beta_1, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_i))^2 \right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \right).
\end{aligned}$$

The log-likelihood function is:

$$\log(\mathcal{L}(\beta_0, \beta_1, \sigma^2; y_1, y_2, \ldots, y_n)) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

Taking partial derivatives and setting the expressions equal to zero gives

$$\frac{\partial}{\partial \beta_0} \log(\mathcal{L}(\beta_0, \beta_1, \sigma^2; y_1, y_2, \ldots, y_n)) = -\frac{1}{2\sigma^2} \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \overset{set}{=} 0, \tag{2.5}$$

$$\frac{\partial}{\partial \beta_1} \log(\mathcal{L}(\beta_0, \beta_1, \sigma^2; y_1, y_2, \ldots, y_n)) = -\frac{1}{2\sigma^2} \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \overset{set}{=} 0, \tag{2.6}$$

and

$$\frac{\partial}{\partial \sigma^2} \log(\mathcal{L}(\beta_0, \beta_1, \sigma^2; y_1, y_2, \ldots, y_n)) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \overset{set}{=} 0. \tag{2.7}$$

Notice, to solve for β_1 and β_0 using expressions (2.5) and (2.6), this is **exactly** the same optimization problem as deriving least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_0$. To see this solution, look at the derivation following Proposition 2.1.

By substituting maximum likelihood estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ into Equation (2.7) and solving for σ^2 yields the maximum likelihood estimator of the population variance. That is,

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = 0$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

Thus, maximizing $\log(\mathcal{L})$ with respect to β_0, β_1 and σ^2 yields maximum likelihood estimators

$$\hat{\beta}_{1,MLE} = \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{2.8}$$

$$\hat{\beta}_{0,MLE} = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.9}$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{n-2}{n}MSE \tag{2.10}$$

Note:

2.6 Inferences About the Slope Parameter

To assess the statistical relationship between response variable Y and covariate x, we want to test the slope parameter. Consider the null hypothesis:

$$H_0 : \beta_1 = \beta_{10}$$

The most common hypothesized value is zero.

$$H_0 : \beta_1 = 0$$

To construct a reasonable test statistic for H_0, we will follow the usual procedure. We want to standardize the slope estimator $\hat{\beta}_1$. I.e.,

$$\text{STAT} = \frac{\hat{\beta}_1 - E[\hat{\beta}_1]}{\sigma_{\hat{\beta}_1}}$$

Before arriving at the test statistic, notice that the estimator $\hat{\beta}_1$ can be expressed as a linear combination of the Y_i's. I.e,

$$\hat{\beta}_1 = \sum_{i=1}^{n} k_i Y_i,$$

where

$$k_i = \frac{x_i - \bar{x}}{S_{xx}}.$$

i. From Proposition 2.3

ii. The variance of estimator $\hat{\beta}_1$. is

iii. The standard error of estimator $\hat{\beta}_1$. is

iv. If we standardize $\hat{\beta}_1$, we get

Note: $Z \sim N(0, 1)$.

v. Consider testing the null hypothesis $H_0 : \beta_1 = \beta_{10}$. Then under the null, the test statistic is given by

$$z_{calc} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\sigma^2}{S_{xx}}}}.$$

vi. In practice we must estimate σ^2. The natural estimator of σ^2 is the mean square error MSE. The estimated standard error is

vii. If we studentize $\hat{\beta}_1$, we get

The appropriate proposition (reference GR5204) implies T has a student's t-distribution with $n - 2$ degrees of freedom.

Under the null hypothesis

$$H_0 : \beta_1 = \beta_{10},$$

the test statistic is given by

$$t_{calc} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSE}{S_{xx}}}}$$

Rejection regions

Alternative Hypothesis	Rejection region for a level α test
$H_A : \beta_1 > \beta_{10}$	$t_{calc} \geq t_{\alpha,n-2}$ (upper-tailed)
$H_A : \beta_1 < \beta_{10}$	$t_{calc} \leq -t_{\alpha,n-2}$ (lower-tailed)
$H_A : \beta_1 \neq \beta_{10}$	$\lvert t_{calc} \rvert \geq t_{\alpha/2,n-2}$ (two-tailed)

The p-values are computed the same manner as any t-hypothesis testing procedure using degrees of freedom $n - 2$.

$100(1 - \alpha)\%$ confidence interval for β_1

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{\frac{MSE}{S_{xx}}}$$

Example 2 continued

R code:

```
model <- lm(y~x)
summary(model)
```

Table 1: R simple linear regression output

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	607.7034	138.7646	4.38	0.0072
x	25.0116	2.1888	11.43	0.0001

Caution

- A statistically significant slope does not always imply a strong correlation.

2.7 Analysis of Variance Approach to Regression Analysis

Partitioning SST

$$Y_i - \bar{Y} \quad = \quad \hat{Y}_i - \bar{Y} \quad + \quad Y_i - \hat{Y}_i$$

$$n - 1 \quad = \quad 1 \quad + \quad n - 2$$

Define the **sums of squares regression** by

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2.$$

Notice that

PROPOSITION 2.4 *Total variation SST can be partitioned into two sources of variation SSR and SSE. This can be represented with the additive identity*

$$SST = SSR + SSE$$

PROOF:

□

Equivalent testing procedure for null hypothesis $H_0 : \beta_1 = 0$

Consider testing

$$H_0 : \beta_1 = 0 \tag{2.11}$$

The F-statistic for testing the above hypothesis is

$$f_{calc} = \frac{SSR/1}{SSE/(n-2)} = \frac{SSR}{MSE} \tag{2.12}$$

Alternative Hypothesis	Rejection region for a level α test
$H_A : \beta_1 \neq 0$	$f_{calc} \geq f_{\alpha,1,n-2}$ (two-tailed)

ANOVA table

Source	df	Sum of Squares	Mean Square	F-value
Regression	1	SSR	SSR	f_{calc}
Error	$n-2$	SSE	MSE	
Total	$n-1$	SST		

Expected mean squares

Next we *informally* motivate how the F-statistic tests the hypothesis stated in (2.11). First consider Proposition 2.5

PROPOSITION 2.5 *The expected value of SSR is*

$$E[SSR] = \sigma^2 + \beta_1^2 S_{xx}$$

PROOF:

□

Motivation of the F–statistic

The F-distribution

Parameters	$\nu_1 > 0, \nu_2 > 0$ (degrees of freedom)
Notation	$F \sim F(\nu_1, \nu_2)$
pdf	$f_F(x) = \frac{1}{xB(\nu_1/2,\nu_2/2)}\sqrt{\frac{(\nu_1 x)^{\nu_1}\nu_2^{\nu_2}}{(\nu_1 x+\nu_2)^{\nu_1+\nu_2}}} \quad x \geq 0$

Note: $B(a, b)$ is the beta function defined by $B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$.

PROPOSITION 2.6 *The mean and variance of the F-distribution are respectively* $E[F] = \frac{\nu_2}{\nu_2-2}$ *for* $\nu_2 > 2$ *and* $\text{Var}[F] = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ *for* $\nu_2 > 4$.

PROPOSITION 2.7 *If X_1 and X_2 are independently distributed chi-squared random variables with respective degrees of freedom ν_1 and ν_2, then the random variable*

$$F = \frac{X_1/\nu_1}{X_2/\nu_2}$$

has an F-distribution with degrees of freedom ν_1 and ν_2.

PROPOSITION 2.8 *Let $X_1, X_2, \ldots, X_m$ be a random sample from a normal distribution with variance σ_1^2, let $Y_1, Y_2, \ldots, Y_n$ be a random sample from another normal distribution with variance σ_2^2, and let S_1^2 and S_2^2 denote the two sample variances. Also assume the random samples are independent of each other. Then the random variable*

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \tag{2.13}$$

has an F-distribution with degrees of freedom $\nu_1 = m - 1$ and $\nu_2 = n - 1$.

PROPOSITION 2.9 *Let T be distributed student's t-distribution with degrees of freedom ν. Then the random variable T^2 has an F-distribution with degrees of freedom 1 and ν. Namely*

$$T^2 \sim F(1, \nu)$$

Example 2 continued

`R` code:

```
model <- lm(y~x)
anova(model)
```

Table 2: Simple linear regression analysis of variance `R` output

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	549097.62	549097.62	130.58	0.0001
Residuals	5	21026.10	4205.22		

2.8 General Linear Test

The analysis of variance test of $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ is an example of the **general test for a linear statistical model**.

General linear test

Let SSE_R be the sum of squares error for the **reduced model** with degrees of freedom df_R and let SSE_F be the sum of squares error for the **full model** with degrees of freedom df_F. Then the **general F-test** uses the following statistic:

$$f_{calc} = \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_F}. \tag{2.14}$$

Note: The F-statistic can be derived through a likelihood ratio test.

Likelihood ratio test

DEFINITION 2.4 Consider a realized dataset $y_1, y_2, \ldots, y_n$. The **likelihood ratio test statistic** for testing $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_0^C$ is

$$\lambda(y_1, y_2, \ldots, y_n) = \frac{\max\limits_{\Theta_0} \mathcal{L}(\theta; y_1, y_2, \ldots, y_n)}{\max\limits_{\Theta} \mathcal{L}(\theta; y_1, y_2, \ldots, y_n)}.$$

Notes:

General F-test statistic

$$f_{calc} = \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_F}$$

Same statistic defined in Equation (2.14).

Rejection rule

	Rejection region for a level α test
Reject H_0 if	$f_{calc} \geq f_{\alpha, df_R - df_F, df_F}$

General F-test for simple linear regression

For simple linear regression, the **full model** is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2). \tag{2.15}$$

Under the null hypothesis $H_0 : \beta_1 = 0$, the **reduced model** is

$$Y_i = \beta_0 + \epsilon_i, \qquad i = 1, 2, \ldots, n, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2). \tag{2.16}$$

The sum of squared errors for the **full model** is:

$$SSE_F = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2,$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimators of β_1 and β_2.

The sum of squared errors for the reduced model is

$$SSE_R = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Notice under the **reduced model** (2.16), β_0 is estimated with $\bar{y}$.

The degrees of freedom for the full and reduced models are respectively

$$df_F = n - 2 \qquad \text{and} \qquad df_R = n - 1.$$

The general linear test gives F-statistic

Notice, f_{calc} above is the same F-statistic as Equation (2.12).

Example 2 continued

`R` code for the full model:

```
model.full <- lm(y~x)
anova(model.full)
```

Table 3: Linear regression analysis of variance `R` output for the **full model**.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	549097.62	549097.62	130.58	0.0001
Residuals	5	21026.10	4205.22		

R code for the reduced model:

```
model.reduced <- lm(y~1)
anova(model.reduced)
```

Table 4: Linear regression analysis of variance R output for the **reduced model**.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	6	570123.71	95020.62		

R code for the general linear F–test:

```
anova(model.reduced,model.full)
```

Table 5: Analysis of variance R output for the general linear F–test.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	6	570123.71				
2	5	21026.10	1	549097.62	130.58	0.0001

Example 5

Consider testing whether the intercept β_0 statistically differs from zero.

Table 6: R simple linear regression output

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	607.7034	138.7646	4.38	0.0072
x	25.0116	2.1888	11.43	0.0001

2.9 Binary Predictor

Consider splitting the response values $y_1, \ldots, y_n$ into two groups with respective sample sizes n_1 and n_2. Define the **dummy** variable

$$x_i = \begin{cases} 1 & \text{if group one} \\ 0 & \text{if group two} \end{cases} \tag{2.17}$$

What will the estimated linear regression model be?

THEOREM 2.9 *Consider the simple linear regression model* (2.1) *using independent variable defined by* (2.17)*. Then the least squares estimators are*

$$\hat{\beta}_1 = \bar{y}_1 - \bar{y}_2 \quad \textit{and} \quad \hat{\beta}_0 = \bar{y}_2,$$

where $\bar{y}_1$ and $\bar{y}_2$ are the respective sample means of each group.

PROOF: Homework

What will the test statistic look like when testing β_1?

Recall the two sample T-test:

When testing the null hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$, the test statistic is

$$t_{calc} = \frac{\bar{y}_1 - \bar{y}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

where $\bar{y}_1, s_1^2$ and n_1 are the respective sample average, sample variance & sample size for group one and $\bar{y}_2, s_2^2$ and n_2 are the respective sample average, sample variance & sample size for group two. To compute p-values, we use the students T-distribution with degrees of freedom

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s^2/n_2)^2}{n_2 - 1}}.$$

When the population variances are assumed to be equal for the two groups ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), then the **pooled** test statistic is

$$t_{calc} = \frac{\bar{y}_1 - \bar{y}_2 - \Delta_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}},$$

where the sample pooled variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

To compute p-values, we use the students T-distribution with degrees of freedom

$$df = n_1 + n_2 - 2.$$

THEOREM 2.10

$$t_{calc} = \frac{\bar{y}_1 - \bar{y}_2 - \Delta_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSE}{S_{xx}}}}$$

Note: $MSE = s_p^2$, $\hat{\beta}_1 = \bar{y}_1 - \bar{y}_2$, $\Delta_0 = \beta_{10}$ (the hypothesized value frequently zero).

Example 6

To investigate the maternal behavior of laboratory rats, we move the rat pup a fixed distance from the mother and record the time (in seconds) required for the mother to retrieve the pup to the nest. We run the study with 5- and 20-day old pups. Note: These are not the same pups being measured twice.

5 Days	20 Days
15	30
10	15
25	20
15	25
20	23
18	20

Use regression techniques with a dummy variable to test if the retrieval time differs per group.

Response (Y)	Covariate (x)
15	1
10	1
25	1
15	1
20	1
18	1
30	0
15	0
20	0
25	0
23	0
20	0

The R code follows:

```
# two groups
y1 <- c(15,10,25,15,20,18)
y2 <- c(30,15,20,25,23,20)
# response
y <- c(y1,y2)
# dummy variable
x <- c(rep(1,6),rep(0,6))
# multiple boxplot
boxplot(y x)
# test
summary(lm(y x))
t.test(y1,y2,var.equal = TRUE)
```

The R output follows:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   22.167      2.088  10.615 9.18e-07 ***
x             -5.000      2.953  -1.693    0.121
---

Residual standard error: 5.115 on 10 degrees of freedom
Multiple R-squared:  0.2228,Adjusted R-squared:  0.145
F-statistic: 2.866 on 1 and 10 DF,  p-value: 0.1213
```

The R output follows:

```
Two Sample t-test

data:  y1 and y2
t = -1.693, df = 10, p-value = 0.1213
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.580454   1.580454
sample estimates:
mean of x mean of y
 17.16667  22.16667
```

2.10 Prediction

In simple linear regression, there are two fundamental goals:

1. Test if there is a relationship between the response variable Y and covariate x.
   The first goal is accomplished by testing null hypothesis $H_0 : \beta_1 = \beta_{10}$.

2. Predict the response Y given a fixed value of x.
   This section describes predictions and confidence intervals on predictions.

Inferences concerning $E[Y_h]$

The parameter of interest is: $\theta = E[Y_h]$

PROPOSITION 2.10 *Let*

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$$

where x_h is some fixed value of x. Then

$$E[\hat{Y}_h] = \beta_0 + \beta_1 x_h$$

and

$$\text{Var}[\hat{Y}_h] = \sigma^2 \left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}} \right]$$

From the above proposition, the standardized score of $\hat{Y}_h$ is

$$Z = \frac{\hat{\beta}_0 + \hat{\beta}_1 x_h - (\beta_0 + \beta_1 x_h)}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}} \right)}}.$$

Since $\hat{Y}_h$ is a linear combination or response variable Y_i, the random variable Z has a standard normal distribution. The studentized score of $\hat{Y}_h$ is

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x_h - (\beta_0 + \beta_1 x_h)}{\sqrt{MSE \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}} \right)}}.$$

The *appropriate* proposition (reference GR5204) implies T has a student's t-distribution with $n - 2$ degrees of freedom. Consequently, the confidence interval of interest follows.

Confidence interval for $E[Y_h]$

The $100(1-\alpha)\%$ confidence interval for $E[Y_h]$ when $x = x_h$ is

$$(\hat{\beta}_0 + \hat{\beta}_1 x_h) \pm t_{\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right)}$$

Inferences concerning the prediction of future Y values ($Y_{h(new)}$)

Goal: For fixed $x = x_h$, we want to find a C.I. for a *single future value* $Y_{h(new)}$ as compared to a C.I for the *true average* $E[Y_h]$.

The **prediction error** for a single future response value is

$$Y_{h(new)} - (\hat{\beta}_0 + \hat{\beta}_1 x_h) \tag{2.18}$$

PROPOSITION 2.11 *The expected value and variance of the prediction error defined Equation* (2.18) *are respectively given by*

$$E[Y_{h(new)} - (\hat{\beta}_0 + \hat{\beta}_1 x_h)] = 0$$

and

$$\text{Var}[Y_{h(new)} - (\hat{\beta}_0 + \hat{\beta}_1 x_h)] = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right]$$

In a similar manner as the confidence interval for $E[Y_h]$, the studentized score of prediction error (2.18) is given by

$$T = \frac{Y_{h(new)} - (\hat{\beta}_0 + \hat{\beta}_1 x_h) - 0}{\sqrt{MSE\Big(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\Big)}}.$$

The *appropriate* proposition (reference GR5204) implies T has a student's t-distribution with $n - 2$ degrees of freedom. Consequently, the interval of interest follows below.

Prediction interval for a single future value $Y_{h(new)}$

The $100(1 - \alpha)\%$ prediction interval for a single future value of $Y_{h(new)}$ when $x = x_h$ is

$$(\hat{\beta}_0 + \hat{\beta}_1 x_h) \pm t_{\alpha/2, n-2} \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right)}$$

Example 2 continued

i. Using the appropriate interval, estimate the **true average** energy expenditure for someone with a fat-free body mass of 65 [kg].

ii. Using the appropriate interval, estimate the energy expenditure for a **single respondent** having a fat-free body mass of 65 [kg].

Example 6 continued

i. Using the appropriate interval, estimate the **true average** retrieval time for both the five day and twenty day old pups.

Example 2 R code follows:

```
# data
x <- c(49.3,59.3,68.3,48.1,57.6,78.1,76.1)
y <- c(1894,2050,2353,1838,1948,2528,2568)
# model
model <- lm(y~x)
# dataframe
x_data <- data.frame(x=65)
# confidence interval
predict(model,newdata=x_data,interval="confidence")
# prediction interval
predict(model,newdata=x_data,interval="prediction")
```

Example 2 R output follows:

```
       fit      lwr     upr
1 2233.459 2168.777 2298.14
```

```
       fit      lwr      upr
1 2233.459 2054.654 2412.264
```

Example 6 R code follows:

```
# two groups
y1 <- c(15,10,25,15,20,18)
y2 <- c(30,15,20,25,23,20)
# response
y <- c(y1,y2)
# dummy variable
x <- c(rep(1,6),rep(0,6))
# prediction
x_data <- data.frame(x=c(1,0))
predict(lm(y x),newdata=x_data,interval="confidence")
```

Example 6 R output follows:

```
       fit      lwr      upr
1 17.16667 12.51358 21.81975
2 22.16667 17.51358 26.81975
```

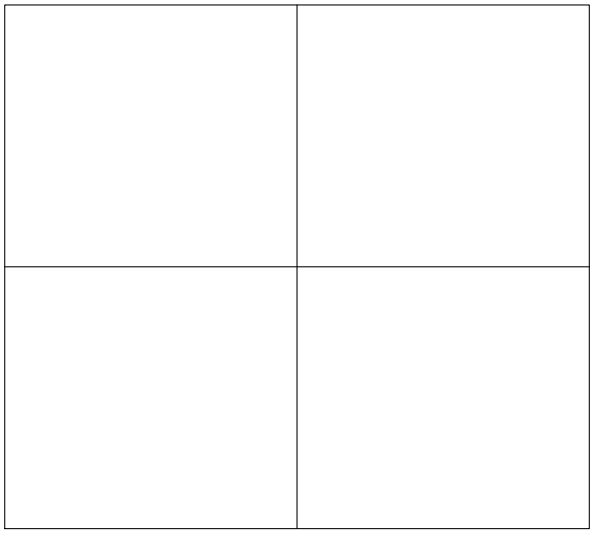
Further Inspection on confidence intervals and prediction intervals

2.11 Linear Correlation

Linear correlation and covariance are both measures of the linear relationship between variables X and Y.

Deviations

- $y_i - \bar{y} =$ the **deviation** of each case y_i from the sample mean of the response variable $\bar{y}$.
- $x_i - \bar{x} =$ the **deviation** of each case x_i from the sample mean of the predictor variable $\bar{x}$.
- $(x_i - \bar{x})(y_i - \bar{y}) =$ the product of the **deviations**.



Quadrant	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})(y_i - \bar{y})$
1			
2			
3			
4			

DEFINITION 2.5 Consider a set of n paired observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The **sample linear correlation coefficient** is

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

Note:

The linear correlation coefficient is computed using the `R` function

```
cor(x,y)
```

Example 2 continued

Subject	1	2	3	4	5	6	7
Fat-free mass (x)	49.3	59.3	68.3	48.1	57.61	78.1	76.1
Energy expenditure (y)	1,894	2,050	2,353	1,838	1,948	2,528	2,568

Properties of the sample correlation

i. for real numbers $a, c > 0$ or $a, c < 0$, define $w_1 = ax_1 + b, w_2 = ax_2 + b, ..., w_n = ax_n + b$ and $z_1 = cy_1 + d, z_2 = cy_2 + d, ..., z_n = cy_n + d$. Then $r_{wz} = r_{xy}$.

ii. $-1 \leq r_{xy} \leq 1$

iii. $r = 1$ or $r = -1$ iff $y = ax + b$ for some real numbers a, b with $a \neq 0$.

Population covariance and correlation

DEFINITION 2.6 The **covariance** of random variables X and Y is defined by

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

where μ_X and μ_Y are the respective expected values of X and Y.

Note:

DEFINITION 2.7 The **correlation** of random variables X and Y is defined by

$$\rho = Corr(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

where σ_X and σ_Y are the respective standard deviations of X and Y.

Note:

Properties of correlation

i. for $a, c > 0$ or $a, c < 0$, $Corr(aX + b, cY + d) = Corr(X, Y)$

ii. $-1 \leq Corr(X, Y) \leq 1$

iii. if X and Y are independent, then $\rho = 0$

iv. $\rho = 1$ or $\rho = -1$ iff $Y = aX + b$ for some real numbers a, b with $a \neq 0$.

2.12 Normal Correlation Model

The Joint Normal Distribution

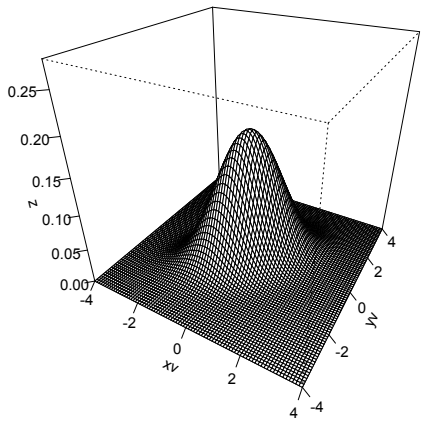
DEFINITION 2.8 The **bivariate normal distribution** of random vector (X, Y) has probability density function defined by

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\},$$

with

$$-\infty < x < \infty \qquad -\infty < y < \infty.$$

Figure 2: Bivariate Normal with parameters $\mu_1 = 0,\ \mu_2 = 0,\ \sigma_1^2 = 1,\ \sigma_2^2 = 1,\ \rho = .25$



Note:

Note: The Bivariate Normal distribution is a special case of a multivariate normal.

Normal correlation model

Recall the simple linear regression model:

Model statement

Suppose $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are random ordered pairs each coming from a bivariate normal distribution. Consequently, the conditional expectation and variance of Y given $X = x$ are

$$E[Y|X = x] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1) \tag{2.19}$$

and

$$Var[Y|X = x] = \sigma_2^2(1 - \rho^2). \tag{2.20}$$

Note:

Estimation of the normal correlation model

Notice for the simple linear regression model,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = r\frac{s_y}{s_x} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \tag{2.21}$$

where s_x and s_y are the sample standard deviations and r is the sample correlation between variables x and y.

Note:

The sample correlation coefficient as an estimator

Assuming the pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are random, the correlation coefficient as an estimator is given by

$$R = \frac{\sum_{i=1}^n X_i Y_i - \frac{1}{n}\Big(\sum_{i=1}^n X_i\Big)\Big(\sum_{i=1}^n Y_i\Big)}{\sqrt{\Big(\sum_{i=1}^n X_i^2 - \frac{1}{n}\Big(\sum_{i=1}^n X_i\Big)^2\Big)\Big(\sum_{i=1}^n Y_i^2 - \frac{1}{n}\Big(\sum_{i=1}^n Y_i\Big)^2\Big)}}.$$

PROPOSITION 2.12 *If $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are random ordered pairs each coming from a bivariate normal distribution, then the quantity*

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

has a student's t-distribution with $n-2$ degrees of freedom.

Note:

Testing the linear correlation coefficient

Consider testing the null hypothesis

$$H_0 : \rho = 0.$$

Under the null, the test statistic is

$$t_{calc} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Note:

Alternative Hypothesis	Rejection region for a level α test
$H_A : \rho > 0$	$t_{calc} \geq t_{\alpha,n-2}$ (upper-tailed)
$H_A : \rho < 0$	$t_{calc} \leq -t_{\alpha,n-2}$ (lower-tailed)
$H_A : \rho \neq 0$	$\lvert t_{calc} \rvert \geq t_{\alpha/2,n-2}$ (two-tailed)

Examples 2 continued

R code follows:

```
# data
x <- c(49.3,59.3,68.3,48.1,57.6,78.1,76.1)
y <- c(1894,2050,2353,1838,1948,2528,2568)
# corr
r <- cor(x,y)
# t-stat
n <- length(x)
t <- r*sqrt(n-2)/sqrt(1-r*r)
# P-value
2(1-pt(t,n-2))
# linear model
summary(lm(yx̃))
```

R output follows:

```
# t

11.42695

# p-value

8.98794e-05
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  607.703    138.765   4.379  0.00716 **
x             25.012      2.189  11.427 8.99e-05 ***
---
Residual standard error: 64.85 on 5 degrees of freedom
Multiple R-squared:  0.9631,Adjusted R-squared:  0.9557
F-statistic: 130.6 on 1 and 5 DF,  p-value: 8.988e-05
```

2.13 History

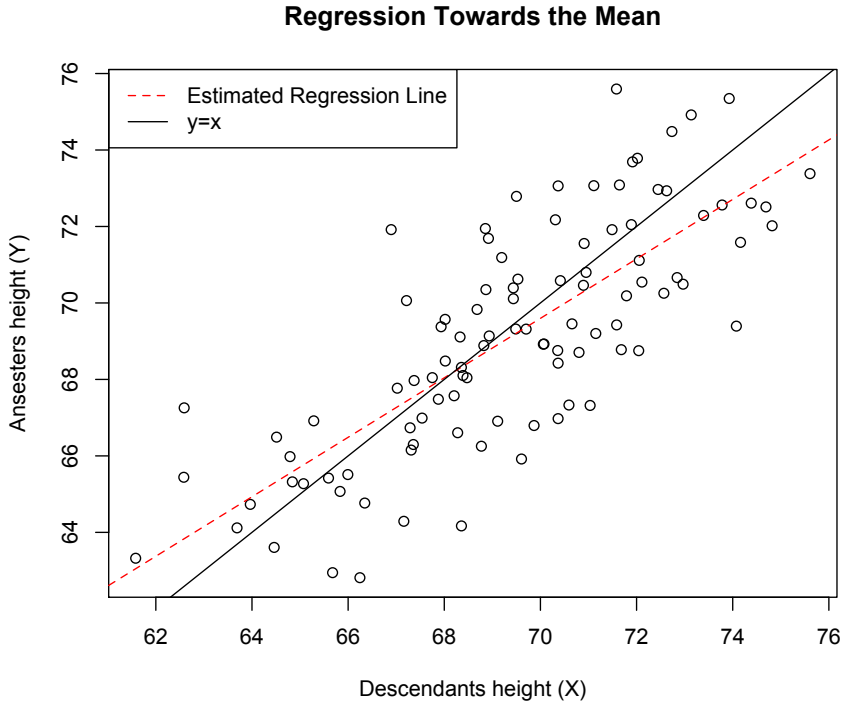
- The method of least squares was first described by Carl Friedrich Gauss around 1794.
- Gauss developed and applied least squares regression techiques to astronomical observations.
- One belief is that Gauss developed least squares to solve a chemistry problem for his friend (when Gauss was about the age of a high-school senior). Gauss however did not publish the method until 1809.
- The idea of least-squares analysis was also independently formulated by the Frenchman Adrien-Marie Legendre in 1805 and the American Robert Adrain in 1808.

What does the term “Regression" mean?

The term “regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. His student Karl Pearson is then accredited to have moved it to a more general statistical context. The biological phenomenon is that the heights of descendants of tall ancestors tend to *regress* down towards a normal average. This phenomenon is also known as regression towards the mean.

Note:

Figure 3: The following Scatter plot illustrates the notion of "regression towards the mean". The ordered pairs are sampled from a bivariate normal distribution with parameters $\mu_Y = 69.5\ [in], \sigma_Y = 3\ [in], \mu_X = 69.5\ [in], \sigma_X = 3\ [in], \rho = .80$. The two lines displayed are the line of best fit and the identity function. If height of the ancestors and decedents are perfectly correlated, then the linear relationship would be a 45 degree line.



2.14 Simultaneous Inferences

Motivation

- Consider making inference with confidence level 95% of both the true slope β_1 and the true intercept β_0.
- The difficulty is that these would not provide 95% confidence that the conclusions of *both* β_1 and β_0 are correct.
- If the inferences were independent, the probability of both being correct would be $(.95)^2 = .9025$.
- The inferences are not independent.

Familywise error rate

Recall, in any hypothesis testing procedure,

$$P(\text{Type I error}) = \alpha.$$

The **familywise error rate** is defined as

$$P(\text{At least one type I error}).$$

To compute the familywise error rate, consider running a pairwise procedure on β_1 and β_0. Then,

Further motivation

- Showing false significance in a testing procedure is a *bad thing*.
- Ideally, researchers want to control for making too many Type I errors.
- There have been many different procedures developed to control for the familywise error rate.

Construction of joint confidence intervals

We start with ordinary confidence limits for β_0 and β_1, with confidence coefficients $1-\alpha$ each. These limits are:

$$\hat{\beta}_0 \pm t_{\alpha/2,n-2}\hat{\sigma}_{\hat{\beta}_0} \quad \text{and} \quad \hat{\beta}_1 \pm t_{\alpha/2,n-2}\hat{\sigma}_{\hat{\beta}_1}$$

Let A_1 denote the event that the first confidence interval does not cover β_0, and let A_2 denote the event that the second confidence interval does not cover β_1. We know

$$P(A_1) = \alpha \quad \text{and} \quad P(A_2) = \alpha.$$

Then

By De Morgan's law

The probability that both intervals are correct is

Using the fact that $P(A_1 \cap A_2) \geq 0$, we obtain the *Bonferroni inequality:*

$$P(A_1^C \cap A_2^C) \geq 1 - P(A_1) - P(A_2),$$

which for our setting is:

$$P(A_1^C \cap A_2^C) \geq 1 - \alpha + \alpha = 1 - 2\alpha, \tag{2.22}$$

We can easily use the Bonferroni inequality (2.22) to obtain a family confidence coefficient of at least $1-\alpha$ for estimating β_0 and β_1. We do this by estimating β_0 and β_1 separately with confidence levels of $1-\alpha/2$ each. Namely,

$$1 - \alpha/2 - \alpha/2 = 1 - \alpha.$$

Note: To find the critical value in two-tailed tests (or centered confidence intervals), we divide the significance level α by 2.

Bonferroni joint confidence intervals

The $1-\alpha$ family intervals for estimating β_0 and β_1 are

$$\hat{\beta}_0 \pm t_{\alpha/4,n-2}\hat{\sigma}_{\hat{\beta}_0} \quad \text{and} \quad \hat{\beta}_1 \pm t_{\alpha/4,n-2}\hat{\sigma}_{\hat{\beta}_1}$$

The above intervals are called **Bonferroni joint confidence intervals** or a **Bonferroni procedure.**

Example 2 continued

Find the Bonferroni joint confidence intervals for estimating both the true intercept and true slope.

Table 7: R simple linear regression output

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	607.7034	138.7646	4.38	0.0072
x	25.0116	2.1888	11.43	0.0001

Extensions of the Bonferroni procedure

- The Bonferroni procedure can also be applied to prediction.
- The critical value can be generalized to

$$t_{\alpha/(2K),df},$$

where K is the number of predictions (or intervals) and df is the degrees of freedom of the linear model.

Example 6 continued

Find Bonferroni joint confidence intervals for estimating the true average retrieval time for both the 5 day old and 20 day old pups.

R code:

```
x1 <- c(15,10,25,15,20,18)
x2 <- c(30,15,20,25,23,20)
y <- c(x1,x2)
x <- c(rep(1,6),rep(0,6))
model <- lm(y~x)
x.data <- data.frame(x=c(0,1))
predict(model,newdata=x.data,interval="confidence",se.fit=TRUE)
```

R output:

```
$fit
       fit      lwr      upr
1 22.16667 17.51358 26.81975
2 17.16667 12.51358 21.81975

$se.fit
       1        2
2.088327 2.088327

$df
[1] 10

$residual.scale
[1] 5.115336
```

Easy way R code:

```
K <- 2
CL <- 1-.05/(K)
predict(model,newdata=x.data,interval="confidence",level=CL)
```

Easy way R output:

```
       fit      lwr      upr
1 22.16667 16.6665 27.66683
2 17.16667 11.6665 22.66683
```

2.15 Confidence Band of the Regression Line

It is often convenient to construct a confidence band for the true line $E[Y] = \beta_0 + \beta_1 x$. This procedure constructs a region that the true line should likely fall in.

Note:

Note:

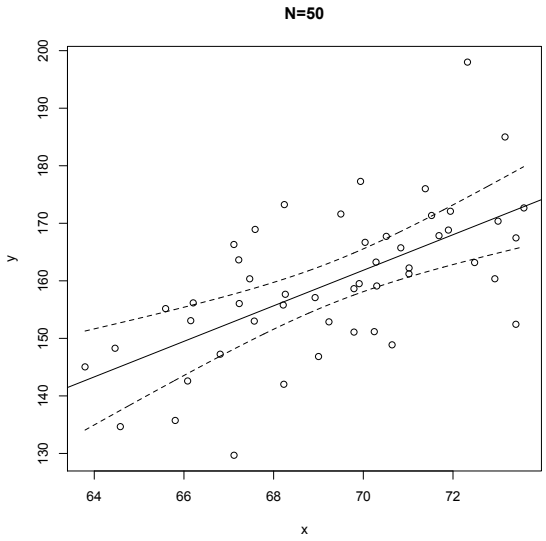
The Working-Hotelling $100(1-\alpha)\%$ confidence band for the simple linear regression model (2.1) has the following boundary values at any level x_h:

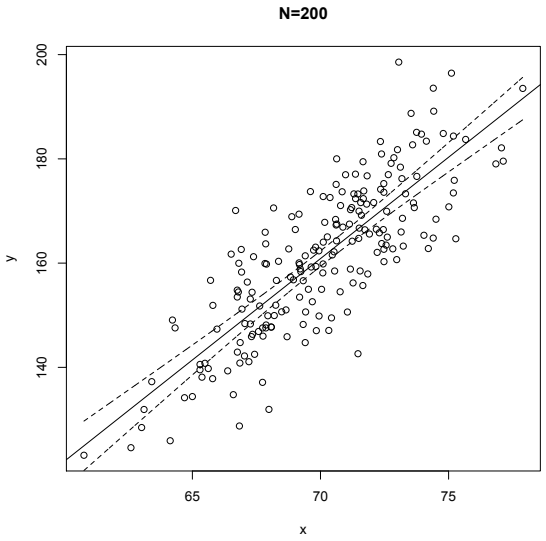
$$(\hat{\beta}_0 + \hat{\beta}_1 x_h) \pm W \sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right)},$$

where

$$W = \sqrt{2 f_{\alpha,2,n-2}}.$$

Note:


N=50
y
x
130
140
150
160
170
180
190
200
64
66
68
70
72


N=200
y
x
140
160
180
200
65
70
75

Notes: