Unit #2: Regression Basics

Linear Models with R, Chapter 1-2

Motivation

So far, we've been trying to estimate or test hypotheses about data generated from a distribution with a constant mean. For example:

X~N(m,s^2) - this is a known constant

But what if the mean depended on another variable? For example:

Y~N(My|x ,s^2) - suggest the mean of Y depends on X and other params

The Context of Linear Regression

Linear regression is used to explain or model the relationship between a single variable Y, and one or more variables $X_1,\ldots,X_{p}$.

Definition: Y is called the *response, outcome, output,* or *dependent* variable.

Definition: $X_1,\ldots,X_p$ are called *predictor*, *input*, *independent*, or *explanatory* variables. also covaraiates, featurs

We will assume, for now, that all variables are continuous (but will soon extend our methods to allow for discrete variables!).

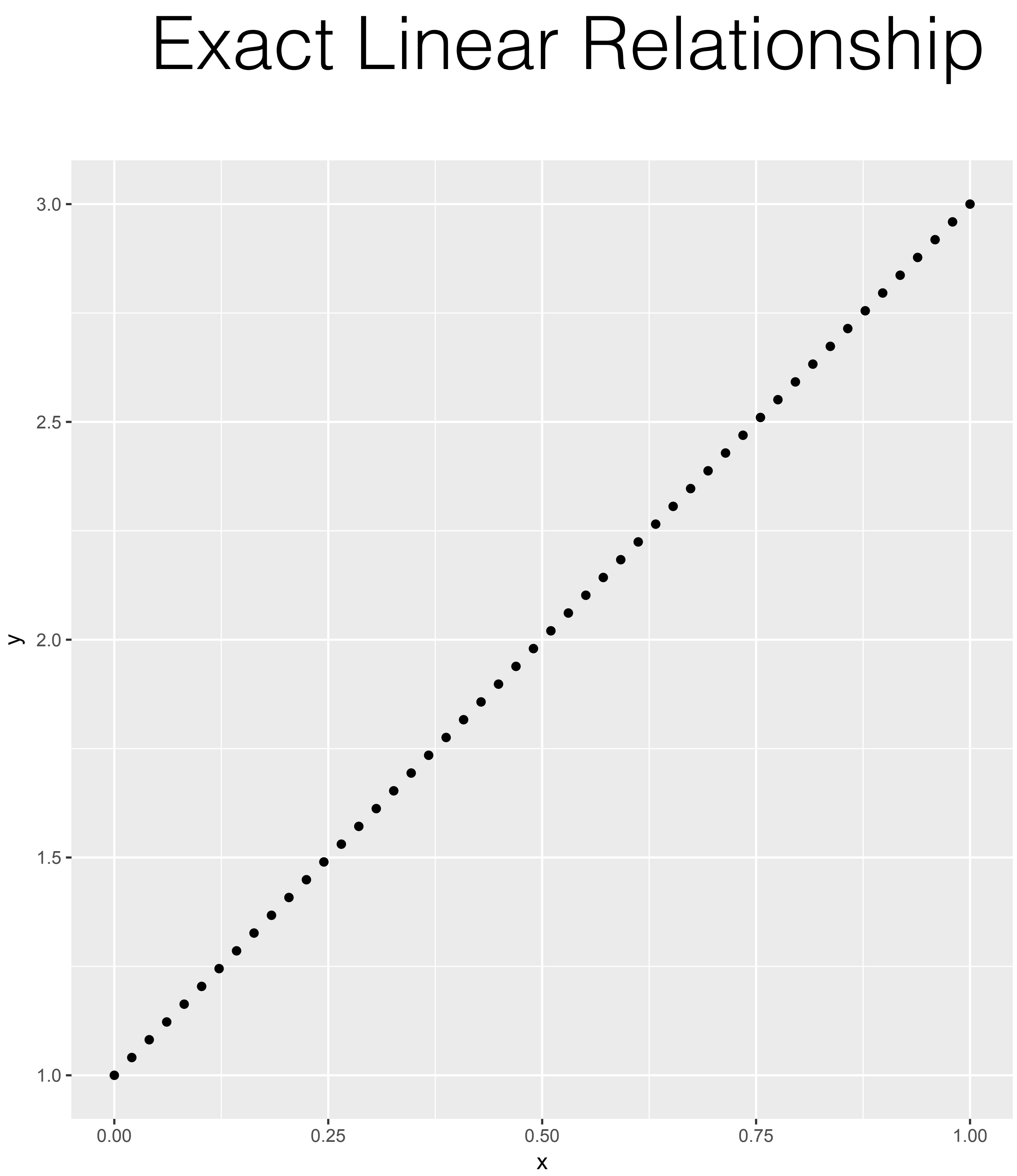
The Context of Linear Regression

The simplest relationship between these variables is a linear relationship:

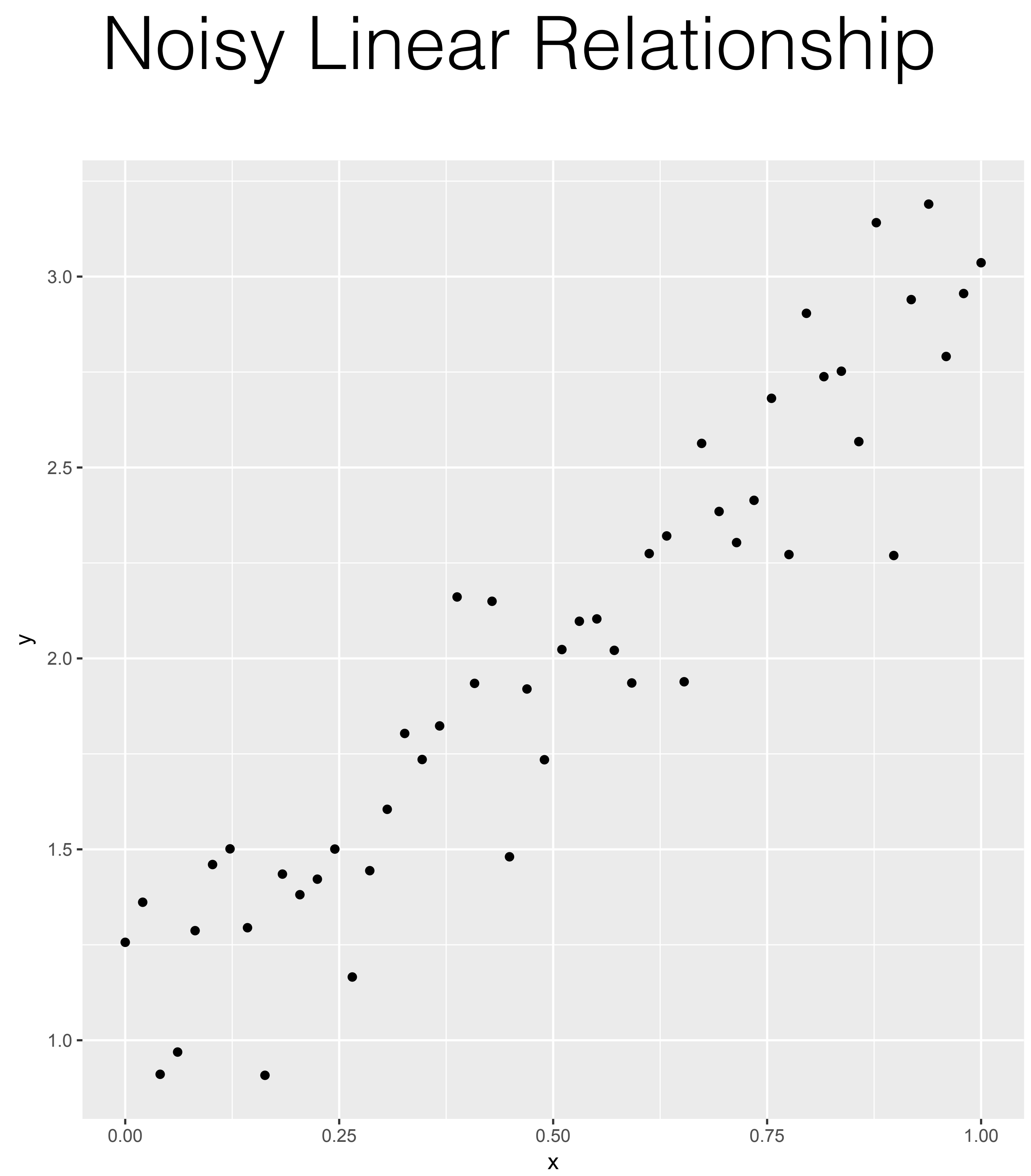
Let $i = 1,\ldots n$
for n = # of measurements, then
$Y_i = B_o + B_1X_{i,1}+..+B_pX_{i,p}$

But note that, when we actually measure data, measurements aren't perfect—there is error. So, we might use the following to model our data:

$Y_i = B_o + B_1X_{i,1}+..+B_pX_{i,p}+E_i$
Where E_i is a random variable



Exact Linear Relationship
y
x
3.0
2.5
2.0
1.5
1.0
0.00
0.25
0.50
0.75
1.00



Noisy Linear Relationship
y
x
3.0
2.5
2.0
1.5
1.0
0.00
0.25
0.50
0.75
1.00

The Context of Linear Regression

Regression analysis has two main objectives:

1. To make predictions about an unmeasured/unseen y using measured $x_1,\ldots,x_p$.

2. To assess the effect of, or the relationship between y and $x_1,\ldots,x_p$.

Can we infer causality?

Without further assumptions/designs… no!

What does "Linear" Mean?

Let ___Y = Y1,…Yn^T___ be the response variable and ___X1 = X11,…,Xn1^T___ be predictors.

Examples of linear models (i.e., models that linear methods can handle):

Yi = Bo + B1sin(Xi) + ei
Yi = exp(Bo +B1log(Xi))

X = [[1111], [x11,x21, …,xn1],…, [x1p,…,xnp]]
e = [e1,e2, …,en]

Examples of nonlinear models:

Y = Bolog(X1+B1X2)

Matrix Representation

Let $Y = (y1,\ldots,yn)^T$ be the response variable and $X1 = (x11,\ldots,xn1)^T$ be predictors. Let $Bo = (Bo,\ldots,Bn)^T$ be a vector of parameters. Finally, let $e^T = (e1,\ldots,en)^T$ be a vector of error terms. Then we can write our model as:

The matrix representation of X follows three rules:
1.) Each row is an observation/member of a sample/unit
2.) Each column is a predictor variable measured for each unit
3.) Ex: Let Xj = weight in lbs and Xij is the weight of the ith unit in the sample

The Linear Regression Model

Definition/Assumptions of the linear regression model:

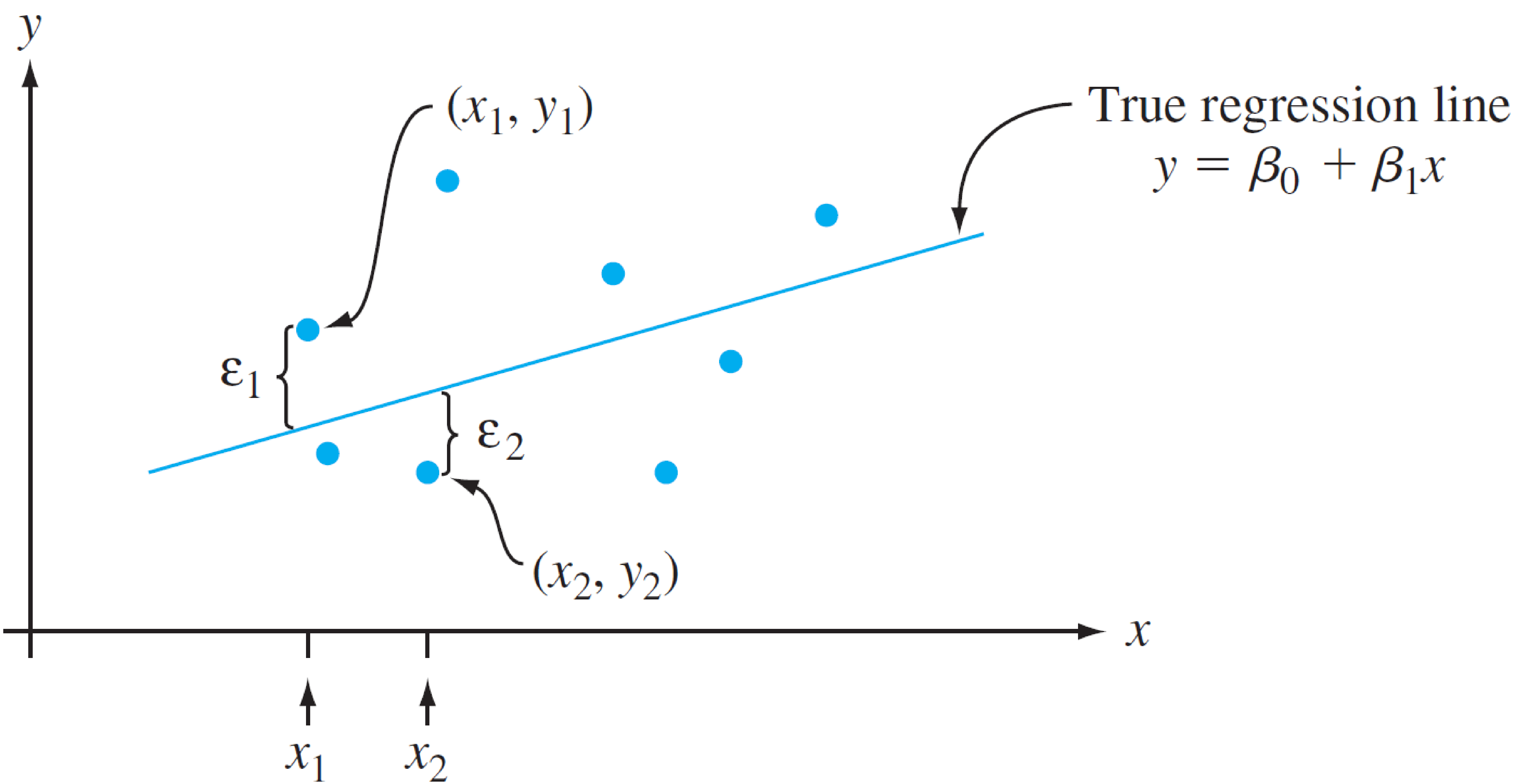
1.

2.

3.

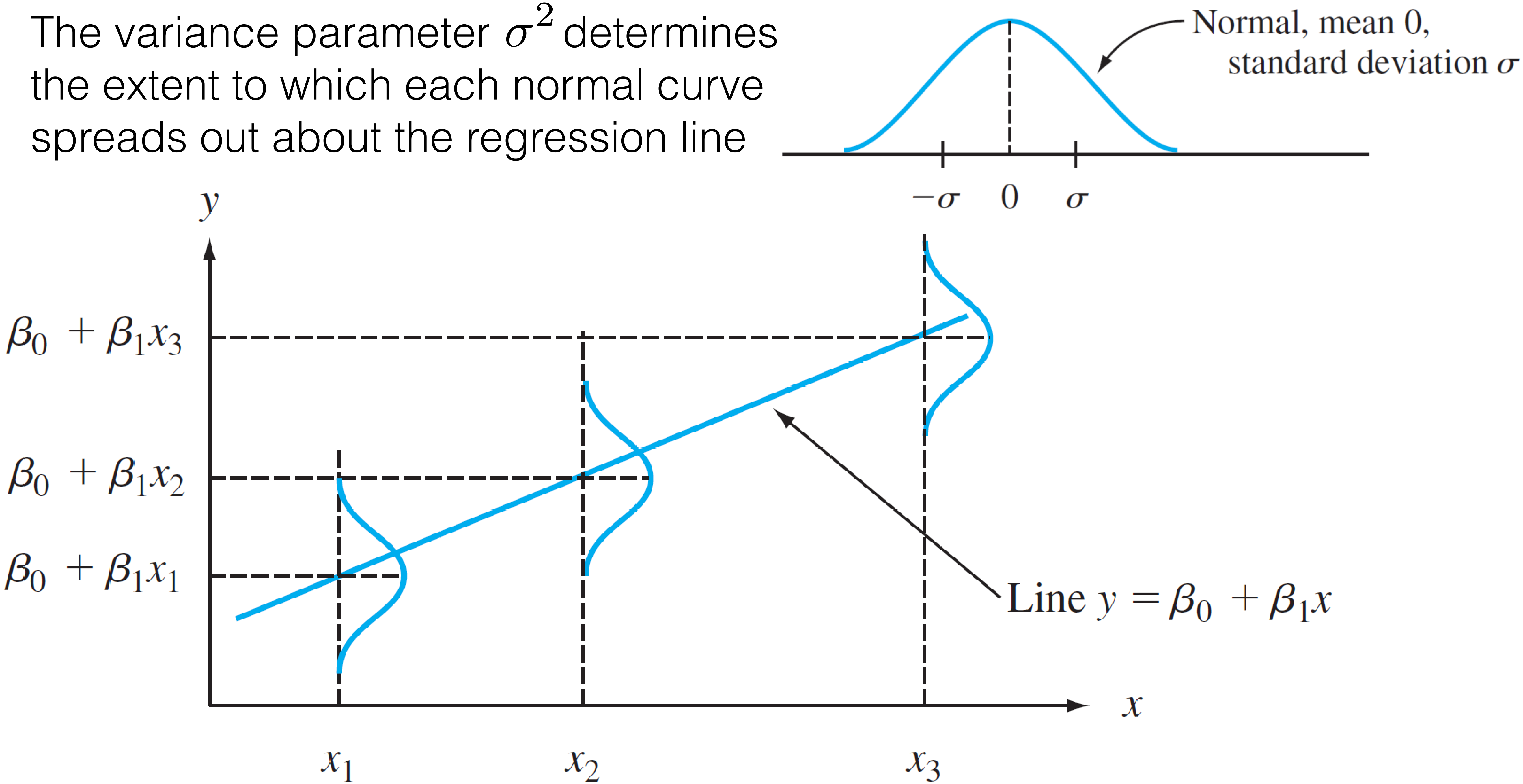
4.

The Simple Linear Regression Model



The Error Term

The variance parameter σ^2 determines the extent to which each normal curve spreads out about the regression line



The Linear Regression Model

How do we know linear regression is appropriate?

1. Theoretical considerations:

2. Experience with past data:

3. Exploratory data analysis:

Interpreting the Regression Parameters

Interpreting *simple* linear regression parameters:

β_0: the intercept of the true regression line:

The average value of Y when x is zero. Usually this is called the "baseline average".

β_1 : the slope of the true regression line:

The average change in Y associated with a 1-unit increase in the value of x.

Interpreting the Regression Parameters

Interpreting *multiple* linear regression parameters:

β_0: the intercept of the true regression surface. We interpret this as *the average value of Y when all of the x's are zero.*

β_j: the slope of the true regression surface. We interpret this as *the average change in Y associated with a 1-unit increase in the value of x_j assuming all other predictors are held constant. ($j = 1,\ldots p$)*

Thus, these "slope" parameters are called *partial* or *adjusted* regression parameters/coefficients.

In contrast, the *simple regression* slope is called the marginal (or *unadjusted*) coefficient.

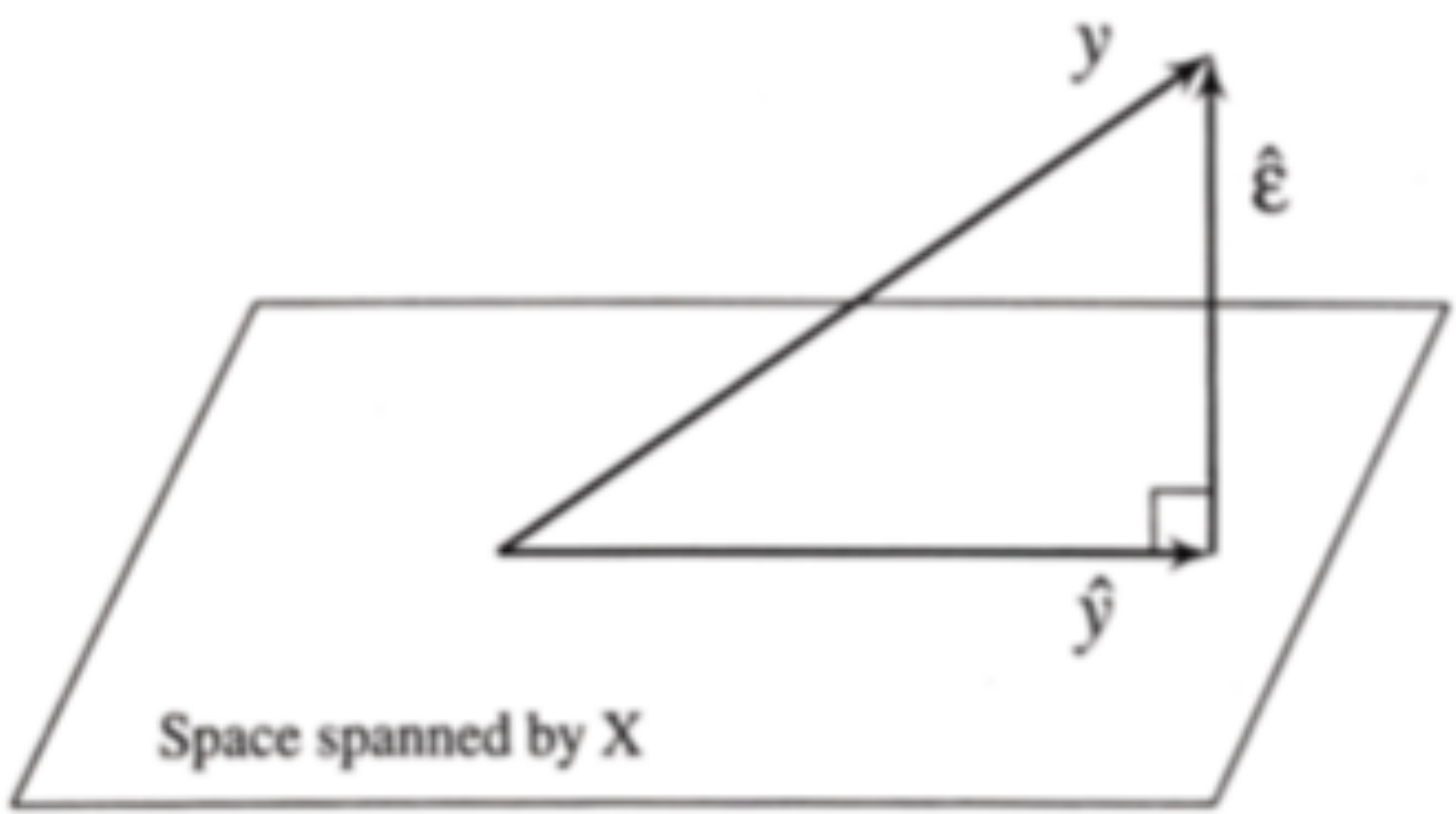
Estimating Model Parameters

The regression model _______________ partitions the response into a systematic component ______ and a random component ______. We want to choose ______ so that the systematic part explains the response as much as possible, without “overfitting” (trying to explain random variability/ measurement error).

The problem is the find a _____ so that ______ is as close as possible to _____.

Estimating Model Parameters

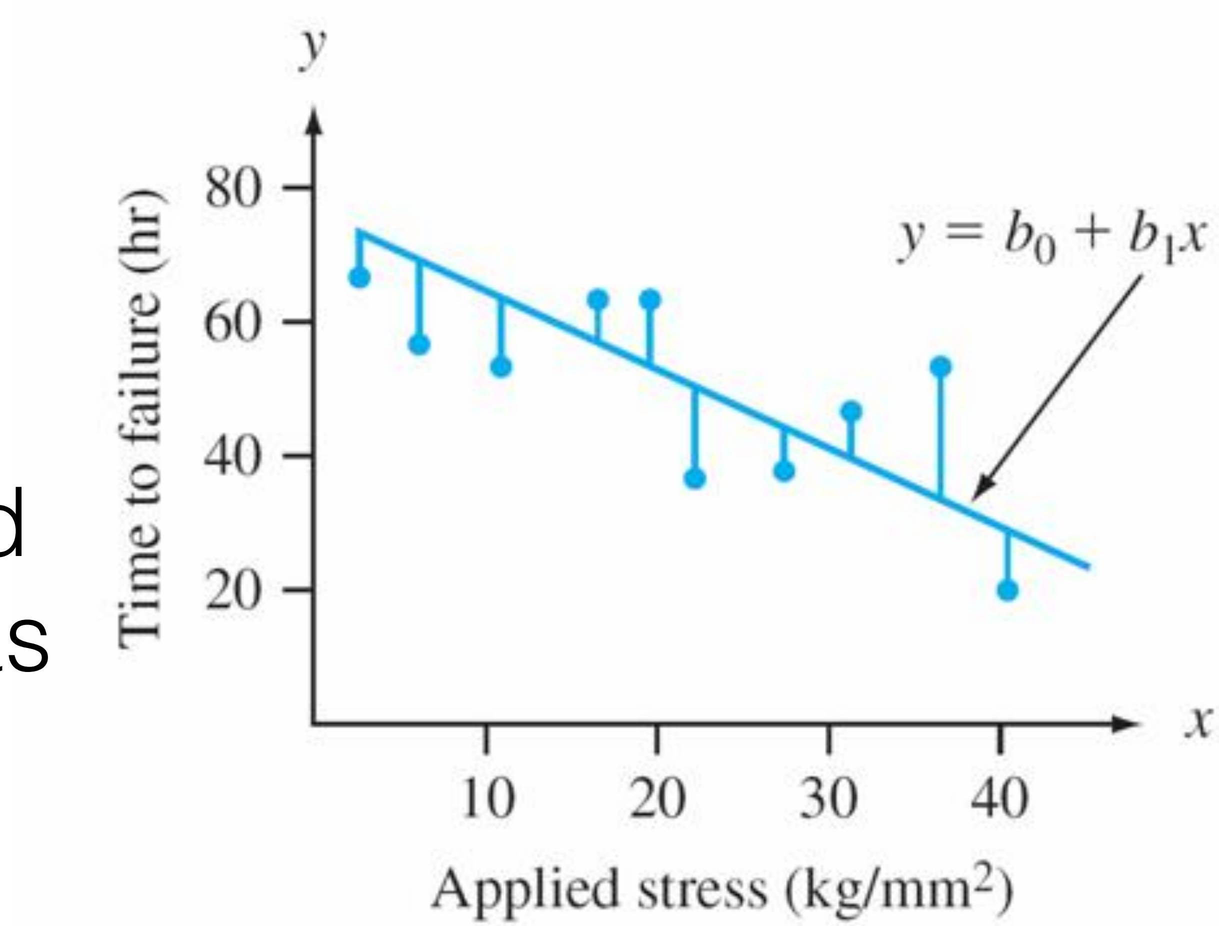
Geometric representation:



Estimating Model Parameters: SLR

The “best fit” line is motivated by the principle of **least squares**, which can be traced back to the German mathematician Gauss (1777–1855):

“A line provides the **best fit** to the data if the sum of the squared vertical distances (deviations) from the observed points to that line is as small as it can be.”



Estimating Model Parameters

Aside: Some linear algebra

1. Lemma 1:

2. Lemma 2:

3. Lemma 3:

Estimating Model Parameters

Least Squares Estimation: We define the best estimate of _____ as the one that minimized the sum of the squared errors:

Differentiating with respect to _____, we get:

Estimating Model Parameters

The Gauss-Markov Theorem: Suppose that:

1.

2.

3.

4.

Then ________________ is the “best” *unbiased* estimator of ______.

Estimating Model Parameters

Definition: The *hat matrix*, *H*, is defined as:

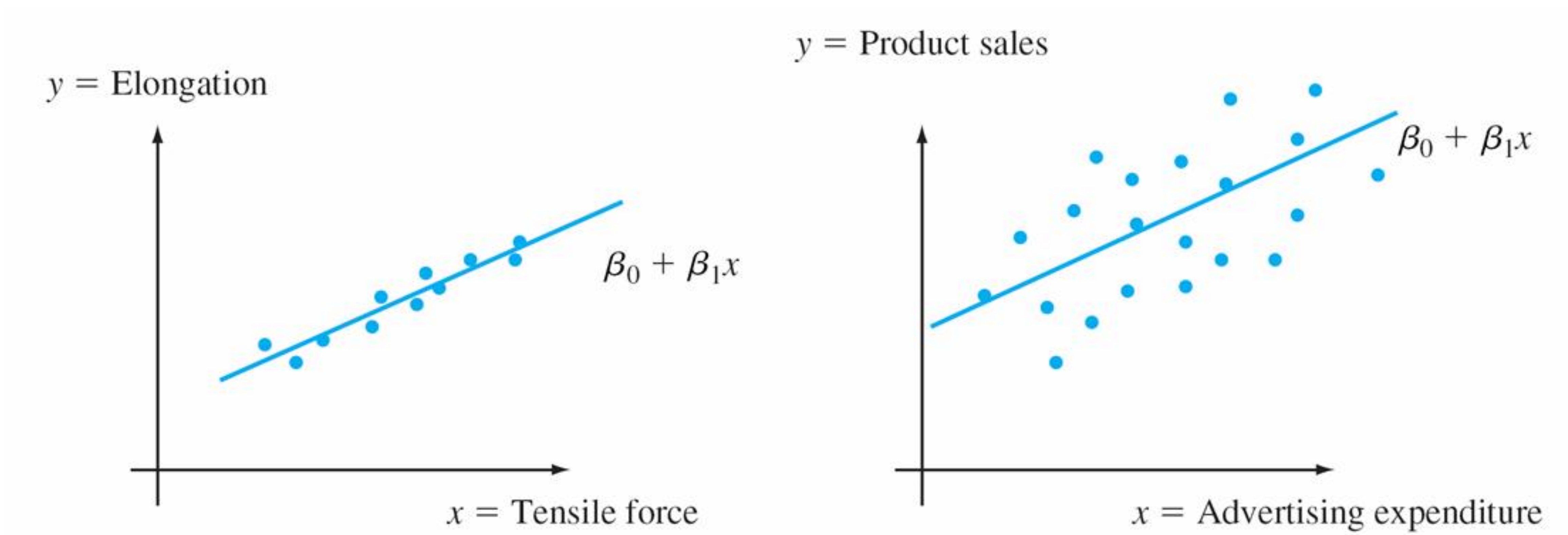
The hat matrix is useful in theoretical calculations.

Definition: The *fitted values* are defined as:

Definition: The *residuals* are defined as:

Estimating the errors

The parameter σ^2 determines the amount of spread about the true regression line. Two separate examples:



Estimating the errors

An estimate of σ^2 will be used in confidence interval formulas and hypothesis testing procedures presented in the next two sections. Recall that the residual sum of squares (RSS) is:

So, our estimate of the variance of the model is:

Estimating the errors

<u>Note</u>:

1. The divisor $n-(p+1)$ in is the number of degrees of freedom (df) associated with RSS and $\widehat{\sigma}^2$.

2. This is because to obtain $\widehat{\sigma}^2$, $p+1$ parameters must first be estimated, which results in a loss of $p+1$ df.

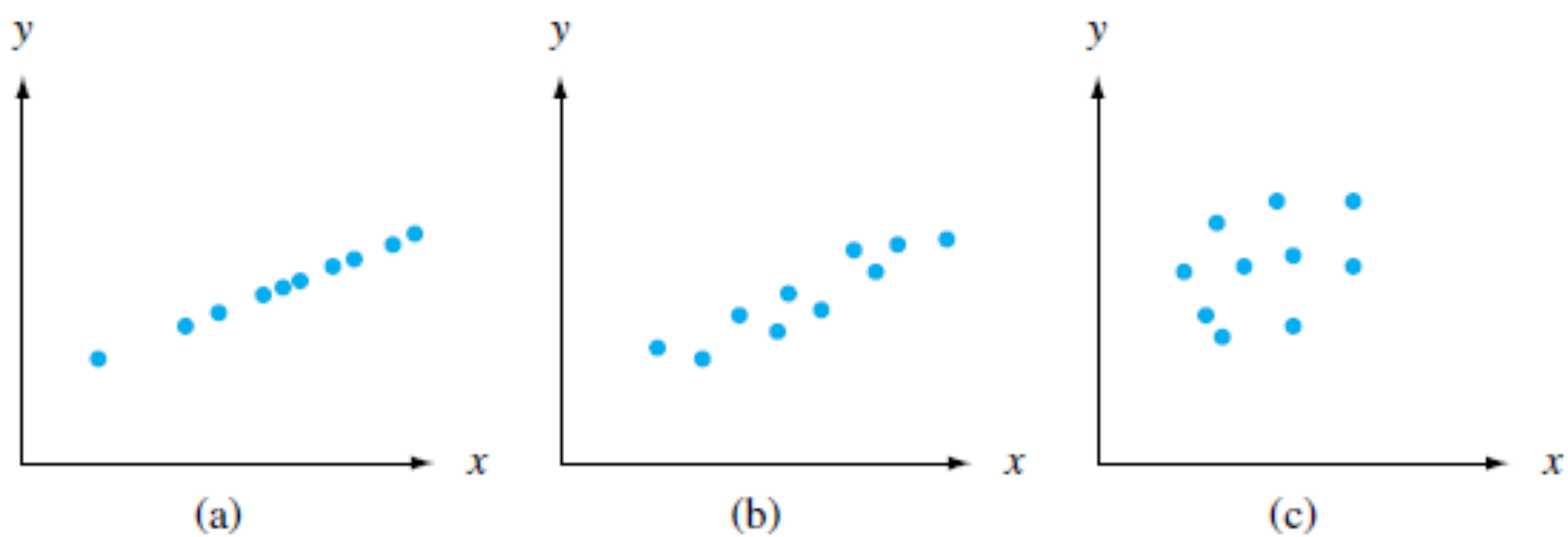
3. Replacing each y_i in the formula for $\widehat{\sigma}^2$ by the r.v. Y_i gives a random variable.

4. It can be shown that the r.v. $\widehat{\sigma}^2$ is an *unbiased estimator* for σ^2.

Sums of Squares

The residual sum of squares RSS can be interpreted as a measure of how much variation in *y* is left *unexplained by the model*—that is, how much cannot be attributed to a linear relationship.

In the first plot RSS = 0, and there is no unexplained variation, whereas unexplained variation is small for second, and large for the third plot.



Sums of Squares

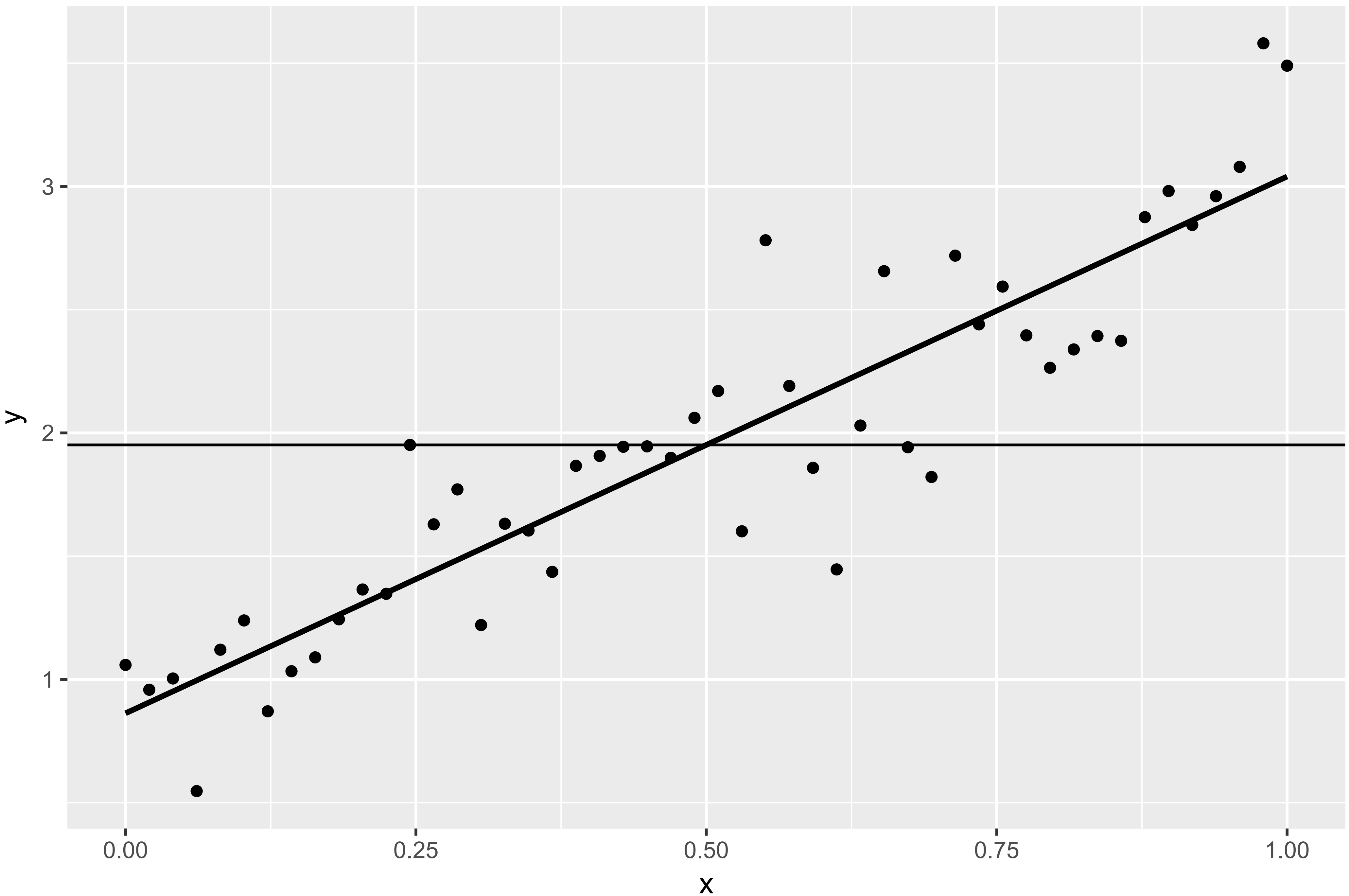
Other sums of squares:

1. RSS: Residual sum of squares:

2. ESS: *Explained* (or regression) sum of squares:

3. TSS: *Total* sum of squares:

Sums of Squares



Coefficient of Determination

The sum of squared deviations about the least squares line is smaller than the sum of squared deviations about *any* other line, i.e. RSS < TSS unless the horizontal line itself is the least squares line.

The ratio RSS/TSS is the proportion of total variation that cannot be explained by the simple linear regression model. The *coefficient of determination* is:

This coefficient is a number between 0 and 1 and is the *proportion of observed y variation explained by the model*.

Coefficient of Determination

Again, R^2 is the proportion of observed *y* variation <u>explained by the model</u>.

The higher the value of *R^2*, the more successful is the linear regression model in explaining *y* variation, **assuming the linear model is correct.**

Identifiability

The least squares estimate is the solution to the *normal* equations:

where _____ is a _______ matrix. If _______ cannot be inverted, then there will be infinitely many solutions to the normal equations and ______ is at last partially *nonidentifiable.* We cannot invert ________ when the columns of _____ are linearly dependent.

Identifiability

Why might we have nonidentifiability?

1. One variable is just a multiple of another.

2. One variable is a *linear combination* of several others.

3. We have more variables than members in the sample, i.e., ________.

Note: nonidentifiability is a relatively easy problem to work with. *Near nonidentifiability* is trickier. We will look into this later on (when we discuss collinearity).