Chapter 1: A Review of Basic Concepts B

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

A Review of Basic Concepts

Testing a Hypothesis About a Population Mean

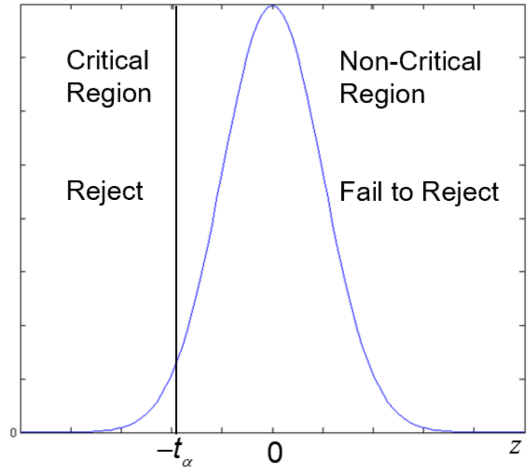
1. **Null Hypothesis (denoted H_0):** This is the hypothesis that is postulated to be true.
2. **Alternative Hypothesis (denoted H_a):** This hypothesis is counter to the null hypothesis and is usually the hypothesis that the researcher wants to support.
3. **Test Statistic:** Calculated from the sample data, this statistic functions as a decision-maker.
4. **Level of significance (denoted α):** This is the probability of a *Type I error* (i.e., the probability of rejecting H_0 given that H_0 is true).
5. **Rejection Region:** Values of the test statistic that lead the researcher to reject H_0 and accept H_a.
6. **p-Value:** Also called the observed significance level, this is the probability of observing a value of the test statistic at least as contradictory to the null hypothesis as the observed test statistic value, assuming the null hypothesis is true.
7. **Conclusion:** The decision to “reject” or “fail to reject” H_0 based on the value of the test statistic, α, the rejection region, and/or the p-value.

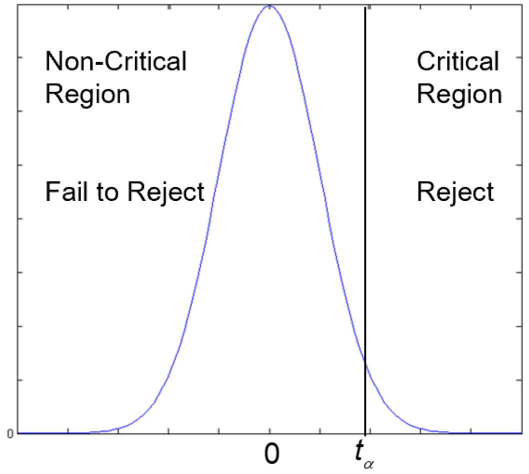
A Review of Basic Concepts

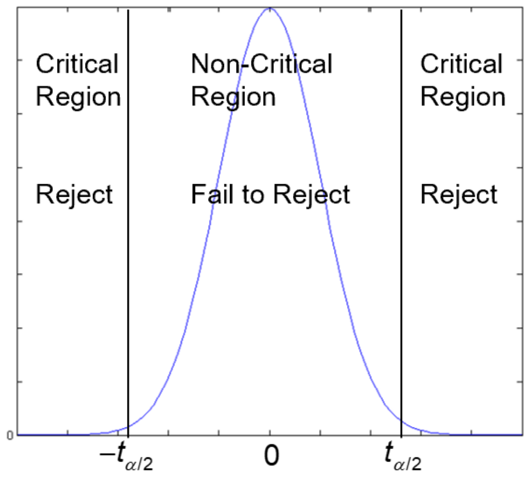
Testing a Hypothesis About a Population Mean
Small-Sample Test of Hypothesis about μ

Test statistic: $t = (\bar{y} - \mu_0) / (s / \sqrt{n})$

	ONE-TAILED TESTS		TWO-TAILED TEST
	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$
	$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu \neq \mu_0$
Rejection region:	$t < -t_\alpha$	$t < t_\alpha$	$\lvert t \rvert > t_{\alpha/2}$
p-*value:*	$\mathrm{P}(t < t_c)$	$\mathrm{P}(t > t_c)$	$2\mathrm{P}(t > t_c)$ if t_c is positive $2\mathrm{P}(t < t_c)$ if t_c is negative







A Review of Basic Concepts

Inferences About the Difference Between Two Population Means
Small-Sample Confidence Interval for (μ_1-μ_2): Independent Samples

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\ , \qquad s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

s_p^2 is a “pooled” estimate of the common population variance and
$t_{\alpha/2}$ is based on df=n_1+n_2-2

Assumptions:
Both sampled populations have distributions that are approximately normal.
The population variances are equal.
The samples are randomly and independently selected from the populations.

A Review of Basic Concepts

Testing a Hypothesis About a Population Mean

Small-Sample Test of Hypothesis About (μ_1-μ_2):

Dependent Samples

Test statistic:

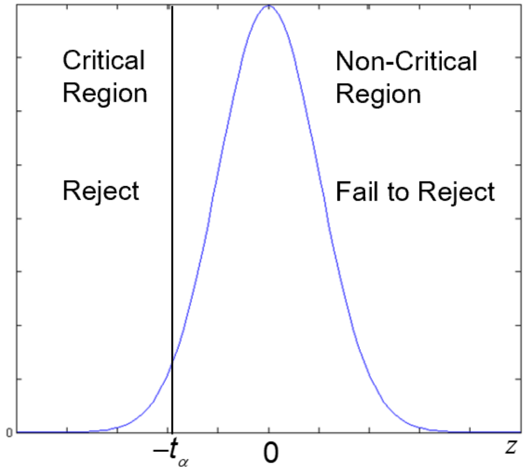
$$t = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

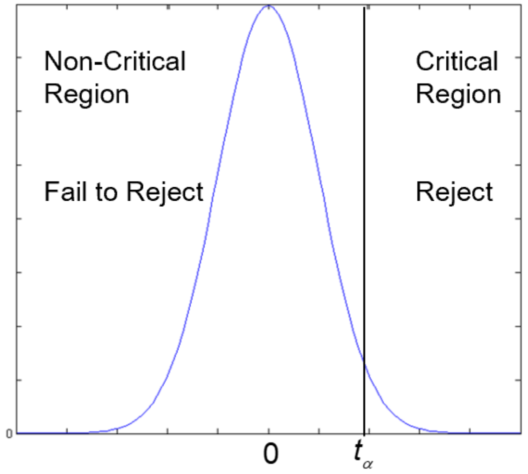
where

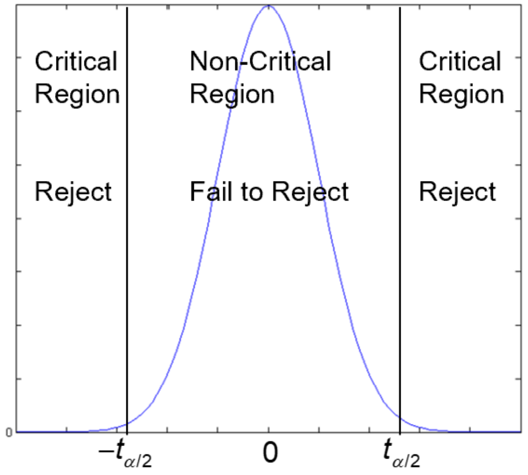
$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}.$$

	ONE-TAILED TESTS		TWO-TAILED TEST
	$H_0: \mu_1 - \mu_2 = D_0$	$H_0: \mu_1 - \mu_2 = D_0$	$H_0: \mu_1 - \mu_2 = D_0$
	$H_a: \mu_1 - \mu_2 < D_0$	$H_a: \mu_1 - \mu_2 > D_0$	$H_a: \mu_1 - \mu_2 \neq D_0$
Rejection region:	$t < -t_\alpha$	$t > t_\alpha$	$\lvert t\rvert > t_{\alpha/2}$
p-*value*:	$P(t < t_c)$	$P(t > t_c)$	$2P(t > t_c)$ if t_c is positve $2P(t < t_c)$ if t_c is negative

Decision: Reject H_0 if $\alpha > p$-value, or, if test statistic falls in rejection region







A Review of Basic Concepts

Inferences About the Difference Between Two Population Means
Paired-Difference Confidence Interval for $\mu_d=\mu_1-\mu_2$: Dependent Samples

$$\bar{y}_d \pm t_{\alpha/2} \frac{s_d}{\sqrt{n_d}} \quad ,$$

$\bar{y}_d$ is the sample mean difference
s_d is the sample standard deviation of differences and
$t_{\alpha/2}$ is based on $df=n_d-1$

Assumptions:
Population of differences has a normal distribution.
The samples differences are randomly selected from the population.

A Review of Basic Concepts

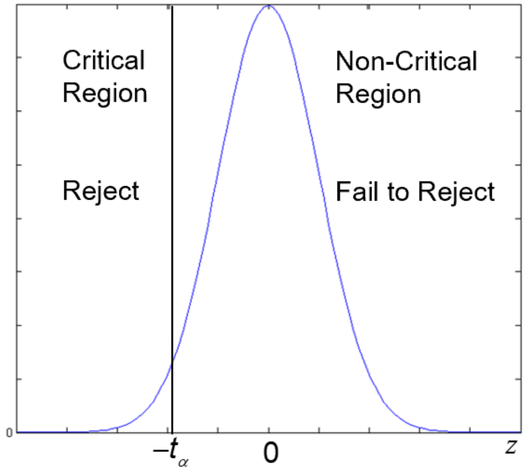
Testing a Hypothesis About a Population Mean
Paired Difference Test of Hypothesis for $\mu_d = \mu_1 - \mu_2$:

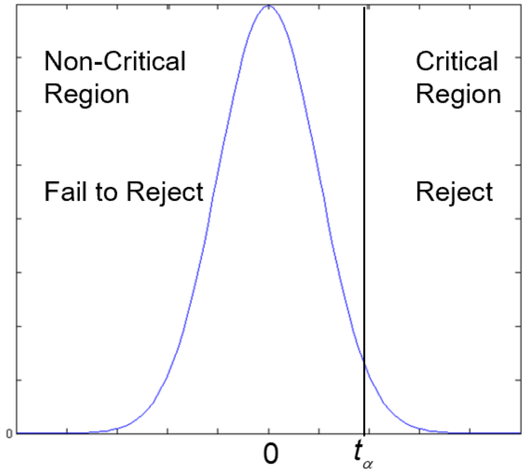
ONE-TAILED TESTS		TWO-TAILED TEST
$H_0: \mu_d = D_0$	$H_0: \mu_d = D_0$	$H_0: \mu_d = D_0$
$H_a: \mu_d < D_0$	$H_a: \mu_d > D_0$	$H_a: \mu_d \neq D_0$

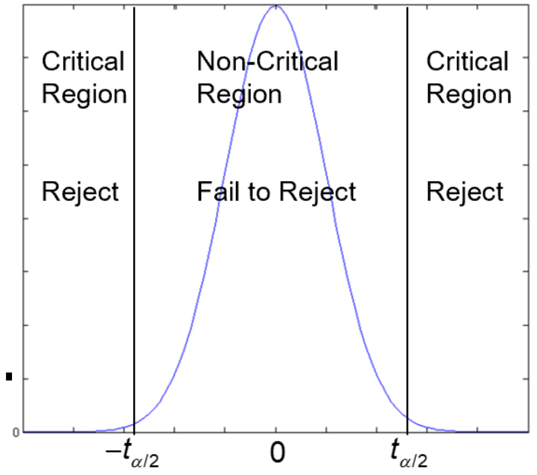
Test statistic: $t = \dfrac{\bar{y}_d - D_0}{s_d / \sqrt{n_d}}$

Rejection Region:	$t < -t_\alpha$	$t < t_\alpha$	$\lvert t\rvert > t_{\alpha/2}$
p-*value*:	$\mathrm{P}(t > t_c)$	$\mathrm{P}(t > t_c)$	$2\mathrm{P}(t > t_c)$ if t_c is positive $2\mathrm{P}(t < t_c)$ if t_c is negative

Assumptions:
The differences are randomly selected from the population of differences.
The relative freq. distribution of the population of differences is normal.







A Review of Basic Concepts

Comparing Two Population Variances

$$H_0\text{:}\ \frac{\sigma_1^2}{\sigma_2^2} = k\ \ (\sigma_1^2 = k\sigma_2^2) \qquad \text{usually } k=1$$

$$H_a\text{:}\ \frac{\sigma_1^2}{\sigma_2^2} \neq k\ \ (\sigma_1^2 \neq k\sigma_2^2)$$

Test Statistic:

$$F = \frac{s_1^2}{ks_2^2}$$

(put larger sample variance in numerator)

Assumptions:

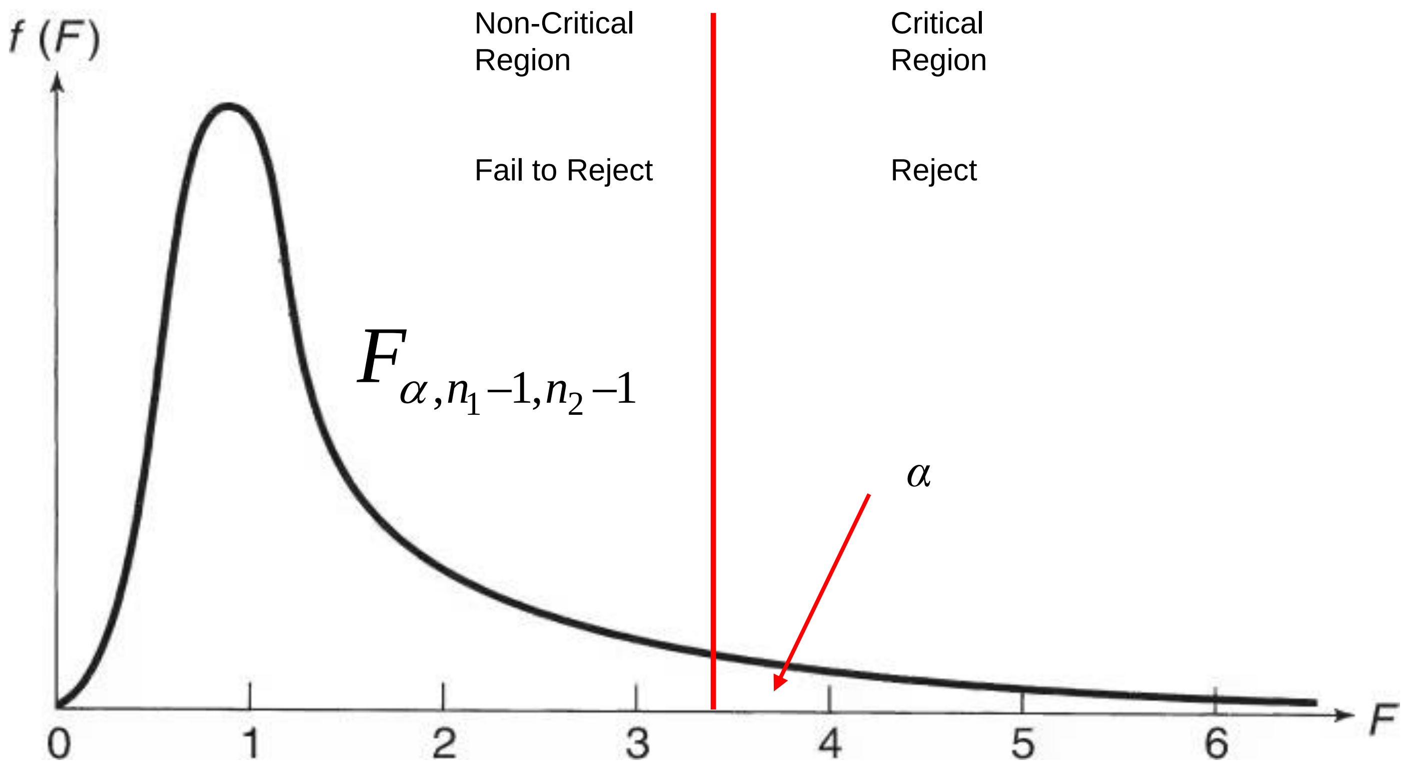
The two sampled populations are normally distributed.

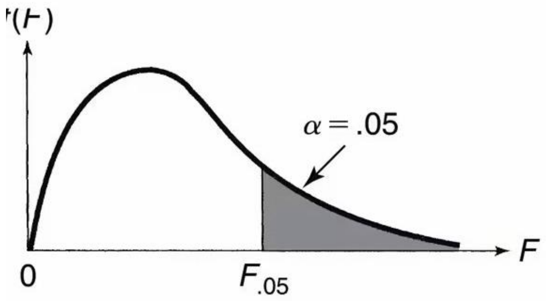
The samples are randomly and independently selected from their respective populations.

A Review of Basic Concepts

Comparing Two Population Variances

$F = \dfrac{s_1^2}{k s_2^2}$ has an F-distribution with $(n_1\text{-}1)$ numerator df and $(n_2\text{-}1)$ denominator df





v_2 \ v_1	Numerator Degress of Freedom						
Denominator Degress of Freedom	1	2	3	4	5	6	7
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01
12	4.75	3.89	3.49	3.25	3.11	3.00	2.91
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76

A Review of Basic Concepts

Comparing Two Population Variances
F-Test for Equal Population Variances: Independent Samples

	ONE-TAILED TESTS		TWO-TAILED TEST
	$H_0: \sigma_1^2 = \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$
	$H_a: \sigma_1^2 < \sigma_2^2$	$H_a: \sigma_1^2 > \sigma_2^2$	$H_a: \sigma_1^2 \neq \sigma_2^2$
Test statistic:	$F = s_2^2/s_1^2$	$F = s_1^2/s_2^2$	$F = \frac{\text{Larger sample variance}}{\text{Smaller sample variance}}$

Rejection Region:	$F > F_\alpha$	$F > F_\alpha$	$F > F_{\alpha/2}$
Numerator df (υ_1):	$n_2 - 1$	$n_1 - 1$	$n - 1$ for large variance
Denominator df (υ_2):	$n_1 - 1$	$n_2 - 1$	$n - 1$ for smaller variance
p-*value*:	$\text{P}(F > F_c)$	$\text{P}(F > F_c)$	$\text{P}(F^* < 1/F_c) + \text{P}(F > F_c)$

Assumptions:

Both sampled populations are normally distributed.

The samples are random and independent.

A Review of Basic Concepts

Homework:

Read Chapter 1

Chapter 2: Introduction to Regression Analysis

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Introduction to Regression Analysis

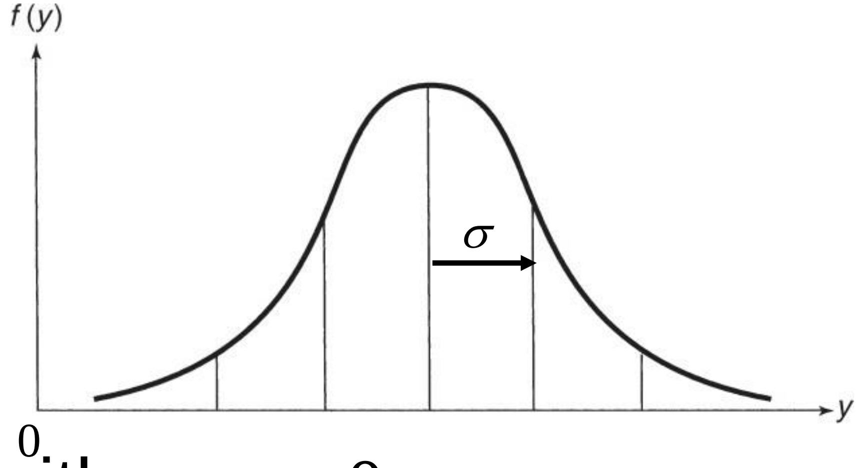
Modeling a Response

Regression analysis is a branch of statistical methodology concerned with relating a response y to a set of independent, or predictor, variables $x_1,\ldots,x_k$.

Our goal is to build a model that mathematically describes the relationship between a value of our independent variable x and our dependent variable y. and allow us to predict the value of y for a given value of x.

$$y = \underbrace{E(y)}_{\uparrow} + Random\ Error$$

expected mean function



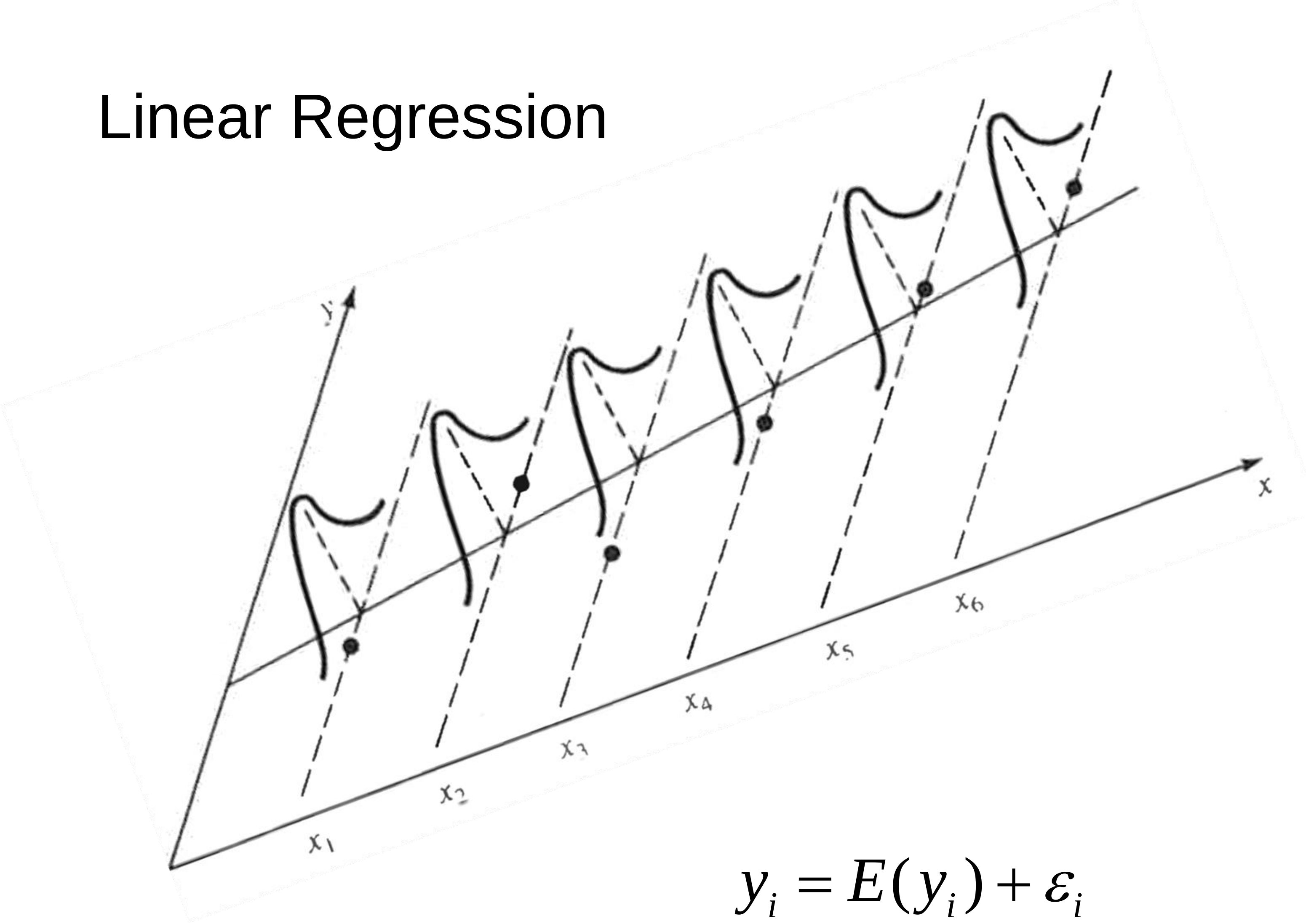
Random Error is a random draw from a normal distribution with mean 0 and variance σ^2.

Introduction to Regression Analysis
Modeling a Response

$$\varepsilon \sim N(0,\sigma^2)$$

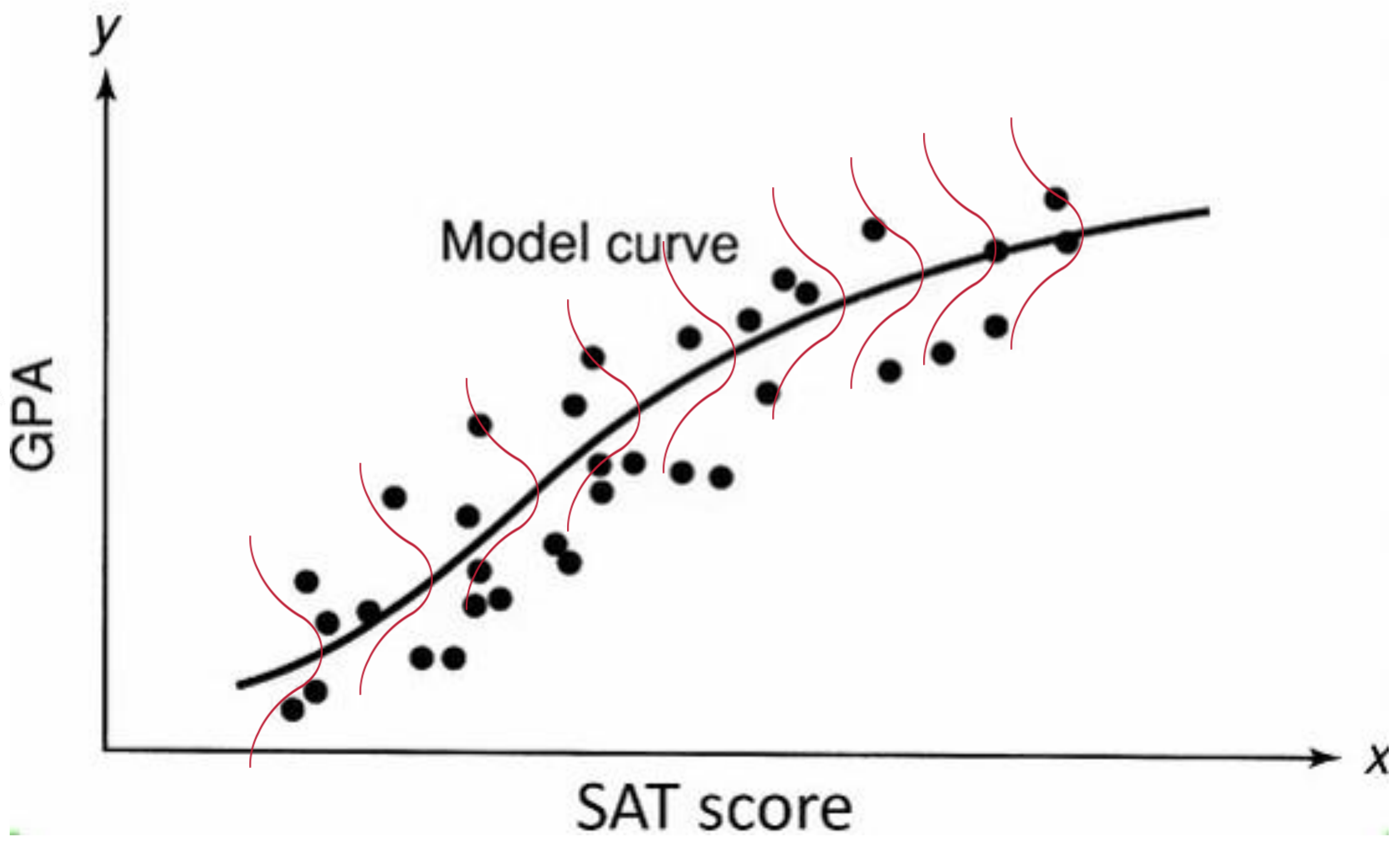
At each point along the curve, observations have additive normal error ε.

Linear Regression



$$y_i = E(y_i) + \varepsilon_i$$

Non-Linear Regression



Introduction to Regression Analysis

Overview of Regression Analysis

If we have a "smooth" function $E(y|x)=f(x)$ that depends on a single independent variable, then we can represent it with a Taylor series expansion around 0 as

$$E(y|x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \ldots$$

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots$$

Linear in the parameters regression.

https://en.wikipedia.org/wiki/Taylor_series

Introduction to Regression Analysis

Overview of Regression Analysis

If we have a “smooth” function $E(y|x)=f(x)$ that depends on a single independent variable, then we can represent it with a Taylor series expansion around 0 as

$$E(y|x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \ldots$$

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots$$

and if $E(y|x_1,x_2)$ depends on two independent variables, then

$$E(y|x_1,x_2) = f(0,0) + \frac{f_{x_1}(0,0)}{1!}x_1 + \frac{f_{x_2}(0,0)}{2!}x_2 + \frac{f_{x_1x_1}(0,0)}{2!}x_1^2 + \frac{f_{x_2x_2}(0,0)}{3!}x_2^2 + \frac{f_{x_1x_2}(0,0)}{3!}x_1x_2 + \ldots$$

$$E(y|x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \ldots$$

Linear in the parameters regression.

https://en.wikipedia.org/wiki/Taylor_series

Introduction to Regression Analysis

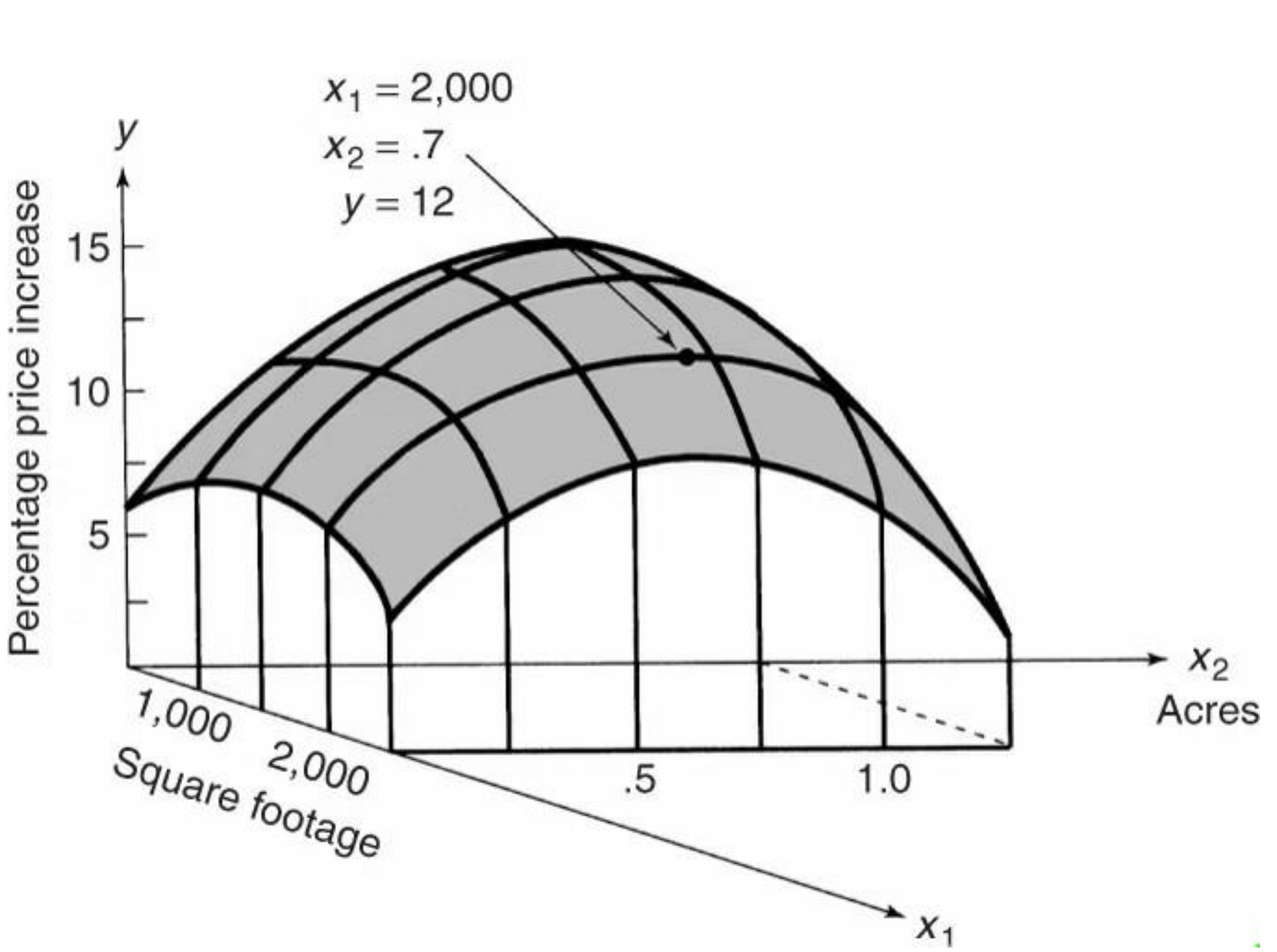
Overview of Regression Analysis

Example: A property appraiser might like to relate percentage price increase y of residential properties to the two quantitative independent variables x_1, square footage of heated space, and x_2, lot size.

This model could be represented by a response surface that traces the mean percentage price increase $E(y\,|x_1,x_2)$ for various combinations of x_1 and x_2.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \hat{\beta}_4 x_1^2 + \hat{\beta}_5 x_2^2$$



Introduction to Regression Analysis

Overview of Regression Analysis

Seven

Regression Modeling: ~~Six~~-Step Procedure

1. Hypothesize the form of the model for $E(y)$.
2. Collect the sample data.
3. Use the sample data to estimate unknown parameters in the model.
4. Specify the probability distribution of the random error term, and estimate any unknown parameters of this distribution.
5. Statistically check the usefulness of the model.
6. Check the validity of the assumptions on the random error term, and make model modifications if necessary.
7. When satisfied that the model is useful, and assumptions are met, use the model to make inferences, i.e., parameter interpretation, prediction, estimation, etc.

Introduction to Regression Analysis

Collecting the Data for Regression

Definition 2.3

If the values of the independent variables (x's) in regression are **uncontrolled** (i.e., not set in advance before the value of y is observed) but are measured without error, the data are observational.

	Executive				
	1	2	3	4	5
Annual compensation, y ($)	85,420	61,333	107,500	59,225	98,400
Experience, x_1 (years)	8	2	7	3	11
College education, x_2 (years)	4	8	6	7	2
No. of employees supervised, x_3	13	6	24	9	4
Corporate assets, x_4 (millions, $)	1.60	0.25	3.14	0.10	2.22
Age, x_5 (years)	42	30	53	36	51
Board of directors, x_6 (1 if yes, 0 if no)	0	0	1	0	1
International responsibility, x_7 (1 if yes, 0 if no)	1	0	1	0	0

Introduction to Regression Analysis

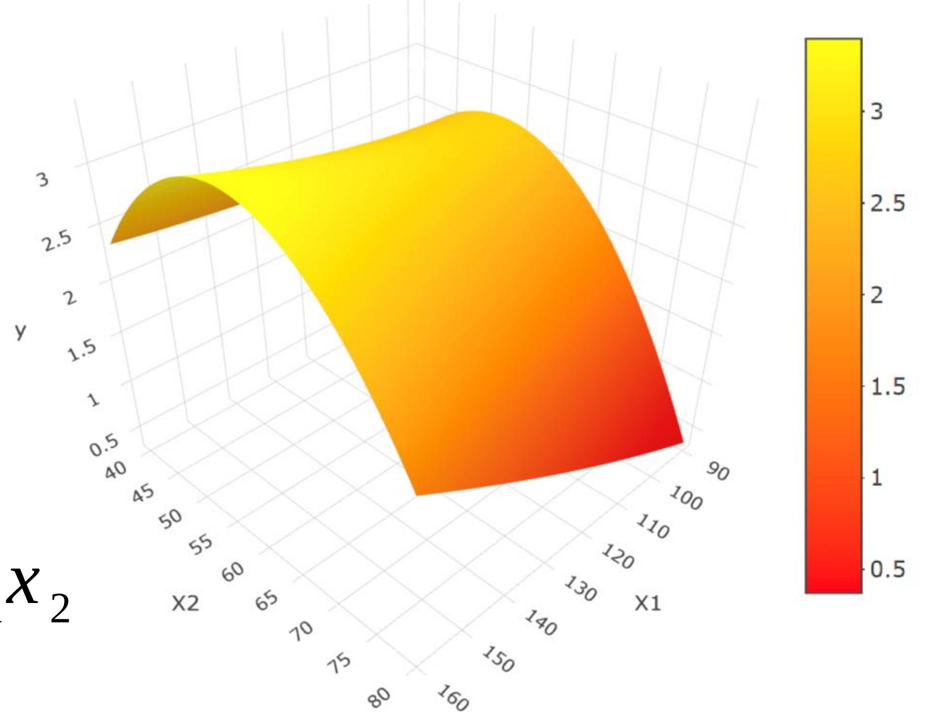
Collecting the Data for Regression

Definition 2.4

If the values of the independent variables (x's) in regression are **controlled** using a designed experiment (i.e., set in advance before the value of y is observed), the data are experimental.

Temperature, x_1	Pressure, x_2	Impurity, y
100	50	2.7
	60	2.4
	70	2.9
125	50	2.6
	60	3.1
	70	3.0
150	50	1.5
	60	1.9
	70	2.2

$$\hat{y} = -7.894 + 0.207x_1 - 0.061x_2$$
$$-0.001x_1^2 + 0.002x_2^2 + 0.005x_1x_2$$



Think of x as dial settings for your science experiment. Every time you fix an x, you run the experiment to get a y. In regression, x is fixed and known.

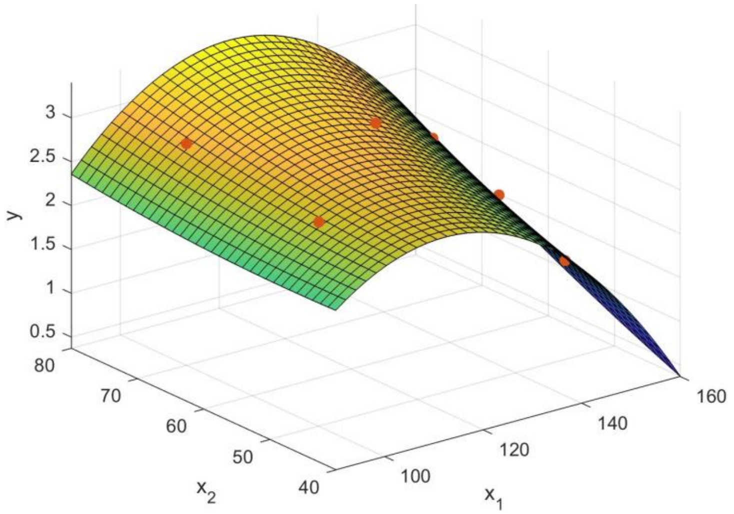
Introduction to Regression Analysis

Collecting the Data for Regression

```
% R code
install.packages("plotly")
library(plotly)

print("Temperature x1, Pressure x2, and Impurity y Data")

% enter x1 and x2 and y dataset
x1 <- c(100,100,100,125,125,125,150,150,150)
x2 <- c(50,60,70,50,60,70,50,60,70)
y  <- c(2.7,2.4,2.9,2.6,3.1,3.0,1.5,1.9,2.2)
x3=x1*x1;
x4=x2*x2;
x5=x1*x2;

# Fit linear model
lm.model_y <- lm(y ~ x1 + x2 + x3 + x4 + x5)
# get fitted coefficients
yhat <- lm.model_y$fitted.values
b0   <- lm.model_y$coefficients[1]
b1   <- lm.model_y$coefficients[2]
b2   <- lm.model_y$coefficients[3]
b3   <- lm.model_y$coefficients[4]
b4   <- lm.model_y$coefficients[5]
b5   <- lm.model_y$coefficients[6]
lm.model_y$coefficients
```

```
% Create surface plot
xx1 <- seq(90, 160, length.out = 100)
xx2 <- seq(40,   80, length.out = 100)
f <- function(xx1, xx2) {b0+b1*xx1+b2*xx2+b3*xx1*xx1+b4*xx2*xx2+b5*xx1*xx2}
fig <- plot_ly(x = xx1, y = xx2, z = outer(xx1, xx2, f),type = "surface",
           colorscale = list(c(0, 1), c("red", "yellow")))
fig <- fig %>%
  layout(scene = list(xaxis = list(title = 'X1'),yaxis = list(title = 'X2')
                ,zaxis = list(title = 'y')))
fig
```

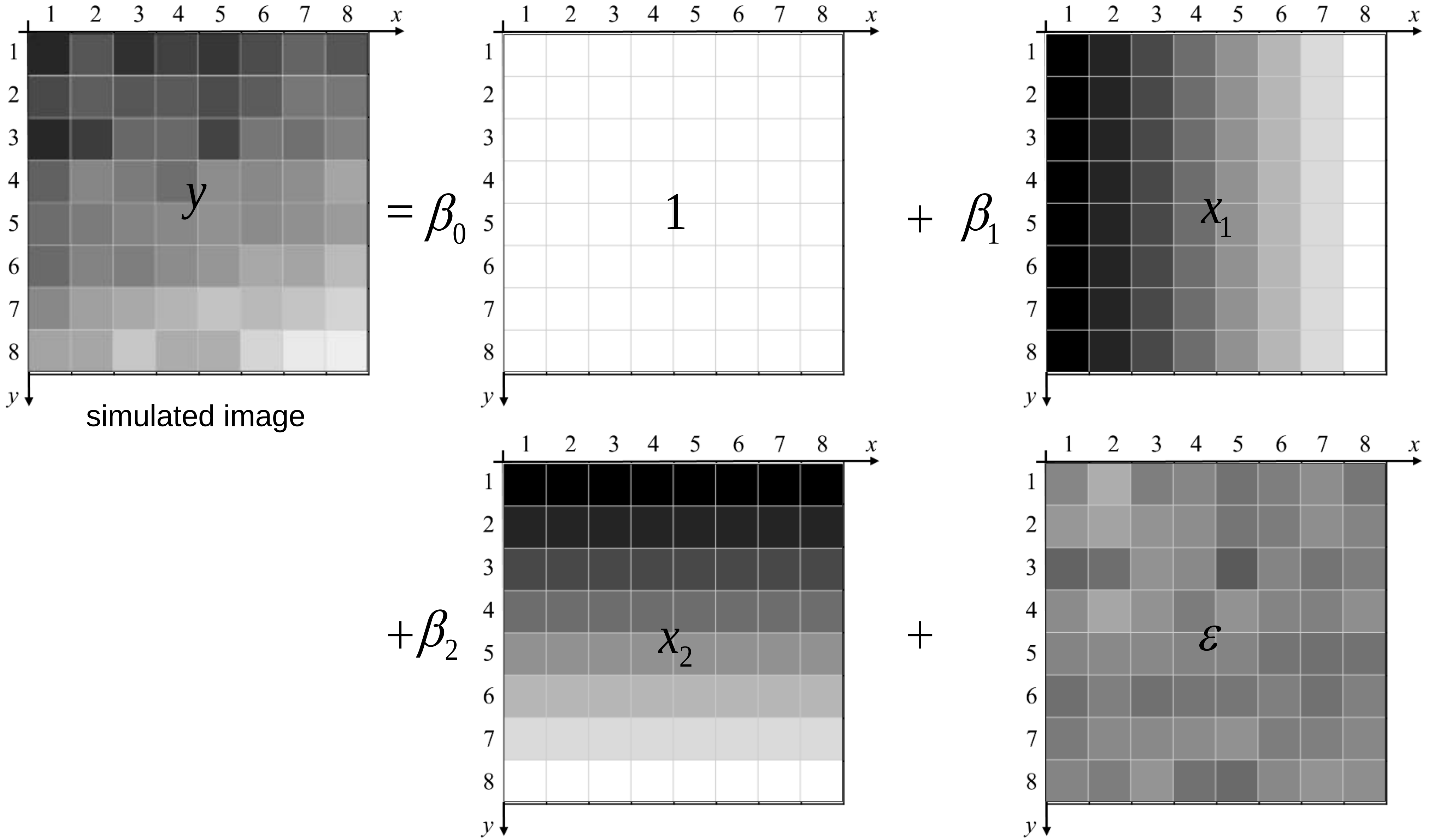
```
% Matlab code
x1 =[100,100,100,125,125,125,150,150,150]';
x2 =[50,60,70,50,60,70,50,60,70]';
y =[2.7,2.4,2.9,2.6,3.1,3.0,1.5,1.9,2.2]';

n=length(y);
X=[ones(n,1),x1,x2,x1.*x1,x2.*x2,x1.*x2];
bhat=inv(X'*X)*X'*y

fxy = @(xx1,xx2) bhat(1,1)+bhat(2,1)*xx1+bhat(3,1)*xx2+...
bhat(4,1)*xx1.^2+bhat(5,1)*xx2.^2+bhat(6,1)*xx1.*xx2;
figure;
fsurf(fxy,[90,160,40,80]), xlabel('x_1'),ylabel('x_2'),zlabel('y')
hold on, scatter3(x1,x2,y,'filled')
```



Introduction to Regression Analysis

$$y = X\beta + \varepsilon$$

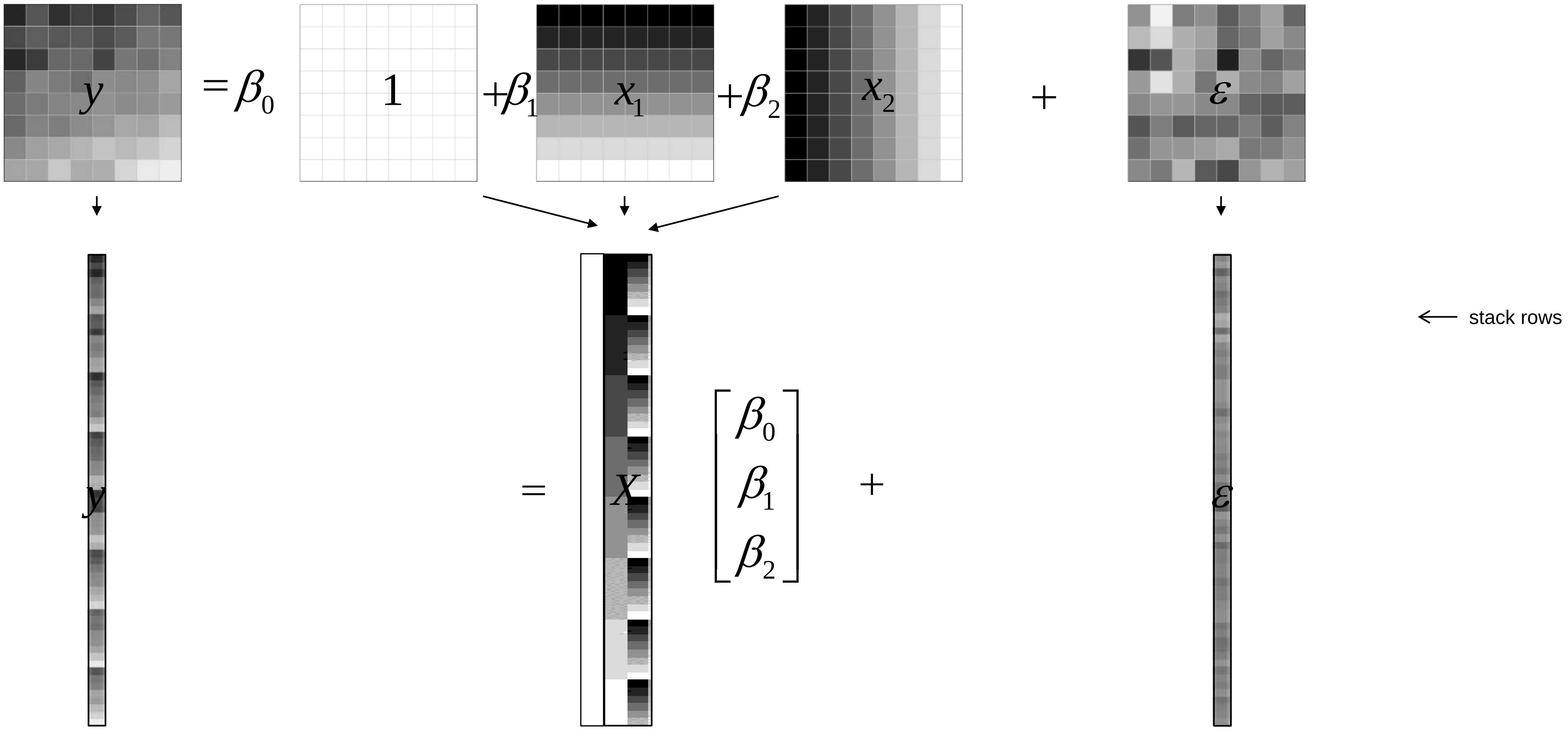
Example: We can use regression to fit a surface to (x,y) data.



Introduction to Regression Analysis

$$y = X\beta + \varepsilon$$

Example: We can use regression to fit a surface to (x,y) data.

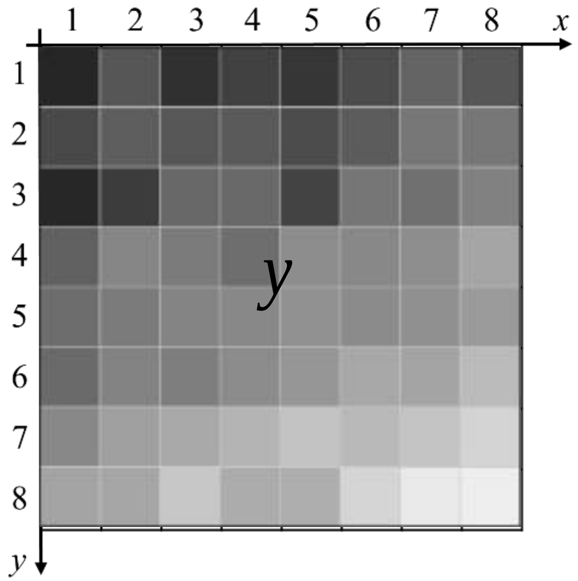


x1	x2	y
-1	-1	86.0753
-1	-0.7143	91.5249
-1	-0.4286	86.1966
-1	-0.1429	95.2958
-1	0.1429	97.0661
-1	0.4286	96.6703
-1	0.7143	101.2757
-1	1	105.6852
-0.7143	-1	93.5854
-0.7143	-0.7143	94.8246
-0.7143	-0.4286	89.4431
-0.7143	-0.1429	101.0698
-0.7143	0.1429	99.308
-0.7143	0.4286	100.5882
-0.7143	0.7143	105.0009
-0.7143	1	106.0186
-0.4286	-1	87.6089
-0.4286	-0.7143	93.6937
-0.4286	-0.4286	96.3895
-0.4286	-0.1429	99.263
-0.4286	0.1429	100.6287
-0.4286	0.4286	99.7279
-0.4286	0.7143	106.4345
-0.4286	1	111.1176
-0.1429	-1	90.2635
-0.1429	-0.7143	94.2122
-0.1429	-0.4286	96.4538
-0.1429	-0.1429	97.2503
-0.1429	0.1429	101.302
-0.1429	0.4286	101.9969
-0.1429	0.7143	108.2054
-0.1429	1	106.9916
0.1429	-1	88.5765
0.1429	-0.7143	91.9524
0.1429	-0.4286	90.54
0.1429	-0.1429	102.1625
0.1429	0.1429	102.7932
0.1429	0.4286	103.4901
0.1429	0.7143	110.5977
0.1429	1	107.2913
0.4286	-1	91.9384
0.4286	-0.7143	94.5171
0.4286	-0.4286	98.4956
0.4286	-0.1429	101.34
0.4286	0.1429	101.8417
0.4286	0.4286	106.3685
0.4286	0.7143	108.956
0.4286	1	113.3983
0.7143	-1	95.758
0.7143	-0.7143	98.6471
0.7143	-0.4286	97.5584
0.7143	-0.1429	102.2976
0.7143	0.1429	102.5718
0.7143	0.4286	105.6301
0.7143	0.7143	110.7006
0.7143	1	116.6367
1	-1	93.4607
1	-0.7143	98.5999
1	-0.4286	100.2631
1	-0.1429	105.8061
1	0.1429	104.2504
1	0.4286	109.3508
1	0.7143	113.2479
1	1	117.2012

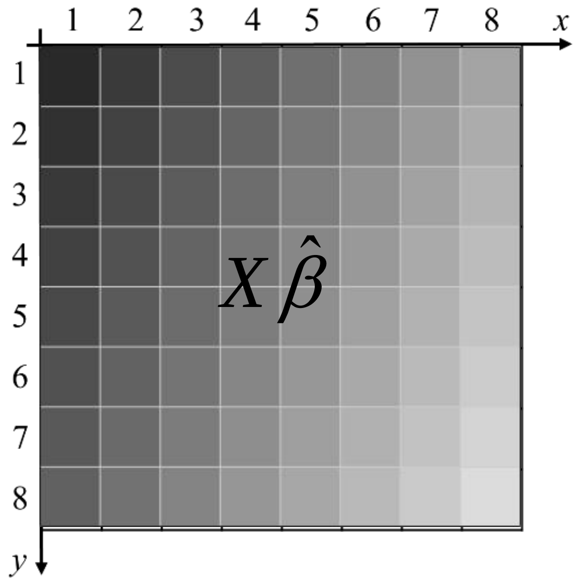
Introduction to Regression Analysis

$$y = X\beta + \varepsilon$$

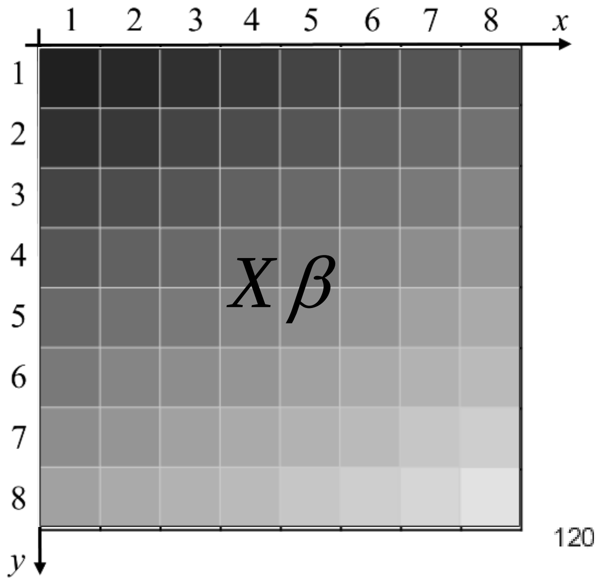
Example: We can use regression to fit a surface to (x,y) data.



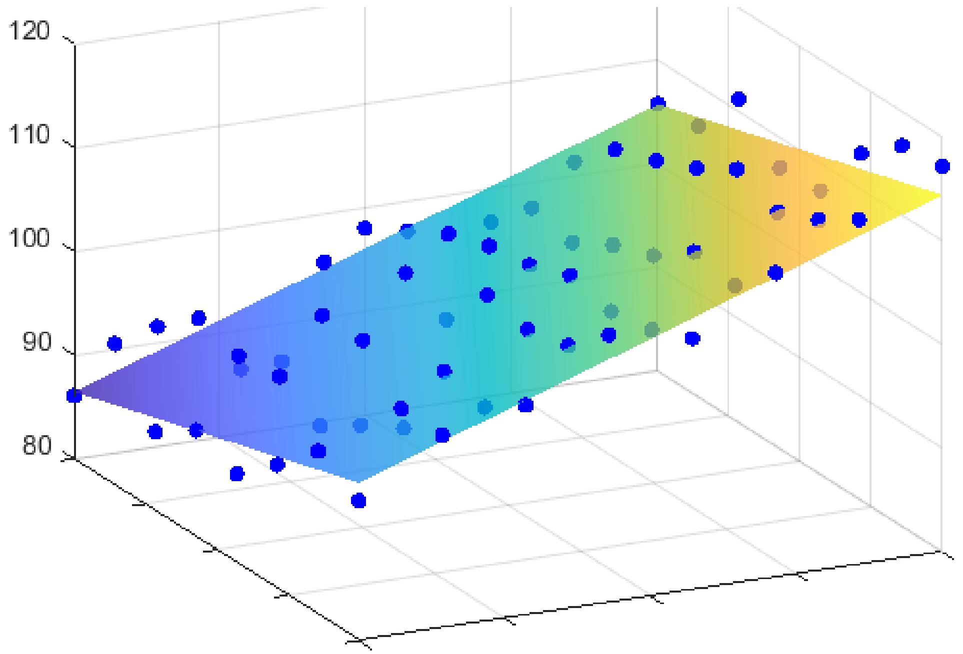
Observed



Estimated



True



$$\hat{\beta} = (X'X)^{-1}X'y$$

Introduction to Regression Analysis

Homework:

Read Chapter 2

Introduction to Regression Analysis

Questions?