GENERAL ASSEMBLY

# INTRO TO DATA SCIENCE
## LINEAR REGRESSION

# I. LINEAR REGRESSION
# II. POLYNOMIAL REGRESSION

# EXERCISES:
# III. LINEAR REGRESSION IN R
# IV. PREDICTING BASEBALL SALARIES

# I. LINEAR REGRESSION

| | *continuous* | *categorical* |
|---|---|---|
| *supervised* | ??? | ??? |
| *unsupervised* | ??? | ??? |

| | *continuous* | *categorical* |
| --- | --- | --- |
| ***supervised*** | regression | classification |
| ***unsupervised*** | dimension reduction | clustering |

Q: What is a **regression** model?

Q: What is a **regression** model?

A: A functional relationship between input & response variables.

Q: What is a **regression** model?
A: A functional relationship between input & response variables

The **simple linear regression** model captures a linear relationship between a single input variable $x$ and a response variable $y$:

Q: What is a **regression** model?
A: A functional relationship between input & response variables

The **simple linear regression** model captures a linear relationship between a single input variable $x$ and a response variable $y$:

$$y = \alpha + \beta x + \varepsilon$$

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

A: $y$ = **response variable** (the one we want to predict)

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

A: $y$ = **response variable** (the one we want to predict)

$x$ = **input variable** (the one we use to train the model)

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

A: $y$ = **response variable** (the one we want to predict)

$x$ = **input variable** (the one we use to train the model)

$\alpha$ = **intercept** (where the line crosses the y-axis)

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

A: $y$ = **response variable** (the one we want to predict)

$x$ = **input variable** (the one we use to train the model)

$\alpha$ = **intercept** (where the line crosses the y-axis)

$\beta$ = **regression coefficient** (the model "parameter")

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

A: $y$ = **response variable** (the one we want to predict)

$x$ = **input variable** (the one we use to train the model)

$\alpha$ = **intercept** (where the line crosses the y-axis)

$\beta$ = **regression coefficient** (the model "parameter")

$\varepsilon$ = **residual** (the prediction error)

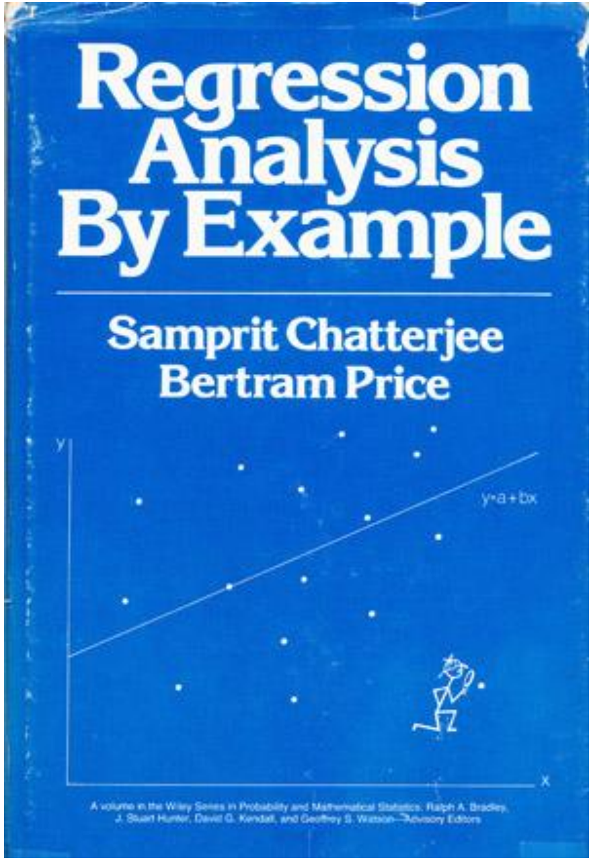
We can extend this model to several input variables, giving us the **multiple linear regression** model:

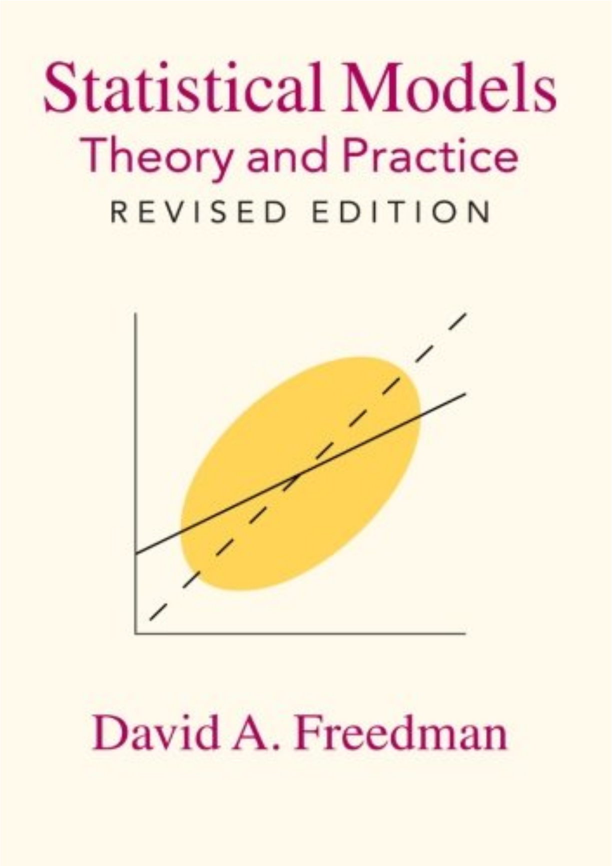
We can extend this model to several input variables, giving us the **multiple linear regression** model:

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$

Linear regression involves several technical assumptions and is often presented with lots of mathematical formality.

The details are not very important for our purposes, but you can check them out if you're interested.


Regression Analysis By Example
Samprit Chatterjee
Bertram Price
y
y=a+bx
x
A volume in the Wiley Series in Probability and Mathematical Statistics, Ralph A. Bradley,
J. Stuart Hunter, David G. Kendall, and Geoffrey S. Watson—Advisory Editors


Statistical Models
Theory and Practice
REVISED EDITION
David A. Freedman

Q: How do we fit a regression model to a dataset?

Q: How do we fit a regression model to a dataset?

A: Minimize the sum of the squared residuals (Ordinary Least Squares).

Q: How do we fit a regression model to a dataset?

A: Minimize the sum of the squared residuals (Ordinary Least Squares).

In practice, any respectable piece of software will do this for you.

Q: How do we fit a regression model to a dataset?
A: Minimize the sum of the squared residuals (Ordinary Least Squares).

In practice, any respectable piece of software will do this for you.

And there are other ways.

Q: How do we fit a regression model to a dataset?
A: Minimize the sum of the squared residuals (Ordinary Least Squares).

In practice, any respectable piece of software will do this for you.

And there are other ways.

And software implements them as well.

# II: POLYNOMIAL REGRESSION

Consider the following **polynomial regression** model:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Consider the following **polynomial regression** model:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Q: This represents a nonlinear relationship. Is it still a linear model?

Consider the following **polynomial regression** model:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Q: This represents a nonlinear relationship. Is it still a linear model?

A: Yes, because it's linear in the $\beta$'s!

Consider the following **polynomial regression** model:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Q: This represents a nonlinear relationship. Is it still a linear model?

A: Yes, because it's linear in the $\beta$'s!

"Although polynomial regression fits a *nonlinear* model to the data, as a statistical estimation problem it is *linear*, in the sense that the regression function E(y|x) is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression." -- Wikipedia

Polynomial regression allows us to fit very complex curves to data.

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$$

Polynomial regression allows us to fit very complex curves to data.

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$$

But there is a problem with the model we've written down so far.

This model displays **collinearity**, which means the predictor variables are highly correlated with each other.

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \dots + \beta_n x^n + \varepsilon$$

```
> x <- seq(1, 10, 0.1)
> cor(x^9, x^10)
[1] 0.9987608
```

This model displays **collinearity**, which means the predictor variables are highly correlated with each other.

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$$

Collinearity causes the linear regression model to “break down”, because it can’t tell the predictor variables apart.

This model displays **collinearity**, which means the predictor variables are highly correlated with each other.

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \ldots + \beta$$

NOTE

For identical features, this results in a *singularity*. We will see an example of this in just a minute!

Collinearity causes the linear regression model to “break down”, because it can’t tell the predictor variables apart.

Q: What can we do about this?

Q: What can we do about this?

A: Replace the correlated predictors with uncorrelated predictors.

Q: What can we do about this?

A: Replace the correlated predictors with uncorrelated predictors.

$$y = \alpha + \beta_1 f_1(x) + \beta_2 f_2(x^2) + \ldots + \beta_n f_n(x^n) + \varepsilon$$

Q: What can we do about this?

A: Replace the correlated predictors with uncorrelated predictors.

$$y = \alpha + \beta_1 f_1(x) + \beta_2 f_2(x^2) + \ldots + \beta_n f_n(x^n) + \varepsilon$$

OPTIONAL NOTE

These polynomial functions form an *orthogonal basis* of the function space.

# EXERCISES