Lecture 11: An intro to linear models in R

Dr. Greg Chism

Machine Learning

- **Supervised:** We are given input samples (X) and output samples (y) of a function $y = f(X)$. We would like to "learn" f, and evaluate it on new data. Types:
  - **Classification:** y is discrete (class labels).
  - **Regression:** y is continuous, e.g. linear regression

Objectives

- Learn different types of regression
- Fit and test assumptions of regression models
- Interpret the meaning of parameters in simple linear models.
- Compare different models
- Think critically about regression models

Regressions are everywhere

Regressions are central to statistics and are part of our daily lives.

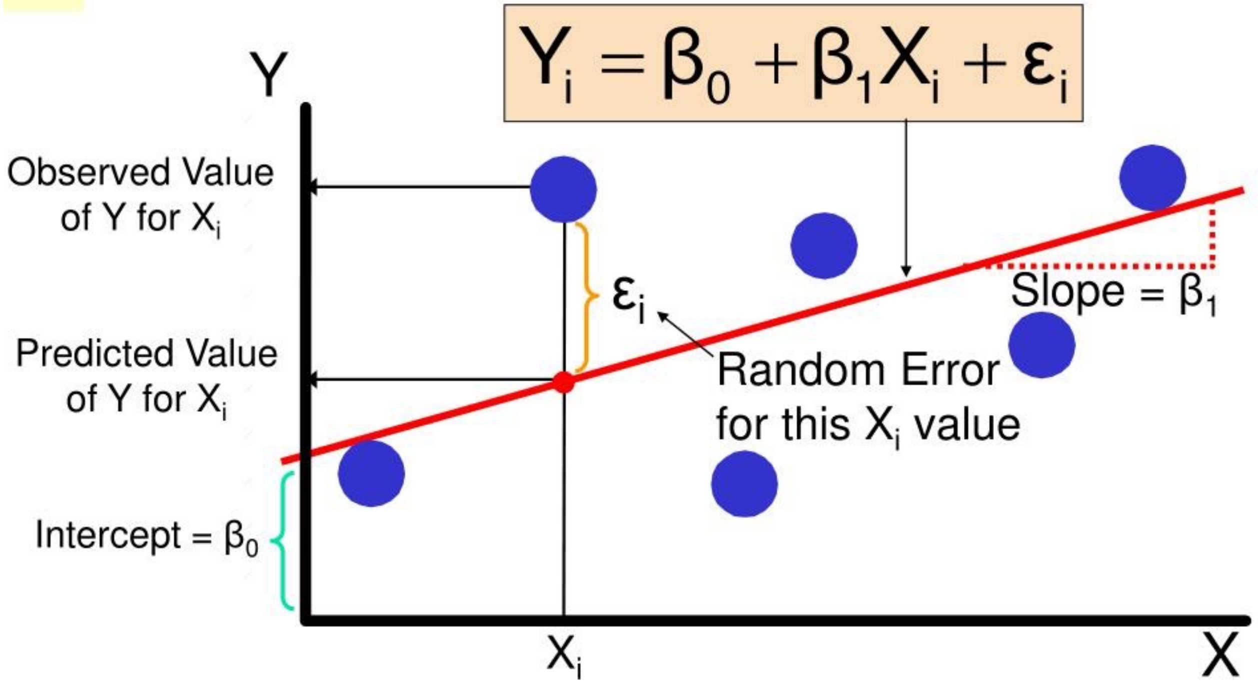
- What is the relationship between surface stream salinity and paved road surface area?
- What factors account for interstate differences in the price of beer?

Part 1. What are regression models?

What are regression models?

Set of methods that are used to predict a response variable from one or more predictor variables

Key terms: dependent and independent variables!



Activity 1. What are regression models?

What are regressions used for?

Today we are going to review specific examples of how to use regression models.

- For now, remember that regressions can be used to:

  identify explanatory variables
  describe the form of the relationships involved
  predicting the response variable from the explanatory variables

Types of regression models?

A few examples*!

- **Simple linear:** Predicting a quantitative response variable from a quantitative explanatory variable.
- **Polynomial:** Predicting a quantitative response variable from a quantitative explanatory variable, where the relationship is modeled as an *n*th order polynomial.
- **Multiple linear:** Predicting a quantitative response variable from two or more explanatory variables.
- **Multilevel:** Predicting a response variable from data that have a hierarchical structure
- **Logistic:** Predicting a categorical response variable from one or more explanatory variables.
- **Poisson:** Predicting a response variable representing counts from one or more explanatory variables.
- …

*There were 205+ regression-related functions implemented in R by 2005!

Part 2. How do we fit regression models in R or Python?

Simple linear model (using R)

In R, the basic function for fitting a linear model is lm(). The format is

model1 <- lm(*formula*, *data*)

We're going to focus on the *formula* component for now!

[Again, there were 205+ regression-related functions implemented in R by 2005!]

Simple linear model (using Python)

In Python, the basic function for fitting a linear model is LinearRegression().

The format is:

```
import numpy as np
from sklearn.linear_model import LinearRegression

model = LinearRegression().fit(x, y)
```

[Linear regression in Python]

Simple linear model (Formula)

Different symbols can be used to indicate alternative aspects within a formula. Below are some of the most important ones!

~ Separates response variables on the left from the explanatory variables on the right.

+ Separates additive predictor variables.

: Denotes an interaction between predictor variables.

***** A shortcut for denoting all possible interactions

Part 3. What are the assumptions of regression models?

Simple linear model (assumptions)

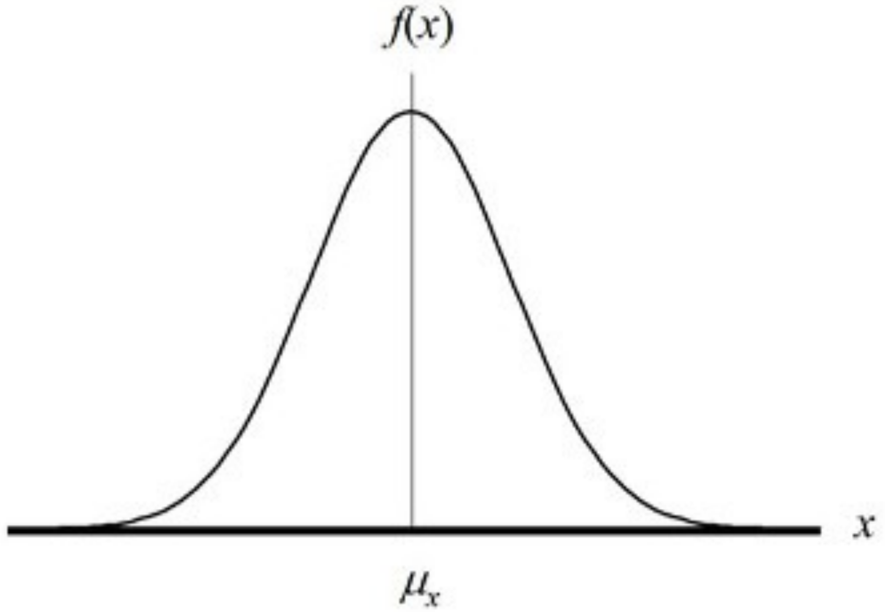
To properly interpret the coefficients of the linear model, specific statistical assumptions must be met:

Normality
Independence
Linearity
Homoscedasticity

Activity 3. What are their assumptions?

Simple linear model (assumptions)

- ***Normality***— For fixed values of the independent variables, the dependent variable is normally distributed.



Activity 3. What are their assumptions?

Simple linear model (assumptions)

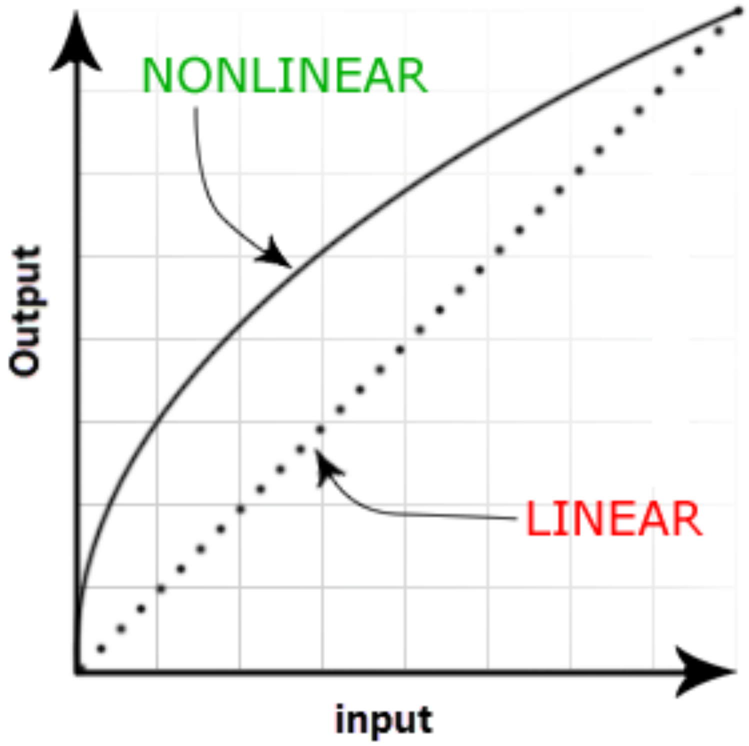
- ***Independence***— The Y_i values are independent of each other.

  It could be understood as a conceptual assumption. However, different tests have been proposed! See the markdown file.

Activity 3. What are their assumptions?

Simple linear model (assumptions)

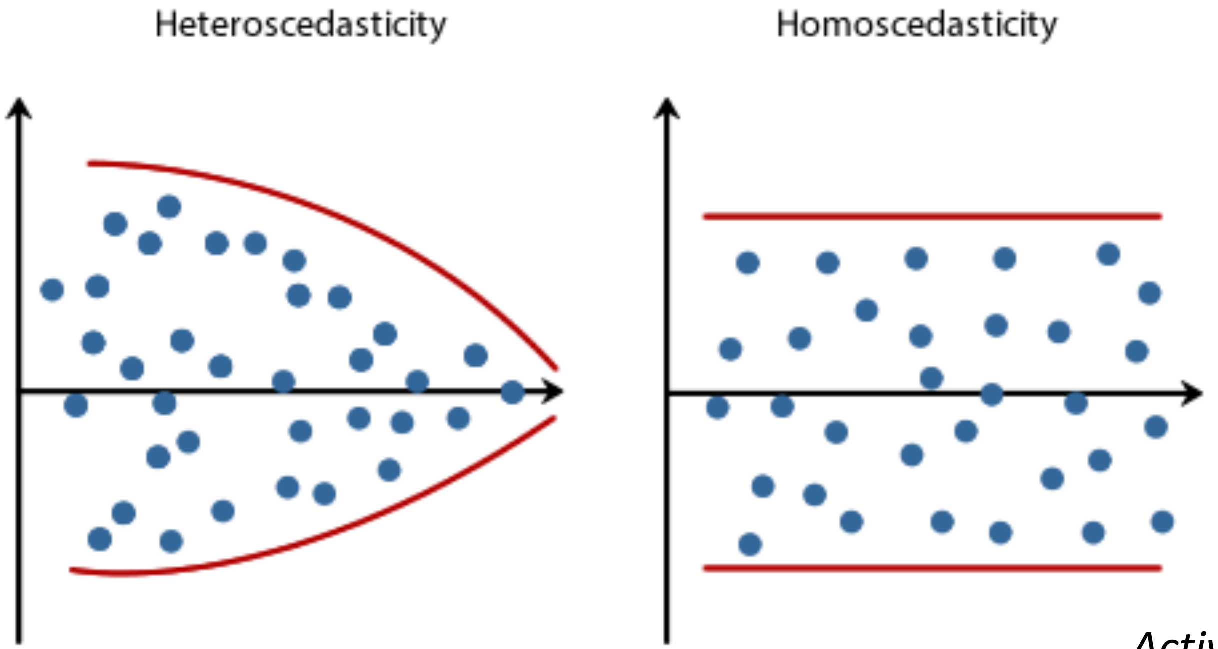
- ***Linearity***— The dependent variable is linearly related to the independent variables.



Activity 3. What are their assumptions?

Simple linear model (assumptions)

- ***Homoscedasticity*****—** The variance of the dependent variable doesn’t vary with the levels of the independent variables. (I could call this constant *variance*, but saying *homoscedasticity* makes me feel smarter.)



Activity 3. What are their assumptions?

Part 4. How to interpret regression models?

Simple linear model

Let's discuss its mathematical structure:

$$\widehat{Y}_i = \widehat{\beta_0} + \widehat{\beta_1} X_{li} + \cdots + \widehat{\beta_k} X_{ki} \quad i = 1 \ldots n$$

$\hat{Y}_i$ is the predicted value of the dependent variable for observation i
$\widehat{\beta_0}$ is the intercept
$\widehat{\beta_k}$ is the regression coefficient for the jth predictor

n is the number of observations
k is the number of predictor variables

Simple linear model (model parameters)

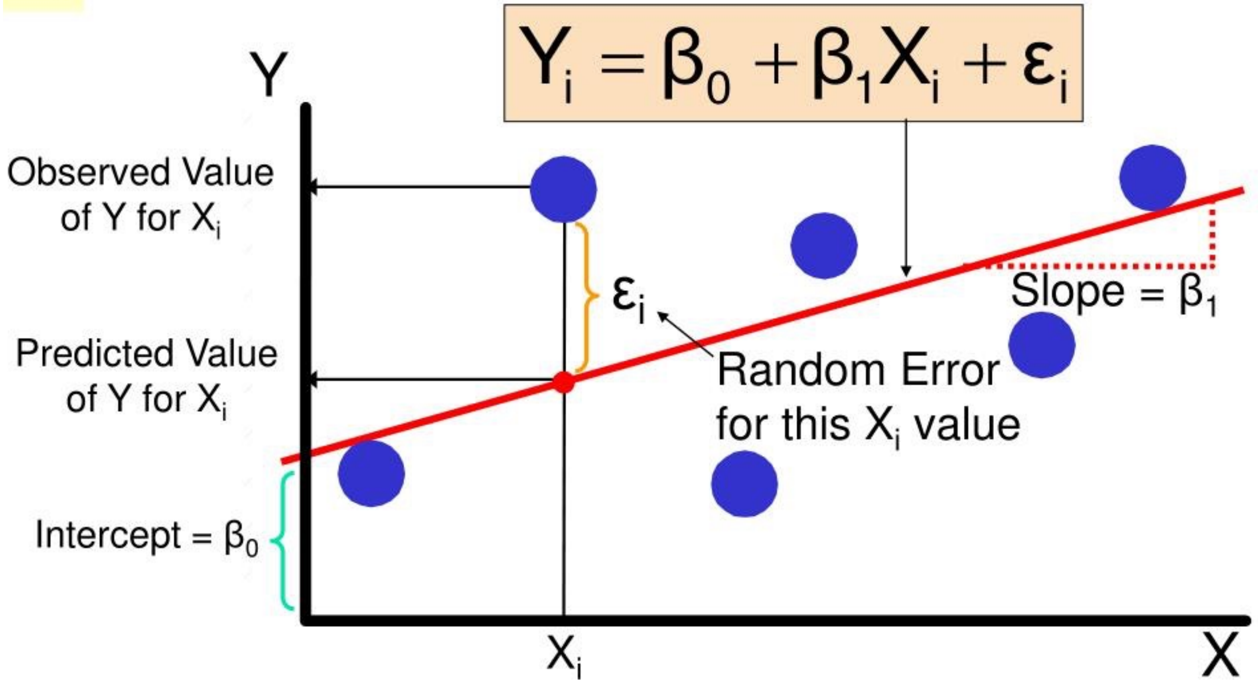
We are going to focus on four main regression parameters

Slope
Intercept
R^2
P-values

We will review these concepts using a simple linear regression

Simple linear model (model parameters)

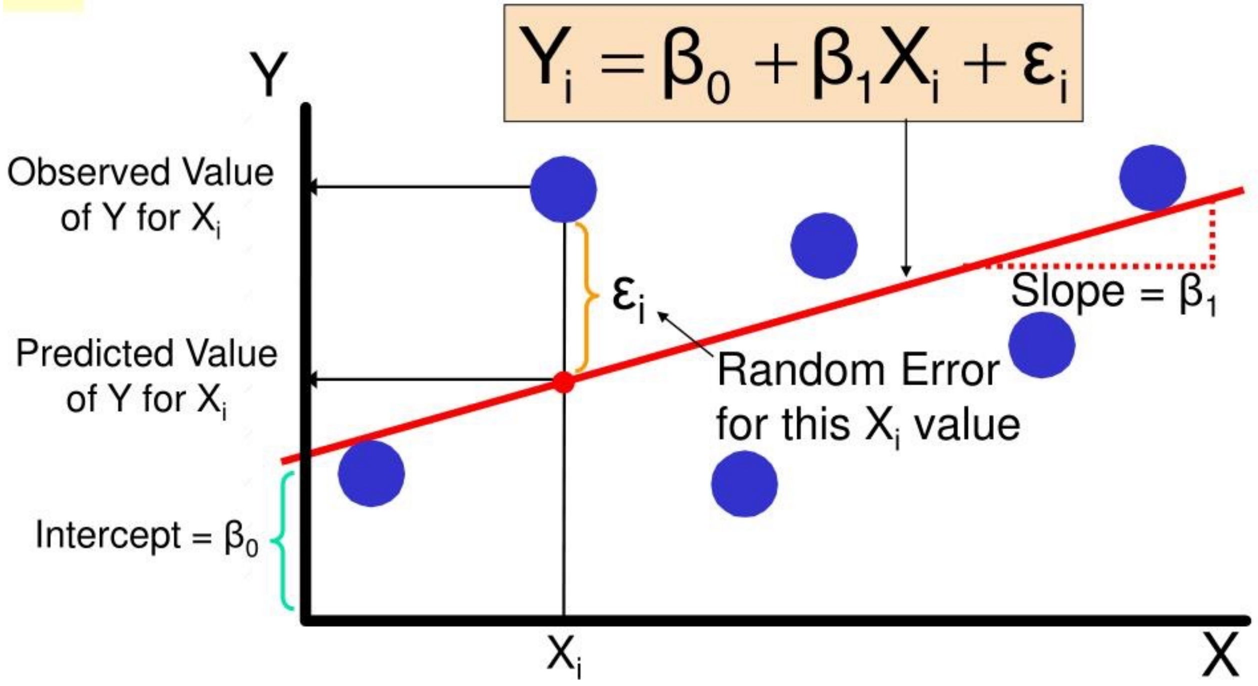
- **Slope:** Change in Y units given a change in X of a single unit.



Activity 4. Interpreting simple regression models?

Simple linear model (model parameters)

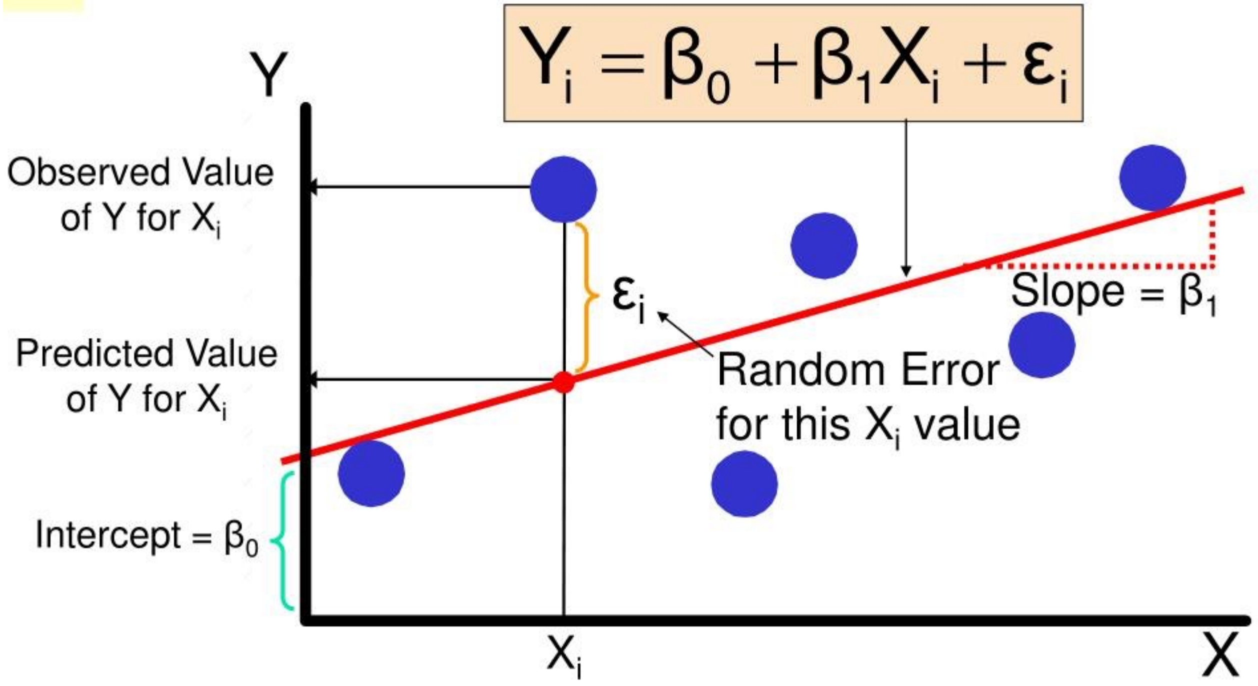
- **Intercept**: Adjustment constant. The Y value when X, the predictor is 0.



Activity 4. Interpreting simple regression models?

Simple linear model (model parameters)

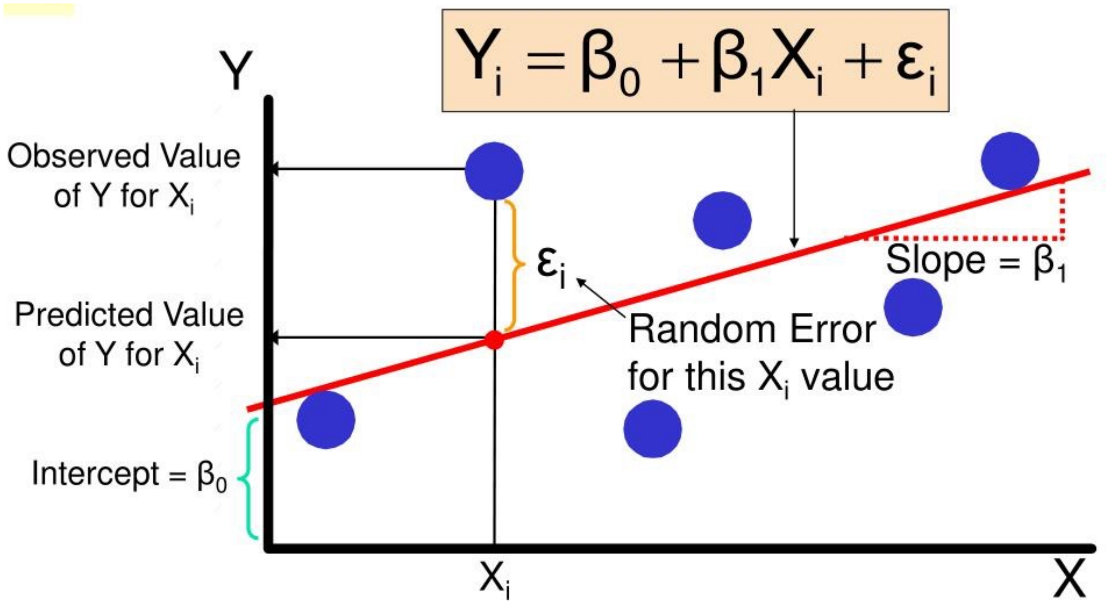
- **R^2:** The fraction of variance in Y that is explained by X.



Activity 4. Interpreting simple regression models?

Simple linear model (model parameters)

- **P-values:** Indicates whether the slope is significantly different from zero.



Activity 4. Interpreting simple regression models?

Part 5. Selecting among competing models!

Selecting among competing models

We will review a few commonly used methods!

AIC
Stepwise regression
All subset regression

Selecting among competing models

AIC (Akaike Information Criterion)

AIC = 2k – 2 ln(L), where k is the number of parameters and L the likelihood.

Index that takes into account model complexity (number of parameters) and fit. The lowest the AIC score, the better!

Selecting among competing models

Stepwise regression

Y might depend on many variables! Which combination of variables explains better Y?

Two alternative approaches. First, in a ”forward” approach, variables are added until no improvement is noted. In a “backward” approach, variables reducing model quality are deleted from a full model.

Selecting among competing models

All subset regression

Exhaustive approach that is likely only useful when the number of variable combinations is reduced. All the possible models are examined.

Similar to stepwise regressions but instead of examining predictor combinations in an alternative fashion, this approach examines *all* possible combinations!