Stat 340
Applied Regression
Class Overview

Learning Task 1: Regression

For given values of covariates/features/explanatory variables $X_1, \ldots, X_p$, what is the expected value of the quantitative response variable Y?

- X_1 = Years of Education, X_2 = Seniority
- Y = Income

Goal: $Y = f(X_1, X_2) + \varepsilon$

Learning Task 1: Regression

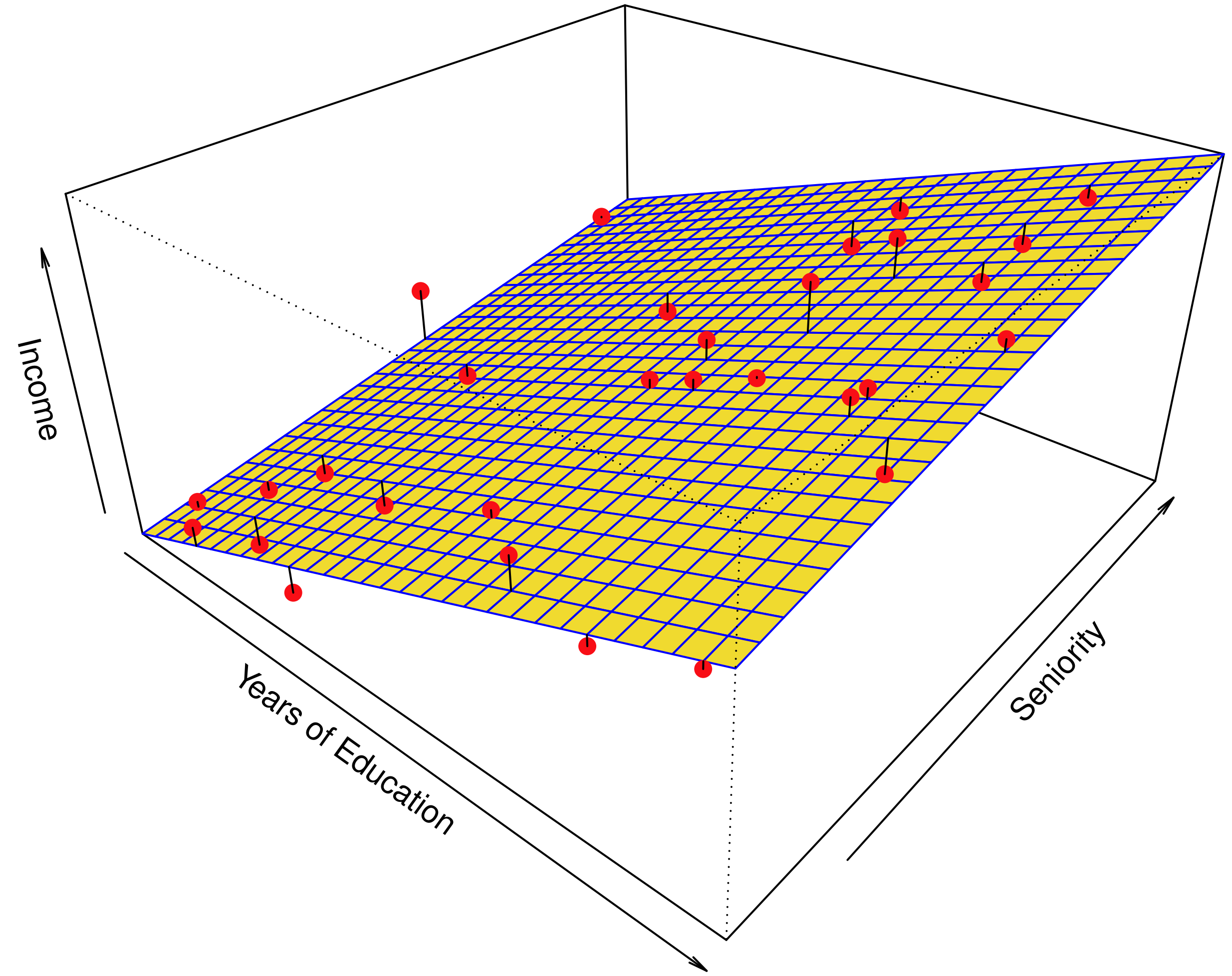
For given values of covariates/features/explanatory variables $X_1, \ldots, X_p$, what is the expected value of the quantitative response variable Y?

- X_1 = Years of Education, X_2 = Seniority
- Y = Income

Goal: $Y = f(X_1, X_2) + \varepsilon$

Approach 1:
Multiple Linear Regression

$$f(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$



Learning Task 1: Regression

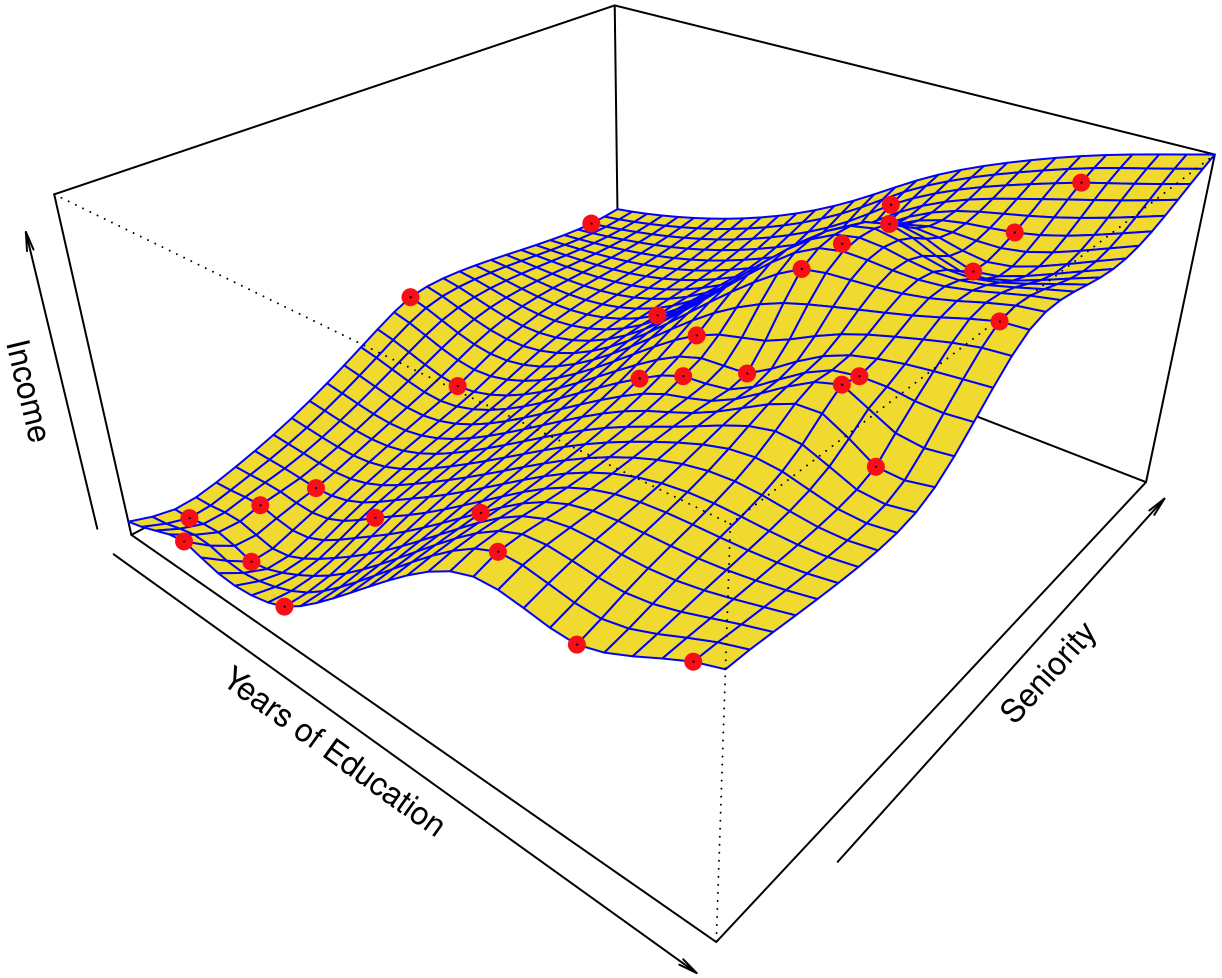
For given values of covariates/features/explanatory variables $X_1, \ldots, X_p$, what is the expected value of the quantitative response variable Y?

- X_1 = Years of Education, X_2 = Seniority
- Y = Income

Goal: $Y = f(X_1, X_2) + \varepsilon$

Additional Approaches:

- Splines (pictured right)
- Regression Trees
- K Nearest Neighbors
- Generalized Additive Models



Learning Task 2: Classification

For given values of covariates/features/explanatory variables $X_1, \ldots, X_p$, what is the probability that a categorical response variable Y falls in category k?

Learning Task 2: Classification

For given values of covariates/features/explanatory variables $X_1, \ldots, X_p$, , what is the probability that a categorical response variable Y falls in category k?

- $X_1, \ldots, X_{57}$ = Summaries of contents of email messages (e.g., percent of words in message that are "business", average length of sequences of capital letters, …)
- Y = Indicator of whether a given email was spam.

Goal: $P(Y = k \,|\, X_1, X_2) = f_k(X_1, X_2)$ for each category k

Learning Task 2: Classification

For given values of covariates/features/explanatory variables $X_1, \ldots, X_p$, , what is the probability that a categorical response variable Y falls in category k?

- $X_1, \ldots, X_{57}$ = Summaries of contents of email messages (e.g., percent of words in message that are "business", average length of sequences of capital letters, …)
- Y = Indicator of whether a given email was spam.

Goal: $P(Y = k \,|\, X_1, X_2) = f_k(X_1, X_2)$ for each category k

Approach 1:
Logistic Regression

$$f(X_1, \ldots, X_{57}) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_{57} X_{57})}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_{57} X_{57})}$$

Learning Task 2: Classification

For given values of covariates/features/explanatory variables $X_1, \ldots, X_p$, what is the probability that a categorical response variable Y falls in category k?

- $X_1, \ldots, X_{57}$ = Summaries of contents of email messages (e.g., percent of words in message that are "business", average length of sequences of capital letters, …)
- Y = Indicator of whether a given email was spam.

Goal: $P(Y = k \mid X_1, X_2) = f_k(X_1, X_2)$ for each category k

Additional Approaches:

- Linear Discriminant Analysis
- Classification Trees
- K Nearest Neighbors
- Generalized Additive Models

Goals for this Class

1. Understand statistical models listed on previous slides
2. Perform model estimation using implementations in R packages
3. Visualize data and model outputs in R
4. Compare/evaluate different models graphically and quantitatively
5. Collaborative coding and writing with Git and GitHub
6. Generally - what to do when presented with a new data set?

Prerequisites

Stat 242 (Intermediate Statistics):

- Multiple Linear Regression
    - Estimating and interpreting linear regression models with multiple explanatory variables
    - Confidence intervals for parameters, interpretations
    - Hypothesis tests (F and t) about parameters
    - **Maybe:** a first look at polynomial regression
    - **Maybe:** data transformations with log and square root
- Some familiarity with R

Math 211 (Linear Algebra)

- Matrix Multiplication
- Bases, Linear Span, Column Space
- Orthogonal Projections
- Diagonalization/spectral theorem (briefly, later in the semester, not by hand)

Some of the figures in this presentation are taken from
"An Introduction to Statistical Learning, with applications in R" (Springer, 2013)
with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani