Lecture 5: Multiple Linear Regression

Module 2: part 1

Spring 2024

Logistics

- Start of Module 2 (3 lectures total)
- Assessment for Module 1 is due 11:59pm on Feb 11 (Wed)
- See Canvas Announcement (ask TAs if any question)

Linear Regression

In Module 1, we discussed **simple linear regression**, the setting where we observe two variables:

- One dependant variable (predicted variable): Y_i
- One independent variable (predictor variable, covariate, regressor): X_i

Linear Regression

In Module 1, we discussed **simple linear regression**, the setting where we observe two variables:

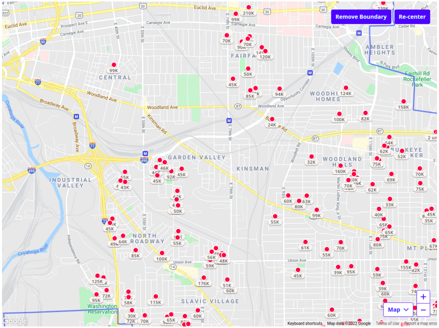
- One dependant variable (predicted variable): Y_i
- One independent variable (predictor variable, covariate, regressor): X_i

In Module 2, we will consider **Multiple Linear Regression**, the setting where we have:

- Multiple independent variables (predictor variable, covariate, regressor): X_i
- Allows for better predictive power
- Allows for more flexible, richer models
- Allows to "adjust" for other variables

Example: Housing prices

Data contains the sale price of 522 houses in a Midwestern city in 2002[1].



- Y_i is sale price of home
- What covariates would you use to predict or model the price of a home?

[1] Dataset from 'Applied Linear Statistical Models' by Kutner, Nachtsheim, Neter, and Li

Example: Housing prices

In this data set we have recorded: Square footage, bedrooms, bathrooms, AC, garage, pool, Age of home, quality, lot size, home style

$$\widehat{\text{Home Price}}_i = \hat{b}_0 + \hat{b}_1 \text{Sq Ft}_i$$

$$R^2 = .67$$

Example: Housing prices

In this data set we have recorded: Square footage, bedrooms, bathrooms, AC, garage, pool, Age of home, quality, lot size, home style

$$\widehat{\text{Home Price}}_i = \hat{b}_0 + \hat{b}_1 \text{Beds}_i$$

$$R^2 = .17$$

Example: Housing prices

In this data set we have recorded: Square footage, bedrooms, bathrooms, AC, garage, pool, Age of home, quality, lot size, home style

$$\widehat{\text{Home Price}}_i = \hat{b}_0 + \hat{b}_1 \text{Baths}_i$$

$$R^2 = .47$$

Example: Housing prices

In this data set we have recorded: Square footage, bedrooms, bathrooms, AC, garage, pool, Age of home, quality, lot size, home style

$$\widehat{\text{Home Price}}_i = \hat{b}_0 + \hat{b}_1 \text{Sq Ft}_i + \hat{b}_2 \text{Beds}_i + \hat{b}_3 \text{Baths}_i$$

$$R^2 = .69$$

Multiple Linear Regression

We will predict $\hat{y}_i$ using p different covariates

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i,1} + \hat{b}_2 x_{i,2} \dots \hat{b}_p x_{i,p} = \hat{b}_0 + \sum_{j=1}^{p} \hat{b}_j x_{i,j}$$

Notation:

- We will typically use bold face to denote vectors
- Observations will typically be $i = 1, \dots, n$ and covariates will be $j = 1, \dots, p$
- $x_{i,j}$ denotes the value of the jth covariate for the ith observation
- The covariates for the ith observation: $\mathbf{X_i} = \{X_{i,1}, X_{i,2}, \dots, X_{i,p}\}$
- $\mathbf{X}$ table (or matrix) where each row is one observation and each column is a covariate
- $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_n\}$
- Vector of linear coefficients : $\hat{\mathbf{b}} = \{\hat{b}_0, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_p\}$

Multiple Linear Regression

We will predict $\hat{y}_i$ using p different covariates

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i,1} + \hat{b}_2 x_{i,2} \ldots \hat{b}_p x_{i,p} = \hat{b}_0 + \sum_{j=1}^{p} \hat{b}_j x_{i,j}$$

Multiple Linear Regression

We will predict $\hat{y}_i$ using p different covariates

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i,1} + \hat{b}_2 x_{i,2} \ldots \hat{b}_p x_{i,p} = \hat{b}_0 + \sum_{j=1}^{p} \hat{b}_j x_{i,j}$$

Select $\hat{\mathbf{b}}$ by minimizing the residual sum of squares:

$$RSS(\hat{\mathbf{b}}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\hat{b}_0 + \sum_{j=1}^{p} \hat{b}_j x_{i,j})]^2$$

Potentially helpful but not necessary math

In matrix vector notation,

$$RSS(\hat{\mathbf{b}}) = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})$$

so to minimize this quantity, we take the derivative and solve for 0.
Taking the derivative with respect to vectors is a bit more complex, but intuitively similar

$$\frac{\partial RSS(\hat{\mathbf{b}})}{\partial \hat{\mathbf{b}}} = -2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})$$

$$0 = -2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})$$

$$0 = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\hat{\mathbf{b}}$$

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \approx \frac{cov(X, Y)}{var(X)}$$

Multiple Linear Regression

The population model we are trying to recover is

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

Multiple Linear Regression

The population model we are trying to recover is

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

Interpretation:

- b_0 is the expected value of Y_i when **all** observed covariates are 0
- For $k \neq 0$, b_k is the difference in the expected value of Y_i and Y_j when $x_{i,k}$ and $x_{j,k}$ differ by 1 unit (i.e., $x_{i,k} - x_{j,k} = 1$), **but** the value of all other observed covariates are the same (holding all the other $x_{i,l}$ constant).

Example

In the housing example:

- In simple linear regression, the coefficient of Beds captures association of an additional bedroom (which may also be associated with additional square footage)
- In multiple linear regression, the coefficient of Beds captures association of an additional bedroom (when Sq footage stays the same)

Example

Simple Linear Regression: $E(\text{Home Price}_i \mid \text{Sq Ft, Beds, Baths}) = b_0 + b_1 \text{Beds}_i$,

$$\hat{b}_1 = 56,200$$

If House 1 has two bedroom and House 2 has three bedrooms, we would expect House 2 to be $56,200$ more expensive than House 1.

Example

Simple Linear Regression: $E(\text{Home Price}_i \mid \text{Sq Ft, Beds, Baths}) = b_0 + b_1 \text{Beds}_i$,

$$\hat{b}_1 = 56,200$$

If House 1 has two bedroom and House 2 has three bedrooms, we would expect House 2 to be $56,200$ more expensive than House 1.

Multiple Linear Regression:
$E(\text{Home Price}_i \mid \text{Sq Ft, Beds, Baths}) = b_0 + b_1 \text{Sq Ft}_i + b_2 \text{Beds}_i + b_3 \text{Baths}_i$,

$$\hat{b}_1 = 143; \hat{b}_2 = -14,786$$

If House 1 has two bedroom and House 2 has three bedrooms but the two houses have the same Sq Footage and the same number of bathrooms, we would expect House 2 to be $14,786$ less expensive than House 1.

Example: Productivity, Coffee, and Caffeine

In the example:

$$E(\text{Productivity}_i \mid \text{Coffee}) = b_0 + b_1 \text{Coffee}_i$$

$$E(\text{Productivity}_i \mid \text{Caffeine}) = b_0 + b_1 \text{Caffeine}_i$$

Example: Productivity, Coffee, and Caffeine

In the example:

$$E(\text{Productivity}_i \mid \text{Coffee}) = b_0 + b_1 \text{Coffee}_i$$

$$E(\text{Productivity}_i \mid \text{Caffeine}) = b_0 + b_1 \text{Caffeine}_i$$

$$E(\text{Productivity}_i \mid \text{Coffee, Caffeine}) = b_0 + b_1 \text{Coffee}_i + b_2 \text{Caffeine}_i$$

Interpreting Coefficients

- Each coefficient captures the association of a single covariate when all other covariates are fixed
- The coefficient (in the population model and the estimated coefficients) will change depending on what other covariates are included
- The size of coefficients can only be compared with respect to the units of the covariates

  e.g., coefficient of Sq Ft has a much smaller magnitude than the coefficient of Beds because 1 additional sq ft is very different than 1 additional bedroom

Interpreting Coefficients

- Each coefficient captures the association of a single covariate when all other covariates are fixed
- The coefficient (in the population model and the estimated coefficients) will change depending on what other covariates are included
- The size of coefficients can only be compared with respect to the units of the covariates

  e.g., coefficient of Sq Ft has a much smaller magnitude than the coefficient of Beds because 1 additional sq ft is very different than 1 additional bedroom
- Discuss a problem in your field where you are interested in measuring association between a covariate and an outcome when holding other covariates fixed

Specific types of covariates

Polynomial regression

The big assumption in linear regression is the conditional expectation of Y is linear in the covariates

$$E(Y \mid X = x) = b_0 + b_1 x$$

But what if the conditional expectation is not linear in x?

$$E(Y \mid X = x) = b_0 + b_1 x + b_2 x^2$$

Polynomial regression

The big assumption in linear regression is the conditional expectation of Y is linear in the covariates

$$E(Y \mid X = x) = b_0 + b_1 x$$

But what if the conditional expectation is not linear in x?

$$E(Y \mid X = x) = b_0 + b_1 x + b_2 x^2$$

We can always include additional terms and estimate:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i + \hat{b}_2 x_i^2$$

- Covariate 1 is x_i; Covariate 2 is just the square of the first covariate

Polynomial regression

- The "degree" is the largest exponent in a polynomial
- You can fit a higher degree polynomial if your data is large enough (bias variance trade-off)
- The model is still *linear* in the covariates (the covariates just happen to be non-linear terms of x_i)
- In practice, because x_i and $x_i^2, x_i^3, \ldots$ can be very correlated and the higher order terms can be large, rescaling the higher order terms is very helpful

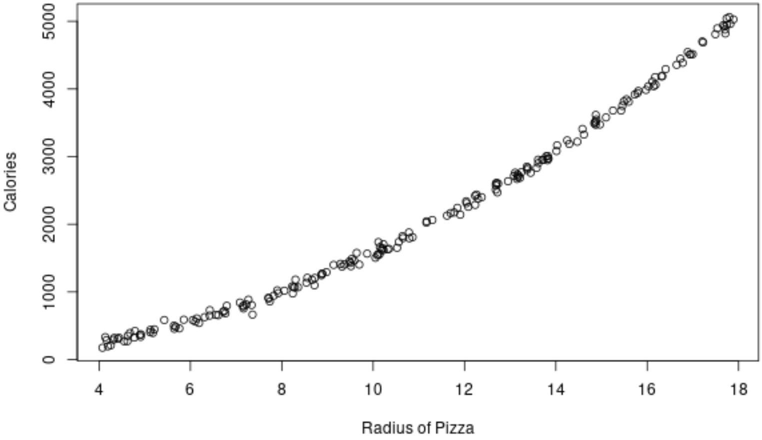
Polynomial regression

Example: Suppose I'm interested in predicting the number of calories in a pizza based on the radius of the pizza

$$\widehat{\text{Calories}}_i = \hat{b}_0 + \hat{b}_1 \text{Radius of Pizza}$$

Polynomial regression

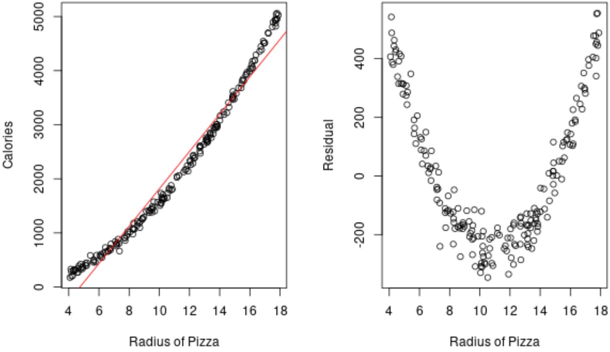
Example: Suppose I'm interested in predicting the number of calories in a pizza based on the radius of the pizza



Polynomial regression

Example: Suppose I'm interested in predicting the number of calories in a pizza based on the radius of the pizza

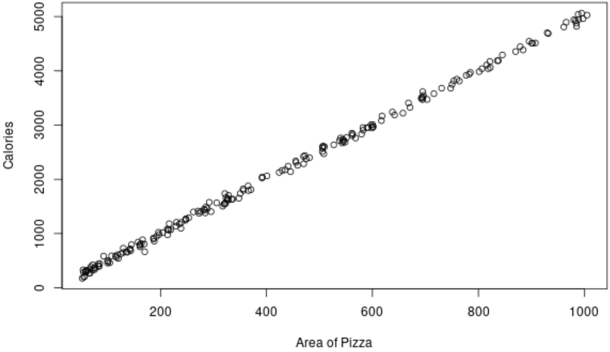
$$\widehat{\text{Calories}}_i = -1.6 + 344.2 \times \text{Radius of Pizza}$$



Polynomial regression

Example: Suppose I'm interested in predicting the number of calories in a pizza based on the radius of the pizza

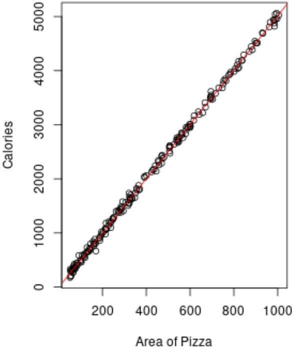
$$\widehat{\text{Calories}}_i = -2.5 + 5.0 \times \text{Area of Pizza}$$

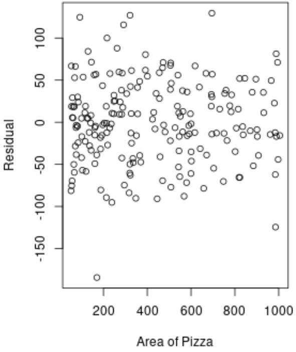


Polynomial regression

Example: Suppose I'm interested in predicting the number of calories in a pizza based on the radius of the pizza

$$\widehat{\text{Calories}}_i = -2.5 + 5.0 \times \text{Area of Pizza}$$

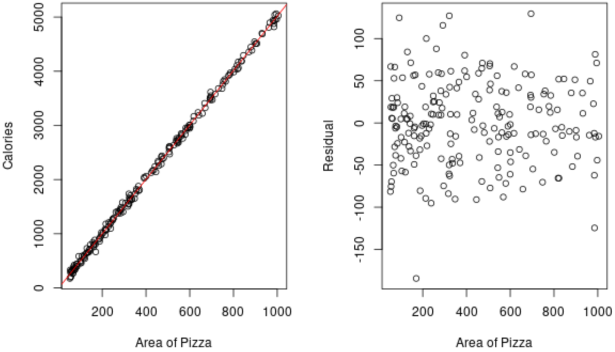




Polynomial regression

Example: Suppose I'm interested in predicting the number of calories in a pizza based on the radius of the pizza

$$\widehat{\text{Calories}}_i = -2.5 + 5.0 \times \text{Area of Pizza} = -2.5 + 0 \times \text{Radius} + 5.0 \times \pi\text{Radius}^2$$



Polynomial regression

Interpretation of coefficients in polynomial regression is more complicated

$$E(Y_i \mid X_i = x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3$$

- **Incorrect**: A 1 unit increase in x is associated with a b_1 increase in Y when holding x^2, x^3, ... constant
- **Correct**: A change of x from 4 to 5 is associated with a

$$[b_1(5) + b_2(5)^2 + b_3(5)^3] - [b_1(4) + b_2(4)^2 2 + b_3(4)^3]$$

  increase in Y
- We must account for the fact that changing x also changes x^2 and x^3 (you CANNOT change x while keeping x^2 constant).
- Association of Y and X not constant everywhere, but depends on specific value of $X = x$

Wrap up

- We can model the conditional expectation of Y with multiple covariates
- Fit coefficients by minimizing the residual sum of squares
- Each coefficient describes the association between covariate and Y when holding all other covariates fixed
- Can include covariates which are polynomials of other covariates
- Lab will consider modeling home prices and predicting Brexit votes