Regression Refresher

Administrivia

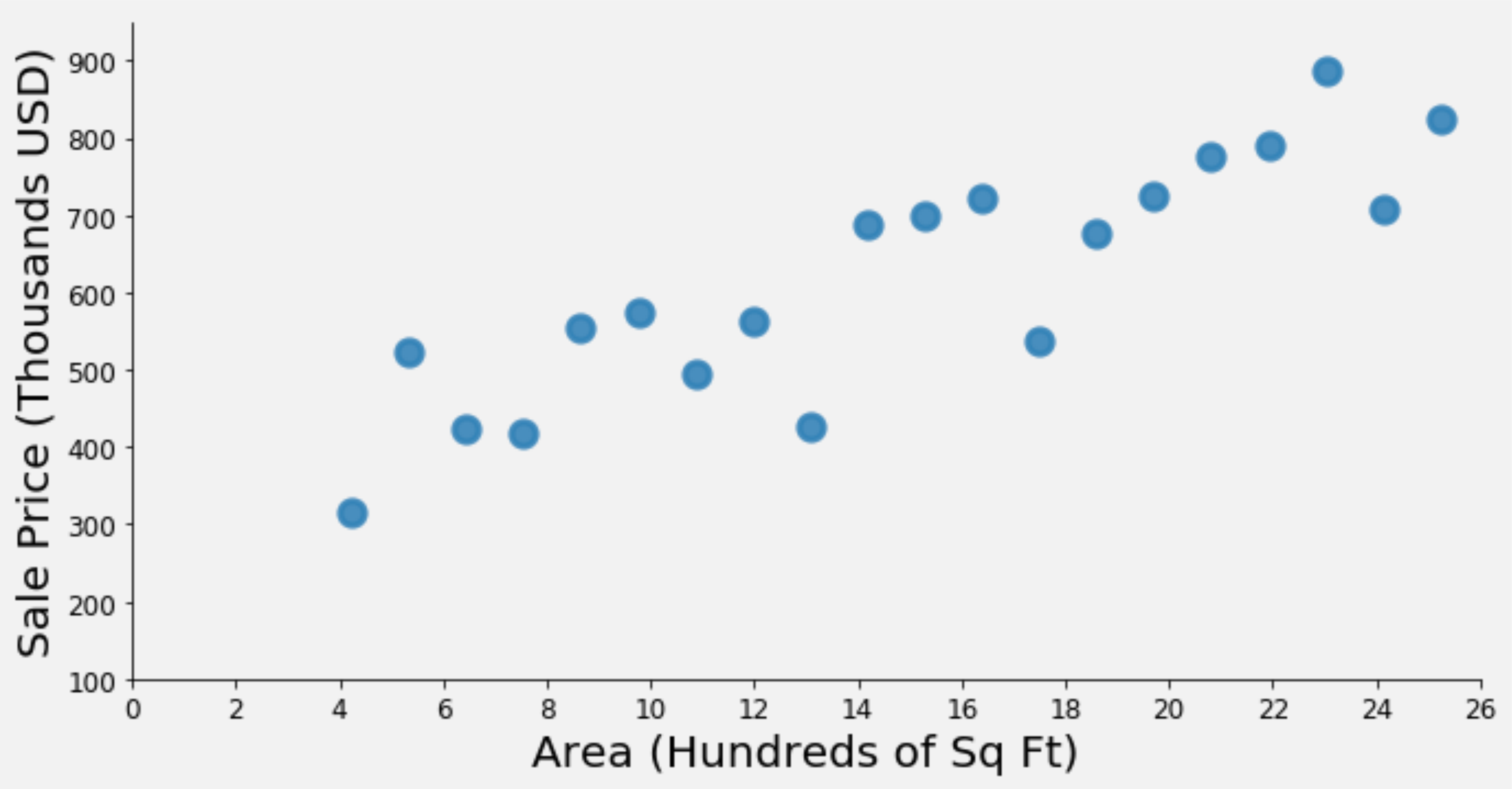
- Make sure you enroll in Moodle soon.
    - Enrollment key on Piazza
    - First Reading Quiz is posted. Due before class on Monday
    - If joined course after Wednesday, email me to get added to Piazza

Linear Regression for Prediction

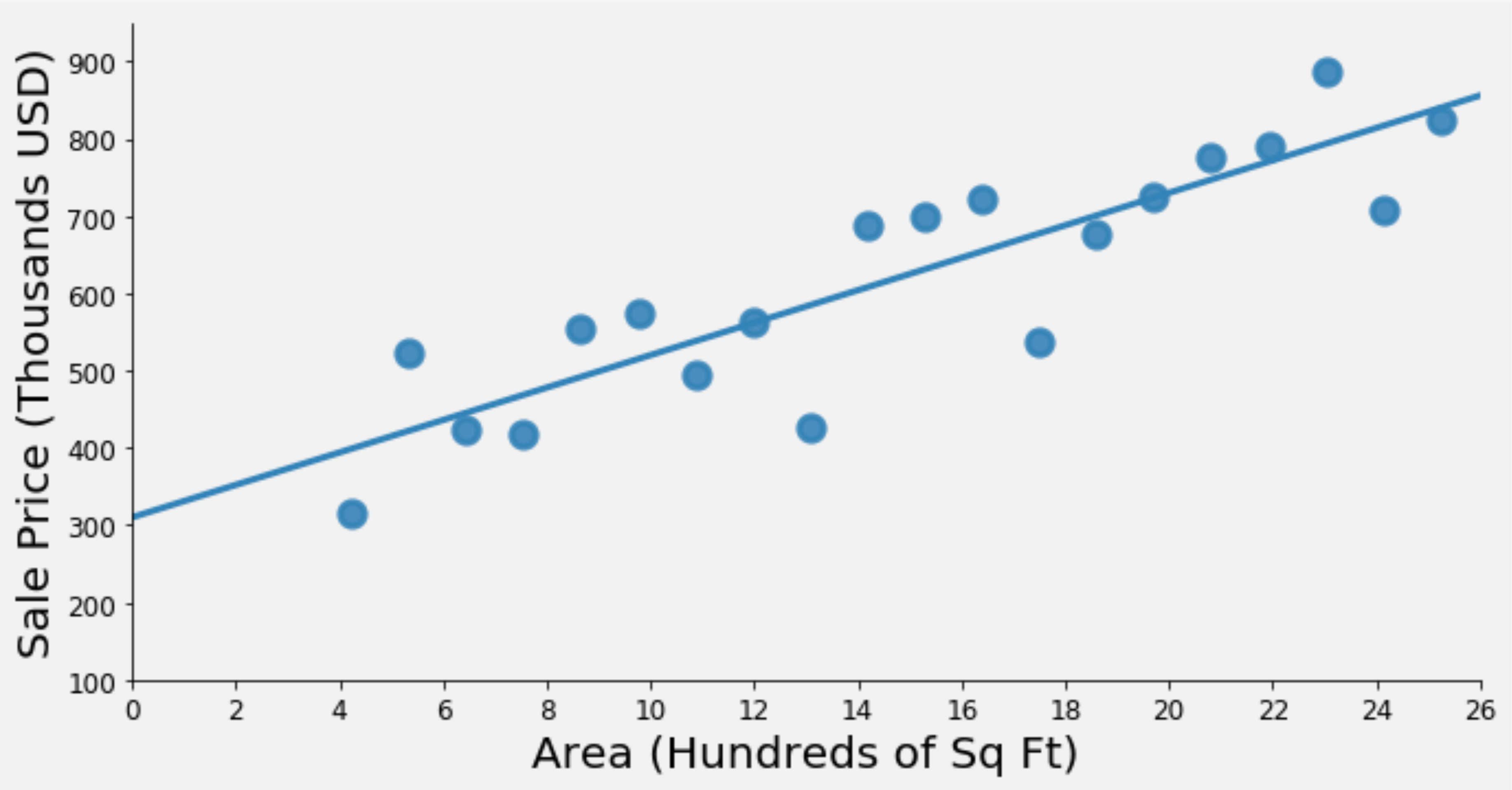
Examples:

- Given a person's age and gender, predict their height
- **Given the area of a house, predict its sale price**
- Given unemployment, inflation, number of wars, and economic growth, predict the president's approval rating.
- Given a person's browser history, predict how long they'll stay on a product page
- Given the advertising budget expenditures in various media markets, predict the number of products a company will sell

Area as Predictor for House Price



Area as Predictor for House Price



Simple Linear Regression

- In this case we have one feature $X =$ AREA

- The response is $Y =$ SALE PRICE

- We **assume** that the true relationship is given by

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Assume we've collected data about n houses. Each house example satisfies

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Simple Linear Regression

Assumptions of the Simple Linear Regression:

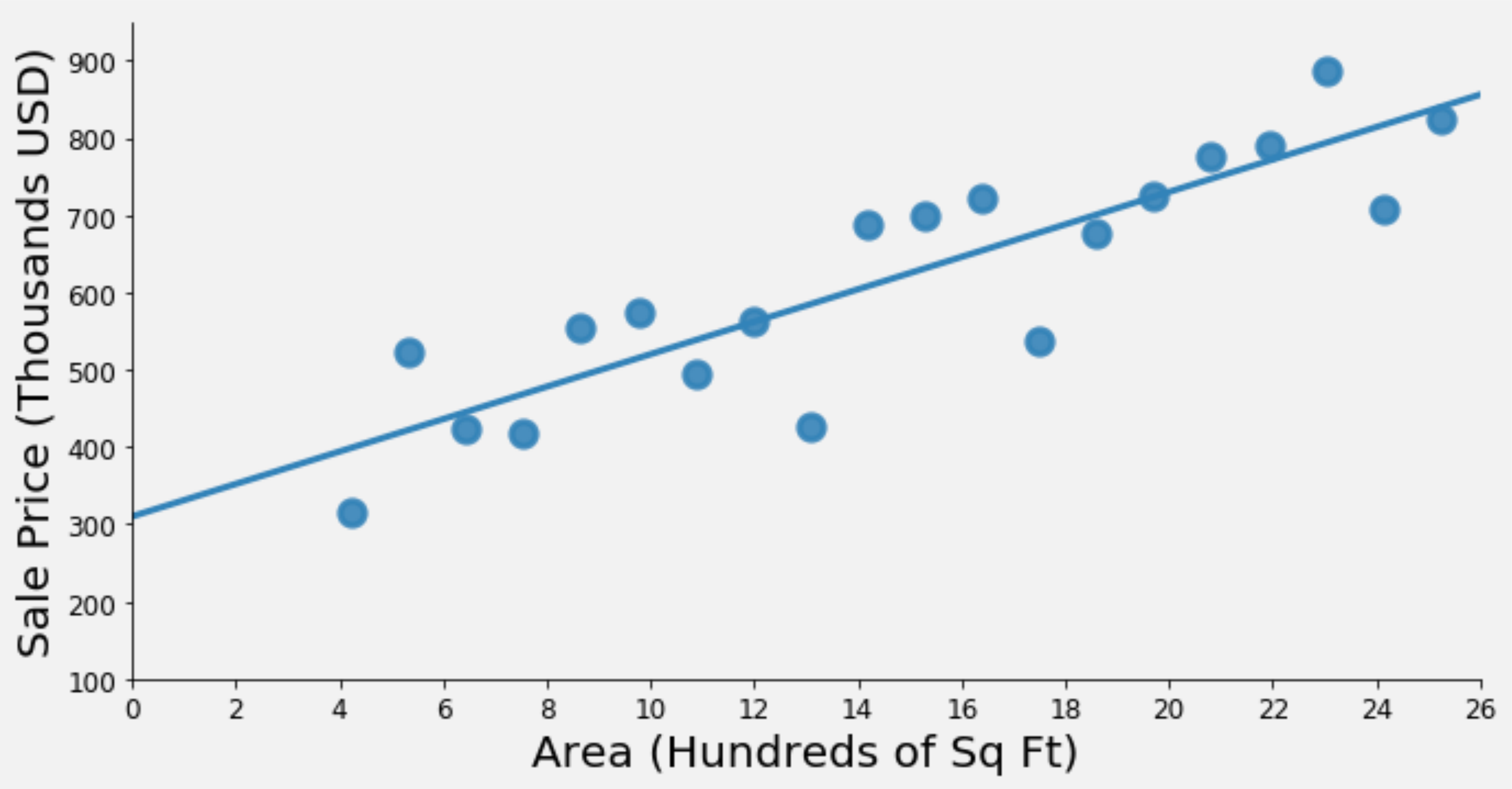
1. $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$

2. ϵ_i's ARE INDEPENDENT

3. $\epsilon_i \sim N(0, \sigma^2)$

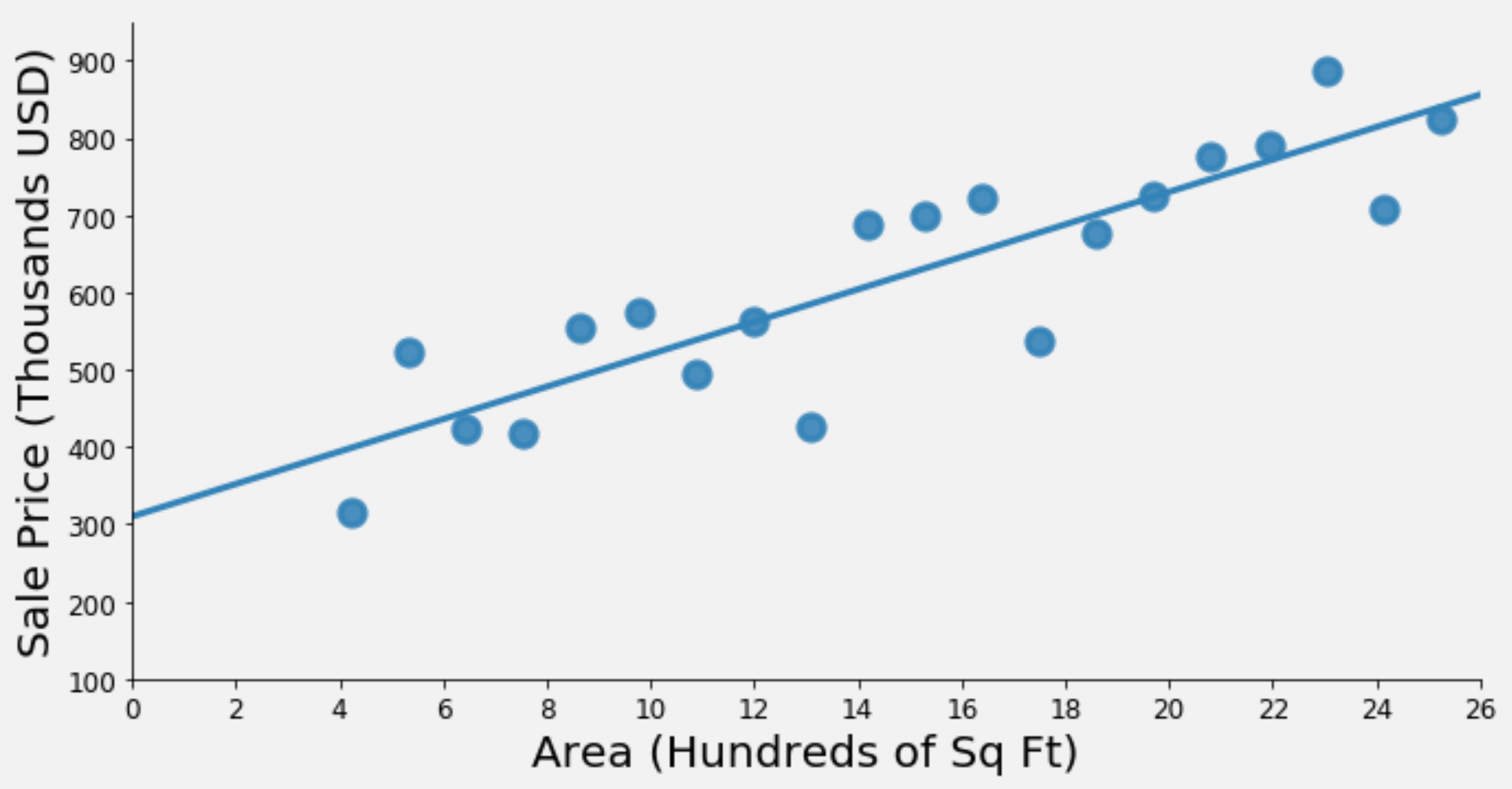
Simple Linear Regression

The points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ resulting from n independent observations will be scattered about the true regression line



Interpreting SLR Parameters

The data shown below comes from the model $Y = 310 + 21X + \epsilon$



Interpreting SLR Parameters

The data shown below comes from the model $Y = 310 + 21X + \epsilon$

- What is the interpretation of $\beta_1 = 21$?

FOR EACH 1 UNIT INCREASE IN X
Y INCREASES BY 21 UNITS

- What is the interpretation of $\beta_0 = 310$?

BIAS: SALE PRICE OF HOUSE W/
AREA $X=0$.

Area and Age as Predictors for Price

- Now we have two features: $X = (X_1, X_2) =$ (AREA, AGE)

- The response is still $Y =$ SALE PRICE

- We **assume** that the true relationship is given by

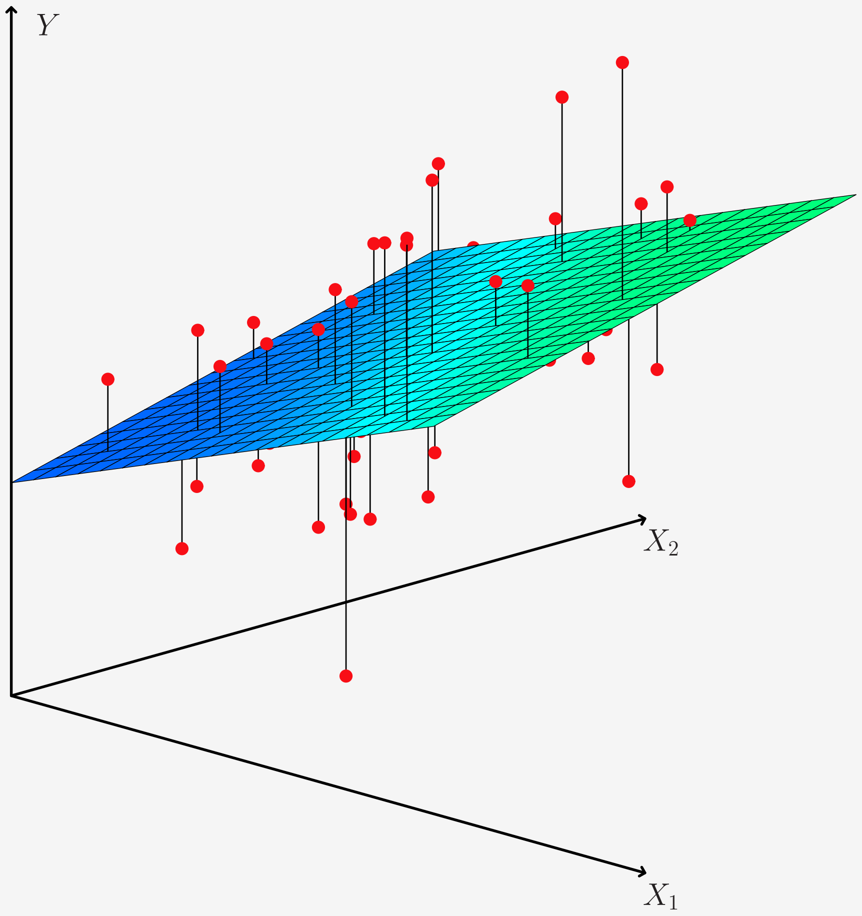
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- Assume we've collected data about n houses. Each house example satisfies

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

Multiple Linear Regression

Note that our model is no longer a simple line. Instead it is a linear surface.



Interpreting MLR Parameters

Suppose the true housing model with features Area and Age is

$$Y = 310 + 22X_1 - 1.25X_2 + \epsilon$$

- What is the interpretation of $\beta_1 = 22$?

W/ X_2 HELD FIXED, IF X_1 INCREASES BY 1 UNIT, Y INCREASES BY 22 UNITS

- What is the interpretation of $\beta_2 = -1.25$?

W/ X_1 HELD FIXED, IF X_2 INCREASES BY 1 UNIT, Y DECREASES BY 1.25 UNITS

Estimating the Regression Parameters

In real life, we have no hope of discovering the true model parameters, and have to estimate them from the **training** data. Our estimated model will be

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Once we've **learned** estimates of the parameters, we can use the regression model to make predictions

Suppose the learned regression model for the housing data is $\hat{y} = 300 + 20x_1 - x_2$

What price would we predict for a 2000 sqft house built 50 years ago?

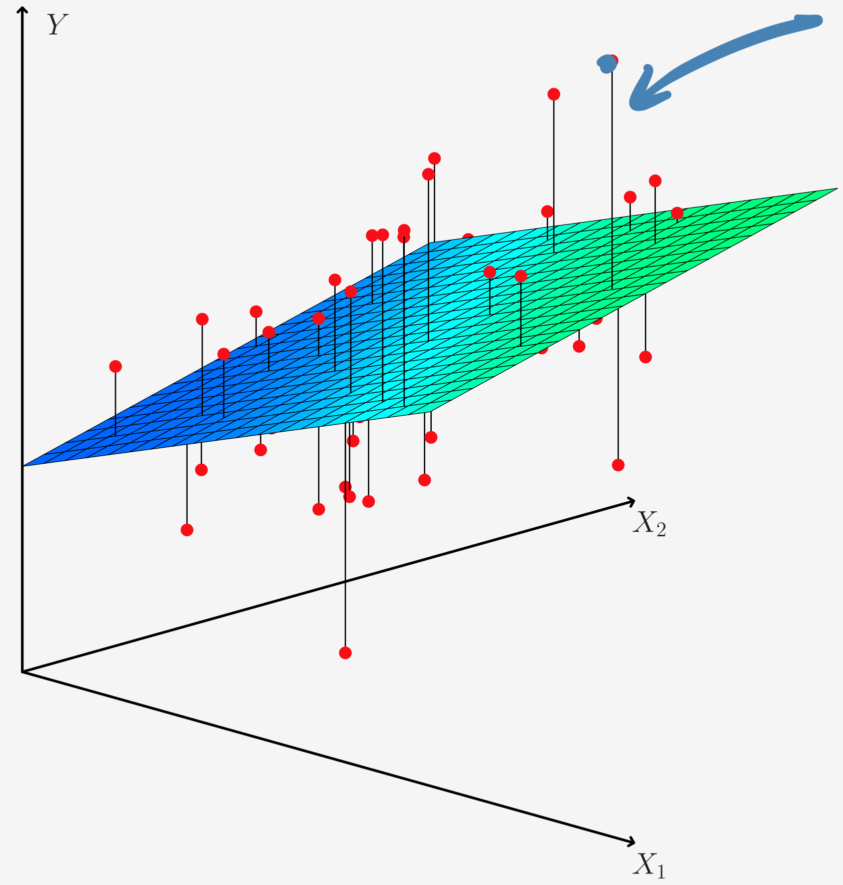
$x_1 = 20, \; x_2 = 50$

$\hat{y} = 300 + 20 \cdot 20 - 1 \cdot 50$

$= 650\,\text{K}$

Estimate Parameters from Data

Given training data, choose the $\hat{\beta}_k$'s to minimize distance from data to estimated model



$y_i - \hat{y}_i = e_i$ (i^{th} ERROR OR i^{th} RESIDUAL)

SQUARE AND SUM UP

$$(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+\cdots+(y_n-\hat{y}_n)^2$$

Estimate Parameters from Data

- The **residual sum of squares** for the points $\{(x_{1i}, x_{2i}, \ldots, x_{pi}, y_i)\}_{i=1}^{n}$ to the regression model is given by

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})\right]^2$$

- The estimates of the $\hat{\beta}_k$’s are called the **least-square estimates**, and are defined to be the values that **minimize** the RSS

- The RSS is a measure of the amount of variation NOT explained by the model

Estimate Parameters from Data

How do we choose the parameters that minimize the RSS?

- **Soon**: Use a powerful optimization technique called Stochastic Gradient Descent
- **For Now**: Remember some linear algebra

Think of our model as a function. Data points go in, predictions come out

$$y = f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Plug in each data point and see how we do …

Estimate Parameters from Data

This gives us n equations with the p+1 parameters as unknowns

$$\begin{aligned}
\beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} &= y_1 \\
\beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} &= y_2 \\
\beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \cdots + \beta_p x_{3p} &= y_3 \\
\vdots \quad &\quad \vdots \\
\beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} &= y_n
\end{aligned}$$

Remember that the x's and y's are **NUMBERS**!

What does this kinda look like?

Estimate Parameters from Data

This gives us n equations with the p+1 parameters as unknowns

$$\begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

We write the equivalent matrix formula as $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$

Note: $\mathbf{X}$ is a matrix where each row is a training example and we've added a column of 1's

We call $\mathbf{X}$ the **Design Matrix**

Estimate Parameters from Data

$$\begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \quad \Leftrightarrow \quad \mathbf{X}\boldsymbol{\beta} = \mathbf{y}$$

What are the dimensions of:	$\mathbf{X}$	$\boldsymbol{\beta}$	$\mathbf{y}$
	n x (p+1)	(p+1) x 1	n x 1

OK, we have a matrix equation for the parameters.

Question: Can we ever hope to solve this exactly? Nope!

Estimate Parameters from Data

$$\begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \quad \Leftrightarrow \quad \mathbf{X}\boldsymbol{\beta} = \mathbf{y}$$

This is a vastly over-determined system.

And what do we do with vastly over-determined systems?

The Least-Squares Problem

We solve the least-squares problem:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \; \|\mathbf{X}\beta - \mathbf{y}\|_2^2$$

Does the thing we're minimizing look familiar?

$$[X\beta - y]_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} - y_i$$

$$\|X\beta - y\|_2^2 = \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} - y_i)^2$$

$$= \text{RSS}$$

The Least-Squares Problem

We solve the least-squares problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \; \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$$

So how do we solve this thing?

The Least-Squares Problem

We solve the least-squares problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \; \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$$

So how do we solve this thing?

- **Theoretical** way… The **Normal Equations**: $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$

- (Sorta) **Practical** way… Methods based on QR factorization

- (**Actual**) Practical way… Stochastic Gradient Descent!

Predictions in Matrix Setting

So you've found your vector of estimated parameters $\hat{\beta}$ using your training data

Now suppose someone gives you a new data point you haven't seen yet

$$\mathbf{x} = (x_1, x_2, \ldots, x_p)$$

How do you predict the response for this new point?

Predictions in Matrix Setting

So you've found your vector of estimated parameters $\hat{\beta}$ using your training data

Now suppose someone gives you a new data point you haven't seen yet

$$\mathbf{x} = (x_1, x_2, \dots, x_p)$$

How do you predict the response for this new point?

$$\hat{y} = \hat{f}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Predictions in Matrix Setting

So you've found your vector of estimated parameters $\hat{\beta}$ using your training data

Now suppose someone gives you a new data point you haven't seen yet

$$\mathbf{x} = (x_1, x_2, \ldots, x_p)$$

How do you predict the response for this new point?

$$\hat{y} = \hat{f}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Vectorized Method: Prepend a 1 onto the data-point … $\mathbf{x} = (1, x_1, x_2, \ldots, x_p)$

and take dot-product: $\hat{y} = \hat{f}(\mathbf{x}) = \mathbf{x} \cdot \hat{\boldsymbol{\beta}} = \mathbf{x}^T \hat{\boldsymbol{\beta}}$

Predictions in Matrix Setting

Vectorized Method: Prepend a 1 onto the data-point … $\mathbf{x} = (1, x_1, x_2, \ldots, x_p)$

and take dot-product: $\hat{y} = \hat{f}(\mathbf{x}) = \mathbf{x} \cdot \hat{\boldsymbol{\beta}} = \mathbf{x}^T \hat{\boldsymbol{\beta}}$

House Price Example: Estimated regression model was $\hat{y} = 300 + 20x_1 - x_2$

What price would we predict for a 2000 sqft house built 50 years ago?

$$\hat{\beta} = \begin{bmatrix} 300 \\ 20 \\ -1 \end{bmatrix} \qquad x = \begin{bmatrix} 1 \\ 20 \\ 50 \end{bmatrix} \qquad \hat{y} = x^T \hat{\beta} = \begin{bmatrix} 1 & 20 & 50 \end{bmatrix} \begin{bmatrix} 300 \\ 20 \\ -1 \end{bmatrix}$$

$$= 1 \cdot 300 + 20 \cdot 20 + 50 \cdot (-1)$$

$$= 650$$

Predictions in Matrix Setting

= n points (n for mess)

What if you have a whole mess of points you want to make predictions for?

STACK IN A MATRIX

$$X_{TEST} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ & & \vdots & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\hat{y}_{TEST} = X_{TEST}\hat{\beta} \quad \text{(MAT-VEC)}$$

Regression Refresher Wrap-Up

OK, there's your regression refresher.

If any of this seems foggy, make sure to do the ISL reading.

Also, see links to 3022 slides posted on Piazza.

Next Time:

- See how we can extend this to nonlinear models with **Polynomial Regression**
- Talk about this **SUPER** important thing called **Regularization**

Don't forget that you have your first **Reading Quiz** due before class Monday

If-Time Bonus: Categorical Features

Suppose we want to add **COLOR** as a feature to our regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Is the color feature the same or different from Area and Age?

If-Time Bonus: Categorical Features

Suppose we want to add **COLOR** as a feature to our regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Is the color feature the same or different from Area and Age?

Area and Age are **continuous** features, but COLOR is a **categorical** feature

Current form of model is fine if there are only two colors, say brown and gray

But what if we have houses in our data set that are brown, gray, & neon green?

If-Time Bonus: Categorical Features

No natural ordering to color variable. Assigning 0, 1, and 2 imposes false ordering.

Instead create additional color features using **one-hot encoding**

$$X_3 \quad =$$

$$X_4 \quad =$$

New model becomes $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$

If-Time Bonus: Categorical Features

Note that the one-hot encoded features really give us three different models

$$\text{price} \approx \beta_0 + \beta_1 \times \texttt{area} + \beta_2 \times \texttt{age} + \begin{cases} \beta_3 & \text{if house is brown} \\ \beta_4 & \text{if house is grey} \\ 0 & \text{if house is green} \end{cases}$$