

LECTURE 13

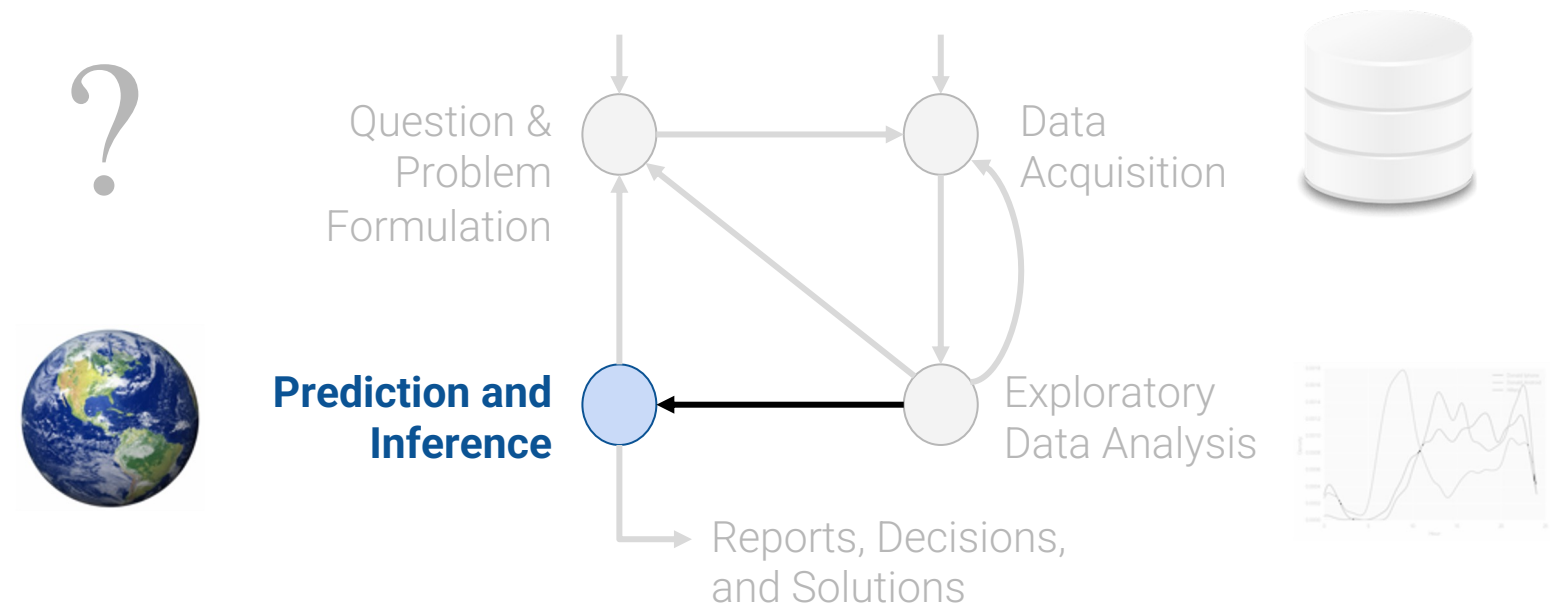
Constant Model, Loss, and Transformations

Adjusting the Modeling Process: different models, loss functions, and data transformations.

Data Science, Spring 2024 @ Knowledge Stream

Sana Jabbar

Plan for Next Few Lectures: Modeling



Modeling I:
Intro to Modeling, Simple
Linear Regression



Modeling III:
Multiple Linear
Regression



(today)

Modeling II:
Different models, loss
functions, linearization

Today's Roadmap

Lecture 13

Modeling Process Reiteration

- Evaluating Model the SLR Model
- Constant Model + MSE

Notation for Multiple Linear Regression

Transformations to Fit Linear Models

Evaluating the Model

Lecture 13

Modeling Process Reiteration

- **Evaluating Model the SLR Model**
- Constant Model + MSE

Transformations to Fit Linear Models

Notation for Multiple Linear Regression

Recap From Last Time...

1. Choose a model



How should we represent the world?

$$\hat{y} = \theta_0 + \theta_1 x \quad \text{SLR model}$$

2. Choose a loss function



How do we quantify prediction error?

$$L(y, \hat{y}) = (y - \hat{y})^2 \quad \text{Squared loss}$$

3. Fit the model



How do we choose the best parameters of our model given our data?

$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 x))^2 \quad \text{MSE for SLR}$$

4. Evaluate model performance

How do we evaluate whether this process gave rise to a good model?

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x \quad \left\{ \begin{array}{l} \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \\ \hat{\theta}_1 = r \frac{\sigma_y}{\sigma_x} \end{array} \right.$$

What are some ways to determine if our model was a good fit to our data?

1. Visualize data, compute statistics:

Plot original data.

Compute column means, standard deviation.

If we want to fit a linear model, compute correlation r .

2. Performance metrics:

Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- It is the square root of MSE, which is the average loss that we've been minimizing to determine optimal model parameters.
- RMSE is in the same units as y .
- A lower RMSE indicates more "accurate" predictions (lower "average loss" across data)

3. Visualization:

Look at a residual plot of $e_i = y_i - \hat{y}_i$ the difference between actual and predicted values.

Four Mysterious Datasets (Anscombe's quartet)

Ideal model evaluation steps, in order:


1. Visualize original data, Compute Statistics

2. Performance Metrics

For our simple linear least square model, use RMSE (we'll see more metrics later)

3. Residual Visualization

4 datasets could have similar aggregate statistics but still be wildly different:

 `x_mean : 9.00, y_mean : 7.50
x_stdev: 3.16, y_stdev: 1.94
r = Correlation(x, y): 0.816
theta_0_hat: 3.00, theta_1_hat: 0.50
RMSE: 1.119`

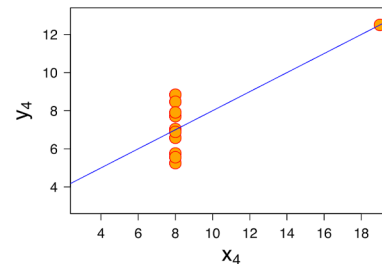
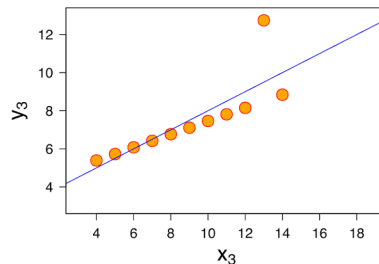
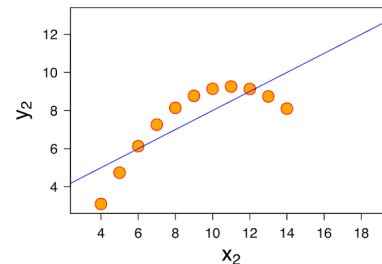
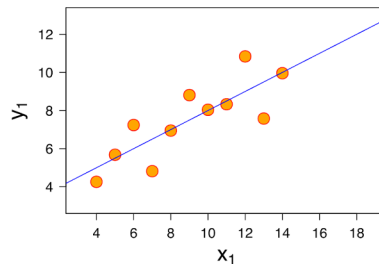
Demo

Four Mysterious Datasets (Anscombe's quartet)

- **The four dataset** each have the same mean of x , mean of y , SD of x , SD of y , and r value.
- Since our optimal Least Squares SLR model only depends on those quantities, they all have the **same regression line** and RMSE.

However, only one of these four sets of data makes sense to model using SLR.

Before modeling, you should always **visualize** your data first!



Demo

Anscombe's quartet: Residuals

Ideal model evaluation steps, in order:

1. **Visualize original data, Compute Statistics**

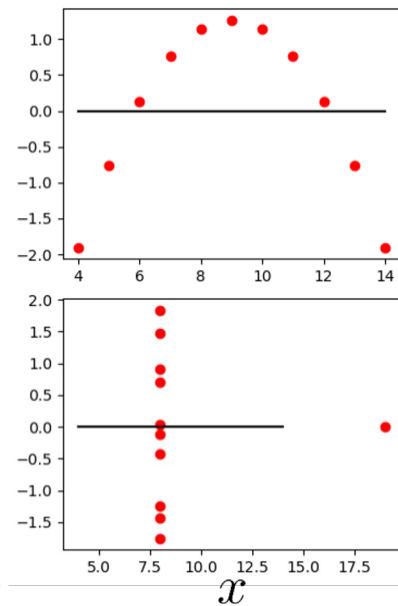
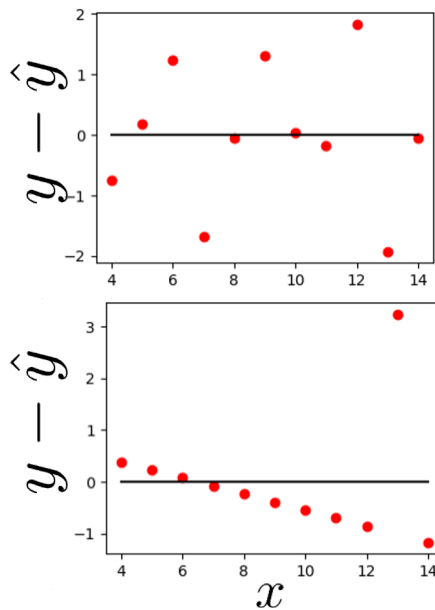
2. **Performance Metrics**

For our simple linear least square model, use RMSE (we'll see more metrics later)

3. **Residual Visualization**

Demo

The residual plot of a good regression shows **no pattern**.



The Modeling Process

1. Choose a model

How should we represent the world?

2. Choose a loss function

How do we quantify prediction error?

3. Fit the model

How do we choose the best parameters of our model given our data?

4. Evaluate model performance

How do we evaluate whether this process gave rise to a good model?

Review of the The Modeling Process (Simple Linear Regression)

1. Choose a model

SLR model

$$\hat{y} = \theta_0 + \theta_1 x$$

2. Choose a loss function

L2 Loss

Mean Squared Error (MSE)

$$L(y, \hat{y}) = (y - \hat{y})^2$$

$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \overbrace{(\theta_0 + \theta_1 x)}^{\hat{y}_i \text{ (SLR)}})^2$$

3. Fit the model

Minimize average loss with calculus

4. Evaluate model performance

Visualize, Root MSE

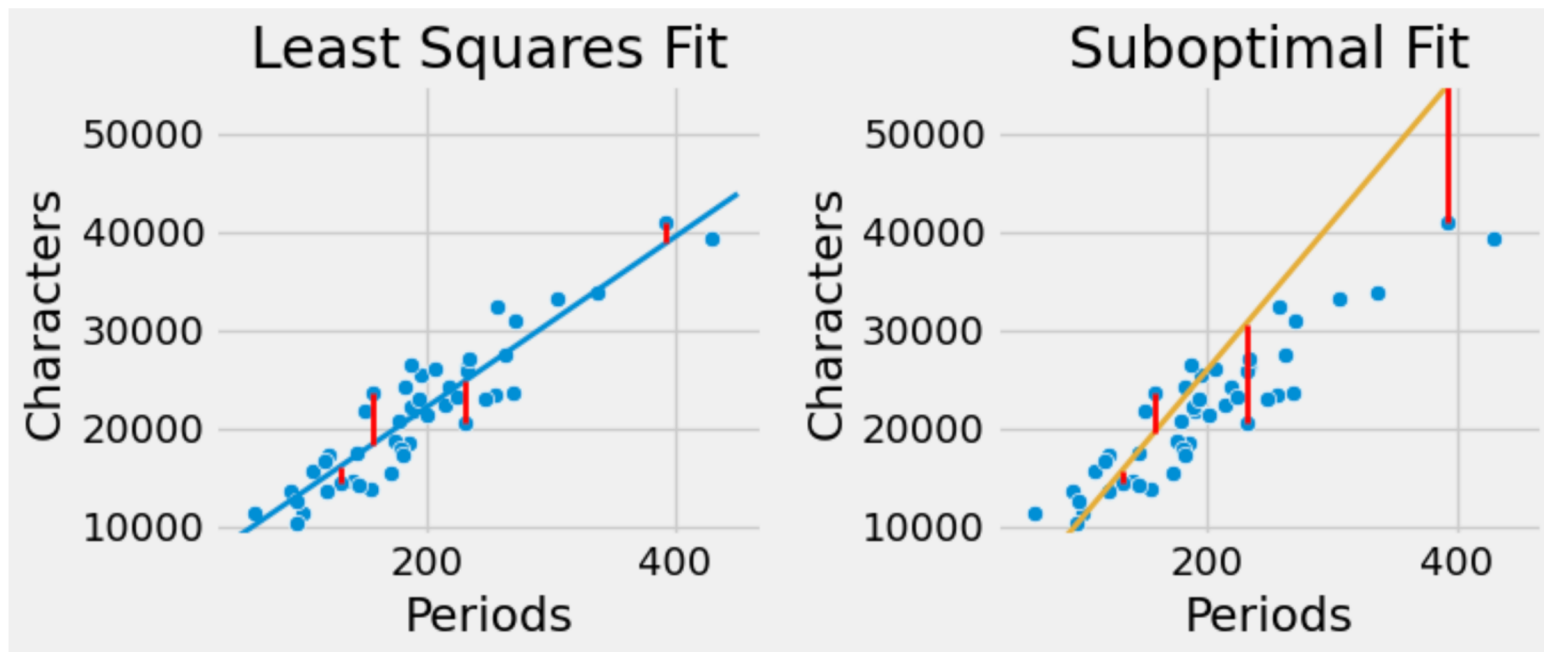
$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x \quad \left\{ \begin{array}{l} \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \\ \hat{\theta}_1 = r \frac{\sigma_y}{\sigma_x} \end{array} \right.$$

Minimizing MSE is Minimizing Squared Residuals

$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Residual ("error") in prediction

Lower residuals = better regression fit!



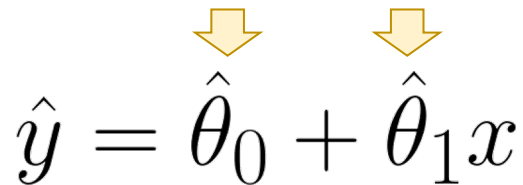
Terminology: Prediction vs. Estimation

These terms are often used somewhat interchangeably, but there is a subtle difference between them.

Estimation is the task of using data to calculate model parameters.

Prediction is the task of using a model to predict outputs for unseen data.

We **estimate** parameters by minimizing average loss...


$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$$

...then we **predict** using these estimates.

Least Squares Estimation

is when we choose the parameters that minimize MSE.

Constant Model + MSE

Lecture 13

Modeling Process Reiteration

- Evaluating Model the SLR Model
- **Constant Model + MSE**

Transformations to Fit Linear Models

Notation for Multiple Linear Regression

The Modeling Process: Using a Different Model

1. Choose a model

SLR model

~~$\hat{y} = \theta_0 + \theta_1 x$~~

Constant Model?

$$\hat{y} = ??$$

2. Choose a loss function

L2 Loss

Mean Squared Error
(MSE)

3. Fit the model

Minimize
average loss
with calculus

4. Evaluate model performance

Visualize,
Root MSE

The Constant Model

You work at a local boba shop and want to estimate the sales each day.

Here's your data from 5 randomly selected previous days, arbitrarily sorted by number of drinks sold:

$\{20, 21, 22, 29, 33\}$

How many drinks will you sell tomorrow?



- A. 0
- B. 25
- C. 22
- D. 100
- E. Something else



The Constant Model

You work at a local boba shop and want to estimate the sales each day.

Here's your data from 5 randomly selected previous days, arbitrarily sorted by number of drinks sold:

$\{20, 21, 22, 29, 33\}$

How many drinks will you sell tomorrow?



- A. 0
- B. 25
- C. 22
- D. 100
- E. Something else

This is a **constant model**.

The Constant Model

The **constant model**, also known as a **summary statistic**, summarizes the data by always "predicting" the same number—i.e., predicting a constant.

It ignores any relationships between variables:

- For instance, boba tea sales likely depend on the time of year, the weather, how the customers feel, whether school is in session, etc.
- Ignoring these factors is a **simplifying assumption**.

The constant model is also a parametric, statistical model:

$$\hat{y} = \theta_0$$

The Constant Model

The **constant model**, also known as a **summary statistic**, summarizes the data by always "predicting" the same number—i.e., predicting a constant.

It ignores any relationships between variables.

- For instance, boba tea sales likely depend on the time of year, the weather, how the customers feel, whether school is in session, etc.
- Ignoring these factors is a **simplifying assumption**.

The constant model is also a parametric, statistical model:

$$\hat{y} = \theta_0$$

- Our parameter θ_0 is 1-dimensional. $\theta_0 \in \mathbb{R}$
- We now have no input into our model; we predict $\hat{y} = \theta_0$
- Like before, we can still determine the best θ_0 that minimizes **average loss** on our data.



The Modeling Process: Using a Different Model



1. Choose a model

~~SLR model~~
 ~~$\hat{y} = \theta_0 + \theta_1 x$~~

Constant Model $\hat{y} = \theta_0$

2. Choose a loss function

L2 Loss

Mean Squared Error
(MSE)

(Let's stick with MSE.)

3. Fit the model

Minimize
average loss
with calculus

4. Evaluate model
performance

Visualize,
Root MSE

The Modeling Process: Using a Different Model

1. Choose a model



~~SLR model~~

~~$\hat{y} = \theta_0 + \theta_1 x$~~

Constant Model $\hat{y} = \theta_0$

2. Choose a loss function



L2 Loss

Mean Squared Error (MSE)

3. Fit the model

Minimize average loss with calculus

How does this step change?

4. Evaluate model performance

Visualize, Root MSE

Fit the Model: Rewrite MSE for the Constant Model

Recall that Mean Squared Error (MSE) is average squared loss (L2 loss) over the data $\mathcal{D} = \{y_1, y_2, \dots, y_n\}$:

$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^n \underbrace{(y_i - \hat{y}_i)^2}_{\text{L2 loss on a single datapoint}}$$

Given the **constant model** $\hat{y} = \theta_0$:

$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \theta_0)^2$$

We **fit the model** by finding the optimal $\hat{\theta}_0$ that minimizes the MSE.

$$\hat{R}(\theta_0) = \frac{1}{n} \sum_{i=1}^n (y_i - \theta_0)^2$$

Approach If you want to prove the general case for any data, you could directly minimize the objective. We can show that average loss is minimized by

$$\hat{\theta}_0 = \text{mean}(\mathbf{y}) = \bar{y} \qquad \mathcal{D} = \{20, 21, 22, 29, 33\}$$

Fit the Model: Calculus for the General Case

1. Differentiate with respect to θ_0 :

$$\frac{d}{d\theta_0} R(\theta) = \frac{d}{d\theta_0} \left(\frac{1}{n} \sum_{i=1}^n (y_i - \theta_0)^2 \right)$$

Derivative of sum is
sum of derivatives

Chain rule

Simplify constants

2. Set equal to 0.

3. Solve for $\hat{\theta}_0$.

Fit the Model: Calculus for the General Case

1. Differentiate with respect to θ_0 :

$$\begin{aligned}\frac{d}{d\theta_0}R(\theta) &= \frac{d}{d\theta_0}\left(\frac{1}{n}\sum_{i=1}^n(y_i - \theta_0)^2\right) \\ &= \frac{1}{n}\sum_{i=1}^n \underbrace{\frac{d}{d\theta_0}(y_i - \theta_0)^2}_{\text{Chain rule}} \quad \text{Derivative of sum is sum of derivatives} \\ &= \frac{1}{n}\sum_{i=1}^n 2(y_i - \theta_0)(-1) \quad \text{Chain rule} \\ &= \frac{-2}{n}\sum_{i=1}^n (y_i - \theta_0) \quad \text{Simplify constants}\end{aligned}$$

2. Set equal to 0.

$$0 = \frac{-2}{n}\sum_{i=1}^n (y_i - \hat{\theta}_0)$$

3. Solve for $\hat{\theta}_0$.

Fit the Model: Calculus for the General Case

1. Differentiate with respect to θ_0 :

$$\begin{aligned}\frac{d}{d\theta_0}R(\theta) &= \frac{d}{d\theta_0}\left(\frac{1}{n}\sum_{i=1}^n(y_i - \theta_0)^2\right) \\&= \frac{1}{n}\sum_{i=1}^n \frac{d}{d\theta_0}(y_i - \theta_0)^2 && \text{Derivative of sum is sum of derivatives} \\&= \frac{1}{n}\sum_{i=1}^n \underbrace{2(y_i - \theta_0)(-1)}_{\text{Chain rule}} \\&= \frac{-2}{n}\sum_{i=1}^n(y_i - \theta_0) && \text{Simplify constants}\end{aligned}$$

2. Set equal to 0.

$$0 = \frac{-2}{n}\sum_{i=1}^n(y_i - \hat{\theta}_0)$$

3. Solve for $\hat{\theta}_0$.

$$\begin{aligned}0 &= \cancel{\frac{-2}{n}}\sum_{i=1}^n(y_i - \hat{\theta}_0) = \sum_{i=1}^n(y_i - \hat{\theta}_0) \\&= \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\theta}_0 && \text{Separate sums} \\&= \sum_{i=1}^n y_i - n \cdot \hat{\theta}_0 && c + c + \dots + c = n \cdot c \\n \cdot \hat{\theta}_0 &= \sum_{i=1}^n y_i \\ \hat{\theta}_0 &= \frac{1}{n}\left(\sum_{i=1}^n y_i\right) \implies \boxed{\hat{\theta}_0 = \bar{y}}\end{aligned}$$

Interpreting $\hat{\theta}_0 = \bar{y}$

This is the optimal parameter for constant model + MSE.

- It holds true regardless of what data sample you have.
- It provides some formal reasoning as to why the mean is such a common summary statistic. Fun fact:

The minimum MSE is the **sample variance**.

$$R(\hat{\theta}_0) = R(\bar{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \sigma_y^2$$

Note the difference:

$$R(\theta_0) = \min_{\theta_0} R(\theta_0) = \sigma_y^2 \quad \text{vs} \quad \hat{\theta}_0 = \operatorname{argmin}_{\theta_0} R(\theta_0) = \bar{y}$$

The **minimum value** of
constant + MSE

The **argument** that **minimizes**
constant + MSE

In modeling, we care less about **minimum loss** $R(\hat{\theta}_0)$ and more about the **minimizer** of loss $\hat{\theta}_0$.

Revisit the Boba Shop Example

You work at a local boba shop and want to estimate the sales each day.

Here's your data from 5 randomly selected previous days, arbitrarily sorted by number of drinks sold:

$\{20, 21, 22, 29, 33\}$

How many drinks will you sell tomorrow?

We will predict the mean of the previous five days' sale:

$$(20 + 21 + 22 + 29 + 33)/5 = 25.$$



A. 0

B. 25

C. 22

D. 100

E. Something else

The Modeling Process: Using a Different Model

1. Choose a model



Constant Model

Constant Model $\hat{y} = \theta_0$

2. Choose a loss function



L2 Loss

Mean Squared Error (MSE)

3. Fit the model



Minimize average loss with calculus

$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \theta_0)^2$$

$$\hat{\theta}_0 = \text{mean}(y) = \bar{y}$$

4. Evaluate model performance

Visualize, Root MSE

Suppose we wanted to predict dugong ages.



A Dugong [\[image source\]](#)



Not a Dugong, a Dewgong [\[image source\]](#)

Constant Model

$$\hat{y} = \theta_0$$

Data: Sample of ages.

$$\mathcal{D} = \{y_1, y_2, \dots, y_n\}$$

Simple Linear Regression

$$\hat{y} = \theta_0 + \theta_1 x$$

Data: Sample of (length, age)s.

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

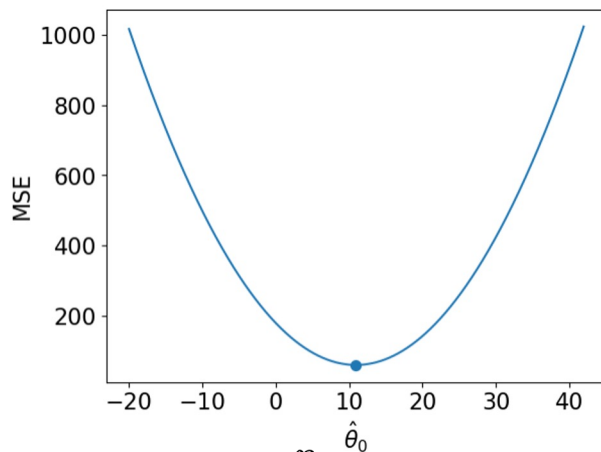
Demo

Constant Model

$$\hat{y} = \theta_0$$

$\hat{\theta}_0$ is **1-D**.

Loss surface is **2-D**.



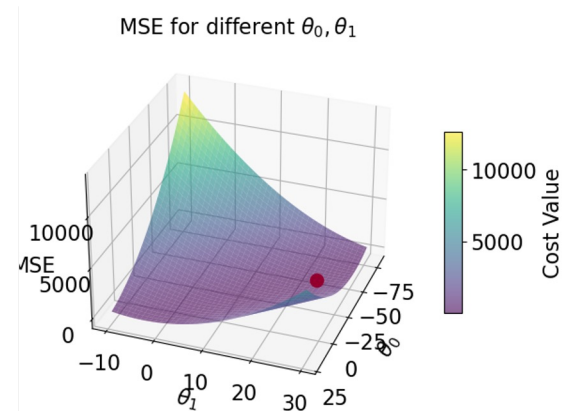
$$\hat{R}(\theta_0) = \frac{1}{n} \sum_{i=1}^n (y_i - \theta_0)^2$$

Simple Linear Regression

$$\hat{y} = \theta_0 + \theta_1 x$$

$\hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1]$ is **2-D**.

Loss surface is **3-D**.



$$\hat{R}(\theta_0, \theta_1) = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 x))^2$$

Demo

[Fit] Comparing Two Different Models, Both Fit with MSE

Constant Model

$$\hat{y} = \theta_0$$

RMSE: **7.72**

Simple Linear Regression

$$\hat{y} = \theta_0 + \theta_1 x$$

RMSE **4.31**

Interpret the RMSE (Root Mean Square Error):

- Constant error is **HIGHER** than linear error
- Constant model is **WORSE** than linear model (at least for this metric)

Demo


See notebook for code

Summary: Loss Optimization, Calculus, and...Critical Points?

First, define the **objective function** as average loss.

- Plug in L1 or L2 loss.
- Plug in model so that resulting expression is a function of θ .

Then, find the **minimum** of the objective function:

1. Differentiate with respect to θ .
 2. Set equal to 0.
 3. Solve for $\hat{\theta}$.
- 
- Repeat w/partial derivatives
if multiple parameters

Recall **critical points** from calculus: $R(\hat{\theta})$ could be a minimum, maximum, or saddle point!

- We should technically also perform the second derivative test, i.e., show $R''(\hat{\theta}) > 0$.

Notation for Multiple Linear Regression

Lecture 13

Modeling Process Reiteration

- Evaluating Model the SLR Model
- Iteration 2: Constant Model + MSE

Notation for Multiple Linear Regression

Transformations to Fit Linear Models

A Note on Terminology

There are several equivalent terms in the context of regression.

Feature(s)

Covariate(s)

Independent variable(s)

Explanatory variable(s)

Predictor(s)

Input(s)

Regressor(s)

Output

Outcome

Response

Dependent variable

Weight(s)

Parameter(s)

Coefficient(s)

Prediction

Predicted response

Estimated value

Estimator(s)

Optimal parameter(s)

Bolded terms are the most common in this course.

Match each column
with the appropriate term: $x, y, \hat{y}, \theta, \hat{\theta}$

A Note on Terminology

There are several equivalent terms in the context of regression.

Feature(s)

Covariate(s)

Independent variable(s)

Explanatory variable(s)

Predictor(s)

Input(s)

Regressor(s)

x

Output

Outcome

Response

Dependent variable

y

Prediction

Predicted response

Estimated value

\hat{y}

Weight(s)

Parameter(s)

Coefficient(s)

θ

Estimator(s)

Optimal parameter(s)

$\hat{\theta}$

Bolded terms are the most common in this course.

A datapoint (x, y) is also called an **observation**.

Multiple Linear Regression

Define the **multiple linear regression** model:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

Parameters are $\theta = [\theta_0, \theta_1, \dots, \theta_p]$

Is this linear in θ ?

- A. no
- B. yes
- C. maybe

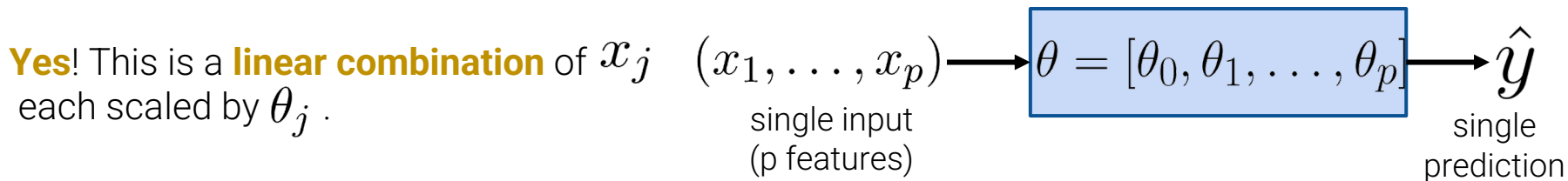
Multiple Linear Regression

Define the **multiple linear regression** model:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

Parameters are $\theta = [\theta_0, \theta_1, \dots, \theta_p]$

Yes! This is a **linear combination** of x_j each scaled by θ_j .



Example: Predict dugong ages \hat{y} as a linear model of 2 features: length x_1 **and** weight x_2 .

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

intercept parameter for length parameter for weight

Does a Unique Solution Always Exist?

	Model	Estimate	Unique?
Constant Model + MSE	$\hat{y} = \theta_0$	$\hat{\theta}_0 = mean(y) = \bar{y}$	Yes. Any set of values has a unique mean.
Constant Model + MAE	$\hat{y} = \theta_0$	$\hat{\theta}_0 = median(y)$	Yes , if odd. No , if even. Return average of middle 2 values.
Simple Linear Regression + MSE	$\hat{y} = \theta_0 + \theta_1 x$	$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$ $\hat{\theta}_1 = r \frac{\sigma_y}{\sigma_x}$	Yes. Any set of non-constant* values has a unique mean, SD, and correlation coefficient.
Ordinary Least Squares (Linear Model + MSE)	$\hat{\mathbf{Y}} = \mathbf{X}\theta$	$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$???

Does a Unique Solution Always Exist?

	Model	Estimate	Unique?
Constant Model + MSE	$\hat{y} = \theta_0$	$\hat{\theta}_0 = mean(y) = \bar{y}$	Yes. Any set of values has a unique mean.
Constant Model + MAE	$\hat{y} = \theta_0$	$\hat{\theta}_0 = median(y)$	Yes , if odd. No , if even. Return average of middle 2 values.
Simple Linear Regression + MSE	$\hat{y} = \theta_0 + \theta_1 x$	$\begin{aligned}\hat{\theta}_0 &= \bar{y} - \hat{\theta}_1 \bar{x} \\ \hat{\theta}_1 &= r \frac{\sigma_y}{\sigma_x}\end{aligned}$	Yes. Any set of non-constant* values has a unique mean, SD, and correlation coefficient.
Ordinary Least Squares (Linear Model + MSE)	$\hat{\mathbf{Y}} = \mathbf{X}\theta$	$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$	Yes , if \mathbf{X} is full col rank (all cols lin independent, #datapoints>> #features)

Transformations to Fit Linear Models

Lecture 12

Modeling Process Reiteration

- Evaluating Model the SLR Model
- Iteration 2: Constant Model + MSE

Notation for Multiple Linear Regression

Transformations to Fit Linear Models

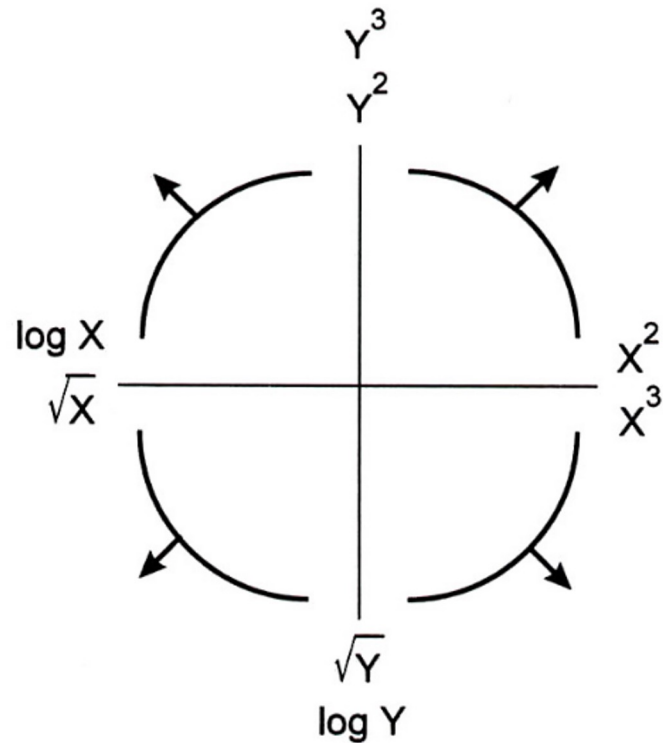
Tukey-Mosteller Bulge Diagram

The **Tukey-Mosteller Bulge Diagram** is a guide to possible transforms to try to get linearity.

- There are multiple solutions. Some will fit better than others.
- sqrt and log make a value “smaller”.
- Raising a value to a power makes it “bigger”.
- Each of these transformations equates to increasing or decreasing the scale of an axis.

Other goals other than linearity are possible

- E.g. make data appear more symmetric.
- Linearity allows us to fit lines to the transformed data

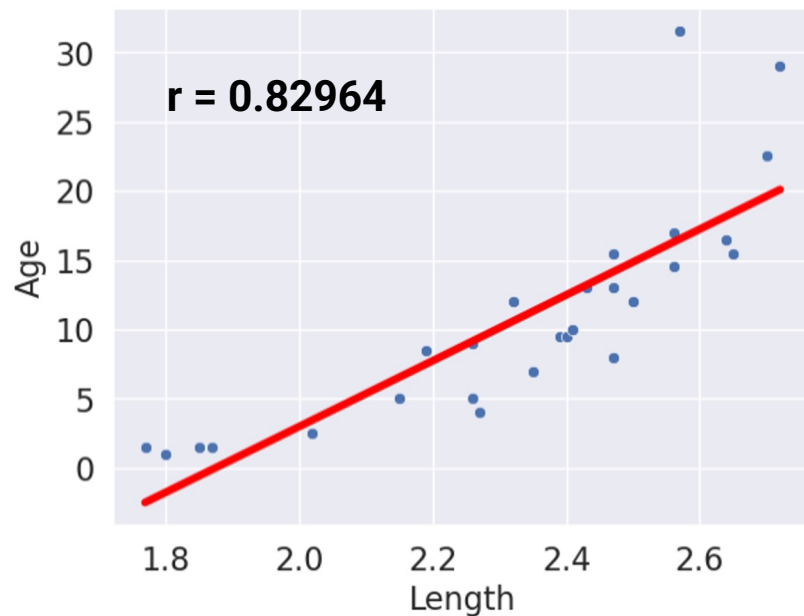




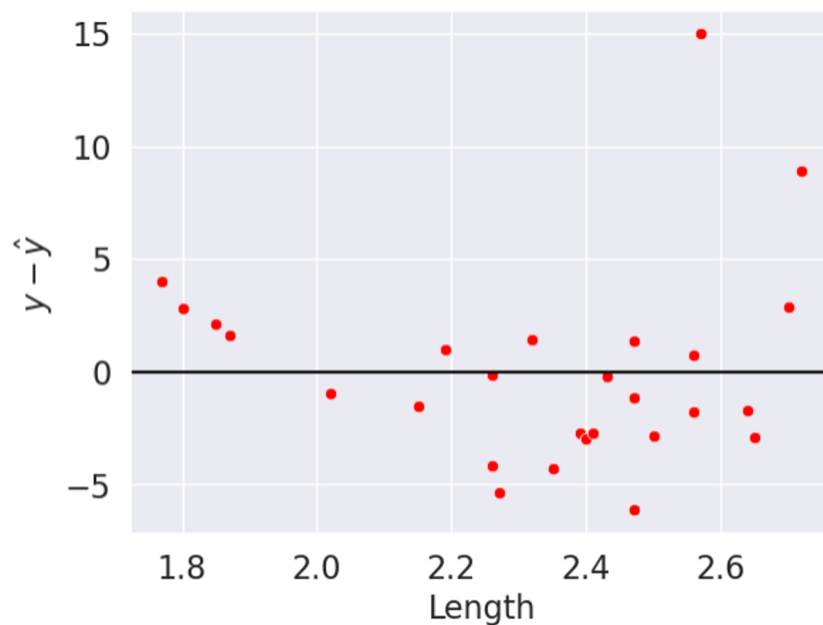
The residual plot of a good regression shows no pattern.

Back to Least Squares Regression with Dugongs

Age by Length



Residual Plot



Residual plot shows a clear pattern! On closer inspection, the scatter plot **curves upward**.

Q: How can we fit a curve to this data with the tools we have?

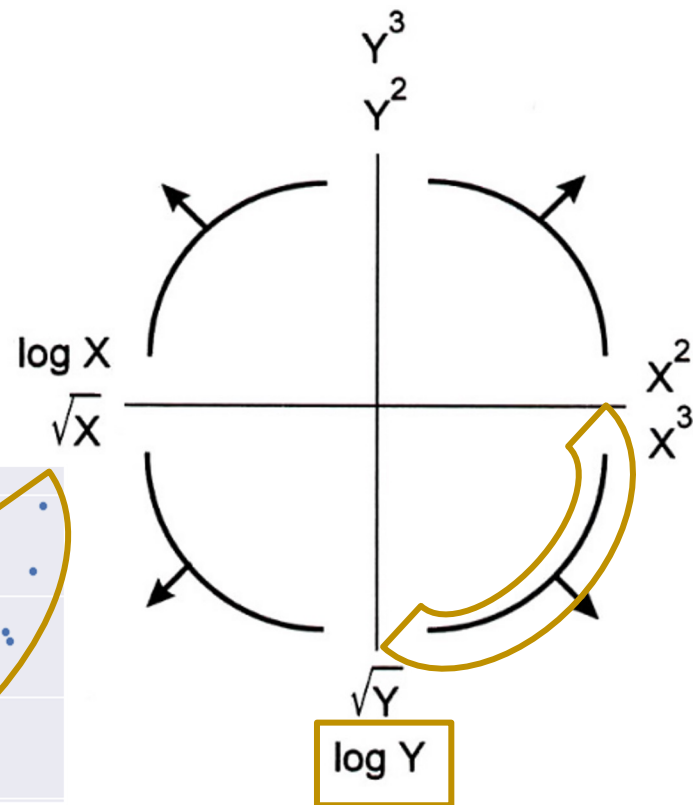
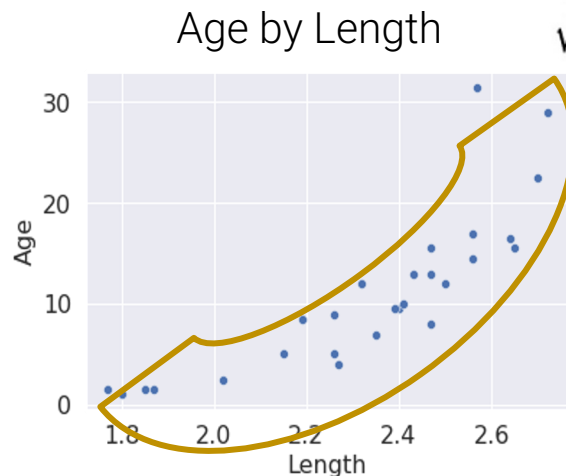
A: **Transform the Data.**

Tukey-Mosteller Bulge Diagram

If your data “bulges” in a direction, transform x and/or y in that direction.

- Each of these transformations equates to increasing or decreasing the scale of an axis.
- Roots and logs make a value “smaller”.
- Raising to a power makes a value “bigger”.

There are multiple solutions!
Some will fit better than others.



Lecture 12

Constant Model, Loss, and Transformations