Introduction to Modeling, SLR

Understanding the usefulness of models and the simple linear regression model

Content credit: Acknowledgments

Population vs. **Sample**

- The **population** is a set of individuals (people / bacteria / villages) in the **real world** that we are interested in learning about.
- The **sample** is a (usually smaller) subset of data we can actually collect & analyze.
- If we pick a good sampling scheme, the sample is **representative** of the population (in practice that is often hard to accomplish!).
- If so, we can hope that patterns we find in the data will reflect patterns in the wider world (but **how closely?** Question of **inference** which we will discuss later).

What Is A Model?

What Is A Model?

A model is an **idealized representation** of a system.

Example:
We model the fall of an object on Earth as subject to a constant acceleration of 9.81 m/s^2 due to gravity.

- While this describes the behavior of our system, it is merely an approximation.
- It doesn’t account for the effects of air resistance, local variations in gravity, etc.
- But in practice, it’s accurate enough to be useful!

Two Reasons for Building Models

Reason 1:

To understand **complex phenomena** occurring in the world we live in.

- What factors play a role in the growth of COVID-19?
- How do an object's velocity and acceleration impact how far it travels? (Physics: $d = d_0 + vt + \frac{1}{2}at^2$)

Often times, we care about creating models that are **simple and interpretable**, allowing us to understand what the relationships between our variables are.

Reason 2:

To make **accurate predictions** about unseen data.

- Can we predict if an email is spam or not?
- Can we generate a one-sentence summary of this 10-page long article?

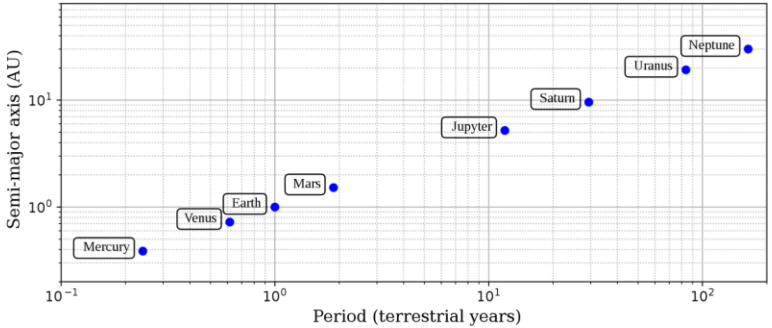
Other times, we care more about making extremely accurate predictions, at the cost of having an uninterpretable model. These are sometimes called **black-box models**, and are common in fields like deep learning.

Most of the time, we want to strike a balance between interpretability and accuracy.

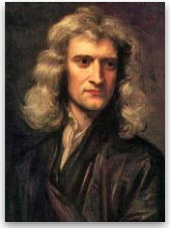
Deterministic physical (mechanistic) models: Laws that govern how the world works.

Kepler's Third Law of Planetary Motion (1619)
The ratio of the square of an object's orbital period with the cube of the semi-major axis of its orbit is the same for all objects orbiting the same primary.

$$T^2 \propto R^3$$

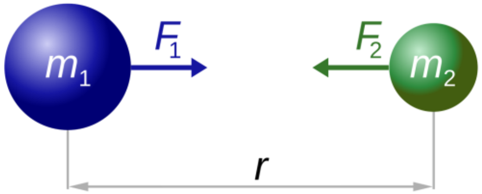


Newton's Laws: motion and gravitation (1687)
Newton's second law of motion models the relationship between the mass of an object and the force required to accelerate it.



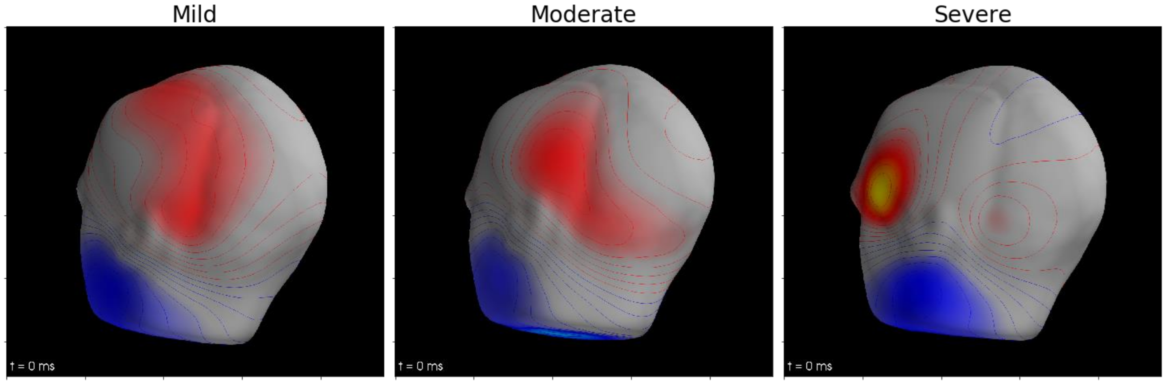
$$\mathbf{F} = m\mathbf{a}$$

$$F = G\frac{m_1 m_2}{r^2}$$



Probabilistic models

- Models of how random processes evolve.
- Often motivated by understanding of an unpredictable system.



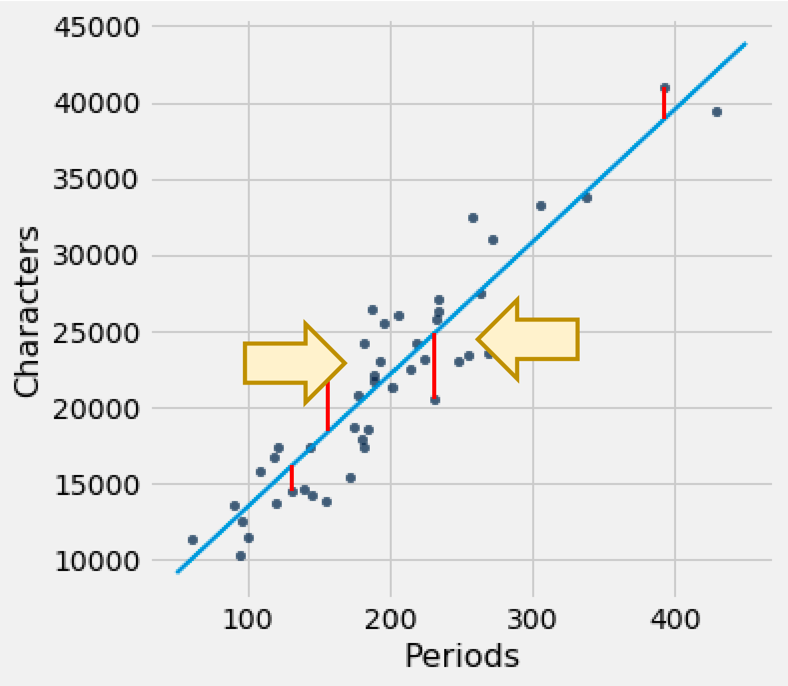
Simple Linear Regression & Correlation

The **regression line** is the unique straight line that minimizes the **mean squared error** of estimation among all straight lines.

$$\text{slope} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\text{intercept} = \text{average of } y - \text{slope} \cdot \text{average of } x$$

residual $= \text{observed value} - \text{regression estimate}$



For every chapter of the novel *Little Women*, Estimate the **# of characters** $\hat{y}$ based on the **# of periods** x in that chapter.

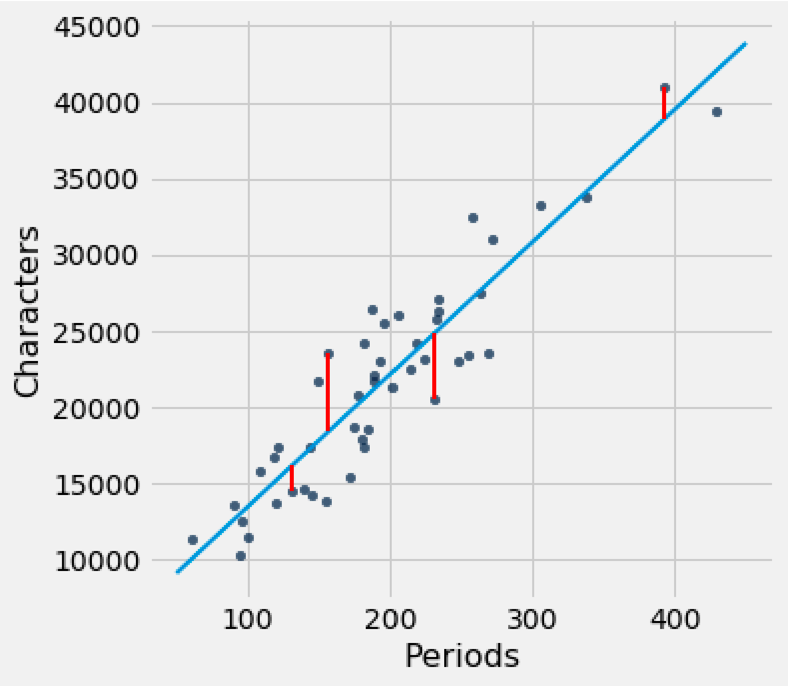
> The **regression line** is the unique straight line that minimizes the **mean squared error** of estimation among all straight lines.

correlation

slope $= r \cdot \dfrac{\text{SD of } y}{\text{SD of } x}$

intercept $=$ $\text{average of } y - \text{slope} \cdot \text{average of } x$

residual $=$ $\text{observed value} - \text{regression estimate}$



For every chapter of the novel *Little Women*, Estimate the **# of characters** $\hat{y}$ based on the **# of periods** x in that chapter.

Correlation

The **correlation** r is the average of the product of x and y, both measured in standard units.

Define the following:

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

$\bar{x}, \bar{y}$ means; σ_x, σ_y standard deviations

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

- x_i in standard units: $\frac{x_i - \bar{x}}{\sigma_x}$
- r is also known as Pearson's correlation coefficient.
- Side note: **covariance** is $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = r\sigma_x\sigma_y$

The **correlation** r is the average of the product of x and y, both measured in standard units.

Define the following:

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

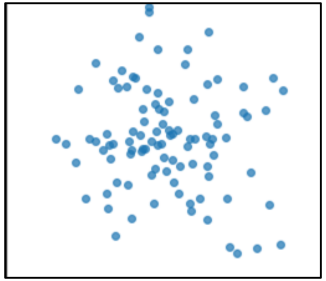
$\bar{x}, \bar{y}$ Means; σ_x, σ_y standard deviations

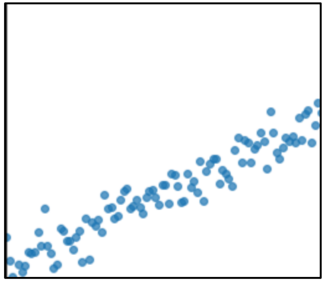
$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

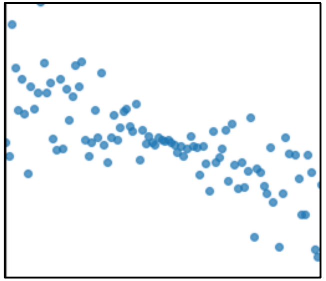
- x_i in standard units: $\frac{x_i - \bar{x}}{\sigma_x}$
- r is also known as Pearson's correlation coefficient.

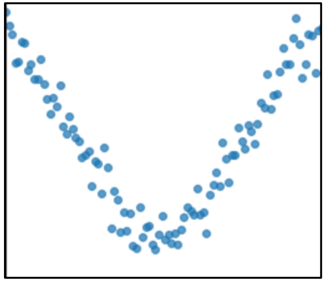
Correlation measures the strength of a **linear association** between two variables.

$|r| < 1$


r = −0.121


r = 0.951

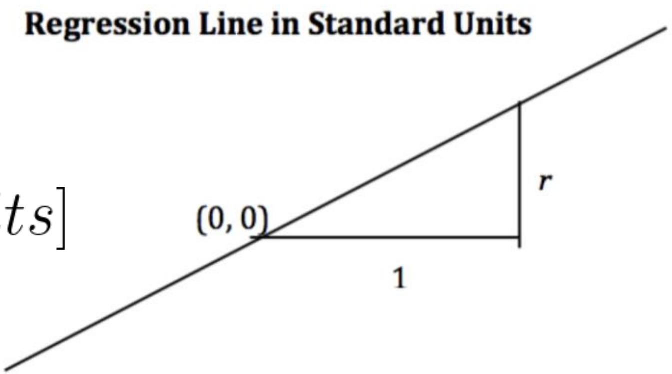

r = −0.723


⚠ **r = 0.056**

[The Regression Line

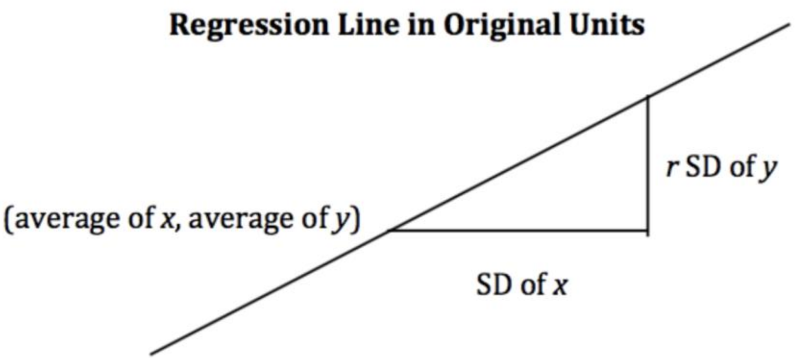
- When the variables x and y are measured in standard units, the regression line for predicting y based on x has slope r passes through the origin and the equation will be:

$$\hat{y} = r \times x \ [both\ measured\ in\ standard\ units]$$



- In the original units of the data, this becomes:

$$\frac{\hat{y} - \bar{y}}{\sigma_y} = r \times \frac{x - \bar{x}}{\sigma_x}$$



The Regression Line

$$\frac{\hat{y}-\bar{y}}{\sigma_y} = r \times \frac{x-\bar{x}}{\sigma_x}$$

$$\hat{y} = \sigma_y \times r \times \frac{x-\bar{x}}{\sigma_x} + \bar{y}$$

$$\hat{y} = \boxed{(\frac{r\sigma_y}{\sigma_x})} \times x + \boxed{(\bar{y} - \frac{r\sigma_y}{\sigma_x}\bar{x})}$$

Recall regression line equation is defined as:

$$\hat{y} = \boxed{\hat{a}} + \boxed{\hat{b}}x$$

slope: $r\frac{SD\ of\ y}{SD\ of\ x} = r\frac{\sigma_y}{\sigma_x}$

intercept: $\bar{y} - slope \times \bar{x}$

Error for the i-th data point: $e_i = y_i - \hat{y}_i$

The Modeling Process: Definitions

Parametric Models

Simple Linear Regression Model (SLR)

notation $\hat{y} = a + bx$ → Another notation: $\hat{y} = \theta_0 + \theta_1 x$

SLR is a **parametric model**: it is described by a few **parameters** (in this case θ_0, θ_1)

- No one tells us the parameters: the data informs us about them.
- The x values are **not** parameters because we directly observe them.
- Sample-based **estimate** of θ_0, θ_1 written as $\hat{\theta}_0, \hat{\theta}_1$.

Usually, we pick the parameters that appear "best" according to some criterion we choose

- Usually standing in as a proxy for fit to new data.

Parametric Model Notation

y True outputs

$\hat{y}$ Predicted outputs

For data:

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

The i-th datapoint is an **observation**:

- y_i is the i-th **output** (aka dependent variable)
- x_i is the i-th **feature** (aka independent variable)
- $\hat{y}_i$ is the i-th **prediction** (aka estimation).

θ Model parameter(s)

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

Any linear model with parameters $\theta = [\theta_0, \theta_1]$

$\hat{\theta}$ Estimated parameter(s), "best" fit to data in some sense

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

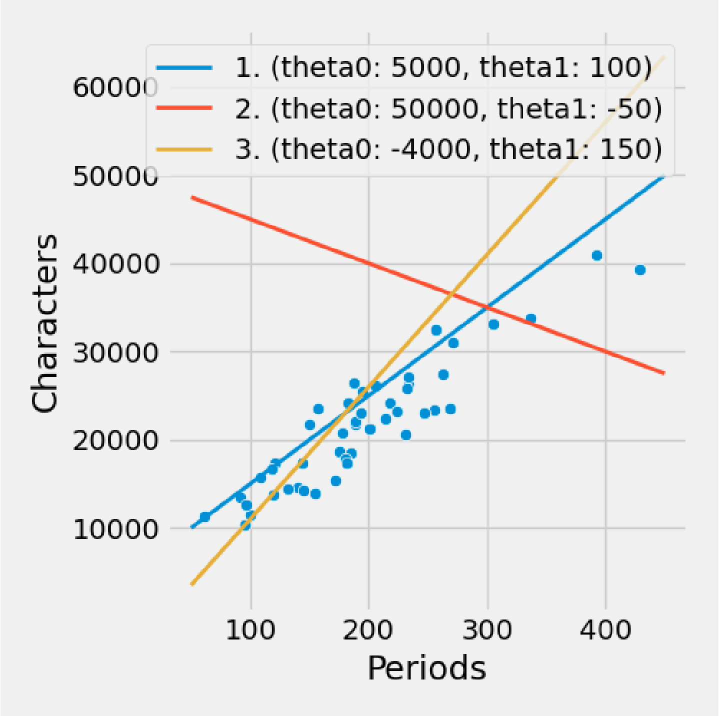
The "best" fitting linear model with parameters $\hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1]$

Which θ is best?

Based on your interpretation of the data, which are the "optimal parameters" for this linear model?

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

$$\hat{\theta}_0 = ? \quad \hat{\theta}_1 = ?$$



We only had 3 values to choose from to find the optimal parameter. In practice, our parameter domain is all reals, i.e., $\theta = [\theta_0, \theta_1] \in \mathbb{R}^2$

For every chapter of the novel *Little Women*, Estimate the **# of characters** $\hat{y}$ based on the **# of periods** x in that chapter.

Simple Linear Regression: Our First Model

Simple Linear Regression Model (SLR)

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

SLR is a **parametric model**, meaning
we choose the "best" **parameters**
for slope and intercept based on data.

- We often express θ as a single parameter vector.
- x is **not** a parameter! It is input to our model.

$$x \longrightarrow \boxed{\text{SLR} \quad \theta = [\theta_0, \theta_1]} \longrightarrow \hat{y}$$

- Note that the true relationship between x and y is usually non-linear.
  This is why $\hat{y}$ (and not y) appears in our **estimated linear model** expression.

- Other parametric models we'll see soon: $\hat{y} = \theta$ $\quad \hat{y} = x^T\theta$ $\quad \hat{y} = \dfrac{1}{1 + \exp(-x^T\vec{\theta})}$
- Note: Not all statistical models have parameters! KDEs are non-parametric models.

1. Choose a model

How should we represent the world?

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

SLR model

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?

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

4. Evaluate model performance

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

Reflect

Loss Functions

1. Choose a model

How should we represent the world?

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

SLR model

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?

We need some metric of how "good" or "bad" our predictions are.

A **loss function** characterizes the **cost**, error, or fit resulting from a particular choice of model or model parameters.

- Loss quantifies how bad a prediction is for a **single** observation.
- If our prediction $\hat{y}$ is **close** to the actual value y, we want **low loss**.
- If our prediction $\hat{y}$ is **far** from the actual value y, we want **high loss**.

$$L(y, \hat{y})$$

There are many definitions of loss functions!

The choice of loss function:

- Affects the accuracy and computational cost of estimation.
- Depends on the estimation task:
  - Are outputs quantitative or qualitative?
  - Do we care about outliers?
  - Are all errors equally costly? (e.g., false negative on cancer test)

L2 Loss or Squared Loss

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

- Widely used.
- Also called "L2 loss".
- Reasonable:
  - $\hat{y} = y$ → good prediction → good fit → no loss
  - $\hat{y}$ far from y → bad prediction → bad fit → *lots of loss*

L1 Loss or Absolute Loss

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

- Sounds worse than it is.
- Also called "L1 loss".
- Reasonable:
  - $\hat{y} = y$ → good prediction → good fit → no loss
  - $\hat{y}$ far from y → bad prediction → bad fit → *some loss*

Squared Loss (L2 Loss)

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

For an SLR model $\hat{y} = \theta_0 + \theta_1 x$:

$$L(y, \hat{y}) = (y - (\theta_0 + \theta_1 x))^2$$

Absolute Loss (L1 Loss)

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

For an SLR model $\hat{y} = \theta_0 + \theta_1 x$:

1. What is the SLR L1 Loss?

2. Why don't we directly use residual error as the loss function? $(y - \hat{y})$

3. Which loss function is better: L1 or L2?

L2 and L1 Loss for SLR

Squared Loss (L2 Loss)

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

For an SLR model $\hat{y} = \theta_0 + \theta_1 x$:

$$L(y, \hat{y}) = (y - (\theta_0 + \theta_1 x))^2$$

Absolute Loss (L1 Loss)

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

For an SLR model $\hat{y} = \theta_0 + \theta_1 x$:

1

$$L(y, \hat{y}) = |y - (\theta_0 + \theta_1 x)|$$

Squared Loss (L2 Loss)

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

For an SLR model $\hat{y} = \theta_0 + \theta_1 x$:

$$L(y, \hat{y}) = (y - (\theta_0 + \theta_1 x))^2$$

Absolute Loss (L1 Loss)

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

For an SLR model $\hat{y} = \theta_0 + \theta_1 x$:

1 $$L(y, \hat{y}) = |y - (\theta_0 + \theta_1 x)|$$

Why don't we directly use residual error as the loss function? $e = (y - \hat{y})$ 2

L2 and L1 Loss for SLR

Squared Loss (L2 Loss)

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

For an SLR model $\hat{y} = \theta_0 + \theta_1 x$:

$$L(y, \hat{y}) = (y - (\theta_0 + \theta_1 x))^2$$

Absolute Loss (L1 Loss)

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

For an SLR model $\hat{y} = \theta_0 + \theta_1 x$:

1 $$L(y, \hat{y}) = |y - (\theta_0 + \theta_1 x)|$$

Why don't we directly use residual error as the loss function? $e = (y - \hat{y})$ 2

- Doesn't work: big negative residuals shouldn't cancel out big positive residuals!

L2 and L1 Loss for SLR

Squared Loss (L2 Loss)

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

For an SLR model $\hat{y} = \theta_0 + \theta_1 x$:

$$L(y, \hat{y}) = (y - (\theta_0 + \theta_1 x))^2$$

Absolute Loss (L1 Loss)

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

For an SLR model $\hat{y} = \theta_0 + \theta_1 x$:

1

$$L(y, \hat{y}) = |y - (\theta_0 + \theta_1 x)|$$

Why don't we directly use residual error as the loss function? $e = (y - \hat{y})$

2

- Doesn't work: big negative residuals shouldn't cancel out big positive residuals!

Which loss function is better: L1 or L2?

L2 penalizes larger residuals more.

3

Empirical Risk is Average Loss over Data

We care about how bad our model's predictions are for our entire data set, not just for one point. A natural measure, then, is of the **average loss** (aka **empirical risk**) across all points**.**

Given data $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$:

$$\hat{R}(\theta) = \frac{1}{n}\sum_i L(y_i, \hat{y}_i)$$

Function of the parameter θ (holding the data fixed) because θ determines $\hat{y}$.

The average loss on the sample tells us how well it fits the data (not the population).

But hopefully these are close.

Empirical Risk is Average Loss over Data

We care about how bad our model's predictions are for our entire data set, not just for one point. A natural measure, then, is of the **average loss** (aka **empirical risk**) across all points.

Given data $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$:

$$\hat{R}(\theta) = \frac{1}{n} \sum_i L(y_i, \hat{y}_i)$$

L2 loss → **Mean Squared Error (MSE)** $\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

L1 loss → **Mean Absolute Error (MAE)** $\frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$

The colloquial term for average loss depends on which loss function we choose.

1. Choose a model

How should we represent the world?

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

SLR model

2. Choose a loss function

How do we quantify prediction error?

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

Squared loss

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

MSE for SLR

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?

The combination of model + loss that we focus on today is known as **least squares regression**.

1. Choose a model

How should we represent the world?

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

SLR model

2. Choose a loss function

How do we quantify prediction error?

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

Squared loss

3. Fit the model

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

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

We want to find $\hat{\theta}_0, \hat{\theta}_1$ that minimize this **objective function**.

4. Evaluate model performance

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

Minimizing Average Loss on Data

Minimizing MSE for the SLR Model

Recall: we wanted to pick the **regression line**

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

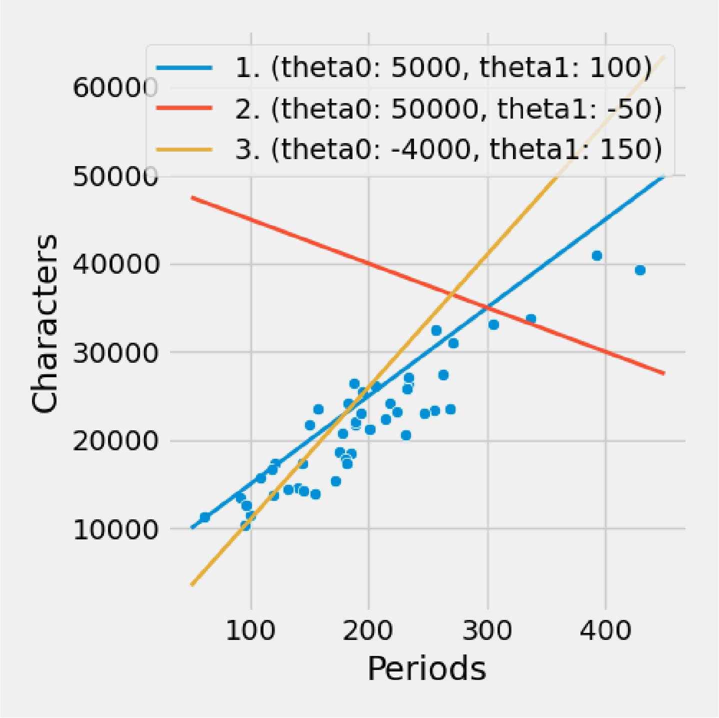
To minimize the (sample) **Mean Squared Error**:

$$\hat{R}(\theta) = \frac{1}{n}\sum_i L(y_i, \hat{y}_i)$$

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

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1 x_i))^2$$

To find the best values, we **take derivatives** with respect to the choice variables θ_0, θ_1



For every chapter of the novel *Little Women*, Estimate the **# of characters** $\hat{y}$ based on the **# of periods** in that chapter.

Minimizing MSE for the SLR Model

Recall: we wanted to pick the **regression line** $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$

To minimize the (sample) **Mean Squared Error**: $MSE(\theta_0, \theta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1 x_i))^2$

To find the best values, we set derivatives equal to zero to **obtain the optimality conditions:**

$$\frac{\partial}{\partial \theta_0} MSE = 0$$

$$\frac{\partial}{\partial \theta_1} MSE = 0$$

Recall: we wanted to pick the **regression line** $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$

To minimize the (sample) **Mean Squared Error**: $MSE(\theta_0, \theta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1 x_i))^2$

To find the best values, we set derivatives equal to zero to **obtain the optimality conditions:**

$$0 = \frac{\partial}{\partial \theta_0} MSE = -\frac{2}{n}\sum_{i=1}^{n} y_i - \theta_0 - \theta_1 x_i \iff \frac{1}{n}\sum_{i} y_i - \hat{y}_i = 0 \qquad \boxed{1}$$

"Equivalent"

$$0 = \frac{\partial}{\partial \theta_1} MSE = -\frac{2}{n}\sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i) x_i \iff \frac{1}{n}\sum_{i} (y_i - \hat{y}_i) x_i = 0 \qquad \boxed{2}$$

To find the best θ_0, θ_1, we need to solve the **estimating equations** on the right.

Goal: Choose θ_0, θ_1 to solve two estimating equations:

$$\frac{1}{n}\sum_i y_i - \hat{y}_i = 0 \quad \boxed{1} \qquad \text{and} \qquad \frac{1}{n}\sum_i (y_i - \hat{y}_i)x_i = 0 \quad \boxed{2}$$

$\boxed{1}$

$$\frac{1}{n}\sum_i (y_i - \theta_0 - \theta_1 x_i) = 0 \iff \overbrace{\left(\frac{1}{n}\sum_i y_i\right)}^{\bar{y}} - \theta_0 - \theta_1 \overbrace{\left(\frac{1}{n}\sum_i x_i\right)}^{\bar{x}} = 0$$

$$\iff \bar{y} - \theta_0 - \theta_1 \bar{x} = 0$$

$$\iff \theta_0 = \bar{y} - \theta_1 \bar{x}$$

Goal: Choose θ_0, θ_1 to solve two estimating equations:

$$\frac{1}{n}\sum_i y_i - \hat{y}_i = 0 \quad \boxed{1} \qquad \text{and} \qquad \frac{1}{n}\sum_i (y_i - \hat{y}_i)x_i = 0 \quad \boxed{2}$$

Now, let's try: $\boxed{2}$ - $\boxed{1}$ * $\bar{x}$

$$\frac{1}{n}\sum_i (y_i - \hat{y}_i)x_i - \frac{1}{n}\sum_i (y_i - \hat{y}_i)\bar{x} = 0 \iff \frac{1}{n}\sum_i (y_i - \hat{y}_i)(x_i - \bar{x}) = 0$$

$$(using\ \hat{y}_i = \theta_0 + \theta_1 x_i) \Rightarrow \frac{1}{n}\sum_i (y_i - \theta_0 - \theta_1 x_i)(x_i - \bar{x}) = 0$$

$$(using\ \theta_0 = \bar{y} - \theta_1 \bar{x}) \Rightarrow \frac{1}{n}\sum_i (y_i - \bar{y} + \theta_1 \bar{x} - \theta_1 x_i)(x_i - \bar{x}) = 0$$

$$\Rightarrow \frac{1}{n}\sum_i (y_i - \bar{y} - \theta_1 (x_i - \bar{x}))(x_i - \bar{x}) = 0$$

1. Choose a model

How should we represent the world?

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

SLR model

2. Choose a loss function

How do we quantify prediction error?

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

Squared loss

3. Fit the model

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

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

MSE for SLR

4. Evaluate model performance

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

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x \quad \begin{cases} \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \\ \hat{\theta}_1 = r\dfrac{\sigma_y}{\sigma_x} \end{cases}$$

Interpreting SLR: Slope

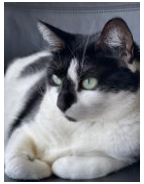
Interpreting the Least Squares Linear Regression Model

You may sometimes hear the prediction task defined as: "**regressing** y on x."

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x \quad \begin{cases} \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \\ \hat{\theta}_1 = r \dfrac{\sigma_y}{\sigma_x} \end{cases}$$

Suppose we fit a model that predicts a Chihuahua's weight (in pounds) given its length (in inches).

```
predicted weight = 2 + 0.5 * length
```



Interpreting the slope?

By definition, the slope measures the increase in y (pounds) for a 1 unit increase in x (1 inch).

1. Does this mean that if a cat in the dataset grows 1 inch, we estimate that they will get 0.5 pounds heavier? What does it actually mean? **No!**

- The model we created shows **association**, not causation.
- The data we collected is a snapshot of several cats at one instance of time (**cross-sectional**), not snapshots of cats over time (**longitudinal**).

Slope interpretation: If two cats have a 1 inch height difference, their estimated weight difference is 0.5 lbs.

Evaluating the Model: RMSE, Residual Plot

Lecture 10, Data 100 Spring 2023

Evaluating Models

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.

1. **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)

1. **Visualization:**

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