Simple Linear Regression

Motivation

- Have data (sample, x’s)
- Want to know likely value of next observation
  - E.g., playtime versus skins owned
- A – reasonable to compute mean (with confidence interval)
- B – could do same, but there appears to be relationship between X and Y!
  ⇒ Predict B e.g., “trendline” (regression)
Motivation

• Have data (sample, x’s)
• Want to know likely value of next observation
  – E.g., playtime versus skins owned
• A – reasonable to compute mean (with confidence interval)
• B – could do same, but there appears to be relationship between X and Y!

→ Predict B
  e.g., “trendline” (regression)
Motivation

• Have data (sample, x’s)
• Want to know likely value of next observation
  – E.g., playtime versus skins owned
• A – reasonable to compute mean (with confidence interval)
• B – could do same, but there appears to be relationship between X and Y!
  → Predict B
  e.g., “trendline” (regression)

Overview

• Broadly, two types of prediction techniques:
  1. Regression – mathematical equation to model, use model for predictions
     – We’ll discuss simple linear regression
  2. Machine learning – branch of AI, use computer algorithms to determine relationships (predictions)
     – CS 453X Machine Learning
Types of Regression Models

- Explanatory variable *explains* dependent variable
  - Variable X (e.g., skill level) explains Y (e.g., KDA)
  - Can have 1 or 2+
- Linear if coefficients added, else Non-linear

Outline

- Introduction *(done)*
- Simple Linear Regression *(next)*
  - Linear relationship
  - Residual analysis
  - Fitting parameters
- Measures of Variation
- Misc
Simple Linear Regression

- Goal – find a linear relationship between to values
  - E.g., kills and skill, time and car speed
- First, make sure relationship is linear! How?

\[ \text{Scatterplot} \]

- (c) no clear relationship
- (b) not a linear relationship
- (a) linear relationship – proceed with linear regression
Linear Relationship

- From algebra: line in form $Y = mX + b$
  - $m$ is slope, $b$ is y-intercept
- Slope ($m$) is amount $Y$ increases when $X$ increases by 1 unit
- Intercept ($b$) is where line crosses y-axis, or where $y$-value when $x = 0$

$$Y = mX + b$$

Simple Linear Regression Example

- Size of house related to its market value.
  - $X =$ square footage
  - $Y =$ market value ($$
- Scatter plot (42 homes) indicates linear trend
Simple Linear Regression Example

• Two possible lines shown below (A and B)
• Want to determine best regression line
• Line A looks a better fit to data
  – But how to know?

Line that gives best fit to data is one that minimizes prediction error
→ Least squares line (more later)
Simple Linear Regression Example

**Chart**

- Scatterplot
- Right click ➔ Add Trendline

![Chart Image]

**Formulas**

=\text{SLOPE}(C4:C45, B4:B45)

- Slope = 35.036

=\text{INTERCEPT}(C4:C45, B4:B45)

- Intercept = 32,673

- Estimate \( Y \) when \( X = 1800 \) square feet

\[ Y = 32,673 + 35.036 \times (1800) = \$95,737.80 \]
Simple Linear Regression Example

• Market value = 32673 + 35.036 x (square feet)
• Predicts market value better than just average

But before use, examine residuals

Outline

• Introduction (done)
• Simple Linear Regression
  – Linear relationship (done)
  – Residual analysis (next)
  – Fitting parameters
• Measures of Variation
• Misc
Residual Analysis

• Before predicting, confirm that linear regression assumptions hold
  – Variation around line is normally distributed
  – Variation equal for all $X$
  – Variation independent for all $X$

• How? Compute residuals (error in prediction) → Chart

[Diagram showing predicted vs actual and residuals plots]

Note that we've colored a few dots in orange so you can get the sense of how this transformation works.
Residual Analysis – **Good**

- Clustered towards middle
- Symmetrically distributed
- No clear pattern

Residual Analysis – **Bad**

- Clear shape
- Outliers
- Patterns

Note: could do normality test (QQ plot)
Residual Analysis – Summary

- Regression assumptions:
  - Normality of variation around regression
  - Equal variation for all y values
  - Independence of variation

  (a) ok
  (b) funnel
  (c) double bow
  (d) nonlinear

Outline

- Introduction (done)
- Simple Linear Regression
  - Linear relationship (done)
  - Residual analysis (done)
  - Fitting parameters (next)
- Measures of Variation
- Misc
Linear Regression Model

\[ Y_i = b_0 + mX_i + \varepsilon_i \]

\[ \varepsilon_i = \text{random error} \]

Fitting the Best Line

- Plot all \((X_i, Y_i)\) Pairs
Fitting the Best Line

• Plot all \((X_i, Y_i)\) Pairs
• Draw a line. But how do we know it is best?

Fitting the Best Line

• Plot all \((X_i, Y_i)\) Pairs
• Draw a line. But how do we know it is best?
Fitting the Best Line

- Plot all \((X_i, Y_i)\) Pairs
- Draw a line. But how do we know it is best?

Slope unchanged

Intercept changed

Slope changed

Intercept changed
Linear Regression Model

- Relationship between variables is linear function

\[ Y_i = b_0 + mX_i + \varepsilon_i \]

- Dependent (response) Variable (e.g., kills)
- Independent (explanatory) Variable (e.g., skill level)

Least Squares Line

- Want to minimize difference between actual y and predicted ŷ
  - Add up \( \varepsilon_i \) for all observed y's
  - But positive differences offset negative ones
  - (remember when this happened for variance?)

\[ \text{Minimize: } \sum (Y_i - \hat{Y}_i)^2 \]

Take derivative
Set to 0 and solve

https://cdn-images-1.medium.com/max/1600/1*AwC1WRm7jtldUcNMJTWmiA.png

Take derivative
Set to 0 and solve

https://i.imgur.com/3vDxDG7.png
Least Squares Line Graphically

Least Squares minimizes

$$\sum_{i=1}^{n} \hat{e}_i^2 = \hat{e}_1^2 + \hat{e}_2^2 + \hat{e}_3^2 + \hat{e}_4^2$$

Least Squares Line Graphically

- Line of Best Fit: Click the circle at the left to Show/Hide. Drag RED dots to position the line.
- Residuals: Click the circle at the left to Show/Hide.
- Squares: Click the circle at the left to Show/Hide.
- Least Squares Regression Line: Click the circle at the left to Show/Hide.

https://www.desmos.com/calculator/zvrc4lg3cr
Outline

• Introduction (done)
• Simple Linear Regression (done)
• Measures of Variation (next)
  – Coefficient of Determination
  – Correlation
• Misc

Measures of Variation

• Several sources of variation in y
  – Error in prediction (unexplained)
  – Variation from model (explained)
Sum of Squares of Error

- Least squares regression selects line with lowest total sum of squared prediction errors
- Sum of Squares of Error, or \( \text{SSE} \)
- Measure of unexplained variation

Sum of Squares Regression

- Differences between prediction and population mean
  - Gets at variation due to \( X \) & \( Y \)
- Sum of Squares Regression, or \( \text{SSR} \)
- Measure of explained variation
**Sum of Squares Total**

- Total Sum of Squares, or SST = SSR + SSE

\[
SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \quad \text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2
\]

\[
SSR = SST - SSE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2
\]

**Coefficient of Determination**

- Proportion of total variation (SST) explained by the regression (SSR) is known as the **Coefficient of Determination** \((R^2)\)

\[
R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}
\]

- Ranges from 0 to 1 (often said as a percent)
  - 1 – regression explains all of variation
  - 0 – regression explains none of variation
Coefficient of Determination – Visual Representation

\[ R^2 = 1 - \text{Variation in observed data model cannot explain (error)} \]

Total variation in observed data

Coefficient of Determination Example

• How “good” is regression model? Roughly:
  
  \[
  0.8 \leq R^2 \leq 1 \quad \text{strong}
  
  0.5 \leq R^2 < 0.8 \quad \text{medium}
  
  0 \leq R^2 < 0.5 \quad \text{weak}
  \]
How “good” is the Regression Model?

I don’t trust linear regressions when it’s harder to guess the direction of the correlation from the scatter plot than to find new constellations on it.

https://xkcd.com/1725/

Relationships Between X & Y

Strong relationships

Weak relationships
Relationship Strength and Direction – Correlation

- **Correlation** measures strength and direction of linear relationship
  - -1 perfect neg. to +1 perfect pos.
  - Sign is same as regression slope
  - Denoted $R$. Why? $R = \sqrt{R^2}$

### Pearson’s Correlation Coefficient

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Where, $\bar{x}$. mean of X variable

$\bar{y}$. mean of Y variable

### Correlation Examples (1 of 3)

- $r = -1$
- $r = -0.6$
- $r = 0$
- $r = +0.3$
- $r = +1$
Correlation Examples (3 of 3)

Anscombe’s Quartet

Summary stats:
Mean_x 9
Mean_y 7.5
Var_x 11
Var_y 4.125
Model: y=0.5x+3

Correlation Summary
Correlation is not Causation

Buying sunglasses causes people to buy ice cream?

Correlation is not Causation

Importing lemons causes fewer highway fatalities?
Correlation is not Causation

https://science.sciencemag.org/content/sci/348/6238/980.2/F1.large.jpg?width=800&height=600&carousel=1

Correlation is not Causation

I USED TO THINK
CORRELATION IMPLIED
CAUSATION.

THEN I TOOK A
STATISTICS CLASS.
NOW I DON'T.

SOUNDS LIKE THE
CLASS HELPED.
WELL, MAYBE.

https://xkcd.com/552/
Outline

• Introduction (done)
• Simple Linear Regression (done)
• Measures of Variation (done)
• Misc (next)

Extrapolation versus Interpolation

• Prediction
  – Interpolation – within measured X-range
  – Extrapolation – outside measured X-range
Be Careful When Extrapolating

If extrapolate, make sure have reason to assume model continues.

Prediction and Confidence Intervals (1 of 2)
Prediction and Confidence Intervals (2 of 2)

Beyond Simple Linear Regression

- Multiple regression – more parameters beyond just $X$
  - Book Chapter 11
- More complex models – beyond just $Y = mX + b$
More Complex Models

• Higher order polynomial model has less error
  → A “perfect” fit (no error)
• How does a polynomial do this?

Graphs of Polynomial Functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Degree</th>
<th>Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant Function</td>
<td>0</td>
<td><img src="https://cdn-images-1.medium.com/max/2400/1*pjIp920-MZdS_3fLVhf-Dw.jpeg" alt="Constant Function Graph" /></td>
</tr>
<tr>
<td>Linear Function</td>
<td>1</td>
<td><img src="https://cdn-images-1.medium.com/max/2400/1*pjIp920-MZdS_3fLVhf-Dw.jpeg" alt="Linear Function Graph" /></td>
</tr>
<tr>
<td>Quadratic Function</td>
<td>2</td>
<td><img src="https://cdn-images-1.medium.com/max/2400/1*pjIp920-MZdS_3fLVhf-Dw.jpeg" alt="Quadratic Function Graph" /></td>
</tr>
<tr>
<td>Cubic Function</td>
<td>3</td>
<td><img src="https://cdn-images-1.medium.com/max/2400/1*pjIp920-MZdS_3fLVhf-Dw.jpeg" alt="Cubic Function Graph" /></td>
</tr>
<tr>
<td>Quartic Function</td>
<td>4</td>
<td><img src="https://cdn-images-1.medium.com/max/2400/1*pjIp920-MZdS_3fLVhf-Dw.jpeg" alt="Quartic Function Graph" /></td>
</tr>
<tr>
<td>Quintic Function</td>
<td>5</td>
<td><img src="https://cdn-images-1.medium.com/max/2400/1*pjIp920-MZdS_3fLVhf-Dw.jpeg" alt="Quintic Function Graph" /></td>
</tr>
</tbody>
</table>

Higher degree, more potential “wiggles”
But should you use?
Underfit and Overfit

- **Overfit** analysis matches data too closely with more parameters than can be justified
- **Underfit** analysis does not adequately match data since parameters are missing
  - Both models do not predict well (i.e., for non-observed values)
- **Just right** – fit data well “enough” with as few parameters as possible