Simple Linear Regression

Chapter 10

Even You Can Learn STATISTICS and ANALYTICS
An Easy to Understand Guide to Statistics and Analytics

David M. Levine and David F. Stephan
Motivation

- Have data (sample, x’s. e.g., playtime)
- Want to know likely value of next observation (A)

A. Compute mean y-value (with confidence interval)
→ Predict A

- But what if have additional information?
  E.g., playtime versus skins owned
→ Better prediction!
Motivation

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

Broadly, two types of prediction techniques:

1. **Regression** – mathematical equation to model, then use model for predictions
   – We’ll discuss *simple linear regression*

2. **Machine learning** – branch of AI, use computer algorithms to determine relationships (predictions)
   – CS 4342 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 (simple) or 2+ (multiple)

• 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** (line) relationship between two values
  – E.g., *travel time* and *car speed*, *KDA* and *skill*,
• First, make sure relationship is linear! How?
  → Scatterplot
  
  (c) no clear relationship
  
  (b) not a linear relationship
  
  (a) **linear relationship** – proceed with linear regression

![Graphs](image-url)
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 (specifying units important!)
- Intercept ($b$) is where line crosses y-axis, or where $y$-value when $x = 0$
Simple Linear Regression Example

• Market value of house is related to size
  \( X = \text{square footage} \)
  \( Y = \text{market value (\$)} \)

• 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?

\[ Y = mX + b \]
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)

\[ Y = mX + b \]
Simple Linear Regression Example

- Scatterplot
- Right click → Add Trendline
Simple Linear Regression Example

Formulas

=SLOPE(C4:C45,B4:B45)

→ Slope = 35.04

=INTERCEPT(C4:C45,B4:B45)

→ Intercept = 32,600

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

\[
Y = 32600 + 35.04 \times (1800) = \$95,672
\]
Simple Linear Regression Example

**Market value** = 32600 + 35.04 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)

→ Residuals Chart
Residual Analysis

Variation around line normally distributed?
Variation equal for all X
Variation independent for all X?

No clear pattern
Residual Analysis – Good


Symmetrically distributed
No clear pattern
Clustered towards middle, ok since need normally distributed
Residual Analysis – Bad


- Clear shape
- Outliers
- Patterns!

Note: could do normality test (QQ plot)
Outline

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

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

\( \varepsilon_i = \) random error

Random error associated with each observation (Residual)

https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression
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?

https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression
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?
Linear Regression Model

- Relationship between variables is linear function

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

- Population Y-Intercept
- Population Slope
- Random Prediction Error
- Dependent (response) Variable (e.g., kills)
- Independent (explanatory) Variable (e.g., skill level)

Want error as small as possible
Least Squares Line

• Want to minimize difference between actual $y$ and predicted $\hat{y}$
  – Add up $\varepsilon_i$ for all observed $y$’s
  – But positive differences offset negative ones!
  – (remember when this happened for variance?)
→ Square the errors! Then, minimize (using Calculus)

![Simple Linear Regression Diagram](https://cdn-images-1.medium.com/max/1600/1*AwC1WRm7jtdUcNMjTWmIA.png)

Minimize:

$$
\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 
$$

Take derivative
Set to 0 and solve
Least Squares (LS) Line Graphically

Least Squares minimizes

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

$$Y_2 = \hat{\beta}_0 + \hat{\beta}_1 X_2 + \hat{\varepsilon}_2$$

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$
Least Squares Line Graphically – Interactive Demo

Create new situations moving the green data points about the graph.

- 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

Try it out yourself!
https://web.cs.wpi.edu/~imgd2905/d24/groupwork/12-least-squares/handout.html
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 (SSE)

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

- Differences between prediction and population mean
  - Gets at variation due to X & Y
- Sum of Squares Regression, or SSR
- Measure of explained variation
Sum of Squares Total

- Total Sum of Squares, or \( \text{SST} = \text{SSR} + \text{SSE} \)

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

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

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

Smaller is better

Total variation in observed data

Larger is better

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

Smaller is better

Total variation in observed data

Larger is better

• How “good” is regression model? Roughly:

\[ 0.8 \leq R^2 \leq 1 \quad \text{strong} \]
Coefficient of Determination Example

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

$I \text{ DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.}$

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? Square \( R = R^2 \)

Pearson’s Correlation Coefficient

\[
\rho = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}
\]

Where, \( \bar{x} \) = mean of X variable
\( \bar{y} \) = mean of Y variable

- Vary together
- Vary Separately

Correlation Examples

- $r = -1$
- $r = -0.6$
- $r = 0$
- $r = +0.3$
- $r = +1$
Groupwork

• Introduction
  – Icebreaker: What game are you looking forward to playing this summer?

• Groupwork
  – Think, discuss, write down – qualtrics

• Correlation
  – Consider scatterplots
  – Estimate correlation

Correlation Examples

\[
\begin{array}{cccccccc}
1 & 0.8 & 0.4 & 0 & -0.4 & -0.8 & -1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\end{array}
\]

https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Correlation_examples2.svg/1200px-Correlation_examples2.svg.png
Correlation Examples
Correlation Examples

https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Correlation_examples2.svg/1200px-Correlation_examples2.svg.png
Correlation Examples

(Note, would want to use residual analysis before using predictions!)
Correlation Examples

Anscombe’s Quartet

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

R^2 = 0.69

(Note, would want to use residual analysis before using predictions!)
Correlation Summary

No Correlation

Perfect Positive Correlation

High Positive Correlation

Low Positive Correlation

Low Negative Correlation

High Negative Correlation

Perfect Negative Correlation

https://www.mathsisfun.com/data/correlation.html
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

- Aggregate comic book sales
- Computer science doctorates awarded
- Injuries related to falling televisions
- Undergrad enrollment at U.S. universities
- Tornadoes
- Shark attacks
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)

95% of values within

95% of regressions within
Prediction and Confidence Intervals
(2 of 2)
Multiple Independent Variables

• Chronic heart disease (CHD) correlates with smoking
  – $R^2 = 0.5$

• But what about other 50%?

• Correlation with exercise? Cholesterol?
Single Linear Regression → Multiple Linear Regression

- Use several independent variables to predict dependent variable
- Weights each predictor based on strength of relationship
- Makes adjustments for inter-relationships among predictors
- Gives overall fit ($R^2$)
- Note: Need independent variables not highly related to each other

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \ldots + b_nX_n$$
Multiple Linear Regression

Example: hours studied and pre-tests affect final score

\[ y = b_0 + b_1 \times X_1 + b_2 \times X_2 + E \]
Multiple Linear Regression Example (1 of 2)

- Hours studied and prep exams taken → exam score

20 students
Multiple Linear Regression Example (2 of 2)

- Independent variable
- Covers both independent variables
Interpret

- $R^2 = 0.734$
- Overall significant ($p < 0.05$)
- Hours significant
- Prep exams not significant
- Base score without prep 67.67
- Each hour gains 5.56 percent

Score = $67.67 + 5.56 \times \text{hours} - 0.60 \times \text{prep\_exams}$
Beyond Linear Regression

- More complex models – beyond just linear

\[ Y = mX + b \]
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

- Constant Function (degree = 0)
- Linear Function (degree = 1)
- Quadratic Function (degree = 2)
- Cubic Function (deg. = 3)
- Quartic Function (deg. = 4)
- Quintic Function (deg. = 5)

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 fit well, but do not *predict* well (i.e., for non-observed values).

- **Just right** – fit data well “enough” with as few parameters as possible (*parsimonious* - desired level of prediction with as few terms as possible).

**Test → Cross Validation**
Cross Validation (1 of 2)

Use to build model

Compute accuracy
Cross Validation (2 of 2)

Repeat for different slices

- **Overfit** and **Underfit** will both have lower accuracy than “just right”
Summary

• Can use **regression** to predict unmeasured values

• Before fit
  – Visual relationship (**scatter plot**) and residual analysis

• Strength of fit – **$R^2$** and correlation (**$R$**)

• Beware
  – Correlation is not causation
  – Extrapolation

• Higher order, more complex models can fit better
  – Beware of overfit $\rightarrow$ less predictive power