Linear Regression

David Armstrong

UCI

Linear Regression

Scatterplots

- May show a relationship or an association between two quantitative variables.
- We are looking for a LINEAR relationship between our two variables.
- Explanatory Variable

  x-variable

  predictor
- Response Variable

  y-variable

  variable of interest

The Linear Regression Model

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$

- β_0

  y-intercept for the population (the true y-intercept)

  When x is equal to 0, then y is equal to β_0.

- β_1

  slope for the population (the true slope)

  A one "unit" increase in x, is associated with a β_1 "unit" INCREASE/DECREASE in y, on average.

- ϵ_i

  random error for the ith subject

  This includes other factors that affect the response variable

GIVEN: Suppose we have data (x_i, y_i) on n individuals.

GOAL: To find a line relating y to x.

NOTE: We need to estimate our parameters based on this dataset.

LEAST SQUARES REGRESSION LINE: LSRL

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, \ldots, n$

- $\hat{\beta}_0$

  the estimated y-intercept

- $\hat{\beta}_1$

  the estimated slope

- $e_i = y_i - \hat{y}_i$

  the residual for the ith subject (the estimated random error for the ith subject)

The Linear Regression Model

- The LSRL describes how a response variable y changes as an explanatory variable x changes.
- Best-fitting line to the data
- Minimizes the (vertical) distances of your observations (data) from your line
  $SSE = \sum_{i=1}^{n} e_i^2$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values to make the SSE the smallest
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimates of β_0 and β_1

Formulas

estimated slope	$\hat{\beta}_1 = \dfrac{s_{xy}}{s_{xx}}$
estimated y-intercept	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
sum of x deviations squared	$s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \dfrac{(\sum_{i=1}^{n} x_i)^2}{n}$
sum of y deviations squared	$s_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \dfrac{(\sum_{i=1}^{n} y_i)^2}{n}$
sum of x deviations times y deviations	$s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \dfrac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$
sum of squared errors	$SSE = \sum_{i=1}^{n} e_i^2 = s_{yy} - \dfrac{s_{xy}^2}{s_{xx}}$
sample mean of x	$\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
sample mean of y	$\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$
residual for the ith subject	$e_i = y_i - \hat{y}_i$
LSRL	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Linear Regression Model Assumptions

The residuals from our model are used to check the model assumptions.

- ϵ_i are independent

  Check that the experimental units are independent

- $\epsilon_i \sim Normal(0, \sigma_\epsilon)$

- When we do a residual plot, it should look random.

Residual Standard Deviation

The residual standard deviation measures the typical vertical distance of the data points from the regression line. In practice we do not know the value of σ_ϵ, but we can estimate the residual standard deviation:

$$\hat{\sigma}_\epsilon = s_\epsilon = \sqrt{\frac{SSE}{n-(k+1)}}$$

$k =$ the number of predictors in the model.

For Simple Linear Regression, there is only one predictor (i.e. $k = 1$).

The Distribution of the y-intercept

The Distribution of the Y-INTERCEPT:

$$\hat{\beta}_0 \sim ApproximatelyNormal\left(\mu_{\hat{\beta}_0} = \beta_0, \sigma_{\hat{\beta}_0} = \sigma_\epsilon\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}\right)$$

If we estimate σ_ϵ using s_ϵ, then we have the following result:

$$\hat{\beta}_0 \sim t\left(\mu_{\hat{\beta}_0} = \beta_0, SE_{\hat{\beta}_0} = s_\epsilon\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}, df = n-(k+1)\right)$$

The Distribution of the slope

The Distribution of the SLOPE:

$$\hat{\beta}_1 \sim ApproximatelyNormal\left(\mu_{\hat{\beta}_1} = \beta_1, \sigma_{\hat{\beta}_1} = \sigma_\epsilon \sqrt{\frac{1}{s_{xx}}}\right)$$

If we estimate σ_ϵ using s_ϵ, then we have the following result:

$$\hat{\beta}_1 \sim t\left(\mu_{\hat{\beta}_1} = \beta_1, SE_{\hat{\beta}_1} = s_\epsilon \sqrt{\frac{1}{s_{xx}}}, df = n - (k+1)\right)$$

Confidence Interval for β_1

We are _% confident that a 1 “unit” increase in x is associated with a somewhere between a CI_{lower} “unit” increase/decrease and CI_{upper} “unit” increase/decrease in y.

$$\hat{\beta}_1 \pm t_{crit} SE_{\hat{\beta}_1}$$

Hypothesis Test for β_1

$H_o : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

$$t_{TS} = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}}$$

Fail to Reject H_o: There is not a linear relationship between x and y (x is not predictive of y).

Reject H_o: There is a linear relationship between x and y (x is predictive of y).

Linear Regression

Scatterplots

- Strength of the relationship

  Strong

  Moderate

  Weak

- Direction of the relationship

  Positive

  Negative

- Form

  Linear

  Non-Linear

  Outliers

  Sub-groups

Do the points follow a single stream that is tight to the line?

Is there considerable spread (or variability) around the line?

Correlation Coefficient

r

- unit-less measurement

  Because we use z-scores, the correlation coefficient does not change when converting to different units.

- Correlation is sensitive to outliers.

  Outliers can dramatically change r.

  NOT RESISTANT to OUTLIERS

- Values between (and including) $-1 \leq r \leq 1$

  $r = 1$ means all the data points lie on a line. They have a positive slope and a positive association.

  $r = -1$ means all the data points lie on a line. They have a Negative slope and a negative association.

  If $r = 0$ the best fitting line has a slope of zero

- Values of r close to 0 means that the linear relationship is not a good fit to the data.

  There may be a general linear trend, but there is a lot of variability around that trend.

  There may be a relationship but it is NOT LINEAR

Describing the relationship based on the correlation

- Strong Negative

  $-1 \leq r < -0.75$

  "There is a strong negative linear relationship between **x** and **y**."

  "As **x** increases, **y** decreases on average."

- Moderate Negative

  $-0.75 \leq r < -0.35$

  "There is a moderate negative linear relationship between **x** and **y**."

  "As **x** increases, **y** decreases on average."

- Weak Negative

  $-0.35 \leq r < 0$

  "There is a weak negative linear relationship between **x** and **y**."

  "As **x** increases, **y** decreases on average."

- Weak Positive

  $0 < r \leq 0.35$

  "There is a weak positive linear relationship between **x** and **y**."

  "As **x** increases, **y** increases on average."

- Moderate Positive

  $0.35 < r \leq 0.75$

  "There is a moderate positive linear relationship between **x** and **y**."

  "As **x** increases, **y** increases on average."

- Strong Positive

  $0.75 < r \leq 1$

  "There is a strong positive linear relationship between **x** and **y**."

  "As **x** increases, **y** increases on average."

$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$

Cautions

- CORRELATION simply does NOT imply CAUSATION:

  May be a coincidence

  Both variables might be directly influenced by some common underlying lurking or confounding variable

- If the correlation is not strong, predictions will not be accurate.

- Extrapolation: making predictions outside of the range for which you have data.

  Do NOT extrapolate ever!

  Do not make predictions outside of the range for which you have data.

- A regression line is a straight line that models the relationship between an explanatory variable and a response variable.

  Therefore, it is only useful when one variable helps to predict the other.

  If the relationship is NOT linear and the correlation is NOT strong, then predictions will NOT be accurate.

- Be careful of different groups (subgroups) being combined in your regression.

- Look for unusual points

  High Leverage Points = X values far from $\bar{X}$.

  Influential Points = Removing this point from the data set results in a

  very different regression model.

  Outliers = Any data point that stands away from the others.

Residuals

- e_i = error for the ith subject

  Difference between observed y and predicted y

  $y_i - \hat{y}_i$

- For every given value of x

  We have an observed y

  We have a predicted $y = \hat{y}$

- Fact: $\sum_{i=1}^{n} e_i = 0$

- It is a "model" which is not perfect

  Some of the data points will be below the line.

  Overpredictions

  Some of the data points will be above the line.

  Underpredictions

Residual Plots

- Scatterplot of the (x_i, e_i) pairs.

- Horizontal line at $e = 0$

- The model is a GOOD fit if:

  the plot looks random

- Do NOT use linear regression if:

  Unusually large values for your residuals

  Non-linear patterns (curvature)

  Uneven variation (Fanning)

  Influential observations

Coefficient of Determination: r^2

- Measuring Predictive Power

  Is your line a good predictor of reality?

  How accurate predictions will be.

- "___% of the variation in **y** can be explained by **x**"

  Describes the connection between x and y

- Excellent

  $0.80 \leq r^2 \leq 1$

  "There is **EXCELLENT** predictive power."

- Good

  $0.50 \leq r^2 < 0.80$

  "There is **GOOD** predictive power."

- Fair

  $0.25 \leq r^2 < 0.50$

  "There is **FAIR** predictive power."

- Weak

  $0 \leq r^2 < 0.25$

  "There is **WEAK** predictive power."

$$r^2 = \frac{SSR}{SST} = \frac{s_{xy}^2}{s_{xx}s_{yy}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

EX: Education was measured as the percentage of residents aged at least 25 in the county who had at least a high school degree. Crime rate was measured as the number of crimes in Florida County in the past year per 1000 residents. The correlation coefficient between these variables is -0.67.

a) What is the coefficient of determination?

b) Give the definition and describe the strength of the predictive power based on the guidelines above.

Ex: Scatterplot of Systolic Blood Pressure versus Weight (Sample of 12 American Adults) yields a Pearson correlation of SBP and Weight = 0.971. The regression equation is: $\hat{y} = 1.1 + 0.764x$.

a) Suppose we know that one person in the sample weighs 188 pounds and has a systolic blood pressure of 136. Describe this data as an ordered pair.

b) Predict the SBP for a person weighing 188 pounds using your regression model.

d) Find the residual.

e) Interpret the correlation coefficient in context of the problem.

Ex: An international distance triathlon consists of a 1.5 km swim, a 40 km bike ride and a 10 km run. Triathletes are ranked based on their overall finishing times, and some people suggest that an athlete's time for the swim has the largest influence on his overall performance. Data from 10 male triathletes was analyzed to produce the results below:
The regression equation is: Overall Finishing Time = 122 + 1.56 (Swim Time)
1. The correct interpretation of the slope is:

A. For every one minute increase in overall finishing time, there is a 122 minute increase in swim time
B. For every one minute increase in swim time, there is a 1.56 minute increase in overall finishing time
C. For every one minute increase in swim time, there is a 122 minute increase in overall finishing time
D. For every one minute increase in overall finishing time, there is a 1.56 minute increase in swim time

2. One athlete completed the swim in 34 minutes. a. Calculate the athlete's predicted finishing time.

b. The athlete actually finished the race in 186.04 minutes. What is the difference between his observed finishing time and the predicted finishing time?