**Learning Objectives**

* What is linear regression?
* What are the goals of linear regression?
* What can linear regression accomplish?
* How are other procedures related to linear regression?
* <https://lindeloev.github.io/tests-as-linear/>

**What is simple linear regression?**

1. Equation for line: *y*=*β*0+*β*1*x*
2. Chart, line chart

   Description automatically generated
3. Have cloud of points

Chart, scatter chart

Description automatically generated

1. Fit line to cloud of points

Chart, scatter chart

Description automatically generated

1. Infer slope from fitted line

Chart, scatter chart

Description automatically generated

1. Inference:
   1. Test if slopes are 0
   2. Confidence intervals on slopes.
   3. Interpret sign/magnitude of slopes.
2. Infer slopes from fitted plane.
3. Inference:
   1. Test if slopes are 0
   2. Confidence intervals on slopes.
   3. Interpret sign/magnitude of slopes.

**Steps of a Regression Analysis**

* The above procedures assume that:
  1. The cloud of points roughly follows a line (or plane).
  2. All predictors (the *x*’s) are associated with the response (the *y*
  3. ). We might have many predictors and we need to choose which ones to include.
* We typically need to transform the data or try out a few models.
* Steps:

START Exploratory Data Analysis Develop one or more tentative regression models Is one or more of the regression models suitable for the data at hand? YES NO Revise regression models and/or develop new ones Identify most suitable model Make inferences one basis of regression model STOP

**What can you use it for?**

* Detecting trends.
  + Easy to see trends if you have two variables. Harder if you have more. Need something more sophisticated.
  + Linear regression allows us to say “folks that have bigger x tend to have bigger y”.
* Control for other variables.
  + “Folks that have the *same* z but *bigger* x tend to have bigger y.”
* Prediction
  + Most machine learning tasks in the read world are “small data”.
  + The fancy ML methods have many parameters that require lots of data to estimate.
  + Linear regression is often the best you can do in small data tasks.

**Generality**

* Many statistical procedures are special cases of (or approximations to) linear regression.
* Understanding linear regression really well will give you a deeper understanding of statistics in general.
* Procedures that are special cases of linear regression, or can be well approximated by linear regression:
  + One/two sample *t*
  + -test.
  + ANOVA
  + Correlation tests
  + Rank tests
  + Chi-square tests
  + Many others

**Learning Objectives**

* Chapter 1 of KNNL.
* Goals of regression analysis.
* The simple linear regression model.
* Least-squares approach to estimating parameters.
* The Ordinary Least Squares (OLS) estimates.

**Overview**

* **Observational/experimental Units**: The people/places/things/animals/groups that we collect information about. Also known as “individuals” or “cases”. Sometimes people just say “units”.
* **Variable**: A property of the observational/experimental units.
  + E.g.: height of a person, area of a country, marital status.
* **Value**: The specific level of a variable for an observational/experimental unit.
  + E.g.: Bob is 5’11’’, China has an area of 3,705,407 square miles, Jane is divorced.
* **Quantitative Variable**: The variable takes on numerical values where arithmetic operations (+/−/×/÷ ) make sense.
* E.g.: height, weight, area, income.
* Counterexample: Phone numbers, social security numbers.

 **Regression Analysis**: Study relationship between one or more *quantitative* variables.

 **Response Variable**:

* What we think is either caused by or explained by the predictor variable.
* Also called “outcome variable” and “dependent variable”.
* Usually denote this with the letter *y*
*  .

 **Predictor Variable**:

* What we think causes or explains the outcome variable.
* Also called a “feature”, “explanatory variable”, “independent variable”, and (when doing an experiment) a “treatment variable”.
* Typically have more than one predictor.
* Usually denote these with the letters *x*1, *x*2, *x*3, …, *xp.*

 Two quantitative variables can have either a **functional** or a **statistical** relationship.

 **Functional relationship**: There is an exact mathematical formula relating the value of one quantitative variable *x* (the predictor) to the other *y* (the response).

*y*=*f*(*x*)

*f*() is some function relating the correspondence of *x* to *y*

* E.g.: *x* = the radius of a circle and *y* = the area of a circle then

*y*=*πx*2

Chart, line chart

Description automatically generated

 **Statistical Relationship**:

* Is not a perfect relation.
* Functional + noise.
* E.g. “bread and peace” data from Chapter 7 of [ROS](https://avehtari.github.io/ROS-Examples/) looking at the statistical relationship between economic growth and vote-share of the incumbant for president.

 library(readr)

library(ggplot2)

hibbs <- read\_csv("https://dcgerard.github.io/stat\_415\_615/data/hibbs.csv")

qplot(x = growth, y = vote, data = hibbs) +

geom\_smooth(method = "lm", se = FALSE)

Chart, scatter chart

Description automatically generated

* Can say higher growth tends to have higher vote-share. But relationship is not perfect (but still pretty good!)
* The scattering of points about the line represents variation in vote share that is not associated with economic growth.

 Goals of Regression:

1. **Description**:
   1. Gene expression (*x*) is associated with height (*y* ).
   2. Biological sex (*x*) is associated with salary (*y*).

 **Prediction**:

1. Predict sales (*y*) based on product attributes (*x* ’s)
2. Predict crop yield (*y*) based on genomic markers (*x*’s)

* When we describe relationships, this does not imply causation.
  1. You need very special settings for “causal inference”, which we might cover later in the course.
  2. One special case where we can make causal claims is when we have a completely randomized experiment, where predictor values are randomly assigned.
* E.g. A researcher noticed that murder rates went up whenever ice cream consumption increased.

**The Simple Linear Regression Model**

* The model:

*Yi*=*β*0+*βiXi*+*ϵi*

* *Yi*: The response value for unit *i*
* *Xi*: The predictor value for unit *i*.
* *β*0: The *y* -intercept of the regression line.
* *β*1: The slope of the regression line.
* *ϵ*: The random noise of individual *i*
  + This is a random variable.
  + *E*[*ϵi*]=0 (mean zero).
  + *var*(*ϵi*)=*σ*2 (variance is the same for all *i*).
  + *cor*(*ϵi*,*ϵj*)=0 for all *i*≠*j* (uncorrelated errors).

 *Xi* and *Yi* are typically known. We usually have a sample of (*X*1,*Y*1),(*X*2,*Y*2),…,(*Xn*,*Yn*)

E.g. *Xi* could be the economic growth in year *i*, and *Yi* could be the incumbent vote-share for year *i*

 *β*0, *β*1 are called **parameters** and are typically not observed. They must be inferred from a sample of values (*X*1,*Y*1),(*X*2,*Y*2),…,(*Xn*,*Yn*)

 **Regression line**: *y*=*β*0+*β*1*x*

 Assumptions in decreasing order of importance:

1. **Linearity**: *E*[*Yi*|*Xi*]=*β*0+*β*1*Xi*
2. **Uncorrelated errors**: *cor*(*ϵi*,*ϵj*)=0 for all *i*≠*j*
3.  **Constant Variance**: *var*(*ϵi*)=*σ*2

 Note: Distribution of *Yi* is conditional on *Xi*

Chart, line chart

Description automatically generated

 **Exercise**: Suppose the regression model between two variables is

*Yi*=3+2*xi*+*ϵi*,  *var*(*ϵi*)=*σ*2

1. What is the mean of *Yi* if *Xi*=−1 ?
2. Suppose *Xi*=1 and *Yi*=4. What is *ϵi* ?

**Review of interpretation**

* *β*0: *Y*-intercept of the regression line.

 If 0 is in the range of the *Xi*’s, then can also interpret this as the value of *E*[*Yi*|*Xi*=0]. But cannot use this interpretation if 0 is outside of the *Xi* ’s.

* *β*1: Difference in average of *Yi*’s when *Xi*’s differ by 1.

Do **not** use implicitely causal language like “change” or “increase” or “decrease”.

 Relationships are **positive** if *β*1>0

(larger *x* tend to correspond to larger *y*).

Chart, line chart

Description automatically generated

 Relationships are **negative** if *β*1<0

(larger *x* tend to correspond to smaller *y*).

Chart, line chart

Description automatically generated

 Two variables are uncorrelated if *β*1=0

(the value of *x* does not matter, the value of *y* tends to stay the same).

Chart

Description automatically generated

 **Example**: In the bread and peace example, a regression line that fits the data well is *y*=46.25+3.06

* **Correct**: Years that show 1 percentage point more growth tend to have 3 percentage points larger vote shares for the incumbant. **Incorrect**: Incumbant vote-share increases 3 percentage points for each 1 percentage point increase in growth.

**Estimating Parameters**

* We have data (*X*1,*Y*1),(*X*2,*Y*2),…,(*Xn*,*Yn*)

, and we want to estimate *β*0 and *β*1

in the equation

*Yi*=*β*0+*β*1*Xi*+*ϵi*

 Idea: Try to get *Yi* as close to its mean. So we want each *Yi*−(*β*0+*β*1*Xi*) to be close to 0.

 To make all of these differences on average close to zero, consider minimizing the *sum of squares*:

∑*i*=(1,n),[*Yi*−(*β*0+*β*1*Xi*)]2

**Residuals**

* We often evaluate the performance of a model by looking at how far the fitted values are from the observed values.
* The *i*

th **residual** are

*ei*=*Yi*−*Y*^*i*.

 These are different from the model error terms, which are

*ϵi*=*Yi*−*E*[*Yi*|*Xi*]

* The residuals are the deviation from an *estimated* regression function, and so are known.
* The error terms are the deviation from the *unknown true* regression function, and so are unknown.
* Graphic:

Chart, scatter chart

Description automatically generated

* **Exercise**: In the above graphic, is the residual *positive* or *negative*? Did we underestimate or over estimate here?

**Properties of Fitted Regression Line**

**Text, letter

Description automatically generated**

**Estimating *σ*2**

**Text, letter

Description automatically generated**

**Example:**

**Let’s find the MSE for the following bivariate data sample.**

**x<- c(100,102,103,101,105,100,99,105)**

**y<- c(257,264,274,266,277,263,258,275)**

**-55.797 3.166**

**Y(hat) = -55.797 + 3.166X**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **i** | **Xi** | **Yi** | **Mean Response**  **Y(hat)i** | **Residual**  **Yi – Y(hat)i = ei** | **Squared Residual**  **ei2** |
| **1** | **100** | **257** | **260.803** | **-3.803** | **14.463** |
| **2** | **102** | **264** | **267.135** | **-3.135** | **9.828** |
| **3** | **103** | **274** | **270.301** | **3.699** | **13.683** |
| **4** | **101** | **266** | **263.969** | **2.031** | **4.125** |
| **5** | **105** | **277** | **276.633** | **.367** | **.135** |
| **6** | **100** | **263** | **260.83** | **2.17** | **4.709** |
| **7** | **99** | **258** | **257.637** | **.363** | **.132** |
| **8** | **105** | **275** | **276.633** | **-1.633** | **2.667** |
| **Total** |  |  |  |  | **49.742** |

**Calculations:**

**df = n – 2 = 8 – 2 = 6**

**MSE = SSE / df = 49.742/6 = 8.29**

**The same results can be obtained by using the following R coding and procedures. Results may differ slightly due to rounding differences.**

**anova(lm(y ~ x))**

**R output**

**Analysis of Variance Table**

**Response: y Df Sum Sq Mean Sq F value Pr(>F) x 1 369.64 369.64 44.484 0.0005498 \*\*\*Residuals 6 49.86 8.31**

**---**

**The Normal Linear Model**

Text

Description automatically generated