**Regression**

Regression analysis is used to estimate relationship between dependent variable (Y) and one or more independent variables (X).

Our theory states Y=f(X)

Regression is used to test theory

We estimate the magnitude of relationships and evaluate how confident we are the relationship exists within the population

**Summarizing a Body of Data**

Consider the variable, [total library expenditure](http://milesfinney.net/491/hand/lib_data.xlsx) in cities within Los Angeles County in 1999

The library expenditure data can be summarized by the [distribution](http://milesfinney.net/491/hand/lib_data.pdf) of the variable

A distribution assigns the chance a variable equals a value or range of values

It may illustrate patterns in the data

**Explaining the Variation in Data**

What accounts for the differences in expenditures across cities?

For example, what is causing library expenditure in Alhambra to be greater than in Arcadia?

Our theory attempts to answer question and regression attempts to verify theory.

What is meant by data being a **population**? A **sample**?

Define Population, Sample

**Population Characteristics**

Let’s assume the data represent a population.

The expected value of library expenditures, E(Y), would be the population mean expenditure, µ.

E(Y) = $1,571,126.093 – interpret.

If library expenditures are related to other variables, the conditional expected value of Y will differ from the unconditional.

The average value of Y will vary for different values of X

E(Y) is the unconditional expected value of Y.

E(Y|X) – expected value of Y conditional on the variable X.

E(Y|X) ≠ E(Y) indicates there is relationship between Y and X.

Suppose X indicates whether the library is run by the individual city or is part of the county library system:

X=1 if city run; = 0 if county run.

E(Y|X=1) – expected value of library expenditures conditional on the library being city-run.

E(Y|X=0) – expected value of library expenditures conditional on the library being county-run.

E(Y|X=1) = 2,450,547.42; E(Y|X=0) = 951,533.80.

1. Libraries run by individual cities have greater mean expenditures than the average library.
2. Libraries run by individual cities have greater mean expenditures than libraries in the county system.

Given that we have defined the data as the population we can definitely say the results indicate a relationship between Y and X within the population.

Our analysis however doesn’t necessarily mean the relationship is causal.

Causation is stated only by our theory.

If data represents a sample, we do not know if the relationship that exists in the sample also exists within the population.

**Regression Analysis**

Regression estimates the determinants of the variation in a dependent variable.

Y = F(X)

What or who determines which variable is the dependent variable? What determines which independent variables go into the model?

|  |  |
| --- | --- |
| **Population Regression Equation** | **Sample Regression Equation** |
| E(Y|X) = ß0+ ß1X1+ ß2X2+…. | ŷ = *b*0 + *b*1X1 + *b*2X2+…… |

What is the relationship between the population and sample regression equations?

What is the relationship between E(Y|X) and ŷ?

Explain why the values of the ß parameters are normally never observed.

What is the relationship between the *b* statistics in the sample equation and the ß parameters in the population equation?

In our class example, data were presented on [library expenditures by city](http://milesfinney.net/491/hand/lib_data.xlsx) in Los Angeles County.

Our first sample regression model:

ŷ = *b*0 + *b*1X1

where y is library expenditures by the sampled cities

X represents the number of residents in the sampled cities.

[Scatter Diagram](http://milesfinney.net/491/hand/scatter.pdf)

**ŷ = 49667 + 24.30X1**

(Y is in dollars and cents; X is the untransformed number of residents)

interpret b1

∆ŷ = b1 ∆X1

∆ŷ = 24.30∆X1

What is the change in predicted expenditure if city size increased by 10 residents? 100 residents?

We can use the equation to predict the level of library expenditure (as opposed to a change) for a city of a given size.

The number of residents in Agoura Hills in 1999 was 21,900. Use the sample regression equation to estimate library expenditure in Agoura Hills.

**ŷ = 49667 + 24.30X1**

The residual term, ê, is the difference between the actual and predicted value of the dependent variable:

yi = ŷi + êi

Actual value (y) = predicted value (ŷi) + residual (êi)

Rearrange terms: **êi= yi- ŷi**

Calculate the residual term for Agoura Hills

Did our model overpredict or underpredict Agoura Hills’ actual library expenditure?

Explain why our model did not perfectly predict Agoura Hills’ expenditures.

Interpret the residuals for the sampled cities from our regression:

|  |  |  |  |
| --- | --- | --- | --- |
| Sample of residual terms | | |  |
| City | Actual Library Expenditure | Predicted Y | Residual |
| AgouraHills | 943,436 | 581,889.45 | 361,546.55 |
| Alhambra | 1,703,193 | 2,275,765.62 | -572,572.62 |
| Arcadia | 1,535,420 | 1,342,554.07 | 192,865.93 |
| Artesia | 262,885 | 462,807.76 | -199,922.76 |
| Asuza | 835,707 | 1,160,286.19 | -324,579.19 |

Is it possible to come up with a model that would perfectly predict each of the sampled city’s expenditures?

**R**2

R2 measures the proportion of the variation in dependent variable that is explained by the model.

How much of the variation in library expenditures across cities is explained by differences in city size?

**R**2 for the model is .6555 Interpret.

Given the model, the calculated **R**2 seems high. What do I mean by this?

How much of the variation in library expenditures across the sampled cities remains to be explained?

If **R**2 equals 1 (100%) explain what that would imply about the relationship between Y and X.

Why should we never calculate an **R**2 equal to 1 (100%) for a regression?

For our model, the population regression equation E(Y|X) = ß0+ ß1X1

The sample equation: ŷ = *b*0 + *b*1X1 ŷ = 49667 + 24.30X1

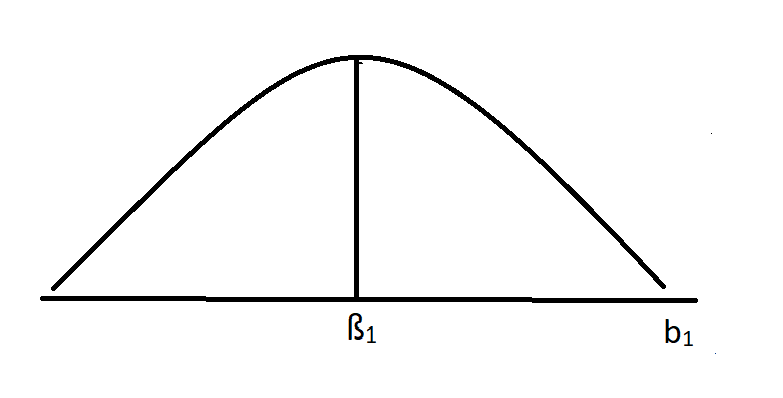
*b*1 is an estimate for ß1.

ß1 isthe “true” relationship between X and Y within the population.

We want to know the value of ß1 but we only have its estimate.

*b*1 is a variable. Why? Define variable.

Given assumptions, *b*1 follows a normal distribution:



For a specific population the parameter ß1is not a variable. Why is ß1 not considered a variable but *b1* is?

E(b1)=ß1 What does this mean?

This difference between a specific b1 and ß1 is called sampling error.

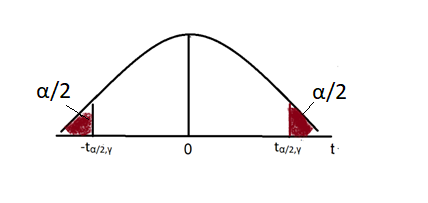
The slope estimate b1 follows a sampling distribution with a standard deviation equal to Sb1 (=2.062 in our regression output).

Population Equation: E(Y|X)=ß0+ ß1X1

Interpret hypotheses:

H0: ß1=0

H1: ß1≠0



Steps to perform hypothesis test.

1. State null and alternative hypotheses, H0 and H1.
2. Use t-distribution. ([t-table](http://milesfinney.net/491/hand/tdist.pdf))
3. Set level of significance, α. This gives the size of the rejection region.
4. Find the critical values. For a two tailed test, the critical values are ± tα/2,γ where γ is degrees of freedom n-k-1.
5. Calculate test statistic t=(b1-ß1)/Sb1.
6. Reject H0 if test statistic, t<- t α/2,,γ or t> t α/2,,γ

**Multiple Regression**

The “true” model would have all the X’s on the right hand side that theory suggests have a systematic relationship with Y.

Example of linear model

ŷ = b0+b1X1+b2X2+b3X3+b4X4

whereŷ is predicted library expenditure; X1 is number of residents in city; X2=1 if library run by city =0 if library run by county; X3 is percent of city residents who are school aged children; X4 is median household income by city.

* 1. Interpret each of the *b* coefficients (be careful in interpreting b2, the coefficient for the dummy variable X2)
  2. Interpret R2 (why is R2 higher in the multiple regression compared to the simple regressions?)
  3. Perform and interpret the hypothesis tests for ß

