Simple Linear Regression: An Introduction

BIOS 6611

CU Anschutz

Week 7

1. Correlation

2. Simple Linear Regression

3. The Regression Line

4. SLR Assumptions

Correlation

Pearson Correlation

When given two *continuous* variables, one way to measure their association is through correlation. This is especially useful if there is not an identified response variable.

Pearson's correlation coefficient is used to measure the strength and direction of the *linear* association between two variables. It can be used to estimate the population correlation, ρ, and is defined by:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

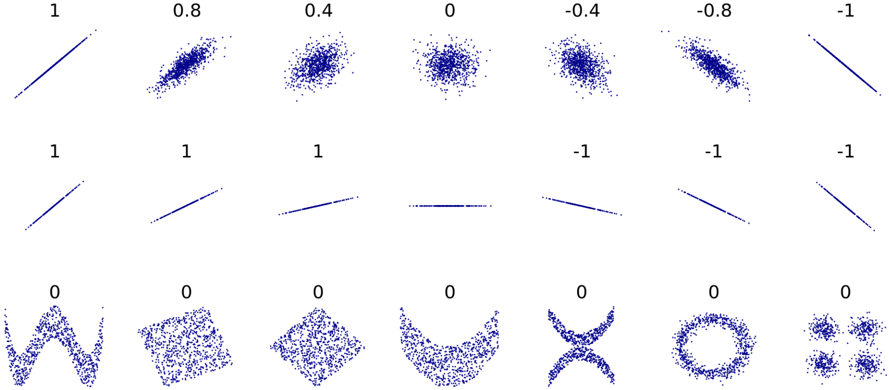
r is a dimensionless quantity; i.e., r is independent of the units of measurement of X and Y

Properties of Correlation

A correlation can range in value between -1 and 1:

- If $r > 0$, then as X increases Y increases and the two variables are said to be positively correlated. $r = 1$ is perfect positive correlation.
- If $r < 0$, then as X increases Y decreases and the two variables are said to be negatively correlated. $r = -1$ is perfect negative correlation.
- If $r = 0$, then there is no linear relationship between X and Y. The two variables are said to be *uncorrelated*. The correlation is 0 when the covariance of X and Y is 0 (i.e., Cov(X,Y)=0 $\implies$ Corr(X,Y)=0).
- The correlation coefficient is a measure of the strength of the linear trend relative to the variability of the data around that trend. Thus, it is dependent both on the magnitude of the trend and the magnitude of the variability in the data.

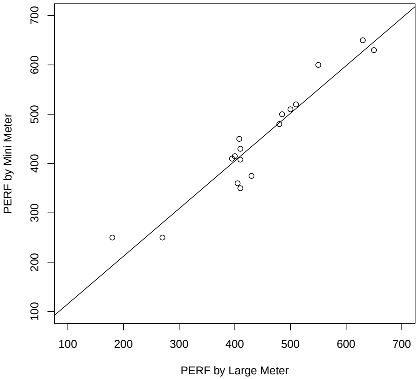
Correlation Examples



Source: Wikipedia

Correlation Example: Correlating Two Lung Function Meters

Peak expiratory flow rate (PERF) can be measured using either the Wright Peak Flow Meter or the Mini-Wright Peak Flow Meter. Do their measurements "correlate"?



- For this set of data, $r = 0.953$, a strong, positive linear correlation.
- *Note that this is a very different question than do their measurements "agree".*

Simple Linear Regression

Motivating Example: Lung Function in Children

Study Objective: To describe how lung function develops in children, and how smoking affects development.

Study Design: Cross-sectional survey. A random sample of children ages 3 to 19 from the East Boston area from which 654 had usable data.

Variables Measured: FEV (forced expiratory volume), age, sex, height, current smoking status. (FEV) measures how much air a person can exhale during a forced breath. Higher FEV indicates better lung function.

Outcome Variable (Y): FEV

Primary Explanatory Variable (X): age, sex, height, smoking status (depending on the question of interest)

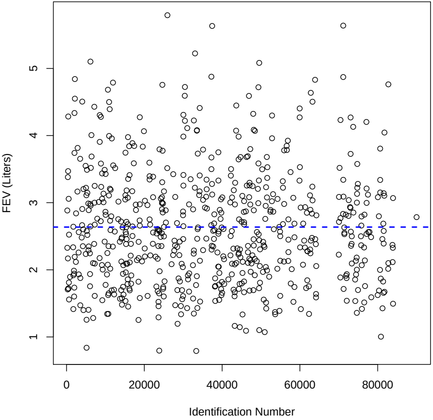
Covariates (C): age, sex, height, smoking status (depending on the question of interest)

Source: Lung function in children (FEV data) [Am J Epidemiology, 110(1): 15-26, 1980.]

Continuous Outcome and No Covariates

If the only information you have about a child is their study ID number and FEV measurement, what would be your "best guess" for the FEV of this individual ("best guess" for Y)?

$$
\begin{aligned}
E(Y) =& E(FEV) \\
=& \frac{\sum_{i=1}^{n} Y_i}{n} \\
=& \bar{Y} \\
=& 2.64 \text{ liters}
\end{aligned}
$$

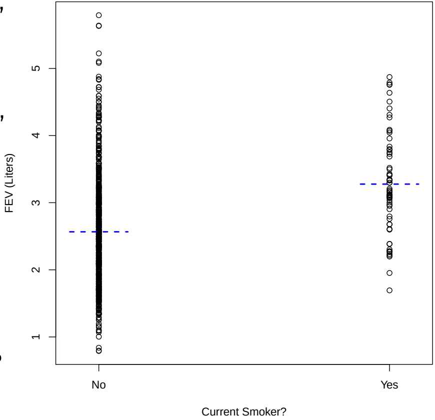


Continuous Outcome and Binary Covariate

If you knew the child is a non-smoker, what would be your "best guess" for the FEV of this individual? Let's call this the expected value of Y given X, or $E(Y|X)$:

$$E(Y|X=0) = \bar{Y}_{ns} = 2.57 \text{ liters}$$
$$E(Y|X=1) = \bar{Y}_{s} = 3.28 \text{ liters}$$

How could we test whether FEV differed for smokers and nonsmokers?

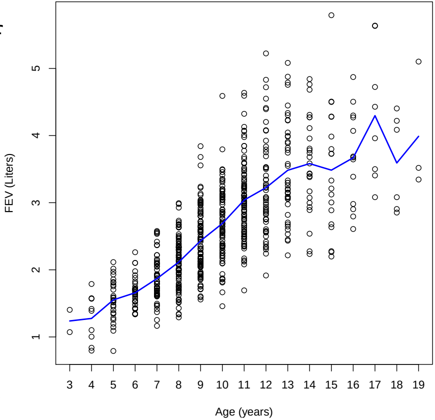


Continuous Outcome and Continuous Covariate

If we knew the child's age, what would our "best guess" for the FEV of this individual be?

If the child is 15, our "best guess" would be
$E(FEV|Age = 15) = 3.5$ liters

If the child is 3, our "best guess" would be
$E(FEV|Age = 3) = 1.2$ liters,
however we can note that this is only based on 2 observations!



Simple Linear Regression (SLR)

Goal: We wish to model the distribution of some continuous response variable (e.g., FEV) across groups defined by a single predictor (e.g., age).

The regression line (i.e. $FEV = \beta_0 + \beta_1 \times Age$) is a linear approximation to the *graph of averages*, which shows the average value of Y (FEV) for each X (age).

This model can be used to answer commonly encountered statistical questions:

- Prediction: Estimating a future observation of response Y.
- Quantifying distributions: Describing the distribution of response Y within groups defined by X.
- Comparing distributions across groups defined by X.

SLR Motivation

The regression model allows us to make inferences about groups that have few (if any) subjects by “borrowing” information from other groups.

Interpolation to unobserved groups is less risky than extrapolation outside the range of predictors included in the regression model.

Different *mathematical models* may be appropriate to model the distribution of Y across X: a straight line, a parabola, a log function, etc.

Ultimately, we are interested in finding the “best” model that describes our data while also being *parsimonious* (i.e., using simpler models and/or as few predictors as possible). Sometimes we choose a model suggested from experience or theory.

We will begin by assuming a straight line model with one predictor:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

The Regression Line

The Regression Line and Its Components

The line $Y = \beta_0 + \beta_1 X_1 + \epsilon$ is known as a **regression model**. The components include:

- β_0, the **intercept** of the line
- β_1, the **slope** of the line
- ϵ, the **error term**

We will walk through their properties on the following slides.

The predicted value of Y for a given value of X is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1,$$

where the hats represent the estimated values for our regression coefficients and the resulting FEV predicted value.

β_0 - The Intercept

The intercept is the expected value of Y when X is zero: $E(Y|X=0)$

Oftentimes we do not actually observe values of zero for X (e.g., age of 0). This may be problematic for a few reasons:

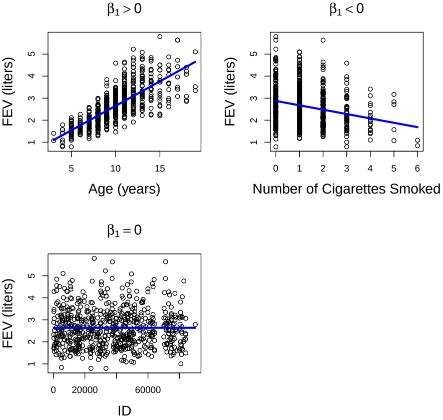
1. We are extrapolating to the estimated mean when $X = 0$, and we should generally avoid extrapolation.
2. It may make no biological/clinical/contextual sense.

β_1 - The Slope

β_1 is the expected change in Y associated with a *one unit* increase in X.

- If $\beta_1 > 0$, then as X increases, the expected value of Y increases.
- If $\beta_1 < 0$, then as X increases, the expected value of Y decreases.
- If $\beta_1 = 0$, then there is no *linear* relationship between X and Y.

These interpretations hold for the "continuous" predictors to the right, and also for "categorical" predictors (interpreted with respect to the reference category).



ϵ - The Error Term

We don't expect the linear relationship to hold exactly for all individuals, so the error term is included to represent this variability. It represents the variance of Y given a value of X.

The error is assumed to follow a normal distribution with mean 0 and variance σ_e^2: $\epsilon \sim N(0, \sigma_e^2)$.

If the model perfectly predicts all individuals, then $\sigma_e^2 = \sigma_{Y|X}^2$, else $\sigma_e^2 > \sigma_{Y|X}^2$.

In reality we do not know the error, but we can estimate this population parameter by calculating the **residual**. This is the difference between a fitted and observed value:

$$e_i = Y_i - \hat{Y}_i$$

SLR Assumptions

SLR Assumptions

Existence: For any *fixed* value of the variable X, Y is a random variable with a certain probability distribution having finite mean and variance.

Linearity: The mean value of Y (or a transformation of Y), $\mu_{Y|X} = E(Y)$, is a straight-line function of X (or a transformation of X).

Independence: The errors, ϵ_i, are independent (i.e., Y-values are statistically independent of one another).

Homoscedasticity: The errors, ϵ_i, at each value of the predictor, x_i, have equal variance (i.e., the variance of Y is the same for any X). That is,

$$\sigma^2_{Y|X} = \sigma^2_{Y|X=1} = \sigma^2_{Y|X=2} = ... = \sigma^2_{Y|X=x}$$

Normality: The errors, ϵ_i, at each value of the predictor, x_i, are normally distributed (i.e., for any fixed value of X, Y has a normal distribution). *(Note this assumption does not state that Y is normally distributed.)*

Illustration of the Linearity, Homoscedasticity, and Normality Assumptions

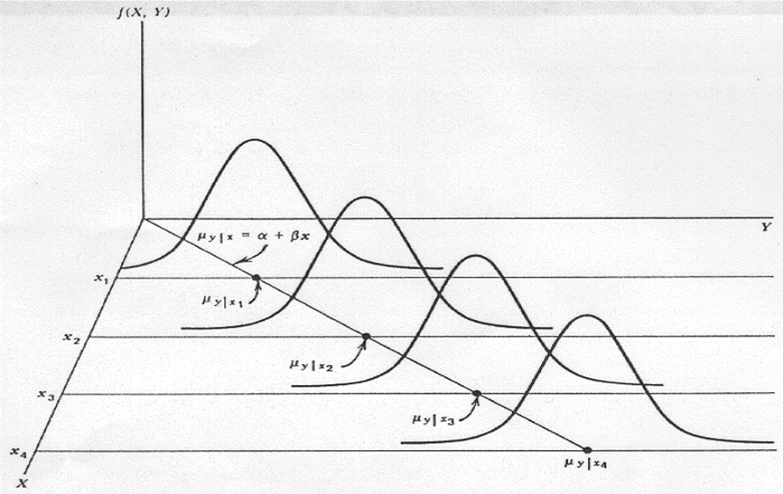


Figure 8.2.1. Representation of the simple linear regression model.