PSTAT 126: Regression Analysis

Department of Statistics and Applied Probability

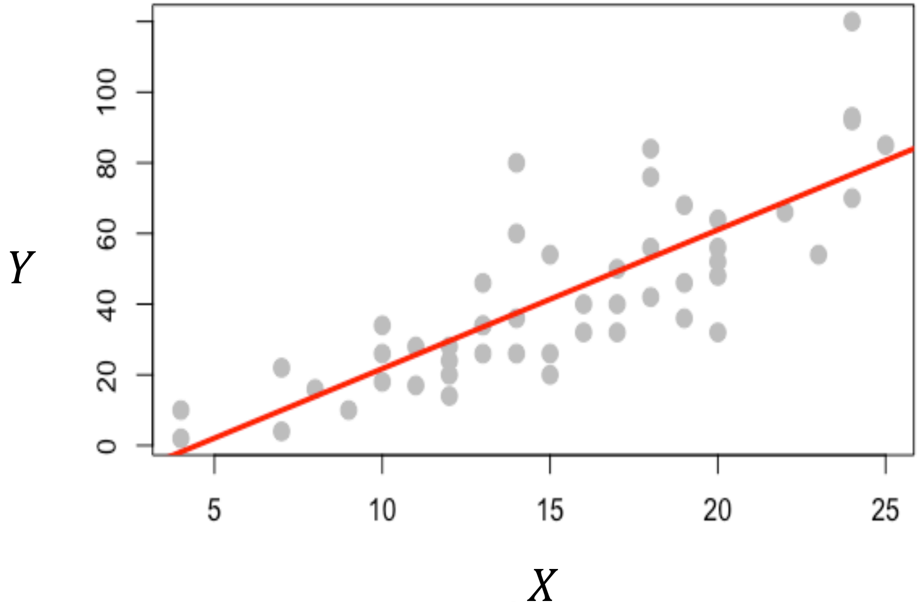
University of California, Santa Barbara

Regression and its Applications

There are many areas of human endeavor in which we would like to learn and model, from relevant but noisy data, an unknown functional relationship between a variable *X* (or variables) and a variable *Y*, the values of which we think of as dependent, in some sense, on those of *X*. The ability to do this has key applications in such areas as, among others:

- Science & Medicine
- Technology & Industry
- Economics & Finance
- Sociology & Behavioral Sciences
- Public Policy

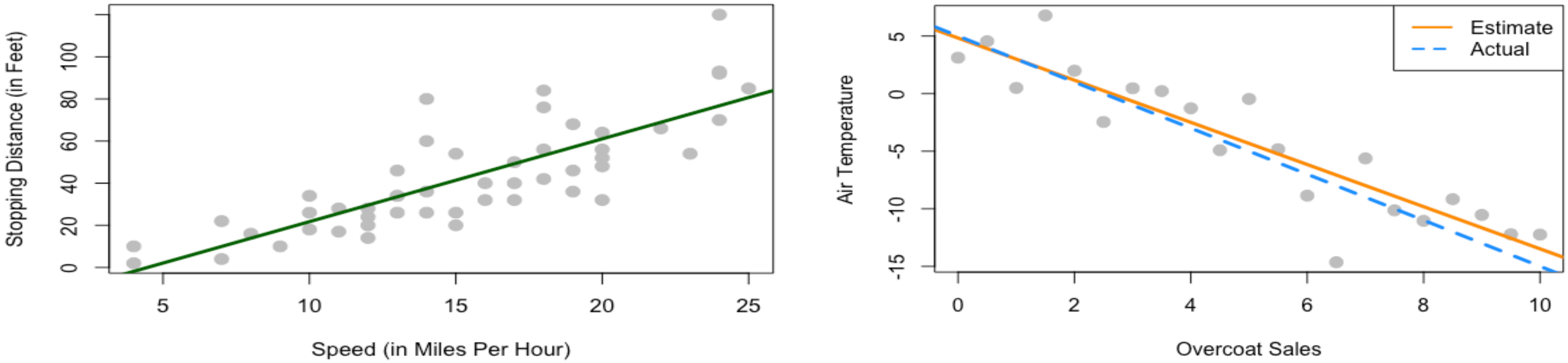
The study of how best to do this, including which mathematical and statistical methods and algorithms to use, is the subject of **Regression**.



Uses of Regression: Explanation and Insight

Explanation and insight:
Modeling the relationship between an input/inputs and an outcome, given observed, sampled data, in order to gain deeper understanding into that relationship.

What is the functional relationship between the stopping distance of a car (that is, the safe stopping distance, without the driver's loss of control) and the car's speed?



The graphic shows an example of linear regression – regression for which the functional relationship between X and Y is, or is presumed to be, linear in an appropriate sense.

Uses of Regression: Prediction

Prediction:
Given a new input value, not previously sampled, estimate the corresponding outcome/output value using the trained regression model.

Given one’s high school and/or college GPA, can SAT and/or GRE scores be predicted?

History of Regression

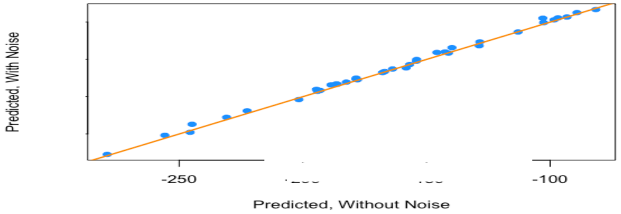
- The mathematicians Legendre (1805) and Gauss (1809) were the first known to have used the technique of statistical regression (that is, the method of least squares) as such, in order to find the best linear fit to a finite set of data points.
- They applied the method to analyze and predict planetary motion.
- Using the normal (or Gaussian) distribution to describe the behavior of errors, Gauss also developed a formula for this distribution, which plays such an important role in modeling errors in (linear) regression.
- Techniques for Linear Regression can rightly be viewed as Artificial Intelligence/Machine Learning methods and indeed as, historically speaking, perhaps the original versions of the types of Machine Learning algorithms so widely used today.

NASA / Johnson Space Center

Goals of Regression

Let P be any population. This population could be virtually any set of objects of interest, including people, cities, companies, biological cells, or stars in the night sky, for example. For any given population, we may be interested in the relationship between two variables of interest, a so-called predictor variable X -- also called the explanatory or independent variable -- and a response variable Y (also known as the dependent or target variable). For example, X and Y could be the respective

- Height and weight of people in P
- Distance from Earth of a set of stars and their corresponding brightness
- Education level and average income in the population of a given city.



In order to understand and explain the interaction between X and Y, which we think of as random variables, we would like to find an approximate functional relationship $f(X) \approx Y$ between them. Note that, for us, the function f we will attempt to learn will be assumed deterministic (non-random), and we will have

$$Y = f(X) + \epsilon,$$

with the noise term ϵ, also a random variable and such that the conditional expectation

$\boldsymbol{E}[\epsilon|X = x] = 0$, representing random error or variation in the model.

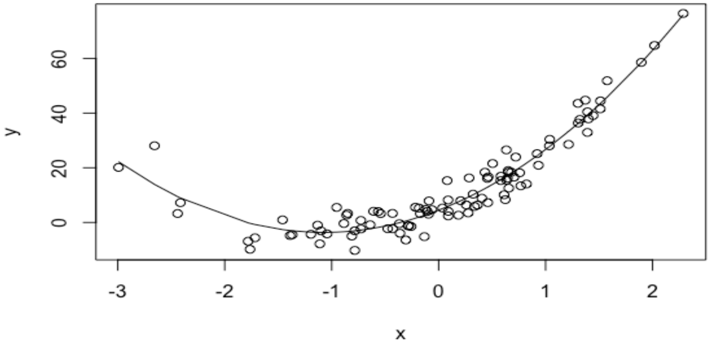
- We are essentially always interested in determining f in the context of regression models, but interest in determining more about the random error ϵ may depend on context.
- Indeed, for the purpose of explanation and insight concerning the relationship between $X$and Y, more information about the nature of ϵ, including its variance, may be of significant interest, whereas, when applying the model expressly for prediction, additional information about ϵ may be of less value.

Regression and the Mean Function

To determine a functional relationship between predictor X and response Y, our goal is to learn the conditional expectation function $\boldsymbol{E}[Y|X]$ – or, at least, a reasonably close approximation of it. We call $\boldsymbol{E}[Y|X]$ the **regression** or **mean function**.

Why is the mean function $\boldsymbol{E}[Y|X]$ so important here?

It clearly gives you the mean value of Y given $X=x$. But we can go further than this. We want to find a a function minimizing the difference between $f(X)$ and Y, on average. So, this would suggest looking at the absolute value of the difference $f(X) - Y$, i.e., $|f(X) - Y|$, and then considering the mean or expectation $\boldsymbol{E}[|f(X) - Y|]$. However, in part because the absolute value function is not smooth as it is not differentiable at 0 (spaces of functions defined by the square having other nice mathematical properties as well), it is more convenient to consider $\boldsymbol{E}[(f(X) - Y)^2]$.



$\boldsymbol{E}[Y|X]$ is the function that minimizes this squared error among all candidate functions f.
In fact it can be shown that

$$\boldsymbol{E}[(f(X) - Y)^2] = \boldsymbol{E}[(f(X) - \boldsymbol{E}[Y|X])^2] + \boldsymbol{E}[(Y - \boldsymbol{E}[Y|X])^2], \qquad (1)$$

for any candidate function f, where $\boldsymbol{E}[(Y - \boldsymbol{E}[Y|X])^2]$ depends on X and Y but not f. Equation (1) holds whether X is a scalar or vector-valued variable. Equation (1) says that that, for any function f, the expectation of the square of the difference between $f(X)$ and Y is equal to the expectation of the square of the difference between f and the mean function (plus a nonnegative constant, as shown in (1)).

- Since $(f(X) - Y)^2 \geq 0$ for any function f we can minimize the magnitude of the error of approximating f by Y on the left-hand side of (1) by in fact taking $f(X) = \boldsymbol{E}[Y|X]$.
- This means that the function of $X=x$ that approximates the behavior of the response Y with the smallest error on average is in fact the mean $\boldsymbol{E}[Y|X]$ function itself.
- So, it is the mean function which gives us the "best" representation of the functional relationship between X and Y in the sense described.

Hence, it is the mean function $\boldsymbol{E}[Y|X]$ that we would like to use regression methods and algorithms to determine or at least closely approximate in order to identify and understand any functional relationship between X and Y.

Linear Regression

Our goal in this course is to study specifically **Linear Regression**, which is regression for which $\boldsymbol{E}[Y|X]$ is or may be presumed to be closely approximated by a linear function (i.e., more technically, a function selected from a finite-dimensional, linear space of candidate functions).

The linear case is of great interest because

- from the point-of-view of mathematical structure, it is relatively simple (shades of Occam's razor)
- it robustly describes many situations arising in applications
- it is the model base case for investigations into nonlinear regression (indeed, somewhat paradoxically, the linear regression model itself encompasses many seemingly "nonlinear" cases as well, as we shall see).

So, for the first part of the course we will be considering models of the relatively simple form

$$\boldsymbol{E}[Y|X = x] = \beta_0 + \beta_1 x, \qquad (2)$$

where x is a fixed, scalar value (real number), and Y is a scalar-valued continuous random variable. The numbers β_0, β_1 are parameters which, as we shall see, it is the goal of canonical regression algorithms to compute. When the regression function can be represented as in (2) it is called **Simple Linear Regression** (see the next slide) because only one predictor variable is involved and the predictor appears within a linear term only. Later on, we will augment this framework by adding additional predictor variable terms on the right in (2). This is called **Multiple Linear Regression.** Note that any representation of the function $\boldsymbol{E}[Y|X]$ in the form as on the RHS of (2) will be unique for either simple -- or multiple – regression (at least for the kinds of typical continuous probability distributions we are interested in in this course).

Simple Linear Regression (SLR) Model

But what are the methods of regression that enable us to determine the parameters β_0 and β_1 (or close approximations of these parameters), given that in general we have no ready or direct access to the actual values of the function $\boldsymbol{E}[Y|X]$?

The answer of course involves sampling. For this, let $x_1, x_2,...,x_N$ be N given fixed, real numbers. We could think of these numbers as sampled from the predictor X, but, in keeping with what seems to be fairly standard expository practice in textbooks on basic regression, we usually downplay or suppress the explicit role of the underlying variable X. Now, given these N values $x_n, n = 1, \ldots, N$, write

$$Y_n = \beta_0 + \beta_1 x_n + \epsilon_n, n = 1, \ldots, N, \qquad (3)$$

$$\epsilon_n \sim N(0, \sigma^2), n = 1, \ldots, N. \qquad (4)$$

Here, the ϵ_n are N independent, real-valued, normally-distributed random variables (i.i.d.), with $N(0, \sigma^2)$ being the normal (Gaussian) distribution with mean 0 and variance σ^2 . The ϵ_n represent random variation or noise in the model, and we shall have more to say later about our assumptions concerning the ϵ_n. We call (3)-(4) are our **Simple Linear Regression (SLR) Model**. The goal of regression is it to identify the scalar parameters β_0 and β_1 and also, often, σ as well, or, more commonly, close approximations of these three parameters. The SLR model above in (3)-(4) is the formal model we will now generally work with until we get to Multiple Linear Regression.

In (3)-(4) we assume, as already noted, that each x_n is a known constant (say the outcome of an experiment after the nth trial). So, for each n, we actually can write

$$\boldsymbol{E}[Y_n] = \boldsymbol{E}[Y_n|X = x_n] = \beta_0 + \beta_1 x_n. \qquad (5)$$

Simple Linear Regression Model (cont'd)

Our SLR model: Given N values $x_n, n = 1, \ldots, N$, write

$$Y_n = \beta_0 + \beta_1 x_n + \epsilon_n, n = 1, \ldots, N, \qquad (6)$$

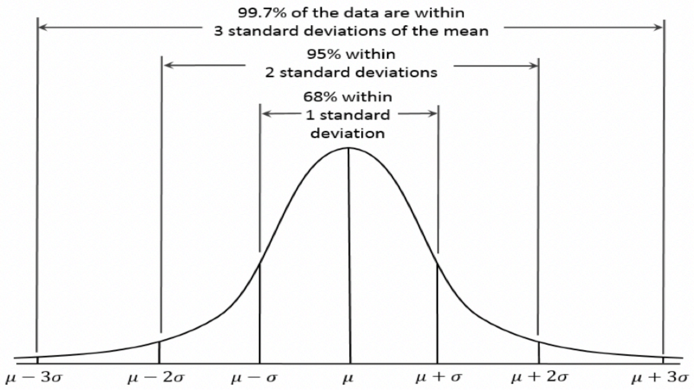
$$\epsilon_n \sim N(0, \sigma^2), n = 1, \ldots, N, \qquad (7)$$

the ϵ_n being N independent, normally-distributed random variables (i.i.d.), with $N(0, \sigma^2)$ being the normal distribution with mean 0 and variance σ^2 . So independence of the ϵ_n for us means *mutual independence* so that the corresponding joint and respective individual probability density functions satisfy

$$f_{\epsilon_1, \ldots, \epsilon_N}(z_1, \ldots, z_N) = f_{\epsilon_1}(z_1) \ldots f_{\epsilon_N}(z_N), \qquad (8)$$

where

$$f_{\epsilon_n}(z) = N(0, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{z^2}{2\sigma^2}), \text{ for each } n. \qquad (9)$$



Normal Distribution

The Y_n satisfy similar conditions but with different means. Note that the error ϵ_n is distributed symmetrically about $\boldsymbol{E}[Y_n | X = x_n] = \beta_0 + \beta_1 x_n$. We also note that the i.i.d. assumption is, while a common assumption, a strong assumption and its full strength is not always necessary in the context of regression analysis as we study in this course.

First Steps with R

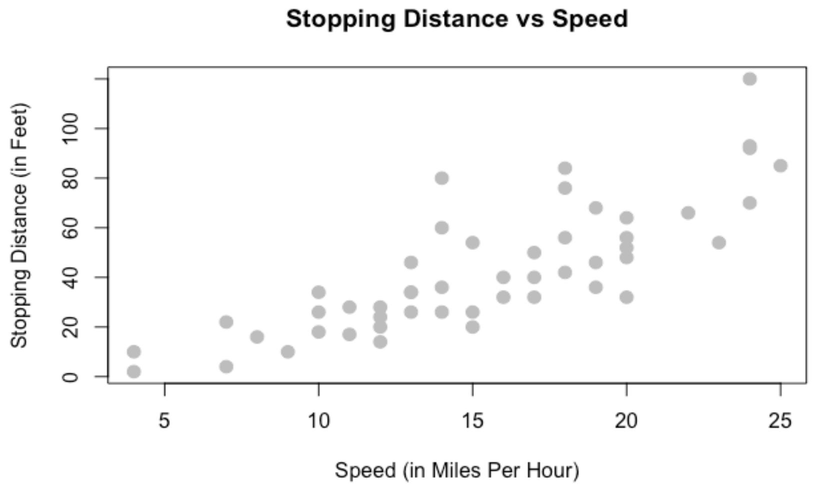
At this point, let’s see how the R language can be applied in the context of an actual data set to generate a simple linear regression model.

R is a language and environment for statistical computing and graphics, an integrated suite of software facilities for data manipulation. R is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

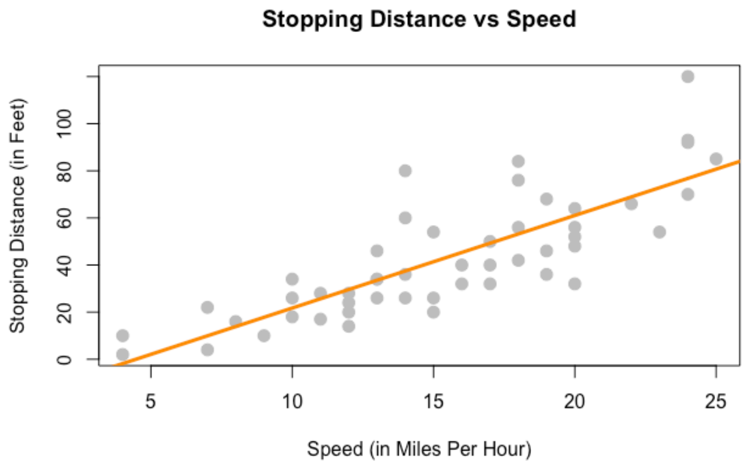
We use the “cars” data set, which is “built-in” to R. It contains data gathered during the 1920s about the speed of cars and the resulting distance it takes for the car to safely come to a stop, without loss of vehicle control.
Thinking of Speed as our predictor variable X and Stopping Distance as our Response Y, we can plot the stopping distance against the speed using the R code below.

```
plot(dist ~ speed, data = cars,
    xlab = "Speed (in Miles Per Hour)",
    ylab = "Stopping Distance (in Feet)",
    main = "Stopping Distance vs Speed",
    pch  = 20,
    cex  = 2,
    col  = "grey")
```



First Steps with R (cont’d)

```
stop_dist_model = lm(dist ~ speed, data = cars)
  stop_dist_model
## Call:
## lm(formula = dist ~ speed, data = cars)
## Coefficients:
## (Intercept)    speed
## -17.579        3.932
```



In order to compute the regression function (regression line) for the cars example we use the lm( ) function in R. The initials stand for ”linear model”, and it will be perhaps our most commonly used R function in this course. We will concern ourselves with how estimates of the model parameters are computed in forthcoming slides, but for now note that R gives

$$\beta_0 = Intercept \approx -17.579$$
$$\beta_1 = Slope \approx 3.932$$

```
plot(dist ~ speed, data = cars,
    xlab = "Speed (in Miles Per Hour)",
    ylab = "Stopping Distance (in Feet)",
    main = "Stopping Distance vs Speed",
    pch  = 20,
    cex  = 2,
    col  = "grey")
abline(stop_dist_model, lwd = 3, col = "darkorange")
```

Lecture 2 Overview

- Some computations with simulated data in R

  (it may be somewhat helpful for this to review the last part of the previous lecture video)
- Method of Least Squares for Simple Linear Regression (SLR)
- Gauss-Markov Theorem
- Behavior of the Mean Function Estimate as $N \to \infty$
- LINE Assumptions for SLR
- The residuals
- Sampling distributions for the SLR regression coefficients

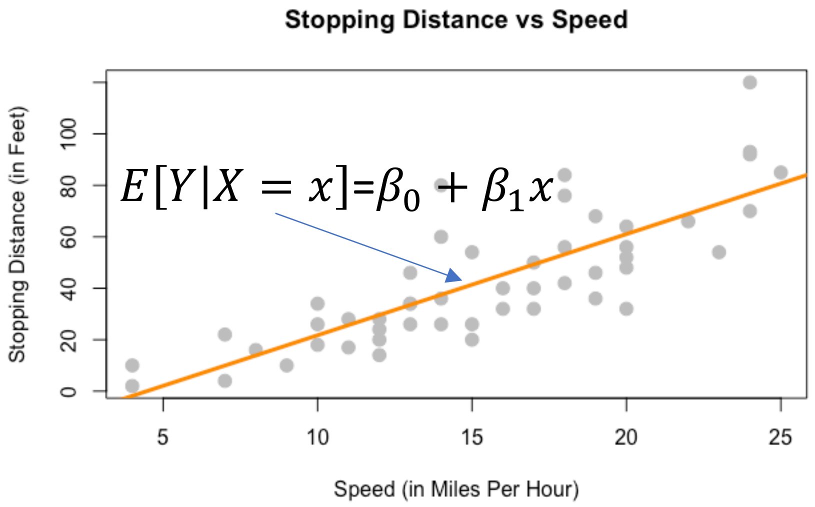
Method of Least-Squares for SLR

How do we approximate the parameters $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$?

Let $x_1, x_2,\ldots,x_N$ be *N* given fixed values as before. Now, for each *n=1,…,N*, we also sample a random value from the variable Y_n (in (3)-(4) on prior slide) corresponding to n. So denote by

$$(x_1, y_1),\ (x_2, y_3),\ldots,\ (x_N, y_N) \qquad (10)$$

the resulting *N* sample data points (*N* ordered pairs).



$E[Y|X = x]$=$\beta_0 + \beta_1 x$

To compute estimates for the true parameters β_0 and β_1 and solve for the model under the linearity assumption, we use the classic Method of Least Squares:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(\alpha_0,\alpha_1)\epsilon\mathbb{R}^2} \sum_{n=1}^{N} \big(y_n - (\alpha_0 + \alpha_1 x_n)\big)^2 \qquad (11)$$

Numbers $\hat{\beta}_0$ and $\hat{\beta}_1$ minimizing (11) will always exist. Our approximation for the mean function $\boldsymbol{E}[Y|X = x]$=$\beta_0 + \beta_1 x$ is then $\boldsymbol{E}[Y|X = x] \approx \widehat{\boldsymbol{E}}[Y|X = x] = \hat{\beta}_0 + \hat{\beta}_1 x$, assuming we can compute $\hat{\beta}_0$ and $\hat{\beta}_1$ (more on that below).

The minimizers $\hat{\beta}_0$ and $\hat{\beta}_1$ of the function $F(\alpha_0, \alpha_1) = \sum_{n=1}^{N} \big(y_n - (\alpha_0 + \alpha_1 x_n)\big)^2$ in (11) can be determined by computing the partial derivatives of F and setting them equal to 0. The resulting system of linear equations can then be solved for $\hat{\beta}_0$ and $\hat{\beta}_1$. In fact it follows that

$$\hat{\beta}_1 = \frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^{N}(x_n - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \frac{1}{N}\left(\sum_{n=1}^{N} y_n - \hat{\beta}_1 \sum_{n=1}^{N} x_n\right), \qquad (12)$$

where $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$ and similarly for the *y*-variable.

Gauss-Markov Theorem

Recall our Simple Linear Regression Model. Given *N* values x_1, x_2,…,x_N, we have

$$Y_n{=}\beta_0 + \beta_1 x_n + \epsilon_n, n = 1, \ldots, N, \qquad (13)$$

$$\epsilon_n \sim N(0, \sigma^2), n = 1, \ldots, N, \qquad (14)$$

where the ϵ_n are N independent, normally-distributed random variables, as well as our respective estimates $(\hat{\beta}_0, \hat{\beta}_1)$ for $(\beta_0, \beta_1\ )$. In the previous slide we defined these estimators in terms of the fixed, deterministic samples (x_1, y_1), (x_2, y_3),…, (x_N, y_N) in part in order to make concrete how they can be defined and calculated. However, it can also be useful, in order to assess their performance and behavior, to view the Y_n as random in this context as well (as if they had not yet already been computed). So, using upper case *Y*-values to denote their instantiation as random variables as in (13)-(14), we rewrite (12) in the form

$$\hat{\beta}_1{=}\frac{\sum_{n=1}^{N}(x_n - \bar{x})(Y_n - \bar{Y})}{\sum_{n=1}^{N}(x_n - \bar{x})^2}, \qquad \hat{\beta}_0{=}\frac{1}{N}\left(\sum_{n=1}^{N} Y_n - \hat{\beta}_1 \sum_{n=1}^{N} x_n\right). \qquad (15)$$

In the setting of our simple linear regression model (13)-(14) above, the **Gauss-Markov Theorem** then asserts that

(1) The respective estimators $\hat{\beta}_0$ and $\hat{\beta}_1$for the regression parameters β_0 and β_1 are unbiased, i.e. $\boldsymbol{E}[\hat{\beta}_0]$= β_0 and $\boldsymbol{E}[\hat{\beta}_1]$= β_1.

(2) $\hat{\beta}_0$ and $\hat{\beta}_1$are of minimum variance among all unbiased, linear estimators for β_0, β_1 , respectively. This implies that, among all unbiased, linear estimators α_0, α_1, the error $\boldsymbol{E}[(\alpha_i - \beta_i)^2]$ = $\boldsymbol{E}[(\alpha_i - \boldsymbol{E}[\alpha_i])^2]$,$i = 1{,}2$, is minimized when (α_1, α_2) =$(\hat{\beta}_0, \hat{\beta}_1)$.

This shows that the respective estimates $\hat{\beta}_0, \hat{\beta}_1$ are in an important sense the optimal ones for a fixed number *N* of samples.

Note that a linear estimator in this context means that both $\hat{\beta}_0$ and $\hat{\beta}_1$can be written as finite, linear combinations of the Y_n(that is, in this context, that we can write $\hat{\beta}_i{=}\sum_{n=1}^{N} k_{in} Y_n, i = 1{,}2$, for some constant coefficients k_{in} -- which follows from (15) since the x_n are assumed to be fixed, constant values).

Behavior of the Mean Function Estimate as $N \to \infty$

But why should the solution of the SLR least-squares minimization problem (Equ. (11) in a previous slide) – an optimization problem that after all only involves minimizing over a finite number of discrete points, however many, give a good estimate of the true mean function $\boldsymbol{E}[Y|X]$ over the entire underlying distribution, if we take $N \to \infty$? If we do know or can assume a priori that $\boldsymbol{E}[Y|X]$ really is linear (and furthermore in our simplified SLR setting right now has the very simple form $\boldsymbol{E}[Y|X = x] = \beta_0 + \beta_1 x$) and we think of the ordered pairs (X_1, Y_1), (X_2, Y_3),…, (X_N, Y_N) as i.i.d.-generated from some

random process, we can give some of the underlying intuition as to why right here, without formal statements or proofs.

Under suitable, quite general conditions, the answer has to do with the Law of Large Numbers (LLN) from Probability Theory and its extensions. From so-called “uniform” versions of the LLN, it follows that, for any small number $\varepsilon > 0$ and all *N* sufficiently large, we have, for all choices α_0, α_1 of the parameters,

$$\left|\boldsymbol{E}[(Y - (\alpha_0 + \alpha_1 X))^2] - \frac{1}{N}\sum_{i=1}^{N}\big(Y_i - (\alpha_0 + \alpha_1 X_i)\big)^2\right| \leq \varepsilon, \text{ with arbitrarily high probability.} \qquad (16)$$

This suggests that for a sufficiently large number *N* of random samples

$$(X_1, Y_1),\ (X_2, Y_3),\ldots,\ (X_N, Y_N)$$

the minimizing least-squares regression parameters $(\hat{\beta}_0, \hat{\beta}_1)$ in (7) also give rise to a function $f_{min}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ that, up to high probability, approximately minimizes

$$\boldsymbol{E}[(f(X) - Y)^2] = \boldsymbol{E}[(f(X) - \boldsymbol{E}[Y|X])^2] + \boldsymbol{E}[(Y - \boldsymbol{E}[Y|X])^2] \qquad (17)$$

among all functions of the form $f(x) = \alpha_0 + \alpha_1 x$ for some choice of α_0, α_1.

Since, as we have seen, the exact or true mean function $\boldsymbol{E}[Y|X]$, which we assume also has

the simple linear form $\boldsymbol{E}[Y|X = x] = \beta_0 + \beta_1 x$ with parameters β_0 and β_1, is a minimizer of (17) it follows that f_{min} is close to $\boldsymbol{E}[Y|X]$ in the sense that $\boldsymbol{E}[(f_{min}(X) - \boldsymbol{E}[Y|X])^2]$ must be small.

LINE Assumptions for Simple Linear Regression

Recall our Simple Linear Regression Model. Given *N* fixed values $x_1, x_2,…,x_N$, we have

$$Y_n = \beta_0 + \beta_1 x_n + \epsilon_n, n = 1, \dots, N, \qquad (18)$$

$$\epsilon_n \sim N(0, \sigma^2), n = 1, \dots, N, \qquad (19)$$

where the error terms ϵ_n -- representing noise or natural stochastic (statistical) variation -- are N independent, normally-distributed random variables.

The main assumptions of this model are frequently denoted by means of the mnemonic acronym **LINE**:

Linearity: The relationship between each Y_n and each x_n, respectively, is linear, and $\boldsymbol{E}[Y_n] = \boldsymbol{E}[Y_n | X_n = x_n] = \beta_0 + \beta_1 x_n$ for all $n = 1, \dots, N$.

Independence: The errors $\epsilon_n, n = 1, \dots, N$, are independent random variables.

Normality: The errors $\epsilon_n, n = 1, \dots, N$, follow a normal distribution. That is, the error across the regression line at any point x_n is described by a normal distribution.

Equal Variance: The normal distribution describing the behavior of the ϵ_n has the same variance, σ^2, for all n. This property is called *homoscedasticity.*

Note that the first or “L” assumption implies that $\boldsymbol{E}[\epsilon_n] = \boldsymbol{E}[Y_n - (\beta_0 + \beta_1 x_n)] = 0$.

Some Comments on the LINE Assumptions

Some observations/comments on the **LINE** assumptions:

- How valid is it to specify that the errors $\epsilon_n, n = 1, \ldots, N$, should be normally distributed? It is known that this frequently tends to be the case for random noise as well as random natural variation. One reason could have to with the Central Limit Theorem, which says that, roughly speaking, a large sum of i.i.d. random variables, whatever distribution these individual random variables may follow, will be approximately normally distributed. This suggests that superpositions of large amounts of random noise will tend to be approximately normally-distributed.
- Gauss-Markov Theorem: Assuming the **LINE** hypotheses enables us to know that the Gauss-Markov Theorem holds, which means that we obtain unbiased, minimal variance estimators for the coefficients of the regression function.

We will see in the rest of the course that we will actually be using various methods – including formal statistical tests as well as graphical ones -- to verify or provide evidence for the **LINE** assumptions – or more precisely the latter three "I-N-E" assumptions -- on the random error terms. Successfully verifying those in a specific situation can provide strong evidence that the linearity assumption on the model itself holds as well, in particular in cases in which any knowledge one may have about the particular application domain involved does not give sufficient insight into the nature of the relationship between *X* and *Y*.

The Residuals and Residual Standard Error

Recall once again our SLR model

$$Y_n = \beta_0 + \beta_1 x_n + \epsilon_n, n = 1, \dots, N, \qquad (20)$$

$$\epsilon_n \sim N(0, \sigma^2), n = 1, \dots, N. \qquad (21)$$

We are not able to sample the errors ϵ_n, $n = 1, \dots, N$, in any direct way, only the Y_n. However, we would want to use the error values to support the validity of our model, as pointed out in the previous slide.

So, consider the so-called **residuals** instead:

$$e_n := y_n - \hat{y}_n, \text{ where } \hat{y}_n := \hat{\beta}_0 + \hat{\beta}_1 x_n, n = 1, \dots, N, \qquad (22)$$

We will in essence use the residuals in key ways in place of the errors ϵ_n, in essence as proxies for the errors ϵ_n whose values we do not have access to, to help justify the validity of our linear regression models, as we will see.

First, we use them to define an estimator for σ^2 in the form

$$\hat{\sigma}^2 = s_e^2 = \frac{1}{N-2} \sum_{n=1}^{N} e_n^2, \qquad (23)$$

where $\hat{\sigma} = s_e$, the square root of the value in (23), is known as the **Residual Standard Error (RSE).** Note the factor $\frac{1}{N-2}$ appearing in (23).

It can be shown that this is actually the right factor to make $\hat{\sigma}^2$ an unbiased estimator for σ^2, so that $E[\hat{\sigma}^2] = \sigma^2$.

In R, we can find the value of the RSE using the following:

```
car_model=lm(dist ~ speed, data = cars)
summary(car_model)$sigma
```

The following further command outputs the residuals for this model:

```
residuals(car_model)
```

Normality of the Residuals

Note that, assuming as we wish to, that the errors ϵ_n, $n = 1, \ldots$,*N*, are normally-distributed according to some distribution $N(0, \sigma^2)$, the Y_n must be normally-distributed as well (with a different mean but the same variance). But, more interestingly, it can be shown (see Sec. 3.2.5 in Sheather (2009) reference) that, for each $n = 1, \ldots$,*N,*

$$e_n = \epsilon_n - \sum_{i=1}^{N} h_{ni}\epsilon_i = (1\text{-}h_{nn})\epsilon_n - \sum_{i=1}^{n-1} h_{ni}\epsilon_i - \sum_{i=n+1}^{N} h_{ni}\epsilon_i, \qquad (24)$$

where $h_{ni} = \frac{1}{N} + \frac{(x_n - \bar{x})(x_i - \bar{x})}{\sum_{j=1}^{N}(x_j - \bar{x})^2}$. Since a (finite) linear combination of independent normally-distributed random variables is also normally distributed, this means that the residuals e_n are themselves also normally distributed if the original noise terms ϵ_n are. Note that, by a linear combination of random variables $Z_1, \ldots, Z_J$, we mean any random variable of the form $\sum_{j=1}^{J} c_j Z_j$, where the c_j are any fixed constants.

But interestingly we can go further than this using (24). Indeed it is argued in Sheather (2009) (again see Sec. 3.2.5 Sheather) that sums of random variables as in (24) can behave approximately like normally-distributed variables even when the ϵ_i,*i=1,…,N*, are not each assumed normally-distributed. Indeed there are extensions of the classical Central Limit Theorem that assert that large, weighted sums of i.i.d. random variables (similar to the sum in (24) above) ,for which the random variables in the sum need not necessarily be normally-distributed themselves, exhibit behavior that approximately follows a normal distribution for very large values of *N.*

Note that our SLR model, as we have defined it, presupposes of course the LINE assumptions, including that of normality of the errors and/or the residuals. However, a key aim of Linear Regression is still try to check that these assumptions are indeed valid for each specific regression model we consider. We will consider methods for this.

Sampling Distributions for $\hat{\beta}_0$ and $\hat{\beta}_1$

Explicitly thinking once again of the regression parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ as random variables, we can discuss their **sampling distributions**, the sampling distribution being the probability distribution that results when a statistic is considered as a random variable. Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are both finite, linear combinations of the Y_n (which are independent) and each Y_n is normally distributed, both $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed as well. In fact, we have

$\hat{\beta}_0 = \sum_{n=1}^{N} c_n y_n$, where $c_n = \frac{x_n - \bar{x}}{S_{xx}}$ and $S_{xx} = \sum_{n=1}^{N}(x_n - \bar{x})^2$, and $\hat{\beta}_1 = \sum_{n=1}^{N} d_n y_n$, where $d_n = \frac{1}{N} - c_n x_n$.

It can be shown (see Appendix A.4 in Weisberg (2014)) that

$$\hat{\beta}_0 \sim N\left(\boldsymbol{E}[\hat{\beta}_0], \sigma_{\hat{\beta}_0}^2\right) = N\left(\beta_0, \sigma^2\left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right)\right), \quad (25)$$

$$\hat{\beta}_1 \sim N\left(\boldsymbol{E}[\hat{\beta}_1], \sigma_{\hat{\beta}_1}^2\right) = N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right), \quad (26)$$

where $S_{xx} = \sum_{n=1}^{N}(x_n - \bar{x})^2$, and $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$. So, Var($\hat{\beta}_0$)= $\sigma^2\left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right)$ and Var($\hat{\beta}_1$)= $\frac{\sigma^2}{S_{xx}}$, where σ is as in the definition of a SLR model in (3)-(4) in a prior slide. Of course as we have seen we must estimate the variance σ^2, so as already observed we can estimate it using the RSE s_e=$\hat{\sigma}$: $\hat{\sigma}^2 = \frac{1}{N-2}\sum_{n=1}^{N} e_n^2$. So we can in turn obtain estimates for the respective variances $\sigma_{\hat{\beta}_0}^2$ = Var($\hat{\beta}_0$) and

$\sigma_{\hat{\beta}_1}^2$ =Var($\hat{\beta}_1$) (that is, for the respective standard deviations, taking square roots) via:

$$\sigma_{\hat{\beta}_0} \approx \mathrm{SE}[\hat{\beta}_0] := \hat{\sigma}\left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right)^{\frac{1}{2}}, \; \sigma_{\hat{\beta}_1} \approx \mathrm{SE}[\hat{\beta}_1] := \frac{\hat{\sigma}}{(S_{xx})^{\frac{1}{2}}},$$

where “SE” refers to “Standard Error” and “:=“ denotes for us “is defined as” and “ ≈” denotes “is approximately equal to”.