PSTAT 126: Regression Analysis

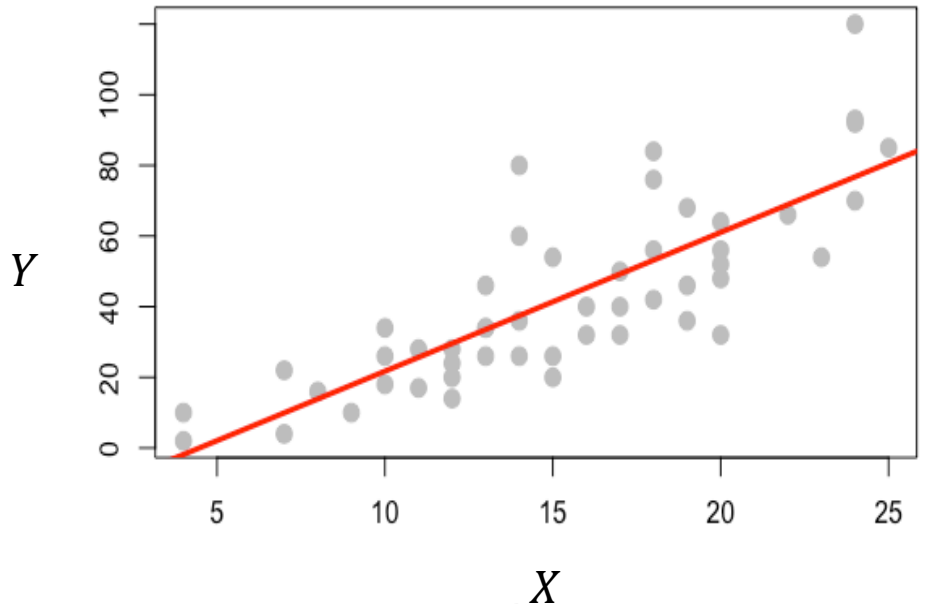
Department of Statistics and Applied Probability
University of California, Santa Barbara

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

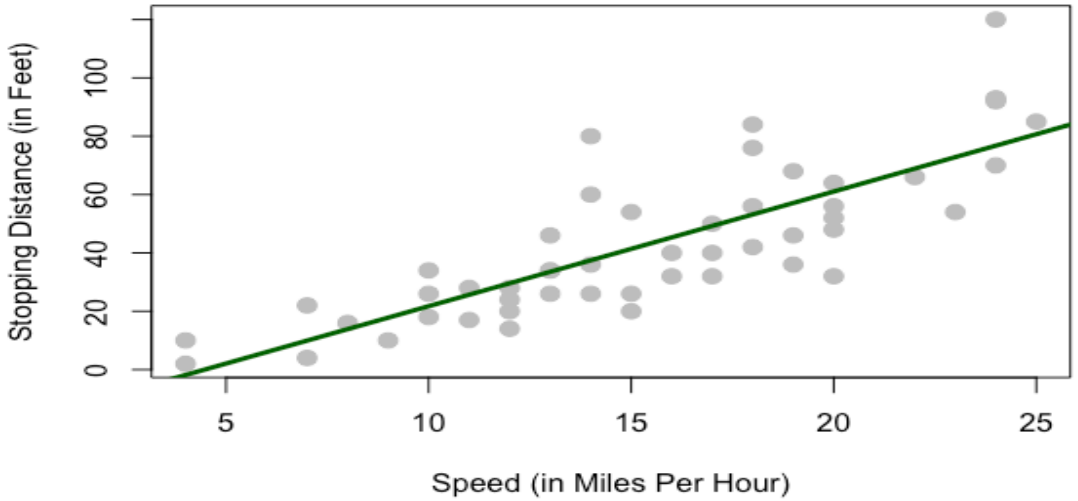
The study of how best to do this, including which mathematical and statistical methods and algorithms to use, is the subject of **Regression**.

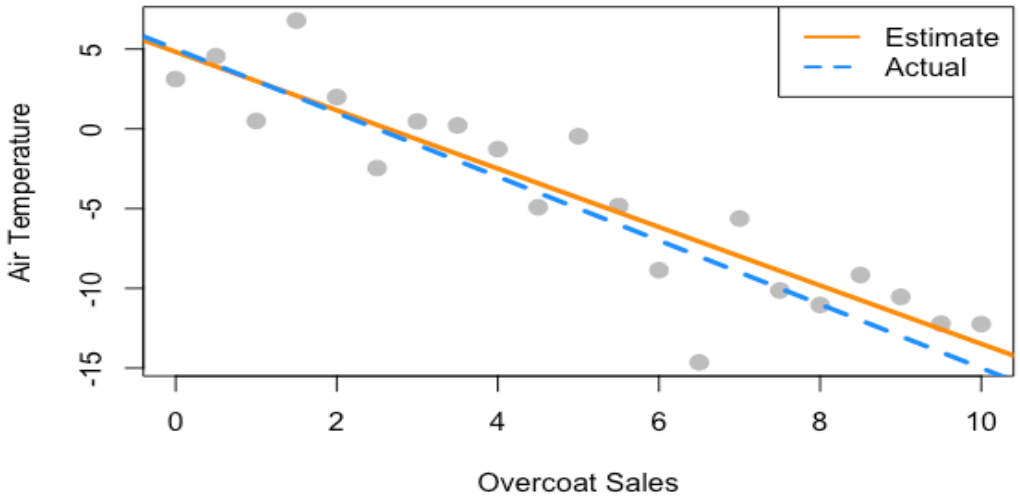


Uses of Regression: Explanation and Insight

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

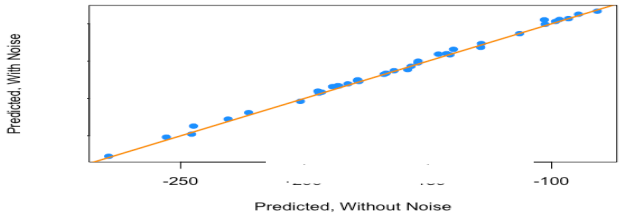
- The mathematicians Legendre (1805) and Gauss (1809) were the first known to have used the technique of statistical regression (that is, the method of least squares) as such, in order to find the best linear fit to a finite set of data points.
- They applied the method to analyze and predict planetary 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.

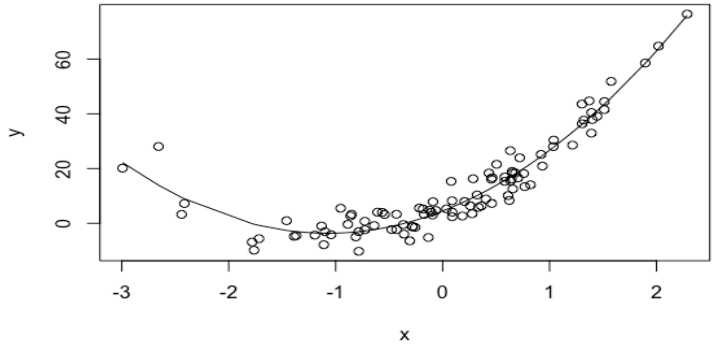
- We are essentially always interested in determining f in the context of regression models, but interest in determining more about the random error ϵ may depend on context.
- Indeed, for the purpose of explanation and insight concerning the relationship between X and Y, more information about the nature 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 $x_1, x_2, \ldots, x_N$ 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 $x_n, n = 1, \dots, N$, write

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

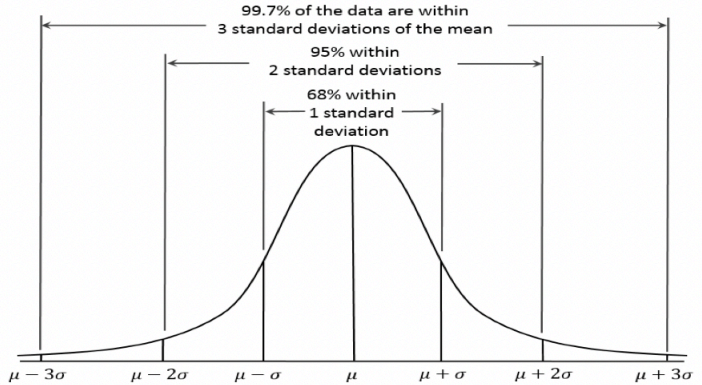
$$\epsilon_n \sim N(0, \sigma^2), n = 1, \dots, N, \quad (7)$$

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

$$f_{\epsilon_1,\dots,\epsilon_N}(z_1, \dots, z_N) = f_{\epsilon_1}(z_1) \dots f_{\epsilon_N}(z_N), \quad (8)$$

where

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



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

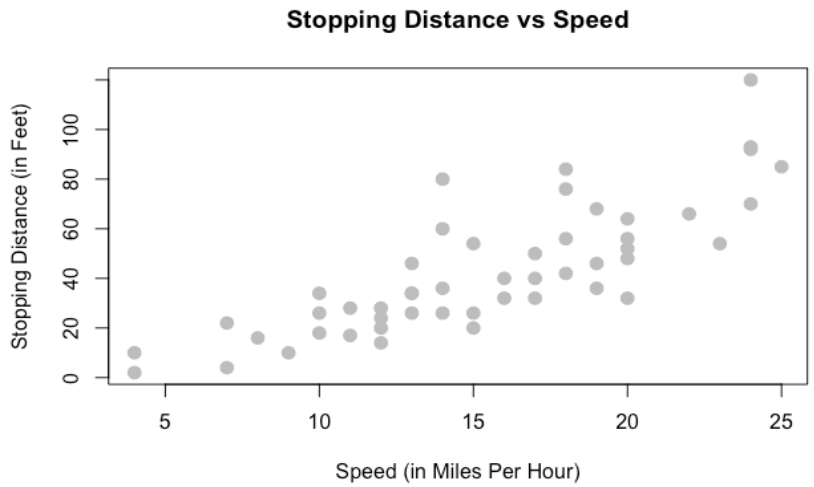
At this point, let's see how the R language can be applied in the context of an actual data set to generate a simple linear regression model.

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.

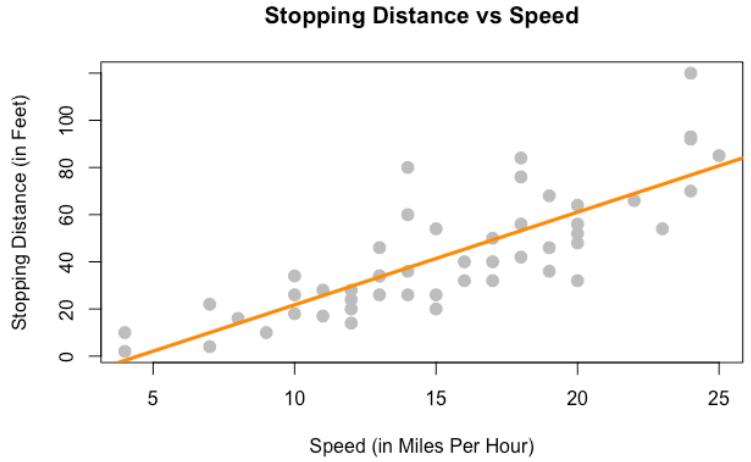
We use the "cars" data set, which is "built-in" to R. It contains data gathered during the 1920s about the speed of cars and the resulting distance it takes for the car to safely come to a stop, without loss of vehicle control.
Thinking of Speed as our predictor variable X and Stopping Distance as our 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

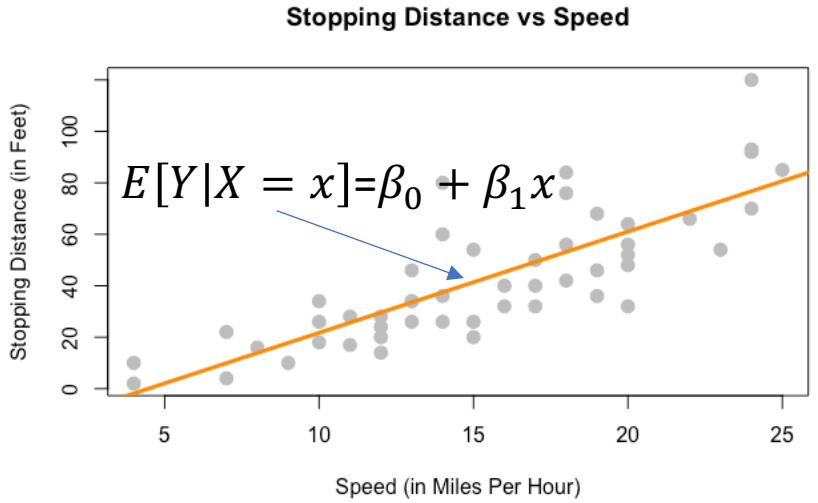
Method of Least-Squares for SLR

How do we approximate the parameters β_0 and β_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).



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:

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

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 $F(\alpha_0, \alpha_1) = \sum_{n=1}^{N} \left(y_n - (\alpha_0 + \alpha_1 x_n)\right)^2$ 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 $x_1, x_2, \ldots, x_N$, 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 (β_0, β_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. $\boldsymbol{E}[\hat{\beta}_0] = \beta_0$ and $\boldsymbol{E}[\hat{\beta}_1] = \beta_1$.

(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 $(\alpha_1, \alpha_2) = (\hat{\beta}_0, \hat{\beta}_1)$.

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,

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

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 $x_1, x_2, \ldots, x_N$, we have

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

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

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 $n = 1, \ldots, N$.

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

Normality**:** The errors $\epsilon_n, n = 1, \ldots, N$, 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 $\epsilon_n, n=1,\dots,N$, 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 $n = 1, \ldots, N$,

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

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.

Lecture 3 Overview

- Sampling distributions for the SLR regression coefficient estimates (review)
- Confidence intervals for intercept and slope (review)
- Distribution of new observations
- Confidence interval for new observations
- Some general background on hypothesis tests
- Interpretation of the SLR model summary output in R
- Some illustrative examples/computations with R

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

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

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

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 $\sigma^2_{\hat{\beta}_0}$ = Var($\hat{\beta}_0$) and

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

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

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

Confidence Intervals for Intercept and Slope

We can obtain confidence intervals for the true values of the intercept and slope β_0 and β_1 as well:

$$\hat{\beta}_0 - t\left(\frac{\alpha}{2}, N-2\right)\mathrm{SE}\left[\hat{\beta}_0\right] \le \beta_0 \le \hat{\beta}_0 + t\left(\frac{\alpha}{2}, N-2\right)\mathrm{SE}\left[\hat{\beta}_0\right], \qquad (27)$$

$$\hat{\beta}_1 - t\left(\frac{\alpha}{2}, N-2\right)\mathrm{SE}\left[\hat{\beta}_1\right] \le \beta_1 \le \hat{\beta}_0 + t\left(\frac{\alpha}{2}, N-2\right)\mathrm{SE}\left[\hat{\beta}_1\right], \qquad (28)$$

with $(1-\alpha)$x100% confidence, where $\mathrm{SE}\left[\hat{\beta}_0\right] = \hat{\sigma}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)^{\frac{1}{2}}$, $\mathrm{SE}\left[\hat{\beta}_1\right] = \frac{\hat{\sigma}}{(S_{xx})^{\frac{1}{2}}}$, $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$.

Here, $t(\alpha/2, N\text{-}2)$ is the value that cuts off $\alpha/2 \times 100\%$ in the *upper tail* of the *t*-distribution for *N-2* degrees of freedom (*N* sample data points along with the 2 parameters, intercept and slope, being estimated). The t-distribution (also called "Student's" t-distribution) is invoked because each $\hat{\beta}_i, i = 0,1$, follows a normal distribution as we saw in the previous slide, except for the fact that the corresponding variance σ^2 is unknown and an estimate $\hat{\sigma}^2$ for it must therefore be used instead. When this is done, the resulting variable follows a t-distribution instead. Note that it also goes under the name "student's" t-distribution because a statistician, William Gosset, who played a key role in developing and promoting it used the *nom de plume* "Student" (and not because it is only good for training purposes or something like that).

```
> stop_distance = lm(dist ~ speed, data = cars)
> confint(stop_distance, level = 0.95)

                2.5 %     97.5 %
(Intercept) -31.167850 -3.990340
speed         3.096964  4.767853
```

R code for generating confidence intervals for β_0, β_1 and example output for 95% confidence.

Distribution of New Observations

We would like to give probabilistic confidence intervals for model predictions for new observations. To do this, let's first describe the corresponding probability distribution. Given *X*=*x* for some new observation *x* (so not one of the *x* values we used for training, i.e., that we used to perform the least-squares regression minimization step in a previous slide in which $\hat{\beta}_0$ and $\hat{\beta}_1$ were defined), we would like to determine confidence intervals for the response variable $Y = \beta_0 + \beta_1 x + \epsilon$ at this new value *x*. For this we need the variance of the variable $\hat{y} + \epsilon = \hat{y}(x) + \epsilon = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon$, where once again we are now viewing $\hat{\beta}_0$ and $\hat{\beta}_1$ here as random variables, that is, as estimators.

Note that $\mathrm{Var}(\hat{y} + \epsilon) = \mathrm{Var}(\hat{y}) + \mathrm{Var}(\epsilon)$, since $\hat{y}$ and ϵ may be presumed independent. Hence, having calculated $\mathrm{Var}(\hat{\beta}_0)$ and $\mathrm{Var}(\hat{\beta}_1)$ in previous slides, it follows that

$$\mathrm{Var}(\hat{y} + \epsilon) = \sigma^2 \left(\frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}} \right) + \sigma^2, \qquad (29)$$

by using the general identity

$$\mathrm{Var}(aZ + bZ') = a^2 \mathrm{Var}(Z) + b^2 \mathrm{Var}(Z') + 2ab\,\mathrm{Cov}(Z, Z'),$$

for any random variables Z, Z' and where $\mathrm{Cov}(\cdot,\cdot)$ denotes the covariance $\mathrm{Cov}(Z, Z') = \boldsymbol{E}[(Z - \boldsymbol{E}[Z])(Z' - \boldsymbol{E}[Z'])]$. We also use the fact that, in this case, $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{S_{xx}}$ (see Equ. (2.12) in Weisberg(2014)). Hence,

$$\hat{y} + \epsilon \sim N\left(\beta_0 + \beta_1 x, \sigma^2 \left(\frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}} + 1 \right) \right), \qquad (30)$$

*S*ince we also need an estimate for σ^2 as we have previously seen, we insert in (30) the RSE and can use, in place of

$$\sigma \left(\frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}} + 1 \right)^{1/2},$$

$$\mathrm{SE}[\hat{y} + \epsilon] = \hat{\sigma} \left(\frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}} + 1 \right)^{1/2}. \qquad (31)$$

Confidence Interval for New Observations

We want to know a confidence interval for the response $Y = \beta_0 + \beta_1 x + \epsilon$ at a given new value *x*. We can use the probability distribution we have for $\hat{y} + \epsilon = \hat{y}(x) + \epsilon = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon$ from the previous slide for this purpose. Our corresponding confidence interval is then

$$\hat{y} + \epsilon - t\left(\frac{\alpha}{2}, N-2\right) \text{SE}[\hat{y} + \epsilon] \leq Y \leq \hat{y} + \epsilon + t\left(\frac{\alpha}{2}, N-2\right) \text{SE}[\hat{y} + \epsilon], \quad (32)$$

with $(1-\alpha)$x100% probabilistic confidence, $0 \leq \alpha \leq 1$ and chosen as desired (e.g., 0.05), and where $t(\cdot,\cdot)$ once again corresponds to a t-distribution (and with *N*– 2 degrees of freedom) as we saw for confidence intervals for β_0, β_1 as well.

R code and output for the 95% confidence interval for a new observation, in this case 16 mph.

```
stop_distance = lm(dist ~ speed, data = cars)
speedsabc = data.frame(speed = c(16))
predict(stop_distance, newdata = speedsabc,
    interval = c("prediction"), level = 0.95)
```

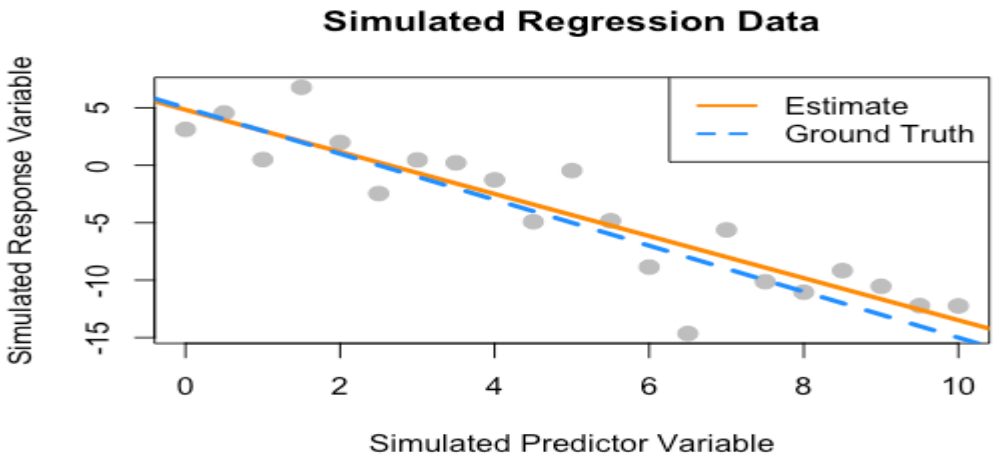
fit	lwr	upr
45.339445	14.10499	76.57390

Simple Linear Regression Simulation Study

We simulate "artificial" data samples

$$Y_n = 5 - 2x_n + \epsilon_n, n = 1, \ldots, 21,$$
$$\epsilon_n \sim N(0,9), n = 1, \ldots, 21,$$

and then can see how closely our computed least-squares regression model can approximate the true, artificially-generated model. Corresponding R code is to the right with some of the output below (in the form of a graphic).




```
num_obs = 21
beta_0  = 5
beta_1  = -2
sigma   = 3
set.seed(1)
epsilon = rnorm(n = num_obs, mean = 0, sd = sigma)
x_vals = seq(from = 0, to = 10, length.out = num_obs)
y_vals = beta_0 + beta_1 * x_vals + epsilon
sim_fit = lm(y_vals ~ x_vals)
coef(sim_fit)
```

```
sim_slr = function(x, beta_0 = 5, beta_1 = -2, sigma = 3) {
  n = length(x)
  epsilon = rnorm(n, mean = 0, sd = sigma)
  y = beta_0 + beta_1 * x + epsilon
  data.frame(predictor = x, response = y)
}
set.seed(1)
sim_data = sim_slr(x = x_vals, beta_0 = 5, beta_1 = -2, sigma = 3)
sim_fit = lm(response ~ predictor, data = sim_data)
coef(sim_fit)
plot(response ~ predictor, data = sim_data,
     xlab = "Simulated Predictor Variable",
     ylab = "Simulated Response Variable",
     main = "Simulated Regression Data",
     pch  = 20,
     cex  = 2,
     col  = "grey")
abline(sim_fit, lwd = 3, lty = 1, col = "darkorange")
abline(beta_0, beta_1, lwd = 3, lty = 2, col = "dodgerblue")
legend("topright", c("Estimate", "Ground Truth"), lty = c(1, 2), lwd = 2,
       col = c("darkorange", "dodgerblue"))
```

Hypothesis Tests

In Hypothesis Tests, we translate a scientific question into a statement concerning hypotheses regarding parameters in a statistical model.

Examples:

1. Are the chances of having a heart attack the same for males and females?

$$H_0 : p_m = p_f \text{ and } H_1 : p_m \neq p_f$$

where p_m is the probability to have a heart attack for males and p_f is the probability to have a heart attack for females.

2. In clinical trials, is the effect of a treatment no different from that of a placebo? Or is the effect different?

Given a possible least squares SLR regression model

$$Y_n = \beta_0 + \beta_1 x_n + \epsilon_n, n = 1, \ldots, N,$$

where x_n is the treatment level, e.g., 0, 1, 2, 3, 4, where 0 means the placebo

level, and Y_n is the observed response. The hypothesis test for significance of the regression relative to an "intercept-only" model is

$$H_0 : \beta_1 = 0 \text{ and } H_1 : \beta_1 \neq 0$$

Steps to Conduct a Hypothesis Test

1) Set up two competing hypotheses: Null hypothesis (H_0) and the alternative hypothesis (H_1 or H_a).

2) Set some significance level called α: the most common α is 0.05 or 5%.

3) Calculate a test statistic ($t\star$): a function of the data whose distribution depends only on the parameter(s) being tested.

4) Calculate probability value (p-value), or equivalently find a rejection region. A p-value is the probability of seeing data at least as extreme as $t\star$ under the assumption of H_0.

5) The rejection region is found by using α to find a critical value; the rejection region is the area that is more extreme than the critical value.

6) Make a test decision about the null hypothesis.
Reject H_0 if the p-value < α.

Interpretation of SLR Model Summary Output in R

Residuals: If the residuals are normally distributed with mean 0 (and constant variance), this should be consistent with the values reported here. The median would likely be close to 0 and the symmetry of the distribution would likely be reflected in the values of the other four numbers here as well, which would be expected to approximately balance.

Estimate: Computed estimates for the intercept and x-variable coefficient (slope).

Std. Error: This is the value $\mathrm{SE}\left[\hat{\beta}_i\right], i = 0,1$, for the corresponding estimator for the standard deviation of $\hat{\beta}_i$ that we have introduced in previous slides.

$$\mathrm{SE}\left[\hat{\beta}_0\right] = \hat{\sigma}\left(\frac{1}{N} + \frac{x^2}{S_{xx}}\right)^{\frac{1}{2}},\ \mathrm{SE}\left[\hat{\beta}_1\right] = \frac{\hat{\sigma}}{(S_{xx})^{\frac{1}{2}}}, \text{where } \hat{\sigma}^2 = \frac{1}{N-2}\sum_{n=1}^{N} e_n^2$$

t value: This is actually the Estimate as above divided by the Std. Error. A larger value implies more confidence in the corresponding parameter estimate.

Pr(>|t|): This is the p-value – probability value – for the associated t-test on the linear model parameters. Since for SLR this test is essentially equivalent to the F-test below, we defer more discussion of it until we consider Multiple Regression.

Signif. Codes: The significance codes indicate how certain we can be that the coefficient has an impact on the dependent variable. For example, a significance level of 0.001 indicates that there is less than a 0.1% chance that the coefficient might be equal to 0 and thus be insignificant. Stated differently, we can be 99.9% sure that it is significant. The significance codes (shown by asterisks) are intended for quickly ranking the significance of each variable.

Residual Standard Error: This is our estimate $\hat{\sigma}$ (introduced in previous slides) for the standard deviation σ of our exact or true SLR regression model:

$$\text{RSE-squared is } \hat{\sigma}^2 = \frac{1}{N-2}\sum_{n=1}^{N} e_n^2.$$

(Multiple) R-squared(Coefficient of Determination): Measures, in a suitable sense, the proportion of the variance in the dependent variable that is predictable from the independent variable(s). Defined as

$$R^2 = 1 - \frac{\sum_{n=1}^{N} e_n^2}{\sum_{n=1}^{N}(y_n - y)^2}, \text{where } \bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n.$$

Adjusted R-squared: We defer discussion on this until we get to multiple regression.

F-statistic: This, here 89.57, is the value of the F-statistic corresponding to an F Test (F Hypothesis Test) for the SLR model intended to assess whether the regression model is significant, that is, whether a non-zero value for $\hat{\beta}_1$ yields a model that better explains the data than simply taking $\hat{\beta}_1$=0, which corresponds to an "intercept-only" model. Note that "1 and 48 DF" corresponds to the difference in the number of parameters between the SLR model and an intercept-only one – which is 1 – and the number of degrees of freedom in the SLR model, which here is 50-2=48.

p-value: This is the p-value for the F-test (Significance of Regression test).

```
stop_distance_model = lm(dist ~ speed, data = cars)
summary(stop_distance_model)
Call:
lm(formula = dist ~ speed, data = cars)
Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464  1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 Degrees of Freedom
Multiple R-squared:  0.6511,  Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Lecture 4 Overview

- Coefficient of Determination
- F-test for significance of Simple Linear Regression
- Multiple Linear Regression
- Least-squares solution of Multiple Regression
- Various computational examples and implementations in R

Coefficient of Determination (R^2)

The **Coefficient of Determination R^2** (R-squared) is a goodness-of-fit measure of the proportion of the variance in the response variable that is explained from the predictor variable(s). The R^2 can be viewed as a measure of regression model accuracy. Define

$$R^2 = 1 - \frac{\sum_{n=1}^{N} e_n^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2}, \text{where } \bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n. \qquad (33)$$

Now the **Decomposition of Variation (DoV)** equation, which holds in the context of Simple Linear Regression models (as well as more generally for Multiple Linear Regression, which we discuss later on) is the following:

$$\sum_{n=1}^{N}(y_n - \bar{y})^2 = \sum_{n=1}^{N}(y_n - \hat{y}_n)^2 + \sum_{n=1}^{N}(\bar{y} - \hat{y}_n)^2 \quad \text{(requires proof)} \qquad (34)$$

So, DoV shows that $0 \leq R^2 \leq 1$. Larger values of R^2 are generally viewed as more promising for the validity of the model. Part of the reason for this is that a lower value for $\sum_{n=1}^{N}(y_n - \hat{y}_n)^2$ clearly implies a better model. But with the DoV we can say more. From (33) and (34), we see that

$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} = \frac{\sum_{n=1}^{N}(\hat{y}_n - \bar{y})^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2}$. Since $\sum_{n=1}^{N}(z_n - \bar{z})^2$ is an estimate (up to a constant factor) of the variance of any given random variable Z and we also have $R^2 \leq 1$, a high value for R^2 appears consistent with a model whose predictor variable *X* (or variables for multiple regression) explains more (or much) of the variance or variation in the response *Y*.

F-test for Significance of Simple Linear Regression

The *F*—test (results of which are at the bottom of the output summary report for the R lm( ) function, see a previous slide) concerns the following statistical Hypothesis Test for SLR:

$$H_0 : \beta_1 = 0 \text{ and } H_1 : \beta_1 \neq 0$$

So, this test in the case of SLR tells us whether the predictor variable adds any explanatory value to the model at all. That is, it tells us whether employing the predictor variable makes the model more complex than it needs to be and simply including the β_0 parameter alone would suffice -- or not. The *F*-statistic for the F-test is defined by

$$F = \frac{\sum_{n=1}^{N}(\hat{y}_n - \bar{y})^2}{\left(\frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{N-2}\right)} \text{ where } \hat{y}_n := \hat{\beta}_0 + \hat{\beta}_1 x_n. \qquad (35)$$

Using Decomposition of Variation, we can rewrite this in the form

$$F = \frac{\sum_{n=1}^{N}(y_n - \bar{y})^2 - \sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\left(\frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{N-2}\right)}, \qquad (36)$$

and this representation appears consistent with the magnitude of the statistic F rising the better H_1 explains the distribution of the sample data points, i.e., the better taking $\hat{y}_n := \hat{\beta}_0 + \hat{\beta}_1 x_n$ explains it. This is because the better H_1 explains the distribution of the sample data points the smaller the denominator is (making F larger) and the smaller in absolute the second term in the numerator is as well (again tending to make *F* larger as the term considered in absolute value is subtracted).

Under the Null Hypothesis, it is known that the *F*-statistic should follow an $F(\cdot,\cdot)$ probability distribution with respective parameters $(d_2 - d_1, N - d_2)$, where $d_2 - d_1$ is the difference in the number of regression function parameters between H_1 and H_0, 2-1=1 in this case, and $N - d_2$ is the number of degrees of freedom for H_1, $N - 2$ in this case. From the summary report we obtain the associated *p*-value corresponding to the model and the data. Indeed, under the assumption that the Null Hypothesis is true, the *p*-value is the probability of the value of the F -statistic being as large as it is or larger. Hence, in essence, a very low p-value (for example, one less or even much less than 0.05, which corresponds to a 95% confidence interval) implies that we should reject the Null Hypothesis H_0, and hence also implies the significance of the SLR model in this case with both parameters – intercept and slope – significant as well.

Multiple Linear Regression

Of course many datasets feature multiple predictor variables. Indeed a response variable may naturally depend on a number (> 1) of explanatory variables. So we extend our current linear model to allow a response to depend on *multiple* predictors. This is called **Multiple Linear Regression**, or simply **Multiple Regression**. Many aspects of our Simple Linear Regression (SLR) model extend fairly naturally to the multiple-predictor setting, and this does in fact hold true for the general definition of the multiple regression model itself:

For any given sample $(x_1, x_2, \dots, x_M)$, we have the general underlying model

$$\boldsymbol{E}[Y] = \boldsymbol{E}[Y|X_1 = x_1, X_2 = x_2, \dots, X_M = x_M] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M, \qquad (37)$$

where Y is a continuous random variable. For N fixed $(\text{non} - \text{random})$ sample vectors $(x_{11}, \dots, x_{1M})$, $(x_{21}, , \dots, x_{2M})$,…,$(x_{N1}, \dots, x_{NM})$ we have, in analogy with SLR, the corresponding set of equations (with random noise terms)

$$Y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_M x_{nM} + \epsilon_n, \qquad (38a)$$

where $\epsilon_n \sim N(0, \sigma^2), n = 1, \dots, N,$ with independent random noise terms ϵ_n. (38b)

Least-Squares Solution of Multiple Linear Regression

Given our Multiple Regression (MR) model

$$Y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_M x_{nM} + \epsilon_n, n = 1, \ldots, N, \quad (39)$$

$$\epsilon_n \sim N(0, \sigma^2), n = 1, \ldots, N, \text{ where the terms } \epsilon_n \text{ are independent r.v.'s} \quad (40)$$

we can, as in the case of simple linear regression, compute estimators

$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_M)$ corresponding to the coefficients $(\beta_0, \beta_1, \ldots, \beta_M)$ as above. This we do, once again by means of least-squares optimization (minimization):

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_M) = \arg\min_{(\alpha_0, \alpha_1, \ldots, \alpha_M) \in \mathbb{R}^{M+1}} \sum_{i=1}^{N} \left(y_i - (\alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_M x_{iM})\right)^2, \quad (41)$$

Perhaps the most natural way to compute a solution to this minimization problem is to rewrite our MR model equations in terms of matrices, so define

$$\mathbb{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \mathbb{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1M} \\ 1 & x_{21} & \cdots & x_{2M} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N1} & \cdots & x_{NM} \end{bmatrix}, \mathbb{B} = \begin{bmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix}, \mathbb{E} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}, \text{ and } \widehat{\mathbb{B}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_M \end{bmatrix} \quad (42)$$

and (39)-(40) can be rewritten as $\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{E}$. It can then be shown, using differential calculus, that a solution $\widehat{\mathbb{B}}$ to (41) always exists and must satisfy

$$(\mathbb{X}^T \mathbb{X})\widehat{\mathbb{B}} = \mathbb{X}^T \mathbb{Y}. \quad (43)$$

Lecture 5 Overview

Topics in this lecture include:

- Basic Concepts and Results in Multiple Linear Regression
- Sampling Distribution for the $\hat{\beta}_m$ for Multiple Regression
- Confidence Intervals for β_m for Multiple Regression

Multiple Linear Regression

Of course many datasets feature multiple predictor variables. Indeed a response variable may naturally depend on a number (> 1) of explanatory variables. So we extend our current linear model to allow a response to depend on *multiple* predictors. This is called **Multiple Linear Regression**, or simply **Multiple Regression**. Many aspects of our Simple Linear Regression (SLR) model extend fairly naturally to the multiple-predictor setting, and this does in fact hold true for the general definition of the multiple regression model itself:

For any given sample $(x_1, x_2, \dots, x_M)$, we have the general underlying model

$$\boldsymbol{E}[Y] = \boldsymbol{E}[Y|X_1 = x_1, X_2 = x_2, \dots, X_M = x_M] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M, \qquad (37)$$

where Y is a continuous random variable. For N fixed $(\text{non} - \text{random})$ sample vectors $(x_{11}, \dots, x_{1M})$, $(x_{21}, , \dots, x_{2M})$,…,$(x_{N1}, \dots, x_{NM})$ we have, in analogy with SLR, the corresponding set of equations (with random noise terms)

$$Y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_M x_{nM} + \epsilon_n, \qquad (38a)$$

where $\epsilon_n \sim N(0, \sigma^2), n = 1, \dots, N,$ with independent random noise terms ϵ_n. (38b)

Least-Squares Solution of Multiple Linear Regression

Given our Multiple Regression (MR) model

$$Y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_M x_{nM} + \epsilon_n, n = 1, \ldots, N, \quad (39)$$

$$\epsilon_n \sim N(0, \sigma^2), n = 1, \ldots, N, \text{ where the terms } \epsilon_n \text{ are independent r.v.'s} \quad (40)$$

we can, as in the case of simple linear regression, compute estimators

$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_M)$ corresponding to the coefficients $(\beta_0, \beta_1, \ldots, \beta_M)$ as above. This we do, once again by means of least-squares optimization (minimization):

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_M) = \arg\min_{(\alpha_0, \alpha_1, \ldots, \alpha_M) \in \mathbb{R}^{M+1}} \sum_{i=1}^{N} \left(y_i - (\alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_M x_{iM})\right)^2, \quad (41)$$

Perhaps the most natural way to compute a solution to this minimization problem is to rewrite our MR model equations in terms of matrices, so define

$$\mathbb{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \mathbb{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1M} \\ 1 & x_{21} & \cdots & x_{2M} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N1} & \cdots & x_{NM} \end{bmatrix}, \mathbb{B} = \begin{bmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix}, \mathbb{E} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}, \text{ and } \widehat{\mathbb{B}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_M \end{bmatrix} \quad (42)$$

and (39)-(40) can be rewritten as $\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{E}$. It can then be shown, using differential calculus, that a solution $\widehat{\mathbb{B}}$ to (41) always exists and must satisfy

$$(\mathbb{X}^T \mathbb{X})\widehat{\mathbb{B}} = \mathbb{X}^T \mathbb{Y}. \quad (43)$$

Least-Squares Solution of Multiple Regression (cont'd)

Given our Multiple Regression model with sample data $\mathbb{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1M} \\ 1 & x_{21} & \cdots & x_{2M} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N1} & \cdots & x_{NM} \end{bmatrix}$, $\mathbb{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$ and solution vector $\widehat{\mathbb{B}}$ of estimators $\widehat{\mathbb{B}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_M \end{bmatrix}$ that satisfies $(\mathbb{X}^T\mathbb{X})\widehat{\mathbb{B}}=\mathbb{X}^T\mathbb{Y}$ from the previous slide, note that we can quite simply take

$$\widehat{\mathbb{B}}=(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y} \qquad (44)$$

provided the matrix inverse of $\mathbb{X}^T\mathbb{X}$ exists. One key case in which it does exist is when the columns of $\mathbb{X}$ are linearly independent. However, even if $\mathbb{X}^T\mathbb{X}$ is not invertible -- or is not invertible in the conventional sense but may require the concept of the "generalized inverse" of a matrix -- there still exist efficient algorithms for computing a matrix $\widehat{\mathbb{B}}$ satisfying $(\mathbb{X}^T\mathbb{X})\widehat{\mathbb{B}}=\mathbb{X}^T\mathbb{Y}$ in any case. For the most part in this course, we will generally make the assumption that the inverse $\mathbb{X}^T\mathbb{X}$ exists or that there does exist a "generalized inverse" that satisfies the key properties of the true inverse that we need here in the context of multiple regression.

Expectation of the $\widehat{\mathbb{B}}$ matrix

Making the assumption that $\mathbb{X}^T\mathbb{X}$ is an invertible matrix, we want to compute $\mathbf{E}(\widehat{\mathbb{B}})=\mathbf{E}((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y})$, where we define the expectation or mean of a random vector $V=(V_1,\ldots,V_d)$ of length d, denoted $\mathbf{E}(V)$, to also be a vector of the same length, having $\mathbf{E}(V_i)$ as its ith component. Also, with respect to such a random vector V, we have, for any matrix A whose entries are fixed (non-random) and with d columns, the identity:

$$\mathbf{E}(AV)=A\mathbf{E}(V), \tag{45}$$

Then,

$$\begin{aligned}\mathbf{E}(\widehat{\mathbb{B}}) &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{E}(\mathbb{Y}) \\ &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T(\mathbb{X}\mathbb{B}) \\ &= \mathbb{B}.\end{aligned} \tag{46}$$

Variance of the $\widehat{\mathbb{B}}$ matrix

Want to compute

$$\mathrm{Var}(\widehat{\mathbb{B}})=\mathrm{Var}((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}), \quad (47)$$

where we define the variance of a random vector $V=(V_1,\ldots,V_d)$ length d to be a dxd matrix, denoted $\mathrm{Var}(V)$, having $\mathrm{Var}(V_i)$ as its (i,i) entry and $\mathrm{Cov}(V_i,V_j)=\mathrm{Cov}(V_j,V_i)$ as both its (i,j) and its (j,i) entries. We have, for any matrix A whose entries are fixed (non-random) and with d columns, the identity:

$$\mathrm{Var}(AV)=A\mathrm{Var}(V)\,A^T, \quad (48)$$

A^T denoting the transpose of A. Hence,

$$\begin{aligned}\mathrm{Var}(\widehat{\mathbb{B}})&=\mathrm{Var}((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y})\\ &=(\mathbb{X}^T\mathbb{X})^{-1}\mathrm{Var}(\mathbb{X}^T\mathbb{Y})\,((\mathbb{X}^T\mathbb{X})^{-1})^T\\ &=(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathrm{Var}(\mathbb{Y})\;\mathbb{X}((\mathbb{X}^T\mathbb{X})^{-1})^T \quad (49)\\ &=(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\sigma^2 I\,\mathbb{X}((\mathbb{X}^T\mathbb{X})^{-1})^T\\ &=\sigma^2\,(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\\ &=\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\end{aligned}$$

using such properties as $(A^{-1})^T=(A^T)^{-1}$, $(AB)^{-1}=B^{-1}A^{-1}$, and $(AB)^T=B^TA^T$, as well as the fact that the entries of $\mathbb{Y}$ are independent (meaning, here, independence in the probabilistic/statistical sense).

Sampling Distribution for the $\hat{\beta}_m$ for Multiple Regression

We update our estimate for σ^2 in the case of Multiple Regression, and now take our estimate for it to be

$$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{n=1}^{N}(y_n - \hat{y})^2}{N-M-1}, \qquad (50)$$

where we had to reduce the value of the denominator to reflect the additional regression parameters. We have $\boldsymbol{E}[s_e^2] = \sigma^2$, so s_e^2 is an unbiased estimator for σ^2.

We have that $\widehat{\mathbb{B}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$ where as usual we assume that the $\mathbb{X}$ matrix is non-random (deterministic), which means that since the entries of $\mathbb{Y}$ are i.i.d. and normally-distributed, the entries of the $\widehat{\mathbb{B}}$ matrix are normally-distributed as well since they are finite, linear combinations of the components of $\mathbb{Y}$.

So, for each $m = 0, \ldots, M$, we have, using (46) and (49), respectively, from prior slides,

$$\hat{\beta}_m \sim N(\beta_m, \sigma^2((\mathbb{X}^T\mathbb{X})^{-1})_{mm}). \qquad (51)$$

Moreover, in analogy with the case of SLR, we have the following estimate $\mathrm{SE}(\hat{\beta}_m)$ for the standard deviation $\sigma\sqrt{((\mathbb{X}^T\mathbb{X})^{-1})_{mm}}$ (which is the square root of the variance value appearing in (51)) of $\hat{\beta}_m$:

$$\mathrm{SE}[\hat{\beta}_m] = s_e\sqrt{((\mathbb{X}^T\mathbb{X})^{-1})_{mm}}, \qquad (52)$$

where $s_e^2 = \frac{\sum_{n=1}^{N}(y_n - \bar{y})^2}{N-M-1}$ is as above.

Confidence Interval for β_m for Multiple Regression

Similarly to the case for SLR, since we must estimate σ^2 and we do so using s_e^2, $\hat{\beta}_m$ as a random variable is known to follow in practice a t-distribution once σ^2 is replaced with the approximation s_e^2 . The t-distribution, like the normal distribution, is symmetric and bell-shaped, but has heavier tails than its more famous cousin so is more apt to generate values that fall further from its mean. It is a special case of the generalized hyperbolic distribution. So we have that, for each, $m = 0, \ldots, M$,

$$\frac{\hat{\beta}_m - \beta_m}{\mathrm{SE}[\hat{\beta}_m]} \sim t_{(N-M-1)} \qquad (53)$$

So, we can obtain confidence intervals for the true values of the β_m as well:

$$\hat{\beta}_m - t\left(\frac{\alpha}{2}, N-M-1\right)\mathrm{SE}\left[\hat{\beta}_m\right] \leq \beta_m \leq \hat{\beta}_m + t\left(\frac{\alpha}{2}, N-M-1\right)\mathrm{SE}\left[\hat{\beta}_m\right], \quad (54)$$

with $(1-\alpha)$x100% confidence, where, as with SLR, $t\left(\frac{\alpha}{2}, N-M-1\right)$ is the value that cuts off $\alpha/2 \times 100\%$ in the *upper tail* of the *t*-distribution for $N-M-1$ degrees of freedom.

R commands: model_147= lm(hp ~ wt + cyl, data = mtcars)

confint(model_147, level = 0.99)

Result: 0.5 % 99.5 %

(Intercept) -124.10755 20.49642

wt -30.57995 33.24087

cyl 13.90506 48.87074

Lecture 6 Overview

Topics in this lecture include:

- The Residuals in Multiple Regression
- F-test for significance of Multiple Regression
- Single parameter significance test for Multiple Regression
- Computational examples with R
- Review of Solutions to Homework Assignment #1

The Residuals for Multiple Regression

Recall of course our Multiple Regression (MR) model

$$Y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_M x_{nM} + \epsilon_n, n = 1, \ldots, N, \quad (55)$$

$$\epsilon_n \sim N(0, \sigma^2), n = 1, \ldots, N, \quad (56)$$

We are not able to sample the errors $\epsilon_n, n = 1, \ldots, N$, 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 for SLR.

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

$$e_n := y_n - \hat{y}_n, \text{ where } \hat{y}_n := \hat{\beta}_0 + \hat{\beta}_1 x_{n1} + \cdots + \hat{\beta}_M x_{nM}, \; n = 1, \ldots, N, \quad (57)$$

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

$$\hat{\sigma}^2 = s_e^2 = \frac{1}{N-M-1} \sum_{n=1}^{N} e_n^2, \quad (58)$$

where $\hat{\sigma} = s_e$ (as also previously defined in (50)) is known as the **Residual Standard Error (RSE).** Note the factor $\frac{1}{N-M-1}$ appearing in (58).

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

Computing the RSE with R:

```
model_147= lm(hp ~ wt + cyl, data = mtcars)
summary(model_147)$sigma
```

F-test for Significance of Multiple Regression

The *F*—test (results of which are at the bottom of the output summary report for the R lm( ) function) concerns the following statistical Hypothesis Test for Multiple Regression:

$H_0 : \beta_1 = \ldots = \beta_M = 0$ (intercept-only model) and $H_1 : \beta_m \neq 0$ for at least one *m=1,…,M*

So, this test tells us whether the predictor variables add any explanatory value to the model at all. That is, it tells us whether including any predictor variables in the regression model makes the model more complex than it needs to be and simply including the β_0 parameter alone would suffice -- or not. The *F*-statistic for the F-test is defined by

$$F = \frac{\frac{\sum_{n=1}^{N}(\hat{y}_n - \bar{y})^2}{M}}{\left(\frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{N-M-1}\right)} \text{ where } \hat{y}_n := \hat{\beta}_0 + \hat{\beta}_1 x_{n1} + \cdots + \hat{\beta}_M x_{nM},\ n = 1, \ldots, N, \quad (59)$$

Using Decomposition of Variation as we have introduced in a previous slide, we can rewrite this in the form

$$F = \frac{\frac{\sum_{n=1}^{N}(y_n - \bar{y})^2 - \sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{M}}{\left(\frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{N-M-1}\right)}, \quad (60)$$

and this representation appears consistent with the magnitude of the statistic F rising the better H_1 explains the distribution of the sample data points, i.e., the better taking $\hat{y}_n$ explains it. This is because the better H_1 explains the distribution of the sample data points the smaller the denominator is (making F larger) and the smaller in absolute the second term in the numerator is as well (again tending to make *F* larger as the term considered in absolute value is subtracted).

Under the Null Hypothesis, it is known that the *F*-statistic should follow an $F(\cdot,\cdot)$ probability distribution with respective parameters $(d_2 - d_1, N - d_2)$, where $d_2 - d_1$ is the difference in the number of regression function parameters between H_1 and H_0, and $N - d_2$ is the number of degrees of freedom for H_1. From the summary report we obtain the associated *p*-value corresponding to the model and the data. Indeed, under the assumption that the Null Hypothesis is true, the *p*-value is the probability of the value of the F -statistic being as large as it is or larger. Hence, in essence, a very low p-value (for example, one less or even much less than 0.05, which corresponds to a 95% confidence interval) implies that we should reject the Null Hypothesis H_0, and hence also implies the significance of the non-intercept only Multiple Regression model in this case.

Single-parameter significance test (t-test) for multiple regression

The R language lm( ) function summary report contains the results of the following Hypothesis Test in the multiple regression context:

For any given $m = 1,2, \ldots, M, H_0: \beta_m = 0$ vs. $H_1: \beta_m \neq 0$.

So, this test assesses whether the response variable Y depends linearly on the m-th predictor x_m in any significant way. The null hypothesis H_0, if true, implies no significant dependence on x_m.

We apply a t-test with test statistic defined by

$$t_{(N-M-1)} = \frac{\hat{\beta}_m - \beta_m}{\mathrm{SE}[\hat{\beta}_m]} = \frac{\hat{\beta}_m}{s_e\sqrt{((\mathbb{X}^T\mathbb{X})^{-1})_{mm}}}, \text{ where } s_e^2 = \frac{\sum_{n=1}^{N}(y_n - \bar{y})^2}{N-M-1}. \qquad (61)$$

(in keeping with the test statistic typically satisfying (estimate – hypothesis)/standard error)

which, under the null hypothesis, follows a t-distribution with *N–M-1* degrees of freedom. The p-value for the corresponding t-test is the probability that a corresponding t-distributed random variable would take on a value greater than or equal to the absolute value of the t_0 statistic as above.