Simple Linear Regression, Linear Models

Stats 203, Winter 2024

Lecture 1 [Last update: January 7, 2024]

1 Introduction

- In linear regression we want to understand the effect of one variable on another, using data $(x_1, y_1), \ldots, (x_n, y_n)$. The xs and ys play different roles: we want to use x to explain or predict the variations of y. For this reason we typically refer to x the *explanatory* or *predictor* variable, and y the *outcome* or *response* variable. We will call this kind of a problem a *regression problem*.
- Some examples of regression problems, from the Elements of Statistical Learning textbook:
  - Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.
  - Identify the numbers in a handwritten ZIP code, from a digitized image.
  - Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data.
  - Estimate the amount of glucose in the blood of a diabetic person, from the infrared absorption spectrum of that person's blood.
  - Identify the risk factors for prostate cancer, based on clinical and demographic variables.

  Notice the diversity in applications, and also in the ultimate goal of the data analysis. These five examples alone cover *predicting* a future event, *estimating* the value of an extant but unobserved response, and *deciding* which of several explanatory variables actually matter for the response.
- We will see that tackling a regression problem involves several kinds of tasks:
  1. *Modeling.* Deciding what the relevant variables are. Expressing a probabilistic relation between the variables, up to a few unknowns (i.e. the **parameters**.)
  2. *Estimation.* Using the data to guess at (i.e. estimate) the unknowns in the model.
  3. *Inference.* Attaching a notion of uncertainty (for instance, a **standard error** or **confidence interval**) to our estimates. Deciding whether there is truly a relationship between predictors and response (i.e. running a **hypothesis test**).
  4. *Prediction.* Given newly observed predictors, using the fitted model to guess at (i.e. predict) what the response will be.

  Our class will cover the most classical method of regression, called linear regression analysis, or just linear regression.
- Regression can be viewed as a particular kind of statistical inference. In a first class on statistical inference (like Stats 200), you will have learned how to guess at (i.e. infer) the value of something unknown (i.e. a parameter) from something observed (i.e. the data). Regression has a similar aim, but now what is unknown is the effect that one set of variables (the xs) has on another (y). For instance, previously we might have estimated the average height of a population using the average height of a

100-person random sample. Now, we might be interested in estimating the effect of malnutrition on height, using the same 100-person sample.

Some other differences are that in regression, we will spend more time on prediction, on data analysis, on checking assumptions, and on communicating results. This means we will have less time to consider different kinds of models.

- Regression can also be viewed as particular kind of predictive data analysis. Compared to another predictive data analyses, regression analysis – and thus this class – will spend more time on inference – standard errors, hypothesis tests, confidence intervals – and checking assumptions. We will spend less time talking about different methods for prediction.
- Our class is divided into two parts. Part I covers the canonical statistical approach for solving linear regression problems, known (fittingly) as *linear regression*. The material here will be quite classical. Part II starts with diagnostics (also classical) and then covers several important more modern developments (modern = 1960s and onwards, mostly) such as the *bootstrap*, *model selection*, *shrinkage*, and *generalized linear models*. Part I will be more mathematical, while Part II will be more applied; this is in part because we have an incomplete mathematical understanding of some of the topics in Part II, and in part because the theory we do have tends to be quite a bit harder.
- Today we will cover simple linear regression, where each x_i is just a single number. Let's look at an example.

2 Example: rocket fuel

- *Background*: this example is from the textbook Introduction to Regression Analysis by Montgomery, Peck and Vining (MPV). Quoting from page 15:

  > A rocket motor is manufactured by bonding an igniter propellant and a sustainer propellant together inside a metal housing. The shear strength of the bond between the two types of propellant is an important characteristic. It is suspected that shear strength is related to the age in weeks of the batch of sustainer propellant.

  We can think of this as a prediction problem. We want to use the age of sustainer propellant (the predictor variable x) to predict how strong the bond between propellants will be (the outcome variable y).

- *Exploratory data analysis, and the model*: We are given $n = 20$ observations of two variables: age of propellant (in weeks), and shear strength (in pounds per square inch). As always, we start with some *exploratory data analysis*, by plotting the data in Figure 1. The relationship between the variables seems to approximately follow a straight line up to some minor error So the following model seems reasonable:

$$ShearStrength = \beta_0 + \beta_1 PropellantAge + \epsilon, \tag{1}$$

  where ϵ represents the error. The notation "ϵ" is used to suggest that the error is small, but it does not have to be. The intercept β_0 and slope β_1 are unknown parameters.

  For some examples where EDA shows that the linear model is *not* appropriate, see Section 6.

- *Estimation and prediction.* We estimate β_0, β_1 by passing a line "through" the observed data points. The resulting line – sometimes called the line of best fit, for obvious reasons – plotted on the right hand side of Figure 1. Mathematically, the line of best fit satisfies the equation:

$$\widehat{ShearStrength} = 2627.82 - 37.15 \times PropellantAge.$$

  The "hat" over Shear Strength is used to mark it as a prediction rather than an observed data point. So, for example, if a 15 week old sustainer propellant was used, we would predict a shear strength of $2627.82 - 37.15 \times 15 = 2070.52$ psi. If a brand new sustainer propellant was used, we would predict a shear strength of $2627.82 - 37.15 \times 0 = 2627.82$ psi.

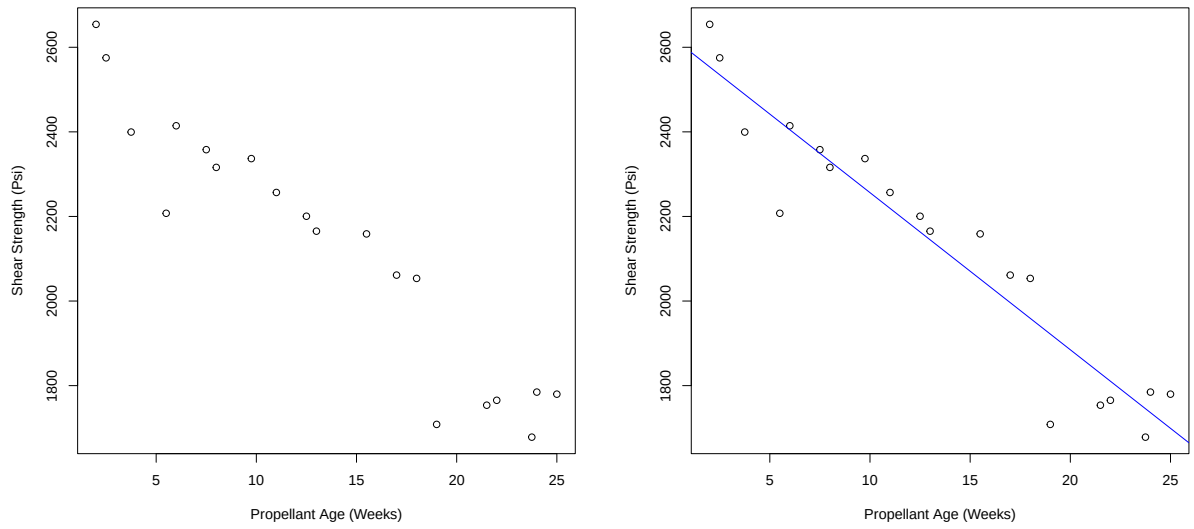

Figure 1: Rocket data.

- We've just fit our first linear model. Although the example was particular to rocket fuel, it is hopefully clear that the same steps could in principle be applied to any dataset. In fitting the model I glossed over three important aspects: (1) a precise definition of the line of best fit, and more abstractly, (2) what the linear model actually means. The rest of class will cover these two points.
- The third aspect that I omitted was how the line of best fit was actually calculated on the computer. In many scientific programming languages there is a function that will do this (and much more) for you (e.g. the function `lm` in *R*). But it is also simple enough to do it "by hand" using the formulas below.

3 Linear models

- In the classical approach to regression analysis, one begins by specifying a regression model. Informally, we can think of this as a set of mathematical equations that describe an underlying process by which the data was generated.
- For the rocket fuel example, we modeled the data by (1). Here is a similar looking but more abstract model, where we have replaced the variable names "*ShearStrength*" and "*PropellantAge*" with the dummy variables Y and X: we suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent observations, distributed according to

$$\begin{aligned} &X_i \text{ comes from an unknown (possibly deterministic) distribution} \\ &Y_i = \beta_1 X_i + \beta_0 + \epsilon_i, \text{where} \\ &\epsilon_i \perp\!\!\!\perp X_i, \mathbb{E}[\epsilon_i] = 0, \mathrm{Var}[\epsilon_i] = \sigma^2. \end{aligned} \tag{2}$$

  This is known as the *simple linear model*. The notation $\epsilon \perp\!\!\!\perp X$ means the random variables ϵ and X are independent, $\mathbb{E}$ means expectation, and Var means variance. We will review independence, expectation, and variance next class.
- In (2) there are three kinds of objects:
  1. The *data*, predictors X_i and responses Y_i, which are observed.
  2. The *parameters* $\beta_0, \beta_1, \sigma^2$ which are fixed but not observed.
  3. The *errors* ϵ_i which are random and not observed.

The random errors are included to make the model more realistic: in real data the response is rarely a perfect linear function of the explanatory variable. The predictors $X_1, \ldots, X_n$ and responses $Y_1, \ldots, Y_n$ are capitalized to indicate they are random variables. Blue Latin letters are used to refer to things that are observed, and red Greek letters to things that are not observed.

- Model (2) imposes several restrictions on the data generating process. It says that each observation is independent, that the errors are independent of the predictors, and that the errors all have the same variance. Most restrictively, it asserts that the **conditional mean** function $m(x) := \mathbb{E}[Y|X = x]$ is linear:

$$m(x) = \beta_1 x + \beta_0.$$

  Freedman gives an intuitive way of understanding the conditional mean via the "graph of averages", for discussion of this point see Section 6.

- Model (2) is only one of several different models. Here are some other examples: in each $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent observations, the errors $\epsilon_i \perp\!\!\!\perp X_i, \mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2$, and additionally
  - *Intercept-only model*: $Y_i = \beta_0 + \epsilon_i$,
  - *No unknown parameters*: $Y_i = 2.07X_i + \epsilon_i$, and $\sigma^2 = 1$. In this model there are no unknown parameters, hence no role for statistical inference.
  - *Normal predictors*: $Y_i = \beta_1 X_i + \beta_0 + \epsilon_i$, $\mathbb{E}[\epsilon_i] = 0$, and $X_i \sim N(\mu_X, \sigma_X^2)$.
  - *Normal errors*: $Y_i = \beta_1 X_i + \beta_0 + \epsilon_i$, where $\epsilon_i \perp\!\!\!\perp X_i$ and $\epsilon_i \sim N(0, \sigma^2)$.
  - *Nonparametric*: $Y_i = m(X_i) + \epsilon_i$. Here there is no parametric form specified for the conditional mean. This allows for more flexible relationships between predictors and response. Data analysis in this model is known as *nonparametric regression*. We won't cover that in this class.

- In class we will focus on regression analysis in the linear model. For what data should we use the linear model? The best answer is that the linear model should be used only when we have reason to believe it is plausible for the data that we have. There are a couple of cases when this might be true:
  - We have concluded that the linear model is reasonable by performing some EDA. This was the case for our rocket fuel example.
  - Prior knowledge tells us that the variables we have measured truly follow some exact (linear) physical model, with observations deviating slightly from this law due to (e.g.) measurement error. Freedman gives the example of Hooke's law

$$stretch_i = \beta_0 + \beta_1 weight_i + \epsilon_i,$$

    as an example of a direct physical (and causal) model.
  - We have reason to believe that the linear model is a reasonable approximation to the truth. For example, if $\mathbb{E}[Y|X = x] = m(x)$ for some "smooth" function m, and most values of the predictor are near $x = 0$ (say), then Taylor's theorem tells us that

$$m(x) \approx m(0) + m'(0)x.$$

    In this case, the errors represent the "missing terms" in a Taylor expansion, which we (optimistically) treat as mean-zero, independent random variables. In fact, regression was originally discovered and used for geodesy and astronomy, where the linear model is exactly such a "first-order" approximation to the truth.
  - The linear model is a direct consequence of a stronger assumption that we are already willing to commit to. For example, later on we will see that the linear model follows from multivariate Normality of the Xs and Ys.

- Particularly in modern applied work, many (perhaps the majority) of regression problems do not fall neatly into any of the above. Why do we nevertheless spend a whole course on linear models? There are several reasons:

  - The traditional statistics curriculum starts with linear regression because historically, it came first.
  - Linear functions are "simple" and easier to understand than more complicated regression models. Despite this, if we work hard enough, we can cast many other regression methods as "linear regression plus bells and whistles."
  - Our understanding of linear regression is more complete than our understanding of other regression methods.
  - Thanks in part to points 1-3, in the "real world" you will frequently be called upon to engage with linear regression, whether the analysis was done by you or someone else. We feel it is important that you understand the assumptions underlying these analyses.

- Let me concluding by making a philosophical but ultimately crucial point. All of the models above are simply mathematical constructs. They are not real, and thus neither true nor false in any real sense. The data, on the other hand, are observed quantities and (a priori) have no connection to the model. What we as statisticians do is to force a connection between model and data by "assuming" that the observed data are realized values of random variables which obey the model. This is really a gigantic leap of faith, which is only justified by the "unreasonable effectiveness" (Eugene Wigner) of (mathematical) models in describing and understanding real-world phenomena.

  The reason this is important is that in the end, you will have to get over your unease about whether the model is ever correct – philosophically, this is kind of nonsense anyways – and make the leap of faith. Eventually, you will come to realize that the question to ask is not whether the model is correct but whether it is plausible: that is, would realized values of random variables drawn from the model "look like" the observed data we actually have. Tools like *exploratory data analysis* and *diagnostics* help answer these questions.

  The statistician George Box has very famous quote which captures this more succinctly: "all models are wrong, but some models are useful."

4 The least-squares line

- Having written down one or the other of the linear models, we are faced with estimating the unknown parameters using the observed data. The most common approach for getting estimates is the **method of least squares**. The resulting line defined by these estimates is known as the least-squares line, or the least-squares regression line, or the line of best fit. Freedman just calls it the regression line.
- The method of least squares is an example of a common approach to estimating parameters in statistical problems. First, introduce a measure of the discrepancy between a candidate value of parameters and the observed data, called a **loss function**. Then, estimate the parameters by minimizing the loss.
- In the method of least squares, the loss function we use is called the (sample) **mean squared error (MSE)**. In the case of the linear model, the sample MSE is

$$\mathrm{MSE}_n(b_0, b_1) = \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - (b_0 + b_1 x_i)\Big)^2$$

  Notice that MSE depends explicitly only on b_0, b_1 but is clearly also a function of the data. This is indicated notationally by the subscript n. The **least-squares line** is defined as the line that minimizes the MSE:

$$\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x, \quad \text{where} \quad (\hat{\beta}_0, \hat{\beta}_1) := \operatorname*{argmin}_{b_0, b_1} \mathrm{MSE}_n(b_0, b_1). \tag{3}$$

- Warning: it is common to mix up estimates $(\hat{\beta}_0, \hat{\beta}_1)$ with parameters (β_0, β_1). Don't do that! Parameters are unknown quantities, and cannot be computed from the data, whereas estimates are purely functions of the observed data.

- To compute the least-squares estimates, we take partial derivatives of the sample MSE with respect to b_0, b_1, and set them equal to 0. Letting $\bar{x} := \frac{1}{n}\sum_{i=1}^n x_i$ be the sample of the xs, and $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i$ be the sample mean of the ys, we have that for the intercept,

$$\begin{aligned}
&\frac{\partial \text{MSE}_n}{\partial b_0} = 2(\bar{y} - (b_0 + b_1\bar{x})) \\
&\implies \bar{y} - (\hat{\beta}_0 + \hat{\beta}_1\bar{x}) = 0 \\
&\iff \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}.
\end{aligned}$$

  For the slope,

$$\begin{aligned}
&\frac{\partial \text{MSE}_n}{\partial b_1} = \frac{2}{n}\sum_{i=1}^n (y_i - (b_0 + b_1 x_i)))x_i \\
&\implies \sum_{i=1}^n (y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}))x_i = 0 \\
&\iff \sum_{i=1}^n (y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}))(x_i - \bar{x}) = 0 \\
&\iff \hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}$$

  Note that since MSE is quadratic, we do not have to check a second derivative condition. Also, we are implicitly assuming that $\sum_{i=1}^n (x_i - \bar{x})^2 > 0$. If this is not so then the least squares estimates for intercept and slope are not well-defined.

- Why do we choose MSE as our loss function? There are several answers.
  - It is easy to find a closed-form solution for the least-squares parameters, which is not true for most other loss functions. This was (naturally) very appealing in the days before large computers existed.
  - Those of you who have taken Stats 200, or an equivalent, will be familiar with an alternative way of getting estimates for parameters from data: the method of *maximum likelihood*. Later, we will see that least squares is exactly maximum likelihood when the errors are Normal.
  - It turns out the conditional means are intimately related to squared error. Specifically, the conditional mean gives the best possible prediction for Y from X, in the sense of minimizing the population MSE. That is,

$$m(x) = \text{argmin}\, \mathbb{E}[(Y - m)^2 | X = x].$$

    Of course, we do not know the population MSE. So we approximate the best possible prediction by minimizing MSE over the samples, i.e. we do least squares.
  - Next class, we will prove that least-squares is unbiased for the true parameters. Later, we will see that the least-squares is in a sense the best possible among all unbiased estimators.

5 Correlation coefficient

- After having run a simple linear regression, naturally you will want to know: how "well" do the xs predict the ys? Or the related question: how strong is the association between x and y? A common way to answer this question is to report the **correlation coefficient**. There are more statistically principled and intrinsically meaningful measures, but because correlation is so often seen in the wild it is important that you know what it is. Later we will learn about some of these better measures.
- Notice that it is not immediately obvious how to extract the information we are looking for from the least-squares estimates. For example, if we take every value of y, multiply it by 2, and recomputed the

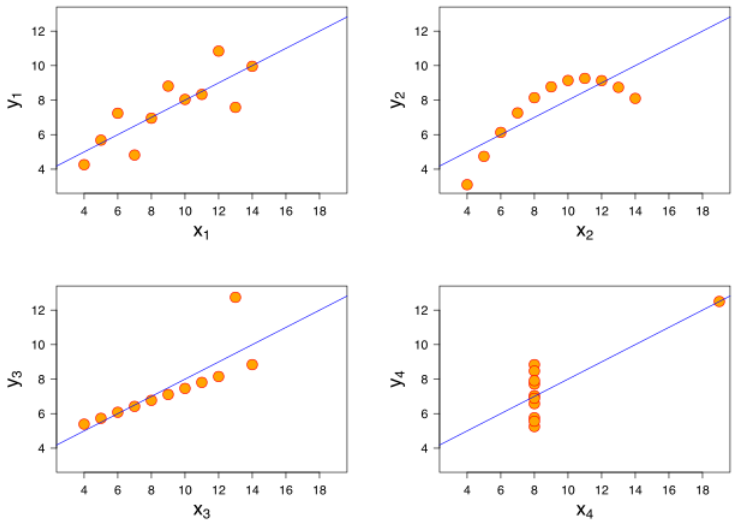

Figure 2: Anscombe's quartet, taken from Wikipedia.

line of best fit, the new estimate for slope (and intercept) will be twice as large as the old estimate, even though the underlying association between predictor and response clearly hasn't changed.

One way to resolve this is to take the slope and make it a "unitless" quantity. This is done through multiplying by the standard deviation of the xs, $s_x := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, and dividing by the standard deviation of the ys:

$$r := \hat{\beta}_1 \cdot \frac{s_x}{s_y} = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_x} \cdot \frac{(y_i - \bar{y})}{s_y}.$$

The quantity r is called the **sample correlation**. It satisfies the following properties:

- Correlation is always between -1 and 1.
- Correlation is a unitless quantity: multiplying or dividing all of the xs or ys by a positive number does not change the correlation.
- The sign of the correlation determines the direction of the linear relationship between x and y: if $r > 0$ then x and y have a positive linear association, meaning that if $x_2 > x_1$ then we would expect $y_2 > y_1$, while if $r < 0$ then x and y have a negative linear association.
- The square of the sample correlation r^2 determines the strength of the linear association: $r^2 = 0$ if we observe no linear relationship between x and y, and $r^2 = 1$ if we observe a perfect linear relationship between x and y, i.e. the line of best fit passes through every observed data point.

- People often divide correlation into a few qualitative categories. For example, an r^2 of at least .5 represents a "strong association", while an r^2 of at least .1 is a "weak association" that could easily be due to pure chance. This is good enough for government work, but later, we will learn how to use correlation, and correlation-like quantities, to make much more rigorous statements. This is what we mean by doing inference.

6 Additional remarks

- It is crucial to plot the data before performing a regression analysis. This is memorably demonstrated by an example of four data sets known as **Anscombe's quartet** (after the statistician Francis Anscombe), and plotted in Figure 2. The line of best fit is the same for all four data sets, but the data are completely

different. In only one out of the four cases is it clear that the assumptions of the linear model are plausible.

- There is a somewhat legendary dispute about who actually invented linear regression, between Gauss and Legendre, another preeminent 18th century mathematician. For an entertaining account of the dispute, see the article "Gauss and the Invention of Least Squares", by Steve Stigler.
- Conditional expectations, and hence the conditional mean $m(x) = \mathbb{E}[Y|X = x]$, are somewhat tricky to define properly. One way to heuristically understand $m(x)$ is by thinking about the "graph of averages". To do this, imagine you received a very large (read, infinite) set of (X, Y) pairs, each independently sampled from some distribution. Then, for many different x values, imagine taking a very small (read, infinitesimal) window around x, finding the mean of all the values of Y for which the corresponding X fell into this window, and plotting the xs against these "local" means. As the number of samples grows very large, and the size of the window gets very small, this "graph of averages" will tend towards the conditional mean.

  For an example that actually constructs the graph of averages using some data, see Freedman 2.1.
- Freedman suggests the following way of remembering the least-squares line, which follows from rewriting the equation of a line in point-slope form:
  - The least-squares passes through $(\bar{x}, \bar{y})$. When x is average, expect y to be average.
  - For all numbers k, the least-squares line passes through $(\bar{x} + ks_x, \bar{y} + krs_y)$. When x is k SDs above (or below) the mean of the xs, expect y to be kr SDs above (or below) the mean of the ys.

  We see an interesting phenomenon: unless the data are perfectly correlated ($r^2 = 1$) the least-squares line always produces a fitted value $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ that, after dividing by standard deviation, is closer to the mean of the ys than x is to the mean of the xs. This is an example of *regression to the mean*, which is where the term "linear regression" comes from.