Regression as a Toolkit: The Descriptive-Predictive-Causal Framework

Ashley I Naimi

September 2026

Contents

1 Introduction

Over the first three weeks of this course, we covered target trial emulation, and how to build a protocol defining who belongs in an analytic dataset and when the follow-up clock starts (Week 1), what type of information we need to collect and why (Week 2), and we covered some sampling designs we can use under certain circumstances for more efficient study designs (Week 3). This week we turn to the focus for the rest of the semester: regression.

The thesis of this lecture can be stated in one sentence. Regression is a single machine with distinct uses including description, prediction, and causal inference. And, the use is decided by BOTH the research question AND the assumptions we attach to the machinery in order to answer that question. These assumptions come, mostly, from outside the machinery of regression itself. What changes across the three uses is the set of assumptions we must add, from outside the model, for the output to answer our question. This point is essential to understand how to use regression in practice.

Recall the note box from Week 2: an estimand is defined by the scientific question, and then paired with an estimator. This week we cover the taxonomy of question types that generate estimands in the first place, and discuss an important framework for understanding how regression fits into this taxonomy. We also open the machine's hood for the first time this week, walking through how least squares and maximum likelihood actually fit a model to data. Seeing the mechanics makes the central point plain: the research question appears nowhere in them.

What to read: Carlin & Moreno-Betancur. On the Uses and Abuses of Regression Models. *Stat Med* 2025;44:e10244
Greenland. Some Ways to Make Regression Modeling More Helpful Than Misleading. *Stat Med* 2025;44:e10313

This framework is a central organizing principle for the entire course. When we study the "anatomy" of a regression model, or variance estimation, or flexible modeling, or survival models, the first question will always be the same: what type of question is this analysis meant to answer?

2 Three Questions, One Machine

Research questions that regression analysis serves can be generally classified into one of three types (Carlin and Moreno-Betancur, 2025; Hernán, 2019):

1) Descriptive questions characterize the distribution of a random variable (health state or outcome or some other construct) in a population: a preva-

lence, a burden, a difference in means or some other function of the distribution, between subgroups.

2) Predictive questions seek to build a model or algorithm that forecasts an individual's outcome from measured characteristics.

3) Causal questions ask whether, and by how much, outcomes in a population would differ under a different (counter-to-fact) exposure or intervention.

There are other uses of regression, but these three capture most of what we typically do. Sometimes, whether an analysis falls into one classification or another is not always clear, and this is itself informative. If the type of research question is not immediately clear, this can serve as a signal that you need to do more work to bring the question into sharper focus (Carlin and Moreno-Betancur, 2025).

A useful operational test is to ask what will be done with the results: who will use them, and what actions will be taken? Broadly speaking, an analysis whose results would inform resource allocation, or one that is just meant to see what the distributions of things in a population actually is, is answering a descriptive question. An analysis whose output will be used to flag high- or low-risk patients is answering a predictive one. An analysis meant to tell us whether a treatment policy should change is answering a causal one.

To make the trichotomy concrete, consider the three examples from Carlin and Moreno-Betancur (2025).[1]

[1] Note that the first example will be revisited throughout this lecture. Carlin and Moreno-Betancur provide a `data/kidneys.csv` file along with their paper, which we include in the course repository.

The first example is descriptive. Young children who acquire a urinary tract infection sometimes develop pyelonephritis, a kidney infection that enlarges the kidney and complicates the use of ultrasound measurement for growth assessment. In a consecutive series of 180 children (350 measured kidneys, 99 with a scintigraphic defect indicating infection), the question was how much larger infected kidneys are than uninfected ones. The published analysis fit a linear regression of kidney length on infection status and a cubic function of age:

$$E(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2 + \beta_4 X_2^3$$

where Y is kidney length in mm, X_1 indicates infection, and X_2 is age in months. We can reproduce this analysis directly:

```
kidneys <- read_csv(here("data", "kidneys.csv"),
    skip = 1) |>
    mutate(infected = as.numeric(infected ==
        "yes"))

kidney_model <- lm(length ~ infected + age +
    I(age^2) + I(age^3), data = kidneys)

round(summary(kidney_model)$coefficients["infected",
    1:2], 2)
```

```
##   Estimate Std. Error
##       4.13       0.71
```

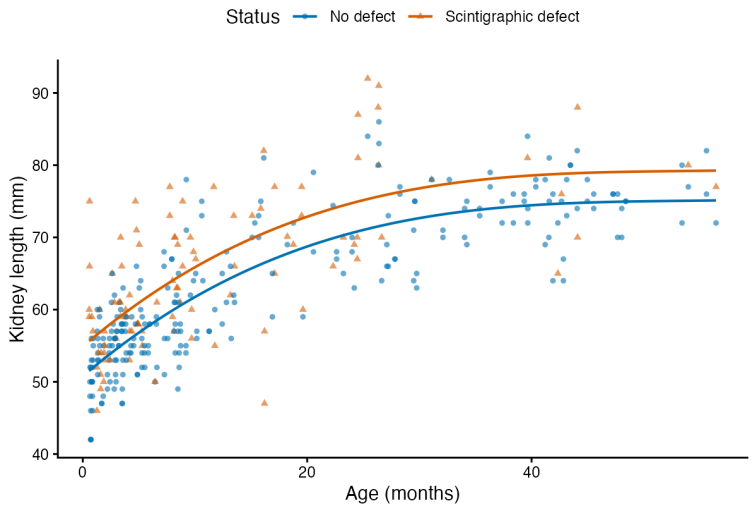


Figure 1: Kidney length against age in 350 kidneys from 180 young children, with fitted values from a linear regression of length on infection status and a cubic function of age. The model constrains the two fitted curves to be parallel: the estimated difference of 4.1 mm applies at every age. Data from Carlin and Moreno-Betancur (2025), Example 1.

The model estimates that infected kidneys are about 4.1 mm longer than uninfected kidneys, **at every age.** The figure above shows the importance of the statement "at every age". The cubic age curves for the two groups are exactly parallel. Why? Because the model contains no product term between infection and age, and thus encodes within it a "parallel trend" constraint. That constant difference is an assumption/constraint the model imposes, and has little to nothing to do with the data supplied.[2] We return to this point again later, when we discuss what regression contributes to descriptive questions;

[2] We'll see more of why this is the case when we discuss the right hand side of a regression model in the next lecture.

the more general question of how much flexibility a model should be given returns in Week 9, when we cover flexible regression.

In the second example, researchers were interested in predicting which children might not respond to a gas enema treatment for intussusception. Intussusception is an acute bowel obstruction in infants, and is usually treated first with a gas enema, which sometimes fails. In 282 consecutive cases, the question was how well success could be forecast from a child's presenting characteristics. The published analysis fit a logistic regression, discarded predictors by backwards selection on p-values, and reported the surviving coefficients with confidence intervals, but no assessment of predictive performance and no validation (each of those choices is problematic when it comes to using regression for prediction, and we discuss why later).

The third example is causal. In a cross-sectional study of 8,585 Tasmanian schoolchildren, the aim was to identify "risk factors" for childhood asthma. The analysis used forward selection and reported the surviving odds ratios as "independent risk factors." The phrase "risk factor" is, as usual, doing some potentially hidden work: it blends prediction and causation into a single ill-defined notion, and the presentation invites readers to interpret every coefficient in the model as an effect (the Table 2 fallacy we met in Week 2 (Westreich and Greenland, 2013)).

Note that for these three examples, nothing really essential distinguishes them at the level of software / regression machinery: all three are calls to a fitting function and optimization algorithm. All three return a table of coefficients. What distinguishes them is the type of question that they answer. We will now see exactly what that fitting function is doing.

3 One Job: Fit a Family to Data

Why can't the machinery itself tell these three uses apart? Because the machinery, considered on its own, only ever does one job. When we fit a regression model, we hand the fitting routine two things: a dataset, and a candidate function for the conditional mean of the outcome. More precisely, we hand it a *family* of candidate functions, indexed by parameters: the form of the regression model we choose (for example, all straight lines if we specify a linear model, or something more complicated like the curvilinear family in the kid-

ney example above). The routine then returns the member of the family that sits closest to the data, with closeness measured by a criterion such as least squares, maximum likelihood, or some other optimization function.

Let's walk through exactly how this works.

4 How the Machine Fits I: Least Squares

Consider a scenario in which we are interested in regressing an outcome Y against a set of covariates X. These covariates can be an exposure combined with a set of confounders needed for identification, a set of predictors used to create a prediction algorithm via regression, or an indicator of subgroups whose outcomes we want to summarize. In its most basic formulation, a regression model can be written as:

$$E(Y \mid X) = f(X)$$

In principle, this model is most flexible in that it states that the conditional mean of Y is simply an *arbitrary function* of the covariates X.[3] We are not stating (or assuming) precisely **how** the conditional mean is related to these covariates. Using this model, we might generate predictions to facilitate a decision making process, or obtain a contrast of expected means between two groups.

[3] The conditional mean is one choice of target. Conditional quantiles, hazards, and entire conditional distributions are others, and we will meet regressions for them in later weeks. However, the reasoning in this section carries over in general.

Because of the flexibility of this model, we may be interested in fitting it to a given dataset. But we can't. There is simply not enough information in this equation for us to quantify $f(X)$, even if we had all the data we could use. In addition to data, we need some "traction" or "leverage" to be able to quantify the function of interest.

The earliest attempt to find some "traction" to quantify $f(X)$ was proposed in the 1800s (Stigler, 1981; Shalizi, 2019). The idea starts from a simple requirement: fitted values should sit close to observed outcomes. Squaring the differences between the two keeps errors on either side of the fitted values from canceling each other out, and averaging the squared differences gives a single criterion to make small, the mean squared error: $E[(Y - f(X))^2]$.

In machine learning, squared error is known as the **L2 loss**, and a criterion like $E[(Y - f(X))^2]$ is called a *loss function*. Fitting a model by minimizing average loss over a dataset is the core of supervised learning; least squares is its oldest special case.

Thus, we can define the "optimal" $f(X)$ as the function of X that minimizes the mean squared error. Recall from calculus (see section 2.2.2 of the math foundations lecture) that finding the $f(X)$ that minimizes mean squared error

can be achieved by taking the derivative of the mean squared error with respect to $f(X)$, setting it to zero, then solving for $f(X)$. Carrying that minimization out over *all* possible functions gives a famous answer: the function that minimizes mean squared error is the conditional mean itself, $f(X) = E(Y \mid X)$. The conditional expectation is therefore not an arbitrary choice of regression target. It is the mean squared error optimal summary of Y given X, which is why the machinery of regression is built around it.

To see why, note that for any candidate f, conditional on X we can write

$$E[(Y-f(X))^2 \mid X] = [E(Y \mid X)-f(X)]^2+Var(Y \mid X).$$

The second term does not involve f at all; it is the irreducible noise in Y among individuals with the same covariate values. The first term is zero exactly when $f(X) = E(Y \mid X)$.

We've made some progress, but a problem remains: the solution names the target, it does not hand us a formula. $E(Y \mid X)$ is exactly the unknown object we are trying to estimate. And with continuous covariates, no dataset of any size contains enough information to pin down an arbitrary function; there are infinitely many curves consistent with any finite scatter of points. To move forward, we need to constrain the search.

Early on, it was recognized that if we select $f(X)$ to be *linear* (more technically, *affine*) the problem of finding the optimal $f(X)$ becomes much easier. That is, if we can simplify $f(X) = b_0 + b_1X$, then we can use calculus and simple algebra to find an optimal *linear* solution set $b_0 = \beta_0, b_1 = \beta_1$ that minimizes MSE. Note what happened in that step: we restricted f to a family and reduced the problem to estimating the finite set of parameters that indexes it. Every such constraint is an assumption the data did not supply. Greenland notes that "parametric models add more assumptions in the form of constraints imposed by the chosen model," and those constraints should come from background knowledge about the setting rather than from software defaults or habit (Greenland, 2025). Unfortunately, there is often insufficient background knowledge to pick all of the constraints we need in a principled way.

Specifically, we can re-write the mean squared error as a function of $b_0 + b_1X$ to be $MSE(b_0, b_1) = E[(Y - (b_0 + b_1X))^2]$. Taking the partial derivatives of $MSE(b_0, b_1)$ with respect to b_0 and b_1 gives us the ordinary least squares estimator for the coefficients in the model (Rencher, 2000; Shalizi, 2019):

$$\hat{b}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

To make this concrete, we can apply these estimators to real data. We will use the NHEFS data, a cohort study of smoking and weight change used throughout Hernán and Robins (2020), which is included in the course repository. Suppose we want to summarize how a respondent's weight in 1982 (`wt82`, in kg) relates to their weight measured eleven years earlier (`wt71`):

```
nhefs <- read_csv(here("data", "nhefs.csv")) %>%
  select(wt82, wt71) %>%
  na.omit()

nrow(nhefs)
```

```
## [1] 1566
```

The closed form estimator asks for nothing more than means and cross-products, so for the 1,566 respondents with weight recorded at both visits we can compute $\hat{b}_1$ and $\hat{b}_0$ directly, with basic arithmetic:

```
x <- nhefs$wt71
y <- nhefs$wt82

b1 <- sum((x - mean(x)) * (y - mean(y))) /
  sum((x - mean(x))^2)
b0 <- mean(y) - b1 * mean(x)

round(c(b0 = b0, b1 = b1), 3)
```

```
##    b0    b1
## 8.007 0.924
```

Of course, in practice no one computes these quantities by hand. R's `lm()` function implements the same least squares machinery, wrapped in an interface that handles model formulas, factors, and much more. Handing it the same data:

```
mod <- lm(wt82 ~ wt71, data = nhefs)

round(coef(mod), 3)
```

```
## (Intercept)        wt71
##       8.007       0.924
```

The results are identical. There is nothing inside `lm()` beyond the optimization we just carried out by hand: it minimizes squared error over the linear family and reports the coefficients of the winning member. These closed form solutions are a luxury of the linear family paired with squared error loss (see the technical note below). The figure below shows the fit at work in these data: the fitted line minimizes the sum of squared residuals, and the mean squared error, viewed as a function of the candidate slope, bottoms out exactly at the estimate we just computed twice.

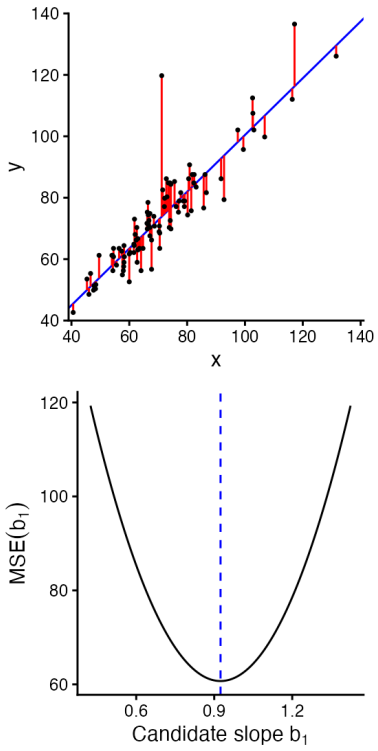


Figure 2: Top: line of best fit (blue) minimizing the sum of squared residuals (red segments) in a sample of 100 observations from the NHEFS data. Bottom: mean squared error for each candidate slope, with the intercept set to its optimal value for that slope. The minimum (dashed line) sits at the least squares estimate that defines the fitted line in the top panel.

Technical Note: Nonlinear Models

For many other families the same calculus applies, but the resulting equations have no algebraic solution. In logistic regression, for example, the analogous equations (the score equations we meet in the next section) can be solved by hand only in special cases. Software instead optimizes iteratively, using algorithms such as Newton-Raphson or iteratively reweighted least squares, which start from a guess and improve it until the criterion stops changing. This is what `glm()` does under the hood every time you call it, and we will see more of this machinery when we cover generalized linear models in Week 6.

5 How the Machine Fits II: Maximum Likelihood

5.1 Basic Illustration

Maximum likelihood estimation is another more general technique that we can use to fit a model $f(X)$ to data. To illustrate, let's start with a simpler example where we want to estimate the 10 day risk of diarrhea ($Y \in [0, 1]$) among 30 infants infected with *Vibrio cholerae* who are also being breastfed. The scientific question of interest is the role that antibiotic concentrations in breastmilk ($X \in [0 = \text{low}, 1 = \text{high}]$) can play in reducing the incidence of diarrheal disease. The data are in the table below (taken from Cole et al., 2013):

Antibiotic Level	Cases ($Y = 1$)	non-Cases ($Y = 0$)	Total
Low ($X = 0$)	12	2	14
High ($X = 1$)	7	9	16
Total	19	11	30

If we assume that diarrhea was an independent occurrence across the 30 infants, we can define its probability mass function as:

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y} \tag{1}$$

In a direct problem, the known portion p is held fixed, and when we change the data inputs we obtain different probabilities of seeing y cases of diarrhea out of n observations.

In this case, p is the probability of observing a single diarrheal event. For the moment, let's ignore the fact that we have exposed and unexposed infants in the cohort that this equation is meant to represent. We are working in an indirect problem setting (we have the data, we don't know the underlying proba-

bility p). This will enable us to re-write equation 1 as:

$$P(Y = y) = \binom{30}{19} p^{19}(1-p)^{30-19}$$

which still doesn't help us, because all we have is data. However, if we guessed that $p = 0.6$, we could then get a predicted outcome:

$$P(Y = y; p = 0.6) = 0.14 = \binom{30}{19} 0.6^{19}(1-0.6)^{30-19}$$

This predicted value of 0.14 is no longer a probability, because the variable in the original equation is p, and not (as would usually be the case) the data (i.e., the *variables*). Instead, we call this the **likelihood of the parameter** p. We can try another guess at the parameter, say $p = 0.4$:

$$P(Y = y; p = 0.4) = 0.005 = \binom{30}{19} 0.4^{19}(1-0.4)^{30-19}$$

Here, the likelihood of this parameter is much lower than the first. What does this suggest?: if the equation governing the data generating mechanism is correct, then the data are **more compatible** with $p = 0.6$ than they are with $p = 0.4$. Thus, it is more likely that the true p is closer to 0.6 than it is to 0.4, given the data we have.

Let's look at this likelihood function a bit more systematically. R's `dbinom` function computes exactly the equation above, so we can evaluate the likelihood over a whole grid of candidate parameter values in two lines, holding the data fixed at 19 cases out of 30 and letting the parameter vary:

```
p_grid <- seq(0, 1, 0.01)

likelihood <- dbinom(19, size = 30, prob = p_grid)
```

This likelihood function tells us that the most likely parameter value is 0.633.

The key takeaway is that a distribution function and a likelihood function usually look identical; what differs is which argument is held fixed. In a direct problem we hold the parameters fixed and compute probabilities for data, $f(y; \theta)$. In maximum likelihood estimation we hold the data fixed and search for the parameters with the highest likelihood, $L(\theta; y)$.

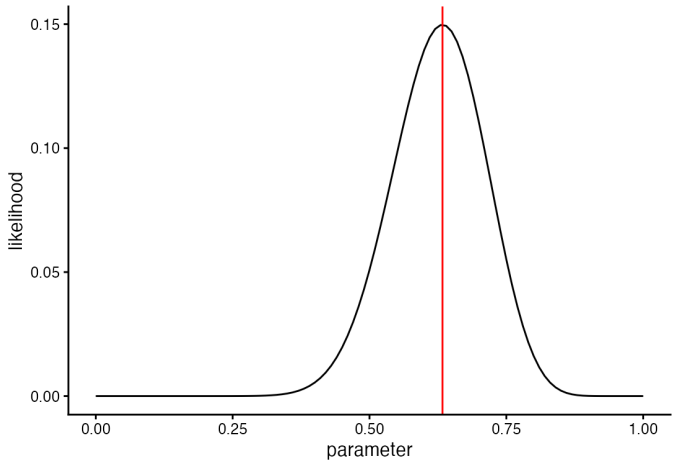


Figure 3: The likelihood of each candidate value of the parameter p, given 19 cases among 30 infants. The red line marks the maximum at 19/30, which is the maximum likelihood estimate.

5.2 MLE for Regression

How might this work if we are interested in understanding the role that antibiotic concentrations in breastmilk play in the incidence of diarrhea?

Like in our OLS example, if we assume independence across observations, we can define the probability mass function for observing y cases of diarrhea among those with $X = x$ as:

$$P(Y = y \mid X = x) = \binom{n_x}{y_x} p_x^{y_x} (1 - p_x)^{n - y_x}$$

where (again) $X \in [0, 1]$. However, this time we can define $p_x = \text{expit}[\beta_0 + \beta_1 x]$ where $\text{expit}(a) = \frac{1}{[1+\exp(-a)]}$. Now that we've made p_x a function of β_0 and β_1, we can define the likelihood function here as:

$$L(\beta_0, \beta_1; y_x) = \prod_{x=0,1} \binom{n_x}{y_x} p_x^{y_x} (1 - p_x)^{n_x - y_x}$$

Once again, the task is to find values of β_0 and β_1 that maximize this likelihood function, and that are thus most compatible with the data. We could plug different values in and evaluate the likelihood as we did above. However, in practice, we would once again use calculus to find where the slope of the likelihood function is equal to zero.

To make the math easier, we often simplify the likelihood function by ignoring factors like $\binom{n}{y}$, since the derivative of the likelihood function will not depend on this scaling factor. Furthermore, if we take the log of the likelihood function, computing derivatives is much simpler:

When these equations have no closed form, machine learning's workhorse optimizer, **gradient descent**, does the same job numerically: repeatedly nudge the parameters in the direction that most improves the criterion.

The **cross-entropy loss** used to train machine learning classifiers, neural networks included, is exactly this binomial log-likelihood with the sign flipped. Minimizing cross-entropy and maximizing the log-likelihood are the same computation.

$$\mathcal{L}(\beta_0, \beta_1; y_x) = \ln[L(\beta_0, \beta_1; y_x)] = \sum_{x=0,1} \left[y_x \ln p_x + (n_x - y_x) \ln(1 - p_x) \right]$$

The (partial) derivatives of this log-likelihood function with respect to the parameters is often referred to as the score function, the gradient, or (less commonly) the informant. If we set the score function for each parameter to zero and solve for the parameter values, we get the score equations for β_0 and β_1, which are our maximum likelihood estimators. In this simple example, solutions for the score equations are easy to compute, and become (Cole et al., 2013):

$$\hat{\beta}_0 = \ln\left(\frac{y_0}{n_0 - y_0}\right) = \ln(12/2) = 1.79$$

$$\hat{\beta}_1 = \ln\left[\frac{y_1(n_0 - y_0)}{y_0(n_1 - y_1)}\right] = \ln\left[\frac{7 \times (14 - 12)}{12 \times (16 - 7)}\right] = -2.04$$

We can again compare these to what we would get from an actual regression analysis of these data:

```
d <- data.frame(
  y = c(1, 1, 0, 0),
  x = c(0, 1, 0, 1),
  freq = c(12, 7, 2, 9)
)

mod_glm <- glm(y ~ x, data = d, weights = freq, family = binomial("logit"))

summary(mod_glm)$coefficients
```

```
##               Estimate Std. Error   z value   Pr(>|z|)
## (Intercept)  1.791759  0.7637224  2.346087 0.01897166
## x           -2.043074  0.9150083 -2.232847 0.02555901
```

The estimates match the closed form solutions. And the coefficient for x answers the scientific question we started with: the log odds ratio is -2.04, so the odds of diarrhea under high antibiotic concentrations are $\exp(-2.04) \approx 0.13$ times the odds under low concentrations, roughly an eight-fold reduction.

Now, step back and notice what appeared nowhere in either derivation: the question. The least squares solution is a property of the data's means, variances, and cross-products; the score equations are properties of counts. At no point did the machinery ask whether X was an exposure whose effect we want, a confounder we must adjust for, or a predictor we hope will forecast. Notice, too, what both procedures were estimating. The least squares target was the conditional mean $f(X) = E(Y \mid X)$, and the likelihood target p_x is also a conditional mean, since for a binary outcome $P(Y = 1 \mid X = x) = E(Y \mid X = x)$.[4] However the machine is driven, it estimates conditional means within the family we chose.

[4] The two criteria are not as different as they look, either: if we assume Y given X is normally distributed with constant variance, maximizing the likelihood yields exactly the least squares solution. Distributions and their paired link functions are a large part of next week's topic.

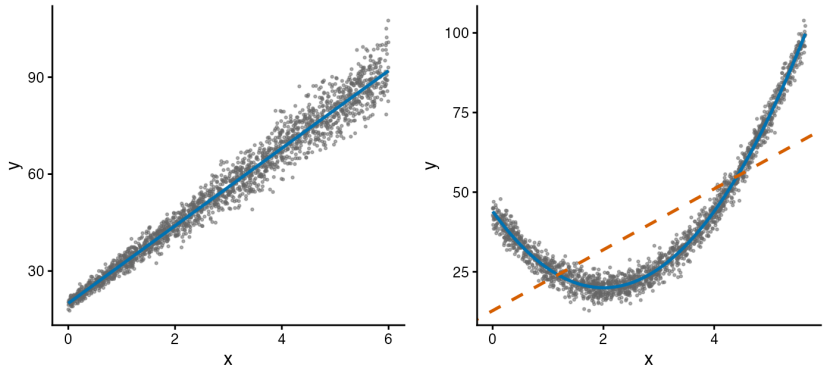
6 Regression Machinery Is Always Descriptive

A summary of conditional means is a description. It describes patterns in the data that the model form permits it to capture. Greenland's framing is useful here: a parametric regression acts like a filter, so that "linear regression extracts and summarizes linear components of data patterns and filters out the rest as residuals" (Greenland, 2025).

A fitted model is in this sense a "lossy compression" of the dataset. Whether that compression is adequate or insufficient depends entirely on what we want to use it for.

Note: Model Projections

What happens when we fit a linear model to a relation that is not linear? The fit does not fail, but rather returns a "projection": the member of the model family that sits closest to the true conditional mean, with closeness measured by (e.g.) mean squared error and averaged over the distribution of the covariates. This idea can be made exact. Without assuming any model at all, the population least squares coefficient is a well defined feature of the joint distribution, $\beta(P) = \text{argmin}_b \, E\left[(Y - b'\mathrm{X})^2\right]$, the best linear approximation to $E(Y \mid \mathrm{X})$ (Buja et al., 2019). When the truth happens to be linear, $\beta(P)$ is the usual model parameter. When it is not, $\beta(P)$ still exists, and it is what the fitted model estimates. A linear coefficient for a curvilinear relation is therefore not meaningless or useless, but it is of scientific interest only if this projection is the summary we actually want (see, e.g., Hubbard et al., 2010).



Left: a linear conditional mean, where the model form and the mean function coincide. Right: a curvilinear conditional mean (solid) and its projection onto the linear family (dashed).

The figure above shows this. In the left panel the conditional mean really is linear: a one unit difference in x corresponds to the same difference in the mean of y everywhere, and the mean function and the true model form coincide. In the right panel the mean is curvilinear, and the dashed line is its projection onto the linear family. The projected slope is a **weighted average** of the local slopes of the curve, with weights implicitly determined by where the covariate values fall (i.e., there's no actual weighting happening explicitly, it's just the covariate distributions shaping how the line of best fit aligns with the data). Importantly, under misspecification (i.e., linear specification of a curvilinear relation), the coefficient depends on the covariate distribution. Two studies of the same underlying relation, fitting the same model in populations whose covariate values sit in different ranges, can arrive at genuinely different population slopes (Buja et al., 2019). Therefore, some of what looks like inconsistency across studies is just the same curve projected from different vantage points (different covariate distributions).

Two consequences follow. The first is what Carlin and Moreno-Betancur call the *true model myth*: the idea that the analyst's central task is to find

the model that best approximates the true data-generating process, after which conclusions of any kind can be read off the coefficients. The myth is embedded in how regression is often taught. A widely used logistic regression text states that "the goal of an analysis using this model is the same as that of any other regression model used in statistics, that is, to find the best fitting and most parsimonious, clinically interpretable model to describe the relationship between an outcome … and a set of independent (predictor or explanatory) variables" (Hosmer et al., 2013), and then moves directly to interpreting the fitted coefficients. As Carlin and Moreno-Betancur (2025) observe, "the logic here seems entirely backwards … unless the model is believed to provide a full and faithful representation of the true data generating process, then its coefficients may be completely uninterpretable." No regression model of health outcomes in human populations is such a representation.

The second consequence concerns whose distribution the fitted summary describes. Greenland clarifies that "conventional regression analysis provides inferences about the data generator, not the assumed target population" (Greenland, 2025). By "data generator", he means here the target population we care about, plus everything Weeks 1 through 3 (and hopefully, all your prior training) taught us to worry about: who was sampled, who was measured and how well were they measured, who was censored, who was selected in or out, how much confounding is there, and a host of other questions. For causal questions, this includes whatever threatens identification.

Standard software output quietly treats the data as a perfect random draw from the target population. The gap between generator and target that is caused by all of these potential "problems" is not something a regression model can see; it is closed, or not, by design knowledge and by assumptions we state and defend.

As an example of this issue, suppose the kidney data was obtained from an archive that had a selection bias embedded within it: scans showing a defect were retained preferentially when the kidney looked enlarged, while scans of unaffected kidneys were kept as a matter of routine. We can impose that history on the data and refit the published model:

```
set.seed(63)
p_keep <- ifelse(kidneys$infected == 1,
```

```
                  plogis(-6 + 0.12*kidneys$length), 1)
kept <- kidneys[rbinom(nrow(kidneys), 1, p_keep) == 1, ]

mod_selected <- lm(length ~ infected + age + I(age^2) + I(age^3),
                   data = kept)
round(summary(mod_selected)$coefficients["infected", 1:2], 2)
```

```
##   Estimate Std. Error
##       4.94       0.75
```

The selection removed 17 of 350 kidneys, all from the defect group and mostly the shorter ones, and the estimated difference moved from 4.1 mm to 4.9 mm, a shift larger than the original standard error. The model specification is identical and nothing in the output signals a problem. The regression is doing its job, describing the generator, and the generator now includes a bias embedded into it, possibly an archivist who thought that what they were doing was helping clinicians identify the problem cases.

Note: The Take-Home Statement

At the level of the fitted object, every regression output is a summary of conditional distributions produced by the data-generating mechanism. Calling that summary descriptive, predictive, or causal is a claim about assumptions that reside outside the model (i.e., that have nothing to do with regression): that sampling and measurement connect the generator to the target population (description), that the joint distribution transports to the setting where forecasts will be deployed (prediction), or that identification conditions license a counterfactual contrast (causation).

None of this is new, but these issues tend not to be presented very clearly in introductory courses. Box's aphorism that "all models are wrong, but some are useful" is a generically used but specifically misunderstood saying that dates to his 1976 reflections on how Fisher kept mathematics tethered to scientific questions (Box, 1976). In doing so, Box argued that a correct model can never be obtained by excessive elaboration, and that since all models are wrong, "the scientist must be alert to what is importantly wrong." (Box, 1976). Leo Breiman's famous "two cultures" paper diagnosed a habit in statistics of treating a fitted parametric regression model as if it were the mechanism itself (Breiman, 2001), and proposed an alternative via machine learning. Galit

Shmueli formalized the gulf between explaining and predicting, noting the different considerations that must be present in each domain (Shmueli, 2010). However, most of these problems boil down to one issue. Stijn Vansteelandt notes that "many traditional statistical analyses focus more on the model than on the problem one is trying to solve," with "great danger in drawing false conclusions when viewing the fitted model as a representation of the ground truth" (Vansteelandt and Steen, 2025).

One goal of this course is to convince you that you need to let the question, and not the machinery of regression, lead you.

7 Three Roadmaps for One Machine

There's an important set of steps we need to take when using the machinery of regression. This is true whether the questions we have are descriptive, predictive, or causal (though the specifics differ).

First, we define the question we have as a target parameter, an estimand, stated without reference to any regression model.

Second, ask whether this estimand is identifiable in principle from the study data, given how the data were sampled and measured.

Third, and only then, write the estimation plan, which is where regression may enter.

Often, the estimand is conflated with the model (for example, the coefficient for an exposure variable in a linear regression model). The corrective is that we define our estimands based on the research questions we have (whether these are descriptive, predictive, or causal) and then pick a modeling approach that allows us to estimate this target in the "best" way possible.

Let's consider this procedure for each type of question:

7.1 The descriptive roadmap

A well-specified descriptive question names a target population anchored in person, place, and time, an outcome, and a measure of occurrence (Lesko et al., 2022). Consider estimating the proportion of a country's population with SARS-CoV-2 antibodies in 2021. The estimand is a population prevalence, and identification considerations still matter here (can our data/methods tell us what we want to know?): no random sample of a national population was

feasible, participation was voluntary, antibody tests are subject to important errors, and all these affect whether we can actually answer the question. Whether the observed proportion can be connected to the target prevalence depends on how sampling and measurement are handled.

Regression can enter into descriptive analyses in several ways. It can compute subgroup summaries with pooled inference. It can fit curves as parsimonious summaries of how an outcome varies with a continuous covariate. It can smooth for precision (as in the kidney example, by assuming a constant infected-uninfected difference and letting the cubic in age absorb the outcome variation due to growth). It can be used to adjust a nonrepresentative sample toward the population, by reweighting or standardizing over covariates whose population distribution is known.

The smoothing role is worth checking in the data rather than taking on faith. The crude comparison is itself a regression, the two-group model with a single indicator:

```
mod_crude <- lm(length ~ infected, data = kidneys)
round(summary(mod_crude)$coefficients["infected",
    1:2], 2)
```

```
##   Estimate Std. Error
##       2.56       1.20
```

The crude difference is 2.6 mm with a standard error of 1.2, against the adjusted 4.1 mm with a standard error of 0.7. Adjustment does two things: It makes the estimate more precise, cutting the standard error nearly in half (by letting the cubic in age absorb growth variation that the crude comparison leaves in the error term). It also changes the target: children with defect kidneys were younger on average than the rest of the series (12.3 versus 16.5 months), and younger children have shorter kidneys, so the crude contrast folds a growth deficit into the infection difference and understates it. The 4.1 mm figure answers the question about how much longer an infected kidney is than an uninfected kidney of the same age, which is the relevant comparison when enlargement must be judged against normal growth. Choosing between 2.6 and 4.1 is not a statistical decision, but one about which descriptive estimand the question targets.

One discipline from earlier weeks carries over to descriptive analyses: descriptive comparisons need a time zero. Vansteelandt and Steen (2025) recommend that researchers run descriptive comparisons with the habits of a trial analysis: start follow-up at a defined entry point and compare groups as defined at that point.

7.2 The predictive roadmap

Prediction is not a primary focus of this course, and we do not return to it in a dedicated week. But much of what we need to do in prediction aligns with descriptive and causal questions. Vansteelandt and Steen refer to the "predictimand," which is the estimand we might target in the context of a predictive analysis. The predictimand is a summary of the population in which predictions will be made, which outcome will be predicted, which predictors will be available to use at the moment of prediction, and exactly how we might measure the performance of the predictive algorithm (i.e., the loss function) (Vansteelandt and Steen, 2025). Note the connection to the fitting machinery we walked through earlier: the closeness criterion the routine optimizes, squared error or a likelihood, is itself a choice, and in a predictive analysis that choice belongs to the predictimand rather than to habit. For instance, a model built to flag ICU deterioration cannot use afternoon laboratory results to make a morning prediction, and a model tuned in one hospital system may not perform well (defined by the loss function) in another hospital section.

Machine learning makes this choice explicit: the analyst picks the loss before training, and the loss determines what is learned. Squared error targets the conditional mean; absolute error targets the conditional median.

One of the key practices needed in the context of a predictive analysis is the out-of-sample validation step, which does not often happen in empirical work. There are so many papers that apply (e.g.) logistic regression models to a dataset, in order to identify whether a variable or set of variables, and determine whether these are good "predictors" on the basis of a p value, model deviance statistic, or R squared. Unfortunately, none of these are good indicators of predictive performance (Pepe et al., 2004), and all can be subject to severe "overfitting" if they are not computed out-of-sample.

7.3 The causal roadmap

In causal inference, the estimand step is a contrast of potential outcomes, such as $E(Y^{a=1}) - E(Y^{a=0})$, made concrete by specifying the target trial

(Week 1). The identification step is often taken to be the triad of conditional exchangeability, positivity, and consistency (Week 1, though other identification strategies are available), interrogated against the data source the way our emulation exercises demanded (Weeks 1 and 2). When those conditions hold given a covariate set Z, the g-formula expresses the counterfactual mean as a function of observable data:

$$E\left(Y^{X=x}\right) = \sum_{z} E(Y \mid X = x, Z = z)\, P(Z = z),$$

and in an ideal randomized trial the whole expression collapses to $E\left(Y^{X=x}\right) = E(Y \mid X = x)$, which is why a trial's primary analysis can be a comparison of two means (Hernán and Robins, 2020).

Only at the third step does regression appear, as one estimation strategy among several. The conditioning approach fits an outcome model such as

$$E(Y \mid X, Z) = \beta_0 + \beta_1 X + \gamma Z,$$

and reports β_1 as the effect estimate. It is worth being exact about what this requires, because the requirements come in distinct layers. The first layer is identification, which belongs to the question and the design: exchangeability given Z, positivity, consistency. The second layer belongs to the model, that is, to the family we handed the machinery: that the effect of X is constant across the strata defined by Z, and that the confounder-outcome relationship follows the parametric form we wrote down.

Neither of these assumptions is part of the causal question, but they are added when we chose this estimator, and they can fail even if causal identification holds.

These model-layer assumptions will be taken up in the rest of the course. Marginal standardization and g-computation, inverse probability weighting, and (maybe) doubly robust methods (Week 8) relax the constant-effect and parametric constraints. The finite-sample dangers of rich adjustment, sparse-data bias among them, connect to variance estimation (Week 7) and penalization (Week 10). However, the causal question and its identification conditions are considered largely separate issues.[5]

[5] Greenland argues that causal thinking belongs even further upstream than this roadmap suggests: paraphrasing Hill, "no contextually sensible probability model can be constructed without asking what caused this data set." The data generator is a causal object, whatever the question type (Greenland, 2025).

8 Conclusion

This week installed the frame the rest of the course hangs on. Regression is one machine whose fitted output is always a description of conditional distributions in the data-generating process. We watched the machine fit, by least squares and by maximum likelihood, and saw that the research question appears nowhere in either procedure. Whether the fitted description answers a descriptive, predictive, or causal question is decided outside the model, by an estimand stated before fitting, an identification argument connecting the data to the target, and an estimation plan whose own assumptions are priced separately. The three roadmaps share their structure and differ entirely in their middle step, which is why the question type must come first. And every choice of model family embeds a tension between rigidity, which risks missing real structure, and flexibility, which risks chasing noise; this bias-variance tradeoff is taken up in earnest with flexible regression in Week 9 and penalized regression in Week 10.

Lab 3 puts the frame to work: starting from a single dataset, you will pose one question of each type, specify its estimand, and identify the distinct assumptions each requires before any model is fit. Having watched the machine fit this week, next week we study its anatomy: the left-hand side, the right-hand side, link functions, distributions, and the distinction between target and nuisance that Week 2 previewed.

References

G. E. P. Box. Science and Statistics. *JASA*, 71(356):791–99, 1976.

Leo Breiman. Statistical modeling: The two cultures. *Stat Sci*, 16(3):199–215, 2001.

Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4): 523–544, 2019.

John B Carlin and Margarita Moreno-Betancur. On the uses and abuses of regression models: A call for reform of statistical practice and teaching. *Statistics in Medicine*, 44(13-14):e10244, 2025.

Stephen R. Cole, David B. Richardson, Haitao Chu, and Ashley I. Naimi. Analysis of occupational asbestos exposure and lung cancer mortality using the g formula. *Am J Epidemiol*, 177(9):989–996, 2013.

Sander Greenland. Some ways to make regression modeling more helpful than misleading. *Statistics in Medicine*, 44(13-14):e10313, 2025.

M. A. Hernán and JM Robins. *Causal Inference*. Chapman & Hall/CRC, Boca Raton, FL, 2020.

MA Hernán. Spherical cows in a vacuum: Data analysis competitions for causal inference. *Statistical Science*, 34(1):69–71, 2019.

David W Hosmer, Stanley Lemeshow, and Rodney X Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, New York, 3rd edition, 2013.

Alan E Hubbard, Jennifer Ahern, Nancy L Fleischer, Mark J van der Laan, Sheri A Lippman, Nicholas Jewell, Tim Bruckner, and William A Satariano. To gee or not to gee: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiol*, 21(4):467–474, 2010.

Catherine R Lesko, Matthew P Fox, and Jessie K Edwards. A framework for descriptive epidemiology. *American Journal of Epidemiology*, 191(12):2063–2070, 2022.

Margaret Sullivan Pepe, Holly Janes, Gary Longton, Wendy Leisenring, and Polly Newcomb. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*, 159(9): 882–890, May 2004.

Alvin C. Rencher. *Linear Models in Statistics*. Wiley, New York, 2000.

Cosma Rohilla Shalizi. *The Truth About Linear Regression*. https://www.stat.cmu.edu/ cshalizi/TALR/, 2019.

Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

Stephen M. Stigler. Gauss and the invention of least squares. *The Annals of Statistics*, 9(3):465–474, 1981.

Stijn Vansteelandt and Johan Steen. Discussion of "on the uses and abuses of regression models: A call for reform of statistical practice and teaching". *Statistics in Medicine*, 44(13-14):e10312, 2025.

Daniel Westreich and Sander Greenland. The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, 177(4):292–298, 2013.