Review of basic probability and statistics
Bias and standard error in simple linear regression

Stats 203, Winter 2024

Lecture 2 [Last update: January 10, 2024]

Today's material begins by reviewing some definitions and basic properties of random variables and estimators. Then we apply what we've reviewed to calculate the expectation and standard error of the least-squares estimator in a simple linear regression.

1 Review of basic probability: random variables, expectation, variance, joint distributions

- We begin with some definitions that you will likely have seen before.
- A **sample space** is a set Ω of potential outcomes, and a **probability** P, which takes in events $E \subseteq \Omega$ and outputs a number between 0 and 1. A **random variable** (r.v.) $X : \Omega \to \mathbb{R}$ is a function mapping potential outcomes $\omega \in \Omega$ to numbers on the real line. The **probability distribution** of X is given by the probabilities
$$P(X \in S) = P(\{\omega : X(\omega) \in S\}),$$
for subsets $S \subseteq \mathbb{R}$. The distribution is determined by the **cumulative distribution function** (CDF) $F_X(x) = P(X \leq x)$.
- A discrete r.v. X can only take on a countable number of values, meaning there exists a countable set $S = \{x_1, x_2, \ldots\}$ for which $P(X \in S) = 1$. The **probability mass function** of a discrete r.v. is $P(X = x_j)$. The **expectation** of a discrete r.v. is
$$\mathbb{E}[X] = \sum_i x_i \cdot P(X = x_i). \tag{1}$$
The **variance** of a discrete r.v. is
$$\mathrm{Var}[X] = \sum_i (x_i - \mathbb{E}[X])^2 \cdot P(X = x_i). \tag{2}$$
- *Example: flipping a fair coin.* Consider the sample space with potential outcomes $\Omega = \{H, T\}$ and probability $P(H) = \frac{1}{2}, P(T) = \frac{1}{2}$. Define the r.v. X according to $X(H) = 1, X(T) = -1$. X has probability mass function $P(X = 1) = P(X = -1) = \frac{1}{2}$. The expectation and variance of X are
$$\mathbb{E}[X] = \frac{1}{2} - \frac{1}{2} = 0, \quad \mathrm{Var}[X] = \frac{1}{2}(-1 - 0)^2 + \frac{1}{2}(1 - 0)^2 = 1.$$
Obviously, this is a model of a game where a single fair coin is flipped.
- A **continuous r.v.** X assigns probability 0 to any single number. So the probability mass function of X is uninformative; instead, its distribution is described by the **density function** p_X, which satisfies
$$P(X \in S) = \int_S p_X(x)\, dx. \tag{3}$$

The expectation and variance of a continuous r.v. X with density p_X are

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p_X(x)\,dx, \quad \mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}X)^2 p_X(x)\,dx.$$

- A function $f(X)$ of a random variable is also a random variable, which (if continuous) has expectation and variance

$$\mathbb{E}[f(X)] = \int f(x) \cdot p_X(x)\,dx, \quad \mathrm{Var}[f(X)] = \int (f(x) - \mathbb{E}[f(X)])^2 p_X(x)\,dx.$$

- Two common continuous are the uniform and Normal distributions.
  - A continuous r.v. is **uniformly distributed between** 0 **and** 1, $U \sim \mathrm{Unif}(0,1)$, if it has probability density
  $$p_U(u) = \begin{cases} 1, & \text{if } 0 \le u \le 1 \\ 0, & \text{otherwise.} \end{cases}$$
  - A continuous r.v. Z is **Normally distributed** with mean μ and variance σ^2, $Z \sim N(\mu, \sigma^2)$, if it has probability density
  $$p_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(z-\mu)^2 \Big).$$

- So far we have considered one random variable at a time. Regression studies the distribution of multiple random variables jointly. A pair of continuous r.v.s X, Y have **joint density** $p_{X,Y}$ if for all $u, v \in \mathbb{R}$:

$$P(X \le u, Y \le v) = \int_{-\infty}^{u} \int_{-\infty}^{v} p_{X,Y}(x,y)\,dx\,dy$$

  If X, Y have joint density $p_{X,Y}$, the **marginal density** of X is $p_X(x) = \int p_{X,Y}(x,y)\,dy$, and the **marginal density** of Y is $p_Y(y) = \int p_{X,Y}(x,y)\,dx$. The **conditional density** of Y given $X = x$ is $p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$. The **conditional expectation** and **conditional variance** of Y given $X = x$ are

$$\mathbb{E}[Y|X=x] = \int y \cdot p_{Y|X}(y|x)\,dy, \quad \mathrm{Var}[Y|X=x] = \int (y - \mathbb{E}[Y|X=x])^2 \cdot p_{Y|X}(y|x)\,dy.$$

  The **covariance** of X, Y is $\mathrm{Cov}[X,Y] := \mathbb{E}[(X - \mathbb{E}X) \cdot (Y - \mathbb{E}Y)]$. Covariance describes the linear relationship between X and Y and so naturally plays an important role in linear modeling.

  Sometimes, we will use the notation $\mathbb{E}[Y|X]$. This means take $\varphi(x) := \mathbb{E}[Y|X=x]$, which is a function of x, and plug in the random variable X. In other words, $E[Y|X]$ is the random variable $\varphi(X)$. Likewise with $\mathrm{Var}[Y|X]$.

- Independence is a key property that is assumed (in one way or another) in many statistical inferences. We say that (X,Y) are **independent**, denoted $X \perp\!\!\!\perp Y$, if $P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$ for all sets $A, B \subseteq \mathbb{R}$. More generally, random variables $(X_1, \ldots, X_n)$ are **jointly independent** if $P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \times \cdots \times P(X_n \in A_n)$, for all sets $A_1, \ldots, A_n \subseteq \mathbb{R}$.

  We say that (X,Y) are **conditionally independent** given Z if $P(X \in A, Y \in B | Z = z) = P(X \in A | Z = z) \cdot P(Y \in B | Z = z)$ for all sets $A, B \subseteq \mathbb{R}$ and numbers z.

- Here are some basic but important properties of sums of random variables. Let $a, b \in \mathbb{R}$.
  - *Linearity of expectation*: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$. Also, $\mathbb{E}[f(X)Y|X] = f(X)\mathbb{E}[Y|X]$.
  - *Properties of covariance*: (i) $\mathrm{Cov}[Y,Y] = \mathrm{Var}[Y]$. $\mathrm{Cov}[Y,X] = \mathrm{Cov}[X,Y]$. $\mathrm{Cov}[X_1 + Y_1, X_2 + Y_2] = \mathrm{Cov}[X_1, X_2] + \mathrm{Cov}[X_1, Y_2] + \mathrm{Cov}[X_2, Y_1] + \mathrm{Cov}[X_2, Y_2]$. $\mathrm{Cov}[aY, bX] = ab \cdot \mathrm{Cov}[Y,X]$. $\mathrm{Cov}[Y, X + a] = \mathrm{Cov}[Y,X]$.

– *Properties of independent r.v.s*: If X and Y are independent then (i) $\mathrm{Cov}[X, Y] = 0$ though the reverse is not true in general; (ii) $p_{Y|X}(y|x) = p_Y(y)$, (iii) $\mathbb{E}[Y|X = x] = \mathbb{E}[Y]$, (iv) $\mathrm{Var}[Y|X = x] = \mathrm{Var}[Y]$.

– *Variance is not linear*: $\mathrm{Var}[aX + bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y] + 2ab\mathrm{Cov}[X, Y]$. More generally,

$$\mathrm{Var}[\sum_{i=1}^{n} a_i X_i] = \sum_{i,j=1}^{n} a_i a_j \mathrm{Cov}[X_i, X_j]$$

If $X_1, \ldots, X_n$ are independent, then the variance of a sum is the sum of variances

$$\mathrm{Var}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathrm{Var}[X_i].$$

Analogous statements hold for conditional variance.

– *Law of total expectation and variance.* Conditional expectation and variance satisfy the identities,

$$\mathbb{E}[Y] = \mathbb{E}\Big[\mathbb{E}[Y|X]\Big], \quad \mathrm{Var}[Y] = \mathbb{E}\Big[\mathrm{Var}[Y|X]\Big] + \mathrm{Var}\Big[\mathbb{E}[Y|X]\Big].$$

The first identity is called the law of total expectation, the second is called the law of total variance.

2 Review of basic statistics: estimators, bias, and standard error

- Consider the following (very basic) statistical model: $X_1, \ldots, X_n \sim P$ are independent. This model is an idealization of the common situation where the data are collected by drawing a simple random sample from a large population.
- We are interested in estimating some aspect of the unknown distribution P. The quantity we are interested in estimating is called the **parameter** or **estimand**, usually denoted by a Greek letter such as $\theta(P)$, often abbreviated to just θ. The rule we use to estimate the parameter is called the **estimator**. Estimators are functions of the data, and so are sometimes written as $\hat{\theta}(X_1, \ldots, X_n)$, but this is usually abbreviated to $\hat{\theta}_n$ or just $\hat{\theta}$. Notice that an estimator is a function of r.v.s, so it too is a random variable. The distribution of an estimator is called the **sampling distribution**.
- We can evaluate the quality of an estimator by looking at its expectation and variance. The **bias** of $\hat{\theta}$ is the difference between its expectation and the truth,

$$\mathbb{E}[\hat{\theta}] - \theta.$$

The **standard error** of $\hat{\theta}$ is the square-root of its variance,

$$\mathrm{SE}[\hat{\theta}] = \sqrt{\mathrm{Var}[\hat{\theta}]}.$$

The name "standard error" is used to distinguish between the standard deviation of an estimator, and the standard deviation of the data.

It should be intuitive that we want estimators that are unbiased with small standard error.

- A very rough rule of thumb is that the true value of a parameter is about equal to the observed value of an unbiased estimator for that parameter, "give or take" a standard error (or two). Soon, we will learn to quantify error in ways that are better than that rule of thumb.
- *Example: estimating the mean.* Define $\mu := \mathbb{E}_P[X]$ and $\sigma^2 := \mathrm{Var}_P[X]$. Suppose we want to estimate the mean μ. Then an obvious estimator is the sample mean $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$. You should remind or convince yourself that the sample mean is *unbiased*, $\mathbb{E}[\bar{X}] = \mu$, and has standard error *inversely proportional to* $\sqrt{n}$: $\mathrm{SE}[\bar{X}] = \frac{\sigma}{\sqrt{n}}$. This latter fact is sometimes known as the $\sqrt{n}$*-law.*

Notice that $\mathrm{SE}[\bar{X}] \to 0$ as $n \to \infty$. This – along with the fact that the sample mean is unbiased – means that $\bar{X}$ converges (in probability) to the true mean μ: as $n \to \infty$, $\bar{X} \to \mu$. This obviously nice property is called **consistency** of an estimator.

- *Example: estimating the variance.* Not all reasonable estimators are unbiased. Suppose we are interested in estimating the variance σ^2. Then a reasonable estimator is the sample variance $s_X^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. But the sample variance is not unbiased for the true variance, as $\mathbb{E}[s_X^2] = \frac{n-1}{n}\sigma^2$.
- Notice that in all of this there was no data analysis and, indeed, no data. When it comes time for the analysis we will not have random variables but rather just the observed data, which clearly are not random (they are just numbers). What is the connection between data and model? As Freedman puts it

  > In a regression model, as a rule, the data are observed values of random variables.

  Once we treat the data as observed values of random variables, it makes sense to use the bias and standard error of an estimator to assess the quality of an estimate calculated from the data.

  Convention is to reserve the word estimator for a rule applied to random variables, and estimate for that same rule applied to observed data (as in the previous sentence). I will try to adhere to this but may slip up sometimes.

3 Bias and standard error of least-squares in simple linear regression

- Now we return to the simple linear model, in which independent r.v.s $(X_1, Y_1), \ldots, (X_n, Y_n)$ are distributed according to the model

$$\begin{aligned} &X_i \sim P_i \text{ where } P_i \text{ is an unknown (possibly deterministic) distribution,} \\ &Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \\ &\epsilon_i \perp\!\!\!\perp X_i, \mathbb{E}[\epsilon_i] = 0, \mathrm{Var}[\epsilon_i] = \sigma^2. \end{aligned} \tag{4}$$

  Notice that the simple linear model assumes nothing about the distribution of each X_i. For example, it could be that

  - $P_i = N(0, \tau_i^2)$. (Normally distributed predictors).
  - $P_i = \mathrm{Unif}(0, 1)$. (Uniform predictors).
  - $P_i(X_i = i) = 1$ (fixed, evenly-spaced predictors).

  Instead, all the assumptions are placed on the distribution of $Y_i | X_i$.

- Our least-squares estimators for the unknown slope and intercept are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y}) \cdot (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

  Notice that I am using capital letters here, to emphasize that these are estimators (functions of r.v.s) and not estimates (functions of observed data).

- Let us focus on computing the bias and standard error of the slope estimator $\hat{\beta}_1$. It turns out to be quite tricky to do this directly. (If you don't believe me, try it!) But once we condition on the predictor variables $X_1, \ldots, X_n$, it becomes a comparatively simple calculation. So we will do just that.
- First we compute the expectation of $\hat{\beta}_1$. We begin with a preliminary calculation for the conditional expectation of $(Y_i - \bar{Y})$. Expanding $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, we see that $(Y_i - \bar{Y}) = \frac{1}{n}\sum_{j=1}^n (Y_i - Y_j) =$

$\frac{1}{n}\sum_{j=1}^n \beta_1 X_i + \epsilon_i - \beta_1 X_j - \epsilon_j$. So

$$\begin{aligned}
\mathbb{E}[(Y_i - \bar{Y})|X_1, \ldots, X_n] &= \frac{1}{n}\mathbb{E}[\sum_{j=1}^n (\beta_1 X_i + \epsilon_i - \beta_1 X_j - \epsilon_j)|X_1, \ldots, X_n] && \text{(linearity of expectation)} \\
&= \frac{1}{n}\sum_{j=1}^n (\beta_1 X_i - \mathbb{E}[\epsilon_i|X_1, \ldots, X_n] - \beta_1 X_j - \mathbb{E}[\epsilon_j|X_1, \ldots, X_n]) \\
&= \frac{1}{n}\sum_{j=1}^n (\beta_1 X_i - \beta_1 X_j) && (\epsilon \perp\!\!\!\perp X, \text{ and } \mathbb{E}\epsilon = 0) \\
&= \beta_1(X_i - \bar{X}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_1|X_1, \ldots, X_n] &= \mathbb{E}\Big[\frac{\sum_{i=1}^n (Y_i - \bar{Y}) \cdot (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}|X_1, \ldots, X_n\Big] \\
&= \frac{\sum_{i=1}^n \mathbb{E}\Big[(Y_i - \bar{Y})|X_1, \ldots, X_n\Big] \cdot (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} && \text{(linearity of expectation)} \\
&= \frac{\sum_{i=1}^n \beta_1 (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \beta_1.
\end{aligned}$$

It follows from the law of total expectation that $\mathbb{E}[\hat{\beta}_1] = \mathbb{E}[\mathbb{E}[\hat{\beta}_1|X]] = \beta_1$.

- Now we compute the variance of $\hat{\beta}_1$. Again, it is useful to begin with some preliminary calculations: using the properties of (conditional) variance and covariance, and the expansion $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, we have

$$\begin{aligned}
\mathrm{Var}[Y_i - \bar{Y}|X_1, \ldots, X_n] &= \mathrm{Var}[Y_i|X_1, \ldots, X_n] - 2\mathrm{Cov}[Y_i, \bar{Y}|X_1, \ldots, X_n] + \mathrm{Var}[\bar{Y}|X_1, \ldots, X_n] \\
&= \mathrm{Var}[\epsilon_i|X_1, \ldots, X_n] - 2\mathrm{Cov}[\epsilon_i, \bar{\epsilon}|X_1, \ldots, X_n] + \mathrm{Var}[\bar{\epsilon}|X_1, \ldots, X_n] \\
& && (\bar{\epsilon} = \tfrac{1}{n}\textstyle\sum_{i=1}^n \epsilon_i) \\
&= \sigma^2 - \frac{2\sigma^2}{n} + \frac{\sigma^2}{n} \\
&= \sigma^2 - \frac{\sigma^2}{n},
\end{aligned}$$

and for all $i \neq j$,

$$\begin{aligned}
&\mathrm{Cov}[Y_i - \bar{Y}, Y_j - \bar{Y}|X_1, \ldots, X_n] \\
&= \mathrm{Cov}[Y_i, Y_j|X_1, \ldots, X_n] - \mathrm{Cov}[Y_i, \bar{Y}|X_1, \ldots, X_n] - \mathrm{Cov}[\bar{Y}, Y_j|X_1, \ldots, X_n] + \mathrm{Var}[\bar{Y}|X_1, \ldots, X_n] \\
&= \mathrm{Cov}[\epsilon_i, \epsilon_j|X_1, \ldots, X_n] - \mathrm{Cov}[\epsilon_i, \bar{\epsilon}|X_1, \ldots, X_n] - \mathrm{Cov}[\bar{\epsilon}, \epsilon_j|X_1, \ldots, X_n] + \mathrm{Var}[\bar{\epsilon}|X_1, \ldots, X_n] \\
&= 0 - \frac{\sigma^2}{n} - \frac{\sigma^2}{n} + \frac{\sigma^2}{n} \\
&= -\frac{\sigma^2}{n}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Var}[\hat{\beta}_1 | X_1, \ldots, X_n] &= \frac{\sum_{i,j=1}^n (X_i - \bar{X})(X_j - \bar{X}) \mathrm{Cov}[Y_i - \bar{Y}, Y_j - \bar{Y}]}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} \\
&= \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \sigma^2 - \sum_{i,j=1}^n (X_i - \bar{X})(X_j - \bar{X}) \frac{\sigma^2}{n}}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} \\
&= \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \sigma^2}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \frac{\sigma^2}{n s_X^2}.
\end{aligned}$$

We can see that $\mathrm{SE}[\hat{\beta}_1 | X_1, \ldots, X_n] = \sigma / \sqrt{n} s_X$.

- We conclude that $\hat{\beta}_1$ is unbiased for the true slope β_1. As we saw for the sample mean, the (conditional) variance of $\hat{\beta}_1$ increases with σ and decreases with $\sqrt{n}$: in other words, the estimator becomes less accurate with larger errors, and more accurate with more samples.

  Unlike for the sample mean, however, there is now the term s_X in the denominator of the standard error of $\hat{\beta}_1$. This means that the more spread out the Xs are, the more accurate our estimator for slope.

- The rule of thumb discussed previously tells us that the true slope β_1 is likely "about equal to the least-squares estimate $\hat{\beta}_1$, give or take $\frac{\sigma}{\sqrt{n} s_X}$."

4 Additional remarks: plug-in method

- Last class, we arrived at our estimators $\hat{\beta}_1, \hat{\beta}_0$ through the method of least squares. Now let's give another way of motivating them.

- Suppose $X_1, \ldots, X_n \sim P$ are identically distributed. (Notice that this is not an assumption in the linear model.) Letting (X, Y) be random variables with the same distribution as $(X_1, Y_1), \ldots, (X_n, Y_n)$, we can define the intercept and slope in terms of variances, covariances, and expectations of (X, Y). For the slope,

$$\mathrm{Cov}[Y, X] = \mathrm{Cov}[\beta_0 + \beta_1 X + \epsilon, X] = \beta_1 \mathrm{Var}[X] \iff \beta_1 = \frac{\mathrm{Cov}[Y, X]}{\mathrm{Var}[X]}, \tag{5}$$

  and for the intercept,

$$\mathbb{E}[Y] = \mathbb{E}[\beta_0 + \beta_1 X + \epsilon] = \beta_0 + \beta_1 \mathbb{E}[X] \iff \beta_0 = \mathbb{E}[Y] - \frac{\mathrm{Cov}[Y, X]}{\mathrm{Var}[X]} \mathbb{E}[X]. \tag{6}$$

- This suggests an alternative method for fitting the regression model: simply plug-in estimates for mean, variance, and covariance into the equations for β_0 and β_1. We already know how to estimate the mean of X, Y, by the sample mean $\bar{X}, \bar{Y}$. Likewise, we can estimate the variance of X, Y by the sample variance s_X^2, s_Y^2. Finally, the covariance of X, Y can be similarly estimated by the empirical covariance:

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

  If in (5) and (6), we substitute sample means/variances/covariance for the true (unknown) means/variances/covariance, we get back

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y}) \cdot (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

  But this is exactly what we got from the method of least squares!

- The previous approach is called the *plug-in method*, since it builds up estimates for "complicated" parameters by plugging in estimates for simpler ones. It is a special property of linear models that the least-squares and plug-in methods give the same result.