Unit 3: Statistical inference for single-predictor linear regression models

Ethan Levien

October 29, 2025

Contents

3.1 Introduction

Now we are almost ready to begin working with regression models, but we first discuss some concepts from statistical inference. These include the idea of a sample distribution (already alluded to in our discussion of the central limit theorem), bias, consistency and confidence intervals. Then we work define linear regression models, which have also already appeared in different forms. The basic idea is that we have some predictors $X_1, X_2, \ldots, X_k$ and conditional on these variables, the distribution of another variable – the response variable Y – is Normal. Moreover, it's mean is a linear function of these variables. These two assumptions allow us to obtain relatively simple formula (at least for a computer) for the coefficients of the X_i in the linear formula. We will start, in this section, with a single-predictor learn how to assess a linear regression model in this context. In particular, we will learn about correlation coefficients, R^2 and p-values, standard errors, confidence intervals. You will need to understand what all these quantities tell us and how they are related.

3.2 Estimators

We have danced around the concept of statistical inference and parameter estimation for a bit and we saw an example of an estimator: $\hat{q} = \overline{Y}$ is an estimator of the parameter q for a Bernoulli distribution. We also talked a lot about going between the world of data (e.g. sample means, histogram) and math (e.g. expectation, density) with sample averages.

Statistical inference is the process of estimating the parameters (e.g. μ and σ) based on samples of Y AND expressing our uncertainty in these estimates. The expressing the uncertainty part is what we have not yet discussed formally.

In general, an estimator $\hat{\theta}$ of a parameter θ is just something we compute from the data which is meant to approximate θ. This could in principle be any quantity we can compute from the data (like the maximum value), but it will almost always be represented as a sample average in some way. We will see examples of this soon.

The point that should be emphasized is that the estimator is a function of the data. That is, $\hat{\theta}$ **depends on the specific data we collect** or simulation we run. It is meant to approximate a parameter which does not depend on the data and is (in classical statistics) a fixed number. For example, in the instance of a YES/NO survey or election with two candidates, the "true" quantity we are interested in measuring is the fraction of people answering YES to some question. Our estimate, $\hat{q}$, is a variable which depends on the specific subset of the population we sample and it will change if we look at a different subset.

3.2.1 Sample distribution and standard errors

We call the distribution of $\hat{\theta}$ over many replications of our data the sample distribution. I will use replicate to mean different realizations of our data (as opposed to the different samples within the data). The distinction is shown in Figure 1 (left panel). The terminology gets a bit confusing: The sample distribution is the distribution of $\hat{\theta}$ over many replicates, but each replicate involves many samples.

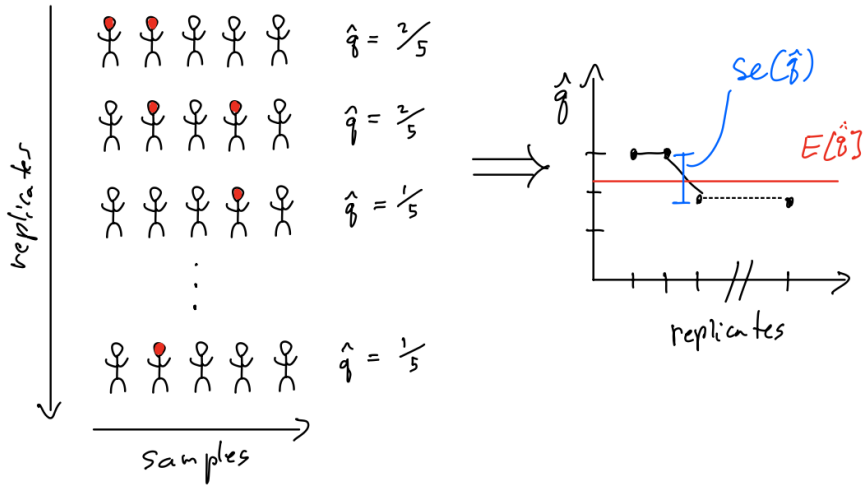

Figure 1: Replicates and samples

Example 1 (sample distribution of normal mean)**.** Suppose

$$Y \sim \text{Normal}(\mu, \sigma)$$

Question: What is the sample distribution of $\hat{\mu}$ (our estimate of μ)?

Solution:

$$\hat{\mu} = \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

The CLT tell us (informally speaking) that

$$\sum_{i=1}^{N} Y_i \sim \text{Normal}\left(\mu n, n\sigma^2\right)$$

where by $\sim$ we really mean "approximately distributed as". Dividing by N,

$$\hat{\mu} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{N}\right)$$

This assumes σ is known.

A natural way to quantify the uncertainty in our estimate is the standard deviation of the $\hat{\mu}$ under the sample distribution. We call the resulting quantity the standard error, which is our estimate of the standard deviation of the sample distribution. For the Normal model, if we are estimating the mean and happen to know σ, then

$$\text{se}(\hat{\mu}) = \frac{\sigma}{\sqrt{N}}. \tag{1}$$

This tells us how much our estimate will vary between different experiments (or surveys/simulations). Importantly, the standard error depends on σ which we may not know!!! Thus, it is common to estimate the standard error using an estimate of σ, $\hat{\sigma}$, leading to an estimator of the standard deviation:

$$\text{se}(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{N}}. \tag{2}$$

It should be clear from the context which one we are talking about: If we are working with data and we don't know what σ is, when we say standard error we mean Equation 2. If we are working with a particular model where we have specified the parameters, we mean Equation 1.

3.2.2 Bias and consistency

There must be some properties we would like the estimator to have. At a minimum, it should be in some way informed by the data, in the sense that having more data should bring our estimate closer to the actual value of the parameter. More precisely, the more data we have (e.g. the larger N) the closer we expect $\hat{\theta}$ to be to the true value θ. Of course, we must define what we mean by "closer" when we are talking about random things. For our purposes we will say $\hat{\theta}$ is consistent if

$$E[\hat{\theta}] \to \theta \quad \text{and } \text{se}(\hat{\theta}) \to 0 \quad \text{as} \quad N \to \infty.$$

This is saying that as we obtain more and more samples, the sample distribution because more concentrated around θ.

To see that consistency is not the only property we look for in an estimator, notice that since $\hat{\mu}_1 = \hat{\mu} + 1/N$ is also consistent, yet clearly seems inferior to $\hat{\mu}$. To this end, we say that an estimator $\hat{\theta}$ is unbiased for some N (not just very large N), the average over the sample distribution is equal to the actual value under the model distribution; that is,

$$E[\hat{\theta}] = \theta.$$

Example 2 (Bias and consistency)**.** For a normal random variable, define the following estimators of the mean:

$$\hat{\mu}_2 = \frac{Y_1 + Y_2}{2}$$

Question: Is $\hat{\mu}_2$ biased and consistent? what is the sample distribution?

Solution: Note that $\hat{\mu}_2$ has the sample distribution

$$\hat{\mu}_2 \sim \text{Normal}(\mu, \sigma/\sqrt{2})$$

It is therefore unbiased but not consistent, since $\text{se}(\hat{\mu}_2) = \sigma/\sqrt{2}$ does not depend on n.

Example 3 (Normal standard deviation)**.** Let now consider estimating the standard deviation of a Normal random variable

$$Y \sim \text{Normal}(\mu, \sigma^2)$$

Given samples $Y_1, Y_2, \ldots, Y_n$, it seems the natural way to estimate σ^2 is using

$$\text{var}(Y) = E[(Y - E[Y])^2] \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

we will call this estimator $\hat{\sigma}_0^2$. It turns out $\hat{\sigma}_0^2$ is biased and in-fact

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{n}{n-1}\hat{\sigma}_0^2$$

is unbiased. The correction by a factor $n/(n-1)$ is called Bessel correction.

Question: Demonstrate with simulated data that $\hat{\sigma}_0^2$ is biased and $\hat{\sigma}^2$ is not.

3.2.3 Confidence intervals

The idea of the confidence interval is, roughly speaking, to describe the range of values where we think the actual value of θ might reasonable be given some estimate $\hat{\theta}$ and its sample distribution. We will mostly work with the 95% confidence interval, or 95%-CI, which is given by

$$[\hat{\theta} - 1.96\text{se}(\hat{\theta}), \hat{\theta} + 1.96\text{se}(\hat{\theta})] \tag{3}$$

The factors 1.96 in front of the standard errors ensure that 95% of samples from the sample distribution will fall in this range. This makes sense if the sample distribution is well approximated by a Normal distribution.

Note that these samples from the sample distribution do not have the same distribution as $\hat{\theta}$ over replicates of our data. Said another way, if we draw many samples from our estimate of the sample distribution, their distribution will not be the same as the distribution of $\hat{\theta}$ we would obtain if we ran an experiment many times and estimated $\hat{\theta}$ each time. The correct interpretation of the 95%-CI is as follows: **If we generate many replicates of the data then θ (the true value) will fall in the CI, for 95% of them.**

Technically speaking, is is NOT the case that there is a 95% chance the true value of θ is in the 95%-CI. To understand why, note that the parameter has a 95% chance to be in the interval

$$[\theta - 1.96\text{std}(\theta), \theta + 1.96\text{std}(\theta)] \tag{4}$$

but this is difference from Equation 3, since we have replaced $\hat{\theta}$ with θ. The distinction, which is shown in Figure 2, is important; however, you don't need to get bogged down by the subtle differences in interpretation. For practical purposes, you can pretty much thing of the 95%-CI as the region where the parameter value is likely to be. We provide alternatives ways to think about these intervals when we discuss Bayesian vs. classical statistics.

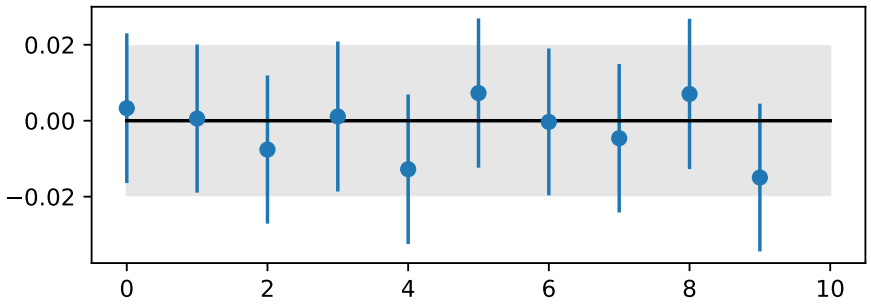

Figure 2: An illustration of the distinction between Equation 4 (gray shaded region) and Equation 3.

Example 4 (Estimating CI)**.** Imagine we are designing an experiment. Our model is a Normal distribution and from previous experience, we have a ballpark estimate of the standard deviation, which is $\sigma = 1$.

Question: Roughly, how many samples do we need to collect to have a 95% chance our estimate is within 1 of the actual value of the variable?

Solution: The standard error of an estimate based on n samples is $\sigma/\sqrt{n} = 1/\sqrt{n}$, so the confidence interval is

$$[\hat{\mu} - 1.96 \times 1/\sqrt{n}, \hat{\mu} + 1.96 \times 1/\sqrt{n}] = [\hat{\mu} - 1.96/\sqrt{n}, \hat{\mu} + 1.96/\sqrt{n}]$$

The width of this interval is $2 \times 1.96/\sqrt{n} = 3.92/\sqrt{n}$. This interval will intersect the true value in 95% of replicates, so we would like it to have a width < 2. It follows that we need

$$\frac{3.92}{\sqrt{n}} < 2 \implies \sqrt{n} > \frac{3.92}{2} = 1.96 \implies n > (1.96)^2 \approx 3.84$$

We can test this by running many replicates for each n, as done in the class notebook.

3.2.4 Bias-variance tradeoff

We introduce the mean-squared error of an estimator $\hat{\theta}$ of some quantity θ. θ could be a parameter, or it could be a value of a function, such as $f(x)$, that we would like to predict.

$$\text{MSE}_{\hat{\theta}} = E\left[(\hat{\theta} - \theta)^2\right] \tag{5}$$

For now, let's just think of $\hat{\theta}$ as any estimator. The following theorem is the key result which will allow us to understand the U-shaped curve.

Theorem 1 (Bias variance decomposition)**.**

$$\text{MSE}_{\hat{\theta}} = \text{var}(\hat{\theta}) + E\left[\hat{\theta} - \theta\right]^2 \tag{6}$$

Proof. Using the definition of variance

$$\text{var}\left(\hat{\theta} - \theta\right) = E\left[(\hat{\theta} - \theta)^2\right] - E\left[\hat{\theta} - \theta\right]^2.$$

Since θ is a constant, $\text{var}\left(\hat{\theta} - \theta\right) = \text{var}\left(\hat{\theta}\right)$, so rearranging terms yields the result. □

3.2.5 Maximum Likelihood (optional)

Sometimes it is quite clear what the estimator for a parameter should be. This is the case for q in the Bernoulli distribution. However, we will find this is not always the case, so it is useful to have a **more systematic way of finding estimators.** Recall that the probability distribution for the binomial distribution is

$$P(Y) = \binom{n}{Y} q^Y (1-q)^{n-Y} \tag{7}$$

In statistics, we sometimes call this the likelihood and denoted $P(Y) = L(Y|q)$. The notation here is suggesting that we think of P as a distribution which is conditioned on a particular value of the parameter. More generally, the likelihood is defined as the probability we say a data set given the parameters. This notation and terminology foreshadows Bayesian thinking, wherein one thinks of the parameter as random variables themselves – more on this later. For now, notice that Equation (7) tells us how likely it is to observe k YES among n people surveyed. Then, it seems reasonable that this number should not be very small, since that would mean our survey results are an anomaly. More generally, the larger $L(Y|q)$ is the more likelihood our results are. This suggests one a way to estimate determine q: We can take as our estimate $\hat{q}$ the value which makes $L(Y|q)$ largest. In other words, we are finding the value of q which makes the data the most likely, and we will call this the maximum likelihood estimate. You can do this using calculus (try it!) to determine that the value of q which makes (7) largest is

$$\hat{q}_{\text{MLE}} = \frac{Y}{n}$$

For a Normal distribution with mean and variance μ and σ, the MLE estimators are the usual sample mean and standard deviation which we have already been exposed to.

3.3 Single-predictor linear regression model

Regression models model input-output relationships. The input is the predictor (X) and the output is the response variable (Y). Recall from the previous unit that a linear regression model is defined by

$$Y = \beta_0 + \beta_1 X_1 + \epsilon, \quad \epsilon \sim \text{Normal}(0, \sigma_\xi^2).$$

Written another way

$$Y|X \sim \text{Normal}(\beta_0 + \beta_1 X, \sigma^2). \tag{8}$$

It should be noted that in regression modeling the distribution of the predictor is often not specified. However, we should always think about what this is, because it plays an important role in the inference process as I will discuss below. For this reason, I often think of a linear regression model as a model of both variables:

$$\begin{aligned} X &\sim \text{some distribution with mean } \mu_x \text{ and variance } \sigma_x^2 \\ Y|X &\sim \text{Normal}(\beta_1 X + \beta_0, \sigma^2). \end{aligned} \tag{9}$$

We saw examples with Bernoulli and Normal predictors in Unit 2.

3.3.1 Covariance

The example above motivates the definition of covariance

$$\text{cov}(X, Y) = E[XY] - E[X]E[Y] \tag{10}$$

Note that another way to write this is

$$E[(Y - E[Y])(X - E[X])] = E[XY] - 2E[X]E[Y] + E[X]E[Y] = \text{cov}(X, Y)$$

so if we replaced X with Y, this becomes the variance.

The relationships above can easily be generalized. Recall that we can write $E[X^2]$ as

$$E[X^2] = \text{var}(X) + E[X]^2 = \sigma_x^2 + \mu_x^2.$$

Now for any linear regression model

$$\begin{aligned} E[XY] &= E[XE[Y|X]] = E[X(\beta_1 X + \beta_0)] = \beta_1 E[X^2] + \beta_0 E[X] \\ &= \beta_1 \sigma_x^2 + \beta_1 \mu_x^2 + \beta_0 \mu_x \\ E[Y] &= \beta_1 \mu_x + \beta_0 \end{aligned}$$

so

$$\text{cov}(X, Y) = \beta_1 \sigma_x^2$$

regardless of the distribution of X (assuming $\sigma_x^2 < \infty$).

A crucial observation is that the covariance allows us to relate the parameter β_1 (the slope) in the model above to averages over X and Y. In other words, it provides us with a means to estimate the slope from samples $(x_1, y_2), \ldots, (x_n, y_n)$.

$$\beta_1 \approx \frac{\sum_{i=1}^n \left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right)}{\sum_{i=1}^n \left(x_i - \bar{X}\right)^2}$$

The is what the function

```
> np.cov(x,y)[0,1]
```

computes in Python. The reason for the $[0, 1]$ is that the covariance function in numpy actually computes a 2D array (a Matrix), where the off diagonal entries are the covariance. The diagonal entries are the variances.

We can also estimate β_0. Using $E[Y] = \beta_1 \mu_x + \beta_0$ we have

$$\beta_0 = E[Y] - \beta_1 \mu_x \approx \hat{\beta}_0 = \overline{Y} - \left(\frac{\sum_{i=1}^n \left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right)}{\sum_{i=1}^n \left(x_i - \bar{X}\right)^2} \right) \overline{X}$$

3.3.2 Least square interpretation

Suppose we plot X and Y points in a place. Regardless of where these X and Y points come from (Normal model or not), we can compute $\hat{\beta}_1$ and $\hat{\beta}_0$. These estimators are known as least squares estimators (note that we haven't formally defined what an estimator is) because it happens that these values minimize the sum of the squared difference between our data points and the line $\hat{a}x + \hat{b}$. That is, they are the values that make the residual sum of squares(RSS) smallest

$$RSS = \sum_{i=1}^n r_i^2, \quad r_i = Y_i - (\hat{\beta}_1 X_i + \hat{\beta}_0)$$

smallest. The R

There are many other ways we could draw a line through a set of (x, y) points. This particular way of estimating the slope – by minimizing RSS – happen to make sense under the assumption that the data is sampled from a Linear regression model (Equation 9).

Example 5 (Marketing data)**.** Here we consider the some on advertising budgets and sales for a company. We will explore whether the budget for TV advertisements is associated with higher sales.

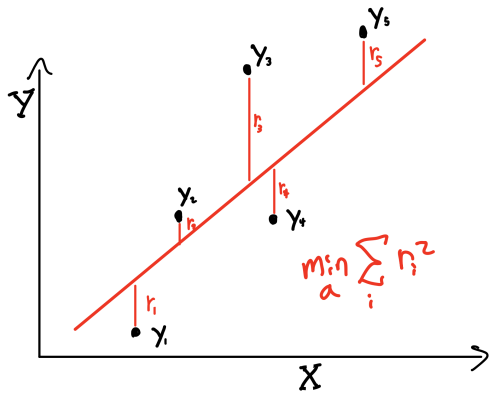

Figure 3:

Questions: Fit the data to a linear regression model with the TV budget at the predictor and sales as the response variable.

(a) **Fit linear regression model:** What are the estimates of β_1 and β_0?

(b) **Visualize the data:** Plot the regression long along with a scatter plot of the data.

(c) **Accessing model assumptions:** Using the fitted values of β_1 and β_0, simulate 10 "fake" data sets which have the same number of points as the real data set and the same x values. Make plots of these and compare to the real data.

Solution: see colab notebook.

3.4 Coefficient of determination and correlation

3.4.1 Coefficient of determination

In most applications the goal of regression modeling is predict the response (Y values) given the predictor (X values). Thus, it is natural to look for metrics which assess how well these predictions can be made. To gain an intuition about this, let's think about the best case scenario for making predictions. This would be that $\sigma_\epsilon = 0$ (and $\beta_1 \neq 0$), because then Y is determined *deterministically* for a given X. Predictions should get worse as σ_ϵ grows, because σ_ϵ measure the variability in the Y values for a fixed X. So the worst case is that σ_ϵ is very large (or $\beta_1 = 0$), but large relative to what? σ_ϵ^2 has the same units as Y. This means we need to compare it to something that has Y units, so we compare to the overall (marginal) variance in Y, $\sigma_Y^2 = \text{var}(Y)$. Note that (check yourself!)

$$\sigma_Y^2 = \beta_1^2 \sigma_X^2 + \sigma_\epsilon^2 \geq \sigma_\epsilon^2 \tag{11}$$

Notice that the overall variance in Y is the sum of the variation from ϵ (random factors independent of X) and $\beta_1^2 \sigma_X^2$ (the spread of the X values, which is weighted by the squared slope).

This discussion motivates coefficient of determination, which is

$$\rho^2 = 1 - \frac{\sigma_\epsilon^2}{\sigma_Y^2}. \tag{12}$$

Another way to write ρ^2 is

$$\rho^2 = 1 - \frac{\text{var}(Y|X)}{\text{var}(Y)}. \tag{13}$$

This is between zero and one. When $\rho^2 = 1$, this is saying nearly all the variation in Y values is a result of variation in X. Notice that $\rho^2 = 0$ when $\sigma_{\epsilon^2} = \sigma_Y^2$. In this case, knowing the X values (e.g. knowing whether some is in the control or treatment group) does not change how variable the Y values are (e.g. the blood pressure). This means all the variation in Y values is due to things other than X.

When we estimate ρ^2 from data, we call it R^2. The estimator has the form

$$R^2 = 1 - \frac{\sum_{i=1}^N r_i^2}{\sum_{i=1}^N (Y_i - \overline{Y})^2} \tag{14}$$

It should make sense why this is the estimator of ρ^2.

3.4.2 Correlation

There is another way to arrive at ρ^2, which is to consider the regression model after standardizing the predictor and response variable. This is natural thing to do since we want to measure the association between two variables on a dimensionless scale. Let

$$Z_X = \frac{X - \mu_X}{\sigma_X}, \quad Z_Y = \frac{Y - \mu_Y}{\sigma_Y} \tag{15}$$

We then have

$$Z_Y = \frac{\beta_0 + \beta_1 X + \epsilon - \mu_Y}{\sigma_Y} = \tag{16}$$

Because $\mu_Y = \beta_0 + \beta_1 \mu_X$, we simplify:

$$Z_Y = \frac{\beta_1 (X - \mu_X)}{\sigma_Y} + \frac{\epsilon}{\sigma_Y} = \frac{\beta_1 \sigma_X}{\sigma_Y} Z_X + \frac{\epsilon}{\sigma_Y}.$$

Thus,

$$Z_Y \big| Z_X \sim \text{Normal}\left(\frac{\beta_1 \sigma_X}{\sigma_Y} Z_X, \frac{\sigma_\epsilon^2}{\sigma_Y^2} \right).$$

This is a simple linear regression of standardized response Z_Y on standardized predictor Z_X with slope

$$b = \frac{\beta_1 \sigma_X}{\sigma_Y}.$$

The slope b represents the expected change in standard deviations of Y associated with a one standard deviation change in X.

Motivated by this, we define the correlation coefficient ρ as this regression slope:

$$\rho := b = \frac{\beta_1 \sigma_X}{\sigma_Y}.$$

Using the earlier result that

$$\text{cov}(X, Y) = \beta_1 \sigma_X^2,$$

we can write

$$\rho = \frac{\beta_1 \sigma_X}{\sigma_Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Usually it is the last equation that is used to define covariance, because this can be defined for any two random variables and doesn't require an underlying regression model (which is where the β_1 comes from). Notice that if we begin with two random variables which have standard deviations 1, then ρ and covariance are the same thing. There is a very import result in math called the Cauchy-Schwarz inequality, which tells us $-1 \leq \rho \leq 1$.

Now to show that this is indeed the same as ρ^2. To do so, we recall that $\sigma_Y = \beta_1^2\sigma_X^2 + \sigma_\epsilon^2$. Therefore,

$$\rho^2 = \frac{\beta_1^2\sigma_X^2}{\sigma_Y^2} = \frac{\sigma_Y^2 - \sigma_\epsilon^2}{\sigma_Y^2} \tag{17}$$

Thus, the approach of standardizing and computing a unitless regression slope ultimately leads to the same metric for the strength of association as comparing the variance conditioned on the predictor to the overall variance.

3.4.3 Regression to the Mean

Consider two standardized random variables X and Y, both with mean zero and variance one. Suppose the conditional distribution of Y given X follows a Normal distribution

$$Y \mid X \sim \text{Normal}(\rho X,\, 1 - \rho^2),$$

where the correlation coefficient ρ satisfies $-1 \leq \rho \leq 1$.

Because X and Y share the same marginal distribution (both standardized with mean zero and variance one), the conditional expectation is

$$\mathbb{E}[Y \mid X] = \rho X.$$

This implies a phenomenon known as **regression to the mean**: if X takes an unusually high (positive) value, then the expected value of Y given X is closer to zero (the mean) than X itself, provided $|\rho| < 1$. In other words, when the correlation is not perfect, extreme values of X tend to be associated with less extreme values of Y on average.

The interpretation can be generalized as:

- When X is above its mean (zero), $\mathbb{E}[Y|X]$ is also above zero but shrunk towards zero by factor ρ.
- When X is below its mean, $\mathbb{E}[Y|X]$ is below zero but again closer to zero than X.

This regression effect arises naturally in linear regression models and is a central concept in understanding the relationship between two correlated variables in standardized units. As discussed in class, regression to the mean can also be understood in a way where once everything is standardized, if things are up, they are likely go down. On the other hand, if things are down, they are likely to go up, thus leading us to the general idea of regression.

3.5 Hypothesis testing (optional)

In statistics, we might infer parameters not because we are interested in specific values, but rather because we would like to use them to make a decision. For example, in a clinical trial, we might be interested in deciding whether a candidate drug is worth moving forward with. This problem is often framed in terms of hypothesis testing, in which we assign a probability to a particular hypothesis or its converse. In rather abstract terms, the basic procedure of hypothesis testing is as follows:

1. Come up with a null hypothesis. For example, this might be that the mean of some variable is zero. We are interested in determining whether we can rule this hypothesis out.
2. Compute something called a test statistic, denoted $\hat{T}$, which like any estimator is simply some quantity we compute from our data.
3. Next, we do a sort of probabilistic thought experiment and ask: What is the chance that we would observe a value of $\hat{T}$ at least as large as the value we measured IF our hypothesis was in-fact true. The result is the p-value.

Example 6 (hypothesis testing for a linear regression model with binary predictor)**.** Consider the example of a clinical trial again. The effect of a drug, denoted Y (e.g. blood pressure is measured in two groups) is measured in two groups. One group is given a placebo, the other (the treatment group) is given a drug. Let $X = 0$ for people in the control group and $X = 2$ for those in the treatment group. For simplicity we will assume that there are $N/2$ people in each group. We can model Y with a regression model

$$Y|X \sim \text{Normal}(\mu_C(1-X) + \mu_T X, \sigma^2)$$

We will assume σ^2 is know! This greatly simplifies the calculations! This is just a linear regression model since we could write

$$\mu_C(1-X) + \mu_T X = \mu_C + (\mu_T - \mu_C)X = \beta_0 + \beta_1 X$$

where

$$\begin{aligned}\beta_0 &= \mu_C \\ \beta_1 &= \mu_T - \mu_C.\end{aligned}$$

We could estimate β_0 and β_1 as we always do in a linear regression model. We could also simply perform inference on the mean and of a Normal distribution within each group to obtain estimators of μ_C and μ_T. For simplicity, let's pretend σ is known for simplicity. This makes things simple, because then the sample distributions are

$$\begin{aligned}\hat{\mu}_C &\sim \text{Normal}\left(\hat{\mu}_C, \frac{\sigma^2}{N/2}\right) \\ \hat{\mu}_T &\sim \text{Normal}\left(\hat{\mu}_T, \frac{\sigma^2}{N/2}\right).\end{aligned}$$

Thus the (estimated) sample distribution of β_1 is

$$\hat{\beta}_1 \sim \text{Normal}\left(\hat{\beta}, \frac{4\sigma^2}{N}\right).$$

In this case, our null hypothesis will be that $\beta_1 = 0$; that is, there is no effect of the drug. As our test statistic, we measure how far β_1 is from zero in standard deviations:

$$\hat{T} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$$

Remember that since we know σ, $\text{se}(\hat{\beta})$ is known and therefore, from the perspective of the sample distribution, this is just dividing by a constant. Now, let $\hat{\beta}_1^*$ be the random variable representing the measured effect under the null hypothesis. Another way to say this is that $\hat{\beta}_1^*$ represents a measurement of β_1 from a replica generated under the assumption that $\beta_1 = 0$. Therefore, $\hat{\beta}_1^*$ will have a distribution centered at zero and with a standard deviation $\text{se}(\hat{\beta}_1)$. This means the distribution of $\hat{\beta}_1^*$ is nothing but the sample distribution shifted to zero, or

$$\hat{\beta}_1^* \sim \text{Normal}\left(0, \frac{4\sigma^2}{N}\right)$$

At this point we can answer the question posed in step 3: **If the null hypothesis was true, how likely would we be to observe a value of $\hat{T}$ larger than the one we did?** This is determined by the p-value:

$$p_v = P(|\hat{T}^*| > |\hat{T}||\hat{T}) \tag{18}$$

where $\hat{T}^*$ is the test statistic computed from $\hat{\beta}_1^*$ and the probability is taken over all the distribution of $\hat{T}^*$, while $\hat{T}$ is given by our data (hence why I use the conditioning notation). p_v, like $\hat{T}$, is a function of the data. See the python notebook were we compute p_v with simulations.

The above example is very simple because we assume that σ is known and we have only a binary predictor. In reality, the computation of p-values is much more complex, however the principle and interpretation is the same! **Interpreting the p-value** If the p-value is very small, then it is highly unlikely we would have observed what we did when the null hypothesis was true. In this case, we can REJECT the null hypothesis as false. Usually some threshold is set for this, and if the p_v is below that threshold we say our result in statistically significant. On the other hand, **if p_v is not small, it does not necessarily mean the null hypothesis is true.** A result is said to be statistically significant if $p_v < 0.05$. Visually, we can see that β_1 is statistically significant exactly if 0 is not contained in the confidence interval! **Relationship between p-values and confidence intervals**. The p-value is all about the "tail" of the sample distribution – "tail" usually just means the ends of the distribution. Naturally there is connection between between p-values and confidence intervals, which also measure the width of the sample distribution. To illustrate the connection, we will again assume σ **is known**. Since the the sample distribution can be obtained by shifting the distribution of $\hat{\beta}_1^*$ to $\hat{\beta}_1$, the p-value, p_v, is exactly the chance of being outside the interval $[\hat{\beta}_1 - |\hat{\beta}_1|, \hat{\beta}_1 + |\hat{\beta}_1|]$. Therefore, recalling the interpretation of confidence intervals, $\hat{\beta}_1$ will fall in the confidence interval with probability p_v when the null hypothesis is true. If σ isn't known all this is only approximately true, but intuition is still.

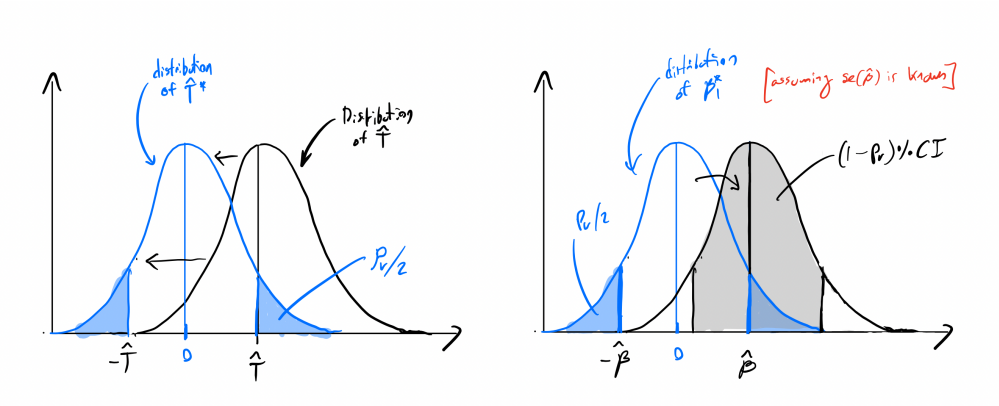

Figure 4: (left) The (two-sided) p-value and (right) the relationship between p_v and the confidence interval.

Exercises

Exercise 1 (Bias and consistency)**:** Let

$$X \sim \text{Bernoulli}(q)$$

and $X_1, \ldots, X_N$ denote N samples of X. For each of the following estimators of q, (i) write down the standard error and (ii) state whether they are un-biased and/or consistent. (In each case, you can write down an exact formula for the standard error, so you do NOT need to use the CLT.)

(a)

$$\hat{q}_0 = \frac{1}{N}\sum_{i=1}^{N} X_i$$

(b)

$$\hat{q}_1 = \frac{Y}{N} + \frac{1}{\sqrt{N}}, \quad Y = \sum_{i=1}^{N} X_i$$

(c)

$$\hat{q}_2 = \frac{1}{\lfloor N/2 \rfloor}\sum_{i=1}^{\lfloor N/2 \rfloor} X_i$$

The notation $\lfloor n \rfloor$ means the floor; that is, the largest integer less than n. For example, $\lfloor 101/2 \rfloor = \lfloor 50.5 \rfloor = 50$.

Exercise 2 (Estimator of mean in exponential model)**:** Let

$$T \sim \text{Exponential}(\lambda).$$

Recall that $E[T] = 1/\lambda$. We can estimate $E[T]$ via the sample average of measurements $T_1, \ldots, T_n$,

$$E[T] \approx \overline{T} = \frac{1}{n}\sum_{i=1}^{n} T_i.$$

This suggests that a natural way to estimate λ is by

$$\hat{\lambda} = \frac{1}{\overline{T}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} T_i}.$$

(a) The goal of the first part of this problem is to show, using simulations, that this is in-fact a biased estimator of λ, although the bias decreases with n. To achieve this, you should do the following:

- Make a list of 100 values of λ. You could use any range, but I picked between 0.2 and 2.
- For each value of λ,
  - simulate 10000 replicates of an experiment, where each replicate includes $n = 5$ values of T.
  - For each of these replicates, compute $\hat{\lambda}$ as defined above.
  - Then estimate the average $E[\hat{\lambda}]$ and save this value in a list.
- Make a plot of λ vs. $\left|E[\hat{\lambda}] - \lambda\right|$.

(b) (**optional – ungraded**) Consider the case $n = 2$. Prove that

$$E[\hat{\lambda}] = E\left[\frac{1}{\overline{T}}\right] \geq \lambda$$

This is a special case of Jenson's inequality.

Exercise 3 (Earnings data)**:** Consider the earnings data. This can be loaded with

```
> df = pd.read_csv("https://raw.githubusercontent.com/avehtari
> /ROS-Examples/master/Earnings/data/earnings.csv")
```

In this exercises, you will study the association between earnings and gender. In particular, you will explore how this depends on height. Later we will see there is a better way to answer this question by performing a regression with multiple predictors, but taking this more elementary approach will elucidate some key aspects of regression analysis.

(a) What do you expect the association between gender and earnings to be? Where do you expectations come from (news, intuition, other courses you've taken)?

(b) Using stats models, perform a linear regression on with gender (the column "male") as the predictor and earnings as the response variable. You can either use "earnk" or "earn", just keep track of the units. Then answer the questions

- Is there a statistically significant effect?
- Is the direction and size of the effect what you expected?

(c) Using stats models, perform a linear regression with height as the predictor and earnings as the response variable. Answer the same questions which are posed in part (a).

(d) You should have found there is an association between both gender and earnings, as well as height and earnings. A natural question is whether the association between height and earnings is simply a byproduct of the fact that men are taller on average. To answer this question, separate the data into males and females, then fit the linear regression model with height as a predictor separately for each group.

(e) Based on the results from the previous problem, what do you conclude? Is the association between height and earnings solely due to the association between gender and heights? Do you think it is partially due to the height?

Exercise 4 (Estimating slope ❑)**:** 1. Estimate the regression slope of the following dataset by hand

X	Y
1	2
2	3
3	5
4	6
5	8

2. What happens when you add a new data point at $(X_6, Y_6) = (100, 3)$? The numbers don't work out as nicely, so you can using Python or LLM. Now try changing the X and Y of the new datapoint. How does the slope change?

Exercise 5 (Linear regression model parameters ❑)**:** Suppose that for a given linear regression model it is found that $\hat{\beta}_1 = 1/2$, $\hat{\sigma}_\epsilon = 1$ and $\sigma_X^2 = 4$. What is your estimate of R^2?

Exercise 6 (Statistical significance – **optional challenge**)**:** Show (using math OR simulations) that it is possible to conduct two experiments (let's use clinical trials as an example) so that $\Delta\hat{\mu}$ (using the same notation as my notes) is statistically significant for one experiment and not the other, yet the difference between $\Delta\hat{\mu}$ between the two experiments is not statistically significant. Here, by statistically significant I mean the p-value is < 0.05.

Exercise 7 (Swapping response and predictor variables)**:** Consider the linear regression model

$$X \sim \text{Normal}(\mu_x, \sigma_x^2)$$
$$Y|X \sim \text{Normal}(\beta_1 X + \beta_0, \sigma_\epsilon^2)$$

This is a regression model for Y. The goal of this problem is to understand the distribution of X conditioned on Y. That is, we would like to understand the corresponding regression model for X. This is important in practice and it will also sharpen you understanding of what the covariance really means.

For some additional motivation, suppose that there is no noise in $Y|X$ (meaning $\sigma_\epsilon^2 = 0$). Then

$$Y = \beta_1 X + \beta_0 \implies X = \frac{1}{\beta_1} Y - \frac{\beta_0}{\beta_1}$$

so the slope of X vs. Y is $1/\beta_1$. We could try adding a normal random variable $Z \sim \text{Normal}(0, \sigma_\epsilon^2)$ to represent the noise in $Y|X$ and then solve this again. This would lead us to to

$$Y = \beta_1 X + \beta_0 + Z \implies X = \frac{1}{\beta_1} Y - \frac{\beta_0}{\beta_1} + \frac{Z}{\beta_1}$$

It is tempting to conclude that $Y|X$ follows a Normal distribution with mean $Y/\beta_1 - \beta_0/\beta_1$ and variance $\sigma_\epsilon^2/\beta_1^2$. This is however false – see part (c). In this problem you will derive the correct formula.

(a) Based on the formula for covariance derived in class, we know

$$\text{cov}(X, Y) = \beta_1' \sigma_y^2.$$

where β_1' is the regression slope on X vs. Y and σ_y^2 is the marginal variance of Y. eUsing(1) $\text{cov}(X, Y) = \text{cov}(Y, X)$ (interchanging the role of X and Y doesn't change the covariance) and (2) the marginal variance of Y is $\sigma_y^2 = \beta_1^2 \sigma_x^2 + \sigma_\epsilon^2$, derive a formula for β_1'.

(b) Using the result of part (a), show that when $\sigma_\epsilon^2 \to 0$ we retrieve the "naive" formula $\beta_1' = 1/\beta_1$.

(c) Why is the naive formula $1/\beta_1$ incorrect when $\sigma_\epsilon^2 > 0$? In particular, why can't we simply solve for X in terms of Y to obtain the regression equation? Hint: does $Y|Z$ have the same distribution as Y?