Caution: These lecture notes are under construction. You may find parts that are incomplete.

1 Regression to the Mean

If you have ever sorted a player leaderboard by a rate stat (e.g. on-base percentage in baseball, field goal percentage in basketball, yards per rush in football), you know the importance of filtering the rows on a minimum sample size threshold. If you do not perform this filtering, the top of your leaderboard is sure to be full with players who have performed incredibly in tiny samples. These performances do not reflect the true talent of the player because performance is the combination of talent and luck. The smaller the sample size, the greater the role played by luck in determining the performance. In other words, the smaller the sample, the greater the noise. Informally, we may write:

$$\text{Performance} = \text{Talent} + \text{Luck}. \tag{1}$$

If a player is near the top of the leaderboard (high performance), this is evidence in favor of both high talent and high luck. The smaller the sample size, the more plausible it is to attribute the performance to exceptionally high luck, and the less evidence we have in favor of high talent.

The standard solution to this problem is to filter the leaderboard on some sample size threshold. This is unsatisfactory because it fully discounts performance below the sample size threshold and applys no discount to performance above the threshold. We would prefer a method that smoothly decreases the discount as the sample size increases. Enter *regression to the mean*, which involves padding the observed performance with some amount of weight on the league average (aka *population mean*). Using n to denote the sample size and c to denote the weight on the population mean, the formula for regression to mean is:

$$\frac{n \cdot (\text{Observed Performance}) + c \cdot (\text{Population Mean})}{n + c}. \tag{2}$$

But how do we choose c?

1.1 A Statistical Model for Performance

A rate stat is the average of repeated performance by an athlete. Let Y_1, Y_2, ..., Y_n be random variables representing the performance of the athlete in n repeated trials. We assume[1] that the repeated trials are *i.i.d.* (independent and identically distributed) with mean μ and variance σ^2. We can invoke[2] the Central Limit Theorem (CLT) to approximate the distribution of the average performance $\bar{Y}$ as:

$$\bar{Y} = \frac{Y_1 + Y_2 + ... + Y_n}{n} \sim \text{Normal}(\mu, \sigma^2/n). \tag{3}$$

We interpret μ as the *true talent* of the athlete, and σ^2 is the noise associated with using observed performance to measure true talent in a single trial. The objective of regression to the mean is estimating μ. In the field of statistics, there are broadly two schools of thought for how to do this:

1. The *Frequentist* approach is to think of μ as a fixed, unknown value. Under this paradigm, the most logical estimate for μ is the value that would be most likely to have generated the observed data $\bar{Y}$. In this case, that value is $\bar{Y}$, so the Frequentist estimate is $\hat{\mu} = \bar{Y}$. This is known as the *maximum likelihood estimate* (MLE).

[1] This amounts to assuming that the expected performance on each trial is the same (regardless of competition faced and other context) and that the outcome of one trial does not affect the outcome of another trial. This assumption is not literally true in real life, and the implications of this are worth discussing. A story for another day.

[2] The CLT states that as the sample size increases toward infinity, the sampling distribution of $\bar{Y}$ converges to a normal distribution. We never have infinite sample sizes, but as a rule of thumb, this approximation works well for $n \geq 30$ (depending on multiple factors). Another story for another day.

2. The *Bayesian* approach is to think of μ as an uknown random variable itself. This requires that we specify a probability distribution for μ before observing any data. This is known as the *prior distribution*. Using the prior distribution and the likelihood of the observed data, we use Bayes' Rule to calculate the *posterior distribution* of μ, i.e. a distribution that represents our uncertainty about μ after observing the data.

For the problem of estimating athlete true talent, the Bayesian approach has advantages over the Frequentist approach. For example, consider a basketball player who attempts five 3-point field goals and makes all of them, a 3-point field goal percentage of 100%. Intuitively, 100% (the Frequentist estimate) is an absurd estimate for the player's true talent, especially based on such a small sample. The player would have to maintain that performance over a much larger sample to convince us that their true talent is close to a 100% field goal percentage. This intuition is exactly what the Bayesian approach captures.

1.2 The Bayesian Posterior Distribution

As noted in the previous section, the Bayesian approach requires that we specify a prior distribution on the athlete's true talent μ. Fortunately, there is a natural choice for this prior: the distribution of true talent across a population of relevant population of athletes (e.g. all players in the same league). This corresponds with our intuitive skepticism that an outlier performance represents an outlier talent. The prior distribution is a mathematical formalization of our intuition that outlier true talent is rare.

Let us assume that the population distribution of true talent is normal with mean μ_0 and variance σ_0^2. We use this as our prior distribution for μ:

$$\mu \sim \text{Normal}(\mu_0, \sigma_0^2).$$

Suppose we observe the value $\bar{y}$ for the random variable $\bar{Y}$. Using Bayes' Rule[3], we can combine the prior distribution with the likelihood from equation (3) calculate the posterior distribution of μ given $\bar{Y} = \bar{y}$:

$$\mu \mid \bar{Y} = \bar{y} \sim \text{Normal}\left(\frac{(n/\sigma^2)\bar{y} + (1/\sigma_0^2)\mu_0}{(n/\sigma^2) + 1/\sigma_0^2}, \frac{1}{n/\sigma^2 + 1/\sigma_0^2}\right). \tag{4}$$

The mean of this posterior distribution is our formula for regression to the mean. It is the weighted average of the observed performance ($\bar{y}$) and the population mean (μ_0), and each is weighted according to the inverse of its variance. We make the following observations:

1. As the population variance (σ_0^2) decreases, we put more weight on the population mean (μ_0) because we have greater confidence that all players are close to the population mean.

2. As the noise variance (σ^2/n) decreases, we put more weight on the observed performance ($\bar{y}$) because we have greater confidence that the observed performance reflects true talent.

3. Our posterior mean matches equation (2) with $c = \sigma^2/\sigma_0^2$.

We get $\bar{y}$ and n from our data, but we still need σ^2, μ_0 and σ_0^2 to perform regression to the mean. We can estimate σ^2 using the standard unbiased estimator for variance based on a sample of size n:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}. \tag{5}$$

Estimating the population mean μ_0 is similarly easy: We calculate the average performance across all players in the league and use this as an estimate of the population mean true talent (denoted by $\hat{\mu}_0$). Because it is based on a league's worth of data, $\hat{\mu}_0$ is very low-noise. This method of estimating the parameter of a prior distribution using observed data is called *empirical Bayes*. We use empirical Bayes to estimate σ_0^2 as well.

[3]We omit the details of this calculation. You can learn this in STAT 425.

1.3 Discussion Questions

1. Suppose that, at the beginning of class every day, each of us flips a coin. At the end of the semester, we each calculate our percentage of coin flips that came up heads and call this $\bar{Y}$. If we were to apply equation (4) in this setting, what would the value be for σ_0^2? σ^2? μ_0? n? $\bar{y}$?

2. Suppose that, at the beginning of class every day, each of us measures our height in inches. At the end of the semester, we each calculate the average of our height measurements and call this $\bar{Y}$. If we were to apply equation (4) in this setting, what the value be for σ_0^2? σ^2? μ_0? n? $\bar{y}$?

3. Suppose that the true talent of rushing yards per attempt in a football league is normally distributed with mean 4 yards per attempt and standard deviation 1 yard per attempt, and suppose that the yardage of each rushing attempt is a random variable with a mean equal to the running back's true talent and a standard deviation of 5 yards. How many rushing attempts does it take before a player's observed yards per attempt is a better predictor of their future yards per attempt than simply using the league average as the predictor?