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

5 Regularized Regression

One shortcoming of the plus-minus models from the previous chapter is that we can get very noisy strength estimates for players with small samples. Even for large-sample players, it may be the case that they are generally active at the same time as another player. If we have a small sample for which only one of those two players is active, then both players may have very noisy strength estimates. We face the same problem we faced in Chapter ??, for which we turned to regression to the mean.

Recall the Bayesian model providing the theoretical justification for regression to the mean:

$$\begin{aligned} \bar{Y} &\sim \text{Normal}(\mu, \sigma^2/n) \\ \mu &\sim \text{Normal}(\mu_0, \sigma_0^2). \end{aligned} \tag{1}$$

Estimating player strength is a matter of estimating μ. This model reduces the noise in our estimate of μ by treating it as a random variable (as opposed to a fixed, unknown parameter) and specifying a prior distribution for it. In the preceding chapters, estimating player/team strength was a matter of estimating the regression coefficients $\boldsymbol{\beta}$ in a linear model. In this chapter, we mirror the approach of regression to the mean to reduce the noise in our estimates of $\boldsymbol{\beta}$.

5.1 Linear Mixed-Effects Regression

Let us revisit the Rasch model from Chapter ??. We write a modified version of the model below by treating the regression coefficients as random variables (by specifying their distribution), rather than fixed and unknown parameters. The random-effect Rasch model is:

$$\begin{aligned} Y_i &\sim \text{Normal}(\beta_0 + \beta_{o_i}^O + \beta_{d_i}^D, \sigma^2) \\ \beta_j^O &\sim \text{Normal}(0, \sigma_O^2) \\ \beta_j^D &\sim \text{Normal}(0, \sigma_D^2). \end{aligned}$$

The regression coefficients β_j^O and β_j^D $(j = 1, 2, ..., p)$ are called *random effects* (as opposed to *fixed effects*, which we used in the original Rasch model). The random-effect Rasch model does not have the same identifiability problem as the original Rasch model because we have specified that the random effects have mean zero. The key to fitting this model is estimating the two new variance parameters σ_O^2 and σ_D^2.

The parameter σ_O^2 describes the spread of the offensive random effects. The smaller σ_O^2 is, the more our estimates of offensive team strengths will be shrunken toward zero. The same relationship holds for σ_D^2 and defensive team strengths. Recall that for regression to the mean, the most involved step was estimating the population variance parameter σ_0^2. The parameters σ_O^2 and σ_D^2 are analogous to σ_0^2.

If we treat the variance parameters σ_O^2 and σ_D^2 as fixed and unknown, then this model would be an example of the Frequentist model known as *linear mixed-effects regression* (LMER). We call it mixed-effects regression because the regression coefficients may be fixed effects or random effects. The standard method for estimating the variance parameters in LMER is *restricted maximum likelihood* (REML), the details of which are beyond the scope of this course.[1] For our purposes, we will use the lme4[2] package in R.

From an intuition perspective, REML works by finding the values of σ_O^2, σ_D^2 and σ^2 that best explain the observed within-player variance and between-player variance in outcomes within the dataset on which the model is fit. In that sense, the method is similar to how we estimate the popluation variance for regression to the mean in Chapter ??. In the next section, we see an alternative paradigm for determining the amount of shrinkage in our regression coefficients.

[1] You can learn this in STAT 410.

[2] Bates et al. (2015) "Fitting linear mixed-effects models using lme4" *Journal of Statistical Software* 67 (1-48)

5.2 Ridge Regression

One alternative to LMER for reining in noisy regression coefficients is *regularization*. Let us revisit the Adjusted Plus-Minus model from Chapter ??:

$$\eta_i = \beta_0 + \sum_{j \in H_i} \beta_j - \sum_{j' \in A_i} \beta_{j'}$$
$$Y_i \sim \text{Normal}(\eta_i, \sigma^2 / w_i).$$

Regularization keeps the same model but changes the criterion for estimating $\boldsymbol{\beta}$, instead of least squares. The modified objective is to minimize the sum of squared residuals *plus* a penalty on the regression coefficients. One example of regularization, *ridge regression* penalizes the sum of squared coefficients:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} w_i (y_i - (\beta_0 + \beta_{h_i} - \beta_{a_i}))^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{2}$$

Using ridge regression to estimate the Adjusted Plus-Minus model is known as *Regularized Adjusted Plus-Minus* (RAPM) and is the standard of practice for estimating player strength in basketball and some other sports. Note that, by using ridge regression instead of least squares, we no longer have an identifiability problem in fitting the model. One can show that any $\boldsymbol{\beta}$ satisfying equation (2) will also satisfy $\sum_j \beta_j = 0$, meaning that the estimated regression coefficients will be mean zero.

Here, $\lambda > 0$ is a tuning parameter (to be chosen by the person fitting the model) that controls the amount of shrinkage in the estimated regression coefficients. If $\lambda = 0$, then equation (2) is equivalent to weighted least squares (WLS), and no shrinkage occurs. As λ increases, we put more weight on $\boldsymbol{\beta}$ being close to zero (relative to the importance of explaining the observed data), and we get more shrinkage. In this way, λ corresponds to σ_O^2 and σ_D^2 in the previous section. These parameters control the amount of shrinkage in the estimated regression coefficients.

Fortunately, there is a closed-form solution to the optimization problem in equation (2). Introducing the penalty on $\boldsymbol{\beta}$ has not made our objective any less convex or differentiable, so the method for deriving this solution is (as with OLS) finding where the gradient is zero.[3] Keeping the same matrix notation from the previous chapter, we need only to introduce $\mathbf{I}$ to denote the $p \times p$ identity matrix. The solution is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}.$$

Because λ controls the amount of shrinkage in the model, choosing the right λ is important (analogous to estimating the variance parameters in LMER). Ridge regression comes from the field of statistical machine learning, where the ethos is somewhat different from traditional statistical modeling. In machine learning, we concern ourselves with the predictive accuracy of the model. We might split our data into a training set and a validation set, using the training set to fit the model for many different values of λ and choosing the value of λ that minimizes the sum of squared errors on the validation set. The most popular method for choosing λ is *cross-validation*,[4] which is a generalization of the training/validation split described here.

Intuitively, LMER and ridge regression are very similar. Both methods shrink the regression coefficients toward zero. The primary difference between the two methods is how they determine the amount of shrinkage.

5.3 Regression to the Mean as Regularized Regression

We now come full circle by describing how regression to the mean, introduced in Chapter ??, is a special case of regularized regression. We start by modifying the notation used to describe the model underlying regression to the mean, from equation (1), to make it look more like the linear models we have been using. Assume we observe n outcomes y_i, numbered $i = 1, 2, ..., n$. For each outcome, we observe $p_i \in \{1, ..., p\}$, the player who produced the outcome. We use the random variable Y_i represent the probability distribution

[3] We omit the details of this calculation. You can learn this in STAT 413.

[4] You can learn more about this in STAT 413.

generating the outcome y_i, which we model as:

$$\begin{aligned}\eta_i &= \beta_0 + \beta_{p_i} \\ Y_i &\sim \text{Normal}(\beta_0 + \beta_{p_i}, \sigma^2)\end{aligned}$$

This model follows the same structure as the linear regression models we have been using. As it is written, the model is not identifiable, but if we constrain $\beta_0 = 0$, we can show that the OLS solution is:

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} \mathbb{I}(p_i = j) \cdot y_i}{\sum_{i=1}^{n} \mathbb{I}(p_i = j)} = \bar{y}_j.$$

In other words, the estimated regression coefficient for each player is exactly equal to their average performance. In this case, the regression model is simply calculating the average for each player.

To achieve regression to the mean, we need to introduce some shrinkage on the regression coefficients. In this chapter, we have learned two options for doing so. The first option is to treat the regression coefficients as random effects, as in LMER. Specifically, we would write:

$$\begin{aligned}\eta_i &= \beta_0 + \beta_{p_i} \\ Y_i &\sim \text{Normal}(\eta_i, \sigma^2) \\ \beta_j &\sim \text{Normal}(0, \sigma_\beta^2).\end{aligned}$$

In this case we would use REML to estimate the model. The second option is to keep the model as-is but use ridge regression to estimate the model:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_{p_i}))^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

In this case we would use cross-validation or some other method to choose λ, at which point we have a closed-form expression for the estimated regression coefficients. Given the existence of software packages to implement LMER and ridge regression, framing regression to the mean this way may lead to easier implementation than the method we derived in Chapter ??.

In conclusion, we have covered in this chapter two different methods for reducing the noise in our estimates of team and player strengths. These methods are particularly important for players with small samples. Although we have introduced two very specific models in the random-effects Rasch model and RAPM, the concepts of LMER and ridge regression are broadly applicable. For example, we can combine LMER or ridge regression with logistic regression to estimate player strengths pertaining to binary outcomes (as opposed to normally distributed outcomes). The right combination of these techniques depends on the details of the sports outcome being modeled.