Predicting Salary of Major League Baseball Players

Stephen Su, Thomas Black; MAST90111

1 Introduction

We are interested in predicting the salary (in $,000, log) of major league baseball players by their cumulated number of hits and number of years played in the major league. 263 non-missing observations of major league players from the 1986 and 1987 seasons are recorded in a dataset (James et al., 2013). We attempt to build predictive models using nonparametric and semiparametric methods thus compare the predictive performance with their conventional parametric counterparts.

2 Main procedure

A random sample without replacement of 200 observations are picked from the dataset which serve as the training set. The remaining observations serve as the test set. Each non/semiparametric or parametric model will be fitted via the training set; its predictive performance is evaluated against the test set.

We assume the variables are approximately continuous on $\mathbb{R}$ as salary is log-transformed, and variables hits and years are mostly much greater than zero. The boundary effect around zero is negligible. Also, we assume the conditions outlined in lecture hold for both the training set and test set and for all variables.

3 Univariate prediction

Suppose we only observe a random sample of the salary variable and wish to predict the salary of another randomly selected player, the best (in mean quadratic error) predictor is simply the sample (arithmetic) mean. Instead, we model the density f using some estimator $\hat{f}$, which allows the generation of a (central) $(1-\alpha)$ confidence interval of prediction by

$$\left\{[a,b] : \int_{-\infty}^{a} \hat{f}(x)dx = \frac{\alpha}{2}, \int_{-\infty}^{b} \hat{f}(x)dx = 1 - \frac{\alpha}{2}\right\}$$

3.1 Prediction with Normality assumption

The common parametric approach is to assume the log salary follows a $\mathrm{N}(\mu, \sigma^2)$ distribution for some (μ, σ^2) which are unknown. Under such an assumption, the sample mean is the unbiased, efficient and consistent estimator of μ. However, there are various estimators of σ^2 under different metrics, examples are

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{1}{n-1}\|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2; \hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2; \hat{\sigma}^2_{\text{minMSE}} = \frac{1}{n+1}\|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2$$

The convention is to estimate σ^2 with unbiased sample variance, $\hat{\sigma}^2_{\text{unbiased}}$. As such, the central $(1-\alpha)$ confidence interval of prediction is given by

$$G_\alpha = \left[\bar{Y} + \phi\left(\frac{\alpha}{2}\right)\frac{1}{n-1}\|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2, \bar{Y} + \phi\left(1 - \frac{\alpha}{2}\right)\frac{1}{n-1}\|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2\right]$$

If the assumption holds, $\overline{\mathbf{1}(Y_{n+1} \in G_\alpha)} \to 1 - \alpha$ in probability as $n \to \infty$.

3.2 Prediction with kernel density estimation

Instead of assuming f is a Gaussian density, which is not necessarily true, we estimate f nonparametrically with a kernel density estimator given by

$$\hat{f} : \mathbb{R} \to [0, \infty), x \mapsto \hat{f}(x) \triangleq \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - Y_i}{h}\right)$$

for some $h > 0$, with $K \triangleq \phi$ being the Gaussian kernel. A $(1 - \alpha)$ confidence interval is thus obtainable.

The bandwidth h is selected by (1) the rule-of-thumb method: $\hat{h} = 1.06n^{-0.2} \min\{\hat{\sigma}_{\text{unbiased}}, \text{IQR}/1.34\}$, and (2) cross-validation: $\hat{h} = \arg\min_{h>0} \text{LSCV}(h)$.

3.3 Evaluation of univariate prediction methods

For both the parametric (Normality) and nonparametric (KDE) methods for constructing confidence intervals, we expect approximately $1 - \alpha$ proportion of the test observations to be included inside the interval. As such, for each of the mentioned method, we construct a (central) 50% confidence interval and perform the following hypothesis test:

$$H_0 : p = 0.5; H_1 : p \neq 0.5$$

Here p is the probability of a random test observation being included in the confidence interval constructed using the training set. An exact Binomial test (Clopper & Pearson, 1934) is performed. We find that, 35% (95% CI: $[0.23, 0.48]$, p-value: 0.0226) of test observations are included in the parametric confidence interval, c.f. 44% (95% CI: $[0.32, 0.58]$, p-value: 0.45) included in the nonparametric KDE confidence intervals (same for both ROT/LSCV methods). The nonparametric method is evidently more robust since less assumptions.

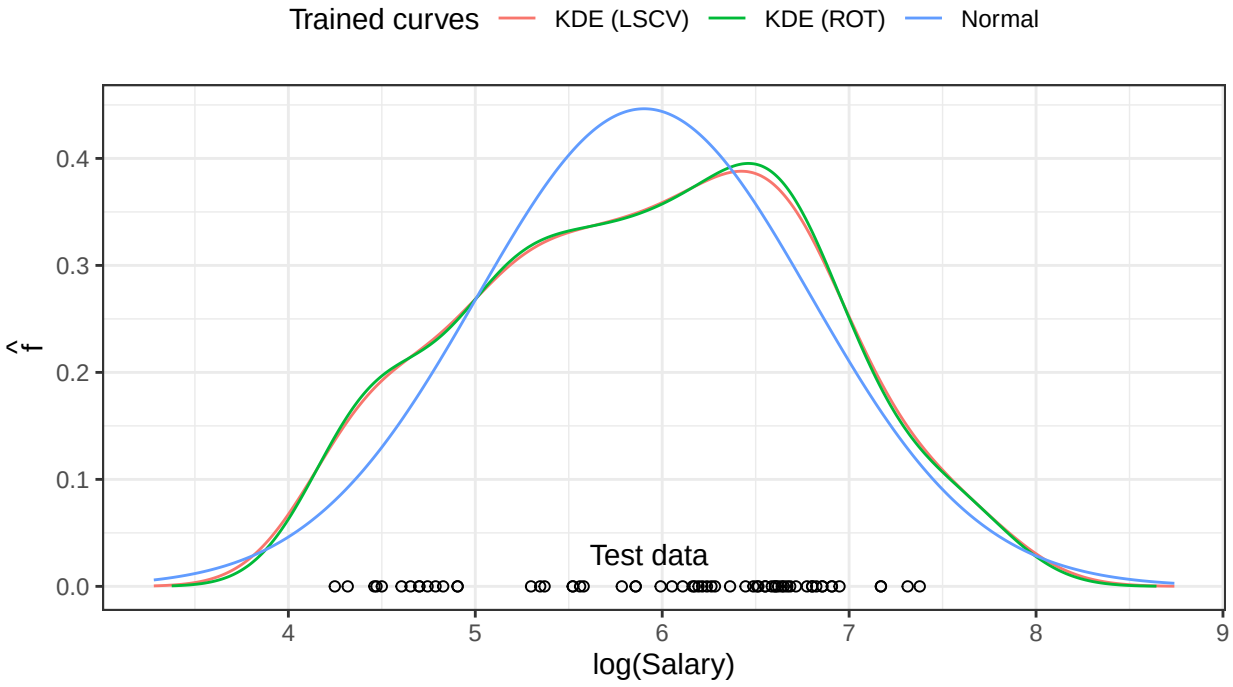

Figure 1: The density curves are above are estimated using the training set, while the data points are from the test set. We can see that the KDE curves approximate the distribution of the test data better.

Nevertheless, the realisation of test performances above subject to the variability of the training/test sampling process. Therefore, we repeat the sampling, fitting and evaluation process above $m = 100$ times.

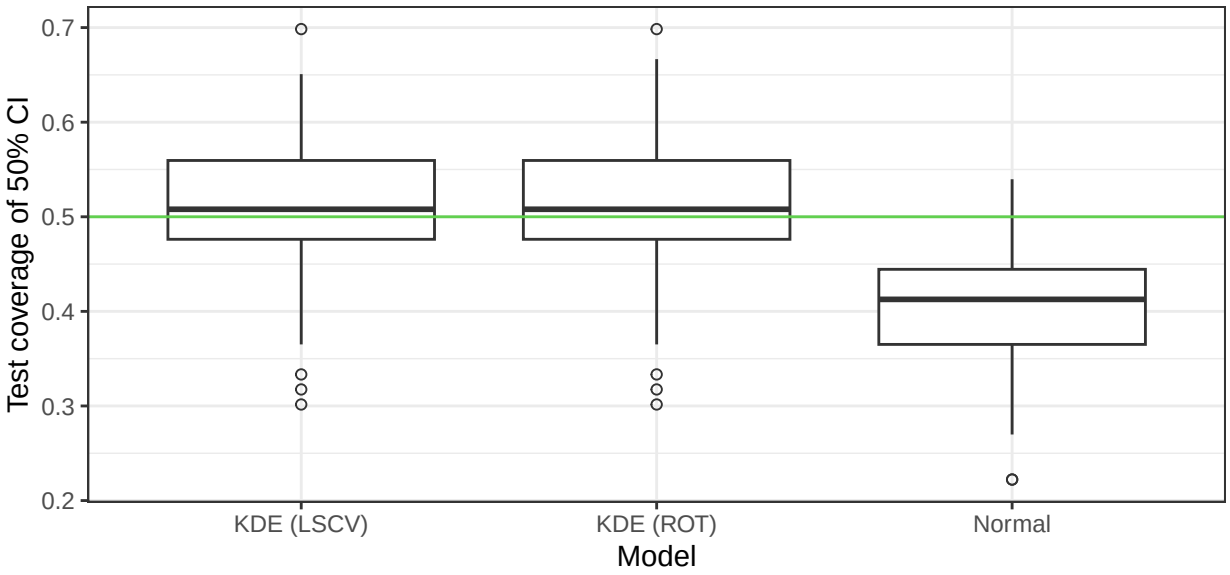

Figure 2: Both the nonparametric KDE confidence intervals consistently capture about 50% of the test data.

References

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*(4), 404–413. https://doi.org/10.2307/2331986

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*. Springer.