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

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

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

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

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.

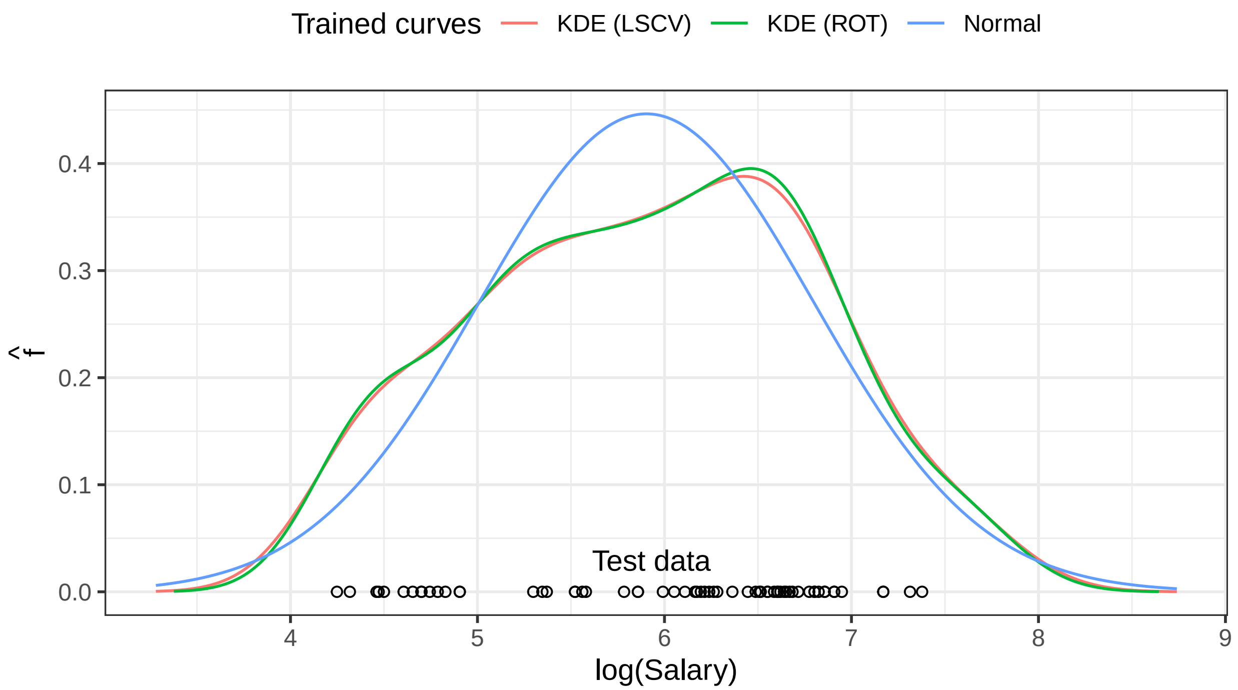

Figure 1: The density curves are above are estimated using the training set, while the data points are from the test set. We can see that the KDE curves approximate the distribution of the test data better.

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.

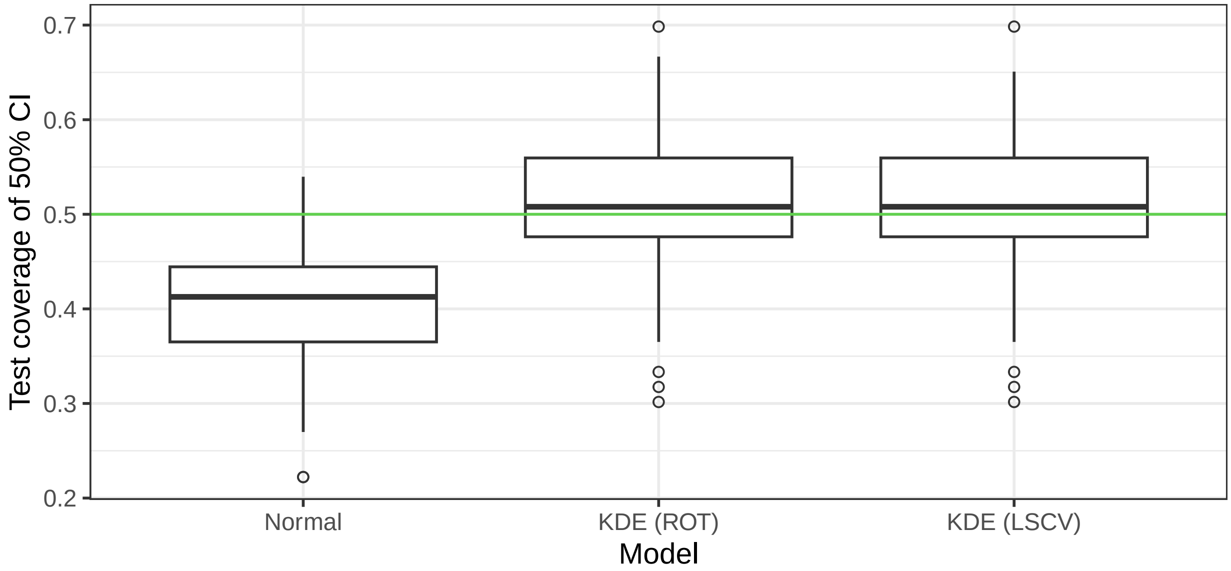

Figure 2: Both the nonparametric KDE confidence intervals consistently capture about 50% of the test data.

4 Prediction via bivariate regression

Suppose, in addition, we observe an additional covariate, the number of cumulative hits: X_1. Then the best predictor becomes the conditional expectation $\mathbb{E}[Y|X_1]$ (if $Y \perp X_1$, then $\mathbb{E}[Y|X_1] = \mathbb{E}[Y]$). A regression $m(x) = \mathbb{E}[Y|X_1 = x]$ is hence used as the predictor for some estimator $\hat{m}$ of m, via both parametric and nonparametric methods fitted with the training set.

4.1 Simple linear regression

As the most ubiquitous parametric regression model, a linear regression assumes the conditional distribution $Y|X_1 = x \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$, hence $m(x) = \beta_0 + \beta_1 x$. Therefore, the predictor of Y given that $X_1 = x$ is

$$\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x; \left(\hat{\beta}_0, \hat{\beta}_1\right)^\top = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}$$

whereby, $\{(\mathbf{X}, \mathbf{Y}) : \mathbf{X} = (\mathbf{1}_n, \mathbf{X}_1)\}$ is the training data set.

4.2 Local linear regression

The aforementioned Normality assumption does not hold in most cases and we therefore need more flexible methods robust against a departure from Normality. The local linear regression (local polynomial, with order $p = 1$) belongs to the class of kernel regression models, that appeals to the Taylor expansion $m(x) = m(x_0) + m'(x_0)(x - x_0)$, with a fitting criterion minimising local weighted least square. The fit is given by

$$\hat{m}(x) = \mathbf{e}_1^\top \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Y}$$

where $\mathbf{X} \triangleq ((X_i - x)^{j-1})_{i=1,j=1}^{n,p+1}$ and $\mathbf{W} \triangleq \mathrm{diag}(K_h(x - X_i) : i = 1, \ldots, n)$ (note: all from training data). The bandwidth h is selected using cross-validation. The fit becomes all linear if $h \to \infty$.

4.3 Penalised cubic spline regression

Spline is yet another nonparametric method that exploits the ability of piecewise polynomials to approximate any smooth function. The penalised cubic spline is among one of the most commonly used ones, by minimising the L^2-loss penalised by integrated square second derivative of fit. The fit is given by

$$\hat{m}_\lambda(x) = \mathbf{B}^\top(x)\left(\mathcal{B}^\top\mathcal{B} + \lambda\mathbf{D}_2\right)^{-1}\mathcal{B}^\top\mathbf{Y}$$

where $\mathbf{B}$ is the B-spline basis spanned by the range of X_1, $\mathcal{B} \triangleq (\mathbf{B}(X_1), \ldots, \mathbf{B}(X_n))^\top$; $\mathbf{D}_2 \triangleq \int \mathbf{B}''(x)\mathbf{B}''(x)^\top dx$. In addition, we constructed the basis with $L = \min\left\{\frac{n}{4}, 35\right\}$ interior knots as suggested by Ruppert (2002). The "smooth" parameter λ is selected via the numerically stable generalised cross-validation (GCV) by

$$\lambda \leftarrow \arg\min_{\lambda>0} \frac{n}{n - \operatorname{tr}\left(\mathcal{B}\left(\mathcal{B}^\top\mathcal{B} + \lambda\mathbf{D}_2\right)^{-1}\mathcal{B}^\top\mathbf{Y}\right)} \sum_{i=1}^{n}(Y_i - \hat{m}_\lambda(X_i))^2$$

All observed quantities used in fitting above comes from training data.

4.4 Evaluation of bivariate regression methods

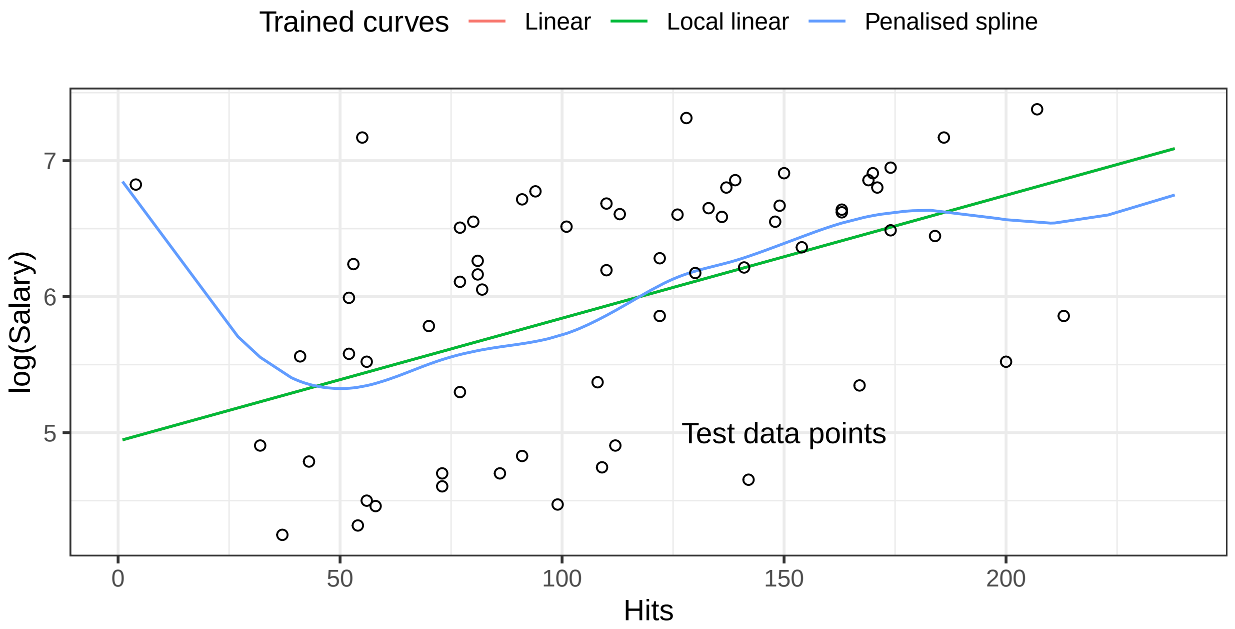

Figure 3: The trained curves c.f. the test data points.

Table 1: Predictive performance of the models.

model	Linear	Local linear	Penalised spline
Test MSEP	0.5995660	0.5995660	0.5407159

In this particular training/test sample, the local linear regression is indistinguishable from the linear regression model for a large selected bandwidth, implying that $m(x)$ is roughly linear. A rather interesting observation is that the spline model successfully captures the boundary effect (note that is not an overfit, as the points are from test data). Nevertheless, The results obtained above subjects to the variability of the sampling process of training/test data. Thus, the sampling, fitting and evaluation process is repeated $m = 100$ times.

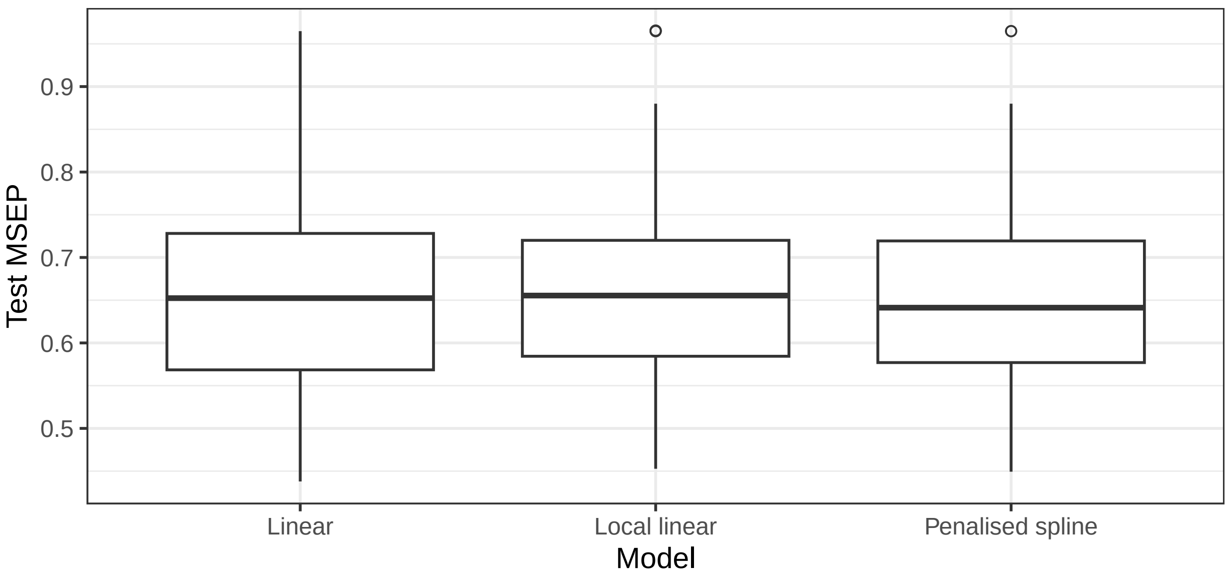

Figure 4: All three parametric and nonparametric models appear to possess similar predictive power under a likely linear dependency between the covariate and the response.

5 Prediction via multivariate regression

Suppose we observe $p > 1$ covariates as vector: $\mathbf{X} \triangleq (X_1, \ldots, X_p)$, thence the conditional expectation now has the form as: $m(\mathbf{x}) \triangleq \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. Therefore, we construct and compare the parametric model and, to ensure generalisability to a larger p, semiparametric models (hence omitting multivariate nonparametric models).

We demonstratively observe another covariate X_2: the number of years each player served in the major league, hence test the predictive power of $\hat{m}(\mathbf{x})$ by different parametric/semiparametric models.

5.1 Multiple linear regression

The multiple linear regression is the higher-dimensional generalisation of simple linear regression, which assumes $Y|\mathbf{X} = \mathbf{x} \sim \mathrm{N}(\beta_0 + \beta_1 x_1 + \beta_2 x_2, \sigma^2)$, therefore, $m(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. The fit becomes

$$\hat{m}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2; \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right)^\top = \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \mathbf{Y}$$

and $\left\{(\mathbb{X}, \mathbf{Y}) : \mathbb{X} = \left(\mathbf{1}_n, (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top\right)\right\}$ is (again) the training data set.

5.2 Semiparametric regression models

The semiparametric models are all fitted using the **mgcv** package. In particular, the single-index model is initiated as an additive model to profile out the spline coefficient and smoothing parameter thus leaving only the linear coefficient $\boldsymbol{\beta}$ to be estimated via some Newton-based general purpose optimiser (Wood, 2011).

5.2.1 Single index model

Again, the conditional Normality assumption for linear regression above does not necessarily hold. Nevertheless, the computation efficiency and interpretability of parametric models remain appealing, especially when the covariate is in very high dimension. The single index model is thus proposed as a semiparametric regression method which uses nonparametric method to estimate some unknown link function (instead of the canonical link) while keeping the linear parametric terms, in form $m(\mathbf{x}) = g(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ for some continuous g.

We propose s, a penalised cubic regression spline, as an estimator of the link function g. The fit is given by

$$\hat{m}(\mathbf{x}) = s\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2\right)$$

For identifiability, we imposed constraints such that

$$\hat{\beta}_0 > 0; \sqrt{\hat{\beta}_0^2 + \hat{\beta}_1^2 + \hat{\beta}_2^2} = 1$$

5.2.2 Partial linear model

Nonetheless, we previously inferred that covariate X_1 influences the response in a linear way. An alternative semiparametric approach we may consider is the partial linear model with form $m(\mathbf{x}) = \beta_0 + \beta_1 x_1 + g(x_2)$. A penalised cubic regression spline s is used as the estimator of some unknown function g. The fit is given by

$$\hat{m}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + s(x_2)$$

The advantage of partial linear models is that it reduces the number of parameters (including those for fitting the splines) to be estimated by assuming linear correlation with response for some covariates.

5.2.3 Additive model

Another (popular) option is not to assume linearity for any covariates. In contrary to single index models, additive models is more flexible by allowing each covariate to have different “trend shapes” related to the response, with form $m(\mathbf{x}) = \beta_0 + g_1(x_1) + g_2(x_2)$ and fit hence given by

$$\hat{m}(\mathbf{x}) = \hat{\beta}_0 + s_1(x_1) + s_2(x_2)$$

where s_1, s_2 are penalised cubic regression splines to estimate the unknown functions g_1, g_2.

5.3 Evaluation of multivariate regression methods

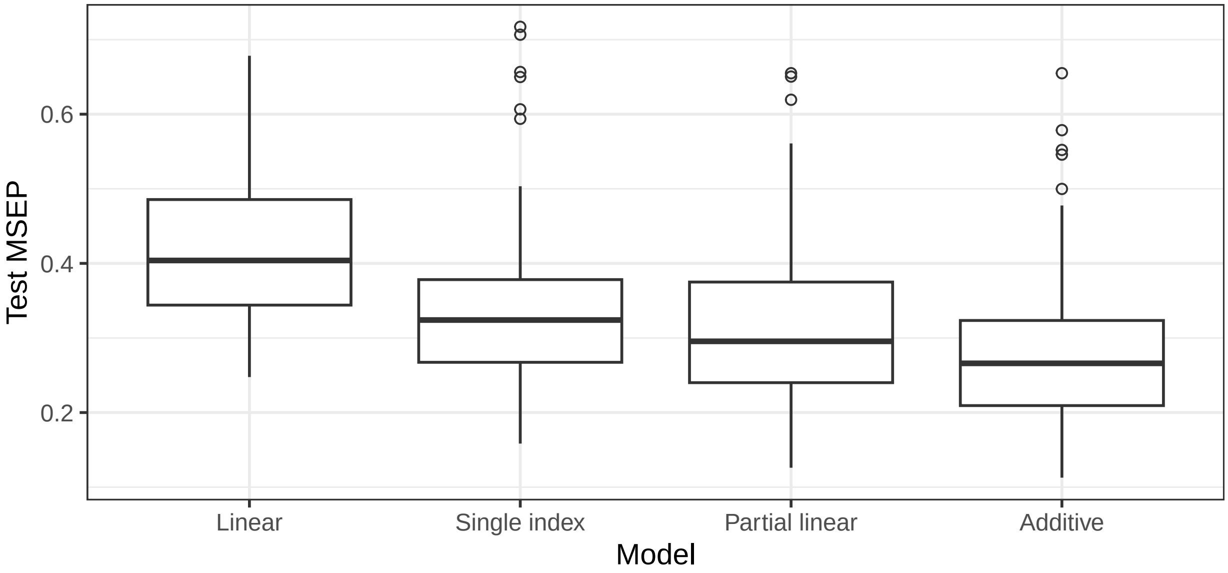

Figure 5: The predictive power of parametric and semiparametric regressions in 100 training/test samples.

6 Conclusion

Our main realisation from this project is that, under almost all cases, non(semi)parametric models are at least non-inferior to parametric ones under purely predictive (non-explanatory) contexts. In contrary to some prejudiced opinions (including myself) that non(semi)parametric methods often overfit linear dependency, we have shown that non(semi)parametric methods are robust even if the underlying relationship is about linear, not mentioning non(semi)parametric models excel at capturing non-linear relationships. It seems like the only major limiting factors for always using non(semi)parametric approaches for prediction are: (1) efficiency of computation of fitting non(semi)parametric model for very large data sets; (2) slower rate of convergence compared to parametric models thus often requires larger sample sizes to achieve satisfactory results.

7 Open and reproducible work

The source code for reproducing the report, slide and the output of the project and history of all edits are available at the GitHub repository szmsu2011/mast90111proj.

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.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, *11*(4), 735–757. https://doi.org/10.1198/106186002853

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x