Predicting Salary of Major League Baseball Players - Draft

Stephen Su, Thomas Black; MAST90111

Introduction

Data: James et al. (2013). Model and predict `log(Salary)` by `Hits` and `Years`.

200 training observations, 63 test observations.

```
data <- ISLR::Hitters |>
  mutate(lSalary = log(Salary)) |>
  select(lSalary, Hits, Years) |>
  drop_na()
set.seed(90111)
train <- slice_sample(data, n = 200)
test <- setdiff(data, train)
```

Model fitting and evaluation

Univariate Normal approximation

```
muhat <- mean(train$lSalary)
sigma2hat <- var(train$lSalary)
## Test coverage of 50% confidence interval of prediction
sum(abs(test$lSalary - muhat) < qnorm(.75) * sqrt(sigma2hat)) |>
  binom.test(63)
```

```
#>
#>  Exact binomial test
#>
#> data:  sum(abs(test$lSalary - muhat) < qnorm(0.75) * sqrt(sigma2hat)) and 63
#> number of successes = 22, number of trials = 63, p-value = 0.02257
#> alternative hypothesis: true probability of success is not equal to 0.5
#> 95 percent confidence interval:
#>  0.2333706 0.4797338
#> sample estimates:
#> probability of success
#>              0.3492063
```

Univariate kernel density estimation

```
kd_rot <- density(train$lSalary, bw = "nrd0") # ROT bandwidth selection
kd_cv <- density(train$lSalary, bw = "ucv") # LOOCV bandwidth selection
## Test coverage of 50% confidence interval of prediction
cdf_rot <- cumsum(kd_rot$y / sum(kd_rot$y))
cdf_cv <- cumsum(kd_cv$y / sum(kd_cv$y))
sum(between(
```

```r
  test$lSalary,
  kd_rot$x[which.min(abs(cdf_rot - .25))],
  kd_rot$x[which.min(abs(cdf_rot - .75))]
)) |>
  binom.test(63)
```

```
#>
#>  Exact binomial test
#>
#> data:  sum(between(test$lSalary, kd_rot$x[which.min(abs(cdf_rot - 0.25))], kd_rot$x[which.min(abs(cd
#> number of successes = 28, number of trials = 63, p-value = 0.45
#> alternative hypothesis: true probability of success is not equal to 0.5
#> 95 percent confidence interval:
#>  0.3191731 0.5751124
#> sample estimates:
#> probability of success
#>              0.4444444
```

```r
sum(between(
  test$lSalary,
  kd_cv$x[which.min(abs(cdf_cv - .25))],
  kd_cv$x[which.min(abs(cdf_cv - .75))]
)) |>
  binom.test(63)
```

```
#>
#>  Exact binomial test
#>
#> data:  sum(between(test$lSalary, kd_cv$x[which.min(abs(cdf_cv - 0.25))], kd_cv$x[which.min(abs(cdf_c
#> number of successes = 28, number of trials = 63, p-value = 0.45
#> alternative hypothesis: true probability of success is not equal to 0.5
#> 95 percent confidence interval:
#>  0.3191731 0.5751124
#> sample estimates:
#> probability of success
#>              0.4444444
```

References

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