# Statistical Learning Comparisons

*Kodyak*

**Problems**

1. Evaluate the methods of classification we have discussed using class-validation for the data at pages.uoregon.edu/~dlevin/DATA/ozo.txt. The response variance is HO, which indicates whether a given day had high ozone levels or not. Discuss the success or failures of these methods.

2. The standard errors in linear regression are based on the assumption that the errors are independent, have the same variance, and have a Normal distribution.

(a) Simulate data where the errors are not Normal, e.g. come from a t-distribution with 3 degrees of freedom. (The t-distribution is like the normal, but has larger probabilities of extreme data points.) (Use rt in R to simulate. Assume the model is y = 3 + 3x + error, where there are 100 x's spread between 0 and 3.)

(b) Fit a linear model, and compare the standard errors of the fitted coefficients to the standard errors estimated by the bootstrap method.

(c) Now simulate where the errors are correlated. (We will discuss in class how to do that.) Compare the standard errors reported by the regression output (for the coefficients) to the bootstrap-estimated standard errors.

. . . .

. . . .

1. We will examine the classification methods: Logistic Regression, LDA (Linear Discriminant Analysis), and $k-$Nearest Neighbors on the ozo dataset (classifying whether days had high or low ozone levels). We will compare the average error rate of cross-validated model fitting from each area, to compare relative success rates of the techniques. In particular, we optimize the error rate in $k-$Nearest Neighbors with respect to choice of neighbor size. For details, see the associated R markdown file.

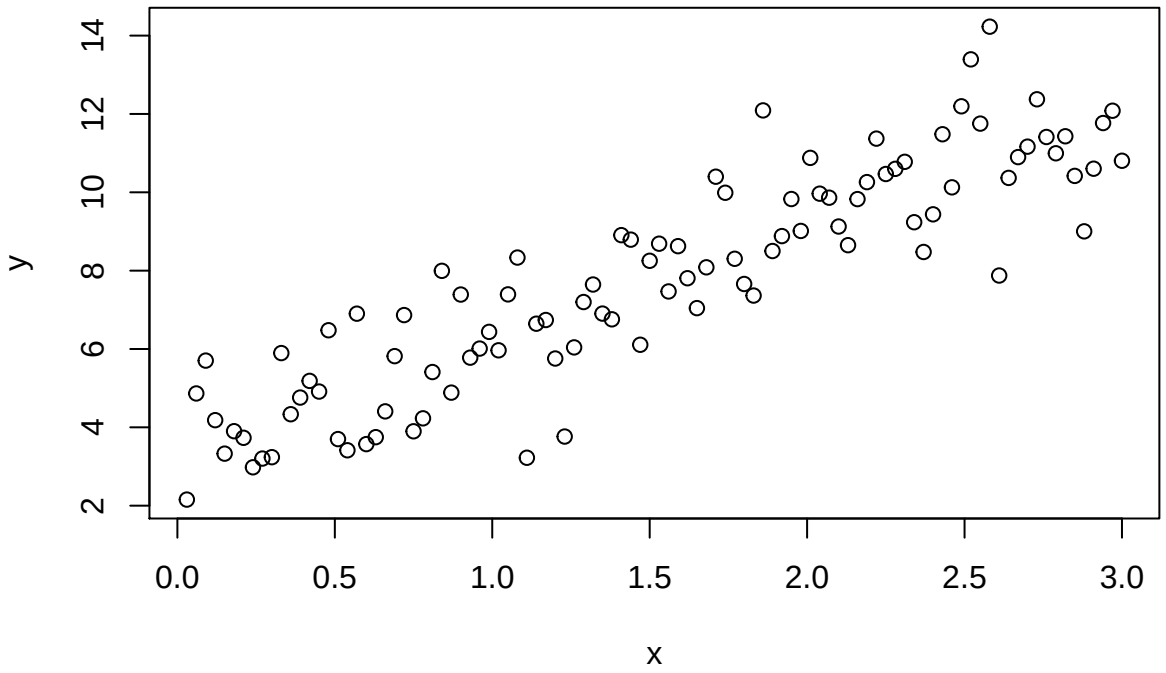

Fitting all 3 ml prediction models, we arrive at a 10-fold cross validated error rates:

### Error Comparison Table

| | Logistic Regression | LDA | k Nearest Neighbors |
|---|---|---|---|
| Average Error Rates | 0.120944741532977 | 0.135650623885918 | 0.150445632798574 |
| Optimal Parameters | | | k = 9 |

2.

(a) Simulate data where the errors are not Normal, e.g. come from a t-distribution with 3 degrees of freedom.(Assume the model is $y = 3 + 3x +$ error, where there are 100 x's spread between 0 and 3.)




Attached below is a plot of the simulated data from the model $y = 3 + 3x +$ error where the vector of error terms are instances drawn from a $t(3)$ distribution:

(b) Fit a linear model, and compare the standard errors of the fitted coefficients to the standard errors estimated by the bootstrap method.

We obtain the regression model

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.363652 | 0.2602120 | 12.92659 | 0 |
| x | 2.907182 | 0.1491155 | 19.49617 | 0 |

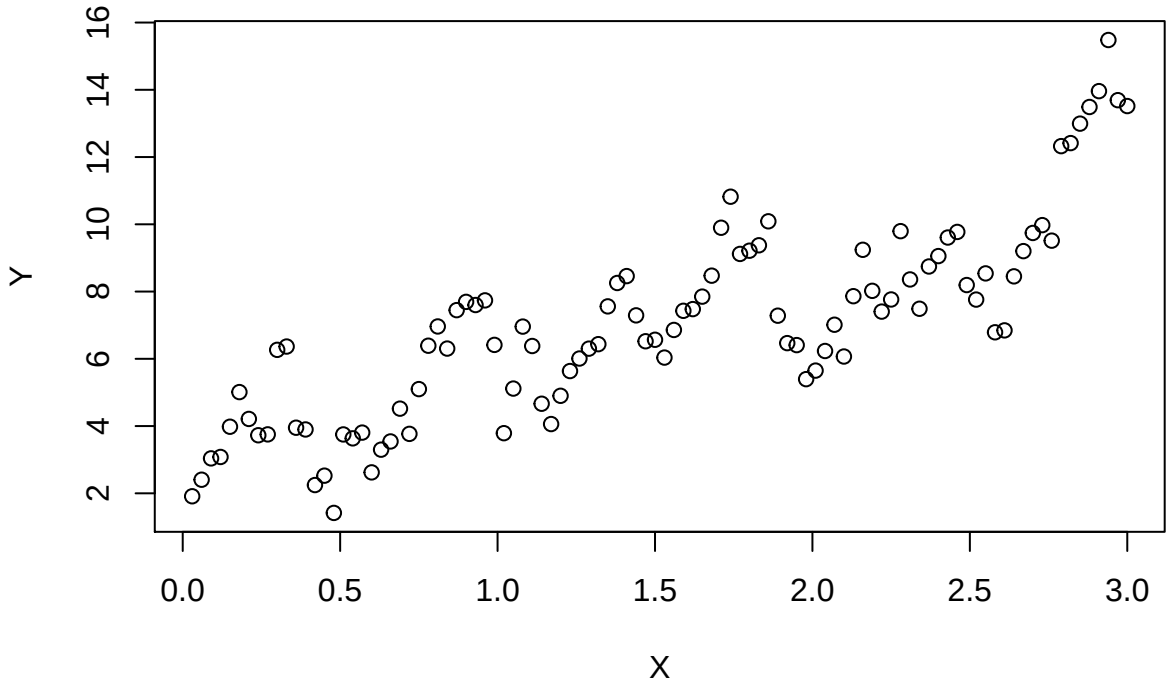
(with violated normality assumptions on the error) achieves a standard error of 0.1491155, and

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = xy, statistic = t.func, R = 100)
##
##
## Bootstrap Statistics :
##      original      bias    std. error
## t1* 0.2734684 0.001954778   0.0148688
```

shows that bootstraping achieves a standard error of .0148688.

(c) Now simulate where the errors are correlated. Compare the standard errors reported by the regression output (for the coefficients) to the bootstrap-estimated standard errors.

We can simulate data $y = 3 + 3x + \epsilon$ similar to as discussed before, where the errors $\epsilon$ are correlated linearly: $\epsilon_1 = \delta_1$ follows a normal distribution about 0, and similarly $\delta_i \sim N(0, 1)$, but $\epsilon_{i+1} = \rho\epsilon_i + \delta_i$ (we pick $\rho = .79$ in the model below).




We obtain the regression model

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 2.849995 | 0.3284760 | 8.676417 | 0 |
| X | 2.755107 | 0.1882346 | 14.636564 | 0 |

(with violated normality assumptions on the error) achieves a standard error of 0.1882346, and

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = XY, statistic = t.func, R = 100)
##
##
## Bootstrap Statistics :
##      original       bias    std. error
## t1* 0.2490385 0.0009752041  0.01323802
```

shows that bootstraping achieves a standard error of 0.01323802.