HW6 q2

10/7/2019

a) predicting arrival delay

We are interested in predicting arrival delay (ARR_DELAY) using the variables that would be available before the flight takes off: departure time, departure airport (ORD or DFW), day of the month, and air carrier (airline). Also, make a "day of week" variable. First, for a baseline comparison, fit a linear model to the data using these variables. (Be naive: assume the effects of time of day or day of week are linear.) Show a table of the coefficients you find. Evaluate the model on the test set (using predict) and report its squared-error loss.

Coefficients are given in Table 1. We see that the mean squared error is 2820.736.

```
# fit on the training data
best.fit=lm(ARR_DELAY~ DEP_TIME +ORIGIN + DAY_OF_MONTH + CARRIER + DAY_OF_WEEK, data=train)
#coef(best.fit, 8)
stargazer(best.fit,
          title = "Q2a",
          header=FALSE, type = "latex", font.size="small",
          column.sep.width = "1pt",
          single.row = TRUE)
```

Table 1: Q2a

	Dependent variable:
	ARR_DELAY
DEP_TIME	0.015^{***} (0.001)
ORIGINORD	5.033^{***} (0.914)
DAY_OF_MONTH	-0.021 (0.043)
CARRIERAS	-5.150 (4.196)
CARRIERB6	1.547 (4.964)
CARRIERDL	-6.439^{***} (2.407)
CARRIEREV	3.838^{***} (1.165)
CARRIERF9	36.971^{***} (3.971)
CARRIERNK	0.969 (2.057)
CARRIEROO	7.317^{***} (1.431)
CARRIERUA	4.092^{***} (1.215)
CARRIERVX	17.159^{***} (5.718)
DAY_OF_WEEK	1.197^{***} (0.194)
Constant	-16.759^{***} (1.571)
Observations	16,639
R^2	0.036
Adjusted R^2	0.036
Residual Std. Error	49.628 (df = 16625)
F Statistic	48.222^{***} (df = 13; 16625)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
# squared-error loss
mean((test$ARR_DELAY - predict.lm(best.fit, newdata = test))^2, na.rm = TRUE)
```

[1] 2820.736

b)

The relationships may be complicated and a linear model may not be appropriate, so use npreg to fit a Nadaraya–Watson kernel regression model; allow npregbw to select all bandwidths with cross-validation. Report the kernel regression's performance on the test set (again using predict and squared-error loss) and compare to the linear model. Does this method seem to do dra- matically better?

Coefficients are given in Table 2. We see that the mean squared error is 3309.925. This method does not seem to do better based on mse.

```
bw =npregbw( ARR_DELAY~ DEP_TIME +ORIGIN + DAY_OF_MONTH + CARRIER + DAY_OF_WEEK,
             data=train,
             subset = 200:500)
```

Multistart 1 of 5 | Multistart 1 of 5 | Multistart 1 of 5 | Multistart 1 of 5 / Multistart 1 of 5 - Multistart 1 of 5
Multistart 1 of 5 | Multistart 1 of 5 | Multistart 1 of 5 / Multistart 1 of 5 - Multistart 1 of 5
Multistart 1 of 5 | Multistart 2 of 5 | Multistart 2 of 5 | Multistart 2 of 5 / Multistart 2 of 5 - Multistart 2 of 5
Multistart 2 of 5 | Multistart 2 of 5 | Multistart 2 of 5 | Multistart 3 of 5 | Multistart 3 of 5 | Multistart 3 of 5 / Multistart 3 of 5 - Multistart 3 of 5 | Multistart 3 of 5 | Multistart 3 of 5 / Multistart 3 of 5 - Multistart 4 of 5 | Multistart 4 of 5 | Multistart 4 of 5 / Multistart 4 of 5 - Multistart 4 of 5 | Multistart 4 of 5 | Multistart 5 of 5 | Multistart 5 of 5 | Multistart 5 of 5 / Multistart 5 of 5 - Multistart 5 of 5
Multistart 5 of 5 | Multistart 5 of 5 | Multistart 5 of 5 |

```
loclinfit <- npreg(bw)
loclinfit
```

Regression Data: 291 training points, in 5 variable(s) DEP_TIME ORIGIN DAY_OF_MONTH CARRIER DAY_OF_WEEK Bandwidth(s): 2625593734 0.2933983 3.373063 0.7194768 0.5220916

Kernel Regression Estimator: Local-Constant Bandwidth Type: Fixed

Continuous Kernel Type: Second-Order Gaussian No. Continuous Explanatory Vars.: 3

Unordered Categorical Kernel Type: Aitchison and Aitken No. Unordered Categorical Explanatory Vars.: 2

```
# squared-error loss
mean((test$ARR_DELAY - predict(loclinfit, newdata = test))^2,
     na.rm = TRUE)
```

[1] 3309.925

c)

The plot function for npregression objects (such as the fit returned by npreg) can plot the marginal association of each variable with the response. If you set the plot.errors.method = "bootstrap" option, it will also plot bootstrap-based standard errors for these. Make the plots with standard errors and interpret the results. Which variables seem strongly related with delay length? What do the plots suggest about the appropriate-ness of linear regression? If you saw major non-linearities in any variable, do these non- linearities appear to harm the predictions enough to make linear regression perform dra- matically worse than kernel regression?

The plots below suggest that delay length (ARR_DELAY) is correlated with the variables ARR_DELAY,

```
dev.new(width=5, height=8)
par(mfrow=c(3,2))
npplot(bw,
       data = data, plot.errors.method= "bootstrap")
```

d) Repeat this analysis, but use a locally linear kernel regression (with regtype = "ll" provided to npregbw). Compare the test error from this model to that for the previous one. Discuss possible reasons for any difference you see.

Here it seems that the mean squared error is 3749.95 which is larger than the previous model.

Multistart 1 of 5 | Multistart 1 of 5 | Multistart 1 of 5 | Multistart 1 of 5 / Multistart 1 of 5 - Multistart 1 of 5
Multistart 1 of 5 | Multistart 1 of 5 | Multistart 1 of 5 / Multistart 1 of 5 - Multistart 1 of 5
Multistart 1 of 5 | Multistart 1 of 5 / Multistart 2 of 5 | Multistart 2 of 5 | Multistart 2 of 5 / Multistart 2 of 5 - Multistart 2 of 5
Multistart 2 of 5 | Multistart 2 of 5 / Multistart 2 of 5 - Multistart 2 of 5
Multistart 2 of 5 | Multistart 2 of 5 | Multistart 3 of 5 | Multistart 3 of 5 | Multistart 3 of 5 / Multistart 3 of 5 - Multistart 3 of 5
Multistart 3 of 5 | Multistart 3 of 5 | Multistart 4 of 5 | Multistart 4 of 5 | Multistart 4 of 5 / Multistart 4 of 5 - Multistart 4 of 5
Multistart 4 of 5 | Multistart 4 of 5 / Multistart 4 of 5 | Multistart 4 of 5 | Multistart 5 of 5 | Multistart 5 of 5 | Multistart 5 of 5 / Multistart 5 of 5 - Multistart 5 of 5
Multistart 5 of 5 | Multistart 5 of 5 / Multistart 5 of 5 - Multistart 5 of 5 | Multistart 5 of 5 |

Regression Data: 291 training points, in 5 variable(s) DEP_TIME ORIGIN DAY_OF_MONTH CARRIER DAY_OF_WEEK Bandwidth(s): 126.4377 0.3067665 27.51952 0.8999998 0.6602434

Kernel Regression Estimator: Local-Linear Bandwidth Type: Fixed

Continuous Kernel Type: Second-Order Gaussian No. Continuous Explanatory Vars.: 3

Unordered Categorical Kernel Type: Aitchison and Aitken No. Unordered Categorical Explanatory Vars.: 2

[1] 3749.95