

## Midterm Exam 4

(Due on class 11/09/2009, Monday)

**Instructions:** This is a take-home exam, which covers model selection and prediction in multiple linear regression. You are required to do it independently. Copy-and-paste only the necessary R output and interpret your results. Remember to include your R codes in an appendix.

To illustrate, we first consider an astronomical data set taken from Ex4.9 in Mendenhall and Sinich (2003). A quasar is a distant celestial object that is at least four billion light-years away from earth. The *Astronomical Journal* (Schmidt, Schneider, and Gunn 1995) reported a study of 90 quasars detected by a deep space survey. Based on the radiations provided by each quasar, astronomers were able to measure several of its quantitative characteristics, including redshift range ( $X_1$ ), line flux in  $\text{erg/cm}^2$  ( $X_2$ ), line luminosity in  $\text{erg/s}$  ( $X_3$ ),  $AB_{1450}$  magnitude ( $X_4$ ), absolute magnitude ( $X_5$ ), and rest frame equivalent width ( $Y$ ). One objective of the study is to model the rest frame equivalent width ( $Y$ ) using other characteristics. The data for a sample of 25 large quasars are used in this project. The following R codes read the dataset into R.

```
quasar <- read.table( file =
  "http://pegasus.cc.ucf.edu/~xsu/CLASS/STA4164/quasar.txt",
  header = F, col.names=c("id", "x1", "x2", "x3", "x4", "x5", "y"))
quasar
```

### Part I. Model Selection

#### 1. All Possible Regressions

First we try out the all possible regressions method. I have written an R function `all.possible.regressions()` for this purpose.

- Copy and paste the following codes in an R script file and run them. Study the output and select your best model choice according to an appropriate criterion. Provide a brief reason for why you want to use that criterion.

```
# =====
# ALL POSSIBLE REGRESSIONS
# =====

# The parameter k is the total number of predictors. To apply the function,
# you need to prepare the data so that all predictors for selection are named
# as x1, x2, ..., and the response is called y.
```

```

all.possible.regressions <- function(dat, k){
  n <- nrow(dat)
  regressors <- paste("x", 1:k, sep="")
  lst <- rep(list(c(T, F)), k)
  regMat <- expand.grid(lst);
  names(regMat) <- regressors
  formular <- apply(regMat, 1, function(x)
    as.character(paste(c("y ~ 1", regressors[x]), collapse="+")))
  allModelsList <- apply(regMat, 1, function(x)
    as.formula(paste(c("y ~ 1", regressors[x]), collapse=" + ")))
  allModelsResults <- lapply(allModelsList,
    function(x, data) lm(x, data=data), data=dat)
  n.models <- length(allModelsResults)
  extract <- function(fit) {
    df.sse <- fit$df.residual
    p <- n - df.sse - 1
    sigma <- summary(fit)$sigma
    MSE <- sigma^2
    R2 <- summary(fit)$r.squared
    R2.adj <- summary(fit)$adj.r.squared
    sse <- MSE*df.sse
    aic <- n*log(sse) + 2*(p+2)
    bic <- n*log(sse) + log(n)*(p+2)
    out <- data.frame(df.sse=df.sse, p=p, SSE=sse, MSE=MSE,
      R2=R2, R2.adj=R2.adj, AIC=aic, BIC=bic)
    return(out)
  }
  result <- lapply(allModelsResults, extract)
  result <- as.data.frame(matrix(unlist(result), nrow=n.models, byrow=T))
  result <- cbind(formular, result)
  rownames(result) <- NULL
  colnames(result) <- c("model", "df.sse", "p", "SSE", "MSE", "R2",
    "R2.adj", "AIC", "BIC")
  return(result)
}

all.possible.regressions(dat=quasar, k=5)

```

- b. Fit this best model. Provide the table of parameter estimates and the analysis of variance (ANOVA) table. Note that the ANOVA table is a little different from the form we presented in class. Just leave it as is. Recall that a linear model with  $X_3$  and  $X_5$ , for example, can be fit using `fit <- lm(y~ x3 + x5, data=quasar)`.

## 2. Stepwise Selection

In R, functions `step()` and `stepAIC()` (in the MASS library) can be used for stepwise procedures. But the implemented procedure, which utilizes AIC, is slightly

different from the procedures outlined in class, which is based on significance testing. In this project, you are asked to NOT use `step()` and `stepAIC()`.

Instead, you are asked to perform both *backward elimination* and *forward addition* step manually step by step with the aid of R, setting both significance thresholds for removal and entry as 0.05.

- a. Provide some details of the fitting results and actions that you take at each step.

For example, to start with the backward elimination, you fit the whole model first.

```
fit <- lm(y~ x1 + x2 + x3 + x4 + x5, data=quasar); summary(fit)
```

You will get the following fitting results:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	18989.03	26569.36	0.715	0.483
x1	113.71	102.19	1.113	0.280
x2	471.36	830.50	0.568	0.577
x3	-254.50	825.25	-0.308	0.761
x4	20.78	573.36	0.036	0.971
x5	64.59	570.65	0.113	0.911

Accordingly, we should consider dropping  $X_3$  because it yields the largest p-value, which is greater than 0.05. Next, we continue with model

```
fit <- lm(y~ x1 + x2 + x3 + x5, data=quasar); summary(fit)
```

and so on. Completing the process, what is the best model choice for backward elimination? For forward addition?

- b. Fit the best models that you find using backward elimination and forward addition. Provide the table of parameter estimates and the analysis of variance (ANOVA) table.

## Part II. Prediction

Suppose now there is a new identified quasar, whose information on  $(X_1, X_2, \dots, X_5)$  is provided as below:

x1	x2	x3	x4	x5
3.3932	-13.8044	45.1432	19.7260	-26.3328

With each of the best models that you find in Part I, construct a 95% *prediction interval* to predict its rest frame equivalent width ( $Y$ ). You might want to study Section 4 in the class handout on Multiple Linear Regression, available at

<http://pegasus.cc.ucf.edu/~xsu/CLASS/STA4164/hd4.pdf>

for how to perform prediction in R.

### References:

- Mendenhall, W. and Sinich, T. (2003). *Regression Analysis: A Second Course in Statistics*, 6th edition. Upper Saddle River, NJ: Prentice Hall.