Stat 794, Example Application of `knitr`

Kristine Dinh

September 3, 2020

An R dump of the summary of the regression fit for application count.

```
##
## Call:
## lm(formula = Apps ~ Private + Elite + Accept + Outstate + Room.Board +
##     Grad.Rate, data = College)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5094.5  -329.7   -22.6   226.8 10114.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -985.95379  204.82380  -4.814 1.78e-06 ***
## PrivateYes  -291.62328  133.28335  -2.188 0.028970 *
## EliteYes    1745.00184  151.74052  11.500  < 2e-16 ***
## Accept         1.42869    0.02024  70.601  < 2e-16 ***
## Outstate      -0.01427    0.01690  -0.845 0.398600
## Room.Board     0.16615    0.04953   3.355 0.000834 ***
## Grad.Rate      8.63483    2.94718   2.930 0.003491 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1128 on 770 degrees of freedom
## Multiple R-squared:  0.9157,Adjusted R-squared:  0.915
## F-statistic:  1394 on 6 and 770 DF,  p-value: < 2.2e-16
```

To predict the number of application recieved every year, a model with six predictors was generated including private school flag, elite, acceptance rate, out of state tuition, room and board costs, and graduation rate. A new set of testing data was used to generate the prediction of application count. Predicted values for the two new schools are: 7000, 4800, 9300 and 2300, 57, 4600

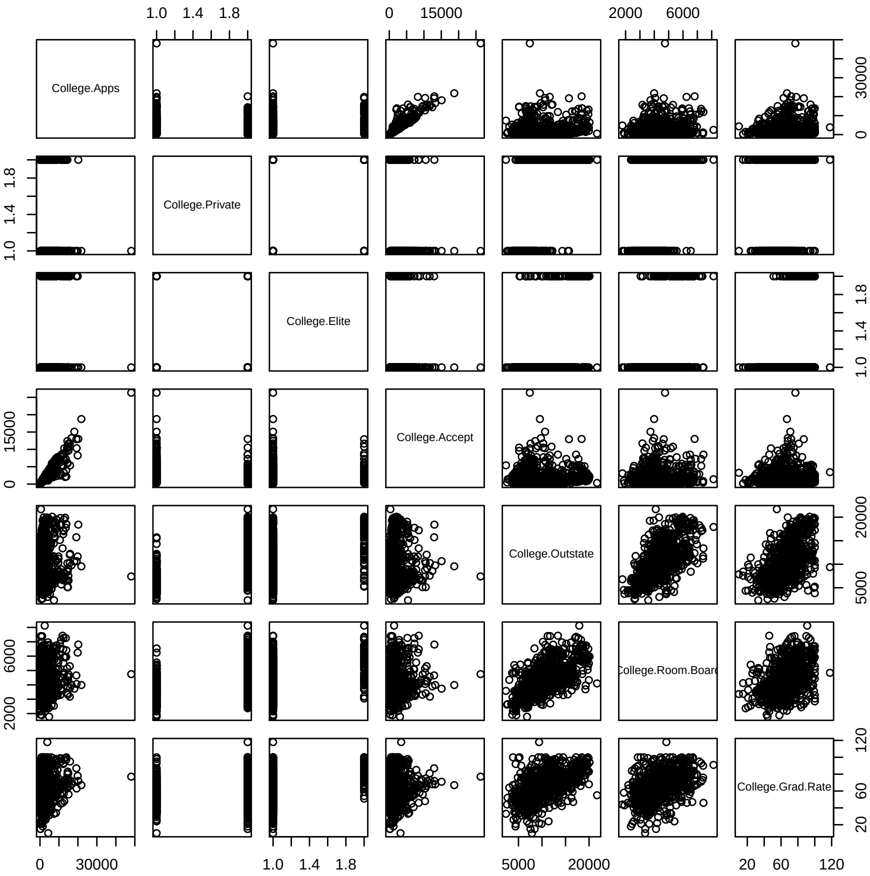

Figure 1: Pairwise scatterplot of variables including college application, college private flag, elite, appceptance rate, out of state tuition, room board cost, and graduation rate.

When exploring the relationship between each variable, a pairwise scatterplot was created to visualize the correlation between each variable in the model. As seen in Figure 1, out of state cost and room board cost are highly correlated with each other. On the other hand, there is a small correlation between out of state tuition and college acceptance rate.

	univ	elite	gradrate	preds	lwr	upr	outstate
1	University 1	No	0.60	7000.00	4800.00	9300.00	8000.00
2	University 2	Yes	0.90	2300.00	57.00	4600.00	16000.00

Table 2: New variables and prediction table of two new University 1 and 2

Using the new variables, we was able to generate predictions of number of application recieved for two universities 1 and 2. As seen in Table 2, University 1 recieved more applications than University 2.

Appendix A: Supplementary Plots

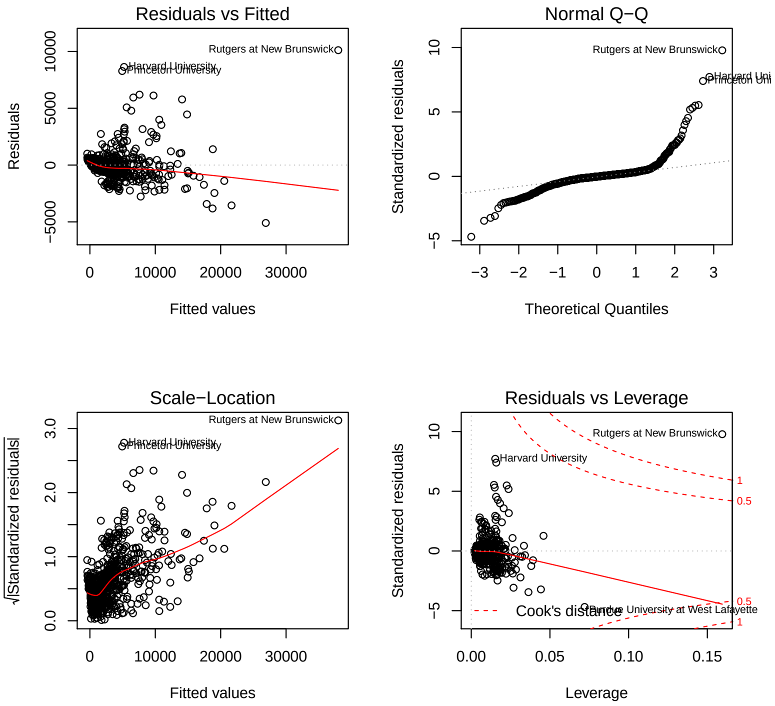

Figure 2: Plot of Regression Diagnostics. Top-left: Residual vs fitted values, top-right: QQ plot with Normal Distribution, bottom-left: Scale-location of root of standardized residuals vs fitted values, bottom-right: Residuals vs leverage.

After developing a model, refression diagnostics were ran to ensure the quality of the model. As seen in Figure 2, residuals vs fitted values have a pattern similar to a fan. This implies that error terms are not independent. In addition, looking at Figure 2 top-right plot, the regression is not normal distributed as standardizard residuals are not close to the dotted line.

Appendix B: R Code

```r

library(ISLR)
library(stargazer)
library(xtable)

rm(list=ls(all=TRUE)) # remove all previous objects from memory

# Set up data for the illustration
# For illustration purposes we will use the College data set from the ISLR text
# Create an indicator of Elite College status (see exercise 8 in Ch. 2 of ISLR text)
Elite=rep("No",nrow(College))
Elite[College$Top10perc >50]="Yes"
Elite=as.factor(Elite)
College=data.frame(College ,Elite)
numvars = length(College) # number of variables in the College data set
n = dim(College)[1]

# Fit a model
fm1 = lm(Apps~Private+Elite+Accept+Outstate+Room.Board+Grad.Rate, data = College)
# predcit a new school not in the data set
new1 = data.frame(Private="No", Elite="No", Accept=5000, Outstate=8000, Room.Board=6000,
    Grad.Rate=0.6)
newpred1 = signif(predict(fm1, new1, interval="prediction"), 2)
new2 = data.frame(Private="Yes", Elite="Yes", Accept=1000, Outstate=16000, Room.Board=4000,
     Grad.Rate=0.90)
newpred2 = signif(predict(fm1, new2, interval="prediction"), 2)

fm.table <- xtable(fm1, digits=2,
                   caption = "Summary of Regression",
                   label="reginf")

align(fm.table) <- "|l|rrrr|"

print(fm.table)

y = data.frame(College$Apps,College$Private,College$Elite,College$Accept,College$Outstate,
    College$Room.Board,College$Grad.Rate)
pairs(y)

stargazer(College,
          title="Summary statistics of all variables for the ISLR College data set.",
          label="descrips",
          summary.stat = c("mean", "median", "sd", "min", "p25", "p75", "max"),
          covariate.labels = c("Private Flag"
                               , "Application Count"
                               , "Acceptance Count"
                               , "Enrollment Count"
                               , "Top 10 Percent in High School"
                               , "Top 25 Percent in High School"
                               , "Full-time Undergrad"
                               , "Part-time Undergrad"
                               , "Out-of-state tuition"
                               , "Room and board Costs"
                               , "Estimated Book Costs"
                               , "Estimated Personal Costs"
                               , "Percent of Faculty with PhD"
                               , "Percent of Faculty with terminal Degree"
                               , "Student/Faculty Ratio"
                               , "Percent of alumni donated"
                               , "Instructional expenditure per student"
                               , "Graduation rate"),
          float.env = "sidewaystable",
          table.placement = "H")
```

```


# create the table and store in 'x'
univ = rbind("University 1", "University 2")
elite = rbind("No", "Yes")
gradrate = rbind(new1[,6], new2[,6])
preds = rbind(newpred1[,1], newpred2[,1])
lwr = rbind(newpred1[,2], newpred2[,2])
upr = rbind(newpred1[,3], newpred2[,3])

x <- data.frame(univ
                 , elite
                 , gradrate
                 , preds
                 , lwr
                 , upr
                 , outstate = rbind(new1[,"Outstate"], new2[,"Outstate"]))

fm.table <- xtable(x,
                    digits = 2,
                    caption = "Prediction Table",
                    label = "pred_table",
                    table.placement="H")

align(fm.table) <- "|l|rrrrrrr|"

print(fm.table)

## diagnostic

par(mfrow=c(2,2))
plot(fm1)
```

Listing 1: List of codes