04MultiRegModel2

Karin reinhold

9/15/2020

Example: Study of CS Students

#####Problem: Computer science majors at Purdue have a large drop-out rate. Potential Solution: Can we find predictors of success? Predictors must be available at time of entry into program.

Data Available: * Grade point average (GPA) after three semesters (Y, the response variable) Five potential predictors (p=6)

- X1=Highschoolmathgrades(HSM)
- X2= High school science grades (HSS)
- X3= High school English grades (HSE)
- X4= SAT Math (SATM)
- X5= SAT Verbal (SATV)
- Gender (1 = male, 2 = female) (not a continuous variable)

We have $n = 224$ observations, so if all five variables are included, the design matrix X has dimension 224x6.

###Look at the individual variables

Our first goal should be to take a look at the variables to see…

- Is there anything that sticks out as unusual for any of the variables?
- How are these variables related to each other (pairwise)?
- If two predictor variables are strongly correlated, we wouldn't want to use them in the same model!

We do this by looking at statistics and plots

```
csdata <- read.table("/cloud/project/csdata.txt", header=TRUE, quote="\"")
#View(csdata)
```

The first column is an id that we don't need. So we'll remove the id column:

```
csdata<- csdata[c(-1)]
```

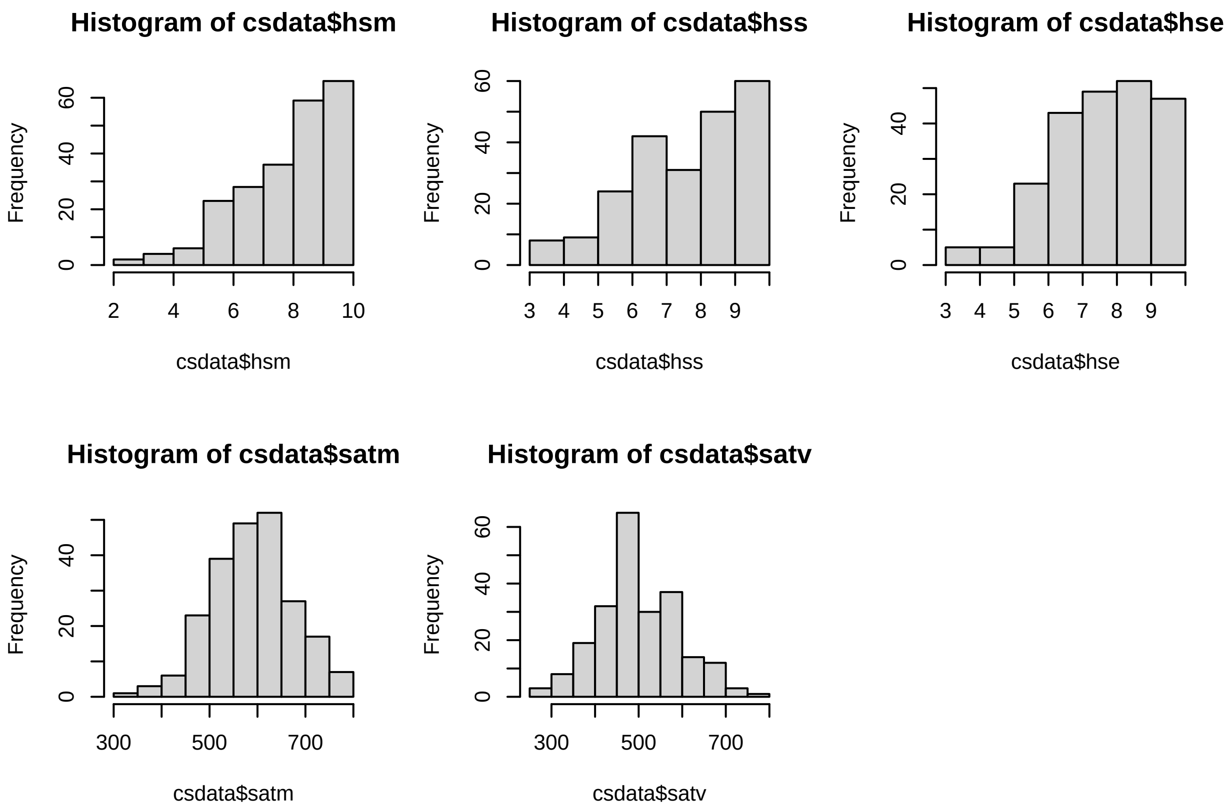
Descriptive Statistics:

```
summary(csdata)
```

```
##       gpa             hsm              hss              hse        
##  Min.   :0.120   Min.   : 2.000   Min.   : 3.000   Min.   : 3.000  
##  1st Qu.:2.167   1st Qu.: 7.000   1st Qu.: 7.000   1st Qu.: 7.000  
##  Median :2.740   Median : 9.000   Median : 8.000   Median : 8.000  
##  Mean   :2.635   Mean   : 8.321   Mean   : 8.089   Mean   : 8.094  
##  3rd Qu.:3.212   3rd Qu.:10.000   3rd Qu.:10.000   3rd Qu.: 9.000  
##  Max.   :4.000   Max.   :10.000   Max.   :10.000   Max.   :10.000  
##       satm            satv            sex       
```

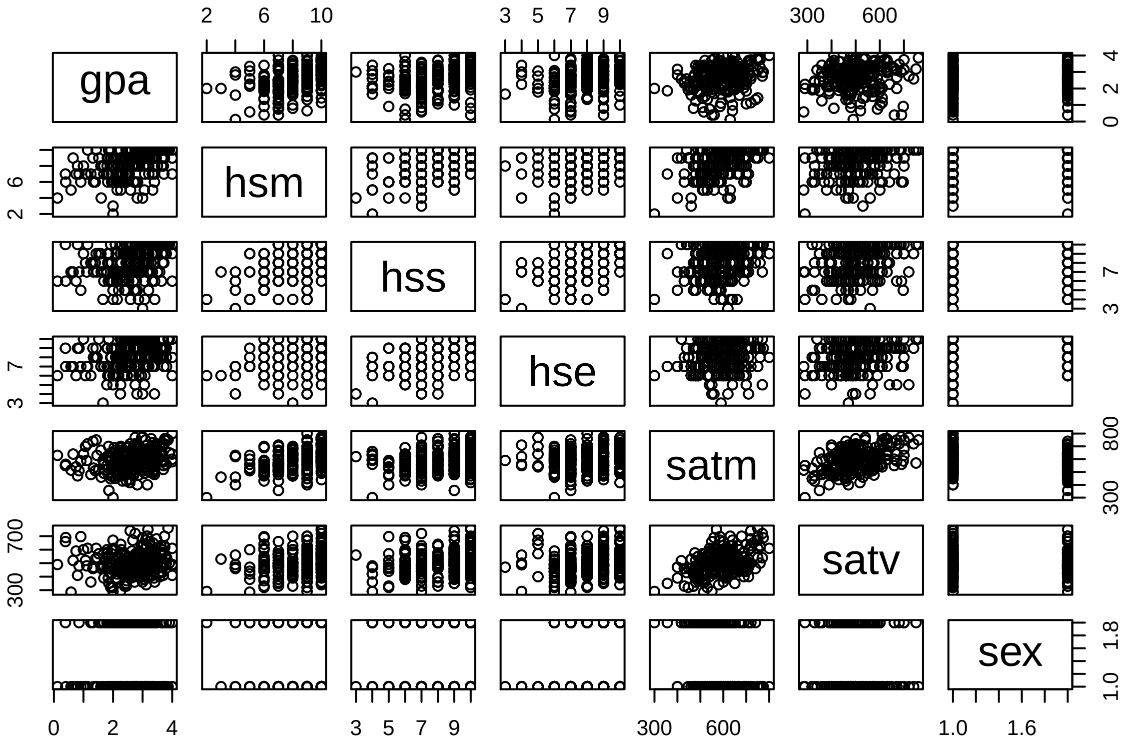
```
##  Min.   :300.0   Min.   :285.0   Min.   :1.000
##  1st Qu.:540.0   1st Qu.:440.0   1st Qu.:1.000
##  Median :600.0   Median :490.0   Median :1.000
##  Mean   :595.3   Mean   :504.5   Mean   :1.353
##  3rd Qu.:650.0   3rd Qu.:570.0   3rd Qu.:2.000
##  Max.   :800.0   Max.   :760.0   Max.   :2.000
```

```r
#t(summary(csdata))   # same info, transposed
par(mfrow=c(2,3))
hist(csdata$hsm)
hist(csdata$hss)
hist(csdata$hse)
hist(csdata$satm)
hist(csdata$satv)
par(mfrow=c(1,1))
```



Scatter plots matrix:

```r
plot(csdata)        #matrix of plots. Removed the 1st column #which is id
```



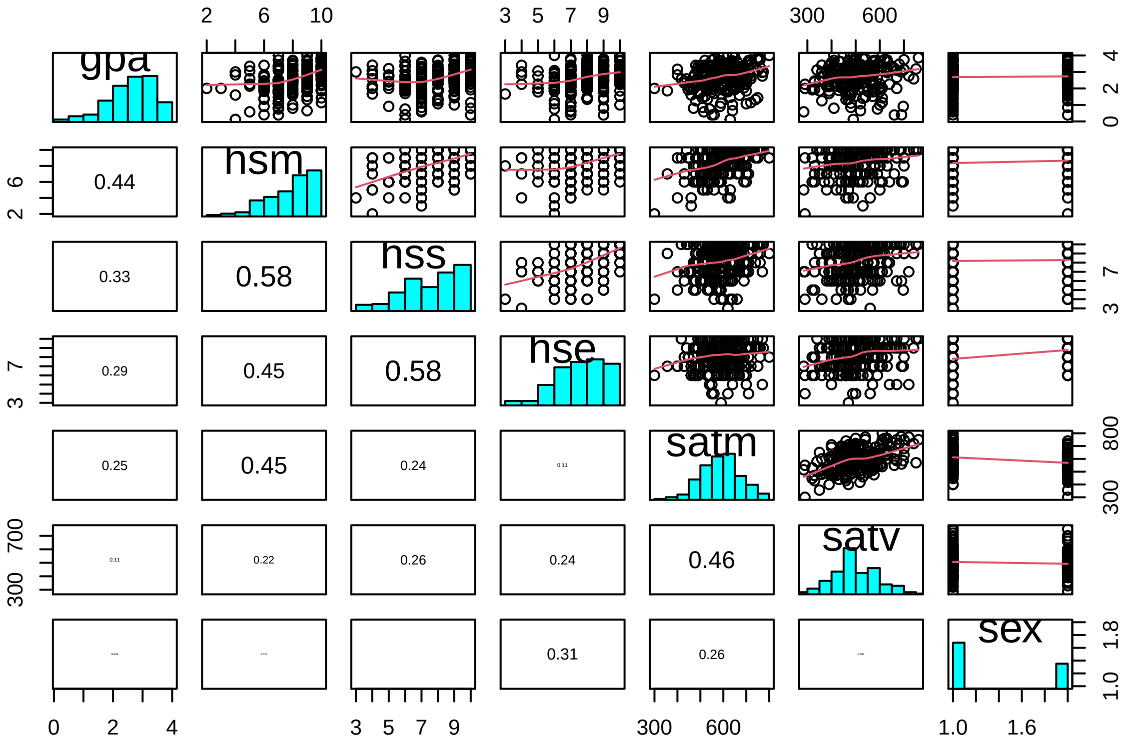
Correlations matrix:

```r
cor(csdata)      #correlation matrix
```

```
##             gpa       hsm        hss       hse       satm        satv
## gpa  1.0000000 0.4364988 0.32942533 0.2890013  0.2517143  0.11449046
## hsm  0.4364988 1.0000000 0.57568646 0.4468865  0.4535139  0.22112029
## hss  0.3294253 0.5756865 1.00000000 0.5793746  0.2404793  0.26169754
## hse  0.2890013 0.4468865 0.57937457 1.0000000  0.1082849  0.24371460
## satm 0.2517143 0.4535139 0.24047931 0.1082849  1.0000000  0.46394188
## satv 0.1144905 0.2211203 0.26169754 0.2437146  0.4639419  1.00000000
## sex  0.0479074 0.0720413 0.01072383 0.3141998 -0.2590924 -0.06293167
##              sex
## gpa   0.04790740
## hsm   0.07204130
## hss   0.01072383
## hse   0.31419985
## satm -0.25909245
## satv -0.06293167
## sex   1.00000000
```

Better display:

```r
source('pairs.r')
pairs(csdata,panel=panel.smooth,diag.panel=panel.hist,lower.panel=panel.cor)
```



Linear Model:

Sex is coded as 0 and 1. But it is a categorical variable. Hence we need to tell R that it is categorical (a factor variable)

```
mod1<-lm(gpa~hsm+hss+hse+satm+satv+as.factor(sex),data=csdata)   #linear model
summary(mod1)
```

```
##
## Call:
## lm(formula = gpa ~ hsm + hss + hse + satm + satv + as.factor(sex),
##     data = csdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.08566 -0.30776  0.07675  0.49203  1.71001
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.3110099  0.4044729   0.769 0.442773
## hsm              0.1442267  0.0397944   3.624 0.000361 ***
## hss              0.0382718  0.0387448   0.988 0.324355
## hse              0.0510335  0.0422777   1.207 0.228707
## satm             0.0010033  0.0007173   1.399 0.163278
## satv            -0.0004109  0.0005932  -0.693 0.489314
## as.factor(sex)2  0.0323725  0.1114797   0.290 0.771796
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7015 on 217 degrees of freedom
```

```
## Multiple R-squared:  0.2118, Adjusted R-squared:   0.19
## F-statistic: 9.716 on 6 and 217 DF,  p-value: 1.781e-09
```

Alternatively, you can do: In the mac: attach(csdata) mod1<-lm(gpa~hsm+hss+hse+satm+satv+as.factor(sex)) #linear model summary(mod1)

####Estimating betas: From the summary we see the test for the β's

$$H_0 : \beta_i = 0 \text{ vs } H_a : \beta_i \neq 0$$

We can also construct confidence intervals:

$$\hat{\beta}_i \pm t_{\alpha/2} se(\hat{\beta}_i), \text{ with df}=n-p-1=217$$

In R:

```r
confint(mod1)     #95% confidence intervasl
```

```
##                         2.5 %       97.5 %
## (Intercept)     -0.4861885501 1.1082084454
## hsm              0.0657936038 0.2226598126
## hss             -0.0380924561 0.1146359767
## hse             -0.0322939907 0.1343609504
## satm            -0.0004103283 0.0024170247
## satv            -0.0015800895 0.0007583722
## as.factor(sex)2 -0.1873490549 0.2520940150
```

R^2 = the square of the correlation of Y with $\hat{Y}$

square of correlation of gpa and fitted values:

```r
(cor(csdata$gpa,fitted(mod1)))^2
```

```
## [1] 0.2117566
```

F-test for global fit

$$\begin{aligned} H_0 &: Y = \beta_0 + \epsilon \\ H_a &: Y = \beta_0 + \beta_1 hsm + \beta_2 hss + \beta_3 hse + \beta_4 satm + \beta_5 satv + \beta_6 sex + \epsilon \end{aligned}$$

$$F = \frac{MSR}{MSE}, \text{ with df} = (6, 217)$$

coming from the ANOVA table. Above, in the summary, we have the end result of the test

F-statistic: 9.716 on 6 and 217 DF, p-value: 1.781e-09

So we reject the null, and we conclude that the linear model with the 6 variables is significant in explaining the variation of GPA.

ANOVA table Just for completion we can do the ANOVA table:

```r
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: gpa
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## hsm               1  25.810 25.8099 52.4524 7.573e-12 ***
## hss               1   1.237  1.2371  2.5141    0.1143
## hse               1   0.665  0.6654  1.3522    0.2462
## satm              1   0.699  0.6987  1.4199    0.2347
## satv              1   0.233  0.2327  0.4728    0.4924
## as.factor(sex)    1   0.041  0.0415  0.0843    0.7718
## Residuals       217 106.778  0.4921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: the sum of the Sum Sq corresponding to the variables = SSR

Source	SS	df	MS	F	p-val
Regression	28.655	6	4.775833	9.705705	1.821345e-09
Residuals	106.778	217	0.4920645		
Total	109.623	223			

This is a one sided test. The p-values is $P(F > 9.705705)$ p-value:

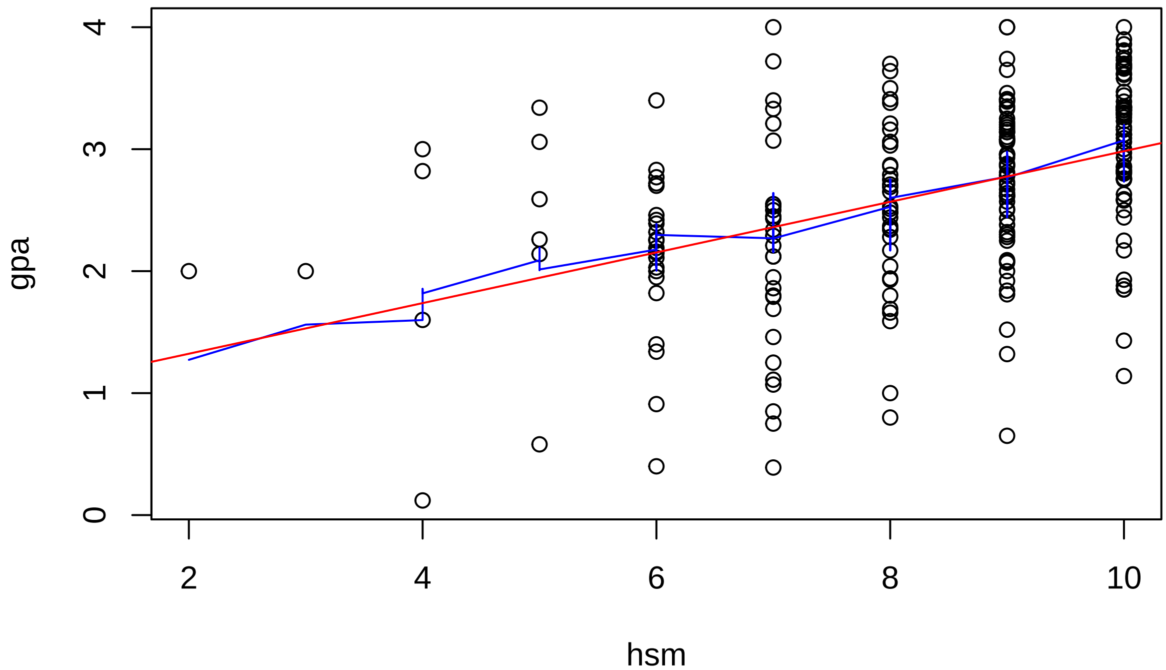
```r
1-pf(9.705705,6,217)
```

```
## [1] 1.821345e-09
```

Summary Now, we have seen the plots and the correlations and adjusted $R^2 = .19$,

- The F tests tells us that the model is capturing a general trend in terms of the other variables (think about the picture of pelican eggs, that there was a linear trend but was not strong),
- but looking at the adjusted $R^2 = .19$ the plots and the correlations, we see the model doesn't capture much, and the fit is quite weak.
- When making predictions, we would get a general estimate but the true estimate would have quite a bit of variability beyond the mean response.

```r
plot(gpa~hsm,data=csdata)
index<-order(csdata$hsm)
lines(csdata$hsm[index],mod1$fitted.values[index],col='blue')
abline(lm(gpa~hsm,data=csdata),col='red')
```



We also saw in the tests for β's that many of the variables were not significant, individually. We are going to investigate about dropping some variables in next class. Notice that satv has a negative sign and that is counter-intuitive. We will see htat this may happen when there is high correlation with other variables. Sex had a high correlation with GPA so an investigation of incorporating interaction terms must be done. We'll look into interaction terms later.

Predicting new observations

Predicting the mean response

Estimate the *expected value of the response* (or *mean response*) for a given predictor score (mean response).

Suppose we want to estimate the mean response for a new value (or at a given value) of the explanatory variables

$$x_h = (1, x_{1,h}, \ldots, x_{p,h}),$$

the fitted value (mean response) is

$$\hat{y}_h = x_h \hat{\beta} = x_h (X'X)^{-1} X'Y$$

so, it is a linear combination of the Y and hence it is normal with variance

$$Var(\hat{y}_h) = \sigma^2 x_h (X'X)^{-1} x_h'$$

Recall that $\text{Var}(Y) = \sigma^2 Id$ and $\text{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$. And in general, the variance of AY is $A\sigma^2 Id A' = \sigma^2 AA'$.

So the standard deviation for $\hat{y}_h$ is

$$\sigma \sqrt{x_h (X'X)^{-1} x_h'}$$

and the standard error is

$$se(\hat{y}_h) = s \sqrt{x_h (X'X)^{-1} x_h'}$$

where s is our estimate for σ, that is $s = \sqrt{MSE}$.

Confidence Interval for the Mean Response The confidence interval for the *mean response* is then

$$\hat{y}_h \pm t_{\alpha/2} se(\hat{y}_h), \text{ with df=}(n-p-1)$$

Predicting an Individual Response

Estimate an *individual value of the response* for a given predictor score.

Racall that

$$y_h = \hat{y}_h + \epsilon$$

is still normal but with larger variance:

$$Var(y_h) = Var(\hat{y}_h) + Var(\epsilon) = \sigma^2 \left[1 + x_h(X'X)^{-1}x_h'\right]$$

So the standard deviation for the individual response y_h is

$$\sigma\sqrt{1 + x_h(X'X)^{-1}x_h'}$$

and the standard error is

$$se(y_h) = s\sqrt{1 + x_h(X'X)^{-1}x_h'}$$

where s is our estimate for σ, that is $s = \sqrt{MSE}$.

Confidence Interval for the Individual Response The confidence interval for the individual response is then

$$\hat{y}_h \pm t_{\alpha/2} se(y_h), \text{ with df=}(n-p-1)$$

```
mynew<-data.frame(1,hsm=8, hss=8, hse=8, satm=600, satv=600, sex=1 )
# Confidence interval for mean response
predict(mod1,mynew,interval="confidence",level=.90)
```

Computation done with R

```
##        fit      lwr      upr
## 1 2.534759 2.399788 2.669731
```

```
# Confidence interval for individual response
predict(mod1,mynew,interval="prediction",level=.90)
```

```
##        fit      lwr      upr
## 1 2.534759 1.368159 3.701359
```

```
XX<-t(model.matrix(mod1)) %*% model.matrix(mod1)  # X'X
mynew<-matrix(c(1,8,8, 8, 600, 600,0 ),nrow=1)    # New observation
yfit<- mynew %*% mod1$coefficients        # fitted value
# next the standard error for the mean prediction
semean<-summary(mod1)$sigma*sqrt(mynew %*% solve(XX) %*% t(mynew))
# critical value for the t-distribution with n-p-1 df
tcrit<- qt(.95,217)
me<-tcrit*semean  # margin of error
upper.mean<- round(yfit + me,2)   # upper bound for mean prediction
```

```r
lower.mean <- round(yfit - me,2)  # lower bound for mean prediction
#cbind(fit=yfit, lower=lower.mean, upper=upper.mean, "mean prediction")
#
# next the standard error for the individual prediction
sepred<-summary(mod1)$sigma*sqrt(1+ mynew %*% solve(XX) %*% t(mynew))
mepred<-tcrit*sepred  # margin of error
upper.pred<- round(yfit + mepred,2)  # upper bound for individual prediction
lower.pred <- round(yfit - mepred,2) # lower bound for individual prediction
results<-data.frame(fit=rep(round(yfit,2),2),lower.bound=c(lower.mean,lower.pred),upper.bound=c(upper.me
#cbind(fit=yfit, lower=lower.pred, upper=upper.pred, "individual prediction")
results
```

Computations by hand

```
##    fit lower.bound upper.bound                  type
## 1 2.53        2.40        2.67       mean prediction
## 2 2.53        1.37        3.70 individual prediction
```