AP Statistics Data Analysis

Kelly Radimer

Dec 16, 2020

## Introduction

The intention of this study is to build a model that students can use to predict their AP Exam Score based on the most relevant test scores. The data used comes from my own students for the years 2016-2020. The majority (approximately 85%) of the students in the class are high school seniors, with the remainder being juniors (14%), and sophomores (1%). AP Exam scores range between 1 and 5, and chapter test scores range from 0 to 40. The Chapters correspond with The Practice of Statistics, 4th edition, by Starnes, Yates and Moore. For each test, there are two different forms that are alternated from year to year, but the question type and difficulty level are consistent from year to year.

## Data Wrangling

Load all relevant packages:

library(pacman)  
p\_load(dplyr, ggplot2, readxl, stringr, MASS, GGally)

Upload my grade books and student AP Exam scores:

for (i in 15:19) {  
 fn <- paste0("./data/",i,i+1," AP Stats S1 Gradebook.xls")  
 nam <- paste("S1Gradebook", i,i+1, sep = "")  
 assign(nam, read\_excel(fn, col\_names = TRUE, skip=1, sheet = "Test"))  
}

for (i in 15:19) {  
 fn <- paste0("./data/",i,i+1," AP Stats S2 Gradebook.xls")  
 nam <- paste("S2Gradebook",i,i+1, sep = "")  
 assign(nam, read\_excel(fn, col\_names = TRUE, skip=1,sheet = "Test"))  
}

for (i in 16:20) {  
 fn <- paste0("./data/20",i," AP Exam Results.xls")  
 nam <- paste("APExam20",i, sep = "")  
 assign(nam, read\_excel(fn, col\_names = TRUE, skip=1))  
 assign(nam, get(nam) %>% dplyr::select(Name, Score))  
}

Filter out the blank rows and students who dropped the class:

for(i in 15:19) {  
 nam <- paste0("S1Gradebook",i,i+1)   
 assign(nam, get(nam) %>%  
 filter(!is.na(Name)) %>%  
 filter(Percent!=0) %>%  
 filter(Name!="Name"))  
 }

for(i in 15:19) {  
 nam <- paste0("S2Gradebook",i,i+1)   
 assign(nam, get(nam) %>%  
 filter(!is.na(Name)) %>%  
 filter(Percent!=0) %>%  
 filter(Name!="Name"))  
 }

Join the two semesters together, add a column for AP Exam score, and filter out the students who didn’t take the AP Exam:

for (i in 15:19) {  
 nam <- paste0("combined",i,i+1)  
 df1 <- paste0("S1Gradebook",i,i+1)  
 df2 <- paste0("S2Gradebook",i,i+1)  
 assign(nam, full\_join(get(df1),get(df2), by=c("Name","ID")))  
}

for (i in 15:19) {  
 nam <- paste0("combined",i,i+1)  
 scores <- paste0("APExam20", i+1)  
 assign(nam, get(scores) %>% right\_join(get(nam), by="Name"))  
}

for (i in 15:19) {  
 nam <- paste0("combined",i,i+1)  
 assign(nam, get(nam) %>% filter(!is.na(Score)))  
}

Rename the relevant variables so that they are the same for all 5 dataframes:

for (i in 1:12) {  
 colnames(combined1516)[grepl(paste0('Ch ',i,' '),colnames(combined1516))] <- paste0('ch',i)  
}  
  
for (i in 1:12) {  
 colnames(combined1617)[grepl(paste0('Ch ',i,' '),colnames(combined1617))] <- paste0('ch',i)  
}  
  
for (i in 1:12) {  
 colnames(combined1718)[grepl(paste0('Ch ',i,' '),colnames(combined1718))] <- paste0('ch',i)  
}  
  
for (i in 1:12) {  
 colnames(combined1819)[grepl(paste0('Ch ',i,' '),colnames(combined1819))] <- paste0('ch',i)  
}  
  
for (i in 1:12) {  
 colnames(combined1920)[grepl(paste0('Ch ',i,' '),colnames(combined1920))] <- paste0('ch',i)  
}

Select AP Exam Scores and chapter test scores (note that due to the pandemic we did not cover Ch 11 and 12 in 2020):

combined1516 <- combined1516 %>%  
 dplyr::select(Score,ch1,ch2,ch3,ch4,ch5,ch6,ch7,ch8,ch9,ch10,ch11,ch12)  
  
combined1617 <- combined1617 %>%  
 dplyr::select(Score,ch1,ch2,ch3,ch4,ch5,ch6,ch7,ch8,ch9,ch10,ch11,ch12)  
  
combined1718 <- combined1718 %>%  
 dplyr::select(Score,ch1,ch2,ch3,ch4,ch5,ch6,ch7,ch8,ch9,ch10,ch11,ch12)  
  
combined1819 <- combined1819 %>%  
 dplyr::select(Score,ch1,ch2,ch3,ch4,ch5,ch6,ch7,ch8,ch9,ch10,ch11,ch12)  
  
combined1920 <- combined1920 %>%  
 dplyr::select(Score,ch1,ch2,ch3,ch4,ch5,ch6,ch7,ch8,ch9,ch10)

Some of the scores are classified as character variables rather than numeric, so we will convert them all to numeric.

dm<-data.matrix(combined1516)  
df16<-as.data.frame(dm)  
dm<-data.matrix(combined1617)  
df17<-as.data.frame(dm)  
dm<-data.matrix(combined1718)  
df18<-as.data.frame(dm)  
dm<-data.matrix(combined1819)  
df19<-as.data.frame(dm)  
dm<-data.matrix(combined1920)  
df20<-as.data.frame(dm)

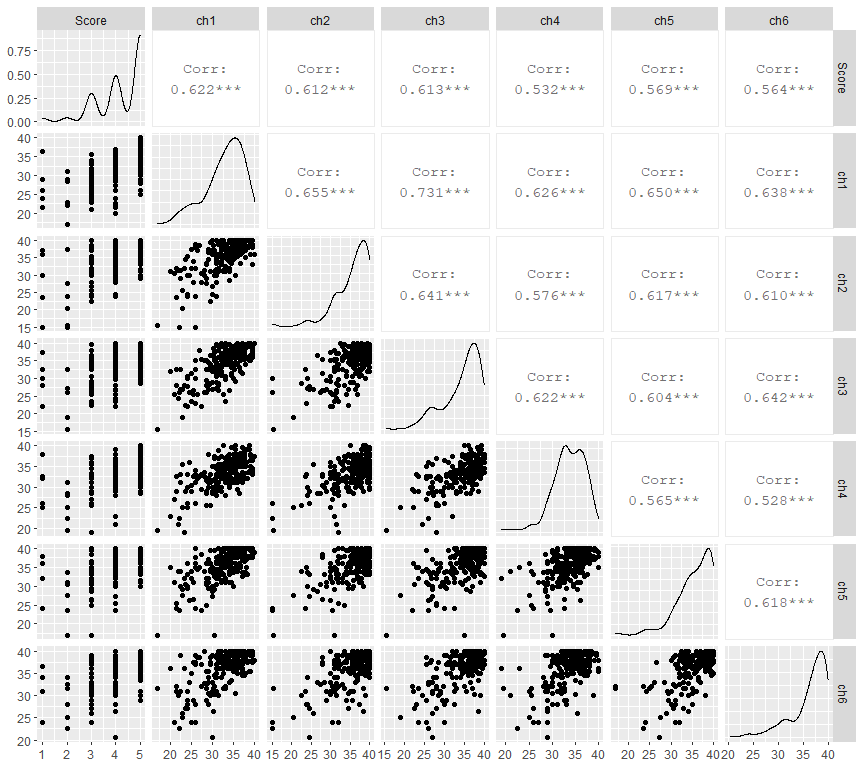
Now we can bind all 5 years into a single data frame and write it into a csv file that we can use for future analysis.

dat <- bind\_rows(df16,df17,df18,df19,df20)  
#write.csv(dat, "CombinedAPStatsData16\_20.csv")

## Exploratory Data Analysis

Now that we have all of our data tidied, let’s do some exploratory data analysis. Here are the paired scatterplots, density curves and correlations for the first seven variables:

ggpairs(dat, columns = 1:7)

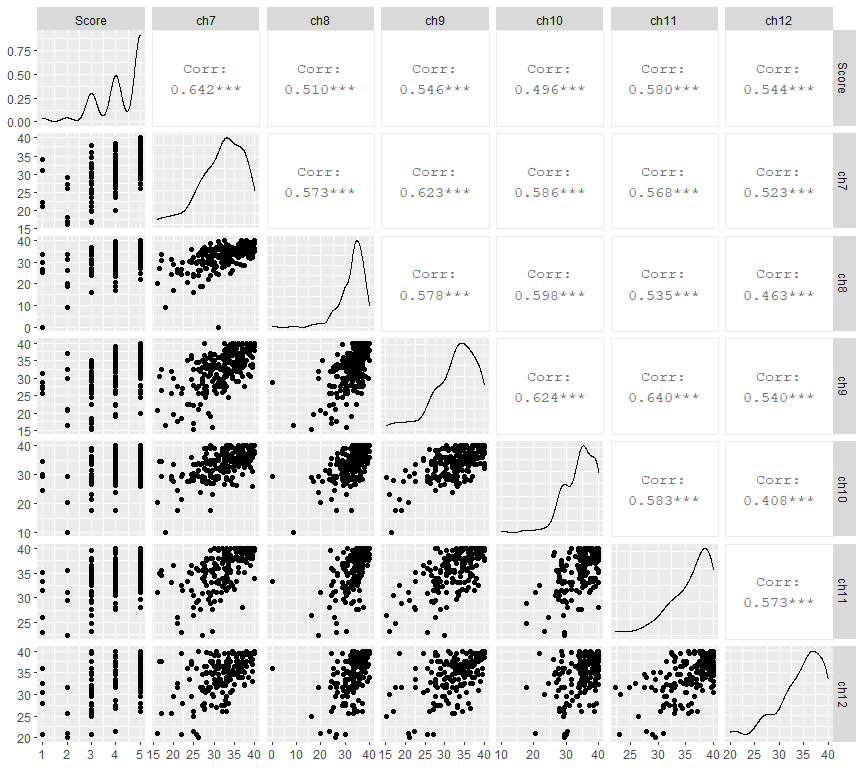


Each of these variables has a moderately strong positive association with the other variables. This makes sense intuitively, as students who do well on one test tend to do well on other tests. Correlation of each test with the AP Exam score is fairly consistent across the first six tests, with the highest correlation being with Chapters 1-3.

Most of the chapter tests have unimodal, left-skewed distributions, since the majority of students tend to perform well on tests. We can see that students tend to perform especially well on the Chapter 2 test, as the scores are concentrated in the 30’s with many students achieving a perfect score of 40/40. Chapter 2 covers Normal Distributions and measures of relative standing. It is largely computational and easy for students to pick up. Chapter 4 was noticeably more difficult, with a very low density of students achieving perfect scores. Chapter 4 is about designing studies, with a large amount of new vocabulary for students to learn, and almost entirely conceptual.

The distribution of AP Exam scores is trimodal, with increasingly large peaks at 3, 4 and 5. There are a small number of outliers with scores of 1 and 2.

ggpairs(dat, columns = c(1,8:13))



Chapters 7-12 have fairly similar relationships with score as the first six chapters. Chapter 7 has a noticeably higher correlation. This makes sense because Chapter 7 is where students are first introduced to sampling distributions, a concept that is key to understanding all the subsequent chapters. Chapter 7 also has greater variability in scores than most of the other chapters.

Chapter 8 in particular has a very low standard deviation. Chapter 8 is about confidence intervals. Most students are able to thoroughly understand the basic mechanics and interpretations of confidence intervals, giving them a solid score. The chapter 8 test has one high-level challenge question asking students to use a simulated sampling distribution to construct a confidence interval for a different parameter, which only a very small number of students are able to successfully complete, which is why the average is high but there are few perfect scores. Chapter 8 also has a low outlier, with a test score of 0. This is a student who was caught cheating on the test.

Chapter 10 has a particularly low correlation with AP Exam Score. I believe the reason for this is that Chapter 10 is taken with a partner, so scores may not be reflective of a student’s actual knowledge.

## Variable selection

Let’s now create a model that students can use to predict their AP Exam score. We will start out with their grades for all twelve tests, and then use Baysian Model selection to create a parsimonious model.

model\_full=lm(Score~ch1+ch2+ch3+ch4+ch5+ch6+ch7+ch8+ch9+ch10+ch11+ch12, data=dat)  
summary(model\_full)

##   
## Call:  
## lm(formula = Score ~ ch1 + ch2 + ch3 + ch4 + ch5 + ch6 + ch7 +   
## ch8 + ch9 + ch10 + ch11 + ch12, data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.3837 -0.3007 0.0536 0.3618 1.6626   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.572782 0.593369 -4.336 2.45e-05 \*\*\*  
## ch1 0.030644 0.019752 1.551 0.122601   
## ch2 0.029682 0.016262 1.825 0.069675 .   
## ch3 0.035070 0.016744 2.094 0.037669 \*   
## ch4 -0.018002 0.018982 -0.948 0.344253   
## ch5 0.012701 0.019409 0.654 0.513717   
## ch6 -0.009835 0.018396 -0.535 0.593610   
## ch7 0.050717 0.014559 3.483 0.000626 \*\*\*  
## ch8 0.022076 0.012801 1.725 0.086390 .   
## ch9 0.003283 0.014283 0.230 0.818481   
## ch10 -0.002430 0.015193 -0.160 0.873123   
## ch11 0.024030 0.017893 1.343 0.181026   
## ch12 0.022767 0.014020 1.624 0.106211   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6324 on 174 degrees of freedom  
## (55 observations deleted due to missingness)  
## Multiple R-squared: 0.6146, Adjusted R-squared: 0.588   
## F-statistic: 23.12 on 12 and 174 DF, p-value: < 2.2e-16

Now we will use the Akaike Information Criterion to strike a balance between over- and under-fitting our model by identifying the model with the lowest AIC value.

model\_AIC=stepAIC(model\_full, direction="backward", k=2, trace=TRUE)

## Start: AIC=-158.84  
## Score ~ ch1 + ch2 + ch3 + ch4 + ch5 + ch6 + ch7 + ch8 + ch9 +   
## ch10 + ch11 + ch12  
##   
## Df Sum of Sq RSS AIC  
## - ch10 1 0.0102 69.603 -160.81  
## - ch9 1 0.0211 69.614 -160.78  
## - ch6 1 0.1143 69.707 -160.53  
## - ch5 1 0.1713 69.764 -160.38  
## - ch4 1 0.3597 69.952 -159.88  
## - ch11 1 0.7214 70.314 -158.91  
## <none> 69.593 -158.84  
## - ch1 1 0.9627 70.555 -158.27  
## - ch12 1 1.0547 70.647 -158.03  
## - ch8 1 1.1895 70.782 -157.67  
## - ch2 1 1.3325 70.925 -157.29  
## - ch3 1 1.7545 71.347 -156.18  
## - ch7 1 4.8532 74.446 -148.23  
##   
## Step: AIC=-160.81  
## Score ~ ch1 + ch2 + ch3 + ch4 + ch5 + ch6 + ch7 + ch8 + ch9 +   
## ch11 + ch12  
##   
## Df Sum of Sq RSS AIC  
## - ch9 1 0.0148 69.618 -162.77  
## - ch6 1 0.1240 69.727 -162.48  
## - ch5 1 0.1671 69.770 -162.36  
## - ch4 1 0.3568 69.960 -161.86  
## - ch11 1 0.7129 70.316 -160.91  
## <none> 69.603 -160.81  
## - ch1 1 0.9968 70.600 -160.15  
## - ch12 1 1.0943 70.697 -159.90  
## - ch8 1 1.1822 70.785 -159.66  
## - ch2 1 1.3306 70.934 -159.27  
## - ch3 1 1.7446 71.348 -158.18  
## - ch7 1 4.8651 74.468 -150.18  
##   
## Step: AIC=-162.77  
## Score ~ ch1 + ch2 + ch3 + ch4 + ch5 + ch6 + ch7 + ch8 + ch11 +   
## ch12  
##   
## Df Sum of Sq RSS AIC  
## - ch6 1 0.1149 69.733 -164.46  
## - ch5 1 0.1730 69.791 -164.31  
## - ch4 1 0.3427 69.960 -163.85  
## <none> 69.618 -162.77  
## - ch11 1 0.8239 70.442 -162.57  
## - ch1 1 1.0339 70.652 -162.02  
## - ch12 1 1.1140 70.732 -161.80  
## - ch8 1 1.2315 70.849 -161.49  
## - ch2 1 1.3621 70.980 -161.15  
## - ch3 1 1.7339 71.352 -160.17  
## - ch7 1 4.9258 74.544 -151.99  
##   
## Step: AIC=-164.46  
## Score ~ ch1 + ch2 + ch3 + ch4 + ch5 + ch7 + ch8 + ch11 + ch12  
##   
## Df Sum of Sq RSS AIC  
## - ch5 1 0.1119 69.845 -166.16  
## - ch4 1 0.3333 70.066 -165.57  
## <none> 69.733 -164.46  
## - ch11 1 0.7805 70.513 -164.38  
## - ch1 1 0.9390 70.672 -163.96  
## - ch12 1 1.0765 70.809 -163.60  
## - ch2 1 1.2616 70.994 -163.11  
## - ch8 1 1.2683 71.001 -163.09  
## - ch3 1 1.6401 71.373 -162.12  
## - ch7 1 4.9834 74.716 -153.56  
##   
## Step: AIC=-166.16  
## Score ~ ch1 + ch2 + ch3 + ch4 + ch7 + ch8 + ch11 + ch12  
##   
## Df Sum of Sq RSS AIC  
## - ch4 1 0.3130 70.157 -167.33  
## <none> 69.845 -166.16  
## - ch11 1 0.7753 70.620 -166.10  
## - ch1 1 1.0889 70.933 -165.27  
## - ch12 1 1.1213 70.966 -165.19  
## - ch8 1 1.3720 71.217 -164.53  
## - ch2 1 1.4547 71.299 -164.31  
## - ch3 1 1.6847 71.529 -163.71  
## - ch7 1 5.5494 75.394 -153.87  
##   
## Step: AIC=-167.33  
## Score ~ ch1 + ch2 + ch3 + ch7 + ch8 + ch11 + ch12  
##   
## Df Sum of Sq RSS AIC  
## - ch11 1 0.7281 70.886 -167.40  
## <none> 70.157 -167.33  
## - ch12 1 0.9406 71.098 -166.84  
## - ch1 1 0.9749 71.132 -166.75  
## - ch8 1 1.2152 71.373 -166.12  
## - ch2 1 1.3604 71.518 -165.74  
## - ch3 1 1.5246 71.682 -165.31  
## - ch7 1 5.2756 75.433 -155.77  
##   
## Step: AIC=-167.4  
## Score ~ ch1 + ch2 + ch3 + ch7 + ch8 + ch12  
##   
## Df Sum of Sq RSS AIC  
## <none> 70.886 -167.40  
## - ch1 1 1.1667 72.052 -166.34  
## - ch12 1 1.5700 72.456 -165.30  
## - ch2 1 1.5991 72.485 -165.23  
## - ch8 1 1.6738 72.559 -165.03  
## - ch3 1 1.6783 72.564 -165.02  
## - ch7 1 5.7916 76.677 -154.71

Bayesian stepwise AIC elimination indicates that we can make our best predictions of a student’s AP Exam score using their grades from the chapter 1, 2, 3, 7, 8 and 12 tests. We noted earlier that Chapter 7 is particularly significant in terms of the material covered. It is interesting to see that Ch 12 (the last chapter of the book) is significant. Chapter 12 is about Inference for Regression and Transformations to Achieve Linearity, topics that are not typically represented heavily on the AP Exam. Given that the majority of students in AP Statistics are seniors, who have already received their acceptance to college, a possible explanation for the importance of Chapter 12 is that it provides an indicator of how strongly a student has fallen prey to “senioritis.” A student who is able to maintain their focus through the final chapter of the book is probably also going to perform better on the AP Exam.

summary(model\_AIC)

##   
## Call:  
## lm(formula = Score ~ ch1 + ch2 + ch3 + ch7 + ch8 + ch12, data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4486 -0.2589 0.0776 0.3749 1.6545   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.46733 0.43718 -5.644 6.37e-08 \*\*\*  
## ch1 0.03182 0.01849 1.721 0.086928 .   
## ch2 0.03058 0.01517 2.015 0.045382 \*   
## ch3 0.03334 0.01615 2.064 0.040413 \*   
## ch7 0.05268 0.01374 3.835 0.000174 \*\*\*  
## ch8 0.02479 0.01203 2.062 0.040681 \*   
## ch12 0.02590 0.01297 1.997 0.047365 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6275 on 180 degrees of freedom  
## (55 observations deleted due to missingness)  
## Multiple R-squared: 0.6074, Adjusted R-squared: 0.5943   
## F-statistic: 46.42 on 6 and 180 DF, p-value: < 2.2e-16

This provides us with the regression equation:

Predicted AP Exam Score = -2.46733 + 0.03182(ch1) + 0.03058(ch2) + 0.03334(ch3) + 0.05268(ch7) + 0.02479(ch8) + 0.0259(ch12).

Approximately 60% of the variability in AP Exam scores is accounted for by our model. On average, our predictions of AP Exam score are off by 0.607.

Let’s have a look at the residual plot.

ggplot(model\_AIC, aes(x=.fitted, y=.resid))+geom\_point()+geom\_hline(yintercept = 0, linetype="dashed")

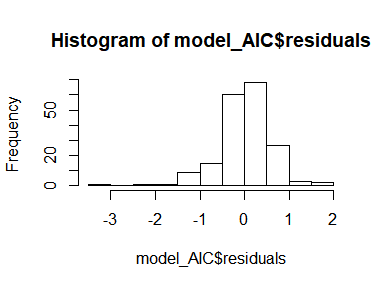


The five diagonal stripes are the result of the fact that students can only receive 1, 2, 3, 4, or 5 on the exam, with the leftmost stripe corresponding with students who obtained a 1, up through the rightmost stripe corresponding with students who obtained a 5. We can see that for most students the prediction was within 1 of their actual score.

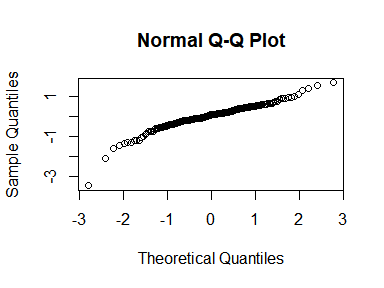
There is one notable outlier, the student who received a score 3.5 below what was predicted. This was a high-achieving senior whose chosen college would not accept AP credit, so he decided to take a nap during the exam. The student with the residual of -2 might have been in a similar situation. Otherwise, our model seems to have performed well. Adding a dummy variable to indicate whether the student had committed to a college that wouldn’t accept AP Exam credit prior to taking the exam would likely make our predictions even more accurate.

Now let’s look at the distribution of the residuals.

hist(model\_AIC$residuals)



qqnorm(model\_AIC$residuals)



The residuals are fairly close to Normal in their distribution, with the two aforementioned under-achieving students ( with residuals of -3.5 and -2) creating a slight left skew.