HW3

Claudius Taylor, Junpu Zhao, Thomas Wilson

November 03, 2018

# 1.

Consider the situation in which the following regression model has been fit to a set of data

Suppose that the hypothesis test based on Analysis of Variance leads to a strong rejection of the joint hypothesis that regression coefficients are all zero (p-value < 0.0001).

Furthermore, suppose that each individual t-test of the form H0: bi = 0 fails to reject the null for each value of i, where i=1,…,3.

Describe why this is likely to have occurred and what plots and which statistics you would look at to diagnose the problem.

, , and are correlated with each other. While there is at least one coefficient that is not zero, non of the three are significantly different from zero in the presence of each other. We can confirm this with a scatterplot matrix and correlation coefficients.

# 2.

Data taken from Howell (2002). “A number of years ago, the student association of a large university published an evaluation of several hundred courses taught during the preceding semester. Students in each course had completed a questionnaire in which they rated a number of different aspects of the course on a 5-point scale (1 = very bad to 5 = excellent)”.

Five variables obtained were: \* overall – overall rating of the course. \* teach – rating of teaching skills of the instructor. \* exams – quality of tests and exams \* knowledge – rating of the instructor’s knowledge of the material \* grade - student’s anticipated grade for the course (1=F to 5=A) \* enroll – enrollment for the course

## overall teach exams knowledge grade enroll  
## 1: 3.4 3.8 3.8 4.5 3.5 21  
## 2: 2.9 2.8 3.2 3.8 3.2 50  
## 3: 2.6 2.2 1.9 3.9 2.8 800  
## 4: 3.8 3.5 3.5 4.1 3.3 221  
## 5: 3.0 3.2 2.8 3.5 3.2 7  
## 6: 2.5 2.7 3.8 4.2 3.2 108

## a

Enter the variables teach, exams, knowledge, grade, and enroll into a multiple regression model predicting scores for overall. Display the regression output.

##   
## Call:  
## lm(formula = overall ~ ., data = data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.45621 -0.23256 -0.03937 0.33722 0.52323   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.2399644 1.4740674 -0.841 0.4220   
## teach 0.6350523 0.2139652 2.968 0.0158 \*  
## exams -0.4134565 0.6471302 -0.639 0.5388   
## knowledge 0.5236860 0.9665941 0.542 0.6011   
## grade 0.5441992 0.6517025 0.835 0.4253   
## enroll -0.0003415 0.0014398 -0.237 0.8178   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4056 on 9 degrees of freedom  
## Multiple R-squared: 0.7046, Adjusted R-squared: 0.5405   
## F-statistic: 4.294 on 5 and 9 DF, p-value: 0.02831

## b

Create a table of R^2 adjusted, AIC, AICC, and BIC for the best subset of each size.

Identify the optimal model or models from the approach based on all possible subsets.

## # A tibble: 5 x 5  
## k `max(adj.r.squared)` `min(AIC)` `min(AICc)` `min(BIC)`  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 0.564 18.6 18.9 20.7  
## 2 2 0.618 17.4 18.4 20.2  
## 3 3 0.607 18.5 20.7 22.0  
## 4 4 0.584 19.9 23.9 24.2  
## 5 5 0.541 21.8 28.5 26.8

## # A tibble: 1 x 6  
## model k adj.r.squared AIC AICc BIC  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 overall ~ teach+grade 2 0.618 17.4 18.4 20.2

## # A tibble: 1 x 6  
## model k adj.r.squared AIC AICc BIC  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 overall ~ grade+teach 2 0.618 17.4 18.4 20.2

## # A tibble: 1 x 6  
## model k adj.r.squared AIC AICc BIC  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 overall ~ grade+teach 2 0.618 17.4 18.4 20.2

## # A tibble: 1 x 6  
## model k adj.r.squared AIC AICc BIC  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 overall ~ teach+grade 2 0.618 17.4 18.4 20.2

## c

Use the Forward method to determine the regression equation when starting with the same predictor variables listed in a. Please describe the steps R went through in generating its regression equation.

At each stage of the process list the variable that was entered or removed from the equation and the R Square for the regression equation up to that point.

Report the final version of the regression equation.

## Start: AIC=-14.44  
## overall ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + teach 1 2.98386 2.0295 -26.004  
## + knowledge 1 1.73097 3.2824 -18.792  
## + grade 1 1.68492 3.3284 -18.583  
## + exams 1 1.26642 3.7469 -16.807  
## <none> 5.0133 -14.439  
## + enroll 1 0.42414 4.5892 -13.765  
##   
## Step: AIC=-26  
## overall ~ teach  
##   
## Df Sum of Sq RSS AIC  
## + grade 1 0.38881 1.6407 -27.194  
## + knowledge 1 0.29895 1.7305 -26.394  
## <none> 2.0295 -26.004  
## + enroll 1 0.01496 2.0145 -24.115  
## + exams 1 0.01217 2.0173 -24.094  
##   
## Step: AIC=-27.19  
## overall ~ teach + grade  
##   
## Df Sum of Sq RSS AIC  
## <none> 1.6407 -27.194  
## + enroll 1 0.091890 1.5488 -26.059  
## + exams 1 0.091096 1.5496 -26.051  
## + knowledge 1 0.024660 1.6160 -25.421

##   
## Call:  
## lm(formula = overall ~ teach + grade, data = data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.57599 -0.28129 -0.04803 0.28262 0.56055   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.6029 1.0757 -0.561 0.58543   
## teach 0.5888 0.1676 3.513 0.00427 \*\*  
## grade 0.6041 0.3582 1.686 0.11753   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3698 on 12 degrees of freedom  
## Multiple R-squared: 0.6727, Adjusted R-squared: 0.6182   
## F-statistic: 12.33 on 2 and 12 DF, p-value: 0.001228

Forward selection process: 1. starting with the null model. 2. fit a separate model adding each potential new feature. 3. select the model which minimizes AIC. 4. repeat 2-3 until AIC cannot be lowered.

## d

Repeat Part c, except use the Backward method (i.e., describe each step R went through).

Is the solution different from the one you got using the Forward method?

## Start: AIC=-22.73  
## overall ~ teach + exams + knowledge + grade + enroll  
##   
## Df Sum of Sq RSS AIC  
## - enroll 1 0.00925 1.4901 -24.638  
## - knowledge 1 0.04830 1.5292 -24.250  
## - exams 1 0.06717 1.5480 -24.066  
## - grade 1 0.11473 1.5956 -23.612  
## <none> 1.4809 -22.731  
## - teach 1 1.44946 2.9303 -14.494  
##   
## Step: AIC=-24.64  
## overall ~ teach + exams + knowledge + grade  
##   
## Df Sum of Sq RSS AIC  
## - knowledge 1 0.05945 1.5496 -26.051  
## - exams 1 0.12588 1.6160 -25.421  
## - grade 1 0.18768 1.6778 -24.858  
## <none> 1.4901 -24.638  
## - teach 1 1.54723 3.0374 -15.956  
##   
## Step: AIC=-26.05  
## overall ~ teach + exams + grade  
##   
## Df Sum of Sq RSS AIC  
## - exams 1 0.09110 1.6407 -27.194  
## <none> 1.5496 -26.051  
## - grade 1 0.46773 2.0173 -24.094  
## - teach 1 1.66039 3.2100 -17.127  
##   
## Step: AIC=-27.19  
## overall ~ teach + grade  
##   
## Df Sum of Sq RSS AIC  
## <none> 1.6407 -27.194  
## - grade 1 0.38881 2.0295 -26.004  
## - teach 1 1.68774 3.3284 -18.583

##   
## Call:  
## lm(formula = overall ~ teach + grade, data = data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.57599 -0.28129 -0.04803 0.28262 0.56055   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.6029 1.0757 -0.561 0.58543   
## teach 0.5888 0.1676 3.513 0.00427 \*\*  
## grade 0.6041 0.3582 1.686 0.11753   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3698 on 12 degrees of freedom  
## Multiple R-squared: 0.6727, Adjusted R-squared: 0.6182   
## F-statistic: 12.33 on 2 and 12 DF, p-value: 0.001228

Backward selection process: 1. starting with the full model. 2. fit a separate model removing each potential feature. 3. select the model which minimizes AIC. 4. repeat 2-3 until AIC cannot be lowered.

In this case, forward and backward selection agree on the best model as overal ~ teach+grade

## e

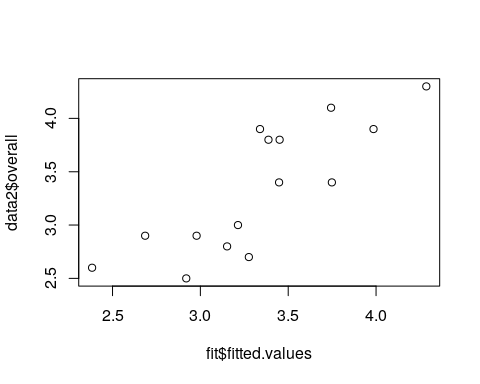
Decide on the model you would recommend.

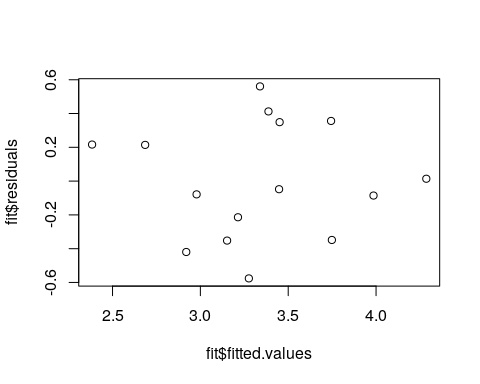
At this point, describe and examine the assumptions of multiple linear regression.

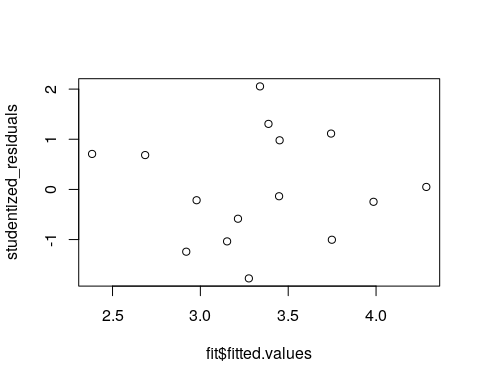
If any assumptions are violated – discuss what steps would/should be performed.

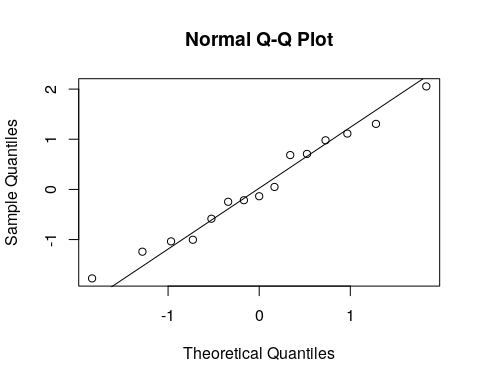
overall ~ teach + grade is best model because it maximizes R^2 while minimzing AIC, AICc, and BIC.

##   
## Call:  
## lm(formula = overall ~ teach + grade, data = data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.57599 -0.28129 -0.04803 0.28262 0.56055   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.6029 1.0757 -0.561 0.58543   
## teach 0.5888 0.1676 3.513 0.00427 \*\*  
## grade 0.6041 0.3582 1.686 0.11753   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3698 on 12 degrees of freedom  
## Multiple R-squared: 0.6727, Adjusted R-squared: 0.6182   
## F-statistic: 12.33 on 2 and 12 DF, p-value: 0.001228









Visually, there is a linear relationship between the actual and predicted values. The residuals are normally distributed with a mean of 0 and a constant variance.

We consider modeling Credit Card Balances.

Data are available on Balances and the following potential predictor variables: 1. Income 2. Limit 3. Rating 4. Cards (Number of Credit Cards) 5. Age (Years) 6. Education (Years) 7. Student (1=Student and 0=Non-student)

Use R to do the following tasks.

Copy the data in excel and run the following syntax my\_data <- read.table(file = “clipboard”, sep = “”, header=TRUE)

Paste the numerical output and plots into a Word document.

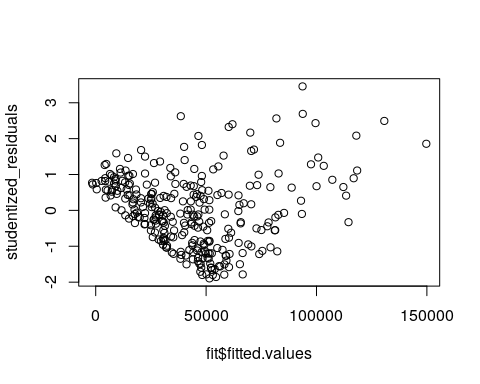
## a

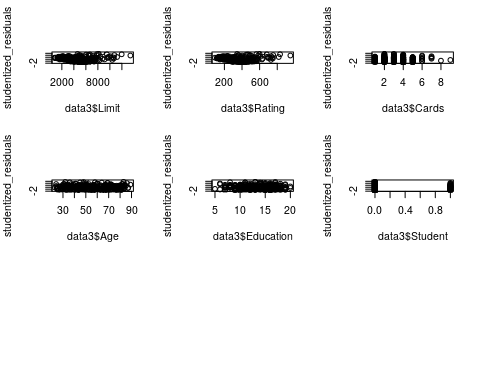
Fit a model to predict Income from Limit, Rating, Cards, Age, Education, and Student. Please do not include any quadratic nor any interaction terms. Ensure that you find the variance inflation factors (VIF).

## Limit Rating Cards Age Education Student   
## 211.862264 212.695215 1.449675 1.015229 1.027733 1.027467

## b

Plot the studentized residuals against the predicted values and each numerical predictor.

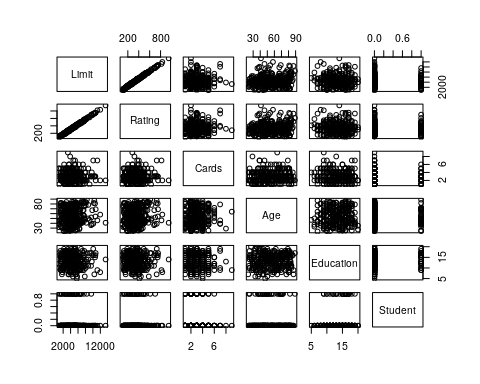




## c

Plot the numerical predictors against each other and obtain the correlations amongst these predictors.

On the basis of the output you produced in (a), (b), and (c) answer the following questions:



## I

Decide whether the assumption of constant error variance is a reasonable one for the fitted model. Give reasons to support your answer.

Based “megaphone” pattern in scatterplots of residual vs limit and residual vs rating, variance in the residual is increasing with both rating and limit.

## II

Decide whether multicollinearity is an issue for the fitted model. Give a reason to support your answer.

based on scatter plot of limit vs rating, there is a strong linear relationship between limit and rating.

## III

Interpret the estimated coefficients. Decide whether these coefficients are statistically significant. Give a reason to support your answer.

##   
## Call:  
## lm(formula = Income ~ Limit + Rating + Cards + Age + Education +   
## Student, data = data3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -39059 -16552 -1364 13912 69645   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -21539.229 7980.708 -2.699 0.00736 \*\*  
## Limit -3.507 8.018 -0.437 0.66217   
## Rating 229.214 119.630 1.916 0.05633 .   
## Cards -1966.809 1045.792 -1.881 0.06101 .   
## Age 137.156 70.884 1.935 0.05396 .   
## Education -63.812 378.877 -0.168 0.86637   
## Student 1412.195 3883.806 0.364 0.71641   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 20760 on 293 degrees of freedom  
## Multiple R-squared: 0.619, Adjusted R-squared: 0.6112   
## F-statistic: 79.35 on 6 and 293 DF, p-value: < 2.2e-16

Based on the p-value of the F-statistic, under the null hypothesis that all coefficients are truely zero, we would expect and f-stat this extreme only 2.2e-14 % of the time. This is strong evidence that there is at lease one non-zero coefficient. Removing the lease significant factors reveals a significant coefficient between income and Rating. A unit change in rating is associated with an increase of 8.212 in Income.

##   
## Call:  
## lm(formula = Income ~ Rating + Cards + Age, data = data3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -39628 -16498 -927 13608 70082   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -20987.812 5310.937 -3.952 9.71e-05 \*\*\*  
## Rating 177.080 8.212 21.564 < 2e-16 \*\*\*  
## Cards -1733.609 866.832 -2.000 0.0464 \*   
## Age 136.264 70.364 1.937 0.0538 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 20670 on 296 degrees of freedom  
## Multiple R-squared: 0.6186, Adjusted R-squared: 0.6147   
## F-statistic: 160 on 3 and 296 DF, p-value: < 2.2e-16