w271: Homework 2

Krysten Thompson

library(knitr)  
opts\_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)

In the live session of week 2, we discussed data analysis, EDA, and binary logistic regression. This homework is designed to review and practice these concepts and techniques. It also covers variable transformation and associated concepts covered in week 3.

For this homework, you will use the dataset *“data\_wk02.csv”*, which contains a small sample of graduate school admission data from a university. The variables are specificed below:

1. admit - the depenent variable that takes two values: where denotes *admitted* and denotes *not admitted*.
2. gre - GRE score
3. gpa - College GPA
4. rank - rank in college major

Suppose you are hired by the University’s Admission Committee and are charged to analyze this data to quantify the effect of GRE, GPA, and college rank on admission probability. We will conduct this analysis by answering the follwing questions:

**Question 1:** Examine the data and conduct EDA.

library(car) #importing libraries

## Loading required package: carData

library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':  
##   
## recode

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

library(knitr)

admission <- read.csv("data\_wk02.csv") #import data

str(admission) #Want to understand variable data types, dimensions, etc.

## 'data.frame': 400 obs. of 5 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ admit: int 0 1 1 1 0 1 1 0 1 0 ...  
## $ gre : int 380 660 800 640 520 760 560 400 540 700 ...  
## $ gpa : num 3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...  
## $ rank : int 3 3 1 4 4 2 1 2 3 2 ...

head(admission) #look at first 5 rows of dataframe

## X admit gre gpa rank  
## 1 1 0 380 3.61 3  
## 2 2 1 660 3.67 3  
## 3 3 1 800 4.00 1  
## 4 4 1 640 3.19 4  
## 5 5 0 520 2.93 4  
## 6 6 1 760 3.00 2

#confirm var types and note that I need to change gpa to "int" for modeling  
sapply(admission, class)

## X admit gre gpa rank   
## "integer" "integer" "integer" "numeric" "integer"

#review missing data, range per variable, number of distinct values per variable  
describe(admission)

## admission   
##   
## 5 Variables 400 Observations  
## ---------------------------------------------------------------------------  
## X   
## n missing distinct Info Mean Gmd .05 .10   
## 400 0 400 1 200.5 133.7 20.95 40.90   
## .25 .50 .75 .90 .95   
## 100.75 200.50 300.25 360.10 380.05   
##   
## lowest : 1 2 3 4 5, highest: 396 397 398 399 400  
## ---------------------------------------------------------------------------  
## admit   
## n missing distinct Info Sum Mean Gmd   
## 400 0 2 0.65 127 0.3175 0.4345   
##   
## ---------------------------------------------------------------------------  
## gre   
## n missing distinct Info Mean Gmd .05 .10   
## 400 0 26 0.997 587.7 131.2 399 440   
## .25 .50 .75 .90 .95   
## 520 580 660 740 800   
##   
## lowest : 220 300 340 360 380, highest: 720 740 760 780 800  
## ---------------------------------------------------------------------------  
## gpa   
## n missing distinct Info Mean Gmd .05 .10   
## 400 0 132 1 3.39 0.4351 2.758 2.900   
## .25 .50 .75 .90 .95   
## 3.130 3.395 3.670 3.940 4.000   
##   
## lowest : 2.26 2.42 2.48 2.52 2.55, highest: 3.95 3.97 3.98 3.99 4.00  
## ---------------------------------------------------------------------------  
## rank   
## n missing distinct Info Mean Gmd   
## 400 0 4 0.91 2.485 1.038   
##   
## Value 1 2 3 4  
## Frequency 61 151 121 67  
## Proportion 0.152 0.378 0.302 0.168  
## ---------------------------------------------------------------------------

#calculate the count and proportion of my predictor variable; note almost 7:3 split   
#for Not Admitted to Admitted  
table(admission$admit)

##   
## 0 1   
## 273 127

prop.table(table(admission$admit))

##   
## 0 1   
## 0.6825 0.3175

prop.table(table(admission$rank)) #review proportion of rank values

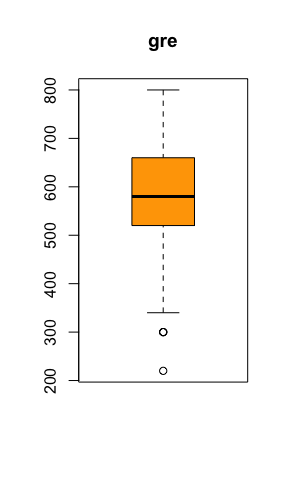
##   
## 1 2 3 4   
## 0.1525 0.3775 0.3025 0.1675

summary(admission) #review high-level statistics for independent variables

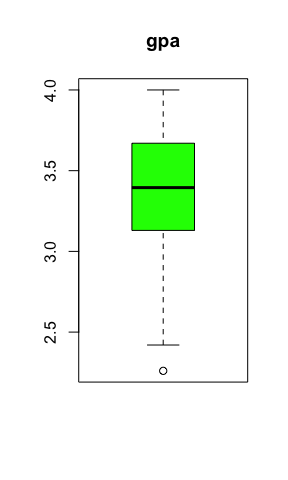
## X admit gre gpa   
## Min. : 1.0 Min. :0.0000 Min. :220.0 Min. :2.260   
## 1st Qu.:100.8 1st Qu.:0.0000 1st Qu.:520.0 1st Qu.:3.130   
## Median :200.5 Median :0.0000 Median :580.0 Median :3.395   
## Mean :200.5 Mean :0.3175 Mean :587.7 Mean :3.390   
## 3rd Qu.:300.2 3rd Qu.:1.0000 3rd Qu.:660.0 3rd Qu.:3.670   
## Max. :400.0 Max. :1.0000 Max. :800.0 Max. :4.000   
## rank   
## Min. :1.000   
## 1st Qu.:2.000   
## Median :2.000   
## Mean :2.485   
## 3rd Qu.:3.000   
## Max. :4.000

Note the outliers for gre and gpa. There appear to be 2 outliers in gre data and one outlier in gpa.

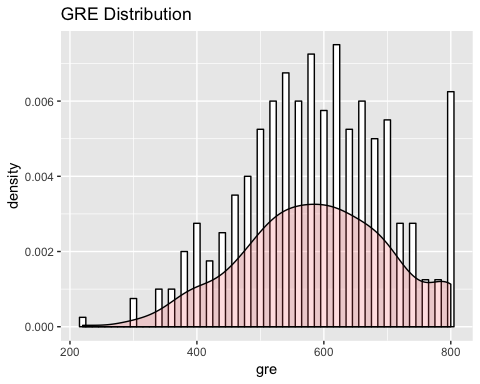
boxplot(admission$gre, main="gre", col=c("orange"))



boxplot(admission$gpa, main="gpa", col=c("green"))

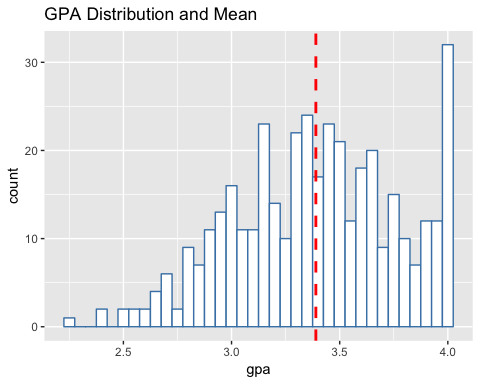
 Review spread of GRE scores with density line. We can see the outlier at ~ 225 and high frequency of perfect scores (800).

ggplot(admission, aes(x=gre)) + ggtitle("GRE Distribution") +  
 geom\_histogram(aes(y=..density..),   
 binwidth=10,  
 colour="black", fill="white") +  
 geom\_density(alpha=.2, fill="#FF6666")



Reviewing the spread of Grade Point Average’s. Red dotted line indicates mean for GPA. Slight left-skewed histogram.

ggplot(admission, aes(x=gpa)) + ggtitle("GPA Distribution and Mean") +  
 geom\_histogram(binwidth=.05, colour="steel blue", fill="white") +  
 geom\_vline(aes(xintercept=mean(gpa)),   
 color="red", linetype="dashed", size=1)



Comparing GRE Score to being admitted. Interesting that the spread of GRE scores is similar between Not Admitted and Admitted. Even students who receive perfect or near pefect score are sometimes not admitted.

ggplot(admission, aes(factor(admit), gre)) +  
 geom\_boxplot(aes(fill = factor(admit))) +   
 geom\_jitter() +  
 ggtitle("Admission Based on GRE Score") +   
 theme(plot.title = element\_text(lineheight=1, face="bold"))



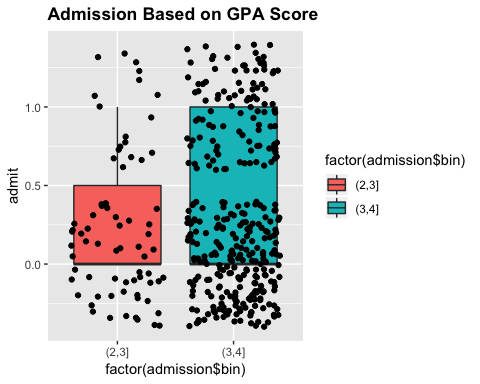
In order to view GPA against Admit, I first binned GPA into three buckets of GPA values: 0-2, 2-3, and 3+.

admission$bin <- cut(admission$gpa, c(0,2,3,4))  
table(admission$bin)

##   
## (0,2] (2,3] (3,4]   
## 0 71 329

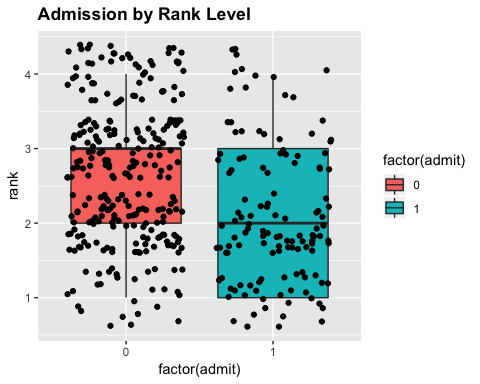
We can see that the spread of Not Admitted to Admitted is similar for all GPA’s, even among students with low GPA’s.

ggplot(admission, aes(factor(admission$bin), admit)) +  
 geom\_boxplot(aes(fill = factor(admission$bin))) +   
 geom\_jitter() +  
 ggtitle("Admission Based on GPA Score") +   
 theme(plot.title = element\_text(lineheight=1, face="bold"))

 Below we see the range of Rank values is similarly spread between Not Admitted and Admitted.

Since the spread of the predictor variables is similar across Not Admitted and Admitted, it begs the question how much does each variable contribute when determining probability of being admitted to graduate school or not?

ggplot(admission, aes(factor(admit), rank)) +   
 geom\_boxplot(aes(fill = factor(admit))) +   
 geom\_jitter() +  
 ggtitle("Admission by Rank Level") +   
 theme(plot.title = element\_text(lineheight=1, face="bold"))



**Question 2:** Estimate a binary logistic regression using the following set of explanatory variables: , , , , , and , where denotes the interaction between and variables.

This is the model that will be used.

admission$gpa <- as.integer(admission$gpa) #change gpa to var type "int" for modeling  
  
sapply(admission, class)

## X admit gre gpa rank bin   
## "integer" "integer" "integer" "integer" "integer" "factor"

First, Rank needs to be factored for modeling so that we can see the estimates for Rank 1, 2, 3, or 4.

#Originally I factored Rank so that the model would be more objective to the Rank values of 1-4;  
#however, the output looked terrible so I commented it out and re-ran model  
  
#admission$rank <- factor(admission$rank)

mod.a <- glm(formula = admit ~ gre + gpa + rank + I(gpa^2) + I(gre^2) + gre:gpa, family = binomial(link=logit), data = admission)  
  
summary(mod.a)

##   
## Call:  
## glm(formula = admit ~ gre + gpa + rank + I(gpa^2) + I(gre^2) +   
## gre:gpa, family = binomial(link = logit), data = admission)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4295 -0.8917 -0.6615 1.1967 2.1858   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.791e+00 3.994e+00 -0.699 0.485   
## gre 5.375e-03 9.836e-03 0.546 0.585   
## gpa 3.638e-01 1.765e+00 0.206 0.837   
## rank -5.506e-01 1.264e-01 -4.355 1.33e-05 \*\*\*  
## I(gpa^2) -6.950e-02 3.133e-01 -0.222 0.824   
## I(gre^2) -3.310e-06 7.758e-06 -0.427 0.670   
## gre:gpa 5.119e-04 2.346e-03 0.218 0.827   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 499.98 on 399 degrees of freedom  
## Residual deviance: 463.75 on 393 degrees of freedom  
## AIC: 477.75  
##   
## Number of Fisher Scoring iterations: 4

**ANSWER**

This model seems to include the “kitchen sink” in terms of independent variables. The only variable with a p-value of < is Rank. Given the significantly high p-values for the other variables makes me want to run a “stepped” version of the model, meaning I would model using only one variable at a time, then add +1.

**Question 3:** Test the hypothesis that GRE has no effect on admission using the likelihood ratio test.

The forumla that will be used:

mod.Ho <- glm(formula = admit ~ gpa + rank, family = binomial(link=logit), data = admission)  
  
mod.Ha <- glm(formula = admit ~ gre + gpa + rank, family = binomial(link=logit), data = admission)  
  
summary(mod.Ha)

##   
## Call:  
## glm(formula = admit ~ gre + gpa + rank, family = binomial(link = logit),   
## data = admission)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4454 -0.8914 -0.6554 1.1846 2.1464   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.993563 0.906898 -2.198 0.02793 \*   
## gre 0.002914 0.001055 2.762 0.00575 \*\*   
## gpa 0.273859 0.249465 1.098 0.27230   
## rank -0.547939 0.125905 -4.352 1.35e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 499.98 on 399 degrees of freedom  
## Residual deviance: 463.98 on 396 degrees of freedom  
## AIC: 471.98  
##   
## Number of Fisher Scoring iterations: 3

Forumula for Likelihood Ration Tests:

Anova(mod.Ha, test="LR")

## Analysis of Deviance Table (Type II tests)  
##   
## Response: admit  
## LR Chisq Df Pr(>Chisq)   
## gre 7.8777 1 0.005005 \*\*   
## gpa 1.2197 1 0.269423   
## rank 20.3062 1 6.599e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(mod.Ho, mod.Ha, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: admit ~ gpa + rank  
## Model 2: admit ~ gre + gpa + rank  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 397 471.86   
## 2 396 463.98 1 7.8777 0.005005 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**ANSWER**

At 95% confidence level, the null hypothesis that GRE scores have no role in determining whether a student is admitted to grad school or not should be rejected. There is strong evidence to support that GRE is important when determining Admitted vs Not Admitted (given that the model includes GPA and Rank).

**Question 4:** What is the estimated effect of college GPA on admission?

mod.Ho1 <- glm(formula = admit ~ gre + rank, family = binomial(link=logit), data = admission)  
  
mod.Ha1 <- glm(formula = admit ~ gre + gpa + rank, family = binomial(link=logit), data = admission)  
  
summary(mod.Ha1)

##   
## Call:  
## glm(formula = admit ~ gre + gpa + rank, family = binomial(link = logit),   
## data = admission)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4454 -0.8914 -0.6554 1.1846 2.1464   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.993563 0.906898 -2.198 0.02793 \*   
## gre 0.002914 0.001055 2.762 0.00575 \*\*   
## gpa 0.273859 0.249465 1.098 0.27230   
## rank -0.547939 0.125905 -4.352 1.35e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 499.98 on 399 degrees of freedom  
## Residual deviance: 463.98 on 396 degrees of freedom  
## AIC: 471.98  
##   
## Number of Fisher Scoring iterations: 3

Anova(mod.Ha1, test="LR")

## Analysis of Deviance Table (Type II tests)  
##   
## Response: admit  
## LR Chisq Df Pr(>Chisq)   
## gre 7.8777 1 0.005005 \*\*   
## gpa 1.2197 1 0.269423   
## rank 20.3062 1 6.599e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(mod.Ho1, mod.Ha1, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: admit ~ gre + rank  
## Model 2: admit ~ gre + gpa + rank  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 397 465.20   
## 2 396 463.98 1 1.2197 0.2694

**ANSWER** GPA has some effect on admission but not much. A one point increase in GPA results in an increase of .

**Question 5:** Construct the confidence interval for the admission probability for the students with , , and .

Formulas beings used for Wald Confidence Interval:

newdf <- data.frame(gpa = 3.3, gre = 720, rank = 1)  
  
predict(object = mod.Ha, newdata = newdf, type = "link")

## 1   
## 0.4601136

predict(object = mod.Ha, newdata = newdf, type = "response")

## 1   
## 0.6130411

alpha <- 0.05  
  
lin.pred <- predict(object = mod.Ha, newdata = newdf, type = "link", se = TRUE)  
  
lin.pred

## $fit  
## 1   
## 0.4601136   
##   
## $se.fit  
## [1] 0.2344009  
##   
## $residual.scale  
## [1] 1

pi.hat <- exp(lin.pred$fit) / (1 + exp(lin.pred$fit))  
  
CI.lin.pred <- lin.pred$fit + qnorm(p = c(alpha/2, 1-alpha/2)) \* lin.pred$se.fit  
  
CI.pi <- exp(CI.lin.pred) / (1 + exp(CI.lin.pred))  
  
(data.frame(newdf, pi.hat, lower = CI.pi[1], upper = CI.pi[2]))

## gpa gre rank pi.hat lower upper  
## 1 3.3 720 1 0.6130411 0.5001741 0.7149465

**ANSWER** For students with , , and , their admission probability is . The confidence interval is between to .