College Admissions Analysis

**1. What variables do you think are pertinent to predicting the two dependent variables of interest. Create a table with these predictors and your reasoning justifying why you think these variables are appropriate.**

Predicting Enrollment:

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Alt Hyp** | **Reasoning** |
| Private | β < 0 | Private colleges are typically more selective and expensive, two factors that would reduce enrollment. |
| AppsRec | β > 0 | Colleges that receive more applications will be able to offer places to more prospective students, and by extension more students will enroll. |
| Acceptance | β > 0 | Colleges that accept more students have a higher capacity for enrollments. |
| Top10perc | β < 0 | The top high school students will choose to go to more selective colleges which have lower enrollments. |
| Top25perc | β < 0 |
| FulltimeUG | β > 0 | Colleges with higher numbers of undergraduates, full or part-time, necessarily have higher enrollments. |
| ParttimeUG | β > 0 |
| OutstateTuition | β < 0 | Higher out of state tuition will be a discouraging factor for prospective students, reducing enrollment. |
| RoomAndBoard | β < 0 | Higher cost of room and board will be a discouraging factor for prospective students, reducing enrollment. |
| Books | β < 0 | Higher cost of books will be a discouraging factor for prospective students, reducing enrollment. |
| PersonalExpense | β < 0 | Higher personal expenses will be a discouraging factor for prospective students, reducing enrollment. |
| PhD | β = 0 | It’s unlikely that the percentage of faculty with a certain level of degree has an effect on enrollment. |
| TerminalDegree | β = 0 |
| StudentFacultyRatio | β > 0 | A higher student to faculty ratio could be due to high enrollment numbers when there is a lag in hiring faculty. In addition, larger schools typically have large classroom sizes. |
| AlumniDonate | β = 0 | It’s unlikely that the percentage of alumni who choose to donate has any effect on enrollment. |
| InstrExpend | β = 0 | It’s unlikely that a per capita measure of spending on students is related to enrollment. |
| GradRate | β < 0 | Colleges with higher graduation rates likely have lower enrollments because their graduation rate is a reflection of selecting only the most qualified students. |

Predicting GradRate:

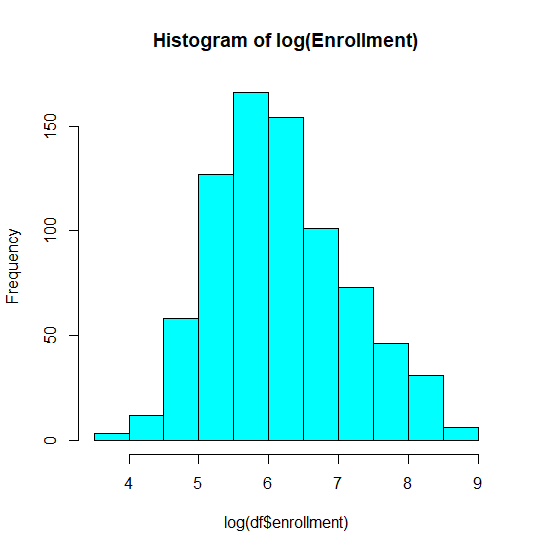
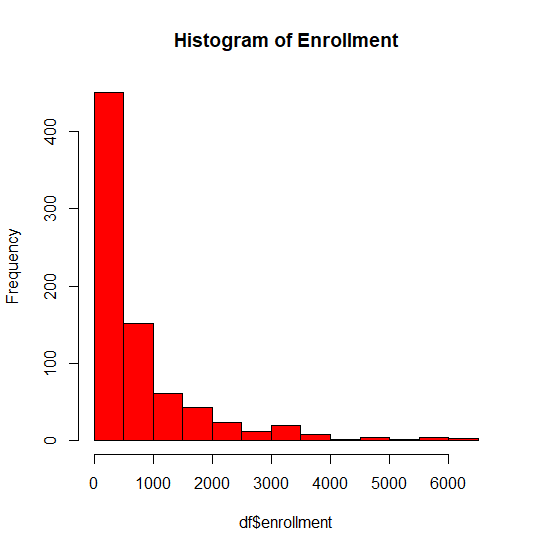
|  |  |  |
| --- | --- | --- |
| **Predictor** | **Alt Hyp** | **Reasoning** |
| Private | β > 0 | Tuition at private colleges is usually higher, so students are incentivized to graduate on time. |
| AppsRec | β < 0 | Colleges that receive a lot of applications are more likely to be larger institutions, which are more likely to admit students who will not graduate on time. |
| Acceptance | β < 0 | Less selective colleges admit more students who are unprepared to graduate on time. |
| Enrollment | β < 0 | Higher enrollment typically means less selectivity, which is likely to result in lower graduation rates. |
| Top10perc | β > 0 | Students that performed well in high school are likely to perform well in college and graduate on time. |
| Top25perc | β > 0 |
| FulltimeUG | β < 0 | Colleges with a higher number of undergraduate students are likely to be less selective, lowering the graduation rate. |
| ParttimeUG | β < 0 |
| OutstateTuition | β > 0 | Higher tuition is an incentive for students to graduate on time. |
| RoomAndBoard | β > 0 | Higher room and board costs is an incentive for students to graduate on time. |
| Books | β > 0 | Higher book costs is an incentive for students to graduate on time. |
| PersonalExpense | β > 0 | Higher personal expenses are an incentive for students to graduate on time. |
| PhD | β > 0 | More selective colleges are more likely to employ faculty with advanced degrees. As mentioned above, selectivity is likely positively correlated with graduation rate. |
| TerminalDegree | β > 0 |
| StudentFacultyRatio | β < 0 | A lower student to faculty ratio means that students get more individual attention from professors, likely improving their graduation rate. It is also likely correlated with selectivity. |
| AlumniDonate | β > 0 | Students that graduate on time are probably more likely to become donors in the future since their job prospects are improved and gratitude to the college is increased. |
| InstrExpend | β > 0 | Colleges that spend more per student are likely improving their educational outcomes, resulting in improved graduation rate. |

**2. Specify the appropriate model. Submit both your R code and your output. Be sure to test for appropriate assumptions and adjust your regression model as necessary.**

Preparation for modelling requires converting the “Private” variable to a factor.

1. **Modeling Enrollment**

A histogram of enrollment shows that a preponderance of colleges enroll fewer than 1000 students. Enrollment is certainly not normally distributed, and seems to fit a Poisson distribution more closely. Applying a logarithmic function to the



Applying a logarithmic function to enrollment will allow us to create a linear regression model.

The first linear model will include all variables except college to see the impact of each.

lm(formula = log(df$enrollment) ~ . - college, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.29669 -0.24994 0.04463 0.28654 1.81886

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.298e+00 1.758e-01 24.453 < 2e-16 \*\*\*

privateYes -4.713e-01 5.996e-02 -7.860 1.32e-14 \*\*\*

appsrec -1.854e-05 1.543e-05 -1.201 0.22994

acceptance 1.392e-04 2.741e-05 5.079 4.77e-07 \*\*\*

top10perc 4.250e-03 2.501e-03 1.699 0.08966 .

top25perc 4.685e-04 1.935e-03 0.242 0.80872

fulltimeug 6.400e-05 8.805e-06 7.269 9.04e-13 \*\*\*

parttimeug 2.572e-05 1.386e-05 1.855 0.06399 .

outstatetuition 1.440e-05 8.318e-06 1.731 0.08379 .

roomandboard -3.175e-05 2.090e-05 -1.519 0.12919

books 2.345e-04 1.029e-04 2.280 0.02287 \*

personalexpense 6.092e-05 2.722e-05 2.238 0.02550 \*

phd 3.465e-03 2.006e-03 1.728 0.08446 .

terminaldegree 1.493e-03 2.198e-03 0.679 0.49723

studentfacultyratio 3.572e-02 5.617e-03 6.360 3.48e-10 \*\*\*

alumnidonate -3.717e-03 1.762e-03 -2.109 0.03523 \*

instrexpend 1.570e-05 5.459e-06 2.876 0.00414 \*\*

gradrate 6.512e-03 1.279e-03 5.092 4.47e-07 \*\*\*

PrivateYes NA NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4493 on 759 degrees of freedom

Multiple R-squared: 0.7828, Adjusted R-squared: 0.778

F-statistic: 160.9 on 17 and 759 DF, p-value: < 2.2e-16

The first ordinary least squares regression has an adjusted R^2 value of .778 and shows that 9 of the variables have a p-value of under .05.

A similar model can be created using a generalized linear model (maximum likelihood estimation) with the Gaussian family.

emle1 = glm(log(df$enrollment)~.-college, data=df, family=gaussian)

========================================================

Dependent variable:

------------------------------------

enrollment)

OLS normal

(1) (2)

--------------------------------------------------------

privateYes -0.471\*\*\* -0.471\*\*\*

(0.060) (0.060)

appsrec -0.00002 -0.00002

(0.00002) (0.00002)

acceptance 0.0001\*\*\* 0.0001\*\*\*

(0.00003) (0.00003)

top10perc 0.004\* 0.004\*

(0.003) (0.003)

top25perc 0.0005 0.0005

(0.002) (0.002)

fulltimeug 0.0001\*\*\* 0.0001\*\*\*

(0.00001) (0.00001)

parttimeug 0.00003\* 0.00003\*

(0.00001) (0.00001)

outstatetuition 0.00001\* 0.00001\*

(0.00001) (0.00001)

roomandboard -0.00003 -0.00003

(0.00002) (0.00002)

books 0.0002\*\* 0.0002\*\*

(0.0001) (0.0001)

personalexpense 0.0001\*\* 0.0001\*\*

(0.00003) (0.00003)

phd 0.003\* 0.003\*

(0.002) (0.002)

terminaldegree 0.001 0.001

(0.002) (0.002)

studentfacultyratio 0.036\*\*\* 0.036\*\*\*

(0.006) (0.006)

alumnidonate -0.004\*\* -0.004\*\*

(0.002) (0.002)

instrexpend 0.00002\*\*\* 0.00002\*\*\*

(0.00001) (0.00001)

gradrate 0.007\*\*\* 0.007\*\*\*

(0.001) (0.001)

PrivateYes

Constant 4.298\*\*\* 4.298\*\*\*

(0.176) (0.176)

--------------------------------------------------------

Observations 777 777

R2 0.783

Adjusted R2 0.778

Log Likelihood -472.740

Akaike Inf. Crit. 981.479

Residual Std. Error 0.449 (df = 759)

F Statistic 160.943\*\*\* (df = 17; 759)

========================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

A stargazer output shows that the two models are identical, but the second model’s measures of Log Likelihood and AIC allow it to be compared to other non-linear models.

Since the distribution of enrollment resembles a Poisson distribution, we can try to remove the log function and apply a GLM with the Poisson family.

epoisson1 = glm(df$enrollment~.-college, data=df, family=poisson)

With a Poisson regression we need to check for overdispersion.

dispersiontest(epoisson1)

Overdispersion test

data: epoisson1

z = 11.33, p-value < 2.2e-16

alternative hypothesis: true dispersion is greater than 1

sample estimates:

dispersion

134.1082

A p-value of less than .05 and a dispersion far away from 1 indicates that there is overdispersion.

This suggests that we should try a negative binomial regression.

enb <- glm.nb(df$enrollment~.-college, data=df)

confint(enb) #confidence intervals

exp(confint(enb)) #exponential function

exp(cbind(coef=coef(enb), confint(enb))) #Effect size

coef 2.5 % 97.5 %

(Intercept) 100.0810860 72.0210773 139.1842897

privateYes 0.6962020 0.6244218 0.7762054

appsrec 1.0000041 0.9999748 1.0000340

acceptance 1.0001679 1.0001151 1.0002203

top10perc 1.0031874 0.9985572 1.0078448

top25perc 1.0013245 0.9979328 1.0047556

fulltimeug 1.0000510 1.0000325 1.0000699

parttimeug 1.0000231 0.9999969 1.0000509

outstatetuition 1.0000035 0.9999883 1.0000187

roomandboard 0.9999657 0.9999273 1.0000042

books 1.0001915 1.0000083 1.0003807

personalexpense 1.0000618 1.0000116 1.0001129

phd 1.0033607 0.9996074 1.0070486

terminaldegree 1.0009388 0.9968788 1.0050735

studentfacultyratio 1.0327502 1.0215935 1.0441462

alumnidonate 0.9974121 0.9941610 1.0006859

instrexpend 1.0000134 1.0000033 1.0000239

gradrate 1.0043156 1.0019979 1.0066393

The effect size of each of the variables is similar, with Private having the lowest and PHD the highest magnitudes.

To see which approach is best for modelling Enrollment, we can use stargazer to do a direct comparison of the GLM, Poisson, and Negative Binomial models:

Enrollment Linear/Poisson/Neg Binom Comparison

=====================================================================================

Dependent variable:

-----------------------------------------------------------------

enrollment) enrollment

OLS normal Poisson negative

binomial

(1) (2) (3) (4)

-------------------------------------------------------------------------------------

privateYes -0.471\*\*\* -0.471\*\*\* -0.526\*\*\* -0.362\*\*\*

(0.060) (0.060) (0.005) (0.056)

appsrec -0.00002 -0.00002 -0.00002\*\*\* 0.00000

(0.00002) (0.00002) (0.00000) (0.00001)

acceptance 0.0001\*\*\* 0.0001\*\*\* 0.0001\*\*\* 0.0002\*\*\*

(0.00003) (0.00003) (0.00000) (0.00003)

top10perc 0.004\* 0.004\* 0.004\*\*\* 0.003

(0.003) (0.003) (0.0002) (0.002)

top25perc 0.0005 0.0005 -0.002\*\*\* 0.001

(0.002) (0.002) (0.0001) (0.002)

fulltimeug 0.0001\*\*\* 0.0001\*\*\* 0.0001\*\*\* 0.0001\*\*\*

(0.00001) (0.00001) (0.00000) (0.00001)

parttimeug 0.00003\* 0.00003\* 0.00002\*\*\* 0.00002\*

(0.00001) (0.00001) (0.00000) (0.00001)

outstatetuition 0.00001\* 0.00001\* 0.00001\*\*\* 0.00000

(0.00001) (0.00001) (0.00000) (0.00001)

roomandboard -0.00003 -0.00003 0.00001\*\*\* -0.00003\*

(0.00002) (0.00002) (0.00000) (0.00002)

books 0.0002\*\* 0.0002\*\* 0.0002\*\*\* 0.0002\*\*

(0.0001) (0.0001) (0.00001) (0.0001)

personalexpense 0.0001\*\* 0.0001\*\* 0.0001\*\*\* 0.0001\*\*

(0.00003) (0.00003) (0.00000) (0.00003)

phd 0.003\* 0.003\* 0.006\*\*\* 0.003\*

(0.002) (0.002) (0.0002) (0.002)

terminaldegree 0.001 0.001 -0.001\*\*\* 0.001

(0.002) (0.002) (0.0002) (0.002)

studentfacultyratio 0.036\*\*\* 0.036\*\*\* 0.015\*\*\* 0.032\*\*\*

(0.006) (0.006) (0.0004) (0.005)

alumnidonate -0.004\*\* -0.004\*\* -0.003\*\*\* -0.003

(0.002) (0.002) (0.0002) (0.002)

instrexpend 0.00002\*\*\* 0.00002\*\*\* 0.00002\*\*\* 0.00001\*\*\*

(0.00001) (0.00001) (0.00000) (0.00001)

gradrate 0.007\*\*\* 0.007\*\*\* 0.003\*\*\* 0.004\*\*\*

(0.001) (0.001) (0.0001) (0.001)

Constant 4.298\*\*\* 4.298\*\*\* 5.200\*\*\* 4.606\*\*\*

(0.176) (0.176) (0.016) (0.163)

-------------------------------------------------------------------------------------

Observations 777 777 777 777

R2 0.783

Adjusted R2 0.778

Log Likelihood -472.740 -52,724.630 -5,244.133

theta 5.829\*\*\* (0.292)

Akaike Inf. Crit. 981.479 105,485.300 10,524.260

Residual Std. Error 0.449 (df = 759)

F Statistic 160.943\*\*\* (df = 17; 759)

=====================================================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

The comparison shows that the AIC of the Poisson and Negative Binomial models is massive compared to the AIC of the GLM model. Therefore we should move forward with the GLM approach.

We can try and improve the GLM model by removing the least significant variables.

emle2 = glm(log(df$enrollment)~private+acceptance+top10perc+fulltimeug+parttimeug+outstatetuition+books+personalexpense+phd+studentfacultyratio+alumnidonate+instrexpend+gradrate , data=df, family=gaussian)

GLM Model Comparison

================================================

Dependent variable:

----------------------------

enrollment)

(1) (2)

------------------------------------------------

privateYes -0.471\*\*\* -0.474\*\*\*

(0.060) (0.059)

appsrec -0.00002

(0.00002)

acceptance 0.0001\*\*\* 0.0001\*\*\*

(0.00003) (0.00001)

top10perc 0.004\* 0.004\*\*\*

(0.003) (0.001)

top25perc 0.0005

(0.002)

fulltimeug 0.0001\*\*\* 0.0001\*\*\*

(0.00001) (0.00001)

parttimeug 0.00003\* 0.00002

(0.00001) (0.00001)

outstatetuition 0.00001\* 0.00001

(0.00001) (0.00001)

roomandboard -0.00003

(0.00002)

books 0.0002\*\* 0.0002\*\*

(0.0001) (0.0001)

personalexpense 0.0001\*\* 0.0001\*\*

(0.00003) (0.00003)

phd 0.003\* 0.005\*\*\*

(0.002) (0.001)

terminaldegree 0.001

(0.002)

studentfacultyratio 0.036\*\*\* 0.035\*\*\*

(0.006) (0.006)

alumnidonate -0.004\*\* -0.003\*

(0.002) (0.002)

instrexpend 0.00002\*\*\* 0.00001\*\*

(0.00001) (0.00001)

gradrate 0.007\*\*\* 0.006\*\*\*

(0.001) (0.001)

PrivateYes

Constant 4.298\*\*\* 4.305\*\*\*

(0.176) (0.163)

------------------------------------------------

Observations 777 777

Log Likelihood -472.740 -475.170

Akaike Inf. Crit. 981.479 978.339

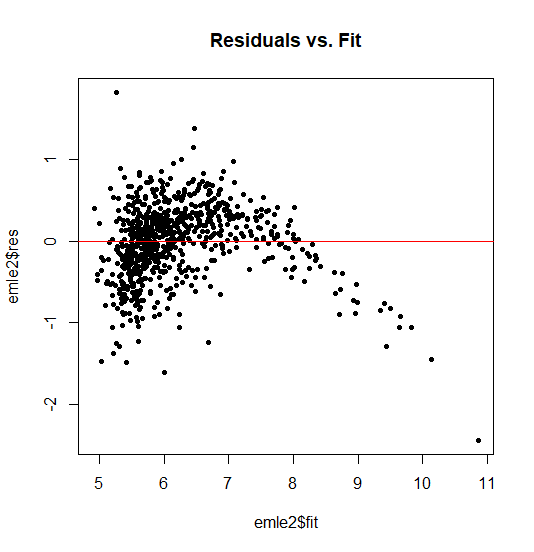
================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

The AIC of the GLM with the reduced variable-set has reduced somewhat, indicating that it is a better model.

Lastly we need to check for linear model assumptions:

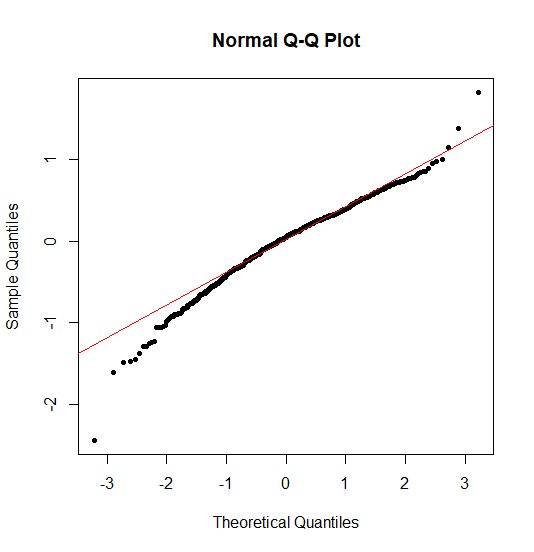
plot(emle2$res ~ emle2$fit)



The residuals are pretty clearly not distributed evenly around the 0,0 line, portraying an arcing shape, indicating lack of linearity.

qqnorm(emle2$res)

qqline(emle2$res, col="red")



The QQ plot shows generally good adherence to a theoretical normal distribution with some divergence in the tails.

shapiro.test(emle2$res)

Shapiro-Wilk normality test

data: emle2$res

W = 0.97485, p-value = 2.633e-10

In the Shapiro test a p-value close to zero indicates that the null hypothesis is rejected and that the sample data is not normally distributed.

norm <- rnorm(777) #create normal dist of same number of observations

ks.test(norm, emle2$res)

Two-sample Kolmogorov-Smirnov test

data: norm and emle2$res

D = 0.19434, p-value = 3.604e-13

alternative hypothesis: two-sided

The KS tests shows a p-value of near zero which means that the null hypothesis is rejected, indicating that the residuals are not normally distributed.

bartlett.test(list(emle2$res, emle2$fit))

Bartlett test of homogeneity of variances

data: list(emle2$res, emle2$fit)

Bartlett's K-squared = 295.56, df = 1, p-value < 2.2e-16

The Bartlett test’s null hypothesis is that the variances of the two compared sets are equal. A p-value close to zero says that we reject the null hypothesis, indicating that variance for the two sets is different, meaning they are heteroskedastic.

vif(emle2)

private acceptance top10perc fulltimeug

2.650717 5.149967 2.491604 6.864932

parttimeug outstatetuition books personalexpense

1.682418 3.625610 1.072585 1.295658

phd studentfacultyratio alumnidonate instrexpend

1.834878 1.894877 1.767110 2.813864

gradrate

1.793441

In the VIF test, a 1 means that there is no correlation among a given predictor and the remaining predictor variables. Most of the inflation factors are fairly low, however acceptance and fulltimeug both have values over 5, indicating problematic multicollinearity.

dwtest(emle2)

Durbin-Watson test

data: emle2

DW = 1.8673, p-value = 0.03052

alternative hypothesis: true autocorrelation is greater than 0

The Durbin-Watson statistic null hypothesis is that there is no autocorrelation. A p value below .05 indicates that there is autocorrelation. The DW statistic of 1.86 shows there is slight positive correlation.

1. **Modelling GradRate**

hist(df$gradrate, main ="Histogram of gradrate", breaks=15, col="green")

A histogram of GradRate shows that the distribution looks fairly normal except for a gap near the high end. Because of this we can proceed with a linear regression.

gm1 = lm(df$gradrate~.-college, data=df)

summary(gm1)

Residuals:

Min 1Q Median 3Q Max

-53.897 -7.132 -0.292 7.213 54.056

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 33.8736716 4.8480858 6.987 6.15e-12 \*\*\*

privateYes 3.3813758 1.6965147 1.993 0.046605 \*

appsrec 0.0012984 0.0004418 2.939 0.003390 \*\*

acceptance -0.0006961 0.0008627 -0.807 0.419995

enrollment 0.0021593 0.0023081 0.936 0.349814

top10perc 0.0548964 0.0717587 0.765 0.444501

top25perc 0.1351288 0.0549667 2.458 0.014179 \*

fulltimeug -0.0004712 0.0004008 -1.176 0.240138

parttimeug -0.0014836 0.0003902 -3.802 0.000155 \*\*\*

outstatetuition 0.0010174 0.0002334 4.359 1.49e-05 \*\*\*

roomandboard 0.0019143 0.0005908 3.240 0.001246 \*\*

books -0.0022205 0.0029168 -0.761 0.446739

personalexpense -0.0016635 0.0007698 -2.161 0.031000 \*

phd 0.0872827 0.0568102 1.536 0.124859

terminaldegree -0.0747023 0.0623172 -1.199 0.231002

studentfacultyratio 0.0758222 0.1593102 0.476 0.634254

alumnidonate 0.2793343 0.0491750 5.680 1.91e-08 \*\*\*

instrexpend -0.0004565 0.0001542 -2.961 0.003163 \*\*

PrivateYes NA NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.75 on 759 degrees of freedom

Multiple R-squared: 0.4615, Adjusted R-squared: 0.4495

F-statistic: 38.27 on 17 and 759 DF, p-value: < 2.2e-16

A summary of the model with all variables included shows that 9 variables have significant p-values.

We can try and improve the model by removing the least significant variables:

gm2 = lm(df$gradrate~private + appsrec + top25perc + parttimeug + outstatetuition +

roomandboard + personalexpense + alumnidonate + instrexpend, data=df)

GradRate Linear Model Comparison

====================================================================

Dependent variable:

------------------------------------------------

gradrate

(1) (2)

--------------------------------------------------------------------

privateYes 3.381\*\* 3.394\*\*

(1.697) (1.525)

appsrec 0.001\*\*\* 0.001\*\*\*

(0.0004) (0.0002)

acceptance -0.001

(0.001)

enrollment 0.002

(0.002)

top10perc 0.055

(0.072)

top25perc 0.135\*\* 0.176\*\*\*

(0.055) (0.030)

fulltimeug -0.0005

(0.0004)

parttimeug -0.001\*\*\* -0.002\*\*\*

(0.0004) (0.0004)

outstatetuition 0.001\*\*\* 0.001\*\*\*

(0.0002) (0.0002)

roomandboard 0.002\*\*\* 0.002\*\*\*

(0.001) (0.001)

books -0.002

(0.003)

personalexpense -0.002\*\* -0.002\*\*

(0.001) (0.001)

phd 0.087

(0.057)

terminaldegree -0.075

(0.062)

studentfacultyratio 0.076

(0.159)

alumnidonate 0.279\*\*\* 0.288\*\*\*

(0.049) (0.048)

instrexpend -0.0005\*\*\* -0.0004\*\*\*

(0.0002) (0.0001)

PrivateYes

Constant 33.874\*\*\* 32.917\*\*\*

(4.848) (2.564)

--------------------------------------------------------------------

Observations 777 777

R2 0.462 0.456

Adjusted R2 0.449 0.450

Residual Std. Error 12.745 (df = 759) 12.743 (df = 767)

F Statistic 38.269\*\*\* (df = 17; 759) 71.465\*\*\* (df = 9; 767)

====================================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

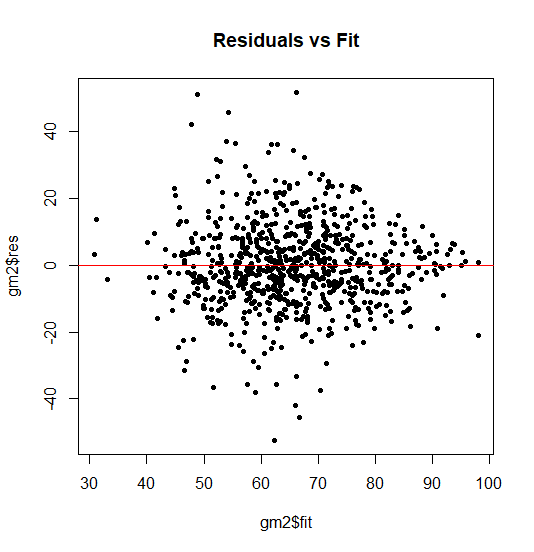
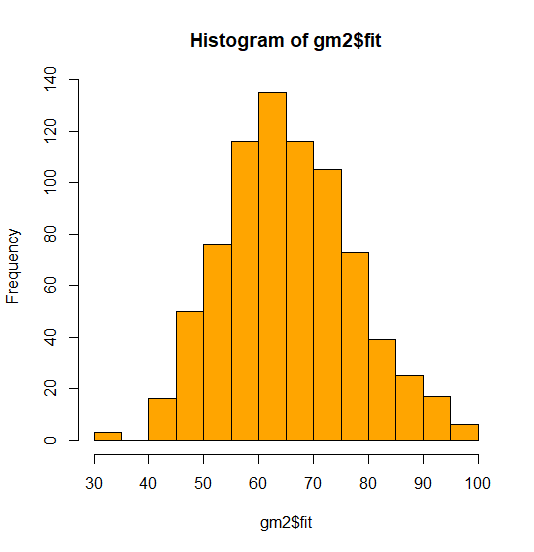
Comparing the two models with stargazer, we see that Adjusted R^2 improved slightly, and the second model has increased degrees of freedom.

Lastly, we check for linear model assumptions:

plot(gm2$res ~ gm2$fit, pch=20, main="Residuals vs Fit")

abline(c(0,0),col ="red")

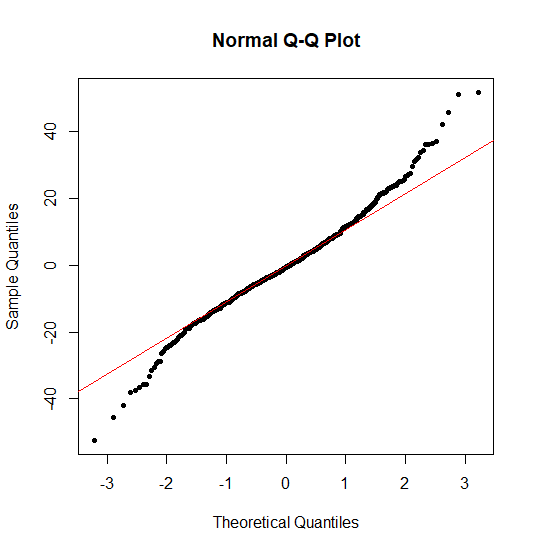
hist(gm2$fit, col="orange")

The residuals are pretty evenly distributed around the 0,0 line, indicating linearity. The histogram of the residuals confirms a fairly normal distribution.

qqnorm(emle2$res)

qqline(emle2$res, col="red")



The QQ plot shows generally good adherence to a theoretical normal distribution with some divergence in the tails.

shapiro.test(gm2$res)

Shapiro-Wilk normality test

data: gm2$res

W = 0.98269, p-value = 5.944e-08

In the Shapiro test a p-value close to zero indicates that the null hypothesis is rejected and that the sample data is not normally distributed.

norm <- rnorm(777) #create normal dist of same number of observations

ks.test(norm, gm2$res)

Two-sample Kolmogorov-Smirnov test

data: norm and gm2$res

D = 0.42857, p-value < 2.2e-16

alternative hypothesis: two-sided

The KS tests shows a p-value of near zero which means that the null hypothesis is rejected, indicating that the residuals are not normally distributed.

bartlett.test(list(emle2$res, emle2$fit))

Bartlett test of homogeneity of variances

data: list(gm2$res, gm2$fit)

Bartlett's K-squared = 6.0016, df = 1, p-value = 0.01429

The Bartlett test’s null hypothesis is that the variances of the two compared sets are equal. A p-value under .05 says that we reject the null hypothesis, indicating that variance for the two sets is different, meaning they are heteroskedastic.

vif(emle2)

private appsrec top25perc parttimeug outstatetuition roomandboard

2.206940 1.854252 1.738981 1.444074 3.754146 1.888706

personalexpense alumnidonate instrexpend

1.229376 1.712828 2.161201

In the VIF test, a 1 means that there is no correlation among a given predictor and the remaining predictor variables. Most of the inflation factors are fairly low, indicating that there is low multicollinearity; however, the outstatetution factor is close to being problematic.

dwtest(emle2)

Durbin-Watson test

data: gm2

DW = 2.0138, p-value = 0.5692

alternative hypothesis: true autocorrelation is greater than 0

The Durbin-Watson statistic null hypothesis is that there is no autocorrelation. A p-value above .05 indicates that there is not autocorrelation. This is supported by a DW statistic close to 2.

**3. Based on your analysis, what recommendations do you have for the college administrator on how to increase enrollment and graduate rate in his/her undergraduate program?**

To increase enrollment, a college administrator should first of all consider becoming a public university (if they are private), since being a private university has the strongest negative impact on enrollment at -47%. If that is not an option or is not relevant, the administrator should first focus on increasing the student to faculty ratio of the school (a 3.53% effect) and then improving the graduate rate (0.61%). Lastly, efforts to hire more PhD educated staff (0.45%) and entice the top 10% of high school graduates (0.42%) should be made.

In direct contrast to enrollment, private schools increase their graduation rate by 3.39 percentage points compared to public schools, so a college administrator might consider going private if that was a key goal. Aside from this, soliciting increased alumni donations (0.29 points) and attracting the top 25% of high school graduates (0.18) should be the administrators first priorities.