Homework 2 – College Football Revenue and Expenses

Data Reading and manipulation

A1. Create two new variables. First, create "total_enroll" which is equal to male and female enrollment combined. **A. = efmalecount_h + effemalecount_h** Second, create "percent_male" which is equal to the percentage of male students (example: 50%=0.5). **A. = efmalecount_h/ total_enroll**

A2. What is the mean, median and standard deviation of "total_enroll" and "percent_male"
FYI, I used sample to calculate StdDev because this is a subsample of all football programs.

	total_enro	percentma
Mean	18503.62	0.48695
Median	16999	0.482623
Std.Dev.	8060.264	0.049935

A3. What is the correlation between expenses and revenues? **A. =CORREL(total_revenue_all_football_h,total_expense_all_football_h) = 0.83 That's pretty highly correlated.**

A4. Create a new variable "percent_female" equal to 1-"percent_male". **A. = 1 – percent_male**

A5. Create year dummy variables. **A. Ugh…. Do I have to? Ok for y_2001 = IF(year=2001,1,0). Repeat until you get to 2009.**

Regression Analytics

B1. What's the R-squared of a simple regression with total_expense_all_football_h as the dependent variable (Y) and the lagged expenses as the only independent variable (X)? What does the R-squared statistic mean here? Is the lagged expenses statistically significant? Is there any evidence for a random walk?

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.798853							
R Square	0.638166							
Adjusted R Square	0.637325							
Standard Error	3.419786							
Observations	432							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regression	1	8869.331	8869.331	758.3906	5.77E-97			
Residual	430	5028.823	11.69494					
Total	431	13898.15						
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	3.586348	0.298487	12.01509	7.12E-29	2.999673	4.173023	2.999673	4.173023
total_expense_all_football_l1_h	0.790639	0.02871	27.53889	5.77E-97	0.73421	0.847068	0.73421	0.847068

The R-squared suggests that previous expenses can explain roughly 63% of current expenses. This is a highly autoregressive variable. Lagged expenses are statistically significant and with that high of a t-stat there's likely a non-linear component to the autoregressive effect. As for random walks we need to avoid coefficients on the lagged variable that are either -1, 0 or 1. Looking at the upper and lower bound of the 95% confidence interval (.73-.84) I can see that we're confident that the coefficient is not -1, 0 or 1. No random walk here.

B2. Run a simple regression with total_expense_all_football_h as the dependent variable (Y) and use three independent variables(X): the lagged expenses, "percent_male" and "efmalecount_h". Are the "male" variables statistically significant? What are the "male" coefficients? What problem are you possibly running into and why?

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.819926							
R Square	0.672278							
Adjusted R Square	0.669981							
Standard Error	3.262192							
Observations	432							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regression	3	9343.423	3114.474	292.6616	2.8E-103			
Residual	428	4554.731	10.6419					
Total	431	13898.15						
	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-0.09498	1.543933	-0.06152	0.950974	-3.12962	2.939653	-3.12962	2.939653
percentmale	4.053138	3.274166	1.237915	0.216426	-2.38231	10.48858	-2.38231	10.48858
efmalecount_h	0.000249	4.12E-05	6.050796	3.15E-09	0.000168	0.000331	0.000168	0.000331
total_expense_all_football_l1_h	0.727234	0.028992	25.0838	4.55E-86	0.670249	0.784219	0.670249	0.784219

The male count variable is statistically significant but the percentage of the student body that is male is not statistically significant. What the HEY? This suggests that if we just keep enrolling more men then our football program will make more money but that would also impact the percentage of the student body that is male? This looks like a multicollinearity problem to me. Generally, if you can use one X variable to calculate another X variable then you're introducing some level of multicollinearity. Try to avoid using X variables that help calculate another X variable. Pick one or the other but not both.

B3. Run a simple regression with total_expense_all_football_h as the dependent variable (Y) and use total_revenue_all_football_h as the only independent variable (X). How does the R-squared compare to question B1? Is the coefficient on revenue statistically significant? What problem are you possibly running into and why?

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.830525							
R Square	0.689772							
Adjusted R Square	0.689051							
Standard Error	3.166538							
Observations	432							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regression	1	9586.561	9586.561	956.0785	2.4E-111			
Residual	430	4311.593	10.02696					
Total	431	13898.15						
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	4.888143	0.235594	20.74813	9.13E-67	4.425083	5.351203	4.425083	5.351203
total_revenue_all_football_h	0.278133	0.008995	30.92052	2.4E-111	0.260453	0.295813	0.260453	0.295813

Wow! The R-squared is even better than the .63 that we had in B1. The revenue variable is super significant. We're geniuses! Oh wait. Revenues and Expenses are determined at the same time. If a team makes a bowl game then they get a payout from the bowl organizers (revenues go up) and they have additional expenses to travel to the bowl game (expenses go up). These two variables occur simultaneously and as a result we have introduced endogeneity into our regression. The best way to fix this would be to lag the X variable by one year. The past can't be simultaneous (unless you're a philosophy major). Endogeneity problem solved.

B4. How could you solve the problem in B3 with the data that is already included in the dataset?
I just answered that?! Weren't you paying attention Word Doc?

B5. Run a simple regression with total_net_all_football_h as the dependent variable and include the lagged net_all_football and "percent_female" as independent variables. What is the sign and significance of "percent_female"? What does the coefficient on "percent_female" imply and what problem are you possibly running into and why?

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.766877							
R Square	0.588101							
Adjusted R Square	0.586181							
Standard Error	8.132748							
Observations	432							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regression	2	40512.89	20256.45	306.2588	2.36E-83			
Residual	429	28374.74	66.14159					
Total	431	68887.63						
	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	11.59215	4.146248	2.795816	0.005409	3.442659	19.74164	3.442659	19.74164
total_net_all_football_l1_h	0.732684	0.030507	24.01699	2.12E-81	0.672722	0.792645	0.672722	0.792645
percent_female	-15.4621	7.953341	-1.94411	0.052536	-31.0945	0.170221	-31.0945	0.170221

The sign of percent_female is negative and suggest that for every 1% (0.01) increase in women at the school we see a decline of $154k in net football revenue. Quick! Athletic Directors should get rid of all the women so the football team can make more money! And look the effect is statistically significant at the 90% confidence interval! Wait… this sounds like a spurious correlation to me. If you think you may have a spurious effect then get rid of that spurious X variable.

B6. Run a regression with total_expense_all_football_h as the dependent variable and use only year dummy variables and conference dummy variables as your independent variables. How does the R-squared compare to question B1? What's interesting (or not) about this particular regression formation?

OH NO! I tried and there are two many X variables! SHAME!

Hmmm… do we need all those X variables? Let's see.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.471671
R Square	0.222474
Adjusted R Square	0.207769
Standard Error	5.054357
Observations	432

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	8	3091.973	386.4967	15.12913	1.44E-19
Residual	423	10806.18	25.54653		
Total	431	13898.15			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	6.971059	0.669466	10.41286	9.32E-23	5.655164	8.286953	5.655164	8.286953
y_2002	0.456006	0.920516	0.49538	0.620589	-1.35335	2.26536	-1.35335	2.26536
y_2003	1.302034	0.959827	1.356529	0.175654	-0.58459	3.188659	-0.58459	3.188659
y_2004	2.922294	1.014292	2.881118	0.004164	0.928615	4.915974	0.928615	4.915974
y_2005	4.458268	1.020926	4.366887	1.59E-05	2.451549	6.464988	2.451549	6.464988
y_2006	5.530563	1.020926	5.417203	1.02E-07	3.523843	7.537283	3.523843	7.537283
y_2007	6.265145	0.995855	6.29122	7.87E-10	4.307704	8.222586	4.307704	8.222586
y_2008	6.085391	1.04252	5.837194	1.06E-08	4.036226	8.134556	4.036226	8.134556
y_2009	7.628944	1.04252	7.317791	1.28E-12	5.579779	9.678108	5.579779	9.678108

Check out this SWEET regression using only the year dummies. Do you see how the coefficients are getting larger every year? This means that we don't need year dummies. By simply including the "year" as a variable we can control for the fact that over time schools are spending more money. Hmmm. Let's do it again for conferences!

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.659567							
R Square	0.435029							
Adjusted R	0.420232							
Standard E	4.323815							
Observatic	432							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regressior	11	6046.097	549.6452	29.40007	1.17E-45			
Residual	420	7852.057	18.69537					
Total	431	13898.15						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	3.64166	1.303679	2.793371	0.005455	1.079111	6.204208	1.079111	6.204208
acc_h	7.838007	1.464482	5.352068	1.43E-07	4.95938	10.71663	4.95938	10.71663
bigeast_h	8.370114	1.505359	5.560211	4.8E-08	5.411137	11.32909	5.411137	11.32909
bigten_h	10.26122	1.418157	7.2356	2.22E-12	7.473646	13.04878	7.473646	13.04878
bigtwelve_	9.687733	1.434958	6.751229	4.89E-11	6.867138	12.50833	6.867138	12.50833
cusa_h	3.279626	1.546609	2.120526	0.034547	0.239566	6.319685	0.239566	6.319685
ind_h	7.835695	2.332092	3.359942	0.000851	3.251668	12.41972	3.251668	12.41972
mac_h	0.574534	1.454315	0.395055	0.693003	-2.28411	3.433177	-2.28411	3.433177
mntwest_h	2.786821	1.499814	1.858111	0.063853	-0.16126	5.734899	-0.16126	5.734899
pacten_h	9.263237	1.468184	6.309316	7.12E-10	6.377333	12.14914	6.377333	12.14914
sec_h	9.773476	1.42194	6.873339	2.28E-11	6.978471	12.56848	6.978471	12.56848
wac_h	0.795692	1.574342	0.505412	0.613534	-2.29888	3.890264	-2.29888	3.890264

Sunbelt is our comparison variable here. Again, let's look at the coefficients ACC, BigEast, Big10, Big12, Ind, Pac10 and SEC are all pretty much the same. Why don't we just group these into Power 5 conference teams and everybody else? Power5= ACC + Big10 + Big12 + Pac10 + SEC. Then we'll include Power5 and year. And because we've already seen a strong autoregressive effect on expenses we can include the lagged expenses.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.845576							
R Square	0.714998							
Adjusted R Square	0.713001							
Standard Error	3.042148							
Observations	432							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regression	3	9937.159	3312.386	357.9155	3E-116			
Residual	428	3960.995	9.254661					
Total	431	13898.15						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-695.439	129.8596	-5.35531	1.4E-07	-950.681	-440.197	-950.681	-440.197
Power5	3.355724	0.333359	10.06639	1.56E-21	2.700499	4.010948	2.700499	4.010948
year	0.348582	0.064832	5.376663	1.25E-07	0.221152	0.476011	0.221152	0.476011
total_expense_all_football_	0.592221	0.032474	18.23677	2.13E-55	0.528392	0.656049	0.528392	0.656049

So Power5 conferences spend significantly more than other schools (about $3 million). Each additional year creates another $348k of expenses and the expenses are still highly autoregressive with no random walk in sight. Three variables, all significant, creating an R-squared of .71. This is a good, simple model to use as a baseline for data mining.

Data Mining

C1. Do your best. Forecast total_expense_all_football_h using any of the information here and any combination/transformation of the data you desire. **A. DO YOUR BEST! HAVE FUN!**