Hunter Geise
Collaborators: Nolan Pittman, Aidan Myers

1. a. I think the β coefficients for FG%, 3Pt%, DReb, TRB, and STL to be positive. I think the β coefficient for TOV will be negative. Potential imperfect multicollinearity can impact my expectations for FG% and 3Pt%. This is due to 3Pt% being a part of the calculation for FG%. The better 3Pt% a team has, it will skew their FG% and make it more positive, causing a larger coefficient. This is also prevalent in Team DREB and Team TRB where the more DREB a team has, the more TRB they will have which will create a larger coefficient. Also with this, there will be wider confidence intervals due to larger standard errors.

b.

	Dependent variable:
	WIN%
FG%	0.015^{***}
	0.005
3Pt%	0.009^{***}
	0.002
DREB	0.030^{**}
	0.013
TRB	0.005
	0.006
STL	-0.005
	0.012
TOV	-0.052
	0.009
Constant	0.501^{***}
	0.006
Observations	108
R^2	0.863
Adjusted R^2	0.855
Residual Std. Error	0.064 (df = 101)
F Statistic	106.228^{***} (df = 6; 101)
Note:	*p<0.1; **p<0.05; ***p<0.01

c.

Stat Category	FG%	3Pt%	DREB	TRB	STL	TOV
VIF	9.449	1.365	22.395	6.987	4.811	6.592

d. FG%, TRB, and TOV are problematic since their VIF is all over 5, but less than 10. DREB is heavily impacted by collinearity because its VIF is over 10. STL is impacted by VIF, but it isn't quite problematic since it is at 4.811.

e. To eliminate multicollinearity from DREB and TRB, I would eliminate the DREB variable and just compare overall rebounding abilities between teams. Keeping TRB is better than keeping DREB because offensive rebounds are a key part to winning games and are kept under consideration when using the TRB variable. Also there is multicollinearity in FG% because it is affected by a team's 3Pt%. This can be eliminated by getting rid of both 3Pt% and FG% and replacing them with eFG% which still accounts for three pointers a team takes and 2Pt%.

	Dependent variable:
	WIN%
Team FG%	0.028***
	0.005
Team 3P%	0.013***
	0.004
Team DREB	0.027**
	0.011
Team TRB	0.011
	0.008
Team STL	0.061***
	0.011
Team TOV	-0.033***
	0.007
Constant	-2.217***
	0.250
Observations	108
R^2	0.735
Adjusted R^2	0.719
Residual Std. Error	0.090 (df = 101)
F Statistic	46.656*** (df = 6; 101)
Note:	*p<0.1; **p<0.05; ***p<0.01

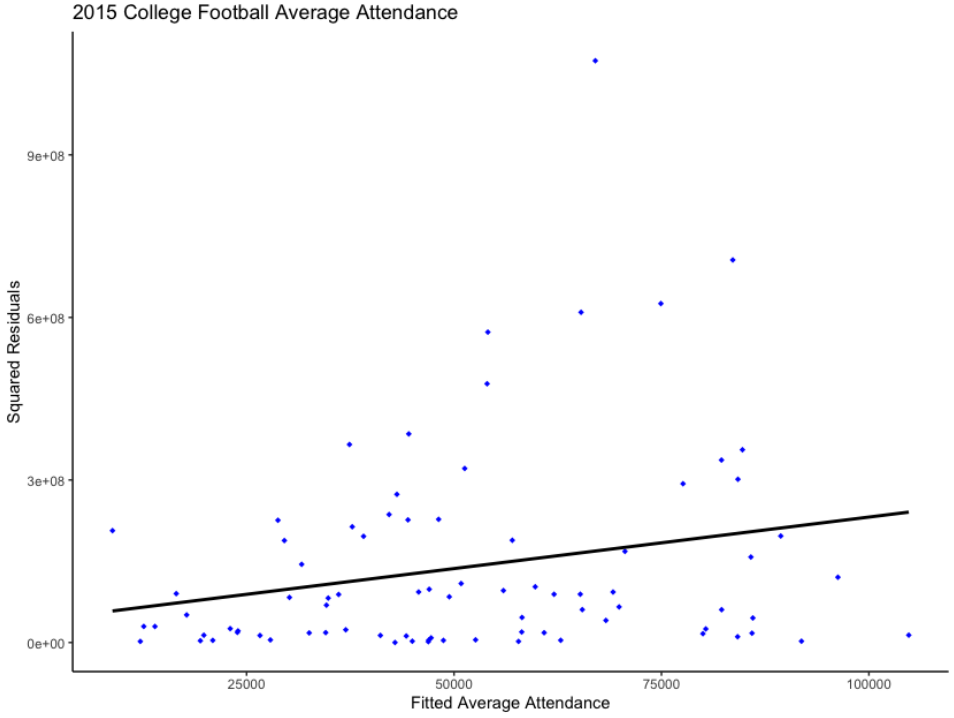
2. a.

   b. Based off of R^2 measures, the regression ran in question one is better since more variation in Win% can be explained by the X variables. I think the first model outperformed the second because it used the difference between teams' percentages and stats which almost doubles as a stat which gives more information on how certain teams performed vs the rest of the league, rather than the second model which is the average among teams.
3. It is important to test for heteroscedasticity to make sure that there are no outliers and that the OLS estimates are BLUE instead of LUE, leading to p-values not being reliable and affecting the decision to reject/fail to reject the null hypothesis. It is important to test for autocorrelation because it can determine whether a model is not efficient, therefore not being BLUE. This can lead to skewing the standard errors and making the hypothesis test not reliable.

4. a.

	Dependent variable:
	AVG Attendance
Stadium Age	32.073
	198.556
(Stadium Age)^2	0.697
	1.748
Recruiting Strength	390.250***
	47.227
Ranked Opp	1,781.111
	2,009.864
Nationally Televised Home Games	435.237
	967.537
Power 5 School	-2,492.525
	4,915.245
Season Win Total	246.896
	535.190
Constant	-31,007.100***
	7,223.662
Observations	82
R^2	0.790
Adjusted R^2	0.770
Residual Std. Error	12,433.040 (df = 74)
F Statistic	39.775*** (df = 7; 74)
Note:	*p<0.1; **p<0.05; ***p<0.01

b.



There is heteroscedasticity present because there is more error and variation as values of X increase. This is noticeable where points on the scatter plot venture further away from the line of best fit.

c.

```
> #c#
> bptest(cfb2015model)

        studentized Breusch-Pagan test

data:  cfb2015model
BP = 20.224, df = 7, p-value = 0.005105

> White_Model <- lm(I(cfbResiduals ^ 2) ~ cfbFitted + I(cfbFitted ^ 2))
> summary(White_Model)

Call:
lm(formula = I(cfbResiduals^2) ~ cfbFitted + I(cfbFitted^2))

Residuals:
       Min         1Q     Median         3Q        Max
-178230621 -126930799  -49484984   69269120  887925883

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -7.174e+07  9.922e+07  -0.723    0.472
cfbFitted       7.052e+03  3.990e+03   1.767    0.081 .
I(cfbFitted^2) -4.787e-02  3.614e-02  -1.325    0.189
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 185400000 on 79 degrees of freedom
Multiple R-squared:  0.07365,   Adjusted R-squared:  0.0502
F-statistic: 3.141 on 2 and 79 DF,  p-value: 0.04871
```

In both tests, the p-values are below the .05 level (BP is 0.005105, White is 0.04871) which gives us enough evidence to reject the null hypothesis of homoscedasticity in favor of the alternative hypothesis of heteroscedasticity.

d.

```
> cfbResiduals2 <- cfb2015model2$residuals
> cfbFitted2 <- cfb2015model2$fitted.values
> bptest(cfb2015model2)

        studentized Breusch-Pagan test

data:  cfb2015model2
BP = 11.668, df = 7, p-value = 0.112

> White_Model2 <- lm(I(cfbResiduals2 ^ 2) ~ cfbFitted2 + I(cfbFitted2 ^ 2))
> summary(White_Model2)

Call:
lm(formula = I(cfbResiduals2^2) ~ cfbFitted2 + I(cfbFitted2^2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.06434 -0.04705 -0.02162  0.01912  0.21511

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -5.21490    3.49752  -1.491    0.140
cfbFitted2       1.00335    0.65514   1.532    0.130
I(cfbFitted2^2) -0.04767    0.03063  -1.556    0.124

Residual standard error: 0.06631 on 79 degrees of freedom
Multiple R-squared:  0.04218,   Adjusted R-squared:  0.01793
F-statistic: 1.739 on 2 and 79 DF,  p-value: 0.1823
```

Using the log of attendance does reduce the presence of heteroscedasticity. In both tests, the p-values are above the .05 level (BP is 0.112, White is 0.1823) which does not gives us enough evidence to reject the null hypothesis of homoscedasticity.

e.

Independent Variables	Regular Standard Errors	Robust Standard Errors
Stadium Age	173.006	198.556
(Stadium Age)^2	1.733	1.748
Recruiting Strength	46.042	47.227
Ranked Opponents	2424.663	2009.864
Nationally Televised Home Games	1036.69	967.537
Power 5 Schools	4576.826	4915.245
Season Win Total	532.789	535.19
Intercept	6122.088	7223.662

Heteroscedasticity gives less correct standard errors which then would affect the confidence interval and p-values of the experiment, making it LUE instead of BLUE.

5. a.

ACC Team	lag1	lag2	lag3	lag4	lag5	Average
Boston College	0.23360	-0.21088	-0.45268	-0.27967	-0.08099	-0.15813
Clemson	0.08393	-0.04913	-0.17404	-0.52635	-0.01992	-0.13710
Duke	-0.49028	0.22250	-0.38561	0.45602	-0.34226	-0.10793
Floriida St	0.26599	0.18364	-0.06541	-0.08034	-0.33318	-0.00586
Georgia Tech	0.19800	0.08860	-0.45713	-0.22632	-0.20959	-0.12129
Louisville	-0.34036	0.14824	-0.15347	-0.18752	-0.13530	-0.13368
Maryland	0.01988	-0.24278	0.13530	-0.38110	-0.07975	-0.10969
North Carolina	0.03321	0.04011	0.27065	-0.34989	-0.08926	-0.01904
Pittsburgh	-0.25504	0.03509	0.04807	-0.26968	0.07943	-0.07243
Syracuse	0.26124	-0.11789	-0.28116	-0.47436	-0.01238	-0.12491
Virginia	0.21388	0.07509	-0.23633	-0.07449	-0.48926	-0.10222
Virginia Tech	0.46169	-0.12332	-0.13642	0.12538	0.09233	0.08393
Wake Forest	0.27940	0.09728	-0.09498	-0.17725	-0.22479	-0.02407

b.

```
> allcfb_bgtest <- bgtest(allcfbmodel, order = 3) #plm package
> allcfb_bgtest

        Breusch-Godfrey test for serial correlation of order up to 3

data:  allcfbmodel
LM test = 435.94, df = 3, p-value < 2.2e-16
```

Since the p-value is under the .05 level, there is enough evidence to reject the null hypothesis of no autocorrelation in favor of the alternative hypothesis which is autocorrelation.

c.

Error Type	Intercept	Stadium Age	(Stadium_Age^2)	Recruiting Strength	Ranked Opp	National TV. Home Games	Power 5 School	Season Win Total
OLS Standard Errors	297565.2661	58.1504	0.5097	13.8082	482.3491	329.3188	2623.6978	150.7005
HAC Standard Errors	483080.3000	154.0050	1.4890	31.3170	460.7700	448.2550	3758.8440	209.0870
Error Type	**American Conference**	**Big 10 Conference**	**Big 12 Conference**	**Conference USA**	**Mountain West Conference**	**PAC-12 Conference**	**Southeast Conference**	**Season**
OLS Standard Errors	2521.0016	1465.6329	1774.6093	2419.4491	2578.3822	1545.4805	1495.6781	147.6911
HAC Standard Errors	4038.0280	4407.8500	3939.1200	3911.7720	2743.2940	2727.6530	3376.8970	239.4150

In terms of this model, autocorrelation created smaller standard errors for all but

one variable which then leave hypothesis tests as unreliable. It also makes the OLS estimates inefficient, making them BLU instead of BLUE.

d.

```
> #d#
> no_age_quadratic <- lm(log(AVG_Attend)~ stadium_age + Recruiting_Strength +
+           Ranked_Opp + Nationally_Televised_Home_Games + Power5_School +
+           Season_Win_Total + home_conference + season, data=allcfb)
> #d#
> no_age_quadratic <- lm(AVG_Attend~ stadium_age + Recruiting_Strength +
+           Ranked_Opp + Nationally_Televised_Home_Games + Power5_School +
+           Season_Win_Total + home_conference + season, data=allcfb)
> age_quadratic <- lm(AVG_Attend~ stadium_age + I(stadium_age^2) + Recruiting_Strength +
+            Ranked_Opp + Nationally_Televised_Home_Games + Power5_School +
+            Season_Win_Total + home_conference + season, data=allcfb)
> anova(no_age_quadratic, age_quadratic)
Analysis of Variance Table

Model 1: AVG_Attend ~ stadium_age + Recruiting_Strength + Ranked_Opp +
    Nationally_Televised_Home_Games + Power5_School + Season_Win_Total +
    home_conference + season
Model 2: AVG_Attend ~ stadium_age + I(stadium_age^2) + Recruiting_Strength +
    Ranked_Opp + Nationally_Televised_Home_Games + Power5_School +
    Season_Win_Total + home_conference + season
  Res.Df         RSS Df Sum of Sq      F Pr(>F)
1    723 75676123639
2    722 75597471876  1  78651763 0.7512 0.3864
```

The p-value of λ is 0.983 (calculated in R and Excel) which means at the .05 level, there is not enough evidence that the quadratic model is superior to the linear model. Therefore, it should be kept.

e. The p-value of λ for the conference variable is 0.709 (calculated in R and Excel) which at the .05 level means there is not enough evidence that the having the conference variable is better than excluding the conference interval. The p-value of λ for the season variable is 0.88483755 (calculated in R and Excel) which at the .05 level means there is not enough evidence that the having the season variable is better than excluding the season variable. Therefore, both variables should be kept in the model.