SAL 213 Module 6 Submission Template

Name: Hunter Geise

Collaborators: NONE. Should work on this solo.

Module 4, Step 1:

	Dependent variable:
	Attendance
Minimum Temperature	-13.370
	27.348
Maximum Temperature	-19.549
	26.726
Precipitation mm	-13.390^{**}
	5.839
Feels Like	-2.113
	44.099
Total Snow cm	76.854^{***}
	29.421
Wind Speed Kmph	6.432
	21.789
Wind Gusts Kmph	-8.488
	15.273
Humidity	8.460^{***}
	2.965
Constant	$17,158.310^{***}$
	269.904
Observations	2,532
R^2	0.044
Adjusted R^2	0.041
Residual Std. Error	2,348.967 (df = 2523)
F Statistic	14.512^{***} (df = 8; 2523)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Dependent variable:			
	Attendance		
Friday	77.340	Start Time Locally 6:00 p.m.	-582.248**
	204.127		257.534
Saturday	716.191***	Start Time Locally 6:30 p.m.	493.839
	180.707		776.458
Sunday	466.811**	Start Time Locally 7:30 p.m.	1,072.534***
	236.742		116.385
Thursday	-66.518	Start Time Locally 8:00 p.m.	249.261
	180.632		283.396
Tuesday	-216.423	Start Time Locally 8:30 p.m.	-121.592
	184.649		2,299.83
Wednesday	158.595	Away Quality	-848.361**
	219.331		358.277
Start Time Locally 1:00 p.m.	241.82	Home Quality	-268.346
	227.955		365.49
Start Time Locally 11:30 a.m.	3,073.664**	Game Importance	777.713**
	1,339.79		321.988
Start Time Locally 12:00 p.m.	-138.439	Intra Division	211.247**
	1,627.98		97.727
Start Time Locally 12:30 p.m.	702.12	Canadien Home Team	1,043.123***
	598.129		114.76
Start Time Locally 2:00 p.m.	-258.33	Total Star Players	42.366
	346.522		36.159
Start Time Locally 3:00 p.m.	306.085	Home Quality x Canadien Home Team	-800.743***
	457.436		230.708
Start Time Locally 4:00 p.m.	541.173	Constant	16,541.170***
	699.064		175.027
Start Time Locally 4:30 p.m.	-187.774	Observations	2,532
	1,627.36	R^2	0.095
Start Time Locally 5:00 p.m.	-390.051	Adjusted R^2	0.085
	320.404	Residual Std. Error	2,294.256 (df = 2503)
Start Time Locally 5:30 p.m.	-219.998	F Statistic	9.409*** (df = 28; 2503)
	2,308.62	*Note:*	*p<0.1; **p<0.05; ***p<0.01

Module 4, Step 2, Part 1:

In the weather model, I think multicollinearity can exist between multiple variables. First, I think it can occur between min/max temp, wind speed/gust, humidity, and the feels like variables. All of the variables mentioned play into what the feels like temperature is. If it is hot and humid, the feels like temp is high. If it is cold with high wind speeds/gusts, the feels like temperature will be low. I also think there can be multicollinearity between the wind speed and wind gust variables because they both involve how fast/hard the wind is.

In the hockey model, I think multicollinearity can exist between the home/away quality and game importance. This is due to high importance game potentially being examples of two high quality teams squaring off on the ice. I do expect there to be a high

Module 4, Step 2, Part 2:

Minimum Temperature	Maximum Temperature	Precipitation mm	Feels Like
41.895	41.749	1.274	141.104
Total Snow cm	Wind Speed Kmph	Wind Gusts Kmph	Humidity
1.119	7.39	7.796	1.536

Variable	GVIF
Day of Week	2.084
Start Time Locally	2.231
Away Quality	13.486
Home Quality	14.016
Game Importance	24.367
Intra Division	1.021
Canadien Home Team	1.051
Total Star Players	1.607
Home Quality x Canadien Home Team	1.399

Module 4, Step 2, Part 3:

Due to getting an error for regressing the interaction term, I did a p-test to see if it added predictive power to the model

```
> interaction <- lm(Home.quality:Canadien.Home.Team ~ Day.of.Week + Start.Time..locally. + Away.quality +
+                   Home.quality + Game.Importance + Intra.Division + Canadien.Home.Team +
+                   Total_Star_Players, data = nhlatt )
Error in model.frame.default(formula = Home.quality:Canadien.Home.Team ~  :
  variable lengths differ (found for 'Day.of.Week')
>
```

```
Analysis of Variance Table

Model 1: Att. ~ Day.of.Week + Start.Time..locally. + Away.quality + Home.quality +
    Game.Importance + Intra.Division + Canadien.Home.Team + Total_Star_Players
Model 2: Att. ~ Day.of.Week + Start.Time..locally. + Away.quality + Home.quality +
    Game.Importance + Intra.Division + Canadien.Home.Team + Total_Star_Players +
    Home.quality:Canadien.Home.Team
  Res.Df        RSS Df Sum of Sq      F    Pr(>F)
1   2504 1.3238e+10
2   2503 1.3175e+10  1  63408022 12.046 0.0005277 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Looking at the p-value in the F test, it is statistically significant, so therefore it does add predictive power to the model.

Module 4, Step 2, Part 4:

For the weather, I would drop wind gust, wind speed, humidity, min and max temperature, and feels like because it doesn't have any effect on transportation which would affect the attendance. People will show up for a hockey game whether it's freezing out, scorching hot, or when there's wind that can knock you over.

I would drop the home/away quality because they have a major effect on the game importance VIF. I feel that hockey fans will show up regardless of how good their team is or how good the opponent is, but when the game is important there is going to be a lot more fans.

Module 4, Step 3:

	Dependent variable:
	Attendance
Precipitation mm	-13.103**
	5.259
Total Snow cm	121.440***
	28.212
Friday	71.492
	204.013
Saturday	737.116***
	180.449
Sunday	483.039**
	236.528
Thursday	-58.262
	180.542
Tuesday	-230.29
	184.606
Wednesday	110.664
	219.368
Start Time Locally 1:00 p.m.	212.81
	227.605
Start Time Locally 11:30 a.m.	2,964.164**
	1,338.72
Start Time Locally 12:00 p.m.	2.698
	1,629.72
Start Time Locally 12:30 p.m.	680.323
	597.796
Start Time Locally 2:00 p.m.	-357.36
	346.321
Start Time Locally 3:00 p.m.	294.153
	457.227
Start Time Locally 4:00 p.m.	525.665
	698.334
Start Time Locally 4:30 p.m.	80.178
	1,629.31
Start Time Locally 5:00 p.m.	-402.552
	320.245
Start Time Locally 5:30 p.m.	-249.852
	2,307.41
Start Time Locally 6:00 p.m.	-687.470***
	256.795
Start Time Locally 6:30 p.m.	403.759
	776.307
Start Time Locally 7:30 p.m.	1,070.434***
	116.353
Start Time Locally 8:00 p.m.	172.046
	283.131
Start Time Locally 8:30 p.m.	-581.972
	2,302.19
Game Importance	242.510***
	85.456
Intra Division	215.675**
	97.443
Canadien Home Team	996.761***
	115.495
Total Star Players	37.141
	35.507
Canadien Home Team x Home Quality	-480.025**
	212.001
Constant	16,610.760***
	173.413
Observations	2,532
R^2	0.096
Adjusted R^2	0.086
Residual Std. Error	2,293.057 (df = 2503)
F Statistic	9.513*** (df = 28; 2503)
Note:	*p<0.1; **p<0.05; ***p<0.01

Module 4, Step 4, Part 1:

```
> anova(allnhlmodel, nhl2)
Analysis of Variance Table

Model 1: Att. ~ Precipitation_mm + total_snow_cm + Day.of.Week + Start.Time..locally. +
    Game.Importance + Intra.Division + Canadien.Home.Team + Total_Star_Players +
    Home.quality:Canadien.Home.Team
Model 2: Att. ~ Precipitation_mm + total_snow_cm + I(total_snow_cm^2) +
    Day.of.Week + Start.Time..locally. + Game.Importance + Intra.Division +
    Canadien.Home.Team + Total_Star_Players + Home.quality:Canadien.Home.Team
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1   2503 1.3161e+10
2   2502 1.3155e+10  1   5977426 1.1369 0.2864
>
```

When adding a quadratic to total snow in cm, it does not add any predictive power to my model because it is not statistically significant at the 5% level with a p-value of 0.2864.

Module 4, Step 4, Part 2:

Found sum of squared residuals in R and calculated chi squared and chi squared p-value in Excel.

```
> sum(resid(allnhlmodel)^2)
[1] 13161045221
> sum(resid(nhl3)^2)
[1] 54.97956
>
```

After plugging in the sum of squared residuals in Excel, I got a p-value of 1.12083E-05. Therefore, taking the log of my dependent variable does add predictive power because it is statistically significant at the 5% level

Module 4, Step 5:

I am going to take the log of my dependent variable of attendance.

	Dependent variable:
	log(Att.)
Precipitation mm	-0.001**
	0.0003
Total Snow cm	0.007***
	0.002
Friday	0.01
	0.013
Saturday	0.051***
	0.012
Sunday	0.031**
	0.015
Thursday	-0.0001
	0.012
Tuesday	-0.013
	0.012
Wednesday	0.012
	0.014
Start Time Locally 1:00 p.m.	0.013
	0.015
Start Time Locally 11:30 a.m.	0.166*
	0.087
Start Time Locally 12:00 p.m.	0.01
	0.105
Start Time Locally 12:30 p.m.	0.046
	0.039
Start Time Locally 2:00 p.m.	-0.021
	0.022
Start Time Locally 3:00 p.m.	0.023
	0.03
Start Time Locally 4:00 p.m.	0.036
	0.045
Start Time Locally 4:30 p.m.	0.011
	0.105
Start Time Locally 5:00 p.m.	-0.025
	0.021
Start Time Locally 5:30 p.m.	-0.005
	0.149
Start Time Locally 6:00 p.m.	-0.043***
	0.017
Start Time Locally 6:30 p.m.	0.025
	0.05
Start Time Locally 7:30 p.m.	0.062***
	0.008
Start Time Locally 8:00 p.m.	0.013
	0.018
Start Time Locally 8:30 p.m.	-0.028
	0.149
Game Importance	0.017***
	0.006
Intra Division	0.014**
	0.006
Canadien Home Team	0.061***
	0.007
Total Star Players	0.002
	0.002
Canadien Home Team x Home Quality	-0.026*
	0.014
Constant	9.702***
	-0.011
Observations	2,532
R^2	0.088
Adjusted R^2	0.078
Residual Std. Error	0.148 (df = 2503)
F Statistic	8.638*** (df = 28; 2503)
Note:	*p<0.1; **p<0.05; ***p<0.01

Module 4, Step 6:

```
> #Step 6#
> bptest(nhl3)

	studentized Breusch-Pagan test

data:  nhl3
BP = 116.98, df = 28, p-value = 7.621e-13

>
```

The Breusch-Pagan tests rejects the null hypothesis of homoscedasticity in favor of the alternative hypothesis of heteroscedasticity because of the p-value being 7.621e-13.

Module 4, Step 7:

```
> #Step 7#
> bgtest(nhl3)

	Breusch-Godfrey test for serial correlation of order up to 1

data:  nhl3
LM test = 5.3461, df = 1, p-value = 0.02077

>
```

The Breusch-Godfrey test rejects the null hypothesis of no autocorrelation at the 5% level in favor of autocorrelation because of the p-value being 0.02077.

Module 4, Step 8:

To take care of autocorrelation and heteroscedasticity I used the HAC Standard Errors.

	Dependent variable:	
	log(Attendance)	
	OLS Normal Standard Errors	*HAC Standard Errors*
Precipitation mm	-0.001**	-0.001**
	0.0003	0.0004
Total Snow cm	0.007***	0.007***
	0.002	0.001
Friday	0.01	0.01
	0.013	0.014
Saturday	0.051***	0.051***
	0.012	0.012
Sunday	0.031**	0.031*
	0.015	0.017
Thursday	-0.0001	-0.0001
	0.012	0.013
Tuesday	-0.013	-0.013
	0.012	0.014
Wednesday	0.012	0.012
	0.014	0.015
Start Time Locally 1:00 p.m.	0.013	0.013
	0.015	0.013
Start Time Locally 11:30 a.m.	0.166*	0.166***
	0.087	0.055
Start Time Locally 12:00 p.m.	0.01	0.01
	0.105	0.031
Start Time Locally 12:30 p.m.	0.046	0.046*
	0.039	0.027
Start Time Locally 2:00 p.m.	-0.021	-0.021
	0.022	0.022
Start Time Locally 3:00 p.m.	0.023	0.023
	0.03	0.024
Start Time Locally 4:00 p.m.	0.036	0.036
	0.045	0.026
Start Time Locally 4:30 p.m.	0.011	0.011
	0.105	0.026
Start Time Locally 5:00 p.m.	-0.025	-0.025
	0.021	0.022
Start Time Locally 5:30 p.m.	-0.005	-0.005
	0.149	0.018
Start Time Locally 6:00 p.m.	-0.043***	-0.043**
	0.017	0.018
Start Time Locally 6:30 p.m.	0.025	0.025
	0.05	0.044
Start Time Locally 7:30 p.m.	0.062***	0.062***
	0.008	0.008
Start Time Locally 8:00 p.m.	0.013	0.013
	0.018	0.012
Start Time Locally 8:30 p.m.	-0.028	-0.028***
	0.149	0.011
Game Importance	0.017***	0.017***
	0.006	0.006
Intra Division	0.014**	0.014**
	0.006	0.006
Canadien Home Team	0.061***	0.061***
	0.007	0.006
Total Star Players	0.002	0.002
	0.002	0.002
Canadien Home Team x Home Quality	-0.026*	-0.026**
	0.014	0.013
Constant	9.702***	9.702***
	0.011	0.013
Observations	2,532	
R^2	0.088	
Adjusted R^2	0.078	
Residual Std. Error	0.148 (df = 2503)	
F Statistic	8.638*** (df = 28; 2503)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Module 4, Step 9:

With all the changes and tests that were done, the interpretation of all the coefficient estimates changes. Originally going from “a 1 unit increase in X changes Y by the coefficient estimate”, the log-lin model makes the interpretation “a one unit increase in X changes Y by $(e^{\beta_1} - 1) * 100\%$”. Looking at the weather variables, the coefficients decreased greatly, but were still statistically significant. Precipitation went from an OLS estimate of -13.390 to-0.001 and total snow went from 76.854 to 0.007. For example, an extra inch of snow adds 0.702 fans. With the hockey variables, every coefficient also decreased by a large amount. Originally all the absolute values of the OLS estimators were greater than 50, but they were all under one besides the constant. All of the days and game times from the original model stayed statistically significant at different levels, but new estimators 12:30 start and 8:30 start became statistically significant.

Module 5, Step 1:

	Dependent variable:
	Num.W.L
(Home vs Away)Home	0.052*
	0.027
(Opp.Division)AL East	-0.03
	0.035
(Opp.Division)AL West	-0.026
	0.036
(Opp.Division)NL Central	-0.098*
	0.057
(Opp.Division)NL East	-0.097
	0.086
(Opp.Division)NL West	-0.111
	0.069
Days Rested	0.0004
	0.001
Innings Pitched	-0.019
	0.028
Earned Runs	-0.049**
	0.02
Strikeouts	0.0002
	0.006
Batters Faced	0.007
	0.009
Pitches Thrown	-0.002
	0.002
Average Leverage Index	-0.091*
	0.053
Base-Outs Runs Saved	0.075***
	0.02
Constant	0.829***
	-0.107
Observations	1,037
R^2	0.258
Adjusted R^2	0.248
Residual Std. Error	0.427 (df = 1022)
F Statistic	25.447*** (df = 14; 1022)
Note:	*p<0.1; **p<0.05; ***p<0.01

<u>Module 5, Step 2, Part 1</u>:

I think that Base-Outs Runs Saved and Earned Runs can have multicollinearity because both of them involve earned runs that a pitcher can allow. Besides that, I don't know of any other variables that can have multicollinearity.

Module 5, Step 2, Part 2:

Variable	GVIF
Home vs Away	1.025
Opponent Division	1.084
Days Rested	1.036
Innings Pitched	11.052
Earned Runs	7.503
Strikeouts	1.692
Batters Faced	8.332
Pitches Thrown	4.295
Average Leverage Index	1.14
Base-Outs Runs Saved	13.435

Module 5, Step 2, Part 3:

My model had no quadratic or interaction term.

Module 5, Step 2, Part 4:

I would probably eliminate Base-Outs Runs Saved since it has multicollinearity with earned runs. It also more than likely has multicollinearity with batters faced, innings pitched, and pitches thrown because it takes it play by play which would involve each pitch/batter faced into consideration. I think that would drive down most of the high GVIF values.

Module 5, Step 3:

	Dependent variable:
	Num.W.L
(Home vs Away)Home	0.065**
	0.027
(Opp.Division)AL East	-0.028
	0.036
(Opp.Division)AL West	-0.030
	0.036
(Opp.Division)NL Central	-0.093
	0.057
(Opp.Division)NL East	-0.091
	0.087
(Opp.Division)NL West	-0.123*
	0.070
Days Rested	0.001
	0.001
Innings Pitched	0.043*
	0.022
Earned Runs	-0.111***
	0.010
Strikeouts	0.0004
	0.006
Batters Faced	-0.002
	0.008
Pitches Thrown	-0.001
	0.002
Average Leverage Index	-0.075
	0.053
Constant	0.763***
	0.106
Observations	1,037
R^2	0.249
Adjusted R^2	0.239
Residual Std. Error	0.429 (df = 1023)
F Statistic	26.038*** (df = 13; 1023)
Note:	*p<0.1; **p<0.05; ***p<0.01

Module 5, Step 4, Part 1:

```
#Step 4#
guardians3 <- lm(Num.W.L ~ as.factor(Player) + as.factor(Home.Away) + as.factor(Opp.Division) + Days.Rested +
                    IP + ER + SO + BF + Pit + aLI, data = guardians_data)
summary(guardians3)

anova(guardians2, guardians3)
```

```
58:30  (Top Level)
Console  Terminal  Background Jobs
R 4.1.1 · ~/Desktop/SAL 213 R Folder/Module 6/
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4291 on 1007 degrees of freedom
Multiple R-squared:  0.2614,	Adjusted R-squared:  0.2401
F-statistic: 12.29 on 29 and 1007 DF,  p-value: < 2.2e-16

> anova(guardians2, guardians3)
Analysis of Variance Table

Model 1: Num.W.L ~ as.factor(Home.Away) + as.factor(Opp.Division) + Days.Rested +
    IP + ER + SO + BF + Pit + aLI
Model 2: Num.W.L ~ as.factor(Player) + as.factor(Home.Away) + as.factor(Opp.Division) +
    Days.Rested + IP + ER + SO + BF + Pit + aLI
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1023 188.59
2   1007 185.40 16    3.1988 1.0859 0.3636
>
```

When adding the dummy variable of pitcher, it does not add predictive power to my model because it is not statistically significant at the 5% level with a p-value of 0.3636.

<u>Module 5, Step 4, Part 2</u>:

```
> guardians4 <- lm(log(Num.W.L) ~ as.factor(Home.Away) + as.factor(Opp.Division) + Days.Rested +
+                    IP + ER + SO + BF + Pit + aLI, data = guardians_data)
Error in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
  NA/NaN/Inf in 'y'
>
```

I couldn't run the log of my dependent variable because my dependent variable is either a 1 or 0, where a 1 is a win and 0 is a loss. Therefore, I will not include the log of my dependent variable.

<u>Module 5, Step 5</u>:

My model will remain the same as Step 3.

	Dependent variable:
	Num.W.L
(Home vs Away)Home	0.065^{**}
	0.027
(Opp.Division)AL East	-0.028
	0.036
(Opp.Division)AL West	-0.030
	0.036
(Opp.Division)NL Central	-0.093
	0.057
(Opp.Division)NL East	-0.091
	0.087
(Opp.Division)NL West	-0.123^{*}
	0.070
Days Rested	0.001
	0.001
Innings Pitched	0.043^{*}
	0.022
Earned Runs	-0.111^{***}
	0.010
Strikeouts	0.0004
	0.006
Batters Faced	-0.002
	0.008
Pitches Thrown	-0.001
	0.002
Average Leverage Index	-0.075
	0.053
Constant	0.763^{***}
	0.106
Observations	1,037
R^2	0.249
Adjusted R^2	0.239
Residual Std. Error	0.429 (df = 1023)
F Statistic	26.038^{***} (df = 13; 1023)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

<u>**Module 5, Step 6**</u>:

```
> setwd("~/Desktop/SAL 215 R Folder/Module 6")
> #Step 6#
> bptest(guardians2)

        studentized Breusch-Pagan test

data:  guardians2
BP = 53.712, df = 13, p-value = 6.79e-07

>
```

The Breusch-Pagan tests rejects the null hypothesis of homoscedasticity in favor of the alternative hypothesis of heteroscedasticity because of the p-value being 6.79e-07.

Module 5, Step 7:

```
> #Step 7#
> bgtest(guardians2)

	Breusch-Godfrey test for serial correlation of order up to 1

data:  guardians2
LM test = 1.1941, df = 1, p-value = 0.2745

> |
```

The Breusch-Godfrey test fails to reject the null hypothesis of no autocorrelation at the 5% level because of the p-value being 0.2745.

Module 5, Step 8:

	Dependent variable:	
	Num.W.L	
	OLS Standard Errors	*Robust Standard Errors*
(Home.Away)Home	0.065**	0.065**
	0.027	0.027
(Opp.Division)AL East	-0.028	-0.028
	0.036	0.034
(Opp.Division)AL West	-0.030	-0.030
	0.036	-0.037
(Opp.Division)NL Central	-0.093	-0.093
	0.057	0.060
(Opp.Division)NL East	-0.091	-0.091
	0.087	0.085
(Opp.Division)NL West	-0.123*	-0.123
	0.070	0.079
Days Rested	0.001	0.001
	0.001	0.001
Innings Pitched	0.043*	0.043*
	0.022	0.022
Earned Runs	-0.111***	-0.111***
	0.01	0.01
Strikeouts	0.0004	0.0004
	0.006	0.006
Batters Faced	-0.002	-0.002
	0.008	0.008
Pitches Thrown	-0.001	-0.001
	0.002	0.002
Average Leverage Index	-0.075	-0.075
	0.053	0.049
Constant	0.763***	0.763***
	0.106	0.101
Observations	1,037	
R^2	0.249	
Adjusted R^2	0.239	
Residual Std. Error	0.429 (df = 1023)	
F Statistic	26.038*** (df = 13; 1023)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

To correct heteroscedasticity, I used Robust Standard errors.

Module 5, Step 9:

My initial analysis doesn't change at all. I ended up eliminating one variable (Base-Outs Runs Saved) because it had similar components to other variables in the model that I thought would explain better despite making the model larger. The intercepts and p-values stayed the same, and all I had to do was adjust for heteroscedasticity by using Robust Standard Errors. Like I originally concluded when I first ran this model, there is still much randomness that comes from a team's hitting and bullpen performance that affects whether they win the game. The constant (Away and AL Central games) and earned runs were both statistically significant at the 1% level in both the normal OLS standard errors and robust standard errors models. AL Central was statistically significant because it is the Guardians division many of their series consist of games within the division, but I'm not sure why away games are significant. Home games were significant at the 5% level in both the normal OLS standard errors and robust standard errors models because approximately half of the games are played at home. Lastly, NL West games and Innings Pitched were statistically significant at the 10% level in both the normal OLS standard errors and robust standard errors models. Innings pitched is statistically significant because a starting pitcher that goes deep into games often leads to the team winning, but I'm not sure why NL West games are statistically significant.