A Good Walk Ruined

An Analysis of Golf Success Measures and Statistics

Outline

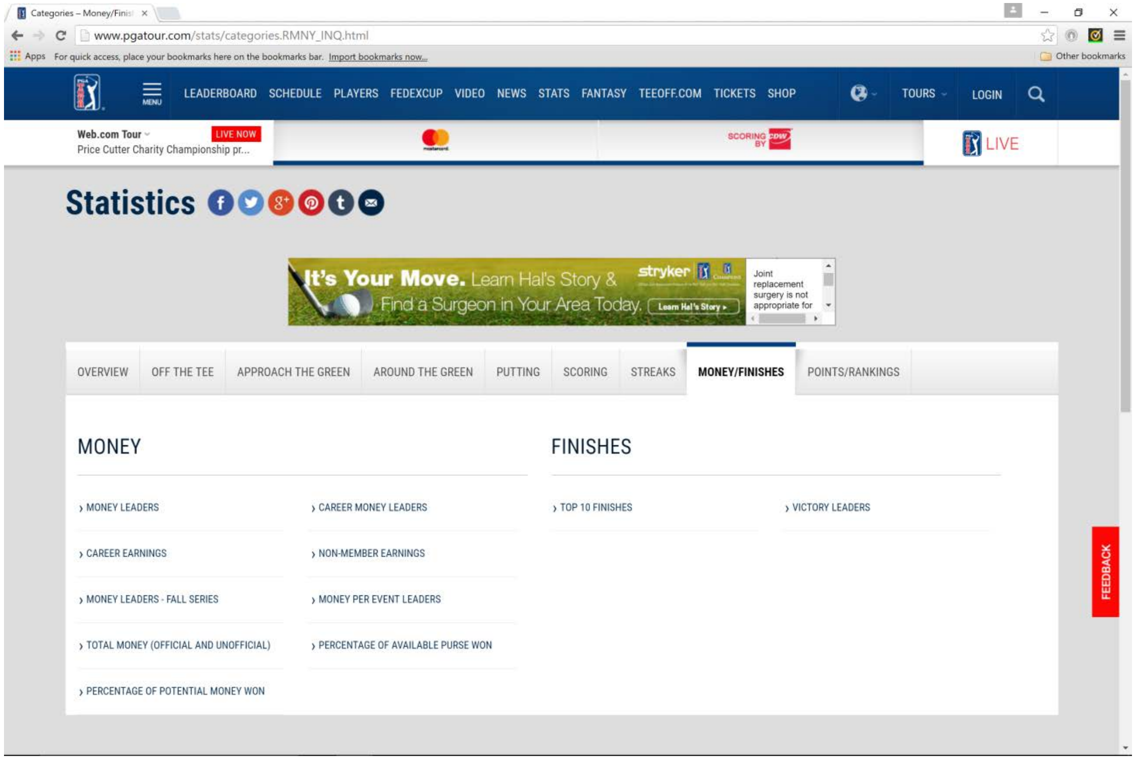
- Practice Range: Collecting the Data
- Front Nine: Traditional Statistics
- Back Nine: New Statistics
- 19th Hole: Time For Fun

Outline

- Practice Range: Collecting the Data
- Front Nine: Traditional Statistics
- Back Nine: New Statistics
- 19th Hole: Time For Fun

Practice Range: The Data

- Data Scraped from www.pgatour.com/stats using Beautiful Soup


Categories – Money/Finis
www.pgatour.com/stats/categories.RMNY_INQ.html
Apps For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Other bookmarks
MENU
LEADERBOARD
SCHEDULE
PLAYERS
FEDEXCUP
VIDEO
NEWS
STATS
FANTASY
TEEOFF.COM
TICKETS
SHOP
TOURS
LOGIN
Web.com Tour
LIVE NOW
Price Cutter Charity Championship pr...
SCORING BY CDW
LIVE
Statistics
It's Your Move. Learn Hal's Story & Find a Surgeon in Your Area Today.
stryker
Learn Hal's Story
Joint replacement surgery is not appropriate for
OVERVIEW
OFF THE TEE
APPROACH THE GREEN
AROUND THE GREEN
PUTTING
SCORING
STREAKS
MONEY/FINISHES
POINTS/RANKINGS
MONEY
FINISHES
› MONEY LEADERS
› CAREER MONEY LEADERS
› TOP 10 FINISHES
› VICTORY LEADERS
› CAREER EARNINGS
› NON-MEMBER EARNINGS
› MONEY LEADERS - FALL SERIES
› MONEY PER EVENT LEADERS
› TOTAL MONEY (OFFICIAL AND UNOFFICIAL)
› PERCENTAGE OF AVAILABLE PURSE WON
› PERCENTAGE OF POTENTIAL MONEY WON
FEEDBACK

Practice Range: Categories

- Traditional Performance Statistics:
  - Off the Tee
  - Approaching the Green
  - Around the Green
  - Putting
- New Performance Statistics:
  - Strokes Gained Off the Tee
  - Strokes Gained Approaching the Green
  - Strokes Gained Around the Green
  - Strokes Gained Putting
- Success Metrics:
  - Money Earned
  - Fedex Cup Points

Outline

- Practice Range: Collecting the Data
- **Front Nine: Traditional Statistics**
- Back Nine: New Statistics
- 19th Hole: Time For Fun

Front Nine: What To Test

First test: Multiple linear regression on traditional statistics

Fedex Cup points on:

drd – Driving Distance
dra – Driving Accuracy
gir – Greens in Regulation
ssv – Sand Saves
scr – Scrambling
pth – Proximity to Hole
pthatg – Proximity to Hole from Around The Green
pmd – Putts Made Distance
ppr – Putts Per Round

Front Nine: Predictive Model

Regress Fedex Cup Points on all* variables

Coefficient*	Estimate	Pr(>|t|)
Intercept	210.2	0.946
drd	21.3	1.43E-06
dra	13.5	0.053
gir	72.5	3.54E-06
ssv	7.5	0.133
scr	-14.4	0.244
pth	-2.2	0.154
pthatg	-16.6	2.00E-03
pmd	25.6	0.02
ppr	-332.7	6.00E-04

R-Squared:	0.5062
Adj R-Sq:	0.4823
p-value:	<2.2E-16

Conclusions

Lots of coefficients with p-values less than .05

Reasonable R-Squared

Perhaps just use highlighted variables

Front Nine: Reduced Model

Regress Fedex Cup Points on reduced variable set

Coefficient*	Estimate	Pr(>|t|)
Intercept	-31.9	0.988
drd	14.4	3.05E-06
gir	88.2	5.17E-14
pthatg	-16.6	.000456
ppr	-312.2	.000150
pmd	-25.9	.016766

R-Squared:	0.4831
Adj R-Sq:	0.4695
p-value:	<2.2E-16

VIF				
drd	gir	pthatg	ppr	pmd
1.23	1.76	1.61	2.94	1.49

Conclusions

All p-values are small, we can say this model is reasonable

However, VIF is troublesome and we should be wary

Further analysis reveals failure on assumptions

Front Nine: Optimal Model

After Box-Cox Transform and Stepwise Regression

AIC yields: gir + ppr + drd + pthatg + dra + pmd

BIC yields: gir + ppr + drd + pthatg
Choose this for its simplicity and lack of overlap of variables

Coefficient*	Estimate	Pr(>|t|)
Intercept	51.927	6.27E-05
drd	0.116	1.45E-06
gir	0.973	<2E-16
pthatg	-0.093	0.00745
ppr	-4.302	4.01E-14

VIF			
drd	gir	pthatg	ppr
1.21	1.69	1.43	2.02

R-Squared:	0.5653
Adj R-Sq:	0.5562
p-value:	<2.2E-16

Conclusions

All variables should be included

VIF no longer an issue

Model appears to meet assumptions

R-Squared of 0.5562

Outline

- Practice Range: Collecting the Data
- Front Nine: Traditional Statistics
- Back Nine: New Statistics
- 19th Hole: Time For Fun

Back Nine: Strokes Gained

Strokes Gained: Statistic describing how well a golfer did compared to the "baseline" from a specific distance and lie

Example: Golf player takes shot from 400 yards out in fairway to 100 yards out in the rough

Position	Lie	Baseline Strokes to Hole	Shots	
400	Fairway	4.05	1	
100	Rough	3.20	2	**Strokes Gained**
	Difference	**0.85**	**1**	**-0.15**

Back Nine: What To Test

Multiple Linear Regression

Fedex Cup points on:

sg_ott – Shots Gained Off the Tee
sg_aptg – Shots Gained Approaching the Green
sg_artg – Shots Gained Around the Green
sg_putt – Shots Gained Putting

Back Nine: SG Optimal Model

After Box-Cox Transform and Stepwise Regression

Optimization reveals that all variables should be used

Coefficient*	Estimate	Pr(>|t|)
Intercept	15.2045	<2E-16
sg_ott	4.369	<2E-16
sg_aptg	4.117	<2E-16
sg_artg	4.579	1.97E-10
sg_putt	4.268	<2E-16

R-Squared:	0.7108
Adj R-Sq:	0.7047
p-value:	<2.2E-16

VIF			
sg_ott	**sg_aptg**	**sg_artg**	**sg_putt**
1.17	1.15	1.07	1.07

Conclusions

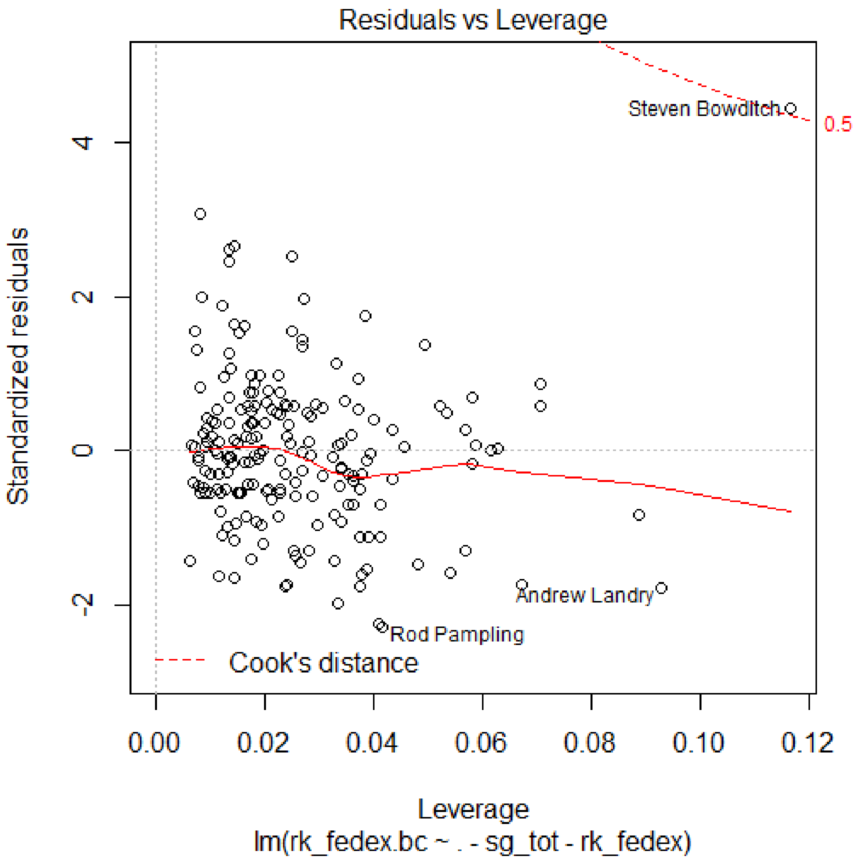
All variables should be included

Multi-collinearity not an issue

All variables meet assumptions independently

R-Squared of 0.7047

Back Nine: Leverage Issue



Back Nine: Comparing Models

- **R-Squared:** Strokes Gained Model R-Squared of 0.7047 beats Traditional Model R-Squared of 0.5562

- **AIC/BIC Confirm:** AIC and BIC both confirm the Strokes Gained Model is better fit

Outline

- Practice Range: Collecting the Data
- Front Nine: Traditional Statistics
- Back Nine: New Statistics
- 19th Hole: Time For Fun

19th Hole: Prediction

Can we use model to predict?

- Strokes Gained Issue: Derived statistic, requires knowledge of past performance and accurate measurement
- Strokes Gained Issue: Was not computed before 2002, altered statistic before 2004
- Traditional Model Issue: Not as accurate!

19th Hole: Tiger Woods

How would Tiger of old fare today?

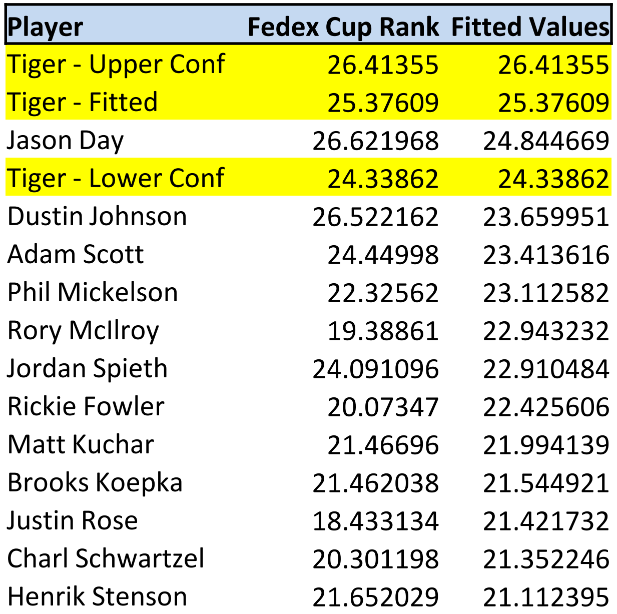
Player	Fedex Cup Rank	Fitted Values
Jason Day	26.62197	24.39598
Dustin Johnson	26.52216	23.84958
Tiger - Upper Conf	23.36207	23.36207
Tiger - Fit	22.37369	22.37369
Phil Mickelson	22.32562	22.23771
Tiger - Lower Conf	21.38531	21.38531
Jordan Spieth	24.0911	21.37981
Rickie Fowler	20.07347	20.93607
Brooks Koepka	21.46204	20.48021
Adam Scott	24.44998	20.21326
Henrik Stenson	21.65203	20.15537
Charl Schwartzel	20.3012	19.66504
Hideki Matsuyama	21.2478	19.4972
Brandt Snedeker	22.42809	19.13311
Matt Kuchar	21.46696	19.10473

Tiger of 2004, using traditional statistics and traditional stats model.

Tiger would have fared well, but would have fallen behind Jason Day and Dustin Johnson

19th Hole: Tiger Woods

Comparing Tiger on Strokes Gained Model



Player	Fedex Cup Rank	Fitted Values
Tiger - Upper Conf	26.41355	26.41355
Tiger - Fitted	25.37609	25.37609
Jason Day	26.621968	24.844669
Tiger - Lower Conf	24.33862	24.33862
Dustin Johnson	26.522162	23.659951
Adam Scott	24.44998	23.413616
Phil Mickelson	22.32562	23.112582
Rory McIlroy	19.38861	22.943232
Jordan Spieth	24.091096	22.910484
Rickie Fowler	20.07347	22.425606
Matt Kuchar	21.46696	21.994139
Brooks Koepka	21.462038	21.544921
Justin Rose	18.433134	21.421732
Charl Schwartzel	20.301198	21.352246
Henrik Stenson	21.652029	21.112395

Tiger of 2004, using Strokes Gained statistics and model

We might have expected Tiger of old to perform well beyond the field.

19th Hole: Ben Townson

How would I fare today?

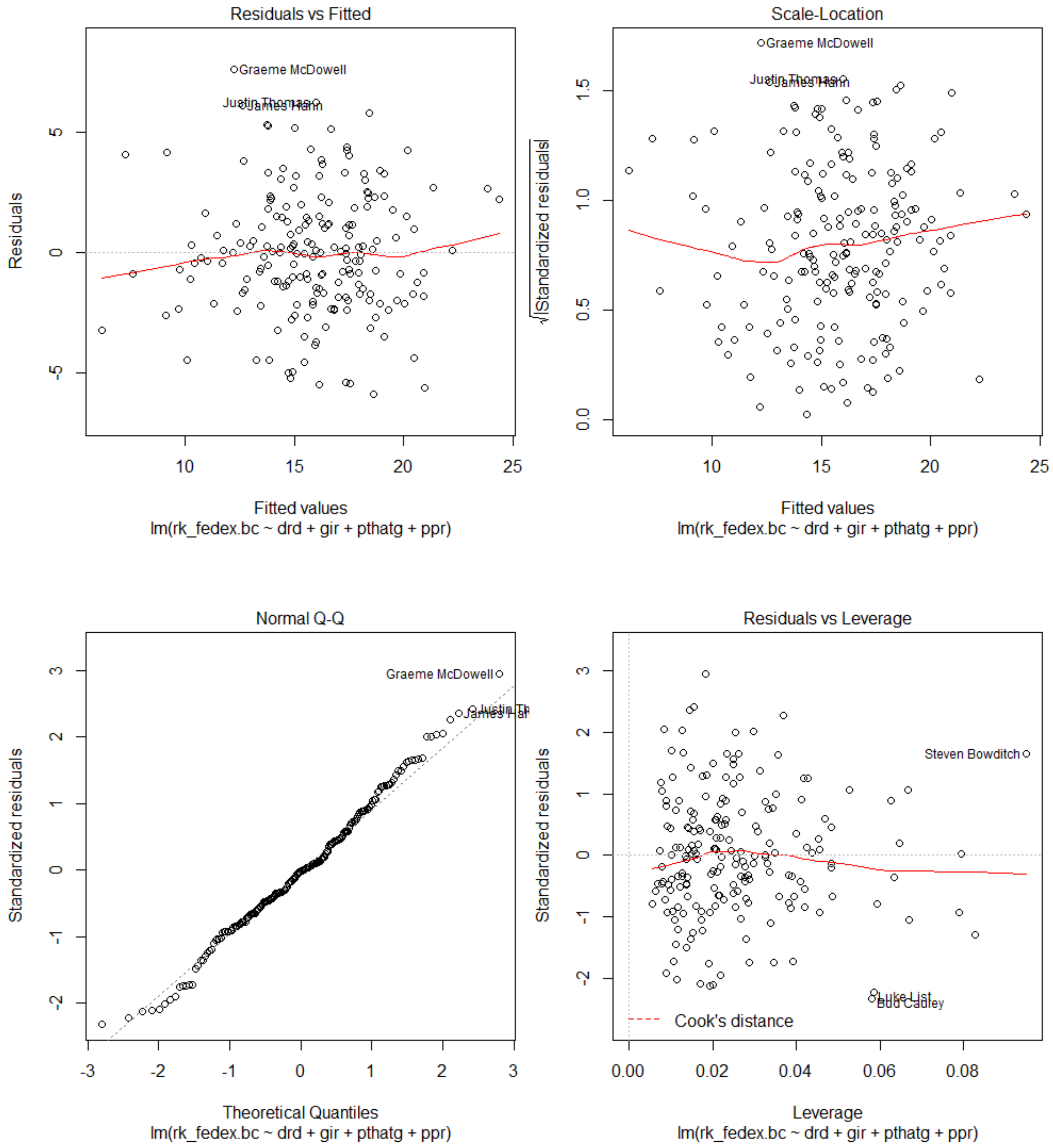
Player	Fedex Cup Rank	Fitted Values
Jason Day	26.62197	24.39598
Dustin Johnson	26.52216	23.84958
Tiger - Upper Conf	23.36207	23.36207
Tiger - Fit	22.37369	22.37369
Phil Mickelson	22.32562	22.23771
Tiger - Lower Conf	21.38531	21.38531
Jordan Spieth	24.0911	21.37981
Rickie Fowler	20.07347	20.93607
Brooks Koepka	21.46204	20.48021
Adam Scott	24.44998	20.21326
Henrik Stenson	21.65203	20.15537
Charl Schwartzel	20.3012	19.66504
Hideki Matsuyama	21.2478	19.4972
Brandt Snedeker	22.42809	19.13311
Matt Kuchar	21.46696	19.10473

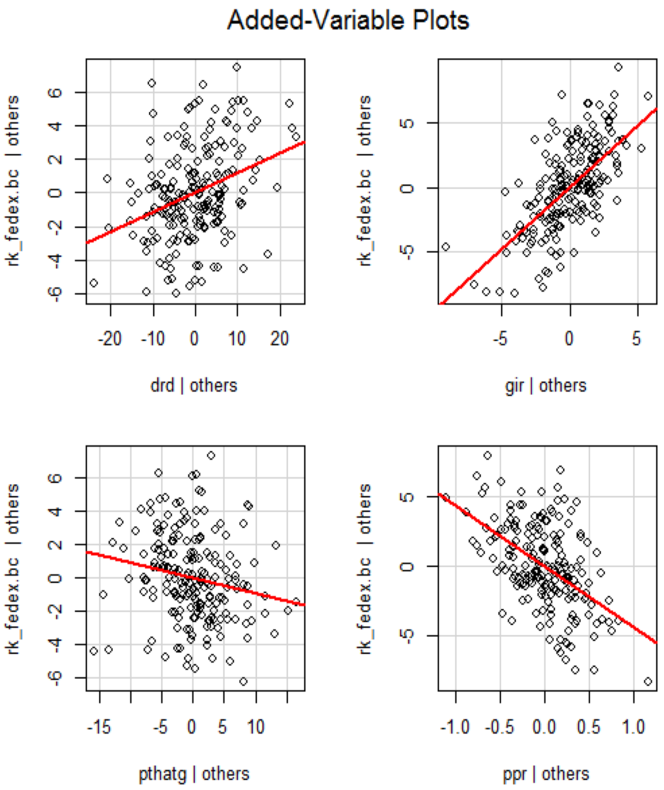
Player	Fedex Cup Rank	Fitted Values
Ben - Upper Conf	-29.16589	-29.16589
Ben - Fit	-37.7505	-37.7505
Ben - Lower Conf	-46.33512	-46.33512

Thanks!

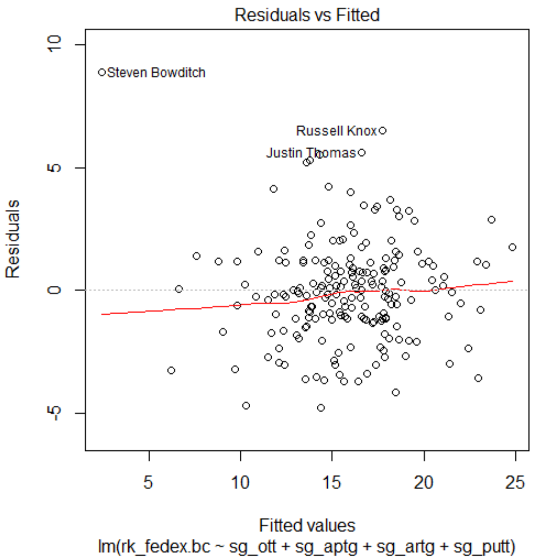
Unless…

Assumptions Plots – Std Model

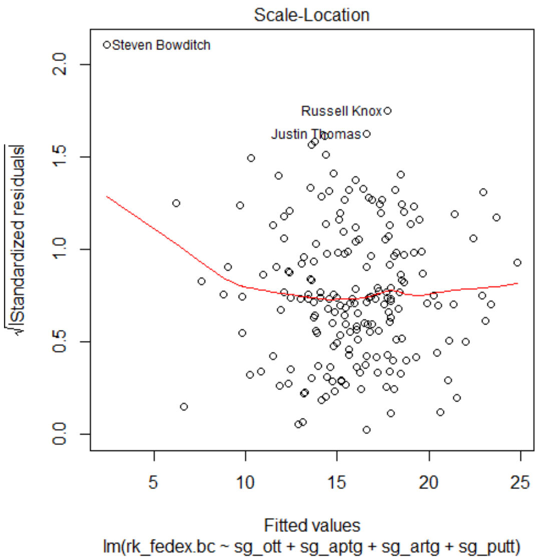




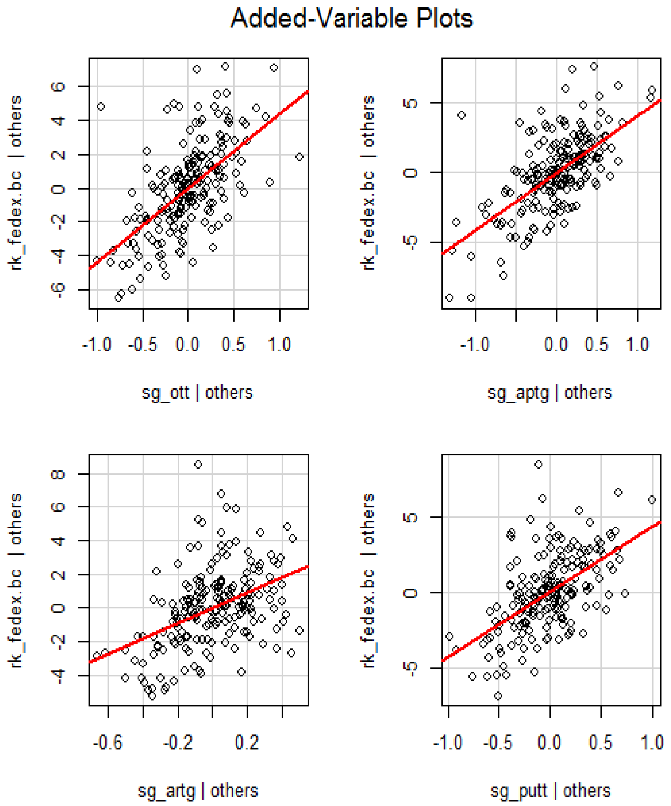
Assumptions Plots – SG Model



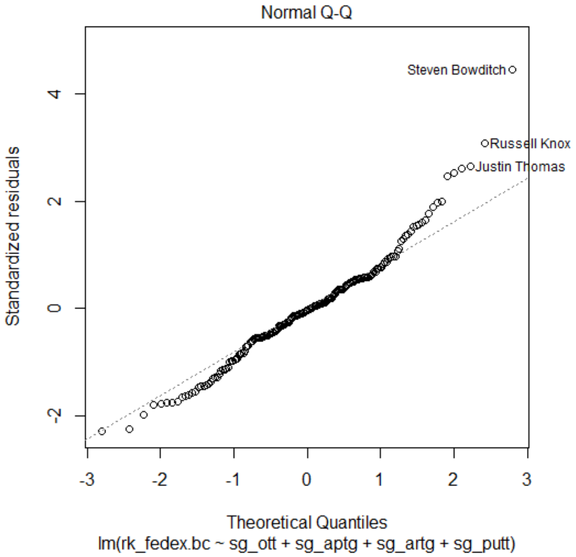
Residuals vs Fitted



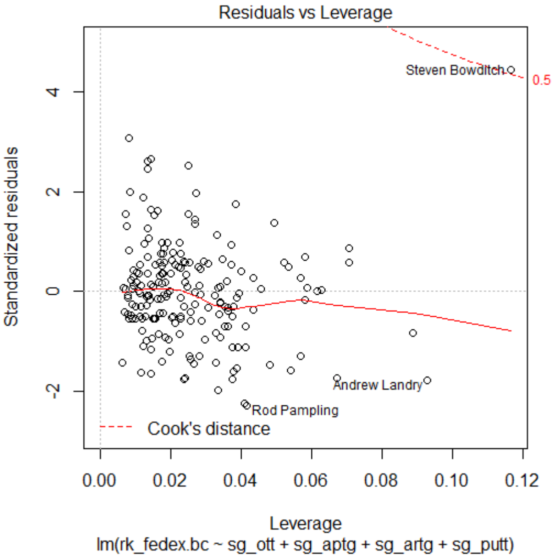
Scale-Location



Added-Variable Plots



Normal Q-Q



Residuals vs Leverage