Abbreviation Code: LS, least squares; GD, Greatest Deviation Correlation Coefficient; lts, least trimmed squares robust regression; L1, absolute value or l1; rreg, M-estimation , default weights (file name: gamefits)

Tables of fitting Atlanta Brave game time in minutes against 6 factors; runs, hits, number of pitchers used, number left on base LOB, walks BB, and strikeouts Ks.
Major league baseball is trying to reduce the games times by a few minutes this year by implementing (supposedly) several time saving features. So it might be fun to see how much time each of the above factors contribute to the game length. At the same time it is of interest to compare several multiple regression methods. After each week the data and fitting will be updated to compare the coefficients of the regressor variables. n will be the number of games in the fit. Since the true answer is not known, it should be interesting to see how the various regressions compare on the coefficients. Each table includes all preceding games, they are in chronologically order, no games deleted.
Games through 11 April 2002

n=9	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	44.335	-1.652	2.519	-1.123	5.724	4.364	0.238
GD	17.619	-1.589	2.541	-1.921	8.478	4.983	-0.462
L1	45.239	-2.052	2.645	-0.343	4.126	5.678	0.523
lts	need	more	observ	twice	as many	as variab	
rreg	45.267	-2.146	2.652	-0.360	4.882	4.562	0.524
remark: rreg failed to converge							

Games through 16 April 2002

n=14	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	24.29	-2.296	2.896	6.621	4.487	-0.402	0.727
GD	16.60	-0.326	1.549	6.429	5.032	-1.189	1.739
L1	26.47	-0.467	1.395	7.116	5.889	-2.382	0.752
lts	need	more	observ	twice	as many	as variab	
rreg	22.46	-0.723	1.700	6.5969	4.687	0.886	1.496

Games through 23 April 2002

n=20	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	35.16	-1.601	2.422	6.332	3.922	0.712	0.332
GD	45.49	-2.248	2.951	3.099	1.657	3.342	1.988
L1	38.077	-1.175	1.688	7.026	3.949	0.400	0.613
lts	37.10	1.011	0.401	6.840	5.158	-1.491	0.971
rreg	34.95	-1.429	2.274	6.385	3.990	0.608	0.392
remark: lts used only 18 of 20 observations							

Games through 29 April 2002

n=26	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	42.187	-1.051	1.800	5.635	3.402	0.993	1.133
GD	44.829	-1.973	1.891	4.807	3.178	2.168	1.320
L1	44.603	-0.053	0.456	6.257	4.085	0.192	1.347
lts	42.550	0.420	0.197	6.505	4.553	-0.571	1.383
rreg	39.792	-0.361	1.265	5.939	3.698	0.432	1.407
remark: Number observation used in lts is 23/26							

Note games above, the intercepts seemed to move toward the GD intercept, The GD BB coefficient remains different from others. Ks coefficients nearly the same!

Games through 5 May 2002, the Sunday game had the time missing so the regressions were used to predict the value and then the time was obtained elsewhere. The results: LS: 199,GDCC: 196, L1: 196, lts: 193, rreg: 197. The actual was 174 so none of them did very well.

n=31	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	43.41	-1.336	1.884	5.440	3.158	1.662	1.135
GD	47.384	-2.215	1.905	3.885	3.344	2.600	1.414
L1	42.790	-1.554	1.577	5.320	3.467	1.674	1.282
lts	34.073	0.433	0.690	6.367	4.204	-0.534	1.736
rreg	41.875	-0.773	1.386	5.645	3.435	1.227	1.373
remark: Number observation used in lts is 27/31							

Note that lts, evidently by deleting games, remains most different.

Games through 14 May 2002. There was one 16 innings game in this group with a game time of 5 hours and 19 minutes. Note the intercepts of LS and lts changed quite a bit. Fairly large changes on runs coefficients for all methods. GD became like the others on the pitchers coefficient. lts changed its sign on the BB coefficient. Estimates of scale from LS, lts, and GD are respectively, 12.48, 8.74, 9.95. Thus games can now be estimated with an standard deviation error of about 10 minutes. The importance of LOB on game times is becoming apparent.

n=39	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	36.845	-0.977	1.827	5.375	3.300	1.724	1.412
GD	46.579	-0.0350	0.816	5.884	3.579	0.378	1.364
L1	43.748	0.041	0.473	6.502	4.274	0.089	1.090
lts	44.874	0.3061	0.527	5.912	4.050	0.604	1.214
rreg	45.955	-0.095	0.773	5.613	3.846	0.648	1.348
remark: Number observation used in lts is 35/39							

Games through 22 May 2002. After 46 games the median values of the variables are: time, 173.5 minutes; runs, 7; hits, 16; number of pitchers used, 7; LOB, 14.5;

BB, 6.5; Ks, 13. Using the median data, and using the multiple regression model,LS game time is 175 minutes, and GD is also 175, although GD is slightly lower without rounding. All methods had a large increase in the intercept values. All runs coefficients became negative. All BB coefficients increased. It is beginning to look like there is not much difference between the robust methods and so maybe the choice would be the method that gives the best-related inference. A game with 1 run, 3 hits, 2 pitchers(how unlikely in today's game), 0 LOB, 0 BB, and 13 Ks gives an estimated game time of 85 minutes. The minimum game time has been 126 minutes, with 4 runs, 11 hits, 5 pitchers used, 8 LOB, 5 BB, and 13 Ks.

n=46	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	46.73	-1.103	1.626	5.176	2.860	2.309	1.323
GD	57.109	-1.290	1.322	3.769	3.078	2.536	1.378
L1	57.826	-1.585	1.305	5.329	3.080	2.090	0.971
lts	51.445	-0.575	0.889	5.725	3.579	0.895	1.301
rreg	56.588	-0.533	0.819	5.145	3.318	1.505	1.171
remark: Number observation used in lts is 41/46							

Games through 30 May 2002, this actually means, the boxscores that appear by then and not the game on the 30th (same for all of above). GD refused to budge from its 3.7 coefficient on pitchers which is somewhat different than other estimators. The Pearson multiple correlation is 0.911 whereas GD's is 0.923. The LS residual SD is 12.05 and GD's is 10.02. The LS estimate of standard error on LOB is 0.816. So all coefficients except lts are within about ½ SE. The LS estimate of standard error on pitchers is 1.11, so only GD and lts differ by about one SE. LS thinks all coefficients are significantly different from 0 at the 1% level or better except runs where the P-value is 0.09

n=53	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	48.81	-1.405	1.955	4.740	2.580	2.494	1.433
GD	60.011	-1.489	1.624	3.778	2.922	2.363	1.209
L1	57.436	-1.428	1.344	4.880	3.037	2.274	1.051
lts	54.524	-0.180	0.462	5.877	3.847	0.568	1.175
rreg	57.949	-0.835	1.179	4.737	3.013	1.736	1.287
remark: Number observation used in lts is 47/53							

Games through 5 of June 2002. Almost all intercepts increased. Lts coefficients increased in runs and decreased in hits opposite of other four methods. GD and lts remain at opposite ends on pitchers coefficient. Lts is the odd ball on BB. P-value on runs for LS increased to 0.299.

n=58	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	52.200	-0.891	1.465	4.913	2.937	2.140	1.209
GD	62.094	-1.190	1.303	3.803	3.179	2.154	1.115
L1	60.075	-1.303	1.149	4.923	3.174	2.128	0.936
lts	56.403	-1.062	1.175	5.265	3.000	1.683	1.381
rreg	60.377	-0.592	0.932	4.764	3.178	1.590	1.144
remark: Number observation used in lts is 52/58							

Below are the coefficients for box scores through June12, 2002. This week there were unusually long games as Atlanta played American League teams in their ballparks and there were extra inning games. Most dramatic changes occurred in the intercept (10 minutes) and hits coefficient of LS (1.465 to 2.432). The lts methods had a large change in the pitchers coefficient. The estimated SD's of the slopes for LS and GD were(same order as table)
LS: 0.853, 0.785, 1.219, 0.855, 0.997, 0.506
GD:0.843, 0.759, 1.501, 1.076, 1.016, 0.523

n=65	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	43.324	-1.398	2.432	5.027	2.049	2.643	1.671
GD	62.056	-0.811	1.348	3.209	2.996	2.686	1.172
L1	61.982	-0.811	1.046	4.668	3.113	2.181	0.913
lts	59.096	-0.937	1.343	4.316	3.299	1.853	0.992
rreg	57.879	-0.552	1.029	4.472	3.081	1.756	1.180
remark: Number of observations used in lts is 58/65							

Below are the multiple regression results for boxscores of games listed by 26 June, 2002. There was a continuation of slow games but the slope estimates did not change dramatically in any of the methods.
The USA article said a pitching change took 4 minutes on the average. Currently the GDCC method at 4.186 is closest to that time. In all 6 slopes and the intercept the magnitude of the LS estimate is either the largest or smallest. LS seems to be the outlier method of the five methods.

n=77	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	45.006	-1.591	2.407	5.426	2.015	2.548	1.510
GD	59.758	-1.299	1.308	4.186	3.252	2.224	1.053
L1	58.872	-1.376	1.280	4.846	3.091	2.231	0.988
lts	59.930	-0.942	1.347	5.144	3.051	1.814	0.985
rreg	59.244	-0.738	1.216	4.915	2.937	1.736	1.065
remark: Number observation used in lts is 69/77							

Below is the analysis and results for boxscores occurring on 2 July 2002. These are the 82 games for the Atlanta Braves and one-half of the regular season. A more thorough analysis follows. In the tradition of baseball, the second half of the season will be analyzed with a new beginning. Sort of pre and post all-star game analysis.

n=82	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	45.612	-1.356	2.343	5.746	1.799	2.350	1.571
GD	60.847	-0.923	1.097	4.159	3.262	2.131	1.107
L1	57.258	-1.421	1.342	4.912	3.028	2.269	1.060
lts	61.311	-1.116	1.373	5.557	2.441	2.234	0.929
rreg	57.700	-0.687	1.287	5.374	2.664	1.624	1.191
remark: Number observation used in lts is 73/82							

Below is a table of classical normal (LS) statistics, standard error, t-value, and Prob-value. The analogous statistics are computed for GDCC so a visual comparison can be made.

	inter	runs	hits	pitchers	LOB	BB	Ks
LS:std	7.86	0.804	0.756	1.129	0.783	0.955	0.437
t-value	5.800	-1.686	3.097	5.087	2.296	2.460	3.591
Prob-V	0.000	0.095	0.002	0.000	0.024	0.016	0.001
GD:std	-	0.755	0.682	1.224	0.883	0.972	0.460
normal-asympt	-	-1.222	1.608	3.396	3.693	2.192	2.405
Prob-V	-	0.110	0.053	0.000	0.000	0.014	0.008

The biggest difference is in the significance of “hits.” This is the variable in which all the robust methods disagree somewhat with LS. The GD is less significant and the LS estimate of the slope, 2.343, is 1.82 std (GD std) away from the GD estimate, 1.097. So does a hit cost the game 2.343 or 1.097 minutes of game time? (2.343-1.097)/0.682=1.82.

Since the correlations among the variables are the leading measure of the relationship between variables, listed below is Pearson's (r) and the sine of (pi time GDCC/2),(sgd). These are comparable quantities. If r is smaller than sgd then there are "outliers" which diminish the relationship. If r is much much greater than sgd, then there are "outliers" which exaggerate the relationship.

upper, r lower sdg	time	runs	hits	pitchers	LOB	BB	Ks
time	1.000	0.359 0.355	0.531 0.494	0.816 0.733	0.808 0.771	0.727 0.636	0.549 0.355
runs		1.000	0.772 0.665	0.382 0.409	0.305 0.319	0.280 0.319	-0.085 0.038
hits			1.000	0.440 0.477	0.519 0.636	0.229 0.355	-0.025 0.114
pitchers				1.000	0.624 0.527	0.657 0.460	0.405 0.227
LOB					1.000	0.774 0.650	0.284 0.264
BB						1.000	0.248 0.152
Ks							1.000

Note that there are several correlations in which r is much bigger that sdg, (Ks and time), (runs and hits), (pitchers and BB), (LOB and BB), so some classical correlations have been inflated by some unusual games. There are two correlations in which sgd is much bigger than r, (hits and LOB) and (hits and BB), so two correlations have been diminished by some non-typical games.

From the USA today article, last year's National League games averaged 173 minutes and the goal for this year is 160 minutes. From the first 82 games (1/2 of a season), the Braves games averaged 178.8 minutes and the median was 170. If games are restricted to 9 innings, then the average was 168.2 and the median 166.5. The GDCC average using a GDCC normal quantile plot fit on "time" using all games (12 were extra innings) was 173.6 minutes. The time goal is not being met by the Braves.
The medians of the regressor variables are:
runs 7,hits 16, pitchers 7, LOB 14, BB 6, Ks 13. If the GDCC regression equation is used(3^{rd} table above, GD line) with these medians, an average game time of 173.9 minutes is obtained. To achieve an 160 minutes average one less in each of the 5 variables; hits, pitchers, LOB,BB, and Ks would give 11.8 minutes, which is only 2.2 minutes short of the goal. Note, the normal GDCC quantile plot reveals that at least 12 games are abnormal; they deviate from the fitted line substantially at the upper end. GDCC is robust so the fit is not influenced by these "outliers".

We now start with the 2nd half of the season. A fit after 7 games gave very unreasonable results. LS, GDCC, rreg all gave about the same worthless estimates. lts and L1 would not give fit. So the first run will be after 17 games.

2nd half of season

n=17	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	102.85	5.897	-4.871	0.569	4.976	0.936	1.311
GD	75.871	9.084	-7.959	-1.027	8.108	-0.675	3.467
L1	113.68	6.914	-6.054	0.601	5.599	0.169	1.002
lts	133.22	6.187	-6.327	0.982	4.873	1.147	0.455
rreg	105.01	6.121	-5.096	0.614	5.066	0.750	1.245
remark: Number observation used in lts is 15/17							

It can be seen that if the first half of season with 82 games gives reasonable estimates then the second half of season with 17 games gives very different results. It is clear that this multiple regression is a tough one to estimate even though there is not a high correlation between any of the variables.

Below is the fit after 44 games of the 2nd half of the 2002 season.

n=44	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	71.185	0.111	0.512	4.134	3.191	0.536	1.372
GD	79.498	-1.261	2.177	2.593	2.035	1.053	1.296
L1	79.499	-0.755	1.589	3.809	1.880	1.053	1.226
lts	102.854	-0.185	0.924	3.481	2.422	-0.316	0.199
rreg	81.108	0.243	0.551	3.648	2.910	0.629	0.961
remark: Number observation used in lts is 39/44							

It can be seen that there seems to be some difference between first and second half of the season with this game time data. After 35 games into the 2nd half of the season the residuals using the models from the 1st 82 games were computed. This was done only for LS and GD. From the 1st half , LS had a residual SE of 14.1 and GD of 11.1 Fitting the old model to the new data, the residuals of each model had their SE estimated by the two different methods. LS the usual sum of squares method gave for LS 13.2 and for GD 13.6. The GD method uses a GD fit on a quantile plot, and this method gave for LS 12.9 and for GD 12.7. Thus, LS is slightly better with sums of squares, but GD is slightly better with the quantile method. In either case, the SE's are reasonably good. So even thought the slopes of the fits of the six variables seem to be quite different after 44 games, the first half model fits fairly well. A plot of LS residuals minus GD residuals against normal quantiles reveals that there are about 11 games with a systematic difference in the prediction. However, none are very large, under two minutes.

Below are the fitting results for the first 55 games of the 2nd half of the season. From 44 to 55 games there were large changes in the coefficients of LOB from LS and GD; hits

from LS, GD, and rreg; BB from lts. L1 had the least changes, but the coefficients are not yet really close to the first half of the season.

n=55	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	71.425	-0.504	1.507	3.997	1.846	1.294	1.506
GD	75.616	-1.782	3.421	2.950	0.166	2.842	1.062
L1	75.812	-0.500	1.825	3.419	1.481	1.289	1.425
lts	101.15	0.132	0.999	3.298	1.759	0.827	0.255
rreg	83.716	-0.204	1.430	3.478	1.641	1.328	0.934
remark: Number observation used in lts is 49/55							

Below is the run of the first 69 games of the 2nd half of the season. Much smaller changes were obtained in all of the coefficients. Most of the methods have similar values. The biggest differences were LS:Pitchers from other four methods. lts:Ks value was not consistent with the other methods. GD:pitchers remains lower. In comparing this run with the coefficients of the first 82 games, only the coefficients from Ks and hits are reasonably close.

n=69	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	72.193	-0.652	1.540	4.221	1.547	1.474	1.511
GD	79.301	-0.525	2.021	2.670	1.154	1.632	1.421
L1	82.306	-0.925	2.132	3.428	0.803	1.823	1.136
lts	93.218	-0.462	1.760	3.083	1.153	1.169	0.662
rreg	82.891	-0.360	1.487	3.560	1.346	1.588	1.044
remark: Number observation used in lts is 62/69							

Several games either were not played or tied and not completed, so there are only 79 games for the 2nd half of the season. Some dramatic changes in the slope coefficients: 4 of the 5 methods changed sign on the runs coefficient, and GD became fairly large and its hits coefficients turned negative. These changes are in the two most highly correlated factors, runs and hits. Since in general there are two hits for each run, one way to compare the hits-runs coefficients is to take 2 times hits plus runs to estimate the average time jointly. This gives in same order as table; 2.550, 1.728, 3.929, 3.743, 2.652. GD and L1 are over two minutes apart. The pitchers coefficient dropped substantially on all methods and most for LS. There was a large increase in the GD LOB coefficient accompanied by a large decrease in the BB coefficient. The lts method also had a fairly large increase in the BB coefficient. The Ks factor was the only one without dramatic changes. Since the 2nd half of season seems different than the first half in the relationship between game time and these 6 dependent variables, some more analysis needs to be done before merging the data for a combined estimate.

n=79	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	74.685	0.1184	1.216	2.729	1.975	1.667	1.409
GD	81.862	1.9469	-0.109	1.767	3.270	0.181	1.221
L1	87.141	-0.301	2.115	1.772	0.892	2.409	0.845
lts	82.636	0.167	1.788	1.646	1.371	2.079	0.972
rreg	88.544	0.454	1.099	2.154	1.737	1.639	0.818
remark: Number observations used in lts is 71/79							

Finally, all games are pooled into one large data set of 161 games and the multiple regressions are compared. First comes the table of regression coefficients.

n=161	intercept	runs	hits	pitchers	LOB	BB	Ks
LS	57.770	-1.0024	2.1346	4.4377	1.3399	2.6926	1.5837
GD	70.775	-0.0978	1.4260	2.5922	2.0770	2.3418	1.3104
L1	65.271	-0.9782	1.8744	4.2854	1.6160	2.1295	1.3907
lts	70.741	-0.6007	2.0306	2.9072	1.5964	2.2312	1.0973
rreg	70.308	-0.4002	1.4613	3.8278	1.7824	2.0900	1.2012
remark: Number observations used in lts is 144/161							

Below is a table of classical normal (LS) statistics, standard error, t-value, and Prob-value. The analogous statistics are computed for GDCC so a visual comparison can be made. A earlier table was produced after the first 82 games.

	inter	runs	hits	pitchers	LOB	BB	Ks
LS:std	5.710	0.6161	0.5831	0.7560	0.5975	0.6864	0.288
t-value	10.11	-1.627	3.661	5.869	2.242	3.923	5.483
Prob-V	0.000	0.105	0.003	0.000	0.026	0.001	0.000
GD:std	-	0.874	0.718	0.967	0.832	0.887	0.374
normal-asympt	-	-0.111	1.984	2.679	2.494	2.639	3.498
Prob-V	-	0.911	0.047	0.007	0.012	0.008	0.000

Overall Summary

The whole point of this summer long study was to compare some S+ multiple regression routines with GD by seeing how well they compared on baseball data that does not fit most fo the theoretical assumptions. All the regressor variables were discrete. To come to a conclusion is very difficult. First, the first half and second half of the seasons were compared by looking at the regression coefficients. All five of the methods had some substantial differences in some of the regression slope estimates. Probably the most stable coefficient was for the Ks factor, the one which would not have been guessed to be very important. Yet each strike out seems to add at least a minute to each game. To help in the understanding of this data, the Braves games on ESPN on Wednesdays were noted

to be almost 30 seconds longer between each inning than a game appearing on TBS. This is only based on several games in April of 2003. Also in three different games the time it took to relieve a pitcher was timed. This task is difficult because the start time is sometimes unclear. Anyway in 6 times the average was 2 minutes and 30 seconds with a standard deviation of 15 seconds. This is a long way from the reported 4 minutes in the USA Today article of last year. In the combined data the GD and lts methods came the closest to this time. However if 4 minutes is correct, then LS, L1, and rreg came the closest.

Overall with this data it does not appear that throwing out games like the lts method does added anything in better estimation. This method threw out 17 games. In examining the quantile plot of the multiple regression residuals with the GD method there were only 15 games that deviated from a straight line fit of the residuals by the GD method. This deviation was in the right tail (large residuals). Games 33, 64, and 112 were the furthest away and these were 16, 15, and 13 inning games, respectively. Closer and about the same distance away were games 58, 81, and 6. Game 6 was 14 innings, but game 81 has only 10 innings and game 6 looks pretty ordinary. Game 6 does have a large number for LOB, 28. The next seven games (111, 78, 62, 12, 14, 123, 45) also were clustered about the same distance from a fit to the normal quantile plot. Games 62 was 10, game 12 was 14, and game 14 was 12 innings. Game 111 was short and less than the first quartile of game times with 155 minutes with Q_1 =159. Games 111, 123, and 45 also had very few BB's 2,3, and 4 respectively. The median BB number was 7 and the minimum was 1. Of the 15 outlier games, a disproportionate number were from games on Mondays (6), Tuesdays (3), and Wednesdays (3). Remember that Mondays is the most common day off, and Wednesday games are sometimes broadcast by ESPN that may take longer time for commercials.

Now we list the Pearson and the sin(GDCC*pi/2) estimates of correlations for 161 games.

upper, r lower sdg	time	runs	hits	pitchers	LOB	BB	Ks
time	1.000	0.359 0.445	0.546 0.555	0.745 0.700	0.754 0.714	0.665 0.641	0.389 0.271
runs		1.000	0.758 0.740	0.430 0.436	0.219 0.098	0.246 0.261	-0.065 -0.049
hits			1.000	0.487 0.418	0.533 0.488	0.228 0.271	-0.042 0.029
pitchers				1.000	0.545 0.418	0.521 0.462	0.332 0.290
LOB					1.000	0.738 0.700	0.148 0.166
BB						1.000	0.140 0.068
Ks							1.000

In order to better appreciate the statistics of this work, the table below gives basic summary statistics for the variables of the multiple regression. Time is in minutes and the only continuous variable.

	min	1st Q	median	mean	3rd Q	max	SD
time	126	159	173	177.5	188	319	30.1
runs	1	5	7	7.91	10	20	3.7
hits	7	14	16	16.98	20	31	4.6
pitchers	3	6	7	7.44	9	15	2.1
LOB	6	12	14	14.57	17	31	4.4
BB	1	4	7	6.86	9	19	3.2
Ks	4	10	13	12.88	16	28	4.3

We now go over all of the regression variables from right to left starting with Ks. All of the methods give similar results and GD and LS have very low P-values. LS gives the highest coefficient. Again on BB all of the methods are similar with both GD and LS having very low P-values. LS and GD give the largest coefficients. On LOB LS and GD are furthest separated with the lowest and highest values (LS low) and about a standard deviation apart. Again on pitchers LS and GD have coefficients furthest apart, about 2.4 standard deviations by the LS estimate of SD. This was also discussed earlier. The pitcher variable is again by GD and LS very significant. LOB was significant at the 1-2 % level with GD being the most significant. Hits and Runs seem to be the most difficult variables to study because of their high correlation. Again LS and GD differ the most on hits at about 1.22 standard deviations apart, and again on runs they are furthest apart of all the methods. On Runs GD is not significant but LS is marginally significant. Finally on intercepts, again LS and GD are furthest apart. So if a different look is wanted on the data it is apparent that LS and GD methods should be used. Because of the difference in the first and second halves of the season, it seems impossible to say which method is preferred.

Look now at the correlation table among the variables. The comparison is only between GD and LS (or by my thoughts, better known as Pearson regression). There is no correlations for the other methods except if one uses the L1 correlation in my paper 1 on this same WEB site. Since I am working alone, I did not do this. Anyway it would have been better to do the regression with the L1 correlation coefficient rather than the L1 method. (they would by some tentative work give similar results). In general there is a remarkable agreement between Pearson's correlation coefficient and the sine transfer of GD, sin(pi*GD/2). Remember that GD only used the ranks of the data. So maybe this agreement for this mostly discrete data is quit remarkable. When an average game, using medians, in terms of runs, hits, pitchers, LOB,BB, Ks (7,16,7,14,7,13) is inserted in the GD regression equation, one gets a game of 173 minutes which is the median game time.

The residual SD of LS in predicting game time is 14 minutes, but for GD it is only 12 minutes. This is typical of a regression in which there are outliers, but the outliers are reliable and not erroneous results.

Finally, it is clear for 2002 for the Atlanta Braves there was no time reduction in game lengths. It is also clear that the GD technique is as good as if not better than the

other data fitting methods. Its versatility in providing SD, correlation, residual SE, etc. as well as its use in other area of statistics, makes it a potent statistical estimator.

The following are comments and quotes from a "USA Today" 27 February, Friday newspaper. It was from "On Baseball" by Hal Bodley. A few comments on the Braves in 2002 were added along with a few thoughts on the implications of the comments.

The average length of a MLB game in 2003 was 2:46 (2 hours and 46 minutes) just missing the goal of 2:45. Sixty years ago the games averaged 1:58 minutes which eventually ballooned to 2:38 in 1960. Intervals between half-innings should be 2:05(2 minutes and 5 seconds) for regular season games and 2:25 for those on national TV. When bases are empty, a pitcher has 12 seconds between pitches, down from the 20 seconds. Otherwise he will be charged with a "ball".

Last season, 2003, of 2430 MLB games 464 were 2:30 or under(compared to only 324 in 2002, 13%) . This is 19%, but for the Braves in 2002 it was only 14% under 2:30. This evidently compares well to the other teams in 2002.

For games over three hours in 2003 there were 524 down from 766 in 2002. Percentages 22% and 32%, respectively. The Braves in 2002 had 37% of their games last over 3 hours.

Post season games averaged in 2003 3:05 and World Series games 3:13.

There are 17 half inning breaks in a nine-inning game. The rule book allows one minute between innings, so and extra 1:05 in in MLB adds 17*1:05 = 18 minutes and 25 seconds to a game. Thus, maybe from sixty years ago (1:58)to 1960 (2:38) , the extra 40 minutes is partly due to the "commercial time" of slightly over 18 minutes. This is almost half of the time. The game times last year averaged 2:46 and so it appears there is only 48-18=30 "slack time" that can be improved on.

In conclusion a table is given of the Least Square and GD predictions of game times with the entries from the summary data; min, Q_1,median, mean, Q_3, max as given in the above table. For example at the median for GD

$$predictedtime = 70.775 - 0.0978(7) + 1.4260(16) + 2.5922(7) + 2.0770(14) + 2.3418(7) + 1.3104(13)$$

=173.6

Predicted games times for summary statistics with LS and GD						
	min	1st Q	median	mean	3rd Q	max
LS	102.0	152.0	174.2	177.5	202.7	307.5
GD	108.5	153.2	173.6	176.7	199.0	297.5
Actual	126	159	173	177.5	188	319

Note that GD is closer to the actual min, Q_1, median, Q_3 than LS, but LS by necessity equals the mean, and by nonrobustness is closer to the max value.