```
dm 'log;clear;output;clear';
options ps=45 ls=105 nocenter nodate nonumber nolabel
        FORMCHAR="|----|+|---+=|-/\<>*";
ODS listing;
ods graphics off;
ods html close;
ODS HTML style=minimal body='rsquare demo - high variance.html';

title1 'Monte carlo study to examine the effect of variance on RSquare values
';
data demo;
   seed = 3394903;
   do nobs = 1000, 5000;              * Generate two sample sizes;
   do stddev = 10, 50, 100, 500, 1000, 1500, 2000, 2500, 3000;
                                      * Generate an increasing series of
variances;
   do obs = 1 to nobs by 1;           * Loop to generate observations;
      xvar = INT(ranuni(seed)*100)+1; * Randomly generate X variable
(uniform);
      Y = 10000 + 25 * xvar;          *Generate regression b0=10000, b1=25;
      Yvar = Y + rannor(seed)*stddev; *Add randomly distributed variation;
   output;
   end; end; end;
run;

proc sort data=demo; by nobs stddev; run;
proc glm data=demo; by nobs stddev;
   title2 'Fitting the generated data to a regression';
   title3 'GLM was used because the rsquare is easy to output';
   model yvar = xvar;
   ods output ParameterEstimates=est FitStatistics=fits;
run;


* Parmest dataset contains estimates and p values;
* The following gets the values from two lines to one;
data parmest; set est;
   format b0p b1p 7.5;
   if parameter ne 'xvar' then do;
      b0 = (estimate);
      b0p = fuzz(probt);
    end;
  if parameter eq 'xvar' then do;
      b1 = estimate;
      b1p = fuzz(probt);
   end;
        retain b0 b0p;
   if b1 eq . then delete;
   keep nobs stddev b0 b0p b1  b1p;
run;

* combine parmest dataset with fits dataset that contains rsquare values;
data combo; merge parmest fits; by nobs stddev;
   FitTo = 'Raw Data';
   keep nobs stddev b0 b0p b1  b1p rsquare rootMSE FitTo;
run;


*Calculate means and repeat analysis;

proc sort data=demo; by nobs stddev XVAR; run;
```

```
proc means data=demo noprint; by nobs stddev XVAR;
   var yvar;
   output out=next02 n=n mean=YMean;
run;

proc sort data=next02; by nobs stddev; run;
proc glm data=next02; by nobs stddev; weight n;
   title2 'Fitting the generated data to a regression';
   title3 'GLM was used because the rsquare is easy to output';
   model YMean = XVAR;
   ods output ParameterEstimates=est2 FitStatistics=fits2;
run;

* Parmest dataset contains estimates and p values;
data parmest2; set est2;
   format b0p b1p 7.5;
   if parameter ne 'xvar' then do;
      b0 = (estimate);
      b0p = fuzz(probt);
    end;
  if parameter eq 'xvar' then do;
      b1 = estimate;
      b1p = fuzz(probt);
   end;
        retain b0 b0p;
   if b1 eq . then delete;
   keep nobs stddev b0 b0p b1  b1p;
run;

* combine parmest dataset with fits dataset that contains rsquare values;
data combo2; merge parmest2 fits2; by nobs stddev;
   FitTo = 'Means @X';
   keep nobs stddev b0 b0p b1  b1p rsquare rootMSE FitTo;
run;

data all; set combo combo2; run;
proc sort data=all; BY nobs descending FitTo stddev; run;

proc print data=all;
   title2 'Results to all fits of the generated data';
   var FitTo stddev nobs b0 b0p b1 b1p rootMSE rsquare;
run;


ods html close;
quit;
```

Monte Carlo generated data to examine the effect of variance on RSquare values
Results to all fits of the generated data for the model "Y = 10000 + 25 * xvar" where xvar values are randomly generated integers between 1 and 100. In most Monte Carlo studies numerous replicates would have been generated to observe patterns and consistency. Here I included only a single sample for demonstration purposes.

Obs	FitTo	stddev	nobs	b_0	b_0 P=	b_1	b_1 P=	RootMSE	RSquare
1	Raw Data	10	1000	10000.19	**0.00000**	24.9973	**0.00000**	10.057	**0.999809**
2	Raw Data	50	1000	10000.87	**0.00000**	25.0383	**0.00000**	50.759	**0.994977**
3	Raw Data	100	1000	10010.90	**0.00000**	24.7832	**0.00000**	99.096	**0.981575**
4	Raw Data	500	1000	10046.82	**0.00000**	24.8381	**0.00000**	478.202	**0.704850**
5	Raw Data	1000	1000	9998.72	**0.00000**	24.5740	**0.00000**	989.588	**0.341024**
6	Raw Data	1500	1000	9986.95	**0.00000**	24.8846	**0.00000**	1481.111	**0.192970**
7	Raw Data	2000	1000	9796.84	**0.00000**	27.5388	**0.00000**	2026.934	**0.130349**
8	Raw Data	2500	1000	9812.89	**0.00000**	29.3624	**0.00000**	2515.937	**0.100578**
9	Raw Data	3000	1000	10074.13	**0.00000**	26.2960	**0.00000**	2916.948	**0.064187**
10	Means @X	10	1000	10000.19	**0.00000**	24.9973	**0.00000**	9.897	**0.999982**
11	Means @X	50	1000	10000.87	**0.00000**	25.0383	**0.00000**	52.199	**0.999476**
12	Means @X	100	1000	10010.90	**0.00000**	24.7832	**0.00000**	95.688	**0.998284**
13	Means @X	500	1000	10046.82	**0.00000**	24.8381	**0.00000**	550.288	**0.948361**
14	Means @X	1000	1000	9998.72	**0.00000**	24.5740	**0.00000**	968.722	**0.846144**
15	Means @X	1500	1000	9986.95	**0.00000**	24.8846	**0.00000**	1519.915	**0.698094**
16	Means @X	2000	1000	9796.84	**0.00000**	27.5388	**0.00000**	2095.414	**0.588182**
17	Means @X	2500	1000	9812.89	**0.00000**	29.3624	**0.00000**	2372.036	**0.561625**
18	Means @X	3000	1000	10074.13	**0.00000**	26.2960	**0.00000**	2735.743	**0.442616**
19	Raw Data	10	5000	10000.24	**0.00000**	24.9970	**0.00000**	9.822	**0.999814**
20	Raw Data	50	5000	9999.77	**0.00000**	25.0096	**0.00000**	50.013	**0.995213**
21	Raw Data	100	5000	9997.99	**0.00000**	25.0384	**0.00000**	99.637	**0.981405**
22	Raw Data	500	5000	10005.02	**0.00000**	24.9416	**0.00000**	506.177	**0.673171**
23	Raw Data	1000	5000	9995.89	**0.00000**	24.6553	**0.00000**	976.386	**0.352368**
24	Raw Data	1500	5000	9995.71	**0.00000**	25.6092	**0.00000**	1508.451	**0.194871**
25	Raw Data	2000	5000	9966.97	**0.00000**	26.4152	**0.00000**	1987.082	**0.127592**
26	Raw Data	2500	5000	10043.72	**0.00000**	25.1749	**0.00000**	2523.095	**0.076964**
27	Raw Data	3000	5000	9997.24	**0.00000**	25.0297	**0.00000**	2950.834	**0.055928**
28	Means @X	10	5000	10000.24	**0.00000**	24.9970	**0.00000**	9.729	**0.999996**
29	Means @X	50	5000	9999.77	**0.00000**	25.0096	**0.00000**	53.157	**0.999893**
30	Means @X	100	5000	9997.99	**0.00000**	25.0384	**0.00000**	96.387	**0.999652**
31	Means @X	500	5000	10005.02	**0.00000**	24.9416	**0.00000**	478.122	**0.991578**
32	Means @X	1000	5000	9995.89	**0.00000**	24.6553	**0.00000**	962.175	**0.966187**
33	Means @X	1500	5000	9995.71	**0.00000**	25.6092	**0.00000**	1696.120	**0.907092**
34	Means @X	2000	5000	9966.97	**0.00000**	26.4152	**0.00000**	2326.593	**0.844741**
35	Means @X	2500	5000	10043.72	**0.00000**	25.1749	**0.00000**	2614.196	**0.798438**
36	Means @X	3000	5000	9997.24	**0.00000**	25.0297	**0.00000**	2959.723	**0.750198**