1.16 Suppose that we are interested in estimating the number of home runs based on the other numerical variables in the data set. So all the other numeric variables will be our predictors. Investigate whether sufﬁcient variability exists among the predictors to perform PCA.

Answer:

Using all other predictors to estimate model for homeruns:

**proc** **reg** data =baseball\_z outest = est\_Baseball;

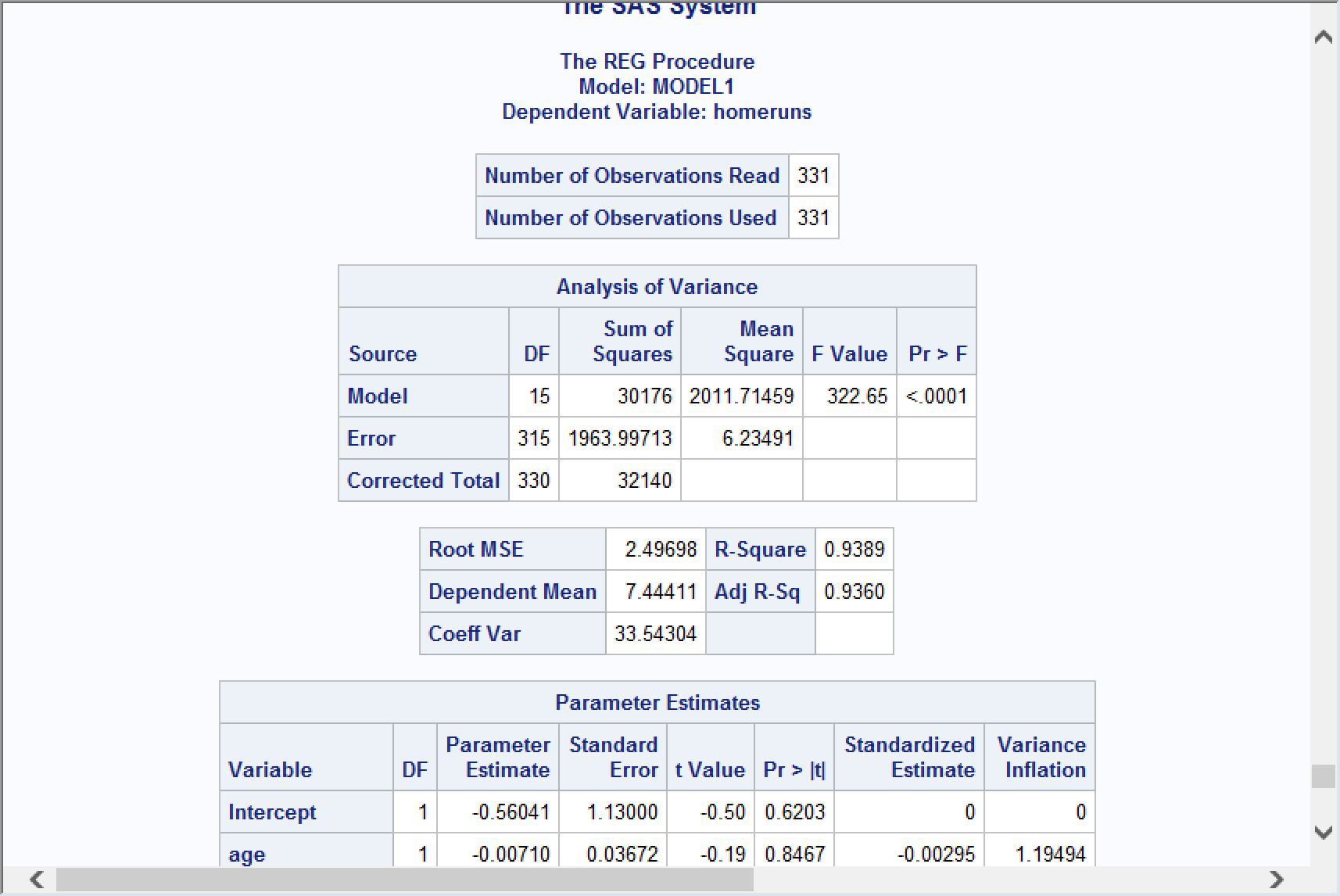
model homeruns = age games at\_bats runs hits doubles triples RBIs walks strikeouts bat\_ave on\_base\_pct slugging\_pct stolen\_bases caught\_stealing / stb vif dwProb dw;

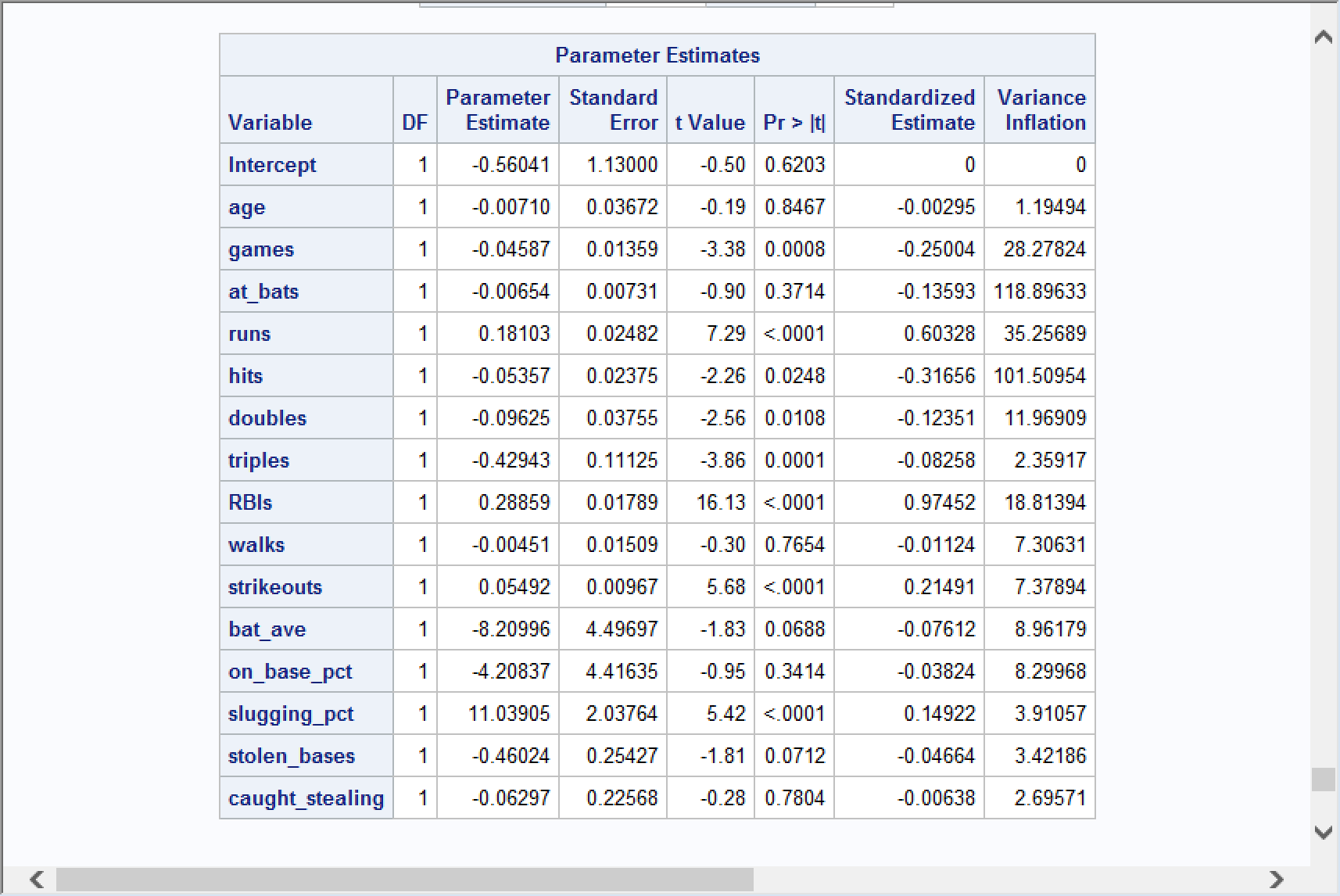
output out=reg\_BaseOUT

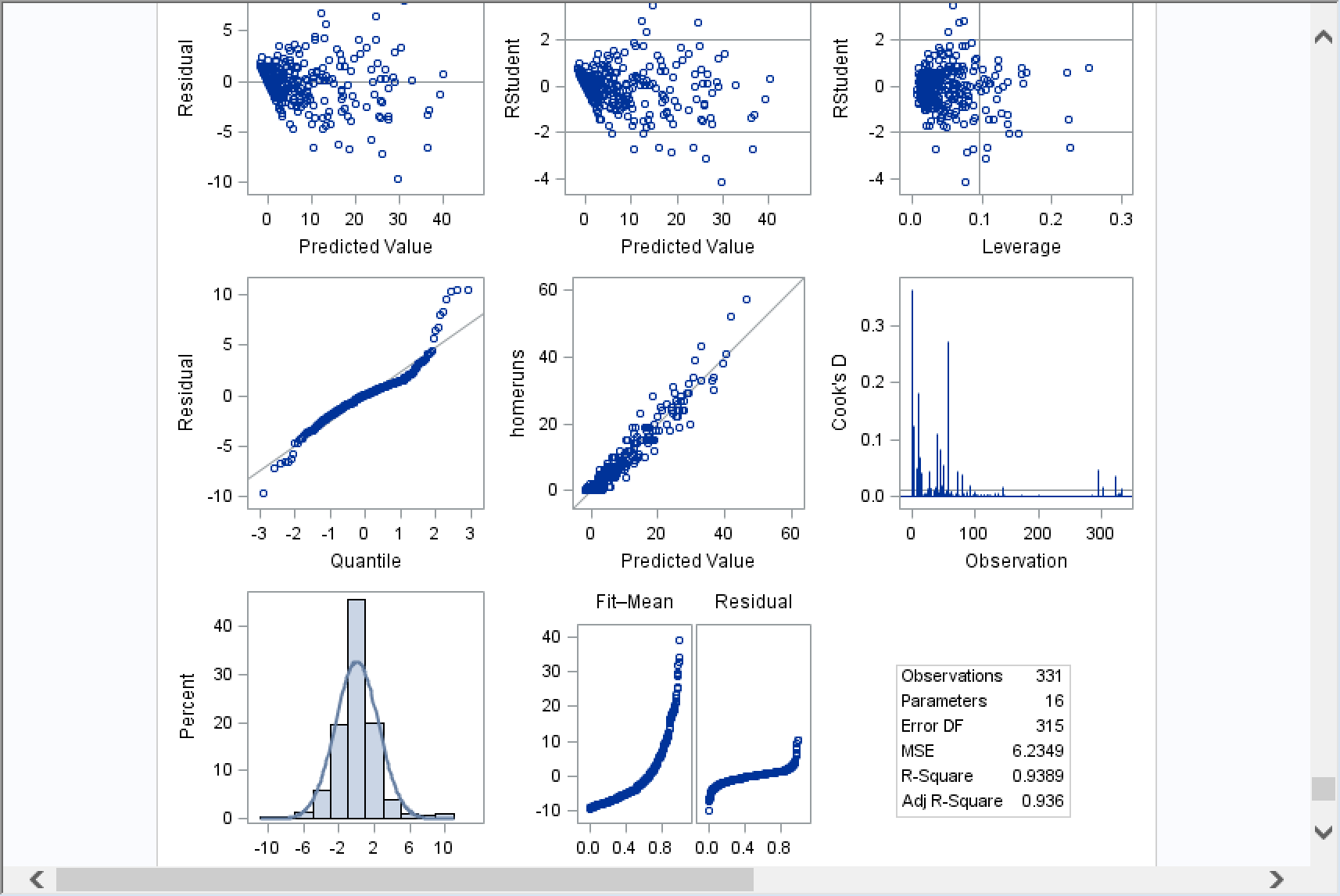
h = lev cookd = Cookd dffits = dffit ;

**quit**;

the result is as follows:

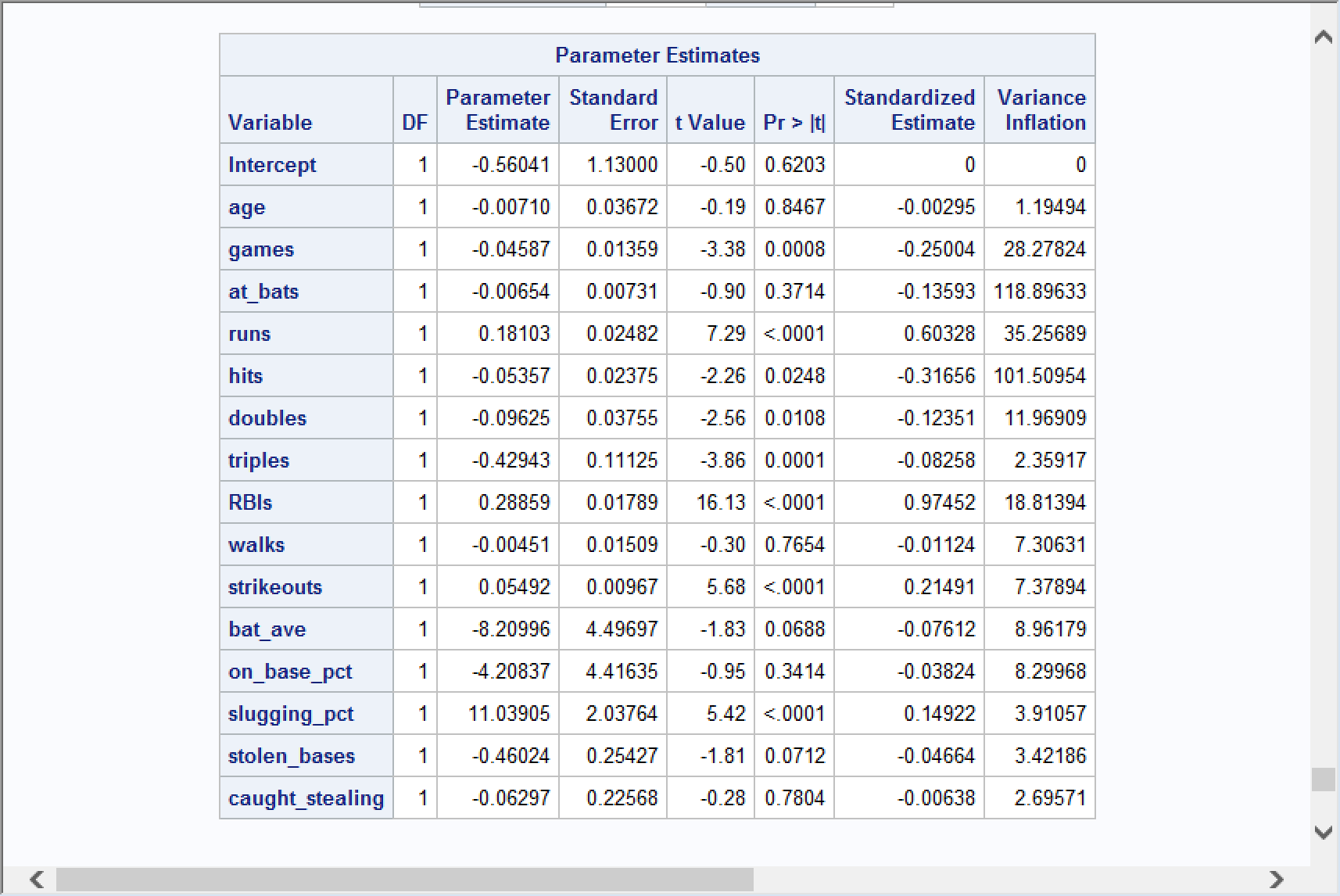






The R-square is 0.9389 . so the variable is sufﬁcient variability exists among the predictors to perform PCA.

1.17 How many components should be extracted?



The runs, RBIs, strikeouts and slugging\_pct should be extracted.

1. Answer:

**proc** **princomp** data=baseball\_z out=PCA\_class\_z;

VAR age games at\_bats runs hits doubles triples RBIs walks strikeouts bat\_ave on\_base\_pct slugging\_pct stolen\_bases caught\_stealing;

**run**;

1. Choose top 4 :

