1. We are interested in investigating whether there is a linear relationship between the number of times a player has been caught stealing and the number of stolen bases the player has.  Construct a scatter plot with caught as the response. Is there evidence of a linear relationship?

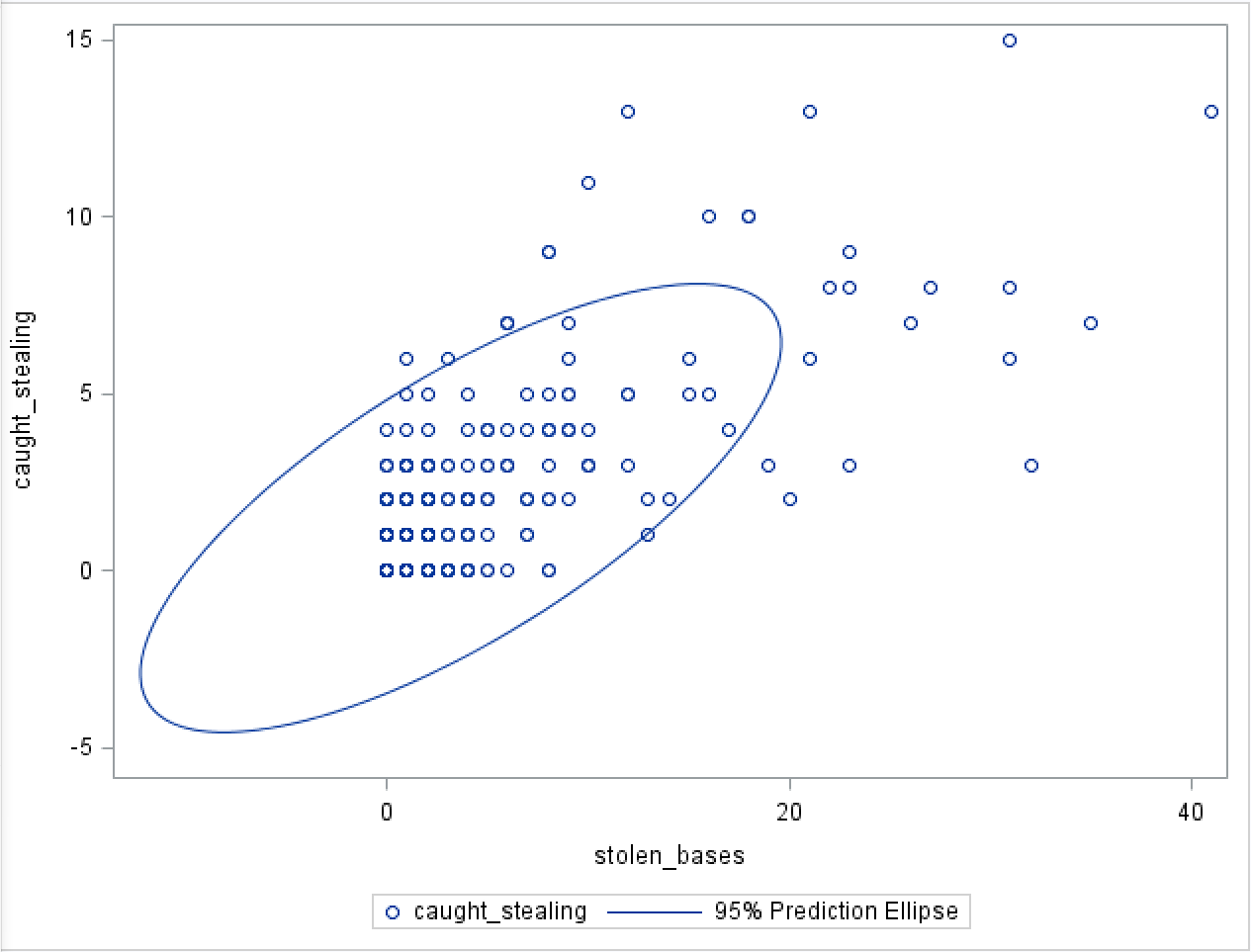
Run this code:

**proc** **sgplot** data=baseball;

scatter x=stolen\_bases y=caught\_stealing;

ellipse x=stolen\_bases y=caught\_stealing;

**run**;



Yes, this is an evidence of a linear relationship.

1. Based on the scatter plot, is a transformation to linearity called for? Why or why not?Do not perform any transformation.

No , because the transformation is called for the situation when the relationship is not linear, which is not like the graph above.

1. Perform the regression of the number of times a player has been caught stealing versus the number of stolen bases the player has.

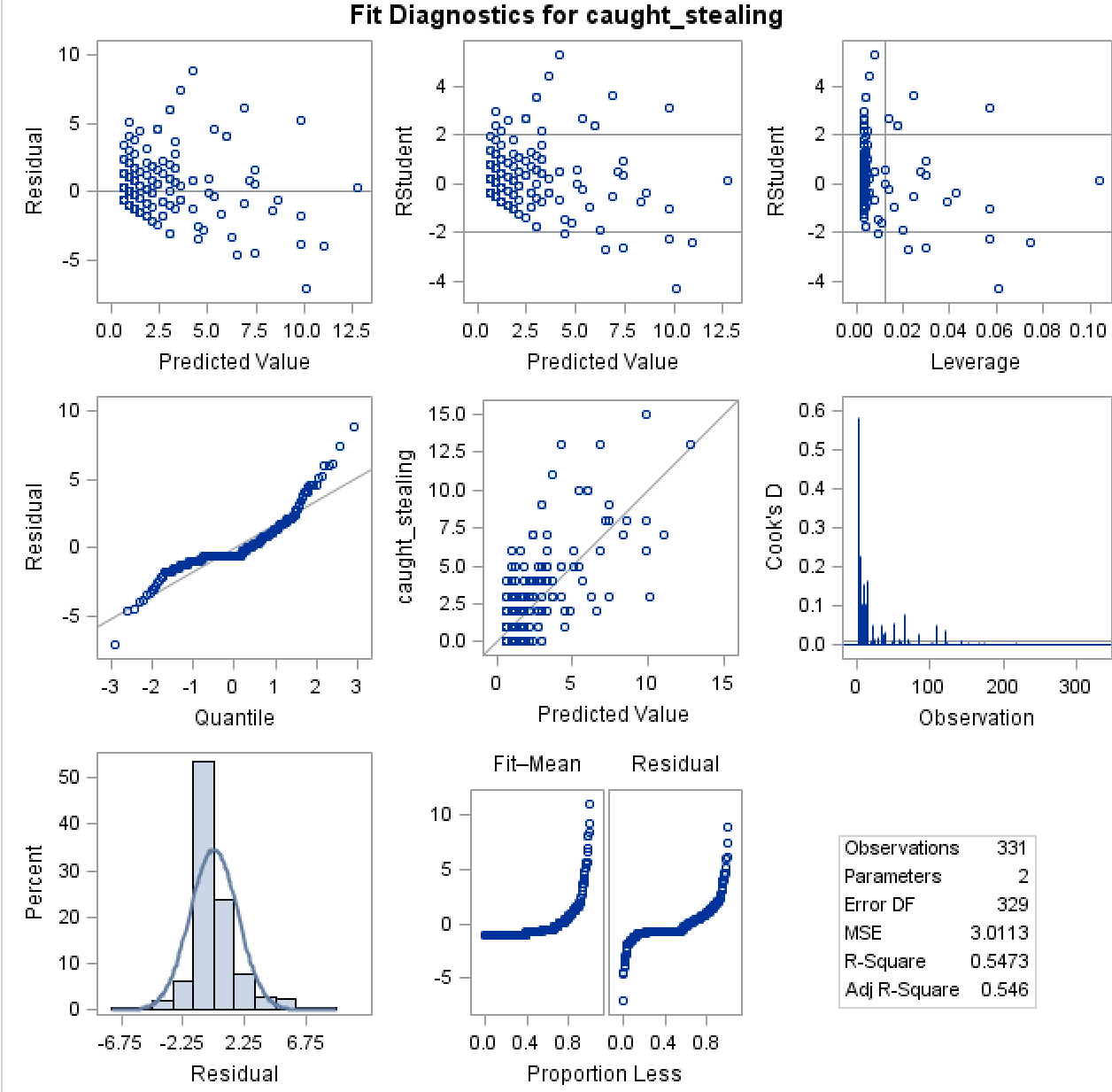
**proc** **reg** data =baseball outest = est\_Baseball;

model caught\_stealing = stolen\_bases / stb vif dwProb dw;

output out=reg\_BaseOUT

h = lev cookd = Cookd dffits = dffit ;

**quit**;



1. Find and interpret the statistic which tells you how well the data ﬁt the model.

R-SQUARE is 0.5473, that tells the degree of the data fit the model . that means not very good, but not very bad.

1. What is the typical error in predicting the number of times a player is caught stealing given his number of stolen bases?

The typical error in predicting the number of times a player is caught stealing is sqrt(MSE)= 1.73531.

1. Interpret the y-intercept. Does this make sense? Why or why not?

The Y-intercept is 0, it makes sense, because it is not easy to predict the behavior of human-being. It fits the behavior of the player.

1. Inferentially, is there a signiﬁcant relationship between the two variables? What tells you this?  
   Pr>|t| which is <0.0001, that means the relationship between 2 variables has linear

Relationship.

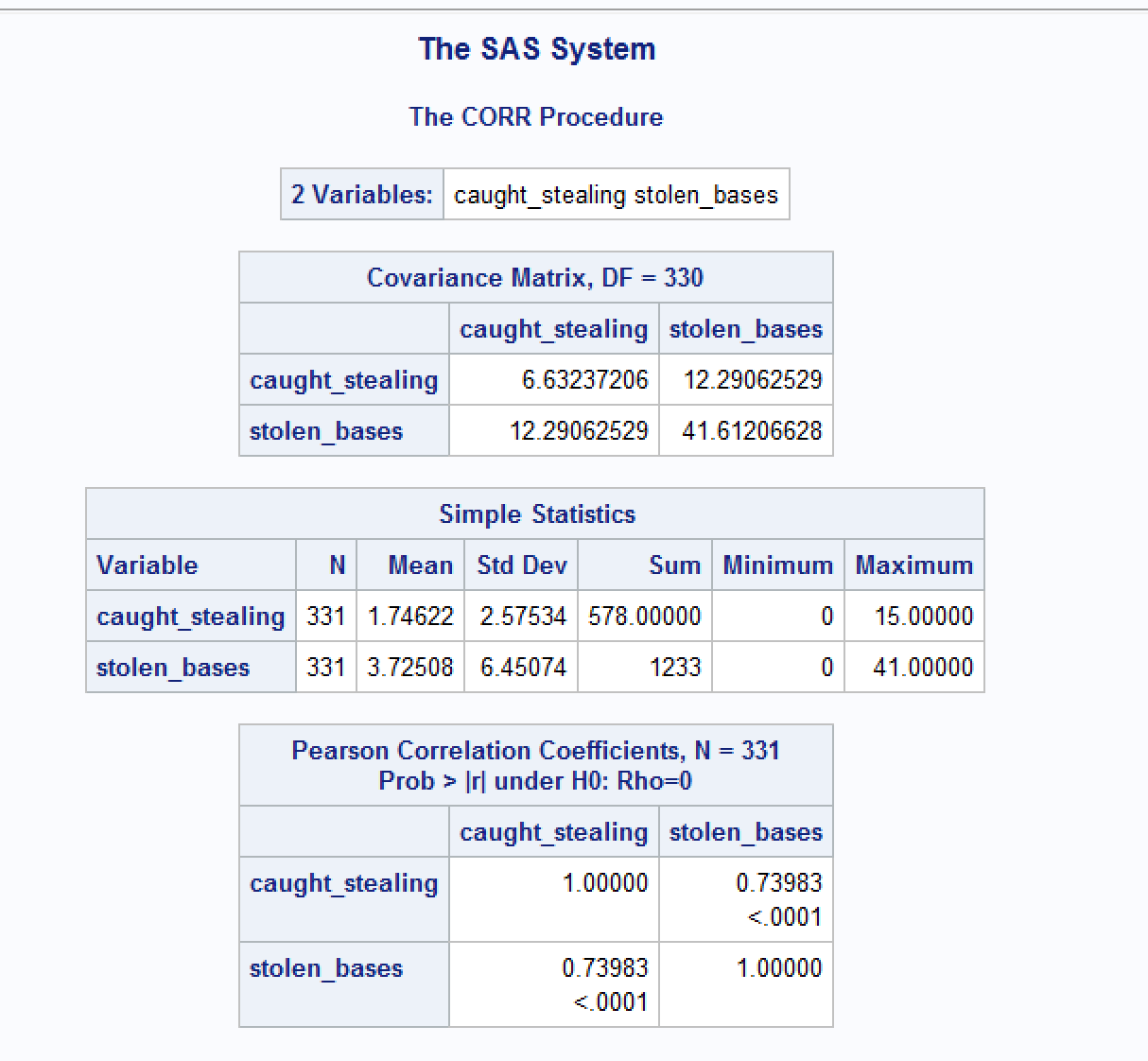
1. Calculate and interpret the correlation coefﬁcient.

**the code is :**

**proc** **corr** data=Baseball cov;

var caught\_stealing stolen\_bases;

**run**;



1. Clearly interpret the meaning of the slope coefﬁcient.

The **slope coefficient** usually refers to the **coefficient** of any independent variable, x, in a regression equation. It tells the amount of change in y that can be expected to result from a unit increase in x.

1. Suppose someone said that knowing the number of stolen bases a player has explains most of the variability in the number of times the player gets caught stealing. What would you say?

I agree, According to the linear regression model ,we can see that two variables has a strong reference , and we can compute two factors B0,B1 by regression model, therefore, we can predict caught stealing by the regression model.