Statistics 621
Robert Stine

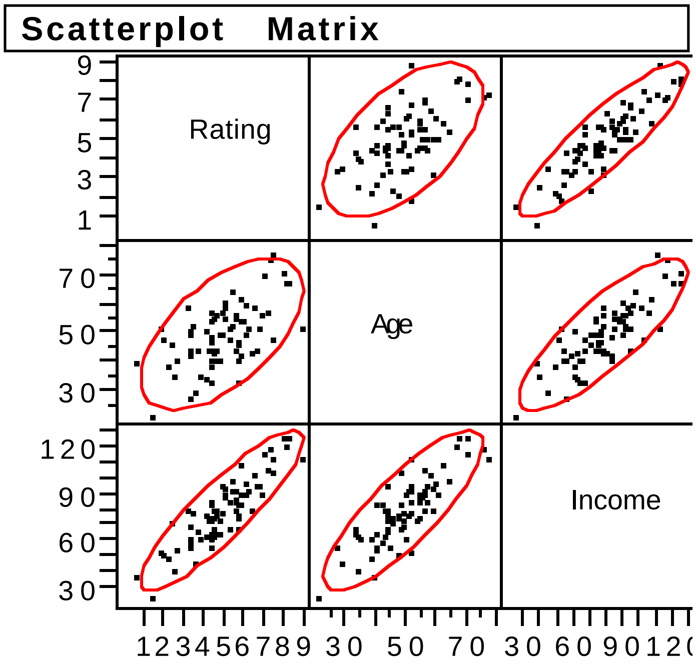
Collinearity in Regression
An Example with Changing Signs

A marketing project has identified a list of affluent customers for its new PDA. Should it focus on the younger or older members of this list?

To answer this question, the marketing firm performed a second study to measure interest in its new type of personal digital assistant (PDA). The firm obtained a sample of 75 consumers and showed each of these consumers individually the new device. (Individual viewing makes the assumption of independence more plausible.) It then asked each consumer to rate their “likelihood of purchase” on a scale of 1–10, with 1 implying little chance of purchase and 10 indicating almost certain purchase. The two predictors of how these consumers rate the PDA that we will consider here are the age (in years) and income (in thousands of dollars) of the consumers. Other factors, such as the sex of the consumer, are not considered here.

Correlations

	Rating	**Age**	**Income**
Rating	1.000	0.587	0.885
Age	0.587	1.000	0.829
Income	0.885	0.829	1.000



The marginal view of the relationship between *Age, Income* and *Rating* shows that both of these predictors are positively related to the rating of the PDA given by consumers. As shown in the following simple regressions, the both marginal relationships are significantly positive.

Regression of Rating on Age

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Intercept	2.067	0.487	4.24	<.0001
Age	0.059	0.009	6.19	<.0001

Regression of Rating on Income

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Intercept	-0.596	0.352	-1.69	0.0951
Income	0.070	0.004	16.20	<.0001

The prior scatterplot matrix also shows that *Age* and *Income* are rather highly correlated in these data (corr = 0.83). This correlation leads to collinearity, complicating the interpretation of a regression model that uses both of these factors. The marketing firm seeks to know how age affects the opinions of those who have high income, implying that we need a partial coefficient – the coefficient of *Age* in a multiple regression with *Income*. Here is the associated multiple regression.

Multiple Regression Estimates

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Intercept	-0.736	0.295	-2.50	0.0149
Age	-0.047	0.008	-5.74	<.0001
Income	0.101	0.006	15.63	<.0001

The estimated effect for *Age* is significantly negative in the multiple regression. For consumers of a given income, their rating of the product declines with age.

How are we to resolve this change in the sign of the marginal and partial coefficients? Rather than just say “collinearity”, observe that as consumers age, they also tend to have more income. Income and age grow together among these consumers. Consequently, we can interpret the marginal relationships between *Age* and *Rating* as an artifact of the underlying simultaneous change in *Income*.

Finally, to answer the initial question of how to direct the marketing, younger customers on this list are more likely to like the PDA. We might devote our attention to them as “first adopters” who will spread the good word. On the other hand, the older members of the list need some convincing, and we may want to devote special advertising to gain their approval.