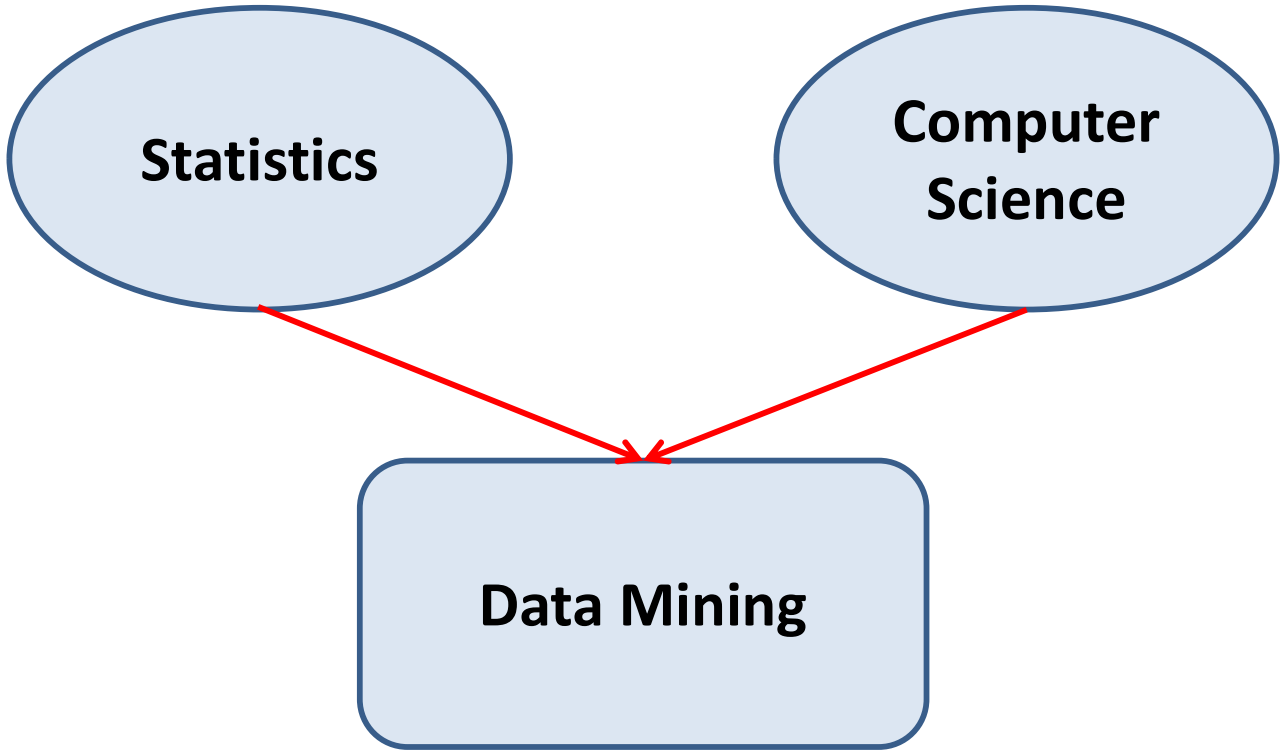
Introduction to Data Mining

What is Data Mining?

Data Mining is the methodology to extract useful information from large data sets.

(Hand et al, 2001)

Origins of Data Mining



Data Mining Methods according to the Nature of the Data

	Quantitative Response	**Categorical Response**	**No Response**
Quantitative Predictors	Linear regression Regression trees	Logistic regression k nearest neighbors Classification trees	Cluster analysis
Categorical Predictors	Linear regression Regression trees	Logistic regression Naïve Bayes Classification trees	Association rules

Core Ideas in Data Mining

- Classification
- Prediction
- Association Rules
- Predictive Analytics
- Data Reduction
- Data Exploration
- Data Visualization

Preprocessing and Cleaning the Data

Type of Variables

- Categorical (qualitative) variables
- Quantitative variables

Qualitative variable vs. Quantitative variable

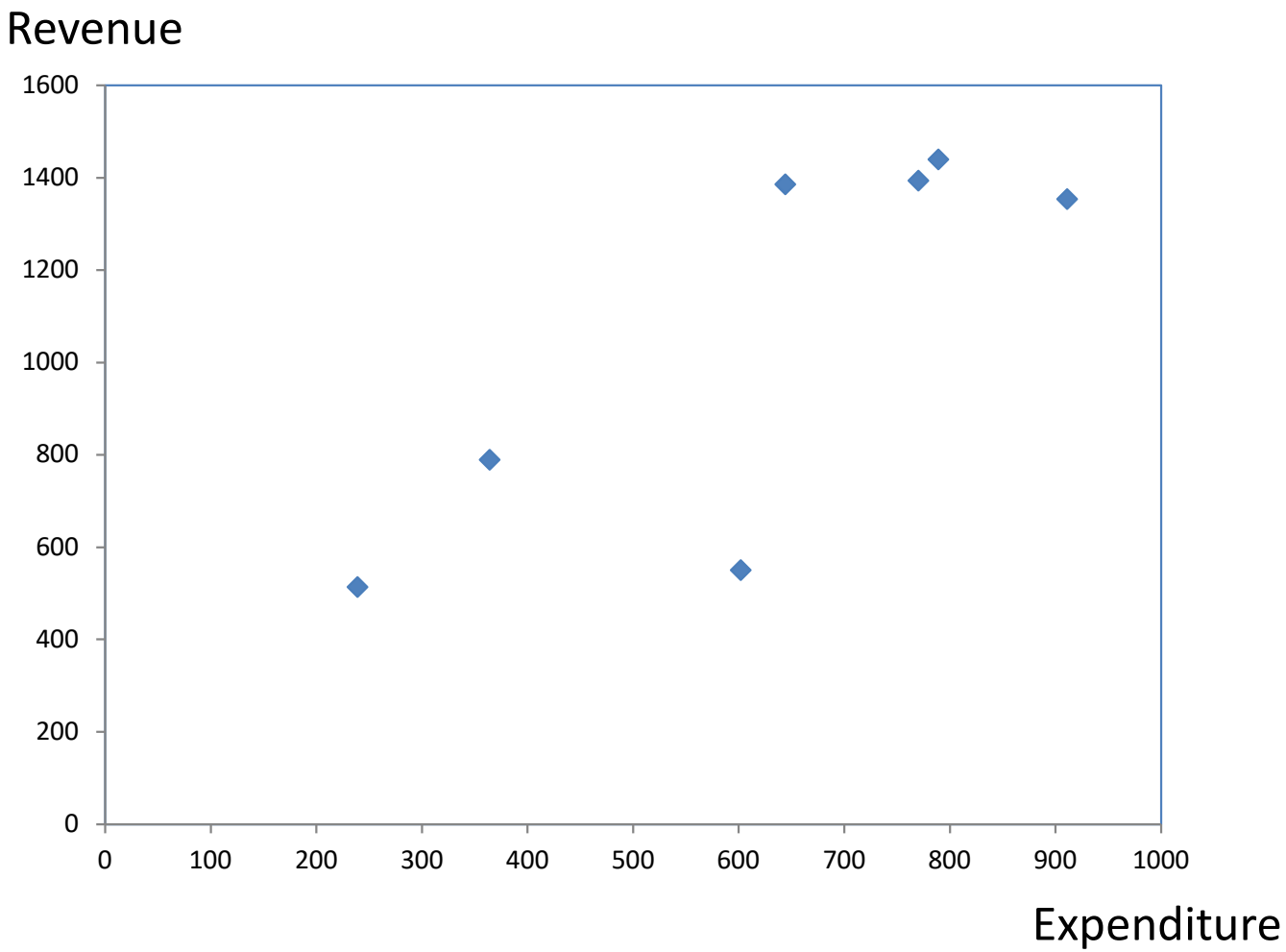
- Gender
- Age
- Temperature
- GPA
- Hours of sleep last night
- Brand of computer used
- Hometown area
- Number of TV at home
- Year of study
- SAT score
- Type of Vehicle

Variable Selection and Overfitting

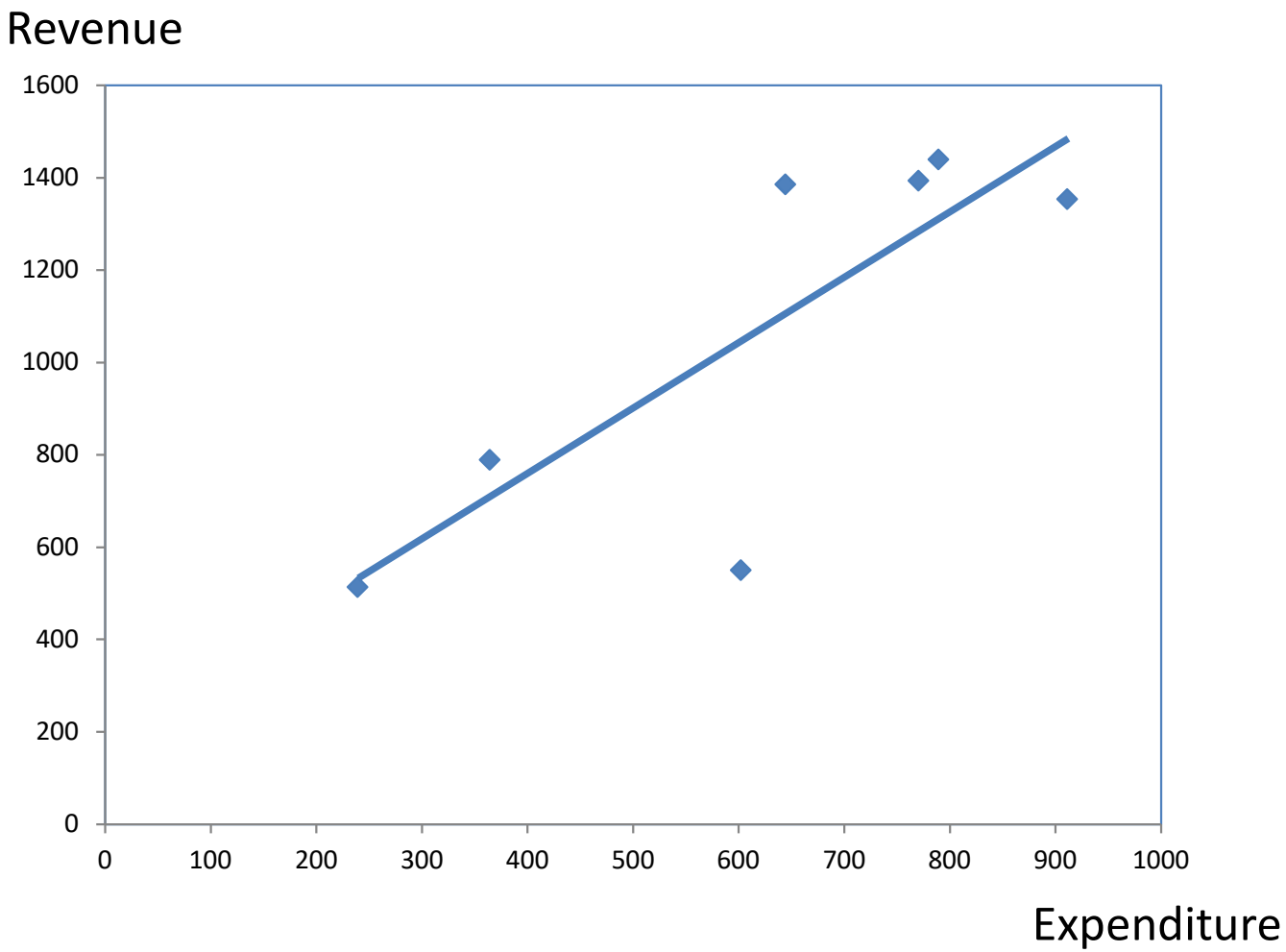
The more variables we include, the greater the risk of overfitting the data.

What is overfitting?

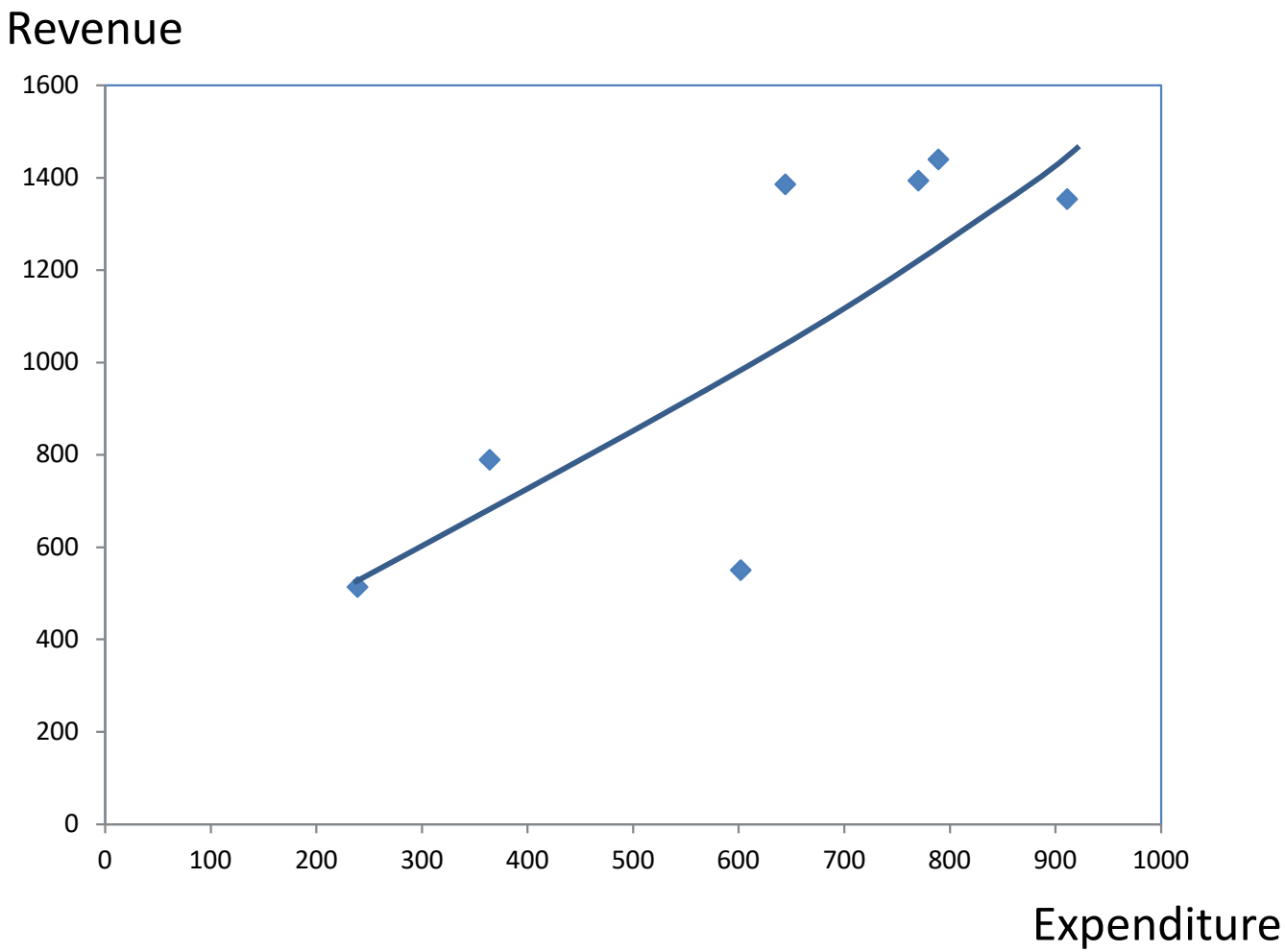
Variable Selection and Overfitting



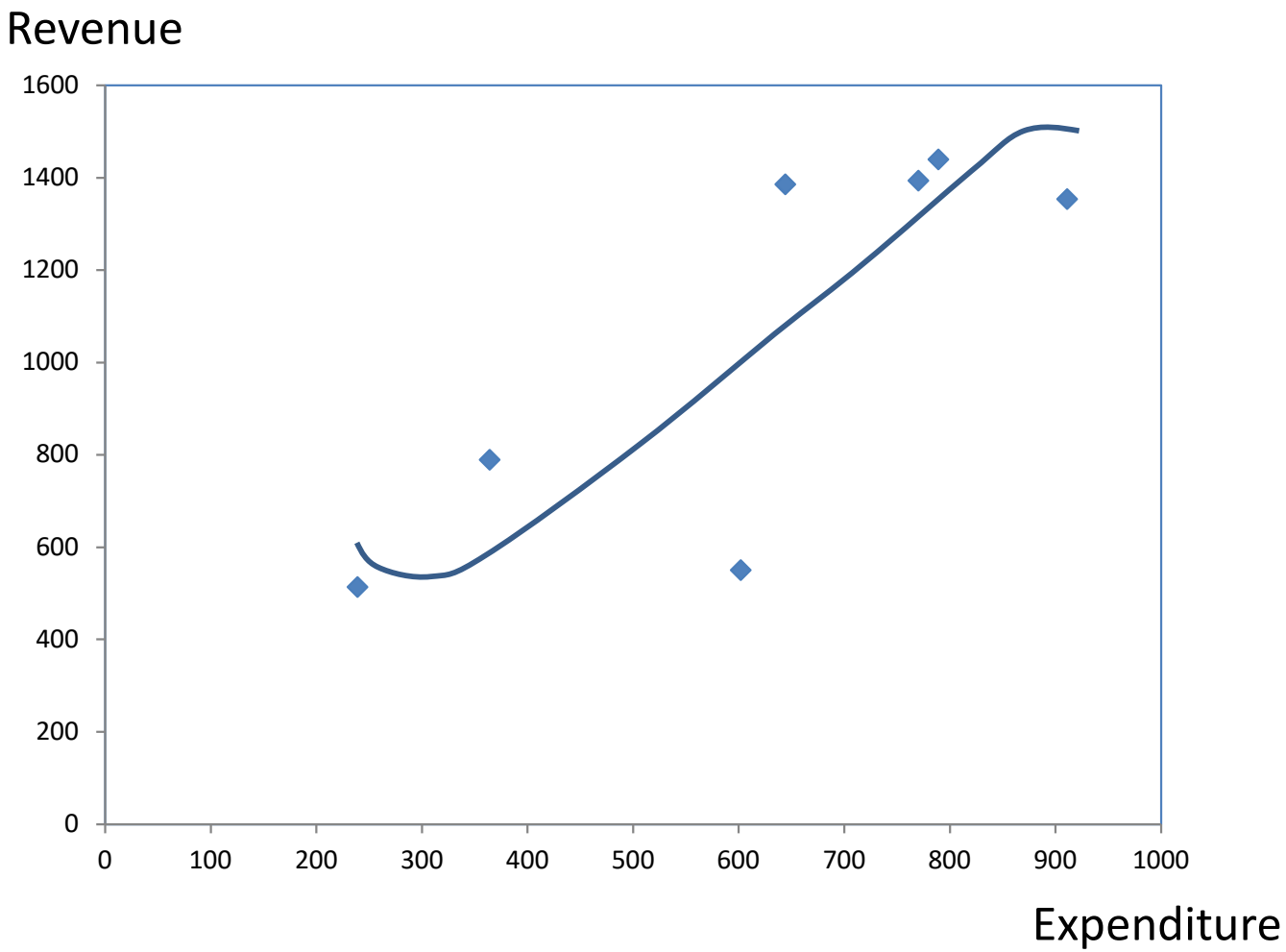
Variable Selection and Overfitting



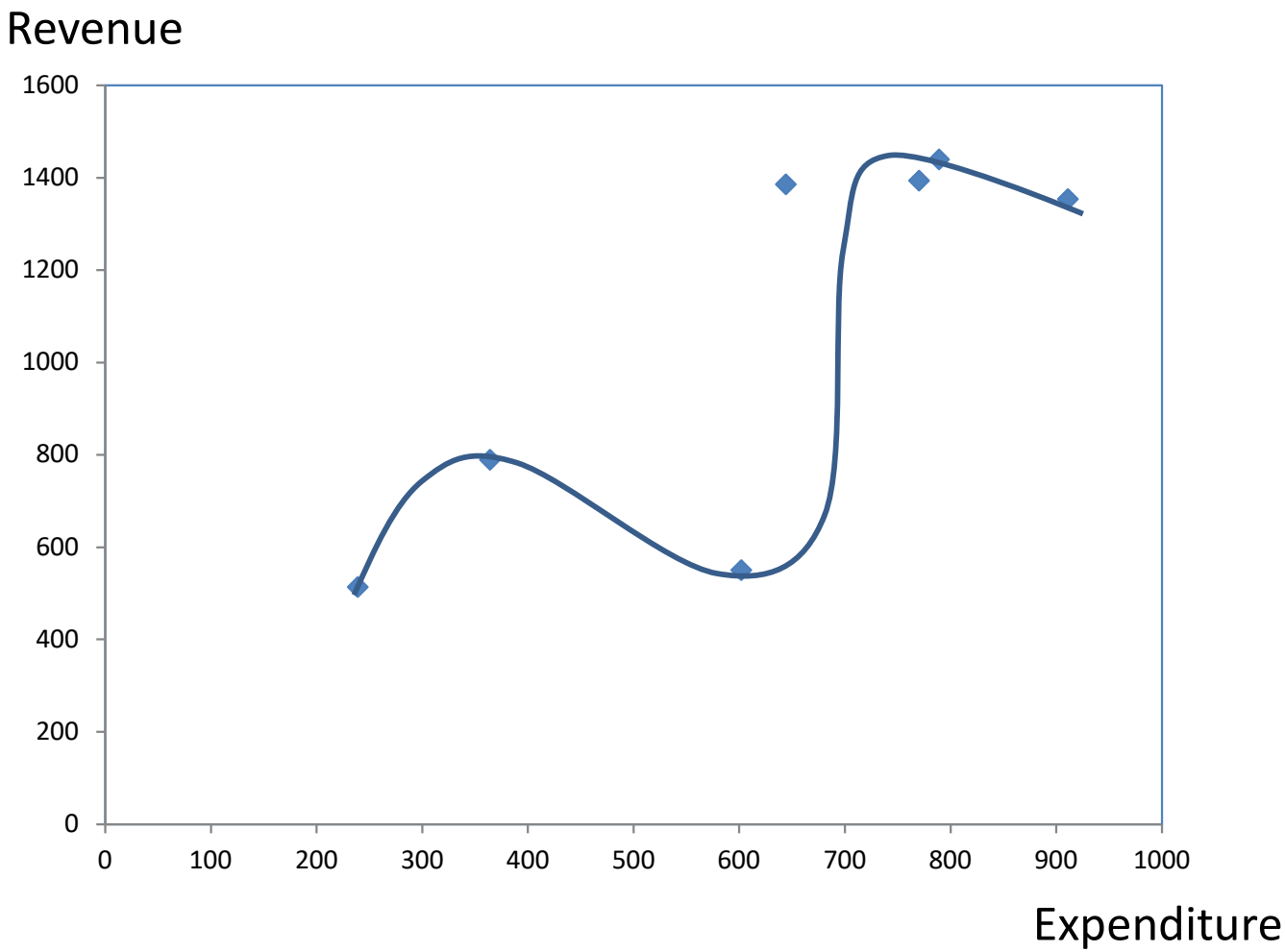
Variable Selection and Overfitting



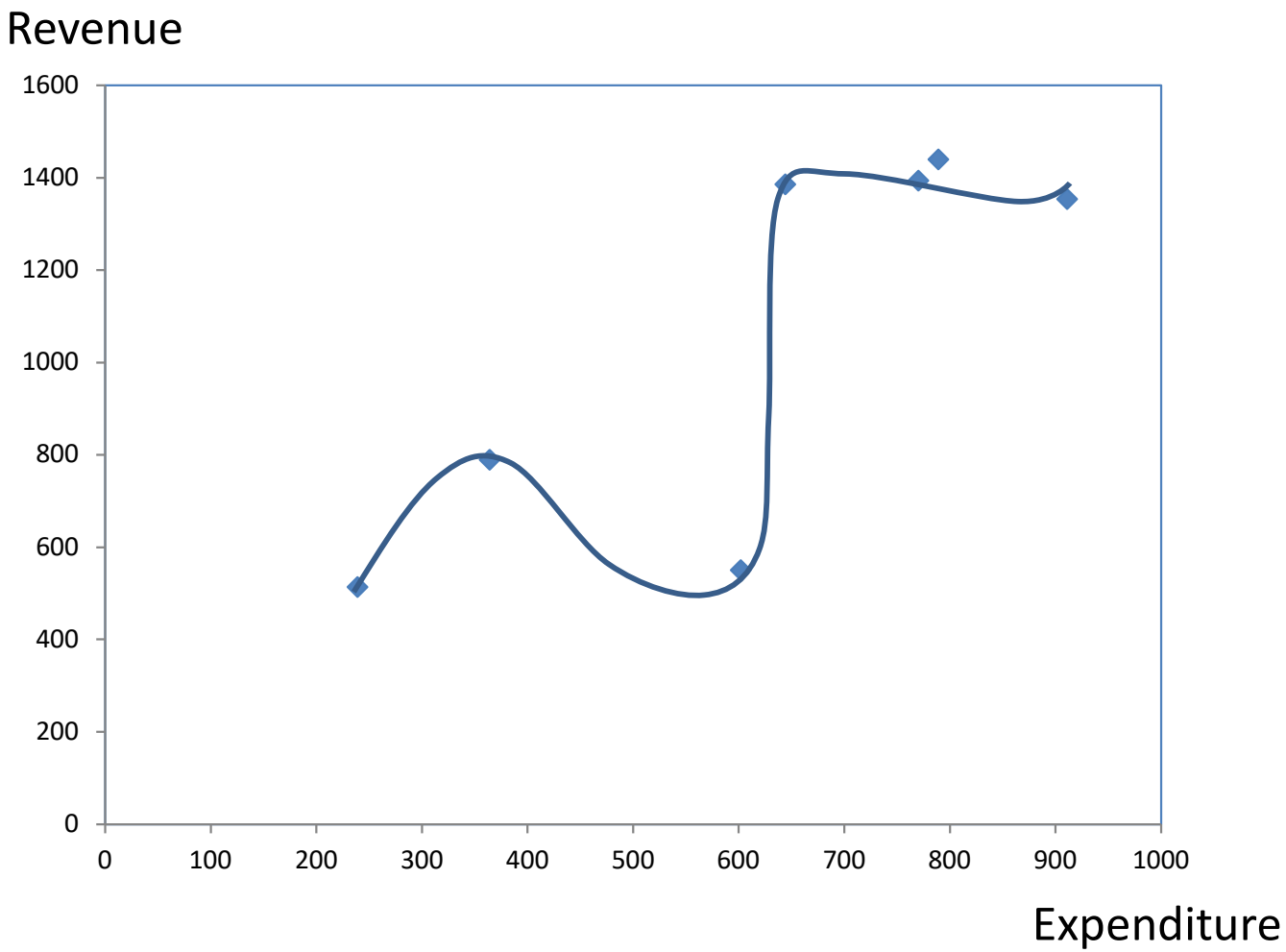
Variable Selection and Overfitting



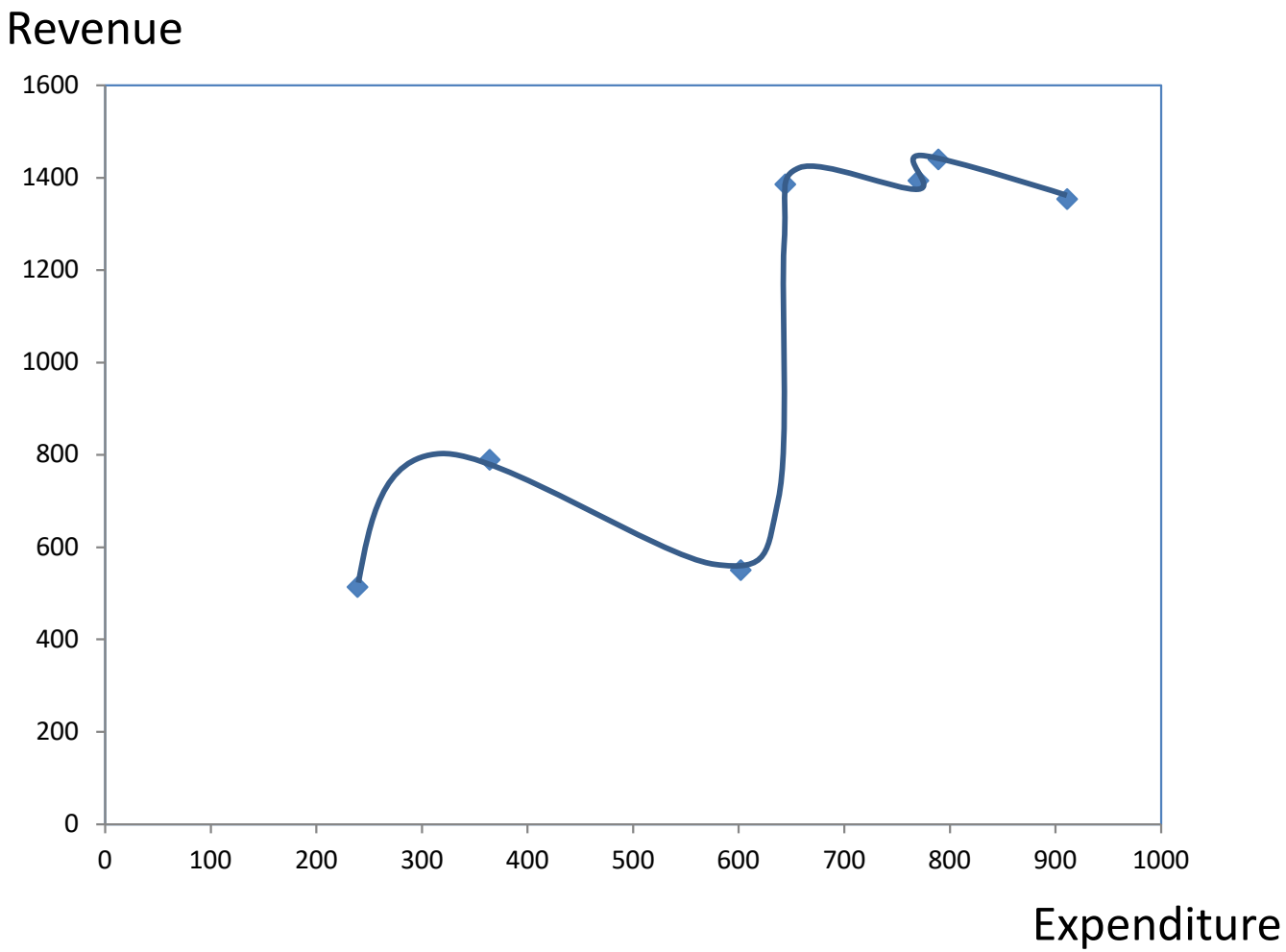
Variable Selection and Overfitting



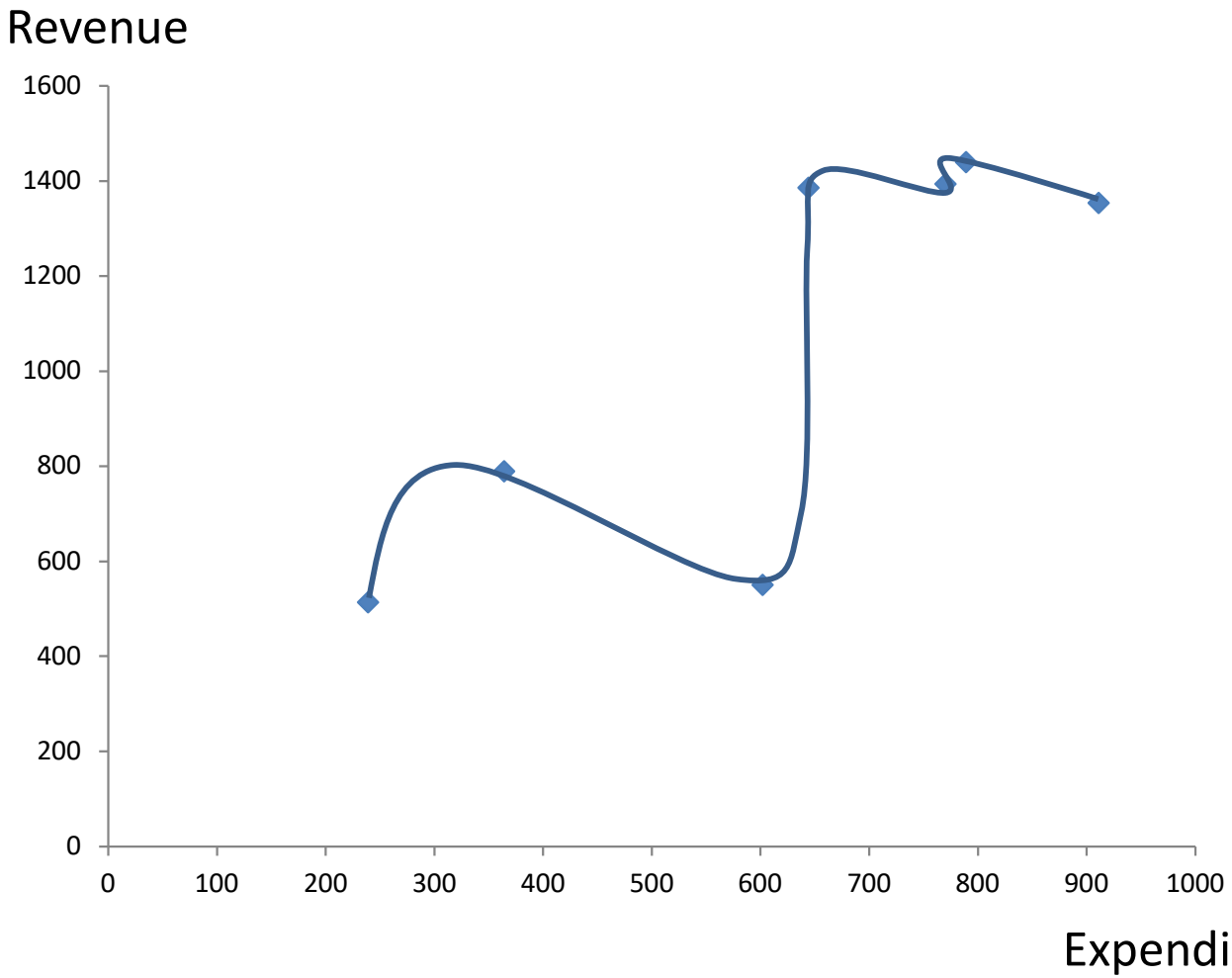
Variable Selection and Overfitting



Variable Selection and Overfitting

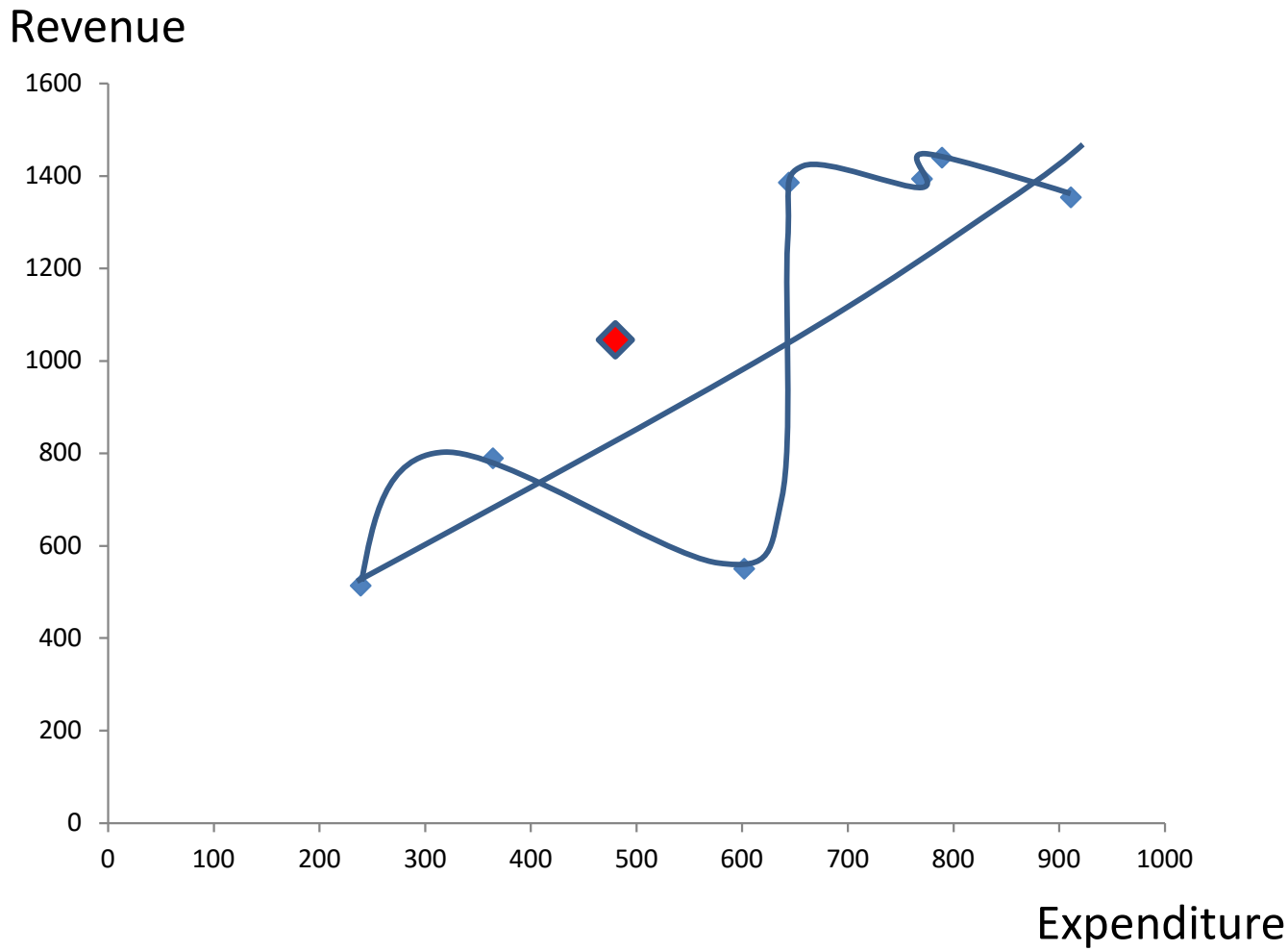


Variable Selection and Overfitting



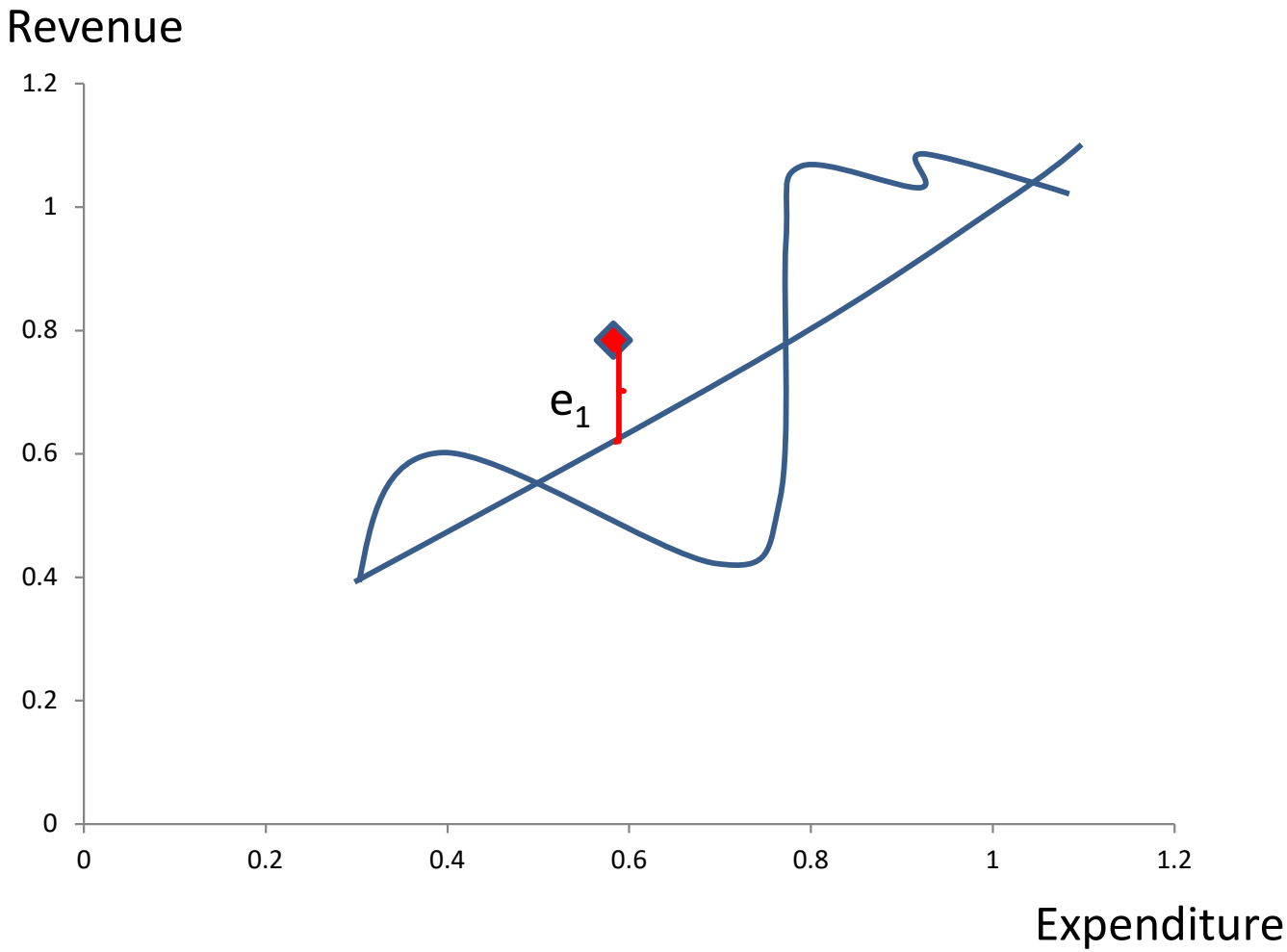
This smooth curve from complicated function connects all the points perfectly and leave no error (residuals).

Variable Selection and Overfitting



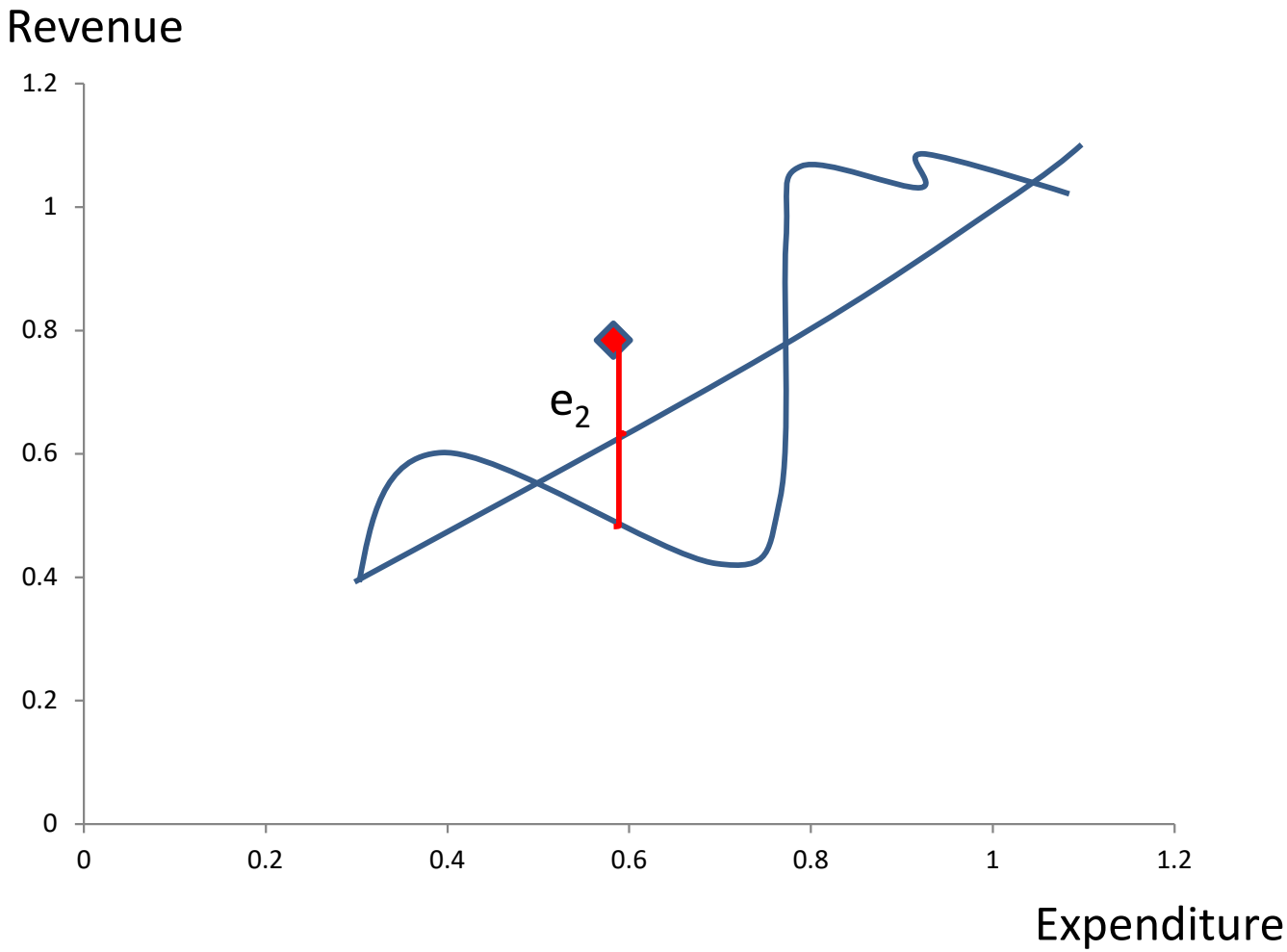
If we have the new data point as shown, which model seems to be better for prediction?

Variable Selection and Overfitting



If we have the new data point as shown, which model seems to be better for prediction?

Variable Selection and Overfitting



If we have the new data point as shown, which model seems to be better for prediction?

Variable Selection and Overfitting

This situation is called "overfitting" since we explain some variation in the data that was nothing more than chance variation.

"We mislabeled the noise in the data if it were a signal."

How many variables and how much data?

- Limit the number of predictors based on the sample size.
- Suggested rules are:

  Evan's Rule (conservative): $n/p \geq 10$ (at least 10 observations per predictor)

  Doane's Rule (relaxed): $n/p \geq 5$ (at least 5 observations per predictor)

- For classification procedures, Delmaster and Hancock (2001) suggested to have at least *6*m*p* records.

Outliers

- Values that lie far away from the bulk of the data are called *outliers*.

- Rule of thumb → anything over 3 standard deviation away from the mean is an outlier.

- How to treat outliers?

Missing values

Dealing with missing values

- Omit the missing records.
- Replace the missing value with an imputed value.
- Drop the predictor with a lot of missing values.

Question?

Given that we have 20 variables. If 1% of the values for each variable are missing independently, what is a probability that a record does not contain a missing value?

Standardizing the data

From the effect of scale, we normalize continuous measurements by

$$z = \frac{observed\ value - mean}{std\ deviation}.$$