Chapter 5:
Fitting Curves to Data

Terry Dielman

Applied Regression Analysis:
A Second Course in Business and Economic Statistics, fourth edition

5.1 Introduction

In Chapter 4 , the model was presented as:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + e_i$$

where we assumed linear relationships between *y* and the *x* variables.

In this chapter we find that this may not be true and consider curvilinear relationships between the variables.

Modeling curvature

- In general, we regress Y on some X which is not a linear function.
- Common functions are X^2 , 1/X or log(X)
- In economics, sometimes regress log(y) on log(x)

5.2 Fitting Curvilinear Relationships

- Polynomial Regression – a common correction for nonlinearity is to add powers of the explanatory variable

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + e_i$$

- In practice a second-order model is often sufficient to describe the relationship

Example 5.1: Telemarketing

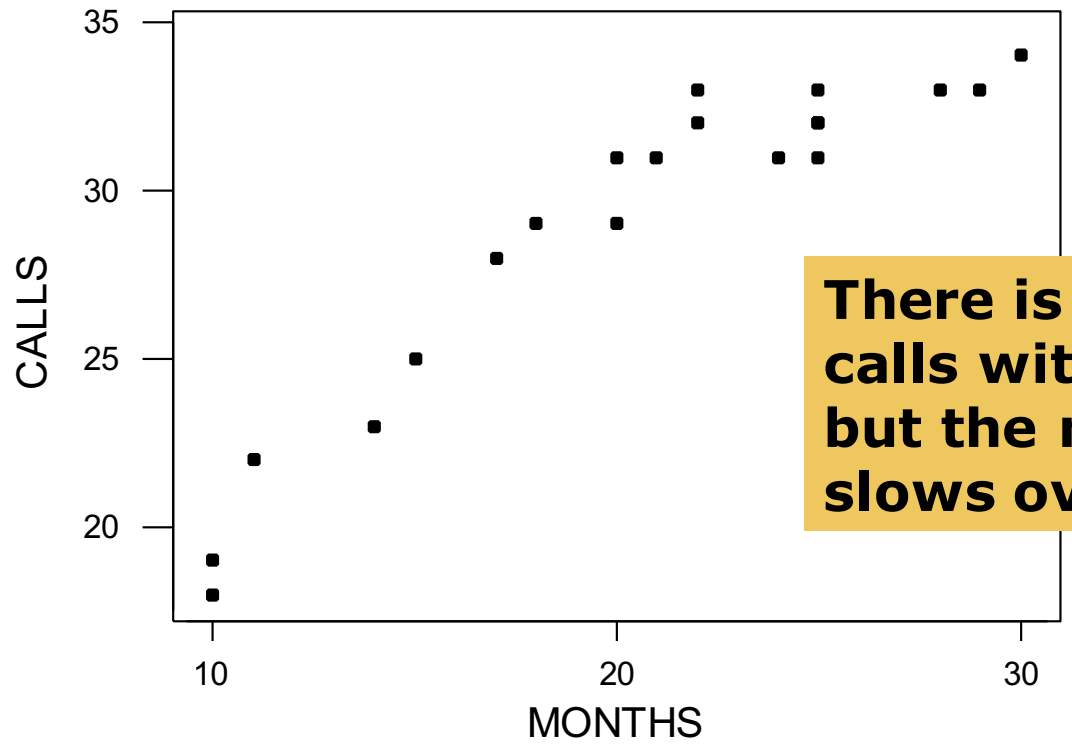
n = 20 telemarketing employees

Y = average calls per day over 20 workdays

X = Months on the job

Data set TELEMARKET5

Plot of Calls versus Months



There is an increase in calls with experience, but the rate of increase slows over time.

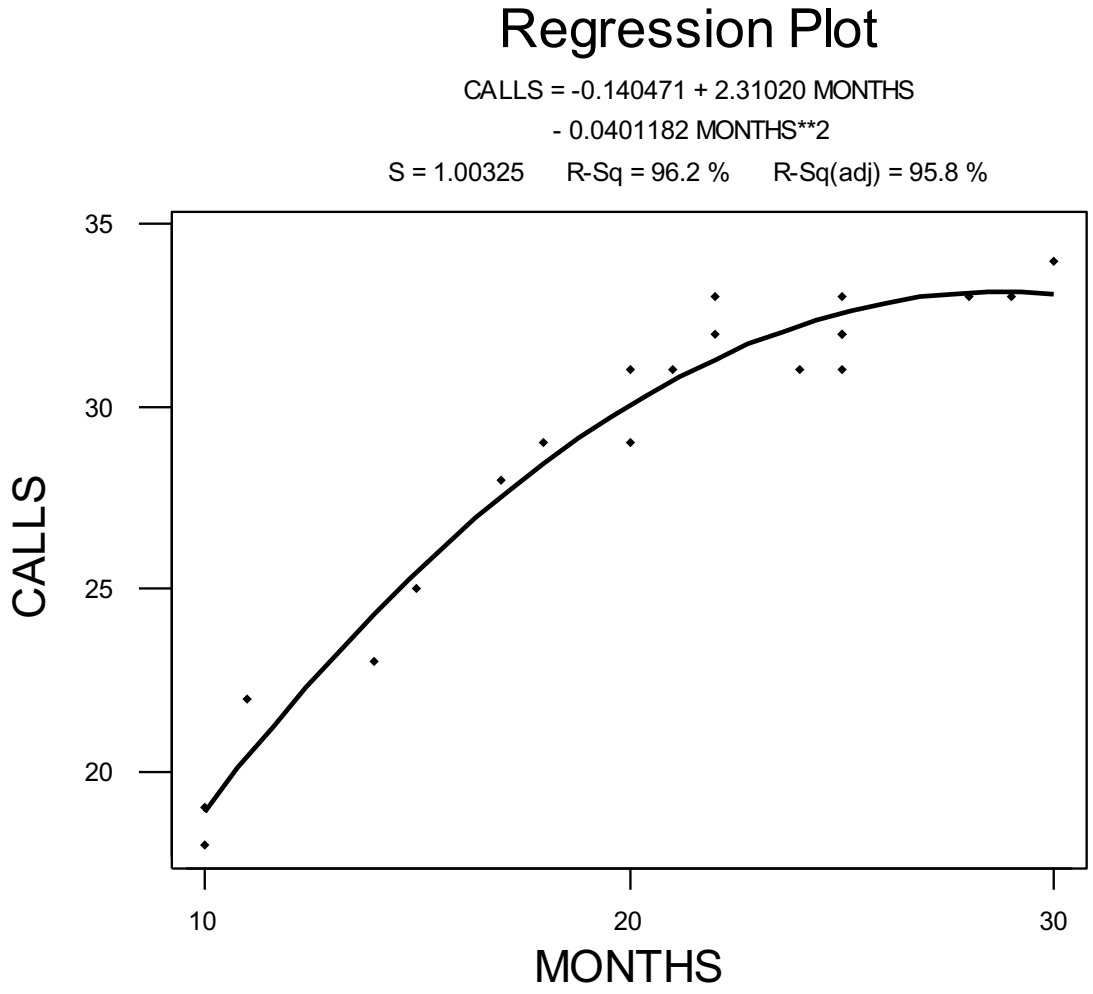
Fit of a First-Order Model

- For comparison purposes, we first fit the linear equation and obtained:

CALLS = 13.6708 + .7435 MONTHS

- This equation, which has an R^2 of 87.4%, implies that each month of experience leads to .7435 more calls per day.

Fitting a Second-Order Model



Regression Output

Regression Analysis: CALLS versus MONTHS, MonthSQ

```
The regression equation is
CALLS = - 0.14 + 2.31 MONTHS - 0.0401 MonthSQ

Predictor        Coef     SE Coef          T        P
Constant       -0.140       2.323      -0.06    0.952
MONTHS         2.3102      0.2501       9.24    0.000
MonthSQ     -0.040118    0.006333      -6.33    0.000

S = 1.003       R-Sq = 96.2%     R-Sq(adj) = 95.8%

Analysis of Variance

Source            DF          SS          MS          F        P
Regression         2      437.84      218.92     217.50    0.000
Residual Error    17       17.11        1.01
Total             19      454.95
```

Hypothesis Test on β_2

H_0: $\beta_2 = 0$ (Use the linear equation)

H_a: $\beta_2 \neq 0$ (Quadratic has improved fit)

Test as usual with $t = b_2/SE(b_2)$

Here $t = -.0402/.00633 = -6.33$ is significant with p-value = .000

Not surprising since R^2 increased 9%

Hypothesis Tests "Top Down"

- The usual practice is to keep lower-order terms when a high-order term is significant.
- In $b_0 + b_1 x + b_2 x^2$ we would retain the b_1 term even if it had an insignificant t-ratio, if the b_2 term was significant.

Higher and higher?

- To see if the polynomial has even a higher order, we fit a cubic equation.
- The table below shows the second-order model was sufficient.

Model	p for highest order term	R^2	Adj R^2	S_e
Linear	0.000	87.4%	86.7%	1.787
Quadratic	**0.000**	96.2%	**95.8%**	**1.003**
Cubic	0.509	96.3%	95.7%	1.020

Centering the X

- When polynomial regression is used, multicollinearity often results because x and x^2 are correlated.
- This can be eliminated by subtracting x-bar (the mean) from each x

$$\text{Use} \quad x - \bar{x} \qquad \text{and} \qquad (x - \bar{x})^2$$

5.2.2 Reciprocal Transformation of the *x* Variable

- Another curvilinear relationship that is in common use is:

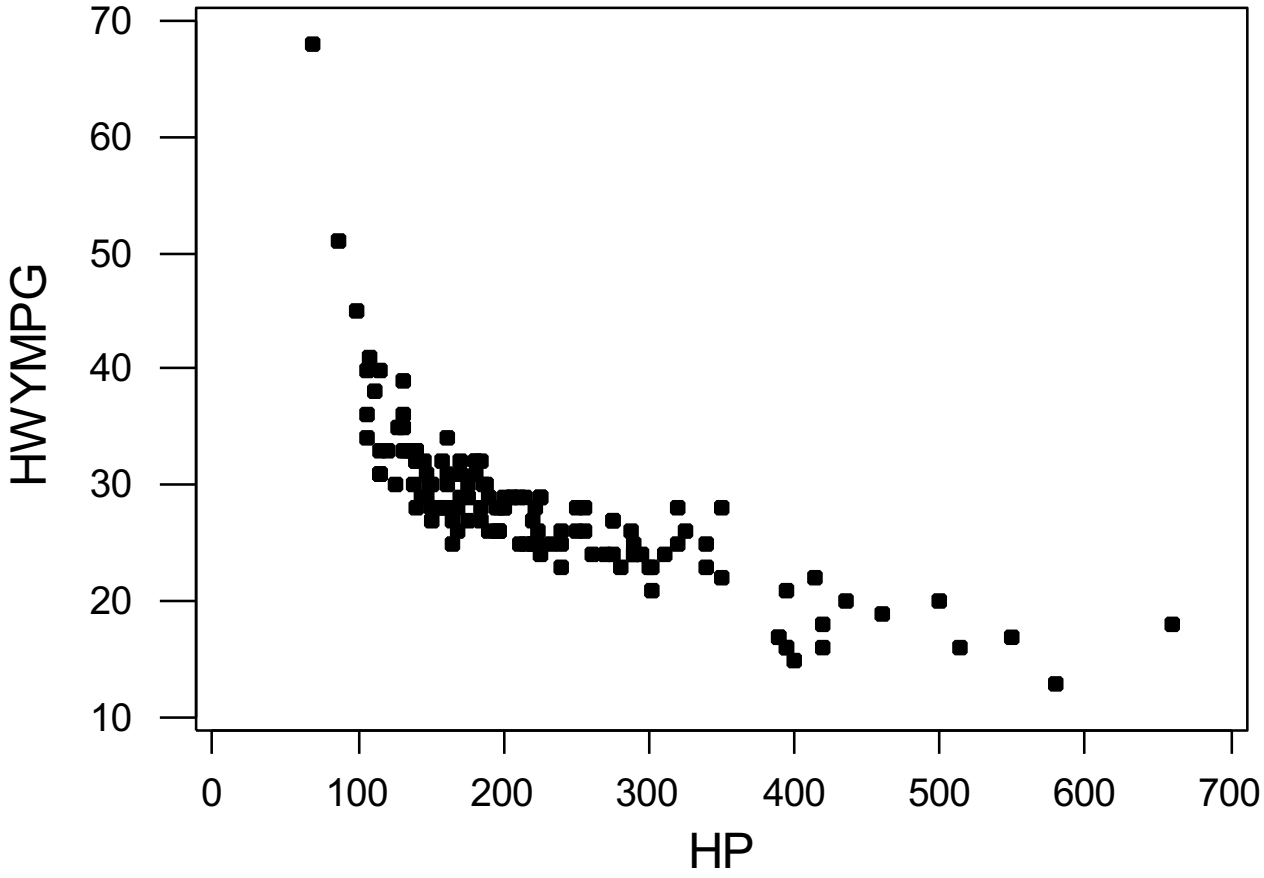
$$y_i = \beta_0 + \beta_1\left(\frac{1}{x_i}\right) + e_i$$

- Here *y* and *x* are inversely related but the relationship is not linear.

Example 5.2

- We are interested in the relationship between gas mileage and a car's horsepower.
- An the next page is a plot of the highway mpg (HWYMPG) and horsepower (HP) for 147 cars listed in the October 2002 *Road and Track*.

Highway MPG versus Horsepower



Modeling the Relationship

- A regression of HWYMPG on HP yields
  HWYMPG = 38.73 - .0477 HP with R^2 = 59.4%
- This does not fit too well because as horsepower increases, mileage decreases, but the rate of decrease is slower for more-powerful cars.
- Although other models, including a quadratic, might work, we regressed HWYMPG on 1/HP.

Regression Results

The regression equation is
HWYMPG = 13.6 + 2692 HPINV

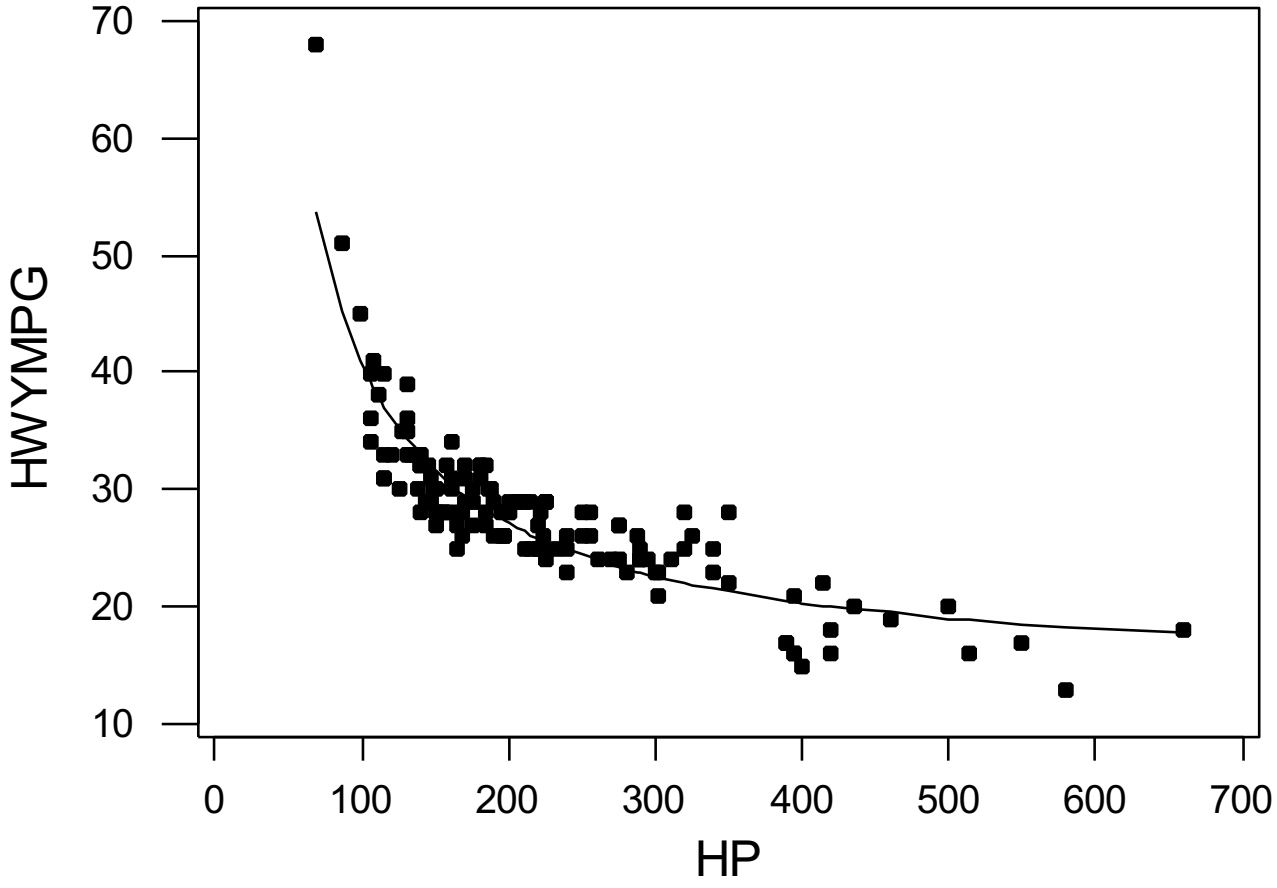
Predictor	Coef	SE Coef	T	P
Constant	13.6310	0.6493	20.99	0.000
HPINV	2962.4675	111.7526	24.09	0.000

S = 2.93107 R-Sq = 80.0% R-Sq(adj) = 79.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	4987.0	4987.0	580.48	0.000
Residual Error	145	1245.1	8.6		
Total	146	6232.7			

Data and Reciprocal Fit



5.2.3 Log Transformation of the *x* Variable

- Yet another curvilinear equation is:

$$y_i = \beta_0 + \beta_1 \ln(x_i) + e_i$$

  where ln(*x*) is the natural logarithm of *x*.

- It is assumed that the *x* values are positive because ln(0) is undefined.

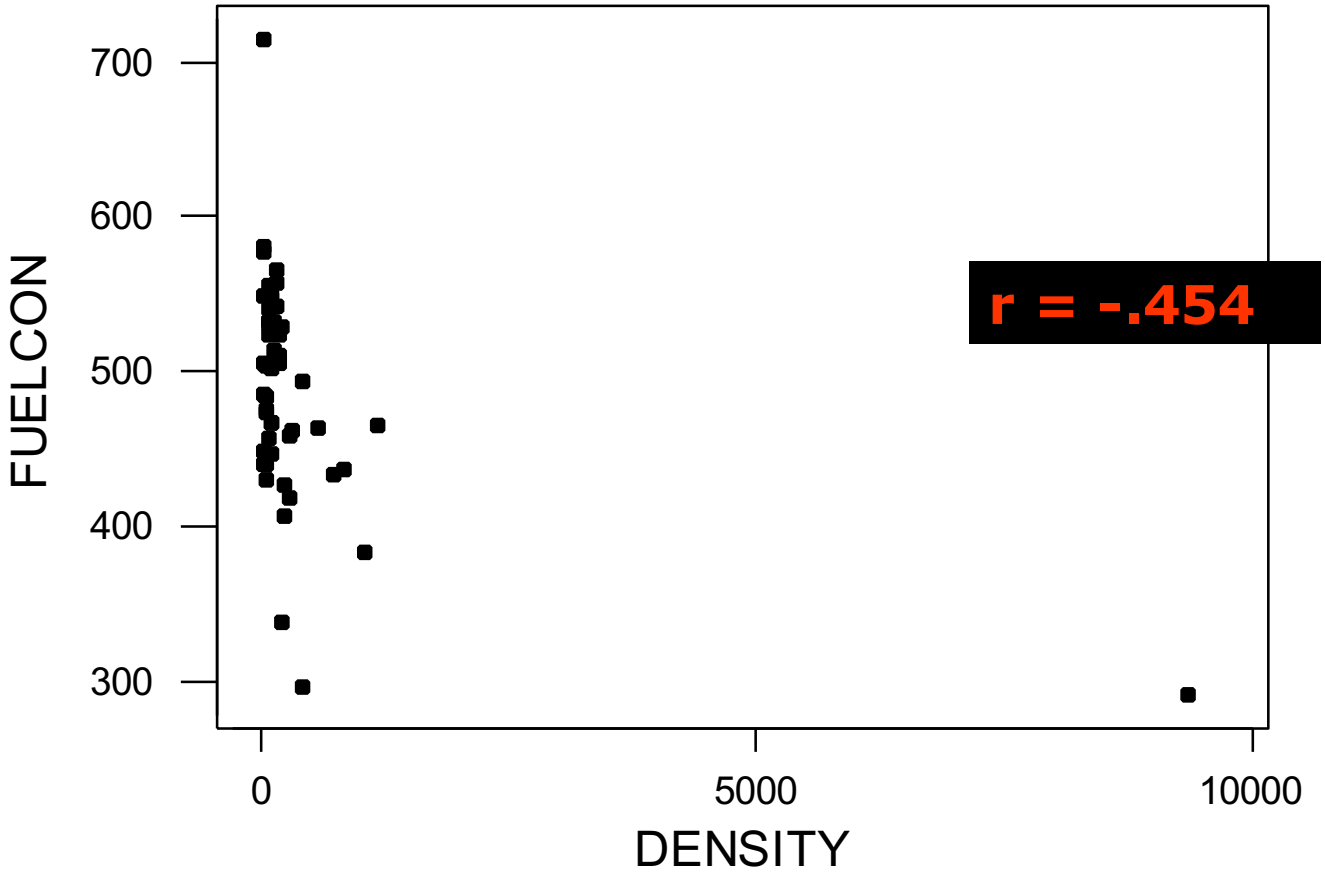
Example 5.4 Fuel Consumption

n = 51 (50 states plus Washington, D.C.)

FUELCON = fuel consumption per capita
POP = state population
AREA = area of state in square miles
POPDENS = population density

Data Set FUELCON5

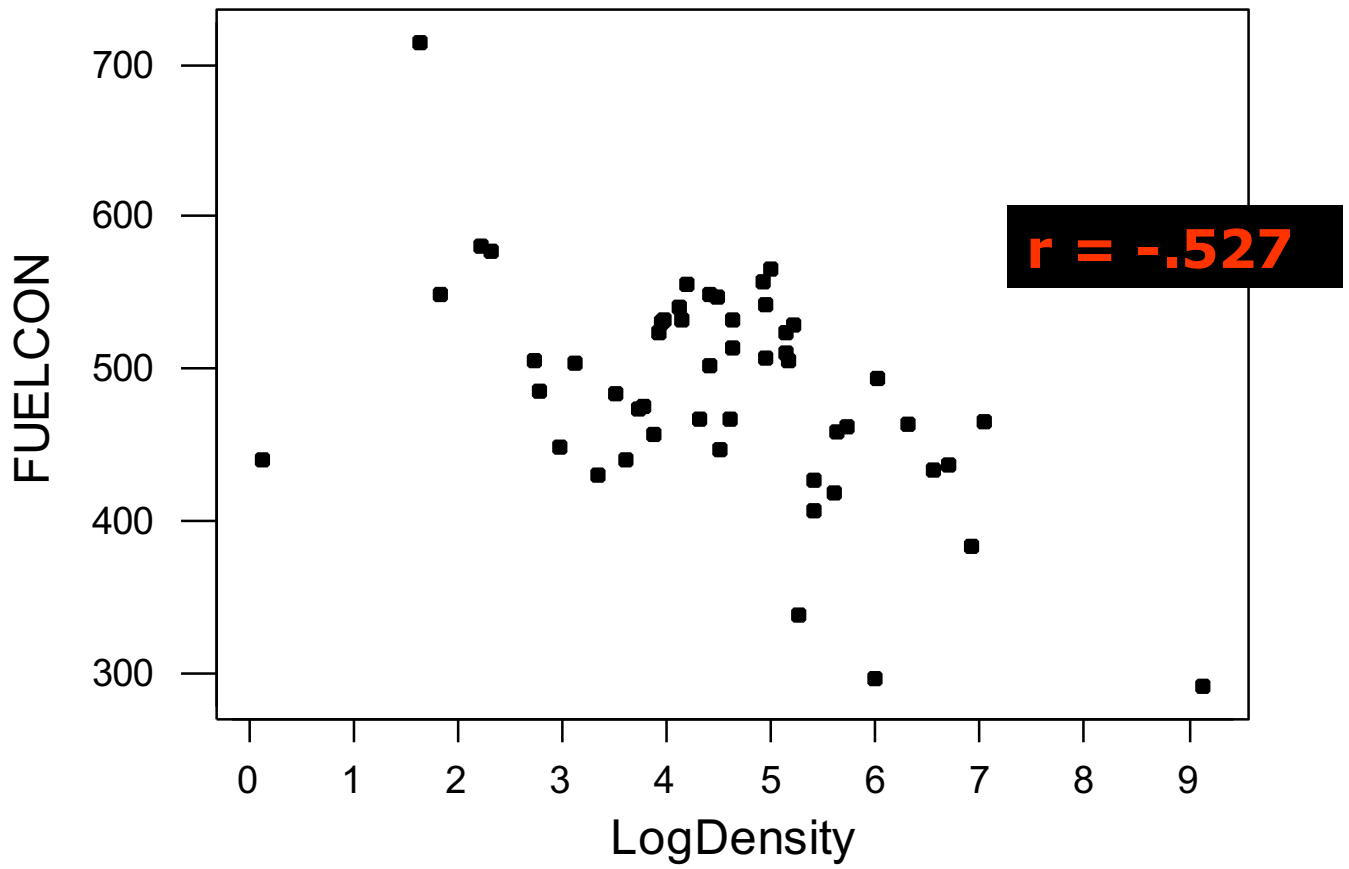
Plot of Fuelcon versus Density



Effect of the Transformation

- The graph has one point (D.C.) on the right with all others clumped to the left.
- It is hard to see what type of relationship there is until some adjustments are made.
- Here take the natural log of density to "pull" the extreme point back in.

Consumption versus Logdensity



Linear and Log Regressions

```
The regression equation is
FUELCON = 495 - 0.025 DENSITY

Predictor        Coef     SE Coef           T         P
Constant      465.628       9.481       52.28     0.000
DENSITY        -0.025       0.007       -3.56     0.001

S = 65.1675        R-Sq = 20.6%     R-Sq(adj) = 19.0%

The regression equation is
FUELCON = 597 - 24.5 LOGDENS

Predictor        Coef     SE Coef           T         P
Constant       597.19       29.96       22.15     0.000
LOGDENS        -24.53        5.65       -4.34     0.000


S = 62.1561        R-Sq = 27.8%     R-Sq(adj) = 26.3%
```

5.2.4 Log Transformations of Both the *y* and *x* Variables

- Here the natural log of *y* is the dependent variable and the natural log of *x* is the independent variable:

$$\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + e_i$$

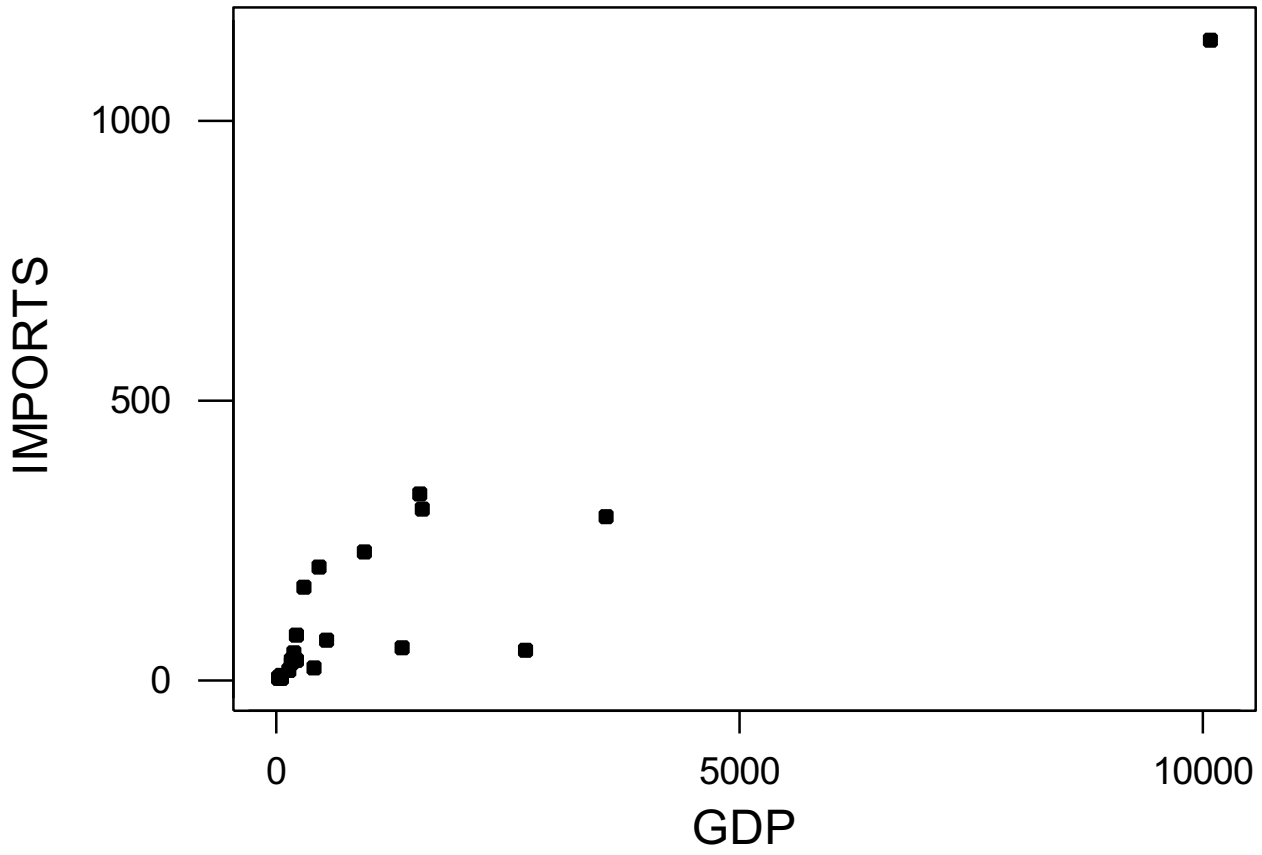
- Comparing results with other models may be difficult since we are not modeling *y* itself.
- Economists sometimes use this to estimate price elasticity (*y* is demand and *x* price; b_1 is estimated elasticity).

Example 5.4 Imports and GDP

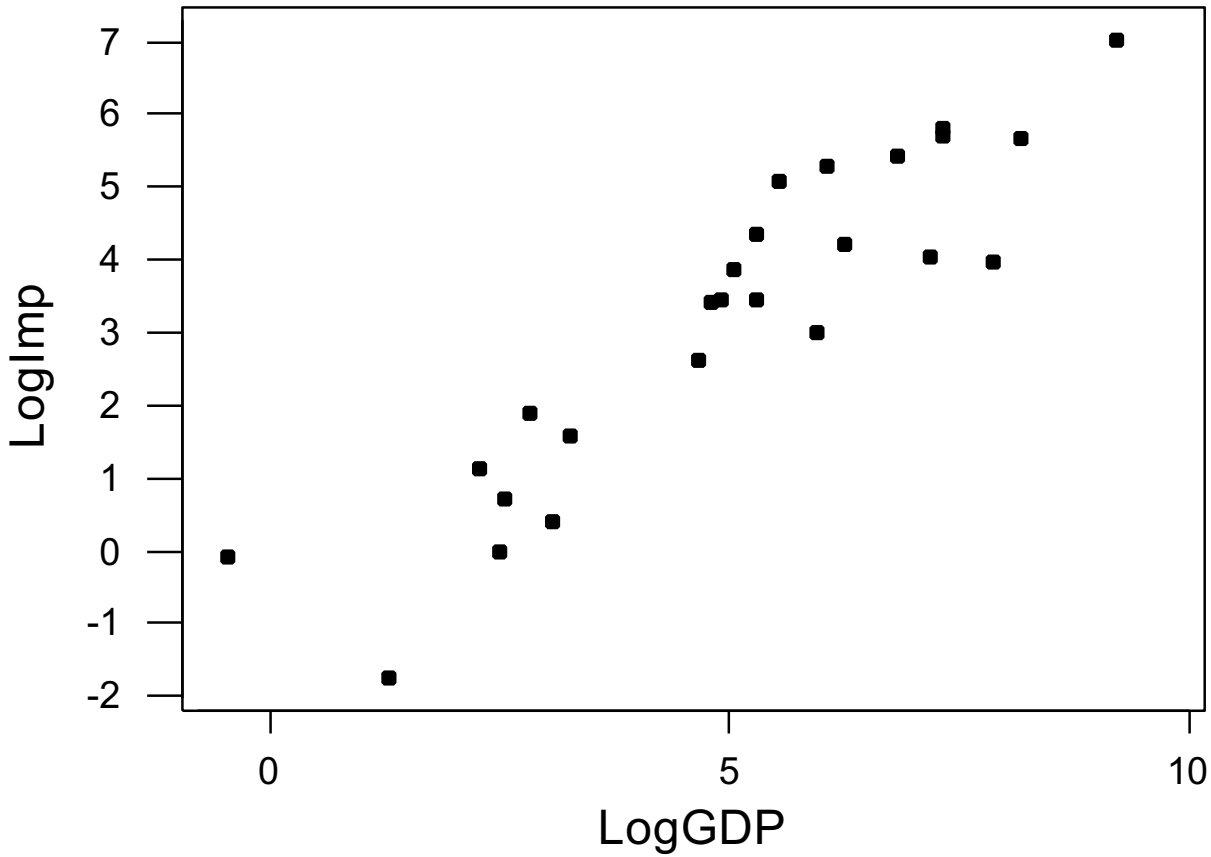
The gross domestic product (GDP) and dollar amount of total imports (IMPORTS) for 25 countries was obtained from the World Fact Book.

For both variables, low values clump together and higher values spread out, suggesting log transformations for both.

Scatterplot of Imports vs GDP



Scatterplot of LogImp vs LogGDP



Two Regression Models

Regression Analysis: IMPORTS versus GDP

Predictor	Coef	SE Coef	T	P
Constant	22.32	19.24	1.16	0.258
GDP	0.105671	0.008452	12.50	0.000

S = 87.00 R-Sq = 87.2% R-Sq(adj) = 86.6%

Not directly comparable

Regression Analysis: LogImp versus LogGDP

Predictor	Coef	SE Coef	T	P
Constant	-1.1275	0.4346	-2.59	0.016
LogGDP	0.86703	0.07877	11.01	0.000

S = 0.9142 R-Sq = 84.0% R-Sq(adj) = 83.4%

The R^2 Compare Different Things

- The 87.2 % R^2 for the no-log model is the percentage of variation in *Imports* explained.
- The 84.0% for the second model is the percentage of variation in *ln(Imports)* explained.
- If you converted the fitted values of the second model back to Imports you might find the log model better.

What Transformation to Use

- It is probably best to try several.
- A quadratic is most flexible because it uses two parameters to fit the relationship between to fit the relationship between y and x.
- Some further analysis is in Chapter 6 where tests for nonlinearity are discussed.

5.2.5 Fitting Curved Trends

If the data is collected over time, we may want to consider variations on the linear trend model of Chapter 3.

$$\text{Quadratic trend:} \quad y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + e_t$$

Another is the S-Curve trend:

$$y_t = \exp\left(\beta_0 + \beta_1 \left(\frac{1}{t} \right) + e_t \right)$$

S Curve Model

Many products have a demand curve like this.

1. Initial demand increases slowly
2. As product matures, demand picks up and steadily grows.
3. At some saturation point demand levels off.

Exponential Growth Model

Another alternative is an exponential trend:

$$y_t = \exp(\beta_0 + \beta_1 t + e_t)$$

This can be fit by least squares if you model ln(y).