

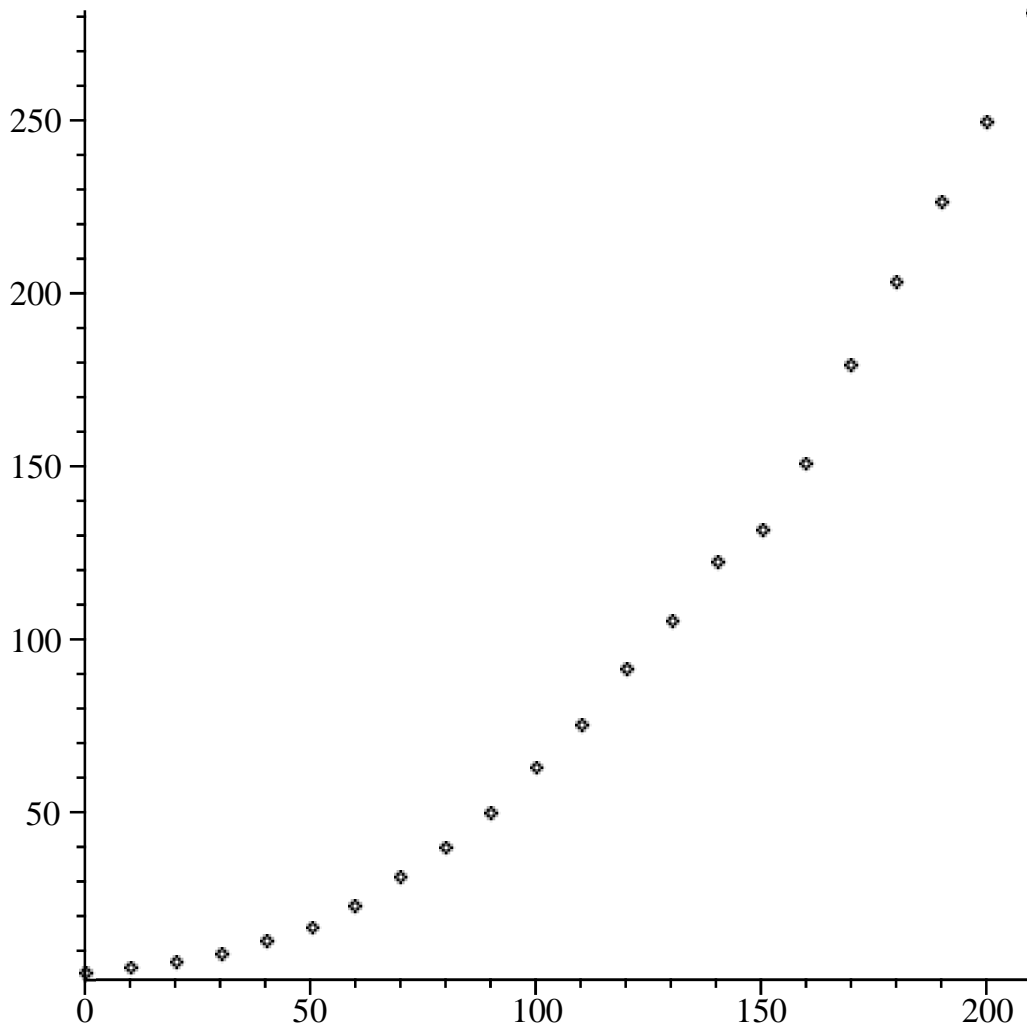
## Regression with more general functions

Of course, not all phenomena which we want to model can be described by polynomials. For example, suppose that we are trying to model the population of the United States. As we've seen, the simplest model is exponential growth, and a more sophisticated model is given by logistic growth. Let's try to fit both to census data from 1790 to 2000. I'll define the vector  $P$  to be the population (in millions) measured every 10 years.

```
> P := [3.93, 5.31, 7.24, 9.64, 12.87, 17.07, 23.19, 31.44, 39.82, 50.16, 62.95, 75.99, 91.97,
        105.71, 122.78, 131.67, 151.33, 179.32, 203.21, 226.50, 249.63, 281.42];
P := [3.93, 5.31, 7.24, 9.64, 12.87, 17.07, 23.19, 31.44, 39.82, 50.16, 62.95, 75.99, 91.97,
      105.71, 122.78, 131.67, 151.33, 179.32, 203.21, 226.50, 249.63, 281.42] (1)
```

The vector  $Y$  will be years (setting 1790 to be year 0). I'm using the "seq" command rather than typing out each one.

```
> Y := [seq(10·i, i = 0..21)];
Y := [0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190,
      200, 210] (2)
> with(Statistics) :
> ScatterPlot(Y, P);
```



So, the above is our plot of the data. Now let's try a regression with an exponential. Unfortunately, this is not automated in Maple; we have to actually understand what's going on. We want  $A$  and  $r$  so that

$A \cdot e^{rt}$  is the best least squares fit, so that the error is minimized. Define the error by

$$\begin{aligned}
 > E := (A, r) \rightarrow \sum_{i=1}^{22} (A \cdot \exp(r \cdot Y_i) - P_i)^2; \\
 E := (A, r) \rightarrow \sum_{i=1}^{22} \left( A e^{(r Y_i)} - P_i \right)^2 \tag{3}
 \end{aligned}$$

Note that I can refer to individual elements in the lists  $Y$  and  $P$  by using subscripts. We want  $A$  and  $r$  so that this is a minimum. Take partials with respect to  $A$  and  $r$  and set these to zero to find critical points.

$$\begin{aligned}
 > eq1 := \text{diff}(E(A, r), A) = 0; \\
 eq1 := 2 A - 7.86 + 2 (A e^{(10 r)} - 5.31) e^{(10 r)} + 2 (A e^{(20 r)} - 7.24) e^{(20 r)} + 2 (A e^{(30 r)} - 9.64) e^{(30 r)} + 2 (A e^{(40 r)} - 12.87) e^{(40 r)} + 2 (A e^{(50 r)} - 17.07) e^{(50 r)} + 2 (A e^{(60 r)} - 23.19) e^{(60 r)} + 2 (A e^{(70 r)} - 31.44) e^{(70 r)} + 2 (A e^{(80 r)} - 39.82) e^{(80 r)} + 2 (A e^{(90 r)} - 50.16) e^{(90 r)} + 2 (A e^{(100 r)} - 62.95) e^{(100 r)} + 2 (A e^{(110 r)} - 75.99) e^{(110 r)} + 2 (A e^{(120 r)} - 91.97) e^{(120 r)} + 2 (A e^{(130 r)} - 105.71) e^{(130 r)} + 2 (A e^{(140 r)} - 122.78) e^{(140 r)} + 2 (A e^{(150 r)} - 131.67) e^{(150 r)} + 2 (A e^{(160 r)} - 151.33) e^{(160 r)} + 2 (A e^{(170 r)} - 179.32) e^{(170 r)} + 2 (A e^{(180 r)} - 203.21) e^{(180 r)} + 2 (A e^{(190 r)} - 226.50) e^{(190 r)} + 2 (A e^{(200 r)} - 249.63) e^{(200 r)} + 2 (A e^{(210 r)} - 281.42) e^{(210 r)} = 0 \tag{4}
 \end{aligned}$$

$$\begin{aligned}
 > eq2 := \text{diff}(E(A, r), r) = 0; \\
 eq2 := 20 (A e^{(10 r)} - 5.31) A e^{(10 r)} + 40 (A e^{(20 r)} - 7.24) A e^{(20 r)} + 60 (A e^{(30 r)} - 9.64) A e^{(30 r)} + 80 (A e^{(40 r)} - 12.87) A e^{(40 r)} + 100 (A e^{(50 r)} - 17.07) A e^{(50 r)} + 120 (A e^{(60 r)} - 23.19) A e^{(60 r)} + 140 (A e^{(70 r)} - 31.44) A e^{(70 r)} + 160 (A e^{(80 r)} - 39.82) A e^{(80 r)} + 180 (A e^{(90 r)} - 50.16) A e^{(90 r)} + 200 (A e^{(100 r)} - 62.95) A e^{(100 r)} + 220 (A e^{(110 r)} - 75.99) A e^{(110 r)} + 240 (A e^{(120 r)} - 91.97) A e^{(120 r)} + 260 (A e^{(130 r)} - 105.71) A e^{(130 r)} + 280 (A e^{(140 r)} - 122.78) A e^{(140 r)} + 300 (A e^{(150 r)} - 131.67) A e^{(150 r)} + 320 (A e^{(160 r)} - 151.33) A e^{(160 r)} + 340 (A e^{(170 r)} - 179.32) A e^{(170 r)} + 360 (A e^{(180 r)} - 203.21) A e^{(180 r)} + 380 (A e^{(190 r)} - 226.50) A e^{(190 r)} + 400 (A e^{(200 r)} - 249.63) A e^{(200 r)} + 420 (A e^{(210 r)} - 281.42) A e^{(210 r)} = 0 \tag{5}
 \end{aligned}$$

We have two equations in the two unknowns  $A$  and  $r$ . Unfortunately, these are nonlinear, and it's unlikely that we'll be able to solve them exactly. I'll try fsolve instead. We need to give a range for  $A$  and  $r$  to give it a start. The exponential which goes exactly through the first and last points is  $3.93$

$$e^{\frac{1}{210} \ln\left(\frac{281.42}{3.93}\right) \cdot y}$$

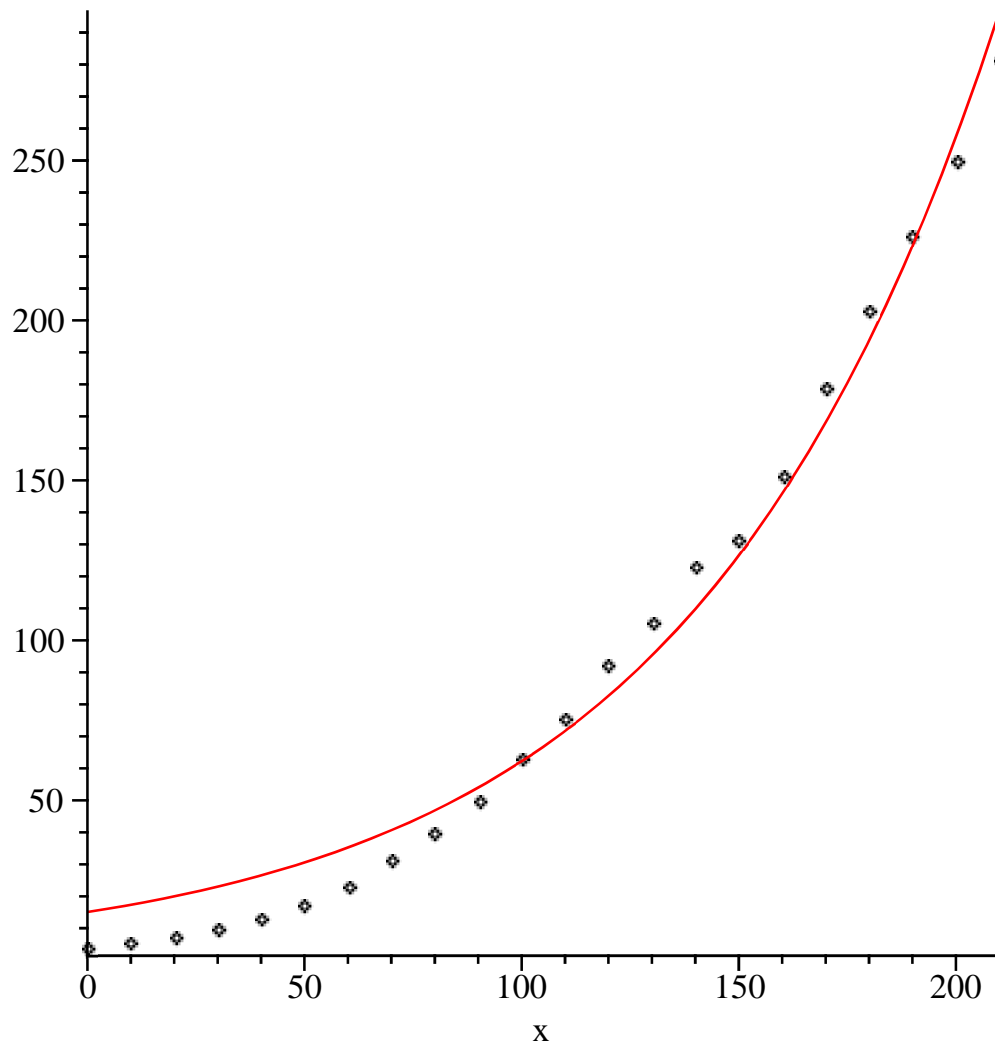
, so perhaps  $A$  is about 3.93 and  $r$  is about 0.02 (which is what I calculated the exponent to be). Let's try intervals around those values. It took me a bit of playing with the intervals until I got Maple to give me an answer. (Another possibility for finding an initial guess is to look at the log of the population versus  $Y$ . If the population is growing exponentially, this should be linear. Then use Maple to do a least squares fit to this line. You'll look at this in homework.)

$$\begin{aligned}
 > sol := \text{fsolve}(\{eq1, eq2\}, \{A, r\}, \{A = 1 .. 20, r = 0.005 .. 0.2\}); \\
 sol := \{r = 0.01419893514, A = 15.03487133\} \tag{6}
 \end{aligned}$$

Let's plot the exponential with those values on top of the scatter plot to see how well we've fit the data.

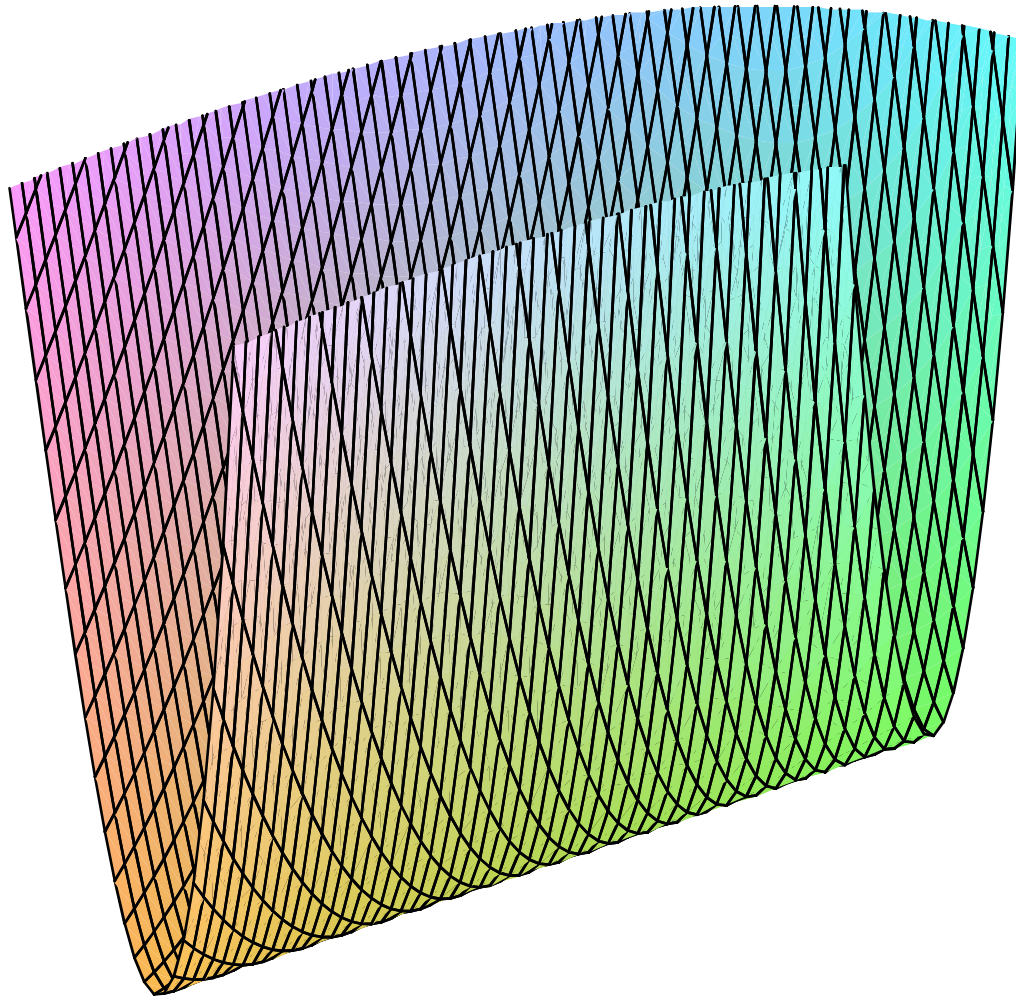
$$\begin{aligned}
 > \text{with}(plots) : \\
 > p1 := \text{ScatterPlot}(Y, P) :
 \end{aligned}$$

```
> p2 := plot(subs(sol, A·exp(r·x)), x=0..210) :  
> display([p1, p2]);
```



Let's see what's going on with the error function. First, let's plot  $E(A, r)$  near the critical point.

```
> plot3d(E(A, r), A=0..20, r=0..0.02, view=0..100000, numpoints=5000);
```



The "view=" part is to restrict the z coordinate. It looks like we have a minimum there. Let's check the value of the error at the critical point.

```
> subs(sol, E(A, r)) : evalf(%);
```

$$2230.463944 \quad (7)$$

This is the sum of the squares of the errors. The square root of this would be more informative.

```
> sqrt(%);
```

$$47.22778784 \quad (8)$$

The units of this are "millions of people".

It looks like the exponential model isn't a great fit. Maybe the logistic model is better. In that model, the

population will be of the form  $p(t) = \frac{p_0 \cdot K}{(K - p_0)e^{-st} + p_0}$ . Let's define a general logistic function in

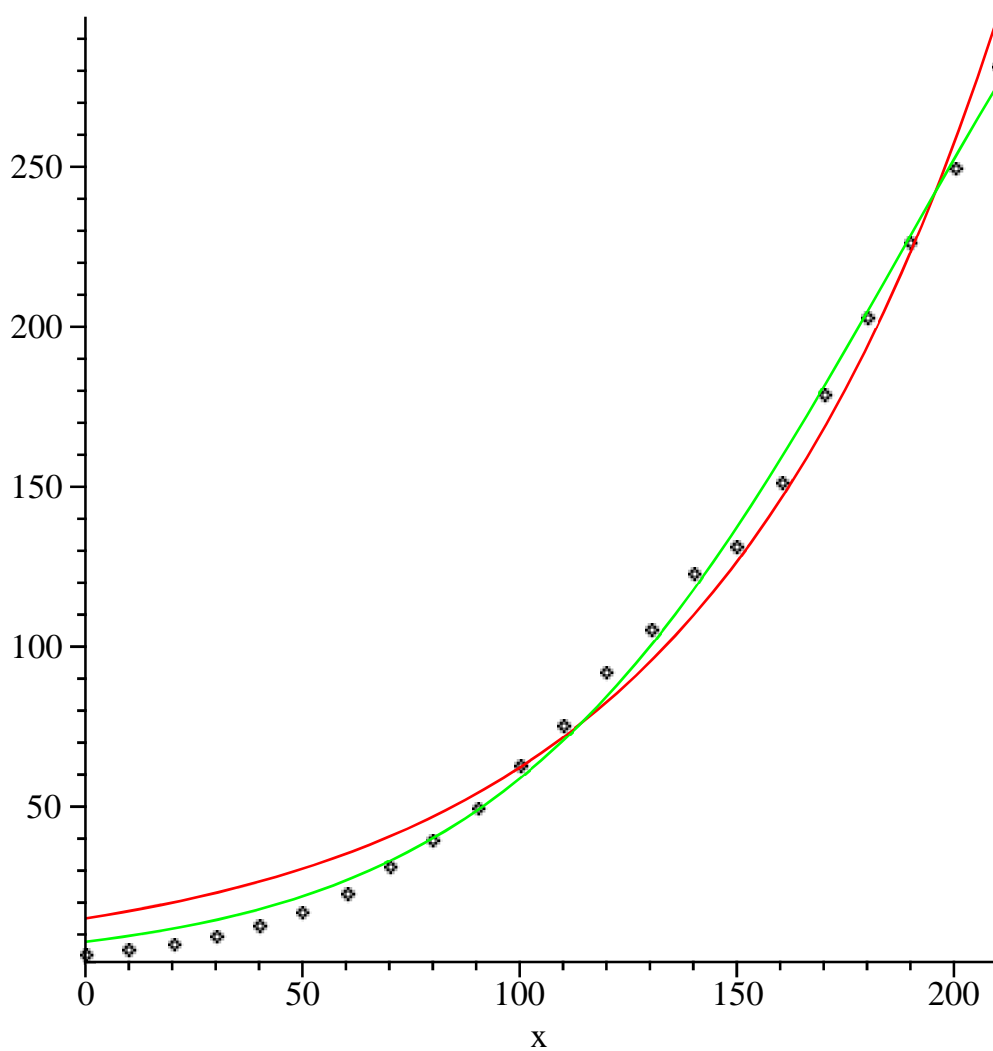
this way. Here  $p_0$  is the initial population, and  $K$  is the carrying capacity.

```
> p := (B, K, s, t) -> \frac{B \cdot K}{(K - B) \cdot \exp(-s \cdot t) + B};
```

$$p := (B, K, s, t) \rightarrow \frac{KB}{(K - B)e^{(-st)} + B} \quad (9)$$

where I'm using  $B$  instead of  $p_0$ . Our error function will be a function of  $B$ ,  $K$ , and  $s$ . I'll call it  $E1$  so as





Definitely the logistic model looks better. Let's compare the error with the exponential fit.

```
> evalf(subs(sol1, E1)); sqrt(%);
440.9982880
20.99995924
```

(12)

```
>
This error is half of the error for the exponential model.
```