Looking for Patterns and Applying the Method of Least Squares to Real Data

Introduction

OBJECTIVE: Minimize the sum of squared residuals to fit an arbitrary function to a set of data.

We look for patterns in real data sets including medical, socio-economic, meteorological, and pollution statistics. Does the least squares method apply only to finding lines of best fit, or does it apply to any type of curve that you wish to fit to a set of data points? We begin by analyzing data with standard fit functions and then utilize the calculus definition of best fit to find a nonstandard fit function.

▪ Technology Guidelines

NOTE: If you have just finished a module, restart *Mathematica* or close the *Kernel* before executing a new module.
TO OPEN CELLS, put your cursor on the right cell bracket and double click.
TO STOP AN EXECUTION
Select the *Kernel* pull-down menu and click on *Abort Evaluation.*
ORDER OF EXECUTION
Execute cells in the order given. Do not skip any Input cells within a given notebook.
SAVING NOTEBOOKS
You can save anytime to any directory you choose, and it is wise to save often.
However, before you do your final save, it is a good idea to delete all your output by selecting the
Delete All Output selection under the *Kernel* pull-down menu.
EXPERIENCING MAJOR PROBLEMS
Save if appropriate, then shut down *Mathematica* and start it up again.

Part I: Time Series Data

Thirty-Two Years of Salaries - Men Versus Women

The following data represent the mean income for men and women in the United States from 1967 to 1998. The study includes people 15 years old and over. All income is in 1998 consumer price index adjusted dollars.
SOURCE: March Current Population Survey
PREPARED BY:
Income Statistics Branch/HHES Division, U.S. Bureau of the Census, U.S. Department of Commerce
Washington, D.C. 20233-8500, (301) 457-3242

In[1]:=

```
Off[General::spell]

Off[General::spell1]

menearnings = {"Men's Salaries", 27232, 28209
  29290, 29389, 31214, 31587, 30495, 29784, 30
  31180, 31039, 29916, 29419, 29180, 29182, 30
  31956, 32210, 32635, 33324, 31978, 31074, 30
  32887, 33126, 33553, 34794, 36315};

womenearnings = {"Women's Salaries", 11500, 1
  12195, 12412, 12930, 12918, 12868, 12889, 13
  13313, 13070, 13207, 13253, 13758, 14148, 14
  15729, 16301, 16703, 17119, 17085, 17027, 17
  17846, 18183, 18790, 19511, 20462};

years = {Years, 1967, 1968, 1969, 1970, 1971, 1
  1974, 1975, 1976, 1977, 1978, 1979, 1980, 19
  1983, 1984, 1985, 1986, 1987, 1988, 1989, 19
  1992, 1993, 1994, 1995, 1996, 1997, 1998};

data =
 Table[{years[[i]], menearnings[[i]], womene
   {i, 1, Length[years]}] // TableForm
```

Out[6]//TableForm=

Years	Men's Salaries	Women's Salaries
1967	27232	11500
1968	28209	11631
1969	29338	11997
1970	29290	12195
1971	29389	12412

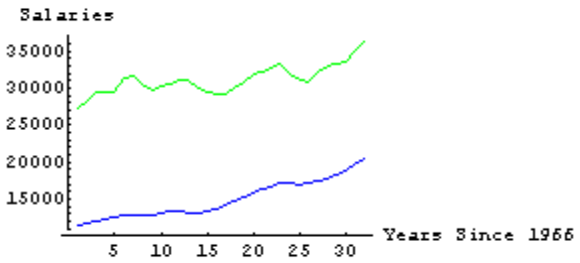
1972	31214	12930
1973	31587	12918
1974	30495	12868
1975	29784	12889
1976	30168	13172
1977	30635	13437
1978	31180	13313
1979	31039	13070
1980	29916	13207
1981	29419	13253
1982	29180	13758
1983	29182	14148
1984	30027	14805
1985	30805	15174
1986	31956	15729
1987	32210	16301
1988	32635	16703
1989	33324	17119
1990	31978	17085
1991	31074	17027
1992	30670	17070
1993	32143	17506
1994	32887	17846
1995	33126	18183
1996	33553	18790
1997	34794	19511
1998	36315	20462

As was probably expected, men's mean income exceeds women's mean income, but we can get a better perspective if we visualize these data through plots.

In[7]:=

```
men = Table[data[[1, i, 2]], {i, 2, Length[yea

women = Table[data[[1, i, 3]], {i, 2, Length[y

pmen := ListPlot[men, PlotJoined → True,
  PlotStyle -> RGBColor[0, 1, 0], DisplayFunct
  AxesLabel -> {"Years Since 1966", "Salaries
  AxesOrigin -> {0, 10000}];
```

```
pwomen := ListPlot[women, PlotJoined → True,
  PlotStyle -> RGBColor[0, 0, 1], DisplayFunct
  AxesLabel -> {"Years Since 1966", "Salaries
  AxesOrigin -> {0, 10000}];

Show[pmen, pwomen, DisplayFunction -> $Displa

Print["\!\(\*
StyleBox[\"Men\",\nFontColor->RGBColor[0, 1
   s salaries are in green and \!\(\*
StyleBox[\"women\",\nFontColor->RGBColor[0,
   s are in blue."]
```




```
Men's salaries are in
  green and women's are in blue.
```

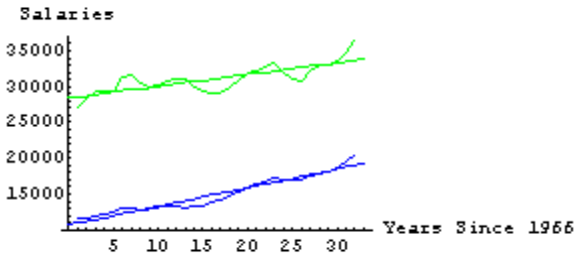
Neither data set follows a straight line, but both have an upward trend. One way we can examine those trends is to compute the best-fit lines for each data set. We will do this using the **Fit** function in *Mathematica*. Note that you choose a linear fit by specifying the {x,1}.

In[13]:=

```
Print["The average salary of males is appro
 fmen = Fit[men, {x, 1}, x], " and"]

Print["the average salary of females is appr
 fwomen = Fit[women, {x, 1}, x], ","]

Print["where x represents the years since 1
```

```
fitplots = Plot[{fmen, fwomen}, {x, 0, 33},
  PlotStyle -> {RGBColor[0, 1, 0], RGBColor[0,
  DisplayFunction -> Identity];

Show[pmen, pwomen, fitplots,
 DisplayFunction -> $DisplayFunction];
```

```
The average salary of males is
   approximately 28386.9 + 163.586 x and

the average salary of females is
   approximately 10641.4 + 260.389 x,

where x represents the years since 1966
```



What do these lines tell you about the rate at which salaries are increasing for men versus women? You could verify that these lines would intersect in about 183 years. Why could this be considered a worthless estimate?

You Try It: Part I

Lake Pollution

The following data sets represent weekly readings of aluminum pollutant levels in Lake Erie over a two-year period. What observations can you make about the amounts of aluminum pollutant? Does the amount of pollution follow a pattern? The first data set represents the data in the order that they were recorded. We then order those data to better analyze the

distribution of the amounts of pollutant.

In[18]:=

```
pollutant = {59.7390, 43.6978, 45.4508, 55.697
  56.6556, 49.9503, 59.2501, 59.2778, 30.8085
  59.0104, 57.1471, 48.4696, 47.4461, 37.0215
  53.8839, 35.8155, 43.8406, 58.4949, 56.7300
  37.7449, 45.7960, 44.0860, 55.0693, 30.5343
  30.2706, 42.1799, 44.9115, 40.9325, 57.4171
  46.1233, 36.4139, 46.7199, 36.3994, 42.1318
  34.4747, 49.2455, 43.0440, 58.2225, 57.4492
  30.9330, 48.8363, 32.6732, 53.6791, 31.7362
  39.6966, 48.3676, 42.1876, 53.2427, 47.2459
  43.3965, 44.8305, 46.8731, 39.5781, 37.0135
  37.1184, 36.5494, 41.2847, 48.9843, 50.6743
  33.1112, 54.3590, 56.7471, 34.9079, 43.4007
  42.0995, 55.4403, 42.9701, 56.2188, 39.4222
  38.6973, 59.6348, 38.3582, 46.5821, 36.8290
  56.9938, 33.0390, 37.8167, 51.4268, 37.3356
  37.7568, 43.9597, 50.1189, 47.6849, 53.0739
  36.0122, 31.8986, 49.6269};

orderedpollutant = Sort[pollutant]
```

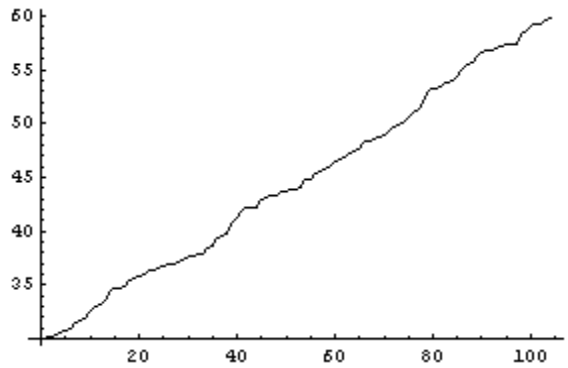
Out[19]=

```
{30.2706, 30.2881, 30.3771, 30.5343,
 30.8085, 30.933, 31.4635, 31.7362, 31.8986,
 32.6732, 33.039, 33.1112, 33.5046, 34.4747,
 34.636, 34.6556, 34.9079, 35.3832, 35.7099,
 35.8155, 36.0122, 36.3994, 36.4139,
 36.5494, 36.829, 37.0135, 37.0215,
 37.1184, 37.3356, 37.7449, 37.7568,
 37.8167, 37.8624, 38.3582, 38.6973,
 39.4222, 39.5781, 39.6966, 40.9325,
 41.2847, 42.0995, 42.1318, 42.1799,
 42.1876, 42.9701, 43.044, 43.3965,
 43.4007, 43.6578, 43.6978, 43.8406,
 43.9597, 44.086, 44.8305, 44.9115,
 45.4508, 45.6032, 45.796, 46.1233,
 46.5821, 46.7199, 46.8731, 47.2459,
 47.4461, 47.6849, 48.3676, 48.4696,
 48.537, 48.8363, 48.9843, 49.2455,
 49.6269, 49.9503, 50.1189, 50.6743,
 50.986, 51.4268, 52.0141, 53.0739,
 53.2136, 53.2427, 53.6791, 53.8839,
 54.0219, 54.359, 55.0693, 55.4403,
 55.6974, 56.2188, 56.6556, 56.73, 56.7471,
 56.9938, 57.1471, 57.3854, 57.4171,
 57.4492, 58.2225, 58.4949, 59.0104,
 59.2501, 59.2778, 59.6348, 59.739}
```

Plot the ordered data, and see what the plot tells you about the pattern of the data.

In[20]:=

```
pollutantplot = ListPlot[orderedpollutant, Pl
```



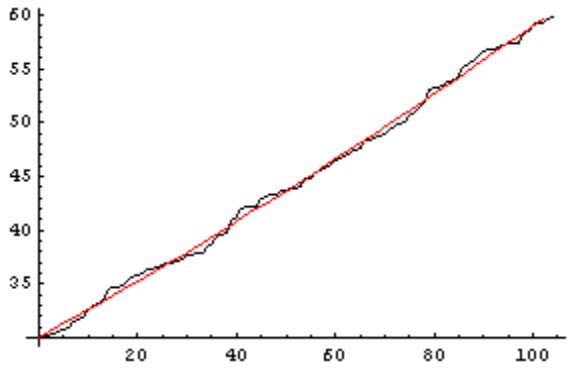
Look for fit functions for these data. Replace the terms in red with an appropriate expression.

In[21]:=

```
pollutantfit = Fit[orderedpollutant, {x^2, x, 1

fitplot = Plot[pollutantfit, {x, 0, 102},
  DisplayFunction -> Identity, PlotStyle -> RG

Show[pollutantplot, fitplot,
 DisplayFunction -> $DisplayFunction];
```

Out[21]=

$$29.9886 + 0.25728\,x + 0.000325239\,x^2$$



The good fit for the linear model could imply that the pollutant distribution is somewhat uniform. What would uniform mean in this context?

Part II: Looking for Cause and Effect: Temperature as a Function of Latitude

Following are data showing 56 cities in the U. S., their January temperatures (^{0}F) averaged over a thirty-year period, and their latitudes.

In[24]:=

```
Off[General::spell]

Off[General::spell1]

Off[Set::write]
```

```
city = {City, MobileAL, MontgomeryAL, PhoenixA
  LittleRockAR, Los AngelesCA, San FranciscoC
  NewHavenCT, WilmingtonDE, WashingtonDC, Jac
  KeyWestFL, MiamiFL, AtlantaGA, BoiseID, Chi
  IndianapolisIN, DesMoinesIA, WichitaKS, Lou
  NewOrleansLA, PortlandME, BaltimoreMD, Bost
  DetroitMI, MinneapolisMN, St.LouisMO, Hele
  ConcordNH, AtlanticCityNJ, AlbuquerqueNM, A
  NewYorkNY, CharlotteNC, RaleighNC, Bismarck
  CincinnatiOH, ClevelandOH, OklahomaCityOK,
  HarrisburgPA, PhiladelphiaPA, CharlestonSC
  AmarilloTX, GalvestonTX, HoustonTX, SaltLak
  BurlingtonVT, NorfolkVA, SeattleWA, Spokane
  MilwaukeeWI, CheyenneWY};

januarytemp = {January Temperature, 44, 38, 35,
  42, 15, 22, 26, 30, 45, 65, 58, 37, 22, 19, 21
  27, 45, 12, 25, 23, 21, 2, 24, 8, 13, 11, 27, 3
  34, 31, 0, 26, 21, 28, 33, 24, 24, 38, 31, 24,
  7, 32, 33, 19, 9, 13, 14};

latitude = {Latitude, 31.2, 32.9, 33.6, 35.4, 3
  40.7, 41.7, 40.5, 39.7, 31, 25, 26.3, 33.9, 4
  39.8, 41.8, 38.1, 39, 30.8, 44.2, 39.7, 42.7,
  39.3, 47.1, 41.9, 43.5, 39.8, 35.1, 42.6, 40.
  36.4, 47.1, 39.2, 42.3, 35.9, 45.6, 40.9, 40.
  36.7, 35.6, 29.4, 30.1, 41.1, 45, 37, 48.1, 4
  43.3, 41.2};

citydata =
 Table[{city[[i]], januarytemp[[i]], latitude
   {i, 1, Length[city]}] // TableForm
```

Out[30]//TableForm=

City	January Temperature	Latitude
MobileAL	44	31.2`
MontgomeryAL	38	32.9`
PhoenixAZ	35	33.6`
LittleRockAR	31	35.4`
AngelesCA Los	47	34.3`

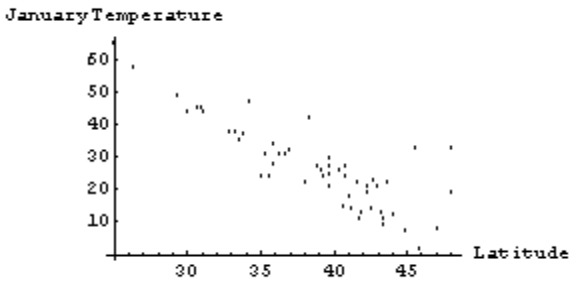
FranciscoCA San	42	38.4`
DenverCO	15	40.7`
NewHavenCT	22	41.7`
WilmingtonDE	26	40.5`
WashingtonDC	30	39.7`
JacksonvilleFL	45	31
KeyWestFL	65	25
MiamiFL	58	26.3`
AtlantaGA	37	33.9`
BoiseID	22	43.7`
ChicagoIL	19	42.3`
IndianapolisIN	21	39.8`
DesMoinesIA	11	41.8`
WichitaKS	22	38.1`
LouisvilleKY	27	39
NewOrleansLA	45	30.8`
PortlandME	12	44.2`
BaltimoreMD	25	39.7`
BostonMA	23	42.7`
DetroitMI	21	43.1`
MinneapolisMN	2	45.9`
St.LouisMO	24	39.3`
HelenaMT	8	47.1`
OmahaNE	13	41.9`
ConcordNH	11	43.5`
AtlanticCityNJ	27	39.8`
AlbuquerqueNM	24	35.1`
AlbanyNY	14	42.6`
NewYorkNY	27	40.8`
CharlotteNC	34	35.9`
RaleighNC	31	36.4`
BismarckND	0	47.1`
CincinnatiOH	26	39.2`
ClevelandOH	21	42.3`
OklahomaCityOK	28	35.9`
PortlandOR	33	45.6`
HarrisburgPA	24	40.9`
PhiladelphiaPA	24	40.9`
CharlestonSC	38	33.3`
NashvilleTN	31	36.7`
AmarilloTX	24	35.6`
GalvestonTX	49	29.4`
HoustonTX	44	30.1`
SaltLakeCityUT	18	41.1`
BurlingtonVT		

	7	45
NorfolkVA	32	37
SeattleWA	33	48.1`
SpokaneWA	19	48.1`
MadisonWI	9	43.4`
MilwaukeeWI	13	43.3`
CheyenneWY	14	41.2`

First, do a scatterplot to see if there is a relationship between the January temperatures and latitude.

In[31]:=

```
scatter =
 ListPlot[Table[{latitude[[i]], januarytemp[
   {i, 2, Length[latitude]}],
  AxesLabel -> {"Latitude", January Temperatu
  PlotStyle -> PointSize[.015]];
```



Since the scatterplot appears to have a somewhat linear pattern, that shows the average January temperature dropping as the latitude increases, we will investigate this relationship further. We will use *Mathematica's* built-in **Fit** function to get results. Note that you choose a linear fit by specifying the {x,1}. We can write the line of best fit as follows.

In[32]:=

```
yfit = Fit[Table[{latitude[[i]], januarytemp[
   {i, 2, Length[latitude]}], {x, 1}, x]
```

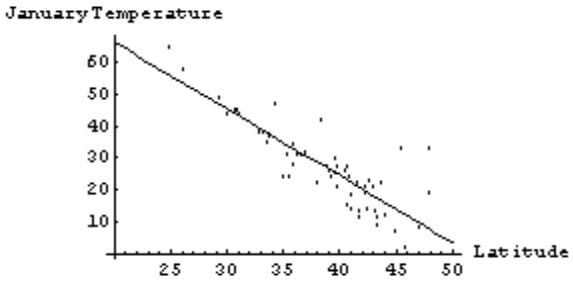
Out[32]=

```
108.728 - 2.10959 x
```

Now we can plot this line with our scatterplot.

In[33]:=

```
linearfit = Plot[yfit, {x, 20, 50},
  DisplayFunction -> Identity];
Show[scatter, linearfit];
```



Can you identify cities that deviate from the pattern? Might you speculate on the reason for the deviation? Name one important factor that might explain at least a part of this deviation.

You Try It: Part II

Here are two data sets you can explore. You can enter your own data sets and then re-execute the computations to explore relationships.

▪ Age and EEG

The electroencephalogram (EEG) is a device used to measure brain waves. Neurologists have found that the peak EEG frequency in children increases with age. The data below represent the results of a study of children age 2 to 16, and the EEG readings (in hertz) represent the average peak EEG frequencies for each age group. It would be reasonable to think of *age* as the independent variable and *eeg* as the dependent variable.

In[34]:=

```
age = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,

eeg = {5.33, 5.75, 5.80, 5.60, 6.00, 5.78, 5.90
  7.28, 7.06, 7.60, 7.45, 8.23, 8.50, 9.38};
```

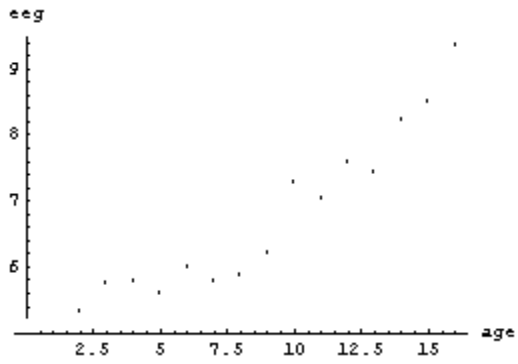
```
eegdata = Table[{age[[i]], eeg[[i]]}, {i, 1, L
```

```
eegdata // TableForm
```

```
scatter = ListPlot[eegdata, AxesOrigin -> {0, !
  AxesLabel -> {"age", "eeg"}, PlotStyle -> Poi
  PlotStyle -> PointSize[.03]];
```

Out[37]//TableForm=

2	5.33`
3	5.75`
4	5.8`
5	5.6`
6	6.`
7	5.78`
8	5.9`
9	6.23`
10	7.28`
11	7.06`
12	7.6`
13	7.45`
14	8.23`
15	8.5`
16	9.38`



Alter the terms in red until you get a good fit.

In[39]:=

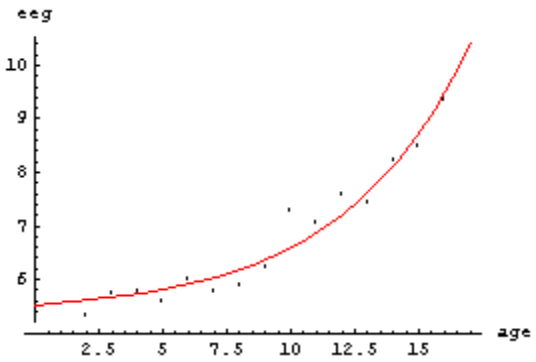
```
eegfit = Fit[eegdata, {Exp[.2 x], 1}, x]
```

```
fitplot = Plot[eegfit, {x, 0, 17}, DisplayFunc
  PlotStyle -> RGBColor[1, 0, 0]];
```

```
Show[scatter, fitplot, DisplayFunction -> $Di
```

Out[39]=

$$5.34803 + 0.168504\, e^{0.2\, x}$$

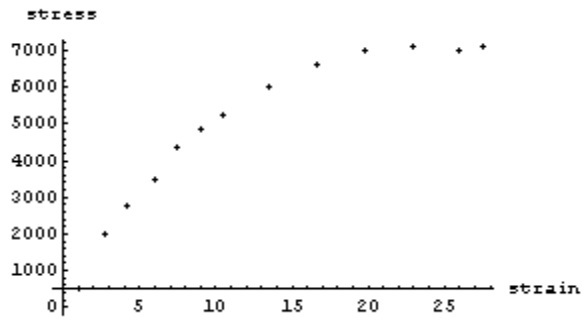


▪ Reliability of Construction Material

The Canadian Geotechnical Journal (Aug., 1985) reported on a study that was conducted to investigate the reliability of the use of fragmented Queenston Shale, a compaction shale, as a rockfill construction material. In particular, the researchers wanted to estimate the stress-strain relationship of the fragmented material. Their study yielded the following results, where axial strain (x) is given as percentagess and deviatoric stress (y) is given in kPa. You might want to consider a quadratic fit for these data.

In[42]:=

```
strain = {1.0, 2.8, 4.2, 6.0, 7.5, 9.0, 10.5, 1
  19.8, 23.0, 26.0, 27.5};

stress = {500, 2000, 2750, 3500, 4375, 4875, 52
  6625, 7000, 7125, 7000, 7125};

ss = Table[{strain[[i]], stress[[i]]}, {i, 1, 1

ssplot = ListPlot[ss, AxesOrigin -> {0, 500},
  AxesLabel -> {"strain", "stress"},
  PlotStyle -> PointSize[.015]];
```



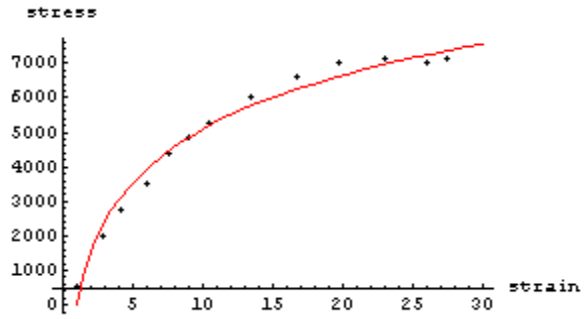
Alter the terms in red until you get a good fit.

In[46]:=

```
ssfit = Fit[ss, {Log[x], 1}, x]

fitplot = Plot[ssfit, {x, 1, 30}, DisplayFunct
  PlotStyle -> RGBColor[1, 0, 0]];

Show[ssplot, fitplot, DisplayFunction -> $Dis
```

Out[46]=

```
-17.3718 + 2224.25 Log[x]
```



Part III: Finding a Least Squares Curve Using Calculus

Data were collected concerning a specific genetic characteristic that researchers believed followed a particular exponential pattern. After observing a scatterplot of the data, they looked for a fit function of the form $y=a\ e^{bx}+c$. Because this is not a standard fit function, it is necessary to use the minimization techniques from multivariable calculus to determine *a, b*,

and *c* from scratch. In all there were 2733 pieces of data collected, and those data were placed into eight categories. In the set defined below, the first entry represents the frequency, the second the categorical variable x, and the third the observed genetic measurement *y*.

In[49]:=

```
Off[General::spell]

Off[General::spell1]

Clear[x, y, a, b, c]

dataset = {{579, 1, 38.08}, {1021, 2, 29.70}, {
  {324, 4, 23.15}, {120, 5, 21.79}, {46, 6, 20.
  {17, 7, 19.37}, {9, 8, 19.36}};

Print["freq. count observations"]

Print[TableForm[dataset]]

y[x_] := a Exp[b x] + c
```

```
freq. count observations
```

579 1 38.08`
1021 2 29.7`
607 3 25.42`
324 4 23.15`
120 5 21.79`
46 6 20.91`
17 7 19.37`
9 8 19.36`

We begin our process by writing the formula for the sum of the squares of the vertical distances between the curve of best fit and the observed data. This will be a function of *a, b*, and *c*, so we will find the first three partial derivatives and solve for the values of *a, b*, and *c* when we set those partial derivatives equal to 0.

In[56]:=

```
Clear[a, b, c, f, fa, fb, fc]
```

```
g[a_, b_, c_] :=
 Sum[dataset[[i, 1]] *
   (dataset[[i, 3]] - a Exp[dataset[[i, 2]] * b
  {i, 1, Length[dataset]}]

ga[a_, b_, c_] = D[g[a, b, c], a];

gb[a_, b_, c_] = D[g[a, b, c], b];

gc[a_, b_, c_] = D[g[a, b, c], c];
```

Now we will set up our equations to solve. Due to previous studies, we had an idea of a range of values for the parameters *a*, *b* and *c*. For that reason, we can use the command **FindRoot** and give some initial starting values, rather than call upon the **Solve** or **NSolve** command.

```
">  About Mathematica
```

In[61]:=

```
Clear[a, b, c]

oursolution = FindRoot[
              {ga[a, b, c] == 0, gb[a, b, c] == 0,
   gc[a, b, c] == 0},
              {a, 40.}, {b, -1.}, {c, 20.}]
```

Out[62]=

```
{a → 33.2221, b → -0.626855, c → 20.2913}
```

Now show y[x] for those values of *a*, *b*, and *c*.

In[63]:=

```
yhat[x_] = y[x] /. oursolution
```

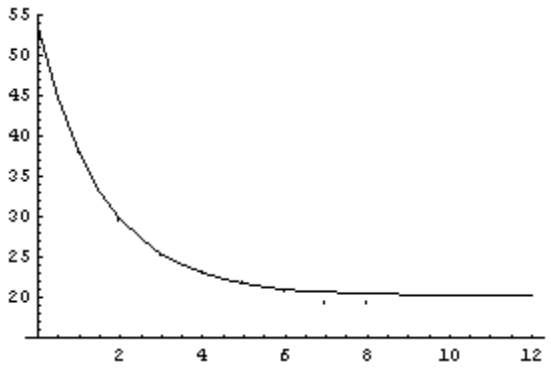
Out[63]=

$$20.2913 + 33.2221\, e^{-0.626855\, x}$$

Look at the results graphically.

In[64]:=

```
p1 = Plot[yhat[x], {x, 0, 12}, PlotRange → {15,
        DisplayFunction -> Identity];

pts = Table[{dataset[[i, 2]], dataset[[i, 3]]}
        {i, 1, Length[dataset]}];

p2 = ListPlot[pts, PlotStyle → PointSize[.01]
    DisplayFunction -> Identity];

Show[p1, p2, DisplayFunction -> $DisplayFunct
```



This looks like a good fit. Compute the sum of the squares of the errors.

In[68]:=

```
yest = Table[yhat[dataset[[i, 2]]], {i, 1, Len

yact = Table[dataset[[i, 3]], {i, 1, Length[da

wts = Table[dataset[[i, 1]], {i, 1, Length[dat

Print[
 "The sum of the squares of the vertical de
   for the 2723 data points is ", sse = wts.(

Print[
 "The sum of the squares of the vertical de
   written as percentages for the 2723 data
 sse = wts.((yact - yest)^2 / yest)]
```

```
The sum of the squares
   of the vertical deviations for
   the 2723 data points is 59.9664


The sum of the squares of the vertical
   deviations written as percentages
   for the 2723 data points is 2.74075
```

This last entry is important for running a statistical test on how good a fit this model is. Because of our calculus work, we know that this is the best fit model for a function of this type.

About *Mathematica*

The disadvantage of using the **Solve** or **NSolve** commands here is that they will look for all solutions, both real and complex. This can take a long time (10 to 15 minutes) because there are many complex solutions to this system of equations. The **FindRoot** command will only look for real solutions if the equations are real and the starting points are real. FindRoot works somewhat like Newton's method. You will notice that it yields only approximate solutions, and it may not converge to a solution at all.

Go back.

Created by *Mathematica* (April 2, 2005)