Problem/Discussion Set for "Curve Fitting/Resampling"

Problem 1. Suppose you want to fit the line $y = mx + b$ to the data points (x_i, y_i), $i = 1, 2, \ldots, N$ and thereby determine the "best-fit" values of the slope, m, and intercept, b. In this problem you will derive formulae for the "least-squares" estimates of the slope and intercept by minimizing the sum of the squares of the deviations (or residuals). [The deviation of the value y_1 is $y_1 - y = y_1 - (mx_1 + b)$.] Proceed by squaring this deviation and summing the squared deviations for all values of y_i. Denote the resulting sum by S. Note that in forming this sum we've regarded all deviations as equally significant (i.e., we've assumed, in effect, that the measurements of the y_i all have an equal uncertainty). Next, consider the sum as a function of m and b and determine the values of the parameters, $\hat{m}$ and $\hat{b}$, for which S is a minimum. Hint: At the minimum, the partial derivatives $\partial S/\partial m$ and $\partial S/\partial b$ will both be zero. Impose this condition and solve the resulting equations. How do your formulae differ if the line is constrained to pass through the origin?

Problem 2. Fit each of the following four data sets (A, B, C, and D) to a straight line using linear regression.[1]

Data Sets							
A		**B**		**C**		**D**	
x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
4.0	4.26	7.0	7.26	8.0	6.77	8.0	6.58
8.0	6.95	9.0	8.77	12.0	8.15	8.0	5.76
7.0	4.82	5.0	4.74	13.0	12.740	8.0	7.71
6.0	7.24	11.0	9.26	11.0	7.81	8.0	8.84
9.0	8.81	4.0	3.1	4.0	5.39	8.0	8.47
13.0	7.58	13.0	8.74	9.0	7.11	8.0	7.04
5.0	5.68	10.0	9.14	6.0	6.08	8.0	5.25
14.0	9.96	12.0	9.13	5.0	5.73	19.0	12.5
12.0	10.840	6.0	6.13	14.0	8.84	8.0	5.56
11.0	8.33	8.0	8.14	7.0	6.42	8.0	7.91
10.0	8.04	14.0	8.1	10.0	7.46	8.0	6.89

For each data set, determine the value of the best-fitting slope (b), intercept (a), their respective uncertainties (σ_b and σ_a), their coefficient of correlation (r_{ab}), the value of chi-square (χ^2), and the probability, Q, that a value of chi-square as poor as this would occur by chance. (Formulae for each can be found in the handout.[2]) Each data set has the form (x_i, y_i); assume that the uncertainties, σ_i, associated with the y_i are constant and equal to 1. Comment on your results.

Problem 3. Imagine that you've made the following measurements (of y as a function of x) to test some current theory of yours. Your theory predicts that the data ought to be well described by a hyperbola of the form $xy = c$, where c is a constant. (For example, your name might be Boyle and the data might represent measurements of the pressure and volume of one mole of gas at room temperature in Ireland.)

Your Measurements													
x	1	2	3	4	6	8	10	12	15	20	30	40	60
y	118	58	38	31	22	13	14	12	6	7	6	2	1

[1] The data are from F. J. Anscombe (1973), *American Statistician* **27**:17.

[2] The function gammq(a,x) used in the handout (from Press et al. (1992), *Numerical Recipes in C*, Cambridge University Press) can be computed using the Matlab function gammainc(x,a) as follows:

$$\text{gammq}(a, x) = 1 - \text{gammainc}(x, a) \; .$$

Given measurements $y_i(x_i)$ you know how to determine the slope of the best-fit line through the origin (Problem 1). To apply your formula you need to "linearize" the problem by transforming the data into a space where the relation between the two variables is linear. Two methods of doing this suggest themselves:

i) You can write a linear relation between y and $1/x$, namely,

$$y = c\,(1/x)\ ;$$

ii) You can write a linear relation between $1/y$ and x, namely,

$$(1/y) = (1/c)\,x\ .$$

1) Transform the data using method (i) and determine the best-fit line through the origin (i.e., the constant c). Plot the transformed data (i.e., y vs. $1/x$) and the resulting best-fit line. Plot the original, untransformed data (i.e., y vs x) and the corresponding best-fit hyperbola. Make a plot of the residuals. What do you conclude about your theory?

2) Transform the data using method (ii) and determine the best-fit line through the origin (i.e., the constant c). Plot the transformed data (i.e., $1/y$ vs. x) and the resulting best-fit line. Plot the original, untransformed data (i.e., y vs x) and the corresponding best-fit hyperbola. Make a plot of the residuals. What do you conclude about your theory?

3) How do you account for any differences between the results of parts (1) and (2).

Problem 4. In the attached excerpt from a recent paper,[3] Voss, Rosowksi, Merchant, and Peake discuss evidence for inter-ear correlations between stapes velocity and impedance at the tympanic membrane. The data shown in their scatterplot are reproduced in the table below. Would their conclusions have changed if they had analyzed a plot of $|V_S/P_{TM}|$ vs $|Y_{TM}|$, where $Y_{TM} = 1/Z_{TM}$? Which procedure would you choose? Justify your choice.

Voss et al. Data		
$\lvert Z_{TM}\rvert/10^6$ [Ns/m^5]	$\lvert V_S/P_{TM}\rvert \cdot 10^3$ [mm/s/Pa]	Bone
48.4	14.66	27
53.7	61.47	9
94.2	38.96	?
95.6	37.90	?
101.6	11.40	?
106.2	13.64	?
110.3	8.658	?
118.4	26.95	?
136.56	25.39	?
136.58	15.69	?
136.85	14.78	?
140.6	8.569	?
207.0	10.84	?

Problem 5. The French "abortion pill" RU-486 was tested as a morning-after contraceptive. According to *Science News* (10/10/92), 398 women were given a standard treatment (high doses of hormone), and 402

[3] S.E. Voss, J.J Rosowski, S.M. Merchant, and W.T.Peake (2000). "Acoustic responses of the human middle ear," *Hearing Res.* **150**:43–69.

women were given RU-486. Four of the women given the "standard treatment" got pregnant, while none of those given RU-486 did. Estimate the probability that such a difference arose by chance. Determine your probability computationally using the Monte Carlo method, i.e., by using a random-number generator to simulate the clinical trial assuming the "null hypothesis" that RU-486 has no effect (i.e., is no more effective at preventing pregnancy than the standard treatment). Provide a computational estimate of the uncertainty of your computed probability.

Problem 6. Do the liquor prices in state-run and private sector systems differ systematically? In the 1960s Julian Simon studied the price of liquor in the sixteen "monopoly" states (where the state government owns the retail liquor stores) compared to the twenty-six states in which retail liquor stores are privately owned. (Some states were omitted for technical reasons.) The representative 1961 prices of a fifth of Seagram 7 Crown whiskey in the two groups of states are given in the table below.

Liquor Prices [1961 dollars]					
State-owned			**Privately-owned**		
4.65	4.55	4.11	4.82	5.29	4.89
4.15	4.20	4.55	4.95	4.55	4.90
3.80	4.00	4.19	5.25	5.30	4.29
4.75	4.74	4.50	4.85	4.54	4.75
4.10	4.00	5.05	4.85	4.85	4.50
4.20			4.75	4.79	4.85
			4.79	4.95	4.95
			4.75	5.20	5.10
			4.80	4.29	
Mean = \$4.35			Mean = \$4.84		

Note that the mean prices differ by 49 cents (or roughly 10%). Use Monte-Carlo methods to estimate the probability that the observed price difference might simply have arisen by chance.

Problem 7. In October 1981, when the World Series was tied at two games apiece, the *New York Times* wrote

> Nobody needed reminders of the urgency of today's game, but reminders were hard to escape. This was the 30th time in the 78 years of the series that the teams had been tied after four games. And 22 times the team that won the fifth game had gone on to win the series.

In this problem you'll use Monte-Carlo methods to examine the hypothesis that winning or losing a game in the World Series has the same odds as tossing a fair coin.

1. Use a random-number generator (or a coin, if you prefer) to simulate the playing of many series. Assume that the probability of winning a game is $\frac{1}{2}$. Use your results to calculate the expected values of the statistics given in the quotation. In other words, how many times in 78 series should one expect the teams to be tied after four games? An in how many of those series should one expect the winner of the fifth game to win the series?

2. Additional statistics appear in the following table, which reports the distribution of lengths of the World Series for the 72 series played during the years 1923–1995 (due to the baseball strike, no series was played in 1994).

Number of Games in the World Series (1923–1995)	
Length of series	Number of occurrences
4 games	12
5 games	14
6 games	16
7 games	30

Use a random-number generator to calculate the expected values for the statistics in the table assuming that the probability of winning a game is $\frac{1}{2}$. Compute the probability that deviations from the predicted distribution as large as those shown in the table could have arisen by chance. Hint: Use your simulations to compute the distribution of some measure of the overall "quality of fit" between the actual data and your predicted distribution (e.g., the sum of the squared deviations). How consistent is the World Series data with the tossing of a coin?

Problem 8. Wetts and Herrup[4] present the following table of Purkinje cell counts from the cerebella of seven assorted chimeric and wild-type mice.

Purkinje Cell Data		
Mouse Id	Number of Purkinje cells	Cell-count ratio (re χ^{11})
χ^{11}	10,200	(1)
χ^{7}	30,200	2.96
χ^{13}	60,300	5.90
C3H^{1}	82,400	8.07
C3H^{2}	83,900	8.22
χ^{19}	102,000	9.97
AKC3	114,000	11.2

The authors estimate their cell counting error to be 3%. Column 3 gives the ratio of the cell counts for each mouse to that for χ^{11}. They write

> When the number of PCs [Purkinje cells] in each of these six animals is divided by the number of PCs in χ^{11}, the quotients are not significantly different (within the 3% counting error) from integral values. ... [The] quotients for most of the animals are well within 3% of an integral multiple, and only one [namely, C3H^{2}] comes close to, but does not exceed, the 3% limit. The probability that these quotients fell within 3% of an integer by chance alone is very small.

They conclude that the number of Purkinje cells in the mouse cerebellum is quantized (apparently in multiples of roughly 10,200). They argue that this quantization arises because cerebellar Purkinje cells are clonal descendants of a small pool of progenitor cells. (The cell-count ratio then estimates the number of these progenitor cells.)

- Comment on their use of 3% to characterize the expected uncertainty in the cell-count ratio. Hint: Use the rules for the propagation of random errors.
- Evaluate the probable significance of their findings by estimating the value of "very small" computationally. Hint: Use a random number generator to estimate the probability that seven numbers chosen

[4] Wetts, R. and Herrup, K. (1982). "Cerebellar Purkinje cells are descended from a small number of progenitors committed during early development: Quantitative analysis of lurcher chimeric mice," *J. Neurosci.*, **2**:1494–1498

at random (from a "reasonable range" of cell-count values, e.g., 4,000–150,000) all fall within 3% of integer multiples of n, where n is the putative quantal unit (i.e., the smallest of the seven random numbers). How sensitively do the results depend on the choice of "reasonable range"? For example, how does the probability change if you allow cell counts in the range 3,000—200,000? How do the results depend on the number of animals examined? And how do the results differ if you used the actual expected uncertainty in the quotients rather than the authors' criterion value of 3%?