Fall 1999 Course 513-697: Applied Linear Models
Assignment 2 Answers and Comments

1.39a Easiest if use form $b_1 = \sum \frac{x - \bar{x}}{\Sigma(x - \bar{x})^2} y$. Since X's are symmetric, only the observations at X=5 and X=15 contribute to the fitted slope. With summary level ("ecologic" in epidemiologic parlance) data, clearly $b_1 = (\bar{y}_3 - \bar{y}_1)/10$. With individual level data, $b_1 = (y_{31} + y_{32} - y_{11} - y_{12}) / 20 = (\bar{y}_3 - \bar{y}_1)/10$. And since line goes through $(\bar{\bar{x}} , \bar{\bar{y}}) = (\bar{x} , \bar{y})$, the lines must be coincident.

NOTE If there are different numbers of observations at each X, this "same line with summary level data as with individual data" no longer holds -- unless one gives more weight to the $\bar{y}$'s based on more observations.

1.39b If $\bar{y}_i$ is based on n_i observations. and if model holds, then the residual will be quite small. The magnitudes will reflect the variation of a mean rather than of an individual point... after all, the model should now be

$$\bar{y} = \beta_0 + \beta_1 X + \bar{\varepsilon} \qquad \text{var}(\bar{\varepsilon}) = \sigma^2 / n_i$$

where σ^2 = var(ε) = variance of <u>individual</u> ε's

So if wanted to get back to σ^2 , would need to enlarge the variation seen in the residuals in the summary level data by the sample sizes on which the means are based.

It is not 100% clear if this is what NKNW were asking, since it would involve looking at the residuals.

What I think they may have been getting at was that if you had access to the individual-level data (y_{11} to y_{32}), then the variation seen in the two data points at each X is a reflection of the individual variation.

so... from y_{11} & y_{32} we have 1 df to estimate σ^2
(after we calculate their mean, that's all they are good for!),

from y_{21} & y_{22} we have 1 df to estimate σ^2 ,

from y_{31} & y_{32} we have 1 df to estimate σ^2 .
----
in all (pooling) we have 3 df to estimate σ^2

I expect NKNW were preparing us for the idea of "pure error" used in Chapter 3 to perform Goodness of Fit tests for models.

1.40 Most of you had the (correct) intuition that removing a point that sat right on the fitted line would not change the fitted line. Some one you tested it empirically. However, I think some of you took a shortcut in saying that because that that point contributed zero to Q, its removal would not affect the line. Maybe I'm missing something, but here is one way I reasoned it out:

The two estimating equations imply

$$\sum_1^n e = 0 \text{ and } \sum_1^n x\, e = 0..$$

But since the e in question is zero, we also have, with the same B0 and B1,

$$\sum_1^{n-1} e = 0 \text{ and } \sum_1^{n-1} x\, e = 0.$$

These two (normal) equations imply that the same (B0,B1) values give the smallest sum of squares for the n-1 datapoints.

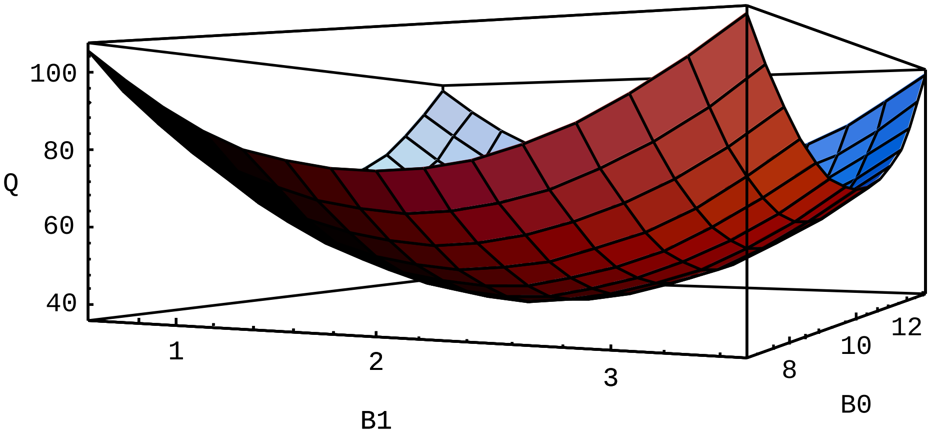
Another way [assume last (n-th) datapoint is the one on the line]

Q is a function of B0, B1 and the n data points. [am using B for parameter]

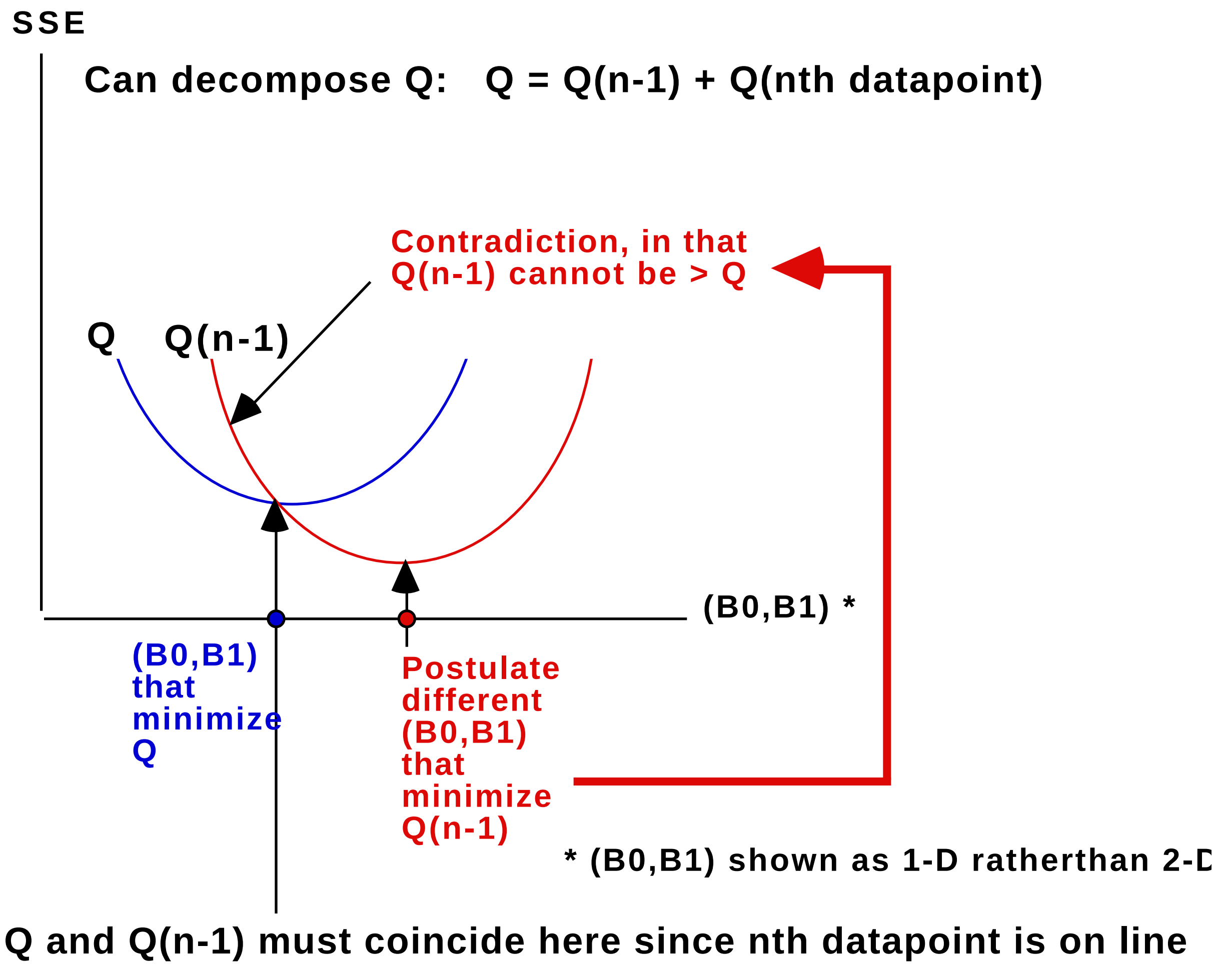
Q is a quadratic form in (B0,B1) with a minimum rather than a maximum (Q is concave up?)

Here is Q for the 4 data points, that give the LS line 10 + 2X and Q_{min}=38

x	y	yhat	e	xe	e^2
-2	9	6	3	-6	9
-1	**8**	**8**	**0**	**0**	**0**
0	5	10	-5	0	25
3	18	16	2	6	4
Σ	40	40	0	0	38



Now imagine that Q(n-1,B0,B1) had a smaller minimum at some (B0,B1) value different from the (10,2) values that minimize Q. This would imply a contradiction (see diagram)


SSE
Can decompose Q: Q = Q(n-1) + Q(nth datapoint)
Contradiction, in that
Q(n-1) cannot be > Q
Q
Q(n-1)
(B0,B1) *
(B0,B1)
that
minimize
Q
Postulate
different
(B0,B1)
that
minimize
Q(n-1)
* (B0,B1) shown as 1-D ratherthan 2-D
Q and Q(n-1) must coincide here since nth datapoint is on line

1.41 **<u>No-Intercept</u> model**, homoskedasticity (constant amplitude of errors)

Least squares estimator of slope (by calculus) ...

$$
\begin{aligned}
b_1 &= \frac{\Sigma\ x\ y}{\Sigma\ x^2} \\
&= \frac{\Sigma\ x^2\ \frac{y}{x}}{\Sigma\ x^2} \\
&= \frac{\Sigma\ w\ \frac{y}{x}}{\Sigma\ w} \\
&= \frac{\Sigma\ w\ \text{slope}}{\Sigma\ w}
\end{aligned}
$$

= weighted average of the n individual slope estimates y/x ,

with weights proportional to x^2.

If $\text{var}(\varepsilon) = \text{var}(\ y \mid x) = \sigma^2$, then $\text{var}(y/x) = \sigma^2 / x^2$.
i.e., slopes based on larger x's are more trustworthy [e.g. estimates of fuel consumption (litres/distance) are more trustworthy if measured over a longer distance] and so should receive more weight in the overall estimate. Indeed, if we weight by the inverse of variance, or by x^2 / σ^2 , we get the weighted form above.

2 (designed somewhat along the lines of exercise 1.42) You are interested in estimating μ = **the average weight of a blank sheet of paper**. You have just 2 "datapoints". One of these comes from a single sheet of paper weighed on its own, the other from two <u>other</u> sheets [without staples!] weighed together.

	x	y
datapoint	# sheets weighed	total weight of the x sheets[s]
1	1	145
2	2	302

a Using only the math tools learned in elementary school, and before going on to part b, make an estimate of μ. Explain the basis for this estimate. Give a rough idea of the "inexactness" of your estimate.a

My first instinct: combine the weights: 447 for 3 sheets => $\hat{\mu}$ =149 .

In a model with no intercept [as in exercise 1.29 and 1.41] it is also possible to directly use y/x as a separate estimate of the slope from each observation (this is not possible with non-zero intercept models) i.e. could also make separate estimates from each set: 145 per page in set 1 and 302/2=151 from set 2, then average the two estimates to get an average slope of 148. Or, might want to weight the151 a bit more, since it is based on a bigger set and so maybe more precise. Weighting would depend on the nature of the errors (see below):

As for precision estimates, we have two observations from which we must estimate 1 parameter. That leaves 1 df (not a lot!) for estimating error variation.

b Besides the difference in x, two possible explanations for the variations in y are (i) the sheets of paper are of uniform weight, but the measuring instrument, while corrected calibrated, cannot reproduce the same result from weighing to weighing, instead giving readings that fluctuate around the true value with a standard deviation of σ (ii) the instrument is perfect calibrated and produces perfectly(!!) reproducible measurements, but because of fluctuations during manufacture, there is slight random variation [standard deviation = σ, of the same magnitude as in (i)] in the weights of individual sheets.

Suppose you knew which of the two explanations was the correct one. Should that knowledge influence the estimation? Explain your reasoning [no formal models or calculations required at this stage!].

If I thought (as in exercise 1.41) that the errors were in the measurement per se, then even though y_2 and y_1 are equally precise, $y_2/2$ should be more precise than $y_1/1$. On the other hand, if the errors arein the sheets per se, and not in the measurement, then each "sheet" should get equal weight -- as it does in the 447/3 estimate from elementary school.

c Carefully write out separate statistical models for the observed y's under (i) and (ii). They may look somewhat like the models in exercises 1.41 and 1.42 but you need to pay particular attention to the variance of each of the 2 ε's !!

(i) true weight of each sheet is exactly μ ; measurement errors ~ $N(0, \sigma^2)$

$$y_1 = \mu + \varepsilon_1 \qquad = x_1 \mu + \varepsilon_1$$

$$\varepsilon_1 \ \& \ \varepsilon_2 \sim N(0, \sigma^2)$$

$$y_2 = (\mu + \mu) + \varepsilon_2 \qquad = x_2 \mu + \varepsilon_2$$

(ii) weight of a sheet differs from μ by manufacturing variation $\varepsilon \sim N(0, \sigma^2)$

(but no measurement error)

$y_1 = \mu + \varepsilon_1$ $= x_1 \mu + \varepsilon_1$ $\varepsilon_1 \sim N(0, \sigma^2)$

$y_2 = (\mu + \varepsilon_{21}) + (\mu + \varepsilon_{22})$ $= x_2 \mu + \varepsilon_{21}+\varepsilon_{22}$ $\varepsilon_{21}+\varepsilon_{22} \sim N(0, \underline{\mathbf{2}}\sigma^2)$

d Explain whether your two proposed models meet the specifications of model 1.1 on page 10.

(i) does ; (ii) does not [heteroskedasticity: var(ε) larger with larger X]

e Do you think the variance part of the model suggested in exercise 1.42 is realistic? Explain your reasoning.

I would expect that time to correct errors is page specific and thus that the variations accumulate, just as in model (ii).

Estimation based on **model that assumes variation is in true weight of sheets**:

weight of sheet = μ + [REAL] weight variation (ε) ; var{ε} = σ^2

so E and var of y_1=weight of set with 1 sheet = μ and σ^2

E and var of y_2=weight of set with 2 sheets = 2μ and $2\sigma^2$

so $\text{Var}\{y_1 / 1\} = \sigma^2$ $\text{Var}\{y_2 / 2\} = 2\sigma^2 / 4 = \sigma^2 / 2$

so slope estimate from set 2 should get a weight of 2/3 and estimate from set 1 weight of 1/3, i.e. (2/3) of 151 and (1/3) of 145 or 149 as our first instinct above.

3 Fill in the **missing values in the 6 situations** given in section 5-6 "SSE & the estimator of the (common) X-specific variation of E" of the "Chapter 5" notes for course 678.

Main message was that n e's are tied together by 2 (estimating) equations

Σ e = 0 and Σ x e = 0

or

$\Sigma\ e = 0$ and $\Sigma\ \hat{y}\ e = 0$

or

$\Sigma\ y = \Sigma\ \hat{y}$ and $\Sigma\ x\ y = \Sigma\ x\ \hat{y}$

It is not enough to use $\Sigma\ e = 0$; this constraint simply ensures that the line goes through $\bar{y}$. The constraint $\Sigma\ x\ e = 0$ ensures that the line has the correct slope. when we go to multiple regression, the residuals will be tied together by as many equations as there are terms in the regression model.

4 Hurricane data

b Round to 1 decimal the fitted intercept and slope of the uncentered model and project "forward" from the rounded intercept to the 1980-1989 decade. Compare this fitted value with that obtained by the same amount of rounding applied to coefficients fitted to a centered model, and to the fit with that obtained with no rounding of the coefficients of the uncentered model. Comment.

Errors (even just rounding) in model coefficients have more impact if start far away from where the X's are!!

c Although the text shows that the last decade is "(1990-1999)" the data (Y=2) are only for the 5-years 1990-1994. Describe some ways to incorporate this "complication' into the analysis.

It is not legitimate to create 'synthetic' data by say doubling the value (2 hurricanes) for the 5 years 1990-94. If you do this, the regression software/model has no way to know that the last datapoint is partly synthetic: it can't distinguish a legitimate 4 from a "2 doubled" and so cannot accurately assess what is actual from induced variation.

Instead, here are some options:

- take "average hurricanes per year" for each decade, and give only 1/.2 weight to the value of 0.4 for the last decade.
- write out the model carefully... starting with the null model (no temporal trend) so interest is in μ = average for a 10-year period

	y	= x β	+ ε	Var(ε)
$y_{1900-1909}$ is an estimate of μ	$y_{1900-1909}$	= 1 × μ	+ ε_0	σ^2
...	...	...	...	

$y_{1980\text{-}1989}$ is an estimate of μ $\qquad y_{1980\text{-}1989} = 1 \times \mu + \varepsilon_8$
σ^2

$y_{1990\text{-}1994}$ is an estimate of $(1/2)\mu$ $\qquad y_{1990\text{-}1994} = (1/2) \times \mu + \varepsilon_9 \qquad (1/2)\sigma^2$

The variance assumption is speculative, and assumes that yearly counts are independent; it yields the same estimate as the average per year approach.

5 Sleeping through the night

a Draw in the regression line. Hint: use the "centered" form: $\mu_{y|x} = \mu_y + \beta (x - \mu_x)$
[also, you may wish to use (general???) relationship 15.10b on page 638, or look at page 7 of "Notes on M&M Chapters 2 and 9" under Chapter 5 of www.epi.mcgill.ca/~web2/hanley/c678/]

- *negative correlation so low birthweight<--> later age and vice versa*

- *r = –0.6 means that the variance of y's at any x*

= 1 – 0.6^2 = 64% of overall variance in y

i.e. SD(Y | x) = 80% of SD in all y's

i.e 80% of 10 or 8 days. still much scatter in indiv. y's at each x

Line goes through the point x=xbar=3.5, y=ybar=50

slope = r • SD(y) / SD(x) = –0.6 • 10 / 0.5 = –12 days per Kg of birthweight.
so if take x = 2.5 Kg, then E(y | x=2.5) = 50 + (–12)(–1) = 62 days

b If the birthweights were in grams rather than Kg, what would β be? What would the correlation be? Likewise, if the age was measured in weeks, what would change?

If the birthweights were in grams rather than Kg, etc..

β is –12 days per Kg or –12/1000 i.e. –0.012 days/gram; Correlation unchanged
If age in weeks, slope is –12/7 weeks per Kg; correlation again unchanged

c If we consider a baby that weighed 2.5 Kg at birth, what is the probability that it will sleep through the night before it is 10 weeks (70 days) old? before it is 10 weeks old? You don't need to DO the calculation, just indicate HOW to. What distributional assumptions do you have to make?

If we knew that mean is 62 days, SD is 8 days, then can use
Prob(Z > (70 – 62) / 8) i.e. Prob(Z>1)

We are assuming Gaussian-ness of the individual variations here.

Since unsure about where mean y is for infants with x=2.5 (after all the line is just and estimate of where the mean is), would need to add this to the uncertainty via the formula { I am assuming n is large, and using n= 800 just to illustrate]

$$70 = 62 + z \bullet 8 \bullet \sqrt{1 + \frac{1}{800} + \frac{[2.5-3.5]^2}{\Sigma\{x - 3.5\}^2}}$$

i.e. get $\text{Prob}\left(Z > \frac{70 - 62}{8\sqrt{1 + \frac{1}{800} + \frac{[2.5-3.5]^2}{\Sigma\{x - 3.5\}^2}}}\right).$

With n so large here, the second and third terms under the square root sign are negligible.

If we consider all babies that weigh 2.5 Kg at birth, what is the probability that the average age at which they will first sleep through the night is less than 70 days?

This is a question dealing with where we think $\mu_{y|x=2.5}$ *is. Our best estimate is 62 days. The uncertainty is given by its SE i.e. by* $8 \bullet \sqrt{\frac{1}{800} + \frac{[2.5-3.5]^2}{\Sigma\{x - 3.5\}^2}}$ *so we can use CLT and assume Gaussian uncertainty. Notice that SE is very close to 8/√800*

BUT NOTE that giving this latter estimate for the average of all such babies to a parent is not very helpful... babies are individualists, no matter what size n we use to estimate the average for babies born 2.5 Kg.

6 Alcohol and impairment

a Paired t-test (or non-parametric analogues such as signed rank test or straight sign test) is most appropriate. We do not have the absolute differences but we do have the % change in each person.

b Can use $t=r[(n-2)^{1/2}] / [(1-r^2)^{1/2}]$. Note that this just tests the H_0: rho=0

c I would say nonzero intercept because (1) might have threshold effect (ii) expt did not measure baseline alcohol (which might not be exactly zero)

d Point to ss #3,5 and 6. Point out that the line is an estimated line, with a lot of uncertainty (different in the next 12 ss) and that even then it refers only to the average person; one must still allow for individual variation.

e Measure x and y before and during (that way could get an estimate of "y at 80" for each person. After all, that is what the main question was. Fit a separate curve for each person and then describe the distribution of "%impairment at 80" estimates.

Measure personal characteristics and see if the estimates of impairment segregate along these lines.

Measure same persons in control situation (non alcohol) to understand how much could be simply due to tiredness etc.