Prep for Exam 3

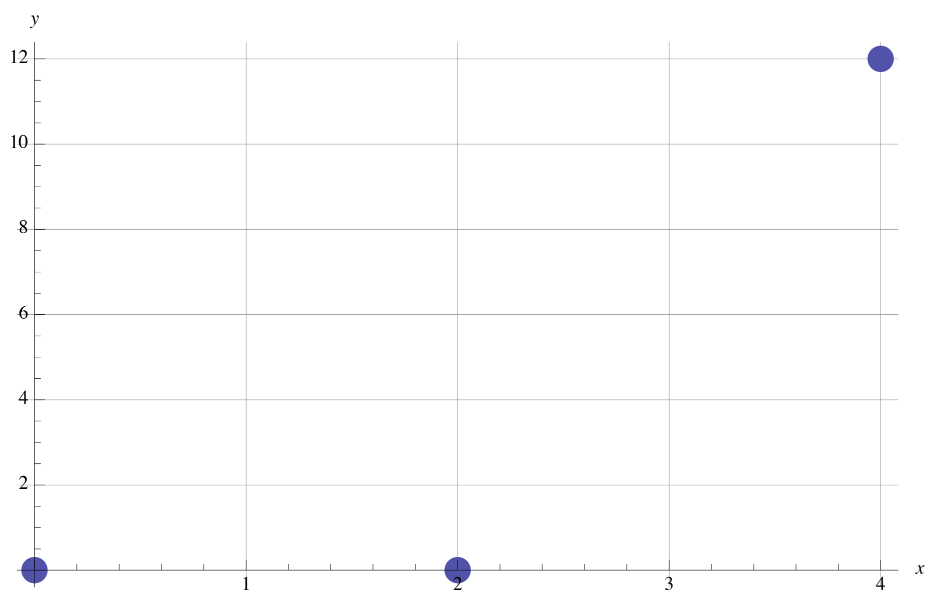
Every solution must, where possible, show the method or formula being used, its evaluation by insertion of available information, and the actual reduced answer. On exam 3, the last part may be dispensed with in some cases in order that too much time need not be spent with such matters.

LR (linear regression)

1. Data

x	y
0	0
2	0
4	12

a. In this plot of the data identify the point of averages $(\overline{x}, \overline{y})$ with a **+**.



b. Calculate slope of regression of y on x. Hint: The math expression of slope having denominator $\overline{x^2}-\overline{x}^2$ will simplify this particular calculation.

c. Plot the regression line in the above grid, clearly indicating two points on the line, or one point and slope.

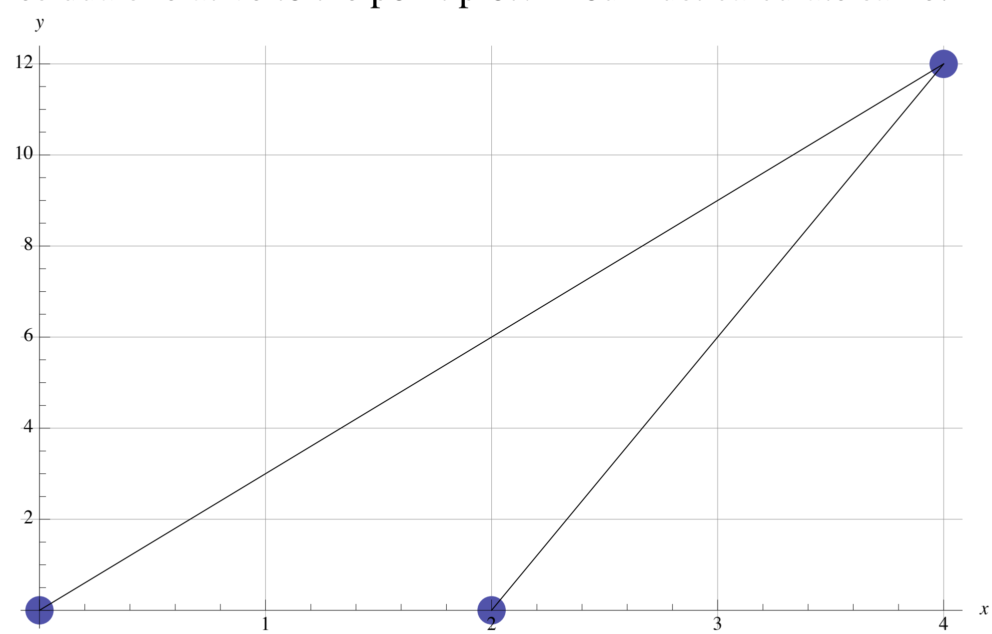
d. Calculate the fitted value $\hat{y}$ for data point (0, 0) and also indicate it in the above plot. Don't neglect to show your method.

e. Calculate the residual y - $\hat{y}$ for data point (0, 0).

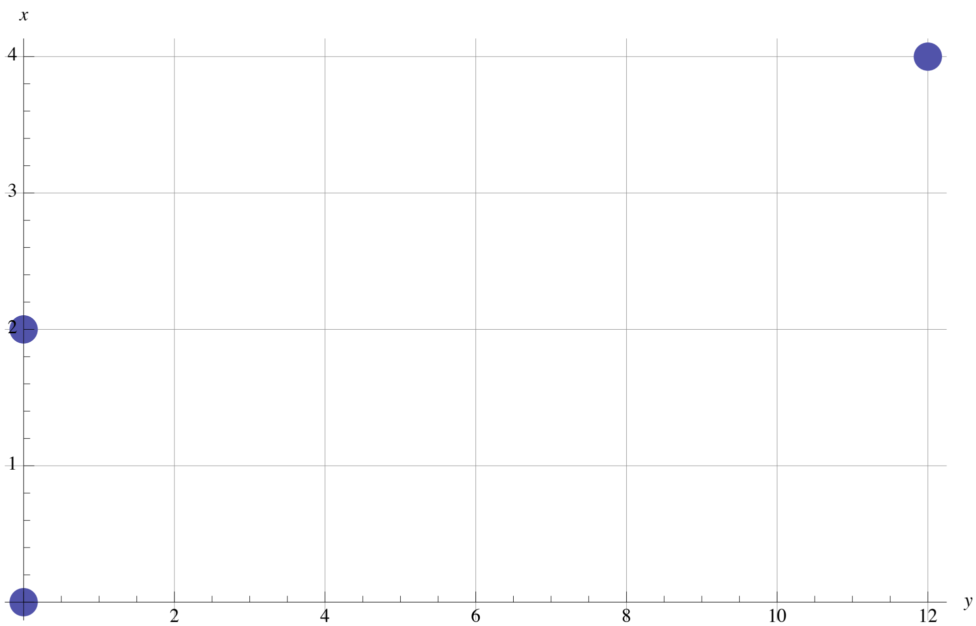
f. Indicate all three residuals in the plot by vertical line segments **|** . With + and - signs indicate which residuals are + and which are -.

2. Two lines are overlaid with a point plot in the figure below.

a. Which line (the upper line or the lower line) has the smaller sum of squares of residuals relative to the point plot? You must calculate same.

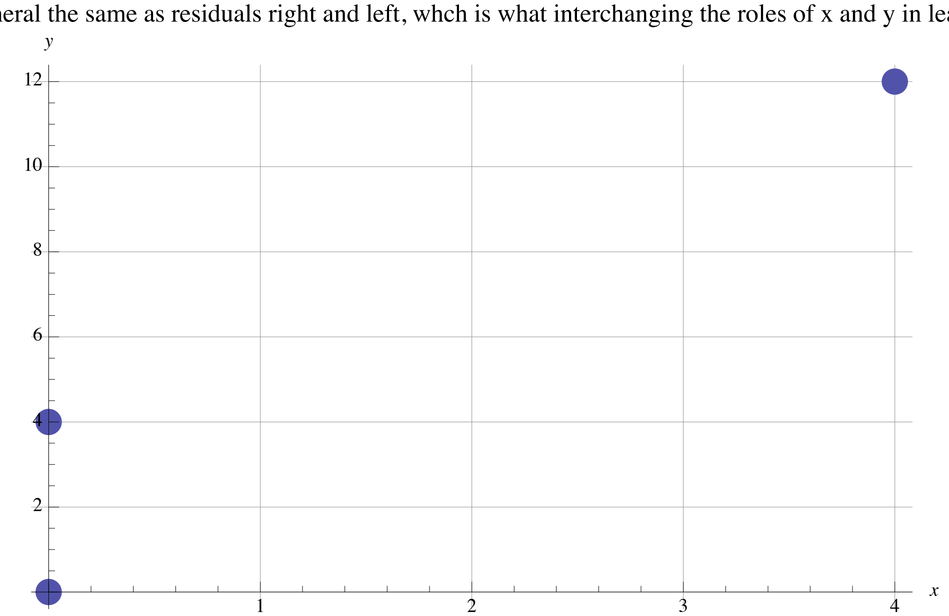


b. The next plot interchanges x with y. It is easy to fit, by eye, the precise line of regression of x on y. Do so.



c. Overlay your (flipped x for y) regression line of y on x from 1(c) on the (flipped x for y) plot just above. Show by this example that regressing x on y is not in general the same as taking the regression of y on x and flipping the axes. *The regression relationship very much depends upon which variable is chosen as the independent variable since (obviously) residuals up and down are not in general the same as residuals right and left, whch is what interchanging the roles of x and y in least squares would amount to.

3.



a. In the plot above, plot the regression line of y on x by eye. By eye also, determine the exact fitted values $\hat{y}$.

b. Calculate the correlation of y with x (same as correlation of x with y).

c. Calculate the correlation of y with $\hat{y}$. It is termed *multiple correlation* and extends to the MLR setting. It is true in general that

$$r[y, \hat{y}] = | r[y, x] |.$$

Verify that this is true in the present example.

d. Using properties of correlation, determine

r[3x-6, 9y+2]

r[3x-6, -9y+2]

r[-3x-6, -9y+2]

without having to, in each case, re-calculate a correlation directly by first transfoming to the new variables such as 3x-6 and 9y+2.

4. Suppose observations $y_i = 4 + 3\,x_i + \epsilon_i$, $i \le n$.

a. If the slope of regression of y on x is 1.78, what is the slope of the regression line of ϵ on x?

b. If the intercept of the regression line of y on x is 5.2 what is the intercept of the regression line of ϵ on x?

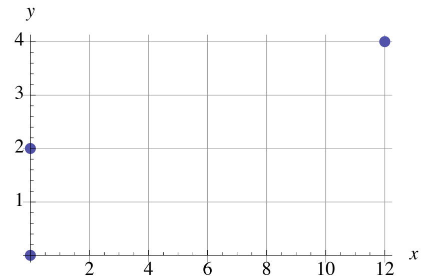
c. If the errors ϵ_i are actually random variables with $E\,\epsilon_i = 0$

$E\,\hat{\beta}_0 =$

$E\,\hat{\beta}_1 =$

MLR

5. For (x, y) data



x	y
0	0
0	2
12	4

a. Determine the *design matrix* (call it xx) of a matrix re-formulation of LR of y on x.

b. PseudoInverse[xx] = $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{24} & -\frac{1}{24} & \frac{1}{12} \end{pmatrix}$. Use this to determine $\hat{\beta}_0$, $\hat{\beta}_1$. See that they agree with what you get by eye.

c. Determine fitted values $\hat{y}$ by feeding your $\hat{\beta}_0$, $\hat{\beta}_1$ from (b) into xx and see that they agree with what you get by eye.

d. Calculate $s[\text{resid}]^2$ from the residuals determined by eye and reduce.

e. Inverse[Transpose[xx].xx] $\frac{n-1}{n-d} s[\text{resid}]^2 = \begin{pmatrix} \frac{3}{2} & -\frac{1}{8} \\ -\frac{1}{8} & \frac{1}{32} \end{pmatrix}$. Use this to determine 95% t-based CI for each of β_0, β_1. *t-based CI are applicable if it is assumed that observations y arise from a model y = xx $\beta + \epsilon$ whose errors ϵ_i are idependent samples from $N(0, \sigma^2)$ for some unknown $\sigma > 0$).

f. In (e) what are n and d?

g. Set up a design matrix xxx for fitting a quadratic model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

What is n-d in this case, and what will the multiple correlation be (no calculation needed but explain it)?

h. For the fit by xxx what is the value of the multiple correlation (no calculation is required but explain it).

i. For the fit by xx, if the correlation is 0.8 (it is not) what is the fraction of $\overline{y^2}-\overline{y}^2$ explained by regression on x?

j. Is it ever possible, in any problem, that the multiple correlation using a richer design (i.e. including all the variables of a sub-design obtained by removing some independent variables) is smaller than that of the sub- design?

k. Can multiple correlation ever be negative?

6. Data and design

		temp	hardness
y_i	1	$x_{1\,i}$	$x_{2\,i}$
4.6	1	2.3	15.7
5.2	1	3.6	12.8
4.1	1	2.8	14.7
6.3	1	5.1	15.9
7.7	1	5.2	17.5
6.3	1	2.9	13.8
3.7	1	1.6	10.4

a. $\hat{\beta} = \{3.61, 0.79, -0.07\}$. Estimate the mean response E y for inputs temp = 3, hardness = 14. Do not reduce.

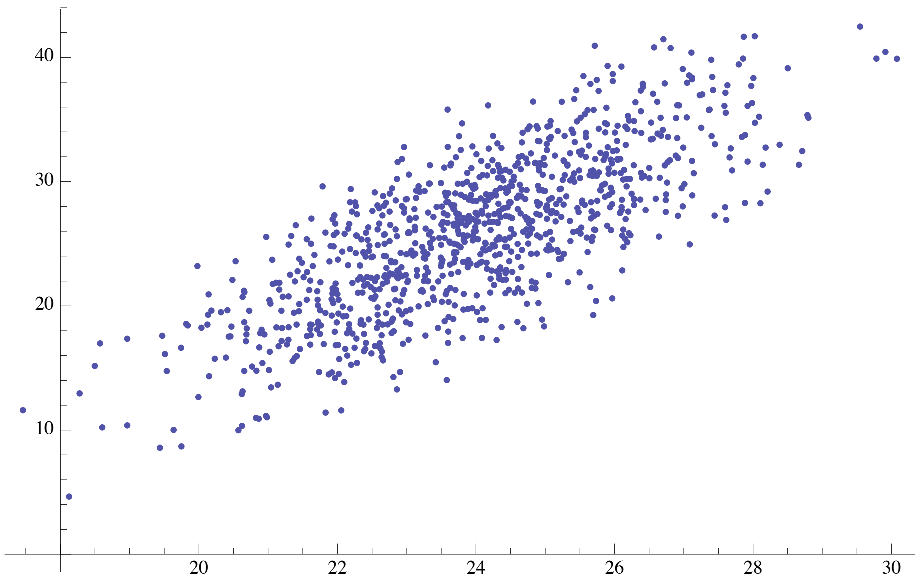
b. Extend the design (work above, do not reduce) to include all second order variables defined in terms of temp and hardness.

c. Suppose the multiple correlations of the original design, and the design as extended in (b), are 0.78 and 0.81 respectively (they are not). What fraction of the variation in y is explained by multiple regression for each of these designs? Would the incorporation of additional varibles in (b) be worthwhile in such case? Explain.

d. If the multiple correlation were 0.78 in the original design (it is not), what would be the new multiple correlation if we were to change temperature from F to C and then re-fit the model?

Elliptical plots

7. For the plot below, estimate by eye the requested quantities. Show how you do it.



a. sd x

b. sd y

c. line of regression

d. slope of regression

e. correlation