Lecture Outline 2-16-09

Part **A** places before you a variety of question types such as might be seen on the 30 minute quiz given in recitation Tuesday, February 17, 2009. You will be able to use notes for this quiz. The recitation is an important part of ongoing preparations for Exam 2 on Tuesday, March 2 (that exam will be closed book, no notes).

Part **B** introduces the basic ideas of multiple linear regression (MLR). The appropriate readings are from chapter 29 of your book. They are found on the DVD provided with the book. My rollout of these ideas will not depend on the book but you may find it helpful to read what the book has to say about MLR.

Part B. In multiple linear regression we attempt to explain or predict "dependent" variable y in terms of a linear expression in several "so-called independent" variables x1, ..., xd. For instance, we may seek a least squares "fit" of y to a model

$$y = b_0 + b_1\, x + b_2\, x^2 \quad \text{(quadratic in x, but linear in the terms x, } x^2\text{)}$$

e.g. calories = const + const time + const time^2

$$y = b_0 + b_1\, x_1 + b_2\, x_2 \quad \text{(linear in two variables } x_1, x_2\text{)}$$

e.g. calories = const + const time + const weight

Elliptical plots play much the same role in models with many variables as well.

a. r_{MLR} is the multiple correlation defined as the ordinary correlation between y-scores and their predicted scores from the x-scores. As it happens, for the straight line linear regression we have been working with $r_{MLR} = |r|$.

b. r^2_{MLR} is the fraction of s_y^2 explained by multiple linear regression on all of the x-variables specified in the model taken together.

c. Multiple linear regression fits that model which minimizes the sum of squares of discrepancies between the y-values and any proposed fit.

d. For elliptical plots vertical means lie on regression, normal with sd = $\sqrt{1 - r^2_{MLR}}\ s_y$.

Part A.

1-12. **Table illustrating calculations of $\bar{x}$, $\bar{y}$, $\overline{x^2}$, $\overline{y^2}$, $\overline{xy}$.**
regrtable[time, calories]

x	y	x^2	y^2	xy
21.4	472	457.96	222 784	10 100.8
30.8	498	948.64	248 004	15 338.4
37.7	465	1421.29	216 225	17 530.5
33.5	456	1122.25	207 936	15 276.
32.8	423	1075.84	178 929	13 874.4
39.5	437	1560.25	190 969	17 261.5
22.8	508	519.84	258 064	11 582.4
34.1	431	1162.81	185 761	14 697.1
33.9	479	1149.21	229 441	16 238.1
43.8	454	1918.44	206 116	19 885.2
42.4	450	1797.76	202 500	19 080.
43.1	410	1857.61	168 100	17 671.
29.2	504	852.64	254 016	14 716.8
31.3	437	979.69	190 969	13 678.1
28.6	489	817.96	239 121	13 985.4
32.9	436	1082.41	190 096	14 344.4
30.6	480	936.36	230 400	14 688.
35.1	439	1232.01	192 721	15 408.9
33.	444	1089.	197 136	14 652.
43.7	408	1909.69	166 464	17 829.6
$\bar{x}$	$\bar{y}$	$\overline{x^2}$	$\overline{y^2}$	$\overline{xy}$
34.01	456.	1194.58	208 788.	15 391.9

Determine the following.

1. s_y.

2. r.

3. The fraction of s_y^2 accounted for by regression on x.

4. The slope of the naive line.

5. The slope of the regression line of y on x.

6. The slope of the regression line of x on y (which would apply if the variables were interchanged).

7. r[2 x - 4, 6 y + 2]

8. r[-x + 2, y - 6]

9. For an ELLIPTICAL plot having the above averages, the average calories for all subjects having time 36.

10. For an ELLIPTICAL plot having the above averages, the best (by least squares) prediction for calories for a student with time 36.

11. The independent variable.

12. The dependent variable.

13-21. A plot has r[x, y] = 0.9, s_x = 2, s_y = 5, $\overline{x}$ = 22, $\overline{y}$ = 54.

13. Determine r[y, x].

14. For points (x, y) on the regression line determine the numerical value of $\frac{y-\overline{y}}{x-\overline{x}}$.

15. For x = $\overline{x}$ + s_x the regression prediction of y is $\overline{y}$ + (?) s_y.

16. For x = 18 the regression prediction of y is?

17. Regression predictions (15), (16) are sometimes useful even if the plot is not elliptical. If the plot IS ELLIPTICAL what is the special nature of the plot of vertical strip averages?

18. If the plot is ELLIPTICAL what is the average y-score for all (x, y) pairs with x = 18?

19. If the plot is ELLIPTICAL what is the standard deviation of y-scores for all (x, y) pairs with x = 18?

20. If the plot is elliptical, sketch the distribution of x, and the distribution of y.

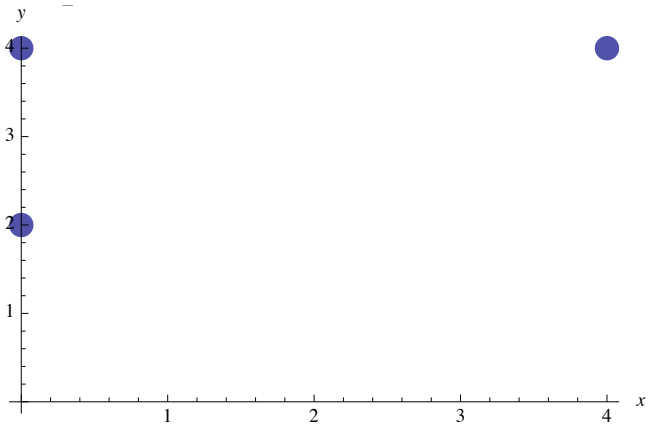
21. Draw a picture illustrating all of (18), (19), (20).

22-23. A plot has r[x, y] = 0.9, s_x = 2, s_y = 5, $\overline{x}$ = 22, $\overline{y}$ = 54. These questions refer not to the plot per-se, but to the distributional properties of $\overline{x}$, $\overline{y}$. In particular, WE DO NOT REQUIRE THAT THE PLOT BE ELLIPTICAL.

22. Using (14), if I tell you that the population mean of x is $\mu_x = 26$ what is the regression-based estimate for μ_x?

23. Give the 95% CI for the estimate (19) if n is large. (The plot need not be elliptical since the estimator (19) is dependent upon $\overline{x}$ and $\overline{y}$ which are approximately jointly normal distributed for large n.)

24. For the plot below, sketch the regression line for y on x (usual). Also, sketch the regression line for x on y. You want to think of flipping the axes (tilt your head?). It helps that these particular regression lines are also the plots of vertical strip averages (resp. horizontal strip averages) for these special but non elliptical plots.



25-29. Calculations.

25. For the plot just above calculate the slope of regression. Confirm it with what you see in the plot.

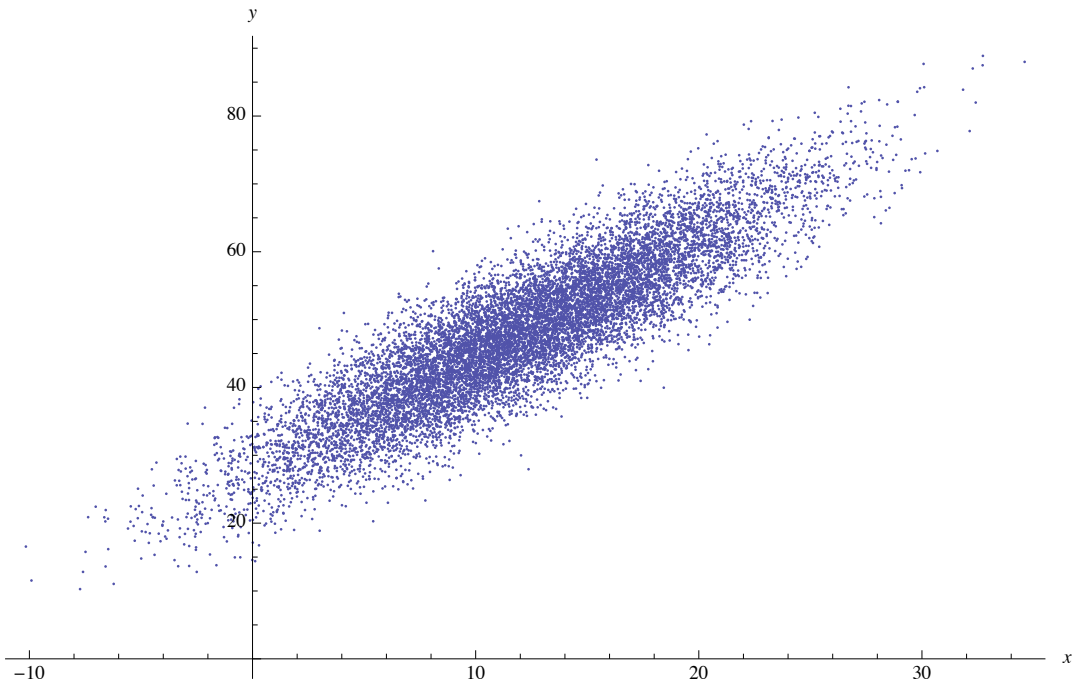
26. From your calculations (25) what is s_y?

27. Calculate s_e.

28. Verify that $r^2 = 1 - \frac{s_e^2}{s_y^2}$ (fraction of s_y^2 explained by regression on x).

29. Interestingly, the correlation of x with the fitted values is exactly equal to |r|. Verify that it is so in this case.

30-36. By eye, in the plot below, determine the requested quantities as best you can. On a quiz or exam you may have to choose between several proposed solutions of elements of this exercise.



30. Draw in the regression of y on x. Identify and label $\overline{x}$, $\overline{y}$, s_x, s_y (using 68% rule).

31. Sketch the bell curve distribution of x just above the x-axis and the bell curve distribution of y just left of the y-axis and tilted on its side. Identify s_x, s_y in these.

32. Sketch the naive line. Identify s_x, s_y in it.

33. Determine the correlation r. Show how you get it from the above.

34. Consistent with (33) what fraction of the variance s_y^2 is explained by regression?

35. On two vertical lines, sketch and label the distribution of y-scores for the given x.