Lecture 13 (Ch 3)

(multiple regression)

So far, simple linear regression

1 predictor, x ↳ in parameters $y = \alpha + \beta_1 x + \beta_2 x^2 + \cdots$

As argued before, this linearity is desirable, but not restrictive.

Now, multiple linear regression.

↳ Several (k) predictors: $x_1, x_2, \cdots, x_k$

E.g. $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_3 (x_1)^2 + \beta_4 (x_2)^{13} + \underbrace{\beta_5 x_1 x_2}_{\text{"Interaction term"}} + \cdots$

2nd variable/predictor, not 2nd case.

E.g.

y = Age at death, x_1 = income, x_2 = health

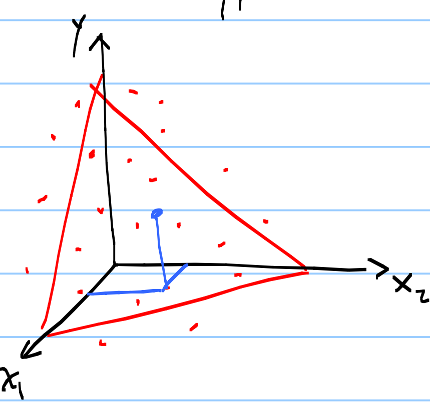
y = ICP, x_1 = blood flow, x_2 = blood pressure.

$y = \Delta Q$ (heat) $x_1 = m$ (mass) $x_2 = \Delta T$ (temper.) $\Delta Q = c\, \underbrace{m \Delta T}_{\text{interaction}}$

specific heat = regression coeff.

Geometry: Instead of a line, we have a hyper-surface

E.g. $y = \alpha + \beta_1 x_1 + \beta_2 x_2$



Meaning of β_i?

Average change in y, for every unit change in x_i,

⇒ IF all other x_i are held constant ⇒ No collinearity
⇒ AND IF there is no $(x_1 \cdot x_2)$ term ⇒ No interaction
} see below

Keep in mind that everytime you add a new term on the R.H.S. (e.g. a new predictor, a non-linear term, an interaction, or even a completely random variable) you increase the chances of overfitting, i.e. R^2 will increase (atleast, it will never decrease). How will you know what to include on the R.H.S., and what not to include?

See bottom of p.7 in last lecture.

How to estimate $\alpha, \beta_1, \beta_2, \cdots \beta_k$?

Same as before, ie. with OLS $\Longrightarrow \hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \cdots \hat{\beta}_k$ See hw

How to do ANOVA? Same, except there is now k, everywhere.

$$SST = SS_{expl.} + SS_{unexplained}$$

$$\sum_i (y_i - \bar{y})^2 \qquad \sum_i (\hat{y}_i - \bar{y})^2 \qquad \sum_i (y_i - \hat{y}_i)^2 \equiv SSE.$$

$$R^2 = \frac{SS_{expl.}}{SST}$$

$$R^2 = 1 - \frac{SSE}{SST}$$

FYI

$$R^2_{adj} = 1 - \frac{SSE/[n-(k+1)]}{SST/(n-1)}$$

$$= 1 - \frac{s_e^2}{s_y^2}$$

Recall $R^2 \to 1$ as model gets more complicated. R^2_{adj} attempts to fix that problem, but only partially. I.e. both R^2 and R^2_{adj} never decrease as the model gets more complex.

$$s_e = \sqrt{\frac{SSE}{\boxed{n-(k+1)}}} \quad \boxed{n-(k+1)} = df$$

One says that SSE has $df = n-(k+1)$. proof, below.

$k = \#$ of β's.

$k+1 =$ total $\#$ of parameters, α, β_i..

E.g.

$y = \alpha + \beta_1 x + \beta_2 x^2$, $k+1 = 3$

$y = \alpha + \beta_1 x + \beta_2 x^4$, $= 3$

$y = \alpha + \beta_1 x_1 + \beta_2 x_2$, $= 3$

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$, 4

Interaction

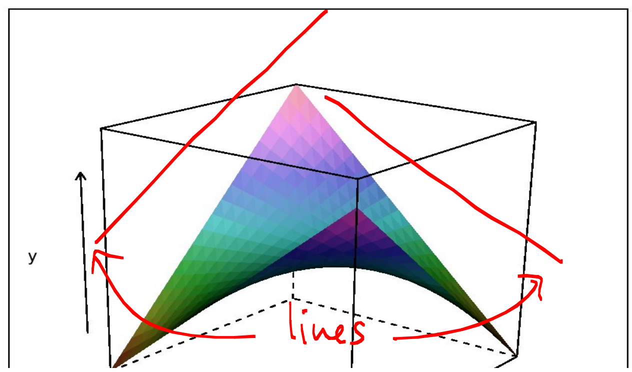
In multiple regression, because of the existence of multiple predictors, there are 2 issues that arise: Collinearity and Interactions.

$\Delta Q = c\, m\, \Delta T$

First, interaction.

Q: What does it look like?

$y = x_1 x_2$



Q: What are the consequences of an interaction term?

A: The effect of one predictor on y, depends on other predictor(s)!

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 = \alpha + \beta_1 x_1 + (\beta_2 + \beta_3 x_1) x_2$

It's like XOR in logic (unimportant).

Example:

Suppose in a certain problem involving y, x_1, x_2, we have found a good model to be $y = 1 + 2x_1 + 3x_2 + 4x_1x_2$. (Assuming there is no collinearity) how much does y change on avg. if x_2 changes 1 unit?

Not 3, not 2+3=5, not 3+4=7, ...

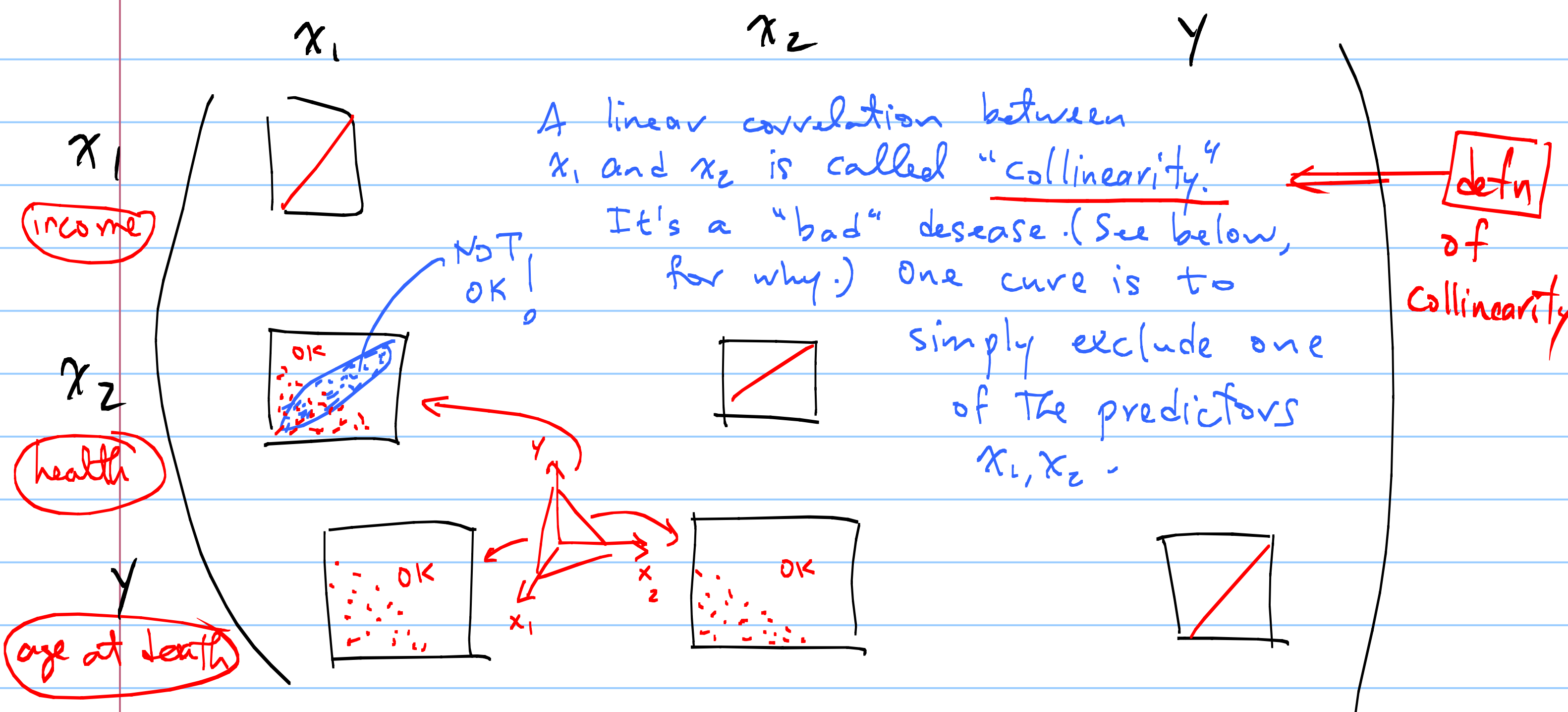
We simply cannot tell, because the answer depends on the specific value of x_2.

Collinearity

In multiple regression, in addition to interaction there is one more thing to worry about: collinearity.

Let's return to the first (important) step: Look at data!

Because there are multiple predictors, there is a matrix of scatterplots:



$\Rightarrow$ A Bad consequence of collinearity is that it renders the β's un-interpretable (as the avg. rate of change of y ...):

Ordinarily, in $y = \alpha + \beta_1 x_1 + \beta_2 x_2$
β_1 = avg. rate of change in y, for 1 unit change in x_1, IF x_2 IS HELD CONSTANT.

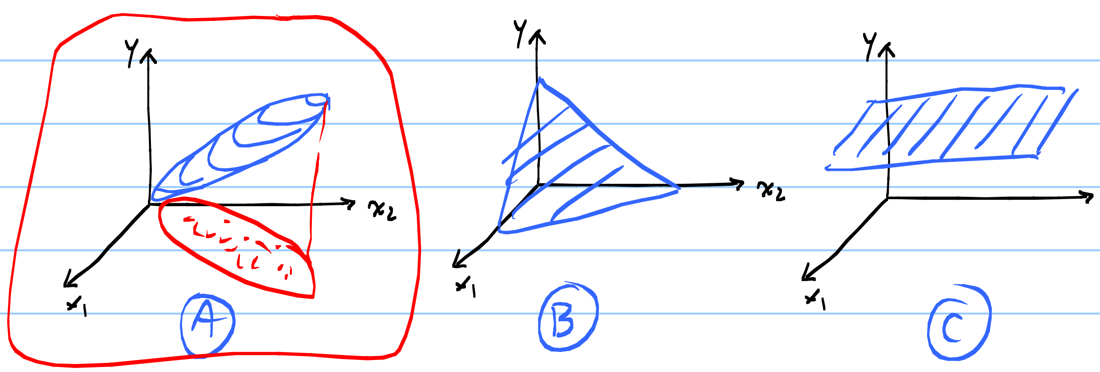
But if x_1 and x_2 are correlated, then one cannot hold one of them fixed.

In fact, in an example $\text{age} = \alpha + \beta_1(\text{health}) + \beta_2(\text{income})$
I once got a value of β_1 that was negative, in spite of the positive association displayed in the scatterplot of age vs. health. The culprit was collinearity.

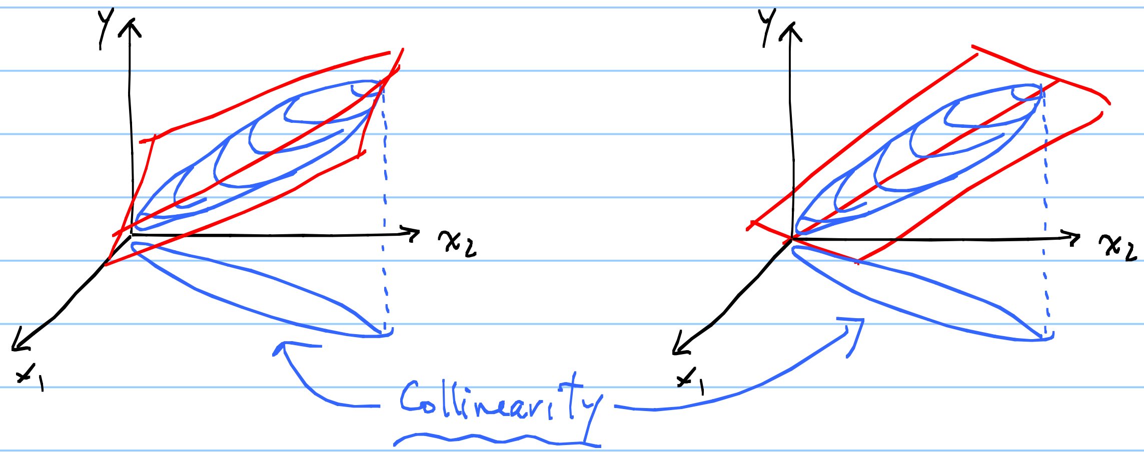
⇒ Another Bad consequence of collinearity is That it effectively reduces The amount of information in The data, which, in turn, leads to more uncertain estimates of The β's and predictions. We'll see That in Ch. 11.

⇒ Another Bad consequence is That it can also lead to overfitting. This is because the various predictors come with params to be estimated from data, but The various predictors are essentially carrying The same information, ie. There is effectively more params. Than data, hence overfitting can happen.

Q1: which of The following display collinearity?



Geometrically, the reason why the β's become uncertain and uninterpretable is that we are then trying to fit a plane through a cigar-shaped cloud in 3D, as opposed to a planar cloud.

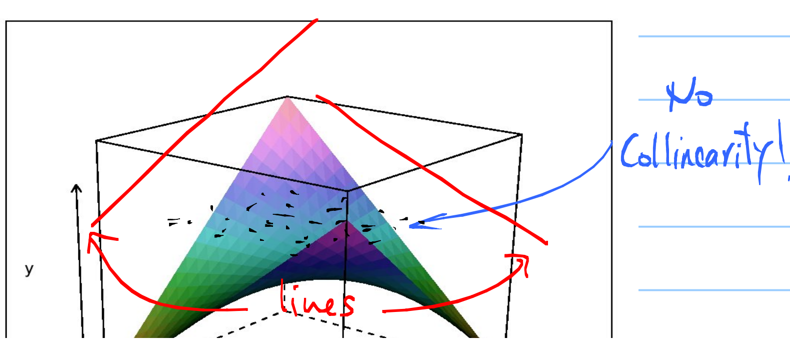


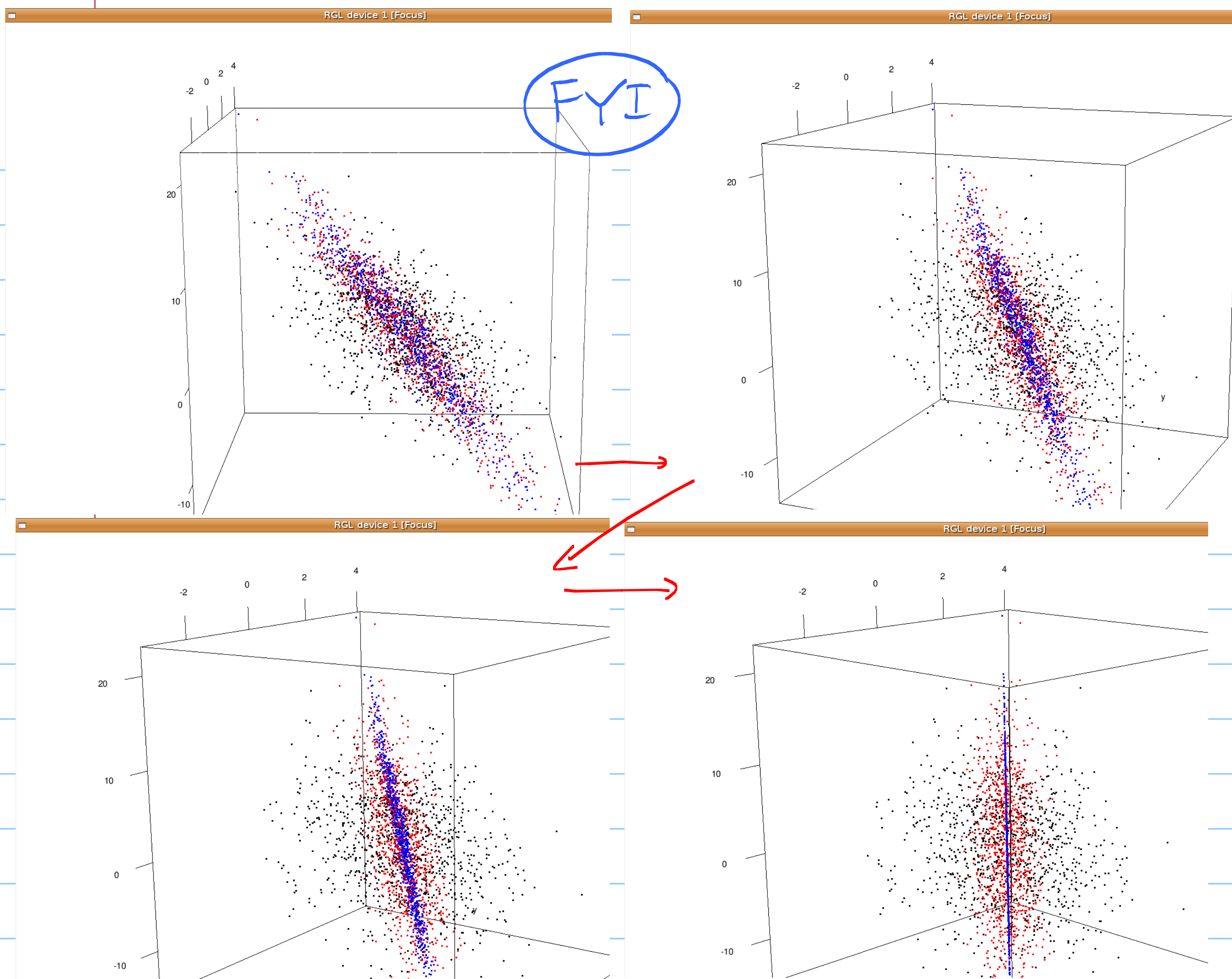
That is ambiguous! There are lots of planes one can fit through a cigar-shaped cloud in 3D. Of course, those different fits differ in their $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2$. That's why they become meaningless. You can also see that the predictions, $\hat{y}$, are affected by collinearity; however, note that the effect is mostly in their uncertainty. (More, in Ch. 11).

In summary, even though both interaction and collinearity make the β's un-interpretable, they are very different concepts.

collinearity $\neq$ interaction.

For example, if in a problem the data look like this, then we have interaction, but no collinearity.





For different levels of collinearity, the problem of uncertain β's and predictions can be qualitatively different.

For very little collinearity, there is a reasonably unique plain one can fit the black dots. For mild collinearity (red) there is no unique surface to fit the "cigar." For extreme collinearity (blue), the "fit" is a "vertical" surface.

Think about what this does to the predictions.

hw_lect 13-1

By R

The article "The Undrained Strength of Some Thawed Permafrost Soils" (Canadian Geotech. J., 1979: 420-427) contained the accompanying data on y shear strength of sandy soil (kPa), x1 depth (m), and x2 water content (%).

Obs	Depth	Content	Strength
1	8.9	31.5	14.7
2	36.6	27.0	48.0
3	36.8	25.9	25.6
4	6.1	39.1	10.0
5	6.9	39.2	16.0
6	6.9	38.3	16.8
7	7.3	33.9	20.7
8	8.4	33.8	38.8
9	6.5	27.9	16.9
10	8.0	33.1	27.0
11	4.5	26.3	16.0
12	9.9	37.0	24.9
13	2.9	34.6	7.3
14	2.0	36.4	12.8

a) Perform regression to predict y from x1, x2, x3 = x1^2, x4 = x2^2, and x5 = x1*x2; and write down the coefficients of the various terms.
b) Can you interpret the regression coeficients? Explain.
c) Compute R^2 and explain what it says about goodness-of-fit ("in English").
d) Compute s_e, and interpret ("in English").
e) Produce the residual plot (residuals vs. *predicted* y), and explain what it suggests, if any.
f) Now perform regression to predict y from x1 and x2 only.
g) Compute R^2 and explain what it says about goodness-of-fit.
h) Compare the above two R^2 values. Does the comparison suggest that at least one of the higher-order terms in the regression eqn provides useful information about strength?
i) Compute s_e for the model in part f, and compare it to that in part d. What do you conclude?

hw_lect 13-2

by R

```
Generate data on x1, x2, and y, such that

1) n (= sample size) = 100,
2) x1 and x2 are uncorrelated, and from a uniform distribution between 0 and 1,

a) Let y be given by y = 2 + 3 x1 + 4 x2 + error, where error is from a
normal distribution with mean = 0 and sigma = 0.5. Fit the model
y = alpha + beta1 x1 + beta2 x2 to the above data, and report R^2 and s_e.

b) Let y be given by y = 2 + 3 x1 + 4 x2 + 50 (x1 x2) + error, where error is
from a normal distribution with mean = 0 and sigma = 0.5. Fit the model
y = alpha + beta1 x1 + beta2 x2 to the above data, and report R^2 and s_e.

c) Fit the model y = alpha + beta1 x1 + beta2 x2 + beta3 (x1 x2) to the data
from part b, and report R^2 and s_e.

d) Install the R package called "rgl" on your computer, by typing
install.packages("rgl",dep=T), and following the instructions. If you
have trouble with this, ask the TAs or I during office hours.
Then, at the R prompt, type
                           library(rgl)
followed by
                           plot3d(x1,x2,y)
The panel you will see is interactive. By holding down the left-button,
and moving the mouse around, you will be able to "turn" the figure around
in different ways. Have some fun with it, THEN based on what you see,
provide an explanation for why the quality (in terms of R2 and/or s_e)
of the fit in part c is better than that in part b.
```

skip This part if you have trouble installing The rgl package.

hw-lect13-3

For each of the data sets a) hw_3_dat1.txt and b) hw_3_dat2.txt, find the "best" (OLS) fit, and report R-squared and the standard deviation of the errors. Do not use some ad hoc criterion to determine what is the "best" fit. Instead, use your knowledge of regression to find the best fit, and explain in words why you think you have the best fit. Specifically, make sure you address 1) collinearity, 2) interaction, and 3) nonlinearity.

Do not do this one,

a) Read the data file transform_data.txt from the course website into R, and make a scatterplot of y versus x. Clearly, the relationship is nonlinear and monotonic. I can tell you that a good transformation that linearizes the relationship is to take the sqrt of both and x and y. Make a scatterplot of the transformed data.

b) Perform regression on the transformed data, and overlay the regression line on the scatterplot of the transformed data in part a)

c) Fit a regression model of the form y = alpha + beta_1 sqrt(x) + beta_2 x to the original (untransformed data).

d) In a clicker question I claimed that these two models are essentially equivalent. To check that, let's see if they make similar predictions. Make a scatterplot to compare their predictions. Just keep in mind that the second model predicts y, but the first model predicts sqrt(y).