Lecture 17 (Ch. 3 – end)

Regression on transformed data, and polynomial regression

e.g. $\sqrt{y} = \alpha + \beta \log x$, $y = \alpha + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$

are some of the most useful things you'll learn here.

But there is more!

So far, simple linear regression

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

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

Today, 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

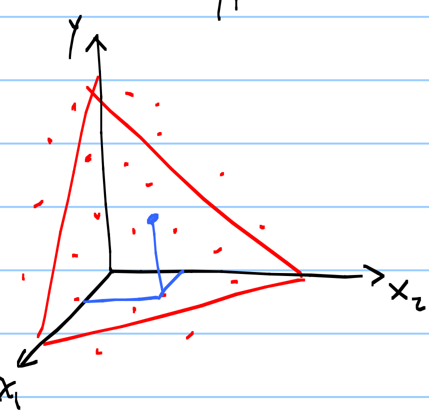
y = ICP , x_1 = blood flow, x_2 = blood pressure.

$y = \Delta Q$ (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. ⇐

⇒ AND IF there is no interaction term. ⇐ (see below).

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

Same as before, i.e. with OLS $\Longrightarrow \hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_k$ 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, later.

k = # of β's.

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

E.g.

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

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

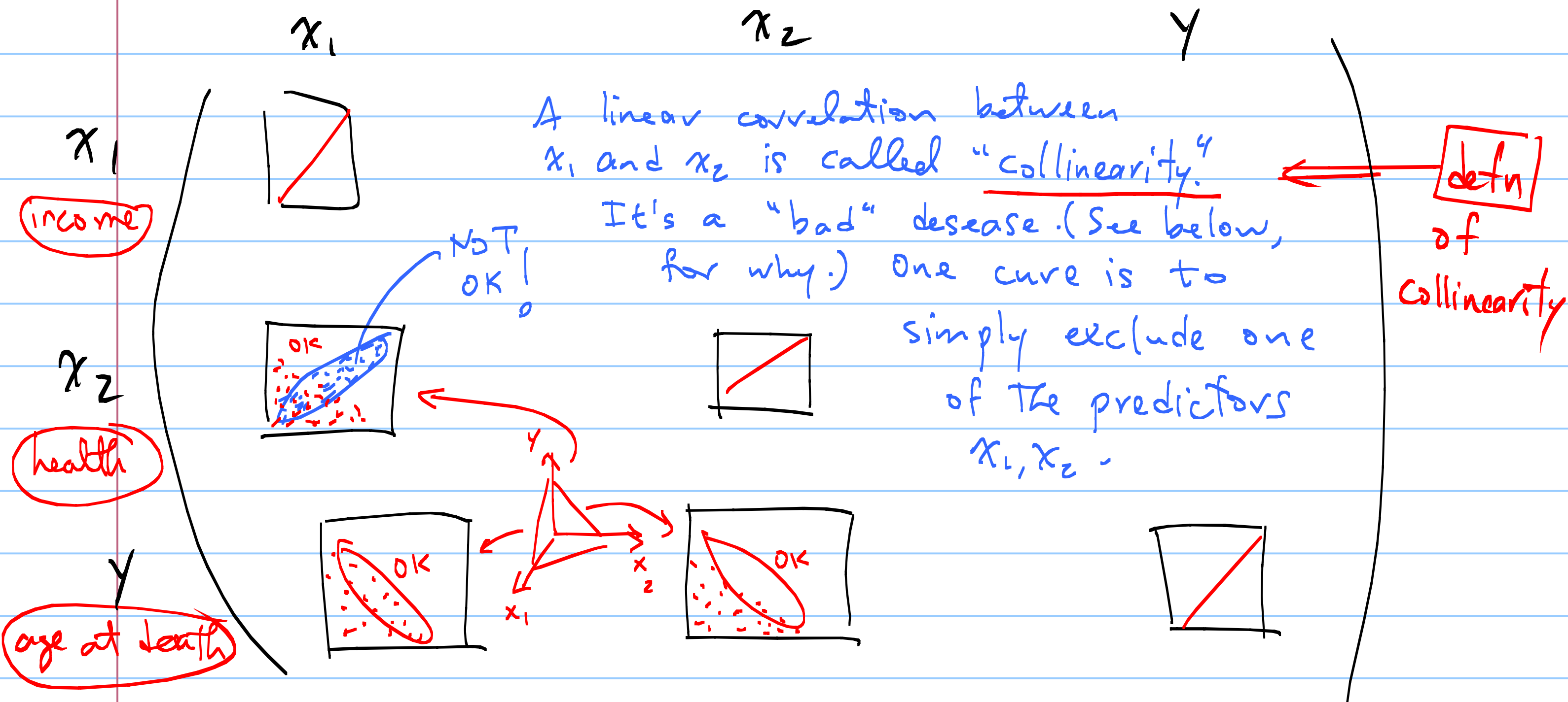
$y = \alpha + \beta_1 x_1 + \beta_2 x_2$, $= 3$

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

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

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

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



A linear correlation between x_1 and x_2 is called "collinearity." It's a "bad" desease. (See below, for why.) One cure is to simply exclude one of the predictors x_1, x_2.

defn of collinearity

⟹ A $\boxed{\text{consequence}}$ 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.

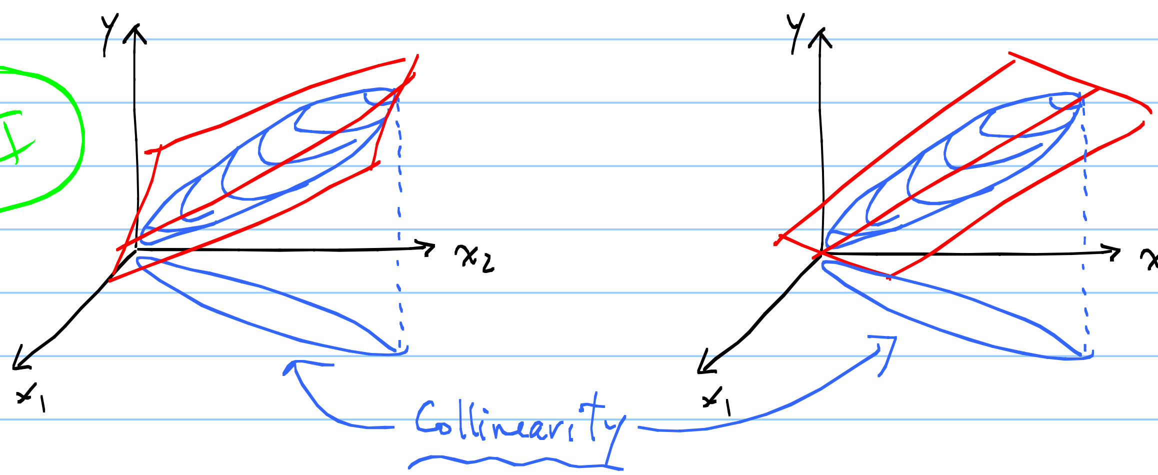
In fact, in an example $\text{age} = \alpha + \beta_1(\text{health}) + \beta_2(\text{income})$
I once got a value of β_1 that was negative, in spite of the positive association displayed in the scatterplot of age vs. health. The culprit was collinearity.

⇒ Another 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 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, i.e. there is effectively more params than data, hence overfitting can happen.

FYI

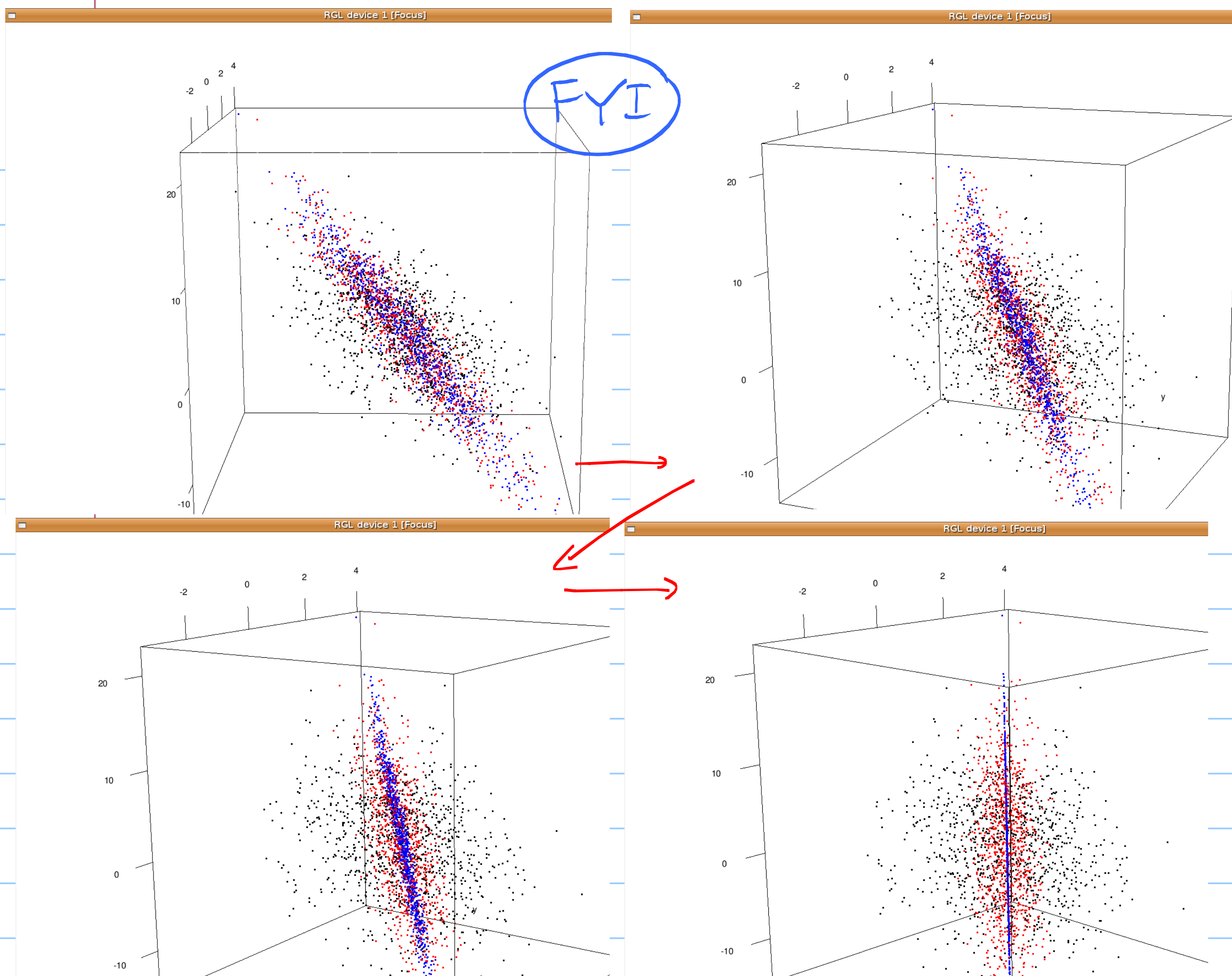
Geometrically, the reason why the β's become uncertain and uninterpretable is that we are then trying to fit a plane through a cigar-shaped cloud in 3D, as opposed to a planar cloud.



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).

FYI



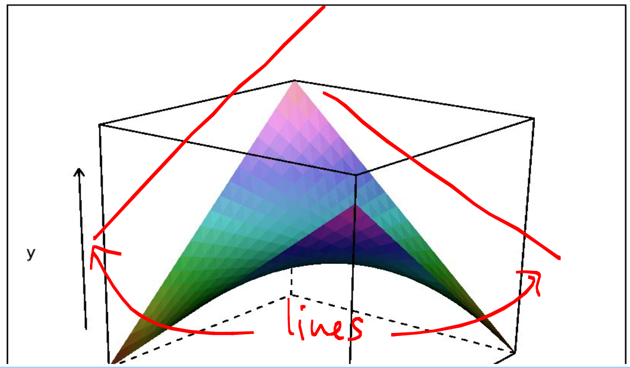
For different levels of collinearity, the problem of uncertain β's and predictions can be qualitatively different.
For very little collinearity, there is a reasonably unique plaine one can fit the black dots. For mild collinearity (red) there is no unique surface to fit the "cigar." For extreme collinearity (blue), the "fit" is a "vertical" surface.
Think about what this does to the predictions.

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

Now, interaction.

Q What does it look like?

$y = x_1 x_2 \Longrightarrow$



Q What does an interaction term mean? (XOR)

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

I.e. the effect on y, of changing x_2, depends on x_1

Q What are the consequences of an interaction term?

It makes the regression coefficients un-interpretable. Why? A

In Summary, collinearity and interaction, both make the β's uninterpretable, but for very different reasons.

collinearity $\neq$ interaction.

Q1: Suppose our data actually look like the saddle shown above. Is there collinearity in our data?

A) Yes $\boxed{\text{B) No}}$ C) The question makes no sense!

look at the x-y plain. Do you see a linear relation?

Warning:

Keep in mind that every time you add a new term on the R.H.S. (whether it's a new variable, or a nonlinear term, or an interaction) you increase the chances of overfitting the data ($R^2 \to 1$). regular or adjusted.

Later we will deal with the question of what's the "best" model (ie. how many terms, and which terms should be kept on the R.H.S.)

skip, for now

C-table

All of Ch. 3 has been about understanding the relationship between several continuous variables. What about categ. vars?

For categorical data the relationship is best captured through the contingency table: C-table

aka confusion matrix.

Data

x	y
Yes	High
Yes	Low
Yes	High
No	High
Yes	High
No	Low
No	Low
perhaps	medium
perhaps	Low

x \ y	High	Low	Medium
Yes	3	1	0
No	1	2	0
perhaps	0	~~0~~ 1	1

$\exists$ Relationship between x and y.

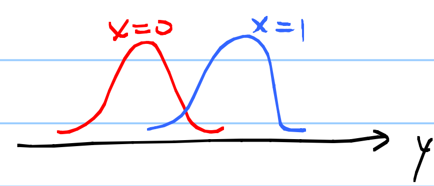
Maybe "positive" or "negative".

3 variables $x, y, z \implies$ Cube = Set of Contingency Tables.

Q: What about mixed (discrete and cont)?

E.g. $\begin{cases} x = 0, 1 \\ y = \text{Continuous} \end{cases}$

A: Conditional histograms $\Rightarrow$



hw-lect17-1 (Revised 3.36). No serious computation is necessary for this problem. Use the printout in the problem as much as possible.

An experiment carried out to study the effect of the mole contents of cobalt (x1) and the calcination temperature (x2) on the surface area of an iron-cobalt hydroxide catalyst (y) resulted in the following data("Structural Changes and Surface Properties of CoxFe3ÂxO4 Spinels," J. of Chemical Tech. and Biotech., 1994: 161-170):

x1: .6 .6 .6 .6 .6 1.0 1.0 1.0 1.0 1.0 2.6 2.6 2.6 2.6
x2: 200 250 400 500 600 200 250 400 500 600 200 250 400 500
y: 90.6 82.7 58.7 43.2 25.0 127.1 112.3 19.6 17.8 9.1 53.1 52.0 43.4 42.4

x1: 2.6 2.8 2.8 2.8 2.8 2.8
x2: 600 200 250 400 500 600
y: 31.6 40.9 37.9 27.5 27.3 19.0

A request to the SAS package to fit a+b1*x1+b2*x2+b3*x3, where x3 = x1*x2 (an interaction predictor), yielded the following output:
Dependent Variable: SURFAREA

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	15223.52829	5074.50943	18.924	0.0001
Error	16	4290.53971	268.15873		
Total	19	19514.06800			

Root MSE 16.37555 R-square 0.7801
Dep Mean 48.06000 Adj R-sq 0.7389
C.V. 34.07314

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	185.485740	21.19747682	8.750	0.0001
COBCON	1	-45.969466	10.61201173	-4.332	0.0005
TEMP	1	-0.301503	0.05074421	-5.942	0.0001
CONTEMP	1	0.088801	0.02540388	3.496	0.0030

a) Interpret the value of the coefficient of multiple determination (R^2).

b) Predict the value of surface area when cobalt content is 2.6 and temperature is 250, and calculate the value of the corresponding residual.

c) Since b1 is about -46.0, is it legitimate to conclude that if cobalt content increases by 1 unit while the values of the other predictors remain fixed, surface area can be expected to decrease by roughly 46 units? Explain your reasoning. Hint: think about collinearity and interaction.

d) What is the typical error about the regression surface? First, find this quantity in the printout, and then reproduce it using the value of SSE given in the printout.

e) Assess collinearity! By computer. For this question, you will have to enter the data into R.

By R.

hw_lect 17-2 (revised 3.37) 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 17-3

For each of the data sets a) hw_3_dat1.txt and b) hw_3_dat2.txt, find the "best" least-square 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.

hw_lect17-4

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.

Ignore!

Consider fitting a model $y_i = \beta x_{1i} x_{2i} + \epsilon_i$, $i=1,\dots,n$.