Lecture 14 (Ch. 3-end)

multiple regression
(3.5 & 11.4)

So far, 1 predictor x

"Simple regression": $y = \alpha + \beta x + \epsilon$

"polynomial regression": $y = \alpha + \beta_1 x + \beta_2 x^2 + \cdots + \epsilon$

Both of these models fall under linear regression, because they are linear in the parameters alpha, beta_1, beta_2, ... This linearity is desirable, because it leads to a system of linear eqns (linear in the parameters alpha, beta_1, beta_2, ...), and so, it guarantees a unique soln. It is not restrictive because the model can be as nonlinear (in x) as you would like it to be.
Even the next topic ("multiple regression") still falls under linear regression, and for the same reason.

multiple predictors $x_1, x_2, \ldots$

Now, multiple linear regression (Several new concepts).

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

"response" 2^{nd} variable/predictor, not 2^{nd} case.

E.g.

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

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

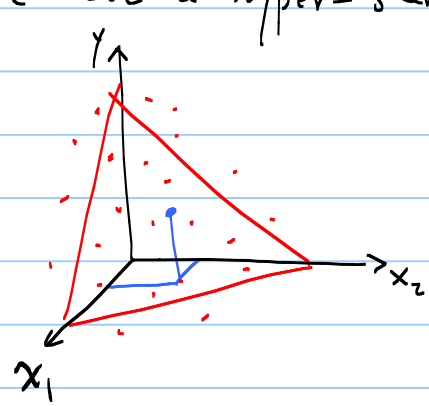
$y = \Delta Q$ (heat) $x_1 = m$ (mass) $x_2 = \Delta T$ (temper.) $\Delta Q = c\, \underbrace{m\, \Delta T}_{\text{interaction}}$

specific heat = regression coeff.

Pattern Recognition models are generally multiple regression.

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

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



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

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

How to do ANOVA?

Same as before, because the decomposition depends on y's, not x's. The only "new" thing is the count of the x's:

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

$$SST = \sum_i (y_i - \bar{y})^2 \qquad SS_{expl.} = \sum_i (\hat{y}_i - \bar{y})^2 \qquad SS_{unexplained} = \sum_i (y_i - \hat{y}_i)^2 \equiv SSE.$$

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

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

Keep in mind that everytime you add a 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 (never decrease).

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.

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

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

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

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

What determines the complexity of a model is $k+1$, not the nonlinearity of the terms.

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

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.

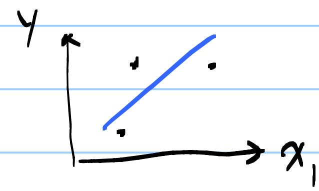
FYI

Read!

Overfitting in multiple regression

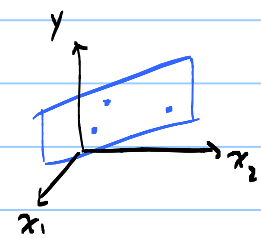
We know that one can overfit data on x and y, if one uses a high-order polynomial in polynomial regression. Recall that the main reason overfitting can happen is because such a regression model will have a lot of parameters, which in turn allow the fit to be more curvy/nonlinear, thereby going through every case.

In multiple regression there is yet another way that overfitting can happen even without high-order (nonlinear or polynomial) terms in the model.



Consider 3 cases on y and x_1.
A model like y = alpha + beta x_1, (a line) cannot over fit that data.

But a model like y = alpha + beta_1 x_1 + beta_2 x_2 (a plain) overfits completely. To see why, visualize the geometry of the problem; the reason for the complete overfit is that now the 3 cases are in 3D (not 2D), and there is always a plain that goes through any 3 points exactly.



Note that the additional predictor x_2 can even be completely unrelated to y; it can even be just random numbers! In other words, by arbitrarily making the space big (by adding another predictor, even a useless one), we opened up the possibility of overfitting. So one can overfit even without any non-linear (e.g. quadratic, cubic, ...) terms.

You way think this is happening only because we are dealing with 3 cases here. But even with more cases, one can still overfit by simply including more (even random) predictors in the model. In general, if there are significantly more parameters/predictors than cases, then overfitting may happen. Furthermore, this overfitting problem is not specific to regression; All models can overfit ifwhen they are too large. AI/Machine Learning students, watch out!

Interpretation of β's

It looks like doing multiple regression is easy.
And it is: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \cdots$

lm(y~ $x_1 + x_2 +$ I(x_1^2) + I($x_1 : x_2$) + ...

But recall that regression is good not only for making predictions, but also for understanding what's going on in the problem. And for that, we interpret the β's.

E.g. if $y = \alpha + 2.3\, x_1 - 2.1\, x_2$, then we can say that on the avg. y decreases by 2.1 units when x_2 increases by 1 unit.

BUT there are 2 situations in which the β's cannot be interpreted at all; they are associated with 2 concepts that we cover in this lecture: Collinearity and Interaction.

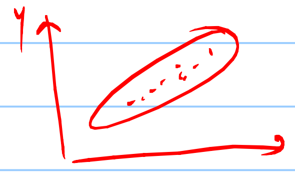
Here is a real example:

Data on
y = life span
x_1 = health
x_2 = wealth

$$\Rightarrow \quad y = \alpha + 2.3\, x_1 - 2.1\, x_2$$

- does y change +2.3 units when x_1 increases by 1 unit?
- does y change -2.1 units when x_2 increases by 1 unit?

How can this sign be right, when in fact there is a positive association between y, x_2



?!

$\therefore$ Something is "wrong" with our interpretation of β!

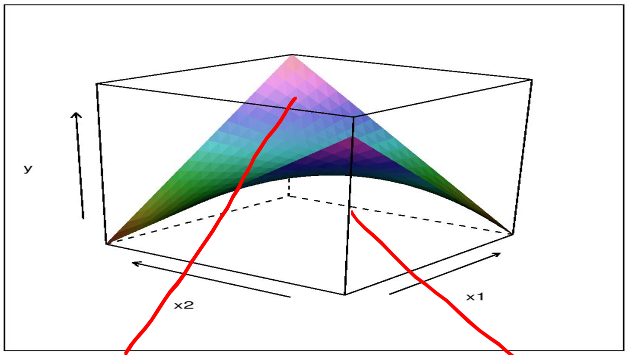
What?

Interaction

Let's deal with interaction, first, because it's easier.

Q: What does it look like?

$y = x_1 x_2$ ⟹



lines

FYI like XOR in logic

Q: What are the consequences of an interaction term?

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

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 = \alpha + \beta_1 x_1 + \underbrace{(\beta_2 + \beta_3 x_1)}_{\text{The effect of } x_2 \text{ on } y \text{ depends on } x_1} x_2$$

For this reason, the β's are uninterpretable, i.e. they don't tell you how much y changes when x is changed by 1 unit.

Guidance

Q: How do we know when to include an interaction term?

A: 1) Look at scatterplots for signs of a saddle surface.

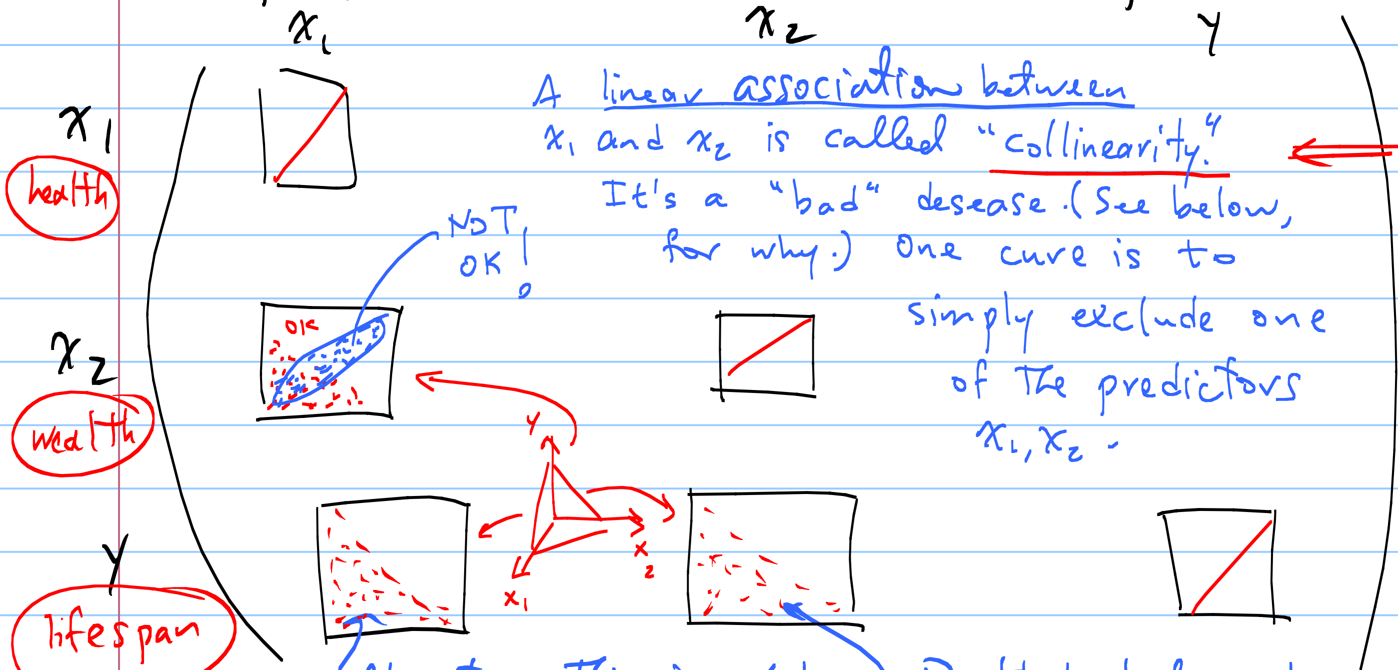
2) Include it, and see if it significantly improves R^2.

3) Ultimately, if it improves predictions, then include it! But always be mindful of overfitting.

4) Look at residual plots for signs of a saddle [Skip this quarter]

5) In Ch. 11 we will learn another way of deciding whether or not an interaction exists.

Let's return to The first (important) step: Look at data!
multiple predictors $\Longrightarrow$ matrix of scatterplots:



A linear association 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

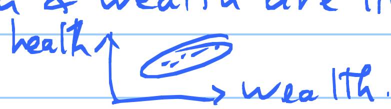
Almost anyThing is OK here. Don't look for a linear relation. In fact, if you get a high-corr. linear pattern in one of Them, Then, you probably don't need The other one at all! See how.

$\Rightarrow$ A 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 The lifespan $= \alpha + 2.3\,(\text{health}) - 2.1\,(\text{wealth})$ example, The coeffs 2.3 and -2.1 cannot be interpreted as rates of change because health & wealth are linearly associated, i.e. There is collinearity.

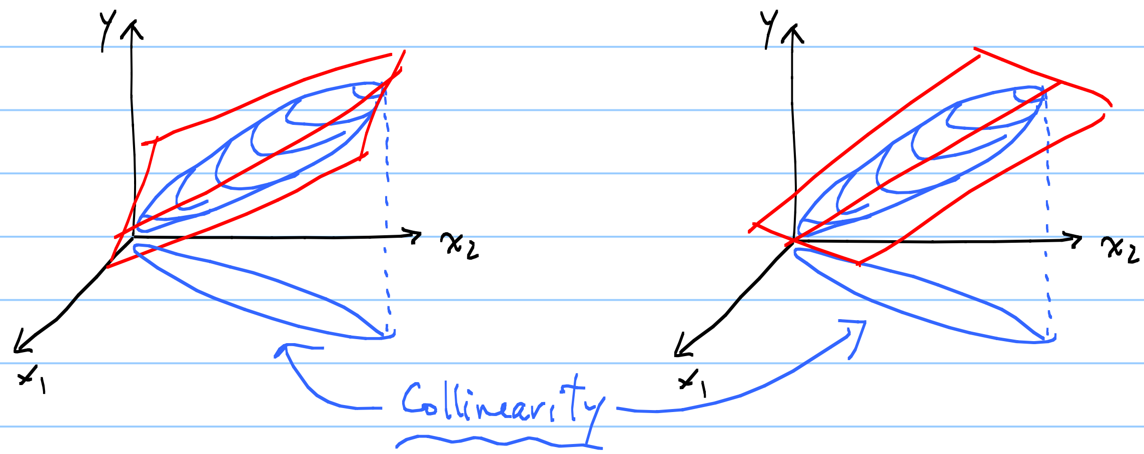


Guidance: when you suspect extreme collinearity, Then just drop one of The predictors from The model.
In less severe cases, ask a statistican!
Or look-up "principal component regression."

FYI

when $\exists$ 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).

$\Rightarrow$ Another bad consequence of coll. 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.

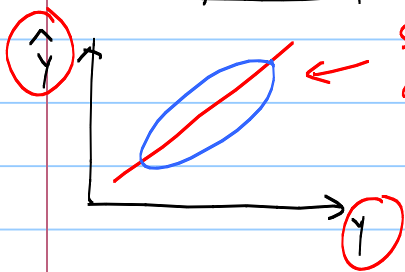
$\Rightarrow$ Another bad consequence of coll. is that it can lead to overfitting. This is because the various predictors come with params to be estimated from data, but the various predictors are essentially carrying the same information, i.e. there is effectively more params than data, hence overfitting can happen.

residual plot

FYI only

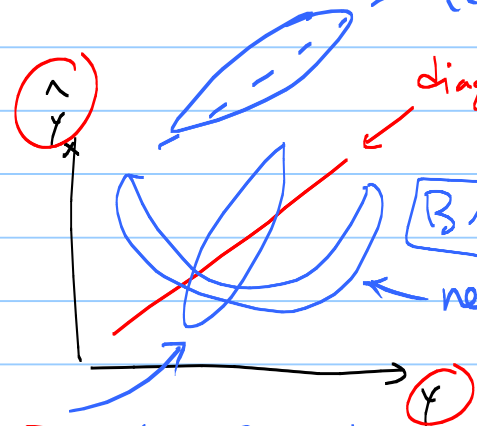
One last Thing in regression (until Ch. 11): The visual assessment of fit quality.

Plot $\hat{y}$ vs. y :



Random, Symmetric about diagonal

GOOD



biased predictions (easy to fix)

diagonal

BAD

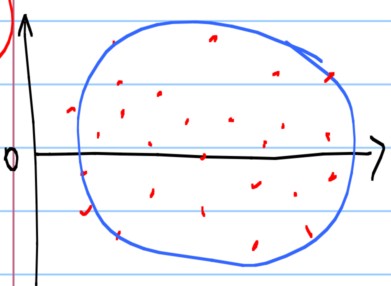
need a quadratic model.

under/over forecasting small/large y.

Residual Plot : (Mona Lisa)

same

errors (last column of "Table" before)



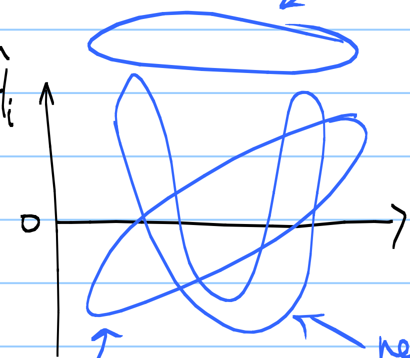
$y_i - \hat{y}_i$

better

$\tilde{y}_i$ or y_i

not recommended

Random, symmetric about error = 0 line



$y_i - \hat{y}_i$

$\hat{y}_i$

biased predictions

need a quadratic model.

under/over forecasting small/large y.

GOOD Model/fit

∴ Nothing more to do

BAD Model/fit

Things to do:

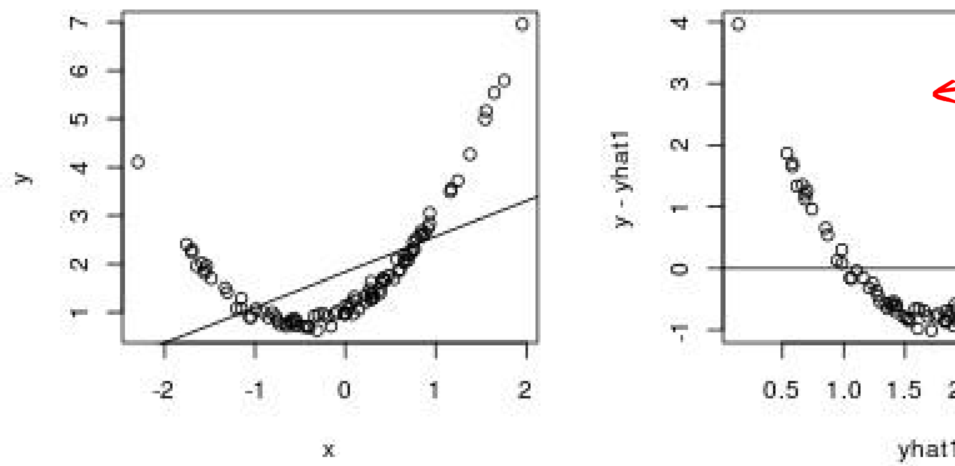
1) Transform data, or

2) fit diff. model (e.g. polynomial or interaction...)

FYI only

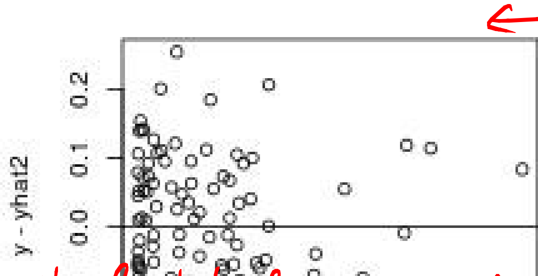
```
n = 100
    x = rnorm(n,0,1)
    y = 1 + x + x^2 + rnorm(n,0,0.1)
    lm.1 = lm(y ~ x)
    yhat1 = predict(lm.1)
    lm.2 = lm(y ~ x + I(x^2))
    yhat2 = predict(lm.2)

    par(mfrow=c(2,2))
    plot(x,y)
    abline(lm.1)
    plot(yhat1, y-yhat1)  ; abline(h=0)
    plot(yhat2, y-yhat2)  ; abline(h=0)
```



residual plot for $y = \alpha + \beta x$

residual plot for $y = \alpha + \beta_1 x + \beta_2 x^2$



The residual plot for $y = \alpha + \beta x$ clearly shows that $y = \alpha + \beta x$ is not a good fit to the data; because it shows a nonlinear pattern, we try $y = \alpha + \beta x + \beta x^2$ and then we see that the residual plot looks good (ie. shows <u>no</u> pattern). This may seem trivial in this case, because you can see that you need $y = \alpha + \beta x + \beta x^2$ from the scatterplot of the data, to begin with! But even in cases where you can<u>not</u> make a scatterplot of data (e.g. in multiple regression) you <u>can</u> still make a residual plot.

In closing Ch. 3:

	best prediction	± uncertainty
Before regression:	$\bar{y}$	s_y
After regression:	$\hat{y}$	$s_e = \sqrt{\dfrac{SSE}{\boxed{n-(k+1)}}}$

$n-(k+1)$ = df of SSE

k = # of β's in

$$\hat{y} = \hat{\alpha} + \underbrace{\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2}_{\text{multiple predictors}} + \cdots \underbrace{\hat{\beta}_3 x_1^3}_{\text{nonlinear terms}} + \cdots + \underbrace{\hat{\beta}_4 x_1 x_2}_{\text{interaction terms}} + \cdots$$

Thanks to ANOVA, $s_e \le s_y$ (ie. regression reduces uncertainty)

Warning: Regression is a powerful tool, and so one can do serious damage with it. At the least, you must be aware of overfitting, interaction, and collinearity. There is a complex interplay between them, even though they are completely different concepts.

For example, even though both interaction and collinearity make the beta's un-interpretable, they are very different concepts: the former refers to a term in the regression model, or equivalently, a saddle surface in the 3d scatterplot. Either way, it involves x1, x2, and y. By contrast, the latter refers to (or is defined as) a linear relationship between the predictors x1 and x2, nothing to do with y. And yet, they both lead to uninterpretable beta's.

One may think that these 2 issues are not important if/when you are not interested in interpreting the beta's (e.g. if/when you are interested in making predictions only). And there is some truth to that thinking. But, these issues can affect your predictions, too! For example, both of them can lead to overfitting. Interaction can lead to overfitting because it introduces one more parameter into the model, allowing the model to twist and turn its way through more points. Collinearity can lead to overfitting because each of the predictors comes with its own beta in the model eqn, and yet, each of the two predictors is essentially carrying the same information. So, don't ignore the three concepts of overfitting, interaction, and collinearity.

Finally, and again, recall that multiple regression (even with polynomial terms), is still an example of linear regression, because the model y = ... is linear in the parameters. As a result of that linearity, when we take derivatives of SSE, we get a set of linear equations (again, linear in the parameters) that can be solved exactly, giving a unique solution. This is the advantage of linear models. Meanwhile, the models can be as nonlinear (in x) as you choose to make them by adding higher powers, or products, of the predictors. So, multiple regression is a win-win!

hw-lect14-1

Consider fitting the model $y_i = \alpha + \beta x_{1i} x_{2i} + \epsilon_i$, $i = 1, \ldots, n$ to data on x_1, x_2, y. Find the 2 normal equations of regression that one gets by differentiating $SSE = \sum_i^n (y_i - \alpha - \beta x_{1i} x_{2i})^2$ with respect to α and β.

Write the 2 equs in "bar" notation; but don't solve them for $\hat{\alpha}, \hat{\beta}$.

hw_lect14_2 (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"). skip part e)
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_lect14_3 (By R)
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 (like maximum R2) to determine what is the "best" model. Instead, use your knowledge of regression to find the best model, and explain in words why you think you have the best model. Specifically, make sure you address 1) collinearity, 2) interaction, and 3) nonlinearity.

hw_lect14_4
In a problem involving x1, x2, y, suppose we have selected x1 such that it has a very (very) strong linear association with y . For simplicity, suppose the linear pattern "goes through" the origin. Similarly for x2, ie. it's highly correlated with y. On the adjacent diagram, and inthe following order, draw the
a) 2d scatterplot in the x1-y plain.
b) 2d scatterplot in the x2-y plain.
c) 3d scatterplot, i.e. cloud of data in 3d. It may be hard to draw, but do your best in showing perspective.
d) 2d scatterplot in x1-x2 plain.
Hint:This requires only 3d visualization of the type we did in the lecture!