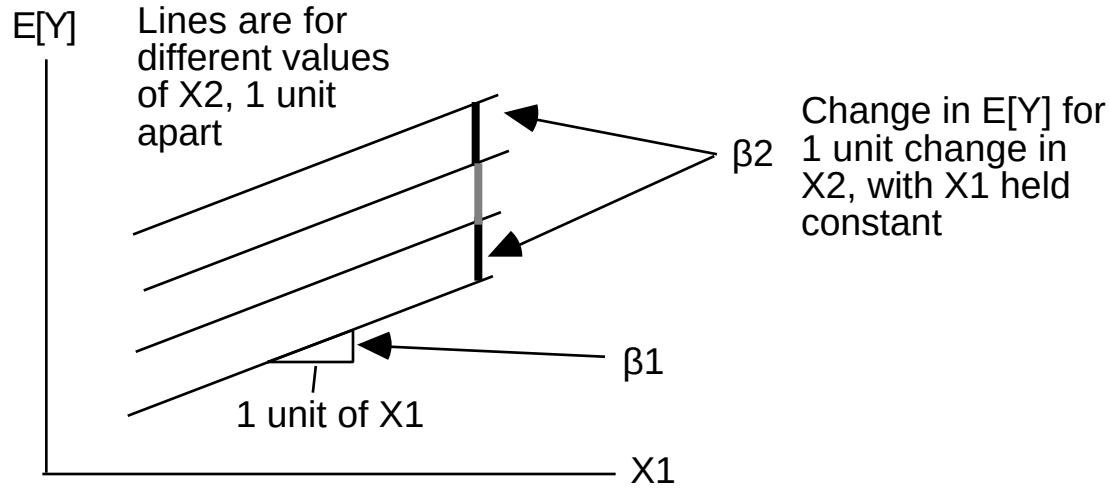
6.1 Multiple Regression Models

Why need multiple terms in model?: 1 term inadequate; too much imprecision. even if have control over (and measure) all factors, still important to use these factors in model

1st order models: linear in the variables (distinguish *variables* from *terms*) response surface a *plane* / *hyperplane* over the *variables.*

meaning of β's: $\beta 1 = \delta E[Y|X_1,X_2]/\delta X_1$ ("freezing" X_2])

for (X_1,X_2) model, can draw as



GENERAL Linear models: linear in the p *terms* (it might have been better if we used X's for *variables* and Z for *terms*), so several possibilities:-

- p-1 separate variables (so response surface is a hyperplane)
- a qualitative variable with more than k > 2 levels, represented as k-1 ***"indicator"*** terms (these terms are often called "***dummy***" variables"; "dummy variables" were originally questions inserted in questionnaires to see if respondent was being consistent or falling asleep or aware of hypothesis being studied)
- terms made from powers X^2, X^3, ... of a variable or products of two or more variables (so that response surface is curvilinear in the variables, but still a hyperplane in the terms)
- transformations of Y variable to have its relationship linear in the predictors
- combinations of the above
- LINEAR means linear in the coefficients (b's). Nothing "wrong" with models where predictor function cannot be made linear in the coefficients; but estimation (even if use the LS criterion) usually requires an iterative solution, since the estimating equations may no longer be linear in the coefficients to be fitted

6.2 General Linear Regression Model in Matrix Terms

- see text

6.3 Estimation of Regression Coefficients

LS criterion -> p estimating equations ("Normal Equations") :
(p linear equations in p β's, if no redundancy!)

ML estimator same as LS estimator (if ε's i.i.d) , since maximize $e^{-(1/2\sigma^2)Q}$

Note that ML estimator of σ^2 involves a divisor of n, rather than n-p, and so is biased.

6.4 Fitted Values and Residuals

Fitted values : substitute **b** for β in the model for E[**y**]: i.e.,

fitted E[**y**] = $\mathbf{X\,b} = \mathbf{X}\,[\,\mathbf{X'\,X})^{-1}\,\mathbf{X'\,y}\,] = \mathbf{H\,y}$,

where $\mathbf{H} = \mathbf{X}\,[\,\mathbf{X'\,X})^{-1}\,\mathbf{X'}$ is called "hat" matrix

Since X is a matrix of "constants", **b** is p linear combinations of the n y's, so
X b is also (another) p linear combinations of the y's, and ultimately -- if model is correct --
of the n (invisible) random ε ' s.

Residuals: substitute **b** for β in the model for E[**y**]: i.e.,

vector of n residuals $\mathbf{e} = y - \mathbf{X\,b} = \mathbf{y} - \mathbf{H\,y} = \mathbf{(\,I - H\,)\,y}$

where $\mathbf{H} = \mathbf{X}\,[\,\mathbf{X'\,X})^{-1}\,\mathbf{X'}$ is called "hat" matrix

Again, **e** is p linear combinations of the n y's, so ultimately -- if model is correct --
p linear combinations of the n (invisible) random ε ' s.

Notice in particular equation 6.31

$\text{var}[\mathbf{e}] = \sigma^2\,\mathbf{(\,I - H\,)}$,

where var is the sampling variance. i.e. what variation -- and covariation -- we would get if we say simulated infinitely many datasets from the same X, just changing the ε vector each time.

*Remember that H involves just the X's. Thus, contrary to what you might expect, the n fitted e's do not all have the same variance .. unlike the ε's . we will try to see this in a spreadsheet. Basically, the reason is that the e's at the edges of the **X** domain are more constrained (i.e. smaller) than those near the centre, since it is the y's at the edge that determine the fit more than the ones at the middle.*

6.5 ANOVA

As with simple linear model, except that df are now p-1 and n-p [1 df removed by fitting mean]

In this context, think of df = number of independent assessments of the variance being estimated.

"F test for regression relation" is a global test

i.e. $\beta_1 = \beta_2 = ... = \beta_{p-1} = 0$ vs. not all p-1 of the β 's are 0 (at least one non-zero)

As such, often not very useful.

Coefficient of Multiple determination (R^2):

$$R^2 = SS_{Regression}/SS_{total} = 1 - SS_{Residual}/SS_{total}$$

Useful to think of R^2 as square of correlation of the n pairs of (y, fitted y) values.

(cf exercise 6.26)

"Adjusted" R^2:

R^2 increases if use more terms in model, to point where $R^2 = 1$ if fit n terms to n y's.

Clearly, a "saturated" or "near-saturated" model that fits a particular n datapoints is unlikely to do quite so well with a new set of y's measured at the same x's. For example, what if there are n=2 y values corresponding to 2 values of X. A straight line model will predict the 2 y's perfectly, so $R^2 = 1$. But this line will not predict where 2 new y values at these same X values will be.

Imagine that for n y's, one created n-1 columns of X's from a random number table.. or the telephone book or the like (or use X_1= last digit of telephone no., X_2 = square of this, X_3=cube of it etc).

With X_1 alone, the R^2 will be positive, since you force the fitting algorithm to find the b_1 that maximizes the correlation of the (y,yhat) pairs [minimize the residuals].

I heard the following argument once, although it does not quite fit the formula. One variable, whether really related or not, will by chance explain roughly 1/nth of the total variance, two variables roughly 2/nths etc [if memory serves me right, the fractions are actually 1/(n-1), 2/(n-1), ...]. Thus, we can quantify how much of the "apparent" predictability is expected even if the variables are -- in the infinite world -- truly unrelated, but in the finite sample, seemingly related. This is the basis for the "adjusted" R^2.

Think of the numerator of an adjusted R^2 as $SS_{regression}$ minus "$SS_{regression}$ expected even if no real relationship with these p-1 variables", so that one candidate might be

$$\text{Hanley's "Adjusted" } R^2 = \frac{SS_{regression} - [p-1]/[n-1]\ SS_{total}}{SS_{total}}$$

Am relying on memory and hurried: my formula doesn't match formula 6.42. Why?

6.6 Inference re β's

Unbiasedness: E[**b**] = β , so **b** is unbiased.

Precision: Var[**b**] = $\Sigma_{\mathbf{b}} = \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}$

Structure of Var[**b**] : the price of estimating **b** from correlated vs uncorrelated X's

e.g. $\mathbf{X} = (\mathbf{1}\ \mathbf{x}_1\ \mathbf{x}_2)$ is an n × 3 matrix consisting of a column of 1's, of the n values of X_1, and the n values for X_2.

Assume w.l.o.g. that ave(X_1) = ave(X_2) = ... = 0, so that ave(X_i^2) = var(X_i) = V_i, then

$\mathbf{X}^T\mathbf{X}$ =

n	0	0
0	n V_1	n Cov_{12}
0	n Cov_{12}	n V_2

$(\mathbf{X}^T\mathbf{X})^{-1}$ =

1/n	0	0
0	$1/[n\{V_1(1-r_{12}^2)\}]$	$-r_{12} / [n\{SD_1SD_2(1-r_{12}^2)\}]$
0	$-r_{12} / [n\{SD_1SD_2(1-r_{12}^2)\}]$	$1/[n\{V_2(1-r_{12}^2)\}]$

If X_1, and X_2 are uncorrelated, so that $Cov_{12} = 0$, then

$(\mathbf{X}^T\mathbf{X})^{-1}$ =

1/n	0	0
0	$1/[nV_1]$	0
0	0	$1/[nV_2]$

with diagonal entries like those in simple linear regression on X_1 and X_2 separately. i.e. there is no price for estimating β_1 and β_2 simultaneously.

Note that the difference in the two cases is the additional $(1-r_{12}^2)$ in the denominator of the variance of each of b_1 and b_2 when X_1 and X_2 are correlated. The bigger the r_{12} , the more this increases.the variance of the estimator: i.e., one pays a price for estimating the β's from positively correlated X's. Note also that the correlation of b_1 and b_2 is opposite in sign from that of the correlation of X_1 and X_2.

6.7 Estimation of mean (and individual) response at a given $\mathbf{X}_h$.

Use the (scalar) dot product $\mathbf{X}_h^T \hat{\beta}$ for both. $\hat{\beta}$ is the random vector and $\mathbf{X}_h$. is the vector of given X values. The Variance of the estimor is larger for the individual response, reflecting both the uncertainty in the estimation of the mean, and the additional uncertainty of the individual response from this mean.

Rules for variance of dot product apply:

$\text{var}\{\mathbf{X}_h^T \hat{\beta}\} = \mathbf{X}_h^T \boldsymbol{\Sigma}_{\mathbf{b}} \mathbf{X}_h$

$\text{var}\{\mathbf{X}_h^T \hat{\beta} + \varepsilon\} = \mathbf{X}_h^T \boldsymbol{\Sigma}_{\mathbf{b}} \mathbf{X}_h + \sigma^2$. (as usual, substitute MSE for σ^2 in $\boldsymbol{\Sigma}_{\mathbf{b}}$ and in σ^2).