§ 3.3

25] §1.4 explains interactions, factors, and indicator variables in an abstract setting.

26] An interaction is a product of 2 or more different variables.

ex] x_2x_3, $x_3x_5x_7$, $x_2x_4^2$

27] x_i^k is a power, eg x_i^2, x_i^3

28] $x_2, \ldots, x_p$ are called main effects.

29] If a power or interaction is in the model, then the main effects that make up the powers and interactions should be in the model.

ex] If x_2^2 and $x_3x_4x_5$ are in the model, x_2, x_3, x_4 and x_5 should be in the model.

30] p.97 * A factor W is a qualitative = categorical variable that takes on C categories $a_1, \ldots, a_C$. Incorporate a factor into the MLR model using C-1 indicator (dummy) variables $x_{w_i} = \begin{cases} 1 & w = a_i \\ 0 & w \neq a_i \end{cases}$ where

one of the "levels" a_i is omitted.

usually omit the 1st or last level.

so $i = 1, \ldots, c-1$ or $i = 2, \ldots, c$.

The degrees of freedom associated with the factor is $c-1$.

ex] gender

$$x = \begin{cases} 0 & x = F \\ 1 & x = m \end{cases}$$

↑ indicator variable

ex] W = brand with 3 brands A, B, C

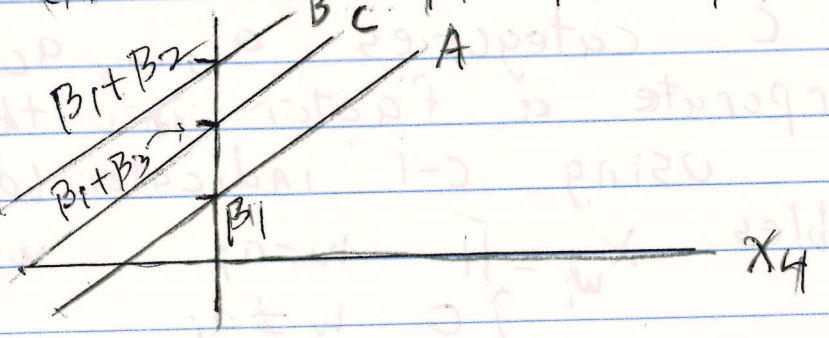
x_2	x_3	Obs W has brand
0	0	A
1	0	B
0	1	C

Suppose x_4 is a quantitative variable

and $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$

parallel lines

$$E(Y \mid \text{brand} = A, x_4) = \beta_1 + \beta_4 x_4$$
$$E(Y \mid \text{brand} = B, x_4) = \beta_1 + \beta_2 + \beta_4 x_4$$
$$E(Y \mid \text{brand} = C, x_4) = \beta_1 + \beta_3 + \beta_4 x_4$$



31] If C indicator variables are used along with a constant, then $X'X$ would be singular.

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

$\underset{\sim}{1} = \underset{\sim}{x}^1 = \underset{\sim}{x}^2 + \underset{\sim}{x}^3 + \underset{\sim}{x}^4$ $\quad \underset{\sim}{x}^1 \; \underset{\sim}{x}^2 \; \underset{\sim}{x}^3 \; \underset{\sim}{x}^4$

32] Suppose x_2 is quantitative and x_3 is qualitative with 2 categories

$$x_3 = \begin{cases} 1 & \text{if A} \\ 0 & \text{if B.} \end{cases}$$

Then a 1st order model with interaction is $Y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \underbrace{x_2 x_3}_{x_4 = x_2 x_3} + e$

4 models:

i) Full model: 2 unrelated lines

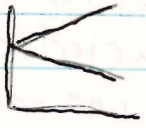

$\begin{cases} A & x_3 = 1, \; E[Y|x] = \beta_1 + \beta_3 + (\beta_2 + \beta_4) x_2 \\ B & x_3 = 0, \; E[Y|x] = \beta_1 + \beta_2 x_2 \end{cases}$

ii) $\beta_4 = 0$: 2 parallel lines


$\begin{cases} A & x_3 = 1, \; E[Y|x] = \beta_1 + \beta_3 + \beta_2 x_2 \\ B & x_3 = 0, \; E[Y|x] = \beta_1 + \beta_2 x_2 \end{cases}$

iii) $\beta_3 = 0$, $\beta_4 = 0$: coincident


$\begin{cases} A & x_3 = 1, \; E[Y|x] = \beta_1 + \beta_2 x_2 \\ B & x_3 = 0, \; E[Y|x] = \beta_1 + \beta_2 x_2 \end{cases}$

iv) $\beta_3 = 0$; lines have same intercept


$\begin{cases} A & x_3 = 1, \; E[Y|x] = \beta_1 + (\beta_2 + \beta_4) x_2 \\ B & x_3 = 0, \; E[Y|x] = \beta_1 + \beta_2 x_2 \end{cases}$

33] Models with interaction can rapidly get complex.

i) interaction between "factor" $x_2, \ldots, x_c$ and x_{c+1}
($c-1$ indicator variables; quant)

has $c-1$ interactions $x_j x_{c+1}$ $j = 2, \ldots, c$

constant, x_2 to x_{c+1}

$1 + c + c-1 = 2c$ parameters

ii) 1 quant, several qualitative

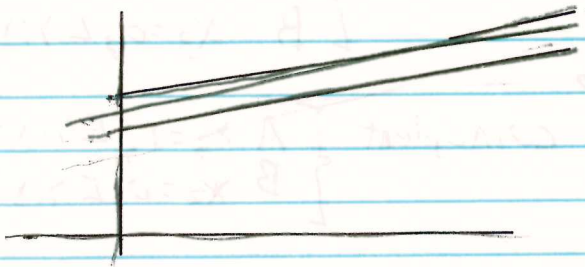
iii) several quant, several qualitative

34] Usually hope that interactions are not needed: use t tests or partial F test.

35] As factors have more levels and interactions have more terms (eg $x_2 x_3 x_4 x_5$), the model becomes more complex.

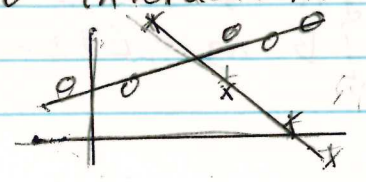
36] One factor and one quant variable x_2:
Use c symbols for the different levels, and fit SLR for each symbol.

If all c lines are close, there may be no interaction (even if the lines cross).



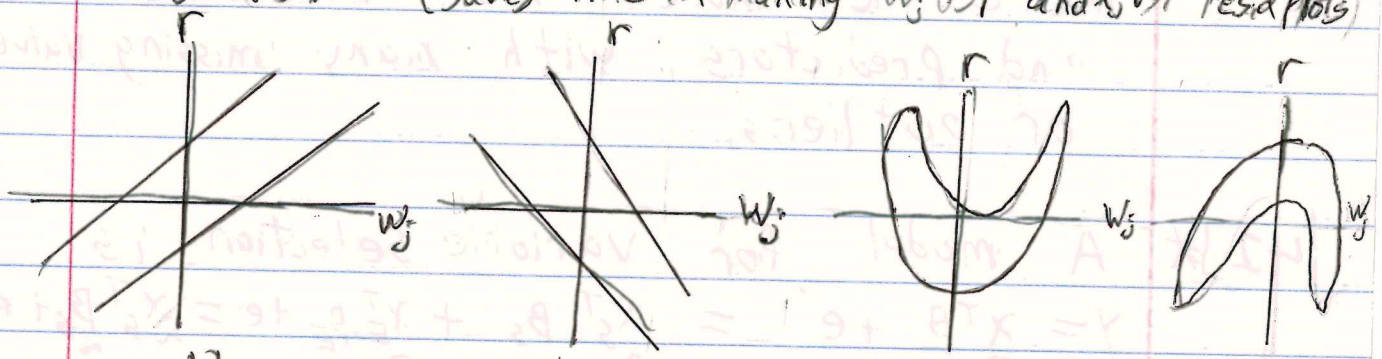
no interaction

In Arc, fit marks general vs fit marks parallel:
If there is little change, then there may be no interaction.



Make plot of x_2 vs Y & x_2 vs r
strongly different lines (relative to symbol variation) suggest interaction. see Ex 3.6

37] After fitting an MLR model, a scatterplot matrix of the predictors, potential predictors, and residuals is useful (saves time in making w_j vs r and x_j vs r resid plots)



add w_j to model

add w_j^2 & w_j^2 to model

§3.4 38] p99 Variable Selection = subset selection = model selection is a search for a subset of predictor variables that can be deleted without important loss of information.

39] * $Y = \underset{\sim}{x}^T \underset{\sim}{\beta} + e$ is the full model

$Y = \underset{\sim}{x}_I^T \underset{\sim}{\beta}_I + e$ is a submodel (like a reduced model)

where $\underset{\sim}{x}_I$ is a subset of the predictors $\underset{\sim}{x}$.

40] * The full model is always a submodel, and may be the final model selected, especially for $p = 2$ or 3.

41] Variable selection is data snooping, but needs to be done since

want $n \geq 5p$. Also, fitting unnecessary predictors increases $V(\hat{Y}_t)$ and $V(\hat{\beta}_i)$. Want to eliminate expensive predictors,

predictors like interactions that make the model hard to understand, and predictors with many missing values or outliers.

42] A model for variable selection is

$$Y = \underset{p\times 1}{\underset{\sim}{x}^T \underset{\sim}{\beta}} + e = \underset{\sim}{x_S}^T \underset{k_S \times 1}{\underset{\sim}{\beta_S}} + \underset{(p-k_S)\times 1}{\underline{x_E}^T \underline{\beta_E}} + e = \underset{\sim}{x_S}^T \underset{\sim}{\beta_S} + e$$

where $\underline{x} = (\underline{x_S}^T, \underline{x_E}^T)^T$, and given $\underline{x_S}$ is in the model, then $\underline{\beta_E} = \underline{0}$ and $\underline{x_E}$ denotes the subset of terms that can be eliminated from the model.

43] $\underset{\sim}{x_S}$ need not be unique; there may be 2 or more subsets of size k_S that work.

44] Often want $k_S \geq 2$ as small as possible, but often certain predictors are required to be in the model.

45] S is unknown. Let $\underset{\sim}{x_I}$ be a $k \times 1$ vector from a candidate submodel and $\underline{x_O}$ be the $(p-k)\times 1$ vector of terms out of the submodel. Then $Y = \underline{x_I}^T \underline{\beta_I} + \underline{x_O}^T \underline{\beta_O} + e$.

46] p103-4

i) want $R^2(I) \approx R^2(F) = R^2_{full}$ ($R^2(F) \geq R^2(I)$ always, adding terms does not decrease R^2.)

ii) Want $MSE(I) \approx MSE(F) = MSE$ or smaller. $MSE(I) < MSE(F)$ is possible.

iii) Let model I have k variables including a constant. want

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k$$

small: $C_p(I) \leq \min(2k, p)$.

iv) want "adjusted R^2" $R^2_A(I)$ high.

$$R^2_A(I) = 1 - (1 - R^2(I))\frac{n}{n-k} = 1 - MSE(I)\frac{n}{SST}.$$

Adding predictors can decrease $R^2_A(I)$ and the model with max $R^2_A(I)$ is also the model with min $MSE(I)$.

47] If $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$, then there are 2^{p-1} submodels that contain β_1 and any number of $x_2, x_3, \ldots, x_p$. So numerical methods are needed to screen out most models.

48] p102 To build a full model, remove strong nonlinearities from the predictors. Do not transform indicator variables.

49] p102-3 If a good full model has c-1 indicator variables corresponding to a factor, either keep or delete all c-1 indicator variables.

50] Forward Selection p. 10, 110

Step 1) $k=1$: Start with a constant $W_1 = X_1$

Step 2) $k=2$: compute C_p for all models with $k=2$ containing a constant and a single predictor X_j. Keep the predictor $W_2 = X_j$ that minimizes C_p.

Step 3) $k=3$: Fit all models with $k=3$ that contain W_1 and W_2. Keep the predictor W_3 that minimizes C_p. This is also the variable that minimizes $SSE = \sum r_i^2$,

⋮

Step j) $k=j$: Fit all models with $k=j$ that contain $W_1, W_2, \ldots, W_{j-1}$. Keep the predictor W_j that minimizes C_p.

Step p) $k=p$: Fit the full model.

51] Backward Elimination p10, 110

Step 0) Full model $U_1, \ldots, U_p$

U_i need not equal X_i. All models contain a constant $U_1 = X_1$.

Step 1) $k=p-1$: Fit each model with $k=p-1$ predictors including a constant. Delete the predictor U_p, say, that corresponds

to the model with the smallest C_p.
Keep $U_1, \ldots, U_{p-1}$.

Step 2) $K = p-2$: Fit each model with $p-2$ predictors, including a constant. Delete the predictor U_{p-1} corresponding to the smallest C_p.
Keep $U_1, \ldots, U_{p-2}$.

⋮

Step j) $K = p-j$: Fit each model with $p-j$ predictors including a constant. Delete the predictor U_{p-j+1} corresponding to the smallest C_p. Keep $U_1, \ldots, U_{p-j}$.

⋮

Step p-2) $K = 2$: The current model contains U_1, U_2, U_3. Fit the model U_1, U_2 and U_1, U_3. Assume the model U_1, U_2 minimizes C_p. Delete U_3.

52) Both backward elimination and forward selection result in models

$K = 2$ U_1, U_2
3 U_1, U_2, U_3
⋮
j $U_1, U_2, \ldots, U_j$
⋮
p $U_1, U_2, \ldots, U_p$ = full model

The sequence of models need not be the same since the predictors could be highly correlated. Forward selection

tries to start with the best SLR model $U_1 = X_1 = 1$, $U_2 = X_j$ say. But in backward elimination, X_j may be highly correlated with other predictors, and X_j could be the 1st variable deleted by backward elimination.

At step 1 of backward elimination, the "least important predictor" (wrt C_p) U_p given the other variables are in the model, is deleted. U_p could be a very important SLR predictor.

53] * p109-10 Rule of thumb for numerical methods like backward elimination and forward selection.

a) Let I_{min} correspond to the model with min C_p. Let I_I be the submodel with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$. Then I_I is the initial submodel to be examined.

b) Models I with fewer predictors than I_I such that $C_p(I) \leq C_p(I_{min}) + 4$ should be examined.

c) Models with k predictors (including a constant) and with fewer predictors than I_I such that $C_p(I_{min}) + 4 < C_p(I) \leq \min(2k, p)$ should be checked, but often underfit important predictors are deleted from the model.

d) If there are no models I with fewer predictors than I_I such that $C_p(I) \leq \min(2k, p)$, then model I_I is a good candidate for the best submodel found by the numerical method.

54] $I_I = I_{min}$ = full model is possible.

55] Forward selection and backward elimination often use $C_p(I)$, but may use $AIC(I)$, $R^2_A(I) \equiv MSE(I)$, delete variable x with largest pval for t test, add variable x with smallest pval for t test, etc.

56] Stepwise elimination is a modification. At each step, 1 predictor that was deleted has a chance to get back in the model for backward stepwise. For forward stepwise, 1 predictor that was added has a chance to be deleted.

57] All subsets: there are programs that will find the smallest 3 or 4 $C_p(I)$ for subsets of size $2, 3, \ldots, p$ if $p \leq 30$.

See ex 3.7 for output from forward selection and backward elimination.

ex] SAT data $Y = f750$ = female 750 or up on verbal
$x_2 = m750$, $x_3 = m700$, $x_4 = f700$

Forward sel (33.5)

		K	CpI
base	intercept		
add	F700	2	22.1
	M750	2	22.2
	M700	2	75.1
base	(F700) (and int)	K	CpI
add	M750	3	2.6 ← $I_{min}=I_I$
	M700	3	23.8
base	(F700, M750)	K	CpI
add	M700	4	4.0

so submodel I_I uses constant, F700, M750

backward elim

			K	CpI
current terms (M750, M700, F700) and constant				
$I_{min}=I_I$	delete	M700	3	2.6
		F700	3	23.4
		M750	3	23.8
current terms (M750, F700)			K	CpI
	delete	M750	2	22.1
		F700	2	38.2

I_I submodel uses constant, M750, F700

For this ex, I_I = final submodel for both forward selection and backward elimination.