



Seleção de Variáveis e Construindo o Modelo

Seleção de modelos candidatos

A idéia é selecionar um conjunto menor de variáveis explanatórias de acordo com algum(s) critério(s), e assim selecionar o modelo de regressão final a ser utilizado.

Contudo existem dois objetivos conflitantes:

- ❖ incluir muitos regressores quanto possível a fim explicar bem a variabilidade de y .
- ❖ incluir poucos regressores quanto possível a fim de não aumentar a variância de $\hat{y}$.

Seleção de variáveis, análise dos resíduos e verificação de pontos de influência e de alavanca devem ser empregados iterativamente.

Consequências

- ❖ removendo variáveis pode melhorar a precisão das estimativas dos parâmetros. Isso inclui a variância da resposta predita.
- ❖ a remoção de variáveis importantes produz viés nas estimativas dos coeficientes e da resposta predita.
- ❖ perigo reter variáveis com coeficientes próximos de zero: produz variâncias maiores para estimativas dos parâmetros e para a resposta predita.
- ❖ removendo variáveis não muito significantes, pode-se ter erro padrão das estimativas viesadas menor do que a variâncias das estimativas não enviesadas.

Critérios:

- Coeficiente de determinação múltiplo: R_p^2
- Coeficiente de determinação múltiplo ajustado: R_a^2
- Critério $Press_p$ (Soma de quadrados da predição)

Notação:

Número de variáveis X : $p-1$

Um modelo com todas as variáveis X , possui p parâmetros (sempre incluído o intercepto).

O número de um subconjunto de variáveis regressoras no modelo é $p-1$.

Assume-se que $n > p$ (quanto maior melhor).

Coeficiente de determinação múltiplo

R_p^2 → É o coef. de determinação múltiplo para um modelo com p parâmetros, ou $p-1$ variáveis regressoras no modelo.

Objetivo: identificar os (vários) modelos candidatos com maior valor de R_p^2

$$R_p^2 = 1 - \frac{SSR_p}{SST}$$

P é o número de parâmetros ou $p-1$ variáveis X .

Coeficiente de determinação múltiplo

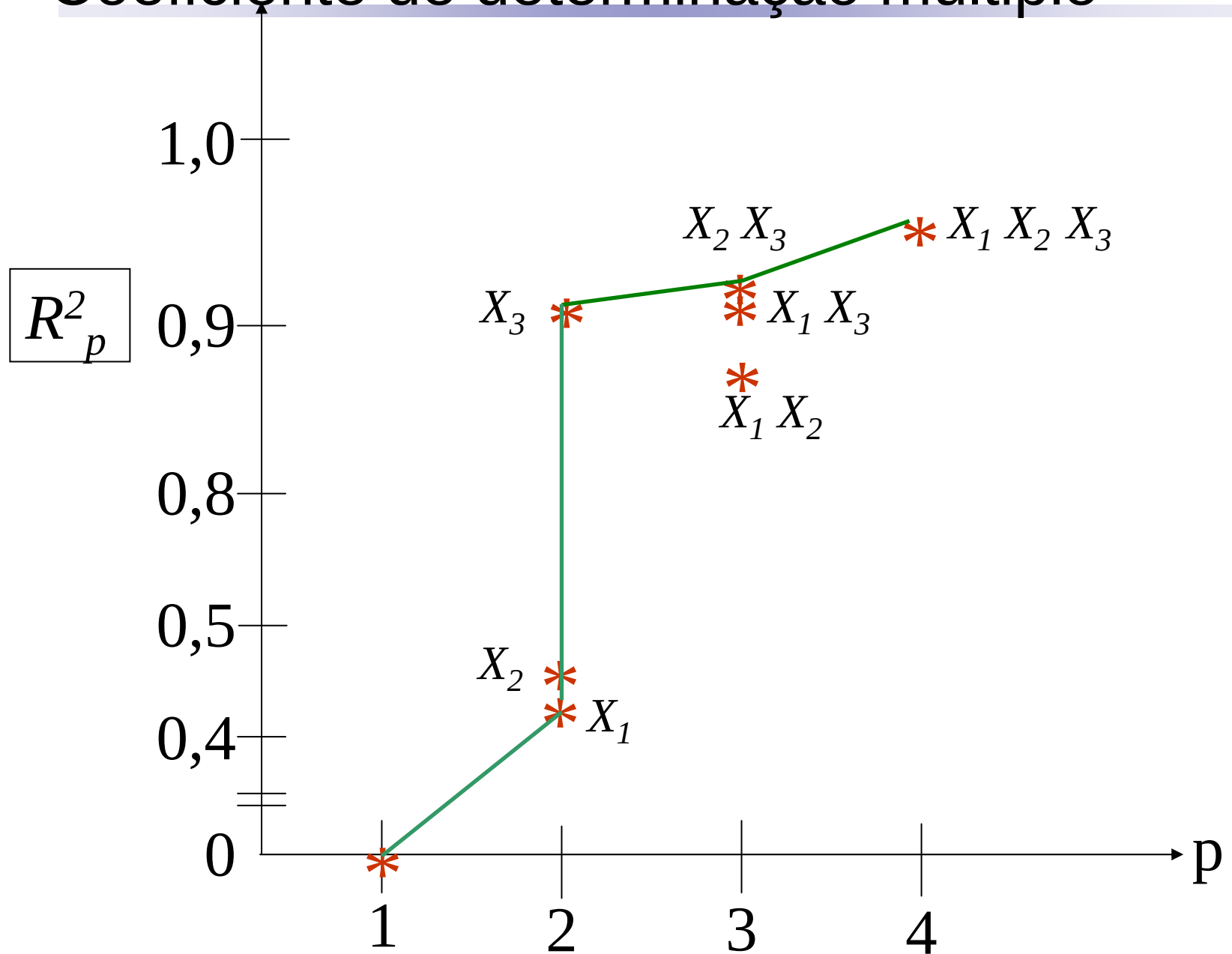
Objetivo: a intenção com a utilização de R^2_p é encontrar o ponto onde a adição de novas variáveis X no modelo não é vantajoso, pois isto conduz a um pequeno acréscimo em R^2_p .

Exemplo: Na tabela a seguir, as colunas 2, 3, 4 e 5 indicam, o número de parâmetros do modelo, graus de liberdade do resíduo, soma de quadrados do erro, e o coeficiente de determinação.

Coeficiente de determinação múltiplo

Variáveis X no modelo	p	gl	SSR_p	R^2_P	MSR_p	$PRESS_P$
Nenhuma	1	17	702,10	0	41,30	787.13
X_1	2	16	390.80	0.4434	24.42	457.09
X_2	2	16	381.34	0.4569	23.83	453.94
X_3	2	16	58.78	0.9163	3.67	70.30
X_1, X_2	3	15	85.02	0.8789	5.67	133.27
X_1, X_3	3	15	57.98	0.9174	3.87	80.21
X_2, X_3	3	15	53.87	0.9233	3.59	71.09
X_1, X_2, X_3	4	14	41.78	0.9405	2.98	64.80

Coeficiente de determinação múltiplo



Coeficiente de determinação múltiplo

O gráfico mostra que existe pouco acréscimo no R^2_p quando o modelo inclui a variável X_3 ou X_2X_3 . Apresenta um pequeno acréscimo, com relação aos dois modelos anteriores, quando inclui as 3 variáveis $X_1X_2X_3$.

Desvantagem: R^2_p não leva em consideração o número de parâmetros no modelo de regressão, assim, ele nunca diminui quando p aumenta.

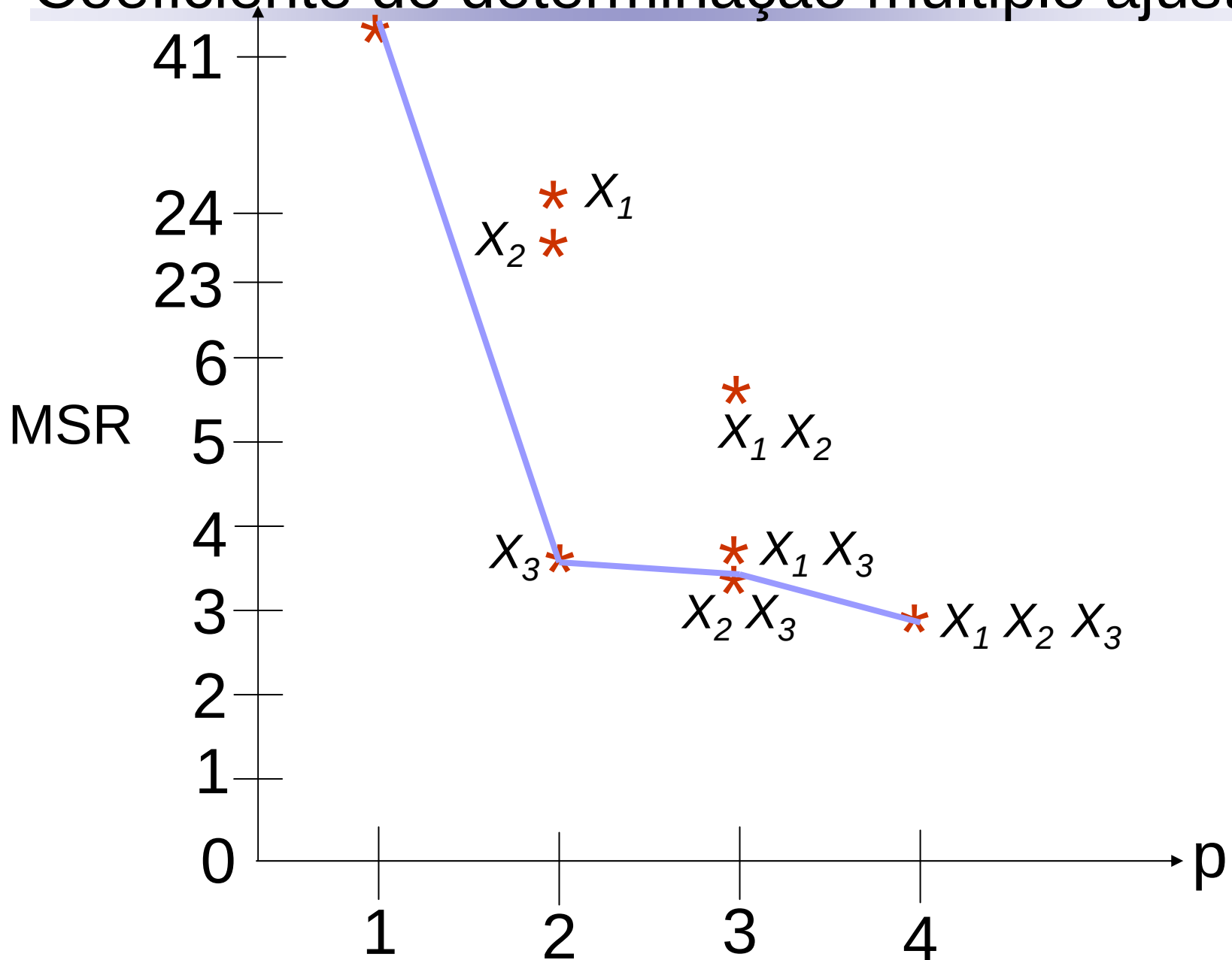
Coeficiente de determinação múltiplo ajustado

$$R_a^2 = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSR}{SST} = 1 - \frac{MSR}{SST / (n-1)}$$

↑
Leva em consideração o número de parâmetros no modelo

Observe que R_a^2 aumenta se e apenas se o quadrado médio do erro diminui, pois a $SST/(n-1)$ é fixa dado os valores observados de y .

Coeficiente de determinação múltiplo ajustado



Coeficiente de determinação múltiplo ajustado

As conclusões são muito semelhantes àquelas feita com o uso do coeficiente de determinação múltiplo. Os quadrados médios dos erros para os subconjuntos X_3 , X_2X_3 , X_1X_3 e $X_1X_2X_3$ são praticamente iguais.

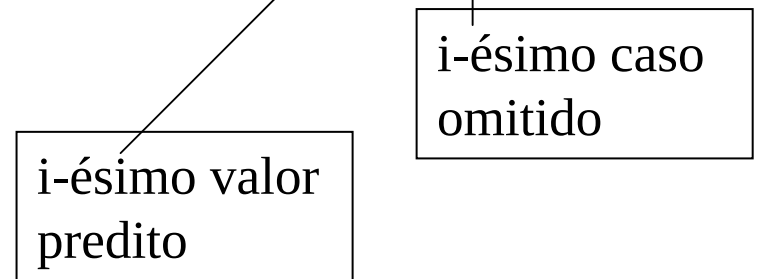
Critério PRESS

O critério $PRESS_p$ (soma de quadrados de predição) é uma medida de quão bem os valores ajustados através de um modelo candidato (de pesquisa) podem prever as respostas observadas Y_i .

É potencialmente útil em discriminação entre modelos alternativos. A complexidade do algoritmo não é linear.

Critério PRESS

Como funciona: cada valor ajustado $Y_i(\text{chapéu})$ para o critério PRESS é obtido deletando-se o i -ésimo caso do conjunto de dados, estimando a função de regressão para o modelo candidato com as $n-1$ observações restantes e, então, usando a equação de regressão ajustada, obter o valor predito $\hat{Y}_{i(i)}$, para a i -ésima observação.



Critério PRESS

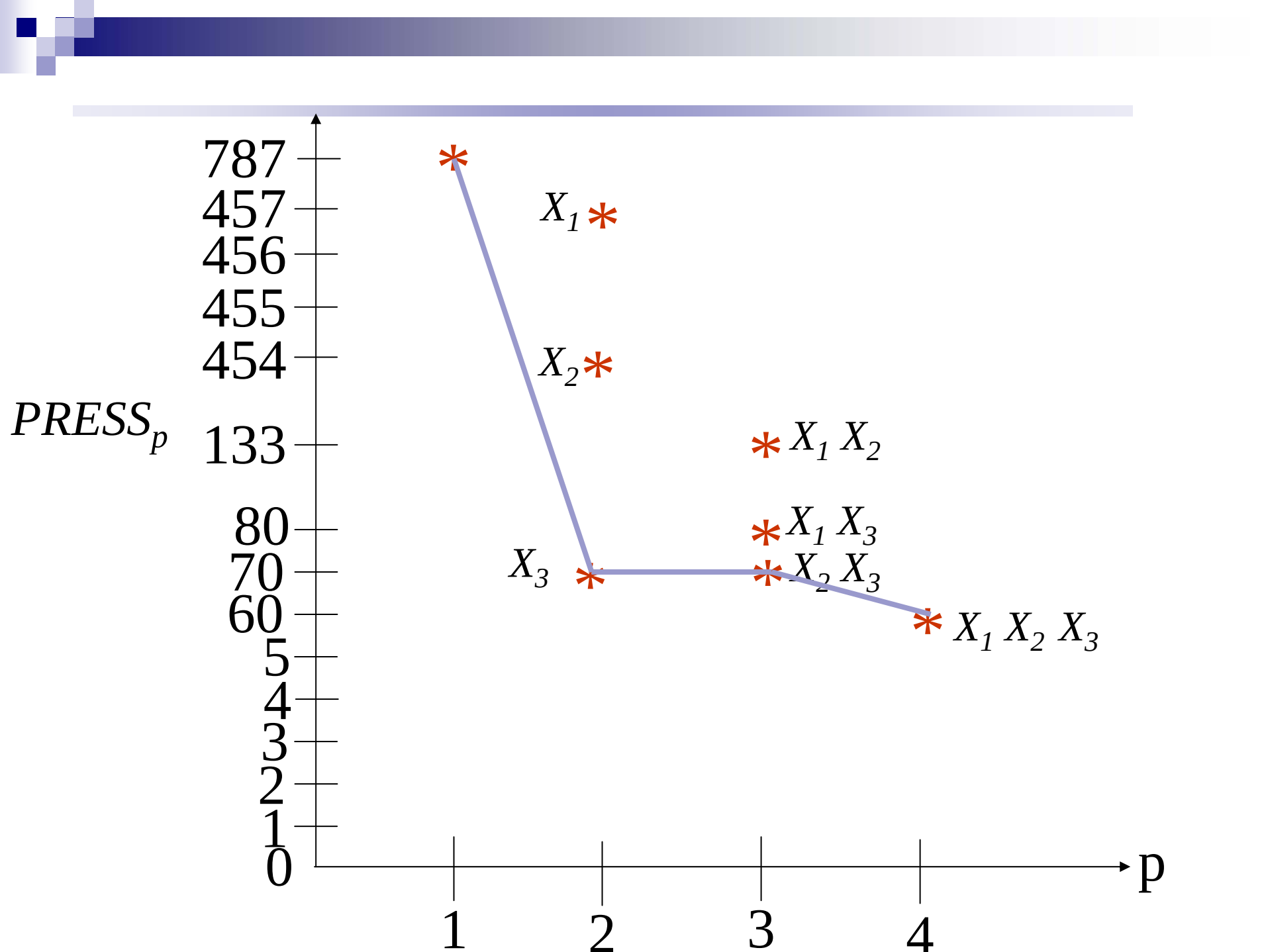
O erro de predição PRESS para o i -ésimo caso é dado por:

$$Y_i - \hat{Y}_{i(i)}$$

O critério $PRESS_p$ é a soma de todos esses n erros de predição quadráticos:

$$PRESS_p = \sum_{i=1}^n \left(Y_i - \hat{Y}_{i(i)} \right)^2$$

Modelos com valores pequenos de $PRESS_p$ são os melhores modelos candidatos (tem menor erro de predição).



Critério PRESS

Observamos que os modelos com X_3 , X_2X_3 e $X_1X_2X_3$ apresentam os menores valores de $PRESS_p$. Na verdade, o modelo com $X_1X_2X_3$ apresenta um valor de $PRESS_p$ levemente inferior aos outros dois.

Regressão passo a passo (Stepwise Regression)

O ideal é computar os critérios apresentados previamente para todos os possíveis modelos que podem ser construídos com um conjunto de variáveis. Existem 2^p regressões possíveis. Contudo o procedimento é computacionalmente inviável para situações envolvendo mais que poucos regressores. Algumas alternativas para isso são os métodos abaixo.

Seleção passo a frente:

É feita uma sequência de modelos de regressão, em cada passo adicionando uma variável preditora X do modelo.

Eliminação passo a atrás:

É feita uma sequência de modelos de regressão, em cada passo retirando uma variável preditora X do modelo.

Regressão passo a passo

É uma modificação do método de seleção passo a frente em que regressores que entram em determinado passo pode ser removido em um passo a frente.

Regressão passo a passo (Stepwise Regression)

Critério para adicionar ou retirar uma variável preditora X do modelo:

- estatística F^*

Deficiência do método: termina com um único modelo como sendo o melhor. A experiência mostra que as vezes se chega a um modelo pobre.

Recomendação:

Usar o subconjunto de variáveis encontrado pelo método de regressão stepwise como sendo o ponto de partida para encontrar outros “bons” subconjuntos. Uma possibilidade é considerar o número de variáveis preditoras encontrado pela regressão stepwise como sendo aproximadamente o número correto e, então, usar um procedimento que realiza todas as regressões possíveis para subconjuntos deste tamanho (R^2_p , R^2_a , $PRESS_p$).

Etapas: (Teste F parcial)

- Inicia-se ajustando-se um modelo de regressão linear simples para cada uma das $P-1$ variáveis explanatórias X . Para cada modelo a estatística F^* é calculada:

$$F_k^* = \frac{MS_M(X_k)}{MS_R(X_k)}$$

- A variável X com o maior valor de F^* é candidata para a primeira adição. Se este valor de F^* ultrapassar determinado valor F_{IN} , então, a variável é adicionada no modelo. Caso contrário, o programa é concluído e não são incluídas variáveis no modelo.

Etapas: (Teste F parcial)

- Vamos assumir que a variável X_3 entrou no modelo no passo 1. A regressão stepwise ajusta todos os modelos com duas variáveis X no modelo, onde X_3 é uma do par. Para cada modelo de regressão calcula-se a estatística do teste F parcial:

$$F_k^* = \frac{MS_M(X_k | X_3)}{MS_R(X_k, X_3)}$$

- A variável X com o maior valor de F^* é candidata a entrar no modelo neste segundo estágio. Se este valor de F^* exceder um determinado valor F_{IN} , a segunda variável X é incluída no modelo, caso contrário, o programa é finalizado.

Etapas: Teste F parcial)

- Suponha que X_1 é incluída no modelo no segundo estágio. Neste passo, a regressão stepwise examina se qualquer das outras variáveis X já incluídas no modelo devem ser retiradas. No nosso caso, só temos uma outra variável no modelo, X_3 , assim, só é feito um teste F parcial:

$$F_3^* = \frac{MS_M(X_3 | X_1)}{MS_R(X_3, X_1)}$$

- Se este valor de F^* for menor do que um determinado valor F_{OUT} , a variável X é deletada do modelo, caso contrário, ela permanece.

Etapas:(Teste F parcial)

- Suponha que ambas as variáveis, X_3 e X_1 , permanecem no modelo. A regressão stepwise examina qual nova variável X é candidata a entrar no modelo, então verifica se qualquer das variáveis já incluídas no modelo deveriam ser deletadas, e assim por diante, até que não se tenha mais variáveis que podem ser adicionadas ou retiradas do modelo. Neste ponto a regressão é encerrada.
- Variáveis para entrar no modelo são significativas pelo menos ao nível de 15% (valor definido pelo pesquisador) (valor p-value $\leq 0,15$). $F_{IN} = F_{0,15;1;n-p}$
- Variáveis para saírem do modelo não alcançaram o nível de significância de (idem)(valor p $> 0,15$). $F_{out} = F_{0,15;1;n-p}$
- As vezes F_{IN} não é igual a F_{out} . Frequentemente escolhe-se $F_{out} > F_{IN}$.
- Em geral usa-se a estatística $t_{\alpha/2,n-p}$ pois é equivalente a estatística $F_{\alpha;1;n-p}$
- Minitab usa $\alpha=0,25$ para $F_{IN}=F_{out}=F_{0,25;1;n-p}$. Kennedy e Bancroft sugerem $F_{IN}=F_{0,25;1;n-p}$ e $F_{out}=F_{0,10;1;n-p}$.

Questões na análise do modelo completo

- Quais regressores são importantes?
- Existem pontos aberrantes?
- Existe necessidade de transformação na variável resposta e/ou em alguma variável preditora?
- Razões básicas para fazer uma transformação na variável resposta:
 - Uso de escala errada nos dados. Uma transformação é a função log.
 - Existência de outliers. Isso pode ser provocado pelo modelo e remoção de regressores não importantes pode atenuar o problema.

Critérios

- A estatística PRESS tende a recomendar modelos menores do que o R^2 ajustado.