Added Variable plot (AVP)

① $\hat{e} = Y - X\hat{\beta} = Y - X(X'X)^{-1}X'Y = (I - \underbrace{X(X'X)^{-1}X'}_{H})Y = (I-H)Y$

② $(I-H)X = 0$, $HX = X(X'X)^{-1}X'X = X$

$H_0 = X_0(X_0'X_0)^{-1}X_0'$: Model $Y = X_i\beta_i + X_0\beta_0 + e$ (interested, others)

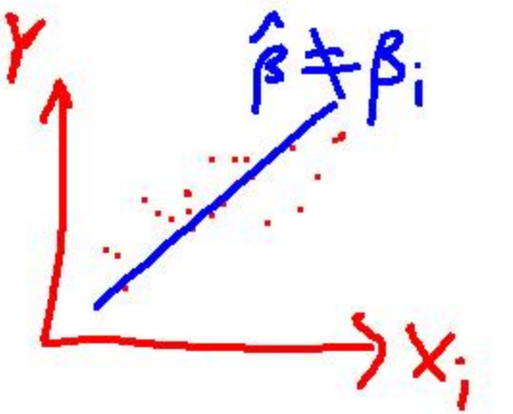
③ AVP: $(I-H_0)Y = (I-H_0)X_i\beta_i + (I-H_0)X_0\beta_0 + (I-H_0)e$

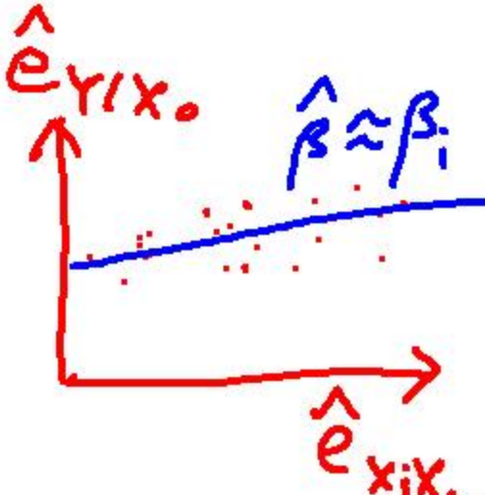
$(I-H_0)X_0\beta_0$: ② ↓

$(I-H_0)Y$: ① , $(I-H_0)X_i\beta_i$: ①

$\hat{e}_{Y|X_0} = \hat{e}_{X_i|X_0}\beta_i + 0 + \hat{e}$

$\Rightarrow$ slope b/w $\hat{e}_{Y|X_0}$ & $\hat{e}_{X_i|X_0}$ is β_i





3.5. Comparing models: Analysis of variance (ANOVA)

- Regression is the study of dependence of variables
  - $Y = X_1\beta_1 + X_O\beta_O + e$,
    - $\beta_1 = 0 \rightarrow X_1$ and y are independent
    - $\beta_1 \neq 0 \rightarrow X_1$ and y are dependent
- Question:
  - Are X_1 and y dependent?
- Answer:
  - Method 1) test whether $\beta_1 = 0$ if β_1 is scalar
  - Method 2) Compare the two models (Here $X = [X_1\ X_O]$)
    - $E(Y|X) = X_O\beta_O$ i.e. $Y = X_O\beta_O + e$
    - $E(Y|X) = X_1\beta_1 + X_O\beta_O$ i.e. $Y = X_1\beta_1 + X_O\beta_O + e$

$RSS = SYY - \frac{SXY^2}{SXX}$ (only for simple linear)

3.5 Comparing models: Analysis of variance (ANOVA)

- Analysis of variance (ANOVA) is a method that compares two models of mean functions
  - NH: $E(Y|X) = X_O\beta_O$
  - AH: $E(Y|X) = X_1\beta_1 + X_O\beta_O$
- For the first model: $E(Y|X) = X_O\,\beta_O$
  - $RSS_{NH} = \min_{\beta_O} \sum (Y_i - X_{Oi}\beta_O)^2 \overset{def}{=} \sum (Y_i - X_{Oi}\tilde{\beta}_O)^2 = \sum \hat{e}_i^2$
- For the second model: $E(y|x) = X_1\,\beta_1 + X_O\,\beta_O$
  - $RSS_{AH} = \min_{\beta_1,\beta_O} \sum (Y_i - X_{1i}\beta_1 - X_{Oi}\beta_O)^2 \overset{def}{=} \sum (Y_i - X_{1i}\hat{\beta}_1 - X_{Oi}\hat{\beta}_O)^2$
- By default, $RSS_{NH} > RSS_{AH}$
  - The second model is useful only if $RSS_{NH} >>> RSS_{AH}$

3.5 Comparing models: Analysis of variance (ANOVA)

- Difference sum of square due to regression
  - RSS_{NH}: $\sum (Y_i - X_{0i}\beta_0)^2$ (p_{NH})
  - RSS_{AH}: $\sum (Y_i - X_{1i}\hat{\beta}_1 - X_{0i}\hat{\beta}_0)^2$ (p_{AH})
  - RSS_{NH}-RSS_{AH}
    - large → model AH explains much more variation
    - Not so large → model NH is already good enough
  - How large is large?
- Study the distribution of RSS_{NH}-RSS_{AH} (idea)
  - RSS_{NH} is a sum of $df_{NH}=n-p_{NH}$ squares of normal r.v.
  - RSS_{AH} is a sum of $df_{AH}=n-p_{AH}$ squares of normal r.v.
  - $RSS_{NH} - RSS_{AH} \sim \chi^2_{df_{NH}-df_{AH}}$ and independent with RSS_{AH}

$$F = \frac{(RSS_{NH} - RSS_{AH})/(df_{NH} - df_{AH})}{RSS_{AH}/df_{AH}} \sim F(df_{NH} - df_{AH}, df_{AH})$$

(σ^2)

H_0: Y & X_1 not related

H_A: Y & X_1 are related

3.5 A Special Case
Overall Analysis of variance (ANOVA)

- Difference sum of square due to regression
  - NH: $E(Y|X)=\beta_0$
  - AH: $E(Y|X)=X\beta$ (X is the matrix formed by p+1-variables)
- RSS_{NH}: $\sum(Y_i-\tilde{\beta}_0)^2=\sum(Y_i-\bar{Y})^2=SYY$
- RSS_{AH}: $\sum(Y_i-X\hat{\beta})^2$
- Study the distribution of RSS_{NH}-RSS_{AH}
  - Define SSreg= RSS_{NH}-RSS_{AH}=SYY-RSS_{AH}
    - This is the variation explained by the multiple regression

$$F=\frac{(RSS_{NH}-RSS_{AH})/(df_{NH}-df_{AH})}{RSS_{AH}/df_{AH}}$$
$$=\frac{SSreg/p}{RSS_{AH}/(n-p-1)}\sim F(p,n-p-1)$$

n-1

n-p-1

ANOVA table: a break-down of squares (variation)

TABLE 3.4 The Overall Analysis of Variance Table

Source	df	SS	MS	F	*p*-value
Regression	p	$SSreg$	$SSreg/p$	$MSreg/\hat{\sigma}^2$	
Residual	$n-(p+1)$	RSS	$\hat{\sigma}^2 = RSS/(n-(p+1))$		
Total	$n-1$	SYY			

$$\sum_{i=1}^{n}[y_i - \bar{y}]^2 = \sum_{i=1}^{n}[y_i - \hat{y}_i]^2 + \sum_{i=1}^{n}[\hat{y}_i - \bar{y}]^2$$

$$TSS = SYY = RSS + SSreg$$

← proved in Ch 8

Variation of the data = **Variation not explained by regression** + **Variation explained by regression**

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	4	201994	50499	11.992	[redacted]07
Residuals	46	193700	4211		
Total	50	395694			

$n-p-1$ $\Sigma \hat{e}_i^2 = \Sigma(Y_i - \hat{Y}_i)^2$

$n-1$ $SYY = \sum_i^n (Y_i - \bar{Y})^2$

$X_i\hat{\beta}$

- Anova F-test – Test if the regression is useful
  - NH: $E(Y|X) = \beta_0$
  - AH: $E(Y|X) = X\beta$
  - F stat=11.992, to compare with F(4,46)
  - p-value = 1-pf(11.992,4,46) =9.33e-07
  - NH is rejected. The regression is considered useful!
- $R^2 = \frac{SSreg}{SYY}$ =201994/395694=0.5105.
  - About half of the variation is explained.