Chapter 3

Multiple Regression

Multiple Regression

- What is multiple regression?
  - Adding more predictors to explain the response variable better.
- Improve $\mathrm{E}(Y|X_1 = x_1) = \beta_0 + \beta_1 x_1$
  by $\mathrm{E}(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
  - Adding X_2 to explain the part of Y that has not already been explained by X_1.

Terms and Predictors (X variables)

- Predictors: Original data you collect
  - e.g. height, weight, color, gender
- Terms: Created from the predictors
  - The X-variables in multiple regression models
  - e.g.height^2, log(weight), height x weight, color
  - In general, terms includes
    - 1)Intercept, 2)predictors, 3)transformation of predictors
    - 4)Polynomials 5) Interaction/combinations of predictors
    - 6) Dummy variable and factor
- **Important Question: Select a `good’ set of terms**

Matrix Notation for Multiple Regression

- Regression Model

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

$$\mathrm{Var}(Y|X) = \sigma^2$$

- Observed value in Matrix form:

case	y	predictor 1		predictor p
1	y_1	x_{11}	$\cdots$	x_{1p}
2	y_2	x_{21}	$\cdots$	x_{2p}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	$\cdots$	x_{np}

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

intercept

Matrix Notation for Multiple Regression

nx1	nx(p+1)	(p+1)x1	nx1
$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$	$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$	$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$	$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$

multiple linear regression in matrix notation

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

$\Rightarrow$ the ith row is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$

about the vector of errors $\mathbf{e}$:

$$\mathrm{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{Var}(\mathbf{e}) = \begin{pmatrix} Var(e_1) & Cov(e_1,e_2) & \ldots & Cov(e_1,e_m) \\ Cov(e_2,e_1) & Var(e_2) & \ldots & Cov(e_2,e_m) \\ \vdots & \vdots & \ldots & \vdots \\ Cov(e_m,e_1) & \ldots & \ldots & Var(e_m) \end{pmatrix} = \sigma^2 \mathbf{I_n}$$

Matrix Notation for Multiple Regression

- multiple linear regression in matrix notation

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

- Least Square estimation for β

$$RSS(\beta) = \sum (y_i - \hat{y}_i)^2$$

$$= \begin{pmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \quad \vdots \\ y_n - \hat{y}_n \end{pmatrix}' \begin{pmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \quad \vdots \\ y_n - \hat{y}_n \end{pmatrix}$$

$$= (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$$

They are scalar

$$= \mathbf{Y}'\mathbf{Y} - 2\,\mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \qquad (\mathbf{Y}'\mathbf{X}\beta = (\mathbf{Y}'\mathbf{X}\beta)' = \beta'\mathbf{X}'\mathbf{Y})$$

Matrix Differentiation

Let

$$\beta = [\beta_1, \beta_2, \cdots, \beta_k]'$$

$$f(\beta) = f([\beta_1, \beta_2, \cdots, \beta_k]')$$

define: the derivative of $f(.)$ wrt β

$$\frac{\partial f(\beta)}{\partial \beta} = \begin{bmatrix} \frac{\partial f(\beta)}{\partial \beta_1} \\ \frac{\partial f(\beta)}{\partial \beta_2} \\ \vdots \\ \frac{\partial f(\beta)}{\partial \beta_k} \end{bmatrix}$$

- e.g.1
  - $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3]$
  - $f(\boldsymbol{\beta}) = (\beta_1 + \beta_2)\, \beta_3$
  - $\frac{\partial f(\beta)}{\partial \beta} = \begin{bmatrix} \beta_3 \\ \beta_3 \\ \beta_1 + \beta_2 \end{bmatrix}$
- e.g.2
  - $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3]$
  - $f(\boldsymbol{\beta}) = \beta^2_1 \beta_2 + \log(\beta_3)$
  -

Matrix Differentiation

Let

$$\beta = [\beta_1, \beta_2, \cdots, \beta_k]'$$

$$f(\beta) = f([\beta_1, \beta_2, \cdots, \beta_k]')$$

define: the derivative of $f(.)$ wrt β

$$\frac{\partial f(\beta)}{\partial \beta} = \begin{bmatrix} \frac{\partial f(\beta)}{\partial \beta_1} \\ \frac{\partial f(\beta)}{\partial \beta_2} \\ \vdots \\ \frac{\partial f(\beta)}{\partial \beta_k} \end{bmatrix}$$

- e.g.1
  - $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3]$
  - $f(\boldsymbol{\beta}) = (\beta_1 + \beta_2)\beta_3$
  - $\frac{\partial f(\beta)}{\partial \beta} = \begin{bmatrix} \beta_3 \\ \beta_3 \\ \beta_1 + \beta_2 \end{bmatrix}$
- e.g.2
  - $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3]$
  - $f(\boldsymbol{\beta}) = \beta^2_1\beta_2 + \log(\beta_3)$
  - $\frac{\partial f(\beta)}{\partial \beta} = \begin{bmatrix} 2\beta_1\beta_2 \\ \beta_1^2 \\ 1/\beta_3 \end{bmatrix}$

Matrix Differentiation

Let

$$\beta = [\beta_1, \beta_2, \cdots, \beta_k]'$$

$$f(\beta) = f([\beta_1, \beta_2, \cdots, \beta_k]')$$

define: the derivative of $f(.)$ wrt β

$$\frac{\partial f(\beta)}{\partial \beta} = \begin{bmatrix} \frac{\partial f(\beta)}{\partial \beta_1} \\ \frac{\partial f(\beta)}{\partial \beta_2} \\ \vdots \\ \frac{\partial f(\beta)}{\partial \beta_k} \end{bmatrix}$$

- e.g.3
  - **β** =[β_1, β_2, β_3]', **c**=[c_1,c_2,c_3]'
  - f(**β**)=**c**'**β**=$\sum c_i \beta_i$
  - $$\frac{\partial f(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta} c' \beta = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = c$$
- e.g. 4
  - **β** =[β_1, β_2, β_3]', **c**=[c_1,c_2,c_3]'
  - f(**β**)=**β'c**=$\sum \beta_i\ c_i$
  - $$\frac{\partial f(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta} \beta' c = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = c$$

Matrix Differentiation

- Remember
  - $$\frac{\partial}{\partial \beta} c' \beta = c$$
  - $$\frac{\partial}{\partial \beta} \beta' c = c$$
- e.g.5
  - **β** =[β_1, β_2, β_3]
  - f(**β**)= **β'M β**
  - By Product Rule,

$$\frac{\partial f(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta} \beta' \mathrm{M} \beta$$
$$= (\beta' \mathrm{M})' + \mathrm{M} \beta$$
$$= (\mathrm{M}' + \mathrm{M}) \beta$$

Key Results

$$\frac{\partial}{\partial \beta} c' \beta = \frac{\partial}{\partial \beta} \beta' c = c$$

$$\frac{\partial}{\partial \beta} \beta' \mathrm{M} \beta = (\mathrm{M}' + \mathrm{M}) \beta$$

Least Square estimator

- Least Square Estimator:
  - Minimizes

$$RSS(\beta) = (Y - X\beta)'(Y - X\beta) = Y'Y - 2Y'X\beta + \beta'X'X\beta$$

  - Find Minimum by differentiation

$$\frac{\partial}{\partial \beta} c'\beta = c \qquad \frac{\partial}{\partial \beta} \beta' \mathrm{M} \beta = (\mathrm{M}' + \mathrm{M})\beta$$

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2(Y'X)' + (X'X + (X'X)')\beta$$

$$= -2X'Y + 2X'X\beta$$

  - Set the derivative equal 0 gives

$$\boxed{\hat{\beta} = (X'X)^{-1}X'Y}$$

Probability Calculation of Matrix

- mx1 Random vector X

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_m \end{pmatrix}^T$$

- Mean

$$E(X) = E\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{pmatrix} = \begin{pmatrix} E(x_1) \\ E(x_2) \\ \dots \\ E(x_m) \end{pmatrix}$$

- Variance

$$Var(X) = E\left((X-\mu)(X-\mu)'\right) = \begin{pmatrix} Var(x_1) & Cov(x_1,x_2) & \dots & Cov(x_1,x_m) \\ Cov(x_2,x_1) & Var(x_2) & \dots & Cov(x_2,x_m) \\ \vdots & \vdots & \dots & \vdots \\ Cov(x_m,x_1) & \dots & \dots & Var(x_m) \end{pmatrix}$$

Probability Calculation of Matrix

- Mean
  - $$E(AX) = E\begin{pmatrix} \sum a_{1i}x_i \\ \sum a_{2i}x_i \\ \cdots \\ \sum a_{mi}x_i \end{pmatrix} = \begin{pmatrix} \sum a_{1i}E(x_i) \\ \sum a_{2i}E(x_i) \\ \cdots \\ \sum a_{mi}E(x_i) \end{pmatrix} = AE(X)$$

A is mxm constant matrix

X is mx1 random vector

- Variance
  - $$\begin{aligned} Var(AX) &= E([AX - E(AX)][AX - E(AX)]') \\ &= E([A(X - E(X))][A(X - E(X))]') \\ &= AE[(X - E(X))(X - E(X))']A' \\ &= AVar(X)A' \end{aligned}$$

$$E(AX) = AE(X)$$

$$Var(AX) = AVar(X)A'$$

Probability Calculation of Matrix

$$E(\mathrm{AX}) = \mathrm{AE}(\mathrm{X})$$

$$\mathrm{Var}(\mathrm{AX}) = \mathrm{AVar}(\mathrm{X})\mathrm{A}'$$

- Example
  - $\mathrm{A} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix},\ \mathrm{X} = \begin{pmatrix} \mathrm{X}_1 \\ \mathrm{X}_2 \end{pmatrix},\ \mathrm{E}(\mathrm{X}_1) = 5, E(\mathrm{X}_2) = 0, Var(\mathrm{X}_1) = 1, Var(\mathrm{X}_2) = 2,\ Cov(\mathrm{X}_1, \mathrm{X}_2) = 0.5$

Probability Calculation of Matrix

$$E(AX) = AE(X)$$

$$\mathrm{Var}(AX) = A\mathrm{Var}(X)A'$$

- Example
  - $A = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix},\ X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},\ E(X_1) = 5, E(X_2) = 0, Var(X_1) = 1, Var(X_2) = 2,\ Cov(X_1, X_2) = 0.5$
  - Method 1 (First principle)

$$AX = \begin{pmatrix} X_1 \\ 2X_1 + X_2 \end{pmatrix},\ E(AX) = \begin{pmatrix} 5 \\ 10 \end{pmatrix},\ \mathrm{Var}(AX) = \begin{pmatrix} Var(X_1) & \mathrm{Cov}(X_1, 2X_1 + X_2) \\ \mathrm{Cov}(X_1, 2X_1 + X_2) & Var(2X_1 + X_2) \end{pmatrix} = \begin{pmatrix} 1 & 2.5 \\ 2.5 & 8 \end{pmatrix}$$

  - Method 2 (Using formula)

$$E(AX) = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 5 \\ 0 \end{pmatrix} = \begin{pmatrix} 5 \\ 10 \end{pmatrix},\ \mathrm{Var}(AX) = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}' = \begin{pmatrix} 1 & 0.5 \\ 2.5 & 3 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2.5 \\ 2.5 & 8 \end{pmatrix}$$

Properties Least Square estimator

- Model

$$Y = X\beta + e, \qquad \mathrm{E}(e) = 0, Var(e) = \sigma^2 I$$

- Least Square Estimator (LSE)

$$\boxed{\hat{\beta} = (X'X)^{-1} X'Y}$$

- Mean of LSE (Unbiasedness: mean of estimate= truth)

$$E(\hat{\beta}) = (X'X)^{-1} X'E(Y) = (X'X)^{-1} X'E(X\beta + e)$$
$$= (X'X)^{-1} X'X\beta = \beta$$

- Variance of LSE

$$Var(\hat{\beta}) = Var[(X'X)^{-1} X'Y] = (X'X)^{-1} X'Var(Y)X(X'X)^{-1}$$
$$= (X'X)^{-1} X'(\sigma^2 I)X(X'X)^{-1} = \sigma^2 (X'X)^{-1}(X'X)(X'X)^{-1}$$
$$= \sigma^2 (X'X)^{-1}$$

A Matrix operator -- Trace

- Trace (tr) is the sum of diagonal element of a Square matrix
  - $$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{pmatrix}$$

    $$tr(A) = \sum_{i=1}^{m} a_{ii}$$
- Properties
  - $$tr(A+B) = \sum_{i=1}^{m} a_{ii} + b_{ii} = tr(A) + tr(B)$$
  - $$tr(AB) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{m} b_{ji}a_{ij} = tr(BA)$$

    Diagonal of AB **Diagonal of BA**
  - $$tr(E(A)) = \sum E(a_{ii}) = E(\sum a_{ii}) = E(tr(A))$$

Properties Least Square estimator

- Model $Y = X\beta + e, \qquad \mathrm{E}(e) = 0, Var(e) = \sigma^2 I$
- Residual sum of square

$$RSS(\hat\beta) = (Y - X\hat\beta)'(Y - X\hat\beta) = Y'Y - 2Y'X\hat\beta + \hat\beta' X' X\hat\beta$$
$$= Y'Y - Y'X(X'X)^{-1}X'Y$$
$$= Y'(I - X(X'X)^{-1}X')Y$$

- Note that

$$E(Y'AY) = E(tr(Y'AY)) = E(tr(AYY')) = tr(AE(YY'))$$
$$= tr(AE[(X\beta + e)(X\beta + e)']) = tr(A(X\beta\beta'X' + \sigma^2 I))$$
$$= tr(A(X\beta\beta'X')) + \sigma^2 tr(A)$$

- Put A = I-H = I-X(X'X)$^{-1}$X'
  - $\mathrm{tr}(AX\beta\beta'X) = (I_n - X(X'X)^{-1}X')X\beta\beta'X = (X - X)\beta\beta'X = 0$
  - $\mathrm{tr}(A) = tr(I_n - X(X'X)^{-1}X') = tr(I_n) - tr(X(X'X)^{-1}X')$

$$= tr(I_n) - tr((X'X)^{-1}X'X) = tr(I_n) - tr(I_{p+1}) = n - (p+1)$$

$$E(RSS(\hat\beta)) = \sigma^2(n - (p+1)) \quad \Rightarrow \quad \hat\sigma^2 = \frac{RSS(\hat\beta)}{n - (p+1)}$$

Distributional properties

- Distribution of $\hat{\beta}$
  - $$\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \cdots\\ \hat{\beta}_p\end{pmatrix} = \hat{\beta} = (X'X)^{-1}X'\begin{pmatrix}Y_1\\ Y_2\\ \cdots\\ Y_n\end{pmatrix} = \begin{pmatrix}\sum M_{1i}Y_i\\ \sum M_{2i}Y_i\\ \cdots\\ \sum M_{(p+1)i}Y_i\end{pmatrix} \sim N(\beta, \sigma^2(X'X)^{-1})$$

    Sum of independent variables

- Distribution of $\hat{\sigma}^2$
  - $$\hat{\sigma}^2 = \frac{RSS(\hat{\beta})}{n-(p+1)} = \frac{\sum \hat{e}_i^2}{n-(p+1)} \sim \frac{\sigma^2 \chi^2_{n-p-1}}{n-(p+1)}$$

    Sum of squares of c-normal variables

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} = \chi^2_{n-p-1}$$

3.1.2. Added-Variable plot

- In Multiple linear regression, plotting graph is difficult

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

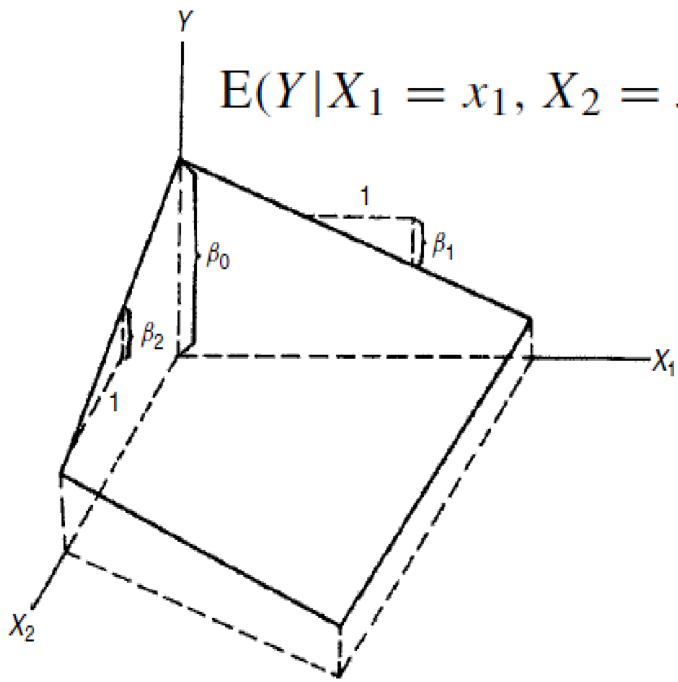


FIG. 3.2 A linear regression surface with $p = 2$ predictors.

- Is there any 2-d way to see the effect of the βs?

3.1.2. Added-Variable plot

- An interesting observation from a computer experiment
  - Data generation
    - x1=rnorm(100); x2=rnorm(100); e=rnorm(100,0,0.1); y=3*x1+2*x2+e
  - Fit regression
    - Fit1=lm(y~x1+x2); summary(Fit1)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.007537   0.010466   -0.72    0.473
x1           3.019023   0.010082  299.44   <2e-16 ***
x2           1.996044   0.010182  196.04   <2e-16 ***
```

    - Fit2=lm(y~x2); summary(Fit2)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1800     0.3163  -0.569     0.57
x2            2.4254     0.3051   7.950 3.24e-12 ***
```

    - Fit3a=lm(y~x1); Fit3b=lm(x2~x1); Fit3c=lm(Fit3a$residual~Fit3b$residual); summary(Fit3c)

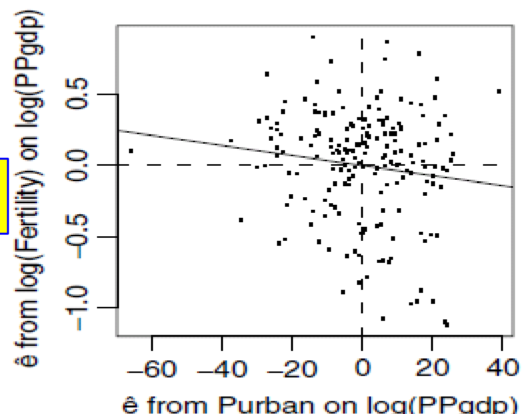
```
Coefficients:
                 Estimate Std. Error   t value Pr(>|t|)
(Intercept)    -2.975e-17  1.037e-02 -2.87e-15        1
Fit3b$residual  1.996e+00  1.013e-02     197.1   <2e-16 ***
```

β_2 measures the relationship between y and x2, after adjusting for the effect of x1

3.1.2. Added-Variable plot

- Added-variable plot (for x_i)
  - Vertical: Residual of the regression of y on all predictors except x_i
  - Horizontal: Residual of the regression of x_i on all other predictors.

$\hat{e}_{Y|X_O}$



$\hat{e}_{X_i|X_O}$

$$Y = X_1\beta_1 + \ldots + \underline{X_i\beta_i} + \ldots + X_p\beta_p + e$$
$$\Rightarrow Y = \underline{X_i\beta_i} + X_O\beta_O + e$$

- Properties
  - The slope equal to $\hat{\beta}_i$ in multiple regression.
    - can see the effect of x after adjusted for the effect of other predictors
    - The plot gives more information than the coefficient in multiple regression.
  - May have different magnitude, sign and significance compare to $\hat{\beta}_i$ in simple linear regression.

3.1.2. Added-Variable plot

- Theory: Why the slope in the added-variable plot = $\hat{\beta}_i$?
- To see why, study the residuals
- The hat matrix H=X(X'X)$^{-1}$X':

$$\hat{e} = Y - \hat{Y} = \left((Y_1 - \hat{Y}_1) \quad (Y_1 - \hat{Y}_1) \ldots (Y_n - \hat{Y}_n)\right)^T$$
$$= Y - X\hat{\beta}$$
$$= Y - X(X'X)^{-1}X'Y$$
$$= (1 - X(X'X)^{-1}X')Y$$
$$= (1 - H)Y$$

- When fitting $Y = X\beta + e$, we have

$$\hat{e} = (1 - H)Y$$

$$(1 - H)X = X - X(X'X)^{-1}X'X = X - X = 0$$

3.1.2. Added-Variable plot

- Added-variable plot (for x_1)
  - Vertical: Residual of the regression of y on all predictors except x_1
  - Horizontal: Residual of the regression of x_1 on all other predictors.

• **Setting**

$$Y = X_1\beta_1 + X_O\beta_O + e$$

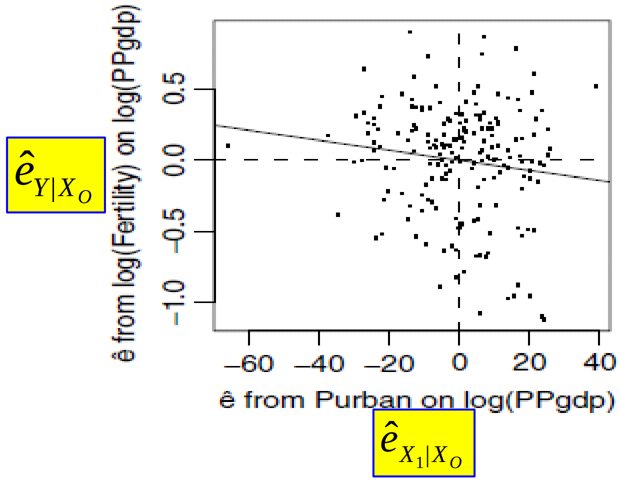
$$H_{Oth} = X_O(X_O^{'}X_O)^{-1}X_O^{'}$$

• **Properties**

$$(I - H_{Oth})X_O = 0$$

$$\hat{e}_{Y|X_O} = (I - H_{Oth})Y$$

$$\hat{e}_{X_1|X_O} = (I - H_{Oth})X_1$$



3.1.2. Added-Variable plot

- Theory: Why the slope in the added-variable plot = $\hat{\beta}_1$?
- Idea
  - Properties $(I - H_{Oth})X_O = 0$

$$\hat{e}_{Y|X_O} = (I - H_{Oth})Y$$

$$\hat{e}_{X_1|X_O} = (I - H_{Oth})X_1$$

  - Adjust for X_O : $Y = X_1\beta_1 + X_O\beta_O + e$

$$\Rightarrow (I - H_{Oth})Y = (I - H_{Oth})X_1\beta_1 + (I - H_{Oth})X_O\beta_O + (I - H_{Oth})e$$

$$= (I - H_{Oth})X_1\beta_1 + \underline{\qquad 0 \qquad} + (I - H_{Oth})e$$

$$\Rightarrow \qquad \hat{e}_{Y|X_O} = \hat{e}_{X_1|X_O}\beta_1 + \tilde{e}$$

  - where $\tilde{e} = (1 - H_{Oth})e$
  - Therefore β_1 is the regression coefficient of $\hat{e}_{Y|X_O}$ against $\hat{e}_{X_1|X_O}$

$$\hat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$$

from chapter 2

$$\hat{\beta}_1 = \frac{\sum \hat{e}_{Y|X_O}\hat{e}_{X_1|X_O}}{\sum \hat{e}^2_{X_1|X_O}} = \frac{X_1'(I - H_{Oth})Y}{X_1'(I - H_{Oth})X_1}$$

2.6 Comparing models: Analysis of variance (ANOVA)

- Regression is the study of dependence of variables
  - $y_i = \beta_0 + \beta_1 x_i + e_i$
    - $\beta_1 = 0 \rightarrow$ x and y are independent
    - $\beta_1 \neq 0 \rightarrow$ x and y are dependent
- Question:
  - Are x and y dependent?
- Answer:
  - Method 1) test whether $\beta_1 = 0$
  - Method 2) Compare the two models
    - $E(y|x) = \beta_0$ i.e. $y_i = \beta_0 + e_i$
    - $E(y|x) = \beta_0 + \beta_1 x$ i.e. $y_i = \beta_0 + \beta_1 x_i + e_i$

3.5. Comparing models: Analysis of variance (ANOVA)

- Regression is the study of dependence of variables
  - $Y= X_1\beta_1 + X_O\beta_O+e$,
    - $\boldsymbol{\beta}_1=0 \rightarrow X_1$ and y are independent
    - $\boldsymbol{\beta}_1\neq 0 \rightarrow X_1$ 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. $Y= X_1\beta_1 +e$
    - $E(Y|X)= X_1\beta_1 + X_O\beta_O$ i.e. $Y= X_1\beta_1 + X_O\beta_O +e$

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: $E(Y|X)= X_O \beta_O$
  - $RSS_{NH}= \min_{\beta_O}\sum(Y_i - X_{Oi}\beta_O)^2 \stackrel{def}{=} \sum(Y_i - X_{Oi}\tilde{\beta}_O)^2$
- For the second model: $E(y|x)=X_1 \beta_1+X_O \beta_O$
  - $RSS_{AH}= \min_{\beta_1,\beta_O}\sum(Y_i - X_{1i}\beta_1 - X_{Oi}\beta_O)^2 \stackrel{def}{=} \sum(Y_i - X_{1i}\hat{\beta}_1 - X_{Oi}\hat{\beta}_O)^2$
- 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}: $\sum (Y_i - X_{Oi}\tilde{\beta}_O)^2$
  - RSS_{AH}: $\sum (Y_i - X_{1i}\hat{\beta}_1 - X_{Oi}\hat{\beta}_O)^2$
  - 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}

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

3.5 A Special Case
Overall Analysis of variance (ANOVA)

- Difference sum of square due to regression
    - NH: E(Y|X)= β_0
    - AH: E(Y|X)= Xβ (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

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

3.5 Comparing models: Analysis of variance (ANOVA)

- 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			

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

$$TSS = SYY = RSS + SSreg$$

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

3.5 Comparing models: Analysis of variance (ANOVA)

TABLE 3.4 The Overall Analysis of Variance Table

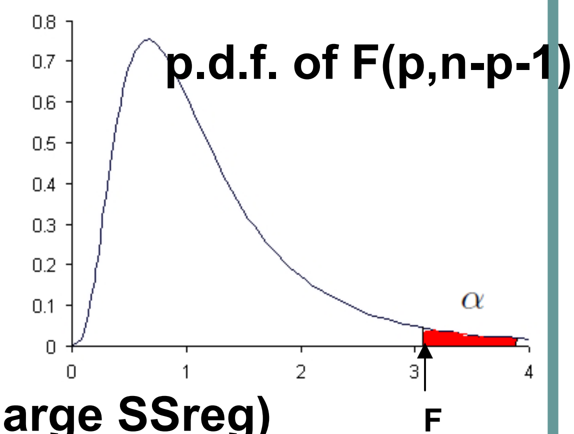
Source	df	SS	MS	F	p-value
Regression	p	$SSreg$	$SSreg/p$	$MSreg/\hat{\sigma}^2$	
Residual	$n-(p+1)$	RSS	$\hat{\sigma}^2 = RSS/(n-(p+1))$		
Total	$n-1$	SYY			

F test for Regression

$$NH : E(Y \mid X = x) = \beta_0$$

$$AH : E(Y \mid X = x) = X\beta$$

F Statistic= $MSreg/\hat{\sigma}^2$ **~ F(p,n-p-1) under NH**



Idea:

- **larger F means regression is effective (large SSreg)**
- **Under NH, F~F(p,n-p-1), it is unlikely to be very big**
- **If the red area (α) is small, F is large → NH is suspicious**

α is the p-value = P(observing a test stat more extreme than F)
If p-value is small, e.g. <0.05, we reject the NH

3.5 Coefficient of Determination, R^2

- Definition

$$R^2 = \frac{SSreg}{SYY}$$

  - Proportion of variability explained by regression
- Scale-free one number summary of strength of relationship between X and Y.
- Connections to the correlation b/w Y and $\hat{Y}$

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \left[\frac{\sum (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y})}{\sqrt{\sum (\hat{Y}_i - \bar{Y})^2 \sum (Y_i - \bar{Y})^2}} \right]^2$$

$$\begin{aligned} &\sum (\hat{Y}_i - \bar{Y})^2 \\ &= \sum (\hat{Y}_i - Y_i + Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) \\ &= \sum (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y}) \end{aligned}$$

- R^2 is always between 0 and 1.
  - Close to 1→ good fit
  - Close to 0→ bad fit

3.5 Example: Fuel Consumption

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	4	201994	50499	11.992	9.33e-07
Residuals	46	193700	4211		
Total	50	395694			

- 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.

Confidence intervals and tests

- Regression model:
  - $E(Y|X=x)= \beta_0 + \beta_1 x_1+\ldots+\beta_p x_p$
- Quantities of interests
  - Intercept: β_0
  - Effect of x_k: β_k
  - Prediction: If we observe $\mathbf{x}_*$, what is the y?
  - Fitted value: $E(Y|\mathbf{X}=\mathbf{x})$ for different values of $\mathbf{x}$
- Confidence intervals give estimates for the above quantities of interests

3.5 Testing one of the terms

- Natural question to ask:
  - Is the k-th variable dependent on y? (after adjusting for the effect of other predictors)
- T-test: (wlog, k=1)

$$\begin{array}{lll} \text{NH:} & \beta_1 = 0, & \beta_0, \beta_2, \beta_3, \beta_4 \text{ arbitrary} \\ \text{AH:} & \beta_1 \neq 0, & \beta_0, \beta_2, \beta_3, \beta_4 \text{ arbitrary} \end{array}$$

  - Recall
    - $\hat{\beta} \sim N(0, \sigma^2 (X'X)^{-1})$
    - T-statistics

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 V_1}} \sim t(n-p-1)$$

where V_1 is the (2,2) entry of $(X'X)^{-1}$

{The (1,1) entry is the variance of $\hat{\beta}_o$}

- Confidence interval

$$\boxed{\hat{\beta}_1 \pm t(n-p-1)\hat{\sigma}\sqrt{V_1}}$$

Two choices in testing β_k=0!

- T-test of coefficient
  - NH: $\beta_k=0$
  - AH: $\beta_k \neq 0$
- Equivalent to F-test of comparing
  - NH: $y_i= \beta_0 + \beta_1 x_1+\ldots+ \beta_{k-1} x_{k-1}+\_\_\_\_\_+ \beta_{k+1} x_{k+1}+\ldots+\beta_p x_p+ e_i$
  - AH: $y_i= \beta_0 + \beta_1 x_1+\ldots+ \beta_{k-1} x_{k-1}+ \beta_k x_k + \beta_{k+1} x_{k+1} +\ldots +\beta_p x_p+ e_i$
- T-stat

$$T = \frac{\hat{\beta}_k}{sd(\hat{\beta}_k)} \sim t(n-p-1)$$

- F-stat

$$F = \frac{(RSS_{NH} - RSS_{AH})/(df_{NH} - df_{AH})}{RSS_{AH}/df_{AH}} = \frac{SSreg/1}{\hat{\sigma}^2} \sim F(1, n-p-1)$$

T-test = F-test in testing β_k=0!

- F-stat=(t-stat)^2 for testing β_2=0
  - Data generation
    - x1=rnorm(100); x2=rnorm(100); e=rnorm(100,0,0.1); y=3*x1+2*x2+e
  - Fit regression
    - Fit.NH=lm(y~x1); summary(Fit.NH)

```
Residual standard error: 1.792 on 98 degrees of freedom
```

    - Fit.AH=lm(y~x1+x2); summary(Fit.AH)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.008814   0.010373   -0.85    0.398
x1           3.010027   0.009787  307.56   <2e-16 ***
x2           2.009821   0.011504  174.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1014 on 97 degrees of freedom
```

$$\hat{\sigma} = \sqrt{\frac{RSS_{AH}}{df_{AH}}}$$

    - t-stat=174.6 ~ t(97)
    - F-stat=(1.792^2*98-0.1014^2*97)/0.1014^2=30510=(174.7)^2=(t-stat)^2

$$F = \frac{(RSS_{NH} - RSS_{AH})/(df_{NH} - df_{AH})}{RSS_{AH}/df_{AH}} = \frac{SSreg/1}{\hat{\sigma}^2} \sim F(1, n-p-1)$$

T-test = F-test in testing β_k=0!

- Theory: F-stat=(t-stat)^2 (optional)

$$H_o \Rightarrow Y = \tilde{\beta}_o + \tilde{\beta}_1 X_1 + ... + \tilde{\beta}_{k-1} X_{k-1} + \_\_\_\_ + \tilde{\beta}_{k+1} X_{k+1} ... + \tilde{\beta}_p X_p \overset{def}{=} X_O \tilde{\beta}_O$$

$$H_A \Rightarrow Y = \hat{\beta}_o + \hat{\beta}_1 X_1 + ... + \hat{\beta}_{k-1} X_{k-1} + \hat{\beta}_k X_k + \hat{\beta}_{k+1} X_{k+1} ... + \hat{\beta}_p X_p \overset{def}{=} X_O \hat{\beta}_O + X_k \hat{\beta}_k$$

  - For F-test, consider SSreg

$$Let \quad H_{all} = X(X'X)^{-1}X', H_k = X_k(X_k'X_k)^{-1}X_k', H_{Oth} = X_O(X_O'X_O)^{-1}X_O'$$

$$RSS_{H_o} = (Y - H_{Oth}Y)'(Y - H_{Oth}Y) = Y'(I - H_{Oth})Y$$

$$RSS_{H_A} = (Y - H_{all}Y)'(Y - H_{all}Y) = Y'(I - H_{all})Y$$

$$SSreg = RSS_{H_o} - RSS_{H_A} = Y'(H_{all} - H_{oth})Y$$

  - For t-test, consider $\hat{\beta}_k$

- The theory of added-variable plot tells us that

$$\hat{\beta}_k = \frac{\sum \hat{e}_{Y|X_O} \hat{e}_{X_k|X_O}}{\sum \hat{e}^2_{X_k|X_O}} = \frac{X_k'(I - H_{Oth})Y}{X_k'(I - H_{Oth})X_k}$$

$$Var(\hat{\beta}_k) = \frac{\hat{\sigma}^2}{X_k'(I - H_{Oth})X_k}$$

k

Coefficient of the added-variable X_k is the regression coefficient between the two residuals: $\hat{e}_{Y|X_O}$ **and** $\hat{e}_{X_k|X_O}$

T-test = F-test in testing $\beta_k=0$!

- Theory: F-stat=(t-stat)^2 (optional)

Model : $Y = X_1\beta_1 + X_O\beta_O + e = X\beta + e$

Using all variables : $\hat{Y} = H_{all}Y$

Idea of added variables :

$$(1-H_{Oth})Y = \hat{e}_{Y|X_{Oth}} = (1-H_{Oth})X_1\beta_1 + (1-H_{Oth})e$$

$$\Rightarrow \quad \hat{E}(\hat{e}_{Y|X_{Oth}} \mid X_1) = (1-H_{Oth})X_1 \frac{X_1'(I-H_{Oth})Y}{X_1'(I-H_{Oth})X_1}$$

$$\hat{\beta}_k = \frac{X_k'(I-H_{Oth})Y}{X_k'(I-H_{Oth})X_k}$$

$$H_{all}Y = \hat{Y} = H_{oth}Y + \hat{E}(\hat{e}_{Y|X_{Oth}} \mid X_1) = \left[H_{oth} + (1-H_{Oth})X_1 \frac{X_1'(I-H_{Oth})}{X_1'(I-H_{Oth})X_1}\right]Y$$

F stat:

$$F = \frac{SSreg}{\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^2}Y'(H_{all}-H_{oth})Y$$

$$= \frac{1}{\hat{\sigma}^2}Y'(1-H_{Oth})X_1 \frac{X_1'(I-H_{Oth})}{X_1'(I-H_{Oth})X_1}Y$$

$$= \frac{\left(X_1'(I-H_{Oth})Y\right)^2}{\hat{\sigma}^2 X_1'(I-H_{Oth})X_1} = \frac{\hat{\beta}_1^2}{Var(\hat{\beta}_1)} = t^2$$

$$Var(\hat{\beta}_k) = \frac{\hat{\sigma}^2}{X_k'(I-H_{Oth})X_k}$$

$$F(1,m) = \frac{\chi_1^2}{\chi_m^2/m} = \left[\frac{N(0,1)}{\sqrt{\chi_m^2/m}}\right]^2 = t^2(m)$$

3.6 Confidence intervals and tests

- Prediction: If we observe x_*, what is the y_*?
- Prediction:

$$\tilde{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_{*1} + \ldots + \hat{\beta}_p x_{*p} = \boldsymbol{x}'_* \hat{\boldsymbol{\beta}}$$

- Prediction uncertainty
  - (predicting a particular observation incorporate the error, giving σ^2)

$$Var(\tilde{y}_* + e_* \mid \boldsymbol{x}_*) = Var(\boldsymbol{x}'_* \hat{\boldsymbol{\beta}} \mid \boldsymbol{x}_*) + \sigma^2 = \sigma^2 \boldsymbol{x}'_* Var(\hat{\boldsymbol{\beta}} \mid \boldsymbol{x}_*) \boldsymbol{x}_* + \sigma^2$$
$$= \sigma^2 \left(1 + \boldsymbol{x}'_* (X'X)^{-1} \boldsymbol{x}_*\right)$$

- Prediction interval for y_* (pointwise)

$$\tilde{y}_* \pm t\left(\frac{\alpha}{2}, n-p-1\right) \hat{\sigma} \sqrt{1 + \boldsymbol{x}'_* (X'X)^{-1} \boldsymbol{x}_*}$$

3.6 Confidence intervals and tests

- Fitted value: E(Y|X=x) for different values of x
- Estimation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p = \boldsymbol{x}' \hat{\boldsymbol{\beta}}$$

- Estimation uncertainty
  - It is not a prediction, no need the error term, no σ^2

$$Var(\hat{y} \mid \boldsymbol{x}) = \sigma^2 \boldsymbol{x}' Var(\hat{\boldsymbol{\beta}} \mid \boldsymbol{x}) \boldsymbol{x} = \sigma^2 \left(\boldsymbol{x}' (X'X)^{-1} \boldsymbol{x} \right)$$

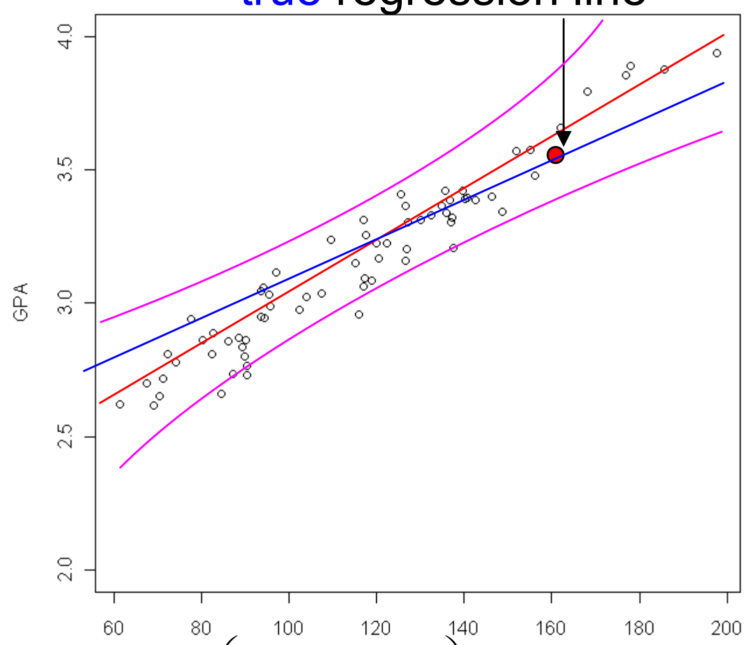
- Confidence interval for E(Y|X=x): (pointwise)

$$\hat{y} \pm t\left(\frac{\alpha}{2}, n-p-1\right) \hat{\sigma} \sqrt{\boldsymbol{x}' (X'X)^{-1} \boldsymbol{x}}$$

3.6 Confidence intervals

- Fitted value: E(Y|X=x)

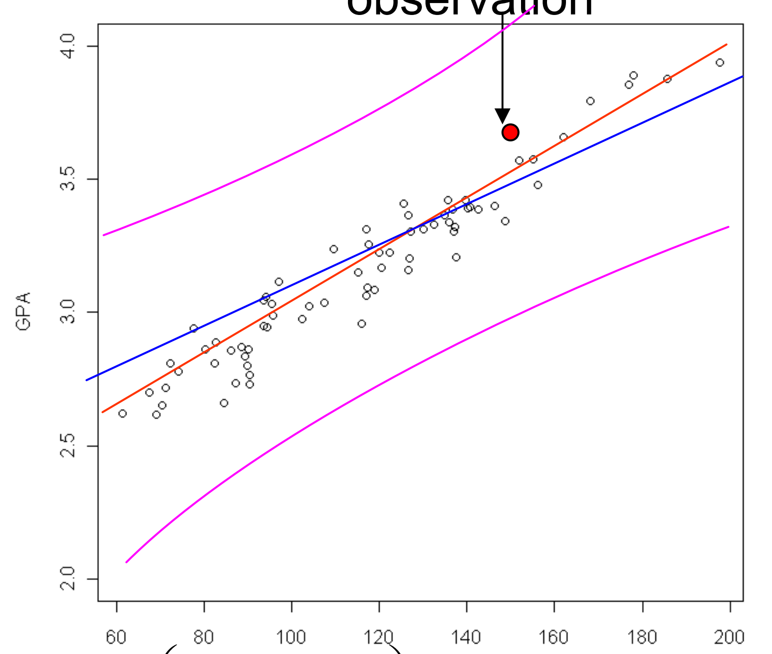
C.I. for a point on the
true regression line



$$\hat{y} \pm t\left(\frac{\alpha}{2}, n-p-1\right)\hat{\sigma}\sqrt{\boldsymbol{x}'(X'X)^{-1}\boldsymbol{x}}$$

- Prediction

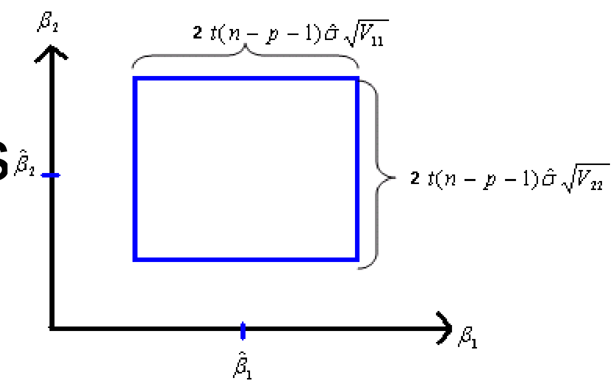
P.I. for a future
observation



$$\tilde{y}_* \pm t\left(\frac{\alpha}{2}, n-p-1\right)\hat{\sigma}\sqrt{1+\boldsymbol{x}_*'(X'X)^{-1}\boldsymbol{x}_*}$$

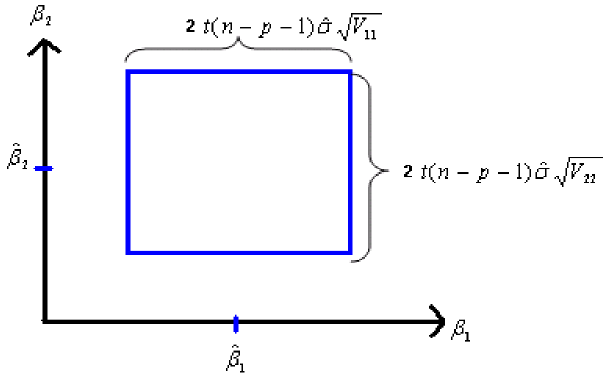
5.5 Joint Confidence Region

- C.I. for β_1:
  - $P(\hat{\beta}_1 - t(n-p-1)\hat{\sigma}\sqrt{V_{11}} \le \beta_1 \le \hat{\beta}_1 + t(n-p-1)\hat{\sigma}\sqrt{V_{11}}) = 1-\alpha$
- C.I. for β_2:
  - $P(\hat{\beta}_2 - t(n-p-1)\hat{\sigma}\sqrt{V_{22}} \le \beta_2 \le \hat{\beta}_2 + t(n-p-1)\hat{\sigma}\sqrt{V_{22}}) = 1-\alpha$
- Question:
  - Does the rectangle covers the truth (β_1, β_2) with probability $1-\alpha$?



5.5 Joint Confidence Region

- Question:
  - Does the rectangle covers the truth (β_1, β_2) with probability 1-α?
  - No…between [1-2α, 1- α]
    - Let A_i={C.I. β_i of covers β_i}
    - $P(A_i)=1-\alpha$ for i=1 and 2
    - $P(A_1 \cap A_2)=P(A_1)+P(A_2)-P(A_1 \cup A_2)$
      $=2(1-\alpha) - P(A_1 \cup A_2)$ ← [1-α,1]
      $\in[1-2\alpha, 1-\alpha]$



Question: How to find a region that covers all parameters with prob1-α?

5.5 Joint Confidence Region

- Answer: (1-α) Confidence ellipse

$$\frac{(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)}{(p+1)\hat{\sigma}^2} \le F(\alpha, p+1, n-p-1)$$

$N(0, \sigma^2(X'X)^{-1})$

$\eta = \frac{1}{\sigma}(X'X)^{1/2}(\hat{\beta}-\beta) \sim N(0, I_{p+1})$

- Idea (optional)

1. $(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta) \approx \sigma^2 \eta'\eta = \sigma^2 \sum_{i=1}^{p+1} \eta_i^2 \sim \sigma^2 \chi^2_{p+1}$

2. $\hat{\sigma}^2 = \frac{1}{n-p-1}\sum (y_i - \hat{y}_i)^2 \sim \frac{\sigma^2 \chi^2_{n-p-1}}{n-p-1}$

3. $\frac{\chi^2_{p+1}/(p+1)}{\chi^2_{n-p-1}/(n-p-1)} \sim F(p+1, n-p-1)$

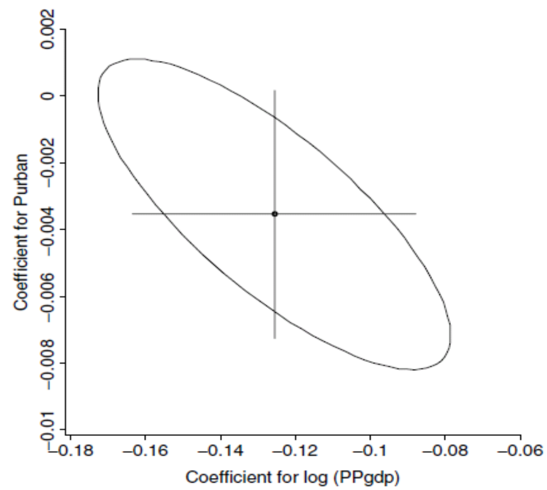
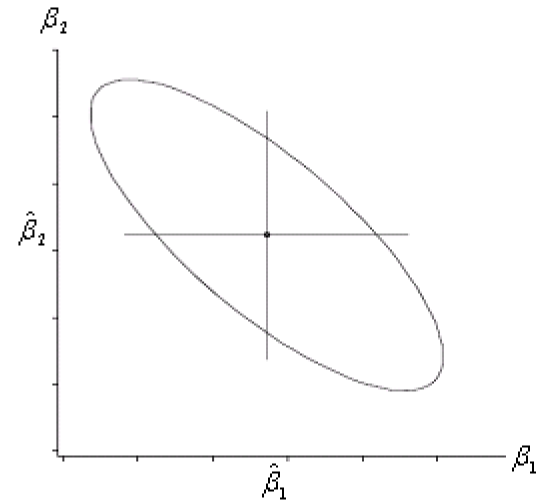


FIG. 5.3 95% confidence region for the UN data.

5.5 Joint Confidence Region

- Example:
  - $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2, n) = (2,3,1,10), \quad (X'X) = \begin{pmatrix} 2 & 1 \\ 1 & 5 \end{pmatrix}$



- (1-α) Confidence ellipse

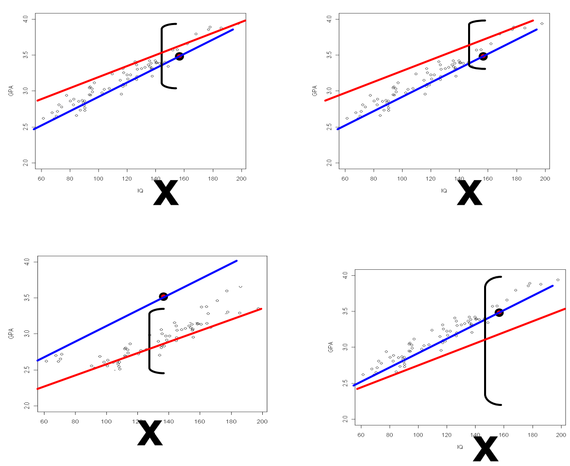
$$\frac{(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)}{(p+1)\hat{\sigma}^2} \le F(\alpha, p+1, n-p-1)$$

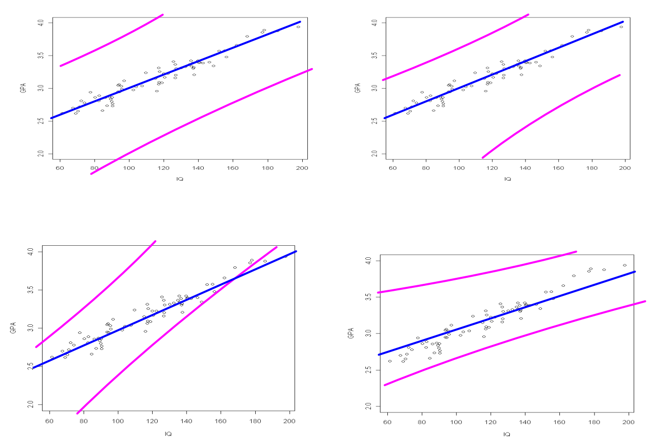
$$\Rightarrow \frac{(2-\beta_0 \quad 3-\beta_1)\begin{pmatrix} 2 & 1 \\ 1 & 5 \end{pmatrix}\begin{pmatrix} 2-\beta_0 \\ 3-\beta_1 \end{pmatrix}}{(1+1)(1)} \le F(0.05, 2, 10-2)$$

$$\Rightarrow 2\beta_0^2 + 2\beta_0\beta_1 + 5\beta_1^2 - 14\beta_0 - 34\beta_1 + 65 \le 2(4.459)$$

3.6 Confidence intervals and bands

- Confidence interval (at each point x)
  - For each of x, P(E(Y|X=x) in C.I.)=1-α
- Confidence band (for the entire line)
  - P(For all x, E(Y|X=x) in C.B.)= 1-α



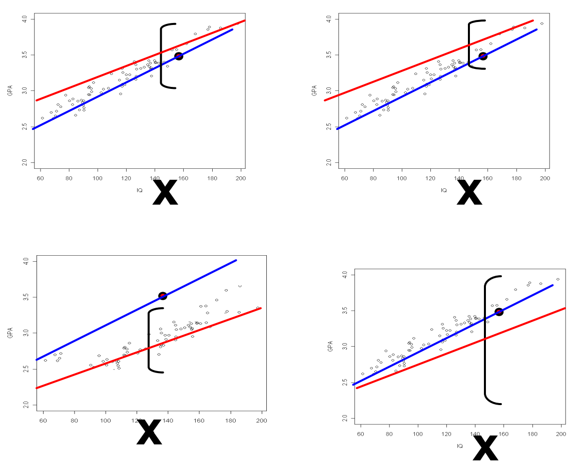


For n C.I.s, n(1-α) of them covers the true value at x

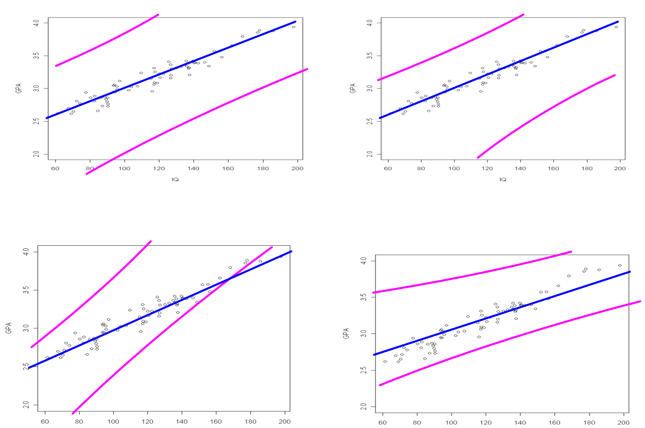
For n C.B.s, n(1-α) of them covers the whole true regression line

3.6 Confidence intervals and bands

- Confidence interval for E(Y|X)
  - $\hat{y} \pm t\left(\frac{\alpha}{2}, n-p-1\right)\hat{\sigma}\sqrt{\boldsymbol{x}'(X'X)^{-1}\boldsymbol{x}}$
- Confidence band for E(Y|X)
  - $\hat{y} \pm \sqrt{(p+1)F(\alpha, p+1, n-p-1)}\hat{\sigma}\sqrt{\boldsymbol{x}'(X'X)^{-1}\boldsymbol{x}}$



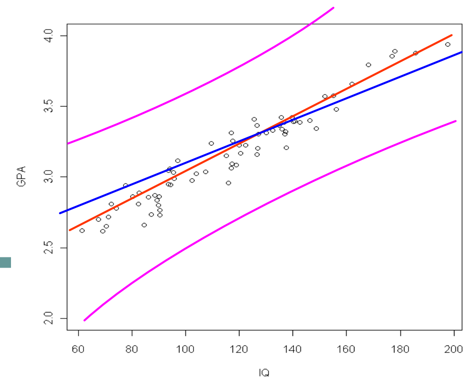
For n C.I.s, n(1-α) of them covers the true value at x



For n C.B.s, n(1-α) of them covers the whole true regression line

3.6 Confidence intervals and bands

- Confidence band for E(Y|X) (Idea)
  - $$\hat{y} \pm \sqrt{(p+1)F\left(\alpha, p+1, n-p-1\right)}\hat{\sigma}\sqrt{\boldsymbol{x}'(X'X)^{-1}\boldsymbol{x}}$$
  - P(For all x, E(Y|X=x) in C.B.)= 1-α
  - $$P\left(x'\beta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \text{ in C.B. for all } x = (1, x_1, \ldots, x_p)'\right) = 1-\alpha$$
    $$\Rightarrow \text{One possibility:}$$
    $$P(x'\hat{\beta} - \text{Error bound} \le x'\beta \le x'\hat{\beta} + \text{Error bound}, \text{ all } x) = 1-\alpha$$
    $$\Rightarrow P(\max_x |x'\beta - x'\hat{\beta}| \le \text{Error bound}) = 1-\alpha$$
    $$\Rightarrow \mathit{Study} \ \max_x x'(\beta - \hat{\beta})$$



3.6 Confidence intervals and bands

- What is this inequality?
  - $$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \le \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2 \qquad \text{or} \qquad (x'y)^2 = \le (x'x)(y'y)$$
  - Proof

$$(x+ty)'(x+ty) \ge 0$$

$$\Rightarrow (y'y)t^2 + 2tx'y + x'x \ge 0$$

$$\Rightarrow \text{Determinant} < 0 \text{ gives the result}$$

  - Note that equality holds iff $x + ty = 0$

3.6 Confidence intervals and bands

- Cauchy Schwartz Inequality

$$(x'y)^2 \le (x'x)(y'y)$$

- Proof of C.B. --- A Super trick! (optional)

$$Put \quad x = (X'X)^{-\frac{1}{2}}\mathrm{x}, \quad y = (X'X)^{\frac{1}{2}}(\hat{\beta}-\beta), \text{ we have}$$

$$\Rightarrow \left(\mathrm{x}'(\hat{\beta}-\beta)\right)^2 \le \left(\mathrm{x}'(X'X)^{-1}\mathrm{x}\right)\left((\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)\right) \quad \text{for any x,}$$

$$\Rightarrow \max_{\mathrm{x}} \frac{\left(\mathrm{x}'(\hat{\beta}-\beta)\right)^2}{\mathrm{x}'(X'X)^{-1}\mathrm{x}} = (\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)$$

$$\Rightarrow P\left\{\max_{\mathrm{x}} \frac{\left(\mathrm{x}'(\hat{\beta}-\beta)\right)^2}{(p+1)\hat{\sigma}^2\mathrm{x}'(X'X)^{-1}\mathrm{x}} \le F^*\right\} = P\left\{\frac{(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)}{(p+1)\hat{\sigma}^2} \le F^*\right\} = 1-\alpha$$

$$\Rightarrow P\left\{ \mathrm{x}'\beta \in \left[\mathrm{x}'\hat{\beta} \pm \sqrt{(p+1)F^*}\,\hat{\sigma}\sqrt{\mathrm{x}'(X'X)^{-1}\mathrm{x}}\right] \text{ for any x}\right\} = 1-\alpha$$

$$F^* = F(\alpha, p+1, n-p-1)$$

3.6 Confidence intervals and bands

- Formula:
  - Prediction Interval. $\hat{y}_* \pm t\left(\frac{\alpha}{2}, n-p-1\right)\hat{\sigma}\sqrt{1 + \boldsymbol{x}_*'(X'X)^{-1}\boldsymbol{x}_*}$
- Example 1
  - Data generation
    - x1=rnorm(100); x2=rnorm(100); e=rnorm(100,0,0.1); y=3*x1+2*x2+e
  - Prediction Interval at x=(1,x_1,x_2)=(1,0,2)
    - X=cbind(1,x1,x2); V=t(X)%*%X; x.p=c(1,0,2)
    - Fit.AH=lm(y~x1+x2); summary(Fit.AH)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.008814   0.010373   -0.85    0.398
x1           3.010027   0.009787  307.56   <2e-16 ***
x2           2.009821   0.011504  174.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1014 on 97 degrees of freedom
```

    - Fit.AH$coef%*%x.p-qt(0.975,97)*0.1014*sqrt(1+x.p%*%solve(V)%*%x.p)
      - lower limit = `[,1]` `[1,] 3.923899`
    - Fit.AH$coef%*%x.p+qt(0.975,97)*0.1014*sqrt(1+x.p%*%solve(V)%*%x.p)
      - upper limit = `[,1]` `[1,] 4.097758`

3.6 Confidence intervals and bands

- Formula:
  - C.I. for fitted value $$\hat{y} \pm t\left(\frac{\alpha}{2}, n-p-1\right)\hat{\sigma}\sqrt{\boldsymbol{x}'(X'X)^{-1}\boldsymbol{x}}$$
- Example 2
  - Data generation
    - x1=rnorm(100); x2=rnorm(100); e=rnorm(100,0,0.1); y=3*x1+2*x2+e
  - Prediction Interval at $x=(1,x_1,x_2)=(1,0,2)$
    - X=cbind(1,x1,x2); V=t(X)%*%X; x.c=c(1,0,2)
    - Fit.AH=lm(y~x1+x2); summary(Fit.AH)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.008814   0.010373   -0.85    0.398
x1           3.010027   0.009787  307.56   <2e-16 ***
x2           2.009821   0.011504  174.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1014 on 97 degrees of freedom
```

    - Fit.AH$coef%*%x.c-qt(0.975,97)*0.1014*sqrt(x.c%*%solve(V)%*%x.c)
      - lower limit =
```
         [,1]
[1,] 3.990707
```
    - Fit.AH$coef%*%x.c+qt(0.975,97)*0.1014*sqrt(x.c%*%solve(V)%*%x.c)
      - upper limit =
```
        [,1]
[1,] 4.03095
```

3.6 Confidence intervals and bands

- Formula:
  - C.B. for fitted value $\hat{y} \pm \sqrt{(p+1)F(\alpha, p+1, n-p-1)}\hat{\sigma}\sqrt{\boldsymbol{x}'(X'X)^{-1}\boldsymbol{x}}$
- Example 3
  - Data generation
    - x1=rnorm(100); x2=rnorm(100); e=rnorm(100,0,0.1); y=3*x1+2*x2+e
  - Prediction Interval at $x=(1,x_1,x_2)=(1,0,2)$
    - X=cbind(1,x1,x2); V=t(X)%*%X; x.c=c(1,0,2)
    - Fit.AH=lm(y~x1+x2); summary(Fit.AH)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.008814   0.010373   -0.85    0.398
x1           3.010027   0.009787  307.56   <2e-16 ***
x2           2.009821   0.011504  174.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1014 on 97 degrees of freedom
```

    - Fit.AH$coef%*%x.c-sqrt(3*qf(0.95,3,97))*0.1014*sqrt(x.c%*%solve(V)%*%x.c)
      - lower limit = `[,1] [1,] 3.978994`
    - Fit.AH$coef%*%x.c+sqrt(3*qf(0.95,3,97))*0.1014*sqrt(x.c%*%solve(V)%*%x.c)
      - upper limit = `[,1] [1,] 4.042662`

Chapter 3 summary

- All you need to know
  - Estimators
    $$\hat{\beta} = (X'X)^{-1}X'Y, \qquad \hat{\sigma}^2 = RSS/(n-p-1)$$
  - Distribution of estimators
    $$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1}) \qquad \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} = \chi^2_{n-p-1}$$
  - Added-variable plot

β_2 measures the relationship b/w y and x2, after adjusting for x1

  - F-test
  - T-test
    $$T = \frac{\hat{\beta}_k}{sd(\hat{\beta}_k)} \sim t(n-p-1)$$
    $$F = \frac{(RSS_{NH} - RSS_{AH})/(df_{NH} - df_{AH})}{RSS_{AH}/df_{AH}} =\sim F(df_{NH} - df_{AH}, df_{AH})$$
  - Prediction Interval.
    $$\hat{y}_* \pm t\left(\frac{\alpha}{2}, n-p-1\right)\hat{\sigma}\sqrt{\boldsymbol{1} + \boldsymbol{x}_*'(X'X)^{-1}\boldsymbol{x}_*}$$
  - C.I. for fitted value
    $$\hat{y} \pm t\left(\frac{\alpha}{2}, n-p-1\right)\hat{\sigma}\sqrt{\boldsymbol{x}'(X'X)^{-1}\boldsymbol{x}}$$
  - C.B. for fitted value
    $$\hat{y} \pm \sqrt{(p+1)F\left(\alpha, p+1, n-p-1\right)}\hat{\sigma}\sqrt{\boldsymbol{x}'(X'X)^{-1}\boldsymbol{x}}$$
  - Confidence Ellipse
    $$\frac{(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)}{(p+1)\hat{\sigma}^2} \le F(\alpha, p+1, n-p-1)$$