Chapter 6 Multiple Regression

Review simple linear regression:

Normal error regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- β_0 and β_1: parameters
- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, 2, \cdots, n$

Example: Growth hormone is used as a prescription drug in medicine to treat children's growth disorders. In the medical study of short children, clinicians want to use the statistical relation to predict growth hormone deficiencies in short children by using simple measurements such as gender, age and various body measurements of the children.

- Gender, age and other body measurements affect the growth hormone in important and distinctive ways
- A single predictor variable in the model would have provided an inadequate description
- In situation of this type, predictions from a simple linear regression model are too imprecise to be useful
- Containing additional predictor variables, typically is more helpful in providing sufficiently precise predictions of the response variable

Multiple Regression

- Multiple–More than one predictor variable
- Y_i is the response variable
- $X_{i1}, X_{i2}, \cdots X_{i,p-1}$ are the $p-1$ explanatory variables for cases $i = 1$ to n
- Potential problem: These predictor variables are likely to be themselves correlated

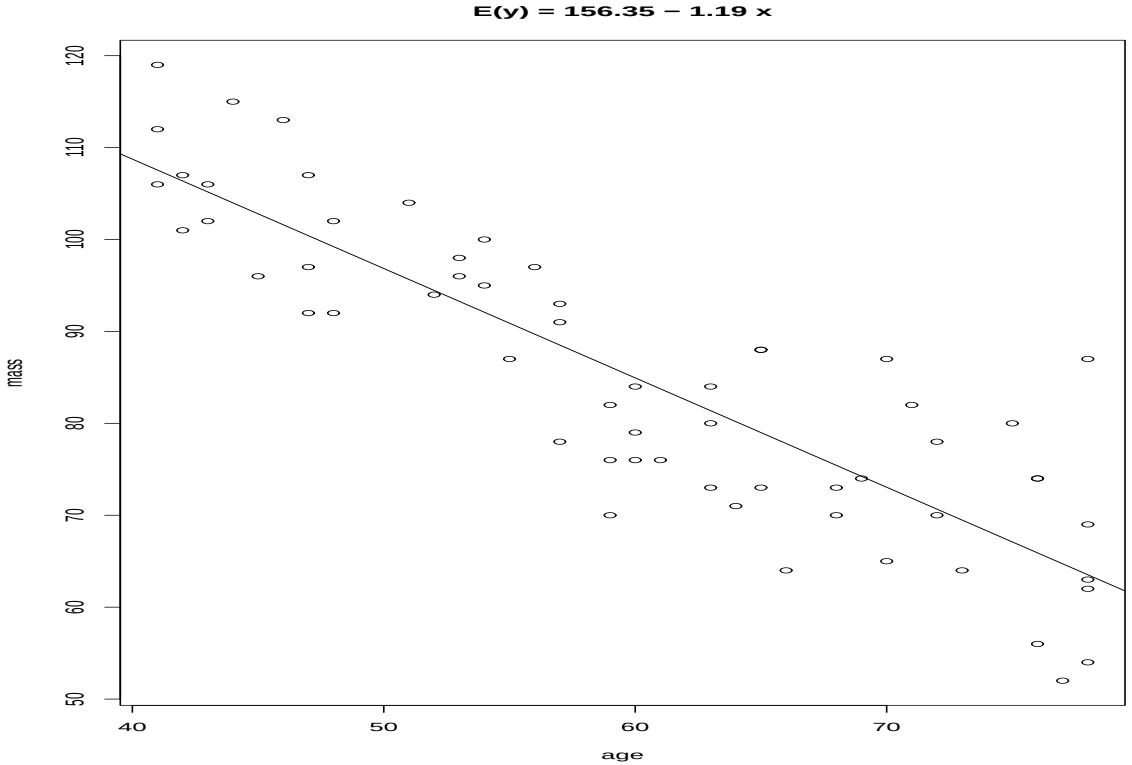
General multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i,$$

- $i = 1, 2, \cdots, n$
- Y_i is the value of the response variable for the ith case
- $X_{i1}, X_{i2}, \cdots, X_{i,p-1}$ are known constants, X_{ik} is the value of the kth explanatory variable for the ith case
- $\beta_0, \beta_1, \cdots, \beta_{p-1}$ are parameters, $p-1$ predictors, p parameters
- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

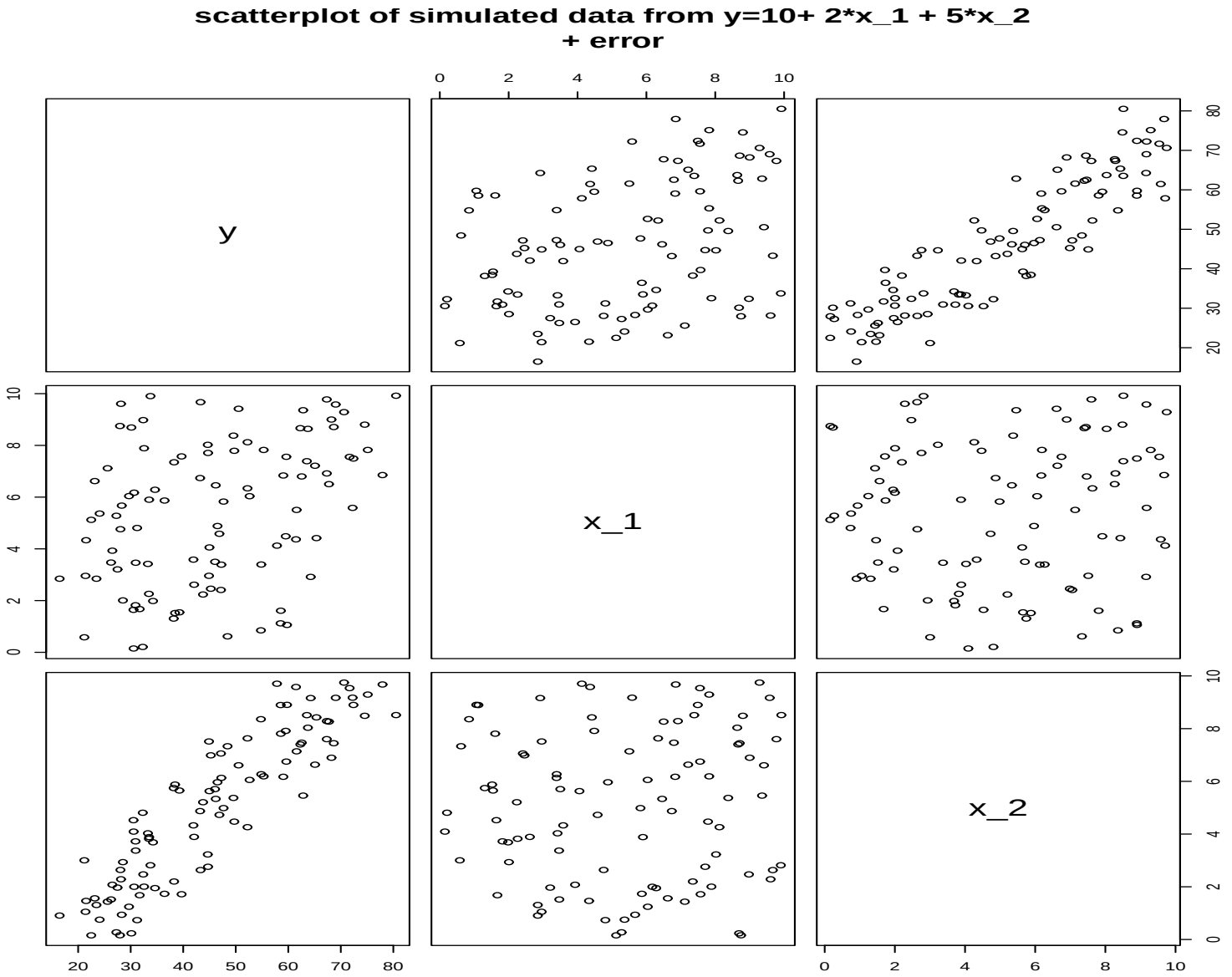
Geometry of one-predictor model (regression function)

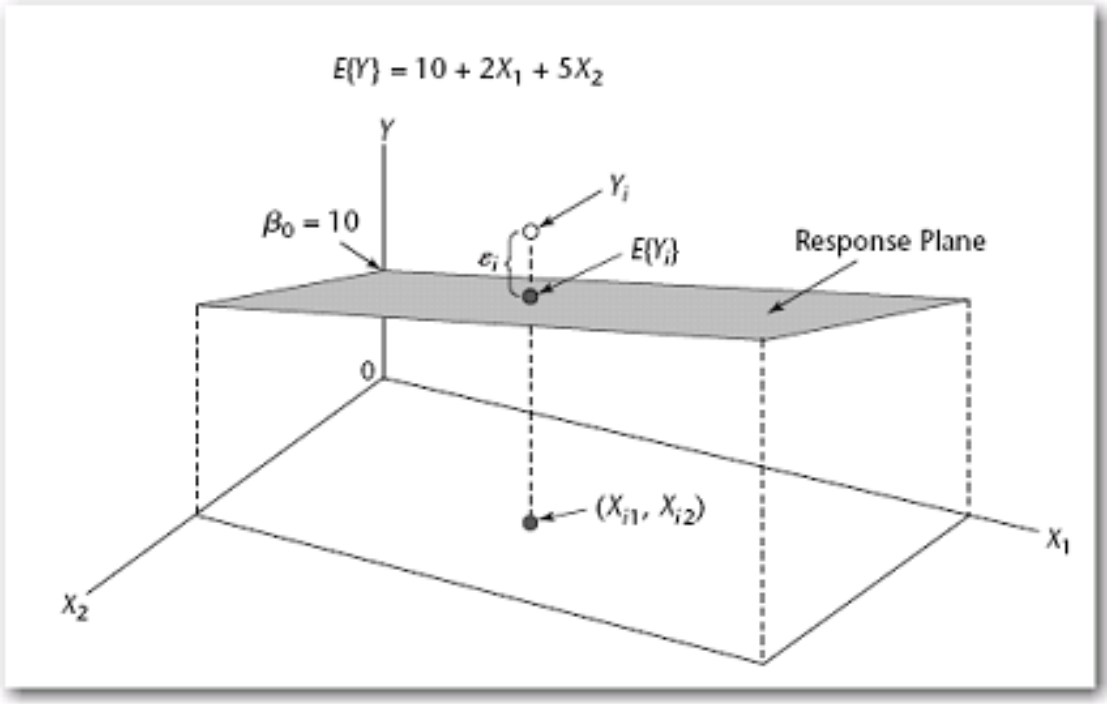
Example: A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each 10-year age group, beginning with age 40 and ending with age 79 with a total number of 60 women.



Geometry of two-predictor model (regression surface or response surface)

Model: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$


scatterplot of simulated data from y=10+ 2*x_1 + 5*x_2
+ error
y
x_1
x_2


$E\{Y\} = 10 + 2X_1 + 5X_2$
Y
Y_i
$\beta_0 = 10$
ε_i
$E\{Y_i\}$
Response Plane
0
(X_{i1}, X_{i2})
X_1
X_2

Meaning of regression coefficients

Model: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

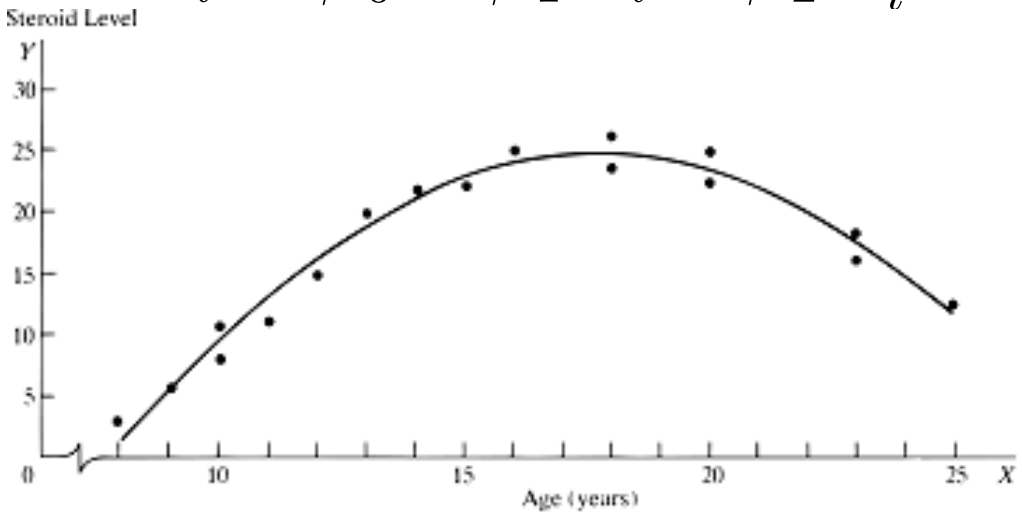
- β_0 is the Y intercept of the regression plane. If the scope of the model includes $X_1 = 0, X_2 = 0$, then β_0 represents the mean response $E(Y)$ at $X_1 = 0, X_2 = 0$. Otherwise, β_0 does not have any particular meaning

- β_k represents the change in the mean response $E(Y)$ for a unit change in X_k while all other X_j's are held constant

Polynomial Regression

Quadratic regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$



- A special case of multiple linear regression if we let $X_2 = X_1^2$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Transform x values

Example: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 e^{X_{i1}} + \beta_3 X_{i2} + \beta_4 \frac{1}{X_{i2}} + \epsilon_i.$

Take

$$X_1^* = X_1$$

$$X_2^* = e^{X_1}$$

$$X_3^* = X_2$$

$$X_4^* = 1/X_2$$

Then the original data after transformation is a multiple linear regression model.

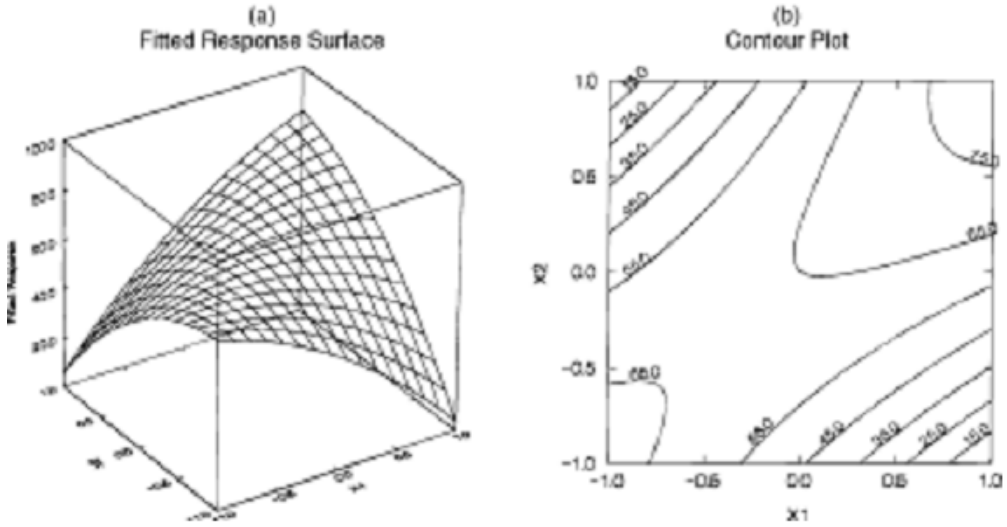
Interaction effects—-cross product

Suppose X_1 and X_2 interact, we can express a form of interaction in the regression model by adding the term X_1X_2

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$$

- Mean of Y at X_1 is $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
- Mean of Y at $X_1 + 1$ is $\beta_0 + \beta_1(X_1 + 1) + \beta_2 X_2 + \beta_3(X_1 + 1)X_2$ = $\beta_0 + \beta_1 X_1 + \beta_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_3 X_2$
- Change in mean of Y is $\beta_1 + \beta_3 X_2$
- X_1 and X_2 interact since the change in the mean of Y for unit change in X_1 depends on X_2

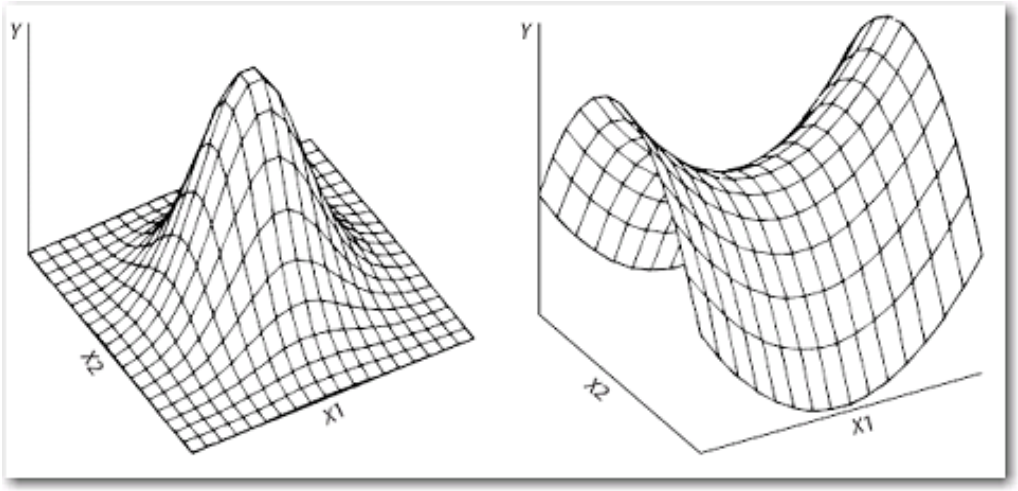
Interaction effects: bends the plane



Interaction plus polynomial terms: full second order model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \epsilon_i$$

Many different shapes are possible, here are two



Matrix approach of multiple linear regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i,$$

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_{p-1} X_{1,p-1} + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_{p-1} X_{2,p-1} + \varepsilon_2$$

$$\vdots$$

$$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_{p-1} X_{n,p-1} + \varepsilon_n$$

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$$

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}.$$

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}),$$

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & & & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Parameter Estimation

Least Squares: Want to minimize the sum of squared residuals:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1})^2$$

Denote the vector of the least squares estimated regression coefficients as

$\mathbf{b} = (b_0, b_1, \cdots, b_{p-1})^T$

- Normal equation

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

solving this equation for $\mathbf{b}$ gives the least squares solution for $\mathbf{b}$

-

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

Note: The method of maximum likelihood leads to the same estimators for normal error regression model

Fitted values and residuals

- Fitted values $\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_{p-1} X_{i,p-1}$
- Residuals $e_i = Y_i - \hat{Y}_i$

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called hat matrix.

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

Estimated covariance matrix

$$\text{var}\{\mathbf{e}\} = \text{cov}(\mathbf{e}) = \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})^T = \sigma^2(\mathbf{I}-\mathbf{H})$$

$$\text{var}\{\mathbf{e}_i\} = \text{var}(\mathbf{e}_i) = \sigma^2(1-h_{ii})$$

where h_{ii} is the ith diagonal element of $\mathbf{H}$

- $\text{cov}(\mathbf{e}_i, \mathbf{e}_j) = -\sigma h_{ij}$

Estimation of σ^2

$$SSE = (\mathbf{Y}-\mathbf{Xb})^T(\mathbf{Y}-\mathbf{Xb}), df_E = n-p$$

$$\hat{\sigma}^2 = SSE/df_E = MSE$$

Variance and estimated variance matrices of $\mathbf{b}$:

$$
\begin{aligned}
\text{var}(\mathbf{b}) &= \begin{bmatrix}
\text{var}(b_0) & \text{cov}(b_0, b_1) & \cdots & \text{cov}(b_0, b_{p-1}) \\
\text{cov}(b_1, b_0) & \text{var}(b_1) & \cdots & \text{cov}(b_1, b_{p-1}) \\
\vdots & \vdots & & \vdots \\
\text{cov}(b_{p-1}, b_0) & \text{cov}(b_{p-1}, b_1) & \cdots & \text{var}(b_{p-1})
\end{bmatrix} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

Estimated variance covariance:

$$s^2(\mathbf{b}) = MSE(\mathbf{X}'\mathbf{X})^{-1}$$

ANOVA Table

- Decomposition of SSTO: SSTO=SSR + SSE

$$SSTO = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_{p-1} x_{i,p-1}$$

Source	SS	df	MS	F-test
Regression	SSR	p-1	MSR = SSR/(p-1)	F =MSR/MSE
Error	SSE	n-p	MSE = SSE/(n-p)	
Total	SSTO	n-1		

F-test for significance of regression:

$$H_0 : \beta_1 = \beta_2 = \cdots \beta_{p-1} = 0$$

$$H_\alpha : \text{not all} \quad \beta_k (k = 1, 2, \cdots p-1) \quad \text{equal zero}$$

Test statistic and decision rule:

$$F^* = MSR/MSE$$

- If H_0 is true, $F = MSR/MSE$ has an F distribution with $(p-1, n-p)$ degrees of freedom.
- Reject H_0, if $F^* > F(1-\alpha, p-1, n-p)$

Coefficient of Multiple Determination R^2

R^2 = proportion of variation in y accounted for by the multiple linear regression model in $x_1, x_2, \cdots, x_{p-1}$

= SSR/SSTO

- $0 \leq R^2 \leq 1$
- R^2 is the square of correlation between y_i and $\hat{y}_i$
- $R^2 = 1$ if $y_i = \hat{y}_i$ for all i
- $R^2 = 0$ if $b_1 = b_2 = \cdots = b_{p-1} = 0$
- A large R^2 value does not necessarily imply that the fitted model is a useful one or that the fit is “good”.
- The addition of more predictors to the regression model will result in an increase in the value of R^2.

- Adjusted R^2

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

The adjusted R^2 can decrease as more predictors are added to the model.

Inference for individual regression coefficient

$$\mathbf{b} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

$$s^2(\mathbf{b}) = MSE(\mathbf{X}^T\mathbf{X})^{-1}$$

$$s^2(b_k) = [s^2(\mathbf{b})]_{k,k}$$

the k-th diagonal element.

- Distribution of b_k: $\dfrac{b_k - \beta_k}{s(b_k)} \sim t(n-p) \quad k = 0, 1, \cdots, p-1$
- a $(1-\alpha)100\%$ confidence interval for β_k

$$b_k \pm t_{n-p}(1-\alpha/2)s\{b_k\}$$

- Significant test for β_k

$$H_0 : \beta_k = 0 \quad \text{v.s} \quad \beta_k \neq 0$$

$$t^* = \frac{b_k}{s\{b_k\}}$$

If H_0 is true, t^* has a t-distribution with $n - p$ degrees of freedom.

Alternative	Reject H_0 if
$H_\alpha : \beta_k > 0$	$t^* > t(1-\alpha; n-p)$
$H_\alpha : \beta_k < 0$	$t^* < -t(1-\alpha; n-p)$
$H_\alpha : \beta_k \neq 0$	$\lvert t^* \rvert > t(1-\alpha/2; n-p)$

Simultaneous Confidence Intervals for $\beta_1, \cdots, \beta_{p-1}$

$$b_k \pm t(1 - \frac{\alpha}{2(p-1)}; n-p)s(b_k),$$

where $j = 1, 2, \cdots, p-1$.

Estimation of $E(Y_h)$

We want a point estimate and a confidence interval for the mean corresponding to the set of explanatory variables $\mathbf{X}_h$.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i,$$

$$\mathbf{X}_h = (1, X_{h,1}, X_{h,2}, \cdots, X_{h,p-1})^T$$

$$Y_h = \beta_0 + \beta_1 X_{h1} + \beta_2 X_{h2} + \cdots + \beta_{p-1} X_{h,p-1} + \epsilon_h$$

$$E(Y_h) = \beta_0 + \beta_1 X_{h1} + \beta_2 X_{h2} + \cdots + \beta_{p-1} X_{h,p-1}$$

or $E(Y_h) = u_h = \mathbf{X}_h^T \boldsymbol{\beta}$

$$\hat{u}_h = \mathbf{X}_h^T \mathbf{b}$$

$$s^2\{\hat{u}_h\} = \mathbf{X}_h^T s^2\{\mathbf{b}\} \mathbf{X}_h = MSE \mathbf{X}_h^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_h$$

95% CI $\hat{u}_h \pm s\{\hat{u}_h\} t_{n-p}(1 - \alpha/2)$

Prediction of $Y_{h(\textbf{new})}$

Predict a new observation Y_h at $\mathbf{X}_h$. We want a prediction of Y_h based on a set of predictor values with an interval that expresses the uncertainty in our prediction. As in SLR this interval is centered at Y_h and is wider than the interval for the mean.

$Y_h = \mathbf{X}_h^T\boldsymbol{\beta} + \epsilon_h$

$\hat{Y}_h = \hat{u}_h = \mathbf{X}_h^T\mathbf{b}$

$$\begin{aligned} s^2\{\text{pred}\} &= \text{var}(Y_{h(\text{new})} - \hat{Y}_h) \\ &= \text{var}(Y_{h(\text{new})}) + \text{var}(\hat{Y}_h) \\ &= MSE(1 + \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h) \end{aligned}$$

CI for $Y_{h(\text{new})}$: $\hat{Y}_h \pm s\{\text{pred}\}t_{n-p}(1-\alpha/2)$

Diagnostics

- Appropriate Regression Model

  –Pairwise Plots Y_i v.s X_{ij} for $j = 1, 2, \cdots, p-1$, i.e, $p-1$ plots of Y v.s $\mathbf{X}$ for each predictor

  –3D plot, plot Y v.s X_j and X_k and look for trends

  –Residual plots, e_i v.s $\hat{Y}_i$, e_i v.s X_{ij}, $j = 1, \cdots, p-1$, e_i v.s pair of $\mathbf{X}$'s

- Constancy of Error Variance

  –Check e_i v.s $\hat{Y}_i$, e_i v.s X_{ij}, $j = 1, 2, \cdots, p-1$

  –Use Brusch-Pagan test, one variable at a time or all variables together

- Normality

  –Histogram Plot

  –Normal probability plot of residuals

  –Tests, Lilliefors' test, correlation test

- Outliers

  –Plot e_i v.s $\hat{Y}_i$, e_i v.s X_{ij}'s, $j = 1, 2, \cdots, p-1$

  –Normal probability Plot

- Independence

  – Check plot of e_i v.s time if possible

- Remedies:

  – Try transformations of Y and/or $\mathbf{X}$'s, polynomials in $\mathbf{X}$'s are often used to deal with curvature

  – May eliminate some of the $\mathbf{X}$'s (this is called variable selection

  Discuss example