Review simple linear regression

Normal error regression model:

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

- y_i: observed response in the ith trial
- x_i: a known constant, the level of the predictor variable in the ith trial
- 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

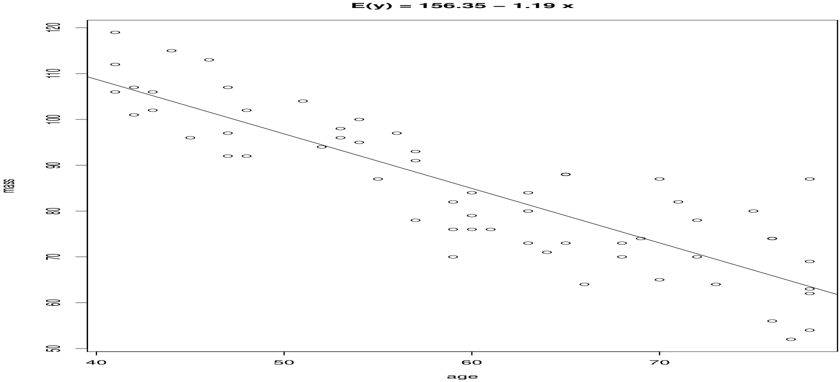
General multiple linear regression model

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

- $i = 1, 2, \cdots, n$
- 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
- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

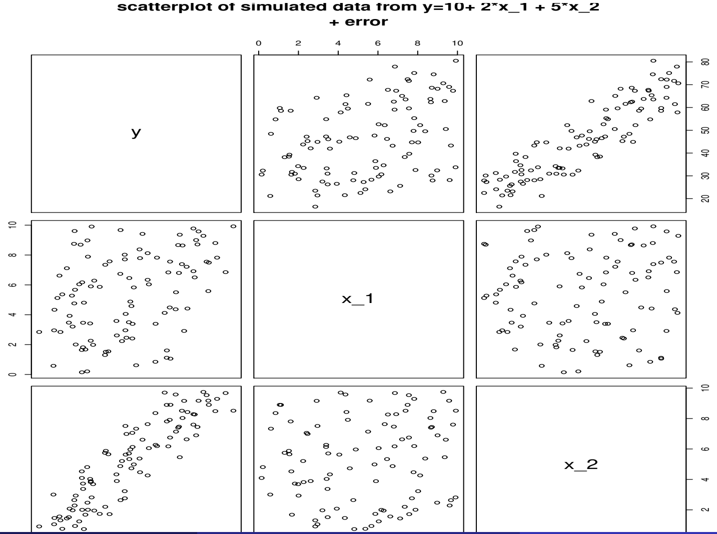
Geometry of one-predictor model

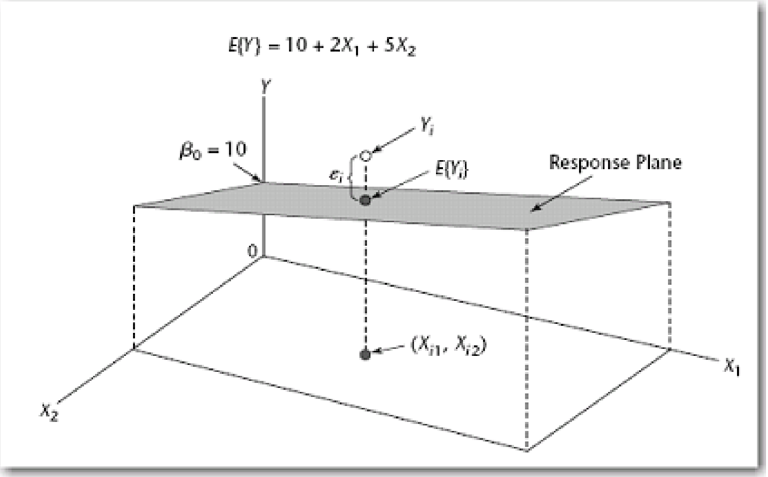
Example: A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each 10-year age group, beginning with age 40 and ending with age 79 with a total number of 60 women.



Geometry of two-predictor model

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




$E\{Y\} = 10 + 2X_1 + 5X_2$
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

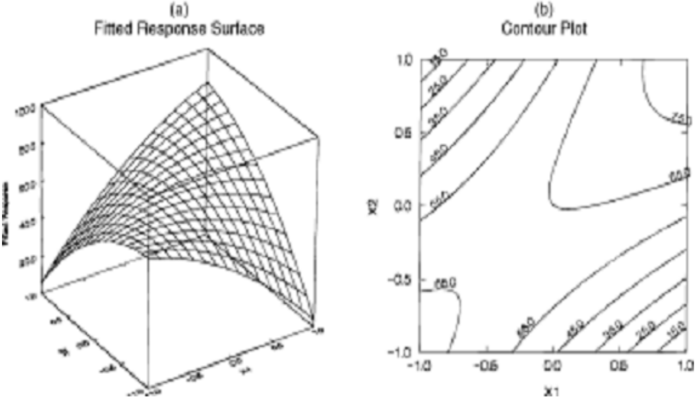
Interaction effects—-cross product

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



Parameter Estimation

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

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

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 $e_i = Y_i - \hat{Y}_i$

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

- 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
$H_\alpha : \beta_k > 0$	$t^* > t(1-\alpha; n-p)$
$H_\alpha : \beta_k < 0$	$t^* < -t(1-\alpha; n-p)$
$H_\alpha : \beta_k \neq 0$	$\|t^*\| > t(1-\alpha/2; n-p)$

Let

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p-1})^T$$

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

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

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 $Y_{h(\mathbf{new})}$

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

Extra sum of squares

Extra sum of squares are used to Measure the effect of adding variables to a regression model, Suppose x_1 is in the regression model and we add the predictor variable x_2, what is the effect of adding x_2 to the model that already contain x_1?

- $SSE(x_1)$ = SSError for the model containing x_1
  $SSR(x_1)$ = SSRegression for the model containing x_1,
  $SSE(x_1) + SSR(x_1) = SSTO$
- Adding x_2 to the model, we get $SSE(x_1, x_2)$ and $SSR(x_1, x_2)$
  ——$SSR(x_1, x_2)$ gives a measure of the effect of both x_1, x_2
  —–The effect of adding x_2 to the model that contains x_1 is measured by

$$SSR(x_2|x_1) = SSR(x_1, x_2) - SSR(x_1)$$

  ——-$SSR(x_2|x_1)$ measures the "marginal" effect of adding x_2 to the model that already contains x_1
- The definition of extra sum of squares can be extended to any number of variables

ANOVA table for decompostion with x_1 first, then x_2 added, then x_3

Source	SS	df	MS
x_1	$SSR(x_1)$	1	$MSR(x_1) = SSR(x_1)/1$
$x_2 \vert x_1$	$SSR(x_2 \vert x_1)$	1	$MSR(x_2 \vert x_1) = SSR(x_2 \vert x_1)/1$
$x_3 \vert x_1, x_2$	$SSR(x_3 \vert x_1, x_2)$	1	$MSR(x_3 \vert x_1, x_2) = SSR(x_3 \vert x_1, x_2)/1$
Regression	SSR	3	MSR $=$ SSR $/3$
Error	SSE	$n-4$	$MSE = SSE/(n-4)$
Total	SSTO	$n-1$	

Three type of sum of squares (SS)

- Type I sum of squares are “sequential. In essence the factors are tested in the order they are listed in the model. $lm(y \sim x_1 + x_2 + x_1 * x_2, data = ex.data)$

$$SSR(x_1), SSR(x_2|x_1), SSR(x_1 * x_2|x_1, x_2)$$

- Type III SS are “partial. In essence, every term in the model is tested in light of every other term in the model. That means the main effects are tested in light of interaction terms as well as in light of other main effects.

$$SSR(x_1|x_2, x_1 * x_2), SSR(x_2|x_1, x_1 * x_2), SSR(x_1 * x_2|x_1, x_2)$$

- Type II SS are similar to Type III, except that they preserve the principle of marginality. This means that main factors are tested in light of one another, but not in light of the interaction term.

Example 1

Goal: Anthropologists conducted a study to determine the long-term effects of an environmental change on systolic blood pressure.
——They measured the blood pressure and several other characteristics of 39 Indians who migrated from a very primitive environment high in the Andes into the mainstream of Peruvian society at a lower altitude.
—— All of the Indians were males at least 21 years of age, and were born at a high altitude.

```
> fn.data <- "http://statacumen.com/teach/ADA2/
ADA2_notes_Ch02_indian.dat"

> indian <- read.table(fn.data, header=TRUE)
> indian
  id age yrmig   wt   ht chin fore calf pulse sysbp diabp
1  1  21     1 71.0 1629  8.0  7.0 12.7    88   170    76
2  2  22     6 56.5 1569  3.3  5.0  8.0    64   120    60
3  3  24     5 56.0 1561  3.3  1.3  4.3    68   125    75
4  4  24     1 61.0 1619  3.7  3.0  4.3    52   148   120
5  5  25     1 65.0 1566  9.0 12.7 20.7    72   140    78
6  6  27    19 62.0 1639  3.0  3.3  5.7    72   106    72
7  7  28     5 53.0 1494  7.3  4.7  8.0    64   120    76
8  8  28    25 53.0 1568  3.7  4.3  0.0    80   108    62
9  9  31     6 65.0 1540 10.3  9.0 10.0    76   124    70
```

Question of interest: is there a relationship between the systolic blood pressure and how long the Indians lived in their new environment as measured by the fraction of their life spent in the new environment?

```
# Create the "fraction of their life" variable
#   yrage = years since migration divided by age
indian$yrage <- indian$yrmig / indian$age
plot(indian$yrage,indian$sysbp,main="scatterplot of
sysbp v.s. yage")
```

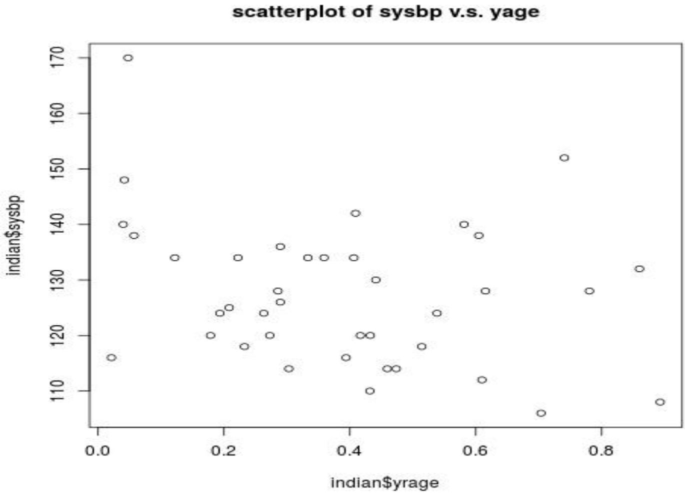


Figure : Scatterplot of sysbp v.s. yage

Fit a simple linear regression model

```
lm.sysbp.yrage <- lm(sysbp ~ yrage, data = indian)
# use Anova() from library(car) to get ANOVA table
library(car)
Anova(lm.sysbp.yrage, type=3)  # (Type 3 SS)
anova(lm.sysbp.yrage)  #Type 1 SS
# use summary() to get t-tests of parameters
summary(lm.sysbp.yrage)
```

```
> summary(lm.sysbp.yrage)

Call:
lm(formula = sysbp ~ yrage, data = indian)

Residuals:
    Min      1Q  Median      3Q     Max
-17.161 -10.987  -1.014   6.851  37.254
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  133.496      4.038  33.060   <2e-16 ***
yrage        -15.752      9.013  -1.748   0.0888 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 12.77 on 37 degrees of freedom
Multiple R-squared:  0.07626, Adjusted R-squared:  0.05129
F-statistic: 3.054 on 1 and 37 DF,  p-value: 0.08881
```

$$\hat{y} = 133.496 - 15.752x$$

- suggests that average systolic blood pressure decreases as the fraction of life spent in modern society increases.
- However, the t-test of $H_0 : \beta_1 = 0$ is not significant at the 5% level (p-value=0.08881).

  ——there is no linear relationship between systolic blood pressure and the fraction of life spent in a modern society
- $R^2 = 0.07626$ suggests that yrage fraction does not explain a substantial amount of the variation in the systolic blood pressures.

```
> Anova(lm.sysbp.yrage, type=3)
Anova Table (Type III tests)

Response: sysbp
            Sum Sq Df   F value  Pr(>F)
(Intercept) 178221  1 1092.9484 < 2e-16 ***
yrage          498  1    3.0544 0.08881 .
Residuals     6033 37
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> anova(lm.sysbp.yrage)  #type I
Analysis of Variance Table

Response: sysbp
          Df Sum Sq Mean Sq F value  Pr(>F)
yrage      1  498.1  498.06  3.0544 0.08881 .
Residuals 37 6033.4  163.06
---
```

Taking Weight Into Consideration

```
# fit the multiple linear regression model, (" + wt" added)
lm.sysbp.yrage.wt <- lm(sysbp ~ yrage + wt, data = indian)
library(car)
Anova(lm.sysbp.yrage.wt, type=3)
anova(lm.sysbp.yrage.wt)  #sequential ss
summary(lm.sysbp.yrage.wt)
```

```
> summary(lm.sysbp.yrage.wt)

Call:
lm(formula = sysbp ~ yrage + wt, data = indian)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.8959    14.2809   4.264 0.000138 ***
yrage       -26.7672     7.2178  -3.708 0.000699 ***
wt            1.2169     0.2337   5.207 7.97e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.777 on 36 degrees of freedom
Multiple R-squared:  0.4731, Adjusted R-squared:  0.4438
F-statistic: 16.16 on 2 and 36 DF,  p-value: 9.795e-06
```

```
> Anova(lm.sysbp.yrage.wt, type=3)
Anova Table (Type III tests)
Response: sysbp
            Sum Sq Df F value    Pr(>F)
(Intercept) 1738.2  1  18.183 0.0001385 ***
yrage       1314.7  1  13.753 0.0006991 ***
wt          2592.0  1  27.115 7.966e-06 ***
Residuals   3441.4 36

> anova(lm.sysbp.yrage.wt)  #sequential ss
Analysis of Variance Table

Response: sysbp
          Df Sum Sq Mean Sq F value    Pr(>F)
yrage      1  498.1  498.06  5.2102   0.02846 *
wt         1 2592.0 2592.01 27.1149 7.966e-06 ***
Residuals 36 3441.4   95.59
---
```

Findings:

- Fitted regression lines
  ——SLR: $\hat{y} = 133.496 - 15.752 yrage$
  ——MLR: $\hat{y} = 60.89 - 26.76 yrage + 1.21 wt$

Table : comparison of some numbers from SLR and MLR

	df_R	df_E	SSE	SSR	R^2	$\hat{\sigma}^2$
SLR	1	39-2=37	6033.37	498.1	0.076	163.06
MLR	2	39-3=36	3441.36	3090	0.473	95.59
Diff			2592	2592		

- SSTO does not depend on the number of predictors so it stays the same.
- SSE, or the part of the variation in the response unexplained by the regression model, never increases when new predictors are added. (You cant add a predictor and explain less variation.)

- Adding the weight variable to the model increases R^2 by 40%. That is, weight explains 40% of the variation in systolic blood pressure not already explained by fraction.

F overall test

$$H_0 : \beta_1 = \beta_2 = 0 \text{ v.s. at least one of them is not equal to } 0$$

Test of no relationship between the average systolic blood pressure and fraction and weight, assuming the relationship is linear.

- $$SSR = 498 + 2592 = 3090, MSR = 3090/2 = 1545$$
  $$MSE = 95.59$$
  $$F_{obs} = MSR/MSE = 1545/95.59 = 16.163$$
  compare with $F(2, 36, 0.95) = 3.259446$, reject H_0, conclude that either fraction or weight, or both, are important for explaining the variation in systolic blood pressure.
- Residual standard error: 9.777 on 36 degrees of freedom Multiple R-squared: 0.4731, Adjusted R-squared: 0.4438 F-statistic: 16.16 on 2 and 36 DF, p-value: 9.795e-06

Individual t test

$$H_0 : \beta_2 = 0 \text{ v.s.} \beta_2 \neq 0$$

whether adding weight to the model explains a significant part of the variation in systolic blood pressure not explained by yrage fraction.

$$t_{obs} = \frac{b_2 - 0}{SE(b_2)} = \frac{1.217}{0.234} = 5.207$$

compare to t critical value with 36 degrees of freedom 2.028094, we reject H_0, conclude that weight is significant in explaining the variation in systolic blood pressure when yrage fraction is already in the model.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.8959    14.2809   4.264 0.000138 ***
yrage       -26.7672     7.2178  -3.708 0.000699 ***
wt            1.2169     0.2337   5.207 7.97e-06 ***
```

Confidence Intervals

$$b_i \pm t_{crit} * SE(b_i)$$

```
> confint(lm.sysbp.yrage.wt, level=.95)
                  2.5 %     97.5 %
(Intercept)  31.9329542  89.858895
yrage       -41.4055934 -12.128836
wt            0.7429167   1.690796
```

Confidence Intervals of the mean

```
> newdata <- data.frame(yrage=c(0.05,0.03), wt=c(58,64))
> predict(lm.sysbp.yrage.wt, newdata=newdata,
interval="confidence", level=.95)
       fit      lwr      upr
1 130.1352 124.3695 135.9010
2 137.9717 131.7534 144.1901
> newdata <- data.frame(yrage=c(0.05,0.03), wt=c(58,64))
## 95% CI's for E(Y) using Bonferroni Correction
predict(lm.sysbp.yrage.wt, newdata=newdata,
interval="confidence", level=1-.05/2)
       fit      lwr      upr
1 130.1352 123.4855 136.7850
2 137.9717 130.7999 145.1435
```

Prediction Intervals of the mean

```
> newdata <- data.frame(yrage=c(0.05,0.03), wt=c(58,64))
> predict(lm.sysbp.yrage.wt, newdata=newdata,
interval="prediction", level=.95)
       fit      lwr      upr
1 130.1352 109.4849 150.7855
2 137.9717 117.1905 158.7529
> predict(lm.sysbp.yrage.wt, newdata=newdata,
interval="prediction", level=1-.05/2)
       fit      lwr      upr
1 130.1352 106.3186 153.9519
2 137.9717 114.0041 161.9393
```

Comments

- The t-test for $H_0 : \beta_1 = 0$ is highly significant (p-value=0.0007)
  —— this implies that fraction is important in explaining the variation in systolic blood pressure by including weight in the model as a predictor.
  ——Weight is called a suppressor variable. Ignoring weight suppresses the relationship between systolic blood pressure and yrage fraction.

  $$\hat{y} = 60.89 - 26.76 yrage + 1.21 wt$$

  Fix $wt = 50kg$, $\hat{y} = 121.39 - 26.76 yrage$
  Fix $wt = 60kg$, $\hat{y} = 133.49 - 26.76 yrage$
  ——-the average systolic blood pressure decreases by 26.76 for each increase of 1 on fraction, regardless of ones weight.
  ——-Recall SLR: $\hat{y} = 133.496 - 15.752 yrage$
  the average systolic blood pressure decreases by 15.75 for each increase of 1 on fraction.

- On the other hand, a predictor that is highly correlated with the response may be unimportant in a multiple regression model once other predictors are included in the model.

  ——— In multiple regression “everything depends on everything else.
- Interpretation: $b_1 = -26.76$ indicates that the predicted systolic blood pressure decreases as yrage fraction increases holding weight constant.

  ——- the predicted systolic blood pressure decreases by 26.76 for each unit increase in fraction, holding weight constant at any value.

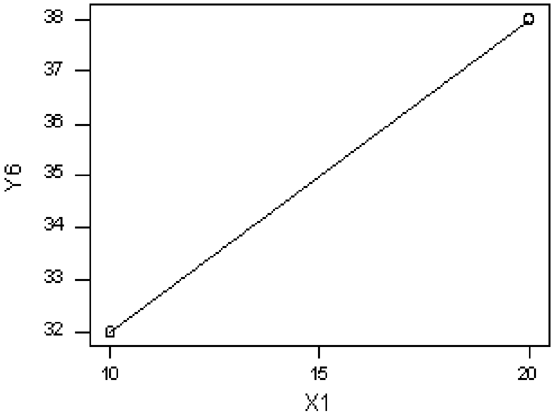
  ——Similarly, the predicted systolic blood pressure increases by 1.21 for each unit increase in weight, holding yrage fraction constant at any level.

- $R^2 = \dfrac{SSR}{SSTO}$ is the ratio of explained and total variation, or proportion of variation in Y that is accounted for by the regression relationship with Xs

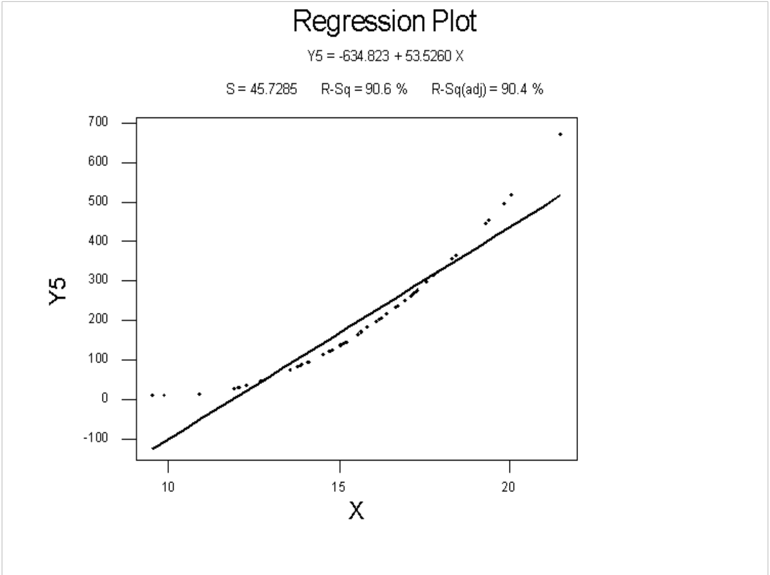
  Limitations of R^2

  Misunderstanding 1: A high coefficient of determination indicates that useful predictions can be made.

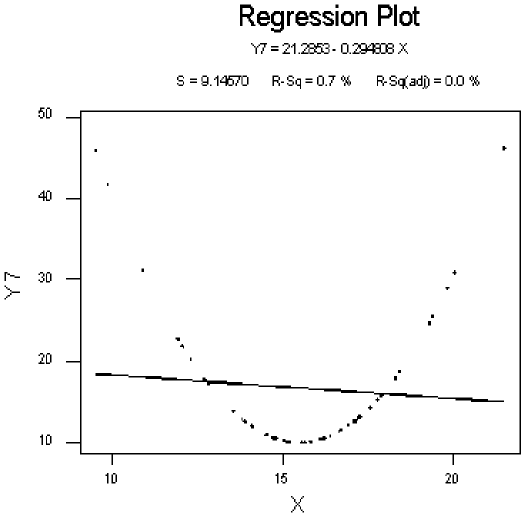
  ———R^2 measures only a relative reduction from SSTO and provides no information about absolute precision for estimating a mean response or predicting a new observation


38
37
36
35
34
33
32
Y6
10
15
20
X1

Misunderstanding 2: A high coefficient of determination indicates that the estimated regression line is a good fit.



Misunderstanding 3: A low coefficient of determination indicates that x and y are not related.



- Adjusted R^2 takes into account the number of predictor variables and the sample size, i.e., it is adjusted based on the df. Adjusted R^2 becomes more relevant as a diagnostic tool when used in multiple regression.

  $$\text{Adjusted } R^2 = 1 - (1 - R^2)\frac{n-1}{n-p}$$

  —n: number of observations in the sample
  —p: number of parameters, or number of predictor variables $+1$
  ——- From this formula, you can see that when the number of observations is small and the number of predictors is large, there will be a much greater difference between R-square and adjusted R-square. By contrast, when the number of observations is very large compared to the number of predictors, the value of R-square and adjusted R-square will be much closer because the ratio of $(n-1)/(n-p)$ will approach 1.

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 studentized deleted residual (rstudent) v.s $\hat{Y}_i$, v.s X_{ij}'s, $j = 1, 2, \cdots, p-1$

- 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

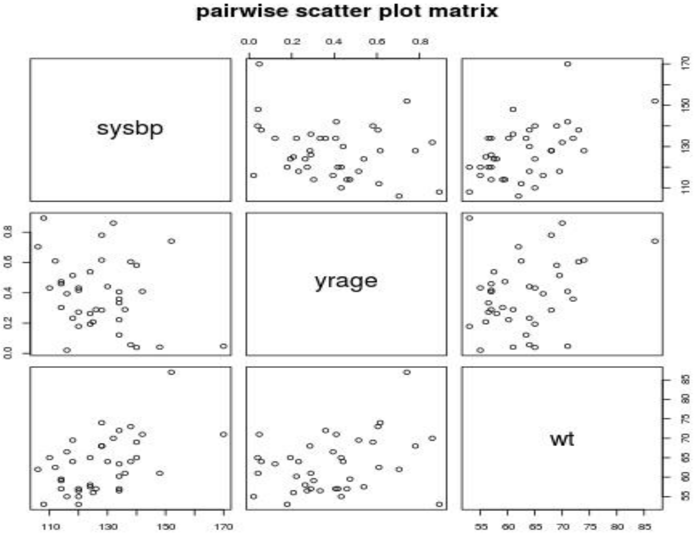
  – May eliminate some of the $\mathbf{X}$'s (this is called variable selection)

Indian example continued

Scatterplot matrix and correlation matrix

```
pairs(sysbp ~ yrage+wt,data=indian,
main="pairwise scatter plot matrix")
cor(cbind(indian$sysbp,indian$yrage,indian$wt))
```

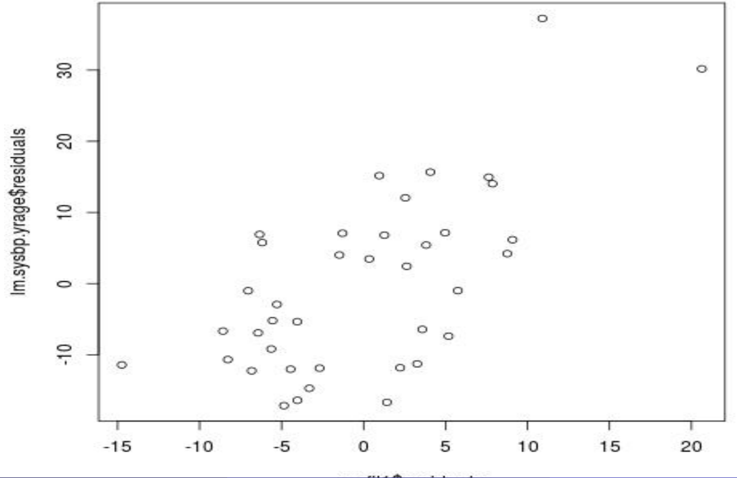
```
           [,1]       [,2]      [,3]
[1,]  1.0000000 -0.2761457 0.5213643
[2,] -0.2761457  1.0000000 0.2930830
[3,]  0.5213643  0.2930830 1.0000000
```


pairwise scatter plot matrix
sysbp
yrage
wt
0.0
0.2
0.4
0.6
0.8
110
130
150
170
55
60
65
70
75
80
85

Added variable plot:

- Fit regression $y \sim x_1$, obtain residual $e_i(y|x_1)$
- Fit regression $x_2 \sim x_1$, obtain residual $e_i(x_2|x_1)$
- plot the residuals $e_i(y|x_1)$ against residuals $e_i(x_2|x_1)$.
  ——If there seems to be no relationship between them, x_2 will not be important;
  —— if the plot looks clearly linear, x_2 will be important.
  ——if the plot looks curved, this may be an indication for transformation on x_2

```
myfit1 <- lm(wt ~ yrage, data = indian)
plot(lm.sysbp.yrage$residuals,myfit1$residuals)
#or library(car) avPlots(lm.sysbp.yrage.wt)
```



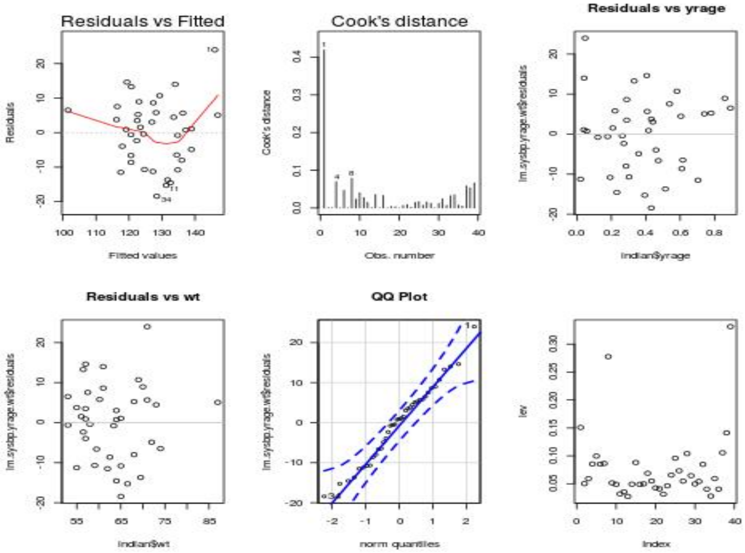
We have seen a linear relationship from the plot of $e_i(sysbp|yrage)$ against residuals $e_i(wt|yrage)$. , therefore, "wt" is important to add to the model.

In general, the added variable plot or the partial regression residual plot compares the residuals from two model fits.

- First, we "adjust Y for all the other predictors in the model except the selected one.
- Then, we "adjust the selected variable X_{sel} for all the other predictors in the model.
- Lastly, plot the residuals from these two models against each other to see what relationship still exists between Y and X_{sel} after accounting for their relationships with the other predictors.

Diagnostic plots

```
par(mfrow=c(2,3))
plot(lm.sysbp.yrage.wt,which=c(1,4)) #residual v.s fitted valu
cooked distance
plot(indian$yrage, lm.sysbp.yrage.wt$residuals,
main="Residuals vs yrage")
# horizontal line at zero
abline(h = 0, col = "gray75")
plot(indian$wt, lm.sysbp.yrage.wt$residuals,
main="Residuals vs wt")
abline(h = 0, col = "gray75")
library(car)
qqPlot(lm.sysbp.yrage.wt$residuals, las = 1, main="QQ Plot")
lev<-hatvalues(lm.sysbp.yrage.wt)
plot(lev)
```

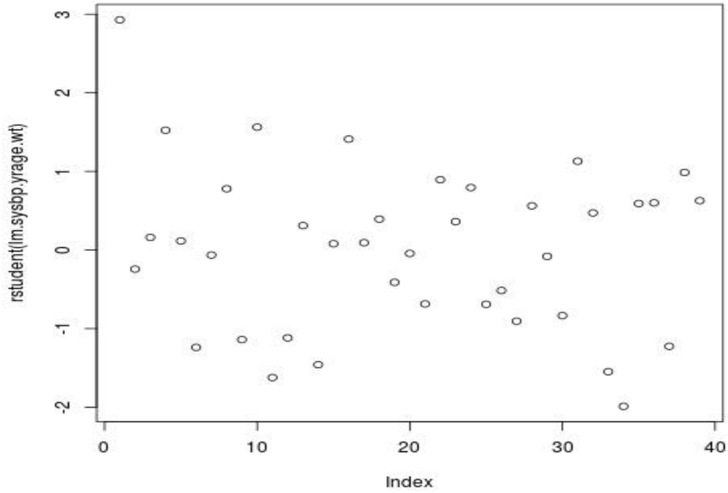

Residuals vs Fitted
Residuals
Fitted values
1
11
34
Cook's distance
Cook's distance
Obs. number
1
4
8
Residuals vs yrage
lm.sysbp.yrage.wt$residuals
indian$yrage
Residuals vs wt
lm.sysbp.yrage.wt$residuals
indian$wt
QQ Plot
lm.sysbp.yrage.wt$residuals
norm quantiles
1
34
lev
Index

```
> library(lmtest)
> bptest(sysbp˜yrage+wt,data=indian,studentize=FALSE)
#test constant variance

Breusch-Pagan test

data:  sysbp ˜ yrage + wt
BP = 2.3797, df = 2, p-value = 0.3043

> shapiro.test(rstandard(lm.sysbp.yrage.wt))

Shapiro-Wilk normality test

data:  rstandard(lm.sysbp.yrage.wt)
W = 0.98081, p-value = 0.733
```

- Residual v.s fitted value plot didn't show obvious pattern
  Residual v.s. yrage and residual v.s. wt didn't show pattern either
  BP testis with a p-value of 0.3043
  Constant variance assumption and linearity assumption seem to be met.
- QQ plot shows that normality assumption is roughly met.
  Shapiro-Wilk test (with a p-value of 0.733) also support the normality assumptions.
- There is no time order provided for the data, we can't check independence assumption. Let's assume independence assumption is met.
- Cooks distance is substantially larger for observation 10, leverages are larger for observations 8 and 39.

Outliers

```
rstudent(lm.sysbp.yrage.wt)  ##gives rstudent values
        1           2           3           4           5
 2.92834234 -0.24318932  0.15980796  1.52347845  0.11470330 -1
outlierTest(lm.sysbp.yrage.wt)  ##Reports the Bonferroni
> outlierTest(lm.sysbp.yrage.wt)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
  rstudent unadjusted p-value Bonferonni p
1 2.928342          0.0059575      0.23234
par(mfrow=c(1,1))
plot(rstudent(lm.sysbp.yrage.wt))
indian[1,]
> indian[1,]
  id age yrmig wt   ht chin fore calf pulse sysbp      yrage
1  1  21     1 71 1629    8    7 12.7    88   170 0.04761905
```


rstudent(lm.sysbp.yrage.wt)
-2
-1
0
1
2
3
0
10
20
30
40
Index

Leverage, x outliers

```
xoutliers <- which(lev > 2*3/39)
xoutliers
lev[xoutliers]
xoutliers2 <- which(lev > 3*3/39)
xoutliers2
 8 39
 8 39
> indian[8,]
  id age yrmig wt    ht chin fore calf pulse sysbp   yrage
8  8  28    25 53 1568  3.7  4.3    0    80   108  0.8928571
> indian[39,]
   id age yrmig wt    ht chin fore calf pulse sysbp  yrage
39 39  54    40 87 1542 11.3 11.7 11.3    92   152  0.7407407
plot(lev)
```

```
cooks.distance(lm.sysbp.yrage.wt)
max(cooks.distance(lm.sysbp.yrage.wt))
order(cooks.distance(lm.sysbp.yrage.wt))[39]
plot(cooks.distance(lm.sysbp.yrage.wt))
> highcook <- which((cooks.distance(lm.sysbp.yrage.wt))
 > qf(0.5,3,36))
> cooks.distance(lm.sysbp.yrage.wt)[highcook]
named numeric(0)
```

Multicolinearity and It's Effect

When the predictor variables are correlated among themselves, we say that there is multicollinearity.

1. The estimate of any parameter, say β_2, depends on all the variables that are included in the model.
2. The sum of squares for any variable, say x_2, depends on all the other variables that are included in the model. For example, none of $SSR(x_2), SSR(x_2|x_1)$, and $SSR(x_2|x_3, x_4)$ would typically be equal.
3. A moderate amount of collinearity has little effect on predictions and therefore little effect on SSE, R^2, and the explanatory power of the model. Collinearity increases the vairance of the $\hat{\beta}_k$s, making the estimates of the parameters less reliable. Sometimes a large amount of collinearity can have an effect on predictions.

4. Suppose the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

is fitted and we obtain t statistics for each parameter.
If the t statistic for testing $H_0 : \beta_1 = 0$ is small, we are led to the model

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i.$$

If the t statistic for testing $H_0 : \beta_2 = 0$ is small, we are led to the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \epsilon_i.$$

However, if the t statistics for both tests are small, we are not led to the model

$$y_i = \beta_0 + \beta_3 x_{i3} + \epsilon_i.$$

Multicollinearity can greatly affect

- the regression coefficient
- the variance of the regression coefficients
- our understanding of the predictor variables and their effect on the response

Two special case (see R handout)

- Uncorrelated Predictor variables: In that case Type I and Type 3 SS will be the same. The contribution of each explanatory variable to the model is the same whether or not the other explanatory variables are in the model.
- Predictor variables are perfectly correlated: The Type 3 SS for the predictor variables involved will be zero because when one is included the other is redundant. It explains NO additional variation over the other variables.

Check for multicollinearity

```
> vif(lm.sysbp.yrage.wt)
   yrage       wt
1.093969 1.093969
```

Variance inflation factor (VIF) is the ratio of variance in a model with multiple terms, divided by the variance of a model with one term alone.
Common rules:
VIF less than 5 are considered as small, no concern of multicollinearity.
VIF between 5 to 10 are considered moderate of multicollinearity.
VIF greater than 10 are considered as high multicollinearity.

Comments:

- As long as points 1 through 4 are kept in mind, a moderate amount of collinearity is not a big problem.
- The overall F statistic is significant, but none of the individual t test is significant. This indicates multicollinearity problem.
- High multicollinearity among the predictor variables does not prevent the mean square error, measuring the variability of the error terms, from being steadily reduced as additional variables are added to the regression model.
- The precision of fitted values within the range of the observations on the predictor variables is not eroded with the addition of correlated predictor variables into the regression model.

Remedies for Multicollinearity

- Variable selection
- Use biased regression methods such as ridge regression or principle components regression, especially there is an interest in the regression coefficients themselves.

Interaction Regression Models

Additive Model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
Interaction Model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$

- When an interaction is present, the effect of the change in the mean response when the value of a predictor variable changes depends on the value of another predictor variable.
- β_1 and β_2 no longer indicate the change in the mean response with a unit increase of the predictor variable, with the other predictor variable held constant at any given level. For example, change in the mean response with a unit increase in x_1 when x_2 is held constant is $\beta_1 + \beta_3 x_2$.

Example:

——No interaction

$$E\{y\} = 10 + 2x_1 + 5x_2$$

- when $x_2 = 1$, the response function $E\{y\}$ as a function of x_1 is

$$E\{y\} = 15 + 2x_1 \tag{1}$$

- when $x_2 = 3$,

$$E\{y\} = 25 + 2x_1 \tag{2}$$

- Line (1) and (2) are parallel.

——Reinforcement Interaction Effect

$$E\{y\} = 10 + 2x_1 + 5x_2 + .5x_1x_2$$

- when $x_2 = 1$, the response function $E\{y\}$ as a function of x_1 is

$$E\{y\} = 15 + 2.5x_1 \tag{3}$$

- when $x_2 = 3$,

$$E\{y\} = 25 + 3.5x_1 \tag{4}$$

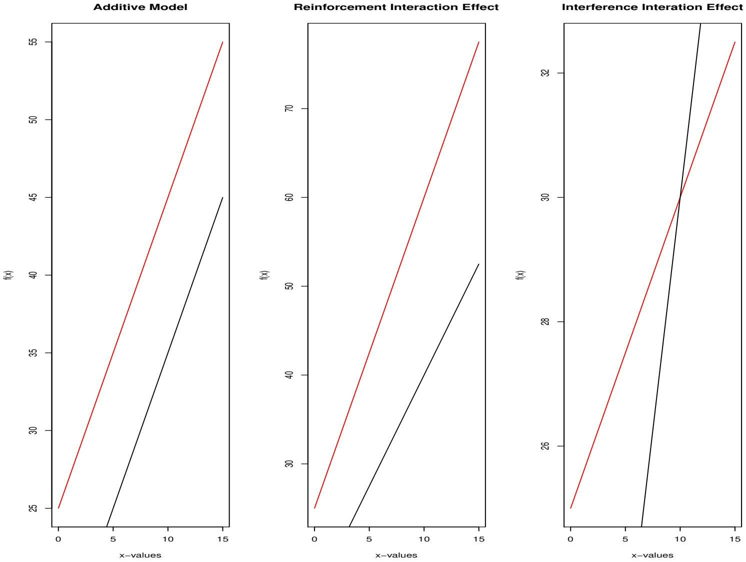
- Line (3) and (4) are not parallel.

——Interference Interaction Effect

$$E\{y\} = 10 + 2x_1 + 5x_2 - .5x_1x_2$$

- when $x_2 = 1$, the response function $E\{y\}$ as a function of x_1 is
$$E\{y\} = 15 + 1.5x_1 \tag{5}$$
- when $x_2 = 3$,
$$E\{y\} = 25 + .5x_1 \tag{6}$$
- Line (5) and (6) are not parallel.

Figure : Interactions



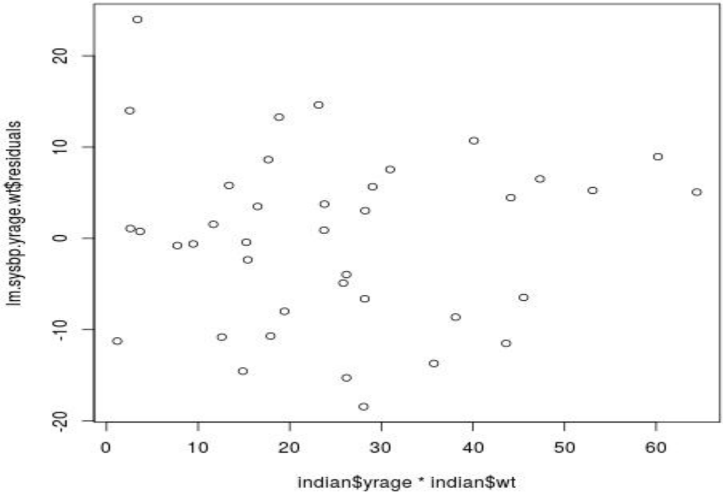
Check if interaction term need to be included

Plot residual from the additive model vs. interaction terms, if no pattern observed, do not include the interaction term.
if a pattern (not random) is observed, may need to include the interaction term.

```
plot(indian$yrage*indian$wt,lm.sysbp.yrage.wt$residuals)
```

No pattern observed, do not suggest adding the interaction term yrage*wt

Figure : Interaction of yrage*wt



Comments:

- High multicollinearity may exist between some of the predictor variables and some of the interaction terms, as well as among some of the interaction terms. A partial remedy to improve computational accuracy is to center the predictor variables
- When the number of predictor variables is large, potential number of interaction terms become very large
- It is desirable to identify in advance, whenever possible, those interactions that are most likely to influence the response variable in important ways. In addition to utilizing a priori knowledge, one can plot the residuals for the additive regression model against the different interaction terms to determine which ones appear to be influential in affecting the response variable
- When the number of predictor variables is large, these plots may need to be limited to interaction terms involving those predictor variables that appear to be the most important on the basis of the initial fit of the additive regression model