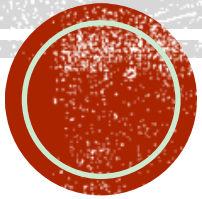
Lecture 10: multiple linear regression



BIG PICTURE

- We seldom run a regression model with just one predictor but for pedagogical reasons is always easier to understand the basics of regression with only one variable
- When modeling, it's useful to start with just one predictor/covariate of interest and then add variables
- We often want to describe relationships "controlling" for other factors. Today, we will see different ways of understanding what it means to "control" or "adjust" for other variables and to "hold them constant" or "taking them into account"

1. Limitation of simple linear regression

Data mining: Torture the data to uncover the useful hidden information.

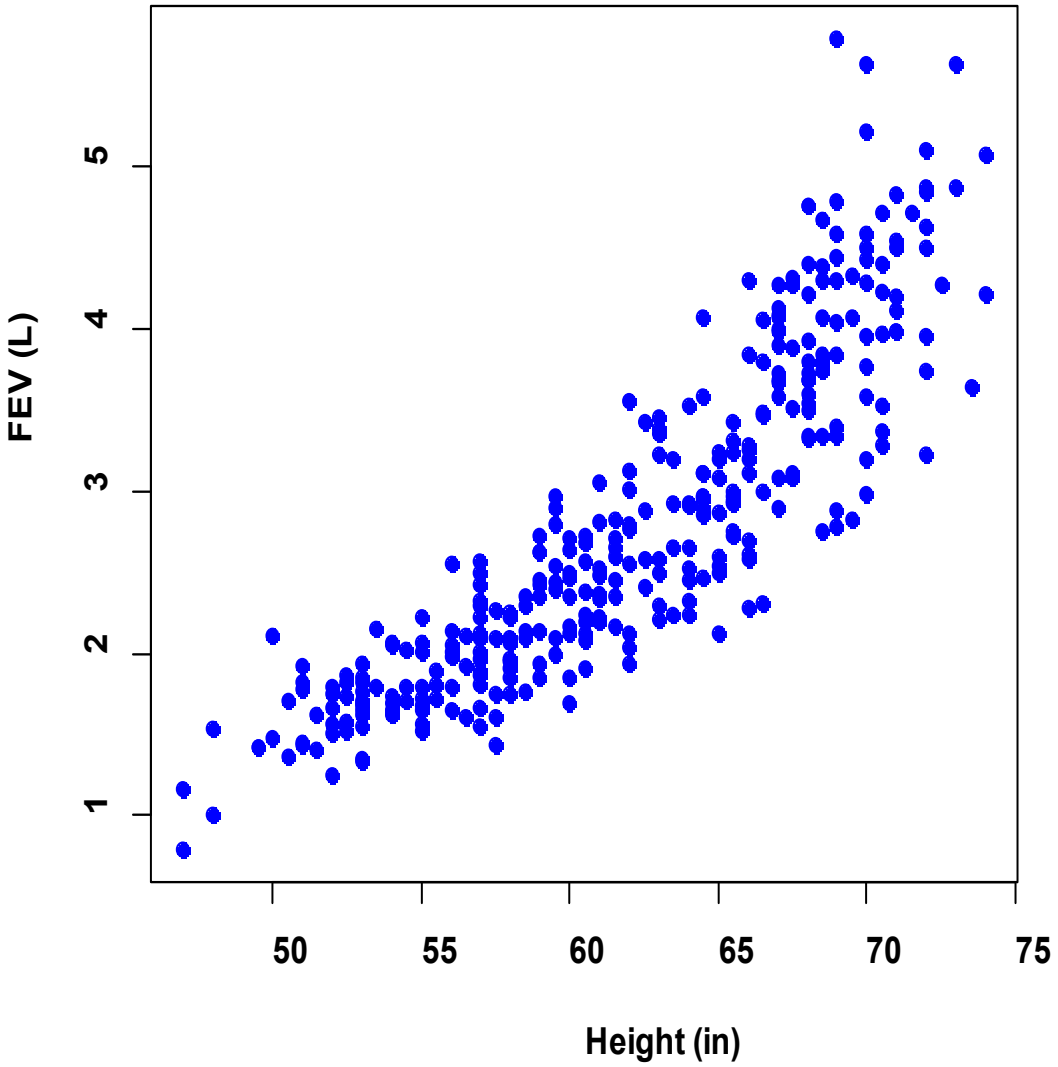
Example (Pulmonary Disease): ***Forced expiratory volume*** **(FEV)** is an index of pulmonary function that measures the volume of air expelled after 1 second of constant effort. The data set (FEV_data.txt) contains the determinations of FEV in 1980 on 654 children ages 3-19, the participants of the Childhood Respiratory Disease (CRD) Study in East Boston, Massachusetts.

Variable	Column	Description	Format or Code
Age	7-8	Age (yrs)	
FEV	10-15	FEV (liters)	X.XXXX
Hgt	17-20	Height (inches)	XX.X
Sex	22	Sex	0=female
			1=male
Smoke	24	Smoking Status	0=non-current
			1=current smoker

Participants includes 336 boys and 318 girls. Here, for illustration purpose, we analyze data of boys only.

Fig. 1: Scatter plot of FEV vs. Height.

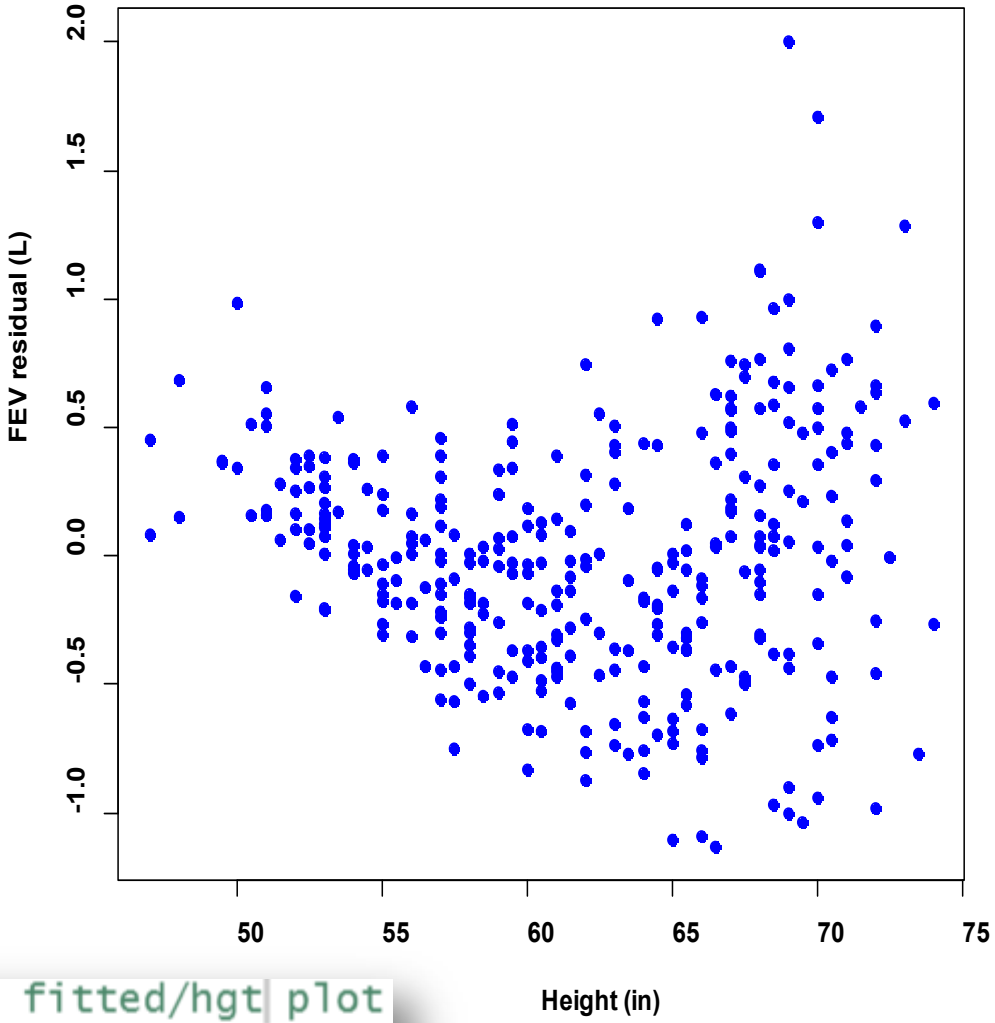
1) Rough linear trend
2) Positive correlation
3) FEV scatters more as height increases
4) Linear trend is violated a little bit more as height increases



RESIDUAL ANALYSIS

Fig. 2: FEV residual vs. Height

1) FEV Residual variation increases as height increases
2) Outliers appear around 70 (in) in height
3) Residual tends to be negative when height is between 55-72; and tend to be positive elsewhere




```
# produce residual vs. fitted/hgt plot
res <- resid(fit)
plot(fitted(fit), res)
plot(x2$hgt,res,col=blues9)
```

Table 1.

Example of *basic statistics*.

	Height x (in)	x^2	FEV y (L)	y^2	Product xy
Sum	20840.5	1306064	944.982	2995.126	60490.87
Mean	62.0253	3887.096	2.812446	8.914066	180.0323

$$\hat{a} = \bar{y} - \bar{x}\hat{\beta} \text{ and } \hat{\beta} = \frac{L_{x,y}}{L_{x,x}}$$

```
lm(formula = fev ~ hgt, data = x2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.13438 -0.30820 -0.00568  0.30821  2.00491

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.863848   0.254470  -23.04   <2e-16 ***
hgt          0.139883   0.004082   34.27   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4729 on 334 degrees of freedom
Multiple R-squared:  0.7786,    Adjusted R-squared:  0.7779
F-statistic:  1175 on 1 and 334 DF,  p-value: < 2.2e-16
```

2. FRAMEWORK

In multiple regression, we will study the relationship between a *single outcome* variable y (dependent) and multiple predictor variables (independent) $x_1,\ldots,x_k\,(k \geq 2)$, see **Figure 3**, where the x's can be accurately measured or observed, the y may depend on both unobserved error (e) and the x's. **We will estimate unknown parameters and test various hypotheses regarding the relationship.**

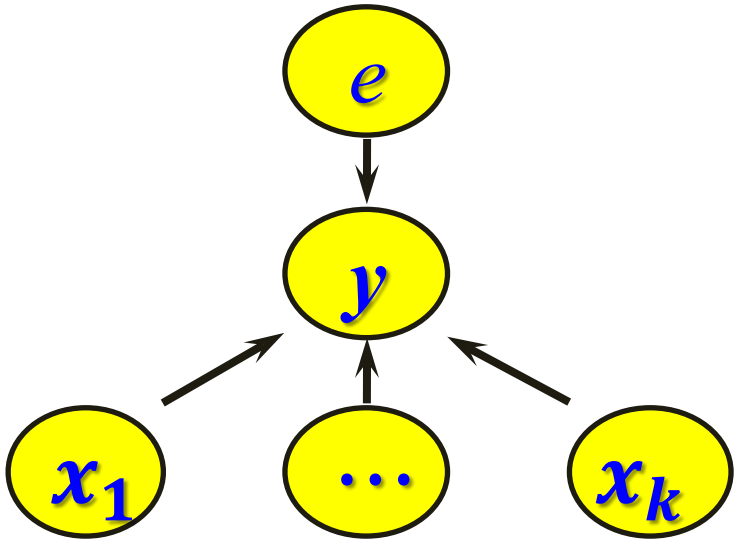


Figure 3: A single outcome variable y depends on multiple ($k \geq 2$) predictor variables x's and random error e.

Table 2: Data layout

Subjects i	**Predictors** x_1 ... x_k	**Outcomes** y
1	x_{11} ... x_{1k}	y_1
2	x_{21} ... x_{2k}	y_2
⋮	⋮ ⋮ ⋮	⋮
n	x_{n1} ... x_{nk}	y_n

CUSTOMAY ~
Can't have missing values, otherwise, the number of observation used for fitting will be reduced.

CONTENTS

1. Ordinary Least Squares Estimator
2. Hypothesis testing
3. Multiple and partial correlations
4. Model evaluation and slections
5. Dummy variables and interactions
6. Model diagnoses

EXAMPLE 1 (HYPERTENSION, PEDIATRICS)

Newborn blood pressure (y) is thought to be affected by weight (x_1) and age (x_2) when both blood pressure and weight are measured.

Table 3 contains the actual data points of 16 infants*. **Figure 4** is the scatter plot of the data points.

*Source: From Rosner, Bernard. Fundamentals of Biostatistics. 7th Edition, 2011 Duxbury, Brooks/Cole, Cengage Learning. Page 469.

Table 3: Sample data for infant birth weight, age and blood pressure for 16 infants.

i	Weight (oz) (x_1)	Age (days) (x_2)	SBP (mm HG) (y)
1	135	3	89
2	120	4	90
3	100	3	83
4	105	2	77
5	130	4	92
6	125	5	98
7	125	2	82
8	105	3	85
9	120	5	96
10	90	4	95
11	120	2	80
12	95	3	79
13	120	3	86
14	150	4	97
15	160	3	92
16	125	3	88

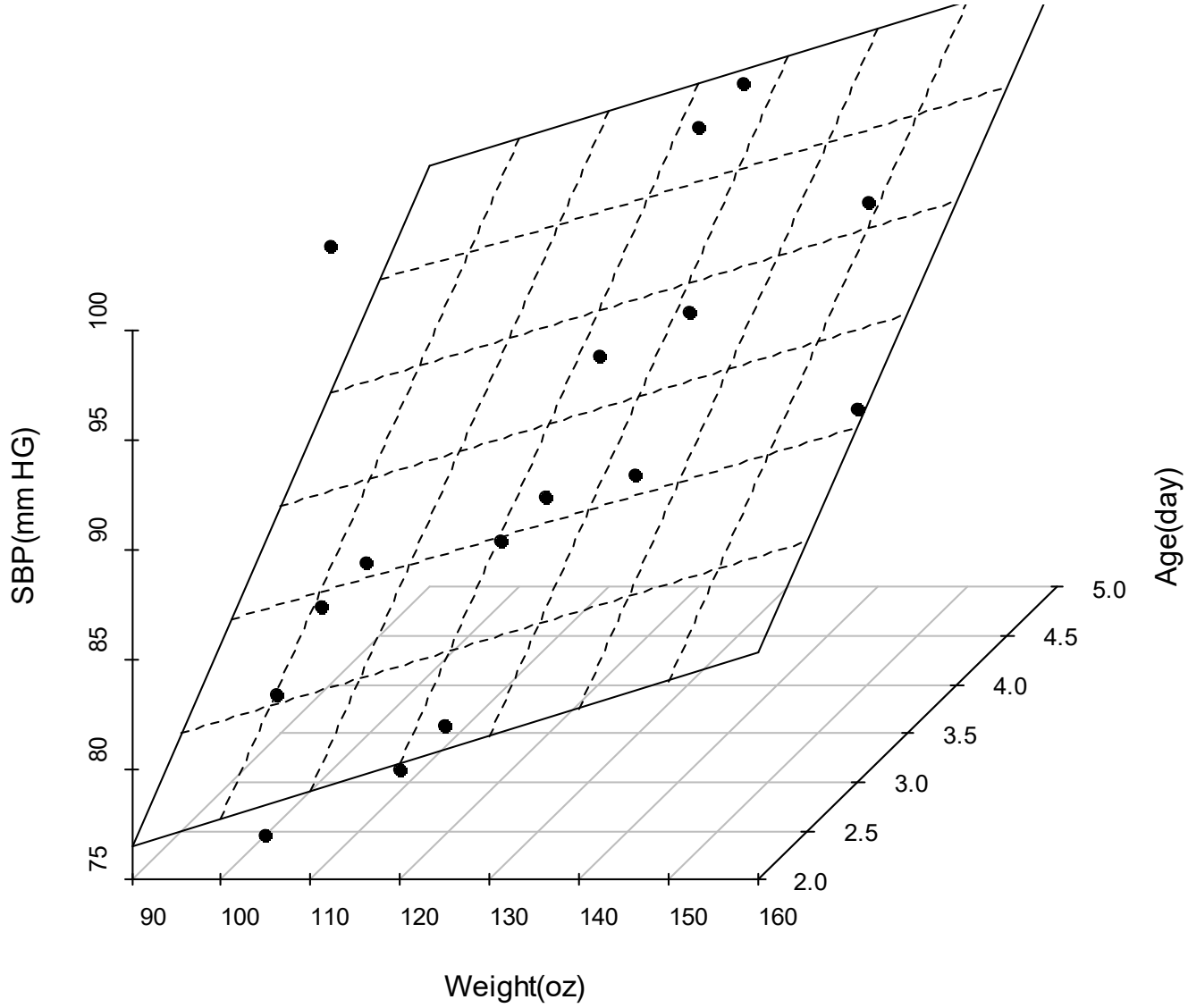


FIGURE 4: SCATTER PLOT OF THE DATA POINTS FOR INFANT (WEIGHT, AGE) AND SBP FOR 16 INFANTS (SEE ALSO THE ANIMATION).

3. MULTIPLE REGRESSION MODEL

As can be seen from the scatter plot (**Figure 4**), there appears to be a linear relationship between (x_1, x_2)=(weight, age) and y = SBP.

- **Linearity:** We *postulate* a linear relationship between y and (x_1, x_2) that is of the following form:

$$E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

That is, for all infants with *identical (weight, age)*= (x_1, x_2), the average SBP is $E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

In addition to (weight, age), there are other (unobserved) factors influencing SBP. So, the relationship $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ is *not* expected to hold *exactly*. Therefore, an error term e, which represents the variance of SBP among *all infants with identical* (x_1, x_2) =(weight, age), is introduced. For an individual infant, we *assume* linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e.$$

In some textbooks, ε

- **Normality:** Often, we assume the residual $e \sim N(0, \sigma^2)$, where the residual variance σ^2 is unknown.

- **Homoscedasticity:** Let e have mean 0 and variance σ^2, regardless of the specific value of (x_1, x_2). In other words, residual variance σ^2 *is assumed to be the same for all* (x_1, x_2) *pairs*.

In summary, for two predictors, the full linear model is of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$, where $e \sim N(0, \sigma^2)$ and $(\beta_0, \beta_1, \beta_2, \sigma^2)$ are unknown parameters.

MULTIPLE LINEAR MODEL

- In general, the full linear model relating a single outcome y to k (≥ 2)predictors$\{x_1, \dots, x_k\}$ is of the form

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e,$$

where $e \sim N(0, \sigma^2)$ and $(\beta_0, \beta_1, \dots, \beta_k, \sigma^2)$ are unknown parameters.

- The regression coefficients β_j's in a multiple regression model are referred to as ***partial regression coefficients***. Each β_j can be interpreted as **the average increase in y per unit increase in x_j, holding all other variables constant.** β_0 is the value of Y when all of the independent variables (X_1 through X_k) are equal to zero.

ceteris paribus

4. ESTIMATION FOR THE UNKNOWNS

For multiple $(k \geq 2)$ predictors, let $\{(y_i, x_i = (x_{i1}, \dots, x_{ik})) : i = 1, \dots, n\}$ be n observations of (y, x) on $n$subjects.

The regression coefficients $\beta_0, \beta_1, \dots, \beta_k$ can be estimated per the popular **Method of Least Squares**. The residual variance σ^2 can be estimated using the individual residual squares.

- To be more explicit, $(\beta_0, \beta_1, \dots, \beta_k)$ will be estimated by minimizing the sum of $(y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2$:

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2 = \sum_{i=1}^{n} \left(y_i - b_0 - \Sigma b_i x_{ij}\right)^2 = \min!$$

- Let $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ be the Least Squares Estimates of $(\beta_0, \beta_1, \dots, \beta_k)$.Define $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$ for an arbitrary $(x_1, \dots, x_k)$tube. In particular, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$. An good estimator for σ^2 is given by

$$\hat{\sigma}^2 = \frac{\text{ResSS}}{n-k-1} = \frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

```
lm(formula = sbp ~ weight + age, data = dt)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0438 -1.3481 -0.2395  0.9688  6.6964

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.45019    4.53189  11.794 2.57e-08 ***
weight       0.12558    0.03434   3.657   0.0029 **
age          5.88772    0.68021   8.656 9.34e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.479 on 13 degrees of freedom
Multiple R-squared:  0.8809,    Adjusted R-squared:  0.8626
F-statistic: 48.08 on 2 and 13 DF,  p-value: 9.844e-07
```

$$y = 53.45019 + 0.12558x_1 + 5.88772x_2$$

the standardized regression coefficient (bs)

- It represents the estimated average increase in *y* (expressed in standard deviation units of *y*) per standard deviation increase in *x*, after adjusting for all other variables in the model.

$$b_s = b \times \frac{S_x}{S_y}$$

```
> sd(dt$sbp)
[1] 6.687987
> sd(dt$weight)
[1] 18.75
> sd(dt$age)
[1] 0.9464847
```

$$b_s(birthweight) = 0.12558 \times \frac{18.75}{6.687987} = 0.352$$

$$b_s(age \text{ in days}) = 5.8887 \times \frac{0.946847}{6.687987} = 0.833$$

```
#adopt lm.beta to get standardized regression coefficient
library(lm.beta)
lm.model.std <- lm.beta(fit2)
summary(lm.model.std)
```

```
Coefficients:
            Estimate Standardized Std. Error t value Pr(>|t|)
(Intercept) 53.45019           NA    4.53189  11.794 2.57e-08 ***
weight       0.12558      0.35208    0.03434   3.657   0.0029 **
age          5.88772      0.83323    0.68021   8.656 9.34e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note 1: For given data points, the computation of the estimates alone does not require the linearity assumption, the normality assumption, or the homoscedasticity assumption.

Note 2: However, the ***data generating mechanism (real model)*** might differ from be ***the postulated linear model (working model)*****,** because our knowledge is limited!

Note 3: The discrepancy between the real model and the linear model determines the properties of the estimators. The estimates and related inference may not be informative or may be misleading the assumptions are severely violated.

4. Multiple Coefficient of Determination

- In *multiple regression*, R^2 is defined as

$$\boldsymbol{R^2 = \frac{RegSS}{TotalSS}},$$

where $\text{TotalSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $\text{RegSS} = (\text{TotalSS}) - (\text{Res SS})$, $\text{ResSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- R^2 is interpreted as the proportion of total variation of y that is explained by *the linear combination of regressors* $x_1, \ldots, x_k$. If the linear fit is perfect, then $R^2 = 1$. If x has *no linear relationship* to y, then $R^2 = 0$.