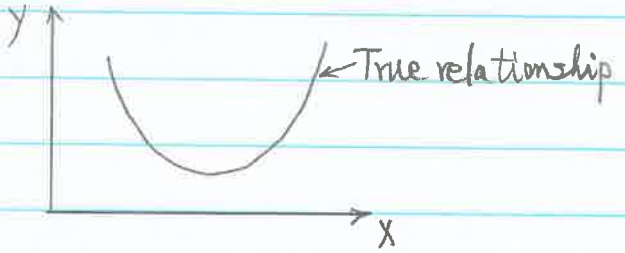
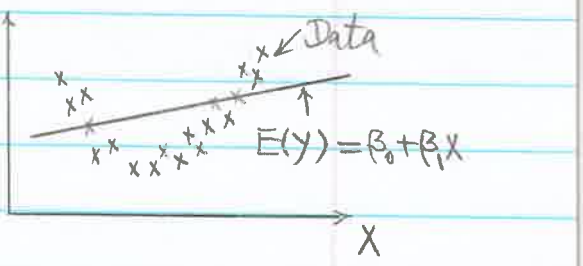
Chapter 2 Assessing the Straight Line Fit

1. Testing for lack of fit

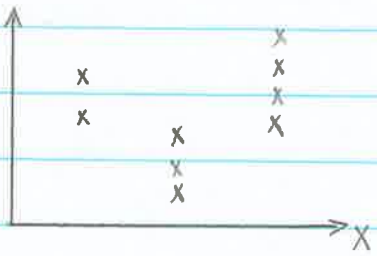
In Chapter 1, we introduced SLR model: $Y = \beta_0 + \beta_1 X + \varepsilon \Leftrightarrow E(Y) = \beta_0 + \beta_1 X$.

Question: Is the model adequate?





Replication:



At X_i, have $Y_{i1}, Y_{i2}, \cdots, Y_{in_i}$, $i = 1, \cdots, m$, i.e.,

X_1	X_2	$\cdots$	X_m
Y_{11}	Y_{21}	$\cdots$	Y_{m1}
Y_{12}	Y_{22}	$\cdots$	Y_{m2}
$\vdots$	$\vdots$		$\vdots$
Y_{1n_1}	Y_{2n_2}	$\cdots$	Y_{mn_m}

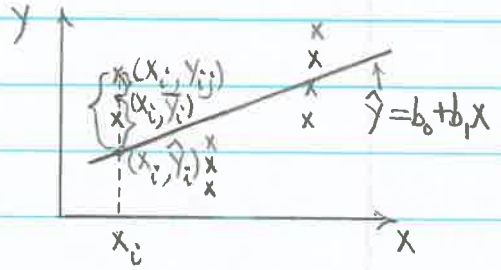
total observations: $n = \sum_{i=1}^{m} n_i$, m = #distinct X's, $Var(Y_{ij}) = \sigma^2$.

Deviation about the fitted line can be decomposed into:

(1) deviation of Y_{ij} from $\bar{Y}_i$, plus

(2) deviation of $\bar{Y}_i$ from $\hat{Y}_{ij} = \hat{Y}_i$, i.e.,

$$e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{Y}_i$$
$$= (Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \hat{Y}_i).$$



— Square both sides

$$(Y_{ij} - \hat{Y}_i)^2 = (Y_{ij} - \bar{Y}_i)^2 + (\bar{Y}_i - \hat{Y}_i)^2 + 2(Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \hat{Y}_i).$$

— Sum over i and j

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_i)^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{Y}_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \hat{Y}_i)$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{m} n_i (\bar{Y}_i - \hat{Y}_i)^2.$$

Note: $\sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \hat{Y}_i) = 0.$

Residual sum of squares = Pure error sum of squares
+ Lack of fit sum of squares.

$RSS = SS(pe) + SS(lof).$

df of $\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$ is $n_i - 1$, $i = 1, 2, \ldots, m$.

$\Rightarrow$ (df of SS(pe)) $= \sum_{i=1}^{m}(n_i - 1) = \sum_{i=1}^{m} n_i - m = n - m.$

Then, (df of SS(lof)) = (df of RSS) − (df of SS(pe))

$= (n-2) - (n-m) = m-2.$

Notes: (1) $SS(pe) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$ does not depend on the model.

(2) $SS(lof) = \sum_{i=1}^{m} n_i(\bar{Y}_i - \hat{Y}_i)^2$ depends on the model fitted and measures model lack of fit.

Modify ANOVA table to "break out" pure error and lack of fit SS.

Source of variation	df	SS	MS
Regression	1	SS_{reg}	$SS_{reg}/1 = SS_{reg}$
Residual	$n-2$	RSS	$RSS/(n-2)$
{ Lack of fit	$m-2$	SS(lof)	$SS(lof)/(m-2)$
{ Pure error	$n-m$	SS(pe)	$SS(pe)/(n-m)$
Total (corrected)	$n-1$	TSS	

Testing H_0: $Y = \beta_0 + \beta_1 X + \varepsilon$ (A SLR model is adequate) vs.

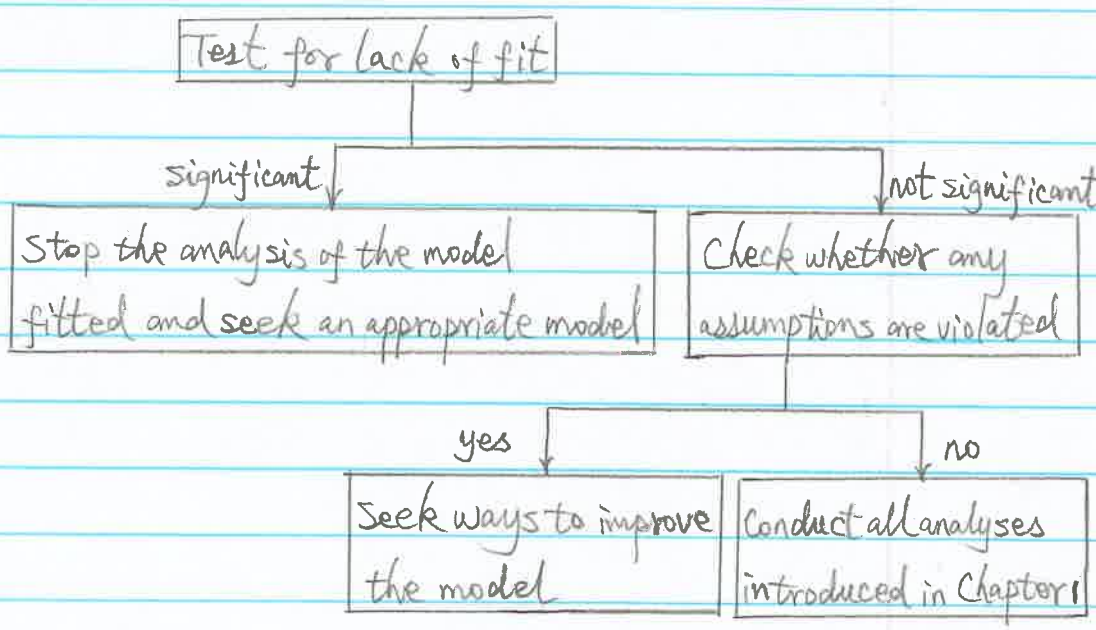
H_a: $Y \neq \beta_0 + \beta_1 X + \varepsilon$ (A SLR model is inadequate).

Test statistic: $F = \frac{MS(lof)}{MS(pe)}$ Under H_0 $F_{m-2,\,n-m}$.

Reject H_0 if p-value $= P(F_{m-2,n-m} \geq F_{obs}) \leq \alpha$ or $F_{obs} \geq F_{m-2,n-m}(1-\alpha)$, where $F_{m-2,n-m}(1-\alpha)$ is the $(1-\alpha)^{th}$ quantile of $F_{m-2,n-m}$.

Notes: (1) For the lack of fit test, we need at least 3 distinct X values and replication at at least one X value.

(2) Analysis steps for data with at least 3 distinct X values and replication at at least one X value.



2. Check Model Assumptions

Model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, 2, \cdots, n$.

Model assumptions: (1) $E(\varepsilon_i) = 0$,

(2) $Var(\varepsilon_i) = \sigma^2$ — constant variance,

(3) $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are independent,

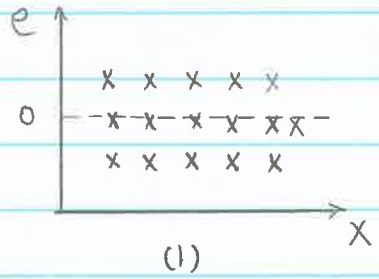
(4) ε_i is normally distributed.

$\Leftrightarrow \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$.

Since ε_i is estimated by $e_i = Y_i - \hat{Y}_i$, we check the assumptions $e_1, e_2, \cdots, e_n$ — residuals.

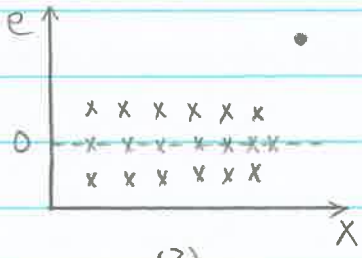
Residual plots: Plot e_i vs. X_i (on X axis)
or e_i vs. $\hat{Y}_i$ $(= b_0 + b_1 X_i)$.

Note: Since $\hat{Y}_i = b_0 + b_1 X_i$ is simply a linear function of X_i, the only real difference between the two types of residual plots is the scale on the horizontal axis. The pattern of points in the residual plots will be the same, and it is the pattern of points that is important, not the scale.



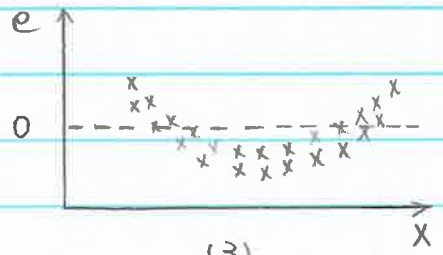
(1)

If model is correct, no pattern.
— residuals are roughly uniformly distributed about the horizontal line $e = 0$.



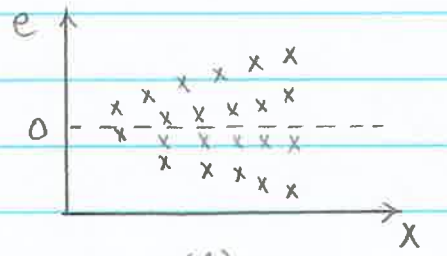
(2)

— a possible outlier.



(3)

— Curvature indicates a curved rather than linear relationship between X and Y. (Note: Sometimes it is too subtle to pick up in the scatter plot of Y_i vs. X_i.)



(4)

— "Funnel" shape indicates non-constant variance (here variance increases with X).

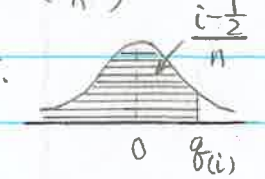
Note: Read question and answer 1, 2, 3 on pages 63-67.

Assessing normality

Normal plot (Quantile-Quantile plot, i.e., Q-Q plot)

(1) Order residuals $e_1, e_2, \cdots, e_n$ from the smallest to the largest to get the ordered residuals: $e_{(1)}, e_{(2)}, \cdots, e_{(n)}$.

(2) Calculate the normal scores: $g_{(i)} = \Phi^{-1}(\frac{i-\frac{1}{2}}{n})$, $i = 1, 2, \cdots n$, where $\Phi^{-1}(\cdot)$ is the inverse function of the standard normal distribution function $\Phi(x) = P(z \le x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$ (i.e., the $(\frac{i-\frac{1}{2}}{n})$th quantile of $N(0,1)$). Thus, $\Phi(g_{(i)}) = P(z \le g_{(i)}) = \frac{i-\frac{1}{2}}{n}$.



(3) Plot the pairs $(g_{(1)}, e_{(1)}), (g_{(2)}, e_{(2)}), \cdots, (g_{(n)}, e_{(n)})$.

A substantial linear pattern in a normal plot indicates that the

normality assumption remains tenable. Otherwise, normality is suspect.

Note: The judgement as to whether a normal plot does or does not show a substantial linear pattern is somewhat subjective.

Test for normality

The Shapiro and Wilk test

H_0: $e_1, e_2, \cdots, e_n$ come from a normal population vs.
(the normality assumption is tenable.)

H_a: $e_1, e_2, \cdots, e_n$ come from a non-normal population.
(the normality assumption is violated.)

Test statistic: $W = \dfrac{\left[\sum_{i=1}^{n} a_i e_{(i)}\right]^2}{\sum_{i=1}^{n} e_i^2}$.

Reject H_0 if p-value $\leq \alpha$.

Notes: (1) Residuals $e_1, e_2, \cdots, e_n$ are not independent. However, the effect of correlations between residuals is usually negligible.

(2) When intercept is in model, $\sum_{i=1}^{n} e_i = 0 \iff \bar{e} = \dfrac{\sum_{i=1}^{n} e_i}{n} = 0$,
Thus, there is need to check $\bar{e} = 0$.

Q: What if the SLR is inadequate or some assumptions are violated?
A: Transform data — Transformations, which are the focuses of Chapter 12 & 13.

Example 2.1: For the data given in problem **F** on page 99 (i.e., the data for HW #1),

X	4.7	5	5.2	5.2	5.9	4.7	5.9	5.2	5.3	5.9	5.6	5
Y	3	3	4	5	10	2	9	3	7	6	6	4

(a) Test to determine whether it is adequate to fit the data by a straight line at $\alpha = 0.05$.
(b) Conduct Shapiro and Wilk test for normality based on residuals using $\alpha = 0.05$.

Example 2.2: In a study, 15 American females aged 30-39 are randomly selected. The following data on X = height (in) and Y = weight (lb) are observed.

X	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
Y	113	115	118	121	124	128	131	134	137	141	145	150	153	159	164

Is the relationship between X and Y linear?