The Lack of Fit Test

Review: use the General Linear Test (GLT) approach to test the linear impact

Ho: $\boldsymbol{\beta_1 = 0}$ versus Ha: $\boldsymbol{\beta_1 \neq 0}$

Full model: $Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$

Under Ha

$$SSE(F) = \Sigma\left(Y_i - \hat{Y}_i\right)^2 = SSE, \quad df_F = n - 2$$

Reduced model: $Y_i = \beta_0 + \epsilon_i = \bar{Y}_{grand\ mean} + \varepsilon_i$

Under Ho

$$SSE(R) = \Sigma\left(Y_i - \bar{Y}_{grand\ mean}\right)^2 = SSTO, \quad df_R = n - 1$$

"Significant reduction in SSE?" →

$$F^* = \frac{\dfrac{SSE(R) - SSE(F)}{df_R - df_F}}{SSE(F)/df_F} = \frac{MSR}{MSE} \sim F\,(1, n-2)$$

In SLR, the global test (the significance of a model test), the ANVOA F test or the T test for the linear impact are equivalent.

The F test for Lack of Fit

- Formal test for determining whether a specific type of regression function adequately fits the data.
- Assumptions (usual):
  - observations $Y|X$ are
    1. i.i.d.
    2. normally distributed
    3. same variance σ^2
- Requires: repeat observations at one or more X levels (called replicates)

The Bank example

- **11** similar branches of a bank offered gifts for setting up money market accounts
- Minimum initial deposits were specific to qualify for the gift
- Value of gift was proportional to the specified minimum deposit
- Interested in: relationship between specified minimum deposit and number of new accounts opened

Notation

Minimum deposit	Number of new accounts
75	28
75	42
100	112
100	136
125	160
125	150
150	152
175	156
175	124
200	124
200	104

- Y_{11} denotes the first measurement (28)made at the first X level (75).
- Y_{12} denotes the second measurement (42)made at the first X level (75).
- $\bar{Y}_1$ denotes the average $\left(\frac{28+42}{2} = 35\right)$ of all y values at the first X level (75).
- $\hat{Y}_{11}$ denotes the predicted response $(b_0 + b_1 X = 87.5)$ for the first measurement at the first X level (75).
- $\hat{Y}_{12}$ denotes the predicted response $(b_0 + b_1 X = 87.5)$ for the second measurement at the first X level (75).

- $\hat{Y}_{ij}$ denotes the predicted response for the jth measurement at the ith X level.
  $\hat{Y}_{ij} = b_0 + b_1 X_i = \hat{Y}_i$ is the same for all j at the same X_i value.
- $\bar{Y}_i$ denotes the average of all y values at the ith X level, or the group mean.
- $\bar{Y}$ denotes the average of all y values at all X levels, or the grand mean.
- C denotes the number of distinct X levels.
  $c = 6, X_1 = 75\ X_2 = 100, X_3 = 125, X_4 = 150, X_5 = 175, X_6 = 200$
- Most Xi has two replicates except X_4

$$X_4 = 150, Y_4 = 152 = \bar{Y}_4 = 152, \hat{Y}_4 = 51 + 0.5(150) = 126$$
$$X_3 = 125, Y_{31} = 160\ , Y_{32} = 150, \bar{Y}_3 = 155\ , \hat{Y}_3 = 51 + 0.5(125) = 114$$

The F test of ANOVA for **Ho: $\beta_1 = 0$ versus Ha: $\beta_1 \neq 0$**

Q: Does X have significant linear impact on Y?

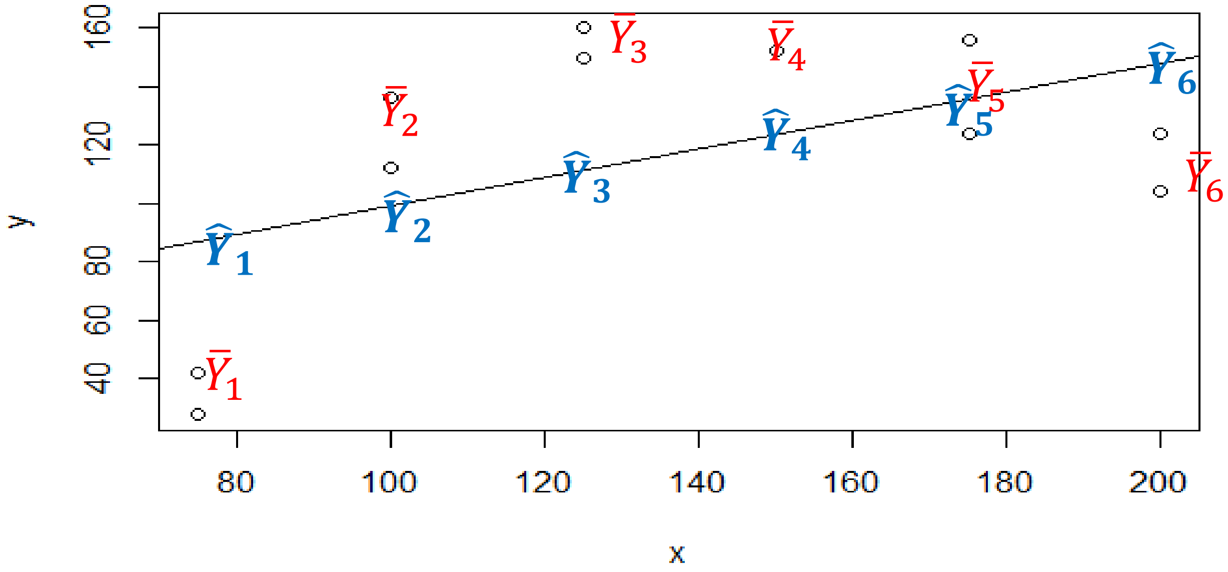
Source of Variation	SS	df	MS	F	Conclusion
Regression	$\mathrm{SSR} = \Sigma(\hat{Y}_i - \bar{Y})^2$	1	MSR= $\frac{\mathrm{SSR}}{1}$	MSR / MSE ~F(1, n-2)	Reject Ho means X has significant Linear impact on Y
Error	$\mathrm{SSE} = \Sigma(Y_i - \hat{Y}_i)^2$	$n-2$	MSE= $\frac{\mathrm{SSE}}{\mathrm{n}-2}$		
Total	$\mathrm{SSTO} = \Sigma(Y_i - \bar{Y})^2$	$n-1$			

The bank example

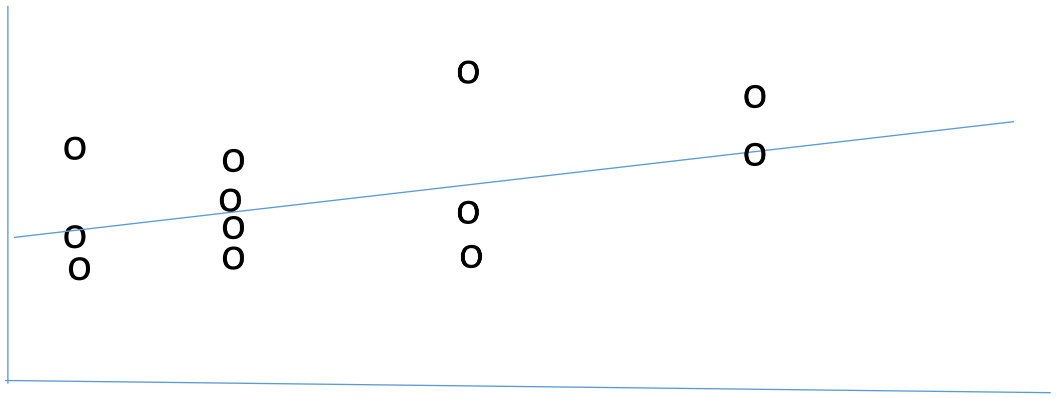
```
Response: y
          Df  Sum Sq Mean Sq F value Pr(>F)
x          1  5141.3  5141.3  3.1389 0.1102
Residuals  9 14741.6  1638.0
```

There is no evidence to reject $\beta_1 = 0$, X seems to have no significant linear impact on Y.

The lack-of-fit property



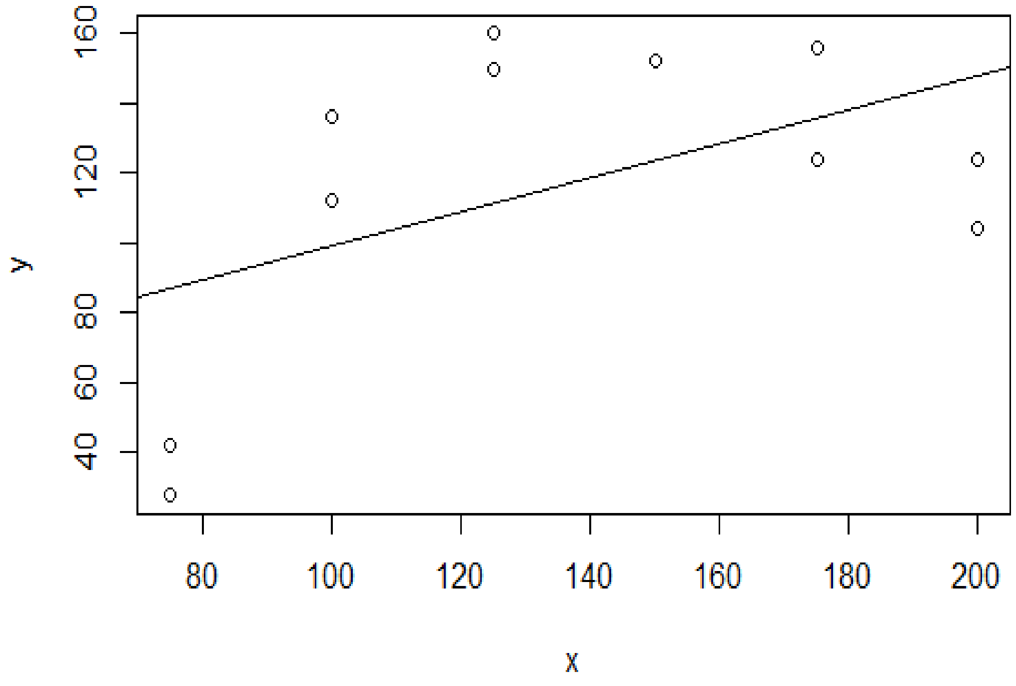
- The linear line is rather flat. But there seems to be more issue found in the scatter plot
- The predictor value $\hat{Y}_i = b_0 + b_1 X_i$ is systematically off from the actual sample mean $\bar{Y}_i$. Such model has a poor fit on the data, or lack of fit.
- This linear model shows X has little impact on Y, and has a lack of fit.



- This model demonstrates X has little impact on Y, but doesn't have a lack of fit issue.

The lack of fit test $\boldsymbol{Ho}$: $\boldsymbol{E\{Y\}(=\mu) = \beta_0 + \beta_1 X}$, $\boldsymbol{Ha}$: $\boldsymbol{E\{Y\}(=\mu) \neq \beta_0 + \beta_1 X}$

Q: Does the linear model fit the data, or is the predicted mean response value the same as the actual mean response value?



Reduced model (Ho) : $\boldsymbol{\hat{Y}_{ij}} = \beta_0 + \beta_1 X_i$

$$SSE(Reduced) = \Sigma\Sigma\left(Y_{ij} - \hat{Y}_i\right)^2 = SSE, \quad dfE_{Reduced} = n - 2$$

Full model (Ha): $\boldsymbol{\hat{Y}_{ij}} = \mu_i + \varepsilon_{ij}$

Specifically,

$\hat{Y}_{1j} = \bar{Y}_1$, the residual $= \mathrm{Y_{1j}} - \bar{Y}_1$ for $j = 1\ or\ 2$

$\hat{Y}_{2j} = \bar{Y}_2$, the residual $= \mathrm{Y_{2j}} - \bar{Y}_2$ for $j = 1\ or\ 2$

$\hat{Y}_{3j} = \bar{Y}_3$, the residual $= \mathrm{Y_{3j}} - \bar{Y}_2$ for $j = 1\ or\ 2$

$\hat{Y}_{4j} = \bar{Y}_4$, the residual $= \mathrm{Y_{4j}} - \bar{Y}_4 =$ 0 for no replicate

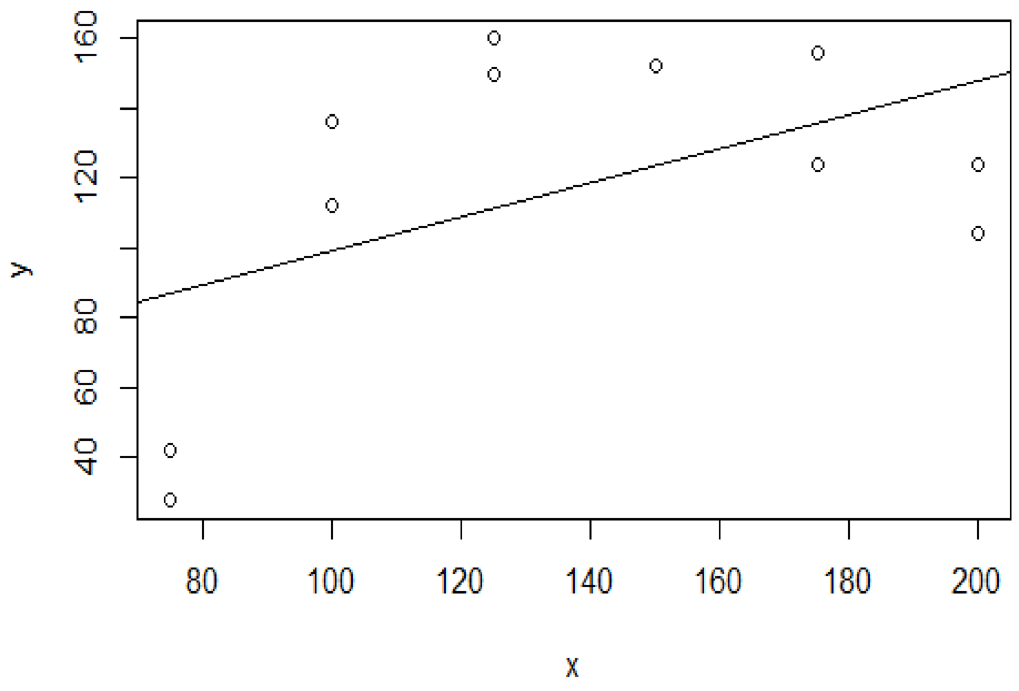
$\hat{Y}_{5j} = \bar{Y}_5$, the residual $= \mathrm{Y_{5j}} - \bar{Y}_5$ for $j = 1\ or\ 2$

$\hat{Y}_{6j} = \bar{Y}_6$, the residual $= \mathrm{Y_{6j}} - \bar{Y}_6$ for $j = 1\ or\ 2$

$$SSE(Full) = \text{Total residuals summing up i and j}$$
$$= \Sigma\Sigma\left(Y_{ij} - \bar{Y}_i\right)^2,$$
$$dfE_{full} = n - 1 + \cdots (n - 1) = n - 6 = n - c$$

The lack of fit test $\boldsymbol{Ho}$: $\boldsymbol{E\{Y\}(=\mu) = \beta_0 + \beta_1 X}$, $\boldsymbol{Ha}$: $\boldsymbol{E\{Y\}(=\mu) \neq \beta_0 + \beta_1 X}$

Q: Does the linear model fit the data, or is the predicted mean response value the same as the actual mean response value?



Reduced model (Ho) : $\hat{\boldsymbol{Y}}_{\boldsymbol{ij}} = \beta_0 + \beta_1 X_i$

$$SSE(Reduced) = \Sigma\Sigma\left(Y_{ij} - \hat{Y}_i\right)^2 = SSE, \quad df_R = n - 2$$

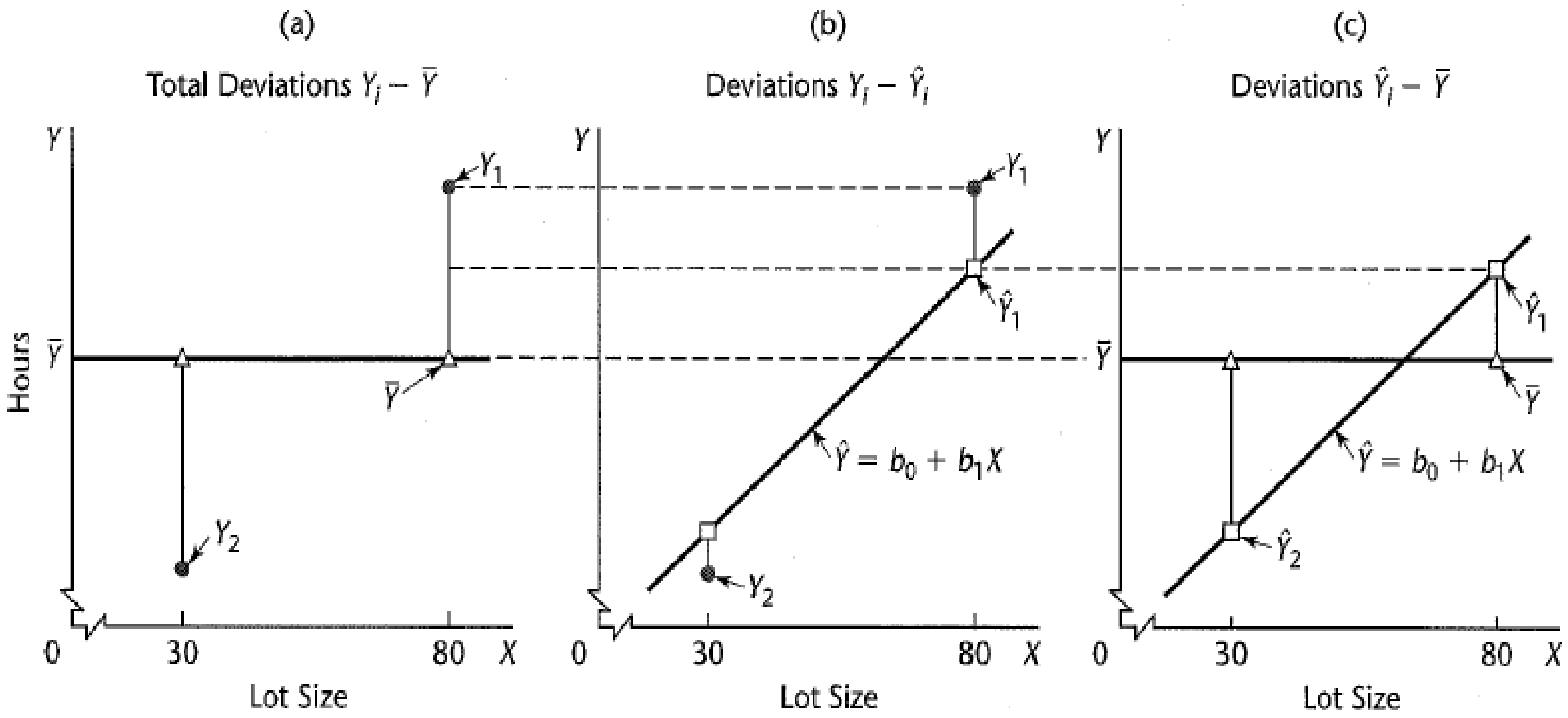
Full model (Ha): $Y_{ij} = \mu_i + \varepsilon_{ij}$

$$SSE(Full) = \Sigma\Sigma\left(Y_{ij} - \bar{Y}_i\right)^2, \quad df_F = n - c$$

$$F = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{SSE(F)/df_F} = \frac{\frac{SSE(R) - SSE(F)}{n - 2 - (n - c)}}{\frac{SSE(F)}{n - c}}$$

$$\sim F\ (c - 2, n - c)$$

Partition the variances



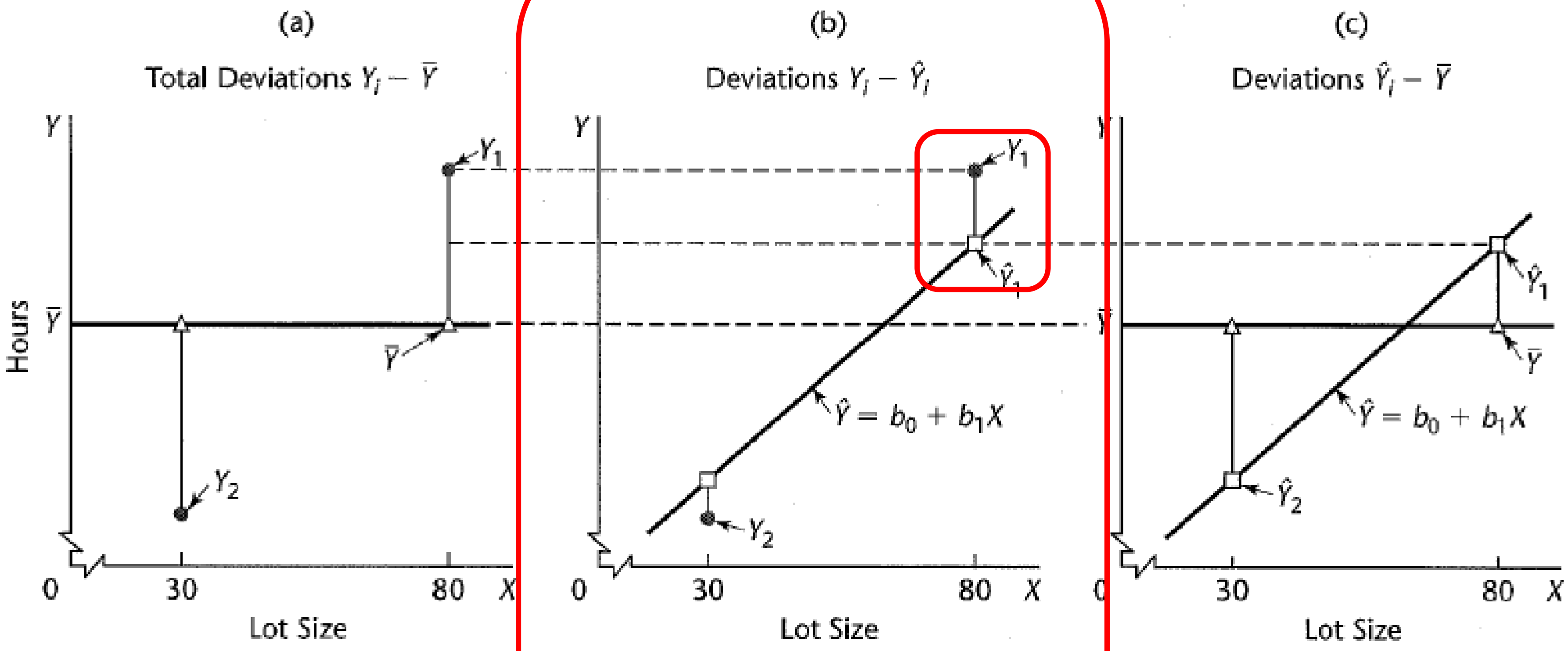
$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(Y_i - \hat{Y}_i)^2 + \Sigma(\hat{Y}_i - \bar{Y})^2$$

$$SSTO = SSE + SSR$$

“Total sum of squares” “ error sum of squares” “regression sum of squares”

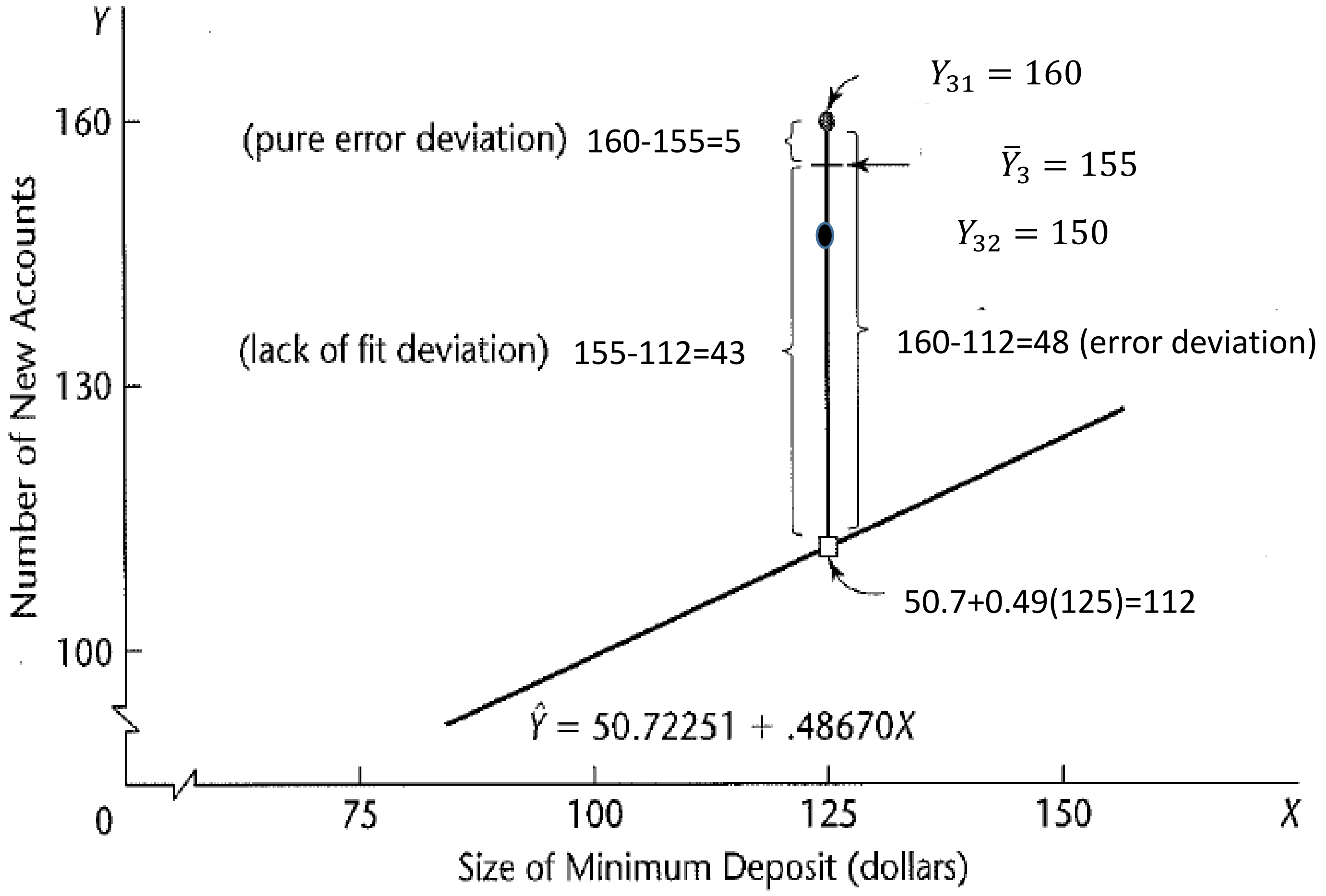
Partition the residual errors for lack of fit

$$Y_i - \bar{Y}$$



$$\mathbf{SSE} = \Sigma(Y_i - \hat{Y}_i)^2$$

Partition the residual errors for lack of fit, SSPE and SSLF



$(Y_{ij} - \hat{Y}_i)$ measures the total error deviation in one observation.

$(Y_{ij} - \bar{Y}_i)$ measure the pure error deviation, which is the randomness result from the data, not from the choice of model.

$(\bar{Y}_i - \hat{Y}_{ij})$ measure the lack of fit deviation, which is the error result from the choice of model and could be improved with a better model.

Do this for every data point, and sum, we have

$$\Sigma\Sigma\left(Y_{ij} - \hat{Y}_{ij}\right)^2 = \Sigma\Sigma\left(Y_{ij} - \bar{Y}_j\right)^2 + \Sigma\Sigma\left(\bar{Y}_j - \hat{Y}_{ij}\right)^2$$

SSE = SSPE + SSLF

Partition the previous ANOVA table on the SSE term further into SSLF and SSPE

Source of Variation	SS	df	MS	F	Conclusion
Regression	$\mathbf{SSR} = \Sigma\Sigma(\hat{Y}_{ij} - \bar{Y})^2$	1	MSR= $\frac{SSR}{1}$	**MSR /MSE ~F(1, n-2)**	**Reject Ho means X has significant Linear impact on Y**
Error	$\mathbf{SSE} = \Sigma\Sigma(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	MSE= $\frac{SSE}{n-2}$		
Lack of fit (in Error)	$\mathbf{SSLF} = \Sigma\Sigma(\bar{Y}_i - \hat{Y}_{ij})^2$	$c-2$	MSLF= $\frac{SSLF}{c-2}$	**MSLF /MSPE ~F(c-2, n-c)**	**Reject Ho means the current model does not fit the data**
Pure error (in Error)	$\mathbf{SSPE} = \Sigma\Sigma(Y_{ij} - \bar{Y}_i)^2$	$n-c$	MSPE= $\frac{SSPF}{n-c}$		
Total	$\mathbf{SSTO} = \Sigma\Sigma(Y_{ij} - \bar{Y})^2$	$n-1$			

Example 1, the R output on the linear impact, or the model significance test

Source of Variation	SS	df	MS	F	Conclusion
Regression	**5141**	1	**5141**	**?**	**?**
Error	**14742**	11-2=9	**1638**		
Lack of fit(in Error)	13594	6-2=4	3398.5		
Pure error(in Error)	**1148**	11-6=5	229.6		
Total	19883	10			

```
bankR.mod<-lm(y~x, bank)
anova(bankR.mod)
```

```
Response: y
          Df  Sum Sq Mean Sq F value Pr(>F)
x          1  5141.3  5141.3  3.1389 0.1102
Residuals  9 14741.6  1638.0
```

Example 2, the R output on the lack of fit test

Source of Variation	SS	df	MS	F	Conclusion
Regression	**5141**	1	**5141**	**3.14 (p=0.11)**	**X does not have significant linear impact on Y**
Error	**14742**	n-2=11-2=9	**1638**		
Lack of fit(in Error)	13594	c-2=6-2=4	3398.5	**?**	**?**
Pure error(in Error)	**1148**	N-c=11-6=5	229.6		
Total	19883	10			

Build the reduced model under Ho: $\hat{Y} = \beta_0 + \beta_1 X$

```
bankR.mod<-lm(y~x, bank)
anova(bankR.mod)
```

Build the full model under Ha: $\hat{Y} = \mu$

```
bankF.mod<-lm(y~as.factor(x),bank)
anova(bankR.mod, bankF.mod)
```

```
Response: y
          Df  Sum Sq Mean Sq F value Pr(>F)
x          1  5141.3  5141.3  3.1389 0.1102
Residuals  9 14741.6  1638.0

Model 1: y ~ x
Model 2: y ~ as.factor(x)
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1      9 14742
2      5  1148  4     13594 14.801 0.005594 **
```

Fs= MSLF/MSPE = $\frac{13594}{4} \div \frac{1148}{5} = 14.801$, this model has a lack of fit issue.

The lack of fit test is not valid without replicates. But we can manually create replicates by grouping.

- SSPE = $\Sigma\Sigma\left(Y_{ij} - \bar{Y}_{ij}\right) = 0$

size	hour
20	113
30	121
40	160
50	221
60	224
70	361
80	399
90	376
100	353
110	435
120	546

```
Model 1: y ~ x
Model 2: y ~ factor(x)
  Res.Df   RSS Df Sum of Sq F Pr(>F)
1      9 16602
2      0     0  9     16602
```

- Solution: grouping

```
g<-c(30,30,30,60,60,60,90,90,90,115,115)
tolucanr$g<-g
tolucanrgR.mod<-lm(y~g, data=tolucanr)
tolucanrgF.mod<-lm(y~factor(g),data=tolucanr)
summary(tolucanrgR.mod)
anova(tolucanrgR.mod)
anova(tolucanrgR.mod,tolucanrgF.mod)
```

```
Model 1: y ~ g
Model 2: y ~ factor(g)
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1      9 21775
2      7 21276  2    498.74 0.082 0.9221
```