## 5 assumptions of Simple Linear Regression (SLR)

- **Existence –** for any fixed value of $X$, $Y$ is a Random Variable with finite mean & variance
  - this defines a set of conditional RVs: $Y|X=x$
- **Independence –** $Y_i$ are independent of each other
  - the $Y_i$ are ***I***ndependent & ***I***dentically ***D***istributed (***iid***) RVs
- **Linearity –** the mean value of $Y$ is a straight-line function of $X$
  - The SLR equation describes $\mu_{Y|X=x}$ in addition to $Y|X=x$

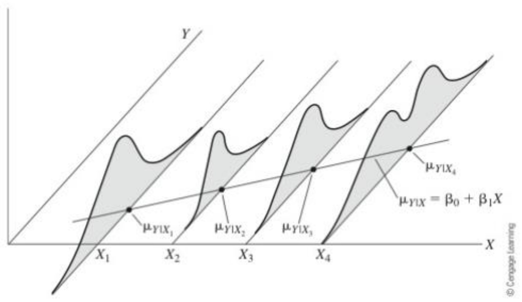
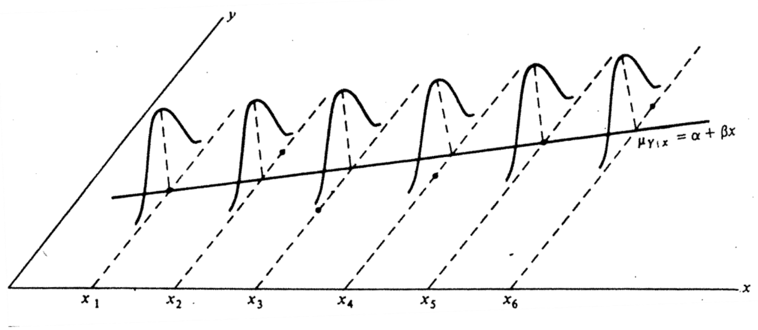



**FIGURE 5.5** Straight-line assumption

- **Homoskedasticity –** the variance of $Y$ is the same for any value of $X$
- **Normality –** for any fixed value of $X$, $Y$ has a normal distribution
  - $Y_i \mid X = x \overset{iid}{\sim} N(\mu_{Y|X=x}, \sigma^2)$ with no subscript on $\sigma^2$




## Statistical Model for Simple Linear Regression

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i\,,\quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)\quad i = 1, 2, \dots, n$
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad e_i = Y_i - \hat{Y}_i$ are the <u>errors</u> or <u>residuals</u>

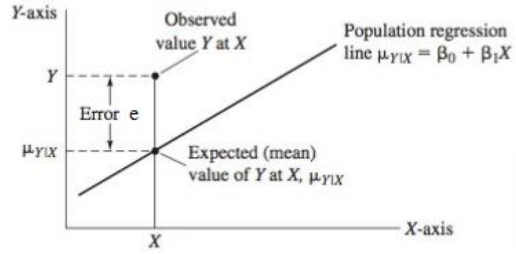



**FIGURE 5.6** Error component e

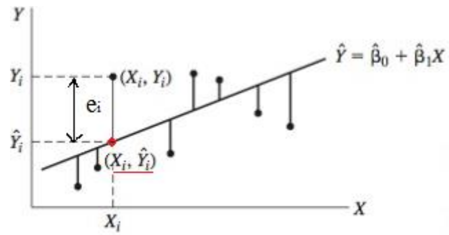



**FIGURE 5.7** Deviations of observed points

- <u>Least Squares Estimates</u> (LSE) for $\beta_0$ and $\beta_1$
  - Minimize the Sum of Squared Errors (SSE)

$$S = SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \left[\sum x_i \sum y_i\right]/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{SSXY}{SSX}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

# Linear Models in R:

```
> model <- lm(dist~speed, data=cars)
> summary(model)
```

```
Call:
lm(formula = dist ~ speed, data = cars)
Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,  Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

$\hat{\beta}_0$ → -17.5791

$\hat{\beta}_1$ → 3.9324

$SE(\hat{\beta}_0)$ → 6.7584

$SE(\hat{\beta}_1)$ → 0.4155

$\hat{\sigma}_e = s_{Y|X}$ → 15.38

$n-2$ → 48

$89.57 = 9.464^2$

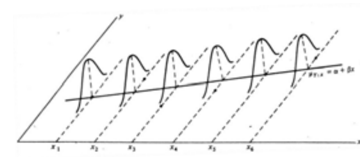
$$dist = -17.5791 + 3.9324 * speed$$

**Prediction of a Conditional Average vs. an Individual Value**

$$\hat{\mu}_{Y|X=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \quad \text{vs.} \quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$$

- Conditional Average:
  - $Y_i \mid X = x_0 \overset{iid}{\sim} N(\beta_0 + \beta_1 x_0, \sigma^2)$
  - $\bar{Y}$ and $\hat{\beta}_1$ are independent RVs
  - CI for $\mu_{Y|X=x_0}$ : $\hat{\mu}_{Y|X=x_0} \pm t_{\frac{\alpha}{2},n-2} SE(\hat{\mu}_{Y|X=x_0})$
  - The collection of CIs at all values for $x_0$ gives an envelope called a *confidence band*



- Individual Value:
  - An individual value has the same prediction, but more variability
  - CI for $\hat{Y}_{X=x_0}$ : $Y_{X=x_0} \pm t_{\frac{\alpha}{2},n-2} SE(\hat{Y}_{X=x_0})$
  - The collection of CIs at all values for $x_0$ gives an envelope called a *prediction band*