PubH 7405:
REGRESSION ANALYSIS

SLR: INFERENCES, Part I

Normal Error Regression Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

The Mean Response:

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

ESTIMATED SLOPE

$$\boxed{\begin{aligned} b_1 &= \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \\ &= \sum \frac{(x_i-\bar{x})}{\sum(x_i-\bar{x})^2} y_i \\ &= \sum k_i y_i \end{aligned}}$$

<u>Reason</u>: Inspecting the numerator of b_1

$$\sum(x_i-\bar{x})(y_i-\bar{y}) = \sum(x_i-\bar{x})(y_i) - (\bar{y})\sum(x_i-\bar{x})$$

$$= \sum(x_i-\bar{x})(y_i)$$

SAMPLING DISTRIBUTION

Under the "Normal Error Regression Model":

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

The sampling distribution of the estimated slope

$\mathbf{b_1}$ is Normal with Mean and Variance :

$$\mathbf{E(b_1) = \boldsymbol{\beta}_1}$$

$$\boldsymbol{\sigma}^2\mathbf{(b_1)} = \frac{\boldsymbol{\sigma}^2}{\sum (\mathbf{x} - \bar{\mathbf{x}})^2}$$

The sampling distribution of the estimated slope, b_1, is "**normal**" because b_1 is a linear combination of the observations y_i and the distribution of each observation is normal under the "normal error regression model".

$$\mathbf{b_1} = \sum \frac{(\mathbf{x_i} - \bar{\mathbf{x}})}{\sum (\mathbf{x_i} - \bar{\mathbf{x}})^2} \mathbf{y_i}$$

$$= \sum \mathbf{k_i} \mathbf{y_i}$$

Next step: $$\mathbf{b_1 = \sum k_i y_i}$$

$$\mathbf{k_i = \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}}$$

We can show that :

$$\sum k_i = 0$$

$$\sum k_i x_i = 1$$

$$\sum k_i^2 = \frac{1}{\sum (x_i - \bar{x})^2}$$

$$\sum (x_i - \bar{x})(x_i) = \sum (x_i - \bar{x})(x_i) - (\bar{x}) \sum (x_i - \bar{x})$$

$$= \sum (x_i - \bar{x})(x_i - \bar{x})$$

$$= \sum (x_i - \bar{x})^2$$

$$\sum k_i x_i = \frac{\sum (x_i - \bar{x})(x_i)}{\sum (x_i - \bar{x})^2} = 1$$

b_1 is an UNBIASED ESTIMATOR

We have :

$$\sum k_i = 0$$

$$\sum k_i x_i = 1$$

$$\sum k_i^2 = \frac{1}{\sum (x_i - \bar{x})^2}$$

$$b_1 = \sum k_i y_i$$

$$k_i = \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$\begin{aligned} E(b_1) &= \sum k_i E(y_i) \\ &= \sum k_i (\beta_0 + \beta_1 x_i) \\ &= \beta_0 \underbrace{\sum k_i}_{0} + \beta_1 \underbrace{\sum k_i x_i}_{1} \\ &= \beta_1 \end{aligned}$$

VARIANCE & STANDARD ERROR

$$\boxed{\begin{aligned} \mathrm{b}_1 &= \sum k_i y_i \\ k_i &= \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \\ Var(b_1) &= \sigma^2(b_1) \\ &= \sum k_i^2 Var(y_i) \\ &= \sigma^2 \sum k_i^2 \\ &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \end{aligned}}$$

$$\sigma^2(b_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$\hat{=} \frac{MSE}{\sum (x_i - \bar{x})^2}$$

$$SE(b_1) = s(b_1)$$

$$= \sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}}$$

$$SE(b_1) = \sqrt{\frac{MSE}{\sum (x - \bar{x})^2}}$$

<u>Design Implication</u>: The larger the sum of squares of X, the more <u>precise</u> the estimate of the Slope.

MORE ON SAMPLING DISTRIBUTION

$$\frac{b_1 - \beta_1}{s(b_1)} = \frac{b_1 - \beta_1}{\sigma(b_1)} \div \frac{s(b_1)}{\sigma(b_1)}$$

distributed as N(0,1)

$$\frac{1}{n-2}\chi^2_{df=n-2}$$

Theorem :

$\dfrac{b_1 - \beta_1}{s(b_1)}$ is distributed as "t" with (n – 2) degrees of freedom

CONFIDENCE INTERVALS

Theorem :

$\dfrac{b_1 - \beta_1}{s(b_1)}$ is distributed as "t" with $(n-2)$ degrees of freedom

$(1-\alpha)100\%$ Confidence Interval for β_1 is :

$$\mathbf{b_1 \pm t(1-\alpha/2; n-2)s(b_1)}$$

$t(1-\alpha/2; n-2)$ is the $(1-\alpha/2)100$ percentile of the "t" distribution with $(n-2)$ degrees of freedom

THE TEST FOR INDEPENDENCE

The Mean Response :

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

$$H_0 : \beta_1 = 0$$

"t" test at (n – 2) degrees of freedom :

$$t = \frac{b_1}{s(b_1)}$$

Theorem :

$\dfrac{b_1 - \beta_1}{s(b_1)}$ is distributed as "t" with (n – 2) degrees of freedom

which is identical to the test using "r":

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

STATISTICAL POWER

The "power", which is (1 – probability of Type II error), of the t-test for independence can be obtained, if needed, from Appendix B5 of the text

$$H_0 : Slope = 0$$

$$H_A : Slope = \beta_1$$

STATISTICAL POWER

The "power", which is (1 – probability of Type II error), of the t-test for independence can be obtained, if needed, from Appendix B5 of the text

$$H_0 : \beta_1 = 0$$

$$t = \frac{b_1}{s(b_1)}; df = n - 2$$

Theorem :

$\frac{b_1 - \beta_1}{s(b_1)}$ is distributed as "t" with (n – 2) degrees of freedom

$$Power = \Pr(|t| > t_{1-\alpha/2} \mid \delta); \delta = \frac{\beta_1}{\sigma(b_1)}$$

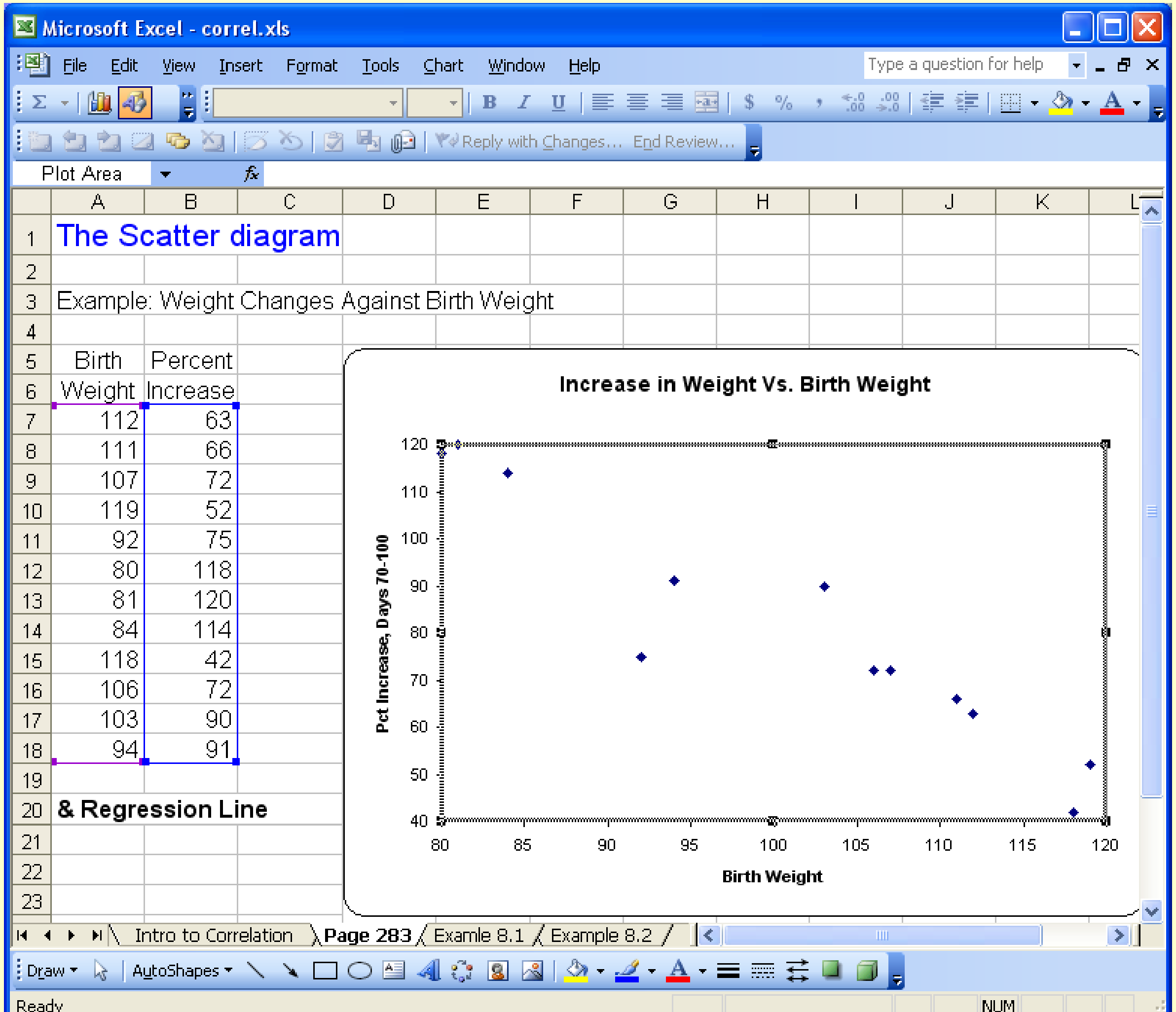
where δ is the noncentrality measure, a standardized measure of how far the true value of the slope is from zero (H_A).

More details on pages 50-51 of your textbook

For practical applications, you will less likely have anything to do with this issue of statistical power. When & if you do, **specification of an Alternative Hypothesis is not easy.**

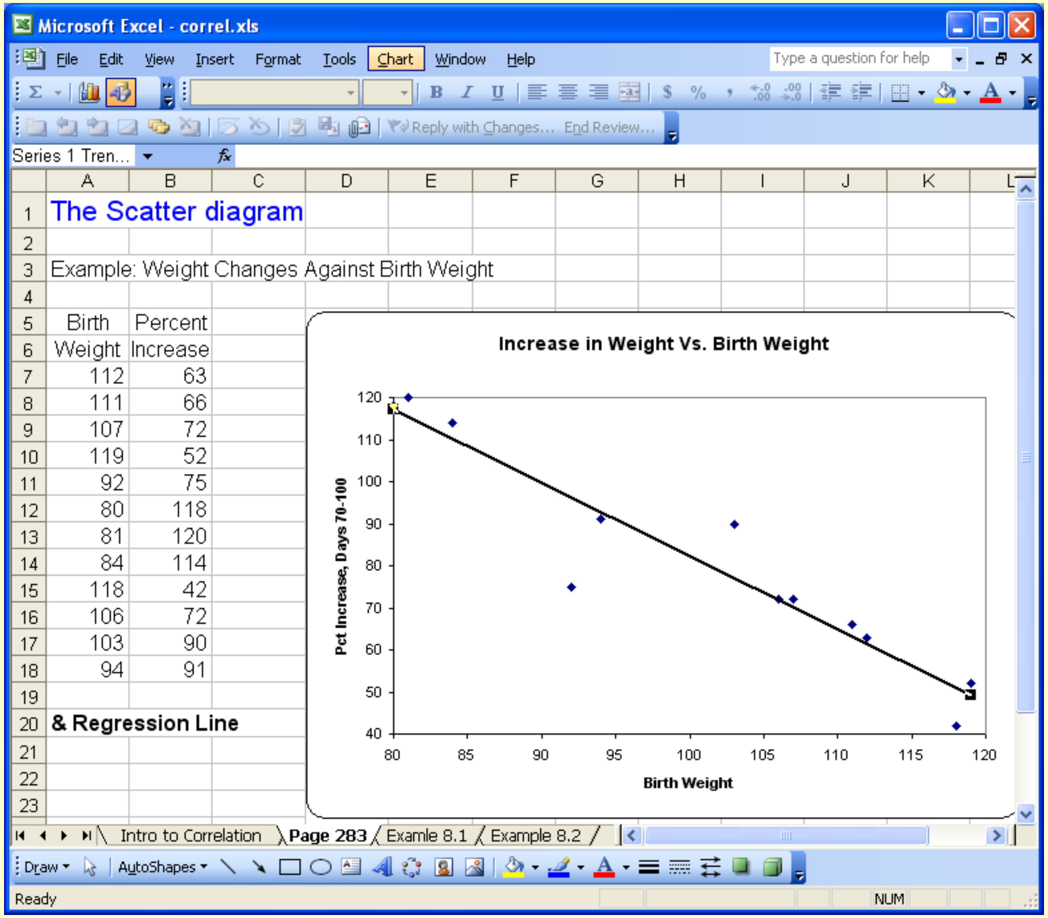
Excel: SCATTER DIAGRAM

Create a scatter diagram using *Chart Wizard*



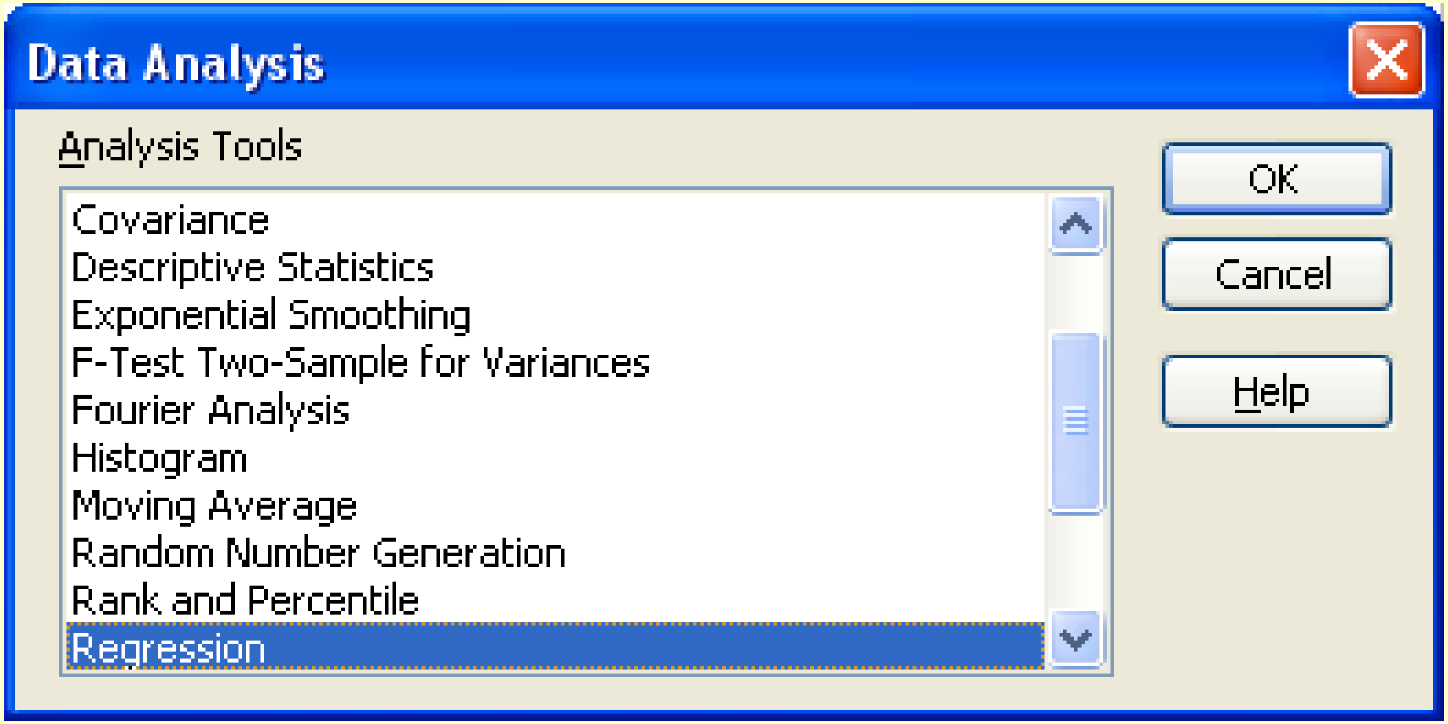
Excel: REGRESSION LINE

Steps: (a) Click on the new Chart (scatter diagram) to make it active, (b) Click on *Chart* (on the top row menu), (c) a box appears to let you choose “Add Trendline”

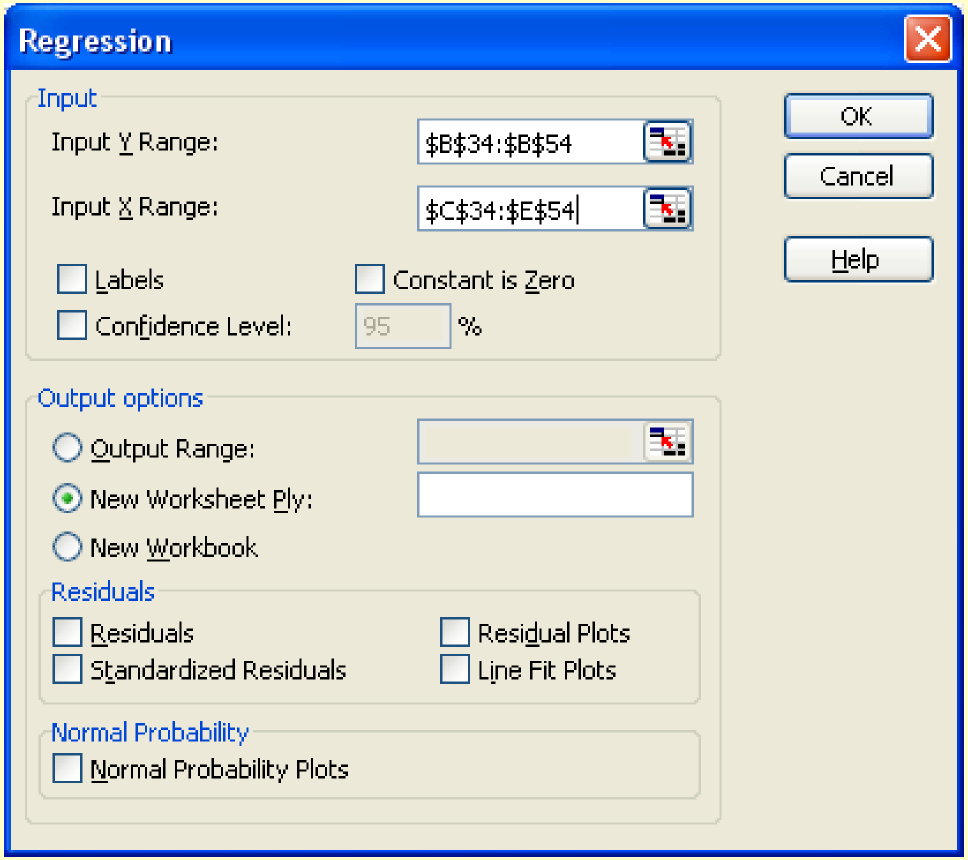


Excel: ANALYSIS

(1) click the *Tools* then (2) *Data Analysis*; among functions available, choose *Regression*.



A box appears, use the cursor to fill in the ranges of Y and X's. The results include all items needed, including **regression estimates of coefficients, their standard errors, and their 95% confidence intervals. And much more.**



BASIC “SAS” PROGRAM

```
data tc;
  input x y;
  label x = 'Lot Size'
        y = 'Work Hrs';
cards;
```

```
 80  399
 30  121
 50  221
 90  376
 70  361
 60  224
120  546
 80  352
100  353
 50  157
 40  160
 70  252
 90  389
 20  113
110  435
100  420
 30  212
 50  268
 90  377
110  421
 30  273
 90  468
 40  244
 80  342
 70  323;
```

(Toluca Company: data go in the middle)

```
proc REG data = tc;
  model y = x;
  plot y*x;
run;
```

EXAMPLE #1: Birth weight data:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	255.971912	19.04537379	13.44011	9.99506E-08	213.536175	298.4076492
X Variable 1	-1.7370861	0.187689258	-9.255117	3.21622E-06	-2.155283843	-1.31888839

x (oz)	y (%)
112	63
111	66
107	72
119	52
92	75
80	118
81	120
84	114
118	42
106	72
103	90
94	91

$$t = \frac{-1.7371}{.1877}$$

$$= -9.255$$

$$95\%\ \text{C.I.} = -1.7371 \pm (2.2281)(.1877)$$

$$= (-2.155, -1.319)$$

EXAMPLE #2: Age and SBP

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	99.9585145	19.25516927	5.191256	0.000174	58.3602504	141.556779
X Variable 1	0.70490069	0.286078656	2.46401	0.028454	0.08686533	1.32293605

Age (x)	SBP (y)
42	130
46	115
42	148
71	100
80	156
74	162
70	151
80	156
85	162
72	158
64	155
81	160
41	125
61	150
75	165

$$t = \frac{.7049}{.2861}$$

$$= 2.464$$

$$95\%\ \text{C.I.} = .7049 \pm (2.1604)(.2861)$$

$$= (.087, 1.323)$$

LotSize	WorkHours
80	399
30	121
50	221
90	376
70	361
60	224
120	546
80	352
100	353
50	157
40	160
70	252
90	389
20	113
110	435
100	420
30	212
50	268
90	377
110	421
30	273
90	468
40	244
80	342
70	323

EXAMPLE #3: Toluca Company Data (Description on page 19 of Text)

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	62.3658586	26.17743389	2.382428	0.025851	8.21371106	116.518006
X Variable 1	3.57020202	0.346972157	10.28959	4.45E-10	2.85243543	4.28796861

$$t = \frac{3.5702}{.3470}$$

$$= 10.290$$

$$95\% \text{ C.I.} = 3.5702 \pm (2.0687)(.3470)$$

$$= (2.852, 4.288)$$

UNBIASED LINEAR ESTIMATORS

We consider the family of all unbiased linear estimators and prove that b_1 is a member of this family with minimum variance:

$$\hat{\beta}_1 = \sum c_i y_i$$

$$(b_1 = \sum k_i y_i)$$

$$\hat{\beta}_1 = \sum c_i y_i$$

$$E(\hat{\beta}_1) = \sum c_i E(y_i)$$

$$= \sum c_i (\beta_0 + \beta_1 x_i)$$

$$= \beta_0 (\sum c_i) + \beta_1 (\sum c_i x_i)$$

Conditions :

$$\sum c_i = 0$$

$$\sum c_i x_i = 1$$

b_1 satisfies these two conditions with $c_i = k_i$

Want to prove that:

$\sum k_i d_i = 0$

$$\mathbf{c_i} = \mathbf{k_i} + \mathbf{d_i}$$

$$\sum k_i d_i = \sum k_i (c_i - k_i)$$

$$= \sum c_i k_i - \sum k_i^2$$

$$= \sum c_i \left[\frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2} \right] - \frac{1}{\sum (x_i - \bar{x})^2}$$

$$\overset{1}{=} \frac{\sum c_i x_i - \bar{x} \overset{0}{\sum c_i}}{\sum (x_i - \bar{x})^2} - \frac{1}{\sum (x_i - \bar{x})^2}$$

$$= 0$$

$$\hat{\beta}_1 = \sum c_i y_i$$

$$\sigma^2(\hat{\beta}_1) = \sigma^2 \sum c_i^2$$

$$= \sigma^2 \sum (k_i + d_i)^2$$

$$= \sigma^2 \left[\sum k_i^2 + \sum d_i^2 + 2 \sum k_i d_i \right]$$

$$= \sigma^2(b_1) + \sigma^2 \sum d_i^2 + (\sigma^2)(0)$$

$$\sigma^2(\hat{\beta}_1) \geq \sigma^2(b_1)$$

b_1 is the member of the family with minimum variance.

ESTIMATED INTERCEPT

Recall:

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

SAMPLING DISTRIBUTION

Under the "Normal Error Regression Model":

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

The sampling distribution of the estimated intercept $\mathbf{b_0}$ is Normal with Mean and Variance :

$$\mathbf{E(b_0) = \beta_0}$$

$$\boldsymbol{\sigma}^2(\mathbf{b_0}) = \boldsymbol{\sigma}^2 \left[\frac{\mathbf{1}}{\mathbf{n}} + \frac{\bar{\mathbf{x}}^2}{\sum (\mathbf{x} - \bar{\mathbf{x}})^2} \right]$$

Next Step :

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$= \frac{1}{n}\sum y_i - \bar{x}\sum k_i y_i$$

$$= \sum \left\{ \frac{1}{n} - \bar{x} k_i \right\} y_i$$

The sampling distribution of b_0 is "normal" because b_0, like b_1, is a linear combination of the observations y_i and the distribution of each observation is normal under the "normal error regression model":

b_0 is an UNBIASED ESTIMATOR

$$E(b_0) = E(\bar{y}) - \bar{x}E(b_1)$$

$$= \frac{\sum E(y_i)}{n} - \left(\frac{\sum x_i}{n}\right)\beta_1$$

$$= \frac{\sum(\beta_0 + \beta_1 x_i)}{n} - \left(\frac{\sum x_i}{n}\right)\beta_1$$

$$= \beta_0$$

VARIANCE & STANDARD ERROR

$$\mathrm{b}_0 = \bar{y} - b_1 \bar{x}$$

$$\mathrm{Var}(\mathrm{b}_0) = Var(\bar{y}) + (\bar{x})^2 Var(b_1)$$

$$\sigma^2(\mathrm{b}_0) = \frac{\sigma^2}{n} + (\bar{x})^2 \frac{\sigma^2}{\sum (x - \bar{x})^2}$$

$$= \sigma^2 \left\{ \frac{1}{n} + \frac{(\bar{x})^2}{\sum (x - \bar{x})^2} \right\}$$

$$\sigma^2(\mathrm{b}_0) = \sigma^2\left\{\frac{1}{n} + \frac{(\bar{x})^2}{\sum(x-\bar{x})^2}\right\}$$

$$s^2(\mathrm{b}_0) = MSE\left\{\frac{1}{n} + \frac{(\bar{x})^2}{\sum(x-\bar{x})^2}\right\}$$

$$\mathbf{SE(b_0)} = \sqrt{\mathbf{MSE}\left\{\frac{\mathbf{1}}{\mathbf{n}} + \frac{(\bar{\mathbf{x}})^{\mathbf{2}}}{\sum(\mathbf{x}-\bar{\mathbf{x}})^{\mathbf{2}}}\right\}}$$

DESIGN IMPLICATION

$$\mathbf{SE(b_1)} = \sqrt{\frac{\mathbf{MSE}}{\sum (\mathbf{x} - \bar{\mathbf{x}})^2}}$$

$$\mathbf{SE(b_0)} = \sqrt{\mathbf{MSE}\left\{\frac{\mathbf{1}}{\mathbf{n}} + \frac{(\bar{\mathbf{x}})^2}{\sum (\mathbf{x} - \bar{\mathbf{x}})^2}\right\}}$$

These **Standard Errors**, for given n, are affected by the spacing of the X's levels in the data. **The larger the sum of squares of X, the more precise the estimates of the Slope and the Intercept.**

MORE ON SAMPLING DISTRIBUTION

$$\frac{b_0 - \beta_0}{s(b_0)} = \frac{b_0 - \beta_0}{\sigma(b_0)} \div \frac{s(b_0)}{\sigma(b_0)}$$

distributed as N(0,1)

$$\frac{1}{n-2}\chi^2_{df=n-2}$$

Theorem :

$\dfrac{b_0 - \beta_0}{s(b_0)}$ is distributed as "t" with (n – 2) degrees of freedom

CONFIDENCE INTERVALS

Theorem :

$\frac{b_0 - \beta_0}{s(b_0)}$ is distributed as "t" with $(n-2)$ degrees of freedom

$(1-\alpha)100\%$ Confidence Interval for β_0 is :

$$\mathbf{b_0 \pm t(1-\alpha/2; n-2)s(b_0)}$$

$t(1-\alpha/2; n-2)$ is the $(1-\alpha/2)100$ percentile of the "t" distribution with $(n-2)$ degrees of freedom

EXAMPLE #1: Birth weight data:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	255.971912	19.04537379	13.44011	9.99506E-08	213.536175	298.4076492
X Variable 1	-1.7370861	0.187689258	-9.255117	3.21622E-06	-2.155283843	-1.31888839

x (oz)	y (%)
112	63
111	66
107	72
119	52
92	75
80	118
81	120
84	114
118	42
106	72
103	90
94	91

$$t = \frac{255.9719}{19.0454}$$

$$= 13.440$$

$$95\%\ \text{C.I.} = 255.9719 \pm (2.2281)(19.0454)$$

$$= (213.536, 298.4076)$$

EXAMPLE #2: Age and SBP

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	99.9585145	19.25516927	5.191256	0.000174	58.3602504	141.556779
X Variable 1	0.70490069	0.286078656	2.46401	0.028454	0.08686533	1.32293605

Age (x)	SBP (y)
42	130
46	115
42	148
71	100
80	156
74	162
70	151
80	156
85	162
72	158
64	155
81	160
41	125
61	150
75	165

$$t = \frac{99.9585}{19.2552}$$

$$= 5.191$$

$$95\%\ \text{C.I.} = 99.9585 \pm (2.1604)(19.2552)$$

$$= (58.360, 141.557)$$

EXAMPLE #3: Toluca Company Data (Description on page 19 of Text)

LotSize	WorkHours
80	399
30	121
50	221
90	376
70	361
60	224
120	546
80	352
100	353
50	157
40	160
70	252
90	389
20	113
110	435
100	420
30	212
50	268
90	377
110	421
30	273
90	468
40	244
80	342
70	323

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	62.3658586	26.17743389	2.382428	0.025851	8.21371106	116.518006
X Variable 1	3.57020202	0.346972157	10.28959	4.45E-10	2.85243543	4.28796861

$$t = \frac{62.3659}{26.1774}$$

$$= 2.382$$

$$95\%\ \text{C.I.} = 62.3659 \pm (2.0687)(26.1774)$$

$$= (8.218, 116.518)$$

DEPARTURE FROM NORMALITY

Normal Error Regression Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

If the probability distributions of Y are not exactly normal but do not depart seriously, the sampling distributions of b_0 and b_1 would still be approximately normal with very little effects on the level of significance of the t-test for independence and the coverage of the confidence intervals. Even if the probability distributions of Y are far from normal, the effects are still minimal provided that the samples sizes are sufficiently large; i.e. **the sampling distributions of b_0 and b_1 are asymptotically normal.**

Sometimes it is known, *a priori*, that the true intercept is zero; the regression function is linear but the line goes through the origin (0,0):

Regression through the origin :

$$Y = \beta_1 x + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

The Mean Response :

$$E(Y \mid X = x) = \beta_1 x$$

RESULTS

Let b_1 be the estimate of slop β_1, "the sum of squared errors" becomes $Q=\Sigma (y - b_1x)^2$. The new "Least Squares Estimate" is:

$$b_1 = \frac{\sum xy}{\sum x^2}$$

$$\hat{Y} = b_1 x$$

All inferences are still drawn through the use of the "t" distribution but with (n-1) degrees of freedom; for example:

$$MSE = \frac{\sum e_i^2}{n-1}$$

$$s^2(b_1) = \frac{MSE}{\sum x^2}$$

Besides “Least Squares”, parameters can be estimated using the method of “Maximum Likelihood”; results are called “MLE” – maximum likelihood estimators/estimates.

Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

Density Function for Y:

$$\mathrm{f(y)} = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(y - \beta_0 - \beta_1 x)^2 \right\}$$

Density Function for Y:

$$f(y) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y - \beta_0 - \beta_1 x)^2\right\}$$

Likelihood Function:

$$L = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

SLOPE & INTERCEPT

- The maximum likelihood estimates of the Intercept and the Slope are identical to the Least Squares estimates.
- **Their variances, obtained from the inverse of the Fisher's Information Matrix, are also identical.**
- As the least squares estimates, Estimated Intercept and Slope have the properties of least squares estimates: (1) they are unbiased, and (2) they have minimum variance in the class of all linear unbiased estimators).
- In addition, as MLEs for the normal error regression model: (3) they are consistent & (4) they are sufficient.

Readings & Exercises

- Readings: A thorough reading of the text's sections 2.1-2.3 (pp. 40-51) is highly recommended.
- Exercises: The following exercises are good for practice, all from chapter 2 of text: 2.1-2.6.
- Due as Homework: 2.4 and 2.6.

Due As Homework

#6.1 Refer to dataset “Infants”, with X = Gestational Weeks and Y = Birth Weight:

a) Obtain the 95% confidence interval for the slope and interpret your result. Does your confidence interval include zero? What would be your conclusion about the possible linear relationship between X and Y.

b) Using the t-statistic test to see whether or not a linear association exist between X and Y.

c) Does your conclusion in part (b) agree with your conclusion in part (a)? Which result, in (a) or in (b), could tell you more about the strength of the relationship between X and Y?

#6.2 Answer the 3 questions of Exercise 6.1 using dataset “Vital Capacity” with X = Age and Y = (100)(Vital Capacity).