

PubH 7405: REGRESSION ANALYSIS



SLR: INFERENCES, Part II

We cover the topic of inference in two sessions; the first session focused on inferences concerning the **slope and the intercept**; this is a continuation on estimating the **mean response** – and more. Applications concerning the slope and the intercept are based on the following four (4) **theorems**

SAMPLING DISTRIBUTION OF SLOPE

Theorem 1A:

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

b_1 is Normal with Mean and Variance :

$$E(b_1) = \beta_1$$

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

IMPLICATION

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

distributed as N(0,1)

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

Theorem 1B :

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

CONFIDENCE INTERVALS

Theorem 1B :

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

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

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

$t(1 - \alpha/2; n - 2)$ is the $(1 - \alpha/2)100$ percentile of

the "t" distribution with $(n - 2)$ degrees of freedom

SAMPLING DISTRIBUTION OF INTERCEPT

Theorem 2A:

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

b_0 is Normal with Mean and Variance :

$$E(b_0) = \beta_0$$

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

IMPLICATION

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

distributed as N(0,1)

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

Theorem 2B :

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

CONFIDENCE INTERVALS

Theorem 2B :

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

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

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

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

The Mean Response :

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

A common objective in regression analysis is to **estimate the mean response**. For example:

(1) we are interested to know the **average blood pressure for women at certain age and how estimate it** using the relationship between SBP and Age, and

(2) in a study of the relationship between level of pay (salary, X) and worker productivity (Y), the **mean productivity** at high, medium, and low levels of pay may be of particular interest for any company.

POINT ESTIMATE

The Mean Response :

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

Let $X = x_h$ denote the level of X for which we wish to estimate the mean response, i.e. $E(Y|X=x_h)$; this x_h may be a value which occurred in the sample, or it may be some other value of the predictor variable within the scope of the model. The point estimate of the response is:

Point Estimate :

$$\begin{aligned} E(Y | X = x_h) &= \hat{Y}_h \\ &= b_0 + b_1 x_h \end{aligned}$$

SAMPLING DISTRIBUTION

Theorem #3A :

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 Mean

Response \hat{Y}_h is Normal with Mean and Variance :

$$E(\hat{Y}_h) = E(Y | X = x_h) = \beta_0 + \beta_1 x_h$$

$$\sigma^2(\hat{Y}_h) = \sigma^2 \left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x - \bar{x})^2} \right]$$

$$\mathbf{b}_0 = \sum \left\{ \frac{1}{n} - \bar{x} k_i \right\} y_i$$

$$\mathbf{b}_1 = \sum k_i y_i$$

$$\hat{Y}_h = \mathbf{b}_0 + \mathbf{b}_1 x_h$$

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

The sampling distribution of \hat{Y}_h is “normal” because this estimated mean response, like the intercept and the slope, **\hat{Y}_h is a linear combination of the observations y_i** and the distribution of each observation is normal under the “normal error regression model”:

The estimated mean response is unbiased because the estimated intercept and estimated slope are both unbiased:

$$\hat{Y}_h = b_0 + b_1 x_h$$

$$\begin{aligned} E(\hat{Y}_h) &= E(b_0) + x_h E(b_1) \\ &= \beta_0 + \beta_1 x_h \\ &= E(Y | X = x_h) \end{aligned}$$

$$\hat{Y}_h = \sum \left\{ \frac{1}{n} + (x_h - \bar{x})k_i \right\} y_i$$

$$\text{Var}(\hat{Y}_h) = \sum \left\{ \frac{1}{n} + (x_h - \bar{x})k_i \right\}^2 \sigma^2$$

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

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

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

$$\text{Var}(\hat{Y}_h) = \sigma^2 \left\{ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}$$

$$s^2(\hat{Y}_h) = \text{MSE} \left\{ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}$$

Taking square root to get Standard Error

$$s^2(\hat{Y}_h) = MSE \left\{ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}$$

$$SE(\hat{Y}_h) = \sqrt{MSE \left\{ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}}$$

Implication:

Our estimates are less precise toward the ends

MORE ON SAMPLING DISTRIBUTION

$$\frac{\hat{Y}_h - E(Y_h)}{\hat{s}(Y_h)} = \frac{\hat{Y}_h - E(Y_h)}{\hat{\sigma}(Y_h)} \div \frac{\hat{s}(Y_h)}{\hat{\sigma}(Y_h)}$$

distributed as $N(0,1)$

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

Theorem #3B :

$\frac{\hat{Y}_h - E(Y_h)}{\hat{s}(Y_h)}$ is distributed as "t" with $(n-2)$ degrees of freedom

CONFIDENCE INTERVALS

Theorem #3B :

$\frac{\hat{Y}_h - E(Y_h)}{\hat{s}(Y_h)}$ is distributed as "t" with $(n - 2)$ degrees of freedom

$(1 - \alpha)100\%$ Confidence Interval for \hat{Y}_h is :

$$\hat{Y}_h \pm t(1 - \alpha / 2; n - 2) \hat{s}(Y_h)$$

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

EXAMPLE #1: Birth weight data:

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

Intercept = 256.972

Slope = -1.737

MSE = 75.982

Mean of X = 100.58

SS of X = **2,156.913**

For **children** with birth weight of $x_h = 95$ ounces, the point estimate and 95%

Confidence Interval for the **Mean growth between 70-100 days as % of BW** is:

$$\hat{Y}_h = 256.972 + (-1.737)(95) = 91.757\%$$

$$s^2(\hat{Y}_h) = (75.982) \left[\frac{1}{12} + \frac{(95 - 100.58)^2}{2,156.913} \right] = 7.429$$

$$91.76 \pm (2.228)\sqrt{7.43} = (85.69\%, 97.83\%)$$

EXAMPLE #2: Age and SBP

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

Intercept = 99.958

Slope = .705

MSE = 278.554

Mean of X = 65.6

SS of X = 3403.6

For $x_h = 60$ years old women, the point estimate and 95% Confidence Interval for the Mean SBP is:

$$\hat{Y}_h = 99.958 + (.705)(60) = 142.26$$

$$s^2(\hat{Y}_h) = (278.554) \left[\frac{1}{15} + \frac{(60 - 65.6)^2}{3403.6} \right] = 21.137$$

$$142.3 \pm (2.160)\sqrt{21.137} = (132.4, 152.2)$$

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

Intercept = 62.366

Slope = 3.570

MSE = 2,384

Mean of X = 70.0

SS of X = 19,800

For the lots' size of $x_h = 65$ units, the point estimate and 90% Confidence Interval for the Mean Work Hours is:

$$\hat{Y}_h = 62.37 + (3.57)(65) = 294.4$$

$$s^2(\hat{Y}_h) = (2,384) \left[\frac{1}{25} + \frac{(65 - 70.0)^2}{19,800} \right] = 98.47$$

$$294.4 \pm (1.714)\sqrt{98.47} = (277.4, 311.4)$$

In regression analysis, besides estimating the mean response, **sometimes** one may want **to estimate a new individual response**. For example:

- (1) In addition to estimating the average blood pressure for women at certain age using the relationship between SBP and Age, **we may be interested in estimating the SBP of a particular woman/patient at that age;** and
- (2) In a study of the relationship between pay (salary, X) and worker productivity (Y), the interest may **focus on the productivity of certain particular worker.**

POINT ESTIMATE

Let $X = x_h$ denote the level of X under investigation, at which the **mean response** is $E(Y|X=x_h)$. Let $Y_{h(\text{new})}$ be the value of the new individual response of interest. This new observation of Y to be predicted is often viewed as **the result of a new trial independent of the trials on which the regression line is formed**. The **point estimate** is still the same as that of the mean response:

$$E(Y | X = x_h) = \beta_0 + \beta_1 x_h$$

$$\hat{Y}_h = b_0 + b_1 x_h$$

$$= \hat{Y}_{h(\text{new})}$$

Same as the mean



VARIANCE

The point estimates of the mean response and of an individual response are the same but the variances are different. In estimating an individual response, there are two layers of variation: (a) variation in the “position of the distribution” (that is of the mean response), and (b) the variation within that distribution (that is **from the individual response to the mean response**)

$$\begin{aligned}
\text{Var}(\hat{Y}_{h(\text{new})}) &= \text{Var}(Y_{h(\text{new})}) + \text{Var}(\hat{Y}_h) \\
&= \sigma^2 + \sigma^2 \left\{ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\} \\
&= \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}
\end{aligned}$$

Theorem #4A : Under the "Normal Error Regression Model,
the sampling distribution of $\hat{Y}_{h(\text{new})}$ is normal.

$$\text{Var}(\hat{Y}_{h(new)}) = \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}$$

$$s^2(\hat{Y}_{h(new)}) = \text{MSE} \left\{ 1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}$$

Taking square root to get Standard Error

MORE ON SAMPLING DISTRIBUTION

Inferences on a new individual response is based on the following results:

Theorem #4B :

$\frac{\hat{Y}_{h(new)} - \hat{Y}_h}{s(\hat{Y}_{h(new)})}$ is distributed as "t" with $(n - 2)$ degrees of freedom

$$s^2(\hat{Y}_{h(new)}) = MSE \left\{ 1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}$$

$$SE(\hat{Y}_{h(new)}) = \sqrt{MSE \left\{ 1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}}$$

Again:

Our estimates are less precise toward the ends

Normal Error Regression Model :

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

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

$\{e_i\}$ is a sample with mean zero :

$$\hat{\sigma}^2 = MSE$$

Theorem #5 :

$\frac{SSE}{\sigma^2}$ is distributed as $\chi_{df=n-2}^2$

$$E(MSE) = \sigma^2$$

THE TEST FOR INDEPENDENCE

The Mean Response :

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

$$H_0 : \beta_1 = 0$$

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

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

which is identical to the test using "r":

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

The method we use most often is this "Test for Independence" which we are now approaching by a different way:

ANOVA

COMPONENTS OF VARIATION

- The variation in Y is conventionally measured in terms of the deviations $(Y_i - \bar{Y})$'s; the total variation, denoted by SST , is the sum of squared deviations: $SST = \sum(Y_i - \bar{Y})^2$. For example, $SST=0$ when all observations are the same; SST is the numerator of the sample variance of Y , the greater SST the greater the variation among Y -values.
- In the regression analysis, the variation in Y is decomposed into two components:
$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

DECOMPOSITION OF SST

- In the decomposition: $(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$
- The first term (RHS) reflects the variation around the regression line; the part that cannot be explained by the regression itself with the **sum of squared errors** $SSE = \sum(Y_i - \hat{Y}_i)^2$.
- The difference between the above two sums of squares, $SSR = SST - SSE = \sum(\hat{Y}_i - \bar{Y})^2$, is called the **regression sum of squares**; SSR may be considered as a **measure of the variation in Y associated with or explained by the regression line**.

Regression helps to “**improve**” the estimate of Y from \bar{Y} (without any information) to \hat{Y} (with information provided by knowing X)

$$\begin{aligned}
SST &= \sum (Y - \bar{Y})^2 \\
&= \sum (Y - \hat{Y})^2 = \sum [(Y - \hat{Y}) + (\hat{Y} - \bar{Y})]^2 \\
&= \sum (Y - \hat{Y})^2 + \sum (\hat{Y} - \bar{Y})^2 + 2 \sum (Y - \hat{Y})(\hat{Y} - \bar{Y}) \\
&= SSE + SSR + 2 \sum e_i \hat{Y}_i + 2 \bar{Y} \sum e_i
\end{aligned}$$

$$SST = SSE + SSR$$

0

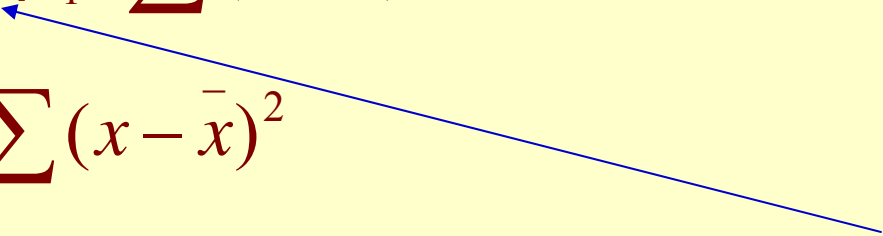


ANALYSIS OF VARIANCE

- SST measures the “total variation” in the sample (of values of the dependent variable) with $(n-1)$ degrees of freedom, n is the sample size. It is decomposed into: $SST=SSE+SSR$
- (1) SSE measures the variation cannot be explained by the regression with $(n-2)$ degrees of freedom, and
- (2) **SSR measures the variation in Y associated with or explained by the regression line with 1 degree of freedom (representing the slope).**

$$\begin{aligned} SSR &= \sum (b_0 + b_1 x - \bar{y})^2 \\ &= \sum [(\bar{y} - b_1 \bar{x}) + b_1 x - \bar{y}]^2 \\ &= b_1^2 \sum (x - \bar{x})^2 \end{aligned}$$

$$E(MSR) = E(SSR)$$

$$\begin{aligned} &= [\sigma^2(b_1) + \beta_1^2] \sum (x - \bar{x})^2 \\ &= \sigma^2 + \beta_1^2 \sum (x - \bar{x})^2 \end{aligned}$$


$$\text{Var}(X) = E(X^2) - \{E(X)\}^2$$

$$E(X^2) = \text{Var}(X) + \{E(X)\}^2$$

“ANOVA” TABLE

- The breakdowns of the total sum of squares and its associated degree of freedom are displayed in the form of an “analysis of variance table” (ANOVA table) for regression analysis as follows:

Source of Variation	SS	df	MS	F Statistic	p-value
Regression	SSR	1	MSR	MSR/MSE	
Error	SSE	n-2	MSE		
Total	SST	n-1			

- Recall**: MSE, the “error mean square”, serves as an estimate of the constant variance σ^2 as stipulated by the regression model.

$$E(MSE) = \sigma^2$$

$$E(MSR) = \sigma^2 + \beta_1^2 \sum (x - \bar{x})^2$$

Under the Null Hypothesis $H_0: \beta_1 = 0$,
 $E(MSE) = E(MSR)$ so that $F = MSR/MSE$ is
expected to be near 1.0

Theorem 6: F is distributed, under H_0 , as
 $F(1, n-2)$ following a theorem by Cochran.

THE F-TEST

The test statistic F for the above analysis of variance approach compares MSR and MSE, a value near 1 supports the null hypothesis of independence. In fact, we have: $F = t^2$, where t is the test statistic for testing whether or not $\beta_1=0$; the F-test is equivalent to the two-sided t-test when referred to the F-table in **Appendix B (Table B.4)** with $(1, n-2)$ degrees of freedom.

THE TEST FOR INDEPENDENCE

The Null Hypothesis :

$$H_0 : \beta_1 = 0$$

Two identical choices :

(1) "t" test at $(n - 2)$ degrees of freedom :

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

which is identical to the test using "r":

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

(2) "F" test at $(1, n - 2)$ degrees of freedom :

$$F = \frac{MSR}{MSE}$$

COEFFICIENT OF DETERMINATION

- We can express the coefficient of determination (the square of the coefficient of correlation r) as:

$$r^2 = \frac{SSR}{SST}$$

- That is the portion of total variation attributable to regression; Regression helps to “**improve**” the estimate of Y from \bar{Y} (without any information) to \hat{Y} (with information provided by knowing X) – reducing the total variation by $(100)(r^2)\%$

EXAMPLE #1: Birth Weight Data

x (oz)	y (%)	SUMMARY OUTPUT					
112	63						
111	66						
107	72	<i>Regression Statistics</i>					
119	52	R Square	0.89546				
92	75	Observations	12				
80	118						
81	120	ANOVA					
84	114		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	
118	42					<i>Significance F</i>	
106	72	Regression	1	6508	6508	85.66	3.21622E-06
103	90	Residual	10	759.8	75.98		
94	91	Total	11	7268			

EXAMPLE #2: AGE & SBP

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

SUMMARY OUTPUT					
<i>Regression Statistics</i>					
R Square	0.3183				
Observations	15				
<i>ANOVA</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	1691	1691	6.071	0.028453563
Residual	13	3621	278.6		
Total	14	5312			

EXAMPLE #3: Toluca Company Data

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

SUMMARY OUTPUT					
<i>Regression Statistics</i>					
R Square	0.3183				
Observations	15				
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	1691	1691	6.071	0.028453563
Residual	13	3621	278.6		
Total	14	5312			

Normal Error Regression Model :

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

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

The normal regression model assumes that the X values are known constants. We do not impose any kind of distribution for the x -values

In many cases, this is not true; for example, if we study the relationship between “height of a person” and weight of a person”, a sample of persons are taken but both measurements are random. Rather than a regression model, one should consider a “correlation model”; the most widely used is the “Bivariate Normal Distribution” with density:

CORRELATION MODEL

“Correlation Data” are often cross-sectional or observational. Instead of a regression model, one should consider a “correlation model”; the most widely used is the “Bivariate Normal Distribution” with density:

$$f(X, Y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{X-\mu_x}{\sigma_x}\right)\left(\frac{Y-\mu_y}{\sigma_y}\right) + \left(\frac{Y-\mu_y}{\sigma_y}\right)^2\right]\right\}$$

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

$$\begin{aligned}\sigma_{xy} &= \text{Cov}(X, Y) \\ &= E[(X - \mu_x)(Y - \mu_y)]\end{aligned}$$

σ_{xy} is the Covariance and ρ is the Coefficient of Correlation between the two random variables X and Y; ρ is estimated by the (sample) Coefficient of Correlation r .

The Coefficient of Correlation ρ between the two random variables X and Y is estimated by the (sample) Coefficient of Correlation r but the sampling distribution of r is far from being normal. Confidence intervals of ρ is by first making the “**Fisher’s z transformation**”; the distribution of z is normal if the sample size is not too small

CONDITIONAL DISTRIBUTION

$$f(X,Y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{X-\mu_x}{\sigma_x}\right)\left(\frac{Y-\mu_y}{\sigma_y}\right) + \left(\frac{Y-\mu_y}{\sigma_y}\right)^2\right]\right\}$$

Theorem :

The conditional distribution of Y for any given $X=x$ is normal with mean $\beta_0 + \beta_1 x$ and standard deviation $\sigma_{y|x}$:

$$\beta_0 = \mu_y - \mu_x \rho \frac{\sigma_y}{\sigma_x}$$

$$\beta_1 = \rho \frac{\sigma_y}{\sigma_x}$$

$$\sigma_{y|x}^2 = (1-\rho^2)\sigma_y^2$$

Again, since $\text{Var}(Y|X) = (1 - \rho^2)\text{Var}(Y)$, ρ is both a measure of linear association and a measure of “variance reduction” (in Y associated with knowledge of X) – that’s why we called r^2 , an estimate of ρ^2 , the “coefficient of determination”.

Readings & Exercises

- Readings: A thorough reading of the text's sections 2.4-2.5 (pp. 52-61), 2.7 (pp. 63-71), and 2.11 (pp. 78-82) is highly recommended.
- Exercises: The following exercises are good for practice, all from chapter 2 of text: 2.13, 2.23, 2.24, 2.28, and 2.29.

Due As Homework

8.1 Refer to dataset “Cigarettes”, $Y = \log(\text{NNAL})$ & $X = \text{CPD}$:

a) Obtain the 95% confidence interval for the mean $\log(\text{NNAL})$ level for subjects who consumed $X = 30$ cigarettes per day and give your interpretation.

b) Obtain the 95% confidence interval for $\log(\text{NNAL})$ level of a subject who consumed 30 cigarettes per day; why is the result is different from (a)?

c) Plot the residual $Y - \hat{Y}$ against X ; What would be your conclusion about their possible linear relationship? What would be the average residual?

d) Set up the ANOVA table and test whether or not a linear association exist between $\log(\text{NNAL})$ and CPD. Does your conclusion agree with your conclusion in problem 5.1 part (b)

8.2 Answer the 4 questions of Exercise 8.1 using dataset “Infants” with $X = \text{Gestational Weeks}$ and $Y = \text{Birth Weight}$.

Only #8.2 is required