Chapter 2

Inference in Regression and Correlation Analysis

Instructor: Li, Caixia

Outline

- Inferences Concerning β_1, β_0, and EY in the Normal Error Regression Model
- Prediction Interval of New Observation
- Confidence Band for Regression Line
- ANVOA (Analysis of Variance) Approach to Regression Analysis
- General linear test approach
- Normal Correlation Models and Inferences

2.1 Inferences Concerning β_1

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i=1,2,\dots n$$

with ε_i are i.i.d and $\varepsilon_i \sim N(0, \sigma^2)$.

$$b_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}, \qquad b_0 = \bar{Y} - b_1\bar{X}$$

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \frac{\sigma^2}{SS_{XX}}\begin{pmatrix} \frac{1}{n}\sum X_i^2 & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix}\right)$$

$$\frac{SSE}{\sigma^2} = \frac{(n-2)MSE}{\sigma^2} \sim \chi^2(n-2), \quad (b_0, b_1) \text{ and } SSE \text{ are independent.}$$

Recaps

- **Structure of t distribution**

$Z \sim N(0, 1)$, $V \sim \chi^2(r)$, Z and V are independent, then

$$\frac{Z}{\sqrt{V/r}} \sim t(r).$$

- **Structure of F distribution**

$U \sim \chi^2(r_1)$, $V \sim \chi^2(r_2)$, U and V are independent, then

$$\frac{U/r_1}{V/r_2} \sim F(r_1, r_2).$$

Sampling distribution of b_1 – Normal Error Model

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{SS_{XX}}\right) \quad \Rightarrow \quad \frac{b_1-\beta_1}{\sqrt{\sigma^2 / SS_{XX}}} = \frac{b_1-\beta_1}{\sigma\{b_1\}} \sim N(0,1)$$

$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2} \quad \text{also: } b_1 \text{ and } MSE \text{ independent}$$

$$\Rightarrow \frac{\left[\dfrac{b_1-\beta_1}{\sqrt{\sigma^2 / SS_{XX}}}\right]}{\sqrt{\dfrac{(n-2)MSE}{\sigma^2} \Big/ (n-2)}} = \frac{b_1-\beta_1}{\sqrt{MSE / SS_{XX}}} = \frac{b_1-\beta_1}{s\{b_1\}} \sim t(n-2)$$

$$where \;\; \sigma\{b_1\} = \sqrt{\sigma^2 / SS_{XX}}, \quad s\{b_1\} = \sqrt{MSE / SS_{XX}}$$

$$\Pr\left\{ t(\alpha/2; n-2) < \frac{b_1-\beta_1}{s\{b_1\}} < t\left(1-(\alpha/2); n-2\right) \right\} = 1-\alpha$$

$$t(\alpha/2; n-2) = -t\left(1-(\alpha/2); n-2\right)$$

Confidence interval of β_1

- $(1-\alpha)$*100% Confidence interval of β_1

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

- Toluca Company example, we obtain:

$SSE = 54825$, SS_{XX}=19800, $MSE = \dfrac{54{,}825}{23} = 2{,}384$

$$s^2\{b_1\} = \frac{MSE}{\sum(X_i - \bar{X})^2} = \frac{2{,}384}{19{,}800} = .12040$$

$$s\{b_1\} = .3470$$

- Giving α =0.05, $t\,(0.975;\,23) = 2.069$.
- 95% Confidence interval of β_1

$$3.5702 - 2.069(.3470) \le \beta_1 \le 3.5702 + 2.069(.3470)$$

$$2.85 \le \beta_1 \le 4.29$$

Hypothesis Test for β_1

- Tests concerning β_1 (the slope) are often of interest, particularly

$$H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0$$

- The null hypothesis model $Y_i = \beta_0 + \varepsilon_i$,
  - implies that there is no linear relationship between Y and X.
  - Note the means of all the Yi 's are equal at all levels of Xi .
- Test statistic

$$t^* = \frac{b_1 - 0}{s\{b_1\}} \overset{H_0}{\sim} t(n-2)$$

- Decision rule

$$\begin{array}{lll} \text{if } |t^*| & \leq & t(1-\alpha/2; n-2), \text{ accept } H_0 \\ \text{if } |t^*| & > & t(1-\alpha/2; n-2), \text{ reject } H_0 \end{array}$$

Hypothesis Test for β_1

- For Toluca Company example, we obtain: $b_1 = 3.5702$, and $s\{b1\} = .3470$
- Giving α =0.05, $t\,(0.975;\,23) = 2.069$.

$$|t^*| = |3.5702/.3470| = 10.29 > 2.069.$$

- P value

$$\Pr\{|t(23)| \geq 10.29\} = 2\Pr\{t(23) \geq 10.29\} = 4.45 \times 10^{-10}$$

- Reject H_0, we conclude that $\beta_1 \neq 0$ or that there is a linear association between work hours and lot size.

Hypothesis Test for β_1

2-sided test: $H_0 : \beta_1 = \beta_{10} \quad H_A : \beta_1 \neq \beta_{10}$ (Almost always $\beta_{10} = 0$)

Test Statistic: $t^* = \dfrac{b_1 - \beta_{10}}{s\{b_1\}}$ Note: if $\beta_{10} = 0 \Rightarrow t^* = \dfrac{b_1}{s\{b_1\}}$

Decision Rule: $|t^*| \geq t(1-\alpha/2; n-2) \Rightarrow$ Reject H_0 otherwise Fail to Reject

P-value: $2\Pr\{t(n-2) \geq |t^*|\}$

Upper-tail test: $H_0 : \beta_1 \leq \beta_{10} \quad H_A : \beta_1 > \beta_{10}$

Decision Rule: $t^* \geq t(1-\alpha; n-2) \Rightarrow$ Reject H_0 otherwise Fail to Reject

P-value: $\Pr\{t(n-2) \geq t^*\}$

Lower-tail test: $H_0 : \beta_1 \geq \beta_{10} \quad H_A : \beta_1 < \beta_{10}$

Decision Rule: $t^* \leq -t(1-\alpha; n-2) \Rightarrow$ Reject H_0 otherwise Fail to Reject

P-value: $\Pr\{t(n-2) \leq t^*\}$

2.2 Inferences Concerning β_0

$$b_0 = \overline{Y} - b_1 \overline{X} \sim N\left(\beta_0, \sigma^2 \frac{\sum X_i^2}{nSS_{XX}}\right) = N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{SS_{XX}}\right)\right)$$

$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2} \quad \text{also: } b_0 \text{ and } MSE \text{ independent}$$

$$\Rightarrow \frac{\dfrac{b_0 - \beta_0}{\sqrt{\sigma^2\left(\dfrac{1}{n} + \dfrac{\overline{X}^2}{SS_{XX}}\right)}}}{\sqrt{\dfrac{(n-2)MSE}{\sigma^2} \Big/ (n-2)}} = \frac{b_0 - \beta_0}{\sqrt{MSE\left(\dfrac{1}{n} + \dfrac{\overline{X}^2}{SS_{XX}}\right)}} = \frac{b_0 - \beta_0}{s\{b_0\}} \sim t(n-2)$$

where estimated standard error: $s\{b_0\} = \sqrt{MSE\left(\dfrac{1}{n} + \dfrac{\overline{X}^2}{SS_{XX}}\right)}$

Inferences Concerning $\boldsymbol{\beta_0}$

$$\Pr\left\{\left|\frac{b_0-\beta_0}{s\{b_0\}}\right|<t\left(1-\alpha/2;n-2\right)\right\}=1-\alpha$$

- (1−α)*100% Confidence interval of β_0

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

- Hypothesis Test for β_0

$$H_0:\beta_0=\beta_{00} \qquad H_A:\beta_0\neq\beta_{00}$$

$$t^*=\frac{b_0-\beta_{00}}{s\{b_0\}} \overset{H_0}{\sim} t(n-2),$$

$$\text{Reject } H_0 \text{ if } \left|t^*\right|\geq t\left(1-\left(\alpha/2\right);n-2\right)$$

Inferences Concerning β_0

- For Toluca Company example, we obtain:

$$s^2\{b_0\} = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right] = 2{,}384\left[\frac{1}{25} + \frac{(70.00)^2}{19{,}800}\right] = 685.34$$

$$s\{b_0\} = 26.18$$

- Giving $\alpha = 0.10$, $t\,(0.95; 23) = 1.714$.
- 90% Confidence interval of β_0

$$62.37 - 1.714(26.18) \leq \beta_0 \leq 62.37 + 1.714(26.18)$$

$$17.5 \leq \beta_0 \leq 107.2$$

- Giving $\alpha = 0.05$, $t\,(0.975; 23) = 2.069$.

$$t^* = \frac{b_0}{s\{b_0\}} = \frac{62.37}{26.18} = 2.382 > 2.069 \quad \text{p=Pr(|t(23)|>2.069)=0.0259}$$

  - Reject H_0: $\beta_0 = 0$, we conclude that $\beta_0 \neq 0$.

2.3 Some Considerations on Making Inferences

- Effects of departures from normality of the Y_i
  - The estimators of β_0 and β_1 have the property of asymptotic normality increases (under general conditions)
- Spacing of the X levels
  - The variances of b_0 and b_1 (for a given n and σ^2) depend on the spacing of X
  - The larger is SS_{XX} and the smaller is the variance
- Power of Tests
  - Power=P{Reject H0} $= P\{|t^*| > t(1-\alpha/2; n-2)|\delta\}$
  - Noncentral t distribution

2.4 Interval Estimation of $E\{Y_h\}$

- Interested in estimating the mean response for particular X_h

$$E\{Y_h\} = \beta_0 + \beta_1 X_h$$

- The unbiased point estimator of $E\{Y_h\}$

$$\hat{Y}_h = b_0 + b_1 X_h = \bar{Y} + b_1\left(X_h - \bar{X}\right)$$

$$E(\hat{Y}_h) = \beta_0 + \beta_1 X_h = E(Y_h),$$

$$\operatorname{var}(\hat{Y}_h) = b_0 + b_1 X_h = \operatorname{var}(\bar{Y}) + \operatorname{var}(b_1)\left(X_h - \bar{X}\right)^2 = \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{SS_{XX}}\right)$$

or $\quad \operatorname{var}(\hat{Y}_h) = \operatorname{var}(b_0) + X_h^2 \operatorname{var}(b_1) + 2X_h \operatorname{cov}(b_0, b_1)$

$$= \left(\frac{1}{n} + \frac{\bar{X}^2}{SS_{XX}}\right)\sigma^2 + \frac{X_h^2}{SS_{XX}}\sigma^2 - \frac{2X_h\bar{X}}{SS_{XX}}\sigma^2 = \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{SS_{XX}}\right]\sigma^2$$

since $\quad \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \frac{\sigma^2}{SS_{XX}}\begin{pmatrix} \frac{1}{n}\sum X_i^2 & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix}\right)$

Interval Estimation of $E\{Y_h\}$

$$\hat{Y}_h = b_0 + b_1 X_h \sim N\left(E(Y_h) = \beta_0 + \beta_1 X_h, \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}} \right] \sigma^2 \right).$$

Remark: The variance is smaller near the mean of X.

$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2} \quad \text{also: } (b_0, b_1, \hat{Y}_h) \text{ and } MSE \text{ independent}$$

$$\Rightarrow \frac{\dfrac{\hat{Y}_h - E(Y_h)}{\sqrt{\sigma^2\left(\dfrac{1}{n} + \dfrac{(X_h - \overline{X})^2}{SS_{XX}}\right)}}}{\sqrt{\dfrac{(n-2)MSE}{\sigma^2} \Big/ (n-2)}} = \frac{\hat{Y}_h - E(Y_h)}{\sqrt{MSE\left(\dfrac{1}{n} + \dfrac{(X_h - \overline{X})^2}{SS_{XX}}\right)}} = \frac{\hat{Y}_h - E(Y_h)}{s\{\hat{Y}_h\}} \sim t(n-2)$$

Interval Estimation of $E\{Y_h\}$

$$\Pr\left\{\left|\frac{\hat{Y}_h - E(Y_h)}{s\{\hat{Y}_h\}}\right| < t\big(1-(\alpha/2);n-2\big)\right\} = 1-\alpha$$

$(1-\alpha)*100\%$ Confidence interval of $E\{Y_h\} = \beta_0 + \beta_1 X_h$

$$\hat{Y}_h \pm t\big(1-\alpha/2;n-2\big)\, s\{\hat{Y}_h\} \qquad s\{\hat{Y}_h\} = \sqrt{MSE\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{SS_{XX}}\right)}$$

Suppose the Toluca Company wishes to estimate $E\{Y_h\}$ at $X_h = 100$ units with a 90 percent confidence interval, $t(.95;23) = 1.714$

$$\hat{Y}_h = 62.37 + 3.5702(100) = 419.4$$

$$s^2\{\hat{Y}_h\} = 2{,}384\left[\frac{1}{25} + \frac{(100 - 70.00)^2}{19{,}800}\right] = 203.72 \qquad s\{\hat{Y}_h\} = 14.27$$

$$419.4 - 1.714(14.27) \le E\{Y_h\} \le 419.4 + 1.714(14.27)$$

$$394.9 \le E\{Y_h\} \le 443.9$$

2.5 Prediction of New Observation

- Interested in predicting new (future) observation when $X = X_h$,

  $Y_{h(\text{new})} = \beta_0 + \beta_1 X_h + \varepsilon_{h(\text{new})}$

- $Y_{h(\text{new})}$ is independent $\{Y_1, Y_2, \ldots, Y_n\}$, and

$$Y_{h(new)} \sim N\left(\beta_0 + \beta_1 X_h, \sigma^2\right)$$

- Prediction of $Y_{h(\text{new})}$

$$\hat{Y}_h = b_0 + b_1 X_h \sim N\left(\beta_0 + \beta_1 X_h, \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right]\sigma^2\right).$$

- Prediction error

$$Y_{h(\text{new})} - \hat{Y}_h \sim N\left(0, \left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right]\sigma^2\right).$$

Prediction interval of New Observation

$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2} \quad \text{also: } (Y_{h(\text{new})}, \hat{Y}_h) \text{ and } MSE \text{ independent}$$

$$\Rightarrow \frac{Y_{h(\text{new})} - \hat{Y}_h}{\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right)}} = \frac{Y_{h(\text{new})} - \hat{Y}_h}{s\left\{Y_{h(\text{new})} - \hat{Y}_h\right\}} = \frac{Y_{h(\text{new})} - \hat{Y}_h}{s\{pred\}} \sim t(n-2)$$

$$\Pr\left\{\left|\frac{Y_{h(\text{new})} - \hat{Y}}{s\{pred\}}\right| < t(1-\alpha/2; n-2)\right\} = 1-\alpha$$

$(1-\alpha)$*100% prediction interval of $Y_{h(\text{new})} = \beta_0 + \beta_1 X_h + \varepsilon_{h(\text{new})}$

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

Prediction interval of New Observation

- Suppose the Toluca Company wishes to predict the required work hours when $X_h = 100$ units
- A 90 percent prediction interval is desired. $t(.95;23) = 1.714$

$\hat{Y}_h = 419.4 \qquad s^2\{\hat{Y}_h\} = 203.72 \qquad MSE = 2{,}384$

$$s^2\left\{Y_{h(\text{new})} - \hat{Y}_h\right\} = s^2\left\{Y_{h(\text{new})}\right\} + s^2\left\{\hat{Y}_h\right\} = 2384 + 203.72 = 2587.72$$

$$s\left\{Y_{h(\text{new})} - \hat{Y}_h\right\} = 50.87$$

$$419.4 - 1.714(50.87) \leq Y_{h(\text{new})} \leq 419.4 + 1.714(50.87)$$

$$332.2 \leq Y_{h(\text{new})} \leq 506.6$$

Comparison

$$419.4 - 1.714(14.27) \leq E\{Y_h\} \leq 419.4 + 1.714(14.27)$$

$$394.9 \leq E\{Y_h\} \leq 443.9$$

Comparisons

- $(1-\alpha)*100\%$ Confidence interval of $E\{Y_h\} = \beta_0+\beta_1 X_h$

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

$$s\left\{\hat{Y}_h\right\}=\sqrt{MSE\left(\frac{1}{n}+\frac{(X_h-\overline{X})^2}{SS_{XX}}\right)}$$

- $(1-\alpha)*100\%$ prediction interval of $Y_{h(\text{new})}$

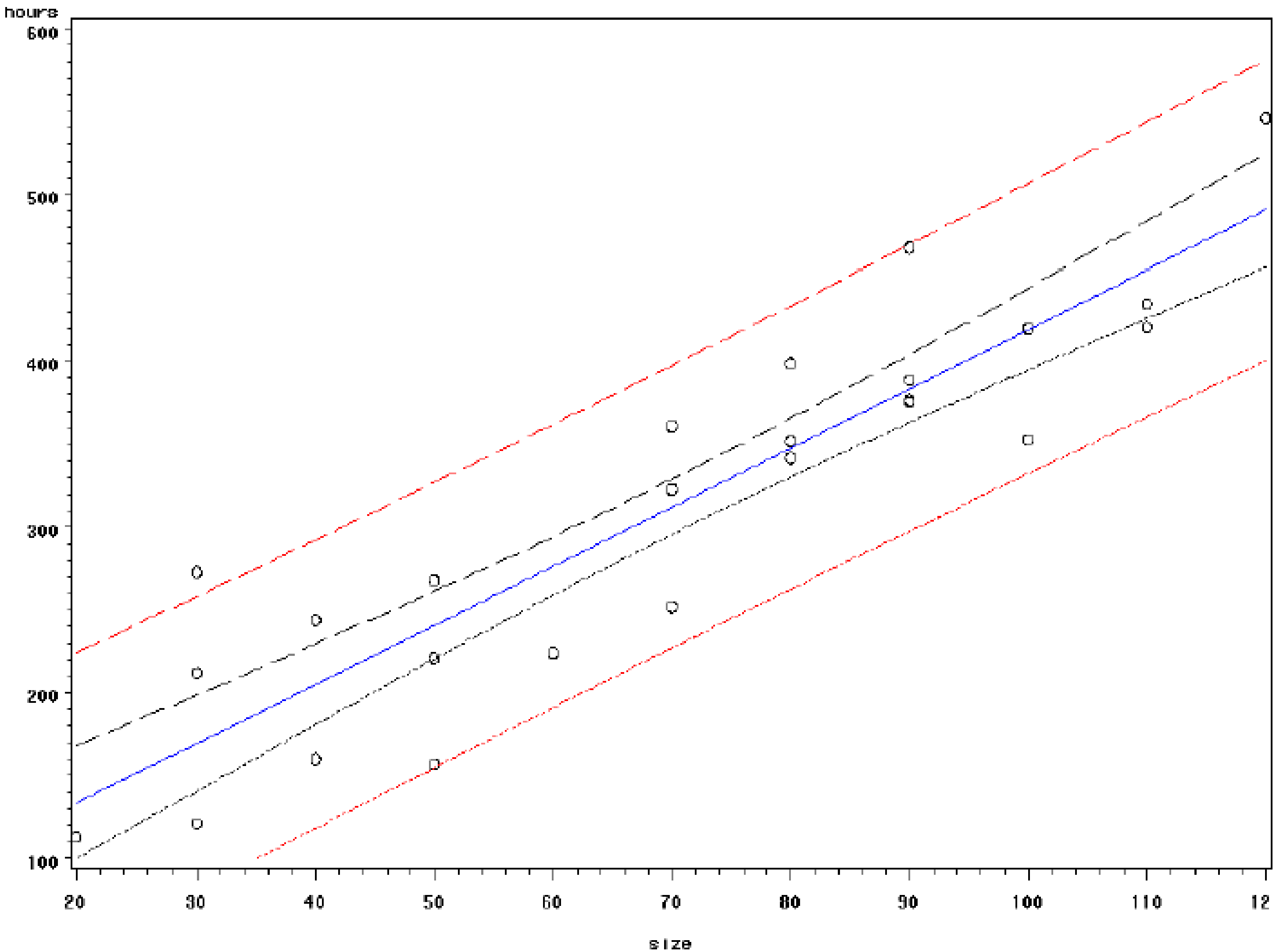
$$\hat{Y}_h \pm t\left(1-\alpha/2;n-2\right)s\left\{pred\right\}$$

$$s\{pred\}=s\left\{Y_{h(\text{new})}-\hat{Y}_h\right\}=\sqrt{MSE\left(1+\frac{1}{n}+\frac{(X_h-\overline{X})^2}{SS_{XX}}\right)}$$

Example: Toluca Company

Obs	size	Dep Var hours	Predicted Value	90% CL Mean		90% CL Predict	
1	80	399.0000	347.9820	330.2215	365.7425	262.4411	433.5230
2	30	121.0000	169.4719	140.3880	198.5559	80.8847	258.0591
3	50	221.0000	240.8760	220.3449	261.4070	154.7171	327.0348
4	90	376.0000	383.6840	363.1530	404.2151	297.5252	469.8429
5	70	361.0000	312.2800	295.5446	329.0154	226.9460	397.6140
6	60	224.0000	276.5780	258.8175	294.3385	191.0370	362.1189
7	120	546.0000	490.7901	456.6706	524.9096	400.4244	581.1558
8	80	352.0000	347.9820	330.2215	365.7425	262.4411	433.5230
9	100	353.0000	419.3861	394.9251	443.8470	332.2072	506.5649
10	50	157.0000	240.8760	220.3449	261.4070	154.7171	327.0348
11	40	160.0000	205.1739	180.7130	229.6349	117.9951	292.3528
12	70	252.0000	312.2800	295.5446	329.0154	226.9460	397.6140
22	90	468.0000	383.6840	363.1530	404.2151	297.5252	469.8429
23	40	244.0000	205.1739	180.7130	229.6349	117.9951	292.3528
24	80	342.0000	347.9820	330.2215	365.7425	262.4411	433.5230
25	70	323.0000	312.2800	295.5446	329.0154	226.9460	397.6140
26	65	.	294.4290	277.4315	311.4264	209.0432	379.8148
27	100	.	419.3861	394.9251	443.8470	332.2072	506.5649

Example: Toluca Company



2.6 Confidence Band for Regression Line

- $(1-\alpha)*100\%$ Confidence interval of $E\{Y_h\} = \beta_0+\beta_1 X_h$

$$\Pr\left\{\left|\frac{\hat{Y}_h - E(Y_h)}{s\{\hat{Y}_h\}}\right| < t(1-\alpha/2; n-2)\right\} = 1-\alpha \qquad s\{\hat{Y}_h\} = \sqrt{MSE\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right)}$$

- Consider looking at entire regression line, want to define likely region where line lies
- Working-Hotelling Confidence Band
  - Replace $t(1-\alpha/2, n-2)$ with Working-Hotelling value W in each confidence interval

$$W = \sqrt{2F(1-\alpha; 2, n-2)} \quad \Longrightarrow \quad \hat{Y}_h \pm W \times s\{\hat{Y}_h\}$$

  - The band is the narrowest at the mean of X

Proof

$$\left\{\left|\frac{\hat{Y}_h - E(Y_h)}{s\left\{\hat{Y}_h\right\}}\right| \le W \text{ for all } x_h\right\} = \left\{\max_{x_h}\left|\frac{\hat{Y}_h - E(Y_h)}{s\left\{\hat{Y}_h\right\}}\right| \le W\right\}$$

$$\frac{\hat{Y}_h - E(Y_h)}{s\left\{\hat{Y}_h\right\}} = \frac{\left(\hat{\beta}_0 + \hat{\beta}_1 X_h\right) - (\beta_0 + \beta_1 X_h)}{\sqrt{MSE\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right)}} \sim t(n-2)$$

$$\left(\frac{\hat{Y}_h - E(Y_h)}{s\left\{\hat{Y}_h\right\}}\right)^2 = \frac{\left[\left(\hat{\beta}_0 - \beta_0\right) + (\hat{\beta}_1 - \beta_1)X_h\right]^2}{MSE\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right)} = \frac{\left[\left(\overline{Y} - E\overline{Y}\right) + (\hat{\beta}_1 - \beta_1)(X_h - \overline{X})\right]^2}{MSE\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right)}$$

$$\boxed{\max_t \frac{(a+bt)^2}{c+dt^2} = \frac{a^2}{c} + \frac{b^2}{d}} \Rightarrow \max_{x_h}\left(\frac{\hat{Y}_h - E(Y_h)}{s\left\{\hat{Y}_h\right\}}\right)^2 = \frac{\left(\overline{Y} - E\overline{Y}\right)^2}{MSE/n} + \frac{\left(\hat{\beta}_1 - \beta_1\right)^2}{MSE/SS_{XX}}$$

It can be shown that (keep as Homework)

$$\frac{1}{2}\max_{x_h}\left(\frac{\hat{Y}_h - E(Y_h)}{s\left\{\hat{Y}_h\right\}}\right)^2 \sim F(2, n-2)$$

$$1-\alpha = P\left\{\frac{1}{2}\max_{x_h}\left(\frac{\hat{Y}_h - E(Y_h)}{s\left\{\hat{Y}_h\right\}}\right)^2 \le F(1-\alpha; 2, n-2)\right\} = P\left\{\max_{x_h}\left|\frac{\hat{Y}_h - E(Y_h)}{s\left\{\hat{Y}_h\right\}}\right| \le \sqrt{2F(1-\alpha; n-2)}\right\}$$

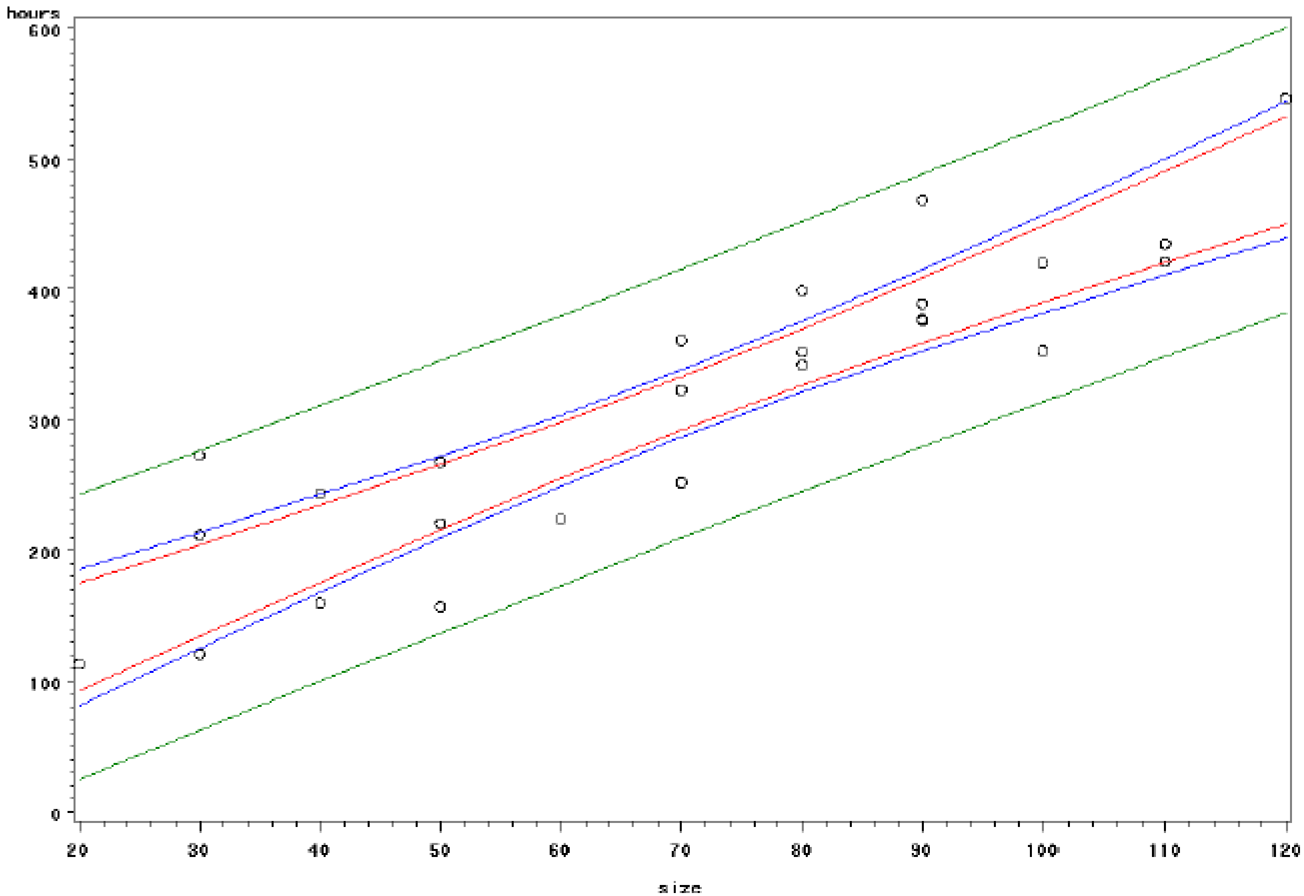
$$P\left\{\left|\frac{\hat{Y}_h - E(Y_h)}{s\left\{\hat{Y}_h\right\}}\right| \le W \text{ for all } x_h\right\} = P\left\{\max_{x_h}\left|\frac{\hat{Y}_h - E(Y_h)}{s\left\{\hat{Y}_h\right\}}\right| \le W\right\} = 1-\alpha$$

$$\text{where } W = \sqrt{2F(1-\alpha; n-2)}$$

- Simultaneous Confidence Band at $(1-\alpha)$ level

$$\left(\hat{Y}_h - W\cdot s\left\{\hat{Y}_h\right\}, \hat{Y}_h + W\cdot s\left\{\hat{Y}_h\right\}\right)$$

Confidence Band for the Toluca example

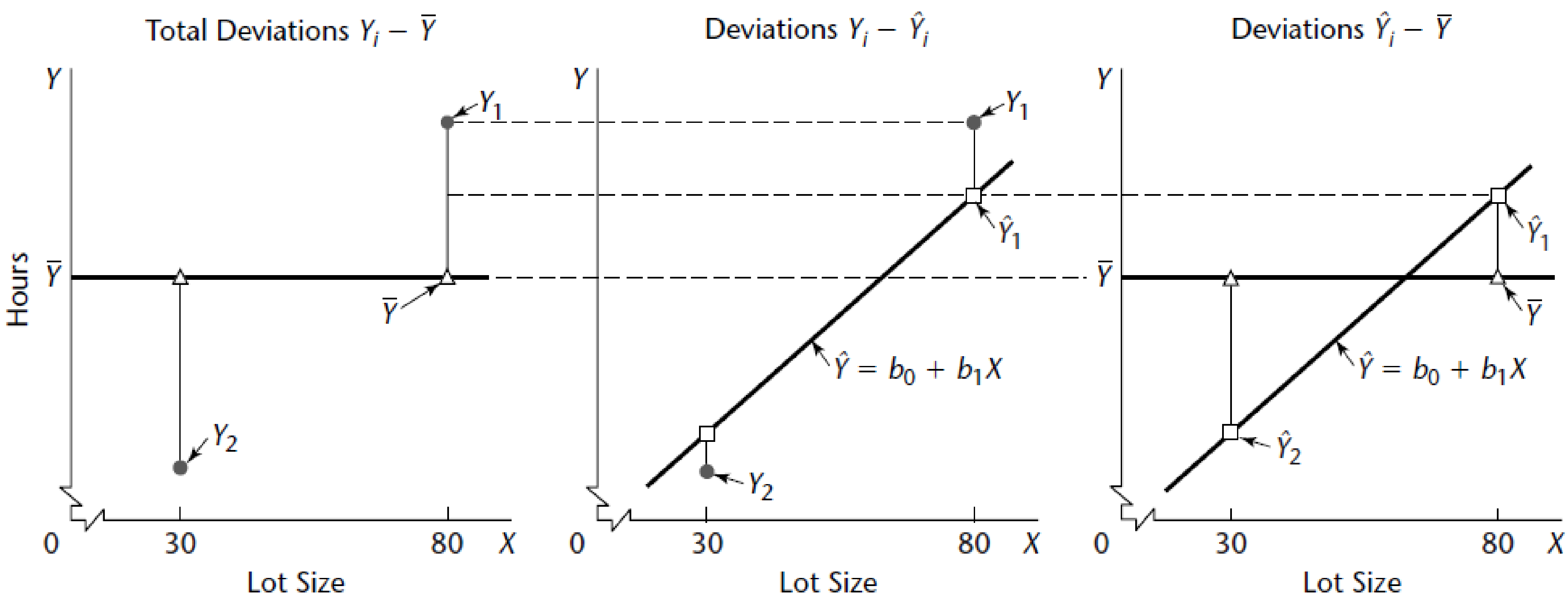


- Blue – 90% Working-Hotelling confidence band
- Red – 90% confidence interval for the mean $E\{Y_h\}$
- Green – 90% prediction interval for the individual observation $Y_{h(\text{new})}$

2.7 ANOVA Approach to Regression

- ANOVA (Analysis of Variance)

$$Y_i - \overline{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$$



ANOVA Partitioning

$$Y_i - \overline{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$$

$$\Rightarrow \left(Y_i - \overline{Y}\right)^2 = \left(Y_i - \hat{Y}_i\right)^2 + \left(\hat{Y}_i - \overline{Y}\right)^2 + 2\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \overline{Y}\right)$$

$$\Rightarrow \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + 2\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \overline{Y}\right)$$

Note(from Chapter 1):

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \overline{Y}\right) = \sum_{i=1}^{n} e_i\left(\hat{Y}_i - \overline{Y}\right) = \sum_{i=1}^{n} e_i\hat{Y}_i - \overline{Y}\sum_{i=1}^{n} e_i = 0 - 0 = 0$$

$$\Rightarrow \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$\Rightarrow SSTO = SSR + SSE$$

ANOVA Partitioning

Partition the total sum of squares SSTO into

- SSR– Model (explained by regression)
- SSE– Error (unexplained / residual)

$$SSTO = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2, \quad SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2, \quad SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$SSR = \sum_{i=1}^{n}\left(b_0 + b_1 X_i - b_0 - b_1 \bar{X}\right)^2 = b_1^2 SS_{XX}$$

- In normal error regression model, we have

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{SS_{XX}}\right) \quad \Rightarrow \quad \frac{b_1 - \beta_1}{\sqrt{\sigma^2 / SS_{XX}}} \sim N(0,1)$$

$$\frac{SSE}{\sigma^2} = \frac{(n-2)MSE}{\sigma^2} \sim \chi^2(n-2).$$

$$(b_0, b_1, \bar{Y}) \perp SSE \quad \Rightarrow \quad SSR = b_1^2 SS_{XX} \perp SSE$$

ANOVA Partitioning

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2) \qquad SSR \perp SSE$$

- Under $H_0 : \beta_1 = 0,$

$$b_1 \overset{H_0}{\sim} N\left(\beta_1 = 0, \frac{\sigma^2}{SS_{XX}}\right) \Rightarrow \frac{SSR}{\sigma^2} = b_1^2 SS_{XX} = \left(\frac{b_1 - 0}{\sqrt{\sigma^2 / SS_{XX}}}\right)^2 \overset{H_0}{\sim} \chi^2(1),$$

$$Y_1, Y_2, \cdots, Y_n \overset{H_0}{\sim} N(\beta_0, \sigma^2), i.i.d. \Rightarrow \frac{SSTO}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{\sigma^2} \overset{H_0}{\sim} \chi^2(n-1)$$

$$SSTO / \sigma^2 = SSR / \sigma^2 + SSE / \sigma^2$$

$$\boxed{\chi^2(n-1) \qquad \chi^2(1) \qquad \chi^2(n-2)} \qquad \boxed{SSR \perp SSE}$$

ANOVA Partitioning

- Generally(H_0 is not required true),

(1) $$\frac{SSE}{\sigma^2} \sim \chi^2(n-2,0)$$

(2) $$\frac{SSR}{\sigma^2} = \frac{b_1^2}{\sigma^2 / SS_{XX}} \sim \chi^2(1,\delta), \quad \text{since } b_1 \sim N\left(\beta_1, \frac{\sigma^2}{SS_{XX}}\right)$$

$$\text{where } \delta = \frac{\beta_1^2}{\sigma^2 / SS_{XX}}.$$

(3) $SSR \perp SSE$

$$(1),(2),(3) \Rightarrow \frac{SSTO}{\sigma^2} = \frac{SSR}{\sigma^2} + \frac{SSE}{\sigma^2} \sim \chi^2(n-1,\delta)$$

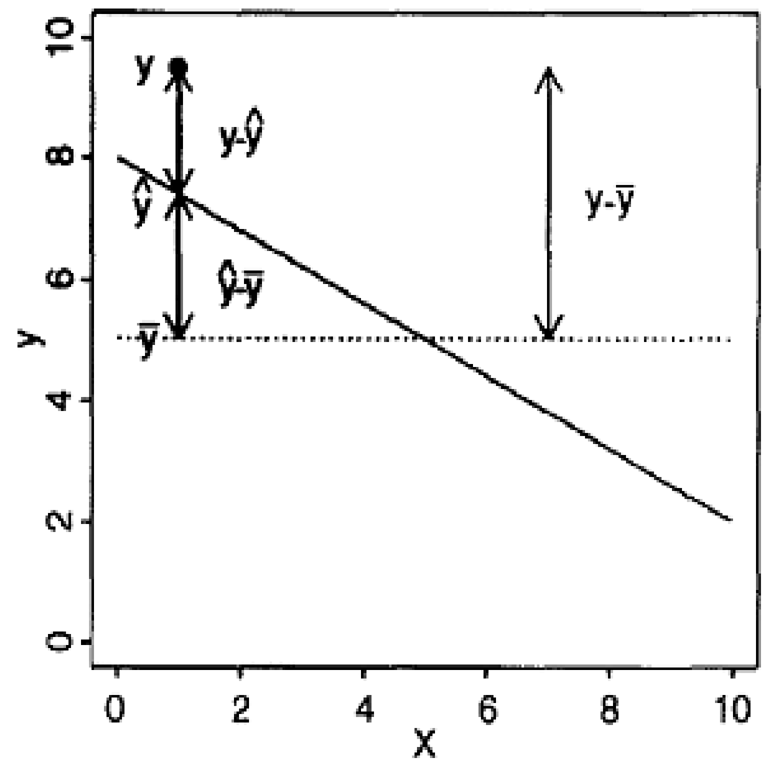
ANOVA Partitioning

$$SSTO = \sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 = SSR + SSE$$

$$SSTO/\sigma^2 \;=\; SSR/\sigma^2 \;+ SSE/\sigma^2$$

$$\boxed{\chi^2(n-1,\delta) \quad \chi^2(1,\delta) \quad \chi^2(n-2,0)}$$

$$\boxed{SSR \perp SSE} \qquad \boxed{\delta = \beta_1^2 SS_{XX}/\sigma^2}$$



- *SSR*— the explained variation or the regression sum of squares.
- *SSE* — sum of squared error
- Coefficient of determination(R^2)—SSR/SSY

Mean Squares

- Mean square= sum of square/ degrees of freedom
- The regression mean square is $MSR = SSR/1,$

$$E(MSR) = E(SSR) = E(b_1^2 SS_{XX}) = SS_{XX}\left(\mathrm{var}(b_1) + E^2(b_1)\right)$$

$$= SS_{XX}\left(\frac{\sigma^2}{SS_{XX}} + \beta_1^2\right) = \sigma^2 + \beta_1^2 SS_{XX}$$

- The mean square error is

$$MSE = \frac{SSE}{n-2}, \qquad E(MSE) = \sigma^2$$

- Then $F^* = \dfrac{SSR/1}{SSE/(n-2)} = \dfrac{MSR}{MSE} \overset{\beta_1=0}{\sim} F(1, n-2)$

$$F^* = \frac{MSR}{MSE} \sim F(1, n-2, \delta = \beta_1^2 SS_{XX}/\sigma^2)$$

F test

$$H_0 : \beta_1 = 0 \;\; vs \;\; H_1 : \beta_1 \neq 0$$

- ANOVA test Statistic

$$F^* = \frac{MSR}{MSE} \overset{H_0}{\sim} F(1, n-2)$$

  - When H_0 is false, $MSR > MSE$. Reject H_0 when F^* large.

- Hypothesis test decision rule
  - Reject H_0 when
  - p-value $=$ Pr($F\,(1, n-2) > F^*$)

ANOVA Table

Source of Variation	SS	df	MS	F	P
Regression (Model)	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$	$P(F(1,n-2)>F)$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$		
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$			

Toluca example

- ANOVA table

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	252378	252378	105.88	<.0001
Error	23	54825	2383.71562		
Cor Total	24	307203			

Root MSE	48.82331	R-Square	0.8215

- t test results

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	62.36586	26.17743	2.38	0.0259
size	1	3.57020	0.34697	10.29	<.0001

- Note that **$10.29^2 = 105.88$**

Equivalence of F test and two-sided t test

$$H_0 : \beta_1 = 0 \;\; vs \;\; H_1 : \beta_1 \neq 0$$

- Test statistics

$$F^* = \frac{MSR}{MSE} = \frac{b_1^2 SS_{XX}}{MSE} = \left(\frac{b_1}{\sqrt{MSE / SS_{XX}}} \right)^2 = \left(\frac{b_1}{s(b_1)} \right)^2 = (t^*)^2$$

- In addition:

$$t^2(n-2) \sim F(1, n-2) \Rightarrow \quad t^2(1-\alpha/2; n-2) = F(1-\alpha; 1, n-2)$$

- Equivalence of rejection regions

$$F^* > F(1-\alpha; 1, n-2) \;\; \Leftrightarrow \;\; |t^*| > t(1-\alpha/2; n-2)$$

- Equivalence of p values

$$\text{p-value} = \Pr(\, F(1, n-2) > F^*)$$
$$\Leftrightarrow \; \text{p-value} = \Pr(\, t(n-2) > |t^*|\,)$$

2.8 General Linear Test Approach

- General Linear Test – Very Flexible Method
- Consider **two** models,
  - Full/unrestricted model, e.g. $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - Reduced/restricted model, e.g. $Y_i = \beta_0 + \varepsilon_i$
- Compare models using SSE's
  - SSE(F): Error sum of squares of the full
  - SSE(R): Error sum of squares of the reduced model
- Can be shown that SSE(F) $\leq$ SSE(R)
  - Idea: more parameters provide better fit

General Linear Test

- If SSE(F) not much smaller than SSE(R), full model doesn't better explain *Y*.
- H_0: Reduced model vs H_1: Full model
- Test Statistic

$$F^* = \frac{\left[\left(SSE(R) - SSE(F)\right) / \left(df_R - df_F\right)\right]}{\left[SSE(F)/df_F\right]}$$

  - Large F^* suggests full model, and small F^* suggests reduced model.
- In normal error models, $F^* \overset{H_0}{\sim} F\left(df_R - df_F, df_F\right)$
- Decision rule
  - Reject H_0 if $F^* \geq F\left(1-\alpha; df_R - df_F, df_F\right)$.

General Linear Test

- Full/unrestricted model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_1, \cdots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$

$$\hat{\beta}_1 = b_1 = \frac{SS_{XY}}{SS_{XX}}, \qquad \hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

$$SSE(F) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i(F)\right)^2 = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i\right)^2 = SSE, \ df_F = n-2$$

- To test $H_0 : \beta_1 = 0 \ vs \ H_1 : \beta_1 \neq 0$
- The model when H_0 holds is a reduced or restricted model.
  - Reduced/restricted model $Y_i = \beta_0 + \varepsilon_i$
  - Under $H_0 : \beta_1 = 0, \ Y_1, \cdots, Y_n \overset{iid}{\sim} N(\beta_0, \sigma^2), \ \hat{\beta}_0 = \bar{Y}$

$$SSE(R) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i(R)\right)^2 = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = SSTO, \ df_R = n-1$$

$$F^* = \frac{\left[\left(SSE(R) - SSE(F)\right) / \left(df_R - df_F\right)\right]}{\left[SSE(F)/df_F\right]} = \frac{(SSTO - SSE)/1}{SSE/(n-2)} = \frac{MSR}{MSE} \overset{H_0}{\sim} F(1, n-2)$$

2.9 Descriptive Measures of Linear Association

- **Linear association measures**
  - Coefficient of determination R^2 in linear regression
  - Estimated Pearson's correlation coefficient r

- SSTO measures the variation in the observations Y_i when X is not considered.
- SSE measures the variation in the Y_i after a predictor variable X is employed.
- SSR =SSTO−SSE measures the effect of X in reducing variation in Y.

Coefficient of Determination

- Coefficient of Determination

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

  - the proportion of total variation in Y explained by X.
- Note that since $0 \leq SSE \leq SSTO$, then $0 \leq R^2 \leq 1$.
- Limitations of and misunderstandings about R^2 (See page 75)
  - High R^2 does not necessarily mean that
    - useful predictions can be made;
    - regression line is a good fit.
  - Low R^2 does not necessarily mean that X and Y are not related.

Pearson's Correlation Coefficient

- Pearson's product-moment correlation coefficient measures the strength of the **linear** relationship between two variables.

$$\rho = corr(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}$$

- ρ can be estimated by

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \bullet \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} = \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}}, \quad -1 \le r \le 1$$

- For simple linear regression

$$b_1 = \frac{SS_{XY}}{SS_{XX}}, \qquad R^2 = \frac{SSR}{SSTO} = \frac{b_1^2 SS_{XX}}{SS_{YY}} = \frac{SS_{XY}^2}{SS_{XX}SS_{YY}}$$

since $SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(b_0 + b_1X_i - b_0 - b_1\overline{X}\right)^2 = b_1^2 SS_{XX}$

correlation coefficient

- Relationship

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX} SS_{YY}}} = \frac{\sqrt{SS_{XX}}}{\sqrt{SS_{YY}}} b_1 = \frac{S_X}{S_Y} b_1, \qquad r = \pm\sqrt{R^2}$$

  - sign of r is the sign of the regression slope.

    $$r = \sqrt{R^2},\ \ if\ \ b_1 \ge 0; \quad r = -\sqrt{R^2},\ \ if\ \ b_1 < 0$$

  - Relationship not true in multiple regression

- r (but not b_1) is not changed by linear transformations of Y and/or X.

- Toluca example, $R^2 = 0.822,\ b_1 > 0,\ \ r = \sqrt{0.822} = 0.907$

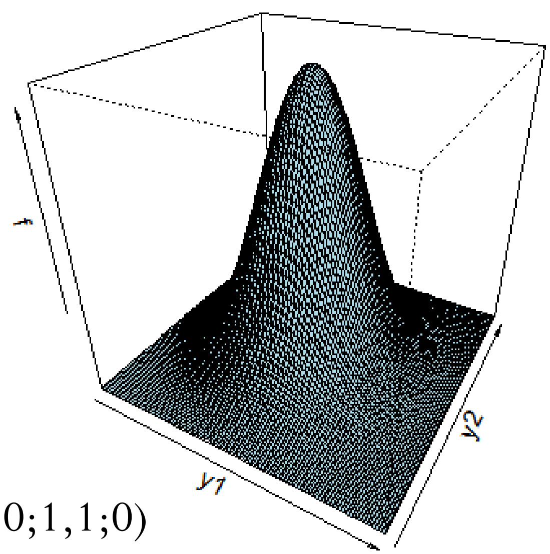
2.11 Normal Correlation Model

- In regression models, we have assumed X_i's are known constants
- Statistical inferences consider repeated sampling with fixed X values
- What if X_i's are random samples from distribution $g(\cdot)$?
- Previous regression results on estimation, testing and prediction still hold if:
  - $Y_i | X_i \sim \mathrm{N}(\beta_0 + \beta_1 X_i, \sigma^2)$, and Y_i's are conditional independent.
  - The X_i are independent and $g(\cdot)$ does not involve the parameters β_0, β_1, and σ^2

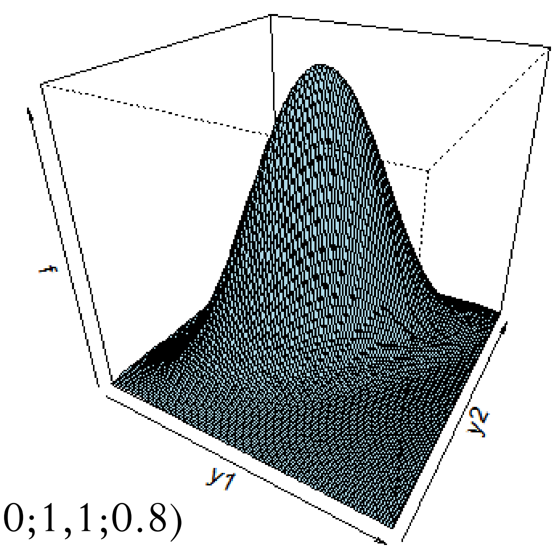
Normal correlation model

- If interest in relation between two variables can use correlation model.
  - Both X and Y are random.
- Normal correlation model uses bivariate normal distribution
- Bivariate normal distribution $N(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2; \rho)$

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} \exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{12}\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right) + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$



$N(0,0;1,1;0)$



$N(0,0;1,1;0.8)$

Bivariate normal distribution

- k-dimensional normal distribution

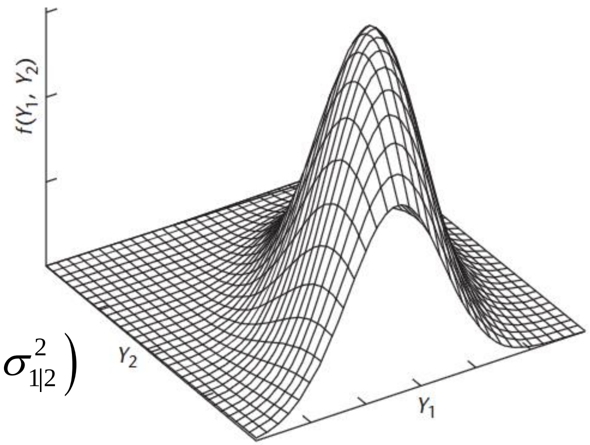
$$f(y_1,\cdots,y_k)=\frac{1}{(2\pi)^{k/2}\sqrt{\det\boldsymbol{\Sigma}}}\exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}$$

$$\text{if } k=2, \mathbf{y}=(y_1,y_2)', \boldsymbol{\mu}=(\mu_1,\mu_2)', \Sigma=\begin{pmatrix}\sigma_1^2 & \rho_{12}\sigma_1\sigma_2\\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$$

- Marginal distributions are normal

$$Y_1\sim N\left(\mu_1,\sigma_1^2\right),\qquad Y_2\sim N\left(\mu_2,\sigma_2^2\right)$$

- Conditional distributions are normal



$$(Y_1\,|\,Y_2=y_2)\sim N\left(\mu_1+\rho_{12}\frac{\sigma_1}{\sigma_2}(y_2-\mu_2),\sigma_1^2\left(1-\rho_{12}^2\right)\right)\equiv N\left(\alpha_{1|2}+\beta_{12}y_2,\sigma_{1|2}^2\right)$$

where: $\alpha_{1|2}=\mu_1-\mu_2\rho_{12}\dfrac{\sigma_1}{\sigma_2},\quad \beta_{12}=\rho_{12}\dfrac{\sigma_1}{\sigma_2},\quad \sigma_{1|2}^2=\sigma_1^2\left(1-\rho_{12}^2\right)$

$$(Y_2\,|\,Y_1=y_1)\sim N\left(\mu_2+\rho_{12}\frac{\sigma_2}{\sigma_1}(y_1-\mu_1),\sigma_2^2\left(1-\rho_{12}^2\right)\right)\equiv N\left(\alpha_{2|1}+\beta_{21}y_2,\sigma_{2|1}^2\right)$$

Inference on ρ_{12}

- Under bivariate normal assumption, the MLE of ρ_{12}

$$\hat{\rho}_{12} = r_{12} = \frac{SS_{XY}}{\sqrt{SS_{XX} SS_{YY}}},$$ just Pearson's correlation coefficient.

- Interest in testing $H_0 : \rho_{12} = 0$
  - $\rho_{12} = 0 \Leftrightarrow \beta_{12} = \beta_{21} = 0$

$$\frac{r_{12}}{\sqrt{(1-r_{12}^2)/(n-2)}} = \frac{b_1}{\sqrt{MSE / SS_{XX}}} = \frac{b_1}{s(b_1)} = t^* \overset{H_0}{\sim} t(n-2)$$

- Test Statistic $t^* = \dfrac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}}$
- Decision rule:
  - For 2-sided test $H_1 : \rho_{12} \neq 0$, reject H_0 if $\left|t^*\right| \geq t\left(1-(\alpha/2); n-2\right)$
  - For 1-sided tests:

$$H_1 : \rho_{12} > 0 : \quad \text{Reject } H_0 \text{ if } t^* \geq t\left(1-\alpha; n-2\right)$$

$$H_1 : \rho_{12} < 0 : \quad \text{Reject } H_0 \text{ if } t^* \leq -t\left(1-\alpha; n-2\right)$$

Inference on ρ_{12}

Interval Estimation of ρ_{12}

- When $\rho_{12} \neq 0$, the distribution of r_{12} is messy.
- Make the Fisher z transformation:

$$z' = \frac{1}{2}\ln\left(\frac{1+r_{12}}{1-r_{12}}\right)$$

- When n is large (n > 25), approximately,

$$z' \overset{approx}{\sim} N\left(\zeta, \frac{1}{n-3}\right) \qquad \zeta = \frac{1}{2}\ln\left(\frac{1+\rho_{12}}{1-\rho_{12}}\right) \Rightarrow \rho_{12} = \frac{e^{2\zeta}-1}{e^{2\zeta}+1} \uparrow in\ \zeta$$

- 100 (1−α)% confidence limits for ζ

$$z' \pm z\left(1-(\alpha/2)\right)\sqrt{\frac{1}{n-3}} \Rightarrow \text{ CI: } (c_1, c_2)$$

- Then 100 (1−α)% confidence interval of ρ_{12} : $\left(\frac{e^{2c_1}-1}{e^{2c_1}+1}, \frac{e^{2c_2}-1}{e^{2c_2}+1}\right)$

Spearman's correlation method

If X and Y non-normal, can use Spearman's correlation coefficient

1) Rank $\left(Y_{11},\ldots,Y_{n1}\right)$ from 1 to n (smallest to largest) and label: $\left(R_{11},\ldots,R_{n1}\right)$

2) Rank $\left(Y_{12},\ldots,Y_{n2}\right)$ from 1 to n (smallest to largest) and label: $\left(R_{12},\ldots,R_{n2}\right)$

3) Compute Spearman's rank correlation coefficient:

$$r_S = \frac{\sum_{i=1}^{n}\left(R_{i1}-\overline{R}_1\right)\left(R_{i2}-\overline{R}_2\right)}{\sqrt{\sum_{i=1}^{n}\left(R_{i1}-\overline{R}_1\right)^2\sum_{i=1}^{n}\left(R_{i2}-\overline{R}_2\right)^2}}$$

To Test: $\mathrm{H}_0:$ No Association Between Y_1, Y_2 vs $\mathrm{H}_A:$ Association Exists

Test Statistic: $t^* = \dfrac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}}$ Reject H_0 if $\left|t^*\right| \geq t\left(1-\left(\alpha/2\right);n-2\right)$

Spearman's rank correlation coefficient

Example: examine whether an association exists between population size (Y_1) and per capita expenditures for a new food product (Y_2).

Test Market i	(1) Population (in thousands) Y_{i1}	(2) Per Capita Expenditure (dollars) Y_{i2}	(3) R_{i1}	(4) R_{i2}
1	29	127	1	2
2	435	214	8	11
3	86	133	3	4
4	1,090	208	11	10
5	219	153	7	6
6	503	184	9	8
7	47	130	2	3
8	3,524	217	12	12
9	185	141	6	5
10	98	154	5	7
11	952	194	10	9
12	89	103	4	1

$r_s = 0.895$

R code

```
toluca = read.table('D:\\Reg_licx\\Data_4e\\CH01TA01.txt',header=F)
names(toluca)<-c("Size", "Hours")
fit = lm(Hours~Size, data=toluca)
summary(fit)  ##Model results
coef(fit)  ##regression coefficients b0 and b1
vcov(fit)  ##Covarince matrix of b0 and b1
confint(fit) ##95% confidence interval of b0 and b1
confint(fit,level=0.99) ##99% CI of b0 and b1
####Inference for E{Y_h}
# 90% confidence interval of E{Y_h} when X_h=100
predict(fit, newdata=data.frame(Size=100), interval="confidence", level=.9)
# 90% CI of E{Y_h} at all Xi's in original trained data
predict(fit, interval="confidence", level=.9)
```

R code

```
####Prediction  for a new observation
predict(fit, newdata=data.frame(Size=100), interval="prediction", level=.9)
predict(fit, interval="prediction ",level=.9)
####ANOVA F test
aov(fit)
####General linear test
fit0 = lm(Hours~1, data=toluca)
anova(fit0, fit, test = "F")
#####Correlation analysis
cor(toluca$Hours,toluca$Size) #Pearson's product-moment correlation
cor.test(toluca$Hours,toluca$Size)  #test for Pearson's correlation
cor(toluca$Hours,toluca$Size,method="spearman") #Spearman's rank correlation
```

Homework

P89~98:

- 2.6; 2.15; 2.25; 2.42; 2.46
- 2.53; 2.57; 2.59; 2.60
- Under the normal error model (2.1), MSE is an unbiased estimator of σ^2. Please calculate $E\left(\sqrt{MSE}\right)$ and show that it is a biased estimator of σ.
- For obtaining W-H confidence band for the regression line (at any level X_h) under the normal error model (2.1), prove that

$$\frac{1}{2}\max_{x_h}\left(\frac{\hat{Y}_h - E(Y_h)}{s\left\{\hat{Y}_h\right\}}\right)^2 \sim F(2, n-2)$$

Appendix： Noncentral distributions

(Noncentral χ^2 distribution) Let $X_1, X_2, \ldots, X_n$ be independent with $X_i \sim N(\mu_i, \sigma^2)$, $i=1,2, \ldots,n$.

$$Y = \sum_{i=1}^{n} X_i^2 / \sigma^2 \sim \chi^2(n, \delta), \text{ with } \delta = \sum_{i=1}^{n} \mu_i^2 / \sigma^2$$

Property: If $U_1 \sim \chi^2(r_1, \delta_1)$ and $U_2 \sim \chi^2(r_2, \delta_2)$ are independent, then $U_1 + U_2 \sim \chi^2(r_1 + r_2, \delta_1 + \delta_2)$

(Noncentral t distribution) If $Z \sim N(0,1)$, $V \sim \chi^2(r)$ are independent, then $$\frac{Z+\delta}{\sqrt{V/r}} \sim t(r, \delta)$$

(Noncentral F distribution) If $U \sim \chi^2(r_1, \delta)$, $V \sim \chi^2(r_2)$ are independent, then $$\frac{U/r_1}{V/r_2} \sim F(r_1, r_2, \delta)$$

Fisher's Theorem

(Fisher's Theorem) Let X_1, X_2, …, X_n be independent $N(\mu_i, \sigma^2)$ distributed random variables, and $Q = Q_1 + Q_2 + \ldots + Q_k$,where $Q, Q_1, Q_2, \ldots, Q_k$ are quadratic forms in X_1, X_2, …, X_n, i.e., $Q = \mathbf{X'AX}$, and $Q_i = \mathbf{X'A}_i\mathbf{X}$, $i = 1, 2, \ldots, k$. If

$$Q/\sigma^2 \sim \chi^2(r, \delta),\ Q_1/\sigma^2 \sim \chi^2(r_1, \delta_1), \cdots, Q_{k-1}/\sigma^2 \sim \chi^2(r_{k-1}, \delta_{k-1}),$$

Then $Q_1, Q_2, \ldots, Q_k$ are independent, and

$$Q_k/\sigma^2 \sim \chi^2(r_k, \delta_k),$$

where $r_k = r - (r_1 + \cdots + r_{k-1}), \delta_k = \delta - (\delta_1 + \cdots + \delta_{k-1})$.