

# Data modeling: CSCI E-106

Applied Linear Statistical Models

Chapter 4 - Simultaneous Inferences and Other Topics in Regression Analysis

# Joint Estimation of $\beta_0$ and $\beta_1$

- To draw inferences with confidence coefficient 0.95 about both the intercept  $\beta_0$  and slope  $\beta_1$ .
  - We could use the methods of Chapter 2 to construct separate 95 percent confidence intervals for  $\beta_0$  and  $\beta_1$ .
  - If the inferences are independent, the probability of both being correct would be  $(.95)^2 = 0.9025$ .
  - These would not provide 95 percent confidence that the conclusions for both  $\beta_0$  and  $\beta_1$ , are correct.
- A family confidence coefficient of 0.95 would indicate that if repeated samples are selected and interval estimates for both  $\beta_0$  and  $\beta_1$  are calculated for each sample by specified procedures would lead to a family of estimates where both confidence intervals are correct.

# Bonferroni Joint Confidence Intervals

- Developing joint confidence intervals for  $\beta_0$  and  $\beta_1$  with a specified family confidence coefficient
  - each statement confidence coefficient is adjusted to be higher than  $1 - \alpha$  so that the family confidence coefficient is at least  $1 - \alpha$
- The ordinary confidence limits for  $\beta_0$  and  $\beta_1$  with  $1 - \alpha$

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

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

# Bonferroni Joint Confidence Intervals, cont'd

- What is the probability that one or both of these intervals are incorrect?
  - $A_1$  : the event that the first C.I. does not cover  $\beta_0$
  - $A_2$  : the event that the second C.I. does not cover  $\beta_1$ $\Rightarrow P(A_1) = \alpha \quad \text{and} \quad P(A_2) = \alpha$
- Bonferroni inequality:
  - $P(\bar{A}_1 \cap \bar{A}_2) \geq 1 - \alpha - \alpha = 1 - 2\alpha$

# Bonferroni Joint Confidence Intervals, cont'd

- $\beta_0$  and  $\beta_1$  are separately estimated with 0.95% C.I.  $\Rightarrow$  the Bonferroni inequality guarantees us a family confidence coefficient of at least 90 percent that both intervals based on the same sample are correct.
- The  $1 - \alpha$  family confidence limits for  $\beta_0$  and  $\beta_1$ , for regression model (2.1) by the Bonferroni procedure are:

$$b_0 \pm Bs\{b_0\} \quad b_1 \pm Bs\{b_1\}$$

$$B = t(1 - \alpha/4; n - 2)$$

# Bonferroni Joint Confidence Intervals, cont'd

- For the Toluca Company example, 90 percent family confidence intervals for  $\beta_0$  and  $\beta_1$  require  $B = t(1 - .10/4; 23) = t(0.975; 23) = 2.069$ .
- We have from Chapter 2:

$$b_0 = 62.37 \quad s\{b_0\} = 26.18$$

$$b_1 = 3.5702 \quad s\{b_1\} = .3470$$

- Hence, the respective confidence limits for  $\beta_0$  and  $\beta_1$ , are

$$8.20 \leq \beta_0 \leq 116.5$$

$$2.85 \leq \beta_1 \leq 4.29$$

# Bonferroni Joint Confidence Intervals, cont'd

- `toluca<-read.table("toluca.txt",header=T)`
- `attach(toluca)`
- `fitreg<-lm(Hrs~Size)`
- `confint(fitreg)`
- `confint(fitreg,level=1-alpha/2) # Bonferroni C.I.`

# Bonferroni Joint Confidence Intervals, cont'd

- The Bonferroni  $1 - \alpha$  family confidence coefficient is actually a lower bound on the true family confidence coefficient.
- If  $g$  interval estimates are desired with family confidence coefficient  $1 - \alpha$ , constructing each interval estimate with statement confidence coefficient  $1 - \alpha / g$  will suffice.

$$P\left(\bigcap_{i=1}^g \bar{A}_i\right) \geq 1 - g\alpha$$

- The Bonferroni technique is ordinarily most useful when the number of simultaneous estimates is not too large.
- It is not necessary with the Bonferroni procedure that the C.I. have the same statement confidence coefficient. ( $P(A_1) + P(A_2) = \alpha$ )
- Joint confidence intervals can be used directly for testing.



# Bonferroni Joint Confidence Intervals, cont'd

- The estimators  $b_0$  and  $b_1$  are usually correlated, but the Bonferroni simultaneous confidence limits recognize this correlation by means of the bound on the family confidence coefficient.

$$\sigma\{b_0, b_1\} = -\bar{X}\sigma^2\{b_1\}$$

- Note that if  $\bar{X}$  is positive,  $b_0$  and  $b_1$  are negatively correlated, implying that if the estimate  $b_1$  is too high, the estimate  $b_0$  is likely to be too low, and vice versa.
- When the independent variable is  $X_i - \bar{X}$ , as in the alternative model (1.6),  $b_0^*$  and  $b_1$  are uncorrelated according to because the mean of the  $X_i - \bar{X}$  observations is zero.

# Simultaneous Estimation of Mean Responses

- The mean responses at a number of  $X$  levels need to be estimated.
  - Example: Toluca Company: the mean number of work hours for  $X = 30, 65, 100$  units
- Two procedures for simultaneous estimation of a number of different mean responses:
  - Working-Hotelling
  - Bonferroni
- A family confidence coefficient is needed for estimating several mean responses:
  - even though all estimates are based on the same fitted regression line, the separate interval estimates of  $E\{Y_h\}$  at the different  $X_h$  levels need not all be correct or all be incorrect.
  - The combination of sampling errors in  $b_0$  and  $b_1$  may be such that the interval estimates of  $E\{Y_h\}$  will be correct over some range of  $X$  levels and incorrect elsewhere.

# Working-Hotelling Procedure

- **Working-Hotelling procedure:** based on the confidence band for the regression line (Chap. 2.6)

$$\hat{Y}_h \pm Ws\{\hat{Y}_h\}, \quad W^2 = 2F(1 - \alpha; 2, n - 2)$$

- The confidence band contains the entire regression line  $\Rightarrow$  contains the mean responses at all  $X$  levels
- The simultaneous confidence limits for  $g$  mean responses  $E\{Y_h\}$  for model (2.1) with the Working-Hotelling procedure:

$$\hat{Y}_h \pm Ws\{\hat{Y}_h\}, \quad W^2 = 2F(1 - \alpha; 2, n - 2)$$

- $\hat{Y}_h = b_0 + b_1 X_h$
- $s\{\hat{Y}_h\} = MSE \left[ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$

# Working-Hotelling Procedure, cont'd

- Example: Toluca Company:
  - $Y_h$  at  $X = 30, 65, 100$  units
  - $1 - \alpha = 0.90$ ;  $F(0.90; 2, 23) = 2.549$
  - $W^2 = 5.098 \Rightarrow W = 2.258$

$X_h$	$\hat{Y}_h$	$s\{\hat{Y}_h\}$
30	169.5	16.97
65	294.4	9.918
100	419.4	14.27

$\Rightarrow$

$$131.2 = 169.5 - 2.258(16.97) \leq E\{Y_h\} \leq 169.5 + 2.258(16.97) = 207.8$$

$$272.0 = 294.4 - 2.258(9.918) \leq E\{Y_h\} \leq 294.4 + 2.258(9.918) = 316.8$$

$$387.2 = 419.4 - 2.258(14.27) \leq E\{Y_h\} \leq 419.4 + 2.258(14.27) = 451.6$$

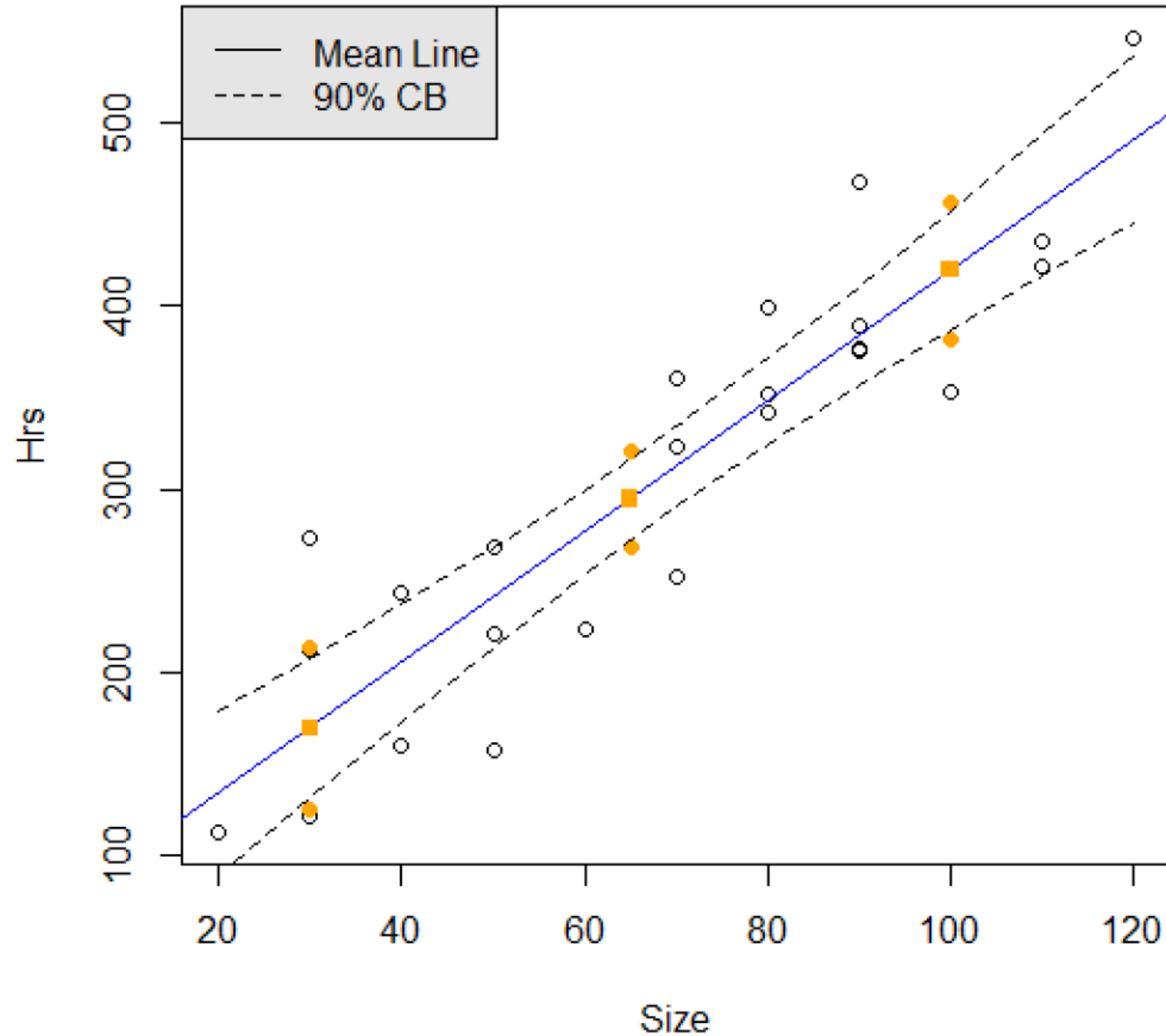
- concluded that the mean number of work hours required is between 131.2 and 207.8 for lots of 30 parts, between 272.0 and 316.8 for lots of 65 parts, and between 387.2 and 451.6 for lots of 100 parts with family confidence coefficient .90.

# Working-Hotelling Procedure, cont'd

- `fitreg<-lm(Hrs~Size)`
- `Xh<-c(30,65,100)`
- `pred<-predict.lm(fitreg,data.frame(Size = c(Xh)),`
- `se.fit = T, level = 0.9)`
- `W <-rep(sqrt( 2 * qf(0.90, 2, n-2) ),length(Xh))`
- `rbind(pred$fit-W * pred$se.fit,pred$fit+W * pred$se.fit)`

	1	2	3
[1,]	131.1542	272.0351	387.1591
[2,]	207.7897	316.8229	451.6130

# Working-Hotelling Procedure, cont'd



# Bonferroni procedure

- When  $E\{Y_h\}$  is to be estimated for  $g$  levels  $X_h$  with family confidence coefficient  $1-\alpha$ , the Bonferroni confidence limits for regression model (2.1) are:

$$\hat{Y}_h \pm Bs\{\hat{Y}_h\}$$

- Where  $B=t(1-\alpha/2g;n-2)$ , and  $g$  is the number of confidence intervals in the family

# Bonferroni procedure, cont'd

- Example: Toluca Company:
  - $Y_h$  at  $X = 30, 65, 100$  units
  - $1 - \alpha = 0.90$ ;  $F(0.90; 2, 23) = 2.549$
  - $B = t(1 - 0.1/(2 \times 3); 23) = 2.263$

$X_h$	$\hat{Y}_h$	$s\{\hat{Y}_h\}$	
30	169.5	16.97	$131.1 = 169.5 - 2.263(16.97) \leq E\{Y_h\} \leq 169.5 + 2.263(16.97) = 207.9$
65	294.4	9.918	$\Rightarrow 272.0 = 294.4 - 2.263(9.918) \leq E\{Y_h\} \leq 294.4 + 2.263(9.918) = 316.8$
100	419.4	14.27	$387.1 = 419.4 - 2.263(14.27) \leq E\{Y_h\} \leq 419.4 + 2.263(14.27) = 451.7$

- concluded that the mean number of work hours required is between 131.1 and 207.9 for lots of 30 parts, between 272.0 and 316.8 for lots of 65 parts, and between 387.1 and 451.7 for lots of 100 parts with family confidence coefficient .90.

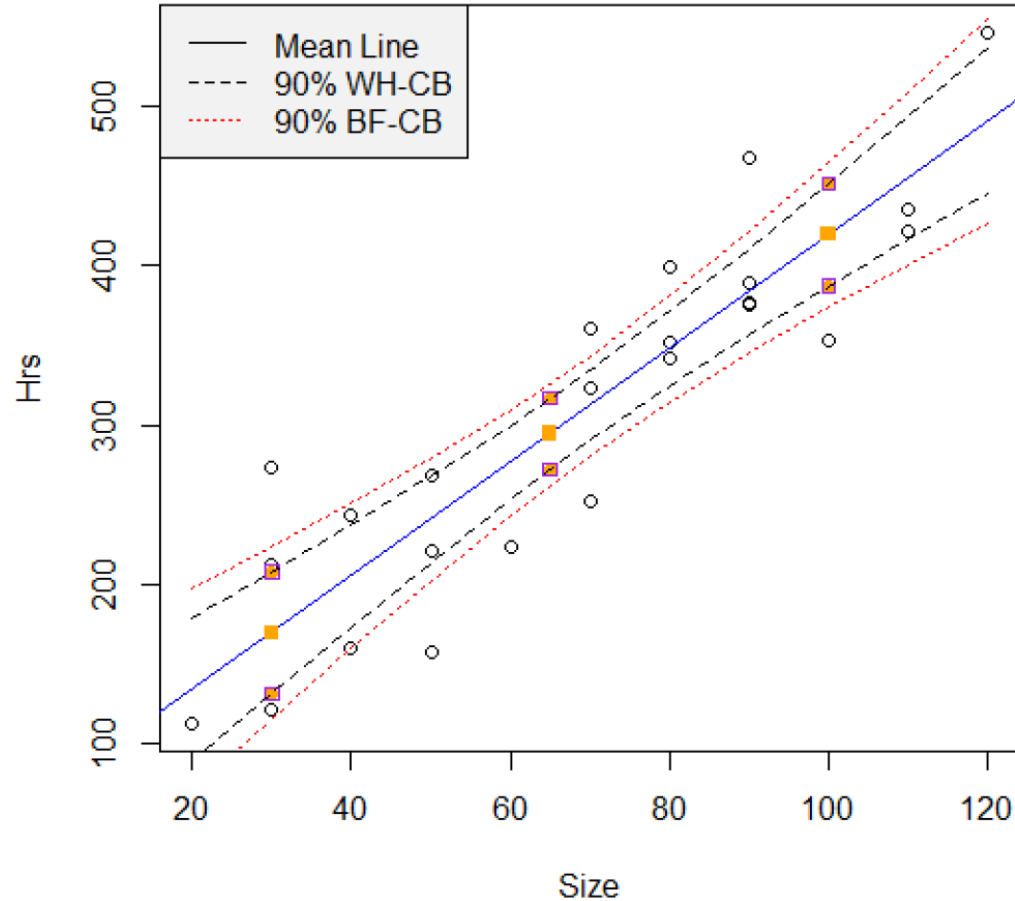


# Bonferroni procedure, cont'd

- `fitreg<-lm(Hrs~Size)`
- `Xh<-c(30,65,100)`
- `pred<-predict.lm(fitreg,data.frame(Size = c(Xh)),`
- `se.fit = T, level = 0.9)`
- `B=rep(qt(1-alpha/(2*length(Xh)),n-2),length(Xh))`
- `rbind(pred$fit-B * pred$se.fit, pred$fit+B * pred$se.fit)`

	1	2	3
[1,]	131.0570	271.9783	387.0774
[2,]	207.8868	316.8797	451.6947

# Bonferroni procedure, cont'd



- The Working-Hotelling confidence limits are slightly tighter than, or the same as, the Bonferroni limits.
- If the number of statements is small, the Bonferroni limits may be tighter. For larger families, the Working-Hotelling confidence limits will always be the tighter, since  $W$  stays the same for any number of statements in the family whereas  $B$  becomes larger as the number of statements increases.
- Both the Working-Hotelling and Bonferroni procedures provide lower bounds to the actual family confidence coefficient.

# Simultaneous Prediction Intervals for New Observations

- The simultaneous predictions of  $g$  new observations on  $Y$  in  $g$  independent trials at  $g$  different levels of  $X$ .
- Two procedures:
  - Scheffé using **F distribution**

$$\hat{Y}_h \pm Ss\{pred\}, \quad S^2 = gF(1-\alpha; g, n-2)$$

- Bonferroni using **t distribution**

$$\hat{Y}_h \pm Bs\{pred\}, \quad B = t(1-\alpha/2g; n-2)$$

where,

$$s^2\{pred\} = MSE \left[ 1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$

# Simultaneous Prediction Intervals for New Observations, cont'd

- Example: Toluca Company:
  - Predict  $Y_{h(\text{new})}$  at  $X_h = 80, 100$  units
  - $1 - \alpha = 0.95$ ;  $S^2 = 2 * F(0.95; 2, 23) = 6.844 \Rightarrow S = 2.616$
  - $B = t(1 - 0.05/(2 \times 2); 23) = 2.398$
  - Bonferroni procedure will yield somewhat tighter prediction limits as (  $B \leq S$  )

$X_h$	$\hat{Y}_h$	$s\{\text{pred}\}$	$Bs\{\text{pred}\}$
80	348.0	49.91	119.7
100	419.4	50.87	122.0



$$228.3 = 348.0 - 119.7 \leq Y_{h(\text{new})} \leq 348.0 + 119.7 = 467.7$$

$$297.4 = 419.4 - 122.0 \leq Y_{h(\text{new})} \leq 419.4 + 122.0 = 541.4$$

# Simultaneous Prediction Intervals for New Observations, cont'd

- `Xh<-c(80,100)`
- `g<-length(Xh)`
- `alpha<-0.05`
- `CI.New <- predict.lm(fitreg,data.frame(Size = c(Xh)),`
- `se.fit = T, level = 1-alpha)`
- `M <-rbind(rep(qt(1 - alpha / (2*g), fitreg$df),g),`
- `rep(sqrt( g * qf( 1 - alpha, g, fitreg$df)),g))`
- `spred <- sqrt( CI.New$residual.scale^2 + (CI.New$se.fit)^2 ) # (2.38)`
- `pred.new <- t(rbind(`

`"Yh" = Xh,`

`"s.pred" = spred,`

`"fit" = CI.New$fit,`

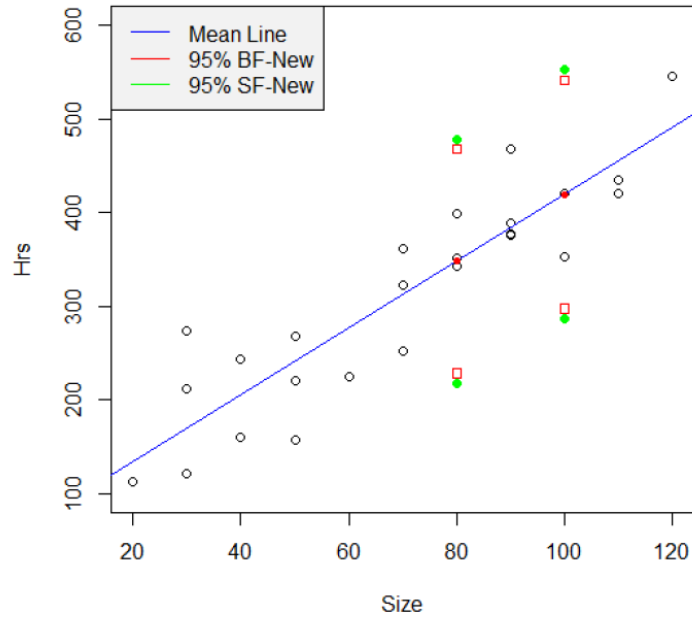
`"lower.B" = CI.New$fit - M[1,] * spred,`

`"upper.B" = CI.New$fit + M[1,] * spred,`

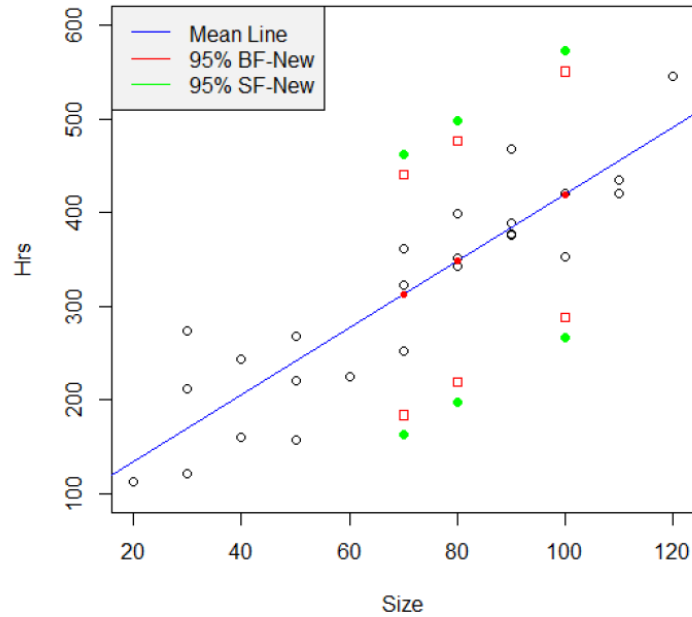
`"lower.S" = CI.New$fit - M[2,] * spred,`

`"upper.S" = CI.New$fit + M[2,] * spred))`

# Simultaneous Prediction Intervals for New Observations, cont'd



(a)  $g = 2$



(b)  $g = 3$

- Simultaneous prediction intervals for  $g$  new observations on  $Y$  at  $g$  different levels of  $X$  with a  $1 - \alpha$  family confidence coefficient are wider than the corresponding single prediction intervals.
- Note that both the B and S multiples for simultaneous predictions become larger as  $g$  increases. This contrasts with simultaneous estimation of mean responses where the B multiple becomes larger but not the W multiple. When  $g$  is large, both the B and S multiples for simultaneous predictions may become so large that the prediction intervals will be too wide to be useful.

# Regression through Origin

- Sometimes the regression function is known to be linear and to go through the origin at (0,0).

$$Y_i = \beta_i X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

- The regression function for model is  $E\{Y\} = \beta_1 X$ , straight line through origin with slope  $\beta_1$
- The point estimator (LSE and MLE):

$$b_1 = \frac{\sum X_i Y_i}{\sum X_i^2}$$

# Regression through Origin, cont'd

---

Estimate of	Estimated Variance	Confidence Limits
$\beta_1$	$s^2\{b_1\} = \frac{MSE}{\sum X_i^2}$	$b_1 \pm ts\{b_1\}$ (4.18)
$E\{Y_h\}$	$s^2\{\hat{Y}_h\} = \frac{X_h^2 MSE}{\sum X_i^2}$	$\hat{Y}_h \pm ts\{\hat{Y}_h\}$ (4.19)
$Y_{h(new)}$	$s^2\{pred\} = MSE \left( 1 + \frac{X_h^2}{\sum X_i^2} \right)$	$\hat{Y}_h \pm ts\{pred\}$ (4.20)

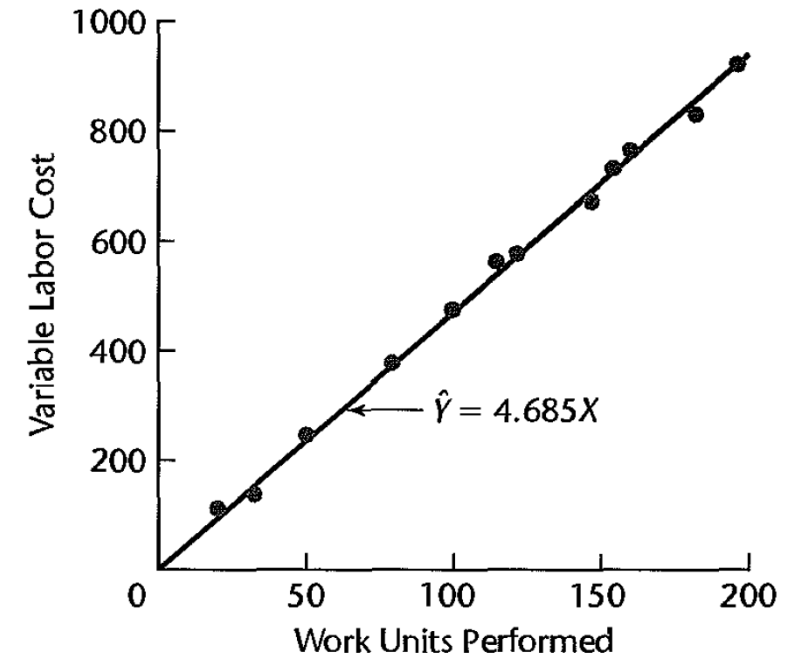
where:  $t = t(1 - \alpha/2; n - 1)$



# Regression through Origin, cont'd

## Example: Warehouse Data

	(1)	(2)	(3)	(4)	(5)	(6)
Warehouse	Work Units Performed	Variable Labor Cost (dollars)				
$i$	$X_i$	$Y_i$	$X_i Y_i$	$X_i^2$	$\hat{Y}_i$	$e_i$
1	20	114	2,280	400	93.71	20.29
2	196	921	180,516	38,416	918.31	2.69
3	115	560	64,400	13,225	538.81	21.19
...	...	...	...	...	...	...
10	147	670	98,490	21,609	688.74	-18.74
11	182	828	150,696	33,124	852.72	-24.72
12	160	762	121,920	25,600	749.64	12.36
Total	1,359	6,390	894,714	190,963	6,367.28	22.72



$$b_1 = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{894,714}{190,963} = 4.68527 \Rightarrow \hat{Y} = 4.68527X$$

# Regression through Origin, cont'd

- Important Caution for Using Regression through Origin:

$$Y_i = \beta_i X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

- The residuals **do not sum to zero** usually
- The only constraint

$$\sum X_i e_i = 0$$

- The residuals plot will usually not be balanced around the zero line.

# Regression through Origin, cont'd

- $SSE = \sum e_i^2$  may exceed  $SSTO = \sum (Y_i - \bar{Y})^2$  when
  - data form a curvilinear pattern or a
  - Linear pattern with an intercept away from the origin.
- If  $SSE \geq SSTO \Rightarrow R^2 \leq 0$  as  $1 - \frac{SSE}{SSTO} \leq 0$
- regression-through-the-origin model (4.10) needs to be evaluated for aptness.
  - it is generally a safe practice not to use regression-through-the-origin model and instead use the intercept regression model.

# Regression through Origin, cont'd

- $SSE = \sum e_i^2$  may exceed  $SSTO = \sum (Y_i - \bar{Y})^2$  when
- Some statistical packages calculate  $R^2$  for regression through the origin and hence will sometimes show a negative value for  $R^2$ . Other statistical packages calculate  $R^2$  using the total uncorrected sum of squares SSTOU (uncorrected SS).
- This procedure avoids obtaining a negative coefficient but lacks any meaningful interpretation.
- The ANOVA tables for regression through the origin shown in the output for many statistical packages are based on  $SSTOU = \sum Y_i^2$ ,  $SSRU = \sum \hat{Y}_i^2 = b_1^2 \sum X_i^2$ , and  $SSE = \sum (Y_i - b_1 X_i)^2$ . It can be shown that these sums of squares are additive:  $SSTOU = SSRU + SSE$ .

# Effects of Measurement Errors

- Measurement Errors in the response variable ( $Y$ ):
  - no new problems are created when these errors are uncorrelated and not biased(positive and negative measurement errors tend to cancel out).
  - absorbed in the model error term  $\varepsilon$
- Measurement Errors in the predictor variable ( $X$ ):
  - pressure in a tank
  - temperature in an oven
  - speed of a production line

# Effects of Measurement Errors, cont'd

Illustration:

- $X_i$ : the true age of the  $i$ th employee
- $X_i^*$ : the age reported by the employee of the employment record
- Measurement Error  $\delta_i$ :

$$\delta_i = X_i^* - X_i$$

- The regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

# Effects of Measurement Errors, cont'd

- Observe only  $X_i^*$ :

$$Y_i = \beta_0 + \beta_1(X_i^* - \delta_i) + \varepsilon_i \quad (4.23)$$

$$\Rightarrow Y_i = \beta_0 + \beta_1 X_i^* + (\varepsilon_i - \beta_1 \delta_i) \quad (4.24)$$

- $X_i^*$  is a random variable is correlated with  $\varepsilon_i - \beta_1 \delta_i$
- $\varepsilon_i - \beta_1 \delta_i$  is not independent of  $X_i^*$  ( $\because \delta_i = X_i^* - X_i$ )
- To determine the dependence: Conditions

$$E\{\delta_i\} = 0$$

$$E\{\varepsilon_i\} = 0$$

$$E\{\delta_i \varepsilon_i\} = 0$$

$$\Rightarrow E\{X_i^*\} = X_i$$

# Effects of Measurement Errors, cont'd

- $E\{\delta_i\} = 0$ : the reported ages would be unbiased estimates of the true ages
- $E\{\varepsilon_i\} = 0$ : the model error term have expectation 0
- $E\{\delta_i\varepsilon_i\} = 0$ : the measurement error  $\delta_i$  not be correlated with  $\varepsilon_i$

$$\Rightarrow \sigma\{\delta_i, \varepsilon_i\} = 0$$

- The covariance between  $X_i^*$  and  $\varepsilon_i - \beta_1\delta_i$ : **not zero**

$$\begin{aligned} E\{\varepsilon_i - \beta_1\delta_i\} &= 0; \\ \sigma\{X_i^*, \varepsilon_i - \beta_1\delta_i\} &= E\{\delta_i\varepsilon_i - \beta_1\delta_i^2\} = -\beta_1\sigma^2\{\delta_i\} \end{aligned}$$



# Effects of Measurement Errors, cont'd

Assume  $(Y, X^*) \sim$  bivariate normal distribution

- The conditional distribution of  $Y_i$  given  $X_i^*$ : normal and independent

$$\text{Conditional mean: } E\{Y_i | X_i^*\} = \beta_0^* + \beta_1^* X_i^*$$

$$\text{Conditional variance: } \sigma_{Y|X^*}^2$$

$$\Rightarrow \beta_1^* = \beta_1 [\sigma_X^2 / (\sigma_X^2 + \sigma_Y^2)] < \beta_1$$

If  $\sigma_Y^2$  is small relative to  $\sigma_X^2$ , the bias would be small; otherwise the bias may be substantial

# Berkson Model

- There is one situation where measurement errors in  $X$  are no problem.  
(Berkson Model)
  - The observation  $X_i^*$  is fixed quantity
  - The unobserved true value  $X$  is a random variable
  - The measurement error:  $\delta_i = X_i^* - X_i$  (no constraint on the relation between  $X_i^*$  and  $\delta_i$ )
  - $E\{\delta_i\} = 0$
  - Still applicable for the Berkson case:

$$Y_i = \beta_0 + \beta_1 X_i^* + (\varepsilon_i - \beta_1 \delta_i) \quad (4.28)$$

# Berkson Model, cont'd

- $E\{\varepsilon_i - \beta_1 \delta_i\} = 0; E\{\varepsilon_i\} = 0; E\{\delta_i\} = 0$

$$\Rightarrow \sigma\{X_i^*, \varepsilon_i - \beta_1 \delta_i\} = 0 \quad (\because X_i^* \text{ is a constant})$$

- Least squares procedures can be applied for the Berkson case: unbiased estimators  $b_0, b_1$ 
  - The error terms have expectation zero
  - The predictor variable is a constant, and the error terms are not correlated with it.

# Inverse Prediction

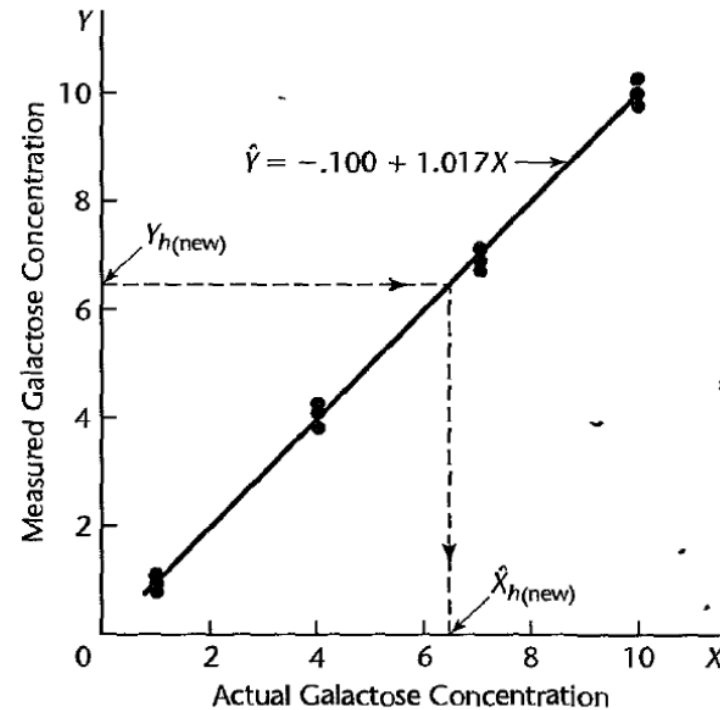


Figure : Scatter Plot and Fitted Regression Line-calibration Example.

# Inverse Prediction, cont'd

- A trade association analyst has regressed the selling price of a product ( $Y$ ) on its cost ( $X$ ). The selling price  $Y_{h(\text{new})}$  for another firm not belonging to the trade association is known, and it is desired to estimate the cost  $X_{h(\text{new})}$  for this firm.
- A regression analysis of the amount of decrease in cholesterol level ( $Y$ ) achieved with a given dosage of a new drug ( $X$ ) has been conducted. A physician is treating a new patient for whom the cholesterol level should decrease by the amount  $Y_{h(\text{new})}$ . It is desired to estimate the appropriate dosage level  $X_{h(\text{new})}$  to be administered to bring about the needed cholesterol decrease  $Y_{h(\text{new})}$ .

# Inverse Prediction, cont'd

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\hat{Y} = b_0 + b_1 X$$

$$\Rightarrow \hat{X}_{h(new)} = \frac{Y_{h(new)} - b_0}{b_1} \quad b_1 \neq 0$$

- $\hat{X}_{h(new)}$  is the MLE of  $X_{h(new)}$  for normal error regression model (2.1)
- Approximate  $1 - \alpha$  confidence limits for  $X_{h(new)}$ :

$$\hat{X}_{h(new)} \pm t(1 - \alpha/2; n - 2) s\{predX\}$$

$$s^2\{predX\} = \frac{MSE}{b_1^2} \left[ 1 + 1/n + \frac{(\hat{X}_{h(new)} - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$

# Inverse Prediction, cont'd

- A medical researcher studied a new, quick method for measuring low concentration of galactose (sugar) in the blood. Linear regression model (2.1) was fitted with the following results:

$$\begin{array}{llll} n = 12 & b_0 = -.100 & b_1 = 1.017 & MSE = .0272 \\ s\{b_1\} = .0142 & \bar{X} = 5.500 & \bar{Y} = 5.492 & \sum(X_i - \bar{X})^2 = 135 \\ & \hat{Y} = -.100 + 1.017X & & \end{array}$$

- $H_o: \beta_1 = 0$  vs  $H_a: \beta_1 \neq 0 \Rightarrow$  conclude  $H_a: \beta_1 \neq 0$
- $Y_{h(new)} = 6.52 \Rightarrow \hat{X}_{h(new)} = 6.509 \Rightarrow 6.13 \leq \hat{X}_{h(new)} \leq 6.89$

# Inverse Prediction, cont'd

- The inverse prediction problem is also known as a calibration problem since it is applicable when inexpensive, quick, and approximate measurements (Y) are related to precise, often expensive, and time-consuming measurements (X) based on n observations. The resulting regression model is then used to estimate the precise measurement  $X_{h(\text{new})}$  for a new approximate measurement  $Y_{h(\text{new})}$ .
- The approximate confidence interval (4.32) is appropriate if the quantity:

$$\frac{[t(1 - \alpha/2; n - 2)]^2 MSE}{b_1^2 \sum (X_i - \bar{X})^2} \text{ is small; less than 0.1}$$



# Choice of X Levels

- By experiment: X's are under control of the experimenter
- Among other things: consider
  1. How many levels of X should be investigated?
  2. What shall the two extreme levels be?
  3. How shall the other levels of X be spaced?
  4. How many observations should be taken at each level of X?
- No single answer: different purposes lead to different answers
  - estimate the intercept
  - predict new observations
  - estimate mean responses
  - other purposes

# Choice of X Levels, cont'd

The variances developed for the regression model(2.1):

$$\sigma^2\{b_0\} = \sigma^2\left[1/n + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right]$$

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \overline{X})^2}$$

$$\sigma^2\{\widehat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]$$

$$\sigma^2\{pred\} = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]$$

# Choice of X Levels, cont'd

- Estimate  $\beta_1$ :  $\min s^2\{b_1\} \Leftrightarrow \max \sum (X_i - \bar{X})^2 \Rightarrow$  using two levels of  $X$  at the two extremes and placing half of the observations at each of the two levels
- Estimate  $Y_h$ :  $\bar{X} = X_h$

Choose: (Cox advices)

- 2 levels: only interested in whether there is an effect and its direction
- 3 levels: goal is describing relation and any possible curvature
- 4 or more levels: further description of response curve and any potential non-linearity such as an asymptotic value