STAT 525

Chapter 2
Inferences in Gaussian Simple Linear Regression

Dr. Qifan Song

Testing for *normal* Linear Relationship

- Term $\beta_1 X_i$ defines the linear association
- Then test $H_0 : \beta_1 = 0$ for the existence of linear relationship
- Test requires
  - Test statistic
  - Sampling distribution of the test statistic

Note that a common t-test is of form

$$\frac{\text{point estimator} - E(\text{point estimator}|H_0)}{\text{std.err(point estimator)}}$$

Sampling Distribution of b_1

- b_1 is a linear estimator, i.e., a linear combination of Y_i's.

$$\begin{aligned} b_1 &= \sum_{i=1}^{n} \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\ &= \sum_{i=1}^{n} \frac{(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\ &= \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} Y_i \\ &= \sum_{i=1}^{n} k_i Y_i \end{aligned}$$

  where $k_i = (X_i - \overline{X})/\sum_{i=1}^{n}(X_i - \overline{X})^2$

- Note that $\sum k_i = 0$, $\sum k_i X_i = 1$.

- Under the normality assumption, b_1 also follows a normal distribution since it is a linear combination of normal r.v.s.

- It is sufficient to figure the first two moments of b_1:

$$\begin{aligned} E(b_1) &= \sum k_i E(Y_i) \\ &= \sum k_i(\beta_0 + \beta_1 X_i) \\ &= \beta_0 \sum k_i + \beta_1 \sum k_i X_i \\ &= \beta_1, \end{aligned}$$

$$\begin{aligned} Var(b_1) &= \sum k_i^2 Var(Y_i) \\ &= \sigma^2 \sum k_i^2 \\ &= \sigma^2 \sum \frac{(X_i - \overline{X})^2}{[\sum_{i=1}^n (X_i - \overline{X})^2]^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}. \end{aligned}$$

- Therefore

$$b_1 \sim N(\beta_1, \sigma^2 / \sum_{i=1}^n (X_i - \overline{X})^2)$$

t-test for $H_0 : \beta_1 = \beta_1^0$

- Consider statistics $\frac{b_1 - \beta_1^0}{se(b_1)}$, where standard error of b_1 means a estimation for the $\sqrt{Var(b_1)}$.
- $se^2(b_1) = \widehat{Var(b_1)} = \hat{\sigma}^2 / \sum_{i=1}^n (X_i - \overline{X})^2 = s^2 / \sum_{i=1}^n (X_i - \overline{X})^2$, that is to replace the unknown σ^2 by its unbiased estimator $s^2 = MSE = \sum (Y_i - \hat{Y}_i)^2/(n-2)$,
- The test statistics is

$$\frac{b_1 - \beta_1^0}{\sqrt{Var(b_1)}} \div \sqrt{\frac{se^2(b_1)}{Var(b_1)}} = \frac{b_1 - \beta_1^0}{\sqrt{Var(b_1)}} \div \sqrt{\frac{s^2}{\sigma^2}}$$

- Under normal assumption:
  - 1st term is $N(0,1)$
  - $(n-2)s^2/\sigma^2 \sim \chi^2_{n-2}$
  - b_1 and s^2 are independent

  Hence $\frac{b_1 - \beta_1^0}{se(b_1)} \sim t_{n-2}$

t-test for $H_0 : \beta_1 = \beta_1^0$

t-test statistics $t^* = \frac{b_1 - \beta_1^0}{se(b_1)}$

- Under H_0, $t^* \sim t_{n-2}$
- Pisa tower example:
  - $H_0 : \beta_1 = 0$(or ≤ 0, ≥ 0)
  - $t^* = \frac{9.31868-0}{0.30991} = 30.0690$
  - p-value $P(|t_{13-2}| \geq |t^*|) = 6.5024 \times 10^{-12} (< .0001)$
  - Reject null since p-value is smaller than $\alpha = 0.05$, we concludes that there is statistical evidence to support the existence of linear relationship *if the Gaussian linear regression model is valid*.

power of t-test

t-test statistics $t^* = \frac{b_1 - \beta_1^0}{se(b_1)}$

- Power $=P$(reject $H_0 | H_a$ is true)
- Under H_a, $t^* \sim t_{n-2}(\delta)$ where δ is the non-centrality parameter

$$\delta = \frac{\beta_1 - \beta_1^0}{\sqrt{\sigma^2 / \sum (X_i - \overline{X})^2}}$$

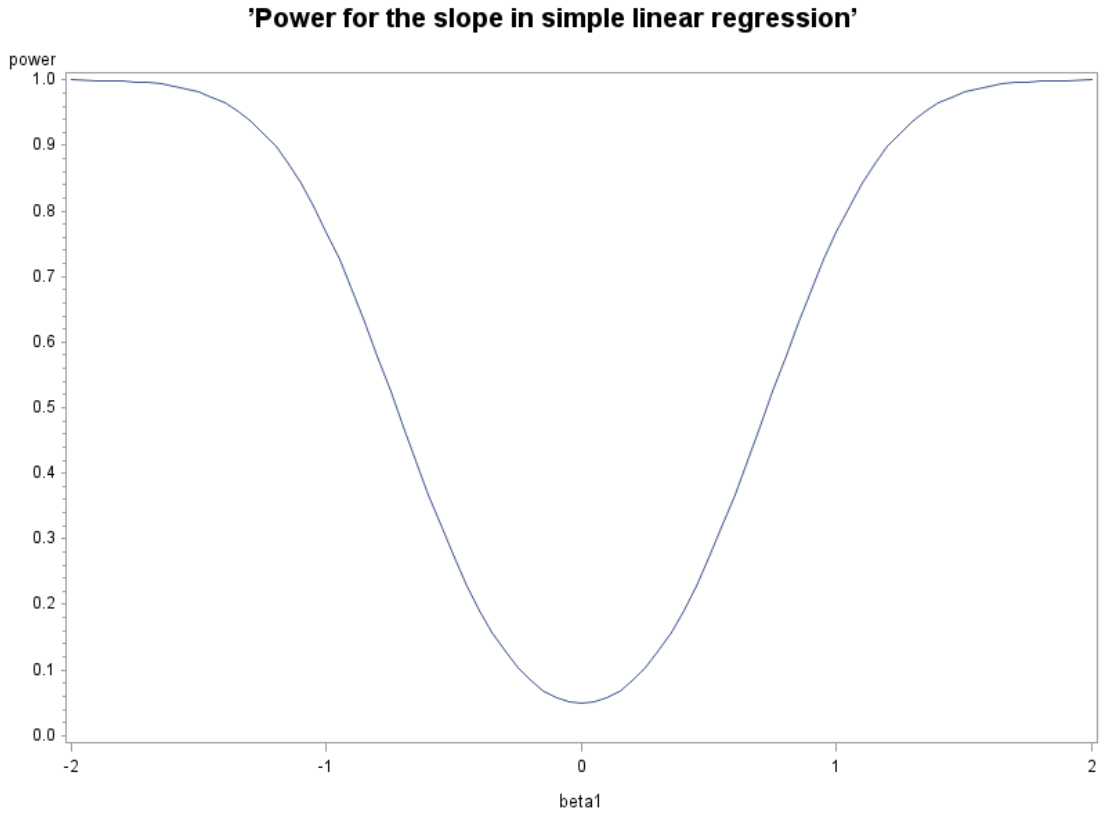
- Power$=P(|t_{n-2}(\delta)| > t(1-\alpha/2, n-2))$ depends on n, δ and α
- Given n, X_i's, α, σ and β_1^0, power is a function of true parameter value β_1 (power curve)

SAS Code for Toluca Company Example (p. 51)

- In practice, one usually needs to estimate the σ^2.
- `tinv` computes the cutoff of the t-distribution such that the area to the left of the cutoff is $1 - \alpha/2$
- `probt` computes the area to the left of the cutoff `t_c`

```
DATA a2;
   n=25; sig2=2500; ssx=19800; alpha=.05;
   sig2b1=sig2/ssx; df=n-2;
   DO beta1=-2.0 TO 2.0 BY .05;
      delta=beta1/sqrt(sig2b1);
      t_c=tinv(1-alpha/2,df);
      power=1-probt(t_c,df,delta)+probt(-t_c,df,delta);
      OUTPUT;
   END;
```

```
/*Generate a power curve based on the data set a2; */
TITLE1 'Power for the slope in simple linear regression';
SYMBOL1 V=NONE I=JOIN;
PROC GPLOT DATA=a2; PLOT power*beta1/FRAME; RUN; QUIT;
```



Alternatively, One can also calculate the power as a function of n (sample size), given a fixed estimated true β_1, to perform a sample size calculation.

Inferences Concerning β_0

- Not always of interests.

Sampling Distribution of b_0

- b_0 is also a linear combination of Y_i's, $b_0 = \sum k_i Y_i$, where

$$k_i = \frac{1}{n} - \frac{\overline{X}(X_i - \overline{X})}{\sum (X_i - \overline{X})^2}, \quad \sum k_i = 1, \quad \sum k_i X_i = 0.$$

- By normal assumption, $b_0 \sim N(E(b_0), Var(b_0))$, where

$$\begin{aligned} E(b_0) &= \sum k_i E(Y_i) = \sum k_i(\beta_0 + \beta_1 X_i) = \beta_0, \\ Var(b_0) &= \sum k_i^2 Var(Y_i) = \sigma^2 \sum k_i^2 \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right]. \end{aligned}$$

t-test for $H_0 : \beta_0 = \beta_0^0$

- Consider statistics $\frac{b_0 - \beta_0^0}{se(b_0)}$.
- $se^2(b_0) = \widehat{Var(b_0)} = \hat{\sigma}^2[1/n + \overline{X}^2 / \sum_{i=1}^n (X_i - \overline{X})^2] = s^2[1/n + \overline{X}^2 / \sum_{i=1}^n (X_i - \overline{X})^2]$
- The test statistics is

$$\frac{b_0 - \beta_0^0}{\sqrt{Var(b_0)}} \div \sqrt{\frac{se^2(b_0)}{Var(b_0)}} = \frac{b_0 - \beta_0^0}{\sqrt{Var(b_0)}} \div \sqrt{\frac{s^2}{\sigma^2}}$$

- Under normal assumption:
  - 1st term is $N(0, 1)$
  - $(n-2)s^2/\sigma^2 \sim \chi^2_{n-2}$
  - b_0 and s^2 are independent

Hence $\frac{b_0 - \beta_0^0}{se(b_0)} \sim t_{n-2}$

t-test for $H_0 : \beta_0 = \beta_0^0$

t-test statistics $t^* = \frac{b_0 - \beta_0^0}{se(b_0)}$

- Under H_0, $t^* \sim t_{n-2}$
- Pisa tower example:
  - $H_0 : \beta_0 = 0$(or ≤ 0, ≥ 0)
  - $t^* = \frac{-61.12-0}{25.13} = -2.43$
  - p-value $P(|t_{13-2}| \geq |t^*|) = 0.0333$
  - Reject null since p-value is smaller than $\alpha = 0.05$, we concludes that is statistical evidence to support that the intercept is not zero.

Confidence Intervals for β_1 and β_0

- Derive the confidence interval formula from the result that

$$\frac{b_i - \beta_i}{se(b_i)} \sim t_{n-2},\ i = 0, 1$$

- CI: $b_i \pm t(1 - \alpha/2, n - 2)se(b_i)$
- Reject $H_0 : \beta_i = \beta_i^0$ if β_i^0 is not in CI.
- These CIs generated in SAS with `clb` option.

Comments

- When errors not normal, procedures are generally reasonable approximations (Bootstrapping as alternative approach)
- Procedures can be modified for one-sided test / confidence intervals
- To obtain an accurate interval estimation, at design stage, choose X_i such that
  - $\sum(X_i - \overline{X})^2$ is large $\rightarrow$ smaller margin of error for β_1
  - $\sum(X_i - \overline{X})^2$ is large and $|\overline{X}|$ is small $\rightarrow$ smaller margin of error for β_0

Interval Estimation of $E(Y_h)$

- Often interested in estimating the mean response for particular X_h, i.e., the parameter of interests is $E(Y_h) = \beta_0 + \beta_1 X_h$.
- Point estimation is $\hat{Y}_h = b_0 + b_1 X_h$.
- Derive the sampling distribution of $b_0 + b_1 X_h$ in order to make test and CI.
  - $\hat{Y}_h = \sum k_i Y_i$ where $k_i = \frac{1}{n} + \frac{(X_h - \overline{X})(X_i - \overline{X})}{\sum (X_i - \overline{X})^2}$
  - $E(\hat{Y}_h) = \beta_0 + \beta_1 X_h$
  - $Var(\hat{Y}_h) = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right]$.
  - $se^2(\hat{Y}_h) = s^2 \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right]$.
  - Test: $(\hat{Y}_h - \text{null value})/se(\hat{Y}_h)$; CI: $\hat{Y}_h \pm t(1-\alpha/2, n-2) se(\hat{Y}_h)$

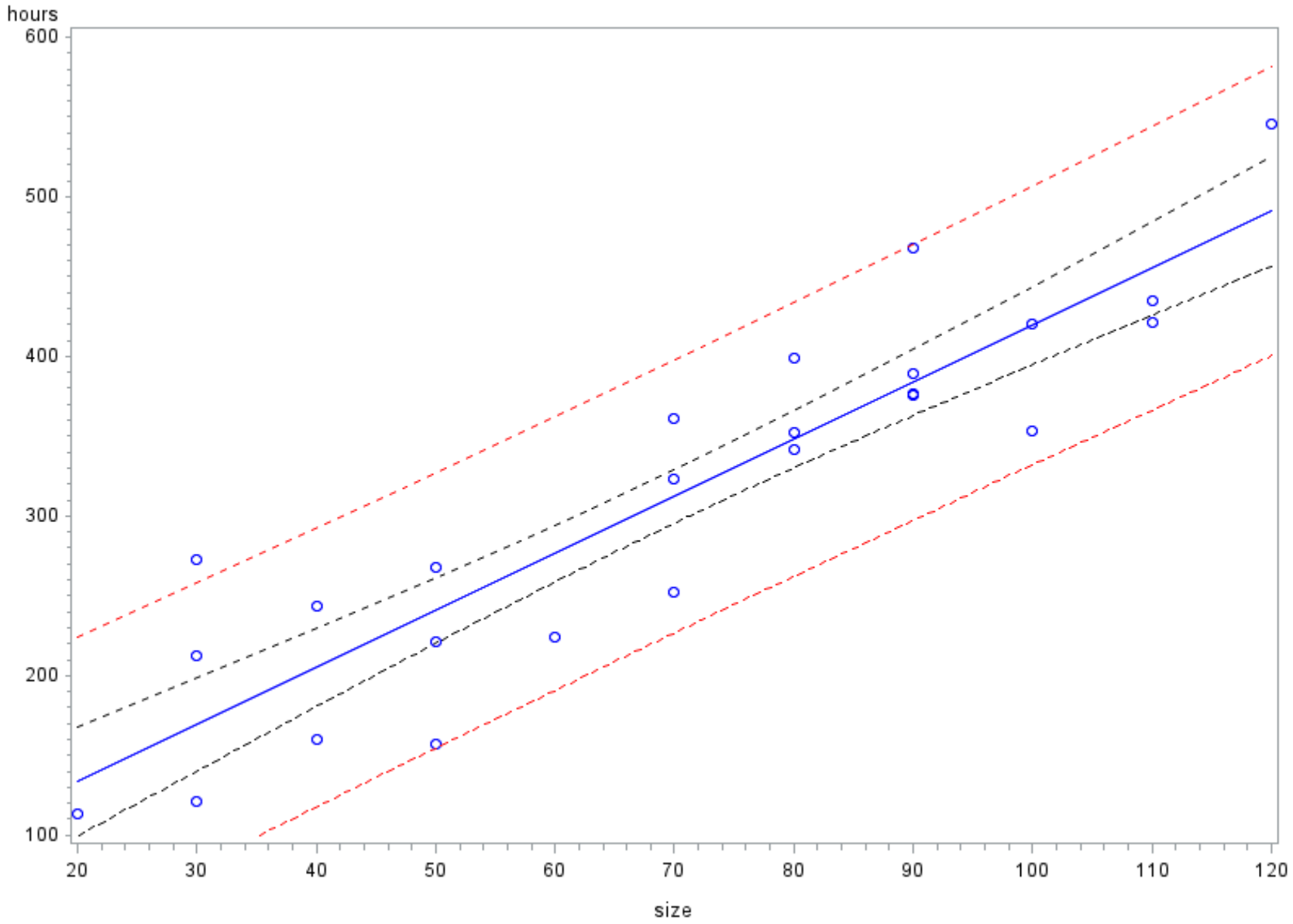
Interval Prediction of $Y_{h(new)}$

- Predicting future observation $Y_{h(new)} = E[Y_h] + \varepsilon_{h(new)}$
- Not a confidence interval for a parameter, but a prediction interval for a unknown r.v., i.e., $P(L < Y_{h(new)} < U) = 1 - \alpha$
- Comparing with CI of $E[Y_h]$, one need to take account of future error $\varepsilon_{h(new)}$.
  - $E(b_0 + b_1 X_h) = E(Y_{h(new)})$
  - $Var(b_0 + b_1 X_h - Y_{h(new)}) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right]$
  - $se^2(b_0 + b_1 X_h - Y_{h(new)}) = s^2 \left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right]$
  - PI: $b_0 + b_1 X_h \pm t(1 - \alpha/2, n-2) \sqrt{s^2 \left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right]}$

Example: Toluca Company (p. 19)

```
/* read data */
DATA a1;
   INFILE 'C:\Textdata\CH01TA01.txt';
   INPUT size hours;
/* add size 65 and 100 for prediction */
DATA a2; size=65; OUTPUT;
         size=100; OUTPUT;
DATA a3; SET a1 a2;
/* plot predicted confidence intervals */
SYMBOL1 V=CIRCLE I=RLCLM90 CI=BLUE CO=BLACK;
SYMBOL2 V=CIRCLE I=RLCLI90 CI=BLUE CO=RED;
PROC GPLOT DATA=a1;
   PLOT hours*size=1 hours*size=2 / OVERLAY;
RUN;
```

Scatterplot



```
/* calculate the actual CI limits */
PROC REG DATA=a3;
   MODEL hours=size / CLM CLI ALPHA=.10;
   ID size;
RUN;
```

Dependent Variable: hours

Analysis of Variance

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
Dependent Mean	312.28000	Adj R-Sq	0.8138
Coeff Var	15.63447		

Parameter Estimates

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

Obs	size	Dep Var hours	Predicted Value	Std Error Mean Predict	90% CL Mean	
		Output Statistics				
1	80	399.0000	347.9820	10.3628	330.2215	365.7425
2	30	121.0000	169.4719	16.9697	140.3880	198.5559
3	50	221.0000	240.8760	11.9793	220.3449	261.4070
4	90	376.0000	383.6840	11.9793	363.1530	404.2151
5	70	361.0000	312.2800	9.7647	295.5446	329.0154
6	60	224.0000	276.5780	10.3628	258.8175	294.3385
7	120	546.0000	490.7901	19.9079	456.6706	524.9096
8	80	352.0000	347.9820	10.3628	330.2215	365.7425
9	100	353.0000	419.3861	14.2723	394.9251	443.8470
10	50	157.0000	240.8760	11.9793	220.3449	261.4070
11	40	160.0000	205.1739	14.2723	180.7130	229.6349
12	70	252.0000	312.2800	9.7647	295.5446	329.0154
22	90	468.0000	383.6840	11.9793	363.1530	404.2151
23	40	244.0000	205.1739	14.2723	180.7130	229.6349
24	80	342.0000	347.9820	10.3628	330.2215	365.7425
25	70	323.0000	312.2800	9.7647	295.5446	329.0154
26	65	.	294.4290	9.9176	277.4315	311.4264
27	100	.	419.3861	14.2723	394.9251	443.8470

Obs	size	Dep Var hours	Predicted Value	Std Error Mean Predict	90% CL Predict	
1	80	399.0000	347.9820	10.3628	262.4411	433.5230
2	30	121.0000	169.4719	16.9697	80.8847	258.0591
3	50	221.0000	240.8760	11.9793	154.7171	327.0348
4	90	376.0000	383.6840	11.9793	297.5252	469.8429
5	70	361.0000	312.2800	9.7647	226.9460	397.6140
6	60	224.0000	276.5780	10.3628	191.0370	362.1189
7	120	546.0000	490.7901	19.9079	400.4244	581.1558
8	80	352.0000	347.9820	10.3628	262.4411	433.5230
9	100	353.0000	419.3861	14.2723	332.2072	506.5649
10	50	157.0000	240.8760	11.9793	154.7171	327.0348
11	40	160.0000	205.1739	14.2723	117.9951	292.3528
12	70	252.0000	312.2800	9.7647	226.9460	397.6140
22	90	468.0000	383.6840	11.9793	297.5252	469.8429
23	40	244.0000	205.1739	14.2723	117.9951	292.3528
24	80	342.0000	347.9820	10.3628	262.4411	433.5230
25	70	323.0000	312.2800	9.7647	226.9460	397.6140
26	65	.	294.4290	9.9176	209.0432	379.8148
27	100	.	419.3861	14.2723	332.2072	506.5649

Output Statistics

Confidence Band for Response Means

- Consider the entire regression line
- Want to define a likely region within which this unknown real line lies
- Rigorously, $P(L(x) < \beta_0 + \beta_1 x < U(x) \text{ for all } x) \geq 1 - \alpha$
- Replace $t(1-\alpha/2, n-2)$ with Working-Hotelling value in each confidence interval

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

- Boundary values define a hyperbola
- Will be discussed more in Chapter 4

Comments

- The band is the narrowest at $\overline{X}$
- Theory comes from fact that (b_0, b_1) is multivariate normal
  - Joint confidence region for (b_0, b_1) is an ellipse.
  - $Cov(b_0, b_1) = -\overline{X} Var(b_1)$
  - It can be shown that

$$\max_x \left[\frac{b_0 + b_1 x - (\beta_0 + \beta_1 x)}{se(\hat{Y}_h)(x)} \right]^2 \sim 2F_{2,n-2}$$

- Band width at $X_h >$ individual CI width of $E[Y_h]$. Joint CI is always larger than the marginal CI.

SAS for Confidence Band

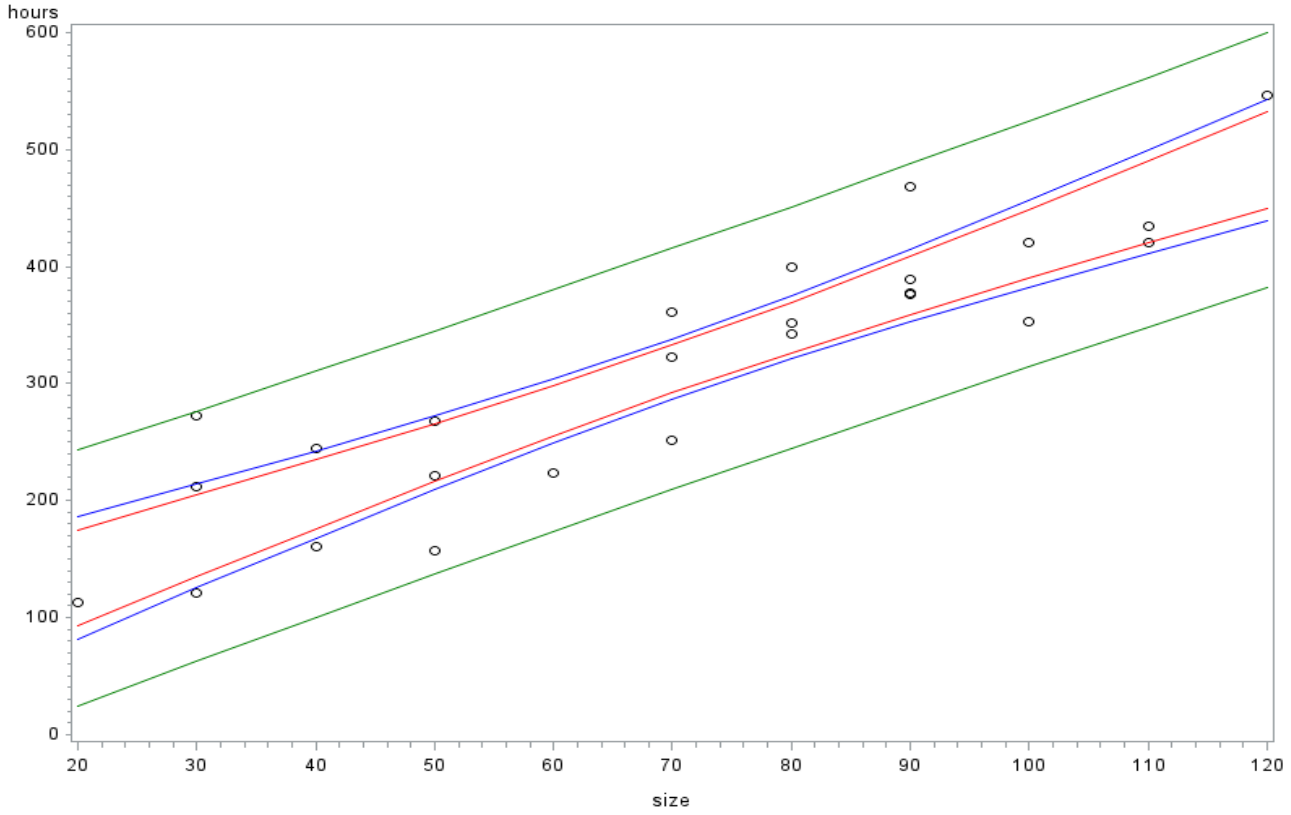
```
proc reg data=a1;
   model hours=size/clm cli alpha=0.05;
   output out=a2 p=predicted stdp=stdp uclm=uclm lclm=lclm ucl=ucl lcl=lcl;
   id size;
run;

/* Calculate Working-Hotelling band */
data a3; set a2;
   whl = predicted - sqrt(2*FINV(1 - 0.05, 2, 25-2))*stdp;
   whu = predicted + sqrt(2*FINV(1 - 0.05, 2, 25-2))*stdp;
run;

proc sort data=a3 out=a4; by size; run;
/* plot comparing the three confidence bands */
symbol1 v=circle i=none c=black; symbol2 v=none i=join c=green;
symbol3 v=none i=join c=red; symbol4 v=none i=join c=blue;
proc gplot data=a4;
   plot hours*size=1 ucl*size=2 lcl*size =2 uclm*size=3
   lclm*size=3 whl*size=4 whu*size=4 / overlay;
run;
```

Comment: `p`: predicted values for the mean; `stdp`: standard error of the mean predicted value; `uclm/lclm`: upper/lower bounds of the CI for the mean; `ucl/lcl`: upper/lower bounds of the CI for a new value

Confidence Band for the Toluca example



- Blue – 95% confidence band; widest when $X_h - \overline{X}$ is large
- Red – 95% confidence interval for the mean; always narrowest
- Green – 95% confidence interval for the individual prediction; widest when $X_h - \overline{X}$ is small

ANOVA Approach to Regression

- A second way to test for linear association
- Equivalent to the t-tests in simple linear regression
- Will have a different use in multiple regression

Partitioning Sums of Squares

- Organizes results arithmetically
- The total sum of squares in Y is defined

$$\text{SSTO} = \sum (Y_i - \overline{Y})^2$$

- Can partition the total sum of squares into
  - Model (explained by regression)
  - Error (unexplained / residual)

$$\begin{aligned} \sum (Y_i - \overline{Y})^2 &= \sum (Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2 \\ &= \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \overline{Y})^2 \\ \text{SSTO} &= \quad \text{SSE} \quad + \quad \text{SSR} \end{aligned}$$

Total Sum of Squares

- If we ignored X, the sample mean $\overline{Y}$ would be the best linear unbiased predictor for the model

$$Y_i = \beta_0 + \varepsilon_i = \mu + \varepsilon_i$$

- SSTO is the sum of squared deviations for this estimated model
  - SAS calls it "Corrected Total" sum of squares
  - "Corrected" means that the sample mean has been subtracted off before squaring
  - "Uncorrected total" sum of squares would be $\sum Y_i^2$
- SSTO has $n-1$ degrees of freedom because we replace β_0 with $\overline{Y}$, and statistically, SSTO $\sim \sigma^2 \chi^2_{n-1}(\delta)$ for some δ.
- The total mean square is SSTO$/(n-1)$ and represents an unbiased estimate of σ^2 under the above model

Model (or Regression) Sum of Squares

$$\text{SSR} = \sum(\widehat{Y}_i - \overline{Y})^2$$

- Can also express

$$\begin{aligned} \text{SSR} &= \sum(\widehat{Y}_i - \overline{Y})^2 \\ &= \sum(b_0 + b_1 X_i - b_0 - b_1 \overline{X})^2 \\ &= b_1^2 \sum(X_i - \overline{X})^2 \end{aligned}$$

- Degrees of freedom is 1 due to the slope estimation, and $\text{SSR} \sim \sigma^2 \chi_1^2(\delta)$ for some δ.

- SSR large when $\widehat{Y}_i$'s are different from $\overline{Y}$ (in other words, when there is a linear trend)

Error Sum of Squares

- Error sum of squares is equal to the sum of squared residuals

$$\text{SSE} = \sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2$$

- Degrees of freedom is $n - 2$ due to using (b_0, b_1) in place of (β_0, β_1), and SSE $\sim \sigma^2 \chi^2_{n-2}$
- SSE large when absolute values of residuals are large, which implies Y_i's vary substantially around line
- The MSE = SSE/$(n-2)$ and represents an unbiased estimate of σ^2 when taking X into account

ANOVA Table

- Table puts these all together

Source of Variation	df	SS	MS
Regression (Model)	1	$b_1^2 \sum (X_i - \overline{X})^2$	SSR/1
Error	n-2	$\sum (Y_i - \hat{Y}_i)^2$	SSE/(n-2)
Total	n-1	$\sum (Y_i - \overline{Y})^2$	

Expected Mean Squares

- All means squares are random variables
- Already showed $E(\text{MSE}) = \sigma^2$
- What about the MSR?

$$\begin{aligned} E(\text{MSR}) &= E(b_1^2 \sum (X_i - \overline{X})^2) \\ &= E(b_1^2) \sum (X_i - \overline{X})^2 \\ &= (Var(b_1) + \beta_1^2) \sum (X_i - \overline{X})^2 \\ &= \sigma^2 + \beta_1^2 \sum (X_i - \overline{X})^2 \end{aligned}$$

- If $\beta_1 = 0$, MSR is also an unbiased estimate of σ^2

F Test

- Can use this structure to test $H_0 : \beta_1 = 0$
- Consider

$$F^* = \frac{\text{MSR}}{\text{MSE}}$$

- If $\beta_1 = 0$, then F^* should be near one, since both denominator and numerator are of mean σ^2.
- Need sampling distribution of F^* under H_0 to obtain p-value.
- By Cochran's Theorem (pg 70)

$$\begin{aligned} F^* &= \frac{\frac{\text{SSR}}{\sigma^2}}{1} \div \frac{\frac{\text{SSE}}{\sigma^2}}{n-2} \\ F^* &\sim \frac{\frac{\chi^2_1}{1}}{\frac{\chi^2_{n-2}}{n-2}} \\ &\sim F_{1,n-2} \end{aligned}$$

- When H_0 is false, MSR $>$ MSE, and F^* tends to be large
- p-value $= Pr(F(1, n-2) > F^*)$
- Reject when F^* large, p-value small
- Recall t-test for $H_0 : \beta_1 = 0$
- Can show $t^2_{n-2} \sim F_{1,n-2}$
- Obtain exactly the same result (p-value)

Example: Toluca Company

```
data a1;
   infile 'C:\Textdata\CH01TA01.txt';
   input size hours;

proc reg data=a1;
   model hours=size;
   id size;
run;
```

Dependent Variable: hours

Analysis of Variance

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
Dependent Mean	312.28000	Adj R-Sq	0.8138
Coeff Var	15.63447		

Parameter Estimates

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 $10.292^2 \approx 105.88$

General Linear Test

- A third way to test for linear association
- Consider **two** models
  - Full model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - Reduced model: $Y_i = \beta_0 + \varepsilon_i$
- Will compare models using SSE's
  - Error sum of squares of the full model will be labeled SSE(F)
  - Error sum of squares of the reduced model will be labeled SSE(R)
- Note: SSTO is the same under each model

- Reduced model corresponds to $H_0 : \beta_1 = 0$
- Can be shown that SSE(F) $\leq$ SSE(R)
- Idea: more parameters provide better fit
- If SSE(F) is not much smaller than SSE(R), full model doesn't better explain Y.

$$\begin{aligned} F^* &= \frac{(\text{SSE(R)} - \text{SSE(F)})/(df_R - df_F)}{\text{SSE(F)}/df_F} \\ &= \frac{(\text{SSTO} - \text{SSE})/1}{\text{SSE(F)}/(n-2)} \end{aligned}$$

- Same test as before, but will have a more general use in multiple regression

Descriptive Measures of Linear Association

- The degree of "linear association", beyond its existence, is often the time of interest
- In simple linear regression,
  - Coefficient of determination R^2
  - Estimated Pearson's correlation coefficient (under random design, see slide 3-41) r

Coefficient of Determination

- Defined as the proportion of total variation explained by the model utilizing X

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

- $0 \leq R^2 \leq 1$
- High R^2 does not necessarily mean that
  - we can make useful predictions
  - regression line is a good fit
- Low R^2 does not necessarily mean that
  - X and Y are not related
- See page 75 for limitations of R^2

Fixed vs Random Design

- Have assumed X_i's are known constants, i.e., fixed design
- Statistical inferences consider repeated sampling with fixed X values
- What if this assumption is not appropriate, i.e., X_i's are random?
- Two possible ways to studies the association:
  - correlation model (joint distribution of X and Y)
  - regression model (conditional distribution of Y given X)

Bivariate Normal Distribution

- Assume that Y and X follow the bivariate normal distribution
- Distribution requires five parameters
  - μ_Y and σ_Y are the mean and std dev of Y
  - μ_X and σ_X are the mean and std dev of X
  - ρ is the coefficient of correlation between Y and X
- Bivariate normal density and marginal distributions given on page 79
- Marginal distributions are normal
- Conditional distributions are also normal

Pearson's Correlation Coefficient

- A number between -1 and 1 which measures the strength of the **linear** relationship between two variables, i.e.,

$$\rho = corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

- For the normal case, $\rho = 0$ indicates independence.

- ρ can be estimated by

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}} = b_1 \sqrt{\frac{\sum(X_i - \overline{X})^2}{\sum(Y_i - \overline{Y})^2}}$$

  - sign of r is the sign of the regression slope

- Trivially,

$$r^2 = b_1^2 \frac{\sum(X_i - \overline{X})^2}{\sum(Y_i - \overline{Y})^2} = \frac{\text{SSR}}{\text{SSTO}} = R^2$$

  - Relationship not true in multiple regression

Inference on ρ_{12}

- Interest in testing the existence of association: $H_0 : \rho = 0$
- Test statistic is

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

- Can also form CI using Fisher z transformation or large sample approximation

$$\frac{1}{2}\log\frac{1+r}{1-r} \sim N\left(\frac{1}{2}\log\frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right)$$

- If X and Y nonnormal, can use Spearman's correlation coefficient (p. 87)

Conditional Distribution

- Consider the distribution of Y_i given $X_i = x$
  - Can show the distribution is normal
  - The mean can be expressed

$$\left(\mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X}\right) + \rho \frac{\sigma_Y}{\sigma_X} x$$

  - With constant variance $\sigma_Y^2(1 - \rho^2)$
- equivalent to normal error regression model
- $\beta_1 \equiv \rho_{12}(\sigma_Y/\sigma_X)$; t-test for $\beta_1 = 0$ and t-test for $\rho = 0$ are actually the same test.

Regression model under random design

- We can still implement regression model for random design problem, and all previous regression results hold if:
  - The conditional distributions of Y_i given X_i are normal and independent with conditional means $\beta_0 + \beta_1 X_i$ and conditional variance σ^2
  - The X_i's are independent and distribution of X, $g(\cdot)$, does not involve the parameters β_0, β_1, and σ^2
- Compared with correlation analysis, regression analysis avoid modeling marginal distribution of X.

Chapter Review

- Inference concerning β_1
- Inference concerning β_0
- Inference concerning prediction
- Analysis of Variance Approach to Regression
  - Partitioning sums of squares
  - Degrees of freedom
  - Expected mean squares
- General linear test
- R^2 and the correlation coefficient
- What if X random variable?