STAT 525 FALL 2018

Chapter 2
Inferences in Simple Linear Regression

Professor Min Zhang

Testing for Linear Relationship

- Term $\beta_1 X_i$ defines linear relationship
- Will then test $H_0 : \beta_1 = 0$
- Test requires
  - Test statistic
  - Sampling distribution of the test statistic

Note: form of test statistic is often

$$\frac{\text{point estimate} - E(\text{point estimate}|H_0)}{s(\text{point estimate})}$$

Sampling Distribution of b_1

- Express b_1 as a linear combination of Y_i
- Can show that

$$\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})=\sum_{i=1}^{n}(X_i-\overline{X})Y_i$$

- Therefore rewrite

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

where k_i fixed constants where $\sum k_i = 0$ and $\sum k_i X_i = 1$

- Since $b_1 = \sum_{i=1}^{n} k_i Y_i$, we can analytically derive its distribution:
  - Normal since linear combination of $i.i.d.$ Y_i's

$$
\begin{aligned}
E(b_1) &= E(\sum k_i Y_i) \\
&= \sum k_i E(Y_i) \\
&= \beta_0 \sum k_i + \beta_1 \sum k_i X_i \\
&= 0 + \beta_1
\end{aligned}
$$

$$
\begin{aligned}
\mathsf{Var}(b_1) &= \mathsf{Var}(\sum k_i Y_i) \\
&= \sum k_i^2 \mathsf{var}(Y_i) \\
&= \sigma^2 \sum k_i^2 \\
&= \sigma^2 / \sum (X_i - \overline{X})^2
\end{aligned}
$$

Test Statistics $\frac{b_1-\beta_1}{s\{b_1\}}$

- An estimator of Var(b_1) is obtained by replacing σ^2 by its unbiased estimator $MSE = \sum (Y_i - \hat{Y}_i)^2/(n-2)$,

$$s^2\{b_1\} = MSE/\sum (X_i - \overline{X})^2$$

- Rewrite as

$$\frac{b_1 - \beta_1}{\sigma\{b_1\}} \div \frac{s\{b_1\}}{\sigma\{b_1\}}$$

- Since Y_i's are $i.i.d.$ normal
  - b_1 is normal $\longrightarrow$ 1st term is standard normal
  - The quantity $\sum (Y_i - \hat{Y}_i)^2/\sigma^2 \sim \chi^2_{n-2}$
  - The variable $s^2\{b_1\}/\sigma^2\{b_1\} \sim \chi^2_{n-2}/(n-2)$
  - The variable $s\{b_1\}/\sigma\{b_1\}$ is independent of b_1

$$\Longrightarrow Test\ Statistics: \quad \frac{b_1 - \beta_1}{s\{b_1\}} \sim t_{n-2}$$

Steps of Hypothesis Test

- $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$ (or $\beta_1 > 0$ or $\beta_1 < 0$)
- Compute the test statistic. In "Leaning Tower of Pisa":

$$t^\star = \frac{b_1 - 0}{s(b_1)} = \frac{9.31868 - 0}{0.30991} = 30.0690$$

- Compute p-value using sampling distribution

$$P(|t_{13-2}| \geq |t^\star|) = 6.5024 \times 10^{-12} (< .0001)$$

  - The above is for two-sided test! What about one-sided test?

- Compare to α
- Reject H_0 at α (usually $= .05$) level, evidence suggests a positive linear relationship

Power of Hypothesis Test

- Power = P{ reject $H_0 : \beta_1 = \beta_1^{H_0}$ | $H_a : \beta_1 \neq \beta_1^{H_0}$ }

- If H_a is true, the test statistic

$$t^\star \sim t_{n-2}(\delta)$$

  where δ is the non-centrality parameter

$$\delta = \frac{\beta_1 - \beta_1^{H_0}}{\sigma(b_1)} = \frac{\beta_1 - \beta_1^{H_0}}{\sqrt{\sigma^2 / \sum (X_i - \bar{X})^2}}$$

- Power calculation requires knowledge of δ, n, and also α.

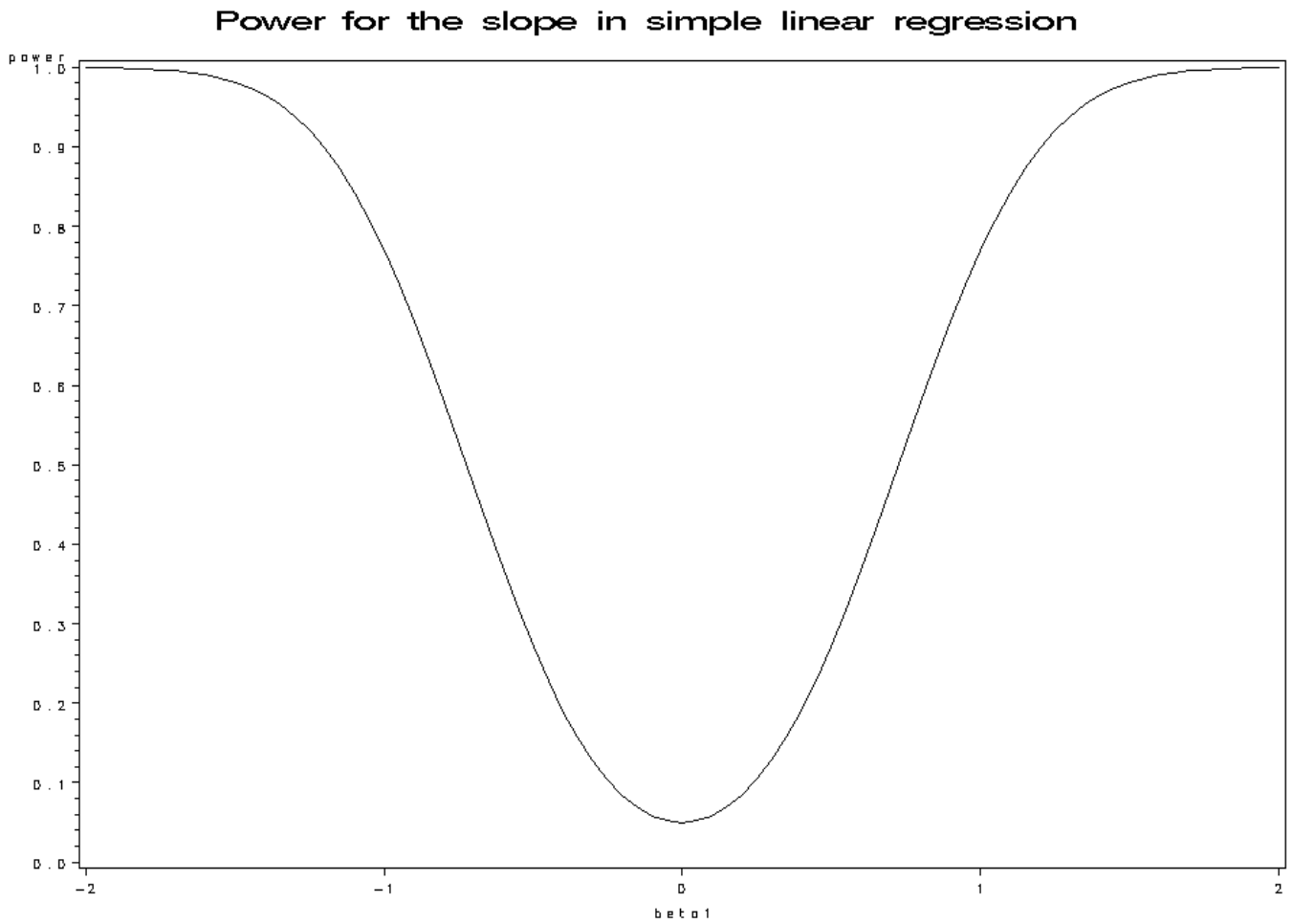
- Can calculate power for a range of input values.

SAS Code for Toluca Company Example (p. 51)

- enter information necessary to compute noncentrality parameter as in example.
- `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;
```



Inferences Concerning β_0

- Test of intercept is usually not of interest

Sampling Distribution of b_0

- Rewrite $b_0 = \sum k_i Y_i$ where

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

- Can now describe distribution of b_0
  - Normal since linear combination of $i.i.d.$ Y_i's

$$\begin{aligned} E(b_0) &= E(\sum k_i Y_i) = \sum k_i E(Y_i) = \sum k_i \beta_0 + \sum k_i \beta_1 X_i = \beta_0 + 0 \\ \text{Var}(b_0) &= \text{Var}(\sum k_i Y_i) = \sum k_i^2 \text{var}(Y_i) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right] \end{aligned}$$

Test Statistics $\frac{b_0-\beta_0}{s\{b_0\}}$

- An estimator of Var(b_0) is obtained by replacing σ^2 by its unbiased estimator $MSE = \sum (Y_i - \hat{Y}_i)^2/(n-2)$,

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

- Rewrite as

$$\frac{b_0 - \beta_0}{\sigma\{b_0\}} \div \frac{s\{b_0\}}{\sigma\{b_0\}}$$

- Since Y_i's are $i.i.d.$ normal
  - b_0 is normal $\longrightarrow$ 1st term is standard normal
  - The quantity $\sum (Y_i - \hat{Y}_i)^2/\sigma^2 \sim \chi^2_{n-2}$
  - The variable $s^2\{b_0\}/\sigma^2\{b_0\} \sim \chi^2_{n-2}/(n-2)$
  - The variable $s\{b_0\}/\sigma\{b_0\}$ is independent of b_0

$$\Longrightarrow Test\ Statistics: \quad \frac{b_0 - \beta_0}{s\{b_0\}} \sim t_{n-2}$$

Steps of Hypothesis Test

- $H_0 : \beta_0 = 0$ and $H_a : \beta_0 \neq 0$
- Compute the test statistic. In "Leaning Tower of Pisa":

$$t^\star = \frac{b_0 - 0}{s\{b_0\}} = \frac{-61.12 - 0}{25.13} = -2.43$$

- Compute p-value using sampling distribution

$$P(|t_{n-2}| \geq |t^\star|) = 0.0333$$

- Compare to α and draw conclusion
  - Reject H_0 at α (usually = .05) level, evidence suggests the intercept is different from zero

Confidence Intervals for β_0 and β_1

- Could also form confidence intervals

$$\frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}$$

  - General form for parameter β_1

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

  - Reject $H_0 : \beta_1 = \beta_1^{H_0}$ if $\beta_1^{H_0}$ is not in CI

- Same procedure for β_0

$$\frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2} \Longrightarrow b_0 \pm t(1 - \alpha/2, n-2)s\{b_0\}$$

- 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
- At design stage, if can choose values of X_i:
  - $\text{Var}(b_1) = \sigma^2 / \sum (X_i - \overline{X})^2$ smaller when $\sum (X_i - \overline{X})^2$ is large
  - $\text{Var}(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right)$ smallest when $\overline{X} = 0$

Interval Estimation of $E(Y_h)$

- Often interested in estimating the mean response for particular X_h

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

- Need sampling distribution of $\hat{Y}_h$ to form CI
  - Rewrite $\hat{Y}_h = \sum k_i Y_i$ where

$$k_i = \frac{1}{n} + \frac{(X_h - \overline{X})(X_i - \overline{X})}{\sum (X_i - \overline{X})^2}$$

  - Similar construction as b_0 (i.e., $X_h = 0$)
  - $E(\hat{Y}_h) = E(Y_h)$
  - $\text{Var}(\hat{Y}_h) = \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\Sigma (X_i - \overline{X})^2} \right)$
  - $s^2\{\hat{Y}_h\} = s^2 \left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\Sigma (X_i - \overline{X})^2} \right)$
  - CI: $\hat{Y}_h \pm t(1 - \alpha/2, n-2) s\{\hat{Y}_h\}$

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

- Predicting future observation $Y_{h(new)} = E[Y_h] + \varepsilon_{h(new)}$

  – Estimate $E[Y_h]$ with $\widehat{Y}_h \Longrightarrow \text{Var}(\widehat{Y}_h) = \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\Sigma (X_i - \overline{X})^2} \right)$

- The prediction error is $Y_{h(new)} - \widehat{Y}_h = (E[Y_h] - \widehat{Y}_h) + \varepsilon_{h(new)}$

  – Unlike the expected value, a new observation does not fall directly on the regression line.

  – Must account for added variability in $\varepsilon_{h(new)} \longrightarrow \sigma^2$.

- The variance of the prediction error

$$\sigma^2\{pred\} = \text{Var}(Y_{h(new)} - \widehat{Y}_h) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\Sigma (X_i - \overline{X})^2} \right)$$

- $s^2\{pred\} = s^2 \left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\Sigma (X_i - \overline{X})^2} \right)$

- CI: $\widehat{Y}_h \pm t(1 - \alpha/2, n-2) s\{pred\}$

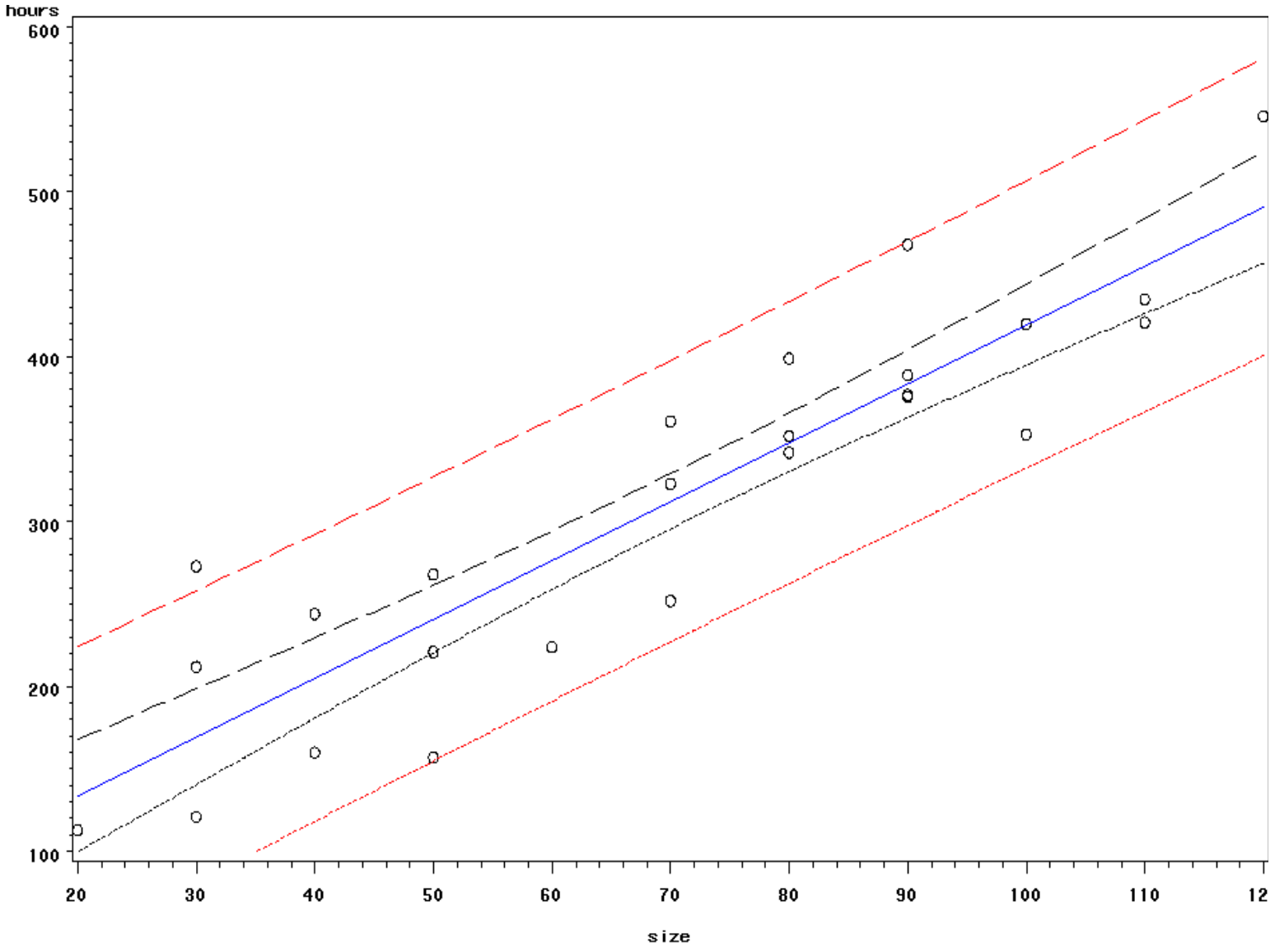
Example: Toluca Company (p. 19)

```
/* 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	
			Output Statistics			
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

Confidence Band

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

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

- Boundary values define a hyperbola
- Confidence level α covers *all* X_h

$$\Pr\left\{|\ \hat{Y}_h - Y_h| \leq W s(\hat{Y}_h),\ \forall X_h\ \right\} \geq 1-\alpha$$

- Will be discussed more in Chapter 4

- The band is the narrowest at $\overline{X}$
- Theory comes from fact that (b_0, b_1) is multivariate normal
  - Joint confidence region for (β_0, β_1) is an ellipse
  - $\text{Cov}(b_0, b_1) = \text{Cov}(\sum k_{i0} Y_i, \sum k_{i1} Y_i) = -\overline{X}\text{Var}(b_1)$
- Band width at $X_h >$ individual CI width of $E[Y_h]$
- Can find α' for individual CIs that gives same results:
  - $t(1 - \alpha'/2, n-2) = \sqrt{2F(1-\alpha; 2, n-2)}$

SAS for Confidence Band

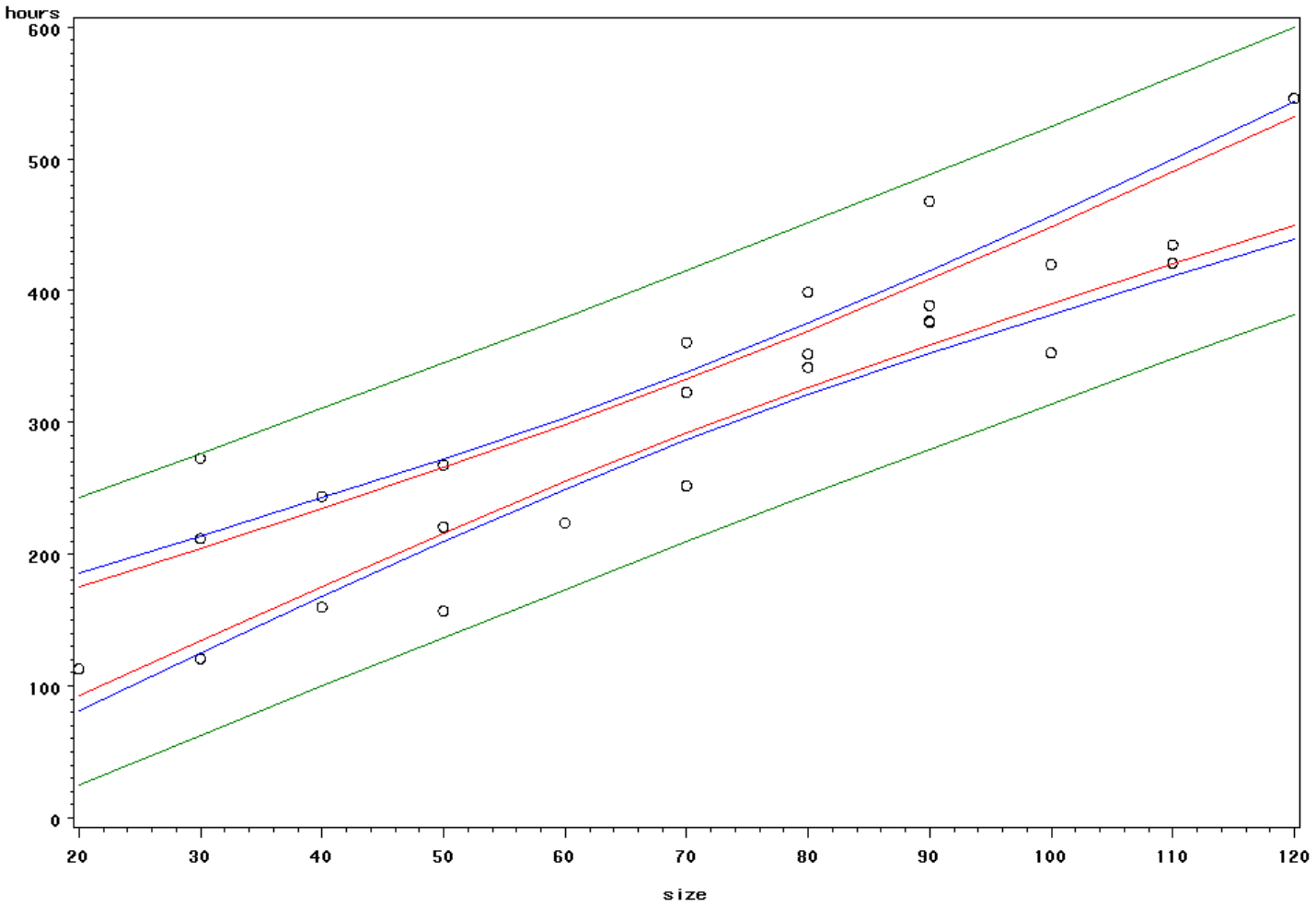
```
/* p: predicted values for the mean
   stdp: sd of the predicted values for the mean
   uclm/lclm: upper/lower bounds of the CI for the mean
   ucl/lcl: upper/lower bounds of the CI for a new value*/
proc reg data=a1;
   model hours=size/clm cli alpha=0.05;
   output out=a2 p=predicted stdp=stdp 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;
```

Confidence Band for the Toluca example



- Blue – 95% confidence band
- Red – 95% confidence interval for the mean
- Green – 95% confidence interval for the individual prediction

ANOVA Approach to Regression

- A second way to test for linear association
- Equivalent to t-test 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)

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

Total Sum of Squares

- If we ignored X_h, 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 $\sum Y_i^2$
- Sum of squares has $n - 1$ degrees of freedom because we replace β_0 with $\overline{Y}$
- 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

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

- Degrees of freedom is 1 due to the addition of the slope
- SSR large when $\hat{Y}_i$'s are different from $\overline{Y}$ (in other words, when there is a linear trend)
- Can also express

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

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)
- SSE large when |residuals| are large
- 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 this all together

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

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

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

F Test

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

$$F^\star = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/df_R}{\text{SSE}/df_E}$$

- If $\beta_1 = 0$ then $F^\star$ should be near one
- Need sampling distribution of $F^\star$ under H_0?
- By Cochran's Theorem (pg 70)

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

- When H_0 is false, MSR $>$ MSE
- p-value $=$ Pr$(F(1, n-2) > F^\star)$
- Reject when $F^\star$ 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 $10.29^2 \approx 105.88$

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

$$F^\star = \frac{(\text{SSE(R)} - \text{SSE(F)})/(df_R - df_F)}{\text{SSE(F)}/df_F}$$

$$= \frac{(\text{SSTO} - \text{SSE})/1}{\text{SSE}/(n-2)}$$

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

Descriptive Measures of Linear Association

- The degree of "linear association" is often the time of interest
- In simple linear regression,
  - Coefficient of determination R^2
  - Estimated Pearson's correlation coefficient r

Coefficient of Determination

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

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

- $0 \leq R^2 \leq 1$
  - often multiplied by 100 and described as a percentage

- 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

Pearson's Correlation Coefficient

- Number between -1 and 1 which measures the strength of the **linear** relationship between two variables, e.g.,

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

- In simple linear regression, 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

- For simple linear regression can show that

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

  - Relationship not true in multiple regression

Normal Correlation Model

- Have assumed X_i's are known constants
- Statistical inferences consider repeated sampling with fixed X values
- What if this assumption is not appropriate?
- In other words, what if X_i's are random?
- If interest still in relation between two variables can use correlation model
- Normal correlation model uses bivariate normal distribution

Bivariate Normal Distribution

- Consider random variables Y_1 and Y_2
- Distribution requires five parameters
  - μ_1 and σ_1 are the mean and std dev of Y_1
  - μ_2 and σ_2 are the mean and std dev of Y_2
  - ρ_{12} is the coefficient of correlation
- Bivariate normal density and marginal distributions given on page 79
- Marginal distributions are normal
- Conditional distributions are also normal

Conditional Distribution

- Consider the distribution of Y_1 given Y_2
  - Can show the distribution is normal
  - The mean can be expressed

$$\left(\mu_1 - \mu_2\rho_{12}\frac{\sigma_1}{\sigma_2}\right) + \rho_{12}\frac{\sigma_1}{\sigma_2}Y_2 \;=\; \alpha_{1|2} + \beta_{12}Y_2$$

  - With constant variance $\sigma_1^2\left(1-\rho_{12}^2\right)$
- Similar properties of normal error regression model
- Can use regression to make inference about Y_1 given Y_2

What if X Random

- What if X_i's are random samples from distribution $g(\cdot)$?
- 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 are independent and $g(\cdot)$ does not involve the parameters β_0, β_1, and σ^2

Inference on ρ_{12}

- Point estimate using $Y = Y_1$ and $X = Y_2$ given on p. 83
- Interest in testing $H_0 : \rho_{12} = 0$
- Test statistic is

$$t^* = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}}$$

- Same result as $H_0 : \beta = 0$
- Can also form CI using Fisher z transformation or large sample approximation (p. 85)
- If X and Y nonnormal, can use Spearman's correlation coefficient (p. 87)

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?