Chapter 2
Inferences in Simple Linear Regression

Professor Dabao 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

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

- Therefore rewrite

$$b_1 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$= \sum_{i=1}^{n} \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2} Y_i = \sum_{i=1}^{n} k_i Y_i$$

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

\[
E(b_1) = E(\sum k_i Y_i) = \sum k_i E(Y_i) = \beta_0 \sum k_i + \beta_1 \sum k_i X_i = 0 + \beta_1
\]

\[
\text{Var}(b_1) = \text{Var}(\sum k_i Y_i) = \sum k_i^2 \text{Var}(Y_i) = \sigma^2 \sum k_i^2 = \sigma^2 / \sum (X_i - \overline{X})^2
\]
Test Statistics $\frac{b_1 - \beta_1}{s\{b_1\}}$

- An estimator of $\text{Var}(b_1)$ is obtained by replacing $\sigma^2$ by its unbiased estimator $MSE = \sum (Y_i - \hat{Y}_i)^2/(n - 2)$,
  
  $$s^2\{b_1\} = \frac{MSE}{\sum (X_i - \bar{X})^2}$$

- Rewrite as
  
  $$\frac{b_1 - \beta_1}{\sigma\{b_1\}} \cdot \frac{s\{b_1\}}{\sigma\{b_1\}}$$

- Since $Y_i$’s are i.i.d. normal
  
  - $b_1$ is normal $\to$ 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 $b_1$

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

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

- Compute p-value using sampling distribution

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

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

- Compare to $\alpha$

- Reject $H_0$ at $\alpha$ (usually = .05) level, evidence suggests a positive linear relationship
Power of Hypothesis Test

• Power = $P\{ \text{reject } H_0 : \beta_1 = \beta_1^{H_0} \mid H_a : \beta_1 \neq \beta_1^{H_0} \}$

• If $H_a$ is true, the test statistic

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

where $\delta$ 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 $\delta$, $n$, and also $\alpha$.

• 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 $t_c$

```sas
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;

Power for the slope in simple linear regression
Inferences Concerning $\beta_0$

- Test of intercept is usually not of interest

Sampling Distribution of $b_0$

- Rewrite $b_0 = \sum k_i Y_i$ where
  \[ k_i = \frac{1}{n} - \frac{\overline{X}(X_i - \overline{X})}{\sum(X_i - \overline{X})^2}, \quad \sum k_i = 1, \quad \sum k_i X_i = 0 \]

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

\[
E(b_0) = E(\sum k_i Y_i) = \sum k_i E(Y_i) = \sum k_i \beta_0 + \sum k_i \beta_1 X_i = \beta_0 + 0
\]

\[
\text{Var}(b_0) = \text{Var}(\sum k_i Y_i) = \sum k_i^2 \text{var}(Y_i) = \sigma^2 \left[ \frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2} \right]
\]
**Test Statistics** \( \frac{b_0 - \beta_0}{s\{b_0\}} \)

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

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

- Rewrite as

\[
\frac{b_0 - \beta_0}{\sigma\{b_0\}} \cdot \frac{s\{b_0\}}{\sigma\{b_0\}}
\]

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

\( \Rightarrow \) Test Statistics : \( \frac{b_0 - \beta_0}{s\{b_0\}} \sim t_{n-2} \)
Steps of Hypothesis Test

- $H_0 : \beta_0 = 0$ and $H_a : \beta_0 \neq 0$

- Compute the test statistic. In “The Leaning Tower of Pisa”:

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

- Compute p-value using sampling distribution

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

- Compare to $\alpha$ and draw conclusion
  - Reject $H_0$ at $\alpha$ (usually $= .05$) level, evidence suggests the intercept is different from zero
Confidence Intervals for $\beta_0$ and $\beta_1$

- Could also form confidence intervals

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

- General form for parameter $\beta_1$

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

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

- Same procedure for $\beta_0$

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

- 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 $X_i$:
  - $\text{Var}(b_1) = \sigma^2 / \sum (X_i - \bar{X})^2$ smaller when $\sum (X_i - \bar{X})^2$ is large
  - $\text{Var}(b_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right)$ smallest when $\bar{X} = 0$
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., $X_h = 0$)
  - $E(\hat{Y}_h) = E(Y_h)$
  - $\text{Var}(\hat{Y}_h) = \sigma^2 \left( \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right)$
  - $s^2\{\hat{Y}_h\} = s^2 \left( \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right)$
  - CI: $\hat{Y}_h \pm t\left(1 - \alpha/2, n - 2\right)s\{\hat{Y}_h\}$
Interval Estimation of $Y_{h(new)}$

- Predicting future observation $Y_{h(new)} = E[Y_h] + \varepsilon_{h(new)}$
  
  - Estimate $E[Y_h]$ with $\hat{Y}_h \implies \text{Var}(\hat{Y}_h) = \sigma^2 \left( \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2} \right)$

- The prediction error is $Y_{h(new)} - \hat{Y}_h = (E[Y_h] - \hat{Y}_h) + \varepsilon_{h(new)}$
  
  - Unlike the expected value, a new observation does not fall directly on the regression line.
  
  - Must account for added variability in $\varepsilon_{h(new)} \rightarrow \sigma^2$.

- The variance of the prediction error
  
  $$\sigma^2\{pred\} = \text{Var}(Y_{h(new)} - \hat{Y}_h) = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2} \right)$$

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

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

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>252378</td>
<td>252378</td>
<td>105.88</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>23</td>
<td>54825</td>
<td>2383.71</td>
<td>562</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Cor Total</td>
<td>24</td>
<td>307203</td>
<td>1271.72</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 48.82331  R-Square 0.8215
Dependent Mean 312.28000  Adj R-Sq 0.8138
Coeff Var 15.63447

Parameter Estimates

<p>| Variable | DF | Estimate | Error | t Value | Pr &gt; |t| |
|----------|----|----------|-------|---------|------|---|
| Intercept| 1  | 62.36586 | 26.17743| 2.38    | 0.0259|
| size     | 1  | 3.57020  | 0.34697| 10.29   | &lt;.0001|</p>
<table>
<thead>
<tr>
<th>Obs size</th>
<th>obs size</th>
<th>hours</th>
<th>Value</th>
<th>Mean</th>
<th>Predict</th>
<th>90% CL</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80</td>
<td>399.000</td>
<td>347.982</td>
<td>10.3628</td>
<td>330.2215</td>
<td>365.7425</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>121.000</td>
<td>169.4719</td>
<td>16.9697</td>
<td>140.3880</td>
<td>198.5559</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>221.000</td>
<td>240.8760</td>
<td>11.9793</td>
<td>220.3449</td>
<td>261.4070</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>90</td>
<td>376.000</td>
<td>383.6840</td>
<td>11.9793</td>
<td>363.1530</td>
<td>404.2151</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>70</td>
<td>361.000</td>
<td>312.2800</td>
<td>9.7647</td>
<td>295.5446</td>
<td>329.0154</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>224.000</td>
<td>276.5780</td>
<td>10.3628</td>
<td>258.8175</td>
<td>294.3385</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>120</td>
<td>546.000</td>
<td>490.7901</td>
<td>19.9079</td>
<td>456.6706</td>
<td>524.9096</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>352.000</td>
<td>347.9820</td>
<td>10.3628</td>
<td>330.2215</td>
<td>365.7425</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>353.000</td>
<td>419.3861</td>
<td>14.2723</td>
<td>394.9251</td>
<td>443.8470</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>157.000</td>
<td>240.8760</td>
<td>11.9793</td>
<td>220.3449</td>
<td>261.4070</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>40</td>
<td>160.000</td>
<td>205.1739</td>
<td>14.2723</td>
<td>180.7130</td>
<td>229.6349</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>70</td>
<td>252.000</td>
<td>312.2800</td>
<td>9.7647</td>
<td>295.5446</td>
<td>329.0154</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>90</td>
<td>468.000</td>
<td>383.6840</td>
<td>11.9793</td>
<td>363.1530</td>
<td>404.2151</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>40</td>
<td>244.000</td>
<td>205.1739</td>
<td>14.2723</td>
<td>180.7130</td>
<td>229.6349</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>80</td>
<td>342.000</td>
<td>347.9820</td>
<td>10.3628</td>
<td>330.2215</td>
<td>365.7425</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>70</td>
<td>323.000</td>
<td>312.2800</td>
<td>9.7647</td>
<td>295.5446</td>
<td>329.0154</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>65</td>
<td>.</td>
<td>294.4290</td>
<td>9.9176</td>
<td>277.4315</td>
<td>311.4264</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>100</td>
<td>.</td>
<td>419.3861</td>
<td>14.2723</td>
<td>394.9251</td>
<td>443.8470</td>
<td></td>
</tr>
<tr>
<td>Obs size</td>
<td>hours</td>
<td>Value</td>
<td>Mean</td>
<td>Predict</td>
<td>90% CL Predict</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>----------</td>
<td>--------</td>
<td>---------</td>
<td>----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>80</td>
<td>399.0000</td>
<td>347.9820</td>
<td>10.3628</td>
<td>262.4411</td>
<td>433.5230</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>121.0000</td>
<td>169.4719</td>
<td>16.9697</td>
<td>80.8847</td>
<td>258.0591</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>221.0000</td>
<td>240.8760</td>
<td>11.9793</td>
<td>154.7171</td>
<td>327.0348</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>90</td>
<td>376.0000</td>
<td>383.6840</td>
<td>11.9793</td>
<td>297.5252</td>
<td>469.8429</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>70</td>
<td>361.0000</td>
<td>312.2800</td>
<td>9.7647</td>
<td>226.9460</td>
<td>397.6140</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>224.0000</td>
<td>276.5780</td>
<td>10.3628</td>
<td>191.0370</td>
<td>362.1189</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>120</td>
<td>546.0000</td>
<td>490.7901</td>
<td>19.9079</td>
<td>400.4244</td>
<td>581.1558</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>352.0000</td>
<td>347.9820</td>
<td>10.3628</td>
<td>262.4411</td>
<td>433.5230</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>353.0000</td>
<td>419.3861</td>
<td>14.2723</td>
<td>332.2072</td>
<td>506.5649</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>157.0000</td>
<td>240.8760</td>
<td>11.9793</td>
<td>154.7171</td>
<td>327.0348</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>40</td>
<td>160.0000</td>
<td>205.1739</td>
<td>14.2723</td>
<td>117.9951</td>
<td>292.3528</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>70</td>
<td>252.0000</td>
<td>312.2800</td>
<td>9.7647</td>
<td>226.9460</td>
<td>397.6140</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>90</td>
<td>468.0000</td>
<td>383.6840</td>
<td>11.9793</td>
<td>297.5252</td>
<td>469.8429</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>40</td>
<td>244.0000</td>
<td>205.1739</td>
<td>14.2723</td>
<td>117.9951</td>
<td>292.3528</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>80</td>
<td>342.0000</td>
<td>347.9820</td>
<td>10.3628</td>
<td>262.4411</td>
<td>433.5230</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>70</td>
<td>323.0000</td>
<td>312.2800</td>
<td>9.7647</td>
<td>226.9460</td>
<td>397.6140</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>65</td>
<td>.</td>
<td>294.4290</td>
<td>9.9176</td>
<td>209.0432</td>
<td>379.8148</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>100</td>
<td>.</td>
<td>419.3861</td>
<td>14.2723</td>
<td>332.2072</td>
<td>506.5649</td>
<td></td>
</tr>
</tbody>
</table>
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

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

- Boundary values define a hyperbola
- Confidence level $\alpha$ covers all $X_h$

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

- Will be discussed more in Chapter 4
The band is the narrowest at $\bar{X}$

Theory comes from fact that $(b_0, b_1)$ is multivariate normal
  
  Joint confidence region for $(\beta_0, \beta_1)$ is an ellipse

  $\text{Cov}(b_0, b_1) = \text{Cov}(\sum k_i Y_i, \sum k_i Y_i) = -\bar{X} \text{Var}(b_1)$

Band width at $X_h >$ individual CI width of $E[Y_h]$  

Can find $\alpha'$ 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

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

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

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

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

$$\text{SSTO} = \text{SSR} + \text{SSE}$$
Total Sum of Squares

- If we ignored $X_h$, the sample mean $\bar{Y}$ 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 $\beta_0$ with $\bar{Y}$

- The total mean square is $\text{SSTO}/(n - 1)$ and represents an unbiased estimate of $\sigma^2$ 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

\[
\begin{align*}
\text{SSR} & = \sum (\hat{Y}_i - \overline{Y})^2 \\
& = \sum (b_0 + b_1X_i - b_0 - b_1\overline{X})^2 \\
& = b_1^2 \sum (X_i - \overline{X})^2
\end{align*}
\]


**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 \((\beta_0, \beta_1)\)

- SSE large when \(|\text{residuals}|\) are large

- Implies \( Y_i \)'s vary substantially around line

- The MSE = \( \text{SSE}/(n-2) \) and represents an unbiased estimate of \( \sigma^2 \) when taking \( X \) into account
**ANOVA Table**

- Table puts this all together

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression (Model)</td>
<td>1</td>
<td>$b_1^2 \sum (X_i - \bar{X})^2$</td>
<td>SSR/1</td>
</tr>
<tr>
<td>Error</td>
<td>$n - 2$</td>
<td>$\sum (Y_i - \hat{Y})^2$</td>
<td>SSE/(n - 2)</td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1$</td>
<td>$\sum (Y_i - \bar{Y})^2$</td>
<td></td>
</tr>
</tbody>
</table>
Expected Mean Squares

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

\[
E(\text{MSR}) = E(b_1^2 \sum (X_i - \bar{X})^2) \\
= E(b_1^2) \sum (X_i - \bar{X})^2 \\
= (\text{Var}(b_1) + \{E(b_1)\}^2) \sum (X_i - \bar{X})^2 \\
= \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2
\]

- If \( \beta_1 = 0 \), MSR unbiased estimate of \( \sigma^2 \)
**F Test**

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

- Consider

\[ F^* = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/df_R}{\text{SSE}/df_E} \]

- If $\beta_1 = 0$ then $F^*$ should be near one

- Need sampling distribution of $F^*$ under $H_0$?

- By Cochran’s Theorem (pg 70)

\[
F^* = \left( \frac{\text{SSR}}{\sigma^2} \right) \div \left( \frac{\text{SSE}}{\sigma^2} \right) \div \left( \frac{1}{n - 2} \right)
\]

\[
F^* \sim \frac{\chi^2_1}{1} \div \frac{\chi^2_{n-2}}{n - 2} \sim F_{1,n-2}
\]
• When $H_0$ is false, MSR > MSE

• p-value $= \Pr(\mathcal{F}(1, n-2) > \mathcal{F}^*)$

• Reject when $\mathcal{F}^*$ large, p-value small

• Recall t-test for $H_0 : \beta_1 = 0$

• Can show $t^2_{n-2} \sim \mathcal{F}_{1, n-2}$

• 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

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>252378</td>
<td>252378</td>
<td>105.88</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>23</td>
<td>54825</td>
<td>2383.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cor Total</td>
<td>24</td>
<td>307203</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 | Estimate | 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
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 $\rightarrow H_0 : \beta_1 = 0$

• 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 $Y$

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

• 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 $X$

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

- $0 \leq R^2 \leq 1$
  - 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.,

\[ \rho = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} \]

- In simple linear regression, \( \rho \) 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
  - $\mu_1$ and $\sigma_1$ are the mean and std dev of $Y_1$
  - $\mu_2$ and $\sigma_2$ are the mean and std dev of $Y_2$
  - $\rho_{12}$ 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
    \[
    (\mu_1 - \mu_2 \rho_{12} \frac{\sigma_1}{\sigma_2}) + \rho_{12} \frac{\sigma_1}{\sigma_2} Y_2 = \alpha_{1|2} + \beta_{12} Y_2
    \]
  - 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 $\sigma^2$
  - The $X_i$ are independent and $g(\cdot)$ does not involve the parameters $\beta_0$, $\beta_1$, and $\sigma^2$
**Inference on \( \rho_{12} \)**

- 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 $\beta_1$
• Inference concerning $\beta_0$
• 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?