Prediction interval for a future $Y$ 

we know $\hat{Y}_* = (Y \text{ at a } x^*)$ is $N(\beta_0 + \beta_1 x^*, \sigma^2)$ 

so if we want to predict the value of $Y$, 

$$P \left[ -1.96 \frac{Y - (\beta_0 + \beta_1 x^*)}{\sigma} \leq 1.96 \right] = 0.95$$

or $P \left[ \beta_0 + \beta_1 x^* - 1.96\sigma \leq Y \leq \beta_0 + \beta_1 x^* + 1.96\sigma \right] = 0.95$

But this interval is not known since we don't know $\beta_0, \beta_1, \sigma$

Instead, want $P \left[ L \leq Y \leq U \right] = 0.95$

$\Phi$ 

Computable from a random sample $x_1, \ldots, x_n$

$Y_1, \ldots, Y_n$

(Note: this is not the same as obtaining a CI for $\beta_0 + \beta_1 x^*$)

Consider $Y_\hat{x}^* - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$

$$\text{Var} \left[ Y_\hat{x}^* \right] = \text{Var} [\hat{Y}_*] + \text{Var} \left[ \hat{\beta}_0 + \hat{\beta}_1 x^* \right]$$
\[ = \sigma^2 + \sigma^2 \left[ \frac{1}{n} + \frac{(\bar{x}^* - \bar{x})^2}{S_{xx}} \right] \]

\[ = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(\bar{x}^* - \bar{x})^2}{S_{xx}} \right] \]

\[ \text{not computable from data} \]

can be shown:

\[ T = \frac{Y^* - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{S \sqrt{1 + \frac{1}{n} + \frac{(\bar{x}^* - \bar{x})^2}{S_{xx}}}} \]

has a \( t \)-distribution with \( n-2 \) d.f.

100(1-\( \alpha \))% Prediction Interval (PI) for a future \( Y \) observation to be made when \( x = x^* \) is

\[ \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\frac{\alpha}{2}, n-2} \cdot S \cdot \sqrt{1 + \frac{1}{n} + \frac{(\bar{x}^* - \bar{x})^2}{S_{xx}}} \]

Note: PI is wider than CI.

Study Example 12.14
Sec 12.5 Correlation

Recall $X, Y$ - r.v.s

- $\text{Cov}(X, Y) = E[(X-\mu_X)(Y-\mu_Y)]$

$$= \begin{cases} \sum_{x} \frac{1}{N}(x-\mu_x)(y-\mu_y) p(x, y) & X, Y \text{ discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_x)(y-\mu_y) f(x, y) \, dx \, dy & X, Y \text{ - continuous} \end{cases}$$

- $\text{P}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \in [-1, 1]$

- $X, Y$ independent $\Rightarrow \rho = 0$

- $\rho = 1 \text{ or } -1 \Leftrightarrow Y = ax + b \text{ for some numbers with } a \neq 0$
Intellectual exercise:

If we draw $X, Y$ many many times and we happen to see

then what does it mean to the joint distribution $(pmf/pdf)$ of $X, Y$?

\[ S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \]

\[ r = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}} = \text{sample correlation} \]

\[ = \frac{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2} \]

Example 12.5

<table>
<thead>
<tr>
<th>$x$</th>
<th>2.4</th>
<th>3.4</th>
<th>4.6</th>
<th>3.7</th>
<th>2.2</th>
<th>3.3</th>
<th>4.0</th>
<th>2.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>1.33</td>
<td>2.12</td>
<td>1.80</td>
<td>1.65</td>
<td>2.00</td>
<td>1.76</td>
<td>2.11</td>
<td>1.63</td>
</tr>
</tbody>
</table>

\[ \sum x_i = 25.7, \quad \sum y_i = 14.4, \quad \sum x_i^2 = 88.31 \]

\[ \sum x_i y_i = 46.856, \quad \sum y_i^2 = 26.4324 \]

\[ S_{xx} = 88.31 - \frac{(25.7)^2}{8} = 5.75 \]

\[ S_{yy} = 26.4324 - \frac{(14.4)^2}{8} = 0.5124 \]
\[ S_{xy} = 46.856 - \frac{(25.7)(4.40)}{8} = 0.5860 \]

\[ r = \frac{0.5860}{\sqrt{55.75} \sqrt{10.5124}} = 0.347 \] 

**Properties of \( r \)**

1. **Value of \( r \) does not depend on which of the two variables under study is labeled \( x \) and which is labeled \( y \).**

   \[( r = r((x_i), (y_i)) ) \]

   \[ \text{ther} \quad r((x_i), (y_i)) = r((y_i), (x_i)) \]

2. **The value of \( r \) is independent of the units in which \( x \) and \( y \) are measured.**

   \[ ( \text{ie.} \quad r((a x_i), (b y_i)) = r((x_i), (y_i)) ) \]

   \[ a, b > 0 \]

3. **\(-1 \leq r \leq 1\)**

4. **\( r = 1 \iff \) all \((x_i, y_i)\) lie on a straight line with positive slope**

   \[ r = -1 \iff \text{all } (x_i, y_i) \text{ lie on a straight line with negative slope.} \]

5. **\( (r^2) = \frac{r^2}{\text{the } r^2 \text{ defined earlier}} \)**
What does $r$ measure?

$r$ near $1$

$r$ near $-1$

$r$ near $0$, no apparent relationship

$r$ near $0$, nonlinear relationship

$r$ measures the degree of linear relationship among variables.
Assume $X_1, \ldots, X_n$ is the marginal distribution $f_X$ of an $f$.

$Y_1, \ldots, Y_n$ is the marginal distribution $f_Y$ of $f$.

\[
\hat{\rho} = R = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \sqrt{\sum(Y_i - \bar{Y})^2}} \quad \text{as an estimator of} \quad \text{CORR}(X, Y) = \frac{\text{COV}(X, Y)}{\sigma_X \sigma_Y}
\]

Can we say something precise about the relationship the two?

**CONSIDER** bivariate normal distribution

\[
f(x, y) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \left( \frac{x - \mu_1}{\sigma_1} \right)^2 - 2\rho \left( \frac{x - \mu_1}{\sigma_1} \right) \left( \frac{y - \mu_2}{\sigma_2} \right) + \left( \frac{y - \mu_2}{\sigma_2} \right)^2 \right] \right\}
\]

$\sigma_1, \sigma_2 > 0$

$\rho \in (-1, 1)$

$\mu_1, \mu_2$ any real numbers

This looks horrendous, but remember:

1. this is the joint pdf of

\[
\begin{bmatrix}
X
Y
\end{bmatrix} = \begin{bmatrix}
\sigma_1 Z_1 + \mu_1
\sigma_2 [\rho Z_1 + \sqrt{1-\rho^2} Z_2] + \mu_2
\end{bmatrix} \quad \text{(see demo for this)}
\]

\[
= \begin{bmatrix}
\sigma_1 & 0 \\
\sigma_2 \rho & \sigma_2 \sqrt{1-\rho^2}
\end{bmatrix} \begin{bmatrix}
Z_1 \\
Z_2
\end{bmatrix} + \begin{bmatrix}
\mu_1 \\
\mu_2
\end{bmatrix}
\]
we are not going to derive this fact.

But note: if \( p = 0 \)

then

\[
\begin{bmatrix}
X \\
Y
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
0 \\
0
\end{bmatrix} + \begin{bmatrix}
\mu_1 \\
\mu_2
\end{bmatrix}
\]

meaning that

\( X \sim N(\mu_1, \sigma_1) \) and \( X, Y \) are

\( Y \sim N(\mu_2, \sigma_2) \) independent

and, indeed, in this special case:

\[
f(x,y) = \frac{1}{\sqrt{2\pi \sigma_1}} \exp\left(-\frac{1}{2\sigma_1}(x-\mu_1)^2\right) \\
\times \\
\frac{1}{\sqrt{2\pi \sigma_2}} \exp\left(-\frac{1}{2\sigma_2}(y-\mu_2)^2\right)
\]

according to (4).

\( \Box \)

2. In general, it's still true that

\[
\begin{cases}
X \sim N(\mu_1, \sigma_1) & \text{just that} \\
Y \sim N(\mu_2, \sigma_2) & X \text{ and } Y \text{ are independent}
\end{cases}
\]

in fact

\[ \text{CORR}(X,Y) = p \]

\( \Box \)

3. \( f_{Y|X}(y|x) \) is the pdf of \( N(\mu_2 - \rho \mu_1 \sigma_2 / \sigma_1, \sqrt{1 - \rho^2} \sigma_2) \)

\[
\frac{f(y|X)}{f_X(x)} \quad \text{(easy to prove, assuming (1))}
\]
consequence: if the observed pairs

\((x_i, y_i)\) are actually drawn

from a bivariate normal distribution

then the simple linear regression

model is an appropriate way

of studying the behaviour of \(Y\)

for fixed \(x\).

\[ \begin{align*}
4) \quad & \text{If } (X_1, Y_1), \ldots, (X_n, Y_n) \text{ are drawn from a \underline{bivariate normal distribution},} \\
& \text{then the r.v.} \\
& V = \frac{1}{2} \ln \left( \frac{1 + R}{1 - R} \right) \quad \text{(called Fisher transformation)} \\
& \text{has approximately a normal distribution with mean} \\
& \frac{1}{2} \ln \left( \frac{1 + \rho}{1 - \rho} \right), \quad \sigma_v^2 = \frac{1}{n - 3} \\
& \text{The test statistics for testing } \text{Ho: } \rho = \rho_0 \text{ is} \\
& Z = \frac{V - \frac{1}{2} \ln \left[ \frac{(1 + \rho_0)/(1 - \rho_0)}{1/\sqrt{n - 3}} \right]}{1/\sqrt{n - 3}} \\
& \text{Alt. Hyp.} \quad \text{Rejection Reg. for Level } \alpha \text{-test} \\
\text{Ha: } \rho > \rho_0 & \quad Z > 3 \alpha \\
\text{Ha: } \rho < \rho_0 & \quad Z < -3 \alpha \\
\text{Ha: } \rho \neq \rho_0 & \quad Z > 3 \alpha_2 \text{ or } Z < -3 \alpha_2 
\end{align*} \]
Example 12.18

CI for $\rho$:

$$
\bar{\nu} \pm \frac{z_{\alpha/2}}{\sqrt{n-3}} \quad \text{is a } 100(1-\alpha)\% \ CI
$$

$$
\frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) \quad \text{for } \mu_v = \frac{1}{2} \ln \left( \frac{1+\rho}{1-\rho} \right)
$$

$$
C_1 = \bar{\nu} - \frac{z_{\alpha/2}}{\sqrt{n-3}} \quad \text{and } C_2 = \bar{\nu} + \frac{z_{\alpha/2}}{\sqrt{n-3}}
$$

So a 100(1-$\alpha$)\% CI for $\rho$ is

$$
\left( \frac{e^{2C_1} - 1}{e^{2C_1} + 1}, \frac{e^{2C_2} - 1}{e^{2C_2} + 1} \right).
$$

Example 12.19