Regression

Ch. 15

Scatter diagram shows linear relationship (if one exists)

\[ -1 < 0 < 1 \]

\[ \text{Perfectly INVERSELY Related} \quad \text{No Linear Relation} \quad \text{Perfectly DIRECTLY Related} \]

\[ \text{Perfectly Inversely Related} \quad \text{No Relation} \quad \text{Perfectly Directly Related} \]

Regression Analysis:
A method/procedure for determining the statistical relationship, if any, between two or more variables, typically for predictive purposes.

3 Phases:
1. Fitting (fitting the line into the data/scatter)
2. Testing (are the variables significantly linear related?)
3. Predicting

"What"  \( \hat{y} \) (fitted/predicted value for \( y \))  \( \hat{y} = b_0 + b_1 x \)

\( y \) = observed value
\( \hat{y} \) = predicted value

SLR: Simple Linear Regression: using only one predictor variable (x). (We will add more later.)
### Regression Example

\( n=5 \)

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( \Delta x )</th>
<th>( \Delta y )</th>
<th>( \frac{\Delta x \cdot \Delta y}{2} )</th>
<th>( \sum (x - \overline{x})^2 )</th>
<th>( \sum (y - \overline{y})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0</td>
<td>-1</td>
<td>4</td>
<td>2.8</td>
<td>3.28</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>5.9</td>
<td>5.59</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>9.0</td>
<td>9.00</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>9.0</td>
<td>9.00</td>
</tr>
</tbody>
</table>

\( b_0 = \text{slope} = \frac{\sum xy}{\sum x^2} = \frac{31}{10} = 3.1 \)

\( b_0 = \text{y-intercept} = \overline{y} - b_0 \overline{x} = 9 - 3.1(2) = 2.8 \)

\( \hat{y} = b_0 + b_1 x \)

\( \hat{y} = 2.8 + 3.1(x) \)

### Characteristics of a Regression Line:
1. Minimize SSE
2. Always pass thru \( \bar{x}, \bar{y} \)
3. **All same**
   - Sum of squared errors
   - Least squares line
   - Regression line
   - Estimating line

### Residual Error
\( e = \text{residual error} = (y - \hat{y}) = \text{observed minus predicted} \)

\( \text{sum of errors} = 0 \) so not best line, so use SSE
\[ \text{SSSTO} = \sum (y - \bar{y})^2 \]: total variation in \( y \) as measured about their mean.

\[ \text{SSE} = \sum (y - \hat{y})^2 \]: Variation in \( y \) about regression line.

- Minimum: 0 if perfect fit then \( y = \hat{y} \) everywhere
- Maximum: SSSTO if \( b = 0 \) then \( \hat{y} = \bar{y} \)

\[ \text{SSR} = \text{SSSTO} - \text{SSE} \]: reduction in SSSTO to SSE, attributed to regression

\[ r = \frac{\text{Cov}(x, y)}{\text{std. error}_x \times \text{std. error}_y} \]: coefficient of correlation - tells if there is relationship or not and if so, the strength: \(-1 < r < 1\)

the closer to -1 or +1, the stronger the relation.

\[ \text{Covariance} = \text{Cov}(x, y) \]: tells direction of linear relation of \( x \) and \( y \)

\[ \text{Cov}(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n-1} \]

\[ \text{Var}(x) = \frac{\sum (x - \bar{x})^2}{n-1} \]

\[ \text{Var}(y) = \frac{\sum (y - \bar{y})^2}{n-1} \]

\[ \text{SSE} = \sum (y - \hat{y})^2 \]

\[ r^2 = \frac{\text{SSE}}{\text{SSSTO}} \]: coefficient of determination - tells what portion (%) of \( y \) is determined by \( x \). \( 0 \leq r^2 \leq 1 \)

\[ r^2 = 1 \] if \( \text{SSE} = 0 \)

\[ \text{SSE} = 0 \] if \( \text{SSSTO} = 0 \)

\[ r^2 = 0 \] if \( \text{SSE} = \text{SSSTO} \)

\[ r^2 = 1 \] (perfect fit)
```
* Cxy, r, b1 all have same sign.

Relationships:

\[ r = \frac{C_{xy}}{\sqrt{C_{xx}C_{yy}}} \]
\[ b_1 = \frac{C_{xy}}{C_{xx}} \]

* If Cxy, r, b1, or SSR = 0 then all = 0.

Regression Example:

\[ SS_{TO} = \sum (y_i - \bar{y})^2 = 98.0 \]
\[ SSE = \sum (y_i - \hat{y}_i)^2 = 1.9 \]

\[ SAS_{R} = SS_{TO} - SSE = 96.0 \]

\[ \hat{y} = \bar{y} + \hat{b_1} \cdot x \]

\[ \hat{b_1} = \frac{SS_{R}}{SS_{xx}} = \frac{96.0}{24.5} = 3.95 \]
\[ \hat{b_0} = \bar{y} - \hat{b_1} \cdot \bar{x} \]

\[ b_1 = \frac{SS_{R}}{SS_{xx}} = \frac{96.0}{24.5} = 3.95 \]

\[ \hat{y} = \bar{y} + \hat{b_1} \cdot \bar{x} \]

\[ r = \frac{C_{xy}}{\sqrt{C_{xx}C_{yy}}} = \frac{7.75}{1.58} = 4.95 \]

\[ \sigma = \sqrt{\frac{MSE}{n-2}} = \sqrt{\frac{1.90}{1.633}} = 0.43 \]

\[ \text{Example ANOVA:} \]

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSR</td>
<td>96.0</td>
<td>1</td>
<td>MSR= \frac{96.0}{1}= 96.0</td>
<td></td>
</tr>
<tr>
<td>SSE</td>
<td>1.90</td>
<td>3</td>
<td>MSE= \frac{1.90}{1.633}= 1.16</td>
<td></td>
</tr>
<tr>
<td>SSTO</td>
<td>98.0</td>
<td>n-1= 4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ F_{0.05, 4, 1} = 151.82 \rightarrow \text{tables} \]

\[ p\text{-value} < 0.01 \]

H0: \[ B_1 = 0 \]
Hi: \[ B_1 \neq 0 \]

\[ \alpha = 0.05 = 10.13 \]
\[ 0.025 = 17.44 \]
\[ 0.01 = 34.12 \]

\[ \beta = 0.05 = 10.13 \]
```
Descriptive Statistics: X, Y

<table>
<thead>
<tr>
<th>Variable</th>
<th>Total Count</th>
<th>Mean</th>
<th>SE Mean</th>
<th>StDev</th>
<th>Variance</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>5</td>
<td>2.000</td>
<td>0.707</td>
<td>1.581</td>
<td>2.500</td>
<td>10.00</td>
</tr>
<tr>
<td>Y</td>
<td>5</td>
<td>9.00</td>
<td>2.21</td>
<td>4.95</td>
<td>24.50</td>
<td>45.00</td>
</tr>
</tbody>
</table>

Correlations: X, Y

Pearson correlation of X and Y = 0.990
P-Value = 0.001

Covariances: X, Y

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>2.50000*</td>
<td>7.75000*</td>
</tr>
<tr>
<td>Y</td>
<td>7.75000*</td>
<td>45.00000*</td>
</tr>
</tbody>
</table>

Regression Analysis: Y versus X

The regression equation is

\[ Y = 2.80 + 3.10 \times X \]

\[ t = \frac{b_1 - 0}{SE_{b_1}} \]

\[ F = 151.74 \]

\[ R^2 = 0.981 \]

Scatterplot of Y vs X
df in regression: # of predictors being used

↑ tightly scattered = ↑ Coef. Correlation ↓ std. error.

Using ‚_related variables = ↓ Error.

* 3-dimensional (contains many) distributions & variability are same.

See p. 554-565

\[ y = b_0 + b_1 x \]

for population

\( b_0 \) & \( b_1 \) are estimates of Beta.

\( X \): Highschool GPA
\( \bar{X} \): Std. Error of Estimate

\( Y \): ACT SCORE

\( \bar{Y} \): All get the same score on the SAT.

\[ H_0: B_1 = 0 \quad \rightarrow \quad \text{Almost ALWAYS the test.} \]

\( \bar{B} \): \( X \) disappeared so \( x \) has no relation to \( y \).

\( \bar{B} \): \( y \) random variable \( \bar{B} \): \( x \) random variable.

\[ E(\bar{B}) = \bar{B} \]

\( \bar{B} = f(x,y) \)

\( \bar{B} = \text{random variable} \quad \bar{B} = \text{random variable.} \]

\( H_0: \bar{B} = 0 \quad \rightarrow \quad \text{Almost Never tested} \)

\( B_0 = \text{y intercept so } x = 0 \)

Ex: What is weight when height is \( x \)? Nonsense!!

so, \( H_0: B_0 = 0 \) is nonsense!
\[ \sqrt{\text{MSE}} = 0 \]
\[ \Delta = \Delta_y - \Delta x = \text{Standard error of estimate} \]
\[ = \text{std. deviation for regression} \]
\[ = \text{std. deviation of residuals} \]
\[ = \text{measure of variation of } Y \text{ about the regression line} = \text{measure of avg. } (\text{if } N \text{ doesn't inflate}) \]

\[ \Delta b_1 = \frac{\Delta}{\sqrt{\sum x^2}} \rightarrow \sqrt{\text{MSE}} \]

**Using Example**

\[ \Delta = \sqrt{\text{MSE}} = \sqrt{0.633} = 0.796 \]

\[ \Delta b_1 = 0.796 = \frac{0.796}{3.162} = 0.2517 \approx 0.253 \]

\[ t = \frac{b_1 - B_1}{\sigma_{b_1}} = \frac{3.1 - 0}{0.253} = 12.3 \]

**Making Predictions:** (will be conditional on X values)

\[ Y = b_0 + b_1x \]

\[ CI = \text{pt. estimate } \pm (\text{Multiplier}) (\text{Std. error}) \]

2 Types of Predictions:

1. **Global Test/Prediction:** The mean of Y for a given X.
   - What is mean/avg. GPA for all who scored 18 on SAT?
   - (Narrow Interval) (Smaller Std. error)

2. **Marginal Test/Prediction:** An individual, new value for Y
   - For a given X
   - What is prediction for indiv. who scored 18 on SAT?
   - (Wider Interval) (Larger Std. error)
\[ \Delta \text{mean} = \begin{vmatrix} \frac{1}{n} + \left( \frac{X_g - \bar{X}}{\sigma_x} \right)^2 \end{vmatrix} \]

- std error
- better predictions

1. \( \Delta \) is directly related to mean. \( \Delta = 0.796 \)

Scatter:
\( \Delta \uparrow \) then \( \hat{y} \uparrow \)

\( r^2 = \text{better fit} \)

2. Inversely related to mean:
Larger sample size = \( \Delta \uparrow \) = better estimate

3. Min. Value = 0 (act-pred)²
   - How to make part 3 to be 0: Both must be same
   - \( \frac{0}{\sigma_x^2} = 0 \)
   - In middle of \( \hat{y} \) is best prediction
   - Move away from middle = less accurate prediction

Ex: \( X_g = 3 \) (we calculated \( \Delta = 0.796 \))

\[ \Delta = 0.796 \left( \frac{1}{5} + \left( \frac{3 - 3}{\sigma_x} \right)^2 \right) = \frac{0.796}{10} \]

\[ = 0.796 \cdot 0.30 = 0.796 \cdot (0.548) = 0.436 = \text{std. error} \]

For 95% Conf:
pt. estm. \( \pm \) (multi) \( \text{std. error} \) for Conf. Invl.

For \( x = 3 \)
\( t \) for 95%
\( \hat{y} = 2.8 \pm 3.1(3) \)
\( w/df = 3 \) (ANOVA)
\( \rightarrow \text{table} = \) 3.182

\[ 12.1 \pm 3.182(0.436) \]

12.1 \pm 1.39 = (10.71, 13.49)
\[ \sigma^2_{\text{new}} = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_g - x)^2}{\hat{\sigma}^2} \right] \leq \sigma^2 \]  
(increases by \( \sigma^2 \))

Ex: \( x_g = 3 \)  
(We calculated \( \sigma = 0.796 \))

\[ \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_g - x)^2}{\hat{\sigma}^2} \right] = 0.796 \sqrt{1 + 0.30} : 0.796 \sqrt{1.30} = 0.796 (1.14) = 0.91 \text{ (std. error)} \]

for 95% conf:

pt. Estim \( \pm \) (mult) (St. Error)  
\[ 12.1 \pm 3.182 \, 0.91 \]

\[ 12.1 \pm 3.182 \, (0.91) \]

Wider interval

\[ 12.1 \pm 2.576 = (9.304, 14.996) \]  
For individual prediction

Wider interval