Descriptive and Inferential Methods in Regression
Overview (Chapters 14 & 15)

14.2 The Regression Equation

Up to now: Inferences about single variables.

In this chapter:

Relate linearly a single dependent variable Y (e.g. house price) to a single independent variable X (e.g. square footage).
Regression with a single predictor (SLR-
Simple Linear Regression)

**Variables**

<table>
<thead>
<tr>
<th>A dependent variable</th>
<th>An independent variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>Denoted Y</td>
<td>Denoted X</td>
</tr>
</tbody>
</table>

(Variable to be modelled) (Predictor variable)

**Examples:** (for SLR)

<table>
<thead>
<tr>
<th>Family expenditure</th>
<th>Size of family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density in plant species</td>
<td>Altitude</td>
</tr>
</tbody>
</table>

Sales | Advertising
A Straight Line Regression Model

Deterministic Model vs Stochastic Model

↑
(Exact)

↑
(Mathematics)

↑
(Has random component)

↑
(Statistics)

Simple linear regression model:

\[ Y = \beta_0 + \beta_1X + \varepsilon \]

where:

Y : dependent variable
X : independent variable
\( \beta_0 \) : intercept of line (beta 0)
\( \beta_1 \) : slope of straight line (beta 1)
Graphically:

Interpretation:

$\beta_1$ amount of increase (or decrease) in $Y$ for every one unit increase in $X$.

$\beta_0$ and $\beta_1$ are usually unknown and are estimated with $b_0$ and $b_1$ calculated from the data. (Sample estimates).
The Method of Least Squares

Example

To understand the pattern of variation in plant species in Mediterranean grasslands, data were collected on $Y=$ density of plant species (# of species per 0.04 m$^2$) and $X=$ altitude of the region (in 1000 m) from 12 experimental plots for a period of five years (*Journal of Vegetation Science*). The results obtained for the average (over the five-year period) species density are as follows:

<table>
<thead>
<tr>
<th>Altitude x (1000 m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plot 0.64 0.86 0.89 1.22 1.45 1.72</td>
</tr>
<tr>
<td>1 15 16 13.5 11 11.5 8</td>
</tr>
<tr>
<td>2 20 18.5 16 19.5 12 8.5</td>
</tr>
</tbody>
</table>
Excel? From pull-down menus:

Tools / Data Analysis / Regression
First step: Scatter plot. Look at data!

Scatter plot: reveals a linear relationship between the mean species density (Y) and altitude (X).

Note. Points are not perfectly aligned.
Task: draw a straight line that gives the “best possible fit”.

How to find the best fit?

Least squares (LS) line which minimizes the sum of squared differences b/w the line and the data points.
That is, minimizes the sum of squares for error (SSE):

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

where

- \( y \): Observed response
- \( \hat{y} \): Predicted response
  \[ (\hat{y} = b_0 + b_1 x) \]

Question:

How are \( b_0 \) and \( b_1 \) calculated?
The regression equation for a set of n data points is
\[ \hat{y} = b_0 + b_1 x, \]
where:

\[ b_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x} \]

Here:

\[ S_{xy} = \sum (x - \bar{x})(y - \bar{y}) \]
\[ = \sum xy - \frac{(\sum x)(\sum y)}{n} \]
\[ S_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} \]

Instead - Use Excel!
Example (Density of Plant Species)

SLR Model:

\[
\text{Species Density} = \beta_0 + \beta_1 \text{Altitude} + \varepsilon
\]

\[Y \quad X\]

LS estimate for this model?
Regression output - Excel:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SUMMARY OUTPUT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression Statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Multiple R</td>
<td>0.7817</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>R Square</td>
<td>0.6110</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Adjusted R Square</td>
<td>0.5721</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Standard Error</td>
<td>2.6539</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Observations</td>
<td>12.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>ANOVA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>df SS MS P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Regression</td>
<td>1 110.6289 110.6289 15.7068 0.0027</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Residual</td>
<td>10 70.4336 7.0434</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Total</td>
<td>11 181.0625</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>Coefficients Standard Error t Stat P-value Lower 95% Upper 95%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Intercept</td>
<td>23.3543 2.4515 3.5264 0.0000 17.8919 28.8167</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Altitude - X</td>
<td>-8.1675 2.0608 -3.9632 0.0027 -12.7594 -3.5757</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thus, from Excel output

\[ b_0 = 23.35 \quad \text{(Cell B17)} \]

\[ b_1 = -8.17 \quad \text{(Cell B18)} \]

and the estimated (prediction) equation is

\[ \hat{\text{density}} = 23.35 - 8.17 \text{ altitude} \]
What is $\hat{y}_{12}$? That is, what is the estimated value for the last observation in the sample?

When $x_{12} = 1.72$ (altitude is 1720 m) then

$$
\hat{y}_{12} = 23.35 - 8.17 x_{12}
$$

$$
= 23.35 - 8.17 \times 1.72 = 9.30
$$

and the estimated residual is

$$
e_{12} = y_{12} - \hat{y}_{12} = 8.5 - 9.3 = -0.80
$$
15.2 Inference for the slope ($\beta_1$) of the Population Regression Line:

Objective:

To assess whether a linear relationship exists b/w X and Y - Inference for $\beta_1$.

Test for linear relationship b/w Y and X?

$H_0 : \beta_1 = 0$ (No linear relationship)

$H_1 : \beta_1 \neq 0$ (linear relationship b/w X and Y exists- Could be +ve or -ve).

$\beta_1 > 0$ (+ve linear relationship)

$\beta_1 < 0$ (-ve linear relationship)
Test statistic?

Estimated slope \[ b \]
Slope under \( H_0 \)

\[ t = \frac{b - \beta_{10}}{s_{b_1}} = \frac{b_1}{s_{b_1}} \]

Standard Error for \( b_1 \)

To set up the rejection region, use t-tables with df=n-2.

**Example** (Species cont’ed)

Is there a linear relationship b/w density species and altitude? Use \( \alpha = .05 \).

\[ H_0 : \beta_1 = 0 \]
\[ H_1 : \beta_1 \neq 0 \]
\[ T = \frac{-8.17}{2.061} = -3.96 \quad \text{(Cell D18)} \]

What is the p-value associated with test?

From Excel printout:

p value is **ALWAYS** for a two tail test

p-value = 0.0027 \quad \text{(Cell E18)}

**Note:** If one tail test then adjust p-value accordingly.

**Decision**

Since p-value = 0.0027 < .05 = \( \alpha \) then reject \( H_0 \). There is evidence of a linear relationship b/w density and altitude.
We could also construct a CI for $\beta_1$:

$$b_1 \pm t_{\alpha/2, n-2} s_{b_1}$$

**Example:** (Species data)

Find a 95% confidence interval for $\beta_1$.

$$-8.17 \pm t_{10, 0.05/2} (2.061)$$

$$= -8.17 \pm 2.228 \times 2.061$$

$$= (-12.76, -3.58)$$

Hence: $-12.76 \leq \beta_1 \leq -3.58$

(Excel Cells: F18 and G18)

**Note:** Zero is not included in the interval, so there is evidence of linear relationship b/w species density and altitude.
14.3 The Coefficient of Determination:

- Defined as the proportion of total variation explained by the regression model and denoted as $R^2$.

\[ R^2 = \frac{\text{Variation explained by regression}}{\text{Total Variation}} \]

- Indicates usefulness of regression model for prediction.

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

- $R^2 \approx 0 \rightarrow$ Model not very useful for making predictions.

- $R^2 \approx 1 \rightarrow$ Indicates that model is very useful for making predictions.
Total Variation? - Denoted SST.

SST - Numerator of the variance of the dependent variable (y)

\[
SST = \sum (y - \bar{y})^2
\]

\[
= \sum y^2 - \frac{(\sum y)^2}{n}
\]

\[
= 181.06 \quad \text{(Excel, cell C14)}
\]

Variation explained by regression? -

Denoted: SSR \quad (= \sum (\hat{y} - \bar{y})^2)

\[
SSR = 110.63
\]

↑

Cell C12 in Excel Output
Coefficient of determination?

\[ R^2 = \frac{\text{Variation explained by regression}}{\text{Total Variation}} \]

\[ = \frac{\text{SSR}}{\text{SST}} \]

\[ = \frac{110.63}{181.06} \]

\[ = 0.6110 \rightarrow \text{Cell B5.} \]

Interpretation?

61.10% of total variation in y (species density) is explained by altitude.