Multiple Regression Analysis

- **Regression Analysis:** Statistical procedure for quantifying (economic) relationships and testing hypothesis about them.
  - A regression model specifies a (usually linear) relationship between a dependent variable and an independent (or explanatory) variable an error term $u$
    $$Y = \beta_0 + \beta_1 X + u$$
  - **Multiple Regression Analysis:** When there is more than one explanatory variable.

Multiple Regression Analysis

- Estimates the relationship between a dependent variable (left-hand side) and independent variables (right-hand side).
  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$
  - $Y$ is the dependent variable, that should be explained by the independent variables $X_1$ and $X_2$.
  - Dependent and independent variables can be observed (there are data available for them).
  - The parameters $\beta_0$, $\beta_1$ and $\beta_2$ cannot be observed, but they can be estimated by linear regression from the data $\rightarrow$ Parameter estimation.

- The parameters $\beta_0$, $\beta_1$ and $\beta_2$ describe the relation between dependent and independent variables:
  $$\frac{\partial Y}{\partial X_1} = \beta_1, \quad \frac{\partial Y}{\partial X_2} = \beta_2$$

  Parameters describe marginal effects, e.g. if $X_2$ increases by 1 unit $Y$ will change by $\beta_2$ units!
  - Can be done with spreadsheet software.
Multiple Regression Analysis

- For example, a demand function
  \[ Q = \beta_0 - \beta_1 P + \beta_2 Adv \]
  \((Q \ldots \text{quantity sold}, P \ldots \text{Price}, \text{Adv} \ldots \text{Spending for Advertisement})\)
- can be estimated, if data for \(Q\), \(P\) and \(Adv\) are available, e.g.
  \[ Q = 116.12 - 1.31 P + 11.25 Adv \]
- describes the relationship between the dependent variable “Quantity demanded” (left-hand side) and the independent variables “Price” and “Advertisement” (right-hand side).

Descriptive Regression Analysis: The Population Regression Function (PRF)

Regression analysis allows to calculate \(\beta_0\) and \(\beta_1\)!

We have observations (data) on variables \(Y\) and \(X\) …

<table>
<thead>
<tr>
<th>(i)</th>
<th>(X)</th>
<th>(Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.2</td>
<td>2.6</td>
</tr>
<tr>
<td>2</td>
<td>3.0</td>
<td>1.6</td>
</tr>
<tr>
<td>3</td>
<td>4.5</td>
<td>4.0</td>
</tr>
<tr>
<td>4</td>
<td>5.8</td>
<td>3.0</td>
</tr>
<tr>
<td>5</td>
<td>7.2</td>
<td>4.9</td>
</tr>
</tbody>
</table>

\[ Y = Y(X) = \beta_0 + \beta_1 X \]

Regression analysis allows to calculate \(\beta_0\) and \(\beta_1\)!

Multiple Regression Analysis

- The coefficient of \(P\) tells us, that – holding Advertisement constant – an increase of price for one unit will decrease (negative sign!) the expected quantity sold by 1.31 units.
- The coefficient of \(Adv\) tells us, that – holding Price constant – an increase of the spending for Advertisement by one unit will increase (positive sign!) the expected quantity sold by 11.25 units.
Regression Analysis

- We decompose the observed variable $Y$ (e.g. quantity sold) in a deterministic and a stochastic component.

$$ Y_i = \beta_0 + \beta_1 X_i + u_i $$

- The deterministic component is explained by the model (i.e. the part of $Y$ that is explained by $X$) and usually denoted with a hat on the variable (e.g. $\hat{Y}$).
- The random component is not explained by the model and called the error term (often denoted with $u$ or $\varepsilon$).

Decomposition of $Y$

$Y_i = f(X_i)$?

$Y_i = \hat{Y}_i + u_i$

$\hat{Y}_i$: systematic

$u_i$: stochastic

Calculation of $\beta_0$ and $\beta_1$

- Regression analysis is based on the minimization of the sum of squared errors!
- We cannot use the simple sum of error terms, because positive and negative errors would compensate for each other. Therefore, we use the sum of squared errors.
- This procedure allows us to calculate the parameters $\beta_0$ and $\beta_1$. Graphically . . .
Calculation of $\beta_0$ and $\beta_1$

Mathematically:
minimizing the sum of the squared error terms

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{N} u_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

Solution to this minimization problem:

$$\beta_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

where $\bar{Y} \equiv 1/N \sum_i Y_i$ and $\bar{X} \equiv 1/N \sum_i X_i$.

Multiple Regression

These results can easily extended to more than 2 variables,

e.g. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$

but we need matrix algebra for the formulas (Excel can do it!).

For multiple regressions the coefficients give the marginal effect of a variable, keeping all other variables constant (ceteris paribus)!

e.g. $$\frac{\partial Y}{\partial X_1} = \beta_1$$

if $X_1$ changes by one unit we expect $Y$ to change by $\beta_1$ units, while $X_2, \ldots, X_k$ are held constant!

Coefficient of Determination

- **Coefficient of Determination ($R^2$):** A measure of how the overall estimating equation fits the data.
- It shows the fraction of the variation in the dependent variable, which is explained statistically by the variables included in the regression.

Coefficient of Determination $R^2$

- The Coefficient of Determination $R^2$ gives the share of the variation in $Y$ which is explained by the regression (i.e., by all $X$ variables).
- It can be shown that $R^2$ is the square of (Pearson’s) correlation coefficient between the observed values of $Y$ and the fitted values $\hat{Y}$.
- If the regression has a constant term (intercept) $R^2$ is always between 1 (best possible fit) and 0 (worst possible fit).
Regressions with different $R^2$:

![Graphs showing different $R^2$ values](image)

Population & Sample

Inferential Regression Analysis:
*The Sample Regression Function (SRF)*

Population & Sample

- Up to now we did only descriptive analysis:
  Which $\beta_0$ and $\beta_1$ describes the relationship in the population 'best'?  
- But: only rarely we can observe the population, usually we observe only samples!
- Question: How can we infer from the sample on the population?
- Example: we assume a population of 20 subjects and draw 3 samples thereof.

<table>
<thead>
<tr>
<th>N</th>
<th>X</th>
<th>Y</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.2</td>
<td>0.8</td>
<td>c</td>
</tr>
<tr>
<td>2</td>
<td>1.3</td>
<td>2.9</td>
<td>b</td>
</tr>
<tr>
<td>3</td>
<td>1.7</td>
<td>0.8</td>
<td>c</td>
</tr>
<tr>
<td>4</td>
<td>1.9</td>
<td>0.7</td>
<td>a</td>
</tr>
<tr>
<td>5</td>
<td>1.9</td>
<td>3.5</td>
<td>b</td>
</tr>
<tr>
<td>6</td>
<td>2.0</td>
<td>3.2</td>
<td>a</td>
</tr>
<tr>
<td>7</td>
<td>2.2</td>
<td>2.6</td>
<td>c</td>
</tr>
<tr>
<td>8</td>
<td>2.4</td>
<td>3.2</td>
<td>c</td>
</tr>
<tr>
<td>9</td>
<td>2.8</td>
<td>3.6</td>
<td>b</td>
</tr>
<tr>
<td>10</td>
<td>2.8</td>
<td>1.1</td>
<td>a</td>
</tr>
</tbody>
</table>

20 observations: $Y = f(X)$?
Sampling & Sampling Error

Population: 20 observations
relationship in the population

\[ \hat{Y}_i = 1.07 + 0.52X_i \]

“True” relation in the population and errors \( u \) of the population

\[ \hat{Y}_i = 1.07 + 0.52X_i + e_i \]

Random-Sample “a” \( (N^a = 6) \)

\[ Y_i^a = 0.98 + 0.39X_i + e_i^a \]

Random-Sample “b” \( (N^b = 7) \)

\[ Y_i^b = -0.12 + 0.91X_i + e_i^b \]
For each random sample we get different coefficients $b_0$ and $b_1$!

Estimated coefficients have a sampling distribution:

This allows us to do statistics!
Regression Analysis & Statistics

Notation: in what follows we use different symbols for values of the population and values estimated out of a sample.

- Coefficients:
  we use $\beta$ for the population and $b$ for the sample. Sometimes you also see $\hat{\beta}$ for sample-estimates.

- Residuals:
  we use $u$ for the population errors and $e$ (or $\hat{u}$) for the sample residuals.

- Residuals are the sample analogue to the population errors!

Single Hypothesis

- Remember, for each sample we draw we get different estimates $b_0$ and $b_1$ for the true – but unknown – $\beta_0$ and $\beta_1$.

- Therefore, the regression coefficients $b_0, b_1, \ldots, b_K$ are itself random variables.

- We have determined the sampling distribution of these random variables: If the population error term is normally distributed, or the sample is sufficiently large (say, $N > 30$), the the estimated coefficients are also (approximately) normally distributed.

- We can calculate the standard errors of these coefficients und do hypothesis tests as usual.

Regression Analysis & Statistics

- Standard Error of coefficient $b_k$: $s_{b_k}$
  A measure of the precision of an estimated regression analysis coefficient. It shows how much the coefficient would vary in regressions from different samples.

- If the standard error is large relative to the coefficient, we should not trust the coefficient!
The $t$-statistic to test a single hypothesis is usually calculated as follows:

$$ t = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error of estimate}} $$

In most cases: $H_0 : \beta_k = 0$, therefore

$$ t = \frac{b_k}{s_b_k} \text{ with } N - K \text{ degrees of freedom.} $$

with $s_{b_k}$: standard error of $b_k$, $N$: number of observations, $K$: number of estimated coefficients

This is the $t$-statistic given in the usual regression output.

**t-Test:** A test based on the ratio of the estimated regression coefficient to its standard error that is used to determine the statistical significance of the coefficient, i.e., to test whether the estimated coefficient is statistically significantly different from zero:

$$ t = \frac{b_k}{s_{b_k}} $$

If the absolute value of the calculated $t$-statistic is larger than the critical value of the theoretical $t$-statistic we can reject the null-hypothesis that there is no relationship between this explanatory variable and the dependent variable, i.e. we say the coefficient is *significant*.

**p-Value:** gives the probability to find a coefficient of the estimated magnitude in a random sample, when in the population is in fact no relationship between the variables (Type I error). $p$-value is the probability we would observe the $t$ statistic we did, if the null hypothesis were true. The $p$-value is usually based in a $t$-test, it gives essentially the same information. For the usual 5 percent significance level the $p$-value should be *smaller than* 0.05!
**F Statistic**

- The F statistic given in the regression output of most programs is used to calculate the probability, that all right-hand (independent or X-) variables except the intercept are *jointly* different from zero, i.e.

  \[ H_0: \beta_1 = \beta_2 = \ldots = \beta_K = 0 \]

- Again, the p-value of this F statistic is the probability we would observe the F statistic we did, if the null were true.

- If this null hypothesis cannot be rejected the regression is of course rather worthless.

---

**Typical Regression Output**

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1826.161</td>
<td>359.131</td>
<td>5.064914</td>
</tr>
<tr>
<td>P</td>
<td>-221.3908</td>
<td>17.14239</td>
<td>-12.91481</td>
</tr>
<tr>
<td>M</td>
<td>0.071657</td>
<td>0.016744</td>
<td>4.279656</td>
</tr>
</tbody>
</table>

R-squared: 0.906682  Mean dependent var: 1730.417
Adjusted R-squared: 0.897795  S.D. dependent var: 181.7550
S.E. of regression: 58.10628  Sum squared resid: 70903.13
F-statistic: 102.0186  Durbin-Watson stat: 1.313079
Prob(F-statistic): 0.000000

For publications usually written as:

\[ Q = 1826.161 - 221.3908 P + 0.072 M \]

\[ R^2 = 0.907, \quad s = 58.106, \quad F-\text{Stat} = 102.019, \quad N = 24 \]

---

**Excel Regression Output**

<table>
<thead>
<tr>
<th>Dependent Variable: Q</th>
<th>Method: Least Squares</th>
<th>Included observations: 24</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Coefficient</strong></td>
<td><strong>Std. Error</strong></td>
<td><strong>t-Statistic</strong></td>
</tr>
<tr>
<td>C</td>
<td>1826.161</td>
<td>359.131</td>
</tr>
<tr>
<td>P</td>
<td>-221.3908</td>
<td>17.14239</td>
</tr>
<tr>
<td>M</td>
<td>0.071657</td>
<td>0.016744</td>
</tr>
</tbody>
</table>

**Excel Regression Output**

**Functional Form**
Functional Form

- OLS can be used for relationships that are not strictly linear in $X$ and $Y$ by using nonlinear functions of $X$ and $Y$ – will still be linear in the parameters.
- You can take . . .
  - the natural log of $X$, $Y$ or both.
  - quadratic forms of $X$.
  - interactions of $X$ variables (e.g. $X_1 \times X_3$)
- However, the interpretation of the coefficient differs

Isoelastic Demand

- If the variables are in logarithms the coefficients are elasticities; i.e. log-linear curves have constant elasticities!
- When $\log(Y) = \beta_0 + \beta_1 \log(X)$ then $\beta_1$ is the constant elasticity between $X$ and $Y$.
- Example: Log-linear demand curve:
  \[
  \log(Q) = -0.23 - 0.34 \log(P) + 1.32 \log(I)
  \]
  Price elasticity $= -0.34$, Income elasticity $= 1.32$

Interpretation of log-models

- If the model is log-log
  \[
  \ln(Y) = b_0 + b_1 \ln(X) + e
  \]
  $b_1$ is the elasticity of $Y$ with respect to $X$.
- If the model is log-lin
  \[
  \ln(Y) = b_0 + b_1 X + e
  \]
  $b_1 \times 100$ is approximately the percentage change in $Y$ given a 1 unit change in $X$. 

Log-Linear Regression Model

[Graph showing log-linear relationship between variables]
Why use log models?

- Log models are invariant to the scale of the variables since measuring percent changes.
- Log-Log models give a direct estimate of elasticity.
- The distribution of \( \ln(Y) \) is more narrow, limiting the effect of outliers and heteroskedasticity.
- Attention: Can only be used for positive \( Y \) (log of negative numbers is not defined!).

Some Rules of Thumb

- What types of variables are often used in log form?
  - Baht or dollar amounts that must be positive.
  - Very large variables, such as population.
  - Variables that have a natural growth rate.
- What types of variables are often used in level form?
  - Variables that are a proportion or percent.
  - Growth rates or changes.

Quadratic Models

- Quadratic regression models are appropriate when the curve fitting the scatter plot is either \( \cap \)-shaped or \( \cup \)-shaped.
- A quadratic equation, \( Y = a + bX + cX^2 \), can be transformed into a linear form by computing a new variable \( Z = X^2 \), which is then substituted for \( X^2 \) in the regression. Then, the regression equation to be estimated is
  \[ Y = a + bX + cZ \]

Quadratic Models

- For a model of the form
  \[ Y = b_0 + b_1X + b_2X^2 + e \]
  we can’t interpret \( b_1 \) alone as measuring the change in \( Y \) with respect to \( X \),
  we need to take into account \( b_2 \) as well, since
  \[ \frac{dY}{dX} = b_1 + 2b_2X \]
  if \( b_1 \) is positive and \( b_2 \) is negative then \( Y \) is increasing in \( X \) at first, but will eventually turn around and be decreasing in \( X \).