# Multiple Regression: Example

# Cobb-Douglas Production Function

The Cobb-Douglas production function for observed economic data $i = 1, \ldots, n$ may be expressed as

$$O_i = e^{\beta_0} l_i^{\beta_1} c_i^{\beta_2} u_i$$

where

- $O_i$ is output
- $l_i$ is labour input
- $c_i$ is capital input
- $u_i$ is a random error term

# Cobb-Douglas Production Function (cont.)

Taking natural logs, we have that

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

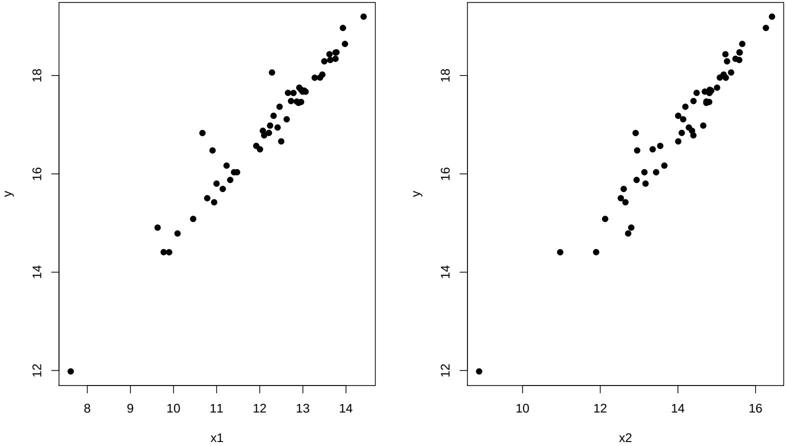
where

- $Y_i = \ln(O_i)$ is log output
- $x_{i1} = \ln(l_i)$ is log labour input
- $x_{i2} = \ln(c_i)$ is log capital input
- $\epsilon_i = \ln(u_i)$ is a random error term

We will term this model the “complete" model.

# Data: 50 US states plus Dist. of Columbia.

Manufacturing sector, 2005.




Note that also $x_1$ and $x_2$ are highly positively correlated:

```
> cor(x1,x2)
[1] 0.960402
```

# Analysis in R

```
> fit12<-lm(y~x1+x2,data=Cobb); summary(fit12)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.88760    0.39623   9.812 4.70e-13 ***
x1           0.46833    0.09893   4.734 1.98e-05 ***
x2           0.52128    0.09689   5.380 2.18e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2668 on 48 degrees of freedom
Multiple R-squared:  0.9642,    Adjusted R-squared:  0.9627
F-statistic: 645.9 on 2 and 48 DF,  p-value: < 2.2e-16

> summary(fit12)$sigma
[1] 0.2667521
```

We see from this analysis that

$$\mathrm{SS}_{\mathrm{Res}} \equiv \mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2) = (n-p)\widehat{\sigma}^2 = 48 \times 0.2667521^2 = 3.41552$$

which can be extracted as

```
> summary(fit12)$df[2]*summary(fit12)$sigma^2
[1] 3.41552
```

# Analysis in R: anova

```
18  > anova(fit12)
19  Analysis of Variance Table
20
21  Response: y
22            Df Sum Sq Mean Sq  F value    Pr(>F)
23  x1         1 89.865  89.865 1262.915 < 2.2e-16 ***
24  x2         1  2.060   2.060   28.947 2.183e-06 ***
25  Residuals 48  3.416   0.071
26  ---
27  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we have the decomposition

$$\overline{\mathrm{SS}}_{\mathrm{R}}(\beta_1, \beta_2 | \beta_0) = \overline{\mathrm{SS}}_{\mathrm{R}}(\beta_1 | \beta_0) + \overline{\mathrm{SS}}_{\mathrm{R}}(\beta_2 | \beta_0, \beta_1)$$

where

- line 23 (`Sum Sq`) : $\overline{\mathrm{SS}}_{\mathrm{R}}(\beta_1 | \beta_0) = 89.865$;
- line 24 (`Sum Sq`) : $\overline{\mathrm{SS}}_{\mathrm{R}}(\beta_2 | \beta_0, \beta_1) = 2.060$

Note from line 25 (`Sum Sq`), $\mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2) = 3.416$ as before.

# Analysis in R: anova

```
28 > fit21<-lm(y~x2+x1,data=Cobb)
29 > anova(fit21)
30 Analysis of Variance Table
31
32 Response: y
33           Df Sum Sq Mean Sq  F value    Pr(>F)
34 x2         1 90.330  90.330 1269.450 < 2.2e-16 ***
35 x1         1  1.595   1.595   22.412 1.981e-05 ***
36 Residuals 48  3.416   0.071
37 ---
38 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we have the decomposition

$$\overline{\text{SS}}_{\text{R}}(\beta_1, \beta_2|\beta_0) = \overline{\text{SS}}_{\text{R}}(\beta_2|\beta_0) + \overline{\text{SS}}_{\text{R}}(\beta_1|\beta_0, \beta_2)$$

where

- line 34 (`Sum Sq`) : $\overline{\text{SS}}_{\text{R}}(\beta_2|\beta_0) = 90.330$;
- line 35 (`Sum Sq`) : $\overline{\text{SS}}_{\text{R}}(\beta_1|\beta_0, \beta_2) = 1.595$

Again from line 36 (`Sum Sq`), $\text{SS}_{\text{Res}}(\beta_0, \beta_1, \beta_2) = 3.416$ as before.

# Analysis in R: *F*-tests

The *F*-tests carried out using `anova` are partial *F*-tests. From the first analysis

```
39  > anova(fit12)
40  Analysis of Variance Table
41  Response: y
42            Df Sum Sq Mean Sq  F value    Pr(>F)
43  x1         1 89.865  89.865 1262.915 < 2.2e-16 ***
44  x2         1  2.060   2.060   28.947 2.183e-06 ***
45  Residuals 48  3.416   0.071
```

The test on line 43 is the comparison of the models

$$
\begin{aligned}
\text{“Reduced"} &: \quad \mathbb{E}[Y_i|\mathbf{x}_i] = \beta_0 \\
\text{“Full"} &: \quad \mathbb{E}[Y_i|\mathbf{x}_i] = \beta_0 + \beta_1 x_{i1}
\end{aligned}
$$

whilst recognizing that $x_2$ may also be used to estimate $\sigma^2$.

# Analysis in R: $F$-tests

We compute

$$F = \frac{(\mathrm{SS}_{\mathrm{Res}}(\beta_0) - \mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1))/r}{\mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2)/(n-p)}$$

where

- $p = 3$ (number of coefficients in the “complete" model)
- $r = 1$ (number of coefficients set to zero in the “full" model to obtain the “reduced" model)

# Analysis in R: *F*-tests

We may access these elements in R as follows:

```
46 >SSRes0<-anova(lm(y~1,data=Cobb))[1,2]
47 >MSRes012<-anova(lm(y~x1+x2,data=Cobb))[3,3]
48 >SSRes01<-anova(lm(y~x1,data=Cobb))[2,2]
49 >F<-((SSRes0-SSRes01)/1)/MSRes012
```

The `anova` function returns a matrix, and we must access elements of the matrix using the R notation `[1,2]`,`[3,3]` and `[2,2]` respectively.

This yields

```
50 > SSRes0
51 [1] 95.34013
52 > MSRes012
53 [1] 0.07115667
54 > SSRes01
55 [1] 5.475317
56 > F
57 [1] 1262.915
```

which matches the result on line 43 (`F value`).

The test on line 44 is the comparison of the models

$$\begin{aligned} \text{“Reduced"} &: \quad \mathbb{E}[Y_i|\mathbf{x}_i] &=& \quad \beta_0 + \beta_1 x_{i1} \\ \text{“Full"} &: \quad \mathbb{E}[Y_i|\mathbf{x}_i] &=& \quad \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} \end{aligned}$$

We compute

$$F = \frac{(\mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1) - \mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2))/r}{\mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2)/(n-p)}$$

where

- $p = 3$ (number of coefficients in the “complete" model)
- $r = 1$ (number of coefficients set to zero in the “full" model to obtain the “reduced" model)

# Analysis in R: *F*-tests

We may access these elements in R as follows:

```
58 > SSRes01<-anova(lm(y~x1,data=Cobb))[2,2]
59 > MSRes012<-anova(lm(y~x1+x2,data=Cobb))[3,3]
60 > SSRes012<-anova(lm(y~x1+x2,data=Cobb))[3,2]
61 > F<-((SSRes01-SSRes012)/1)/MSRes012
62 >
63 > SSRes0
64 [1] 95.34013
65 > MSRes012
66 [1] 0.07115667
67 > SSRes01
68 [1] 5.475317
69 > F
70 [1] 28.94735
```

which matches the result on line 44 (`F value`).

# Analysis in R: $F$-tests

The $F$-value on line 34 performs the partial $F$-test for testing

$$
\begin{aligned}
\text{“Reduced"} &: \quad \mathbb{E}[Y_i|\mathbf{x}_i] &= \quad \beta_0 \\
\text{“Full"} &: \quad \mathbb{E}[Y_i|\mathbf{x}_i] &= \quad \beta_0 + \beta_2 x_{i2}
\end{aligned}
$$

whilst recognizing that $x_1$ may also be used to estimate $\sigma^2$ using the statistic

$$F = \frac{(\mathrm{SS}_{\mathrm{Res}}(\beta_0) - \mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_2))/r}{\mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2)/(n-p)}$$

```
> SSRes0<-anova(lm(y~1,data=Cobb))[1,2]
> MSRes012<-anova(lm(y~x1+x2,data=Cobb))[3,3]
> SSRes02<-anova(lm(y~x2,data=Cobb))[2,2]
> (F<-((SSRes0-SSRes02)/1)/MSRes012)
[1] 1269.45
```

# Analysis in R: $F$-tests

The $F$-value on line 35 performs the partial $F$-test for testing

$$
\begin{array}{lcrcl}
\text{“Reduced"} & : & \mathbb{E}[Y_i|\mathbf{x}_i] & = & \beta_0 + \beta_2 x_{i2} \\
\text{“Full"} & : & \mathbb{E}[Y_i|\mathbf{x}_i] & = & \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}
\end{array}
$$

using the statistic

$$F = \frac{(\mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_2) - \mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2))/r}{\mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2)/(n-p)}$$

```
76  > SSRes02<-anova(lm(y~x2,data=Cobb))[2,2]
77  > MSRes012<-anova(lm(y~x1+x2,data=Cobb))[3,3]
78  > SSRes012<-anova(lm(y~x1+x2,data=Cobb))[3,2]
79  > (F<-((SSRes02-SSRes012)/1)/MSRes012)
80  [1] 22.41237
```

# Analysis in R: $F$-tests

The conclusions of the above analyses are that

- when we start with $x_1$ in the model, and try to add $x_2$, **there is a significant improvement in fit**; we see this from line 44: the $p$-value is `2.183e-06`
- when we start with $x_2$ in the model, and try to add $x_1$, **there is a significant improvement in fit**; we see this from line 35: the $p$-value is `1.981e-05`

# Analysis in R: $F$-tests

Note that, if we considered $x_2$ irrelevant from the start, we might omit it from any analysis and consider the alternative "complete" model.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i.$$

Then to test

$$\begin{array}{lcrcl} \text{"Reduced"} & : & \mathbb{E}[Y_i|\mathbf{x}_i] & = & \beta_0 \\ \text{"Full"} & : & \mathbb{E}[Y_i|\mathbf{x}_i] & = & \beta_0 + \beta_1 x_{i1} \end{array}$$

we would compute

$$F = \frac{(\mathrm{SS_{Res}}(\beta_0) - \mathrm{SS_{Res}}(\beta_0, \beta_1))/r}{\mathrm{SS_{Res}}(\beta_0, \beta_1)/(n-p)}$$

where now $p = 2$.

# Analysis in R: $F$-tests

```
> summary(lm(y~x1,data=Cobb))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.99902    0.42371   11.80 6.29e-16 ***
x1           0.97950    0.03454   28.36  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3343 on 49 degrees of freedom
Multiple R-squared:  0.9426,    Adjusted R-squared:  0.9414
F-statistic: 804.2 on 1 and 49 DF,  p-value: < 2.2e-16

> anova(lm(y~x1,data=Cobb))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 89.865  89.865  804.22 < 2.2e-16 ***
Residuals 49  5.475   0.112
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Analysis in `R`: $F$-tests

The numerical result (`804.22`) on lines 91 (`F-statistic`) and 98 (`F value`) is different from that on lines 43 and 57 (`1262.915`).

Both $F$-tests compare

$$
\begin{aligned}
\text{"Reduced"} &: \quad \mathbb{E}[Y_i|\mathbf{x}_i] &= \quad & \beta_0 \\
\text{"Full"} &: \quad \mathbb{E}[Y_i|\mathbf{x}_i] &= \quad & \beta_0 + \beta_1 x_{i1}
\end{aligned}
$$

however, the results on line 43 and 57 acknowledge a possible influence of $x_2$; this leads to a reduction the $\mathrm{MS}_{\mathrm{Res}}$ quantity which is in the denominator of the $F$-statistic.

# Analysis in R: $F$-tests

To assess the importance of each of the variables $x_1$ and $x_2$ directly, we may use the `drop1` command:

```
> fit12<-lm(y~x1+x2,data=Cobb)
> drop1(fit12,test='F')
Single term deletions

Model:
y ~ x1 + x2
       Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>              3.4155 -131.88
x1      1    1.5948 5.0103 -114.34  22.412 1.981e-05 ***
x2      1    2.0598 5.4753 -109.81  28.947 2.183e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

reproducing the results on lines 35 and 44 respectively.