Lecture 8. Using the CLR Model

Relation between patent applications and R&D spending

Variables

• PATENTS = No. of patents (in 1000) filed
• RDEXP = Expenditure on research & development (in billions of 1992 $)

The data are time series for 1960-1993 (34 observations)

First step in analysis: descriptive statistics and graphs

• Sample mean, standard deviation etc. of variables
• With time-series data, time-series plot (scale!). Note units are chosen to make the range of the variables comparable
• Scatterplot of PATENTS against RDEXP
<table>
<thead>
<tr>
<th></th>
<th>PATENTS</th>
<th>RDEXP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>119.2382</td>
<td>106.9118</td>
</tr>
<tr>
<td>Median</td>
<td>109.5000</td>
<td>89.46000</td>
</tr>
<tr>
<td>Maximum</td>
<td>189.4000</td>
<td>166.7000</td>
</tr>
<tr>
<td>Minimum</td>
<td>84.50000</td>
<td>57.94000</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>29.30583</td>
<td>34.29868</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.227838</td>
<td>0.561985</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>3.409492</td>
<td>1.900218</td>
</tr>
<tr>
<td>Jarque-Bera</td>
<td>8.780538</td>
<td>3.503174</td>
</tr>
<tr>
<td>Probability</td>
<td>0.012397</td>
<td>0.173498</td>
</tr>
<tr>
<td>Observations</td>
<td>34</td>
<td>34</td>
</tr>
</tbody>
</table>
Second step: estimate the relation

\[ PATENTS_t = \alpha + \beta RDEXP_t + u_t, \]

(Note that I use subscript \( t \) to reflect the nature of the data)

We assume that this inexact linear relation satisfies the assumption of the Classical Linear Regression (CLR) model

Assumption 1: \( u_t, t = 1, \ldots, n \) are random variables with \( E(u_t) = 0 \)

Assumption 2: \( X_t, t = 1, \ldots, n \) are deterministic, i.e. non-random, constants.

Assumption 3 (Homoskedasticity)
All \( u_t \)'s have the same variance, i.e. for \( t = 1, \ldots, n \)

\[ Var(u_t) = E(u_t^2) = \sigma^2 \]

Assumption 4 (No serial correlation)
The random errors \( u_t \) and \( u_s \) are not correlated for all \( t \neq s = 1, \ldots, n \)

If these assumptions hold then the best estimator for \( \alpha, \beta \) is the OLS estimator.

The output of a computer program that computes the OLS estimates and other quantities is reproduced next.

The formulas that the computer uses are

OLS estimates

\[ \hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \]

\[ \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} \]
Standard errors of estimates and standard error of regression

\[
\text{Std}(\hat{\alpha}) = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} (X_i - \bar{X})^2}} \times s^2
\]

\[
\text{Std}(\hat{\beta}) = \sqrt{\frac{s^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}
\]

\[
s^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2
\]

t-Statistic

\[
T_\alpha = \frac{\hat{\alpha}}{\text{Std}(\hat{\alpha})} \quad T_\beta = \frac{\hat{\beta}}{\text{Std}(\hat{\beta})}
\]

\[
R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}
\]
Dependent Variable: PATENTS
Method: Least Squares
Date: 09/26/01   Time: 11:27
Sample: 1960 1993
Included observations: 34

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>34.57106</td>
<td>6.357873</td>
<td>5.437521</td>
<td>0.0000</td>
</tr>
<tr>
<td>RDEXP</td>
<td>0.791935</td>
<td>0.056704</td>
<td>13.96621</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

R-squared 0.859065   Mean dependent var 119.2382
Adjusted R-squared 0.854661   S.D. dependent var 29.30583
S.E. of regression 11.17237   Akaike info criterion 7.721787
Sum squared resid 3994.300   Schwarz criterion 7.811573
Log likelihood -129.2704   F-statistic 195.0551
Durbin-Watson stat 0.233951   Prob(F-statistic) 0.000000
Computer output

Estimates:
Interpretation coefficient RDEXP: if RDEXP changes by 1 unit, i.e. 1 billion of 1992 $, then PATENTS changes by .7919, i.e. by about 792.

Estimates and standard errors:
We use the standard error to find a 95% confidence interval for $\hat{\beta}$. The formula for the bounds is

$$\hat{\beta} \pm c \text{Std}(\hat{\beta})$$

with $c$ the value such that $Pr(T > c) = .025$ where $T$ has a $t$ distribution with 34-2=32 degrees of freedom (df). From table in front cover $c \approx 2.038$. Hence, the 95% interval:

$$[.6764, .9075]$$

Note if we use $c = 2$ (rule of thumb) or $c = 1.96$ (standard normal), the result is close.

Test:
We test $H_0: \beta = 0$ against $H_1: \beta \neq 0$ (two-sided test). If $H_0: \beta = 0$ is true, then

$$T_0 = \frac{\hat{\beta}}{\text{Std}(\hat{\beta})}$$

has a $t$ distribution with 34-2=32 degrees of freedom (df). We reject if $|T_0| > c$, with $C$ such that $Pr(T > c) = .025$ for $t$ distribution with 32 df. Hence $c \approx 2.038$. If $H_1: \beta > 0$, then $c$ is such that $Pr(T > c) = .05$ for $t$ distribution with 32 df or $c \approx 1.694$. For both $H_1$’s we reject $H_0$.

Alternative way to report result of test is p-value

$$\text{p-value} = Pr(|T_0| > t_0) = Pr(|T| > 5.4375) = .00000556$$

where $T_0$ has the distribution that holds if $H_0$ is true, i.e. $t$ distribution with 32 df. If the p-value is less than .05, we reject at the 5% level.

Note F-statistic=$(t$-statistic $\hat{\beta})^2$

Fit of linear relation:
The R2 seems high, but as we shall see that does not imply that all is well.
Reporting regression results

Always report

- Estimates of the regression coefficients
- Their standard errors (preferred to t-ratios. Why?)
- Estimate of $\sigma^2$ or $\sigma$, i.e. $s^2$ or $s$
- $R^2$ (least useful, but most users want to know)

If the number variables is small two options in reporting the results

As equation

$$PATENTS = 34.57 + 0.79 \, RDEXP$$

$$s = 11.17 \quad R^2 = 0.86$$

Or in a table

OLS estimates (standard errors) regression
of no. of patents (in 1000) on R&D expenditure (in billion 1992 $); 1960-1993

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>34.57</td>
</tr>
<tr>
<td></td>
<td>(6.37)</td>
</tr>
<tr>
<td>RDEXP</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>(0.057)</td>
</tr>
<tr>
<td>Std. Error regression</td>
<td>11.17</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.86</td>
</tr>
</tbody>
</table>
How well does the model fit?

With time-series we can plot the observed $Y_i$ and fitted/predicted $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ and also $e_i = Y_i - \hat{Y}_i$ (see graph)

The fit seems to be good, but there is a clear pattern in the OLS residuals.

In the scatterplot I plot $e_t$ against $e_{t-1}$, i.e. I investigate whether subsequent OLS residuals are correlated.

Note the clear relation that indicates that Assumption 4 (no serial correlation) may not be correct.

To check assumption 3 I plot $e_t^2$ against $RD\text{EXP}_t$. There is evidence of heteroskedasticity, i.e. the variance of $u_t$ (estimated by $e_t^2$) is related to $X_t$ and not the same for all $t$.

To check the normal distribution of $u_t$ I plot the distribution of $e_t$. 

RESID2 vs. RDEXP
Kernel Density (Epanechnikov, $h = 9.2422$)
Forecasting

If \( RDEXP_{1994} = 170 \) (not in sample), then we predict

\[
PATENTS_{1994} = 34.57 + .792 \times 170 = 169.2
\]

The variance of the prediction error is

\[
s^2_{n+1} = s^2 \left( 1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)
\]

The square root (standard deviation of prediction error) is 11.89 (compare with \( s = 11.17 \))

A 95% prediction interval is, using

\[
\hat{Y}_{n+1} - 2.038 s_{n+1} < Y_{n+1} < \hat{Y}_{n+1} + 2.038 s_{n+1}
\]

equal to

\[
144.97 < PATENTS_{1994} < 193.43
\]
Now a rather curious relation

Births= No. of births (in 10000), West-Germany

Storks=No. of stork couples in Baden-Württemberg

Time-series 1969-1980
BIRTHS vs. STORKS

The graph shows a positive correlation between births and stork sightings, indicating that as the number of storks increases, the number of births also increases. The linear trend suggests a potential ecological or statistical relationship that warrants further investigation.
Dependent Variable: BIRTHS  
Method: Least Squares  
Date: 09/26/01   Time: 23:43  
Sample: 1969 1980  
Included observations: 12

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>38.37181</td>
<td>3.063321</td>
<td>12.52621</td>
<td>0.0000</td>
</tr>
<tr>
<td>STORKS</td>
<td>1.169960</td>
<td>0.119089</td>
<td>9.824221</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

R-squared 0.906117  Mean dependent var 66.84083  
Adjusted R-squared 0.896728  S.D. dependent var 10.70629  
S.E. of regression 3.440562  Akaike info criterion 5.460159  
Sum squared resid 118.3747  Schwarz criterion 5.540976  
Log likelihood -30.76095  F-statistic 96.51533  
Durbin-Watson stat 1.010130  Prob(F-statistic) 0.000002
Dependent Variable: DBIRTHS  
Method: Least Squares  
Date: 09/26/01   Time: 23:51  
Sample(adjusted): 1970 1980  
Included observations: 11 after adjusting endpoints

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>-1.293593</td>
<td>0.948663</td>
<td>-1.363595</td>
<td>0.2058</td>
</tr>
<tr>
<td>DSTORKS</td>
<td>0.669070</td>
<td>0.211647</td>
<td>3.161254</td>
<td>0.0115</td>
</tr>
</tbody>
</table>

R-squared 0.526154  Mean dependent var -2.570909
Adjusted R-squared 0.473505  S.D. dependent var 3.923245
S.E. of regression 2.846706  Akaike info criterion 5.093168
Sum squared resid 72.93361  Schwarz criterion 5.165512
Log likelihood -26.01242  F-statistic 9.993527
Durbin-Watson stat 1.528759  Prob(F-statistic) 0.011527
Lecture 9. The multiple Classical Linear Regression model

Often we want to estimate the relation between a dependent variable $Y$ and say $K$ independent variables $X_1, \ldots, X_K$.

Even with $K$ explanatory variables, the linear relation is not exact:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + u$$

with $U$ the random error term that captures the effect of the omitted variables.

Note if the relation has an intercept, then $X_1 \equiv 1$ and $\beta_1$ is the intercept of the relation.

If we have $n$ observations $Y_i, X_{i1}, \ldots, X_{iK}, i = 1, \ldots, n$, they satisfy

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK} + u_i$$

for $i = 1, \ldots, n$.

Why do we want to include more than 1 explanatory variable?

1. Because we are interested on the relation between all $K - 1$ variables and $Y$
2. Because we want to estimate the direct effect of one variable and include the other variables for that purpose
To see 2. consider the relation \((K = 3)\)

\[ \gamma = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u \]

Even if we are only interested in the effect of \(X_2\) on \(Y\) we cannot omit \(X_3\), if \(X_2\) and \(X_3\) are related. For instance, if

\[ X_3 = \gamma_1 X_1 + \gamma X_2 + \nu \]

then after substitution we find

\[ \gamma = (\beta_1 + \beta_3 \gamma_1) X_1 + (\beta_2 + \beta_3 \gamma_2) X_2 + u + \beta_3 \nu \]

In the relation between \(Y\) and \(X_1, X_2\), the coefficient of \(X_2\) is the sum of the direct effect \(\beta_2\) and the indirect effect \(\beta_3 \gamma_2\) of \(X_2\) on \(Y\).

If we include \(X_3\) in the relation, the coefficient of \(X_2\) is \(\beta_2\), i.e. the direct effect of \(X_2\) on \(Y\).

The assumptions are the same as before

Assumption 1: \(u_i, i = 1, \ldots, n\) are random variables with \(E(u_i) = 0\)

Assumption 2: \(X_{ik}, i = 1, \ldots, n, k = 1, \ldots, K\) are deterministic, i.e. non-random, constants.

Assumption 3 (Homoskedasticity)
All \(u_i\)'s have the same variance, i.e. for \(i = 1, \ldots, n\)

\[ \text{Var}(u_i) = E(u_i^2) = \sigma^2 \]

Assumption 4 (No serial correlation)
The random errors \(u_i\) and \(u_j\) are not correlated for all \(i \neq j = 1, \ldots, n\)

\[ \text{Cov}(u_i, u_j) = E(u_i u_j) = 0 \]
The inexact linear relation for \( i = 1, \ldots, n \)

\[
Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{ik} + u_i
\]

with assumptions 1-4 is the multiple Classical Linear Regression (CLR) model (remember with an intercept \( X_{i1} = 1, i = 1, \ldots, n \))

As in the simple CLR model the estimators of the regression parameters \( \beta_1, \ldots, \beta_K \) are found by minimizing the sum of squared residuals

\[
S(\beta_1, \ldots, \beta_K) = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} - \cdots - \hat{\beta}_K X_{ik})^2
\]

The \( \hat{\beta}_1, \ldots, \hat{\beta}_K \) that minimize the sum of squared residuals are Ordinary Least Squares (OLS) estimators of \( \beta_1, \ldots, \beta_K \).

The OLS residuals are

\[
e_i = Y_i - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} - \cdots - \hat{\beta}_K X_{ik}
\]

The (unbiased) estimator of \( \sigma^2 \) is

\[
s^2 = \frac{1}{n - K} \sum_{i=1}^{n} e_i^2
\]

Note \( n - K = \text{no. of observations} - \text{no. of regression coefficients (including intercept)} \)

The OLS residuals have the same properties as in the regression model with 1 regressor: for \( k = 1, \ldots, K \)

\[\sum_{i=1}^{n} X_{ik} e_i = 0\]  

(1)

In words: The sample covariance of the OLS residuals and all regressors is 0.

For \( k = 1 \) we have \( X_{i1} = 1 \) and hence \( \sum_{i=1}^{n} e_i = 0 \). Note: this holds if the regression model has an intercept.

Goodness of fit
Define the fitted value as before

\[ \hat{Y}_i = \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_K X_{ik} \]

By definition

\[ Y_i = \hat{Y}_i + e_i \]

Because of (1) the sample covariance of the fitted values and the OLS residuals is 0. If the model has an intercept, then

\[ \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2 \]

<table>
<thead>
<tr>
<th>Variation</th>
<th>Explained Variation</th>
<th>Unexplained Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Sum of Squares (TSS)</td>
<td>Regression Sum of Squares (RSS)</td>
<td>Error Sum of Squares (ESS)</td>
</tr>
</tbody>
</table>

The \( R^2 \) or Coefficient of Determination is defined as

\[ R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} \]

The \( R^2 \) increases if a regressor is added to the model. Why? Hint: Consider sum of squared residuals.
Adjusted (for degrees of freedom) \( R^2 \) decreases if added variable does not ‘explain’ much, i.e. if ESS does not decrease much,

\[
\bar{R}^2 = 1 - \frac{ESS / (n - K)}{TSS / (n - 1)}
\]

If \( ESS_{K+1}, ESS_K \) are the ESS for the model with \( K + 1 \) (one added) and \( K \) regressors, respectively, then \( ESS_{K+1} < ESS_K \), and the \( R^2 \) increases if \( X_{K+1} \) is added, but \( \bar{R}^2 \) decreases if

\[
\frac{ESS_{K+1} - ESS_K}{ESS_K} > -\frac{K + 1}{n - K}
\]

i.e. if the relative decrease is not big enough. For instance for \( n = 100, K = 10 \), the relative decrease has to be at least 12.2% to lead to an improvement in \( \bar{R}^2 \). This cutoff value looks arbitrary (and it is!)

In Section 4.3 many more criteria that balance decrease in ESS and degrees of freedom: Forget about them.

Application: Demand for bus travel (Section 4.6 of Ramanathan)

Variables

- BUSTRAVL = Demand for bus travel (1000 of passenger hours)
- FARE = Bus fare in $
- GASPRICE = Price of gallon of gasoline ($)
- INCOME = Average per capita income in $
- DENSITY = Population density (persons/sq. mile)
- LANDAREA = Area of city (sq. miles)

Data for 40 US cities in 1988

Question: Does demand for bus decrease if average income increases? Important for examining effect economic growth on bus system.

Reported are

- Descriptive statistics
- Scatterplot BUSTRAVL and INCOME
- OLS estimates in simple linear regression model
- OLS estimates in multiple linear regression model

Note in simple model INCOME has positive effect (be it that coefficient is not significantly different from 0)
If we include the other variables the effect is negative: 1$ more average income per head gives about 195 fewer passenger hours.

Note that the fit improves dramatically from $R^2 = 0.05$ to $R^2 = 0.92$. 
### Bustravl Density Fare Gasprice Income Landarea

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Maximum</th>
<th>Minimum</th>
<th>Std. Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Jarque-Bera</th>
<th>Probability</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bustravl</td>
<td>1933.175</td>
<td>1589.550</td>
<td>13103.00</td>
<td>18.1000</td>
<td>2431.757</td>
<td>3.066575</td>
<td>13.22568</td>
<td>236.9669</td>
<td>0.00000</td>
<td>40</td>
</tr>
<tr>
<td>Density</td>
<td>6789.100</td>
<td>5127.000</td>
<td>24288.00</td>
<td>1551.00</td>
<td>4559.464</td>
<td>1.701548</td>
<td>6.742500</td>
<td>42.64562</td>
<td>0.00000</td>
<td>40</td>
</tr>
<tr>
<td>Fare</td>
<td>0.882500</td>
<td>0.800000</td>
<td>1.500000</td>
<td>0.500000</td>
<td>0.279319</td>
<td>0.691675</td>
<td>2.732890</td>
<td>3.308340</td>
<td>0.191251</td>
<td>40</td>
</tr>
<tr>
<td>Gasprice</td>
<td>0.914250</td>
<td>0.910000</td>
<td>1.030000</td>
<td>0.790000</td>
<td>0.051285</td>
<td>0.189827</td>
<td>3.379224</td>
<td>0.479913</td>
<td>0.786662</td>
<td>40</td>
</tr>
<tr>
<td>Income</td>
<td>16760.30</td>
<td>17116.00</td>
<td>21886.00</td>
<td>12349.00</td>
<td>2098.046</td>
<td>0.189827</td>
<td>3.491300</td>
<td>0.838778</td>
<td>0.657448</td>
<td>40</td>
</tr>
<tr>
<td>Landarea</td>
<td>153.9775</td>
<td>99.55000</td>
<td>556.4000</td>
<td>18.90000</td>
<td>129.8950</td>
<td>1.284290</td>
<td>3.986898</td>
<td>12.61929</td>
<td>0.001819</td>
<td>40</td>
</tr>
</tbody>
</table>

### Pop

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Maximum</th>
<th>Minimum</th>
<th>Std. Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Jarque-Bera</th>
<th>Probability</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>919.2600</td>
<td>555.8000</td>
<td>7323.300</td>
<td>167.0000</td>
<td>1243.917</td>
<td>3.870887</td>
<td>19.22195</td>
<td>538.4778</td>
<td>0.000000</td>
<td>40</td>
</tr>
<tr>
<td>Median</td>
<td>555.8000</td>
<td>7323.300</td>
<td>919.2600</td>
<td>167.0000</td>
<td>1243.917</td>
<td>3.870887</td>
<td>19.22195</td>
<td>538.4778</td>
<td>0.000000</td>
<td>40</td>
</tr>
<tr>
<td>Maximum</td>
<td>7323.300</td>
<td>919.2600</td>
<td>555.8000</td>
<td>167.0000</td>
<td>1243.917</td>
<td>3.870887</td>
<td>19.22195</td>
<td>538.4778</td>
<td>0.000000</td>
<td>40</td>
</tr>
<tr>
<td>Minimum</td>
<td>167.0000</td>
<td>167.0000</td>
<td>167.0000</td>
<td>167.0000</td>
<td>167.0000</td>
<td>167.0000</td>
<td>167.0000</td>
<td>167.0000</td>
<td>167.0000</td>
<td>40</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>1243.917</td>
<td>1243.917</td>
<td>1243.917</td>
<td>1243.917</td>
<td>1243.917</td>
<td>1243.917</td>
<td>1243.917</td>
<td>1243.917</td>
<td>1243.917</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Jarque-Bera</th>
<th>Probability</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>538.4778</td>
<td>0.000000</td>
<td>40</td>
</tr>
</tbody>
</table>
Dependent Variable: BUSTRAVL  
Method: Least Squares  
Date: 10/01/01   Time: 11:05  
Sample: 1 40  
Included observations: 40

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>-2509.015</td>
<td>3091.188</td>
<td>-0.811667</td>
<td>0.4220</td>
</tr>
<tr>
<td>INCOME</td>
<td>0.265042</td>
<td>0.183042</td>
<td>1.447986</td>
<td>0.1558</td>
</tr>
</tbody>
</table>

R-squared: 0.052290  
Mean dependent var: 1933.175  
Adjusted R-squared: 0.027350  
S.D. dependent var: 2431.757  
S.E. of regression: 2398.272  
Akaike info criterion: 18.45159  
Schwarz criterion: 18.53604  
F-statistic: 2.096662  
Prob(F-statistic): 0.155822
### Regression Results

**Dependent Variable:** BUSTRAVL  
**Method:** Least Squares  
**Date:** 10/01/01  
**Time:** 11:07  
**Sample:** 1 40  
**Included observations:** 40

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2744.680</td>
<td>2641.672</td>
<td>1.038994</td>
<td>0.3064</td>
</tr>
<tr>
<td>INCOME</td>
<td>-0.194744</td>
<td>0.064887</td>
<td>-3.001294</td>
<td>0.0051</td>
</tr>
<tr>
<td>FARE</td>
<td>-238.6544</td>
<td>451.7281</td>
<td>-0.528314</td>
<td>0.6008</td>
</tr>
<tr>
<td>GASPRICE</td>
<td>522.1132</td>
<td>2658.228</td>
<td>0.196414</td>
<td>0.8455</td>
</tr>
<tr>
<td>POP</td>
<td>1.711442</td>
<td>0.231364</td>
<td>7.397176</td>
<td>0.0000</td>
</tr>
<tr>
<td>DENSITY</td>
<td>0.116415</td>
<td>0.059570</td>
<td>1.954253</td>
<td>0.0592</td>
</tr>
<tr>
<td>LANDAREA</td>
<td>-1.155230</td>
<td>1.802638</td>
<td>-0.640855</td>
<td>0.5260</td>
</tr>
</tbody>
</table>

**Summary Statistics**

- **R-squared:** 0.921026  
- **Adjusted R-squared:** 0.906667  
- **S.E. of regression:** 742.9113  
- **Sum squared resid:** 18213267  
- **Log likelihood:** -317.3332  
- **Durbin-Watson stat:** 2.082671  
- **S.D. dependent var:** 2431.757  
- **Mean dependent var:** 1933.175  
- **S.E. of regression:** 742.9113  
- **Akaike info criterion:** 16.21666  
- **Schwarz criterion:** 16.51221  
- **F-statistic:** 64.14338  
- **Prob(F-statistic):** 0.000000
The sampling distribution of $\hat{\beta}_1, \ldots, \hat{\beta}_K$

Under assumptions 1-4 the sampling distribution of $\hat{\beta}_1, \ldots, \hat{\beta}_K$ has mean $\beta_1, \ldots, \beta_K$ and a sampling variance that is proportional to $\sigma^2$. The square root of the estimated (estimate $\sigma^2$ by $s^2$) variance of the individual regression coefficients are called the standard errors of the regression coefficients. Regression programs always report both $\hat{\beta}_k$ and $std(\hat{\beta}_k)$, the standard error of $\hat{\beta}_k$.

The OLS estimator is also the Best Linear Unbiased Estimator (BLUE) in the multiple CLR model.

If we make the additional assumption

Assumption 5. The random error terms $u_i, i = 1, \ldots, n$ are random variables with a normal distribution.

We can derive the sampling distribution of $\hat{\beta}_1, \ldots, \hat{\beta}_K$ and $s^2$.

The sampling distribution of $\hat{\beta}_1, \ldots, \hat{\beta}_K$ multivariate normal with mean $\beta_1, \ldots, \beta_K$.

It can be shown that for all $k = 1, \ldots, K$

$$T_k = \frac{\hat{\beta}_k - \beta_k}{std(\hat{\beta}_k)}$$

has a t-distribution with $n - K$ degrees of freedom.

As in the simple CLR model we can use this to find a confidence interval for $\hat{\beta}_k$. This is done in the same way.
Hypothesis testing in the multiple CLR model

We consider two cases

1. Hypotheses involving 1 coefficient
2. Hypotheses involving 2 or more coefficients

1. Hypotheses involving 1 coefficient

For regression coefficient $\beta_k$ the hypothesis is

$$H_0 : \beta_k = \beta_{k0}$$

with alternative hypothesis

$$H_1 : \beta_k \neq \beta_{k0}$$

two-sided alternative

or

$$H_1 : \beta_k > \beta_{k0}$$

one-sided alternative

In these hypotheses $\beta_{k0}$ is a hypothesized value, e.g. $\beta_{k0} = 0$ if the null hypothesis is no effect of $X_k$ on $Y$.

The test is based on the statistic

$$T_k = \frac{\hat{\beta}_k - \beta_{k0}}{std(\hat{\beta}_k)}$$

If $H_0$ is not true, then $T_k$ is likely either large negative or large positive. For two-sided alternative, $H_1 : \beta_k \neq \beta_{k0}$, we reject if $|T_k| > c$ and for a one-sided alternative, $H_1 : \beta_k > \beta_{k0}$, if $T_k > d$.

For a test with a 5% confidence level, the cutoff values $c,d$ are chosen such that

$$\Pr(|T_k| > c) = .05$$, i.e. $\Pr(T_k > c) = .025$

(two-sided)

or

$$\Pr(T_k > d) = .05$$

(one-sided)

where $T_k$ has a t-distribution with $n - K$ d.f.
In example, \( n = 40, K = 7, \quad c \approx 2.035, \quad d \approx 1.695 \)

Note that the coefficients of INCOME, POP are significantly different from 0 at the 5% level. The coefficient of DENSITY is significantly different from 0 at the 10% level or if we test with a one-sided alternative.
2. Hypothesis involving 2 or more coefficient

As example consider the question whether the model with only INCOME as regressor is adequate.

In the more general model with INCOME, FARE, GASPRICE, POP, DENSITY, LANDAREA this corresponds to the hypothesis that the coefficients of FARE, GASPRICE, POP, DENSITY, LANDAREA, \( \beta_3, \beta_4, \beta_5, \beta_6, \beta_7 \) are all 0, i.e.

\[
H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0, \beta_6 = 0, \beta_7 = 0
\]

with alternative hypothesis

\[
H_1 : \text{One of these coefficients is not 0}
\]

How do we test this?

Idea: If \( H_0 \) is true then the model with only INCOME should fit as well as the model with all regressors.

Measure of fit is sum of squared OLS residuals \( \sum_{i=1}^{n} e_i^2 \).

Denote sum of squared residuals if \( H_0 \) is true by \( ESS_0 \). This is \( ESS \) if only INCOME is included. The \( ESS \) if \( H_1 \) is true is denoted by \( ESS_1 \).

Note: \( ESS_0 \geq ESS_1 \). Why?

We consider

\[
F = \frac{ESS_0 - ESS_1}{ESS_1 / 5}
\]

Numerator: Difference of restricted \( ESS \) (\( ESS_0 \)) and unrestricted \( ESS \) (\( ESS_1 \)) divided by the number of restrictions.

Denominator: Estimator of \( \sigma^2 \) in the model with all variables.

We reject if \( F \) is large, i.e. \( F > c \).

If \( H_0 \) is true then \( F \) has an F-distribution with degrees of freedom 5 (numerator) and 35 (denominator).
From Table of F-distribution, we find that e.g.

\[ \Pr(F > c) = 0.05 \]

if \( c \approx 2.49 \).

In example:

\[ ESS_0 = 218564926.27 \]
\[ ESS_1 = 18213267.59 \]
\[ ESS_0 - ESS_1 = 200351658.68 \]
\[ F = \frac{40070331.736}{551917.20} = 72.60 \]

Conclusion: We reject \( H_0 \) at the 5% level (and at the 1% level; check this) and the model with only INCOME is not adequate.

In general, we consider

\[ F = \frac{ESS_0 - ESS_1}{\frac{ESS_1}{m} / n - K} \]

with \( m \) the number of restrictions and we reject \( H_0 \) if \( F \) is (too) large. This is the F test or Wald test (Wald test is really F test if \( n \) is large).
Special case: Test of overall significance

\[ H_0 : \beta_2 = \cdots = \beta_K = 0 \]

i.e. all regression coefficients except intercept are 0. Hence, \( m = K - 1 \)

In example, the F statistic for this hypothesis is 64.14 with under \( H_0 \), \( m = 6 \) and \( n - K = 33 \) d.f. Hence, we reject \( H_0 \) and we conclude that the regressors are needed to explain BUSTRAVL.
Lecture 10. Multiple CLR model: Specification errors and Multicollinearity

Testing linear restrictions

In the multiple CLR model

\[ Y = \beta_1 X_1 + \cdots + \beta_K X_K + u \]

we may want to test other types of hypotheses involving more than 1 regression coefficient.

Example

\[ Y = \log C \quad \text{with } C = \text{aggregate consumption in a year} \]

\[ X_1 = \log I \quad \text{with } I = \text{aggregate income in a year} \]

\[ X_2 = \log \text{Pop} \quad \text{with } \text{Pop} = \text{Population size in a year} \]

Consider relations

(1) \[ \log C = \beta_1 + \beta_2 \log I + \beta_3 \log \text{Pop} + u \]

and

\[ \log \left( \frac{C}{\text{Pop}} \right) = \beta_1 + \beta_2 \log \left( \frac{I}{\text{Pop}} \right) + u \]

In second equation the relation is between consumption per capita and income per capita

Because e.g. \( \log \left( \frac{I}{\text{Pop}} \right) = \log I - \log \text{Pop} \) we rewrite the second equation as

\[ \log C - \log \text{Pop} = \beta_1 + \beta_2 \log I - \beta_2 \log \text{Pop} + u \]

or

(2) \[ \log C = \beta_1 + \beta_2 \log I + (1 - \beta_2) \log \text{Pop} + u \]

Equations (1) and (2) are the same if
\( \beta_3 = 1 - \beta_2 \) or equivalently \( \beta_2 + \beta_3 = 1 \).

To test whether the relation should be in scaled (by population) variables we test

\[ H_0 : \beta_2 + \beta_3 = 1 \]
against

\[ H_1 : \beta_2 + \beta_3 \neq 1 \]

We use the F-test

\[
F = \frac{ESS_0 - ESS_1}{ESS_1 / m} / \frac{ESS_1}{n - K}
\]

We must compute \( ESS_0 \) and \( ESS_1 \).

\( ESS_1 \) is the sum of the squared residuals for the CLR model

\[
\log C = \beta_1 + \beta_2 \log I + \beta_3 \log Pop + u
\]

\( ESS_0 \) is the sum of the squared residuals for CLR model under \( H_0 \), i.e. if \( \beta_2 + \beta_3 = 1 \).

To find this we estimate the CLR model

\[
\log \left( \frac{C}{Pop} \right) = \beta_1 + \beta_2 \log \left( \frac{I}{Pop} \right) + u
\]

The \( ESS \) of this model is \( ESS_1 \).

What is \( m \)?

If there is 1 restriction we can also use the t-test

Consider CLR model (2)

\[
\log C = \beta_1 + \beta_2 \log I + (1 - \beta_2) \log Pop + u
\]

This can be written as

\[
\log C = \beta_1 + \beta_2 \log \left( \frac{I}{Pop} \right) + \log Pop + u
\]

If we estimate the CLR model
\[ \log C = \gamma_1 + \gamma_2 \log \left( \frac{1}{Pop} \right) + \gamma_3 \log Pop + u \]

we must test

\[ H_0 : \gamma_3 = 1 \quad \text{against} \quad H_1 : \gamma_3 \neq 1 \]

and we can use the t-test.

F-test and t-test will give same conclusion because if

\[ F = \frac{ESS_0 - ESS_1}{\sqrt{\frac{ESS_1}{n-K}}} \]

and

\[ T = \frac{\hat{\gamma}_3 - \gamma_{30}}{\text{std} (\hat{\gamma}_3)} \]

then

\[ T^2 = F \]

Specification errors

Consider the CLR model

\[ Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u \]

In the specification of this model we can make two mistakes

1. Exclude a relevant variable
2. Include an irrelevant variable

An explanatory variable \( X_k \) is relevant if its regression coefficient \( \beta_k \neq 0 \) (and irrelevant if \( \beta_k = 0 \) )

What happens if you exclude a relevant variable?

Consider the correct CLR model

(3) \[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \]
and $\beta_3 \neq 0$.

However we estimate

(4) \[ Y = \gamma_1 + \gamma_2 X_2 + v \]

Note I use different symbols for coefficients and error term. What is the relation between $\beta$’s and $U$ and $\gamma$’s and $V$?

Express the relation between $X_3$ and $X_2$ as

\[ X_3 = \delta_1 + \delta_2 X_2 + w \]

Substitution in (3) gives

\[ Y = (\beta_1 + \delta_1 \beta_3) + (\beta_2 + \delta_2 \beta_3) X_1 + u + \beta_3 w \]

Comparing with (4) we see

\[ \gamma_1 = \beta_1 + \delta_1 \beta_3 \quad \gamma_2 = \beta_2 + \delta_2 \beta_3 \quad v = u + \beta_3 w \]

Hence if $\delta_1 \neq 0$ the intercept is not that of the correct CLR model and if $\delta_2 \neq 0$, i.e. if $X_2$ and $X_3$ are related, then the slope is not the slope of the correct CLR model.

If $X_2$ and $X_3$ are related and we exclude the relevant $X_1$, then we estimate the sum of the direct ($\beta_2$) and indirect (through $X_3$) effect ($\delta_2 \beta_3$) of $X_2$ on $Y$.

If we want the direct effect we estimate the wrong coefficient! This is called omitted variable bias.

See example in CLR model for demand for bus travel: coefficient income was positive if included alone and negative after other (relevant) variables were included.

Because we do not estimate $\beta_2$

- The standard error of $\hat{\gamma}_2$ is not equal to that of $\hat{\beta}_2$.
- The confidence interval for $\gamma_2$ is not equal to that of $\beta_2$.
- The t-test for $H_0 : \gamma_2 = \gamma_{20}$ is not that for $H_0 : \beta_2 = \beta_{20}$.

Is omitted variable bias a problem?
Depends on goal. Remember relation between selling price and area living room. For prediction we may not be interested in direct effect, but if we consider an addition we are.
Including an irrelevant variable

Let the correct CLR model be

\[ Y = \beta_1 + \beta_2 X_2 + u \]

We estimate

(5) \[ Y = \gamma_1 + \gamma_2 X_2 + \gamma_3 X_3 + \nu \]

Now the relation between the two models is

\[ \beta_1 = \gamma_1 \quad \beta_2 = \gamma_2 \quad \gamma_3 = 0 \quad \nu = u \]

Hence the correct model is a special case.

If we estimate we expect that \( \hat{\gamma}_3 \) is close to 0 (and that the t-test for \( \gamma_3 = 0 \) will not reject).

Model (5) is a valid CLR model and the OLS estimator will estimate the direct effects. Standard errors and tests can be used to find confidence intervals and test hypotheses on these direct effects.

It can be shown that if we estimate (5), then the standard error of \( \hat{\gamma}_2 \) is

\[ \text{std}(\hat{\gamma}_2) = \frac{\text{std}(\hat{\beta}_2)}{1 - r_{X_2,X_3}^2} \]

with \( \text{std}(\hat{\beta}_2) \) the standard error of the OLS estimator of \( \beta_2 \) in the correct model and \( r_{X_2,X_3} \) the sample correlation between \( X_2 \) and \( X_3 \).

Hence the standard error increases if we include an irrelevant variable: the estimates are less precise.
Lecture 11. Multicollinearity

Data, US 1963-1985

Housing = new housing units started
         (thousands)

Intrate = interest rate on mortgages

GNP     = GNP (billions of 1982 $)

Pop     = US population (millions)

Regressions with dependent variable housing and independent variables (and intercept)

- Intrate, GNP (model 1)
- Intrate, Pop (model 2)
- Intrate, Pop, GNP (model 3)
- Intrate (model 4)

In regression with all variables the coefficients of Pop and GNP not significantly different from 0

F-test of hypothesis that these coefficients are both 0

\[ F = \frac{\frac{ESS_4 - ESS_3}{2}}{\frac{ESS_3}{23-4}} \]

\[ = \frac{499425.5}{77800.6} = 6.42 \]

Hence we reject this hypothesis

Also regression coefficients of Pop and GNP change from models 1, 2 to model 3.

What is going on?

Consider relations among the explanatory variables

- Scatterplot GNP and Pop
- Scatterplot Intrate and GNP

Relation between GNP and Pop almost exact

Sample correlation coefficients

\[ r(GNP, POP) = .99 \]
## Regression Results

**Dependent Variable:** HOUSING  
**Method:** Least Squares  
**Date:** 10/11/01  
**Time:** 06:51  
**Sample:** 1963 1985  
**Included observations:** 23

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>-3812.933</td>
<td>1588.849</td>
<td>-2.399808</td>
<td>0.0263</td>
</tr>
<tr>
<td>INTRATE</td>
<td>-198.3982</td>
<td>51.29332</td>
<td>-3.867915</td>
<td>0.0010</td>
</tr>
<tr>
<td>POP</td>
<td>33.81927</td>
<td>9.374440</td>
<td>3.607604</td>
<td>0.0018</td>
</tr>
</tbody>
</table>

**Summary Statistics:**  
- **R-squared:** 0.428507  
- **Mean dependent var:** 1601.100  
- **Adjusted R-squared:** 0.371358  
- **S.D. dependent var:** 345.4715  
- **S.E. of regression:** 273.9139  
- **Schwarz criterion:** 14.33272  
- **F-statistic:** 7.498023  
- **Durbin-Watson stat:** 0.845996  
- **Prob(F-statistic):** 0.003716
Dependent Variable: HOUSING
Method: Least Squares
Date: 10/11/01   Time: 06:54
Sample: 1963 1985
Included observations: 23

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>687.8977</td>
<td>382.6818</td>
<td>1.797571</td>
<td>0.0874</td>
</tr>
<tr>
<td>INTRATE</td>
<td>-169.6579</td>
<td>43.83829</td>
<td>-3.870084</td>
<td>0.0010</td>
</tr>
<tr>
<td>GNP</td>
<td>0.905395</td>
<td>0.248978</td>
<td>3.636444</td>
<td>0.0016</td>
</tr>
</tbody>
</table>

R-squared 0.432101
Adjusted R-squared 0.375311
S.E. of regression 273.0513
Sum squared resid 1491140.
Log likelihood -160.0505
Durbin-Watson stat 0.831697
# Regression Analysis

**Dependent Variable:** HOUSING  
**Method:** Least Squares  
**Date:** 10/11/01  
**Time:** 06:55  
**Sample:** 1963 1985  
**Included observations:** 23

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>-1315.752</td>
<td>4930.529</td>
<td>-0.266858</td>
<td>0.7925</td>
</tr>
<tr>
<td>INTRATE</td>
<td>-184.7507</td>
<td>58.10431</td>
<td>-3.179638</td>
<td>0.0049</td>
</tr>
<tr>
<td>POP</td>
<td>14.90113</td>
<td>36.55288</td>
<td>0.407660</td>
<td>0.6881</td>
</tr>
<tr>
<td>GNP</td>
<td>0.522159</td>
<td>0.973884</td>
<td>0.536162</td>
<td>0.5981</td>
</tr>
</tbody>
</table>

- **R-squared:** 0.437025  
- **Adjusted R-squared:** 0.348134  
- **S.E. of regression:** 278.9276  
- **Sum squared resid:** 1478211.  
- **Log likelihood:** -159.9503  
- **Durbin-Watson stat:** 0.831133

- **Mean dependent var:** 1601.100  
- **S.D. dependent var:** 345.4715  
- **Akaike info criterion:** 14.25655  
- **Schwarz criterion:** 14.45403  
- **F-statistic:** 4.916420  
- **Prob(F-statistic):** 0.010775
Dependent Variable: HOUSING  
Method: Least Squares  
Date: 10/11/01   Time: 06:56  
Sample: 1963 1985  
Included observations: 23  

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1872.813</td>
<td>252.4118</td>
<td>7.419671</td>
<td>0.0000</td>
</tr>
<tr>
<td>INTRATE</td>
<td>-29.47410</td>
<td>26.25531</td>
<td>-1.122596</td>
<td>0.2743</td>
</tr>
</tbody>
</table>

R-squared    0.056613  Mean dependent var 1601.100  
Adjusted R-squared 0.011690  S.D. dependent var 345.4715  
S.E. of regression 343.4463  Akaike info criterion 14.59888  
Sum squared resid 2477062  Schwarz criterion 14.69762  
Log likelihood -165.8871  F-statistic 1.260221  
Durbin-Watson stat 0.849696  Prob(F-statistic) 0.274281
INTRATE vs. GNP
\[ r(GNP, Inrate) = .88 \]
\[ r(Inrate, POP) = .91 \]

Consider CLR model

\[ Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + u_i \]

Assume that there is an exact relation between \( X_2 \) and \( X_3 \)

\[ X_{i3} = \gamma_1 + \gamma_2 X_{i2} \]

Substitute this relation in the CLR model

\[ Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 (\gamma_1 + \gamma_2 X_{i2}) + u_i \]
\[ = (\beta_1 + \beta_3 \gamma_1) + (\beta_2 + \beta_3 \gamma_2) X_{i2} + u_i \]

We basically have a model with 1 explanatory variable \( X_2 \)

\[ Y_i = \delta_1 + \delta_2 X_{i2} + u_i \]

and all we can estimate is the intercept \( \delta_1 \) and the slope \( \delta_2 \).

If we have estimates for these coefficients can we get the original ones \( \beta_1, \beta_2, \beta_3 \)?

The relationship between the coefficients in the two models is

\[ \delta_1 = \beta_1 + \beta_3 \gamma_1 \quad \delta_2 = \beta_2 + \beta_3 \gamma_2 \]

Now the coefficients \( \gamma_1, \gamma_2 \) in the exact relation between \( X_2 \) and \( X_3 \) can be easily computed.

However from the estimates from \( \delta_1, \delta_2 \) we cannot solve for \( \beta_1, \beta_2, \beta_3 \): 2 equations for 3 unknowns.

The problem is that we cannot distinguish between the effects of two variables that have an exact linear relation.

This is an example of an identification problem: With the information that we have we cannot estimate all regression coefficients.
With exact linear relation the computer is not able to compute the OLS estimates of the regression coefficients.

What happens if the relation between the explanatory variables is almost exact (as in our example)?

Formulas for sample variance

\[ \text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{S_{22}(1-r_{X_2,X_3}^2)} \]

\[ \text{Var}(\hat{\beta}_3) = \frac{\sigma^2}{S_{33}(1-r_{X_2,X_3}^2)} \]

with \( S_{22}, S_{33} \) the sampling variance of \( X_2, X_3 \) and \( r_{X_2,X_3} \) the sample correlation coefficient of these variables.

If sample correlation close to 1 or −1 (exact relation), then sampling variances large.

If there is an exact relation between some explanatory variables, we have exact multicollinearity.

This is rare. In the example we have multicollinearity, i.e. the relation is not exact, but strong enough that

- Individual coefficients are not individually significantly different from 0
- Jointly they are significantly different from 0

This is also how you discover whether this problem exists.

As with exact multicollinearity we cannot distinguish between the effect of \( X_2 \) and \( X_3 \).

Important: Unless there is exact multicollinearity, the OLS estimators have the usual properties. In particular, they remain the best estimators.

Multicollinearity is a problem with the data, not with the model.

Some signs of this problem

- Low individual t-stats, but high F-stat for omission of these explanatory variables
- High sample correlations between these variables
- Coefficients change is one the variables is dropped
Solution depends on goal

- If multicollinearity among variables that were only included in order to avoid omitted variable bias is estimate of the effect of some other variable, then drop one of them.
- If you are not interested in individual coefficients, as in forecasting, ignore the problem.
- If you are interested in the effects of the variables that are closely related, then get more data or if not possible admit defeat.

Severe multicollinearity as in the example is rare.

Fear of multicollinearity could lead to the omission of explanatory variables. That is much worse. If inclusion leads to multicollinearity, you can always drop the variable again.
Lecture 12. Functional form

Multiple linear regression model

\[ Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_K X_K + u \]

Interpretation of regression coefficient \( \beta_k \):

Change in \( Y \) if \( X_k \) is changed by 1 unit and the other variables are held constant.

This change

- Does not depend on the value of \( X_k \)
- Does not depend on the value of other variables

Examples

- Relation between demand for electricity \( Y \) and its price \( X_2 \). Price effect may depend on temperature \( X_3 \).
- Relation between consumption \( Y \) and income \( X_2 \). Change in consumption due to extra income may decrease with income.

Linear regression model seems not able to allow for this.

Linear in coefficients or linear in variables

Compare the following relations between \( Y \) and \( X \) (omit error term \( u \))

1. \[ Y = \beta_1 + \beta_2 X + \beta_3 X^2 \]

2. \[ Y = \beta_1 + \beta_2 X^{\beta_3} \]

Both relations are nonlinear (see graphs)

Relation 1 is linear in the regression coefficients, i.e. it can be expressed as a linear relation between \( Y \) and independent variables \( X_2, X_3, \ldots \). Define \( X_2 = X, X_3 = X^2 \).

For relation 2 this is not possible (try!).

If a nonlinear relation can be expressed as a linear relation by redefining variables we can estimate that relation using OLS.

Review of (natural) logarithms and the exponential function

(natural) logarithm

\[ \log X \]
exponential function

\[ e^x \]

See graphs.

Note

- Both functions increasing
- Logarithm only defined if \( X \) is positive

Some relations

\[
\begin{align*}
\log e^x &= X \\
\log e^x &= X \\
\log X^a &= a \log X \\
\log aX &= \log a + \log X \\
\log e^{a+x} &= e^a e^x
\end{align*}
\]

Derivatives

\[
\begin{align*}
\frac{d \log X}{d X} &= \frac{1}{X} \\
\frac{d e^x}{d X} &= e^x
\end{align*}
\]

Examples of nonlinear relations that can be expressed as a linear relation (independent variables \( X, Z \); define the \( X_2, X_3, \ldots \) that make the relation linear; in some cases the dependent variable changes as well; \( U \) is omitted)

Quadratic relation

\[ Y = \beta_1 + \beta_2 X + \beta_3 X^2 \]

Relation with interaction term

\[ Y = \beta_1 + \beta_2 X + \beta_3 XZ \]

Reciprocal relation

\[ Y = \beta_1 + \beta_2 \frac{1}{X} \]

Linear log relation

\[ Y = \beta_1 + \beta_2 \log X \]
FIGURE 6.1  Graph of Exponential and Logarithmic Functions

a. Graph of $Y = \exp(X)$

b. Graph of $Y = \ln(X)$
Exponential relation

\[ Y = e^{\beta_1 + \beta_2 X} \]

Power relation (here constant must be redefined as well)

\[ Y = \gamma X^{\beta_2} Z^\beta_3 \]

How do we analyze nonlinear relations?

- Graphs (only for 1 independent variable)
- Derivatives: Change in \( Y \) associated with small change in \( X \)

Remember derivative \( \frac{dY}{dX} \) in point \((X, Y)\) on curve is the slope of the straight line that touches the nonlinear relation in that point, i.e. the slope of the linear relation that approximates the nonlinear relation in that point.

Example: Quadratic relation

\[ Y = \beta_1 + \beta_2 X + \beta_3 X^2 \]

Then

\[ \frac{dY}{dX} = \beta_2 + 2\beta_3 X \]

Hence effect of small change of \( X \) depends on \( X \) (and may even change sign)

Other use: Find value of \( X \) that maximizes \( Y \), e.g \( Y \) is income and \( X \) is age if \( \beta_2 = 20, \beta_3 = -0.25 \) the income maximal at age 40.

In economics often interested in elasticities: relative change in \( Y \) associated with a small relative change in \( X \). Note relative changes do not depend on unit of measurement.

Small relative change in variable \( X \):

\[ \frac{dX}{X} \]

or small relative percentage change
\[ 100 \frac{dX}{X} \]

Elasticity (note 100 cancels if defined for percentage change)

\[ \frac{dY}{Y} \frac{Y}{dX} = X \frac{dY}{Y} \frac{Y}{dX} \]

In other words: Multiply derivative by \( \frac{X}{Y} \)

In quadratic relation

\[ \frac{dY}{Y} \frac{Y}{dX} = X \frac{dY}{Y} \frac{Y}{dX} = X \left( \beta_2 + 2\beta_X \right) \]

Note depends on both \( X \) and \( Y \)

Remember

\[ \frac{d \log X}{dX} = \frac{1}{X} \]

Hence for small change in \( X \)

\[ d \log X = \frac{1}{X} dX \]

With this we can compute an elasticity as

\[ \frac{dY}{Y} \frac{Y}{dX} = \frac{d \log Y}{d \log X} \]

Convenient if the (in)dependent variables are log’s

Log-log relation

\[ \log Y = \beta_1 + \beta_2 \log X \]
In this relation $\beta_2$ is the elasticity of $Y$ with respect to $X$.

Also consider

Semi-log relation

$$\log Y = \beta_1 + \beta_2 X$$

Here

$$\beta_2 = \frac{d \log Y}{d X} = \frac{d Y}{d X}$$

This is relative change in $Y$ associated with 1 unit change in $X$. Note: percentage change is $100\beta_2$.

Relation with interaction

$$Y = \beta_1 + \beta_2 X + \beta_3 XZ$$

If relation has more than 1 independent variable, we look at the effect of $X$ holding $Z$ constant.

Note if $Z$ constant

$$\frac{\partial Y}{\partial X} = \beta_2 + \beta_3 Z$$

(we use $\partial$ instead of $d$ to indicate that some variables are held constant)

Note effect of $X$ depends on $Z$ (compare electricity demand above).

Special exponential relation

$$Y_t = e^{\beta Y_0}$$

with $Y_t$ the value of $Y$ in period $t$ and $Y_0$ the value of $Y$ in the initial period.

Relative change in $Y$
\[
\frac{Y_t - Y_{t-1}}{Y_{t-1}} = \frac{Y_0e^{\beta t} - Y_0e^{\beta(t-1)}}{Y_0e^{\beta(t-1)}} = e^\beta - 1
\]

Hence the coefficient in the relation

\[ \log Y_t = \log Y_0 + \beta t \]

can be used to obtain the relative change/growth rate in \( Y \)

Other relations: Lags in \( Y, X \)

With time series data we can define new variables by lags

\[ Y_{t-1}, Y_{t-2}, \ldots \]

Note in period \( t \) the value of \( Y_{t-1} \) is the value of \( Y \) in period \( t-1 \)

Linear relation with lagged variables

\[ Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + \beta_4 X_{t-1} \]

This expresses

- Delayed response
- Adjustment over time
- Habits/reluctance to change

Some caveats in estimating nonlinear relations

- Graphs with one \( X \) are misleading
- Do not compare \( R^2 \) if dependent variables are different

For rest estimation and testing as before

Consider

\[ Y = \beta_1 + \beta_2 X + \beta_3 X^2 + u = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \]

Then e.g. test of nonlinearity is test of \( H_0 : \beta_3 = 0 \)
Lecture 13. Dummy variables

Types of variables

- Continuous (income, height, weight, etc.)
- Discrete (gender, season, points scored etc.)

Continuous variables have

- Origin, i.e. value is 0
- Unit of measurement

Often obvious, e.g. price in US$.

In regression both origin and unit of measurement can be changed.

Discrete variables: three types

- Counts, e.g. number of runs scored
- Ordinal, e.g. agree/neutral/disagree
- Nominal/categorical, e.g. gender

With counts there is obvious origin and also unit of measurement is obvious

Continuous variables and counts together are called quantitative variables

With ordinal variables there is no origin and no unit of measurement, but there is an order

With nominal variables there is no unit of measurement and no origin and even no order
Ordinal and nominal variables are called qualitative variables.

Discrete variables can be

- Dependent variable
- Independent variable

If dependent variable is discrete various problems, e.g. in

\[ Y = \alpha + \beta X + u \]

random error \( U \) cannot be continuous variable and hence cannot have a normal distribution.

In this lecture we consider qualitative variables as independent variables in linear regression models.

To use a qualitative variable as an independent variables in a linear regression

\[ Y = \alpha + \beta X + u \]

we must first attach numerical values to the categories.

For this dummy/indicator variables are very useful. A dummy/indicator variable \( D \) is a variable that has two values: 0 and 1.
Consider gender with categories female and male. We could choose

\[ D_i = 0 \quad \text{if } i \text{ is female} \]

(1)

\[ D_i = 1 \quad \text{if } i \text{ is male} \]

or

\[ D_i^* = 0 \quad \text{if } i \text{ is male} \]

(2)

\[ D_i^* = 1 \quad \text{if } i \text{ is female} \]

Because the labels are arbitrary this should not make a difference.

Note the 0 is not the origin and 1 is not the unit of measurement. They are just labels and we could have used –2 and 99 instead (but that is not a convenient choice).

The category with label 0 is called the control or reference category (I prefer reference category)

Now consider the regression model

\[ Y = \alpha + \beta D + u \]

with \( D \) as in (1) and with \( Y \) is monthly salary.

What is the interpretation of \( \alpha, \beta \)?

If assumption 2 of the CLR model holds, then

\[ E(u \mid D = 0) = E(u \mid D = 1) = 0 \]

and hence
\[ E(Y \mid D = 0) = \alpha + E(u \mid D = 0) = \alpha \]
\[ E(Y \mid D = 1) = \alpha + \beta + E(u \mid D = 1) = \alpha + \beta \]

with
\[ E(Y \mid D = 0) \text{ is average monthly salary female employees (reference category)} \]
\[ E(Y \mid D = 1) \text{ is average monthly salary male employees} \]

This suggests for OLS estimators \( \hat{\alpha}, \hat{\beta} \)
\[ \hat{\alpha} = \bar{Y}_{\text{female}} \]
\[ \hat{\alpha} + \hat{\beta} = \bar{Y}_{\text{male}} \]
and hence
\[ \hat{\beta} = \bar{Y}_{\text{male}} - \bar{Y}_{\text{female}} \]

Intercept is average for reference category

Example: Sample of 49 employees
\[ n_{\text{male}} = 26, \quad n_{\text{female}} = 23 \]
\[ \bar{Y}_{\text{male}} = 2086.93, \quad \bar{Y}_{\text{female}} = 1518.70 \]

Compare with regression results:
\[ \hat{\alpha} = 1518.70, \quad \hat{\beta} = 568.23 \]
Advantage of regression: direct confidence interval of/test for salary difference between male and female employees

If we replace $D$ by $D^*$, i.e. now 0 indicates male and 1 female we have the regression model

$$Y = \alpha^* + \beta^* D^* + u$$

and

$$E(Y \mid D^* = 0) = \alpha^*$$

$$E(Y \mid D^* = 1) = \alpha^* + \beta^*$$

and hence

$$\hat{\alpha}^* = \bar{Y}_{\text{male}} \quad \quad \quad \quad \quad \quad \quad \hat{\beta}^* = \bar{Y}_{\text{female}} - \bar{Y}_{\text{male}}$$

For the OLS estimates we find

$$\hat{\alpha}^* = 2086.92 \quad \quad \quad \quad \quad \quad \quad \hat{\beta}^* = -568.23$$

Note $\hat{\beta}^* = -\hat{\beta}$ and standard error is identical: tests/confidence intervals give same conclusion.

Is the result a proof of gender discrimination? Why (not)?

Now consider two dummy variables

$$D_{i1} = 0 \quad \quad \quad \text{if } i \text{ is female}$$

$$D_{i1} = 1 \quad \quad \quad \text{if } i \text{ is male}$$
and

\[ D_{i2} = 0 \quad \text{if } i \text{ is nonwhite} \]

\[ D_{i2} = 1 \quad \text{if } i \text{ is white} \]

We consider the following models

(1) \[ Y = \beta_1 + \beta_2 D_1 + \beta_3 D_3 + u \]

(2) \[ Y = \beta_1 + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_1 D_2 + u \]

We consider the salary difference between men and women by ethnicity.

In model (1)

\[ E(Y \mid D_1 = 1, D_2 = 0) - E(Y \mid D_1 = 0, D_2 = 0) = \beta_2 = \]

\[ = E(Y \mid D_1 = 1, D_2 = 1) - E(Y \mid D_1 = 0, D_2 = 1) \]
Restriction: Salary difference the same for whites and nonwhites

In model (2)

\[ E(Y \mid D_1 = 1, D_2 = 0) - E(Y \mid D_1 = 0, D_2 = 0) = \beta_2 \]

and

\[ E(Y \mid D_1 = 1, D_2 = 1) - E(Y \mid D_1 = 0, D_2 = 1) = \beta_2 + \beta_4 \]

Estimation results: Salary difference only for whites.

Also: Race difference only for men.

Model (2) has an interaction term \( D_1D_2 \).

Next, we consider qualitative variable with more than 2 categories

Examples: State of residence, level of education, income category (grouped continuous variable)

\[
S = \begin{cases} 
0 & \text{if no high school diploma} \\
1 & \text{if high school diploma, but no college degree} \\
2 & \text{if college degree} 
\end{cases}
\]

Using \( S \) in this way is bad idea (why?)

Instead we introduce two dummy variables
\[ S_1 = 1 \quad \text{if high school diploma, but no college degree} \]
\[ S_1 = 0 \quad \text{otherwise} \]

and

\[ S_2 = 1 \quad \text{if college degree} \]
\[ S_2 = 0 \quad \text{otherwise} \]

Note: reference group has not a high school diploma

Regression model

\[ Y = \beta_1 + \beta_2 S_1 + \beta_3 S_2 + u \]

Now

\[ \beta_1 \] is average of \( Y \) for reference group (no high school diploma)

\[ \beta_1 + \beta_2 \] is average of \( Y \) for group with high school diploma, but no college degree

\[ \beta_1 + \beta_3 \] is average of \( Y \) for group with college degree

How do you test
• Education has no impact on income
• The return (in income) to having a college degree is 0

Give $H_0$ and indicate which test you want to use.

Define

$$S_3 = 1 \quad \text{if no high school diploma}$$

$$S_3 = 0 \quad \text{otherwise}$$

Consider the regression model

$$Y = \beta_1 + \beta_2 S_1 + \beta_3 S_2 + \beta_4 S_3 + u$$

Why can the coefficients of this model not be estimated?

This is called the dummy variable trap

Example: Monthly salary and type of work

Maint=maintenance work

Crafts=works in crafts

Clerical=clerical work

Reference category is professional

Interpret the constant and the other coefficients.

Combining quantitative and qualitative independent variables

Consider the model
\[ Y = \beta_1 + \beta_2 D + \beta_3 X + u \]

with \( Y \) is log of monthly salary, \( D \) is gender and \( X \) is education (in years of schooling)

In relation between \( Y \) and \( X \) the intercept is \( \beta_1 \) for women and \( \beta_1 + \beta_2 \) for men (see figure)

Estimation results (what is interpretation of coefficient of gender?)

Note that gender difference is not due to difference in level of education.

Consider two other models

(3) \[ Y = \beta_1 + \beta_2 X + \beta_3 D X + u \]

In this model intercept is the same but slope is different for men and women (see figure)

For women slope is \( \beta_2 \)

For men slope is \( \beta_2 + \beta_3 \)

(4) \[ Y = \beta_1 + \beta_2 D + \beta_3 X + \beta_4 D X + u \]

In this model both slope and intercept are different

Model for women

\[ Y = \beta_1 + \beta_3 X + u \]

and for men
\[ Y = \beta_1 + \beta_2 + (\beta_3 + \beta_4)X + u \]

This amounts to splitting the sample and estimating two separate regressions

OLS estimates

Advantage dummy approach: Tests
Dependent Variable: WAGE
Method: Least Squares
Date: 10/30/01  Time: 06:58
Sample: 1 49
Included observations: 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1518.696</td>
<td>122.5373</td>
<td>12.39374</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDER</td>
<td>568.2274</td>
<td>168.2208</td>
<td>3.377868</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

R-squared       0.195343 Mean dependent var 1820.204
Adjusted R-squared 0.178223 S.D. dependent var 648.2687
S.E. of regression 587.6681 Akaike info criterion 15.63016
Sum squared resid 16231629 Schwarz criterion 15.70738
Log likelihood  -380.9390  F-statistic 11.40999
Durbin-Watson stat  1.664603  Prob(F-statistic) 0.001476
FIGURE 7.1 An Example of an Intercept Shift Using a Dummy Variable

\[ \hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\beta}X \]

\[ \hat{\alpha}_2 + \hat{\beta}X \]
<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.305416</td>
<td>0.058705</td>
<td>124.4423</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.281818</td>
<td>0.080591</td>
<td>3.496883</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

- R-squared: 0.206459
- Adjusted R-squared: 0.189575
- S.E. of regression: 0.281541
- Sum squared resid: 3.725458
- Log likelihood: -6.400539
- Durbin-Watson stat: 1.622408

Mean dependent var: 7.454952
S.D. dependent var: 0.312741
Akaike info criterion: 0.342879
Schwarz criterion: 0.420096
Prob(F-statistic): 0.001039
Dependent Variable: LNWAGE
Method: Least Squares
Date: 10/30/01   Time: 08:36
Sample: 149    Included observations: 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.024929</td>
<td>0.111997</td>
<td>62.72427</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.273708</td>
<td>0.075079</td>
<td>3.645624</td>
<td>0.0007</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.045753</td>
<td>0.015946</td>
<td>2.869199</td>
<td>0.0062</td>
</tr>
</tbody>
</table>

R-squared     0.326916    Mean dependent var 7.454952
Adjusted R-squared 0.297652 S.D. dependent var 0.312741
S.E. of regression 0.262096 Akaike info criterion 0.219060
Sum squared resid 3.159944 Schwarz criterion 0.334886
Log likelihood -2.366973 F-statistic 11.17108
Durbin-Watson stat 1.809135 Prob(F-statistic) 0.000111
Dependent Variable: WAGE
Method: Least Squares
Date: 10/30/01   Time: 08:20
Sample: 1 49   Included observations: 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2670.222</td>
<td>152.5465</td>
<td>-17.50431</td>
<td>0.0000</td>
</tr>
<tr>
<td>MAINT</td>
<td>-1052.040</td>
<td>205.6937</td>
<td>-5.114597</td>
<td>0.0000</td>
</tr>
<tr>
<td>CRAFTS</td>
<td>-641.1313</td>
<td>205.6937</td>
<td>-3.116922</td>
<td>0.0032</td>
</tr>
<tr>
<td>CLERICAL</td>
<td>-1279.222</td>
<td>186.8306</td>
<td>-6.846964</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

R-squared 0.532794  Mean dependent var 1820.204
Adjusted R-squared 0.501647  S.D. dependent var 648.2687
S.E. of regression 457.6396  Akaike info criterion 15.16815
Sum squared resid 9424530.  Schwarz criterion 15.32258
Log likelihood -367.6196  F-statistic 17.10576
Durbin-Watson stat 1.780224  Prob(F-statistic) 0.000000
**Dependent Variable:** WAGE  
**Method:** Least Squares  
**Date:** 10/30/01  
**Time:** 07:42  
**Sample:** 1 49  
**Included observations:** 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1305.974</td>
<td>161.7846</td>
<td>8.072299</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDER</td>
<td>592.7722</td>
<td>163.9563</td>
<td>3.615427</td>
<td>0.0007</td>
</tr>
<tr>
<td>RACE</td>
<td>326.1732</td>
<td>167.9351</td>
<td>1.942258</td>
<td>0.0582</td>
</tr>
</tbody>
</table>

| R-squared   | 0.256330    | Mean dependent var | 1820.204 |
| Adjusted R-squared | 0.223997 | S.D. dependent var | 648.2687 |
| S.E. of regression  | 571.0668  | Akaike info criterion | 15.59216 |
| Sum squared resid   | 15001395  | Schwarz criterion | 15.70799 |
| Log likelihood      | -379.0079  | F-statistic | 7.927696 |
| Durbin-Watson stat  | 1.822353   | Prob(F-statistic) | 0.001101 |
### Regression Results

**Dependent Variable:** WAGE  
**Method:** Least Squares  
**Date:** 10/30/01  
**Time:** 07:43

**Sample:** 149  
**Included observations:** 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1635.000</td>
<td>186.9467</td>
<td>8.745809</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDER</td>
<td>24.45455</td>
<td>245.6961</td>
<td>0.099532</td>
<td>0.9212</td>
</tr>
<tr>
<td>RACE</td>
<td>-178.3333</td>
<td>231.4920</td>
<td>-0.770365</td>
<td>0.4451</td>
</tr>
<tr>
<td>RACEGENDER</td>
<td>919.2788</td>
<td>312.4829</td>
<td>2.941853</td>
<td>0.0051</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-squared</td>
<td>0.376284</td>
<td>Mean dependent var</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.334703</td>
<td>S.D. dependent var</td>
</tr>
<tr>
<td>S.E. of regression</td>
<td>528.7651</td>
<td>Akaike info criterion</td>
</tr>
<tr>
<td>Sum squared resid</td>
<td>12581662</td>
<td>Schwarz criterion</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-374.6983</td>
<td>F-statistic</td>
</tr>
<tr>
<td>Durbin-Watson stat</td>
<td>1.780860</td>
<td>Prob(F-statistic)</td>
</tr>
</tbody>
</table>
Dependent Variable: WAGE
Method: Least Squares
Date: 10/30/01   Time: 07:15
Sample: 1 49
Included observations: 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2086.923</td>
<td>115.2512</td>
<td>18.10760</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDERC</td>
<td>-568.2274</td>
<td>168.2208</td>
<td>-3.377868</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

R-squared 0.195343  Mean dependent var 1820.204
Adjusted R-squared 0.178223  S.D. dependent var 648.2687
S.E. of regression 587.6681  Akaike info criterion 15.63016
Sum squared resid 16231629  Schwarz criterion 15.70738
Log likelihood -380.9390  F-statistic 11.40999
Durbin-Watson stat 1.664603  Prob(F-statistic) 0.001476
FIGURE 7.2 An Example of a Slope Shift Using a Dummy Variable

\[ \hat{\alpha} + (\hat{\beta}_1 + \hat{\beta}_2)x \]

\[ \hat{\alpha} + \hat{\beta}_1x \]
FIGURE 7.3  An Example of a Shift in the Intercept and Slope

\[ (\hat{\alpha}_1 + \hat{\alpha}_2) + (\hat{\beta}_1 + \hat{\beta}_2)X \]

\[ \hat{\alpha}_1 + \hat{\beta}_1 X \]
Dependent Variable: LNWAGE  
Method: Least Squares  
Date: 10/30/01  Time: 08.47  
Sample (adjusted): 248 IF GENDER=1  
Included observations: 26 after adjusting endpoints

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.232265</td>
<td>0.145793</td>
<td>49.60624</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.056276</td>
<td>0.021111</td>
<td>2.665669</td>
<td>0.0135</td>
</tr>
</tbody>
</table>

R-squared         | 0.228439     | Mean dependent var | 7.587234   |
Adjusted R-squared| 0.196291     | S.D. dependent var | 0.337607   |
S.E. of regression| 0.302564     | Akaike info criterion | 0.521415  |
Sum squared resid | 2.198528     | Schwarz criterion   | 0.618192   |
Log likelihood    | -4.778394    | F-statistic         | 7.105789   |
Durbin-Watson stat| 1.926152     | Prob(F-statistic)   | 0.013528   |
**Dependent Variable: LNWAGE**

**Method: Least Squares**

**Date: 10/30/01   Time: 08:49**

**Sample: 1 49 IF GENDER=0**

**Included observations: 23**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.230143</td>
<td>0.160588</td>
<td>45.02294</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.012278</td>
<td>0.025268</td>
<td>0.485930</td>
<td>0.6320</td>
</tr>
</tbody>
</table>

**R-squared**

<table>
<thead>
<tr>
<th>R-squared</th>
<th>Mean dependent var</th>
<th>7.305416</th>
</tr>
</thead>
</table>

**Adjusted R-squared**

<table>
<thead>
<tr>
<th>Adjusted R-squared</th>
<th>S.D. dependent var</th>
<th>0.199545</th>
</tr>
</thead>
</table>

**S.E. of regression**

<table>
<thead>
<tr>
<th>S.E. of regression</th>
<th>Akaike info criterion</th>
<th>-0.267274</th>
</tr>
</thead>
</table>

**Sum squared resid**

<table>
<thead>
<tr>
<th>Sum squared resid</th>
<th>Schwarz criterion</th>
<th>-0.168535</th>
</tr>
</thead>
</table>

**Log likelihood**

<table>
<thead>
<tr>
<th>Log likelihood</th>
<th>F-statistic</th>
<th>0.236128</th>
</tr>
</thead>
</table>

**Durbin-Watson stat**

<table>
<thead>
<tr>
<th>Durbin-Watson stat</th>
<th>Prob(F-statistic)</th>
<th>0.632048</th>
</tr>
</thead>
</table>
Dependent Variable: LNWAGE
Method: Least Squares
Date: 10/30/01  Time: 08:50
Sample: 1 49
Included observations: 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.231691</td>
<td>0.106183</td>
<td>68.10581</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.012044</td>
<td>0.018206</td>
<td>0.661531</td>
<td>0.5116</td>
</tr>
<tr>
<td>EDUCGENDER</td>
<td>0.044308</td>
<td>0.011391</td>
<td>3.889899</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

R-squared     0.347184  Mean dependent var  7.454952
Adjusted R-squared  0.318800  S.D. dependent var  0.312741
S.E. of regression   0.258120  Akaike info criterion  0.188486
Sum squared resid    3.064794  Schwarz criterion   0.304312
Log likelihood     -1.617911  F-statistic          12.23196
Durbin-Watson stat  1.832842  Prob(F-statistic)    0.000055
Dependent Variable: LN WAGE
Method: Least Squares
Date: 10/30/01  Time: 08:51
Sample: 149
Included observations: 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.230143</td>
<td>0.206344</td>
<td>35.03922</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.002122</td>
<td>0.241622</td>
<td>0.008780</td>
<td>0.9930</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.012278</td>
<td>0.032467</td>
<td>0.378176</td>
<td>0.7071</td>
</tr>
<tr>
<td>EDUC*GENDER</td>
<td>0.043997</td>
<td>0.037222</td>
<td>1.182013</td>
<td>0.2434</td>
</tr>
</tbody>
</table>

R-squared          0.347185
Adjusted R-squared 0.303664
S.E. of regression 0.260972
Sum squared resid   3.064789
Log likelihood      -1.617869
Durbin-Watson stat  1.832670
Mean dependent var  7.454952
S.D. dependent var  0.312741
Akaike info criterion 0.229301
Schwarz criterion   0.383735
F-statistic         7.977402
Prob(F-statistic)   0.000227