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)
<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>

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  Akaike info criterion 14.18461
Sum squared resid 1500576.  Schwarz criterion 14.33272
Log likelihood -160.1230  F-statistic 7.498023
Durbin-Watson stat 0.845996  Prob(F-statistic) 0.003716
<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  Mean dependent var 1601.100
Adjusted R-squared 0.375311  S.D. dependent var 345.4715
S.E. of regression 273.0513  Akaike info criterion 14.17830
Sum squared resid 1491140.  Schwarz criterion 14.32641
Log likelihood -160.0505  F-statistic 7.608750
Durbin-Watson stat 0.831697  Prob(F-statistic) 0.003489
**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  
- Mean dependent var: 1601.100  
- Adjusted R-squared: 0.348134  
- S.D. dependent var: 345.4715  
- Akaike info criterion: 14.25655  
- Schwarz criterion: 14.45403  
- F-statistic: 4.916420  
- Prob(F-statistic): 0.010775  
- Durbin-Watson stat: 0.831133  
- Prob(Durbin-Watson)
<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
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

\[ F = \frac{ESS_4 - ESS_3}{2 \frac{ESS_3}{23-4}} = \frac{499425.5}{77800.6} = 6.42 \]

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

\[ r(GNP, POP) = .99 \]
\[ r(GNP, Intrate) = .88 \]
\[ r(Intrate, POP) = .91 \]
POP vs. GNP

![Graph showing the relationship between POP and GNP. The graph has a linear trend line indicating a positive correlation.](image-url)
INTRATE vs. GNP

![Graph showing the relationship between INTRATE and GNP.](image)
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

\[
Var(\hat{\beta}_2) = \frac{\sigma^2}{S_{22}(1 - r_{X_2X_3}^2)}
\]

\[
Var(\hat{\beta}_3) = \frac{\sigma^2}{S_{33}(1 - r_{X_2X_3}^2)}
\]

with \( S_{22}, S_{33} \) the sampling variance of \( X_2, X_3 \) and \( r_{X_2X_3} \) 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.