Multiple Regression Analysis

\[ y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots + \beta_kx_k + u \]

4. Further Issues

Ch. 6 Multiple Regression: Further Issues

1. Effects of Data Scaling on OLS Statistics
2. More on Functional Form
3. More on Goodness-of-Fit & Selection of Regressors
4. Prediction & Residual Analysis*

6.1 Effects of Data Scaling

Redefining Variables

Changing the scale of the variable will lead to a corresponding change in the scale of the coefficients and standard errors, but no change in the significance or interpretation. → see table 6.1.

Cont. Redefining Variables

- If \( y \) is multiplied by \( c \), its coefficient is divided by \( c \).
- If \( y \) is multiplied by \( c \), all OLS coefficients is multiplied by \( c \).
- Neither \( t \) nor \( F \) statistics are affected by changing the units of measurement of any variables.
- If the variables appears in logarithmic form, changing unit of measurement does not affect the slope coefficient.

<table>
<thead>
<tr>
<th>Table 6.1: Effects of Data Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent Variable</strong></td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td><strong>Independent Variables</strong></td>
</tr>
<tr>
<td>cigs</td>
</tr>
<tr>
<td>packs</td>
</tr>
<tr>
<td>female</td>
</tr>
<tr>
<td>intercept</td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>R-Squared</td>
</tr>
<tr>
<td>MSE</td>
</tr>
<tr>
<td>SSR</td>
</tr>
</tbody>
</table>
### Effects of Data Scaling

<table>
<thead>
<tr>
<th>Dependent Independent</th>
<th>( y )</th>
<th>( \hat{e}y )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( \beta_1 (s_{x_1}) )</td>
<td>( c \beta_1 (s_{x_1}) )</td>
<td>( \beta_1 (s_{x_1}/d) )</td>
</tr>
<tr>
<td>( d \hat{x}_1 )</td>
<td></td>
<td></td>
<td>( \hat{\beta}<em>1 (s</em>{d}) )</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( \beta_2 (s_{x_2}) )</td>
<td>( c \beta_2 (s_{x_2}) )</td>
<td>( \beta_2 (s_{x_2}/d) )</td>
</tr>
<tr>
<td>Intercept</td>
<td>( \beta_0 (s_{\hat{y}}) )</td>
<td>( c \beta_0 (s_{\hat{y}}) )</td>
<td>( \beta_0 (s_{\hat{y}}/d) )</td>
</tr>
<tr>
<td>R-squared</td>
<td>( R^2 )</td>
<td>( R^2 )</td>
<td>( R^2 )</td>
</tr>
<tr>
<td>SSR</td>
<td>SSR</td>
<td>( c^2 * SSR )</td>
<td>SSR</td>
</tr>
</tbody>
</table>

Standard errors in parentheses.

### Beta Coefficients

- Idea is to replace \( y \) and each \( x \) variable with a standardized version — subtract mean and divide by standard deviation.
- Coefficient reflects standard deviation of \( y \) for a one standard deviation change in \( x \).
  - We can compare the magnitudes of the resulting beta coefficients and conclude that “which variable is most important,” etc.
  - Whether we use standardized or unstandardized variables does not affect statistical significance.

---

### 6.2 More on Functional Form

- **Functional Forms**
  - 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.
    - the natural log of \( x, y \) or both
    - quadratic forms of \( x \)
    - interactions of \( x \) variables

### Interpretation of Log Models

- **6-2a Logarithmic Functional Forms**
  - \( \ln(y) = \beta_0 + \beta_1 \ln(x) + \epsilon \)
    - \( \beta_1 \) is the elasticity of \( y \) with respect to \( x \)
  - \( \ln(y) = \beta_0 + \beta_1 x + \epsilon \)
    - \( \beta_1 \) is approximately the percentage change in \( y \) given a 1 unit change in \( x \)
  - \( y = \beta_0 + \beta_1 \ln(x) + \epsilon \)
    - \( \beta_1 \) is approximately the change in \( y \) for a 100 percent change in \( x \)

### Why use log models?

- Log models are invariant to the scale of the variables since measuring percent changes.
- They give a direct estimate of elasticity.
- For models with \( y > 0 \), the conditional distribution is often heteroskedastic or skewed, while \( \ln(y) \) is much less so.
- The distribution of \( \ln(y) \) is more narrow, limiting the effect of outliers.
Some Rules of Thumb for “log”

- What types of variables are often used in log form?
  - Dollar amounts that must be positive
  - Very large variables, such as population
- What types of variables are often used in level form?
  - Variables measured in years
  - Variables that are a proportion or percent

Quadratic Models

- 6-2b Models with Quadratics
  - For a model of the form \( y = \beta_0 + \beta_1 x + \beta_2 x^2 + u \), we can’t interpret \( \beta_1 \) alone as measuring the change in \( y \) with respect to \( x \), we need to take into account \( \beta_2 \) as well, since \( \frac{\partial \hat{y}}{\partial x} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x \)

More on Quadratic Models

- For the case of the coefficient on \( x > 0 \) and the coefficient on \( x^2 < 0 \), \( y \) is increasing in \( x \) at first, but will eventually turn around and be decreasing in \( x \) (see fig.6.1).
- For the case of the coefficient on \( x < 0 \) and the coefficient on \( x^2 > 0 \), \( y \) is decreasing in \( x \) at first, but will eventually turn around and be increasing in \( x \) (see fig.6.2).
- For both case, the turning point will be at \( x^* = \hat{\beta}_1 / (2\hat{\beta}_2) \)

Interaction Terms

- 6-2c Models with Interaction Terms
  - For a model of the form \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u \) we can’t interpret \( \beta_i \) alone as measuring the change in \( y \) with respect to \( x_i \), we need to take into account \( \beta_i \) as well, since \( \frac{\partial y}{\partial x_i} = \hat{\beta}_i + \beta_i x_j \)
  - To summarize the effect of \( x_i \) on \( y \), we typically evaluate the above at average of \( x_j \).

6.3 More of Goodness-of-Fit

Adjusted R-Squared

- Recall that the \( R^2 \) will always increase as more variables are added to the model.
- The adjusted \( R^2 \) takes into account the number of variables in a model, and may decrease.
  \[
  R^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} = 1 - (1-R^2) \cdot \frac{n-1}{n-k-1} \tag{6.22}
  \]

Cont. Adjusted R-Squared

- Most PC packages will give you both \( R^2 \) and adj-\( R^2 \).
- You can compare the fit of 2 models (with the same \( y \)) by comparing the adj-\( R^2 \).
- However, you cannot use the adj-\( R^2 \) to compare models with different \( y \)’s.
  - e.g. \( y \) vs. \( \ln(y) \).
It is important not to fixate too much on adj-
$R^2$ and lose sight of theory and common sense.

- If economic theory clearly predicts a variable
  belongs, generally leave it in the model.
- Don’t want to include a variable that prohibits a
  sensible interpretation of the variable of interest –
  remember ceteris paribus interpretation of
  multiple regression.

* e.g. housing price = $f(\# \text{ of rooms}, \text{square footage})$

---

In a regression of traffic fatalities
on state beer taxes (and other
factors), one should not directly
control for beer consumption.

\[
\text{fatalities} = f(\text{tax, beercons, ...})
\]

---

In a regression of family health
expenditures on pesticide usage
among farmers, one should not
control for doctor visits.

\[
\text{hexpend} = f(\text{pestusage, docvisit, ...})
\]

---

In a regression of house prices on
house characteristics, one would
include price assessments if the
purpose of the regression is to
study their validity; otherwise
one would not include them.

1. \[
\text{hprice} = f(\text{bdrms, assess, ...})
\]
2. \[
\text{hprice} = f(\text{bdrms, ...})
\]