# Chapter 6: Multiple Regression Analysis: Further Issues

## Introductory Econometrics: A Modern Approach

# 6.1 Effects of data scaling on OLS statistics

### Dependent Variables

$$\widehat{bwght} = \hat{\beta}_0 + \hat{\beta}_1 \text{ cigs } + \hat{\beta}_2 \text{ faminc}$$

- $\text{bwght} =$ child birth weight, in ounces.
- $\text{cigs} =$ number of cigarettes smoked by the mother while pregnant, per day.
- $\text{faminc} =$ annual family income, in thousands of dollars.

$$\widehat{\text{bwght}}\ /16 = \hat{\beta}_0/16 + \left(\hat{\beta}_1/16\right) \text{ cigs } + \left(\hat{\beta}_2/16\right) \text{ faminc }.$$

When variables are re-scaled, the coefficients, standard errors, confidence intervals, t statistic and F statistic changes in a way that preserve the testing outcome.

## Effects of Data Scaling

| Dependent Variable | (1) *bwght* | (2) *bwghtlbs* | (3) *bwght* |
|---|---|---|---|
| Independent Variables | | | |
| *cigs* | −.4634<br>(.0916) | −.0289<br>(.0057) | — |
| *packs* | — | — | −9.268<br>(1.832) |
| *faminc* | .0927<br>(.0292) | .0058<br>(.0018) | .0927<br>(.0292) |
| *intercept* | 116.974<br>(1.049) | 7.3109<br>(.0656) | 116.974<br>(1.049) |
| Observations | 1,388 | 1,388 | 1,388 |
| *R*-Squared | .0298 | .0298 | .0298 |
| SSR | 557,485.51 | 2,177.6778 | 557,485.51 |
| SER | 20.063 | 1.2539 | 20.063 |

# Data scaling of independent variables

Define $packs = \frac{cigs}{20}$

$$\widehat{\text{bwght}} = \hat{\beta}_0 + \left(20\hat{\beta}_1\right)(\text{ cigs }/20) + \hat{\beta}_2 \text{ faminc} = \hat{\beta}_0 + \left(20\hat{\beta}_1\right) \text{ packs } + \hat{\beta}_2 \text{ faminc.}$$

**Effects of Data Scaling**

| Dependent Variable | (1) *bwght* | (2) *bwghtlbs* | (3) *bwght* |
|---|---|---|---|
| Independent Variables | | | |
| *cigs* | −.4634<br>(.0916) | −.0289<br>(.0057) | — |
| *packs* | — | — | −9.268<br>(1.832) |
| *faminc* | .0927<br>(.0292) | .0058<br>(.0018) | .0927<br>(.0292) |
| *intercept* | 116.974<br>(1.049) | 7.3109<br>(.0656) | 116.974<br>(1.049) |
| Observations | 1,388 | 1,388 | 1,388 |
| *R*-Squared | .0298 | .0298 | .0298 |
| SSR | 557,485.51 | 2,177.6778 | 557,485.51 |
| SER | 20.063 | 1.2539 | 20.063 |

- the coefficient on `packs` is 20 times that on `cigs`
- standard error on `packs` is 20 times that on `cigs`

- this means that `t-statistics` is the same in both cases

# Standardizing the variables (Beta coefficients)

The test scores are used in wage equations, and the scale of these test scores is often arbitrary and not easy to interpret

We are interested in how a particular individual's score compares with the population:

- Instead of asking about the effect on hourly wage if say a test score is 10 point higher, it makes more sense to ask what happends when the test score is **one standard deviation** higher.

Standardizing all variables:

1) Original form

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i$$

2) Subtract the mean

$$y_i - \bar{y} = \hat{\beta}_1 \left(x_{i1} - \bar{x}_1\right) + \hat{\beta}_2 \left(x_{i2} - \bar{x}_2\right) + \ldots + \hat{\beta}_k \left(x_{ik} - \bar{x}_k\right) + \hat{u}_i$$

3) Let $\hat{\sigma}_i$ be the sample standard deviation for each variable

$$\left(y_i - \bar{y}\right)/\hat{\sigma}_y = \left(\hat{\sigma}_1/\hat{\sigma}_y\right)\hat{\beta}_1 \left[\left(x_{i1} - \bar{x}_1\right)/\hat{\sigma}_1\right] + \ldots + \left(\hat{\sigma}_k/\hat{\sigma}_y\right)\hat{\beta}_k \left[\left(x_{ik} - \bar{x}_k\right)/\hat{\sigma}_k\right] + \left(\hat{u}_i/\hat{\sigma}_y\right)$$

# Standardizing all variables

$$(y_i - \bar{y})/\hat{\sigma}_y = (\hat{\sigma}_1/\hat{\sigma}_y)\,\hat{\beta}_1\left[(x_{i1} - \bar{x}_1)/\hat{\sigma}_1\right] + \ldots + (\hat{\sigma}_k/\hat{\sigma}_y)\,\hat{\beta}_k\left[(x_{ik} - \bar{x}_k)/\hat{\sigma}_k\right] + (\hat{u}_i/\hat{\sigma}_y)$$

It is useful to rewrite the equation above as:

$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \ldots + \hat{b}_k z_k +$ error

where $z_y$ denotes the $z$-score of $y$, $z_1$ is the $z$-score of $x_1$, and so on. The new coefficients are

$$\hat{b}_j = (\hat{\sigma}_j/\hat{\sigma}_y)\,\hat{\beta}_j \text{ for } j = 1, \ldots, k$$

$\hat{b}_j$ are traditionally called **standardized coefficients** or **beta coefficients**

Beta coefficients receive their interesting meaning from equation above:

- If $x_1$ increases by one standard deviation, then $\hat{y}$ changes by $\hat{b}_1$ standard deviations.
- Thus, we are measuring effects not in terms of the original units of $y$ or the $x_j$, but in **standard deviation units**.

# Example 6.1: Effects of pollution on housing prices

Level- level model:

$\text{price} = \beta_0 + \beta_1 \text{ nox } + \beta_2 \text{ crime } + \beta_3 \text{ rooms } + \beta_4 \text{ dist } + \beta_5 \text{ stratio } + u$

where `crime` is the number of reported crimes per capita

Standardized model:

$\widehat{\text{zprice}} = -.340 \text{ znox } -.143 \text{ zcrime } +.514 \text{ zrooms } -.235 \text{ zdist } -.270 \text{ zstratio}$

Interpretation:

- a one standard deviation increase in `nox` decreases price by .34 standard deviations
- a one standard deviation increase in `crime` reduces price by .14 standard deviations

Whether we use standardize or unstandardized variables does not affect statistical significance: the **t statistics** are the same in both cases

# 6.2 More on functional form

a) More on using logarithmic functional form

$\log(\text{ price }) = \beta_0 + \beta_1 \log(\text{ nox }) + \beta_2 \text{ rooms } + u$

$$\begin{aligned} \widehat{\log(price)} &= \underset{(0.19)}{9.23} - \underset{(.066)}{.718} \log(nox) + \underset{(.019)}{.306}\ rooms \\ n &= 506,\ R^2 = .514. \end{aligned}$$

Thus, when `nox` increases by 1% `price` falls by .718% holding `rooms` fixed.

When `rooms` increases by one, `price` increases by approximately $100(.306) = 30.6\%$

# a) More on using logarithmic functional form

To calculate the exact change:

$$\%\Delta\hat{y} = 100 \cdot \left[\exp\left(\hat{\beta}_2 \Delta x_2\right) - 1\right]$$

when $\Delta x_2 = 1$

$$\%\Delta\hat{y} = 100 \cdot \left[\exp\left(\hat{\beta}_2\right) - 1\right]$$

Applied to the housing price example with $x_2 =$ rooms and $\hat{\beta}_2 = .306$,

$$\%\Delta\ \widehat{\text{price}}\ = 100[\exp(.306) - 1] = 35.8\%$$

- which is notably larger than the approximate percentage change 30.6%.

# More on using logarithmic functional form

Reasons why logarithmic form appears in applied work:

- coefficients have appealing interpretations
- we can ignore the units of measurement of variables
- strictly positive variables often have distribution that are heteroskedastic or skewed; taking the log can mitigate, if not, eliminate both problems
- taking the log of variables often narrows its range which makes the OLS estimates less sensitive to outliers (extreme values)
- it can create extreme values cases. - An example is when a variable $y$ is between zero and one (such as a proportion) and takes on values close to zero. In this case, $\log(y)$ (which is necessarily negative) can be very large in magnitude whereas the original variable, $y$, is bounded between zero and one.

# More on using logarithmic functional form

Unwritten rules:

- When a variable is a positive dollar amount, the log is often taken. We have seen this for variables such as wages, salaries, firm sales, and firm market value.
- Variables such as population, total number of employees, and school enrollment often appear in logarithmic form; these have the common feature of being large integer values.

Review: distinction between a percentage change and a percentage point change

- Remember, if `unem` goes from 8% to 9% , this is an increase of one percentage point, but a $12.5\%$ increase from the initial unemployment level. Using the log means that we are looking at the percentage change in the unemployment rate:
  - $\log(9) - \log(8) \approx .118$ or $11.8\%$, which is the logarithmic approximation to the actual $12.5\%$ increase.

# b) Model with quadratics

Quadratic functions are also used quite often in applied economics to capture decreasing or increasing marginal effects

Model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$$

Remember, that $\beta_1$ does not measure the change in y with respect to x ; it makes no sense to hold $x^2$ fixed while changing $x$

Estimated equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

$$\Delta \hat{y} \approx \left( \hat{\beta}_1 + 2\hat{\beta}_2 x \right) \Delta x,$$

so $\Delta \hat{y} / \Delta x \approx \hat{\beta}_1 + 2\hat{\beta}_2 x$

The estimated slope is $\hat{\beta}_1 + 2\hat{\beta}_2 x$

# Example

Concave experience profile

$$\widehat{wage} = \underset{(.35)}{3.73} + \underset{(.041)}{.298}\ exper - \underset{(.0009)}{.0061}\ exper^2$$

$$n = 526, R^2 = .093$$

- Marginal effect of experience

$$\frac{\Delta\ \text{wage}}{\Delta\ \text{exper}} = .298 - 2(.0061)\ \text{exper}$$

This estimated equation implies that `exper` has a diminishing effect on `wage`. The first year of experience is worth roughly 30¢ per hour ($.298).

The second year of experience is worth less $.298 - 2(.0061)(1) \approx .286$, or $28.6$¢

In going from 10 to 11 years of experience, wage is predicted to increase by about $.298 - 2(.0061)(10) = .176$, or $17.6\%$

When the coefficient on $x$ is positive and the coefficient on $x^2$ is negative, the quadratic has a parabolic shape (concave).

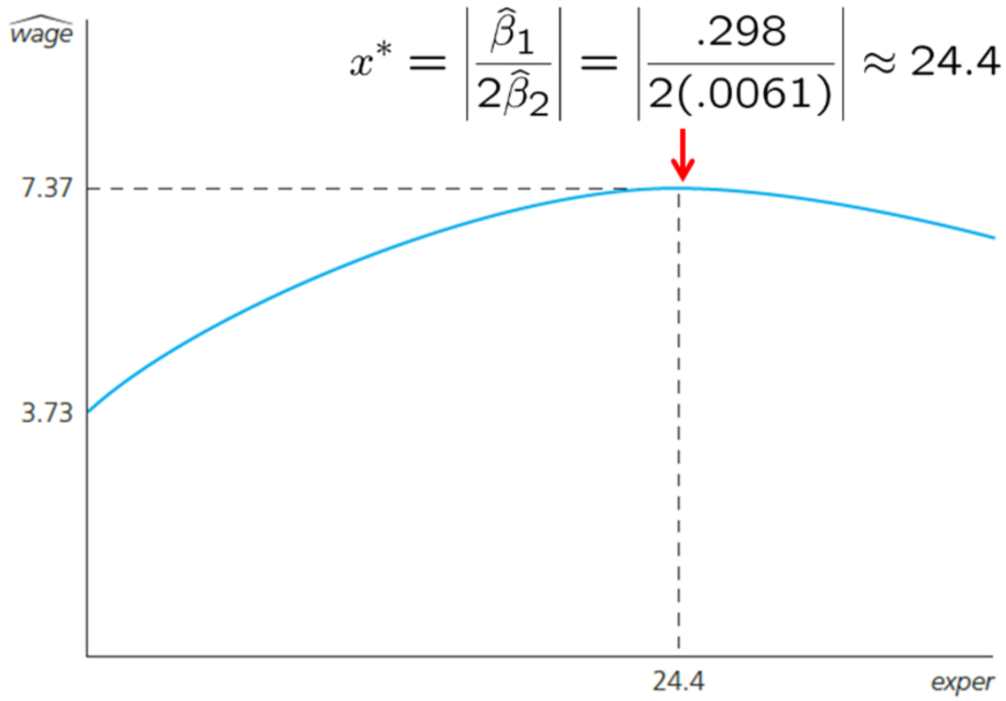
There is always a positive value of $x$ where the effect of $x$ on $y$ is zero; before this point, $x$ has a positive effect on $y$ ; after this point, $x$ has a negative effect on $y$.

How can we find the turning point?

$$x^* = |\frac{\hat{\beta}_1}{2\hat{\beta}_2}|$$

In the wage example, $x^* =$ `exper` $^*$ is $.298/[2(.0061)] \approx 24.4$

- Does this mean the return to experience becomes negative after 24.4 years?
    - Not necessarily. It depends on how many observations in the sample lie to the right of the turnaround point.
    - In the given example, these are about 28% of the observations. There may be a specification problem (e.g. omitted variables).



# Example: Effects of pollution on housing prices

When the coefficient on $x$ is negative and the coefficient on $x^2$ is positive, the quadratic has a convex shape. Increasing effect of $x$ on $y$.

$$\log(\text{ price }) = \beta_0 + \beta_1 \log(\text{ nox }) + \beta_2 \log(\text{ dist }) + \beta_3 \text{ rooms } + \beta_4 \text{ rooms }^2 + \beta_5 \text{ stratio } + u.$$

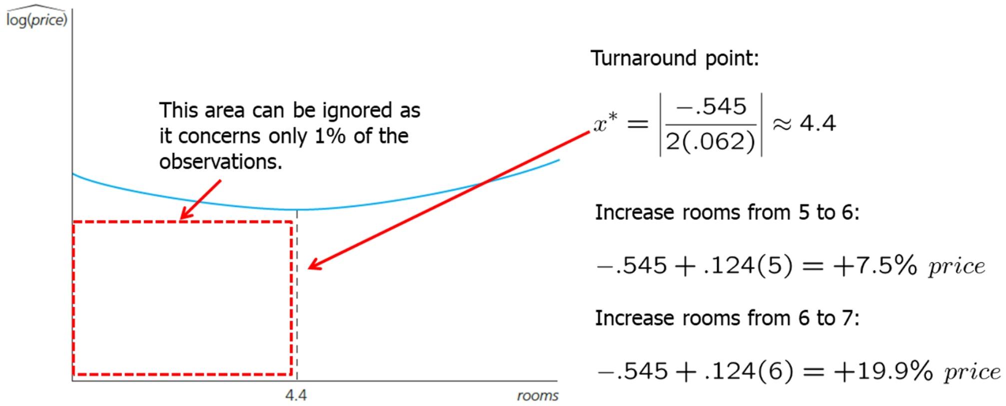
$$\log(\widehat{price}) = \underset{(.57)}{13.39} - \underset{(.115)}{.902} \log(nox) - \underset{(.043)}{.087} \log(dist)$$
$$- \underset{(.165)}{.545}\, rooms + \underset{(.013)}{.062}\, rooms^2 - \underset{(.006)}{048}\, stratio$$
$$n = 506,\ R^2 = .603$$

*nox*: nitrogen oxide in the air
*dist*: distance from employment centers
*rooms*: number of rooms
*stratio*: average student/teacher ratio

Does this mean that, at a low number of rooms, more rooms are associated with lower prices?

$$\frac{\Delta \log(\text{ price })}{\Delta \text{ rooms}} = \frac{\% \Delta \text{ price}}{\Delta \text{ rooms}} = -.545 + .124 \text{ rooms}$$

# Example: Effects of pollution on housing prices




The coefficient on `rooms` is negative and the coefficient on `rooms^2` is positive, this equation literally implies that, at low values of rooms, an additional room has a negative effect on `log(price)`

Do we really believe that starting at three rooms and increasing to four rooms actually reduces a house's expected value? Probably not.

- It turns out that only five of the 506 communities in the sample have houses averaging 4.4 rooms or less, about 1% of the sample. This is so small that the quadratic to the left of 4.4 can, for practical purposes, be ignored.

$$\begin{aligned}\%\Delta \text{ price } &\approx 100\{[-.545 + 2(.062)] \text{ rooms }\}\Delta \text{ rooms}\\ &= (-54.5 + 12.4 \text{ rooms })\Delta \text{ rooms}.\end{aligned}$$

# c) Models with interaction terms

Sometimes, it is natural for the dependent variable with respect to an explanatory variable to depend on the magnitude of yet another explanatory variable.

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft \cdot bdrms + \beta_4 bthrms + u$$

- The partial effect of `bdrms` on `price` (holding all other variables fixed) is

$$\frac{\Delta \text{ price}}{\Delta bdrms} = \beta_2 + \beta_3 sqrft$$

If $\beta_3 > 0$, then it implies that an additional bedroom yields a higher increase in housing price for larger houses

In other words, there is an interaction effect between `square footage` and `number of bedrooms`.

In summarizing the effect of `bdrms` on `price`, we must evaluate the equation above at interesting values of `sqrft`, such as the mean value, or the lower and upper quartiles in the sample

**The parameters on the original variables can be tricky to interpret when we include an interaction term.**

- For example, in the previous housing price equation, equation shows that $\beta_2$ is the effect of `bdrms` on price for *a home with zero square feet*

# Example 6.3: Effects of attendance on final exam performance

$$stndfnl = \beta_0 + \beta_1\, atndrte + \beta_2\, priGPA + \beta_3\, ACT + \beta_4 priGPA^2 + \beta_5 ACT^2 + \beta_6\, priGPA \cdot atndrte + u.$$

$$\begin{aligned} \widehat{stndfnl} = {} & \underset{(1.36)}{2.05} - \underset{(.0102)}{.0067}\, atndrte - \underset{(.48)}{1.63}\, priGPA - \underset{(.098)}{.128}\, ACT \\ & + \underset{(.101)}{.296}\, priGPA^2 + \underset{(.0022)}{.0045}\, ACT^2 + \underset{(.0043)}{.0056}\, priGPA \cdot atndrte \\ & n = 680,\ R^2 = .229,\ \bar{R}^2 = .222. \end{aligned}$$

```r
data(attend, package='wooldridge')

# Regression
model_1 <- feols(data=attend, stndfnl ~ atndrte+priGPA+ACT+priGPA^2+ACT^2+priGPA*atndrte)
summary(model_1)
```

```
## OLS estimation, Dep. Var.: stndfnl
## Observations: 680
## Standard-errors: IID
##                 Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)     2.050293   1.360319  1.507215 0.13222476
## atndrte        -0.006713   0.010232 -0.656067 0.51200517
## priGPA         -1.628540   0.481003 -3.385720 0.00075118 ***
## ACT            -0.128039   0.098492 -1.299998 0.19404671
## I(priGPA^2)     0.295905   0.101049  2.928314 0.00352322 **
## I(ACT^2)        0.004533   0.002176  2.082939 0.03763374 *
## atndrte:priGPA  0.005586   0.004317  1.293817 0.19617261
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.868368   Adj. R2: 0.221777
```

If we add the term $\beta_7$ `ACT·atndrte` to equation (6.18), what is the partial effect of `atndrte` on `stndfnl`?

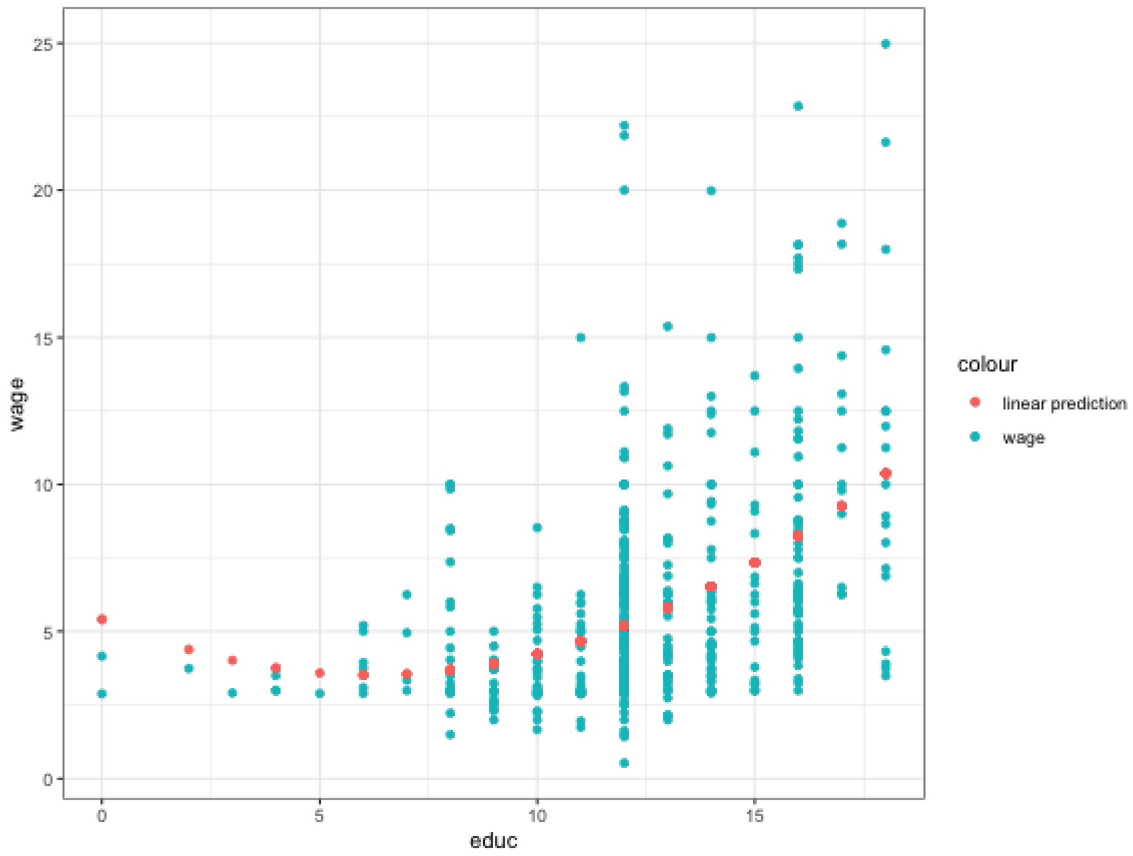
The new model would be:

stndfnl $= \beta_0 + \beta_1$ atndrte $+\beta_2$ priGP $A + \beta_3 ACT + \beta_4$ priGP $A^2 + \beta_5 ACT^2 + \beta_6$ priGPA atndrte $+\beta_7 ACT\cdot$ atndrte $+u$ .

Therefore, the partial effect of atndrte on stndfnl is $\beta_1 + \beta_6$ priGPA $+\beta_7 \mathrm{ACT}$.

```r
data(wage1, package='wooldridge')

# Regression
model_1 <- lm(wage ~ educ, wage1)
#summary(model_1)
wage1 %<>% mutate(wagehat1 = fitted(model_1))
ggplot(data = wage1, mapping = aes(x = educ)) +
  theme_bw() +
  geom_point(mapping = aes(y = wage, col = 'wage')) +
  geom_point(mapping = aes(y = wagehat1, col = 'linear prediction'))
```



25
20
15
10
5
0
wage
0
5
10
15
educ
colour
linear prediction
wage