Key observation:

If we plot $n$ observations

$$Y_i, X_i, i = 1, \ldots, n$$

on a dependent variable $Y$ and an independent variable $X$ in a scatterplot, then points $(X_i, Y_i)$ not on any straight line

$$Y = \alpha + \beta X$$

Reason: Omitted variables that affect $Y$ (beside $X$)
Contribution of omitted variables to $Y$ is the residual $u$

$$u = Y - \alpha - \beta X$$

See figure for illustration
HOUSEPRICE vs. SQUAREFT

$Y = \alpha + \beta X$

$Y_i - \alpha - \beta X_i$

$Y = \tilde{\alpha} + \tilde{\beta} X$

SQUAREFT

HOUSEPRICE
Econ 414

How do we find the hidden straight line if there are omitted variables?

**Ordinary Least Squares (OLS) method**

- **Compute residuals for** \( i = 1, \ldots, n \)
  \[
  u_i = Y_i - \alpha - \beta X_i
  \]

- **Compute the sum of squared residuals**
  \[
  S(\alpha, \beta) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2
  \]
The sum of squared residuals depends on the coefficients $\alpha, \beta$ (as is obvious from figure)

- Minimize $S(\alpha, \beta)$ with respect to $\alpha, \beta$

The last step is not obvious and will be justified later.
Math reminder

Summation

\[ \sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \]

Hence

(1) \[ \sum_{i=1}^{n} c = c + c + \cdots + c = nc \]

(2) \[ \sum_{i=1}^{n} cx_i = cx_1 + cx_2 + \cdots + cx_n = c \sum_{i=1}^{n} x_i \]

Squares

(3) \[ (a + b + c)^2 = a^2 + b^2 + c^2 + 2ab + 2bc + 2ac \]
OLS solution

\[ S(\alpha, \beta) = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2 = \]

by (3)

\[ = \sum_{i=1}^{n} (Y_i^2 + \alpha^2 + \beta^2 X_i^2 - 2\alpha Y_i - 2\beta X_i Y_i + 2\alpha\beta X_i) = \]

by (1) and (2)

\[ = \sum_{i=1}^{n} Y_i^2 + n\alpha^2 + \beta^2 \sum_{i=1}^{n} X_i^2 - 2\alpha \sum_{i=1}^{n} Y_i - 2\beta \sum_{i=1}^{n} X_i Y_i + 2\alpha\beta \sum_{i=1}^{n} X_i \]
Consider first case $\alpha = 0$

$$S(\beta) = \sum_{i=1}^{n} Y_i^2 + \beta^2 \sum_{i=1}^{n} X_i^2 - 2\beta \sum_{i=1}^{n} X_i Y_i$$

This is expression in $\beta$ of form

$$A + B\beta^2 + C\beta$$

with $A, B, C$ constants.

The graph of such a function (if $B > 0$) is a parabola with a minimum (see figure)
Econ 414

\[ S(\beta) \]

\[ A + B\beta^2 + C\beta \]

\[ \frac{C}{2B} \]

Slope=0
How do we find minimum?

At $\beta$ that minimizes $S(\beta)$ the slope of $S(\beta)$ is 0.

Slope of $S(\beta)$ in $\beta$ is derivative of $S(\beta)$

$$\frac{dS(\beta)}{d\beta}$$
Math reminder

\[
\frac{d A}{d \beta} = 0 \quad \frac{d C \beta}{d \beta} = C \frac{d \beta}{d \beta} = C
\]

\[
\frac{d B \beta^2}{d \beta} = B \frac{d \beta^2}{d \beta} = 2B \beta
\]

Using this

\[
\frac{d S(\beta)}{d \beta} = 2B \beta + C
\]

Value of \( \beta \) that minimizes \( S(\beta) \) solution to

\[
2B \beta + C = 0 \quad \text{or} \quad \beta = -\frac{C}{2B}
\]
Econ 414

Now

\[ B = \sum_{i=1}^{n} X_i^2 \quad C = -2 \sum_{i=1}^{n} X_i Y_i \]

and

\[ \hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \]
Econ 414

Next, case with $\alpha, \beta$

$$S(\alpha, \beta) = \sum_{i=1}^{n} Y_i^2 + n\alpha^2 + \beta^2 \sum_{i=1}^{n} X_i^2 - 2\alpha \sum_{i=1}^{n} Y_i - 2\beta \sum_{i=1}^{n} X_iY_i + 2\alpha\beta \sum_{i=1}^{n} X_i$$

This a quadratic function in $\alpha, \beta$. Graph has same shape as parabola, but in 3D.

In minimum slope is again 0.
Now we have slope in $\alpha$-direction and slope in $\beta$-direction. These are the partial derivatives with respect to $\alpha$ and $\beta$:

$$\frac{\partial S(\alpha, \beta)}{\partial \alpha} \quad \text{and} \quad \frac{\partial S(\alpha, \beta)}{\partial \beta}$$

Computed like ordinary derivatives, except that other argument is kept constant: for slope/derivative in $\alpha$-direction we keep $\beta$ constant and for slope/derivative in $\beta$-direction we keep $\alpha$ constant.

With

$$S(\alpha, \beta) = \sum_{i=1}^{n} Y_i^2 + n\alpha^2 + \beta^2 \sum_{i=1}^{n} X_i^2 - 2\alpha \sum_{i=1}^{n} Y_i - 2\beta \sum_{i=1}^{n} X_i Y_i + 2\alpha\beta \sum_{i=1}^{n} X_i$$

we have
Econ 414

\[
\frac{\partial S(\alpha, \beta)}{\partial \alpha} = 2n\alpha - 2\sum_{i=1}^{n} Y_i + 2\beta \sum_{i=1}^{n} X_i
\]

\[
\frac{\partial S(\alpha, \beta)}{\partial \beta} = -2\sum_{i=1}^{n} X_i Y_i + 2\beta \sum_{i=1}^{n} X_i^2 + 2\alpha \sum_{i=1}^{n} X_i
\]

\(S(\alpha, \beta)\) is minimal at the value of \((\alpha, \beta)\) where the slope in both directions is 0.

Hence

\[
\frac{\partial S(\alpha, \beta)}{\partial \alpha} = 2n\alpha - 2\sum_{i=1}^{n} Y_i + 2\beta \sum_{i=1}^{n} X_i = 0
\]

\[
\frac{\partial S(\alpha, \beta)}{\partial \beta} = -2\sum_{i=1}^{n} X_i Y_i + 2\beta \sum_{i=1}^{n} X_i^2 + 2\alpha \sum_{i=1}^{n} X_i = 0
\]
These are two (linear) equations in two unknowns $\hat{\alpha}, \hat{\beta}$. These are called the normal equations.

Solution:

1. Solve first equation for $\alpha$

$$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} Y_i - \hat{\beta} \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{Y} - \hat{\beta} \bar{X}$$

with $\bar{Y}, \bar{X}$ the sample average of $Y, X$
2. Substitute this solution in the second equation

\[- \sum_{i=1}^{n} X_i Y_i + \hat{\beta} \sum_{i=1}^{n} X_i^2 + \left( \frac{1}{n} \sum_{i=1}^{n} Y_i - \hat{\beta} \frac{1}{n} \sum_{i=1}^{n} X_i \right) \sum_{i=1}^{n} X_i = 0\]

and

\[\hat{\beta} \left( \sum_{i=1}^{n} X_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} X_i \right)^2 \right) = \sum_{i=1}^{n} X_i Y_i - \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i\]
Econ 414

Now

\[ \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} (X_i^2 - 2X_i \bar{X} - \bar{X}^2) = \sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 - n\bar{X}^2 = \]

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

and

\[ \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} (X_iY_i - \bar{Y}X_i - \bar{X}Y_i + \bar{X}\bar{Y}) = \]

\[= \sum_{i=1}^{n} X_iY_i - n\bar{Y}\bar{X} - n\bar{X}\bar{Y} + n\bar{X}\bar{Y} = \sum_{i=1}^{n} X_iY_i - \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i \]
Using these results we find for $\hat{\beta}$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$
Hence the values of $\alpha, \beta$ that minimize the sum of squared residuals are

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

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

These are the Ordinary Least Squares (OLS) solutions to the problem of fitting a straight line to the points in a scatterplot. The least squares line

$$Y = \hat{\alpha} + \hat{\beta}X$$

is the straight line that ‘fits’ the scatterplot best (see figure).
\[ Y = \hat{\alpha} + \hat{\beta} X \]

\[ \hat{\alpha} = 52.35091 \]

\[ \hat{\beta} = 0.138750 \]

\[ R^2 = 0.820522 \]
Econ 414

Note

• If we divide numerator and denominator by $n$ or $n - 1$, then

$$
\hat{\beta} = \frac{\text{Sample covariance of } X \text{ and } Y}{\text{Sample variance of } X}
$$

• From the OLS solution for $\hat{\alpha}$

$$
\bar{Y} = \hat{\alpha} + \hat{\beta} \bar{X}
$$

In words: The point $(\bar{Y}, \bar{X})$ is on the least squares line
The residuals with respect to the least squares line are the OLS residuals

\[ e_i = Y_i - \hat{\alpha} - \hat{\beta}X_i \]

From the normal equations

\[ 0 = -n\hat{\alpha} + \sum_{i=1}^{n} Y_i - \hat{\beta} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}X_i) = \sum_{i=1}^{n} e_i \]

\[ 0 = \sum_{i=1}^{n} X_iY_i - \hat{\beta} \sum_{i=1}^{n} X_i^2 - \hat{\alpha} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i) = \sum_{i=1}^{n} X_i e_i \]
Hence

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

\[
\sum_{i=1}^{n} X_i e_i = \frac{1}{n} \sum_{i=1}^{n} X_i e_i = 0
\]

In words: The sample average (and sum) of the OLS residuals is 0 and the sample covariance of these residuals and \( X \) is also 0.
These are all consequences of the fact that we minimize the sum of squared residuals. How good is the fit of the straight line to the scatterplot?

Define the fitted value

\[ \hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i \]

then by the definitions

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

Because the OLS residuals have average 0

\[ \bar{Y} = \bar{\hat{Y}} \]
In words: $Y$ and $\hat{Y}$ have the same sample average

Using this we have

$$(Y_i - \bar{Y})^2 = (\hat{Y}_i - \bar{Y})^2 + e_i^2 + 2e_i(\hat{Y}_i - \bar{Y})$$

If we take the sum over $i$ we first observe

$$\sum_{i=1}^{n} e_i(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^{n} e_i \hat{Y}_i - \bar{Y} \sum_{i=1}^{n} e_i =$$

$$= \sum_{i=1}^{n} e_i (\hat{\alpha} + \hat{\beta}X_i) = \hat{\alpha} \sum_{i=1}^{n} e_i + \hat{\beta} \sum_{i=1}^{n} e_i X_i = 0$$
Econ 414

Note that this implies that the sample covariance between the OLS residuals and the OLS fitted values is 0: OLS decomposes $Y_i$ into two parts (residual and fitted value) that have covariance 0, i.e. are unrelated

Using this we find

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

The sample variance of $Y$ is equal to the sum of the sample variance of $\hat{Y}$ and the sample variance of $e$ or

Total Variance = Explained variance +
  +Unexplained variance
A measure of goodness of fit is

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

This is fraction of total variance that is explained by fitted straight line.

Note \( R^2 = 1 \) if and only if \( Y_i = \hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i \), i.e. if all observations are on the straight line.
Also $R^2 = 0$ if and only if

$$\bar{Y} - \hat{\beta}(\bar{X} - X_i) = \hat{\alpha} + \hat{\beta}X_i = \hat{Y}_i = \bar{Y} = \bar{Y}$$

or

$$\hat{\beta}(\bar{X} - X_i) = 0$$

for all $i = 1, \ldots, n$. If the $X_i$ are not all equal, then this can only be the case if $\hat{\beta} = 0$. In that case $X$ does not help in explaining $Y$.

These are the extreme values for $R^2$. We have

$$0 \leq R^2 \leq 1$$