Multiple Regression Analysis

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

1. Estimation

Multiple Regression Analysis

- The main drawback of the SLR analysis for empirical work is that it is very difficult to draw "ceteris paribus" conclusions about how *x* affects *y*.
- Multiple Linear Regression (MLR) analysis is more amenable to "ceteris paribus" analysis because it allows us to explicitly control for many other factors that simultaneously affect the dependent variable, *y*.

A Model with Two Regressors

- Consider the **population model**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

β_0 the intercept,

β_1 measures the Δy with respect to x_1, holding other factors fixed, and

β_2 measures the Δy with respect to x_2, holding other factors fixed.

A Model with Two Regressors

- In this model the key assumption about how u is related to the regressors is

$$E(u|x_1,x_2) = 0$$

- As in the SLR the important part of the assumption is $E(u|x_1,x_2) = E(u)$, given that, as long as an intercept, β_0, is included in the equation, we can assume that $E(u) = 0$

A Model with Two Regressors

- Note that this is equivalent to

$$E(y|x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

A Model with Two Regressors

- This model can accommodate fairly arbitrary forms of dependence between y and x.
- For example,

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

Now $\Delta y \approx (\beta_1 + 2\beta_2 x)\Delta x$.

So, in a particular application, the definitions of the independent variables are crucial, but for theoretical developments we can ignore these details.

A Model with *k* Regressors

- There is no need to stop with two regressors.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

β_0 the intercept,

β_j, $j = 1,2,\ldots,k$; are usually referred as slope parameters, that measure the Δy with respect to x_j, holding other factors fixed.

The variable u is the error term or disturbance. It contains factors other than $x_1, x_2, \ldots, x_k$ that affect y.

A Model with *k* Regressors

- The MLR has many similarities with the SLR.
- We have the same terminology.
- As before, the "linear" term in MLR means that the population model is linear in parameters, and not necessarily in variables.

A Model with k Regressors

- The key assumption now about how u is related to the regressors is

$$E(u|x_1,x_2,\ldots,x_k) = 0$$

- At a minimum, this requires that all factors in u be uncorrelated with the regressors.
- It also means that we have correctly accounted for the functional relationships between y and $x_1,x_2,\ldots,x_k$.

Ordinary Least Squares

- Basic idea of regression is to estimate the population parameters, (β_0 , β_1,…, β_k), from a sample.
- Let $\{(y_i, x_{ij}): i = 1,\ldots,n;\ j = 1,\ldots,k\}$ denote a random sample of size n from the population.
- For each observation in this sample, it will be the case that

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$$

Deriving OLS Estimates

- To derive the OLS estimates we need to realize that our key assumption implies that

1. $E(u) = 0$
2. $E(x_j u) = 0$, $j = 1,2,\ldots,k$

A set of k+1 population moment conditions that can be imposed on the sample.

This give us a set of k+1 equations in k+1 unknowns.

Deriving OLS Estimates

- An alternate approach is to minimize a sum of squares residuals,

$$\min_{b_0, b_1, b_2, \ldots, b_k} \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik} \right)^2$$

- First order conditions for this problem give us a set of k+1 equations in k+1 unknowns.
- See Appendix 3A.1 for a derivation.

Deriving OLS Estimates

- In any case the system we have to solve is:

$$\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}-\ldots-\hat{\beta}_k x_{ik}\right)=0$$

$$\sum_{i=1}^{n}x_{i1}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}-\ldots-\hat{\beta}_k x_{ik}\right)=0$$

$$\sum_{i=1}^{n}x_{i2}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}-\ldots-\hat{\beta}_k x_{ik}\right)=0$$

$$\vdots$$

$$\sum_{i=1}^{n}x_{ik}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}-\ldots-\hat{\beta}_k x_{ik}\right)=0$$

Deriving OLS Estimates

- A set of k+1 equations in k+1 unknowns.
- This system is known as the **normal equations**.
- We must assume that this system has a unique solution in terms of the $\hat{\beta}_j$'s, $j = 0,1,\ldots,k$.
- Note that for $\hat{\beta}_0$ the solution is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2 - \ldots - \hat{\beta}_k\bar{x}_k$$

More on the OLS estimates

- Given the OLS estimates, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$, the **fitted value** for y when $x_j = x_{ij}, \forall j$ is given by
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}$$
- This is the **OLS regression line** or **Sample Regression Function** (SRF). The value that the model predicts for y when $x_j = x_{ij}, \forall j$.
- There is a fitted value for each observation in the sample.

More on the OLS estimates

- The **residual** for observation *i* is the difference between the actual y_i and its fitted value, $\hat{y}_i$,

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

- Again there are *n* residuals.
- The residual, $\hat{u}$, is an estimate of the error term, *u*, and is the difference between the fitted line (SRF) and the sample point.

More on the OLS estimates

- There is a residual for each observation.
- If $\hat{u}_i > 0$, then $\hat{y}_i < y_i$, which means that, for this observation y_i is underpredicted.
- If $\hat{u}_i < 0$, then $\hat{y}_i > y_i$, which means that, for this observation y_i is overpredicted.

Interpreting Multiple Regression

- More important than the details underlying the computation of the $\hat{\beta}_j$'s is the interpretation of the estimated equation.

- The estimates, $\hat{\beta}_j$'s, have a partial effect, or "ceteris paribus" interpretations.

Interpreting Multiple Regression

- From

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \ldots + \hat{\beta}_k \Delta x_k$$

so holding $x_2,\ldots,x_k$ fixed implies that

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1$$

The coefficient on x_1 measures the change in $\hat{y}$ due to a one-unit increase in x_1, holding $x_2,\ldots,x_k$ fixed.

Interpreting Multiple Regression

- Thus, we have controlled the variables $x_2,\ldots,x_k$ when estimating the effect of x_1 on y.

- That is, each $\hat{\beta}_j$ has a "ceteris paribus" interpretation. So including additional regressors allows us to obtain partial effects.

"Holding other Factors Fixed"

- The power of multiple regression analysis is that it allows us to do in nonexperimental environments what natural scientists are able to do in a controlled laboratory setting: keep other factors fixed.

Algebraic Properties of OLS

- The sum of the OLS residuals is zero.
- Thus, the sample average of the OLS residuals is zero as well.
- The sample covariance between the regressors and the OLS residuals is zero.
- The OLS regression line always goes through the mean of the sample.

Algebraic Properties (precise)

$$(1)\quad \Sigma_{i=1}^{n}\hat{u}_i = 0 \qquad \Rightarrow \qquad \bar{\hat{u}} = \frac{\Sigma_{i=1}^{n}\hat{u}_i}{n} = 0$$

$$(2)\quad \Sigma_{i=1}^{n}x_{ij}\hat{u}_i = 0 \quad \forall j = 1,2,\ldots,k$$

(3) $(\bar{y}, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k)$ is on the regression line

$$\Rightarrow \quad \bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_1\bar{x}_2 + \ldots + \hat{\beta}_k\bar{x}_k$$

Algebraic Properties (precise)

Writing $\quad y_i = \hat{y}_i + \hat{u}_i \quad$ we have

$$(1) \qquad \Sigma_{i=1}^{n} \hat{u}_i = 0 \qquad \Rightarrow \qquad \bar{y} = \bar{\hat{y}}$$

$$(1)+(2) \qquad \left.\begin{array}{l} \Sigma_{i=1}^{n} \hat{u}_i = 0 \\ \Sigma_{i=1}^{n} x_{ij} \hat{u}_i = 0 \quad \forall j \end{array}\right\} \qquad \Rightarrow \qquad \Sigma_{i=1}^{n} \hat{y}_i \hat{u}_i = 0$$

this last one implies that the sample covariance between fitted values, $\hat{y}_i$, and residuals, $\hat{u}_i$, is zero.

Algebraic Properties

- Thinking of each observation as being made up of an explained part, and an unexplained part, $y_i = \hat{y}_i + \hat{u}_i$, we can view OLS as decomposing each y_i into two parts, a fitted value and a residual. The fitted values and residuals are uncorrelated in the sample.

Sum of Squares Decomposition

- Define:

1. Total Sum of Squares (SST)

$$\text{SST} \equiv \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2$$

2. Explained Sum of Squares (SSE)

$$\text{SSE} \equiv \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2$$

3. Residual Sum of Squares (SSR)

$$\text{SSR} \equiv \sum_{i=1}^{n} \hat{u}_i^2$$

Sum of Squares Decomposition

- SST is a measure of the total sample variation in the y_i.
- It can be shown that total variation in y, SST, can always be expressed as the sum of the explained variation, SSE, and the unexplained variation, SSR. Thus

$$SST = SSE + SSR$$

Proof that SST = SSE + SSR

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})]^2$$

$$=\sum_{i=1}^{n}[\hat{u}_i+(\hat{y}_i-\bar{y})]^2$$

$$=\sum_{i=1}^{n}\hat{u}_i^2+2\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i-\bar{y})+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

$$=\mathrm{SSR}+\underbrace{\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i-\bar{y})}_{=0}+\mathrm{SSE}$$

Given the above properties, so SST = SSE + SSR

Goodness-of-Fit

- How well our SRF fits our sample data?
- We can compute the fraction of the total sum of squares (SST) that is explained by the model (SSE), call this the R-squared, R^2, of regression:

$$R^2 = SSE/SST = 1 - SSR/SST$$

Goodness-of-Fit

- $100.R^2$ is the percentage of the sample variation in y that is explained by $\hat{y}$ (the model).
- $R^2 \in [0,1]$
- If $R^2 = 1$, then we have a perfect fit, $\hat{u}_i = 0$ for all observations.
- If $R^2 = 0$, or close to zero, then we have a poor fit: very little variation in y is explained by $\hat{y}_i$.

Goodness-of-Fit

- It can be shown that R^2 is equal to:

1. The square of the sample correlation coefficient between y_i and $\hat{y}_i$.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\left(\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}) \right)^2}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2 \right) \left(\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2 \right)} = r_{y,\hat{y}}^2$$

Please show this as an exercise!.

Goodness-of-Fit

- An important fact about R^2 is that it never decreases, and it usually increases when another independent variable is added to a regression.
- This algebraic fact follows because, by definition, the sum of squared residuals never increases when additional regressors are added to the model.
- The fact that R^2 never decreases when any variable is added to a regression makes it a poor tool for deciding whether one variable or several variables should be added to a model.
- The factor that should determine whether an explanatory variable belongs in a model is whether the explanatory variable has a nonzero partial effect on y in the population.
- For this we need to perform significance statistical tests.

Goodness-of-Fit

- Because we want high explanatory power for our models, we look, other things equal, for high R^2 in our regressions.
- It is worth emphasizing now that a seemingly low R^2 does not necessarily mean that an OLS regression equation is useless.
- It is still possible that the OLS estimates are reliable estimates of the "ceteris paribus" effects of each regressor on *y*.
- Generally, a low R^2 indicates that it is hard to predict individual outcomes on *y* with much accuracy, which is a general feature in the social sciences.
- Goodness of fit is not the only feature we look for in a regression equation.

A “Partialling Out” Interpretation

- When applying OLS, we don’t need to know explicit formulas for the $\hat{\beta}_j$'s that solves the above system of equations.
- The software does the job for you.
- Nevertheless, for certain derivations, it is useful to know explicit formulas for the $\hat{\beta}_j$'s .
- In addition, these formulas also shed light on the workings of OLS.

A "Partialling Out" Interpretation

- Consider the case $k = 2$,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

then

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$$

where $\hat{r}_{i1}$ are the OLS residuals from a SLR of x_1 on x_2, this is, residuals from the estimated regression $\hat{x}_1 = \hat{\gamma}_0 + \hat{\gamma}_2 x_2$.

A "Partialling Out" Interpretation

- As an **exercise** show that the above formula is correct.
- *Hint:*

(i) Consider the second normal equation,

$$\sum_{i=1}^{n} x_{i1}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right) = 0$$

(ii) Use the algebraic properties of the MLR of y on x_1 and x_2 and of the SLR of x_1 on x_2.

A "Partialling Out" Interpretation

- Previous equation implies that regressing y on x_1 and x_2 simultaneously gives same effect of x_1 on y as regressing y on residuals from a previous regression of x_1 on x_2.
- This means that only the part of x_1 that is uncorrelated with x_2 is being related to y, so we're estimating the effect of x_1 on y after x_2 has been "partialled out".

A "Partialling Out" Interpretation

- In the general model with k regressors, $\hat{\beta}_1$ can still be written as in the previous equation, but residuals $\hat{r}_1$ come from the regression of x_1 on $x_2, x_3, \ldots, x_k$.

  See Appendix 3A.2 for a general proof.
- Thus, $\hat{\beta}_1$ measures the effect of x_1 on y after we have discounted the (linear) effect of $x_2, x_3, \ldots, x_k$, so these variables have been netted out.

A “Partialling Out” Interpretation

- Note that the above argument also implies that MLR coefficients can always be estimated in two steps:

1. Regress one independent variables on the others plus a constant and take the residuals.
2. Regress *y* on these residuals.

Simple *versus* Multiple Regression Estimates ($k = 2$)

- If we compare the OLS estimates in the SLR, say $\tilde{\beta}_1$, and in the MLR, say $\hat{\beta}_1$. Generally, $\tilde{\beta}_1 \neq \hat{\beta}_1$ unless:

1. $\hat{\beta}_2 = 0$, this is, the partial effect of x_2 on y is zero, or
2. x_1 and x_2 are uncorrelated, $r_{x_1, x_2} = 0$.

Regression Through the Origin

- Regression through the origin constraints the estimated intercept to be zero.
- If $\beta_0 \neq 0$, then the slope estimates will be biased.
- Another problem is that if R^2 is defined as $1 - SSR/SST$ then R^2 can be negative.
- **Advise**: always include an intercept in your regressions.

Statistical Properties of OLS

- We defined de population model

  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$, and we claimed that the key assumption for the MLR analysis to be useful is that $E(u|x_1,\ldots,x_k) = 0$.
- We now return to the population model and study the statistical properties of OLS estimators, $\hat{\beta}_j$, considered as estimators of the population parameters, β_j.

Assumptions

- **MLR.1: LINEAR IN PARAMETERS**

  The population model is linear in parameters and given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

Assumptions

- **MLR.2: RANDOM SAMPLING**

  We have a random sample from of size n, $\{(y_i, x_{ij}): i = 1,2,3,\ldots,n; j = 1,2,\ldots,k\}$, from the population model.

  Thus we can write the population model in terms of the sample,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$$
$$i = 1,2,3,\ldots,n$$

Assumptions

- **MLR.3: ZERO CONDITIONAL MEAN**

$$E(u|x_1,\dots,x_k) = 0$$

For a random sample, this assumption implies that

$$E(u_i|x_{i1},\dots,x_{ik}) = 0\,, \quad i = 1,2,3,\dots,n$$

NOTE: Derivations will be conditional on the sample values, x's.

Assumption MLR.3

- Assumption MLR.3 can fail if:

1. An important factor that is correlated with any $x_1, x_2, \ldots, x_k$ is omitted from the estimated equation (MLR.3 always fail in this case).
2. The functional relationship between y and the explanatory variables, x's, is misspecified.

Assumption MLR.3: Notation

- When MLR.3 holds, we often say that we have **exogenous explanatory variables**.
- If x_j is correlated with u for any reason, then x_j is said to be an **endogenous explanatory variable**.

- We shall denote $\mathbf{x} = (x_1, x_2, \ldots, x_k)$.

Assumptions

- **MLR.4: NO PERFECT COLLINEARITY**

  In the sample, and therefore in the population, none of the independent variables is constant, and there are no *exact linear* relationships among the independent variables.

Assumption MLR.4

- Assumption MLR.4 concerns only the independent variables.
- If an independent variable is an exact linear combination of the other independent variables, then we say the model suffers from **perfect collinearity**, and it cannot be estimated by OLS.
- Note that Assumption MLR.4 *does* allow the independent variables to be correlated; they just cannot be *perfectly* correlated.

Assumption MLR.4

- Assumption MLR.4 can fail if we are not careful in specifying our model, i.e. if we introduce an accounting relationship between explanatory variables.
- Assumption MLR.4 also fails if the sample size, n, is too small in relation to the number of parameters being estimated. In particular, MLR.4 fails if $n < k + 1$.
- Intuitively, this makes sense: to estimate $k + 1$ parameters, we need at least $k + 1$ observations.

Assumption MLR.4

- If the model is carefully specified and $n \geq k + 1$, Assumption MLR.4 can fail in rare cases only due to bad luck in collecting the sample.

- Under MLR.1 through MLR.4 OLS estimators are unbiased.

Unbiasedness of OLS

- **THEOREM 3.1 UNBIASEDNESS OF OLS**

  Under assumptions MLR.1 to MLR.4

  $$E(\hat{\beta}_j) = \beta_j \qquad j = 0, 1, 2, \ldots, k$$

  PROOF:

  Appendix 3A.3

Unbiasedness of OLS

- Remember that when we say that OLS is unbiased under Assumptions MLR.1 through MLR.4, we mean that the *procedure* by which the OLS estimates are obtained is unbiased when we view the procedure as being applied across all possible random samples.
- This property says nothing about a particular sample.

Misspecification

- We speak of misspecification when we end up estimating a model different from the population model.
- Why are we going to do such a thing?
- Because de population model, at least in social science, is always unknown. So there is always a chance that the estimated model is misspecified.

Misspecification

- There are many types of misspecification, we shall consider now only two:

1. Inclusion of an irrelevant variable.
2. Exclusion a relevant variable.

- Remember that the statistical properties take the population model as benchmark.

Inclusion of an Irrelevant Variable

- One (or more) of the independent variables included in the regression model don't belong to the population model, i.e. it has no partial effect on *y* in the population, that is, its population coefficient is zero.

  Population:

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

  In terms of conditional expectations:

  $$E(y|x_1,x_2,x_3) = E(y|x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Inclusion of an Irrelevant Variable

Estimated model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

- What are the effects on the OLS estimates?

1. In terms of unbiasedness there is no effect, $\hat{\beta}_j$ are all unbiased.
2. The variance, however, will increase with respect to the case in which x_3 is (correctly) omitted.

- This is a general result.

Exclusion of a Relevant Variable

- One variable that actually belongs to the population model is omitted in the regression model.

  Population:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

  Estimated model:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

Exclusion of a Relevant Variable

- Our primary interest is in the partial effect of x_1 on y.
- In order to get un unbiased estimator of β_1, we *should* regress y on x_1 and x_2.
- However, due to ignorance or data unavailability, we estimate the model by *excluding* x_2.
- Then the estimator of β_1 will be biased.

Exclusion of a Relevant Variable

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)y_i}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2} = \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(\overbrace{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+u_i}^{y_i})}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$$

$$= \beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)x_{i2}}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2} + \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)u_i}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$$

Exclusion of a Relevant Variable

- Taking expectations conditional on the sample values of x_1 and x_2

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)x_{i2}}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$

- Thus $E(\tilde{\beta}_1) \neq \beta_1$ in general: so $\tilde{\beta}_1$ is biased for β_1.

Exclusion of a Relevant Variable

- The ratio $\dfrac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)x_{i2}}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$ is just the OLS slope coefficient from the regression of x_2 on x_1:

$$\hat{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1$$

- So $E(\tilde{\beta}_1) = \beta_1 + \beta_2\tilde{\delta}_1$, which implies that the bias in $\tilde{\beta}_1$ is $E(\tilde{\beta}_1) - \beta_1 = \beta_2\tilde{\delta}_1$.
- This is often called the omitted variable bias.

Exclusion of a Relevant Variable

- There are two cases where $\tilde{\beta}_1$ is unbiased:

1. If $\beta_2 = 0$, so there is no misspecification.
2. If $\tilde{\delta}_1 = 0$, so x_1 and x_2 are uncorrelated in the sample.

- The size of the bias is determined by the sizes of β_2 and $\tilde{\delta}_1$.
- The sign of the bias depends on the signs of both β_2 and $\tilde{\delta}_1$.

Summary of Direction of Bias

	$\text{Corr}(x_1,x_2) > 0$	$\text{Corr}(x_1,x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Exclusion of a Relevant Variable

- If $E(\tilde{\beta}_1) > \beta_1$, then we say that $\tilde{\beta}_1$ has un **upward bias**.
- If $E(\tilde{\beta}_1) < \beta_1$, then we say that $\tilde{\beta}_1$ has a **downward bias**.
- The phrase **biased towards zero** refers to cases where $E(\tilde{\beta}_1)$ is closer to zero than β_1.

Omitted Variable Bias: More General Cases

- In a general model we must remember that correlation between a single explanatory variable and the error term generally results in all OLS estimators being biased.
- Beyond that we cannot determine the direction of the bias, except in special cases.
- Technically, can only sign the bias for the more general case if all of the included x's are uncorrelated

Variance of the OLS Estimators

- Now we know that the sampling distribution of our estimator is centered around the true parameter.
- How spread out this distribution is? This will be a measure of uncertainty.
- It is much easier to think about this variance under an additional assumption.

Assumptions

- **MLR.5:** **HOMOSKEDASTICITY**

$$\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2$$

- Assumptions MLR.1-MLR.5 are collectively known as the **Gauss-Markov assumptions**.

Variance of the OLS Estimators

- The homoskedasticity assumption is quite distinct from the zero conditional mean assumption, $E(u|\mathbf{x}) = 0$. MLR.3 involves the expected value of u, while MLR.5 concerns the variance of u.
- Homoskedasticity plays no role in showing that the $\hat{\beta}_j$ are unbiased.
- We add MLR.5 because it simplifies the variance calculations and because it implies that OLS has certain efficiency properties.

Variance of the OLS Estimators

- $\mathrm{Var}(u|\mathbf{x}) = \sigma^2 = \mathrm{E}(u^2|\mathbf{x}) - [\mathrm{E}(u|\mathbf{x})]^2$
- $\mathrm{E}(u|\mathbf{x}) = 0$, so $\sigma^2 = \mathrm{E}(u^2|\mathbf{x}) = \mathrm{E}(u^2) = \mathrm{Var}(u)$
- Thus σ^2 is also the unconditional variance, called the error variance.
- σ, the square root of the error variance, is called the standard deviation of the error.

Variance of the OLS Estimators

- We can say:

$$E(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

$$\text{and} \quad \text{Var}(y|\boldsymbol{x}) = \sigma^2.$$

- So, the conditional expectation of y given $\boldsymbol{x}$ is linear in $\boldsymbol{x}$, but the variance of y given $\boldsymbol{x}$ is constant.
- When $\text{Var}(u|\boldsymbol{x})$ depends on $\boldsymbol{x}$, the error term is said to exhibit heteroskedasticity. Since $\text{Var}(u|\boldsymbol{x}) = \text{Var}(y|\boldsymbol{x})$, heteroskedasticity is present whenever $\text{Var}(y|\boldsymbol{x})$ is a function of $\boldsymbol{x}$.

Variance of OLS estimators

- **THEOREM 3.2 SAMPLING VARIANCES OF OLS SLOPE ESTIMATORS**

Under assumptions MLR.1 to MLR.5

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j\left(1 - R_j^2\right)} \qquad j = 1, 2, 3, \ldots, k$$

where these are conditional on the sample values $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, R_j^2 is the R-squared from regressing x_j on all other x's and

$$SST_j = \sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2$$

Variance of OLS estimators

- **PROOF:** Appendix 3A.5
- All of the Gauss-Markov assumptions are used in obtaining this formula.
- The size of $Var(\hat{\beta}_j)$ is practically important. A larger variance means a less precise estimator, and this translates into larger confidence intervals and less accurate hypotheses tests.
- Theorem 3.2 shows that the variance depends on three factors: σ^2, SST_j and R_j^2 .

Variance of OLS estimators: σ^2

- The larger the error variance, σ^2, the larger the variance of the slope estimates.
- This is not at all surprising: more "noise" in the equation, a larger σ^2, makes it more difficult to estimate the partial effect of any *x*'s on *y*, and this is reflected in higher variances for the OLS slope estimators.
- Since σ^2 is a feature of the population, it has nothing to do with the sample size.

Variance of OLS estimators: SST_j

- The larger the total variation in x_j is, the smaller is $Var(\hat{\beta}_j)$.
- Everything else being equal, for estimating β_j we prefer to have as much sample variation in x_j as possible.
- This is the component of the variance that systematically depends on the sample size.
- So $\uparrow n \;\Rightarrow\; \downarrow Var(\hat{\beta}_j)$.
- $SST_j = 0$ is not allowed by Assumption MLR.4.

Variance of OLS estimators: R_j^2

- R_j^2 is the proportion of the total variation in x_j that can be explained by the *other* independent variables appearing in the equation.
- For a given σ^2 and SST_j, the smallest $Var(\hat{\beta}_j)$ is obtained when $R_j^2 = 0$, which happens if, and only if, x_j has zero sample correlation with *every other* independent variable.
- The case $R_j^2 = 1$ is ruled out by Assumption MLR.4, since $R_j^2 = 1$ means that, in the sample, x_j is an *exact linear combination* of the other x's in the regression.

Variance of OLS estimators: R_j^2

- A more relevant case is when R_j^2 is "close" to 1.
- As $R_j^2 \to 1 \;\Rightarrow\; Var(\hat{\beta}_j) \to \infty$
- High, but not perfect, correlation between two or more independent variables is called **multicollinearity**.
- The case where R_j^2 is "close" to one is *not* a violation of Assumption MLR.4.
- Since multicollinearity violates none of our assumptions, the "problem" of multicollinearity is not really well-defined.

Variance of OLS estimators: R_j^2

- We say that multicollinearity arises for estimating β_j when R_j^2 is "close" to one, but there is no absolute number that we can cite to conclude that multicollinearity is really a problem for the precision of the estimates.
- Although the problem of multicollinearity cannot be clearly defined, it is true, that for estimating β_j, it is better to have less correlation between x_j and the other independent variables.
- The effect of $R_j^2 \rightarrow 1$ is the same as $SST_j \rightarrow 0$.

Misspecified Models

- Consider de population model ($k = 2$)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- Consider two estimators of β_1:

1. From the regression of y on x_1 and x_2

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

2. From the regression of y on x_1 only

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

Misspecified Models

- From the previous results:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_1\left(1 - R_1^2\right)}$$

and

$$Var(\tilde{\beta}_1) = \frac{\sigma^2}{SST_1}$$

Misspecified Models

- Assuming x_1 and x_2 are not uncorrelated, we can draw the following conclusions:

1. When $\beta_2 = 0$, $\tilde{\beta}_1$ and $\hat{\beta}_1$ are both unbiased, and $Var(\tilde{\beta}_1) < Var(\hat{\beta}_1)$.

2. When $\beta_2 \neq 0$, $\tilde{\beta}_1$ is biased, $\hat{\beta}_1$ is unbiased, and $Var(\tilde{\beta}_1) < Var(\hat{\beta}_1)$.

Misspecified Models

- Including irrelevant variables ($\beta_2 = 0$) increases the variance of the estimators, but they are unbiased.
- Excluding relevant variables ($\beta_2 \neq 0$) causes the variance to decrease (assuming we condition on x_1 and x_2), but the estimator is biased. The variance is not centered at the population parameter we are interested in.

Estimating the Error Variance

- We don't know what the error variance, σ^2, is, and we cannot estimate it from the errors, u_i, because we don't observe the errors.
- $\sigma^2 = E(u^2)$, so an unbiased "estimator" would be $n^{-1}\Sigma_{i=1}^{n} u_i^2$.
- Unfortunately, this is not a true estimator, because we don't observe the errors u_i. But, we do have estimates of the u_i, namely the OLS residuals $\hat{u}_i$.

Estimating the Error Variance

- The relation between errors and residuals is given by

$$\hat{u}_i = y_i - \hat{y}_i = u_i - \left(\hat{\beta}_0 - \beta_0\right) - \sum_{j=1}^{k}\left(\hat{\beta}_j - \beta_j\right)x_{ij}$$

- Hence $\hat{u}_i$ is not the same as u_i, although the difference between them does have an expected value of zero.

Estimating the Error Variance

- If we replace the errors with the OLS residuals, we have $n^{-1}\Sigma_{i=1}^{n}\hat{u}_i^2 = \text{SSR}/n$
- This is a true estimator, because it gives a computable rule for any sample of the data, x and y.
- However, this estimator is biased, essentially because it does not account for the $k + 1$ restrictions that must be satisfied by the OLS residuals, $n^{-1}\Sigma_{i=1}^{n}\hat{u}_i = 0$ and $n^{-1}\Sigma_{i=1}^{n}x_{ij}\hat{u}_i = 0 \;\; \forall j$

Estimating the Error Variance

- One way to view these restrictions is this: If we know $n - (k + 1)$ of the residuals, we can get the other $k + 1$ residuals by using the restrictions implied by the moment conditions.
- Thus, there are only $n - (k + 1)$ **degrees of freedom** (*df*) in the OLS residuals, as opposed to n degrees of freedom in the errors.

***df*: observations – parameters estimated**

Estimating the Error Variance

- The unbiased estimator of σ^2 that we will use makes a degrees of freedom adjustment:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} = \frac{\text{SSR}}{n-k-1}$$

- **THEOREM 2.3 UNBIASED ESTIMATOR OF σ^2**

  Under assumptions MLR.1 to MLR.5

$$E(\hat{\sigma}^2) = \sigma^2$$

Estimating the Error Variance

- If $\hat{\sigma}^2$ is plugged into the variance formulas we then have unbiased estimators of $Var(\hat{\beta}_j)$.
- The natural estimator of σ is $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ and is called the **standard error of the regression**.
- Since $sd(\hat{\beta}_j) = \dfrac{\sigma}{\left[\mathrm{SST}_j\left(1-\mathrm{R}_j^2\right)\right]^{\frac{1}{2}}}$,

its natural estimator is $se(\hat{\beta}_j) = \dfrac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2\left(1-\mathrm{R}_j^2\right)}}$

Estimating the Error Variance

- Note that $se(\hat{\beta}_j)$, the **standard error** of $\hat{\beta}_j$, is view as a random variable when we think of running OLS over different samples; this is because $\hat{\sigma}$ varies with different samples.

- The standard error of any estimate gives us an idea of how precise the estimator is.

Efficiency of OLS: Gauss-Markov Theorem

- **THEOREM 3.2 GAUSS-MARKOV THEOREM**

  Under assumptions MLR.1 through MLR.5 OLS are the Best Linear Unbiased Estimators (BLUE) of the population parameters.

  PROOF: Appendix 3A.6

Gauss-Markov Theorem

- What is the meaning of the Gauss-Markov Theorem?
- If we restrict the set of eligible estimators to the estimators that are:

1. Linear, so $b_j = \Sigma_{i=1}^{n} w_{ij} y_i$
2. Unbiased, so the weights, w_j, satisfy some restrictions.

Then, OLS is "best".

Where "best" is defined as the smallest variance, so

$$Var(\hat{\beta}_j) \leq Var(b_j) \qquad j = 0,1,2,\ldots,k$$

Gauss-Markov Theorem

- Thus, if MLR.1 through MLR.5 holds then we use OLS.

Appendix: Algebra for $k = 2$

- The system we have to solve is:

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right) = 0$$

$$\sum_{i=1}^{n} x_{i1}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right) = 0$$

$$\sum_{i=1}^{n} x_{i2}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right) = 0$$

Appendix: Algebra for $k = 2$

- From the first equation:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

and substituting in the other two:

$$\sum_{i=1}^{n} x_{i1}\Big(y_i - \bar{y} - \hat{\beta}_1 (x_{i1} - \bar{x}_1) - \hat{\beta}_2 (x_{i2} - \bar{x}_2)\Big) = 0$$

$$\sum_{i=1}^{n} x_{i2}\Big(y_i - \bar{y} - \hat{\beta}_1 (x_{i1} - \bar{x}_1) - \hat{\beta}_2 (x_{i2} - \bar{x}_2)\Big) = 0$$

Appendix: Algebra for $k = 2$

- Alternatively:

$$\hat{\beta}_1 \sum_{i=1}^{n} x_{i1}(x_{i1} - \bar{x}_1) + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}(x_{i2} - \bar{x}_2) = \sum_{i=1}^{n} x_{i1}(y_i - \bar{y})$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_{i2}(x_{i1} - \bar{x}_1) + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}(x_{i2} - \bar{x}_2) = \sum_{i=1}^{n} x_{i2}(y_i - \bar{y})$$

$$\hat{\beta}_1 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}(x_{i2} - \bar{x}_2) = \sum_{i=1}^{n} x_{i1}(y_i - \bar{y})$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_{i2}(x_{i1} - \bar{x}_1) + \hat{\beta}_2 \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 = \sum_{i=1}^{n} x_{i2}(y_i - \bar{y})$$

Appendix: Algebra for $k = 2$

- Solving for $\hat{\beta}_2$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_{i2}(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n} x_{i2}(x_{i1} - \bar{x}_1)}{\sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2}$$

and substituting into the other equation

$$\hat{\beta}_1 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 + \frac{\sum_{i=1}^{n} x_{i2}(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n} x_{i2}(x_{i1} - \bar{x}_1)}{\sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2} \sum_{i=1}^{n} x_{i1}(x_{i2} - \bar{x}_2) = \sum_{i=1}^{n} x_{i1}(y_i - \bar{y})$$

Appendix: Algebra for $k = 2$

- Solving for $\hat{\beta}_1$

$$\hat{\beta}_1\left(\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2\right)\left(\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2\right)-\hat{\beta}_1\left[\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)\right]^2=$$

$$=\left(\sum_{i=1}^{n}x_{i1}(y_i-\bar{y})\right)\left(\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2\right)-\left(\sum_{i=1}^{n}x_{i2}(y_i-\bar{y})\right)\left(\sum_{i=1}^{n}x_{i1}(x_{i2}-\bar{x}_2)\right)$$

Which eventually lead us to

$$\hat{\beta}_1=\frac{\left(\sum_{i=1}^{n}x_{i1}(y_i-\bar{y})\right)\left(\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2\right)-\left(\sum_{i=1}^{n}x_{i2}(y_i-\bar{y})\right)\left(\sum_{i=1}^{n}x_{i1}(x_{i2}-\bar{x}_2)\right)}{\left(\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2\right)\left(\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2\right)-\left[\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)\right]^2}$$

Appendix: Algebra for $k = 2$

- This shows that for the general case, when $k \geq 2$, ordinary algebra is inadequate. In this case it is necessary to switch to matrix algebra (See Appendix E).
- Defining

$$r^2_{x_1,x_2} = \frac{\left(\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)\left(x_{i2}-\bar{x}_2\right)\right)^2}{\left(\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)^2\right)\left(\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_2\right)^2\right)}$$

we can write $\hat{\beta}_1$ as

Appendix: Algebra for $k = 2$

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y})\right)\left(\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2\right) - \left(\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(y_i-\bar{y})\right)\left(\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)\right)}{\left(\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2\right)\left(\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2\right)\left(1-r_{x_1,x_2}^2\right)}$$

$$= \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y})}{\left(\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2\right)\left(1-r_{x_1,x_2}^2\right)} - \frac{\left(\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(y_i-\bar{y})\right).r_{x_1,x_2}}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}\sqrt{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}\left(1-r_{x_1,x_2}^2\right)}$$

Appendix: Algebra for $k = 2$

1. If $r^2_{x_1,x_2} = 0$ then $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} = \tilde{\beta}_1$$

so the OLS slope estimates in the MLR of y on x_1 and x_2 and the SLR of y on x_1 are the same.

Remember that $r^2_{x_1,x_2}$ is a measure of multicolinearity in this model.

Appendix: Algebra for $k = 2$

2. If $\hat{\beta}_2 = 0$ then $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} = \tilde{\beta}_1$$

so, when the partial effect of x_2 on y is zero, $\hat{\beta}_2 = 0$, then the MLR of y on x_1 and x_2 and the SLR of y on x_1 are the same.

We encountered these two cases before.

Appendix: Algebra for $k = 2$

3. Moreover we can write

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y})}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2} - \hat{\beta}_2 \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$$

so letting $\tilde{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y})}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$,

the OLS slope estimate in the SLR of y on x_1,

Appendix: Algebra for $k = 2$

we see that $\hat{\beta}_1 = \tilde{\beta}_1 - \hat{\beta}_2 \tilde{\delta}_1$, where

$$\tilde{\delta}_1 = \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2}$$

is just the OLS slope coefficient from the SLR of x_2 on x_1.

Hence $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$, which shows the relation between the SLR and the MLR coefficient estimates and it is another way to study the omitted variable bias.

Appendix: $\mathrm{R}^2 = r^2_{y,\hat{y}}$

$$\left.\begin{aligned} \mathrm{R}^2 &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ r^2_{y,\hat{y}} &= \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right)} \end{aligned}\right\}$$

but since $\bar{y} = \bar{\hat{y}}$

$$r^2_{y,\hat{y}} = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})\right)^2}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right)}$$

Appendix: $R^2 = r^2_{y,\hat{y}}$

$$\boxed{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})} = \sum_{i=1}^{n} y_i(\hat{y}_i - \bar{y})$$

$$= \sum_{i=1}^{n}(\hat{y}_i + \hat{u}_i)(\hat{y}_i - \bar{y})$$

$$= \sum_{i=1}^{n}\hat{y}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y})$$

$$= \sum_{i=1}^{n}\hat{y}_i(\hat{y}_i - \bar{y}) + \underbrace{\sum_{i=1}^{n}\hat{u}_i\hat{y}_i}_{=0} - \bar{y}\underbrace{\sum_{i=1}^{n}\hat{u}_i}_{=0}$$

$$= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})(\hat{y}_i - \bar{y}) = \boxed{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}$$

Appendix: $\mathrm{R}^2 = r^2_{y,\hat{y}}$

$$\boxed{\mathrm{R}^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} =$$

$$= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \cdot \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} = \boxed{r^2_{y,\hat{y}}}$$

Nothing in this derivation depends on *k*.

Appendix: $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$

Consider the normal equation for x_1

$$\sum_{i=1}^{n} x_{i1} \underbrace{\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} \right)}_{\hat{u}_i} = 0$$

Regressing x_1 on x_2 we can write

$$x_1 = \hat{x}_1 + \hat{r}_1 = \hat{\gamma}_0 + \hat{\gamma}_2 x_2 + \hat{r}_1 \Rightarrow \begin{cases} \sum_{i=1}^{n} \hat{r}_{i1} = 0 \\ \sum_{i=1}^{n} x_{i2} \hat{r}_{i1} = 0 \end{cases} \Rightarrow \sum_{i=1}^{n} \hat{x}_{i1} \hat{r}_{i1} = 0$$

Appendix: $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$

1. $$\sum_{i=1}^{n} x_{i1}\hat{u}_i = \sum_{i=1}^{n} \left(\hat{x}_{i1} + \hat{r}_{i1}\right)\hat{u}_i = \underbrace{\sum_{i=1}^{n} \hat{x}_{i1}\hat{u}_i}_{=0} + \sum_{i=1}^{n} \hat{r}_{i1}\hat{u}_i = \sum_{i=1}^{n} \hat{r}_{i1}\hat{u}_i$$

since $$\sum_{i=1}^{n} \hat{x}_{i1}\hat{u}_i = \sum_{i=1}^{n} \left(\hat{\gamma}_0 + \hat{\gamma}_2 x_{i2}\right)\hat{u}_i = \hat{\gamma}_0 \underbrace{\sum_{i=1}^{n} \hat{u}_i}_{=0} + \hat{\gamma}_2 \underbrace{\sum_{i=1}^{n} x_{i2}\hat{u}_i}_{=0} = 0$$

by the algebraic properties of the OLS.

Appendix: $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$

2. $$\sum_{i=1}^{n} \hat{r}_{i1}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right) =$$

$$= \sum_{i=1}^{n} \hat{r}_{i1} y_i - \hat{\beta}_0 \underbrace{\sum_{i=1}^{n} \hat{r}_{i1}}_{=0} - \hat{\beta}_1 \sum_{i=1}^{n} \hat{r}_{i1} x_{i1} - \hat{\beta}_2 \underbrace{\sum_{i=1}^{n} \hat{r}_{i1} x_{i2}}_{=0}$$

$$= \sum_{i=1}^{n} \hat{r}_{i1} y_i - \hat{\beta}_1 \sum_{i=1}^{n} \hat{r}_{i1}\left(\hat{x}_{i1} + \hat{r}_{i1}\right)$$

$$= \sum_{i=1}^{n} \hat{r}_{i1} y_i - \hat{\beta}_1 \underbrace{\sum_{i=1}^{n} \hat{r}_{i1} \hat{x}_{i1}}_{=0} - \hat{\beta}_1 \sum_{i=1}^{n} \hat{r}_{i1}^2 = 0$$

Appendix: $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$

Solving $\sum_{i=1}^{n} \hat{r}_{i1} y_i - \hat{\beta}_1 \sum_{i=1}^{n} \hat{r}_{i1}^2 = 0$ we get the formula for $\hat{\beta}_1$.

Note that the argument can be generalized for general *k*. In this case $\hat{r}_1$ are the residuals from the regression of x_1 on $x_2, x_3, \ldots, x_k$.

Appendix: $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$

Of course we can proof this directly from the formula for $\hat{\beta}_1$ we got before.

From the theory of OLS we can write $\sum_{i=1}^{n} \hat{r}_{i1}^2$ as

$$\sum_{i=1}^{n} \hat{r}_{i1}^2 = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 \left(1 - r_{x_1,x_2}^2\right)$$

Given the Sum of Squares Decomposition and since the *R*-squared in the regression of x_1 on x_2 is just r_{x_1,x_2}^2 .

Appendix: $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$

Substituting this into the formula for $\hat{\beta}_1$ we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1}-\overline{x}_1)(y_i-\overline{y}) - \dfrac{\sum_{i=1}^{n}(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2)}{\sum_{i=1}^{n}(x_{i2}-\overline{x}_2)^2}\left(\sum_{i=1}^{n}(x_{i2}-\overline{x}_2)(y_i-\overline{y})\right)}{\sum_{i=1}^{n} r_{i1}^2}$$

$$= \frac{\sum_{i=1}^{n}(x_{i1}-\overline{x}_1)(y_i-\overline{y}) - \hat{\gamma}_2\left(\sum_{i=1}^{n}(x_{i2}-\overline{x}_2)(y_i-\overline{y})\right)}{\sum_{i=1}^{n} r_{i1}^2}$$

Appendix: $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$

where $\hat{\gamma}_2 = \dfrac{\sum_{i=1}^{n} (x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2)}{\sum_{i=1}^{n} (x_{i2} - \overline{x}_2)^2}$, but

$$
\begin{aligned}
\sum_{i=1}^{n} (x_{i1} - \overline{x}_1)(y_i - \overline{y}) - \hat{\gamma}_2 \left(\sum_{i=1}^{n} (x_{i2} - \overline{x}_2)(y_i - \overline{y}) \right) &= \sum_{i=1}^{n} \left((x_{i1} - \overline{x}_1) - \hat{\gamma}_2 (x_{i2} - \overline{x}_2) \right)(y_i - \overline{y}) \\
&= \sum_{i=1}^{n} \left((x_{i1} - \overline{x}_1) - \hat{\gamma}_2 (x_{i2} - \overline{x}_2) \right) y_i \\
&= \sum_{i=1}^{n} \left(x_{i1} - (\overline{x}_1 - \hat{\gamma}_2 \overline{x}_2) - \hat{\gamma}_2 x_{i2} \right) y_i \\
&= \sum_{i=1}^{n} \left(x_{i1} - \hat{\gamma}_0 - \hat{\gamma}_2 x_{i2} \right) y_i
\end{aligned}
$$

Appendix: $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$

where $\hat{\gamma}_0 = \bar{x}_1 - \hat{\gamma}_2 \bar{x}_2$, hence

$$\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(y_i - \bar{y}) - \hat{\gamma}_2 \left(\sum_{i=1}^{n} (x_{i2} - \bar{x}_2)(y_i - \bar{y}) \right) = \sum_{i=1}^{n} \underbrace{(x_{i1} - \hat{\gamma}_0 - \hat{\gamma}_2 x_{i2})}_{\hat{r}_{i1}} y_i$$

$$= \sum_{i=1}^{n} \hat{r}_{i1} y_i$$

So eventually we get $\boxed{\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}}$