Multiple Linear Regression Model

2025/2026, Semester 1

Ralf A. Wilke
Copenhagen Business School

OLS properties

Textbook:

Wooldridge (2025), Chapter 4.

Motivation for Multiple Regression

- Multiple regression analysis allows us to explicitly control for many factors that simultaneously affect the dependent variable.
- For this reason we can hope to infer causality in cases where simple regression analysis would be misleading.
- If we add more variables to explain y, then more of the variation of y can be explained. Thus, we can better predict the dependent variable.
- The model incorporates fairly general functional form relationships.

The Model with *k* Independent Variables

- The general multiple regression model is

$$\begin{aligned} y &= \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + u \\ &= \mathbf{x}\beta + u \end{aligned}$$

- x_1 is typically 1.
- This equation contains *K* unknown population parameters.
- The key assumption for the general multiple regression model is: $E(u|\mathbf{x}) = 0$.
- We still require that *u* is uncorrelated with all independent variables $x_1, \ldots, x_K$.

Estimation by OLS

- We seek estimates $\hat{\beta}_1, \ldots, \hat{\beta}_K$ in the equation

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_K x_K$$

- The OLS estimates are chosen to minimise the squared residuals:

$$\sum_{i=1}^{N} (y_i - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_K x_{iK})^2$$

  with observations *i=1,…,N*.
- Derive the *K* first order conditions which are linear and solve for the *K* unknowns.

Derivation of OLS estimates

- Using matrix notation.
- For

$$\begin{aligned} y_i &= \beta_1 x_{i,1} + \ldots + \beta_K x_{i,K} + u_i \\ &= \mathbf{x}_i \beta + u_i \end{aligned}$$

with

$$\mathbf{x}_i' = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,K} \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$$

- Define:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \text{ and } \quad \mathbf{X}_{N \times K} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K} \\ \vdots & & & \\ x_{N1} & x_{N2} & \dots & x_{NK} \end{pmatrix}.$$

$$\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix}$$

- This then altogether yields:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

$\mathbf{X}\boldsymbol{\beta}$ is *(N* x *1)* because $\boldsymbol{\beta}$ is *K* x *1* and $\mathbf{X}$ is *N* x *K.*

- The *Kx1* vector of OLS estimates, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K)'$, minimises *SSR(**b**)* over all possible vectors **b** and is:

$$\boxed{\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$$

- Define the fitted values and residuals as:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

OLS regression line
Sample regression function

$$\hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

- Then because of $\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{0}$, we have

$$\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$$

  - Because the first column of ***X*** consists of ones, the OLS residuals always sum to zero.
  - The covariance between each independent variable and the OLS residuals is zero

Example: Determinants of College GPA

- Data set: GPA1.dta
- Contains: *colGPA*= college grade point average, *hsGPA*= high school GPA, *ACT*= achievement test score.
- *n*=141 students from a large university
- Both GPAs are on a four point scheme
- We obtain the following OLS regression to predict college GPA:

$$\widehat{colGPA} = 1.29 + 0.453 hsGPA + 0.0094 ACT$$

- How do we interpret this?
- The predicted college GPA if *hsGPA* and *ACT* are zero, is 1.29.
  - Since *hsGPA*=0 or *ACT*=0 does not exist in the sample, the intercept is not meaningful.

- More interesting are the slope coefficients.
  - Positive relationship between *colGPA* and *hsGPA*, holding *ACT* fixed. (a point increase in *hsGPA* predicts an increase of *colGPA by 0.453*)
  - Positive relation between *ACT* and *colGPA*, holding *hsGPA* fixed. The estimated effect is however, very small (an increase of 1000 in *ACT* predicts an increase of *colGPA* by less than one point). Note, that the sample average of *ACT* is about 24.
- What happens if we ignore *hsGPA* in the regression?

$$\widehat{colGPA} = 2.40 + 0.0271ACT$$

thus, the coefficient on *ACT* is now almost three time larger, suggesting a stronger relationship.

  - In this model, we cannot, however, compare two people with the same high school GPA.
  - We will later formally discuss the implication if variables are omitted.

Comparison of Simple and Multiple Regression estimates

- Suppose *k*=2.
- In which cases will a simple regression of *y* on x_1 produce the same results as the regression of *y* on x_1 and x_2?
- Define:

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ \tilde{y} &= \tilde{\beta}_0 + \tilde{\beta}_1 x_1 \\ \tilde{x}_2 &= \tilde{\delta}_0 + \tilde{\delta}_1 x_1 \end{aligned}$$

  One can show that: $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$
- There is equality in two cases:
  - $\hat{\beta}_2 = 0$.
  - $\tilde{\delta}_1 = 0$. Why?

- Example: Determinants of College GPA

$$\widehat{colGPA} = 1.29 + 0.453 hsGPA + 0.0094 ACT$$
$$col\tilde{G}PA = \tilde{\beta}_0 + 0.482 hsGPA$$

  - The correlation between *hsgpa* and *ACT* is about 0.346 but the coefficient $\hat{\beta}_2$is very little.
  - For this reason the two slope estimates for *hsGPA* are quite similar.

- This reasoning can be extended to *k*-independent regressors. Estimates for β_1 are just identical if:
  - $\hat{\beta}_j = 0 \text{ for } j = 2, \ldots, k$
  - if x_1 is uncorrelated with each of $x_2, \ldots, x_k$.

Goodness-of-Fit

- Same as in the simple regression model, because definitions only depend upon y_i, $\hat{y}_i$ and $\hat{u}_i$:
  - Total sum of squares (SST): $\mathrm{SST}=\sum_i (y_i - \bar{y})^2$
  - Explained sum of squares (SSE): $\mathrm{SSE}=\sum_i (\hat{y}_i - \bar{y})^2$
  - Residual sum of squares (SSR):

$$\mathrm{SSR}=\sum_i \hat{u}_i^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

    with SST=SSE+SSR.

  - $R^2=\mathrm{SSE}/\mathrm{SST}=1\text{-}\mathrm{SSR}/\mathrm{SST}$
    which is between *0* and *1*.

The expected value of the OLS estimators

- We state and discuss four assumptions, which are direct extensions of the simple regression model assumptions, under which the OLS estimators are unbiased for the population parameters.

Assumption 1: Linear in Parameters

The model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} ,$$

where ***y*** is an observed (*n x 1*) vector, ***X*** is an *n x k* observed matrix, and ***u*** is an (*n x 1)* vector of unobserved errors or disturbances.

- This is the population or true model.

Assumption 2: Random Sampling

We have a random sample of size *n*, $\{(x_{i1}, x_{i2}, \ldots, x_{iK}, y_i)_{i=1,\ldots,n}\}$, following the population model in Assumption 1.

- This assumption implies that the selection into the sample is random. In particular that it is not related to the error term *u*.

- OPTIONAL MATERIAL:

  Hirschauer et al. (2021), Inference using non-random samples? Stop right there! Significance.

Assumption 3: No Perfect Collinearity

The matrix ***X*** has rank(*K*).

- Under this assumption, ***X'X*** is nonsingular and we can write $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
- This assumption does not preclude some correlation between the independent variables, it rules out perfect correlation.

Assumption 4: Population orthogonality condition

$$E(\mathbf{x}'u) = 0$$

- This assumption rules out that there are independent variables which are correlated with u.
- It is implied by the zero conditional mean assumption:

$$E(u|\mathbf{x}) = 0 \qquad \text{(Assumption 4')}$$

- If ***x*** contains a constant, *u* has zero mean.
- The independent variables are said to be exogenous.

Theorem 1.1 (Unbiasedness of OLS)

Using Assumptions 1 through 4', the OLS estimator $\hat{\beta}$ is unbiased for β .

- <u>Proof</u>:
  - First rewrite $\hat{\beta}$ as a function of β:

$$\boxed{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = I_K}$$

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\end{aligned}$$

  - Then take the expectation conditional on ***X***

$$\begin{aligned}E(\hat{\beta}|\mathbf{X}) &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{0} \\ &= \beta\end{aligned}$$

Assumption 4'

The Variance of the OLS Estimators

- In order to derive the simplest form of the variance-covariance matrix of $\hat{\beta}$, we make an additional assumption.

Assumption 5 (Homoscedasticity)

(i) $\mathrm{Var}(u_i|\mathbf{X}) = \sigma^2$ for *i=1,…,n*.

(ii) $\mathrm{Cov}(u_i, u_j|\mathbf{X}) = 0$ for all $i \neq j$.

In matrix form this is

$$\mathrm{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2 I_N \; ,$$

where I_N is the (*N x N*) identity matrix.

- Part i) says that the variance of *u* cannot depend on any element of ***X***.
- Part ii) says that the errors cannot be correlated across observations. It is implied by Assumption 2.

Theorem 1.2 (Variance Covariance Matrix of the OLS Estimator)

Under Assumptions 1 through 5,

$$\text{var}(\hat{\boldsymbol\beta}\,|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

- Proof: First, remark that $\hat{\boldsymbol\beta} = \boldsymbol\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$.

  Then $\text{var}(\hat{\boldsymbol\beta}) = \text{var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u})$.

  and conditional on ***X*** :

$$\begin{aligned} var(\hat\beta|\mathbf{X}) &= var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[var(\mathbf{u}|\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

$\boxed{var(\mathbf{A}'\mathbf{X}) = \mathbf{A}'var(\mathbf{X})\mathbf{A}}$

Assumption 5

$\boxed{\mathbf{X}'I_n = \mathbf{X}'}$

$\boxed{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = I_K}$

$$\mathrm{Var}(\hat{\beta}_j|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{jj}$$

- the conditional variance of $\hat{\beta}_j$ is obtained by multiplying σ^2 by the j'th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.
- One can show that:

$$SST_j = \sum_i (x_{ij} - \bar{x}_j)^2$$
$$R_j^2\text{: R-squared of } x_j \text{ on other x}$$

$$\mathrm{Var}(\hat{\beta}_j|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{jj} = \frac{\sigma^2}{SST_j(1-R_j^2)}$$

  - The conditional variance is large if
    - σ^2 is large.
    - the sample variation in x_j is small (SST_j).
    - there is a high correlation between x_j and any other $x_l,\ l \neq j : R_j^2$ is large. *"Multicollinearity"*
- The conditional variance decreases with *N.*

Multicollinearity

- If there is high (but not perfect) correlation between two or more variables.
- This implies that the R^2 of a regression of x_j on all other independent variables is large R_j^2.
- In this case $\mathrm{var}(\hat{\beta}_j|\mathbf{X})$ is large. It explodes as R_j^2 goes to one, i.e. $R_j^2 \to 1 \Longrightarrow \mathrm{var}(\hat{\beta}_j|\mathbf{X}) \to \infty$
- Note: if x_j is uncorrelated with all other $x_l,\ l \neq j$: $R_j^2 = 0$.
- Why is $R_j^2 \neq 1$?

- In an application it is better to have less correlation between the regressors.

- As $\mathrm{var}(\hat{\beta}_j|\mathbf{X})$ explodes when $R_j^2 \rightarrow 1$ it could be useful to define an upper ”acceptable” level of multicollinearity.
- This is sometimes by considering the so-called variance inflation factor (VIF):

$$\mathrm{Var}(\hat{\beta}_j|\mathbf{X}) = \frac{\sigma^2}{SST_j(1-R_j^2)} = \frac{\sigma^2}{SST_j} VIF_j$$

with

$$VIF_j = 1/(1-R_j^2).$$

  - Evidently, $VIF_j = 1$ whenever there is no correlation between x_j and the other regressors and VIF_j explodes as $R_j^2 \rightarrow 1$.
  - $VIF_j > 10$ indicates a high a degree of multicollinearity and can be used to explain large variance of OLS estimates.
    - This does not mean, however, that omitting some of the variables will “improve” estimates as this normally leads to omitted variable bias. Trade-off! (Code: multiple_reg_vif.R)

- Alternatively, statistical regularisation methods can be applied that automatically select the relevant regressors.
  - Useful techniques if the regressor set is large.
  - For consistency, all variables of the population model are observed plus some that may not belong to the model.
  - Superior than sequential elimination of variables (Oracle property).
  - Corresponds to an OLS regression under additional inequality constraints on the parameters (Penalised regression).

<u>Example</u>:
What makes employees change their employer?

- To some extent there is some theory suggesting important variables:
  - Pay (absolute/relative)
  - Performance
  - Contract duration

- What else?
  - Possibly a vast number of variables and sources.

- Classical variable selection approaches based on VIF and sequential elimination are not optimal.

- Human resources analytic software produces scores or fitted probabilities for the probability of leaving.
  - Typically based on fitting methods such as Neural Networks. Limited interpretability of results (black box).

- What are the relevant factors and how to catch them?

We wish to have an oracle that tells us what to include.

Each data source may comprise of a large number of variables.

Variable selection techniques

- High-dimensional data cause problems for estimation when
  - K>N (more candidate variables than observations).
  - High degree of multicollinearity.
- Drawbacks of sequential elimination methods (subset selection based on likelihood ratio test, stepwise selection based on AIC/BIC,etc.). Code: stepwise_aic.R
  - Only a small number of variables can be tested.
  - Results are highly affected by sequence of the test.
  - Overfitting.
- Desired statistical property
  - Oracle property: Only the relevant variables are selected, and the estimates of those variables are asymptotically equal to the estimates from a model that only includes the relevant variables.

Penalised Regression

- A penalty is added to the objective function that penalises the use of too many variables.
- Objective function: $min_\beta \, L(\beta|X,y)$
- Add penalty: $P_\lambda(\beta)$, where λ is a penalisation/tuning parameter.
- Penalised regression: $min_\beta \, L(\beta|X,y) + P_\lambda(\beta)$
- Also called:
  - Shrinkage methods
  - Statistical regularisation
  - Unsupervised learning

e.g.
OLS: $L(\beta|X,y)=$
$\sum(y_i - X_i\beta)^2$

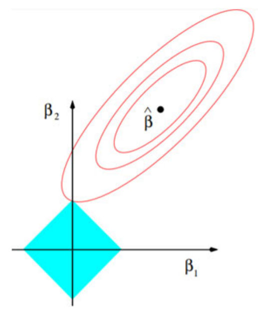
Shrinkage methods

Various methods. Differ in the choice of penalty.

- Individual variable selection
  - Lasso (ℓ_1-type penalty)

$$P_\lambda(\beta) = \lambda \sum_{k=1}^{K} |\beta_k|$$

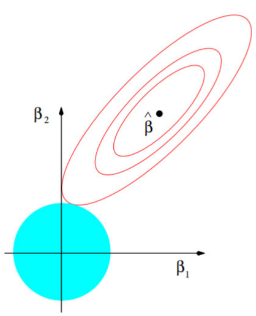
$$min\, L(\beta|X,y)\; s.t. \sum_{k=1}^{K} |\beta_k| < t$$



  - Ridge (ℓ_2-type penalty)

$$P_\lambda(\beta) = \lambda \sum_{k=1}^{K} \beta_k^2$$

$$min\, L(\beta|X,y)\; s.t. \sum_{k=1}^{K} \beta_k^2 < t$$



Source: Hastie et al. (2009)
K=2, t is tuning parameter.

- Group variable selection
  - Group Lasso

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{J} \sqrt{A_j} \sqrt{\sum_{k=1}^{A_j} \beta_{jk}^2}$$

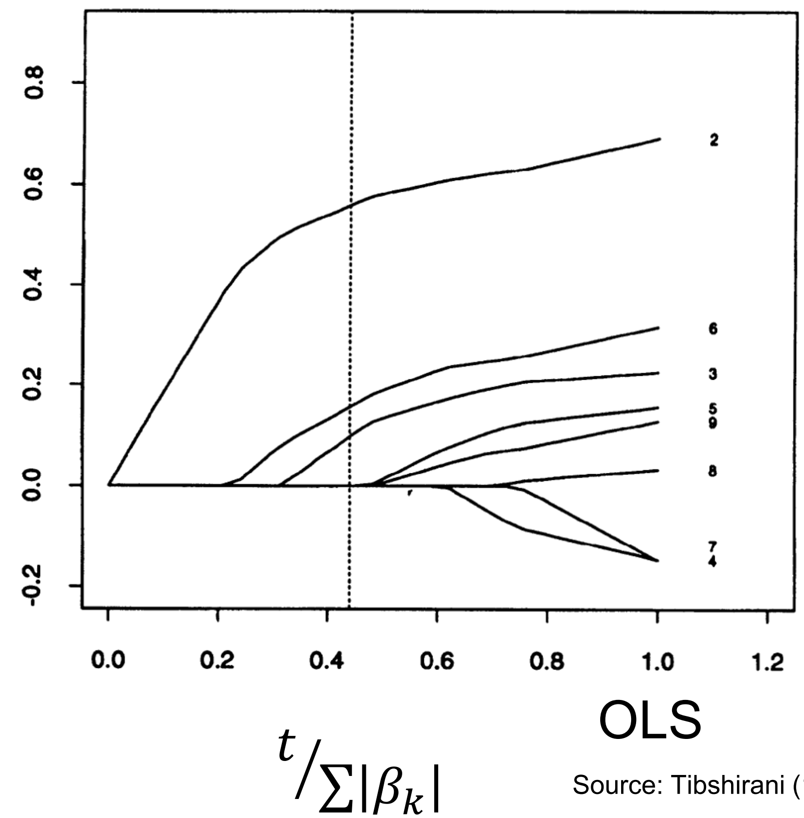
- Bi-level variable selection
  - (Adaptive) group bridge

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{J} \sqrt{A_j} \sqrt{\sum_{k=1}^{A_j} w_{jk} |\beta_{jk}|}$$

A_j: number of variables in a group
w_{jk}: weights

The tuning parameter for LASSO

- The number of selected variables increases with *t* (penalty becomes smaller=λ decreases).
- The choice of the tuning parameter is important:
  - Various methods (optional): AIC (some models), BIC, Cross validation (CV), Generalised Cross Validation (GCV)
- Example code: shrinkage.R



Source: Tibshirani (1996)

Penalty terms (literature optional)

- Penalty terms incorporate different beliefs on the structure and magnitude of the variables and result in different models
  - Individual variable selection: Lasso (Tibshirani, 1996), Elasticity net (Zou and Hastie, 2003), Adaptive Lasso (Zou, 2006), Fused Lasso (Tibshirani, 2005)
  - Group-level variable selection: Group Lasso (Yuan and Lin, 2006), Hierarchical Lasso (Zhao et al., 2009)
  - Bi-level variable selection: Group bridge (Huang et al., 2009), Sparse group lasso (Simon et al., 2013)
- Simultaneous variable selection and inference is challenging. Still a developing field.
  - Sample splitting (Meinshausen et al., 2009), covariance test (Lockhart et al., 2014), exact post-selection inference (Lee et al., 2016), OLS post-Lasso (Belloni and Chernozhukov, 2013), etc.

Selected References (optional)

- Journal references
  - Huang, J., Breheny, P., & Ma, S. (2012). A Selective Review of Group Selection in High-Dimensional Models
  - Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference
- Book references
  - Bühlmann, P., & Van De Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications
  - Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity: the lasso and generalizations.

Variances in misspecified models

- How do the variances change if we omit a variable?
- Remember:
$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ \tilde{y} &= \tilde{\beta}_0 + \tilde{\beta}_1 x_1\end{aligned}$$

- We know: $\mathrm{Var}(\hat{\beta}_1|\mathbf{X}) = \sigma^2/SST_1(1-R_1^2)$
- It can be shown: $\mathrm{Var}(\tilde{\beta}_1|\mathbf{X}) = \sigma^2/SST_1$
- This implies that $\mathrm{Var}(\tilde{\beta}_1|\mathbf{X}) \leq \mathrm{Var}(\hat{\beta}_1|\mathbf{X})$
- They are equal if all independent variables are uncorrelated, otherwise not.

- If $\beta_2 = 0$, do not include x_2 in the regression, because variance of the OLS estimator may increase, while both estimators are unbiased.

- The case $\beta_2 \neq 0$ is more difficult:
  - There is a trade-off between bias and variance
  - For large samples we may, however, prefer to include in $\hat{\beta}$, because the variance becomes less important, while the bias does not depend on the sample size.

Estimating σ^2

- The sampling variance of $\hat{\beta}_j$ depends on σ^2.
- Since $E(u_i^2) = \sigma^2$, it would be natural to estimate σ^2 by $N^{-1}\sum_i u_i^2$ this is, however, not possible because u_i is unknown.
- Instead, we use the estimated u_i, which are unbiased.

Theorem 1.3 (Unbiased Estimation of σ^2)

Under Assumptions 1 through 5,

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

With

$$\begin{aligned} \hat{\sigma}^2 &= (N-K)^{-1}\sum_i \hat{u}_i^2 = (N-K)^{-1}\hat{\mathbf{u}}'\hat{\mathbf{u}} \\ &= SSR/(N-K) \end{aligned}$$

.

- The denominator is *(N-K)* and not *N* because the residuals have to satisfy the *K* conditions:

$$\textstyle\sum_i x_{il}\hat{u}_i = 0 \text{ for } l = 1, \ldots, K$$

- For this reason we have only (*N-K)* degrees of freedom.
- In contrast to the SSR, $\hat{\sigma}$ can increase or decrease when another variable is added, since degrees of freedom decrease.

- $\hat{\sigma}$ is called the standard error of the regression (**SER**) and typically reported by econometric packages.

- The standard error of $\hat{\beta}_j$ is therefore estimated by

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{[SST_j(1-R_j^2)]^{1/2}} = \hat{\sigma}[(\mathbf{X}'\mathbf{X})^{-1}_{jj}]^{1/2}$$

- Since it relies on the homoscedasticity assumption, it is not a valid estimator if Assumption 5 does not hold.

Efficiency of OLS

- It can be shown that under the above assumptions that OLS has another nice property.

Theorem 1.4 (Gauss-Markov Theorem)

Under Assumptions 1 through 5, $\hat{\beta}$ is the best linear unbiased estimator (BLUE).

- This means that for any estimator $\tilde{\beta}_j$ that is linear and unbiased, $\mathrm{var}\hat{\beta}_j \leq \mathrm{var}(\tilde{\beta}_j)$, i.e. OLS has the smallest variance among all unbiased linear estimators.
- For this reason, we don't need to look for a better estimator under Assumptions 1 – 5.
- Assumptions 1 - 5 are known as Gauss-Markov assumptions.

- Finite sample or exact properties of the OLS estimators:
  - Unbiasedness holds for any sample size *N* if the four Assumptions 1-4’ hold.
  - Also, the fact that OLS is the best linear unbiased estimator under Assumptions 1-5 is a finite sample property

- It is also important to know the large sample or asymptotic properties. This is if sample size grows without bound ($N \to \infty$).
  - OLS estimators have nice asymptotic properties (consistent and asymptotically normal distributed).

Theorem 1.5 (Consistency of OLS)

Under Assumptions 1 through 4, the OLS estimator $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

Theorem 1.6 (Asymptotic Normality of OLS)

Under the Gauss-Markov Assumptions 1 through 5,

(i) $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{a}{\sim} N(\mathbf{0}, \sigma^2 \mathbf{A}^{-1})$ with $\mathbf{A} = E(\mathbf{x}_i' \mathbf{x}_i)$

(ii) $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2 = \mathrm{var}(u)$.

(iii) for each j, $(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) \overset{a}{\sim} N(0, 1)$

- From now we mainly focus on asymptotic properties.

- Under the Gauss-Markov assumptions, the OLS estimators are best linear unbiased.
  - OLS is also asymptotically efficient among a certain class of estimators under the Gauss-Markov assumptions.
  - A wide class of estimators are unbiased for β but OLS has the smallest asymptotic variance in this class.
- Suppose *g(x)* is any function of x such that *g(x)* and *u* are uncorrelated. Let $\tilde{\beta}$ be the solution to the *K* conditions:

$$\sum_i g_j(\mathbf{x}_i)(y_i - \tilde{\beta}_1 x_{i1} - \ldots - \tilde{\beta}_K x_{iK}) = 0 \text{ for } j = 0, 1, \ldots, K$$

Theorem 1.7 (Asymptotic Efficiency of OLS)

Under the Gauss-Markov Assumptions 1 through 5, let $\tilde{\beta}_j$ denote estimators that solve the above equations and let $\hat{\beta}_j$ denote the OLS estimators. Then for j=1,2,…,K, the OLS estimators have the smallest asymptotic variances: $\text{avar}\ \sqrt{n}(\hat{\beta}_j - \beta_j) \leq \text{avar}\ \sqrt{n}(\tilde{\beta}_j - \beta_j)$

Large Sample Test: Lagrange Multiplier Statistic

- For most purposes there is little reason to go beyond the usual t and F statistics.

- There are, however, other ways to test multiple exclusion restrictions.

- The Lagrange Multiplier (**LM**) Statistics has achieved some popularity in modern econometrics.

- It does not require estimation of the unrestricted model.
- It does not require the normality of errors.

- A guide to perform the LM Test for q exclusion restrictions:
  1. Regress *y* on the restricted set of independent variables and save the residuals $\tilde{u}$.
  2. Regress $\tilde{u}$ on all independent variables and obtain the R-squared, say R_u^2.
  3. Compute $LM = NR_u^2$.
  4. Compare *LM* to the appropriate critical value, *c*, in a χ_q^2 distribution; if *LM*>*c*, the null hypothesis is rejected. (similarly, one can also compute the p-value and reject if it is too low.) Otherwise, H_0 cannot be rejected.

- Often results are similar compared to the F-test.
- The *F* statistic is usually automatically computed by econometric packages.

Summary

- In the multiple regression model:
  - We can hold several factors fixed while looking at partial effects.
  - Independent variables can be correlated.
  - Selection of independent variables important.
  - Captures a variety of nonlinear relationships between x_j and *y.*
  - OLS is easy to calculate and has nice properties under five assumptions: unbiasedness & efficiency
  - Sample and asymptotic distribution of the OLS estimator, which can be used for inference.