# Multiple Linear Regression & Gauss-Markov Theorem II


Evgeny Sedashov, PhD
esedashov@hse.ru


10/02/2024

## Follow-Up to the Last Class

- During the last class, I gave you the formula for $R^2$:

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

  which has an intuitive interpretation: if model fits data well, then $\hat{y}_i$ will be close to $y_i$, and $R^2$ will be close to 1.

- In this class, we'll use alternative formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = 1 - \frac{SSR}{SST}$$

## Introduction I

- In two previous classes we covered the OLS estimation and saw how Gauss-Markov Assumptions lead to OLS being the BLUE estimator.

- We also covered basic interpretations of OLS parameters, namely slopes and the intercept.

- Today, our main focus will be on **statistical inferences** you can make from OLS regression which is, in many ways, the most important part of regression analysis for us.

## Introduction II

- When we talk about uncertainty pertaining to OLS estimators, we usually mean uncertainty that comes from the *sampling distribution*.
- Suppose your dependent variable is person's income and your independent variable is person's level of education.
- You keep person's level of education fixed and then sample repeatedly values of dependent variable for these fixed values of education.
- If you run bivariate regression for each sample, you'll get lines with slightly varying slopes.

## Normality Assumption I

- So far, we've only made two assumptions about the error terms:

$$\mathbb{E}[u|x_1...x_n] = 0$$

  and

$$\mathbb{V}[u|x_1...x_n] = \sigma^2$$

- To make statistical inferences from OLS, we need to know more about the distribution of error terms: we, therefore, assume that the error terms are independent from $x_1...x_n$ independent variables, and are normally distributed with mean 0 and constant variance $\sigma^2$.
- Normality assumption is much stronger than Gauss-Markov assumptions pertaining to the error terms, but without it we do not really have much to say about the sampling distribution of OLS parameters.

## Normality Assumption II

- You might be wondering: what makes normality assumption reasonable?
- The usual argument goes about like this: $u$ can be decomposed as a sum of many different independent and identically distributed unobserved factors, therefore we can invoke the Central Limit Theorem to claim that the limiting distribution of $u$ is approximately normal.
- If components of $u$ do not have identical distributions, a version of the Central Limit Theorem called Lyapunov CLT still applies, but approximation might not be as "nice" as conventional CLT.
- Much more problematic is the possibility that $u$ is some complicated function of unobserved parameters (i.e., multiplicative instead of additive); in that case, nothing really assures normality.

## Normality Assumption III

- In the end, normality of $u$ is an empirical question the answer to which is dictated by our prior knowledge about the dependent variable.
- In some cases, normality assumption clearly fails: for instance, if your dependent variable is something like protest counts or binary indicator of turning out to vote, normality assumption is clearly wrong.
- Log transformation sometimes provides a way to ensure normality in situations when the actual dependent variable does not satisfy the requirement.
- For instance, income distribution is often left-skewed, but log transformation ensures normality.

## Normality Assumption IV

- Under Normality Assumption, the sampling distribution of $\hat{\mathbf{b}}$ is multivariate normal with the vector of means $\mathbf{b}$ and variance-covariance matrix given as $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.
- Each entry on the main diagonal of variance-covariance matrix gives the variance of individual OLS parameter, therefore implying that $\hat{\beta}_j \sim \mathcal{N}(\beta_j, a_{jj})$ where $a_{jj}$ is the $j$'s entry on the main diagonal of the variance-covariance matrix.
- To get the distribution of $(\hat{\beta}_j - \beta_j)/sd(\hat{\beta}_j)$, we can apply standard properties of expectations and variances and get $\mathcal{N}(0, 1)$, a familiar standard normal distribution.

Distribution of Test Statistic

- Since we've established that $(\hat{\beta}_j - \beta_j)/sd(\hat{\beta}_j) \sim \mathcal{N}(0,1)$, some of you might wonder: if we assume that $\beta_j = \mu_0$, can't we test the null hypothesis the same way we did couple of weeks ago?
- Not quite: computation $sd(\hat{\beta}_j)$ involves $\sigma^2$, unknown population quantity.
- Replacing $\sigma^2$ to $\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{u_i}^2}{K-n-1}$ where $N$ is the number of observations and $k$ is the number of independent variables in the regression model gives $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$ as variance-covariance matrix, and square roots of the main diagonal entries are now standard errors rather than standard deviations.
- We will now prove that $(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) \sim t_{K-n-1} = t_{df}$ which corresponds to the $t$-distribution with $df = K - n - 1$ degrees of freedom.

## Distribution of OLS Estimates

### Theorem II

Suppose G-M Assumptions I-V are all true. Furthermore, suppose conditional distributions of error terms are all normal. Define standard error se of $\hat{\beta}_j$ as $\sqrt{\hat{\sigma}^2 c_{jj}}$ where $c_{jj}$ is the $j$th entry on the main diagonal of a matrix $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$. Then $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$ follows $t$-distribution with $K - n - 1$ degrees of freedom, with $K$ being the number of observations, $n$ being the number of independent variables.

Proof of Theorem II

- **Lemma I**: if $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$ and matrix $\mathbf{A}$ is symmetric and idempotent, then the scalar-valued random variable $\mathbf{x}^T\mathbf{A}\mathbf{x}$ follows $\chi^2$ distribution with $df = rank(\mathbf{A})$.

- **Lemma II**: if $x \sim \mathcal{N}(0, 1)$ and $z \sim \chi^2$ with $k$ degrees of freedom, then $x/\sqrt{z/k} \sim t_k$ ($t$-distribution with $k$ degrees of freedom), provided $x$ and $z$ are independent.

- Consider the matrix $\mathbf{I}_K - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

- This matrix is symmetric and idempotent (why?)

Proof of Theorem II

- As I have shown you, unbiased estimator for $\sigma^2$ is $\hat{\sigma}^2 = \hat{\mathbf{u}}^T \hat{\mathbf{u}} / K - n - 1$.

- However, $\hat{\mathbf{u}}^T \hat{\mathbf{u}}$ can be simplified even further to obtain $\mathbf{u}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{u}$ (also was proved before).

- Define $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

- Vector $\mathbf{u}^T \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ due to G-M assumptions II, IV and normality assumption, hence $\mathbf{u}^T/\sigma \sim \mathcal{N}(0, \mathbf{I})$

Proof of Theorem II

- Let's now consider

$$V = \frac{(K-n-1)\hat{\sigma^2}}{\sigma^2} = \frac{\hat{\mathbf{u}^T}\hat{\mathbf{u}}}{\sigma^2} = (\mathbf{u}^T/\sigma)\mathbf{M}(\mathbf{u}/\sigma)$$

- $V$ follows $\chi^2$ distribution with $df = rank(\mathbf{M})$ by Lemma I.

- Rank of an idempotent matrix is its trace, and we know that trace of **M** equals to $K-n-1$.

- So, $V$ follows $\chi^2$ distribution with $K-n-1$ degrees of freedom.

Proof of Theorem II

- Now consider $z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 c_{jj}}} \sim \mathcal{N}(0,1)$.

- As a final step, define

$$t_j = \frac{z_j}{\sqrt{V/(K-n-1)}} \sim t_{K-n-1}$$

by Lemma II.

Proof of Theorem II

- The remaining part is fairly simple:

$$t_j = \frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 c_{jj}}}}{\sqrt{\left(\frac{(K-n-1)\hat{\sigma^2}}{\sigma^2}\right)/(K-n-1)}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma^2} c_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$$

- Having this result under our belt, it is easy to consider any $t_j$ corresponding to different $\hat{\beta}_j$, giving us the fundamental statement behind statistical inference in the context of OLS:

  *Under the null hypothesis $H_0$: $\beta_j = \mu_j$, test statistic $t_j$ follows t-distribution with $K-n-1$ degrees of freedom.*

## t-test I

- With the knowledge of the distribution of test statistic under our belt, we can perform the null hypothesis testing the same way we did three weeks ago.
- First, we posit that $H_0 : \beta_j = \mu_0$ as a null hypothesis; the common standard is to assume $\mu_0 = 0$ and, hence, test the hypothesis that the independent variable $X_j$ has an effect on the dependent variable $Y$.
- If we test the alternative hypothesis $H_1 : \beta_j > \mu_0$ (or $\beta_j < \mu_0$), then we deal with one-sided (or one-tailed) null hypothesis testing.
- If we test the alternative hypothesis $H_1 : \beta_j \neq \mu_0$, then we deal with two-sided (or two-tailed) null hypothesis testing.

## t-test II

- First step in the null hypothesis testing is to determine the critical rejection level; in practical applications, you will often see stars like this *, **, *** at the regression tables and short note $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.
- What do these numbers-stars actually mean?
- Once we determined $\mu_0$, the actual value for null hypothesis testing, we can compute our test statistic

$$t = (\hat{\beta}_j - \mu_0)/se(\hat{\beta}_j)$$

  which we know follows $t$ distribution with $K - n - 1$ degrees of freedom.
- Contingent on what type (one-sided vs. two-sided) of testing we are conducting, interpretations will differ a little bit.

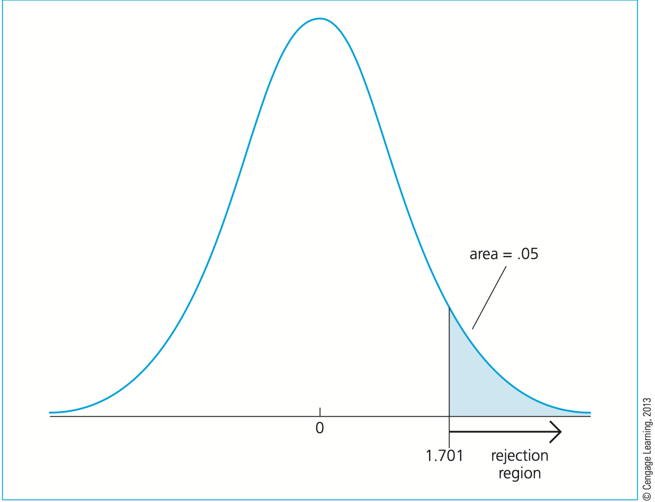
## t-test III

- Suppose we are testing one-sided alternative hypothesis $\beta_j > 0$; once we computed $t$, the test statistic, we can use the properties of the corresponding $t$ distribution to determine the actual $p$-value which is defined as

$$p = \mathbb{P}[T > t]$$

- Intuitively, $p$-value here is the area under the PDF of the $t$ distribution with left bound cut at the computed value of $t$.
- The meaning of the stars now should be clear: if the actual computed $p$-value is lower than a certain rejection level, then we say that the effect is significant at this level; for instance, if our actual computed $p$-value is lower than 0.01 but higher than 0.001 (say, 0.005), then effect will be significant at 0.01 level but not at 0.001 level.

**FIGURE 4.2** 5% rejection rule for the alternative $H_1: \beta_j > 0$ with 28 *df*.



## t-test IV

- For two-sided alternative hypothesis $\beta_j \neq 0$; the logic is essentially the same, but what is actually reported is

$$p = \mathbb{P}[|T| > |t|]$$

- Intuitively, $p$-value here is the sum of two areas under the PDF of the $t$ distribution: one with left bound cut at the value of $|t|$, and the second with right bound cut at the value of $-|t|$.
- The usefulness of $p$-values goes well beyond their relationship to critical rejection levels as it allows to answer the question: given the computed value of a $t$ test statistic, what is the lowest critical level that allows the rejection of the null hypothesis?

## More on Critical Rejection Levels

- Critical rejection approach, strictly speaking, does not require exact $p$-values for implementation, as one can simply look at $t$ table and check whether computed value of $t$ statistic is greater than the one reported as critical for a given rejection level and degrees of freedom.

- Popularity of stars reporting approach is mainly due to the easiness of comparison.

General Algorithm for Null Hypothesis Testing I

- First, determine the value of $\beta_j$ for the null hypothesis testing; normally, this value is set to 0; call this value $\mu_0$.
- Compute the $t$-statistic with the following formula:

$$t = (\hat{\beta}_j - \mu_0)/se(\hat{\beta}_j)$$

- Determine the number of degrees of freedom as $K - n - 1$ where $K$ is the number of observations ans $n$ is the number of independent variables; determine the critical rejection level (e.g., 0.05 or 0.01).
- Use $t$-table to find the critical value of $t - t^*$; if you use one-sided test, then you should look for upper-tail probability = critical rejection level; if you use two-tailed test, then should look for upper-tail probability = critical rejection level/2.

## General Algorithm for Null Hypothesis Testing II

- If you test $\beta_j > \mu_0$, then you should check whether $t > t^*$.
- If you test $\beta_j < \mu_0$, then you should check whether $t < t^*$.
- Finally, if you test $\beta_j \neq 0$, then you should check whether $|t| > t^*$.

## Confidences Intervals in OLS Context I

- Confidence intervals computed in OLS context do not differ all that much in interpretation from the ones we covered three weeks ago.
- Ideas remain virtually the same: you want to have
$$\mathbb{P}[-c \leq (\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) \leq c] = 0.95$$
or some larger number for wider confidence interval (and smaller for tighter confidence interval).
- Transforming this gives the following:
$$\mathbb{P}[\hat{\beta}_j - c * se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c * se(\hat{\beta}_j)] = 0.95$$
with $c$ picked based on degrees of freedom of the $t$ distribution.

## Confidences Intervals in OLS Context II

- I would once again encourage you not to make a common mistake and interpret confidence interval as probability of a population parameter located between upper and lower bounds: population parameter is not a probabilistic quantity because in the population it is defined by population regression function.
- Instead, what we have is the probability that, in repeated sampling, the true population parameter will be located between upper and lower bounds 95 times out of 100; nonetheless, since upper and lower bounds are themselves random variables, we never know for sure whether we "hit" the population parameter by the actual computed interval or "missed" it.

## Testing Relationship Between Multiple Parameters I

- Sometimes we might be interested in the relationship between multiple regression coefficients.
- The null hypothesis can be $H_0 : \beta_1 = \beta_2$ with alternatve $\beta_1 < \beta_2$.
- The test statistic in such a case is $t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)}$.
- The denominator can be computed by:

$$\sqrt{[se(\hat{\beta}_1)]^2 + [se(\hat{\beta}_2)]^2 - 2s_{12}}$$

  where $s_{12}$ denotes the estimate of the $\mathbb{C}[\hat{\beta}_1, \hat{\beta}_2]$.
- We can get the information for computations above from the estimated variance-covariance matrix (statsmodels computes it for you automatically with cov_params method).

## Testing Relationship Between Multiple Parameters II

- We can also test the hypothesis about the link between multiple parameters by defining $\theta_1 = \beta_1 - \beta_2$ and estimating the following equation:

$$y = \beta_0 + (\theta_1 + \beta_2)x_1 + \beta_2 x_2 + u = \beta_0 + \theta_1 x_1 + \beta_2(x_1 + x_2) + u$$

- The alternative hypothesis $\beta_1 < \beta_2$ corresponds to alternative hypothesis $\theta_1 < 0$ which we can now test with the already covered approach.

## How Many Variables are Enough?

- In practical applications we often face the problem: if we have many possible predictors for the dependent variable, which among those are important?
- Mathematically, if the independent variable does not have any effect, the corresponding parameter $\beta$ of the population regression function should be 0.
- It turns out we can test model specifications against each other using the so-called $F$-test.

$F$-test: the Definition

- Suppose you have a model with $n+1$ parameters ($n$ independent variables plus one intercept); you then posit that $q \leq n$ slope parameters are all 0.
- To test this hypothesis, you first estimate the **unrestricted** model, i.e., the one that has all variables included, and then **restricted** model, i.e., the one that excludes $q$ variables you posited having 0 effect.
- To compute the $F$ test statistic, you calculate

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(K-n-1)}$$

  where $SSR_r$ denotes sum of squared residuals for restricted model, $SSR_{ur}$ denotes sum of squared residuals for unrestricted model.
- It can be shown that $F$ test statistic has an $F$ distribution with $q, K-n-1$ parameters (degrees of freedom).

## $F$-test: the Details I

- Since $F$ test statistic follows $F$ distribution, it stands to reason that we can perform inferential analysis similar to the null hypothesis testing.
- $p$-values for the $F$ test statistic are computed as

$$\mathbb{P}[\mathcal{F} > F]$$

- Knowing degrees of freedom parameters, this value can be easily calculated by a computer.
- For critical rejection levels, we simply check whether the computed $p$-value is lower than 0.05, 0.01, or 0.001.
- Sometimes, it is also easier to use alternative formula for $F$ that involves $R^2$:

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(N - K)}$$

## $F$-test: the Details II

- To see how we can get the alternative formula for $F$, first pre-multiply the numerator and denominator by $1/TSS$ where $TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2$:

$$F = \frac{[1/TSS] * [(SSR_r - SSR_{ur})/q]}{[1/TSS] * [SSR_{ur}/(N-K)]}$$

- From that, we get

$$\frac{(SSR_r/TSS - SSR_{ur}/TSS)/q}{[SSR_{ur}/TSS]/(N-K)} =$$

$$\frac{(SSR_r/TSS - SSR_{ur}/TSS + 1 - 1)/q}{[SSR_{ur}/TSS + 1 - 1]/(N-K)} =$$

$$\frac{(1 - SSR_{ur}/TSS - (1 - SSR_r/TSS))/q}{[1 - (1 - SSR_{ur}/TSS)]/(N-K)}$$

## $F$-test: the Details III

- In its "extreme" form, $F$-test can be used to evaluate whether the model has any relevance at all by testing it against restricted model with only the intercept present.
- This is akin to a sanity check: if our independent variables are totally unrelated to dependent variable, then there is not much merit in the model to begin with.
- Another usefulness of $F$-test stems from the famous Occam's razor: if there are multiple explanations available, the simplest possible one should be preferred; $F$-test allows to determine a "minimal" model in terms of explanatory variables.

## $F$-test: the Details IV

- The final thing pertaining to the $F$-test that we need to cover is related to the non-zero exclusion restrictions: for instance, you may imagine $\beta_1 = 1, \beta_2 = 2, \beta_3 = 0, \beta_4 = 0$ hypothesis for the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

- To test this, plug exclusion restrictions from the null hypothesis into the model:

$$y = \beta_0 + x_1 + 2x_2 + u \implies y - x_1 - 2x_2 = \beta_0 + u$$

- Therefore, adding non-zero restrictions is equivalent to estimating the restricted model with a different dependent variable.