25934 Applied Financial Econometrics

Lecture 2
Applications of Linear Models

Vitali Alexeev

Finance Discipline Group
UTS Business School
University of Technology Sydney, Australia

Table of Content

1. Regression Analysis
   - Why do we include a Disturbance term?
   - Assumptions Underlying Linear Regression Model
   - Consistency, Unbiasedness, and Efficiency
2. Measures of Fit
   - Simple Regression: An Example
   - Influential Observations
3. Residual Diagnostics
4. Collinearity
   - Multi-collinearity
5. Statistical Inference
   - Single Coefficient Tests
   - Multiple Coefficients Tests
   - Restricted Regressions
   - F-statistic
6. Appendix

Regression Analysis

$$Y \approx f(X, \beta)$$

- **Regression analysis** is a statistical process for estimating the relationships among variables.
- The focus is on the relationship between a **dependent variable** and one or more **independent variables** (or "predictors").
- Helps understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.
- The estimation target is a function of the independent variables called the **regression function**.
- Regression analysis is widely used for **prediction** and **forecasting**.
- Can be used to infer causal relationships between the independent and dependent variables. However, beware of false relationships (correlation does not imply causation).

Regression is different from Correlation

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \ldots, n$$

- If we say y and x are correlated, it means that we are treating y and x in a completely symmetrical way.
- In regression, we treat the dependent variable (y) and the independent variable(s) (x's) very differently.
- The y variable is assumed to be random or "stochastic" in some way, i.e. to have a probability distribution.
- The x variables are, however, assumed to have fixed ("non-stochastic") values in repeated samples.

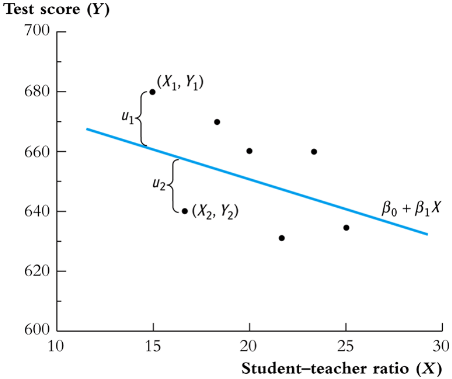
Why do we include a Disturbance term?

Usually denoted by u, e, or ϵ

- The disturbance term can capture a number of features:
  - We always leave out some determinants of y_t
  - There may be errors in the measurement of y_t that cannot be modelled.
  - Random outside influences on y_t which we cannot model

Determining the Regression Coefficients

Choose β_0 and β_1 so that the (vertical) distances from the data points to the fitted lines are minimised (so that the line fits the data as closely as possible):



Mechanics of OLS

The OLS Estimator, Predicted Values, and Residuals

The OLS estimator of the slope β_1 and the intercept β_0 are:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2} \\ \hat{\beta}_0 &= \bar{Y}-\hat{\beta}_1\bar{X} \end{aligned}$$

The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are:

$$\begin{aligned} \hat{Y}_i &= \hat{\beta}_0+\hat{\beta}_1 X_i, \;\; i=1,...,n \\ \hat{u}_i &= Y_i-\hat{Y}_i, \;\; i=1,...,n \end{aligned}$$

The estimated intercept ($\hat{\beta}_0$), slope ($\hat{\beta}_1$), and residual ($\hat{u}_i$) are computed from a sample of n observations of X_i and Y_i, $i=1,...,n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

Assumptions Underlying the Classical Linear Regression Model I

- We observe data for x_t, but since y_t also depends on u_t, we must be specific about how the u_t are generated.
- We usually make the following set of assumptions about the u_t's (the unobservable error terms):

Technical notation	Interpretation
(1) $\mathrm{E}(u_t) = 0$	The errors have zero mean
(2) $\mathrm{var}(u_t) = \sigma^2$	The variance of the errors is constant and finite over all values of x_t
(3) $\mathrm{cov}(u_i,\, u_j) = 0$	The errors are linearly independent of one another
(4) $\mathrm{cov}(u_t,\, x_t) = 0$	There is no relationship between the error and corresponding x variate

Assumptions Underlying the Classical Linear Regression Model II

- A fifth assumption is required if we want to make inferences about the population parameters (the actual β_0 and β_1) from the sample parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$)
- Additional assumption

  (5) u_t is normally distributed

Consistency, Unbiasedness, and Efficiency

Consistent

- That is, the estimates will converge to their true values as the sample size increases to infinity. Need the assumptions $E(x_t u_t) = 0$ and $Var(u_t) = \sigma^2 < \infty$ to prove this.

Unbiased

- That is, on average, the estimated value will be equal to the true values. To show this, we require the assumption that $E(u_t) = 0$. Unbiasedness is a stronger condition than consistency.

Efficient

- An estimator $\hat{\beta}$ of parameter β is said to be efficient if it is unbiased and no other unbiased estimator has a smaller variance.

Measures of Fit

Two regression statistics provide complementary measures of how well the regression line “fits” or explains the data:

- The **regression** R^2 measures the fraction of the variance of Y that is explained by X; it is unit-free and ranges between zero (no fit) and one (perfect fit)
- The **standard error of the regression (SER)** measures the magnitude of a typical regression residual in the units of Y.

R^2

The regression R^2 is the fraction of the sample variance of Y_i "explained" by the regression.

- $Y_i = \hat{Y}_i + \hat{u}_i =$ OLS prediction + OLS residual
- $Var(Y) = Var(\hat{Y}) + Var(\hat{u})$
- Total sum of squares (TSS) = "explained" SS (ESS)+ "residual" SS (RSS)

$$\text{Definition of } R^2: \quad R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$$

- $R^2 = 0$ means $ESS = 0$
- $R^2 = 1$ means $ESS = TSS$
- $0 \leq R^2 \leq 1$
- For regression with a single variable X, R^2 is the square of the correlation coefficient between X and Y.

Measures of Fit for Multiple Regression

Recall, that R^2 is the fraction of the variance explained

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

where $ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$, $SSR = \sum_{i=1}^{n} \hat{u}_i^2$, $TSS = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$.

- R^2 now becomes the square of the correlation coefficient between Y and predicted $\hat{Y}$.
- It is still the proportional reduction in the residual sum of squares as we move from modeling Y with just a sample mean, to modeling it with a group of variables.
- The R^2 always increases when you add another regressor X because SSR has to *decrease to add it*
  - otherwise adding it would actually increase SSR.
- This is a problem for quantifying the measure of “fit”:
  - You can always increase R^2 by just adding another X.

R^2 and $\bar{R}^2$

Adjusted R^2 ($\bar{R}^2$)

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SSR}{TSS} = 1 - \left(\frac{n-1}{n-k-1}\right)\left(1 - R^2\right)$$

where k is the number of regressors.

- The $\bar{R}^2$ (**adjusted** R^2) correct the problem by "penalizing" the inclusion of another regressor.
- The $\bar{R}^2$ does not necessarily increase when you add another regressor.
- Note that $\bar{R}^2 < R^2$. However, if n is large, the two will be very close.

Standard Error of the Regression (SER)

- The SER measures the spread of the distribution of u.
- The SER is (almost) the sample standard deviation of the OLS residuals:

$$
\begin{aligned}
SER &= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(\hat{u}_i - \bar{\hat{u}}\right)^2} \\
&= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}
\end{aligned}
$$

The second equality holds because $\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$.

Root mean squared error (RMSE)

The SER

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

- has the units of u, which are the units of Y
- measures the average "size" of the OLS residual (the average "mistake" made by the OLS regression line)

The **root mean squared error (**$RMSE$**)** is closely related to the SER:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

- This measures the same thing as the SER – the minor difference is division by $\frac{1}{n}$ instead of $\frac{1}{n-2}$.

Simple Regression: An Example

- Suppose that we have the following data on the excess returns on a fund manager's portfolio ("fund XXX") together with the excess returns on a market index:

Year, t	Excess return $= r_{XXX,t} - rf_t$	Excess return on market index $= rm_t - rf_t$
1	17.8	13.7
2	39.0	23.2
3	12.8	6.9
4	24.2	16.8
5	17.2	12.3

- We have some intuition that the beta on this fund is positive, and we therefore want to find whether there appears to be a relationship between x and y given the data that we have. The first stage would be to form a scatter plot of the two variables.

Leverage

- In statistics and, in particular, in regression analysis, **leverage** is a measure of how far away the independent variable values of an observation are from those of the other observations.
- High-leverage points are those observations, made at extreme or outlying values of the independent variables such that the lack of neighboring observations means that the fitted regression model will pass close to that particular observation.

Cook's distance

after the American statistician R. Dennis Cook

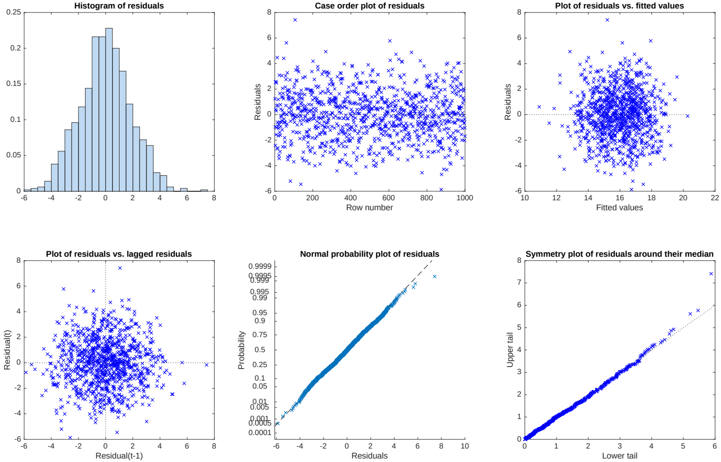
- **Cook's distance** or **Cook's D** is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis.
- In practice, Cook's distance (large values of Cook's D) can be used to indicate influential data points that are particularly worth checking for validity.

Cook's distance measures the effect of deleting a given observation

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p \times s^2}$$

where $\hat{y}_{j(i)}$ is the fitted response value obtained when excluding i; p is number of independent variables, and s^2 is the mean squared error of the regression.

Residual Diagnostics



Perfect multicollinearity

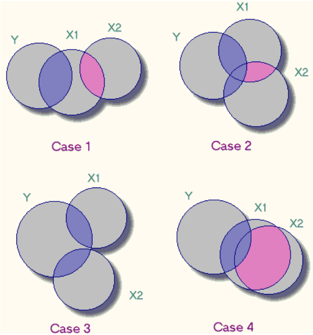
Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

- What would happen if you include *STR* twice in your regression?
- In such a regression, β_1 is the effect on *TestScore* of a unit change in *STR*, holding *STR* constant... 8-(
- The *Standard Errors* become **infinitely large** when perfect multicollinearity exists.
- Solution: Modify your regression model and remove collinear regressors.

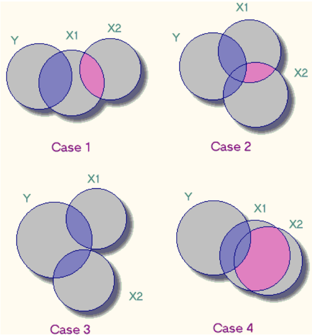
Imperfect multicollinearity

Imperfect multicollinearity is when two or more of the the regressors are highly correlated.

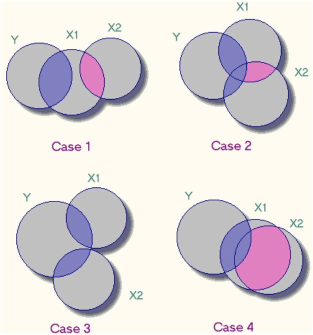
- Why this name?
  - If two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line - they are collinear - but unless the correlation is exactly ± 1, that collinearity is "imperfect".
- The *Standard Errors* become **larger** when imperfect multicollinearity exists.
- Solution: There is no direct solution but **Belsley Collinearity Diagnostics** could be helpful.



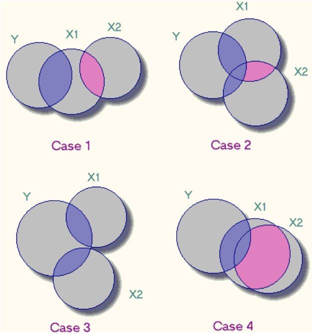
- R^2 alone cannot tell you how well your model is specified.
- Consider the four cases: the overlapping area between Y and X_1, X_2 is the variance explained.
- In all four cases the union of Y and X are almost the same. Numerically you cannot tell much difference when the R^2 are .45, .48, .41, .40.
- Actually, all these models are very different.



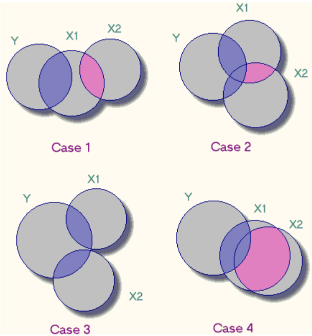
- **Case 1**: X_1 and X_2 are related; X_1 and Y are related, **but** X_2 and Y has no relationship.
- E.g., the number of hours of study is related to test scores, the frequency of going to the restroom is related to study (you drink more coffee to stay up), but going to the bathroom is not related to the test performance.



- **Case 2**: Both X_1 and X_2 contribute to some unique variance explained to Y, but they also have some common variance explained.
- E.g., drinking and smoking may cause cancer, and you smoke more when you drink.



- **Case 3**: Again, both X_1 and X_2 contribute unique variance explained to Y, **but** X_1 and X_2 are totally unrelated (*orthogonal*).
- E.g., For instance, mathematical intelligence and verbal intelligence could predict competence in business, but these two types of intelligence have no relationship. A good speaker may not be able to count from one to ten.

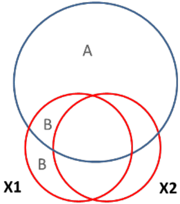


- **Case 4**: Although both X_1 and X_2 could predict Y, the variance explained contributed by X_2 has been covered by X_1 because X_1 and X_2 are too correlated (*collinear*).

Multi-collinearity and OLS estimator standard errors

Visual interpretation

- There is a high degree of collinearity in the model below.
- The higher the collinearity, the smaller the total B area will be and, thus, the higher the standard error of the slope of the regression plane.



$$SE_{b_1} \sim \frac{A}{B}$$

Trade-off between bias and $SE(\hat{\beta})$

- **Leaving out** X_2 risks creating an OVB bias for the coefficient β_1. We risk *mis-attributing causation*, i.e. **wrongly attributing** X_2**'s effect on** Y **to** X_1.
- **Including** X_2 can lead to imperfect multicollinearity, high SE's and to an **incorrect non-rejection of the null** . We therefore risk a *false positive* for insignificance of X_1.
- *Model specification* is deciding which variables to include in a model.
- There's a bias/precision trade-off, and specification is hard and takes experience and understanding of the real world you are modelling.

Testing Hypotheses: The Test of Significance Approach I

Assume the regression equation is given by,

$$y_t = \beta_0 + \beta_1 x_t + u_t \text{ for } t = 1, 2, ..., T$$

The steps involved in doing a test of significance are:

1. Estimate $\hat{\beta}_0$, $\hat{\beta}_1$ and $SE(\hat{\beta}_0)$, $SE(\hat{\beta}_1)$ in the usual way
2. Calculate the test statistic. This is given by the formula

   $$test\ statistic = \frac{\hat{\beta}_1 - \beta_1^*}{SE(\hat{\beta}_1)}$$

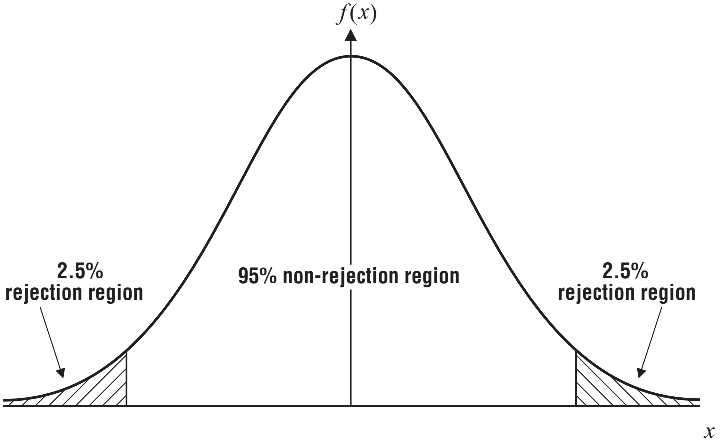
   where β_1^* is the value of β_1 under the null hypothesis.
3. Use tabulated distribution with which to compare the estimated test statistics. E.g., a t-distribution with T-2 degrees of freedom.

Testing Hypotheses: The Test of Significance Approach II

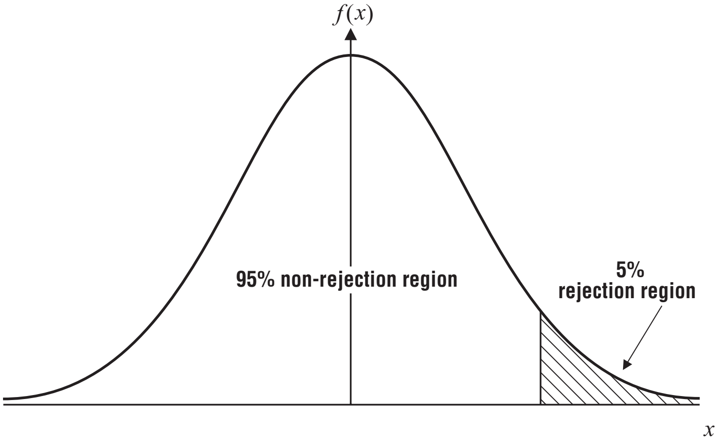
4. Choose a "significance level", often denoted α. This is also sometimes called the size of the test and it determines the region where we will reject or not reject the null hypothesis that we are testing. It is conventional to use a significance level of 5%, but 10% and 1% are also commonly used.

Determining the Rejection Region for a Test of Significance

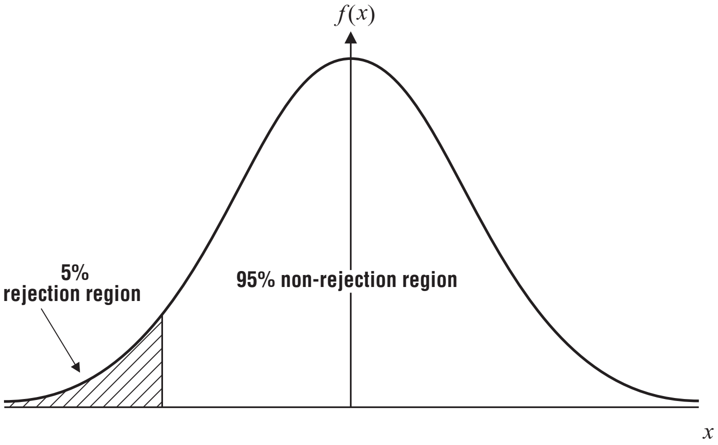
5. Given a significance level, we can determine a rejection region and non-rejection region. For a 2-sided test:



The Rejection Region for a 1-Sided Test (Upper Tail)



The Rejection Region for a 1-Sided Test (Lower Tail)



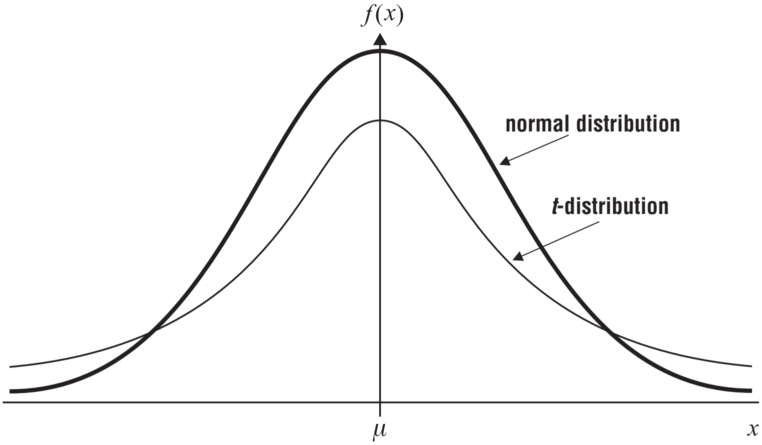
The Test of Significance Approach: Drawing Conclusions

6. Use the t-tables to obtain a critical value or values with which to compare the test statistic.
7. Finally perform the test. If the test statistic lies in the rejection region then reject the null hypothesis (H_0), else do not reject H_0.

A Note on the t and the Normal Distribution

- You should all be familiar with the normal distribution and its characteristic "bell" shape.
- We can scale a normal variate to have zero mean and unit variance by subtracting its mean and dividing by its standard deviation.
- There is, however, a specific relationship between the t- and the standard normal distribution. Both are symmetrical and centred on zero. The t-distribution has another parameter, its degrees of freedom. We will always know this (for the time being from the number of observations -2).

What Does the t-Distribution Look Like?



Comparing the t and the Normal Distribution

- In the limit, a t-distribution with an infinite number of degrees of freedom is a standard normal, i.e. $t(\infty) = N(0, 1)$
- Examples from statistical tables:

Significance level	$N(0,1)$	$t(40)$	$t(4)$
50%	0	0	0
5%	1.64	1.68	2.13
2.5%	1.96	2.02	2.78
0.5%	2.57	2.70	4.60

- The reason for using the t-distribution rather than the standard normal is that we had to estimate σ^2, the variance of the disturbances.

Test of Overall Significance of a model

With the general multiple regression model with n explanatory variables and $n+1$ unknown coefficients:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + u$$

To test if we have a **viable explanatory model** we set up the following null (denoted H_0) and alternative hypotheses (denoted H_1 or H_A):

$$\begin{aligned} H_0 &: \quad \beta_1 = 0, \beta_2 = 0, ..., \beta_n = 0 \\ H_1 &: \quad \text{at least one } \beta_i \neq 0, \quad i = 1, ..., n \end{aligned}$$

Note that the **constant is not included** in the null.

- The null hypothesis has n parts, and it is called a **joint hypothesis**. It amounts to n restrictions!
- If this null hypothesis is true, **none** of the explanatory variables influence Y, and thus our model is of little or no value.

Example: Do School Resources Matter or Not?

Let $Expn$ = expenditures per pupil and consider the population regression model:

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

The null hypothesis, "*school resources do not matter*", and the alternative that they do, corresponds to:

$$\begin{aligned} H_0 &: \beta_1 = 0 \textbf{ and } \beta_2 = 0 \\ H_1 &: \textbf{either } \beta_1 \neq 0 \textbf{ or } \beta_2 \neq 0 \textbf{ or both} \end{aligned}$$

Joint Hypothesis: Do School Resources Matter?

$$H_0 : \beta_1 = 0 \textbf{ and } \beta_2 = 0$$
$$H_1 : \textbf{either } \beta_1 \neq 0 \textbf{ or } \beta_2 \neq 0 \textbf{ or both}$$

- A **joint hypothesis** specifies a value for two or more coefficients. That is, it imposes a restriction on two or more coefficients.
- In general, a joint hypothesis will involve q restrictions. In the above example above, $q = 2$, and the two restrictions are $\beta_1 = 0$ and $\beta_2 = 0$.
- A "common sense" idea is to reject if either of the individual t-statistics exceeds 1.96 in absolute value.
- This "one at a time" test is not actually valid: the resulting test rejects too often under the null hypothesis (more than 5%)!

Joint Hypothesis: Do School Resources Matter?

Can't we just t-test the coefficients one at a time?

- No! Because the rejection rate under the null is not 5%.

Let's calculate the probability of incorrectly rejecting the null using this "common sense" test based on the two individual t-stats.

For simplicity, suppose that $\hat{\beta}_1$ and $\hat{\beta}_2$ are independently distributed. Let t_1 and t_2 be the t-statistics:

$$t_1 = \frac{\hat{\beta}_1 - 0}{SE\left(\hat{\beta}_1\right)} \quad \text{and} \quad t_2 = \frac{\hat{\beta}_2 - 0}{SE\left(\hat{\beta}_2\right)}$$

The "one at time" test is:

$$\text{Reject } H_0\text{: } \beta_1 = \beta_2 = 0 \text{ if } |t_1| > 1.96 \text{ or } |t_2| > 1.96 \text{ (or both).}$$

What is the probability that this "one at a time" test rejects H_0, when H_0 is actually true? (It *should* be 5%.)

Suppose t_1 and t_2 are independent. The probability of incorrectly rejecting H_0 using "one at a time" t-tests is:

$$
\begin{aligned}
&= \Pr_{H_0}\left(|t_1| > 1.96 \text{ and/or } |t_2| > 1.96\right) \\
&= \Pr_{H_0}\left(|t_1| > 1.96, |t_2| > 1.96\right) + \Pr_{H_0}\left(|t_1| > 1.96, |t_2| \leq 1.96\right) \\
&\quad + \Pr_{H_0}\left(|t_1| \leq 1.96, |t_2| > 1.96\right) \\
&= \Pr_{H_0}\left(|t_1| > 1.96\right) \times \Pr_{H_0}\left(|t_2| > 1.96\right) \\
&\quad + \Pr_{H_0}\left(|t_1| > 1.96\right) \times \Pr_{H_0}\left(|t_2| \leq 1.96\right) \\
&\quad + \Pr_{H_0}\left(|t_1| \leq 1.96\right) \times \Pr_{H_0}\left(|t_2| > 1.96\right) \\
&\quad \text{(as } t_1 \text{ and } t_2 \text{ are independent by assumption)} \\
&= 0.05 \times 0.05 + 0.05 \times 0.95 + 0.95 \times 0.05 \\
&= 0.0975 \\
&= 9.75\% \text{ which is } \textbf{not} \text{ the desired } 5\%!
\end{aligned}
$$

The size of a test is the actual rejection rate under the null hypothesis. It should be alpha (α)!

- The size of the "common sense" test is not 5%!
- In fact, its size depends on the correlation between t_1 and t_2 (and thus on the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$).

Solution:
Use a different test statistic to test both β_1 and β_2 at once: the F-statistic (this is common practice).

Restricted Regressions

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots, n$$

- One more thing before we do the F-test, let's consider the effect of imposing a restriction on a regression, and getting a restricted regression.
- This is just the model you get when a hypothesis is assumed true, which places some restrictions on the coefficients.

Restricted Regressions

Example

Given $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i, \;\; i = 1, \ldots, n$, we want to know whether $\beta_1 = 100\beta_2$.

Consider:

$$\begin{aligned} Y &= \beta_0 + 100\beta_2 X_1 + \beta_2 X_2 + \beta_3 X_3 + u \\ &= \beta_0 + \beta_2 \left(100 X_1 + X_2\right) + \beta_3 X_3 + u \\ &= \beta_0 + \beta_2 \widetilde{X}_2 + \beta_3 X_3 + u, \quad \text{where } \widetilde{X}_2 = 100 X_1 + X_2 \end{aligned}$$

The restricted regression is an OLS regression of Y on a constant, $\widetilde{X}_2$, and X_3.

Restricted Regressions

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots, n$$

Example 1:

- Consider the statement that "this whole model is worthless".
- This amounts to the restriction:

$$\beta_1 = \beta_2 = \ldots = \beta_k = 0$$

- Then the restricted model is:

$$Y_i = \beta_0 + u_i, \quad i = 1, \ldots, n$$

- The restricted regression is an OLS regression of Y on a constant.
- The estimate for the constant will just be the **sample mean** of Y.

Restricted Regressions

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i, \quad i = 1, ..., n$$

Example 2:

- Instead of "the whole model being worthless", you want to check if X_1 and X_2 are **jointly insignificant**.
- This amounts to the restriction:

$$\beta_1 = \beta_2 = 0$$

- Then the restricted model (true model under the H_0) is:

$$Y_i = \beta_0 + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i, \quad i = 1, ..., n$$

- The restricted regression is an OLS regression of Y on a constant and X_3 to X_k.

Properties of Restricted Regressions

- Imposing a restriction always **increases the residual sum of squares**, since you're forcing the estimates to take the values implied by the restriction, rather than letting OLS choose the coefficient values to minimize the SSR.
- If the SSR increases a lot, this implies that the restriction is relatively "unbelievable",
  - i.e. the model fits a lot worse with the restriction imposed.
- This is the basic intuition of the F-test:
  - impose the restriction and see if SSR goes up "too much".

F-test and Restricted Least Squares (RLS)

Run regression with the restriction imposed.

$$\begin{aligned} H_0 &: \text{ restriction on } \beta_0, \beta_1, ..., \beta_k \\ H_1 &: \text{ violation of restriction} \end{aligned}$$

- **If the null hypothesis is true**, we expect that the data are compatible with the restrictions placed on the parameters.
  - So, we expect little change in the sum of squared errors (SSR).
- If the Restricted SSRs are **substantially bigger** than the Unrestricted SSRs, then applying the restriction has **significantly reduced** the model's ability to explain the data.
- We call the sum of squared errors in the model that *assumes a null hypothesis to be true* the **restricted sum of squared errors**, or SSR_R.
- Conversely, the sum of squared errors from the *original* model is the **unrestricted sum of squared errors**, or SSR_{UR}.
- It is always true that $SSR_R \geq SSR_{UR}$ and thus $SSR_R - SSR_{UR} \geq 0$.

F-statistic

Introduction

- We need some way to decide if the increase in SSR when we move to a restricted model is "big enough" to reject the restrictions.
- We use the F-statistic to do this.
- We also need to know about the sampling distribution of the F-statistic, under the H_0 assuming the restriction.

F-test and Restricted Least Squares (RLS)

$$
\begin{aligned}
H_0 &: \text{restriction on } \beta_0, \beta_1, ..., \beta_k \\
H_1 &: \text{violation of restriction}
\end{aligned}
$$

Let q be the number of restrictions being tested. Then the general F-statistic is given by

$$F = \frac{(SSR_R - SSR_{UR})\,/q}{SSR_{UR}/\,(n-k-1)}$$

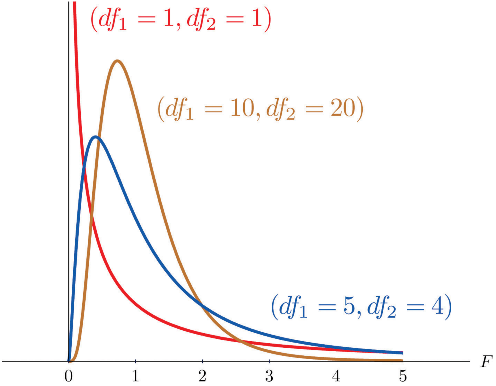
- We **run both** the **restricted regression** and the **unrestricted regression** (i.e. the original regression), n = number of data samples.
- If H_0 is true, then the sampling distribution of the F-statistic is F-distributed, $F\text{-statistic} \sim F_{q,n-k-1}$ with q numerator degrees of freedom and $n-k-1$ denominator degrees.
- We reject H_0 if the value of the F-test statistic **becomes "too large"**, but how much is that?

Some facts about the F-statistic

- The F-statistic is *always positive*, since $SSR_R \geq SSR_{UR}$.
- We compare the calculated value of F_{actual} to a critical value F_{crit} which leaves a probability α in the upper tail of the F-distribution with $(q, n-k-1)$ degrees of freedom (d.f.) as shown on the following slide.
- The F-statistic is essentially measuring the *relative increase* in SSR when moving from the unrestricted to restricted model.
- We want to know if the change in SSR is *big enough* to suggest the restriction is wrong.

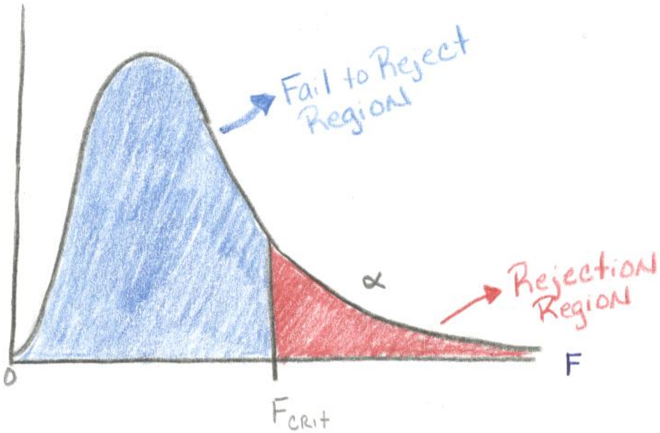
F distribution

The shape of F distribution depends on numerator d.f. ($df_1 = q$) and denominator d.f. ($df_2 = n - k - 1$)



F-test

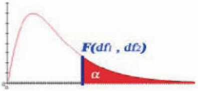
Reject H_0 at α significance level if $F_{actual} > F_{crit}$



Critical values of F distribution

$\alpha = 0.05$, 5% significance level

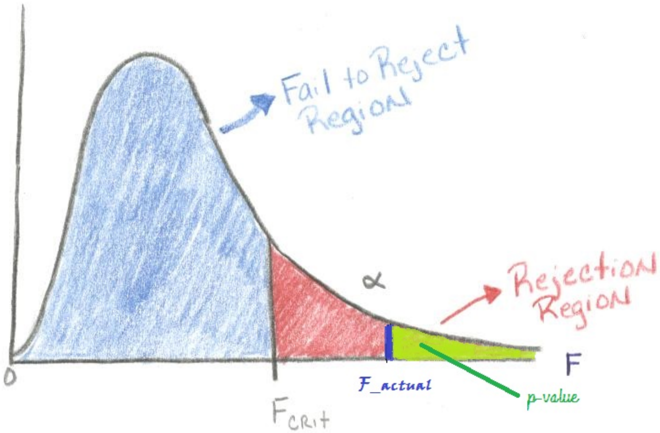
F Table for α = .05



	df_1=1	2	3	4	5	6	7	8	9	10
df_2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45

F-test

Equivalently, reject H_0 if $p - \text{value} < \alpha$



The R^2 form of the F-statistic

Recall that $R^2 = ESS/TSS = 1 - SSR/TSS$, thus $SSR = \left(1 - R^2\right) TSS$. Since TSS is the total variability in Y, it does not depend on whether you have restricted or unrestricted model. We have:

$$SSR_R = \left(1 - R_R^2\right) TSS \quad \text{and} \quad SSR_{UR} = \left(1 - R_{UR}^2\right) TSS$$

F-statistics can then be written as:

$$\begin{aligned} F &= \frac{\left(SSR_R - SSR_{UR}\right)/q}{SSR_{UR}/\left(n-k-1\right)} \\ &= \frac{\left(\left(1 - R_R^2\right) TSS - \left(1 - R_{UR}^2\right) TSS\right)/q}{\left(\left(1 - R_{UR}^2\right) TSS\right)/\left(n-k-1\right)} \\ &= \frac{\left(R_{UR}^2 - R_R^2\right)/q}{\left(1 - R_{UR}^2\right)/\left(n-k-1\right)} \end{aligned}$$

UTS

Overall Significance Test

A special case of exclusion restrictions is to test

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

- $R_R^2 = 0$ for a restricted model with only an intercept coefficient β_0.
- This is because the OLS estimator is just the sample mean, implying the $TSS = SSR$.
- The calculated F-statistic for $q = k$ restrictions is then

$$F = \frac{\left(R_{UR}^2\right) / k}{\left(1 - R_{UR}^2\right) / (n - k - 1)} \sim F_{k,n-k-1}$$

Or simply $F = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k,n-k-1}$

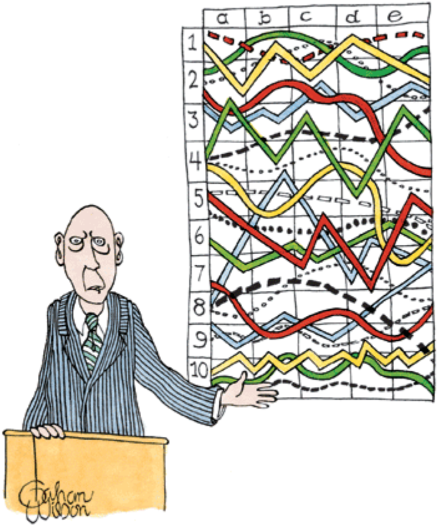
In-class example

TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31** (0.34)	−1.01** (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547** (0.024)		−0.529** (0.038)
Percent on public income assistance (X_4)				−0.790** (0.068)	0.048 (0.059)
Intercept	698.9** (10.4)	686.0** (8.7)	700.2** (5.6)	698.0** (6.9)	700.4** (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\bar{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K-8 school districts in California, described in Appendix 4.1. Heteroskedasticity-robust standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.



"I'll pause for a moment so you can let this information sink in."