Analysis of Economics Data
Chapter 11: Statistical Inference for Multiple Regression

© A. Colin Cameron
Univ. of Calif. Davis

November 2022

CHAPTER 11: Statistical Inference for Multiple Regression

- Consider statistical inference for the relationship between house price and several variables
  - size, number of bedrooms, ...
- Mostly a straight-forward extension of bivariate regression
  - now use $T(n-k)$ distribution where $k =$ number of regressors including intercept.
- New is:
  - tests of joint hypotheses (rather than a single hypothesis).

Outline

Example for this Chapter with dependent variable price

Variable	Coefficient	St. Error	t statistic	p value	95% conf. int.	
Size	68.37	15.39	4.44	0.000	36.45	101.29
Bedrooms	2685	9193	0.29	0.773	-16379	21749
Bathrooms	6833	15721	0.43	0.668	-25771	39437
Lot Size	2303	7227	0.32	0.753	-12684	17290
Age	-833	719	-1.16	0.259	-2325	659
Month Sold	-2089	3521	-0.59	0.559	-9390	5213
Intercept	137791	61464	2.24	0.036	10321	265261
n	29					
F(6,22)	6.83					
p-value for F	0.0003					
R^2	0.651					
Adjusted R^2	0.555					
St. error	24936					

11.1 Properties of the Least Squares Estimator

- Data assumption
  - There is variation in the sample regressors so regressors are not perfectly correlated with each other
  - generalize bivariate regression cannot estimate b_2 if $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$.
- If this data assumption does not hold then it is not possible to estimate all k regression coefficients
  - see chapter 10.8 and later chapter on multicollinearity.

Population Model Assumptions

- These are a straightforward extension of those for bivariate regression.

1. Population model:
   $y = \beta_1 + \beta_2 x_2 + \beta_2 x_3 + \cdots + \beta_k x_k + u.$
2. Error has zero mean conditional on all regressors:
   $\mathrm{E}[u_i | x_{2_i}, ..., x_{ki}] = 0, \quad i = 1, ..., n.$
3. Error has constant variance conditional on the regressors:
   $\mathrm{Var}[u_i | x_{2i}, ..., x_{ki}] = \sigma_u^2, \quad i = 1, ..., n.$
4. Errors for different observations are statistically independent
   u_i is independent of $u_j, \quad i \neq j.$

Population Model Assumptions (continued)

- Key is that Assumptions 1-2 imply the **population regression line** or the **conditional mean** of y given $x_1, ..., x_k$ is

$$\mathrm{E}[y|x_2, ..., x_k] = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k.$$

- Assumptions 2-4 imply

$$u_i \sim [0, \sigma_u^2] \text{ and is independent over } i.$$

- Assumptions 1-4 imply

$$y_i|x_{2i}, ..., x_{ki} \quad \sim \quad [(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki}),\ \sigma_u^2] \text{ and is independent over } i.$$

  - Similar to univariate: μ replaced by $\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k$.
  - Similar to bivariate: $\beta_1 + \beta_2 x_2$ replaced by $\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k$.

Properties of Least Squares Estimates

- **Mean** of b_j is β_j under assumptions 1-2.
- **Variance** of b_j is $\text{Var}[b_j] = \sigma^2_{b_j} = \sigma^2_u / \sum_{i=1}^n \widetilde{x}_{ji}{}^2$
  - where $\widetilde{x}_{ji}$ is the residual from regressing x_{ji} on an intercept and all regressors other than x_{ji}
  - from chapter 10 $b_j = \sum_{i=1}^n \widetilde{x}_{ji} yto_i ensure / \sum_{i=1}^n \widetilde{x}_{ji}{}^2$.
- **Standard error of the regression** is s_e where $s^2_e = \frac{1}{n-k} \sum_{i=1}^n (y_i - \widehat{y}_i)^2$
  - same as for bivariate except divide by $n - k$
  - this ensures $\text{E}[s^2_e] = \sigma^2_u$ given assumptions 1-4.
- **Estimated variance** of b_j is $s^2_e / \sum_{i=1}^n \widetilde{x}_{ji}{}^2$.
- **Standard error of estimator** b_j is $se(b_j) = s_e / \sqrt{\sum_{i=1}^n \widetilde{x}_{ji}{}^2}$.

When is a Slope Coefficient Precisely Estimated?

- **Standard error of estimator** b_j is $se(b_j) = s_e / \sqrt{\sum_{i=1}^{n} \widetilde{x}_{ji}^{\,2}}$.
- So more precise estimate when
  - model fit is good so s_e is small
  - when there are many observations as then $\sum_{i=1}^{n} \widetilde{x}_{ji}^{\,2}$ is big
  - when $|\widetilde{x}_{ji}|$ is big
    - which is the case if there is big dispersion in the j^{th} regressor after controlling for the other regressors.

The t-Statistic

- Confidence intervals and hypothesis tests are based on the *t*-statistic.
- Given assumptions 1-4:

$$t_j = \frac{b_j - \beta_j}{se(b_j)} \sim T(n-k) \text{ approximately}$$

  - now $T(n-k)$ rather than $T(n-2)$.
- The result is exact if additionally the errors are normally distributed.
- How large should the sample be?
  - Larger than in the bivariate regression case.

11.2 Estimators of Model Parameters

- We want OLS estimator b_j for the coefficient j^{th} regressor x_j to be
  - centered on β_j : unbiased and consistent
  - smallest variance (best) among such estimators.
- Centering
  - b_j is **unbiased** for β_j $(\mathrm{E}[b_j] = \beta_j)$ given assumptions 1-2
  - b_j is **consistent for** β_j $(b_j \rightarrow \beta_j$ as $n \rightarrow \infty)$ given assumptions 1-2 plus a little more to ensure $\mathrm{Var}[b_j] \rightarrow 0$ as $n \rightarrow \infty$.
- Smallest variance
  - b_j is **best linear unbiased for** β_j given assumptions 1-4
    - ★ i.e. smallest variance among unbiased estimators that are a weighted average of y_i, $\sum_i a_i y_i$, with weights a_i depending on the regressors.
  - b_j is **best unbiased** for β_j given assumptions 1-4 and normally distributed errors
    - ★ i.e. minimum variance among unbiased estimators.

11.3 Confidence Intervals

- Usual estimate $\pm$ critical t-value $\times$ standard error.
- A $100(1-\alpha)$ **percent confidence interval for** β_j is

$$b_j \pm t_{n-k;\alpha/2} \times se(b_j),$$

where

  - b_j is the slope estimate
  - $se(b_j)$ is the standard error of b_j
  - $t_{n-k,\alpha/2}$ is the critical value
  - e.g. in Stata use `invttail`$(n-k, \alpha/2)$.

- A **95 percent confidence interval** is approximately

$$b_j \pm 2 \times se(b_j).$$

Confidence Interval Example

- Regression of house price on house size and five other regressors
  - output given at start of slides
  - includes a 95% confidence interval for β_{SF} is $(36.45, 100.29)$.
- Manual computation using $b_{SF} = 68.37$ and $se(b_{SF}) = 15.39$:

$$\begin{aligned} & b_{SF} \pm t_{n-k,\alpha/2} \times se(b_{SF}) \\ = & \; 68.37 \pm t_{22,.025} \times 15.39 \\ = & \; 68.37 \pm 2.074 \times 15.39 \\ = & \; 68.37 \pm 31.92 \\ = & \; (36.45, 100.29). \end{aligned}$$

11.4 Tests on Individual Parameters

- Two-sided test that $\beta_j = \beta_j^*$

$$H_0 : \beta_j = \beta_j^* \quad \text{against} \quad H_a : \beta_j \neq \beta_j^*$$

- Use

$$t = \frac{(b_j - \beta_j^*)}{se(b_j)} \sim T(n-k).$$

- Can also do one-sided tests.

Tests of Statistical Significance

- Test whether there is any relationship between y and x_j (after controlling for the other regressors).
- Does $\beta_j = 0$? Formally test

$$H_0 : \beta_j = 0 \quad \text{against} \quad H_a : \beta_j \neq 0$$

- Use t-statistic where $\beta_j = 0$. So simply

$$t = \frac{b_j}{se(b_j)} \sim T(n-k).$$

- Aside: $|t| > 1$ if $\bar{R}^2$ increases when a regressor is added
  - so usual t-test is more demanding than including regressor if adjusted R^2 increases.

Example: House Price

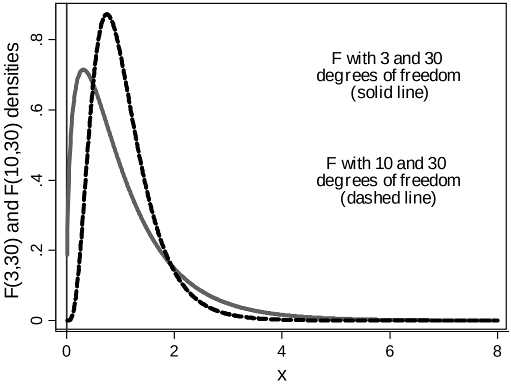
- Test of statistical significance of size for house price example
  - $t = \frac{b_{Size}}{se(b_{Size})} = \frac{68.37}{15.39} = 4.44$
  - so for two-sided test
    - ★ $p = 2 * ttail(22, 4.44) = 0.0002 < 0.05$ so reject H_0
    - ★ or $c = invttail(22, .05) = 1.717$ and $|t| = 4.44 > c$ so reject H_0
  - conclude that *Size* is statistically significance at level 0.05.
- Test of $H_0 : \beta_2 = 50$ against $H_a : \beta_2 \neq 50$
  - $t = \frac{b_{Size} - 50}{se(b_{Size})} = \frac{68.37 - 50}{15.39} = 1.194.$
  - so for two-sided test
    - ★ $p = 2 * ttail(22, 1.194) = 0.245 > 0.05$ so do not reject H_0
    - ★ or $c = invttail(22, .05) = 1.717$ and $|t| = 1.194 < c$ so do not reject H_0
  - conclude that *Size* is statistically significance at level 0.05.

11.5 Joint Hypothesis Tests

- Suppose we wish to test more than one restriction on the parameters.
  - e.g. both $\beta_2 = 0$ and $\beta_3 = 0$
  - e.g. all slope parameters equal zero
  - e.g. $\beta_2 = -\beta_3$ and $2\beta_4 + \beta_6 = 9$.
- Tests of several restrictions are called tests of joint hypotheses.
- t tests can handle test of only one restriction on the parameters.
- Instead use F tests and the F distribution
  - this nests t tests and t distribution as a special case
  - for tests of a single restriction $F = t^2$.

11.5 Joint Hypothesis Tests: F Distribution

- The F distribution is for a random variable that is > 0
  - it is right-skewed
  - it depends on two parameters v_1 and v_2 called degrees of freedom
  - $v_1 =$ number of restrictions; $v_2 = n - k$.



Probabilities and Inverse Probabilities for the F

- General notation is $F(v_1, v_2)$.
- The critical values (and p values) for the F distribution vary with the two degrees of freedom
- For F_{v_1,v_2} the critical value (area in right tail) is
  - decreasing in both v_1 and v_2
- Some representative values
  - 5% and one restriction: $F_{1,30;.05} = 4.17$ and $F_{1,\infty;.05} = 3.84$
  - 5% and ten restrictions: $F_{10,30;.05} = 2.16$ and $F_{10,\infty;.05} = 1.83$.
- Examples in Stata
  - probability: $\Pr[F_{10,30} > 2] =$ `Ftail(10,30,2)`.
  - inverse probability: $F_{10,30;.05} =$ `invFtail(10,30,.05)`

The F Statistic

- Consider two models that are nested in each other.
- General model: **unrestricted model** or **complete model**, is a model with k regressors, so

$$y = \beta_1 + \beta_2 x + \beta_3 x_3 + \cdots + \beta_k x_k + u.$$

- **Restricted model** or **reduced model** places q restrictions on $\beta_1, \beta_2, ..., \beta_k$.
    - e.g. all regressors but the intercept are dropped so $q = k - 1$.
    - e.g. a subset of g regressors is included so $q = k - g$.
    - e.g. one regressor is dropped so $q = 1$.
- In general the formula for the F statistic is complicated
    - just use computer output.

F Tests

- An F **test** is a **two-sided test of**
  - H_0 : The q parameter restrictions implied by the restricted model are correct
  - against H_a : The q parameter restrictions implied by the restricted model are incorrect.
- Define α to be the desired **significance level** of the test.
- **p-value:** $p = \Pr[F_{q,n-k} \geq F\ ]$
  - H_0 is rejected if $p < \alpha$.
- **critical value:** c is such that $c = F_{q,n-k,\alpha}$, equivalently $\Pr[|F_{q,n-k}| \geq c] = \alpha$
  - H_0 is rejected if $F > c$.

Example: Test of Overall Statistical Significance

- Special case that is a test

$$\begin{array}{rl} & H_0 : \beta_2 = 0, ..., \beta_k = 0 \\ \text{against} & H_a : \text{ At least one of } \beta_2, ..., \beta_k \neq 0. \end{array}$$

- Regression programs automatically provide this in regression output.
- For house price example with $k = 7$ regressors including intercept
  - Test statistic is $F(q, n-k) = F(6, 22)$ distributed
  - $F = 6.83$ with $p = 0.0003$
  - so reject H_0 at level 0.05.
  - conclude regressors are jointly statistically significant.
- Test only says that the regressors are jointly statistically significant
  - it does not say which regressors are individually statistically significant
    - ★ in this example only *Size* was individually statistically significant at 5%.

Test of Subsets of Regressors

- Clearly variable *Size* matters
  - suppose we want to test whether the remaining regressors matter.
- The **unrestricted model** or **complete model** has all k regressors

$$y = \beta_1 + \beta_2 x_2 + \cdots \beta_g x_g + \beta_{g+1} x_{g+1} + \cdots + \beta_k x_k + \varepsilon$$

- The **restricted model** or **reduced model** has only the first g regressors

$$y = \beta_1 + \beta_2 x_2 + \cdots \beta_g x_g + \varepsilon.$$

- We test whether the last $(g - k)$ are statistically significant.

$$\begin{array}{rl} & H_0 : \beta_{g+1} = 0, ..., \beta_k = 0 \\ \text{against} & H_a : \text{ At least one of } \beta_{g+1}, ..., \beta_k \neq 0. \end{array}$$

- A specialized test command yields $F = 0.417$ with $p = 0.832 > 0.05$
  - we do not reject $H_0 : \beta_3 = 0, ..., \beta_7 = 0$ at significance level 0.05
  - the additional five regressors are jointly statistically insignificant
  - it is best to just include *Size* as a regressor.

Further Details

- For test of a single restriction $F = t^2$
  - the F test gives the same answer as a two-sided t test
  - the p value is the same
  - the critical value for F equals that for t squared
    - ★ in particular for large n the $F(1, n-k)$ critical value is $1.96^2 = 3.84$.
- Some packages report chisquared tests rather than F tests
  - in large samples with $n \rightarrow \infty$
  - q times $F(q, \infty)$ is $\chi^2(q)$ distributed (chi-squared with q degrees of freedom).
  - to get the F-statistic divide the χ^2-statistic by q.
- Separate tests of many hypotheses
  - with many separate tests there is high probability of erroneously finding a variable statistically significant
  - adjusting for multiple testing is beyond the scope of this text.

11.6 F Statistic under Assumptions 1-4

- The proceeding presentation of the F test also applies following regression with robust standard errors.
- Now specialize to default standard errors (assumptions 1-4)
  - then analysis simplifies and provides some insights.
- Intuitively, reject restrictions if the restricted model has much poorer fit.
  - Reject restrictions if $RSS_r - RSS_u$ is large where
    - RSS_r is residual sum of squares in restricted model
    - RSS_u is residual sum of squares in unrestricted model
- Under assumptions 1-4 the F statistic is a function of $RSS_r - RSS_u$.

F Statistic under Assumptions 1-4

- Under H_0 and assumptions 1-4 the **F-statistic** can be shown to be

$$F = \frac{(RSS_r - RSS_u)/q}{RSS_u/(n-k)} \sim F(q, n-k)$$

- This is a two-sided test - there is no one-sided test.
- Reject H_0 when F is large, since then restricted model fits much worse.
  - reject at level α if $p = \Pr[F_{k-1,n-k} > F]$ is $< \alpha$
    - ★ Stata: $p = \texttt{Ftail}(k-1, n-k, F)$
  - or reject at level α if F$< c = F_{k-1,n-k;\alpha}$
    - ★ Stata: $c = \texttt{invFtail}(k-1, n-k, \alpha)$.

Test Overall Statistical Significance under Assumptions 1-4

- Test $H_0 : \beta_2 = 0, ..., \beta_k = 0$ vs. H_a : At least one of $\beta_2, ..., \beta_k \neq 0$.
- The restricted model is an intercept-only model with $\widehat{y}_i = \bar{y}$
  - so $RSS_r = \sum_{i=1}^{n} (y_i - \bar{y})^2 = TSS$.
- Some algebra then shows that in this special case

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

  - where R^2 is the usual R^2 from the regression of y on all regressors.
- Example: House price regressed on all regressors
  - $R^2 = 0.6506$, $n = 29$, $k = 7$
  - $F = (.6506/6)/(.3494/22) = 6.827$.
  - $p = Ftail(6, 22, 6.827) = 0.000342$
  - reject H_0 at 5% since $p < 0.05$.

Test of Subsets of Regressors under Assumptions 1-4

- Test whether regressors other than house price are statistically significant
  - so test $H_0 : \beta_{bed} = 0, \beta_{bath} = 0..., \beta_{month} = 0$.
- Manual computation
  - Full model: $RSS_u = 13679397855$ ($k = 7$ including intercept).
  - Restricted model: $RSS_r = 14975101655$ ($g = 2$: *Size* plus intercept)
  - $F = \frac{(14975101655 - 13679397855)/5}{13679397855/22} = 0.417$.
  - $p = Ftail(5, 22, .417) = 0.832 < 0.05$
  - $c = invFtail(22, .05, 5, 22) = 2.66$
  - do not reject H_0 at level 0.05.
- The additional five regressors are not jointly statistically significant at 5%.

Relationship between F test and adjusted R-Squared

- Under assumptions 1-4
  - as regressors are added $\bar{R}^2$ increases if and only if $F > 1$
  - if a single regressor is added $\bar{R}^2$ increases if and only if $|t| > 1$.
- So including a regressor or regressors on the basis of increasing $\bar{R}^2$ is a much lower threshold than testing at 5%.

11.7 Presentation of Regression Results

- Save space by not reporting all of b, s_b, t and p.
- **1.** Report just coefficients and standard errors

$$\widehat{Price} = \underset{(21489)}{111691} + \underset{(11.17)}{73.77} \times Size + \underset{(7846)}{1553} \times Bedrooms;\; R^2 = 0.618.$$

- **2.** Report just coefficients and t statistics for $H_0 : \beta_2 = 0$

$$\widehat{Price} = \underset{(5.35)}{111691} + \underset{(6.60)}{72.41} \times Size + \underset{(0.20)}{1553} \times Bedrooms;\; R^2 = 0.618.$$

- **3.** Report just coefficients and p values for $H_0 : \beta_2 = 0$

$$\widehat{Price} = \underset{(0.000)}{111691} + \underset{(0.000)}{72.41} \times Size + \underset{(0.845)}{1553} \times Bedrooms;\; R^2 = 0.618.$$

- **4.** Report just coefficients and 95% confidence intervals.
- **5.** Report just coefficients and asterisks:
  - one if statistically significant at 10%
  - two if statistically significant at 5%
  - three if statistically significant at 1%.

11.7 Presentation of Regression Results

- Different ways to present results from the same regression
  - same coefficients but different quantities in parentheses.

In parentheses:	Results 1 St.errors	Results 2 t statistics	Results 3 p-values	Results 4 95% Conf.int.	Results 5
Size	72.41	72.41	72.41	72.41	72.41***
	(13.29)	(5.44)	(0.000)	(45.07,99.75)	
Bedrooms	1553	1553	1553	1553	1553
	(7847)	(0.20)	(0.845)	(-14576,17682)	
Intercept	11691	11691	11691	11691	11691***
	(27589)	(4.05)	(0.000)	(54981,168401)	
R^2	0.618	0.618	0.618	0.618	0.618
F(2,26)	21.93	21.93	21.93	21.93	21.93
n	29	29	29	29	29

Key Stata Commands

```
clear
use AED_HOUSE.DTA
regress price size bedrooms bathroom lotsize age
       monthsold
test size = 50
test bedrooms bathroom lotsize age monthsold
```

Some in-class Exercises

1. We obtain fitted model $\widehat{y} = \underset{(1.5)}{3.0} + \underset{(2.0)}{5.0} \times x_2 + \underset{(2.0)}{7.0} \times x_3$, $n = 200$, with standard errors given in parentheses. Provide an approximate 95% confidence interval for the population slope parameter.
2. For the preceding data is x_2 statistically significant at level 0.05?
3. For the preceding data test the claim that the coefficient of x_3 equals 10.0 at significance level 0.05.
4. Consider the model $y = \beta_1 + \beta_2 x_2 + +\beta_2 x_2 + \cdots + \beta_k x_k + u$. We wish to test the claim that the only regressors that should be included in the model are x_2 and x_3. State H_0 and H_a for this test, and give the degrees of freedom for the resultant F test.