Week 4: Simple Linear Regression II

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019

Updated notes are here: https://clas.ucdenver.edu/marcelo-perraillon/teaching/health-services-research-methods-i-hsmp-7607

Outline

- Algebraic properties of OLS
- Reminder on hypothesis testing
- The Wald test
- Examples
- Another way at looking at causal inference

Big picture

- We used the method of least squares to find the line that minimizes the sum of square errors (SSE)
- We **made NO assumptions about the distribution** of ϵ or Y
- We saw that the mean of the predicted values is the same as the mean of the observed values and that implies that **predictions regress towards the mean**
- Today, we will assume that ϵ distributes $N(0, \sigma^2)$ and are independent (iid). **We need that assumption for inference** (not to find the best β_j)

Algebraic properties of OLS

- 1) The sum of the residuals is zero: $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$. One implication of this is that if the residuals add up to zero, then we will always get the mean of Y right.
- **Confusion alert**: ϵ is the error term; $\hat{\epsilon}$ is the residual (more on this later)
- Recall the first order conditions:
  $\frac{\partial SSE}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
  $\frac{\partial SSE}{\partial \beta_1} = \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
- From the first one, it's obvious that we choose $\hat{\beta}_0, \hat{\beta}_1$ that satisfy this property: $\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} \hat{\epsilon}_i = 0$, so $\bar{\hat{\epsilon}} = 0$
- In words: On average, the residuals or predicted errors are zero, so on average the predicted outcome $\hat{y}$ is the same as the observed y. On average we don't make prediction errors

Algebraic properties of OLS

- 2) The sample covariance between the residuals and the regressor x is zero
- This follows from the second first first order condition above: $\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = \sum_{i=1}^{n} x_i \hat{\epsilon}_i = 0$
- Note that to calculate the covariance, we would only need to subtract $\bar{x}$ and $\bar{\hat{\epsilon}}$, which are constants. We just saw that $\bar{\hat{\epsilon}} = 0$
- So $cov(\hat{\epsilon}, x) = 0$

Algebraic properties of OLS

- Let's verify these properties with the college grades example

```
qui reg colgpa hsgpa
predict colhat
* Two ways of calculating residuals
gen res = colgpa - colhat
predict resgpa, r
sum resgpa

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      resgpa |        141    1.47e-10    .3388114  -.8521981   .8855062

 corr resgpa hsgpa
(obs=141)

             |   resgpa    hsgpa
-------------+------------------
      resgpa |   1.0000
       hsgpa |   0.0000   1.0000
```

Hypothesis testing reminder

- We start with a **null hypothesis** or a statement about the **population parameter** than you want to test for
- We **assume that the null hypothesis is true** until the data provides evidence that it is not; similar to the way the criminal justice system treats a defendant (innocent until proven guilty by evidence)
- We denote the null hypothesis as H_0. For example, we are interested in testing if the population parameter is equal to zero, so we write: $H_0 : \beta_1 = 0$ (note that we don't write $H_0 : \hat{\beta}_1 = 0$)
- We will **use** $\hat{\beta}_1$ and the precision of the estimate of $\hat{\beta}_1$, $\sqrt{var(\hat{\beta}_1)}$ as **evidence** (what is the "precision" called?)
- We also need an alternative hypothesis, which is usually $H_1 \neq 0$ (two-tailed test) or $H_1 > 0$ or $H_1 < 0$ (one-tailed test)

Hypothesis testing reminder

- We will either accept or reject the null hypothesis given the evidence, which doesn't necessarily make the alternative hypothesis true
- We can make **two mistakes**. We can reject the null when in fact it's true, which is analogous to the mistake of rejecting the innocence of a person on trial when in fact the person is innocent. We call this mistake **Type I error**
- The **significance level** of the test, α, is the **probability of Type I error**: $\alpha = P(\mathit{reject} H_0 | H_0 \mathit{true})$
- We want to **make this error as small as possible** and by convention it's often set at $\alpha = 5\%$, although $\alpha = 10\%$ is acceptable in some disciplines (e.g. economics)
- Think of a clinical trial. Suppose that the drug has no effect. Why do we want to make this error small?

Hypothesis testing reminder

- The other error is failing to reject the null when in fact is false; this is called the **Type II error**
- We also want to minimize this error
- As an alternative, we would like to maximize the **power of the test**, which is $1 - P(TypeII)$
- In other words, we want to make sure that we have enough evidence to reject the null
- Sample size plays a key role. The larger the sample size the more "power" we have, so we are confident that we will be able to reject the null

Hypothesis testing reminder

- We also need to choose a **test statistic** and a **critical value**
- A **test statistic**, denoted by T, is a function of the random sample; a particular number for the test statistic is denoted by t
- A more intuitive explanation for the test statistic is that it's a **standardized value** calculated **from the sample data** that is used as **evidence** to accept or reject a null hypothesis
- We are essentially interested in figuring out: How likely is it that we will find a test statistic t given that the null hypothesis is true? To answer this question given a level of α **we need to know the distribution of t under the null**

Hypothesis testing reminder

- That's how you learned to do a test in stats 101: calculate t statistic and compare it to a critical value from a table at the end of the book (you will do so for the homework). If the t statistic falls outside a rejection region, you reject the null
- With the use of statistical software came the convenience of using **p-values** instead
- **P-value**: What is the largest significance level at which we could carry out the test and still fail to reject the null hypothesis?
- Now we just look at the p-value for a statistic: if p-value ≥ 0.05, then we do not reject the null at $\alpha = 0.05$
- In economics, a p-value of < 0.1 is still quite acceptable
- **So, how do you choose a test statistics? Where do they come from?** Answer: from theory

Wald tests

- **Wald tests** come up in parametric models in which parameters are estimated to model a relationship
- Using the estimated parameters from the sample, we can perform statistical tests of the true population value of a parameter θ
- They have the form:

  $t = \frac{\hat{\theta} - \theta_0}{\sqrt{var(\hat{\theta})}}$,

  where θ_0 is the value of θ we want to test for
- Often, we want to test $\theta_0 = 0$ so the Wald test is

  $t = \frac{\hat{\theta}}{\sqrt{var(\hat{\theta})}} = \frac{\hat{\theta}}{se(\hat{\theta})}$

Intuition

- Look at the formula for the Wald test. When is the null more likely to be rejected? $t = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$
- The largest the t the more likely that the test is rejected given a significance level. So:
  1. If $\hat{\theta} - \theta_0$ is large, or alternatively, if the value of the null hypothesis is far from the estimated value from the sample
  2. If $se(\hat{\theta})$ is small; if $se(\hat{\theta})$ is large we will reject because we don't have enough solid evidence (the estimate of the true parameter is uncertain)
- $se(\hat{\theta})$ is a **function of the sample size** n. With large datasets, $se(\hat{\theta})$ will be small so we will tend to reject the null even for small $\hat{\theta} - \theta_0$
- This is a (hot) current area of research (navel gazing?) in statistics and "big data." Should we use conventional statistical inference? *Most differences are statistically significant* when you have millions of observations

Wald tests

- We need to figure out how t distributes **under the null** hypothesis
- In the linear model, it distributes t-student (in the logistic model it will distribute standard normal)
- How do we know that? This is when we **need to make assumptions about the distribution of the error**: To find the asymptotic properties of the OLS estimators, we assume that ϵ_i are iid and distribute $N(0, \sigma^2)$
- Note that we assume σ^2 and not σ_i^2; all observations have the same variance (**homoskedasticy**)
- **Technical note**: The Wald test and CIs are based on the coefficients β_j distributing normal. Yet, this is an approximation that depends on sample size and other assumptions
- **Confusion alert**: "t" is for **test statistic** in this context, which in the OLS model happens to distribute t-student

Digression and clarification

- The Wald test can also be written as $t = \frac{(\hat{\theta}-\theta_0)^2}{se(\hat{\theta})}$
- The intuition is the same. The difference is that the the test statistics distribution is different (F distribution). Without the square in the numerator, the test statistics distributes t-student
- **Remember from the review of stats class**: the ratio of two chi-squares distributions distributes F. The ratio of a normal to a Chi-square distributes t-student.
- Not the only type of hypothesis test. We will see another way of performing the same test using the **Likelihood Ratio Test** (LRT)

Digression and clarification II

- See Buse (1982) and Engle (1984) for more on Wald tests versus LRT. It may not be clear until we cover MLE so you could skip those for a while (we would need to talk about Fisher information matrix too...)
- You could see the Wald test as the distance weighted by the curvature of the MLE: $t = (\hat{\theta} - \theta_0)^2 C(\theta)$
- But this test statistics is **an approximation** that may not work well when θ and $\hat{\theta}$ are far apart or when the curvature of the log-likelihood (approximated by $C(\hat{\theta})$ is more flat (more uncertain)
- **BUT**: Wald and LRT are **asymptotically** equivalent. The general consensus is that LRTs are better with smaller sample sizes, but in modern applications we seldom use small samples anymore

Examples

- Stata displays Wald tests (and CIs) for all estimation commands

```
reg colgpa hsgpa

      Source |       SS           df       MS      Number of obs   =       141
-------------+----------------------------------   F(1, 139)       =     28.85
       Model |  3.33506006         1  3.33506006   Prob > F        =    0.0000
    Residual |  16.0710394       139  .115618988   R-squared       =    0.1719
-------------+----------------------------------   Adj R-squared   =    0.1659
       Total |  19.4060994       140  .138614996   Root MSE        =    .34003

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsgpa |   .4824346   .0898258     5.37   0.000      .304833    .6600362
       _cons |   1.415434   .3069376     4.61   0.000     .8085635    2.022304
------------------------------------------------------------------------------
```

- The $P > |t|$ column is a two-tailed Wald test for $\beta = 0$. **Please never forget that in that test the null is $H_0 : \beta = 0$**
- A story so you remember...

Examples

- The t-student distribution depends on sample size (degrees of freedom), but with larger samples it **converges to a standard normal**
- If you recall stats 101, the two-tailed rejection at $\alpha = 0.05$ is 1.96 with a standard normal
- With the t-student and 120 df it is 1.98. With 10 it's 2.23; see table A.2 of your textbook
- Of course, we do not need tables now. We use p-values but we can also use Stata

```
 di invttail(10,0.025)
2.2281389
di invttail(120,0.025)
1.9799304
di invttail(300,0.025)
1.967903
```

Tests in Stata

- Stata has a post-estimation command called "test" that is extremely flexible; you can do all sort of tests with the coefficients of a model (type "help test")

```
. reg colgpa hsgpa male
<...ouput omitted...>
------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsgpa |   .4848333    .092151     5.26   0.000     .3026229    .6670437
        male |     .00738    .058826     0.13   0.900    -.1089369    .1236969
       _cons |     1.4034   .3226199     4.35   0.000     .7654822    2.041317
------------------------------------------------------------------------------
. test male
 ( 1)  male = 0
       F(  1,   138) =    0.02
            Prob > F =    0.9003
. test male = 1
 ( 1)  male = 1
       F(  1,   138) =  284.73
            Prob > F =    0.0000
* Replicate p-value
. display ttail(139,0.13)*2
.89675437
. display ttail(139,(.00738/.058826 ))*2
.90034483
```

Tests in Stata

- As you can tell, Stata is not showing a t-student tests but rather F tests
- That's because Stata is performing a Wald test squaring the numerator
- The Wald test in outpout is an F test with 1 numerator degree of freedom and 138 denominator degrees of freedom. The Student's t distribution is directly related to the F distribution. The square of the Student's t distribution with d degrees of freedom is equivalent to the F distribution with 1 numerator degree of freedom and d denominator degrees of freedom

Tests in Stata and confusion alert

- I'm about to confuse you (again) so **pay attention**
- Although we often see the Wald test in the context of testing for the significance of one model parameter, the Wald test can be used to test the significance of a **linear combination** of parameters
- In the model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$, we could use the Wald test for $H_o : \beta_1 = \beta_2 = 0$. Stata syntax would be "test x1 x2"
- This Wald test in the linear model distributes F (in the logistic model, Chi-square)
- We will cover next class the F test of the **overall significance** of a linear model. You have seen this one before: it uses sum of squares of model versus sum of squares of the errors/residuals. It's used in ANOVA, ANCOVA
- Although you arrive at the same place, **not the same tests**. During the semester, we will use **Wald tests and LRTs** because I want to teach you an approach to modeling you can use with the linear model **AND** logit, probit, survival, GLM, and most other common models...

SRL, conditional expectation, and causal inference

- We could just take the conditional expectation of the **population model** $Y = \beta_0 + \beta_1 X + \epsilon$:
- $E[Y|X] = \beta_0 + \beta_1 X + E[\epsilon|X]$
- $E[\epsilon|x]$ is not automatically zero; we need to assume it is: (**zero conditional mean assumption**)
- It turns out,though, that the real assumption is not $E[\epsilon|x] = 0$ but rather $E[\epsilon|x] = E[\epsilon]$
- That is, the error term is **independent** of x; think of this as the error term ϵ being **uncorrelated with any function** of x (remember that correlation only measures a linear association)
- Without this assumption, we can't "hold" other factors constant because the error is correlated with the x

Example

- Suppose we want to estimate $wage = \beta_0 + \beta_1 educ + \epsilon$, where *wage* is measured in dollars per hour and *educ* is years of education
- We expect that β_1 is positive. The average wage is higher with more years of education
- But can we really hold other factors equal? What about ability/intelligence? Let's say $\epsilon = ability = a$
- $E[a|educ] = E[a]$ doesn't hold. The average expected level of ability is not the same for every level of education; alternative, ability and years of education are correlated
- Ability is a **confounder**. An economist would say that education is **endogenous** or the model is not **identified**
- What is the effect on $\hat{\beta}_1$? Probably overestimated... more on **omitted variable bias**

Confusion

- Note how confusing this is. At the beginning of the class we showed that X is uncorrelated with $\hat{\epsilon}$ and that the sum of the residuals is zero. This is **always true**
- Now I'm telling you that we need to *assume* this to interpret SLR as a conditional expectation function
- It's not that confusing when you realize that the **zero conditional mean assumption** is an assumption about the **population**, not a property of the sample
- It's an assumption about whether our estimated model can have a *ceteris paribus* or causal interpretation in the population
- This is another way of understanding no **unmeasured confounder**: $E[\epsilon|x] = E[\epsilon]$ if we didn't omit a confounder in the regression

Summary

- Several algebraic properties of OLS are direct consequences of the first order conditions
- To perform statistical tests, we need to figure out the distribution of the parameters and their standard errors and also the distribution of the test statistic **under the null**
- To derive the statistical properties of the estimated coefficients, we need to assume normality of the errors, which is equivalent to assume that the outcome distributes normal
- To interpret our estimated model as *holding other factors constant* (*ceteris paribus*), the zero conditional mean assumption must hold in the population
- Next class, more about goodness of fit