Dis 4: Hypothesis Testing; Multivariate Linear Regression

1 Hypothesis Testing

1.1 Testing one restriction: Two-tails test

- Say that we have a simple (univariate) linear regression model: $y_i = \beta_0 + \beta_1 x_i + u_i$
- In the true model, β_1 is considered as the degree of which increase in x_i impacts y_i. This means that we are often interested in whether β_1 significantly differs from 0. To formally test this, one constructs the following hypothesis test:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- However, given that we don't know the true β_1 value, one needs to rely on its estimate $\hat{\beta}_1$ to carry out the above test. But $\hat{\beta}_1$ as an estimate is not precise: estimator for $\hat{\beta}_1$ generates variation, so it wouldn't be correct to simply look at the absolute value of $\hat{\beta}_1$ and perform an "eye-ball test" per se. We thus construct this the following statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{se(\hat{\beta}_1)} \quad \sim \quad t_{n-k-1}$$

where k is the number of regressors (Xs), and n is the number of observations within the sample.

- One then needs to know how (often) we can reject the null hypothesis. To do this, **significance level (size)** of the test needs to be specified, so that we know the probability of rejecting the null hypothesis, given that the null hypothesis assumed was true. Often times, 5% is used as the significance level.

  Rejection cutoff value can be found from t-table. Notice that when the alternative hypothesis is $\neq$, this is called a **two-tails test**, which means **upper tail probability = 1/2 of significance level**:

  - If our test statistic $>$ the cutoff value, then we **reject** the null hypothesis, and conclude that $\beta_1 \neq 0$ at specified significance level. This means that β_1 is estimated to be **statistically significant**.
  - If our test statistic $<$ the cutoff value, then we **fail to reject** the null hypothesis. This means that β_1 is estimated to be **statistically insignificant**.

- Doing all this in Stata:

Source	SS	df	MS			
				Number of obs	=	706
				F(1, 704)	=	81.09
Model	14381717.2	1	14381717.2	Prob > F	=	0.0000
Residual	124858119	704	177355.282	R-squared	=	0.1033
				Adj R-squared	=	0.1020
Total	139239836	705	197503.313	Root MSE	=	421.14

sleep	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
totwrk	-.1507458	.0167403	-9.00	0.000	-.1836126	-.117879
_cons	3586.377	38.91243	92.17	0.000	3509.979	3662.775

- Notice that you can also use the $P > |t|$ column to determine whether a coefficient is statistically significant. This column records **two-tails p-value**, which is the **lowest significance level needed to reject the null hypothesis of a two-tails test**.
- Perform a two-tails t-test using the `test` command:

  `test varname`

```
. test totwrk

 ( 1)  totwrk = 0

       F(  1,   704) =   81.09
            Prob > F =    0.0000
```

- Perform a two-tails t-test against nonzero null hypothesis (say, $H_0 = 1$):

  `test varname = 1`

```
. test totwrk = 1

 ( 1)  totwrk = 1

       F(  1,   704) = 4725.36
            Prob > F =    0.0000
```

1.2 Testing one restriction: One-tail test

- Say that one's instead interested in whether β_1 is positive. The hypothesis test is then formalized to be the following:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 > 0$$

  (the test is analogous for alternative hypothesis $\beta_1 < 0$)
- Same test statistic as before:

$$t = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{se(\hat{\beta}_1)} \quad \sim \quad t_{n-k-1}$$

- The only difference in here is that the cutoff value for rejecting the null hypothesis is found by assigning significance level to only one tail of the null distribution. In English, this means that when looking at the t-table, we will set **upper tail probability = significance level**.

  (if alternative hypothesis is $\beta_1 < 0$, set lower tail probability = significance level instead; or, still find cutoff by setting upper tail probability = significance level, but the actual cutoff to use is the negative version: $-1 *$ upper-tail-cutoff)

  Again,
  - When test statistic $>$ the cutoff value, then we **reject** the null hypothesis.
    Notice that when we are doing a one-tail test, if we fail to reject the null hypothesis, it doesn't mean that $\beta_1 = 0$. In fact, we just couldn't say that $\beta_1 > 0$, so $\beta_1 \leq 0$ might be the case.

- When test statistic $<$ the cutoff value, then we **fail to reject** the null hypothesis.

- Doing this in Stata is a bit more complicated:

```
test varname
local sign_wgt = sign(_b[varname])
display "H_1: coef > 0  -> p-value = " ttail(r(df_r), 'sign_wgt' * sqrt(r(F)))
display "H_1: coef < 0  -> p-value = " 1 - ttail(r(df_r), 'sign_wgt' * sqrt(r(F)))
```

*In the above example, the quotation marks surrounding `sign_wgt` are the back quotation mark and a single quotation mark.

```
. test _cons

 ( 1)  _cons = 0

       F(  1,   704) = 8494.45
            Prob > F =    0.0000

. local sign_wgt = sign(_b[_cons])

. display "H_1: coef < 0  -> p-value = " 1 - ttail(r(df_r), `sign_wgt' * sqrt(r(F)))
H_1: coef < 0  -> p-value = 1
```

1.3 Testing multiple restrictions

- We sometimes might want to test more than one restrictions. Say that a hypothesis looks like the following:

$$H_0 : \beta_0 = \beta_1 = 0$$
$$H_1 : \beta_0 \neq 0 \text{ or } \beta_1 \neq 0$$

Notice that this provides different information from performing two separate tests of $\beta_0 = 0$ and $\beta_1 = 0$: not only does our null further includes the scenario of $\beta_0 = \beta_1$, to reject our proposed null hypothesis, only one β is needed to be nonzero (instead of each β needing to be nonzero).

- The general idea of performing such test is by imposing restriction on our regression model so that the coefficients follow the null hypothesis. With this "restricted" regression, we can compare it with the "unrestricted" version and derive test statistic.

- Turns out that our test statistic will follow a F-distribution:

$$\begin{aligned} F &= \frac{(SSR_{\text{restricted}} - SSR_{\text{unrestricted}}/q)}{SSR_{\text{unrestricted}}/(n - k_{\text{unrestricted}} - 1)} \\ &= \frac{(R^2_{\text{unrestricted}} - R^2_{\text{restricted}})/q}{(1 - R^2_{\text{unrestricted}})/(n - k_{\text{unrestricted}} - 1)} \\ &\sim F_{q,n-k-1} \end{aligned}$$

where q is the number of restrictions, and $SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of squared residuals. (In $H_0 : \beta_0 = \beta_1 = 0$, there are 2 restrictions: $\beta_0 = \beta_1$, and $\beta_0 = 0$. So $q = 2$ in this case.)

An F-table is needed to find cutoff values. Notice that there are two separate degrees of freedom here, q is df1, and $n-k-1$ is df2.

- This test is very easy to carry out in Stata:

```
test varname1 varname2 // null hypothesis is coef on varname1 = on varname2 = 0
test varname1 = varname2 = 1 // null is coef on varname 1 = on varname2 = 1
```

```
. test _cons = totwrk = 1

 ( 1)  - totwrk + _cons = 0
 ( 2)  _cons = 1

       F(  2,   704) = 4960.41
            Prob > F =    0.0000
```

2 Multivariate Linear Regression

- Extending our univariate linear regression case from last week:
  - In the univariate case, the true model states that only one variable x_i explains the variation in y_i:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

  - However, one can come up with stories about y_i determined by multiple factors. (For example, amount of sleep doesn't only depend on total hours of work, it also depends on noise level at night, mood at the end of the day, time of first meeting tomorrow, time of last meeting today, etc.)
  - The variables that also affects y_i but didn't get included in our model could cause **omitted variable bias**. This type of bias arises when
    - X is correlated with the omitted variable
    - The omitted variable is a determinant of the dependent variable Y

    when this type of bias arises, $E[u_i|x_i] \neq 0$, so we violated one key assumption of our simple linear regression model.
  - A formula that helps us get a sense of how big the omitted variable bias is

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}$$

- One natural solution is to simply add back the omitted variables to our linear model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i$$

  k here denotes the total number of regressors.

- Stronger assumptions needed for multivariate linear regression:
  1. **Zero conditional mean**: $E[u_i|x_1, x_2, \ldots, x_k] = 0$

2. **I.I.D. Data**: $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ are i.i.d. (independent and identically distributed).
3. **Large outliers are unlikely**: There doesn't exist some $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ that live in a dramatically different region. Could be measured as the fourth moment of each variable is finite (i.e. $0 < E[x_{1i}^4] < \infty$, $0 < E[x_{2i}^4] < \infty$, ..., $0 < E[x_{ki}^4] < \infty$, and $0 < E[y_i^4] < \infty$)
4. **No perfect multicollinearity**: One of the regressors cannot be a perfect linear function of the other regressors.

   Example:

   x_1 is number of students at UW.

   x_2 is number of students at UW age 21 and below.

   x_3 is number of students at UW above 21 years old.

   In this case, $x_1 = x_2 + x_3$, so one cannot include all x_1, x_2, and x_3 in the same regression.
5. ***Homoskedasticity**: $Var(u_i|x_1, x_2, \ldots, x_k) = \sigma^2$ is a constant.

- *Partial and semi-partial correlations: run the following in Stata → `pcorr Y X1 X2 X3`
- *Zero-order (i.e. bivariate) correlations: run the following in Stata → `corr Y X2`

3 Problems

1. Load the following dataset from http://fmwww.bc.edu/ec-p/data/wooldridge/sleep75.dta into Stata (don't forget to first change your working directory).

   Dataset codebook is available at http://fmwww.bc.edu/ec-p/data/wooldridge/sleep75.des

   (a) Suppose we're interested in studying how years of schooling affect hourly wage. Let's start off by runing a simple linear regression:

   $$\text{hrwage}_i = \beta_0 + \beta_1 \text{educ}_i + u_i$$

   In Stata, run the following commands:

```
use "http://fmwww.bc.edu/ec-p/data/wooldridge/sleep75.dta", clear
reg hrwage educ
```

   The output looks like the following:

```
. use "http://fmwww.bc.edu/ec-p/data/wooldridge/sleep75.dta", clear

. reg hrwage educ

      Source |       SS           df       MS      Number of obs   =       532
-------------+----------------------------------   F(1, 530)       =     38.86
       Model |  497.747784         1  497.747784   Prob > F        =    0.0000
    Residual |  6788.88284       530  12.8092129   R-squared       =    0.0683
-------------+----------------------------------   Adj R-squared   =    0.0666
       Total |  7286.63063       531  13.7224682   Root MSE        =     3.579

------------------------------------------------------------------------------
      hrwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .3594428   .0576616     6.23   0.000     .2461696    .4727161
       _cons |   .5066988   .7503211     0.68   0.500    -.9672696    1.980667
------------------------------------------------------------------------------
```

   The coefficient estimates are stored in the `Coef.` column at the bottom table. Our regression estimate is

   $$\widehat{\text{hrwage}}_i = \underset{(0.75)}{0.51} + \underset{(0.06)}{0.36}\text{educ}_i$$

   (the parenthesis under the coefficient estimates are standard errors)

   (b) What's the R^2 of this simple linear regression? What's the unit of R^2?

   (Recall that $R^2 = \frac{SSE}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$)

   Looking at the upper right corner of the regression table result, R^2 is recorded in `R-squared`. Hence, $R^2 = 0.0683$. Since we said that R^2 is a measure of fitness that lies between 0 and 1, with bigger R^2 implying better fit of the regression line to the data, the fact that R^2 is very low here suggests that the linear model doesn't do a great job fitting the data (i.e. the underlying simple linear regression model might be incorrect).

   R^2 is unitless. To see this, we can look at how R^2 is calculated:

- The numerator $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ has unit of squared of whatever unit y has
- The denominator also has the unit of squared of whatever unit y has

So taking the ratio between the two means that the unit in numerator and denominator cancels out, resulting in R^2 being unitless.

This is actually a very appealing feature of R^2: since we want R^2 to measure how well the fitted line fits our data, it shouldn't depend on how we choose the data units during our data collection process. Changing units of x or y shouldn't change R^2, making it an ideal statistic for measuring fitness of the linear model without other sorts of distractions.

(c) Is the slope coefficient statistically significant at 5% of significance level?

Looking at the regression table, the row `educ` reports t-statistic of 6.23, and two-tails p-value ($P > |t|$) to be 0.000. This means that the lowest significance level we need to reject the null hypothesis of coef on educ $= 0$ is 0.000, which means that we can obviously reject the null at an even bigger significance level of 5%. Thus, the slope coefficient is statistically significant at 5% of significance level.

(d) Does the simple linear regression suffer from omitted variable bias? If so, list some examples of omitted variables.

Yes. Omitted variable should be correlated with x (educ in this case), and should also affect value of y (hrwage here). Some examples of such omitted variables include parents' education level, family income, ability, etc.

(e) Now suppose that we want to run a multivariate linear regression instead. Looking at the variable definition, can we include both educ, age, and exper all at once?

No, educ, age, and exper cannot be included all at once. Looking at the definition of variables, we see that

$$\text{exper} = \text{age} - \text{educ} - 6$$

so these variables are perfectly colinear, and thus cannot be included all at once in our regression model.

(To be clear, you can still choose two of the three to include in your linear regression model; you just can't include all three at once.)

(f) Let our multivariate linear regression model be the following:

$$\text{hrwage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{black}_i + \beta_4 \text{male}_i + \beta_5 \text{union}_i + u_i$$

In Stata, run the following commands:

```
reg hrwage educ exper black male union
```

The output looks like the following:

```
. reg hrwage educ exper black male union

      Source |       SS           df       MS      Number of obs   =       532
-------------+----------------------------------   F(5, 526)       =     30.07
       Model |  1619.67016         5  323.934032   Prob > F        =    0.0000
    Residual |  5666.96046       526  10.7736891   R-squared       =    0.2223
-------------+----------------------------------   Adj R-squared   =    0.2149
       Total |  7286.63063       531  13.7224682   Root MSE        =    3.2823

------------------------------------------------------------------------------
      hrwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .4670271   .0600128     7.78   0.000      .349133    .5849213
       exper |   .0621648   .0131918     4.71   0.000     .0362497    .0880799
       black |   .3444818   .6539602     0.53   0.599    -.9402128    1.629176
        male |   2.620406   .2878281     9.10   0.000     2.054972     3.18584
       union |   .2231009   .3471206     0.64   0.521    -.4588121    .9050139
       _cons |  -3.590309   .9490908    -3.78   0.000    -5.454783   -1.725835
------------------------------------------------------------------------------
```

(g) What's the R^2 of the multivariate linear regression? Compare this to the R^2 from simple linear regression. What does the change in R^2 mean?
$R^2 = 0.2223$ here, which is bigger than what we had in (b). This means that by including more regressors, the linear model fits the data better. But one should still caution to say that this is the "correct" underlying model: the R^2 is still on the low end, plus, what we are modelling is always first about correlation, so turning this into a causation story requires additional steps.

(h) With $\hat{\beta}_1 = 0.467$, $se(\hat{\beta}_1) = 0.060$, $n = 532$, construct the t-test statistic by hand, and use a two-tails test to test whether $\beta_1 = 0$ under 5% significance level.
Our hypotheses are the following:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Construct t-statistic to perform the hypothesis testing:

$$t = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{se(\hat{\beta}_1)} = \frac{0.467 - 0}{0.060} = 7.783$$

Since t-statistic follows the t-distribution

$$t_{n-k-1} = t_{532-5-1} = t_{526}$$

we then need to look up the t-table for cutoff value with df (degree of freedom) = 526 and upper-tail probability 0.025 (since significance level is 0.05, and this is a two-tails test). However, the table doesn't have df = 526, but we can use df = 100 just as an approximation.
In this case, cutoff value = 1.984, and our t-statistic (7.783) is certainly greater than this cutoff, so we reject the null hypothesis and conclude that $\beta_1 \neq 0$ under 5% significance level.

(i) Is the coefficient on union statistically significant at 10% significance level?
Looking at the two-tails p-value from the regression table (`union` row, $P > |t|$ column), the p-value is 0.521. Again, p-value reports the lowest significance level to reject null hypothesis of coefficient on union = 0, meaning that the lowest significance level we need for rejection is

52.1%. Hence, we cannot conclude that the coefficient on union statistically significant at 10% significance level.

(j) Interpret the coefficient on black variable.

This is going to differ slightly from our interpretation of continuous variables. black is a variable that only takes 0 and 1 as its value (this type of variable is called a **dummy variable**), so it wouldn't make sense to say something like "increase value of black by 1 unit leads to 0.344 dollars increase in hourly wage prediction". We can instead interpret this as the black variable being "turned on" from 0 to 1, so its coefficient can be interpreted as "being black is associated with 0.344 dollars increase in hourly wage prediction".

(k) Construct the 90% confidence interval on coefficient for exper.

This can be achieved both by hand and using Stata.

By hand: Assume that coef on exper follows a normal distribution, then the confidence interval would look like

$$[\hat{\beta}_2 - z_{0.05} \times se(\hat{\beta}_2), \quad \hat{\beta}_2 + z_{0.05} \times se(\hat{\beta}_2)]$$

where $z_{0.05}$ is associated with upper tail of standard normal distribution of size 0.05 (with both an upper tail and lower tail of 0.05 size are, the main area covered by the confidence interval is the middle 0.9 of the standard normal distribution).

Z score of 0.05 is 1.645, so our confidence interval is constructed as

$$[0.062 - 1.645 \times 0.013, \quad 0.062 + 1.645 \times 0.013] \quad = \quad [0.041, 0.083]$$

Using Stata: Run the following command in Stata:

```
reg hrwage educ exper black male union, level(90)
```

The output looks like the following:

```
. reg hrwage educ exper black male union, level(90)
```

Source	SS	df	MS		
Model	1619.67016	5	323.934032	Number of obs =	532
Residual	5666.96046	526	10.7736891	F(5, 526) =	30.07
Total	7286.63063	531	13.7224682	Prob > F =	0.0000
				R-squared =	0.2223
				Adj R-squared =	0.2149
				Root MSE =	3.2823

hrwage	Coef.	Std. Err.	t	P>|t|	[90% Conf.	Interval]
educ	.4670271	.0600128	7.78	0.000	.3681407	.5659136
exper	.0621648	.0131918	4.71	0.000	.0404279	.0839017
black	.3444818	.6539602	0.53	0.599	-.7330849	1.422048
male	2.620406	.2878281	9.10	0.000	2.146136	3.094676
union	.2231009	.3471206	0.64	0.521	-.3488691	.7950709
_cons	-3.590309	.9490908	-3.78	0.000	-5.154179	-2.026439

Looking at the confidence interval reported in the table, this is very close to what we constructed by hand.

(l) Test whether the coefficient on black = the coefficient on union = 0.5

Running the following command in Stata:

```
test black = union = .5
```

The output looks like the following:

```
. test black = union = .5

 ( 1)  black - union = 0
 ( 2)  black = .5

       F(  2,   526) =    0.35
            Prob > F =    0.7081
```

Looking at the p-value reported (`Prob > F`), since $0.05 < 0.7081$, this means that we cannot reject the null hypothesis of $\beta_3 = \beta_5 = 0$ at 5% significance level.