OLS multivariate regression.

Clement de Chaisemartin, UCSB

Banks have more than 1 variable to predict amount applicants will fail to reimburse

- To predict the amount applicants will fail to reimburse, banks can use their FICO score (score based on their current debts and on their history of loan repayments), and all other variables contained in their application: e.g. their income.
- Will bank be able to make better predictions by using both variables rather than just FICO score?

Yes, provided people with ≠ incomes but same FICO fail to reimburse ≠ amounts on their loan.

- Assume FICO score can take only two values: 0 and 100.
- Assume applicants' income can take two values: 2000 and 4000.
- If average amount people fail to reimburse varies with FICO and income as in table below, adding applicant's income to model improves prediction.

	Income=2000	Income=4000
FICO=0	2000	1000
FICO=100	500	200

- People with different income levels but with same FICO score fail to reimburse different amounts on their loan => adding income to your prediction model will improve quality of your predictions.

Gmail has more than 1 variable to predict whether an email is a spam.

- To predict whether email is spam, Gmail can use variable equal to 1 if “free” appears in email, and variable equal to 1 if “buy” in email.
- If percentage of spams varies as in table below, adding the “buy” variable to the model will improve prediction.

% of spams	Email has “buy” in it	Email doesn’t have “buy” in it
Email has “free” in it	3%	1.5%
Email doesn’t have “free” in it	1%	0.5%

- Emails which have “buy” and “free” in it are more likely to be spams than emails which only have “free” in it.
- Emails which have “buy” but not “free” in it are more likely to be spams than emails which neither have “buy” or “free” in it.
- => adding “buy” variable will improve predictions.

Multivariate regression

- In these lectures, we are going to discuss OLS multivariate regressions, which are OLS regressions with several **independent variables** to predict a **dependent variable**.

Roadmap

1. The OLS multivariate regression function.
2. Estimating the OLS multivariate regression function.
3. Advantages and pitfalls of multivariate regressions.
4. Interpreting coefficients in multivariate OLS regressions.

Set up and notation.

- We consider a population of N units.
  - N could be number of people who apply for a loan in bank A during May 2017.
- Each unit k has a variable y_k attached to it that we do not observe:
  - In loan example, y_k is variable equal to the amount applicant k will fail to reimburse on her loan when her loan expires in May 2018.
- Each unit k also has J variables x_{1k}, x_{2k}, x_{3k},…, x_{Jk} attached to it that we do observe:
  - In the loan example, x_{1k} could be FICO score of applicant k, x_{2k} could be the income of that applicant, etc.

Prediction = function of x_{1k}, x_{2k}, x_{3k},…, x_{Jk}.

- Based on value of x_{1k}, x_{2k}, x_{3k},…, x_{Jk} of each unit, we want to predict her y_k.
- E.g.: in the loan example, we want to predict the amount that unit k will fail to repay based on her FICO score and her income.
- Prediction should be function of x_{1k},…, x_{Jk}, $f(x_{1k},\ldots, x_{Jk})$.
- **In these lectures, we focus on predictions which are affine function of $\boldsymbol{x_{1k}}$,…, $\boldsymbol{x_{Jk}}$: $\boldsymbol{f(x_{1k},\ldots, x_{Jk}) = c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}}$, for $\boldsymbol{J+1}$ real numbers $\boldsymbol{c_0}$, $\boldsymbol{c_1}$,…, $\boldsymbol{c_J}$.**

Prediction error: $y_k - \left(c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}\right)$

- Based on value of x_{1k},…, x_{Jk} of each unit, we predict her y_k.
- Our prediction should be function of x_{1k},…, x_{Jk}, $f(x_{1k}$,…, $x_{Jk})$.
- We focus on affine functions of x_{1k},…, x_{Jk}: $f\left(x_{1k},\ldots, x_{Jk}\right) = c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}$, for $J + 1$ real numbers c_0, c_1,…, c_J.
- $y_k - \left(c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}\right)$, difference between prediction for y_k and actual value of y_k, is prediction error.
- Large positive or neg. $y_k - \left(c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}\right)$ mean bad prediction.
- $y_k - \left(c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}\right)$ close to 0 means good prediction.

Goal: find the value of $(c_0, c_1, \ldots, c_J)$ that minimizes $\sum_{k=1}^{N}\left(y_k - \left(c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}\right)\right)^2$

- $\sum_{k=1}^{N}\left(y_k - \left(c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}\right)\right)^2$ is positive. => minimizing it = making it as close to 0 as possible.
- If $\sum_{k=1}^{N}\left(y_k - \left(c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}\right)\right)^2$ is as close to 0 as possible, means that the sum of the squared value of our prediction errors is as small as possible.
- => we make small errors. That’s good, that’s what we want!

The OLS multivariate regression function

- Let

$(\gamma_0, \gamma_1, \dots, \gamma_J) = argmin_{(c_0,\ c_1, \dots,\ c_J)} \sum_{k=1}^{N} \left(y_k - (c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk})\right)^2$

- $(\gamma_0, \gamma_1, \dots, \gamma_J)$: value of $(c_0$, c_1,..., $c_J)$ minimizing

$\sum_{k=1}^{N} \left(y_k - (c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk})\right)^2$.

- We call $\gamma_0 + \gamma_1 x_{1k} + \cdots + \gamma_J x_{Jk}$ the OLS multivariate regression function of y_k on a constant, x_{1k}, x_{2k}, x_{3k},..., x_{Jk}.
- We let $\tilde{y}_k = \gamma_0 + \gamma_1 x_{1k} + \cdots + \gamma_J x_{Jk}$ denote prediction from multivariate OLS regression.
- We let $e_k = \ y_k - \tilde{y}_k$: prediction error.
- We have $y_k = \tilde{y}_k + e_k$.

How can we find $(\gamma_0, \gamma_1, \dots, \gamma_J)$?

- $(\gamma_0, \gamma_1, \dots, \gamma_J)$: value of $(c_0$, c_1,…, $c_J)$ minimizing

$\sum_{k=1}^{N}\left(y_k - (c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk})\right)^2$.

- To minimize a function of several variables, we differentiate it wrt to each of those variables, and we find the value of $(c_0$, c_1,…, $c_J)$ for which all those derivatives are equal to 0. No need to worry about second derivatives because objective function convex.

- What is derivative of $\sum_{k=1}^{N}\left(y_k - (c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk})\right)^2$ with respect to c_0? Discuss this question with your neighbor for 2mns.

iClicker time

- What is the derivative of $\sum_{k=1}^{N} \left(y_k - (c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk})\right)^2$ with respect to c_0?

a) $2 \sum_{k=1}^{N} \left(y_k - (c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk})\right)$

b) $\sum_{k=1}^{N} -2 \left(y_k - (c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk})\right)$

c) $-\sum_{k=1}^{N} \left(y_k - (c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk})\right)$

$$\sum_{k=1}^{N} -2\left(y_k - \left(c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}\right)\right)$$

- Derivative of $\sum_{k=1}^{N}\left(y_k - \left(c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}\right)\right)^2$ with respect to c_0 is

$\sum_{k=1}^{N} -2\left(y_k - \left(c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}\right)\right)$: P4Sum+chain rule.

- What is the derivative of $\sum_{k=1}^{N}\left(y_k - \left(c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}\right)\right)^2$ with respect to c_1? Discuss this question with your neighbor for 2mns.

iClicker time

- What is the derivative of $\sum_{k=1}^{N}\left(y_k - (c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk})\right)^2$ with respect to c_1?

a) $\sum_{k=1}^{N} -2\left(y_k - (c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk})\right)$

b) $-2\sum_{k=1}^{N} x_{1k}\left(y_k - (c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk})\right)$

$$\sum_{k=1}^{N} -2x_{1k}\left(y_k - (c_0 + c_1x_{1k} + \cdots + c_Jx_{Jk})\right)$$

- Derivative of $\sum_{k=1}^{N}\left(y_k - (c_0 + c_1x_{1k} + \cdots + c_Jx_{Jk})\right)^2$ wrt c_1:

$-2\sum_{k=1}^{N} x_{1k}\left(y_k - (c_0 + c_1x_{1k} + \cdots + c_Jx_{Jk})\right)$: P4Sum+chain rule.

- Derivative of $\sum_{k=1}^{N}\left(y_k - (c_0 + c_1x_{1k} + \cdots + c_Jx_{Jk})\right)^2$ wrt c_2:

$-2\sum_{k=1}^{N} x_{2k}\left(y_k - (c_0 + c_1x_{1k} + \cdots + c_Jx_{Jk})\right)$.

- …

- Derivative of $\sum_{k=1}^{N}\left(y_k - (c_0 + c_1x_{1k} + \cdots + c_Jx_{Jk})\right)^2$ wrt c_J:

$-2\sum_{k=1}^{N} x_{Jk}\left(y_k - (c_0 + c_1x_{1k} + \cdots + c_Jx_{Jk})\right)$.

$(\gamma_0, \gamma_1, \dots, \gamma_J)$ is the solution of a system of $J + 1$ equations with $J + 1$ unknowns.

- $(\gamma_0, \gamma_1, \dots, \gamma_J)$: value of $(c_0$, c_1,…, $c_J)$ for which all those derivatives are equal to 0.
- Thus, we have:

$$-2\sum_{k=1}^{N}\left(y_k - (\gamma_0 + \gamma_1 x_{1k} + \cdots + \gamma_J x_{Jk})\right) = 0$$

$$-2\sum_{k=1}^{N} x_{1k}\left(y_k - (\gamma_0 + \gamma_1 x_{1k} + \cdots + \gamma_J x_{Jk})\right) = 0$$

…

$$-2\sum_{k=1}^{N} x_{Jk}\left(y_k - (\gamma_0 + \gamma_1 x_{1k} + \cdots + \gamma_J x_{Jk})\right) = 0$$

- $(\gamma_0, \gamma_1, \dots, \gamma_J)$ is the solution of a system of $J + 1$ equations with $J + 1$ unknowns.
- If we give the values of the y_ks, of the x_{1k}s, …, of the x_{Jk}s to a computer, can solve this system and give us value of $(\gamma_0, \gamma_1, \dots, \gamma_J)$.

What you need to remember

- Population of N units. Each unit has J+1 variables attached to it: y_k is a variable we do not observe, x_{1k}, x_{2k}, x_{3k},…, x_{Jk} are variables we observe. We want to predict y_k based on x_{1k}, x_{2k}, x_{3k},…, x_{Jk}.
- E.g.: bank wants to predict amount an applicant will fail to reimburse on her loan based on her FICO score and her income.
- Our prediction should be function of x_{1k},…, x_{Jk}, $f(x_{1k},\dots, x_{Jk})$. Affine functions of x_{1k}, x_{2k}, x_{3k},…, x_{Jk}: $c_0 + c_1x_{1k} + \cdots + c_Jx_{Jk}$, for some real numbers $(c_0, c_1,\dots, c_J)$.
- Good prediction should be such that $e_k = y_k - \left(c_0 + c_1x_{1k} + \cdots + c_Jx_{Jk}\right)$, our prediction error, is as small as possible for most units.
- Best $(c_0, c_1,\dots, c_J)$: minimizes $\sum_{k=1}^{N}\left(y_k - \left(c_0 + c_1x_{1k} + \cdots + c_Jx_{Jk}\right)\right)^2$.
- We call that value $\left(\gamma_0, \gamma_1, \dots, \gamma_J\right)$. $\gamma_0 + \gamma_1x_{1k} + \cdots + \gamma_Jx_{Jk}$ is OLS regression function of y_k on a constant, x_{1k}, x_{2k}, x_{3k},…, x_{Jk}.
- $\left(\gamma_0, \gamma_1, \dots, \gamma_J\right)$: solution of system of J+1 equations with J+1 unknowns: derivatives of $\sum_{k=1}^{N}\left(y_k - \left(c_0 + c_1x_{1k} + \cdots + c_Jx_{Jk}\right)\right)^2$ wrt to c_0, c_1,…, c_J must = 0 at $\left(\gamma_0, \gamma_1, \dots, \gamma_J\right)$.

Roadmap

1. The OLS multivariate regression function.
2. Estimating the OLS multivariate regression function.
3. Advantages and pitfalls of multivariate regressions.
4. Interpreting coefficients in multivariate OLS regressions.

We cannot compute $\left(\gamma_0, \gamma_1, \dots, \gamma_J\right)$

- Our prediction for y_k based on a multivariate regression is $\gamma_0 + \gamma_1 x_{1k} + \cdots + \gamma_J x_{Jk}$, the OLS multivariate regression function.
- => to be able to make a prediction for a unit's y_k based on her x_{1k}, x_{2k}, x_{3k},…, x_{Jk}, we need to know the value of $\left(\gamma_0, \gamma_1, \dots, \gamma_J\right)$.
- Under the assumptions we have made so far, we cannot compute $\left(\gamma_0, \gamma_1, \dots, \gamma_J\right)$. Solution of

$$-2\sum_{k=1}^{N}\left(y_k - \left(\gamma_0 + \gamma_1 x_{1k} + \cdots + \gamma_J x_{Jk}\right)\right) = 0$$

$$-2\sum_{k=1}^{N} x_{1k}\left(y_k - \left(\gamma_0 + \gamma_1 x_{1k} + \cdots + \gamma_J x_{Jk}\right)\right) = 0$$

…

$$-2\sum_{k=1}^{N} x_{Jk}\left(y_k - \left(\gamma_0 + \gamma_1 x_{1k} + \cdots + \gamma_J x_{Jk}\right)\right) = 0$$

- To solve this system, we need to know the y_ks, which we don't!
- E.g.: the bank knows the FICO score and income (x_{1k} and x_{2k}) for each applicant, but does not know the amount each applicant will fail to reimburse in April 2018 when loan expires (y_k).

A method to estimate $\left(\gamma_0, \gamma_1, \dots, \gamma_J\right)$

- We draw n units from the population, and we measure the dependent and the independent variable of those n units.
- For i included between 1 and n, Y_i, X_{1i},…, X_{Ji} = value of dependent and independent variables of ith unit we randomly select.
- $\left(\gamma_0, \gamma_1, \dots, \gamma_J\right)$: value of $(c_0$, c_1,…, $c_J)$ minimizing

$$\sum_{k=1}^{N}\left(y_k - \left(c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}\right)\right)^2$$

- => to estimate $\left(\gamma_0, \gamma_1, \dots, \gamma_J\right)$, we use $(c_0$, c_1,…, $c_J)$ minimizing $\sum_{i=1}^{n}\left(Y_i - \left(c_0 + c_1 X_{1i} + \cdots + c_J X_{Ji}\right)\right)^2$.
- $\left(\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_J\right)$ denotes that value.
- Instead of $(c_0$, c_1,…, $c_J)$ minimizing sum of squared errors in population, use $(c_0$, c_1,…, $c_J)$ minimizing sum of squared errors in sample.
- If we find a good prediction function in sample, should also work well in entire population: sample representative of population.

The OLS regression function in the sample.

- Let

$$\left(\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_J\right) = argmin_{(c_0,\ c_1, \dots,\ c_J) \in R^{J+1}} \sum_{i=1}^{n} \left(Y_i - \left(c_0 + c_1 X_{1i} + \dots + c_J X_{Ji}\right)\right)^2$$

- We call $\hat{\gamma}_0 + \hat{\gamma}_1 X_{1i} + \dots + \hat{\gamma}_J X_{Ji}$ the OLS regression function of Y_i on a constant, X_{1i},…, and X_{Ji} **in the sample.**
- $\left(\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_J\right)$: coefficients of the constant, X_{1i},…, and X_{Ji}.
- Let $\hat{Y}_i = \hat{\gamma}_0 + \hat{\gamma}_1 X_{1i} + \dots + \hat{\gamma}_J X_{Ji}$. $\hat{Y}_i$ is the predicted value for Y_i according to the OLS regression function of Y_i on a constant, X_{1i},…, and X_{Ji} **in the sample.**
- Let $\hat{e}_i = Y_i - \hat{Y}_i$. $\hat{e}_i$: error we make when we use OLS regression **in the sample** to predict Y_i.
- We have $Y_i = \hat{Y}_i + \hat{e}_i$.

How can we find $\left(\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_J\right)$?

- $\left(\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_J\right)$: value of $(c_0$, c_1,…, $c_J)$ minimizing

$\sum_{i=1}^{n}\left(Y_i - \left(c_0 + c_1 X_{1i} + \cdots + c_J X_{Ji}\right)\right)^2$.

- To minimize a function of several variables, we differentiate it wrt to each of those variables, and we find the value of $(c_0$, c_1,…, $c_J)$ for which all those derivatives are equal to 0. No need to worry about second derivatives because objective function convex.

The derivatives of the objective function

- Derivative of $\sum_{i=1}^{n}\left(Y_i - \left(c_0 + c_1 X_{1i} + \cdots + c_J X_{Ji}\right)\right)^2$ wrt c_0:

$-2\sum_{i=1}^{n}\left(Y_i - \left(c_0 + c_1 X_{1i} + \cdots + c_J X_{Ji}\right)\right)$:P4Sum+chain rule+ P2Sum

- Derivative of $\sum_{i=1}^{n}\left(Y_i - \left(c_0 + c_1 X_{1i} + \cdots + c_J X_{Ji}\right)\right)^2$ wrt c_1:

$-2\sum_{i=1}^{n} X_{1i}\left(Y_i - \left(c_0 + c_1 X_{1i} + \cdots + c_J X_{Ji}\right)\right)$

- …

- Derivative of $\sum_{i=1}^{n}\left(Y_i - \left(c_0 + c_1 X_{1i} + \cdots + c_J X_{Ji}\right)\right)^2$ wrt c_J:

$-2\sum_{i=1}^{n} X_{Ji}\left(Y_i - \left(c_0 + c_1 X_{1i} + \cdots + c_J X_{Ji}\right)\right)$

$\left(\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_J\right)$= solution of system of $J+1$ equations with $J+1$ unknowns.

- $\left(\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_J\right)$: value of $(c_0$, c_1,…, $c_J)$ for which all derivatives = 0.

$$-2\sum_{i=1}^{n}\left(Y_i-\left(\hat{\gamma}_0+\hat{\gamma}_1 X_{1i}+\cdots+\hat{\gamma}_J X_{Ji}\right)\right) \quad = 0$$

$$-2\sum_{i=1}^{n} X_{1i}\left(Y_i-\left(\hat{\gamma}_0+\hat{\gamma}_1 X_{1i}+\cdots+\hat{\gamma}_J X_{Ji}\right)\right) \quad = 0$$

…

$$-2\sum_{i=1}^{n} X_{Ji}\left(Y_i-\left(\hat{\gamma}_0+\hat{\gamma}_1 X_{1i}+\cdots+\hat{\gamma}_J X_{Ji}\right)\right) \quad = 0$$

- To compute $\left(\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_J\right)$, we:
  - draw n units from population, measure their Y_is and $(X_{1i}, \dots, X_{Ji})$s
  - set up above system plugging in actual values of the Y_is and $(X_{1i}, \dots, X_{Ji})$s
  - Yields system of $J+1$ equations with $J+1$ unknowns, the $\left(\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_J\right)$s: all the remaining quantities are real numbers.
  - Ask a computer to solve that system.

The ls command in E-views solves for you that system of $J+1$ equations with $J+1$ unknowns.

- We have:

$$-2\sum_{i=1}^{n}\left(Y_i-\left(\hat{\gamma}_0+\hat{\gamma}_1X_{1i}+\cdots+\hat{\gamma}_JX_{Ji}\right)\right)\quad=0$$

$$-2\sum_{i=1}^{n}X_{1i}\left(Y_i-\left(\hat{\gamma}_0+\hat{\gamma}_1X_{1i}+\cdots+\hat{\gamma}_JX_{Ji}\right)\right)\quad=0$$

…

$$-2\sum_{i=1}^{n}X_{Ji}\left(Y_i-\left(\hat{\gamma}_0+\hat{\gamma}_1X_{1i}+\cdots+\hat{\gamma}_JX_{Ji}\right)\right)\quad=0$$

- To compute $\left(\hat{\gamma}_0,\hat{\gamma}_1,\ldots,\hat{\gamma}_J\right)$, we:
  - draw n units from population, measure their Y_is and $(X_{1i},\ldots,X_{Ji})$s
  - set up system plugging in values of Y_is and $(X_{1i},\ldots,X_{Ji})$s
  - Ask a computer to solve that system.
- That is what the “ls” command in eviews does, where the Y_is are the values of the first variable after “ls” command, and the $(X_{1i},\ldots,X_{Ji})$s are the values of the variables after “c”.

Doing E-views' job once in our life.

- Gmail example. Assume we sample 4 emails ($n = 4$).
- For each, we measure Y_i: whether spam, X_{1i}: whether has word "free" in it, and X_{2i}: whether has word "buy" in it.

Email	Y_i	X_{1i}	X_{2i}
1	1	1	1
2	0	1	0
3	1	1	0
4	0	0	0

- E.g.: 1st email we sample is a spam, and has words "free" and "buy" in it. 2nd email is not a spam, and has "free" in it but not "buy", etc.
- If you regress Y_i on constant, X_{1i}, and X_{2i}, what is value of $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2)$? Hint: you need to write system of 3 equations and three unknowns solved by $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2)$, plug-in values of Y_i, X_{1i}, and X_{2i} given in table into system, and then solve system. You have 4 minutes to find answer.

iClicker time

Email	Y_i	X_{1i}	X_{2i}
1	1	1	1
2	0	1	0
3	1	1	0
4	0	0	0

- If you regress Y_i on a constant, X_{1i}, and X_{2i}, what will be the value of $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2)$?

$a)\ \hat{\gamma}_0 = 0, \hat{\gamma}_1 = 0.5, \hat{\gamma}_2 = 0.5.$

$b)\ \hat{\gamma}_0 = 0.5, \hat{\gamma}_1 = 1, \hat{\gamma}_2 = 0.$

$c)\ \hat{\gamma}_0 = -1, \hat{\gamma}_1 = 0, \hat{\gamma}_2 = 0.$

$\hat{\gamma}_0 = 0, \hat{\gamma}_1 = 0.5, \hat{\gamma}_2 = 0.5.$

- $n = 4$ and $J = 2$, so we have (we can forget the -2):

$$\sum_{i=1}^{4}\left(Y_i - (\hat{\gamma}_0 + \hat{\gamma}_1 X_{1i} + \hat{\gamma}_2 X_{2i})\right) \quad = 0$$

$$\sum_{i=1}^{4} X_{1i}\left(Y_i - (\hat{\gamma}_0 + \hat{\gamma}_1 X_{1i} + \hat{\gamma}_2 X_{2i})\right) \quad = 0$$

$$\sum_{i=1}^{4} X_{2i}\left(Y_i - (\hat{\gamma}_0 + \hat{\gamma}_1 X_{1i} + \hat{\gamma}_2 X_{2i})\right) \quad = 0$$

- Plugging in values of Y_i, X_{1i}, and X_{2i} in table, yields:

$$2 - 4\hat{\gamma}_0 - 3\hat{\gamma}_1 - \hat{\gamma}_2 = 0$$
$$2 - 3\hat{\gamma}_0 - 3\hat{\gamma}_1 - \hat{\gamma}_2 = 0$$
$$1 - \hat{\gamma}_0 - \hat{\gamma}_1 - \hat{\gamma}_2 = 0$$

- Subtracting equation 1 from equation 2 yields $\hat{\gamma}_0 = 0$.
- Plugging $\hat{\gamma}_0 = 0$ yields system of 2 equations & 2 unknowns:

$$2 - 3\hat{\gamma}_1 - \hat{\gamma}_2 = 0$$
$$1 - \hat{\gamma}_1 - \hat{\gamma}_2 = 0$$

- Subtracting equation 2 from 1 yields

$1 - 2\hat{\gamma}_1 = 0$, which is equivalent to $\hat{\gamma}_1 = 0.5$.

- Plugging $\hat{\gamma}_1 = 0.5$ into $1 - \hat{\gamma}_1 - \hat{\gamma}_2 = 0$ yields $\hat{\gamma}_2 = 0.5$.

$\left(\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_J\right)$ converges towards $\left(\gamma_0, \gamma_1, \ldots, \gamma_J\right)$

- One can show that as for univariate regressions, the estimators of multivariate regression coefficients converge towards the true multivariate regression coefficients (those for the full population).
- $\lim_{n \to +\infty} \left(\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_J\right) = \left(\gamma_0, \gamma_1, \ldots, \gamma_J\right)$.
- Intuition: when the sample size becomes large, the sample becomes similar to the population.

Using central limit theorem for the $\hat{\gamma}_j$s to construct tests and confidence intervals.

- Formula of the variance of multivariate regression coefficients complicated. No need to know it, E-views computes it for you.
- If $n \geq 100$, $\frac{\hat{\gamma}_j - \gamma_j}{\sqrt{V(\hat{\gamma}_j)}}$ follows normal distrib. mean 0 and variance 1.
- We can use this to test null hypothesis. Often, want to test $\gamma_j = 0$.
- If we want to have 5% chances of wrongly rejecting $\gamma_j = 0$, test is:

Reject $\gamma_j = 0$ if $\frac{\hat{\gamma}_j}{\sqrt{V(\hat{\gamma}_j)}} > 1.96$ or $\frac{\hat{\gamma}_j}{\sqrt{V(\hat{\gamma}_j)}} < -1.96$.

Otherwise, do not reject $\gamma_j = 0$.

- We can also construct a 95% confidence interval for γ_j:

$$\left[\hat{\gamma}_j - 1.96\sqrt{V(\hat{\gamma}_j)}, \hat{\gamma}_j + 1.96\sqrt{V(\hat{\gamma}_j)}\right].$$

Assessing quality of our predictions: the R^2.

- To assess the quality of our predictions, we are going to use the same measure as with the OLS affine regression:

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{e_i}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- 1 – MSE / sample variance of the Y_is.
- As in the previous lectures, we have that R^2 is included between 0 and 1.

What you need to remember

- Prediction for y_k based on multivariate regression is $\gamma_0 + \gamma_1 x_{1k} + \cdots + \gamma_J x_{Jk}$, with $\left(\gamma_0, \gamma_1, \ldots, \gamma_J\right)$: value of $(c_0, c_1,\ldots, c_J)$ minimizing $\sum_{k=1}^{N}\left(y_k - \left(c_0 + c_1 x_{1k} + \cdots + c_J x_{Jk}\right)\right)^2$.
- We can estimate $\left(\gamma_0, \gamma_1, \ldots, \gamma_J\right)$, if we measure y_ks for random sample of population.
- For every i between 1 and n, Y_i, X_{1i},…, X_{Ji} = value of dependent and independent variables of ith unit we randomly select.
- To estimate $\left(\gamma_0, \gamma_1, \ldots, \gamma_J\right)$, find $(c_0, c_1,\ldots, c_J)$ minimizing

$\sum_{i=1}^{n}\left(Y_i - \left(c_0 + c_1 X_{1i} + \cdots + c_J X_{Ji}\right)\right)^2$.

- Differentiating this function wrt to $(c_0, c_1,\ldots, c_J)$ yields system of J+1 equations with J+1 unknowns.
- We solved system in simple example, you should know how to do that.
- We used the central limit theorem to propose 5% level test of $\gamma_j = 0$, and to derive a 95% confidence interval for γ_j.

Roadmap

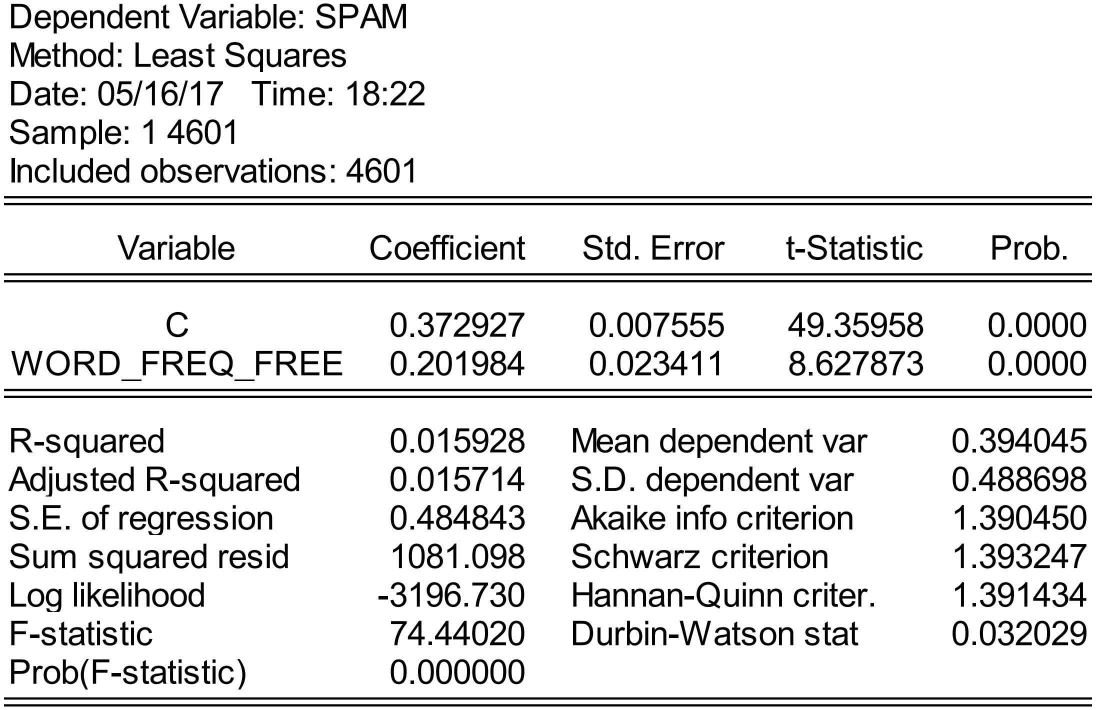
1. The OLS multivariate regression function.
2. Estimating the OLS multivariate regression function.
3. Advantages and pitfalls of multivariate regressions.
4. Interpreting coefficients in multivariate OLS regressions.

Adding variables to a regression always improves the R^2

- Assume you regress a variable Y_i on a constant and on a variable X_{1i}.
- Then, you regress Y_i on a constant and on two variables X_{1i} and X_{2i}.
- The R^2 of your second regression will be at least as high as the R^2 of the first regression.
- Adding variables to a regression always increases its R^2.
- => a regression with many variables gives better predictions for the Y_is in the sample than a regression with few variables.

Example

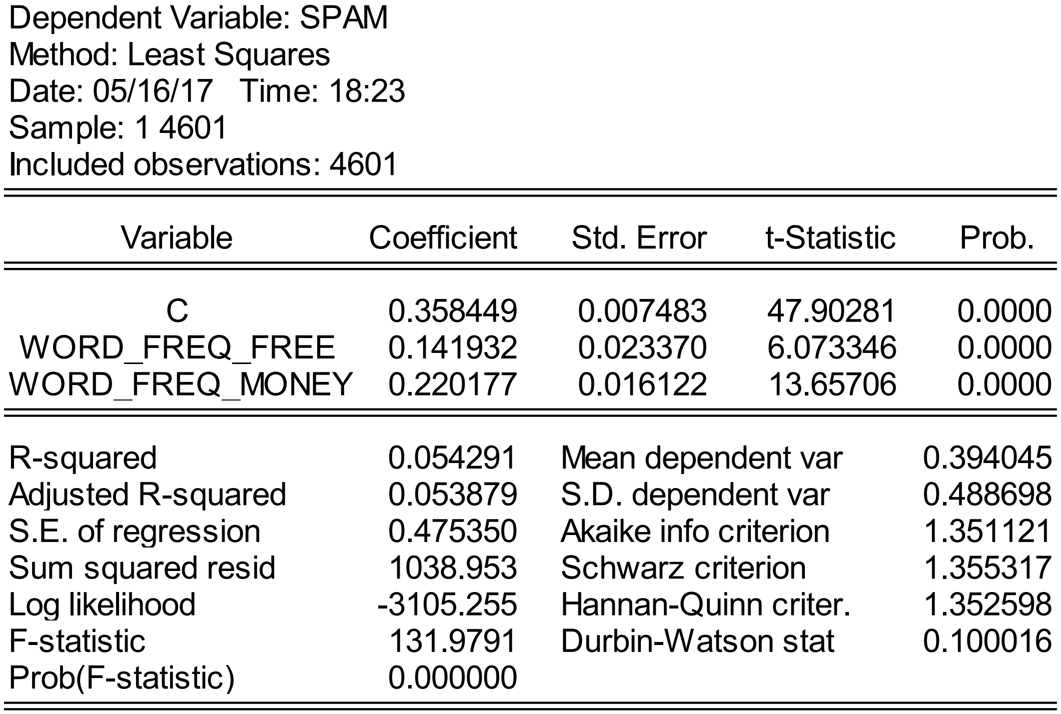
- Sample of 4601 emails, for which you observe whether they are a spam or not.
- You regress spam on constant and variable equal to the percentage of the words of the email that are the word “free”.
- Eviews command: ls spam c word_freq_free.
- Is the R^2 of that regression low or high?



```
Dependent Variable: SPAM
Method: Least Squares
Date: 05/16/17   Time: 18:22
Sample: 1 4601
Included observations: 4601
====================================================================
      Variable         Coefficient    Std. Error    t-Statistic    Prob.
====================================================================
         C              0.372927      0.007555      49.35958     0.0000
 WORD_FREQ_FREE         0.201984      0.023411      8.627873     0.0000
====================================================================
R-squared               0.015928    Mean dependent var        0.394045
Adjusted R-squared      0.015714    S.D. dependent var        0.488698
S.E. of regression      0.484843    Akaike info criterion     1.390450
Sum squared resid       1081.098    Schwarz criterion         1.393247
Log likelihood         -3196.730    Hannan-Quinn criter.      1.391434
F-statistic             74.44020    Durbin-Watson stat        0.032029
Prob(F-statistic)       0.000000
====================================================================
```

Example

- You regress spam variable on a constant, a variable equal to % of words of email that are the word “free”, and a variable equal to % of words of the email that are the word money.
- Eviews command: ls spam c word_freq_free word_freq_money.
- R^2 higher in that regression than in previous one. R^2= 1- average of square prediction errors / variance of the spam variable. => higher R^2 means lower sum of square prediction errors => better predictions.



```
Dependent Variable: SPAM
Method: Least Squares
Date: 05/16/17   Time: 18:23
Sample: 1 4601
Included observations: 4601
==========================================================================
        Variable          Coefficient    Std. Error    t-Statistic    Prob.
==========================================================================
           C               0.358449      0.007483      47.90281     0.0000
    WORD_FREQ_FREE         0.141932      0.023370      6.073346     0.0000
    WORD_FREQ_MONEY        0.220177      0.016122      13.65706     0.0000
==========================================================================
R-squared                  0.054291    Mean dependent var         0.394045
Adjusted R-squared         0.053879    S.D. dependent var         0.488698
S.E. of regression         0.475350    Akaike info criterion      1.351121
Sum squared resid          1038.953    Schwarz criterion          1.355317
Log likelihood            -3105.255    Hannan-Quinn criter.       1.352598
F-statistic                131.9791    Durbin-Watson stat         0.100016
Prob(F-statistic)          0.000000
==========================================================================
```

Should we include all variables in regression?

- Sometimes we have many potential variables we can include in our regression.
- E.g.: Gmail example. We could use whether the words “free”, “buy”, “money” appear in the email, the number of exclamation marks, etc. to predict whether the email is a spam.
- Previous slides suggest we should include as many variables as possible in the regression, to get the highest R^2.
- If we do this, we run into a problem called overfitting: we will make excellent predictions within the sample we use to run the regression (high R^2), but bad predictions when we use regression predict dependent variables of units outside of our sample.
- Issue: we do not care about in-sample prediction: for the units in the sample, we already know their Y_i, no need to predict them. It’s for the units not in the sample, for which we do not know the value of their dependent variable that we want to make good predictions.

Introduction to overfitting, through an example

- Assume that in your data, you only have 3 emails.
- Assume also that for each email, you measure 3 variables:
  - Y_i: whether the email is a spam
  - X_{1i}: whether the minute when email was sent is an odd number
  - X_{2i}: whether the second when email was sent is an odd number.
- X_{1i} and X_{2i} should be poor predictors of whether email is spam: no reason why spams more likely to be sent on odd minutes/seconds.
- Assume that the values of Y_i, X_{1i}, X_{2i} are as in the table below.

Email	Y_i	X_{1i}	X_{2i}
1	1	1	1
2	0	1	0
3	0	0	0

- Find (c_0, c_1, c_2) such that $\sum_{i=1}^{3}\left(Y_i - (c_0 + c_1 X_{1i} + c_2 X_{2i})\right)^2$=0.

iClicker time

- Assume that the values of Y_i, X_{1i}, X_{2i} are as in the table below.

Email	Y_i	X_{1i}	X_{2i}
1	1	1	1
2	0	1	0
3	0	0	0

- (c_0, c_1, c_2) such that $\sum_{i=1}^{3}\left(Y_i - (c_0 + c_1 X_{1i} + c_2 X_{2i})\right)^2$=0 is:

a) $c_0 = 0, c_1 = 0, c_2 = 0.$

b) $c_0 = 0, c_1 = 0, c_2 = 1.$

c) $c_0 = 0, c_1 = 1, c_2 = 1.$

d) $c_0 = 1, c_1 = 1, c_2 = 1.$

$c_0 = 0, c_1 = 0, c_2 = 1.$

Email	Y_i	X_{1i}	X_{2i}
1	1	1	1
2	0	1	0
3	0	0	0

- (c_0, c_1, c_2) such that $\sum_{i=1}^{3}\left(Y_i - (c_0 + c_1X_{1i} + c_2X_{2i})\right)^2$=0 is solution of this system:

$1-(c_0 + c_1 + c_2) = 0$

$0 - (c_0 + c_1) = 0$

$0 - (c_0) = 0$

- You can check that solution is $c_0 = 0, c_1 = 0, c_2 = 1.$
- Now assume that in this example, you regress Y_i on a constant, X_{1i} and X_{2i}. Let $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2$ respectively denote the coefficient of the constant, of X_{1i}, and of X_{2i} in this regression. What will be the value of $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2$? Discuss this question during 2 minutes with your neighbor.

iClicker time

- Values of Y_i, X_{1i}, X_{2i} are as in the table below.

Email	Y_i	X_{1i}	X_{2i}
1	1	1	1
2	0	1	0
3	0	0	0

- You regress Y_i on a constant, X_{1i} and X_{2i}. $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2$ denote the coefficient of the constant, of X_{1i}, and of X_{2i} in this regression. What will be the value of $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2$?

$a)\ \hat{\gamma}_0 = 0, \hat{\gamma}_1 = 0, \hat{\gamma}_2 = 0.$

$b)\ \hat{\gamma}_0 = 0, \hat{\gamma}_1 = 0, \hat{\gamma}_2 = 1.$

$c)\ \hat{\gamma}_0 = 0, \hat{\gamma}_1 = 1, \hat{\gamma}_2 = 1.$

$d)\ \hat{\gamma}_0 = 1, \hat{\gamma}_1 = 1, \hat{\gamma}_2 = 1.$

$\hat{\gamma}_0 = 0, \hat{\gamma}_1 = 0, \hat{\gamma}_2 = 1.$

- $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2)$: minimizer of $\sum_{i=1}^{3}\left(Y_i - (c_0 + c_1 X_{1i} + c_2 X_{2i})\right)^2$.
- For any (c_0, c_1, c_2), $\sum_{i=1}^{3}\left(Y_i - (c_0 + c_1 X_{1i} + c_2 X_{2i})\right)^2 \geq 0$.
- If for a (c_0, c_1, c_2), $\sum_{i=1}^{3}\left(Y_i - (c_0 + c_1 X_{1i} + c_2 X_{2i})\right)^2 = 0$, this (c_0, c_1, c_2) is that minimizing $\sum_{i=1}^{3}\left(Y_i - (c_0 + c_1 X_{1i} + c_2 X_{2i})\right)^2$, so $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2)$ equal to that (c_0, c_1, c_2).
- $\sum_{i=1}^{3}\left(Y_i - (c_0 + c_1 X_{1i} + c_2 X_{2i})\right)^2 = 0$ if $c_0 = 0, c_1 = 0, c_2 = 1$.
- Therefore, $\hat{\gamma}_0 = 0, \hat{\gamma}_1 = 0, \hat{\gamma}_2 = 1$.
- Prediction function for whether email spam: $0 + 0 \times X_{1i} + 1 \times X_{\mathbf{2}i}$
- => you predict that all emails sent on an odd second are spams while all emails sent on an even second are not spams.
- This regression has an $R^2 = 1$: regression predicts perfectly whether email are spams, in your sample of 3 observations.
- Do you think that this regression will give good predictions, when you use it to make predictions for emails outside of the sample of 3 emails in the regression?

iClicker time

- => in this example, prediction function for whether the email is a spam is $0 + 0 \times X_{1i} + 1 \times X_{\mathbf{2}i}$
- => you predict that all emails sent on an odd second are spams while all emails sent on an even second are not spams.
- This regression has an $R^2 = 1$: regression model predicts perfectly whether email are spams, in your sample of 3 observations.
- Do you think that this regression will give good predictions, when you use it make predictions for emails outside of the sample of 3 emails you use in the regression?

a) Yes

b) No

No!

- There is no reason why emails sent on odd seconds would be more likely to be spams than emails sent on even seconds.
- => when you use the regression to make predictions for whether emails out of your sample are spams or not, you will get very bad predictions.
- This is despite the fact that in your sample, your regression yields perfect predictions. $R^2 = 1$.
- So what is going on?

R^2 of reg. with as many variables as units =1…

- You have n observations, and for each observation you measure dependent variable Y_i and $n-1$ independent variables X_{1i},…, X_{n-1i}.
- You regress Y_i on constant, X_{1i},…, X_{n-1i}.
- $(\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_{n-1})$, coefficients of constant, X_{1i},…, X_{n-1i}: value of $(c_0,$ $c_1, \dots, c_{n-1})$ minimizing $\sum_{i=1}^{n}\left(Y_i - (c_0 + c_1 X_{1i} + \cdots + c_{n-1}X_{n-1i})\right)^2$.
- We can make each term in summation =0. Equivalent to solving:

$$Y_1 - \left(c_0 + c_1 X_{11} + \cdots + c_{n-1}X_{n-1,1}\right) = 0$$

$$Y_2 - \left(c_0 + c_1 X_{12} + \cdots + c_{n-1}X_{n-1,2}\right) = 0$$

…

$$Y_n - \left(c_0 + c_1 X_{1n} + \cdots + c_{n-1}X_{n-1,n}\right) = 0$$

- System of n equations with n unknowns => has a solution.

$\Rightarrow (\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_{n-1})$ is the solution of this system, and

$$\sum_{i=1}^{n}\left(Y_i - (\hat{\gamma}_0 + \hat{\gamma}_1 X_{1i} + \cdots + \hat{\gamma}_{n-1}X_{n-1i})\right)^2 = 0$$

$\Rightarrow$ MSE in this regression =0

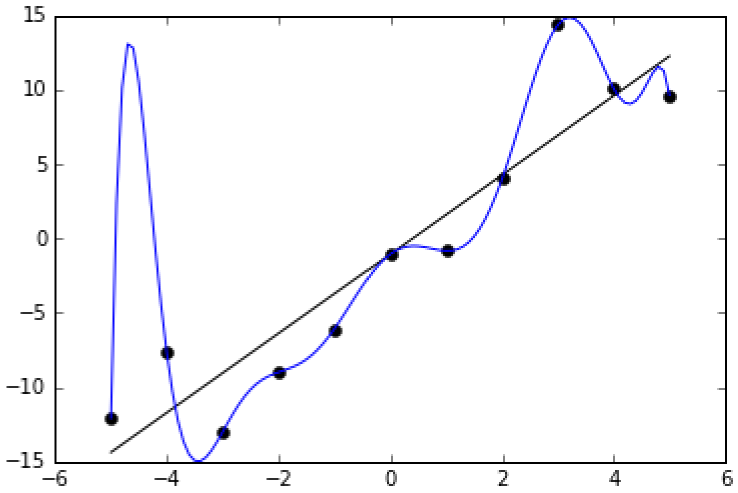
$\Rightarrow R^2$= 1: R^2= 1- MSE/ variance of spam variable.

… Even if independent variables in regression are actually really bad predictors of Y_i

- R^2 of reg. with as many independent variables as units=1.
- Mechanical property, just comes from the fact that a system a n equations with n unknowns has a solution.
- True even if independent variables actually bad predictors of Y_i.
- E.g.: previous example, where we regressed whether an email is spam or not on stupid variables (whether it was sent on an odd second…) still had R^2 of 1.
- R^2=1 means that regression predicts Y_i perfectly well in sample, but probably will yield bad predictions outside of the sample.
- Overfitting: we give ourselves so many parameters we can play with (the coefficients of all the variables in the regression) that we end up fitting perfectly the variable Y_i in our sample, but we will make very large prediction errors outside of our sample.

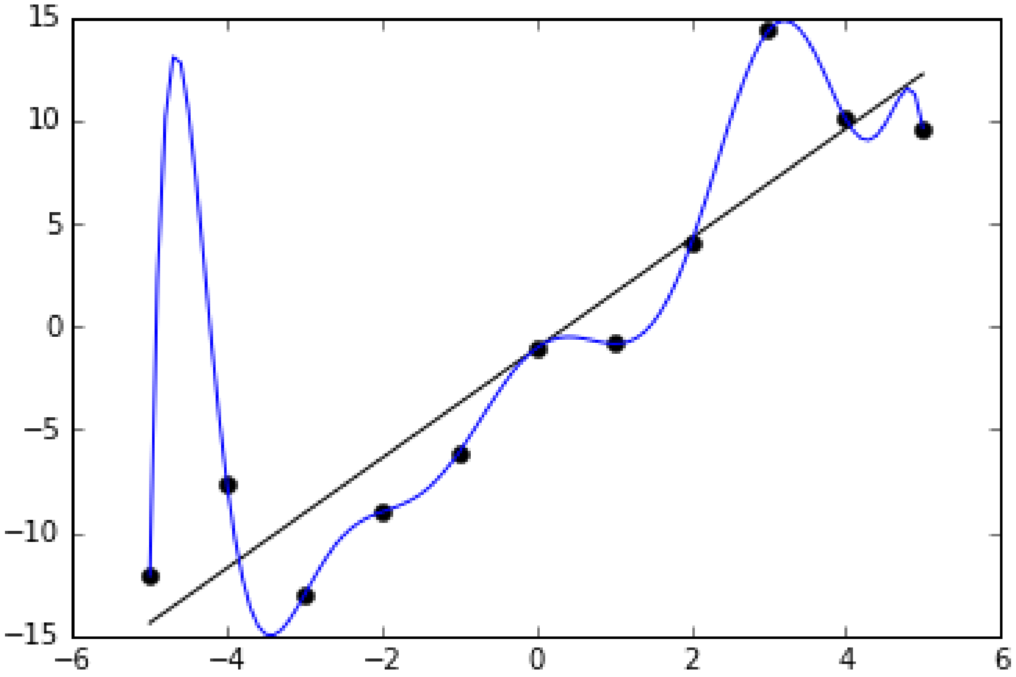
Another example of overfitting

- Figure below: 11 units, with their values of a variable X_{1i} and of a variable Y_i.
- Black line: regression function you obtain when you regress Y_i on a constant and X_{1i}.
- Blue line: regression function you obtain when you regress Y_i on a constant, X_{1i}, ${X_{1i}}^2$, ${X_{1i}}^3$, ${X_{1i}}^4$,…., ${X_{1i}}^{10}$.



- Which of these two regressions will have the highest R^2?

iClicker time



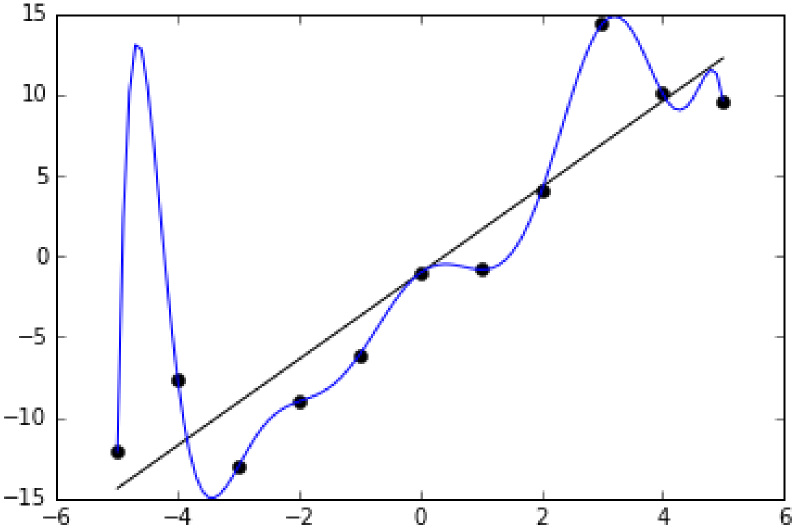
- Which of these two regressions will have the highest R^2?

a) The regression of Y_i on a constant and X_{1i}.

b) The regression of Y_i on a constant, X_{1i}, ${X_{1i}}^2$, ${X_{1i}}^3$, ${X_{1i}}^4$,…., ${X_{1i}}^{10}$.

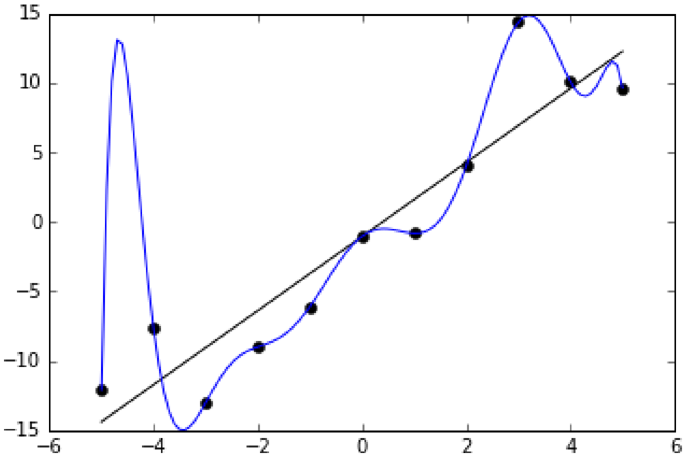
Regression of Y_i on constant, X_{1i}, ${X_{1i}}^2$,...., ${X_{1i}}^{10}$.

- Regression of Y_i on constant, X_{1i}, ${X_{1i}}^2$, ${X_{1i}}^3$, ${X_{1i}}^4$,...., ${X_{1i}}^{10}$ has 11 observations and 11 independent variables. R^2=1. Blue line fits perfectly black dots.
- Black line does not perfectly fit black dots => regression of Y_i on a constant and X_{1i} has R^2 <1.



- Goal of regression is to make prediction for value of dependent variable of units not in your sample, for which you observe the xs but not y.
- Assume that one of these units has $x = -4.5$. Do you think you will get a better prediction for the y of that unit using regression of Y_i on a constant and X_{1i}, or regression of Y_i on a constant, X_{1i}, ${X_{1i}}^2$, ${X_{1i}}^3$, ${X_{1i}}^4$,...., ${X_{1i}}^{10}$?

iClicker time



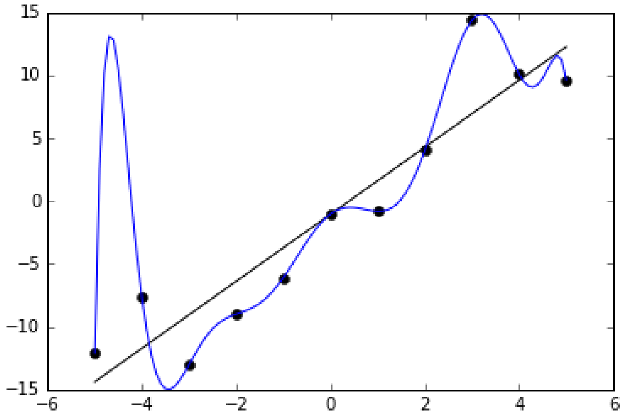
- The goal of a regression is to make a prediction for the value of the dependent variable of units not in sample, for which you observe the xs but not y.
- Assume that one of these units has $x = -4.5$. Do you think you will get a better prediction for the y of that unit using regression of Y_i on constant and X_{1i}, or regression of Y_i on constant, X_{1i}, ${X_{1i}}^2$, ${X_{1i}}^3$, ${X_{1i}}^4$,…., ${X_{1i}}^{10}$?

a) We will get a better prediction using the regression of Y_i on a constant and X_{1i}

b) We will get a better prediction using the regression of Y_i on a constant, X_{1i}, ${X_{1i}}^2$, ${X_{1i}}^3$, ${X_{1i}}^4$,…., ${X_{1i}}^{10}$.

Better prediction using reg. of Y_i on constant & X_{1i}

- Prediction of the y of unit with $x = -4.5$:
  - according to reg. of Y_i on a constant, X_{1i}, ${X_{1i}}^2$,…., ${X_{1i}}^{10}$: 13.
  - according to reg. of Y_i on a constant and X_{1i}: -12.
- In sample, units with x close to -4.5 have y much closer to -12 than to 13=> regression of Y_i on a constant and X_{1i} will give better prediction.



- Again, regression with many independent variables might give very good in-sample prediction but very bad out of sample prediction
- But making good out of sample prediction is goal of regression.
- => Comparing R^2 of 2 regs. is not right way to assess which will give best out of sample predictions. Reg. with many variables always has very high R^2 but might end up making poor out of sample predictions.

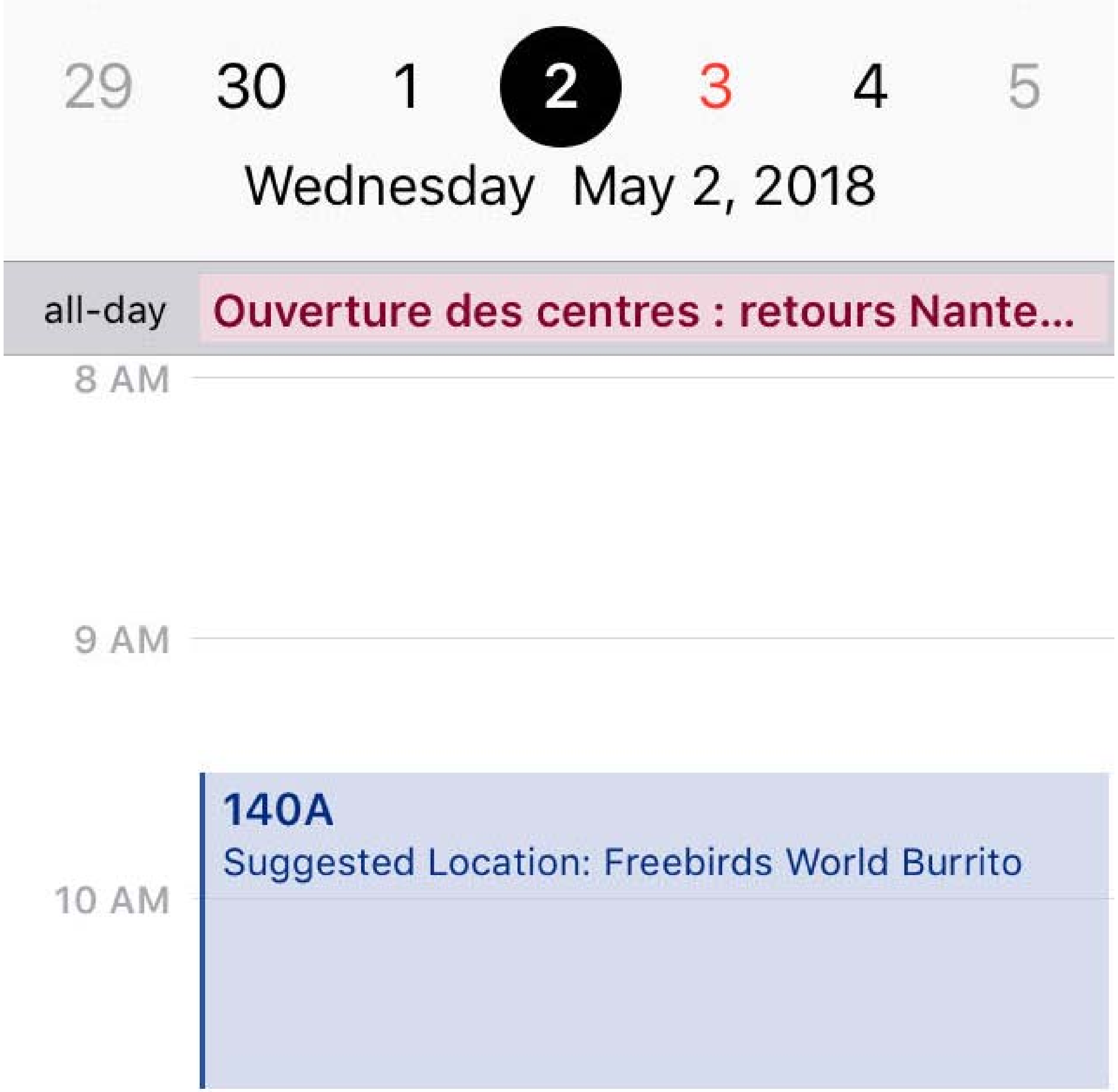
Instead, use a training and validation sample

- You start from sample of n units for which you measure Y_i, dependent variable, and X_{1i},…, X_{Ji}, independent variables.
- Randomly divide sample into two subsamples of $n/2$ units. Subsample 1: training sample. Subsample 2: validation sample.
- **In training sample**, you estimate the regressions you are interested in.
- For instance, **in training sample**:
  - Regression 1: Y_i on a constant and X_{1i},…, X_{Ji}. Coefficients $\left(\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_J\right)$.
  - Regression 2: Y_i on a constant and X_{1i}. Coefficients $\left(\hat{\beta}_0, \hat{\beta}_1\right)$.
- Then, compute squared prediction error according to each regression for units **in validation sample.**
- For instance, for each unit **in validation sample**, compute:
  - $\left(Y_i - (\hat{\gamma}_0 + \hat{\gamma}_1 X_{1i} + \cdots + \hat{\gamma}_J X_{Ji})\right)^2$: squared pred. error with Reg. 1.
  - $\left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i})\right)^2$: squared pred. error with Reg. 2.
- Finally, choose regression for which sum squared prediction errors for units **in validation sample** lowest.
- Intuition: you want to use the reg. that gives best out of sample predictions. By choosing reg. that gives best predictions in validation sample, you ensure that your regression will give good out of sample predictions, because you did not use the validation sample to compute your reg. coefficients.

Machine learning in 2 minutes

- Using training and validation sample = key idea underlying **machine learning methods** (statistical methods more sophisticated than, but inspired from multivariate regressions, and that are used by tech companies to do image recognition, spam detection, etc.)
- **Goal**: teach a computer to recognize whether an email is a Spam, whether a picture of a letter is an “a”, a “b”, etc.
- Train the computer in a sample of emails for which the computer knows whether the email is a spam and many other variables (all the words in the email, etc.).
- The computer finds the model that predicts the best whether the email is a spam given all these variables, in the training sample.
- Then, check whether prediction model works well in validation sample, where you also know which emails are spams or not.
- If the statistical model also works well in the validation sample, implement method in real life to predict whether new emails reaching Gmail accounts are spams or not. If email predicted to be spam, send to junk box. Otherwise, send to regular mail box.

Machine learning often works, but not always



Wednesday May 2, 2018

Ouverture des centres : retours Nante...

140A

Suggested Location: Freebirds World Burrito

What you need to remember

- Great advantage of multivariate regression over univariate regression: improves the quality of our predictions.
- However, putting too many variables in regression might result in overfitting: regression fits very well ys in sample, but gives poor out of sample predictions.
- For instance, a regression with as many independent variables as units will automatically have a $R^2 = 1$, even if those independent variables are actually poor predictors of the independent variable.
- => comparing R^2s not good way to choose between several regs
- Instead, you should:
  - randomly divide sample into training and validation sample
  - estimate your regressions in the training sample only
  - compute squared predicted errors according to each regression in validation sample
  - choose regression for which MSE in validation sample is smallest.
- Training / validation sample idea underlies machine learning models used for spam detection / image recognition, etc. by tech companies.

Roadmap

1. The OLS multivariate regression function.
2. Estimating the OLS multivariate regression function.
3. Advantages and pitfalls of multivariate regressions.
4. Interpreting coefficients in multivariate OLS regressions.

Interpreting coeff. of multivariate regs. An example.

- 6 units ($n = 6$). 3 variables: Y_i, D_i, and X_i. D_i, and X_i: binary.

Unit	Y_i	$\boldsymbol{D_i}$	X_i
1	5	1	1
2	3	1	1
3	4	0	1
4	1	1	0
5	0	0	0
6	2	0	0

- If you regress Y_i on constant and D_i, what will be coeff. of D_i? If you regress Y_i on a constant, D_i, and X_i, what will be coeff. of D_i? Hint: to answer first question, you can use a result you saw during sessions. To answer second question, write system of 3 equations and three unknowns solved by $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2)$, the coefficients of constant, D_i, and X_i, plug-in the values of Y_i, D_i, and X_i in table, and then solve system.

iClicker time

Unit	Y_i	D_i	X_i
1	5	1	1
2	3	1	1
3	4	0	1
4	1	1	0
5	0	0	0
6	2	0	0

If you regress Y_i on constant and D_i, what will be coeff. of D_i? If you regress Y_i on a constant, D_i, and X_i, what will be coeff. of D_i?

a) In reg. of Y_i on constant and D_i, coeff. of D_i is 2. In reg. of Y_i on a constant, D_i, and X_i, coeff. of D_i is 0.5.

b) In reg. of Y_i on constant and D_i, coeff. of D_i is 1. In reg. of Y_i on a constant, D_i, and X_i, coeff. of D_i is 0.5.

c) In reg. of Y_i on constant and D_i, coeff. of D_i is 1. In reg. of Y_i on a constant, D_i, and X_i, coeff. of D_i is 0.

d) In reg. of Y_i on constant and D_i, coeff. of D_i is 1. In reg. of Y_i on a constant, D_i, and X_i, coeff. of D_i is -0.5.

In regression of Y_i on constant + D_i, coeff of D_i is 1. In regression of Y_i on constant, D_i, + X_i, coeff of D_i is 0.

- Coeff of $\mathrm{D_i}$ in reg. of $\mathrm{Y_i}$ on constant, $\mathrm{D_i}$. Result of sessions:

(Average $\mathrm{Y_i}$ for $\mathrm{D_i} = 1$)- (Average $\mathrm{Y_i}$ for $\mathrm{D_i} = 0$)=
1/3(5+3+1)−1/3(4+2+0) $= 1$.

- Coeff $\mathrm{D_i}$ in reg. of $\mathrm{Y_i}$ on constant, $\mathrm{D_i}$, X_i: 3 eqs. with 3 unknowns.
- $n = 6$ and $J = 2$, so we have (we can forget the -2):

$$\sum_{i=1}^{6}\left(Y_i - (\hat{\gamma}_0 + \hat{\gamma}_1 D_i + \hat{\gamma}_2 X_i)\right) \quad = 0$$

$$\sum_{i=1}^{6} D_i\left(Y_i - (\hat{\gamma}_0 + \hat{\gamma}_1 D_i + \hat{\gamma}_2 X_i)\right) \quad = 0$$

$$\sum_{i=1}^{6} X_i\left(Y_i - (\hat{\gamma}_0 + \hat{\gamma}_1 D_i + \hat{\gamma}_2 X_i)\right) \quad = 0$$

- Plugging values of Y_i, D_i, and X_i, yields:

$15 - 6\hat{\gamma}_0 - 3\hat{\gamma}_1 - 3\hat{\gamma}_2 = 0$
$9 - 3\hat{\gamma}_0 - 3\hat{\gamma}_1 - 2\hat{\gamma}_2 = 0$
$12 - 3\hat{\gamma}_0 - 2\hat{\gamma}_1 - 3\hat{\gamma}_2 = 0$

- Subtracting eq 2 to eq 3: $3 + \hat{\gamma}_1 - \hat{\gamma}_2 = 0$
- Multiplying eq 3 by 2 and subtracting eq 1: $9 - \hat{\gamma}_1 - 3\hat{\gamma}_2 = 0$
- Adding the two preceding equations: $12 - 4\hat{\gamma}_2 = 0$, so $\hat{\gamma}_2 = 3$.
- Plugging $\hat{\gamma}_2 = 3$ in $3 + \hat{\gamma}_1 - \hat{\gamma}_2 = 0$: $\hat{\gamma}_1 = 0$.

A general formula for coefficient of binary variable in a regression of Y_i on constant and 2 binary variables.

- Let D_i and X_i be 2 binary variables.
- n_{00}: number of units with $D_i = 0, X_i = 0$. n_{10}: number of units with $D_i = 1, X_i = 0$. n_{01}: number of units with $D_i = 0$, $X_i = 1$. n_{11}: number of units with $D_i = 1, X_i = 1$.
- Coeff of D_i in regression of Y_i on constant, D_i, X_i is:

$$w\left(\frac{1}{n_{10}}\sum_{i:D_i=1,\ X_i=0} Y_i - \frac{1}{n_{00}}\sum_{i:D_i=0,\ X_i=0} Y_i\right) + (1-w)\left(\frac{1}{n_{11}}\sum_{i:D_i=1,\ X_i=1} Y_i - \frac{1}{n_{01}}\sum_{i:D_i=0,\ X_i=1} Y_i\right)$$

w: number included between 0 and 1, no need to know formula.

- $\frac{1}{n_{10}}\sum_{i:D_i=1,\ X_i=0} Y_i - \frac{1}{n_{00}}\sum_{i:D_i=0,\ X_i=0} Y_i$: difference between average Y_i of units with $D_i = 1$ and of units with $D_i = 0$, among units with $X_i = 0$.
- $\frac{1}{n_{11}}\sum_{i:D_i=1,\ X_i=1} Y_i - \frac{1}{n_{01}}\sum_{i:D_i=0,\ X_i=1} Y_i$: difference between average Y_i of units with $D_i = 1$ and of units with $D_i = 0$, among units with $X_i = 1$.
- Coeff of D_i measures difference between average of Y_i across subgroups whose D_i differs by one, but that have same X_i!

Applying formula in example.

- Sample with 6 units. 3 variables: Y_i, D_i, X_i. D_i X_i: binary variables.

Unit	Y_i	$\boldsymbol{D_i}$	X_i
1	5	1	1
2	3	1	1
3	4	0	1
4	1	1	0
5	0	0	0
6	2	0	0

- What is value of $\frac{1}{n_{10}}\sum_{i:D_i=1,\ X_i=0} Y_i - \frac{1}{n_{00}}\sum_{i:D_i=0,\ X_i=0} Y_i$?

Of $\frac{1}{n_{11}}\sum_{i:D_i=1,\ X_i=1} Y_i - \frac{1}{n_{01}}\sum_{i:D_i=0,\ X_i=1} Y_i$?

iClicker time

Unit	Y_i	D_i	X_i
1	5	1	1
2	3	1	1
3	4	0	1
4	1	1	0
5	0	0	0
6	2	0	0

$a)$ $\frac{1}{n_{10}}\sum_{i:D_i=1,\ X_i=0} Y_i - \frac{1}{n_{00}}\sum_{i:D_i=0,\ X_i=0} Y_i$=1

$\frac{1}{n_{11}}\sum_{i:D_i=1,\ X_i=1} Y_i - \frac{1}{n_{01}}\sum_{i:D_i=0,\ X_i=1} Y_i = -1$

b) $\frac{1}{n_{10}}\sum_{i:D_i=1,\ X_i=0} Y_i - \frac{1}{n_{00}}\sum_{i:D_i=0,\ X_i=0} Y_i$=0

$\frac{1}{n_{11}}\sum_{i:D_i=1,\ X_i=1} Y_i - \frac{1}{n_{01}}\sum_{i:D_i=0,\ X_i=1} Y_i = 0$

c) $\frac{1}{n_{10}}\sum_{i:D_i=1,\ X_i=0} Y_i - \frac{1}{n_{00}}\sum_{i:D_i=0,\ X_i=0} Y_i$=-1

$\frac{1}{n_{11}}\sum_{i:D_i=1,\ X_i=1} Y_i - \frac{1}{n_{01}}\sum_{i:D_i=0,\ X_i=1} Y_i = 1$

$$\frac{1}{n_{10}}\sum\nolimits_{i:D_i=1,\ X_i=0} Y_i - \frac{1}{n_{00}}\sum\nolimits_{i:D_i=0,\ X_i=0} Y_i = 0,$$
$$\frac{1}{n_{11}}\sum\nolimits_{i:D_i=1,\ X_i=1} Y_i - \frac{1}{n_{01}}\sum\nolimits_{i:D_i=0,\ X_i=1} Y_i = 0$$

Unit	Y_i	$\boldsymbol{D_i}$	X_i
1	5	1	1
2	3	1	1
3	4	0	1
4	1	1	0
5	0	0	0
6	2	0	0

- $\frac{1}{n_{10}}\sum\nolimits_{i:D_i=1,\ X_i=0} Y_i - \frac{1}{n_{00}}\sum\nolimits_{i:D_i=0,\ X_i=0} Y_i = 1 - \frac{1}{2}(0+2) = 0$
- $\frac{1}{n_{11}}\sum\nolimits_{i:D_i=1,\ X_i=1} Y_i - \frac{1}{n_{01}}\sum\nolimits_{i:D_i=0,\ X_i=1} Y_i = \frac{1}{2}(5+3) - 4 = 0$
- The coeff of D_i in regression of Y_i on constant, $D_i,\ X_i$ is a weighted average of these two numbers, so that’s why it’s equal to 0, as we have shown earlier.

Interpreting coefficients in multivariate regressions.

- Previous slides: in reg. of Y_i on constant, D_i, and X_i, where D_i and X_i binary, coeff of D_i = difference between average of Y_i across groups whose D_i differs by one, but that have same X_i.
- Extends to all multivariate regressions.
- In a multivariate regression of Y_i on constant, D_i, X_{1i},…, X_{Ji}, $\hat{\gamma}_1$, the coeff. of D_i, measures difference between average of Y_i across subgroups whose D_i differs by one, but that have same X_{1i},…, X_{Ji}.
- If $\hat{\gamma}_1 = -0.3$, that means that if you compare average Y_i across units whose D_i differs by one but that have the same value of X_{1i},…, X_{Ji}, the average of Y_i is -0.3 smaller among units whose D_i is 1 unit larger.
- In a multivariate regression of $\ln(Y_i)$ on constant, D_i, X_{1i},…, X_{Ji}, if $\hat{\gamma}_1 = -0.3$, that means that if you compare average Y_i across units whose D_i differs by one but that have the same value of X_{1i},…, X_{Ji}, the average of Y_i is 30% smaller among units whose D_i is 1 unit larger.

Women earn less than males

- Same representative sample of 14086 US wage earners as in Homework 3.
- Regression of ln(weekly wage) on constant and binary variable equal to 1 for females in Stata.

```
. reg ln_weekly_wage female, r

Linear regression                               Number of obs     =     14,086
                                                F(1, 14084)       =     516.54
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0354
                                                Root MSE          =     .84461

------------------------------------------------------------------------------
             |               Robust
ln_weekly_~e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.3235403   .0142357   -22.73   0.000    -.3514442   -.2956365
       _cons |   6.642133   .0099315   668.80   0.000     6.622666      6.6616
------------------------------------------------------------------------------
```

- Women earn 32% less than males, difference very significant.
- From that regression, can we conclude that women are discriminated against in the labor market? Why?

iClicker time

- Women earn 32% less than males, difference very significant.
- Can we conclude that women are discriminated against in the labor market? Why?

a) Yes, we can conclude that women are discriminated against in the labor market, this 32% difference in wages must reflect discrimination.

b) No, we cannot conclude that women are discriminated against in the labor market, because the R2 of the regression is too low.

c) No, we cannot conclude that women are discriminated against in the labor market. Maybe women earn less than men for reasons that have nothing to do with their gender.

Maybe women earn less for reasons that have nothing to do with their gender.

- Women earn less than men.
- But that difference could for instance come from the fact they work less hours per week outside of the home.
- Maybe women not discriminated by their employer, maybe just work fewer hours for their employer => get paid less.
- (Aside: women indeed tend to work fewer hours a week outside of the home than men, but that may be because they also tend to spend more time taking care of children in households with children, another form of gender imbalance, though that imbalance is taking place in the family, not in the labor market).

A more complicated regression

- Regression of ln(weekly wage) on constant, variable for females + years of schooling, age, hours worked per week.

```
. reg ln_weekly_wage female age hours_worked years_schooling, r

Linear regression                               Number of obs     =     14,086
                                                F(4, 14081)       =    1449.64
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3883
                                                Root MSE          =     .67267

----------------------------------------------------------------------------------
                 |               Robust
  ln_weekly_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+----------------------------------------------------------------
          female |  -.2731097   .0117557   -23.23   0.000    -.2961524    -.250067
             age |   .0104635   .0004288    24.40   0.000     .0096231     .011304
    hours_worked |   .0231192   .0005812    39.78   0.000       .02198    .0242583
 years_schooling |   .1024575    .002269    45.15   0.000     .0980099    .1069052
           _cons |   3.931402   .0398147    98.74   0.000      3.85336    4.009444
----------------------------------------------------------------------------------
```

- Interpret coeff. of female variable in that regression.

iClicker time

- Interpret coeff. of female variable reg. on previous slide.

a) On average, women earn 0.27 dollars less than men per week.

b) When we compare women and men that have the same number of years of schooling, the same age, and that work the same number of hours per week, we find that on average, women earn 0.27 dollars less than men per week.

c) When we compare women and men that have the same number of years of schooling, the same age, and that work the same number of hours per week, we find that on average women earn 27% less than men per week.

Answer c) !

- Remember: in multivariate reg. of $\ln(Y_i)$ on constant, D_i, X_{1i},…, X_{Ji}, if $\hat{\gamma}_1 = -0.27$, means that if you compare average Y_i across units whose D_i differs by one but that have the same value of X_{1i},…, X_{Ji}, the average of Y_i is 27% smaller among units whose D_i is 1 unit larger.
- Here: D_i is female variable. Females have $D_i = 1$, males have $D_i = 0$.
- The other variables in the regression are years of schooling, age, and number of hours worked / week.
- => $\hat{\gamma}_1 = -0.27$ means that when we compare women and men that have the same number of years of schooling, the same age, and that work the same number of hours per week, we find that on average women earn 27% less than men per week.

Complicated reg. is stronger, though still imperfect evidence that gender discrimination on labor market.

- Difference between men and women's earnings cannot be explained by differences in education, hours worked per week, and professional experience.
- Even when we compare men and women with same education, hours worked per week, and professional experience, women earn substantially less (27%).
- This is still not definitive evidence of discrimination. Maybe women tend to go into lower paying jobs and industries than men.
- E.g.: less women in finance and engineering.
- But is this because women do not like that type of jobs (if so, no discrimination) or is it because those industries do not want to hire women (if so, discrimination), or because women would like to go into those jobs but do not do so because frowned upon due to social norms (if so, discrimination)?
- Overall, even though there are limits even with the complicated regression, the fact that women earn less even when we compare men and women with same education, hours worked per week, and professional experience, suggests that women discriminated on labor market.

What is econometrics?

- Econometrics is a set of **statistical techniques** that we can use to study **economic questions empirically.**
- The tools we use in econometrics are statistical techniques, which is why the beginning of an intro to econometrics class looks more like a stats class than an econ class: before we can apply the statistical tools to study economics question, we need to master the tools!
- Why do we want to study economic questions empirically? Isn't economic theory enough?
- The issue with economic theory is that on a number of issues, different theories lead to different conclusions.
- E.g.: neo-classical economist will tell you that increasing minimum wage will reduce employment, while a neo-Keynesian will tell you that increasing minimum wage will increase employment.
- Conflicting theories => we need to study these questions empirically (with data) to say which theory is true.
- The wage regressions in homework 3 and in these slides are a first example of how to use statistical tools to study an economic question: "are women discriminated on the labor market?" empirically (with data).
- Other examples coming in the next slides.

What you need to remember

- In a multivariate regression of Y_i on constant, D_i, X_{1i},…, X_{Ji}, $\hat{\gamma}_1$, if $\hat{\gamma}_1$, the coeff of D_i, is equal to x, means that if you compare average Y_i across units whose D_i differs by one but that have the same value of X_{1i},…, X_{Ji}, the average of Y_i is x larger (if $x > 0$) / smaller (if $x < 0$) among units whose D_i is 1 unit larger.
- In a multivariate regression of $\ln(Y_i)$ on constant, D_i, X_{1i},…, X_{Ji}, if $\hat{\gamma}_1$, the coeff of D_i, is equal to x, means that if you compare average Y_i across units whose D_i differs by one but that have the same value of X_{1i},…, X_{Ji}, the average of Y_i is x% larger (if $x > 0$) / smaller (if $x < 0$) among units whose D_i is 1 unit larger.