Introduction to Linear Regression

Guillem Riambau (Yale-NUS)

February 10, 2016

Linear Regression

Suppose we want to know what affects wages:

$wage = f(x_1, x_2, x_3...)$

For instance:

$\omega = f(age, experience, gender, education, ...)$

The model is:

$\omega = \alpha + \beta_1 AGE + \beta_2 EXP + \beta_3 ... + \varepsilon$

Linear Regression

The model is:

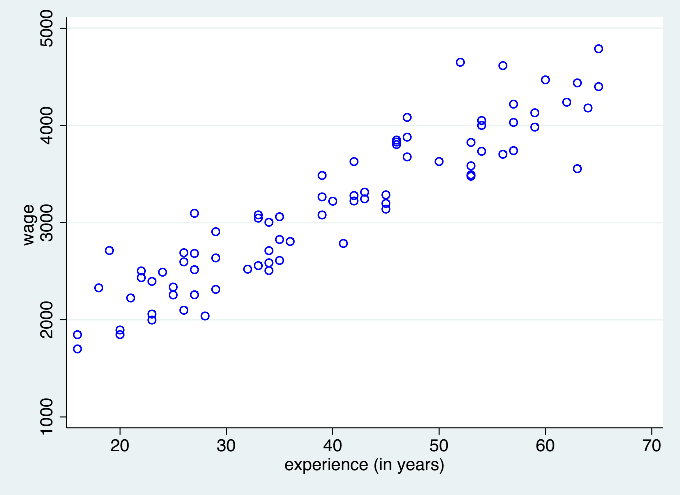
$\omega = \alpha + \beta_1 AGE + \beta_2 EXP + \varepsilon$

- Does AGE have an effect on wages?
- **Is β_1 different from 0? is it positive or negative? is the effect big or small?**
- These are the 3 key questions we want to solve, generally, for most analysis we will undertake: is there an effect, what's the direction of the effect, how big the effect is.

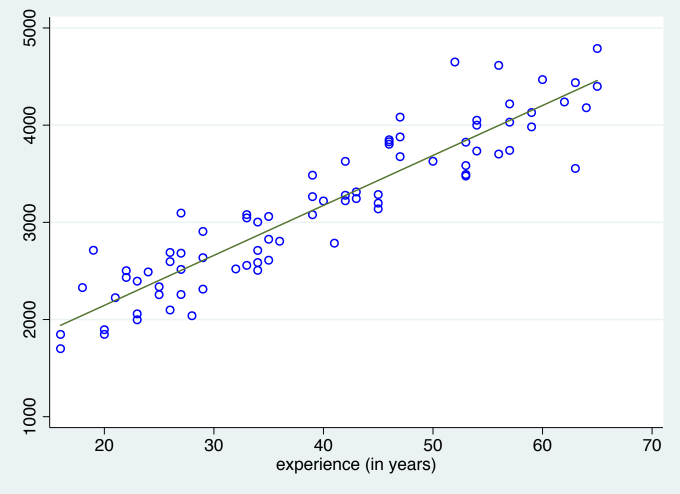
Terminology - general case

- Dependent variable: Our variable of interest (usually "y")
- Regressors or explanatory variables: variables that we believe have a causal effect on "y". Usually denoted with "x_k"
- $\Rightarrow y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_k x_k + \varepsilon$

Example: $\omega = \alpha + \beta EXP + \varepsilon$



$\hat{\beta}_1 = 51 \quad \hat{\alpha} = 1,930$



Interpretation

- $\hat{\beta} \rightarrow$ slope
- $\hat{\alpha} \rightarrow$ intercept or constant

Interpretation

- Results: $\hat{\omega} = \hat{\alpha} + \hat{\beta}_1 \times EXP$
- $\hat{\omega} = 1,117 + 51 \times EXP$
- $\hat{\beta}_1 = 51$: *On average* the predicted increase in wages for any random individual *for each additional year of experience* is $51.
- Technically, $\hat{\beta}_1 = \frac{\partial \omega}{\partial EXP}$
- $\hat{\alpha} = 1,117$ *Average predicted wage* if experience is 0

Question

- What if $\hat{\beta}_1 = -10$: What would that mean?

Question

- What if $\hat{\beta}_1 = -10$: What would that mean?
- *On average* the predicted DEcrease in wages for any random individual *for each additional year of age* is \$10.

Interpretation when we have two (or more) regressors

- Model: $\omega = \alpha + \beta_1 \times EXP + \beta_2 \times EDUC + \varepsilon$

Interpretation when we have two (or more) regressors

- Model: $\omega = \alpha + \beta_1 \times EXP + \beta_2 \times EDUC + \varepsilon$
- Results: $\hat{\omega} = 1,117 + 51 \times EXP + 25 \times EDUC$
- $\hat{\beta}_1 = 51$: *On average, holding everything else constant* the predicted increase in wages for any random individual *for each additional year of age* is \$51.
- $\hat{\beta}_2 = 25$: *On average, holding everything else constant* the predicted increase in wages for any random individual *for each additional year of education* is \$25.
- We are assuming we are not changing anything else: $\hat{\beta}_k$ tells us, *caeteris paribus*, what happens to ω if we increase x_k by one unit.
- Since *effect is linear*, do not need to impose that we hold other variables at their mean.

Generally

- $\hat{\beta}_K$: *On average* the predicted increase in y for any random individual *for each unit increase of* x_k
- $\hat{\alpha}$ *Average predicted* value of y if all explanatory variables are set at 0.
- $\hat{\alpha}$ Usually not reported (uninteresting)

Stars*,**,*** and standard errors (below, typical table)

Dependent variable: wage	(1)	(2)	(3)
EXPERIENCE	38.41***	40.13***	25.77***
	(4.83)	(3,38)	(4.56)
MALE		1,320***	203.23
		(102.12)	(270.13)
MALE ×			28.24***
EXPERIENCE			(6.39)
Constant	1,979***	1,236***	1,813***
	(204.01)	(154.23)	195.84
R^2	0.28	0.65	0.69
Observations	160	160	160

Standard errors in parenthesis.*: significant at 10% level, **: significant at 5% level, ***: significant at 1% level.

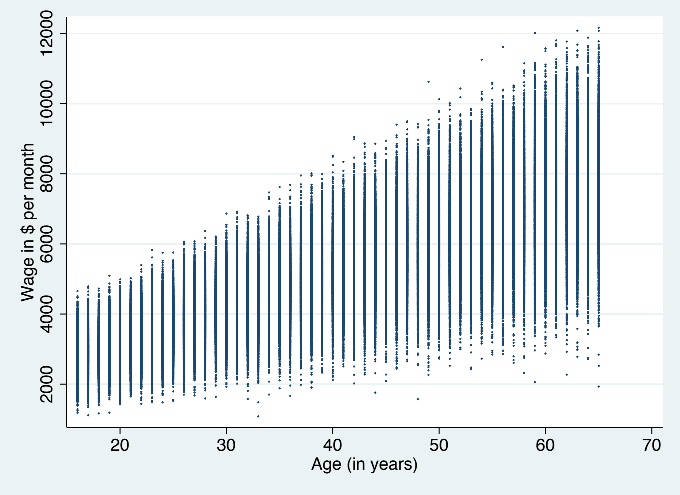
Stars*,**,*** and standard errors

- We want to find the true population β
- We need to take samples, so we can only estimated $\hat{\beta}$
- (we cannot survey every single person in the population, so we take a random sample)

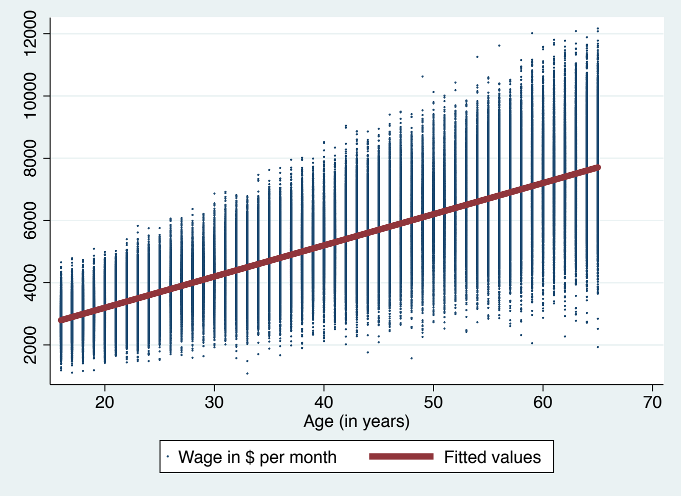
Stars*,**,*** and standard errors

- Why do tables include stars?
- Why do tables include this thing known as standard errors
- SAS rule: *Stars are shiok!*

Whole Population. N=200,000



Whole Population. N=200,000 $\beta_0 = 100.16$ $\alpha_0 = 1,193.34$



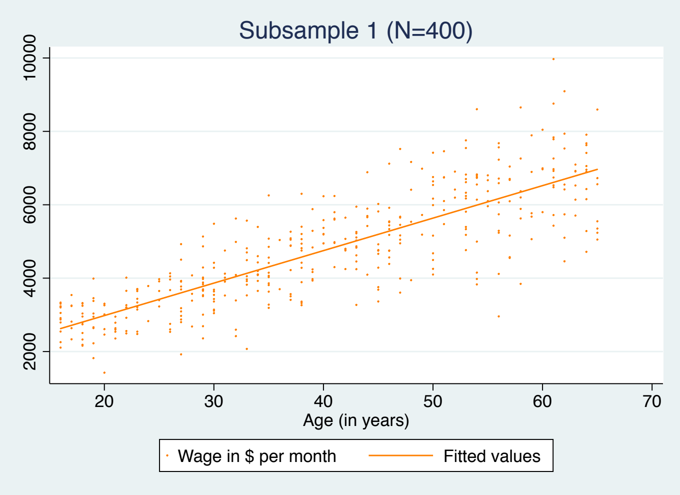
Stars*,**,*** and standard errors

- Note: we do not need standard errors when we have the whole population
- β_0 α_0 are the *true* average effects
- Note: α_0 is average wage at age=0. This is a bit unrealistic.
- How to solve this?

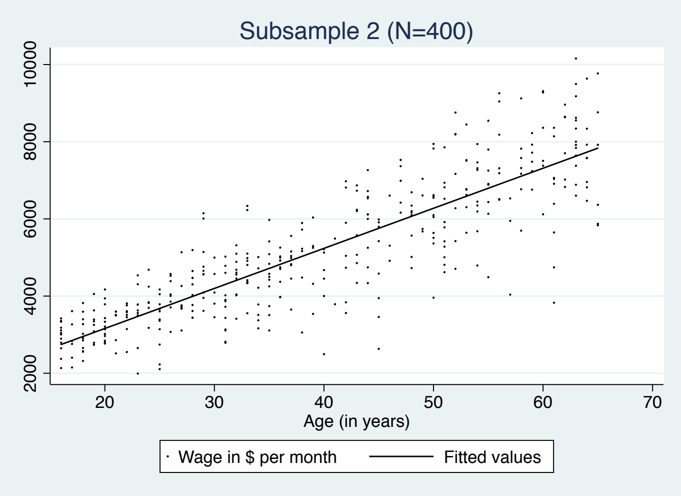
Stars*,**,*** and standard errors

- Note: we do not need standard errors when we have the whole population
- β_0 α_0 are the *true* average effects
- Note: α_0 is average wage at age=0. This is a bit unrealistic.
- How to solve this?
- Age_in_market = Age - 16.
- New α_0 would be average wage at age=16 (labour market entry)

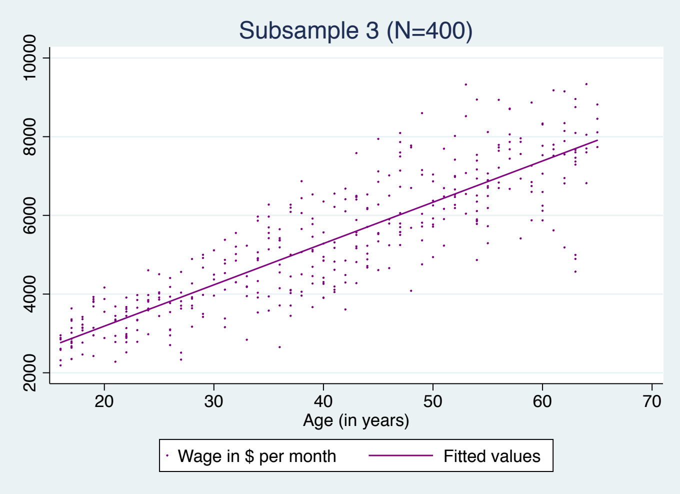
Subsample 1: N=400 $\hat{\beta} = 88.51$ $\hat{\alpha} = 1,209$



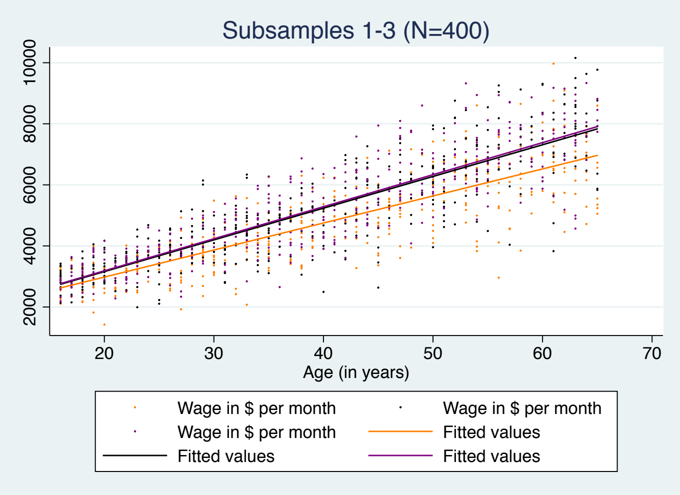
Subsample 2: N=400 $\hat{\beta} = 103.8$ $\hat{\alpha} = 1,082$



Subsample 3: N=400 $\hat{\beta} = 104.9$ $\hat{\alpha} = 1,086$



All three subsamples together



Stars*,**,*** and standard errors

- Every subsample yields a different estimate
- As long as we know how much these estimates may vary from one subsample to the next one, we are good
- We know by CLT (derived in class)
- Intuition: if all subsamples give very similar βs, then the standard errors will be small

Stars*,**,*** and standard errors

- Every subsample yields a different estimate
- As long as we know how much these estimates may vary from one subsample to the next one, we are good
- We know by CLT (derived in class)
- Intuition: if all subsamples give very similar βs, then the standard errors will be small
- Subsample 1: $\hat{\beta} = 88.51$ st.error $\hat{\beta} = 3.09$
- Subsample 2: $\hat{\beta} = 103.8$ st.error $\hat{\beta} = 3.16$
- Subsample 3: $\hat{\beta} = 104.9$ st.error $\hat{\beta} = 3.07$

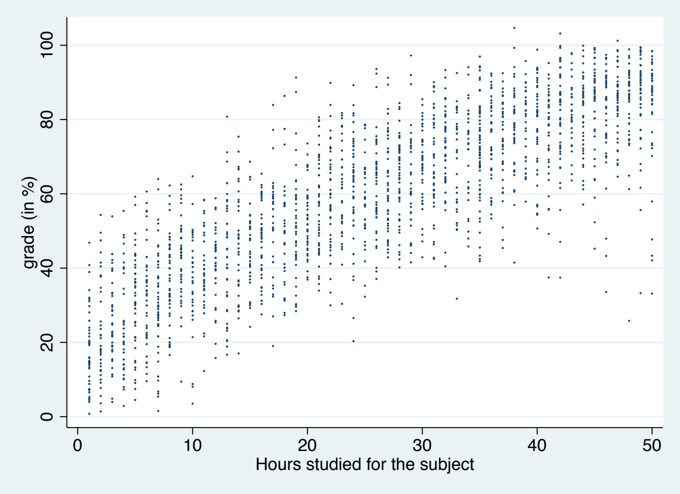
Stars*,**,*** and standard errors

- Subsample 1: $\hat{\beta} = 88.51$ st.error $\hat{\beta} = 3.09$
- Subsample 2: $\hat{\beta} = 103.8$ st.error $\hat{\beta} = 3.16$
- Subsample 3: $\hat{\beta} = 104.9$ st.error $\hat{\beta} = 3.07$
- Note: standard errors are similar. Why?
- Because random sampling: if well done, that's what will happen

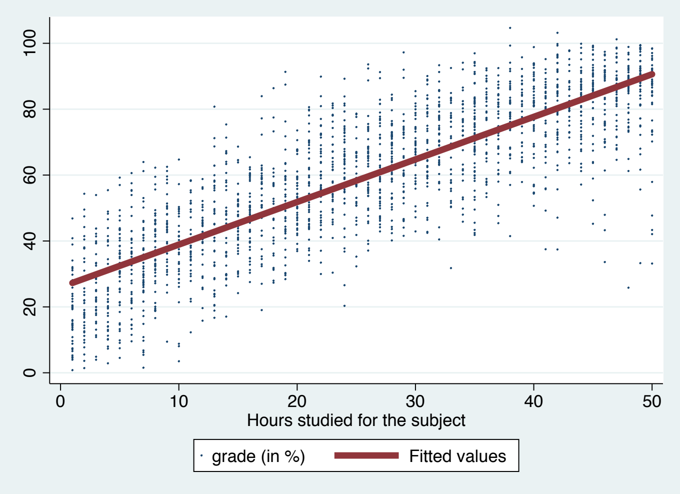
Stars*,**,*** and standard errors

- Now suppose another population
- We can get smaller samples
- Larger variability

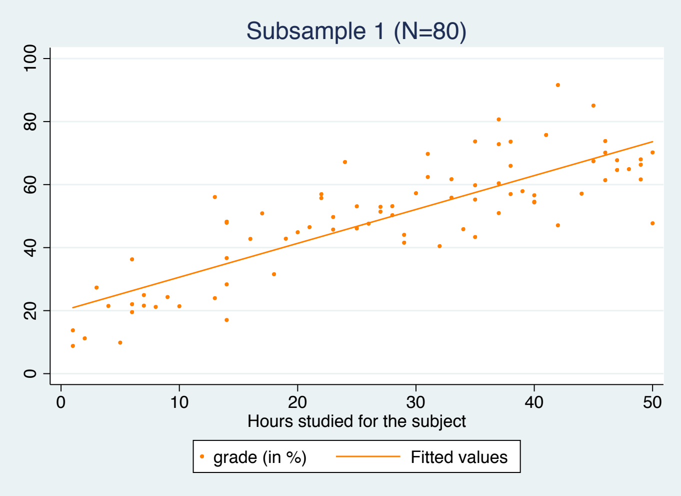
College A. N=2,000



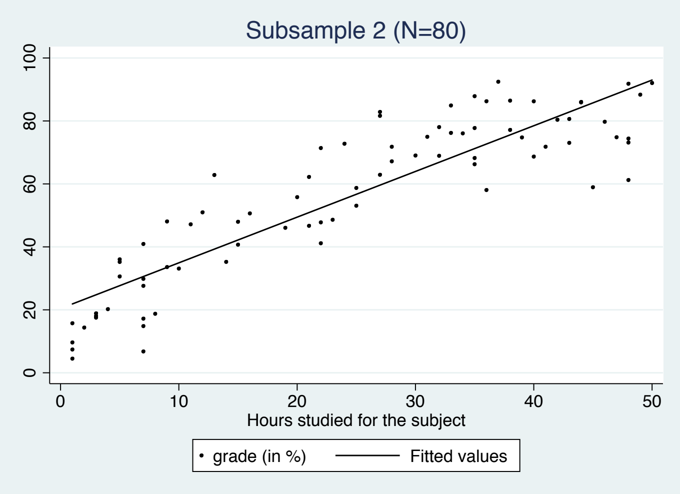
College A. N=2,000 $\beta_0 = 1.29$ $\alpha_0 = 25.98$



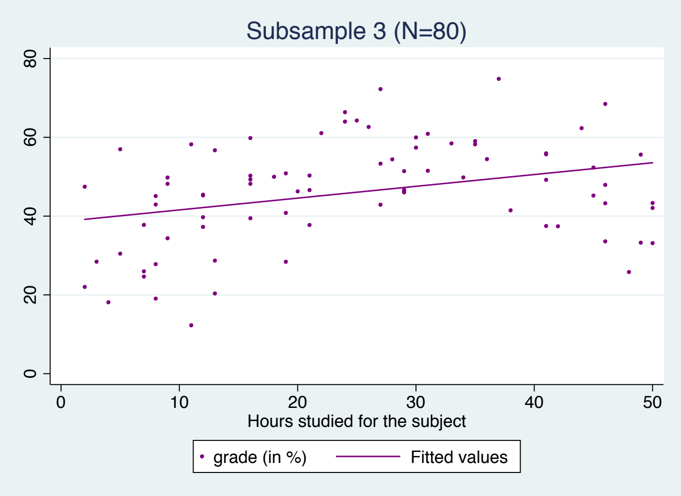
College A. Subsample 1. N=80 $\hat{\beta} = 1.07$ $\hat{\alpha} = 19.87$



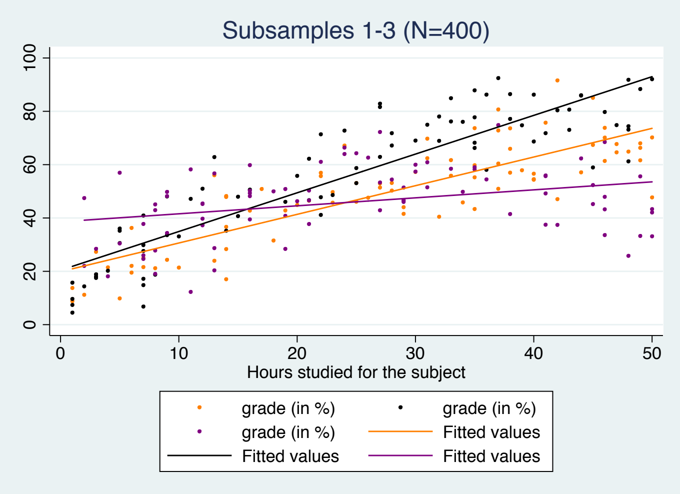
College A. Subsample 2. N=80 $\hat{\beta} = 1.45$ $\hat{\alpha} = 20.41$



College A. Subsample 3. N=80 $\hat{\beta} = 0.30$ $\hat{\alpha} = 38.57$



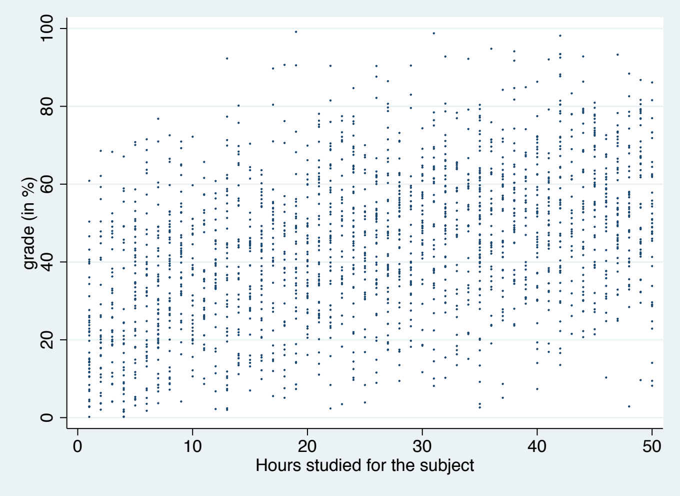
College A. Subsamples 1-3. N=80



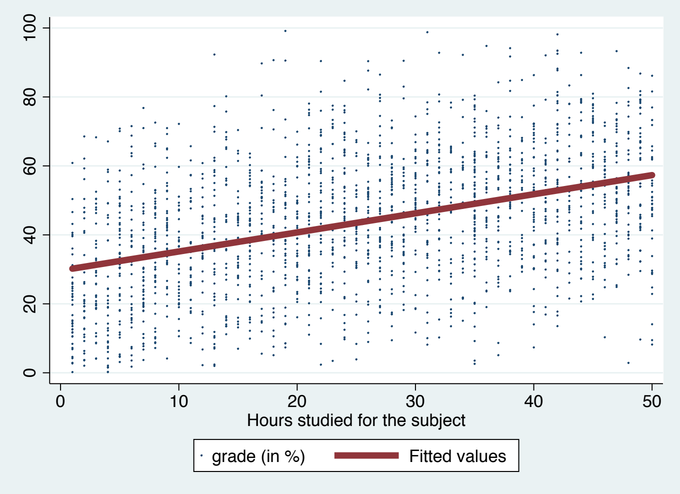
Stars*,**,*** and standard errors

- Subsample 1: $\hat{\beta} = 1.07$ st.error $\hat{\beta} = 0.080$
- Subsample 2: $\hat{\beta} = 1.45$ st.error $\hat{\beta} = 0.084$
- Subsample 3: $\hat{\beta} = 0.30$ st.error $\hat{\beta} = 0.098$
- Note: standard errors roughly the same

College B. N=2,000



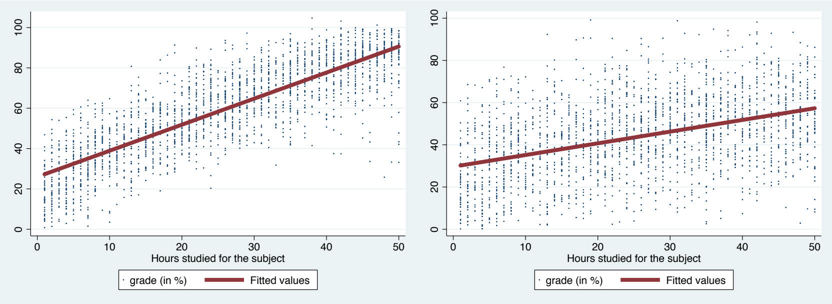
College B. N=2,000 $\beta_0 = 1.29$ $\alpha_0 = 29.64$



College B. N=2,000 $\beta_0 = 0.55$ $\alpha_0 = 29.64$

- Note: slope positive as in College A (i.e. similar "effect")
- However, one can readily see that estimates will not be as precise

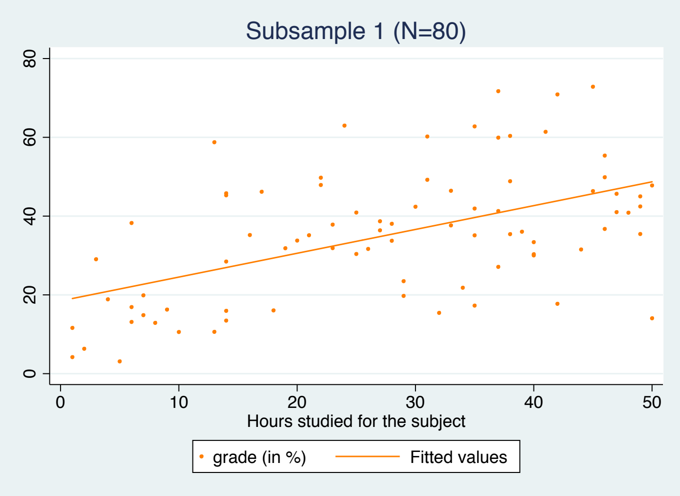
College A vs B. N=2,000 $\beta_0^A = 1.29$ vs $\beta_0^B = 0.55$



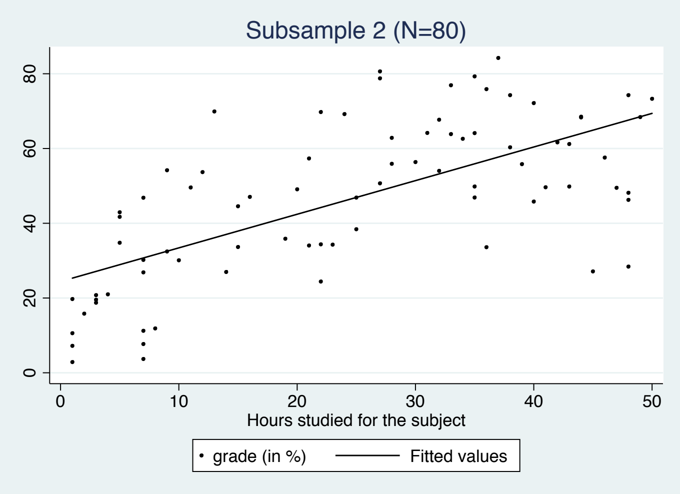
Stars*,**,*** and standard errors

- When we take subsamples, how sure can we be that the effect of studying longer hours is positive in College B?

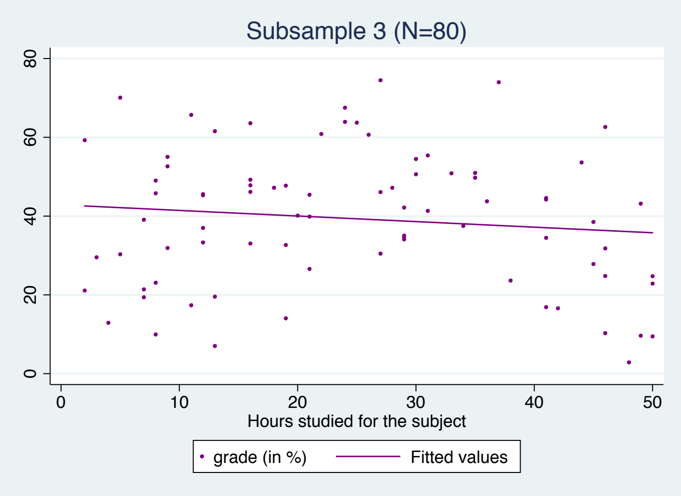
College B. Subsample 1. N=80 $\hat{\beta} = 0.60$ $\hat{\alpha} = 18.47$



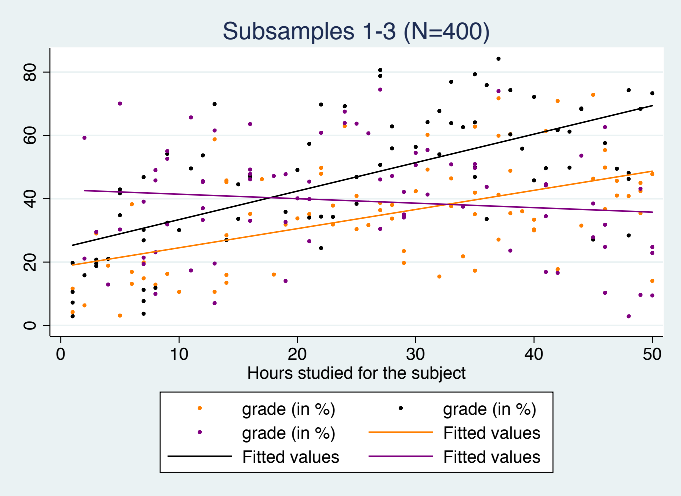
College B. Subsample 2. N=80 $\hat{\beta} = 0.89$ $\hat{\alpha} = 24.41$



College B. Subsample 3. N=80 $\hat{\beta} = -0.141$ $\hat{\alpha} = 42.86$



College B. Subsamples 1-3. N=80



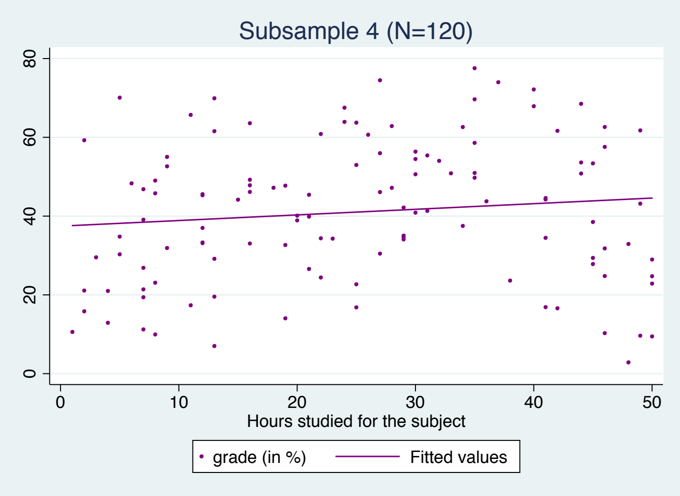
Stars*,**,*** and standard errors

- Subsample 1: $\hat{\beta} = 0.60$ st.error $\hat{\beta} = 0.110$
- Subsample 2: $\hat{\beta} = 0.89$ st.error $\hat{\beta} = 0.116$
- Subsample 3: $\hat{\beta} = -0.141$ st.error $\hat{\beta} = 0.134$
- Note: standard errors roughly the same
- Note: standard errors closer in size to coefficient $\Rightarrow$ coefficient is less precise
- We are less sure about the size and significance of the effect

Stars*,**,*** and standard errors

- Suppose we take the following random sample (N=120).
- What can we conclude?

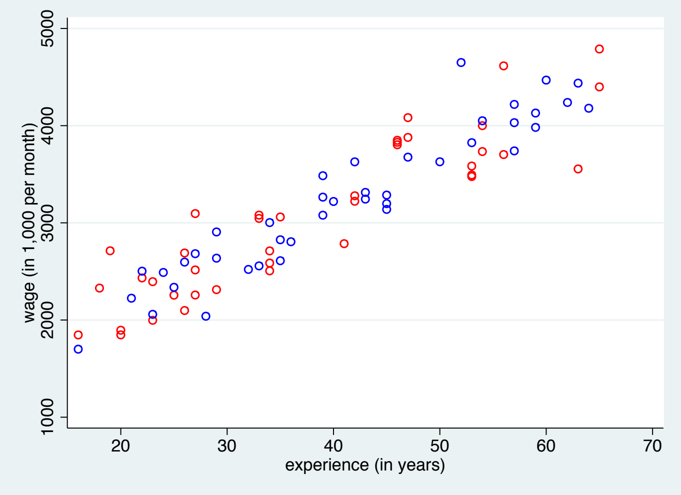
College B. Subsample 4. $\hat{\beta} = 0.142$ st.err.$(\hat{\beta}) = 0.111$



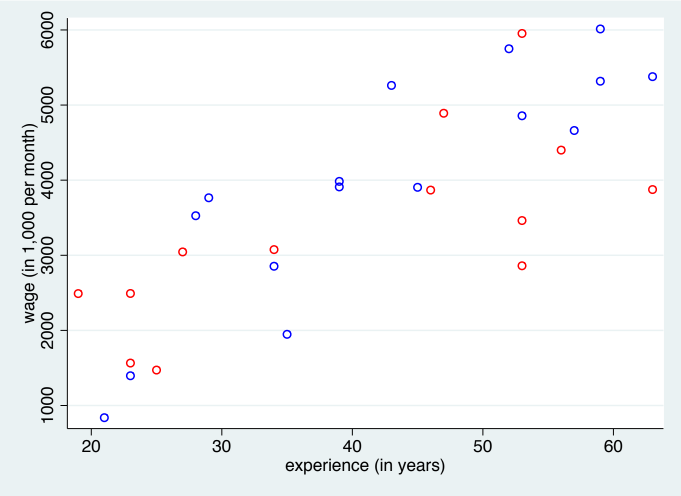
Stars*,**,*** and standard errors

- Suppose we take the following random sample (N=120).
- What can we conclude? Not much in this case
- This is why standard errors are critical: they tell us up to what point we can conclude that the effect truly exists
- The key is NOT the size of the coefficient: the key is the ratio of the estimated coefficient to its estimated standard error
- Roughly: $\frac{\hat{\beta}}{\hat{s}_{\hat{\beta}}} \geq 2$ We can conclude there is an effect

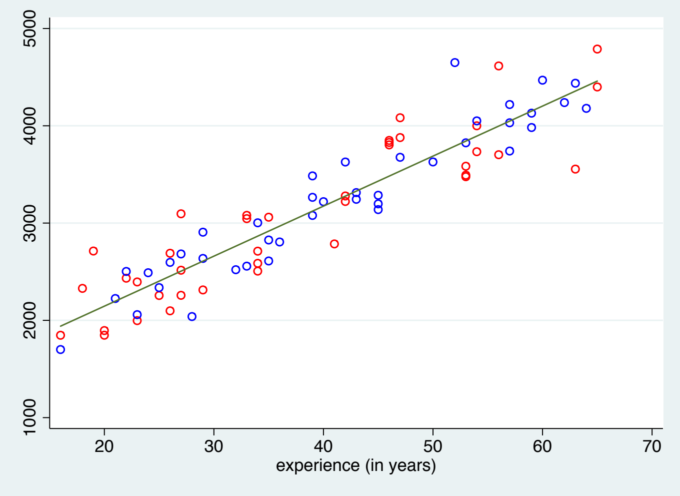
Population 1



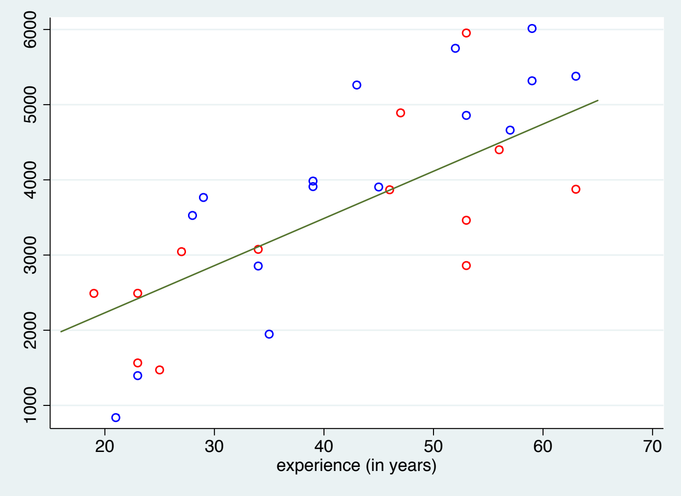
Population 2



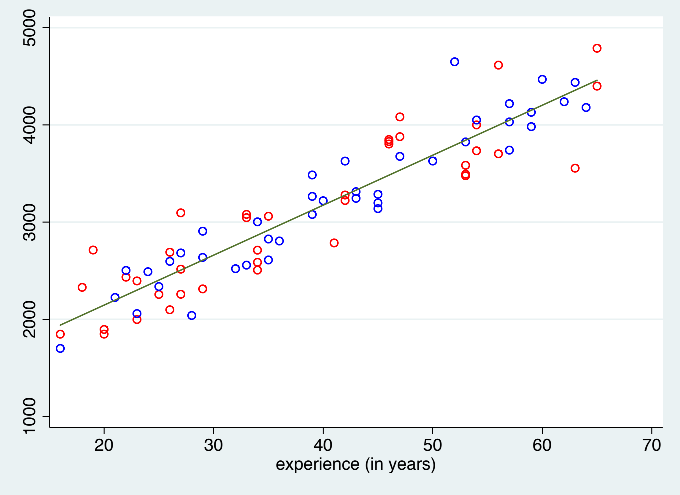
Population 1: linear fit



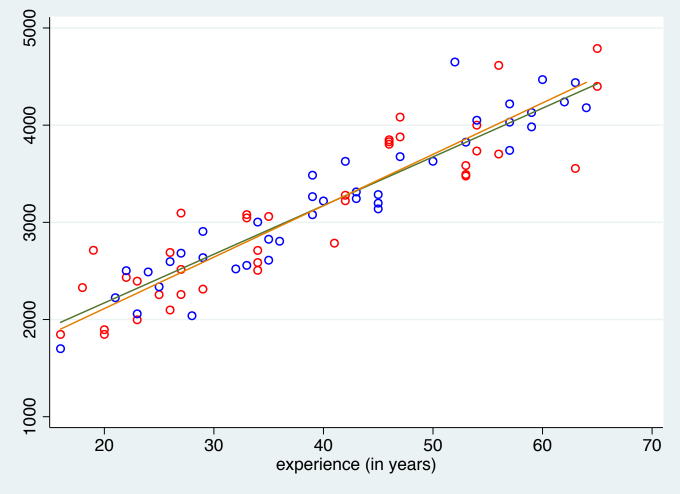
Population 2: linear fit



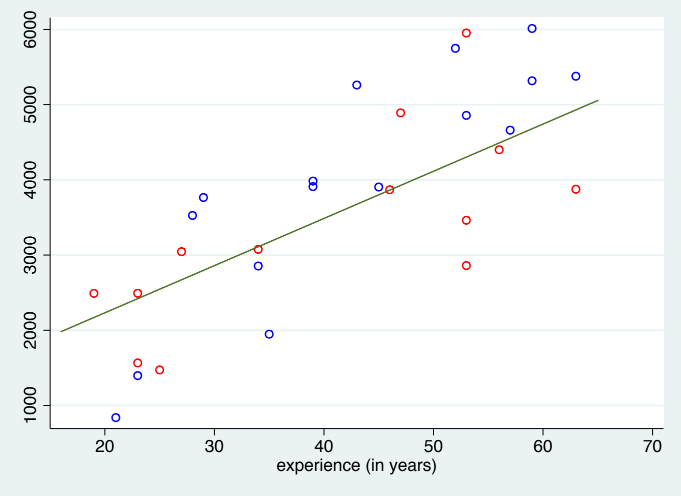
Population 1: linear fit



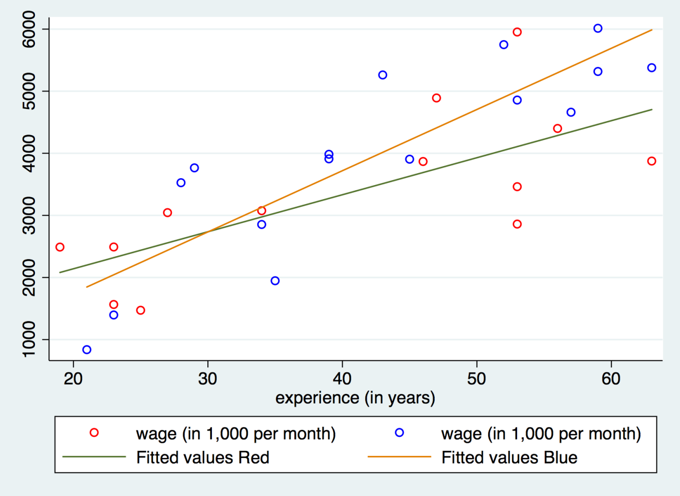
Population 1: linear fit



Population 2: linear fit



Population 2: linear fit



Comments on standard errors

- Population 1: it does not matter whether we take red or blue sample
- Population 2: it *does* matter whether we take red or blue sample
- $\Rightarrow$ Estimator ($\hat{\beta}$) is less precise for second case (it varies more)

Comments on standard errors

- Population 1: it does not matter whether we take red or blue sample
- Population 2: it *does* matter whether we take red or blue sample
- $\Rightarrow$ Estimator ($\hat{\beta}$) is less precise for second case (it varies more)
- Standard errors are a measure of that precision: how sure we can be that the actual β is actually the number we found
- Standard errors $\rightarrow$ 0 we can be quite sure that the *real* value of β is close to the one we estimated
- Standard errors $\rightarrow$ BIG we have no idea whether the *real* value of β is close to the one we estimated

Comments on significance level

- Rule of thumb: the smaller the Standard errors, the better (the more precise the estimate)
- Stars: the smaller the the Standard errors, the more stars we will have.
- Stars, usual meaning
    - *: Significant at 10% level
    - **: Significant at 5% level
    - ***: Significant at 1% level

Comments on significance level

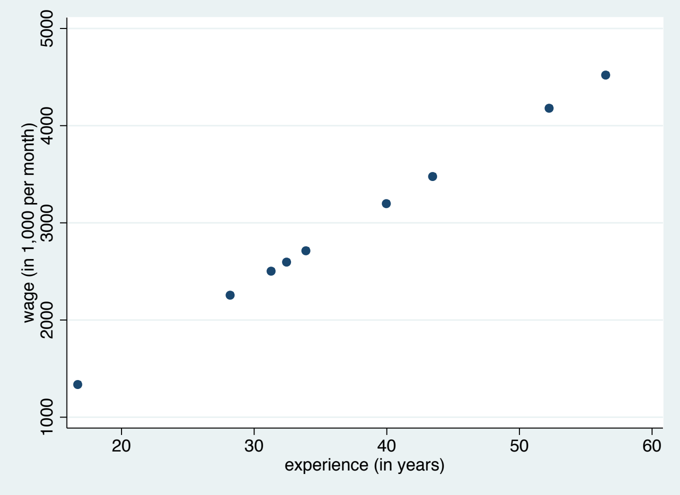
Rough interpretation (not technical one!) of stars

- *: Significant at 10% level: there is a 90% chance that the value we found is different from 0, 10% chances that it is actually 0
- **: Significant at 5% level: there is a 95% chance that the value we found is different from 0, 5% chances that it is actually 0
- ***: Significant at 1% level: there is a 99% chance that the value we found is different from 0, 1% chances that it is actually 0

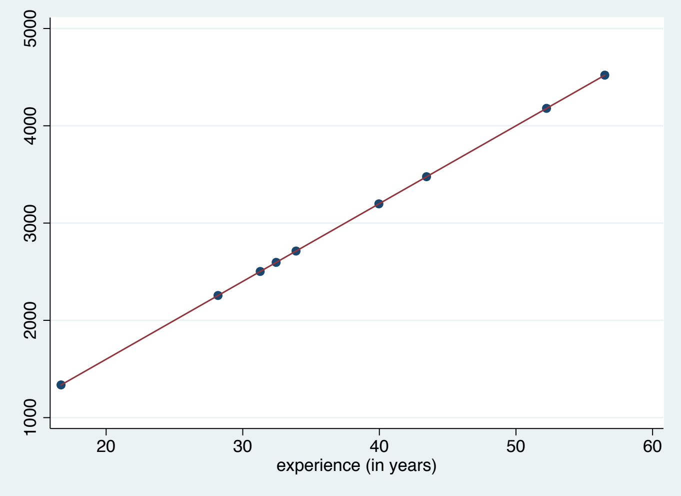
Comments on significance level

- *: Significant at 10% level: Assuming under null hypothesis is correct (that is, that $\beta = 0$), the probability that we find a value of $\hat{\beta}$ at least as different from 0 as the one we have found is at most 10% [1]
- **,***: (alt.) We are confident that β is different from 0 at a 10% significance level.
- Note: with ***, the probability is so small, that you can actually be quite confident that the true value is different than 0.

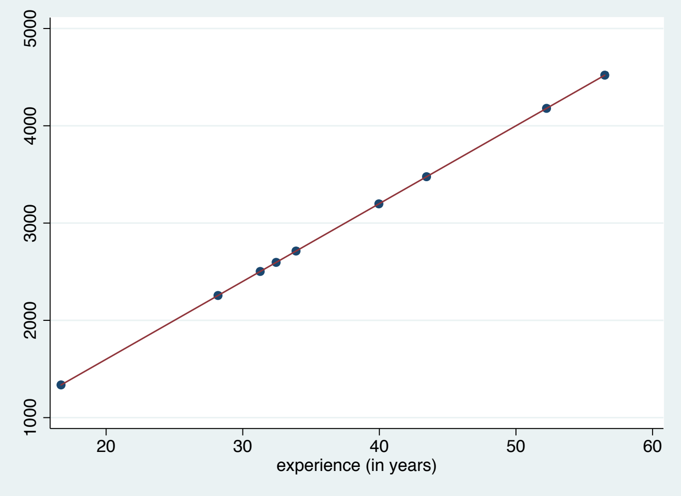
Question: Guess the standard error of $\hat{\beta}$



Guess the standard error of $\hat{\beta}$



Guess the standard error of $\hat{\beta}$: 0



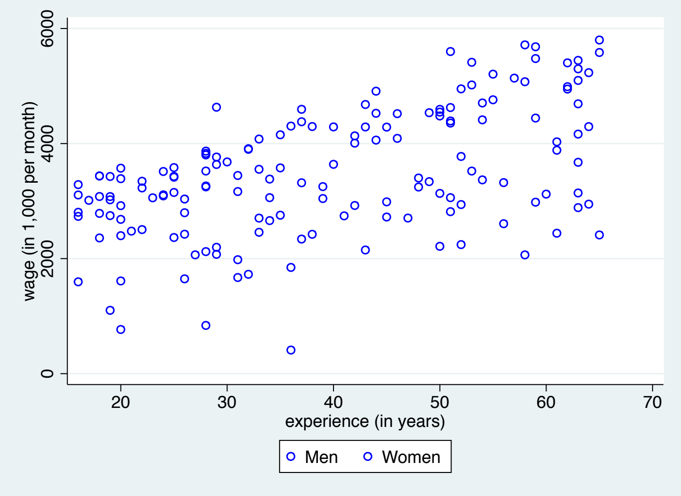
Guess the standard error of $\hat{\beta}$: 0

(Assuming it's a proper random sample, there is nothing that will make you suspect that β can be any different than the $\hat{\beta}$ you have found).

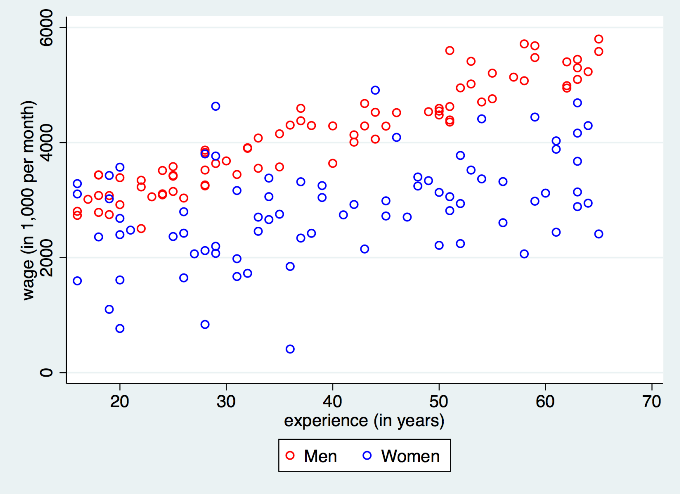
What if you feel men and women don't earn the same?

- You think experience increases wages equally for both
- But you suspect that wages for men are larger than for women overall

What if you feel men and women don't earn the same?



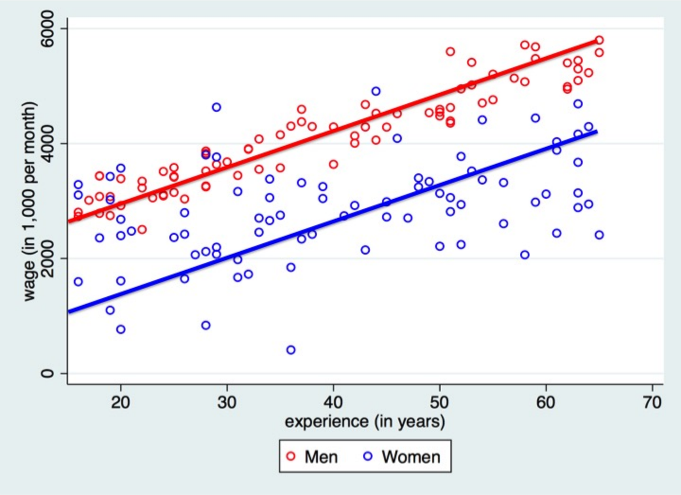
What if you feel men and women don't earn the same?



What if you feel men and women don't earn the same?

- You think experience increases wages equally for both
- But you suspect that wages for men are larger than for women overall

What if you feel men and women don't earn the same?

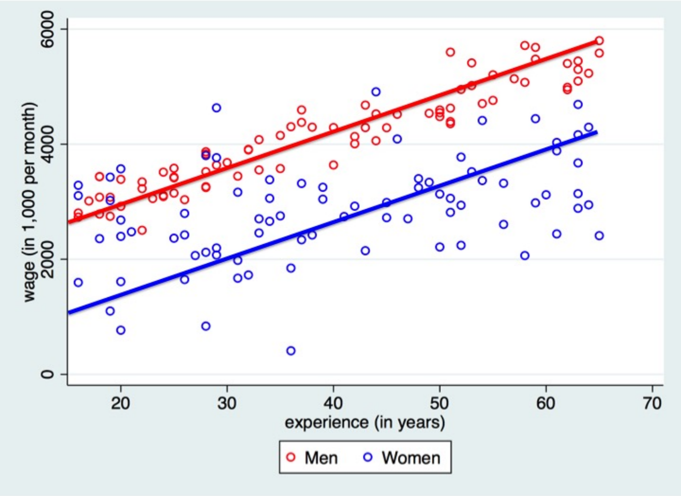


What if you feel men and women don't earn the same?

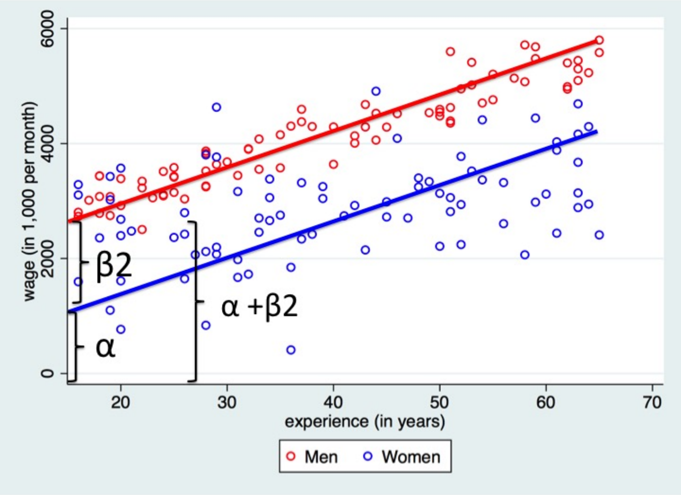
$$\omega = \alpha + \beta_1 \times \mathit{EXP} + \beta_2 \mathit{MALE} + \varepsilon$$

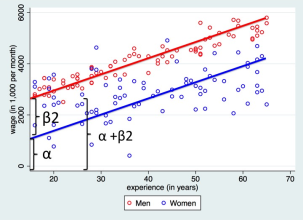
MALE = Dummy for men (=takes value 1 if individual is male, 0 otherwise)

$\omega = \alpha + \beta_1 \times EXP + \beta_2 MALE + \varepsilon$



$\omega = \alpha + \beta_1 \times EXP + \beta_2 MALE + \varepsilon$





$$\omega = \alpha + \beta_1 \times EXP + \beta_2 MALE + \varepsilon$$

$\hat{\alpha}$: Predicted wage if experience is 0 years for women
$\hat{\beta}_2 + \hat{\alpha}$: Predicted wage if experience is 0 years for men
$\hat{\beta}_2$: Predicted wage difference between men and women *at any experience level*

Comment

- Many tables will NOT refer to $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$,...
- They will refer to the name of the variable itself: "constant", "experience", "male",....

Why like this instead of two regressions?

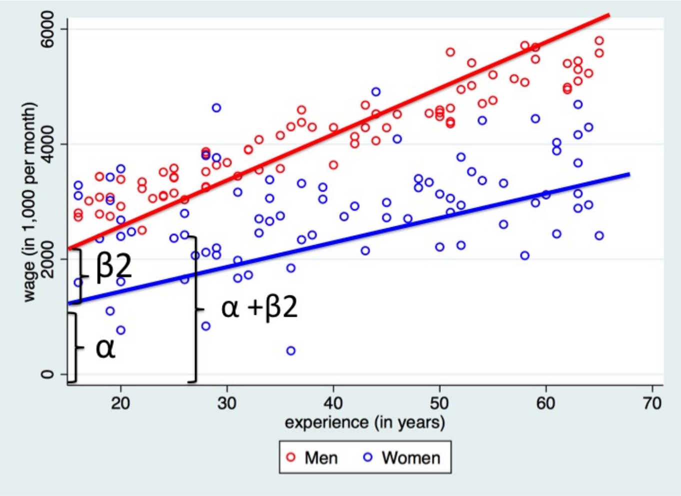
- Why $\omega = \alpha + \beta_1 \times EXP + \beta_2 MALE + \varepsilon$ instead of
- $\omega_f = \alpha_f + \beta_f \times EXP + \varepsilon_f$ AND $\omega_m = \alpha_m + \beta_m \times EXP + \varepsilon_m$?

- Using more information
- Can compare if men and women are different
- As easy as to check whether the associated $\hat{\beta}_2$ is significantly different from 0

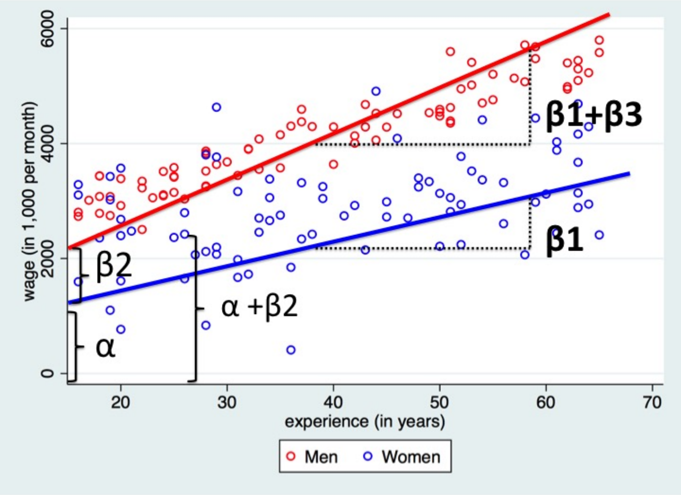
What if you feel men and women don't earn the same?

- Now you no longer think experience increases wages equally for both
- You suspect that men benefit more from experience than women *(experience has a differential effect through gender)*
- But you still suspect that unconditional wages for men are larger than for women overall (i.e. if experience is 0)
- $\Rightarrow \omega = \alpha + \beta_1 \times EXP + \beta_2 MALE + \beta_3 MALE \times EXP + \varepsilon$

$\omega = \alpha + \beta_1 \times EXP + \beta_2 MALE + \beta_3 MALE \times EXP + \varepsilon$



$\omega = \alpha + \beta_1 \times EXP + \beta_2 MALE + \beta_3 MALE \times EXP + \varepsilon$



$\omega = \alpha + \beta_1 \times EXP + \beta_2 MALE + \beta_3 MALE \times EXP + \varepsilon$

- "β_3" or "MALE $\times$ EXP": how much more an extra year of experience yields to men make with respect to non-men

$\omega = \alpha + \beta_1 \times EXP + \beta_2 MALE + \beta_3 MALE \times EXP + \varepsilon$

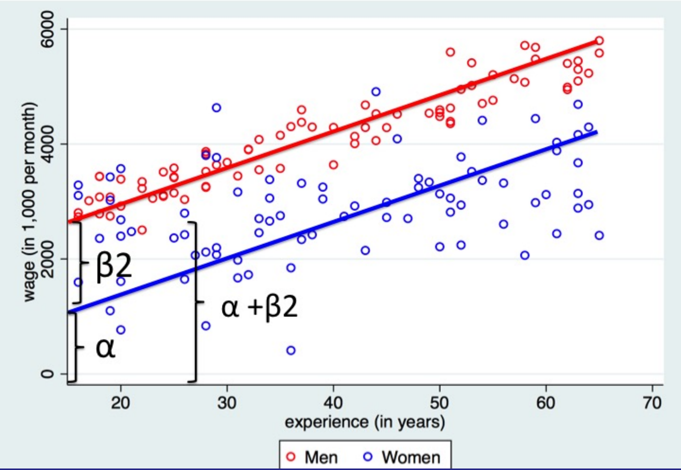
- Q: what happens if there is no differential effect of experience on gender?

$\omega = \alpha + \beta_1 \times EXP + \beta_2 MALE + \beta_3 MALE \times EXP + \varepsilon$

- Q: what happens if there is no differential effect of experience on gender?
- A: Then we are back to the first case, and $\hat{\beta}_3 = 0$
- In practical terms, there are no starts associated to $\hat{\beta}_3$ in the table
- If $\hat{\beta}_3 = 0$ we are in the situation as before (see next slide)

$\omega = \alpha + \beta_1 \times EXP + \beta_2 MALE + \beta_3 MALE \times EXP + \varepsilon$

Case when $\hat{\beta}_3 = 0$



This is how a typical table will look

Dependent variable: wage	(1)	(2)	(3)
EXPERIENCE	38.41***	40.13***	25.77***
	(4.83)	(3,38)	(4.56)
MALE		1,320***	203.23
		(102.12)	(270.13)
MALE $\times$ EXPERIENCE			28.24***
			(6.39)
Constant	1,979***	1,236***	1,813***
	(204.01)	(154.23)	195.84
R^2	0.28	0.65	0.69
Observations	160	160	160

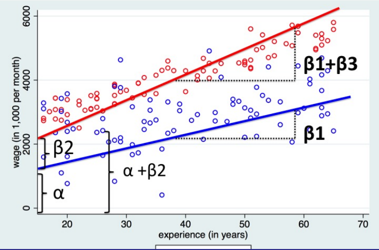
Standard errors in parenthesis.*: significant at 10% level, **: significant at 5% level, ***: significant at 1% level.

Interpretation of (3)

- Experience has a positive effect on people's wages: for each extra year of tenure, predicted monthly salary goes up by $25 on average.
- Experience has *an even larger* positive effect for men: for each extra year of tenure, predicted monthly salary goes up by $25 + $28 on average.
- In other words, in (3) "Experience" refers to the effect of experience *for everyone*, "Experience $\times$ Male" refers to the added effect for men (on top of the general effect)
- We cannot find an gender differential effect independent of experience. That is, if men and women have no experience, there is not enough evidence to state that their wages will be different on average (note the lack of $*$)

Interpretation of (3)

- We cannot find an gender differential effect independent of experience. That is, if men and women have no experience, there is not enough evidence to state that their wages will be different on average (note the lack of $*$)
- As you can see in the graph above, $\hat{\beta}_2$ is indeed positive. What the regression tells us, though, is that we do not have enough information to conclude that it is indeed different than 0



Question

- Suppose we have Singapore data and the model is $\omega = \alpha + \beta_1 \times \mathit{Malay} + \beta_2 \times \mathit{Indian} + \beta_3 \times \mathit{Angmoh} + \varepsilon$
- Say $\hat{\beta}_2 = -35$. What does this mean?

Question

- Suppose we have Singapore data and the model is $\omega = \alpha + \beta_1 \times \mathit{Malay} + \beta_2 \times \mathit{Indian} + \beta_3 \times \mathit{Angmoh} + \varepsilon$
- Say $\hat{\beta}_2 = -35$. What does this mean? That on average, the predicted wage of an Indian is $35 less than for someone who is neither Malay, nor Indian nor Angmoh.

Question

- Suppose we have Singapore data and the model is $\omega = \alpha + \beta_1 \times \mathit{Malay} + \beta_2 \times \mathit{Indian} + \beta_3 \times \mathit{Angmoh} + \varepsilon$
- Say $\hat{\beta}_2 = -35$. What does this mean? That on average, the predicted wage of an Indian is \$35 less than for someone who is neither Malay, nor Indian nor Angmoh.
- Say $\hat{\beta}_2 = -35$. The predicted wage of an Indian is \$35 less than that of the omitted category

Question

- Suppose we have Singapore data and the model is
$\omega = \alpha + \beta_1 \times \mathit{Malay} + \beta_2 \times \mathit{Indian} \ + \beta_3 \times \mathit{Angmoh} + \varepsilon$
- Say $\hat{\beta}_2 = -35$. What does this mean? That on average, the predicted wage of an Indian is \$35 less than for someone who is neither Malay, nor Indian nor Angmoh.
- Say $\hat{\beta}_2 = -35$. The predicted wage of an Indian is \$35 less than that of the omitted category
- Say $\hat{\beta}_2 = -35$. The predicted wage of an Indian is \$35 less than that of the average Chinese Singaporean.