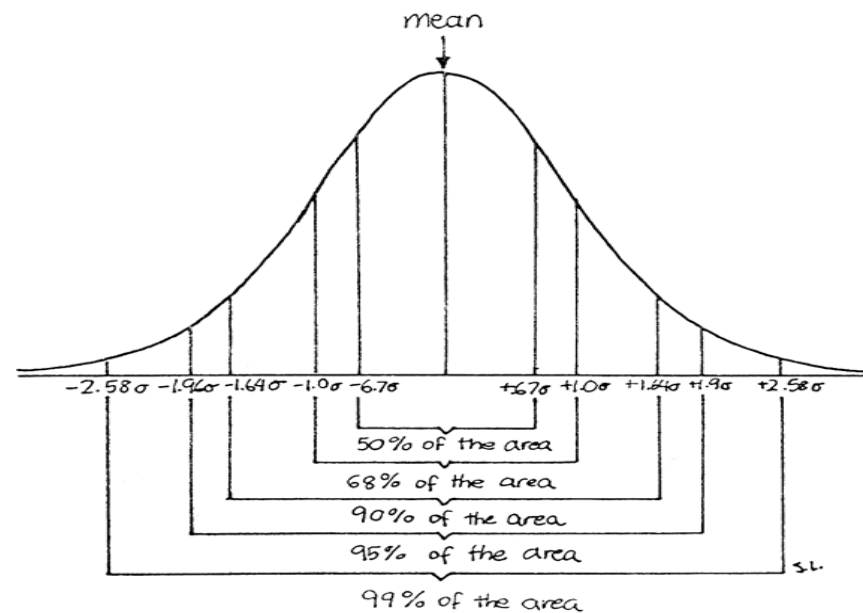


Quantitative Economics - Review

Ian Crawford
Department of Economics & New College



“We have normality. I repeat, we have normality. Anything you still can’t cope with is therefore your own problem.”

Douglas Adams, *The Hitchhikers Guide to the Galaxy*

Outline

1. Understanding Regression Results
 - What is linear regression?
 - Interpreting regression results

2. Estimation MLE
 - The intuition, independence and Bayes rule
 - The recipe.

What is a linear regression?

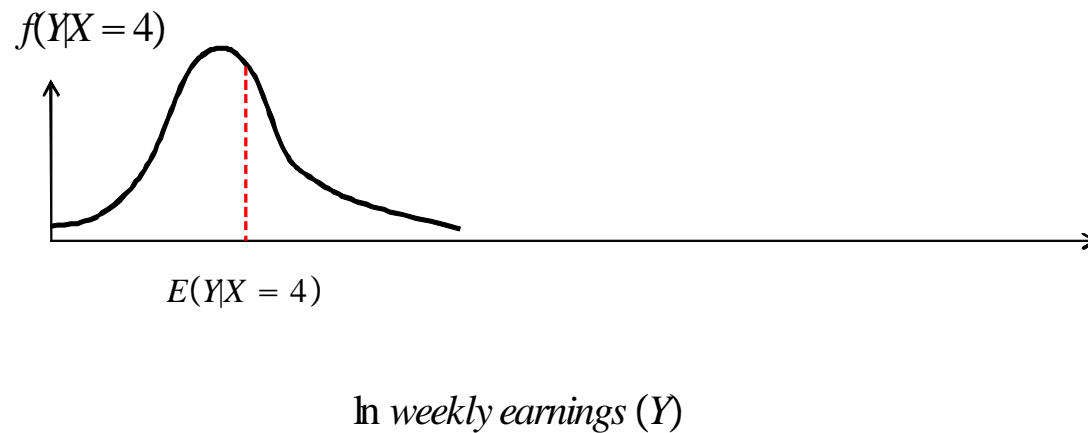
It's a description of how the mean of something (Y) varies as you change something else (X) or a bunch of other things (X_1, X_2, \dots, X_K)

In the lingo of econometrics linear regression tries to capture the “conditional expectation function” (CEF).

This tells you the expected value of Y given X .

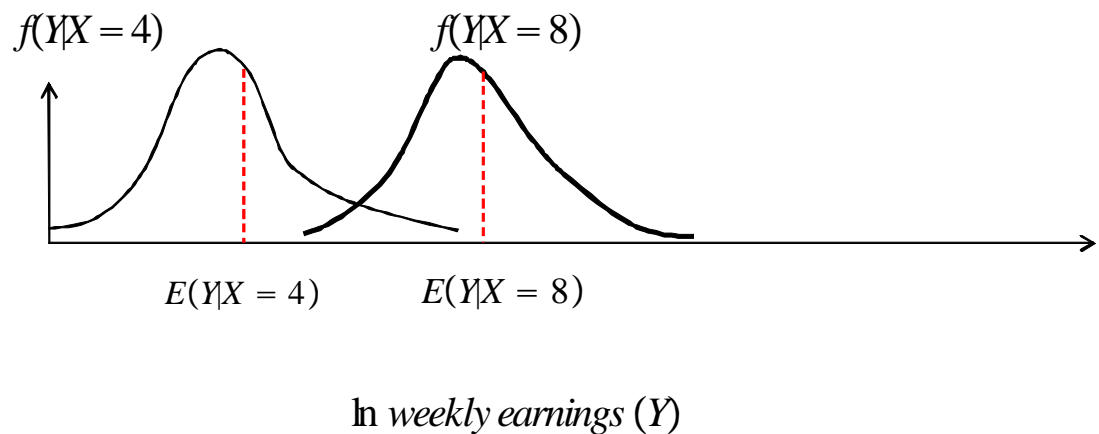
[NB it is quite helpful to forget completely about random variables, statistics etc. for the moment and imagine that we have data on the whole population.]

Example: $E(\log \text{ weekly earning} | \text{ years of completed education})$



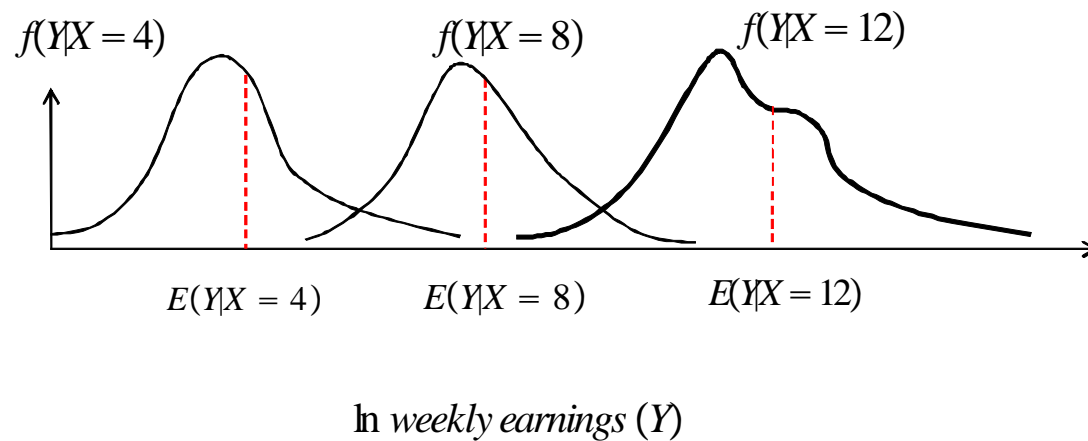
Here we draw the pdf (smoothed histogram) for the distribution of log weekly earnings just for those who have completed 4 years of education.

Example: $E(\log \text{ weekly earning} | \text{ years of completed education})$

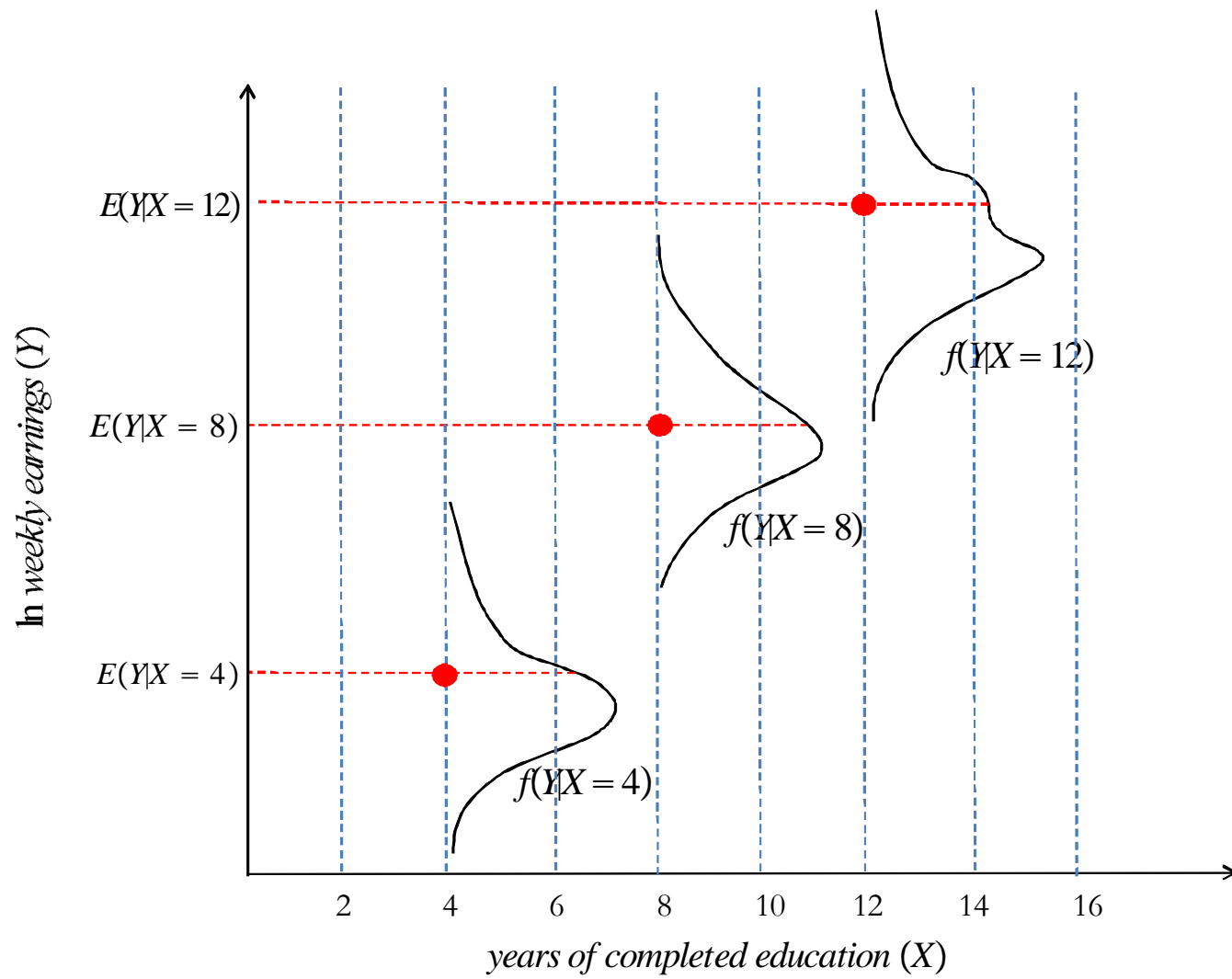


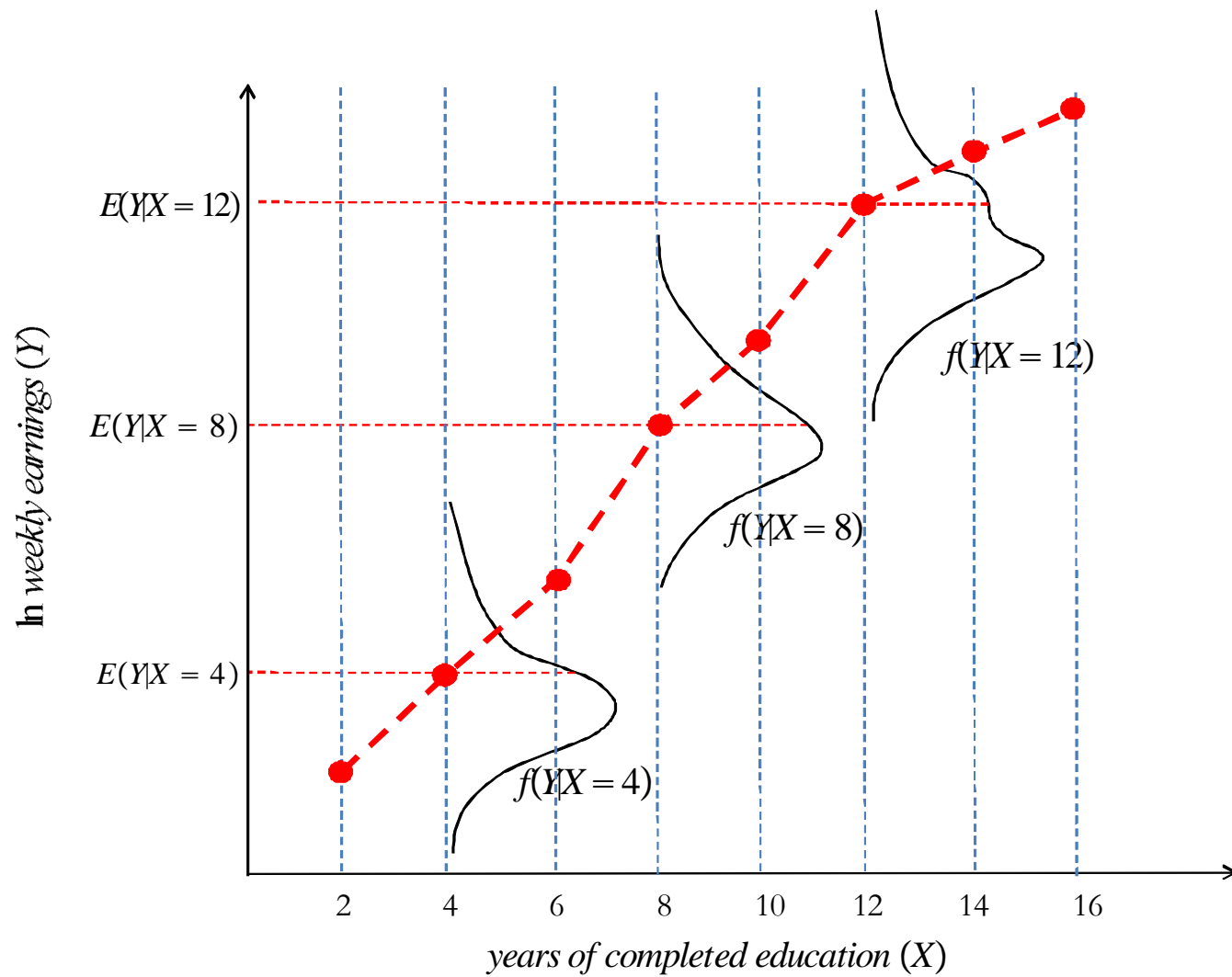
Now we add the pdf of log weekly earnings for those who have completed 8 years of education (it's higher, on average).

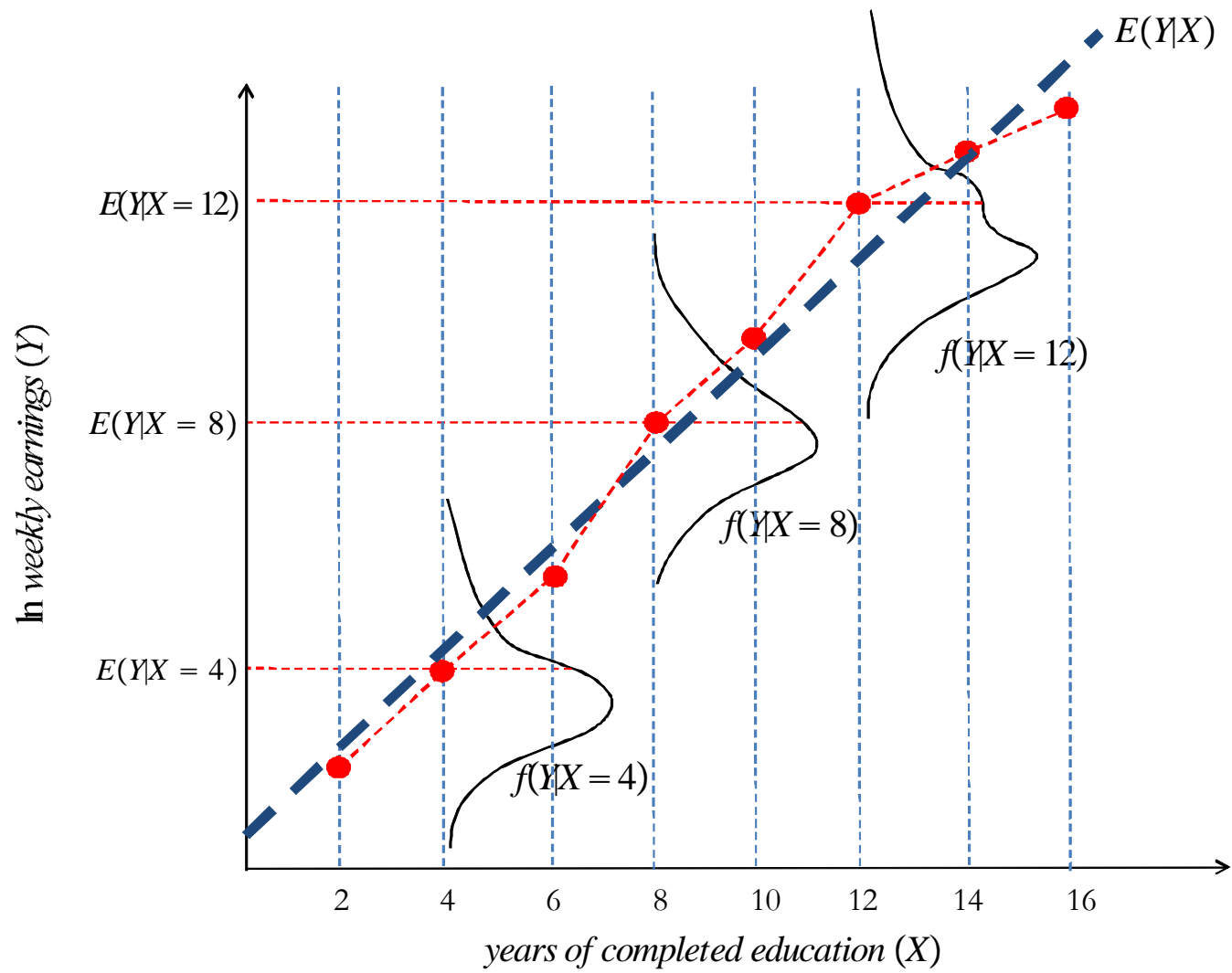
Example: $E(\log \text{ weekly earning} | \text{ years of completed education})$



Now add the pdf of log weekly earnings for those who have completed 12 years of education (it's higher still, on average).







The CEF connects up the red dots (the conditional expectations).

The regression line is a linear approximation of the CEF.

If you're interested in the CEF then you should be interested in the regression line because:

- If the CEF is linear the regression line is it.
- If the CEF is non-linear then the regression line gives the *best* linear approximation to it.
- The regression line is the *best* linear predictor of Y given X in any case.

There are a bunch of ways of actually fitting the regression line

E.g. "least squares", "least absolute deviations", "maximum likelihood", ...there are others.

They all depend on exactly what you mean by "best".

The dominant method in econometrics is probably MLE which has been the main focus in this course.

I'll come back to MLE later.

Understanding and Interpreting Regression Results

It's an important "learning outcome" that you can interpret regression results.

- What are their economic implications?
- Are those implications statistically significant?
- Is the regression reliable?

It is essential that you can discuss regression output coherently/critically.

You might usefully structure any discussion of regression results in this way:

1. Interpretation of the parameters - economic
2. Interpretation of the parameters - statistical
3. Interpretation of the overall robustness of the regression.

In what follows I have used the data from the NSW trial (pre treatment sample) which you may have looked at in your tutorial on programme evaluation.

It's the pre-treatment data only so you can forget about the impact of training etc. and think of this as a simple study of the determinants* of earnings.

There are 433 observations (workers) for whom we observe their annual earnings and a number of "covariates" (translation: things about them).

We going to run a regression - either by MLE/OLS, it hardly matters - to try to capture the conditional expectation function (CEF).

Remember: the CEF will tell us the expected value of their earnings for different values of the covariates. As we change the values of the covariates we can see how *on average* their earnings will change.

*I slipped in a causal interpretation there - we'll come back to this.

Variable	Obs	Mean	Std. Dev.	Min	Max
Earnings	433	5073.837	5701.513	74.34345	37431.66
age	433	23.69053	5.703015	17	50
educ	433	10.42725	1.516849	4	15
black	433	.7713626	.420441	0	1
married	433	.1755196	.380851	0	1

Note the mix of continuous (Earnings), discrete/categorical (age, educ), and binary (black, married) variables.

The regression equation:

$$Earnings_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Educ_i + \beta_4 Black_i + \beta_5 Married_i + u_i$$

The estimation method gives us "best" estimates for $\{\beta_0, \dots, \beta_5\}$.

Source	SS	df	MS			
Model	1.9654e+09	5	393078432	Number of obs =	433	
Residual	1.2078e+10	427	28285101.9	F(5, 427) =	13.90	
				Prob > F =	0.0000	
				R-squared =	0.1400	
				Adj R-squared =	0.1299	
				Root MSE =	5318.4	
Total	1.4043e+10	432	32507247			

Earnings	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_cons	-13116.42	3724.211	-3.52	0.000	-20436.49	-5796.352
age	1273.129	282.1772	4.51	0.000	718.4999	1827.758
age2	-19.58934	5.027359	-3.90	0.000	-29.47079	-9.707887
educ	1.79439	175.4027	0.01	0.992	-342.9658	346.5546
black	-1172.18	612.6281	-1.91	0.056	-2376.322	31.9623
married	3102.903	700.9686	4.43	0.000	1725.125	4480.682

[NB regression output doesn't all ways look exactly like this - but most of what you see here, will be presented in some form or other]

1. Interpretation of the parameters - economic

Source	SS	df	MS	Number of obs = 433		
Model	1.9654e+09	5	393078432	F(5, 427)	=	13.90
Residual	1.2078e+10	427	28285101.9	Prob > F	=	0.0000
Total	1.4043e+10	432	32507247	R-squared	=	0.1400
				Adj R-squared	=	0.1299
				Root MSE	=	5318.4

Earnings	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_cons	-13116.42	3724.211	-3.52	0.000	-20436.49	-5796.352
age	1273.129	282.1772	4.51	0.000	718.4999	1827.758
age2	-19.58934	5.027359	-3.90	0.000	-29.47079	-9.707887
educ	1.79439	175.4027	0.01	0.992	-342.9658	346.5546
black	-1172.18	612.6281	-1.91	0.056	-2376.322	31.9623
married	3102.903	700.9686	4.43	0.000	1725.125	4480.682

$$\begin{aligned}
 Earnings_i = & -13116.42 + 1273.129Age_i - 19.58934Age_i^2 \\
 & + 1.79439Educ_i - 1172.18Black_i + 3102.903Married_i + u_i
 \end{aligned}$$

I'm going to compress the notation a bit and dump some decimal places:

$$Y_i = -13116.4 + 1273.1A_i - 19.6A_i^2 \\ + 1.8S_i - 1172.2B_i + 3102.9M_i + u_i$$

Since the expected value of the error term (u_i) is zero by construction (in OLS and MLE) another way to write the regression results is as the conditional expectation:

$$E(Y_i | A_i, A_i^2, S_i, B_i, M_i) = \hat{Y}_i = \\ -13116.4 + 1273.1A_i - 19.6A_i^2 + 1.8S_i - 1172.2B_i + 3102.9M_i$$

The constant:

The constant tells you the expected value of earnings given everything else is zero:

$$E(Y_i | 0, 0, 0, 0, 0, 0) = -13116.4$$

$S_i = 0 \Rightarrow$ completed no years of education

$B_i = 0 \Rightarrow$ means non-black individuals

$M_i = 0 \Rightarrow$ means unmarried

$A_i = 0 \Rightarrow$ means unborn

–\$13116.4 is the expected earnings of uneducated, non-black, single *foetuses*.

The youngest person in our dataset was 17 so this is far outside of the range of the data.

The slopes:

As a first approximation you interpret them as partial derivatives.

$$E(Y_i|\dots) = -13116.4 + 1273.1A_i - 19.6A_i^2 + 1.8S_i - 1172.2B_i + 3102.9M_i$$

So

$$\frac{\partial E(Y_i|\dots)}{\partial A_i} = 1273.1 - 40.2A_i$$

is the *ceteris paribus* effect of a marginal change in age on expected earnings. Since A_i is quadratic the marginal effects are linear and decreasing. Implication: the effect of age on earnings is increasing to begin with but eventually starts to go down. It is maximised at about 32 years.

But ...

This only makes complete sense for continuous variables.

For example, interpreting

$$\frac{\partial E(Y_i|\dots)}{\partial M_i} = 3102.9$$

as the effect of “a marginal increase in being married” and earnings doesn’t make sense. Marriage is binary (on/off). So for binary explanatory variables the coefficient is

$$\frac{\Delta E(Y_i|\dots)}{\Delta M_i} = 3102.9$$

which is the effect on earnings of being married versus not being married. (Think about plugging in $M_i = 0$ and predicting earnings, then setting $M_i = 1$ and doing it again: the difference between the two predictions would be \$3102.9)

Another caveat concerns *discrete/categorical* variables like education which is measured in lumps (completed years).

In this case the derivative interpretation doesn't make complete sense either as you can only vary education in lumps of one year.

So

$$\frac{\Delta E(Y_i | \dots)}{\Delta S_i} = 1.8$$

is interpreted as the effects of an additional one year of education on earnings.

Source	SS	df	MS	Number of obs =	433
Model	1.9654e+09	5	393078432	F(5, 427) =	13.90
Residual	1.2078e+10	427	28285101.9	Prob > F =	0.0000
Total	1.4043e+10	432	32507247	R-squared =	0.1400
				Adj R-squared =	0.1299
				Root MSE =	5318.4

Earnings	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_cons	-13116.42	3724.211	-3.52	0.000	-20436.49	-5796.352
age	1273.129	282.1772	4.51	0.000	718.4999	1827.758
age2	-19.58934	5.027359	-3.90	0.000	-29.47079	-9.707887
educ	1.79439	175.4027	0.01	0.992	-342.9658	346.5546
black	-1172.18	612.6281	-1.91	0.056	-2376.322	31.9623
married	3102.903	700.9686	4.43	0.000	1725.125	4480.682

The term "model specification" is econometrician-speak for what you decide to put in your model and how you decide to measure/transform your data.

By using different explanatory variables or transforming your data you're changing the specification of the regression equation.

When you have a dependent variable which cannot be negative like Earnings a "popular" specification is to log it Why?

(a) it stops you predicting negative earnings

(b) it imposes a principal implication of the Mincer model - that education has a constant *proportional* effect on earnings.

Source	SS	df	MS	Number of obs = 433		
Model	82.43658	5	16.487316	F(5, 427)	=	14.96
Residual	470.690927	427	1.10232067	Prob > F	=	0.0000
-----				R-squared	=	0.1490
Total	553.127507	432	1.28038775	Adj R-squared	=	0.1391
-----				Root MSE	=	1.0499
ln(Earnings)	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_cons	4.25775	.7352066	5.79	0.000	2.812676	5.702825
age	.2623196	.0557054	4.71	0.000	.1528288	.3718105
age2	-.0038868	.0009925	-3.92	0.000	-.0058375	-.0019361
educ	-.0137658	.0346267	-0.40	0.691	-.0818259	.0542942
black	-.1854204	.1209406	-1.53	0.126	-.4231334	.0522926
married	.5378583	.1383801	3.89	0.000	.2658674	.8098493

$$\ln Y_i = -4.3 + 0.26A_i - 0.003A_i^2 - 0.014S_i - 0.185B_i + 0.54M_i + u_i$$

$$E(\ln Y_i | \dots) = \widehat{\ln Y}_i = -4.3 + 0.26A_i - 0.003A_i^2 - 0.014S_i - 0.185B_i + 0.54M_i$$

We have changed the specification and all of the coefficients have changed.

Worrying?.... Not really.

Think about the interpretation of the coefficients as partial derivatives/*ceteris paribus* effects. The slopes

$$\beta_i = \frac{\partial E(\ln Y_i | \dots)}{\partial X_i}$$

capture the effects of marginal changes in the explanatory variable (X_i) on the expected value of the LOG of earnings. They are therefore interpreted as *proportional* effects on earnings.

E.g. being black reduces expected earnings by 18.54%, whilst being married increases it by 53.78%.

2. Interpretation of the parameters - statistical

We need to recognise now that this regression was performed on a *sample* from a large population.

If we had taken a different sample the results would have been different (thanks to variation in the population).

The question is - would they have been a little bit different, or a lot different?

Put another way - are our results just a matter of luck (or serendipity as Martin Ellison called it), are the economic implications we drew from them robust to sampling variation?

Statistical inference is the art of dealing with this question.

Source	SS	df	MS	Number of obs = 433		
Model	1.9654e+09	5	393078432	F(5, 427) = 13.90		
Residual	1.2078e+10	427	28285101.9	Prob > F = 0.0000		
-----				R-squared = 0.1400		
Total	1.4043e+10	432	32507247	Adj R-squared = 0.1299		
-----				Root MSE = 5318.4		

Earnings	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_cons	-13116.42	3724.211	-3.52	0.000	-20436.49	-5796.352
age	1273.129	282.1772	4.51	0.000	718.4999	1827.758
age2	-19.58934	5.027359	-3.90	0.000	-29.47079	-9.707887
educ	1.79439	175.4027	0.01	0.992	-342.9658	346.5546
black	-1172.18	612.6281	-1.91	0.056	-2376.322	31.9623
married	3102.903	700.9686	4.43	0.000	1725.125	4480.682

The key object of interest is the column of standard errors on each of the estimated coefficients.

The t -ratios, P -values and the confidence intervals are useful and save you time.

Standard errors are measures of the variability/uncertainty attached to the estimate of the coefficient due to sampling variation. Alone they are not amazingly useful as, like variances, they depend on the units in which they are expressed (here it's \$) so "bigness" is hard to judge.

But, if you form the "t-ratio"

$$t = \frac{\hat{\beta}}{\widehat{SE}(\beta)}$$

then you have something which is unit-free and which is very useful.

The bottom line: a coefficient is statistically significantly different from zero at 95% if $|t| > 1.96$ and different at 90% if $|t| > 1.645$

Source	SS	df	MS			
Model	1.9654e+09	5	393078432	Number of obs = 433		
Residual	1.2078e+10	427	28285101.9	F(5, 427) = 13.90		
Total	1.4043e+10	432	32507247	Prob > F = 0.0000		
				R-squared = 0.1400		
				Adj R-squared = 0.1299		
				Root MSE = 5318.4		

Earnings	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_cons	-13116.42	3724.211	-3.52	0.000	-20436.49	-5796.352
age	1273.129	282.1772	4.51	0.000	718.4999	1827.758
age2	-19.58934	5.027359	-3.90	0.000	-29.47079	-9.707887
educ	1.79439	175.4027	0.01	0.992	-342.9658	346.5546
black	-1172.18	612.6281	-1.91	0.056	-2376.322	31.9623
married	3102.903	700.9686	4.43	0.000	1725.125	4480.682

They all statistically significant at 95% are except for Educ and Black (which is significant at 90%).

Confidence intervals are another (more intuitive) way of telling you when a coefficient is different from zero.

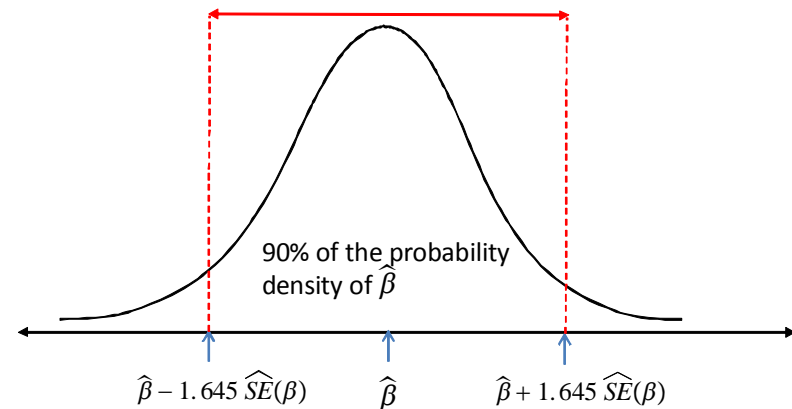
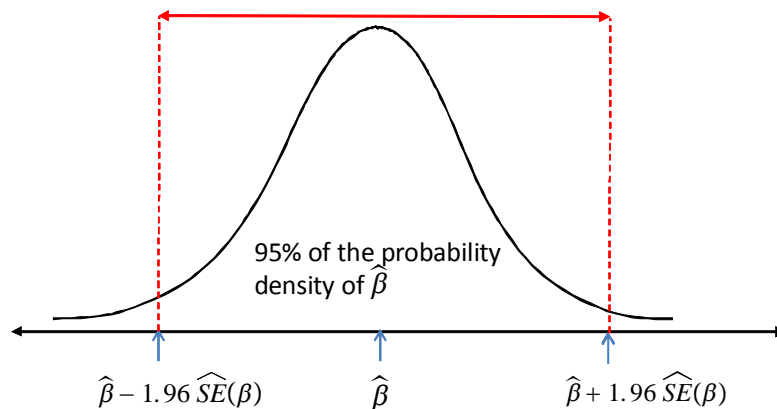
They are formed as

$$\hat{\beta} \pm 1.96 \widehat{SE}(\beta)$$

[95% confidence interval]

$$\hat{\beta} \pm 1.645 \widehat{SE}(\beta)$$

[90% confidence interval]



Source	SS	df	MS	
Model	1.9654e+09	5	393078432	Number of obs = 433
Residual	1.2078e+10	427	28285101.9	F(5, 427) = 13.90
Total	1.4043e+10	432	32507247	Prob > F = 0.0000
				R-squared = 0.1400
				Adj R-squared = 0.1299
				Root MSE = 5318.4

Earnings	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_cons	-13116.42	3724.211	-3.52	0.000	-20436.49 -5796.352
age	1273.129	282.1772	4.51	0.000	718.4999 1827.758
age2	-19.58934	5.027359	-3.90	0.000	-29.47079 -9.707887
educ	1.79439	175.4027	0.01	0.992	-342.9658 346.5546
black	-1172.18	612.6281	-1.91	0.056	-2376.322 31.9623
married	3102.903	700.9686	4.43	0.000	1725.125 4480.682

Curb your enthusiasm

Whilst the signs and magnitudes of the regression coefficients may be economically meaningful (even exciting) *statistical inference* is essential.

Without statistical inference *you have no way of knowing whether your results were driven by chance.*

For example it looks like there is a positive return to education in these results (the coefficient is positive), but the t-ratio and the confidence interval show that this is essentially down to sampling variation and there is really an insignificant (i.e. zero) effect.

Recall that we are thinking about how to structure a discussion of regression results:

1. Interpretation of the parameters - economic [✓]
2. Interpretation of the parameters - statistical [✓]
3. Interpretation of the overall robustness of the regression.

3. Interpretation of the overall robustness of the regression.

Source	SS	df	MS			
Model	1.9654e+09	5	393078432	Number of obs =	433	
Residual	1.2078e+10	427	28285101.9	F(5, 427) =	13.90	
				Prob > F =	0.0000	
				R-squared =	0.1400	
				Adj R-squared =	0.1299	
Total	1.4043e+10	432	32507247	Root MSE =	5318.4	

Earnings	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_cons	-13116.42	3724.211	-3.52	0.000	-20436.49	-5796.352
age	1273.129	282.1772	4.51	0.000	718.4999	1827.758
age2	-19.58934	5.027359	-3.90	0.000	-29.47079	-9.707887
educ	1.79439	175.4027	0.01	0.992	-342.9658	346.5546
black	-1172.18	612.6281	-1.91	0.056	-2376.322	31.9623
married	3102.903	700.9686	4.43	0.000	1725.125	4480.682

This issue here is how well has our regression captured the variation we see in the data. Have we explained most of what we see in the data, or only a small amount?

Recall (this was set as one of the exercises in a tutorial) the decomposition/identity:

$$\begin{aligned} \text{total sum of squares} &= \text{explained sum of squares} + \text{residual sum of squares} \\ TSS &= ESS + RSS \\ \sum (Y_i - \bar{Y})^2 &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum \hat{u}_i^2 \end{aligned}$$

The TSS is just the variance of the dependent variable[†].

The ESS is the variance in the predictions of the model (since $\hat{Y}_i = E(Y_i|X_i)$).

the RSS is what's left in the residuals - i.e. unexplained.

If the model has done a good job the TSS will mostly be accounted for by the ESS with the RSS being relatively unimportant.

[†]Or it would be if you divided the expression by $n - 1$.

The *R* – squared measures the explained variation (*ESS*) relative to the total variation:

$$R^2 = \frac{ESS}{TSS}$$

(there are a bunch of equivalent formulations using *TSS/RSS/ESS* but this one makes to most sense to me).

If $ESS \rightarrow TSS$ then $RSS \rightarrow 0$ and most of the variation is captured by the model and $R^2 \rightarrow 1$.

If $ESS \rightarrow 0$ then $RSS \rightarrow TSS$ and most of the variation remains in the errors (the model explains little) and $R^2 \rightarrow 0$

R^2 is a useful summary of goodness-of-fit of the model: values close to one indicate better fit than values close to zero.

Source	SS	df	MS			
Model	1.9654e+09	5	393078432	Number of obs =	433	
Residual	1.2078e+10	427	28285101.9	F(5, 427) =	13.90	
Total	1.4043e+10	432	32507247	Prob > F =	0.0000	
				R-squared =	0.1400	
				Adj R-squared =	0.1299	
				Root MSE =	5318.4	

Earnings	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_cons	-13116.42	3724.211	-3.52	0.000	-20436.49	-5796.352
age	1273.129	282.1772	4.51	0.000	718.4999	1827.758
age2	-19.58934	5.027359	-3.90	0.000	-29.47079	-9.707887
educ	1.79439	175.4027	0.01	0.992	-342.9658	346.5546
black	-1172.18	612.6281	-1.91	0.056	-2376.322	31.9623
married	3102.903	700.9686	4.43	0.000	1725.125	4480.682

In this example we can see that the RSS (1.2078×10^{10}) is an order of magnitude bigger than the ESS (1.9654×10^9) and the R^2 is consequently low (0.14).

Most of the observed variation remains unexplained by this model.

This low R^2 is not unusual in microdata - the number of observations are typically high and individual behaviour/circumstances are highly heterogeneous.

But is it too low? Does it mean that the regression is rubbish? We need to know: how low is "too low"?

Of course the TSS/ESS/RSS are all subject to *sampling variation* (if we'd sampled different people and done the same regression the results would have been different).

That means that we can formulate another statistical test to tell us whether the R^2 is statistically different from zero.

$$\frac{R^2 / (k - 1)}{(1 - R^2) / (n - k)}$$

where k is the number of coefficients, has an F -distribution.

Number of obs	=	433
F(5, 427)	=	13.90
Prob > F	=	0.0000
R-squared	=	0.1400
Adj R-squared	=	0.1299
Root MSE	=	5318.4

The 95% critical value from the F -distribution is 2.23. Since $13.90 > 2.23$ we can be 95% confident that the R^2 is bigger than zero. That doesn't sound like a big deal but it does mean our regression equation is explaining *something* and that the low R^2 isn't just down to blind luck/serendipity.

A little bit more formally

To anyone actually doing applied work, hypothesis testing is little more than running your eye over the t-ratios and the R^2 and or F value.

However, it's worth being a little more formal about it and showing you how to conduct a proper hypothesis test.

Let's focus on the coefficient on the "Black" binary variable.

$$\begin{aligned}\hat{\beta} &= -1172.18 \\ \widehat{SE}(\beta) &= 612.6281\end{aligned}$$

We don't need to do a hypothesis test about our estimate $\hat{\beta}$ we *know* what that is.

What we want to know about is the true (population) value of β

Consider first the question of whether being Black has any influence at all on earnings. We formulate the null hypothesis and alternative concerning the true value of β

$$H_0 : \beta = 0 \quad \text{[no influence]}$$

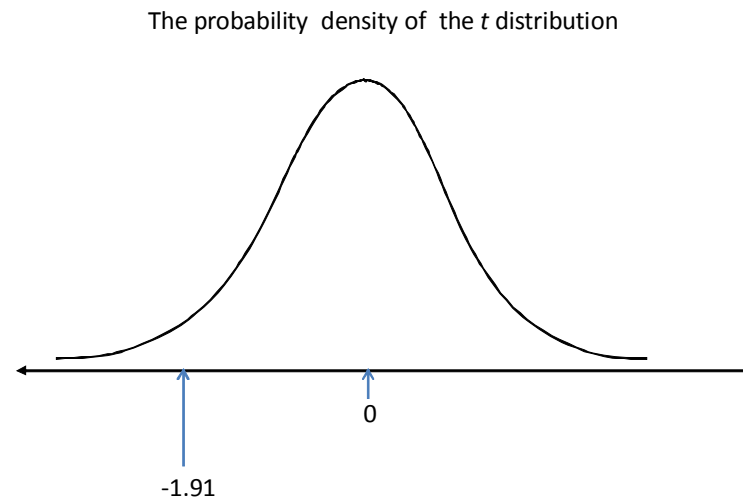
$$H_1 : \beta \neq 0 \quad \text{[influence - in either direction]}$$

We know that $\frac{\hat{\beta} - \beta}{\widehat{SE}(\beta)}$ has a t distribution which is centered at zero (under the null hypothesis).

We also "know" β under our null hypothesis: it's 0.

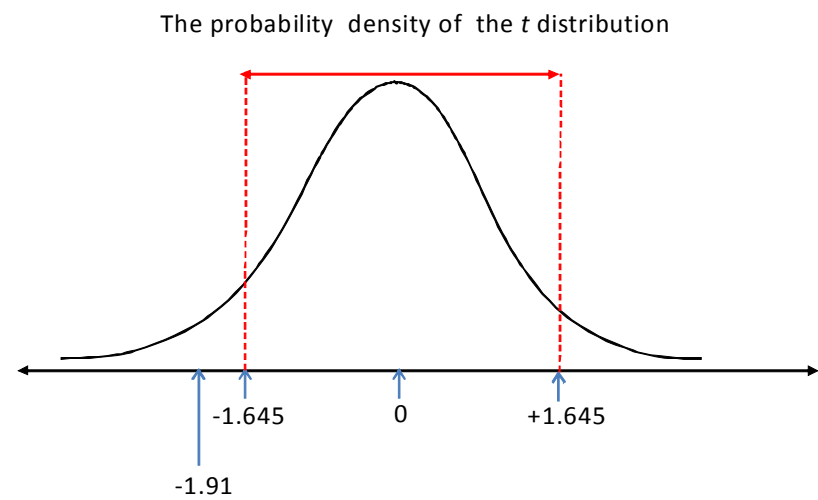
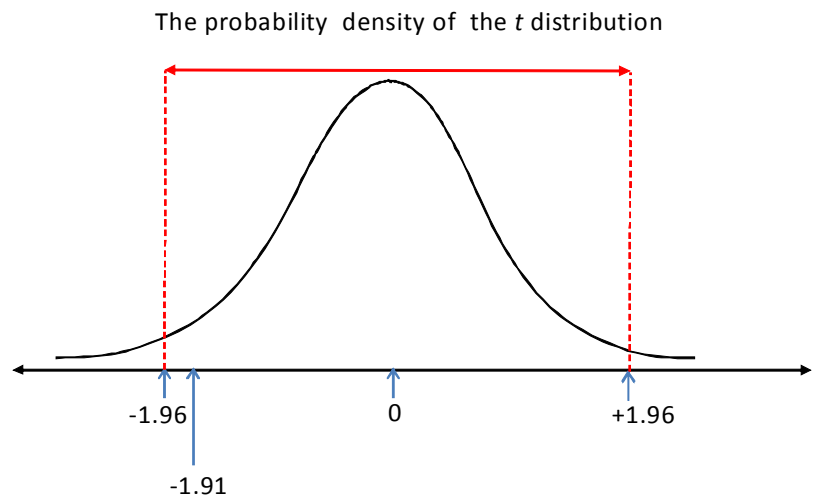
$$\text{So we can calculate } \frac{\hat{\beta} - \beta}{\widehat{SE}(\beta)} = \frac{-1172.18 - 0}{612.6281} = -1.91$$

Remember what a pdf tells you - the probability of each value of the occurring. Under the null hypothesis we can work out the most likely range of values for the t-ratio.



If the null is true then the most likely values are close to zero. If the observed value is along way from 0 then the null is unlikely to be true. If the null is true -1.91 is clearly pretty unlikely. How unlikely?

It turns out that if the true value of β is zero then 95% of the time we would expect to find values of the t-ratio between ± 1.96 or ± 1.645 90% of the time.



Step 1. Set out the null and alternative in terms of the true parameter:

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

Step 2. Calculate:

$$\frac{\hat{\beta} - \beta}{\widehat{SE}(\beta)}$$

Step 3: Decide whether you can reject the null.

In the case of the Black variable we cannot reject the null of no effect at 95% confidence.

Our value of -1.91 is within the range of ± 1.96 in which we'd expect to see 95% of the time if the null is true.

But we can reject the null of no effect at 90% confidence.

Our value of 1.91 is outside of the range of ± 1.645 where it would be 90% of the time if the null were true.

Another type of test looks to see if we can infer the sign of the true parameter.

Once again you set up a null of no effect

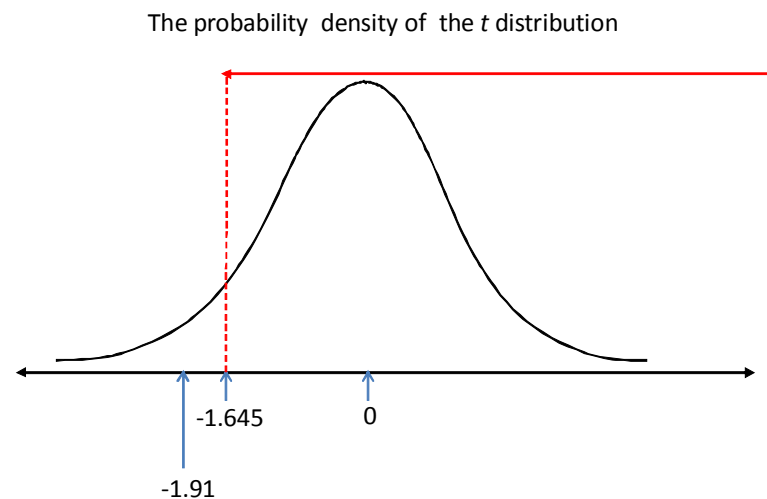
$$H_0 : \beta = 0$$

but this time specify the alternative be an inequality

$$H_1 : \beta < 0$$

In this case we are testing whether there is evidence of discrimination in the sense that observably equivalent individuals (as far as we can tell with our data) get paid less if they are black than if they are not.

Very little changes except the critical value. If the null is true the 95% of the time we would expect t to be bigger than -1.645.



As it is our t of -1.91 outside of this range. So it's very unlikely under the null hypothesis and it's in the "right" direction as far as our alternative is concerned, so we can reject the null. It's "easier" to reject the null with one tailed tests. This is because you are *a priori* ruling out the other "tail".

We've now been through the results:

Source	SS	df	MS	Number of obs = 433		
Model	1.9654e+09	5	393078432	F(5, 427)	=	13.90
Residual	1.2078e+10	427	28285101.9	Prob > F	=	0.0000
-----				R-squared	=	0.1400
Total	1.4043e+10	432	32507247	Adj R-squared	=	0.1299
-----				Root MSE	=	5318.4

Earnings	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_cons	-13116.42	3724.211	-3.52	0.000	-20436.49	-5796.352
age	1273.129	282.1772	4.51	0.000	718.4999	1827.758
age2	-19.58934	5.027359	-3.90	0.000	-29.47079	-9.707887
educ	1.79439	175.4027	0.01	0.992	-342.9658	346.5546
black	-1172.18	612.6281	-1.91	0.056	-2376.322	31.9623
married	3102.903	700.9686	4.43	0.000	1725.125	4480.682

But we're not quite done ...

1. Interpretation of the parameters - economic [✓]
2. Interpretation of the parameters - statistical [✓]
3. Interpretation of the overall robustness of the regression [✓ - as far as the output is concerned].

The final stage is to go beyond the output ... and think.

The regression in itself may be

1. informative about the economic relationship of interest
2. the coefficients may be well-determined (absolute t ratios bigger than 1.96)
3. fit the data quite well (decent R^2 and an F-value bigger than 2.23)

But there remain some important questions ...

1. What's the population of interest here? Are the results generalisable?
2. Did we leave anything important out? And what if we have?
3. Does the regression have a causal interpretation?

1. Generalisability

What was the population from which the sample was drawn?

In this case it was low income men with low labour market attachment amongst whom high-school drop-outs, ex-criminals, and ex-drug addicts were over represented.

How generalisable are the results to middle-income, highly educated women?

The point is that the data were not a random sample of the population at large, but of a particular sub-population. So inferences are only valid for the target sub-population.

Bottom line: *don't extrapolate* [I mean it].

2. Omitted variable bias (*OVB*)

You can't always measure/observe all of the potentially relevant variables.

They may be very important but if you can't put them in the regression what is the effect of leaving them out?

To be a bit more formal about it suppose that

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + e_i \quad [\text{the "true" model}]$$

But you for one reason or another you omit W and run the regression

$$Y_i = b_0 + b_1 X_i + u_i \quad [\text{the model you actually run}]$$

What is the relationship between the coefficient you estimate b_1 , and the true coefficient you'd like to estimate β_1 ?

The OVB formula is one of the most important things to know about regression:

$$E(b_1) = \beta_1 + \beta_2 \frac{\text{Cov}(X, W)}{\text{Var}(X)}$$

This says that the expected value of b_1 is equal to the true value β_1 plus another term.

The other term causes "*omitted variable bias*" if it's not zero.

$$E(b_1) = \beta_1 + \beta_2 \frac{Cov(X, W)}{Var(X)}$$

When is there no OVB? In either of the following two case:

1. When $\beta_2 = 0$ in which case the true model shouldn't have W in it anyway, so there's no harm in omitting it
2. When $Cov(X, W) = 0$ i.e. when the included and the omitted variables are uncorrelated/independent - remember independence means knowing W tells you nothing extra about X so omitting W doesn't lose you any information about X .

$$E(b_1) = \beta_1 + \beta_2 \frac{Cov(X, W)}{Var(X)}$$

Otherwise the direction and extent of the bias depends on the sign and size of β_2 and $Cov(X, W)$.

Example: omitting ability from a wage equation which includes schooling.

You would expect ability to effect wages positively and to be positively associated with levels of schooling. Therefore

$$E(b_1) > \beta_1$$

and the coefficient on schooling would be biased upwards - you'd attribute too high a return to education.

3. Does the regression have a causal interpretation?

(This is important.)

Researchers often implicitly or explicitly give regression results causal interpretations. "The stuff on the RHS determines the thing on the LHS".

Causation is a tricky business, and the potential outcomes framework is the best way we have of trying to understand it.

We will stick with the connection between schooling and earnings as a running example.

The causal relationship between schooling and earnings tells us what people would earn, on average, if we could change their level of schooling randomly so that those being compared with different levels of schooling would be otherwise comparable.

Unlike the treatment/no treatment framework we looked at in the lecture on program evaluation schooling takes on an ordered list of values

$$S_i \in \{0, 1, \dots, 21\}$$

But the potential outcomes framework can easily stretch to accommodate this:

$$\text{Potential Outcome} = \begin{cases} Y_i(0) & \text{if } S_i = 0 \\ Y_i(1) & \text{if } S_i = 1 \\ \vdots & \\ Y_i(25) & \text{if } S_i = 21 \end{cases}$$

Or more simply,

$$Y_i(s)$$

is someone's potential earnings after s years of education.

The *observed* outcome, Y_i , can be linked to *potential* outcomes as follows

$$Y_i = \sum_{s=0}^{s=21} Y_i(s) [S_i = s]$$

$[S_i = s]$ is an indicator (1 if true, 0 otherwise).

Suppose that we run the regression of earnings on the schooling variable:

$$Y_i = \beta_1 + \beta_2 S_i + u_i$$

Regressions give us linear conditional expectations so we can look at the conditional expectations of the outcome for each level of schooling

$$\begin{aligned} E(Y_i|S_i = 0) &= \beta_1 \\ E(Y_i|S_i = 1) &= \beta_1 + \beta_2 \\ E(Y_i|S_i = 2) &= \beta_1 + 2\beta_2 \\ &\vdots \\ E(Y_i|S_i = 21) &= \beta_1 + 21\beta_2 \end{aligned}$$

Let s denote s years of schooling and t denote an extra year on top of s (i.e. $t = s + 1$)

We can interpret the coefficient on the schooling variable as

$$\beta_2 = E(Y_i|S_i = t) - E(Y_i|S_i = s)$$

Since we're only conditioning on one thing (S_i) I'm now going to drop the $S_i = \text{bit}$ and write it more compactly as

$$\beta_2 = E(Y_i|t) - E(Y_i|s)$$

So β_2 measures the comparison in earnings between those with t years of schooling and s years of schooling

$$\beta_2 = E[Y_i(t) | t] - E[Y_i(s) | s]$$

[Remember that for those with t years of schooling the observed outcome Y_i is also the potential outcome for t years of schooling $Y_i(t)$]

Recall from the programme evaluation lecture that this comparison can be written out as

$$E [Y_i (t) | t] - E [Y_i (s) | s] =$$
$$\underbrace{E [Y_i (t) | t] - E [Y_i (s) | t]}_{\text{ATT}} + \underbrace{E [Y_i (s) | t] - E [Y_i (s) | s]}_{\text{Selection effect}}$$

The ATT captures the effects of an additional year of schooling for those who stayed on to year t

The selection bias captures the difference in the earnings for those who dropped out at year s and the earnings at year s of those who decided to stay on to year t .

It seems likely that those who stay on would have earned more anyway so the selection bias term is positive and the coefficient would therefore be bigger than the ATT overstating the returns to school.

In order to make sure that there is no selection bias one of the key ideas we looked at was the CIA - Conditional Independence Assumption.

$$Y_i(s) \perp S_i | X_i$$

This assumes that *potential* outcomes are independent of schooling as long as you first account for the effects of a bunch of other covariates which might have a bearing (the X_i 's: these would be things like family background, income, the opportunity cost of school etc).

In a regression scenario trying to establish the CIA means lobbying in as many exogenous covariates as you can which might account for the selection of higher ability people into staying on.

Incredible Econometrics Again

CIA is an assumption. That assumption has to be credible.

You need to be sure that you have accounted for everything which systematically influences staying on at school so that what is left is as *good as random*.

If and only if that is credible can you credibly interpret the coefficient on schooling as a causal effect.

2. Maximum Likelihood - FAQ!

You've been doing a fair bit of MLE.

Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Let's give names to these events.

A = the event that the parameters of the regression line take certain values.

B = the event that the sample data have the distribution which we observe.

Then we can write

$$P(A|B) = P(\text{parameters taking certain values} \mid \text{data has the observed distribution})$$

$$P(B|A) = P(\text{data has the observed distribution} \mid \text{parameters taking certain values})$$

Or in shorthand:

$$P(A|B) = P(\text{parameters} \mid \text{data})$$

$$P(B|A) = P(\text{data} \mid \text{parameters})$$

So Bayes' Theorem says:

$$P(\text{parameters} | \text{data}) = \frac{P(\text{data} | \text{parameters}) P(\text{parameters})}{P(\text{data})}$$

or

$$P(\text{parameters} | \text{data}) \propto P(\text{data} | \text{parameters}) P(\text{parameters})$$

where \propto means "is proportional to"

$$\underbrace{P(\text{parameters} | \text{data})}_{\text{Posterior}} \propto \underbrace{P(\text{data} | \text{parameters})}_{\text{Likelihood}} \underbrace{P(\text{parameters})}_{\text{Prior}}$$

$$\underbrace{P(\text{parameters} | \text{data})}_{\text{Posterior}} \propto \underbrace{P(\text{data} | \text{parameters})}_{\text{Likelihood}} \underbrace{P(\text{parameters})}_{\text{Prior}}$$

We could ask

“given the data we see, what are the most probable values for the parameters?”

[this focuses on the posterior]

$$\underbrace{P(\text{parameters} | \text{data})}_{\text{Posterior}} \propto \underbrace{P(\text{data} | \text{parameters})}_{\text{Likelihood}} \underbrace{P(\text{parameters})}_{\text{Prior}}$$

Alternatively we can ask

"which values for the parameters would give us the highest probability of observing our data?"

[this focuses on the likelihood]

These two questions and their answers are tied together through Bayes Rule.

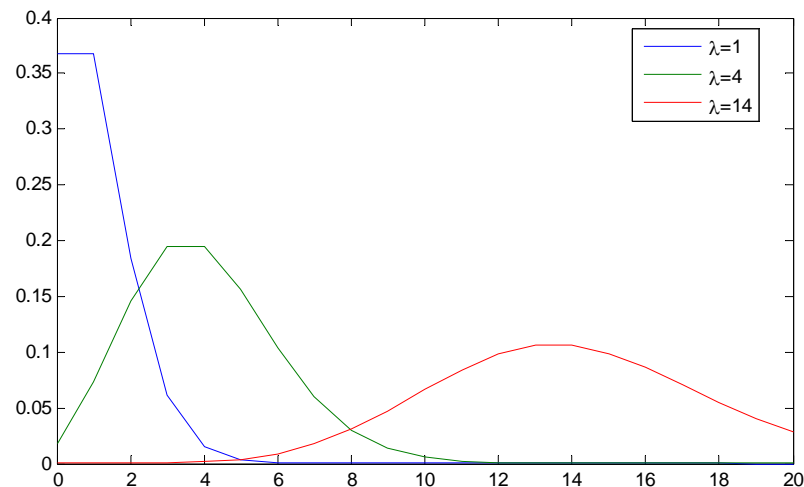
Really, the ONLY difference between $P(\text{parameters} \mid \text{data})$ and $P(\text{data} \mid \text{parameters})$ lies in what you regard as fixed and what you regard as variable.

In the case of the likelihood $P(\text{data} \mid \text{parameters})$ you regard the observations as fixed and the parameters as unknown variables to be worked out.

Example: (Based on Thomas, 11). The arrival rate of buses has a Poisson distribution:

$$P(X|\lambda) = \frac{\lambda^X e^{-\lambda}}{X!}$$

where the variable X is the number of buses arriving in some fixed interval of time (say 10 mins) and λ is an unknown parameter (which governs the location and spread of the distribution).



Suppose we observe 4 buses arrive in the recording period. That's one observation

$$X_1 = 4$$

We write down the probability with $X = 4$ plugged in.

$$P(4|\lambda) = \frac{\lambda^4 e^{-\lambda}}{4!}$$

We then ask the question: “what value of the parameter λ maximise the likelihood of this event?” and we do a mental flip and think about this probability as being a function of the parameter. Presto! - it's now the likelihood.

$$L(\lambda) = \frac{\lambda^4 e^{-\lambda}}{4!}$$

(This is all you've been doing in the MLE work - plugging data into a density formula and then regarding it as a function of unknown parameters to be solved for).

To find the maximising value you log it first (if you like)

$$\log L(\lambda) = 4 \log \lambda - \lambda - \log(4!)$$

Then differentiate it and set it to zero (f.o.c.)

$$\frac{d \log L(\lambda)}{d\lambda} = \frac{4}{\lambda} - 1 = 0$$

and pop a hat on the solution:

$$\hat{\lambda} = 4$$

Suppose that you went out and made two more observations on arrival rates and you noted down

$$X_1 = 4, \quad X_2 = 7, \quad X_3 = 5$$

The individual probabilities of these events are

$$P(4|\lambda) = \frac{\lambda^4 e^{-\lambda}}{4!}, \quad P(7|\lambda) = \frac{\lambda^7 e^{-\lambda}}{7!}, \quad P(5|\lambda) = \frac{\lambda^5 e^{-\lambda}}{5!}$$

Given these three observations you want to ask the question “what value of λ maximises the likelihood of this *joint* event $P(4 \text{ and } 7 \text{ and } 5 | \lambda)$?”

Independence etc.

You know the rule:

"the joint probability of independent events is the product of the probabilities"

In probability notation this is

$$P(A \text{ and } B) = P(A)P(B)$$

But this doesn't say much about the intuitive nature of independence.

A more intuitive definition of independence is this:

$$P(A|B) = P(A)$$

This says the conditional probability of A is the same as the unconditional.

What this **means** is that knowing B tells you nothing about how likely A is.

It's equivalent to the other definition: what's the connection?

The connection between the two definitions is (you guessed it) Bayes rule again:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

If knowing B doesn't help you with A then $\Pr(A|B) = \Pr(A)$ so

$$P(A) = \frac{P(A \text{ and } B)}{P(B)}$$

And rearranging gives us:

$$P(A \text{ and } B) = P(A)P(B)$$

So independent observations means that

$$P(4 \text{ and } 7 \text{ and } 5 \mid \lambda) = \frac{\lambda^4 e^{-\lambda}}{4!} \times \frac{\lambda^7 e^{-\lambda}}{7!} \times \frac{\lambda^5 e^{-\lambda}}{5!} = \frac{\lambda^{16} e^{-3\lambda}}{4!7!5!}$$

Do the mental flip:

$$L(\lambda) = \frac{\lambda^{16} e^{-3\lambda}}{4!7!5!}$$

log it (if you like)

$$\log L(\lambda) = 16 \log \lambda - 3\lambda - \log(4!7!5!)$$

Differentiate and set to zero

$$\frac{d \log L(\lambda)}{d\lambda} = \frac{16}{\lambda} - 3 = 0$$

Solve to the maximising value of $\hat{\lambda} = 16/3$.

Now let's do the most important MLE problem you have come across - the use of MLE to estimate the regression line.

We have n pairs of observations on $\{Y_i, X_i\}$ which are iid (independent and identically distributed).

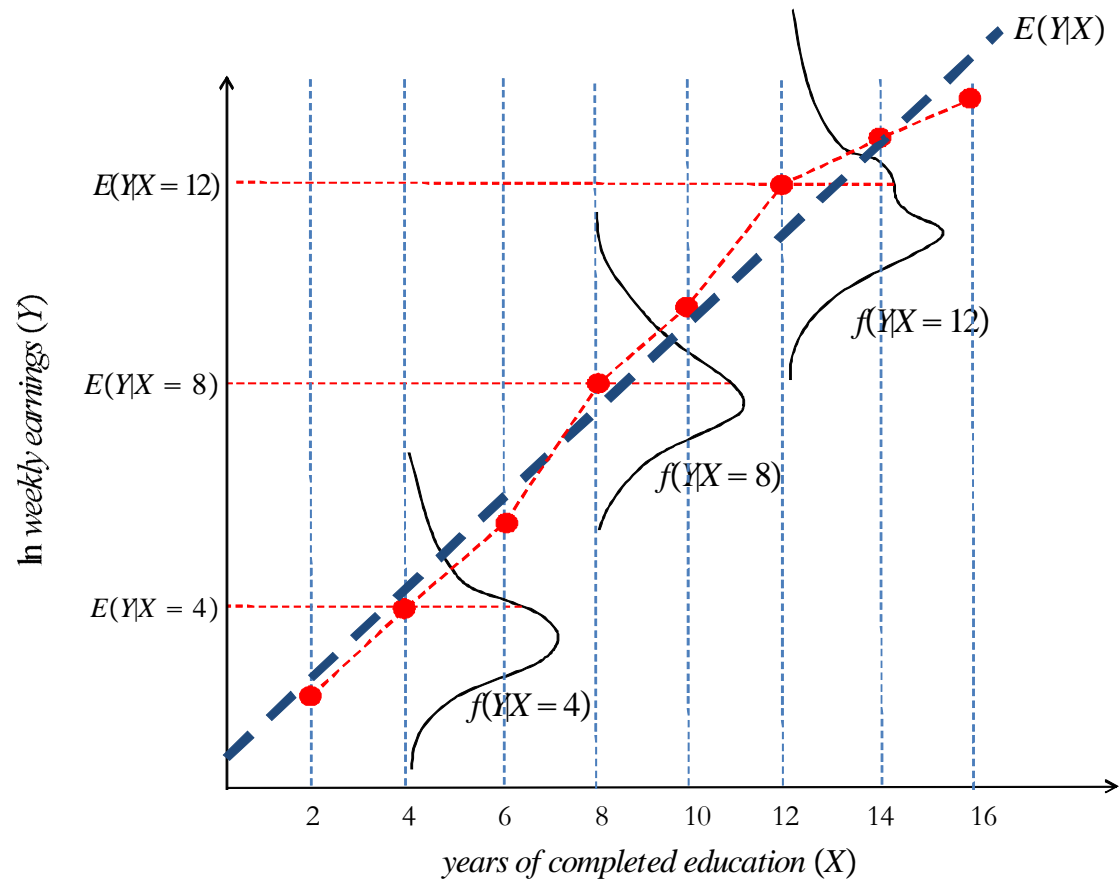
We will assume that Y_i has a normal distribution.

Normal distributions have two parameters which govern their shape - their mean and variance $N(\mu, \sigma^2)$

We will suppose that the mean of the distribution of Y_i depends on X_i and that this relationship is linear

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

This is the conditional expectation function (assuming it's linear) or an approximation to it even if it's not linear.



So Y_i is normal with mean given by $\beta_0 + \beta_1 X_i$ and a variance of σ^2

$$Y_i \stackrel{D}{=} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Consider the first observation $\{Y_1, X_1\}$.

If we write down the probability with this value plugged in we get

$$P(Y_1, X_1 | \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (Y_1 - \beta_0 - \beta_1 X_1)^2\right)$$

The probability for the second observation $\{Y_2, X_2\}$ is

$$P(Y_2, X_2 | \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (Y_2 - \beta_0 - \beta_1 X_2)^2\right)$$

and so on.

To get the joint probability of all of the data we multiply these together (because they are independent)

$$\begin{aligned} P(\text{data}|\beta_0, \beta_1, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (Y_1 - \beta_0 - \beta_1 X_1)^2\right) \\ &\quad \times \\ &\quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (Y_2 - \beta_0 - \beta_1 X_2)^2\right) \\ &\quad \times \\ &\quad \vdots \\ &\quad \times \\ &\quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (Y_n - \beta_0 - \beta_1 X_n)^2\right) \end{aligned}$$

Using “sigma” notation this is:

$$P(\text{data}|\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{i=n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (Y_i - \beta_0 - \beta_1 X_i)^2\right)$$

We then do the mental flip and regard this as a function of the parameters

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{i=n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (Y_i - \beta_0 - \beta_1 X_i)^2\right)$$

and we’ve got the likelihood function.

It’s really that simple.

Deriving the estimators is then just algebra (and highly learnable).

It's worth remembering the estimators which come out of this:

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})} = \left[\frac{Cov(X, Y)}{Var(X)} \right] = \frac{\sum_{i=1}^n Y_i (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2\end{aligned}$$