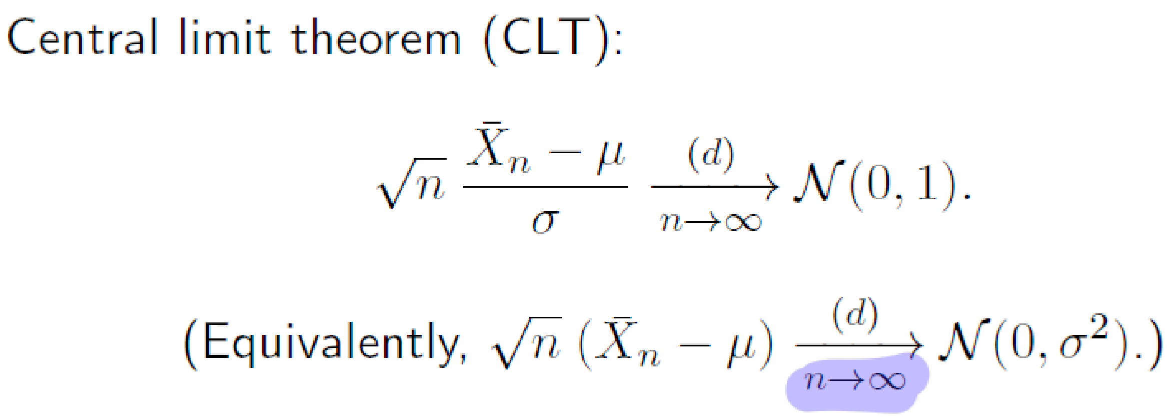
Law of large numbers (LLN)

The average converges towards the expectation.

CLT

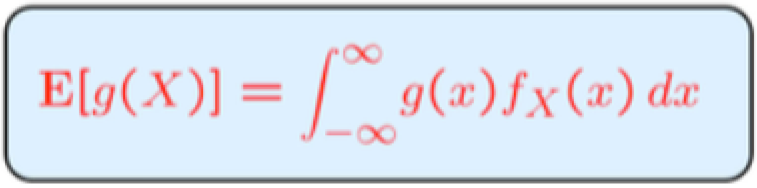


Central limit theorem (CLT):

$$\sqrt{n}\,\frac{\bar{X}_n-\mu}{\sigma}\xrightarrow[n\to\infty]{(d)}\mathcal{N}(0,1).$$

(Equivalently, $\sqrt{n}\,(\bar{X}_n-\mu)\xrightarrow[n\to\infty]{(d)}\mathcal{N}(0,\sigma^2)$.)

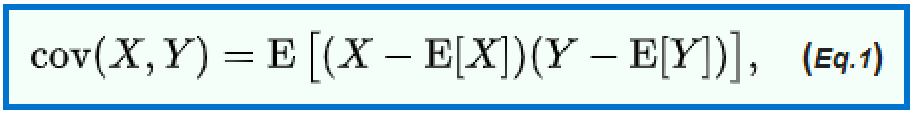
Expectations



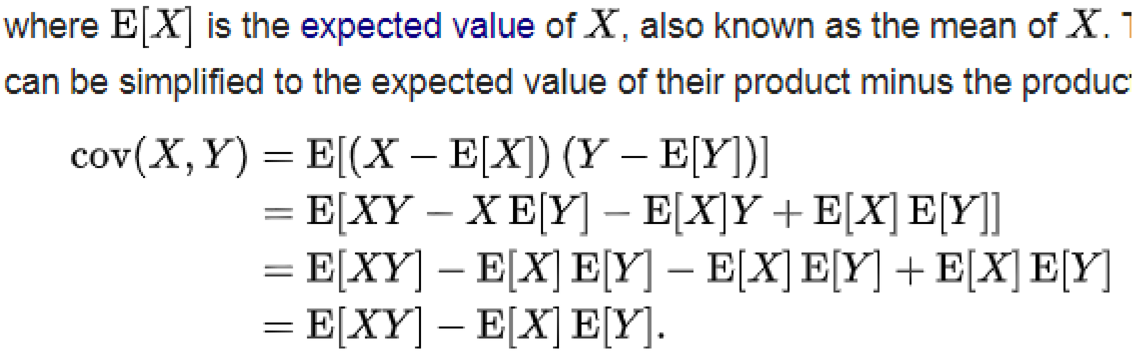
$$\boxed{\mathrm{E}[g(X)]=\int_{-\infty}^{\infty}g(x)f_X(x)\,dx}$$

Variance

Covariance



$$\boxed{\operatorname{cov}(X,Y)=\mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])\right],\quad \textit{(Eq.1)}}$$



where $\mathrm{E}[X]$ is the expected value of X, also known as the mean of X. T
can be simplified to the expected value of their product minus the produc

$$\begin{aligned}\operatorname{cov}(X,Y)&=\mathrm{E}[(X-\mathrm{E}[X])\,(Y-\mathrm{E}[Y])]\\&=\mathrm{E}[XY-X\,\mathrm{E}[Y]-\mathrm{E}[X]Y+\mathrm{E}[X]\,\mathrm{E}[Y]]\\&=\mathrm{E}[XY]-\mathrm{E}[X]\,\mathrm{E}[Y]-\mathrm{E}[X]\,\mathrm{E}[Y]+\mathrm{E}[X]\,\mathrm{E}[Y]\\&=\mathrm{E}[XY]-\mathrm{E}[X]\,\mathrm{E}[Y].\end{aligned}$$

Joint distributions

f(x,y) = f(y|x).f(x)

f(x,y) = f(x|y).f(y)

Conditional distribution

f(y|x) = f(x,y) / f(x)

f(x|y) = f(x,y) / f(y)

Marginal distributions

Marginal distribution f(x) = integrale over y de f(x,y)dy

Marginal distribution f(x) = Ey [f(x|y)]

Bayes’ theorem

P(B|A) = P(A|B)* P(B)/ P(A)

Distributions

Student's t-distribution : continuous probability distribution that arises when estimating the mean of a normally distributed population where sample size n is small and variance is unknown.

T statistic = (beta_hat – true_beta)/ std dev of beta_hat

F-distribution, Fisher–Snedecor distribution, is a continuous probability distribution that arises frequently as the null distribution of a test statistic, most notably in the analysis of variance (ANOVA), e.g., F-test.

Fisher information

$$\mathcal{I}\left(\theta\right) = \mathrm{Cov}\left(\nabla \ell\left(\theta\right)\right) = -\mathbb{E}\left[\mathbf{H}\ell\left(\theta\right)\right].$$

$$\mathcal{I}\left(\theta\right)=\int_{-\infty}^{\infty}\frac{\left(\frac{\partial}{\partial\theta}f_{\theta}\left(x\right)\right)^{2}}{f_{\theta}\left(x\right)}\,dx.$$

ESTIMATION

Hypothesis testing

Tn is the test statistic

p-value is the mass for which X > Tn

t-statistic (1 parameter)

Used a lot in regression

Looks like normal but with fatter tails, so gives more conservative p-values

Can be used for small values of n

Needed for a one sided test

For testing one particular Beta equal a given value, equivalent to F test (the t-test statistics are the square root of the F-test statistics.) but easier to use

*Use the t-statistic when discussing individual coefficients/ parameters

F-statistic

Cannot be used for one sided test

Critical value calculator

https://www.socscistatistics.com/tests/criticalvalues/default.aspx

Maximum likelihood estimator

MLE is the theta for which l'(theta)=0

Test first if l'' is negative (concave function, which means we are the MAX)

Maximum likelihood estimator asymptotic normality

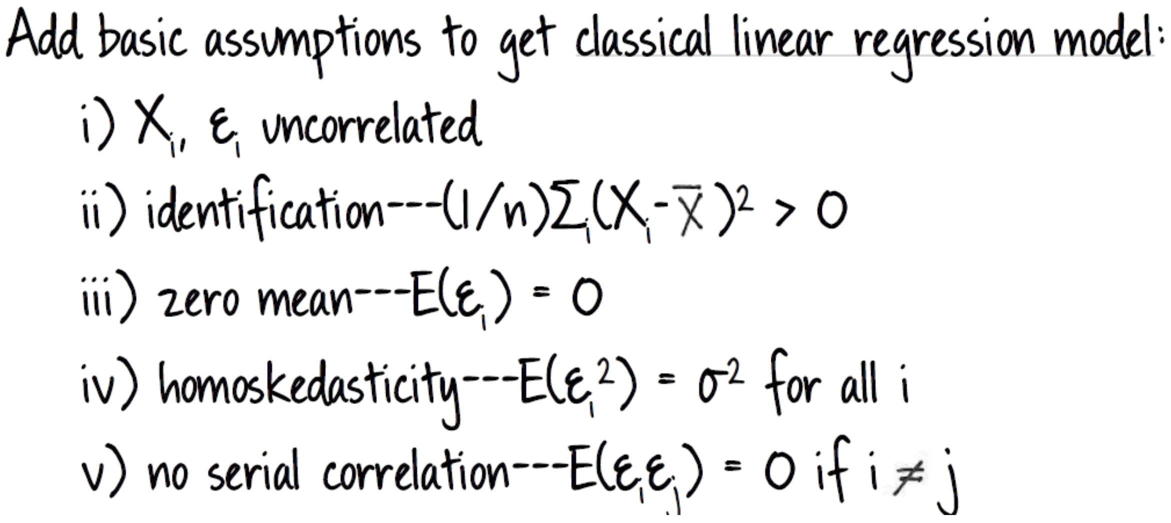
4 Conditions

- Model identified / parameter identifiable: if I give you 2 parameters theta, they give me 2 different distributions and inversely

Linear regression / OLS (ordinary least squares)

Conditions

- Non linearly dependant variables
- - Errors: i.i.d, E(errors) = 0



Add basic assumptions to get classical linear regression model:

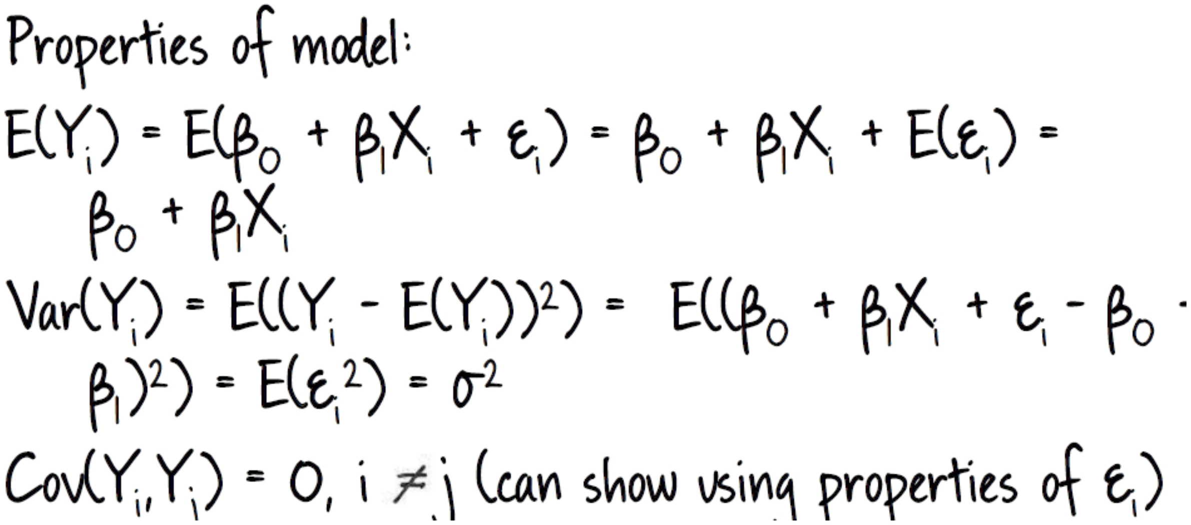
i) X_i, ε_i uncorrelated

ii) identification---$(1/n)\sum_i(X_i - \overline{X})^2 > 0$

iii) zero mean---$E(\varepsilon_i) = 0$

iv) homoskedasticity---$E(\varepsilon_i{}^2) = \sigma^2$ for all i

v) no serial correlation---$E(\varepsilon_i\varepsilon_j) = 0$ if $i \neq j$



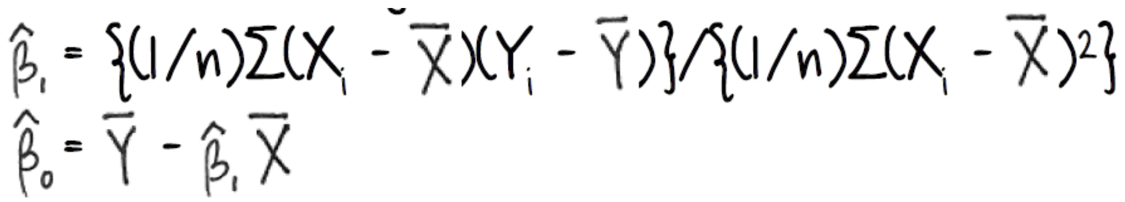
Properties of model:

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

$Var(Y_i) = E((Y_i - E(Y_i))^2) = E((\beta_0 + \beta_1 X_i + \varepsilon_i - \beta_0 -$
$\beta_1)^2) = E(\varepsilon_i{}^2) = \sigma^2$

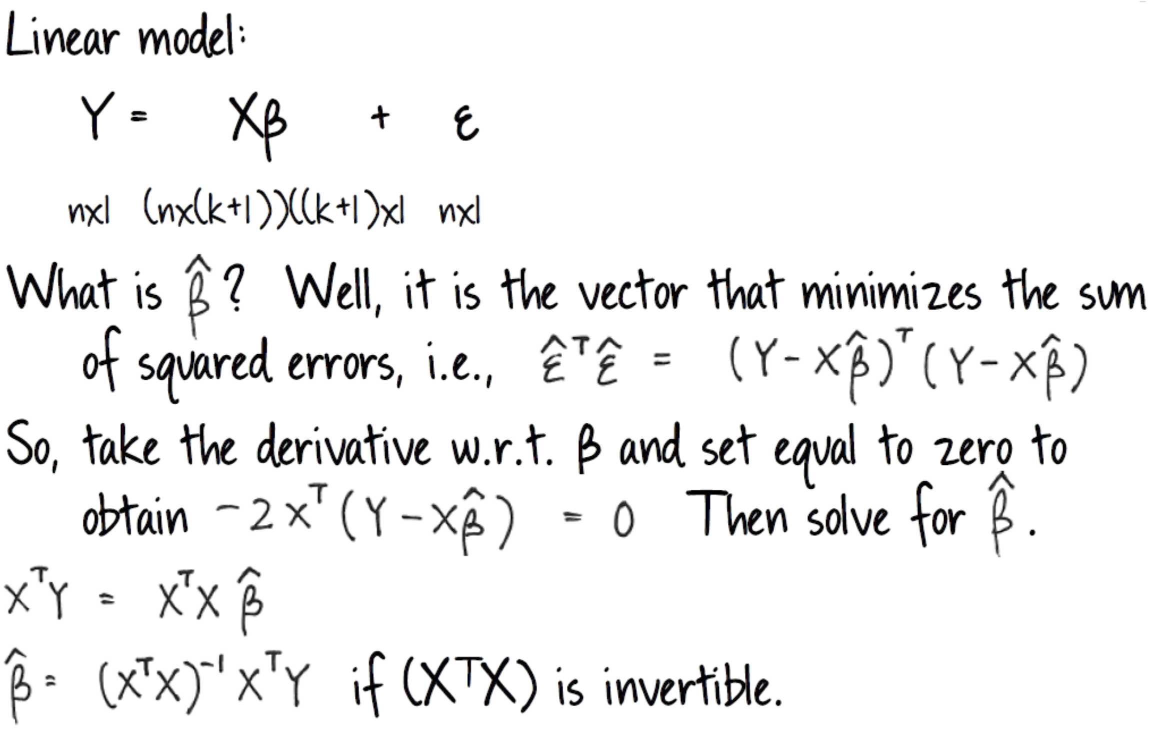
$Cov(Y_i, Y_j) = 0$, $i \neq j$ (can show using properties of ε_i)

Closed forms regression parameters



$$\hat{\beta}_1 = \{(1/n)\Sigma(X_i - \bar{X})(Y_i - \bar{Y})\}/\{(1/n)\Sigma(X_i - \bar{X})^2\}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$



Linear model:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1}{X\beta} + \underset{n\times 1}{\varepsilon}$$

What is $\hat{\beta}$? Well, it is the vector that minimizes the sum of squared errors, i.e., $\hat{\varepsilon}^T\hat{\varepsilon} = (Y - X\hat{\beta})^T(Y - X\hat{\beta})$

So, take the derivative w.r.t. β and set equal to zero to obtain $-2X^T(Y - X\hat{\beta}) = 0$ Then solve for $\hat{\beta}$.

$$X^TY = X^TX\hat{\beta}$$

$\hat{\beta} = (X^TX)^{-1}X^TY$ if (X^TX) is invertible.

Linear model:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1)}{X\beta} + \underset{n\times 1}{\varepsilon}$$

What do we want to know about $\hat{\beta}$? Its distribution!

$E(\hat{\beta}) = \beta$ (Treat Xs as fixed, they come outside of the expectation operator, and it's easy to show.)

$Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$ (Again, not too hard to show if you treat the Xs as fixed---details in notes on website.)

And $\hat{\sigma}^2 = \hat{\varepsilon}^T\hat{\varepsilon}/(n-k)$

And if the errors are normally-distributed, they're also normal.

	mean	variance	covariance
$\hat{\beta}_0$	β_0	$\sigma^2\bar{X}^2/n\hat{\sigma}_x^2 + \sigma^2/n$	$-\sigma^2\bar{X}/n\hat{\sigma}_x^2$
$\hat{\beta}_1$	β_1	$\sigma^2/n\hat{\sigma}_x^2$	

Some comparative statics:

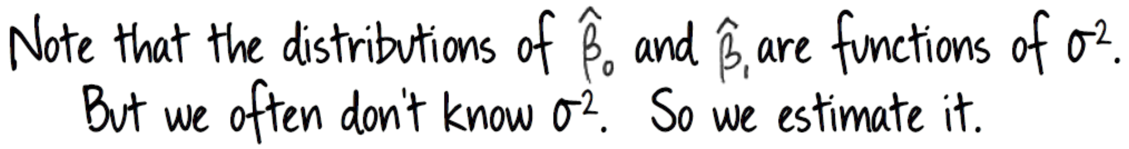
---A larger σ^2 means larger $Var(\hat{\beta})$

---A larger $\hat{\sigma}_x^2$ means smaller $Var(\hat{\beta})$

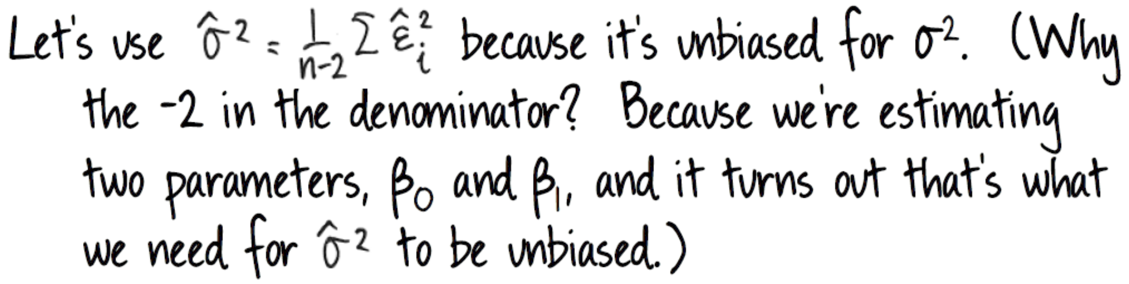
---A larger n means smaller $Var(\hat{\beta})$

---If $\bar{X} > 0$, $Cov(\beta_0, \beta_1) < 0$

Linear model - Student t test for hypothesis testing about a regression coefficient



Note that the distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are functions of σ^2. But we often don't know σ^2. So we estimate it.



Let's use $\hat{\sigma}^2 = \frac{1}{n-2}\sum\hat{\varepsilon}_i^2$ because it's unbiased for σ^2. (Why the -2 in the denominator? Because we're estimating two parameters, β_0 and β_1, and it turns out that's what we need for $\hat{\sigma}^2$ to be unbiased.)

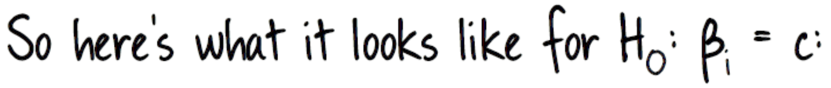
When in the numerator we have a normal (regression coeff is normal if errors are normal, but we use it also if no assumption of normality)

and in the denominator variance replaced it by an estimate:

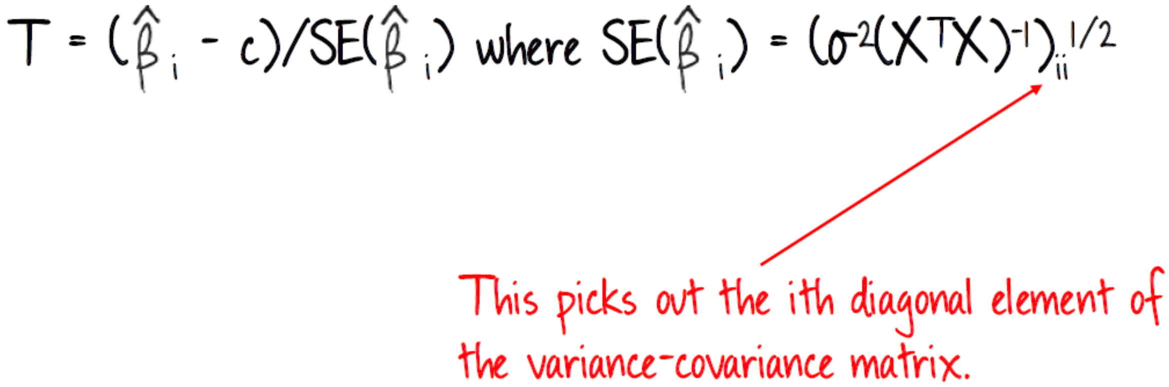
Hypothesis testing - Student's t-test :

Works when sample size n is small

T statistic = (beta_hat – true_beta under the null)/ std dev of beta_hat



So here's what it looks like for $H_0: \beta_i = c$:



$T = (\hat{\beta}_i - c)/SE(\hat{\beta}_i)$ where $SE(\hat{\beta}_i) = (\sigma^2(X^TX)^{-1})_{ii}^{1/2}$

This picks out the ith diagonal element of the variance-covariance matrix.

So here's what it looks like for H_0: $R\beta = c$:

$$T = (R\hat{\beta} - c)/SE(R\hat{\beta})$$

where $SE(R\hat{\beta}) = (\sigma^2 R(X^TX)^{-1}R^T)^{1/2}$

Since this is a t-test, and we can only test one hypothesis (potentially involving multiple parameters), R is a $1 \times (k+1)$ matrix and c is a scalar here.

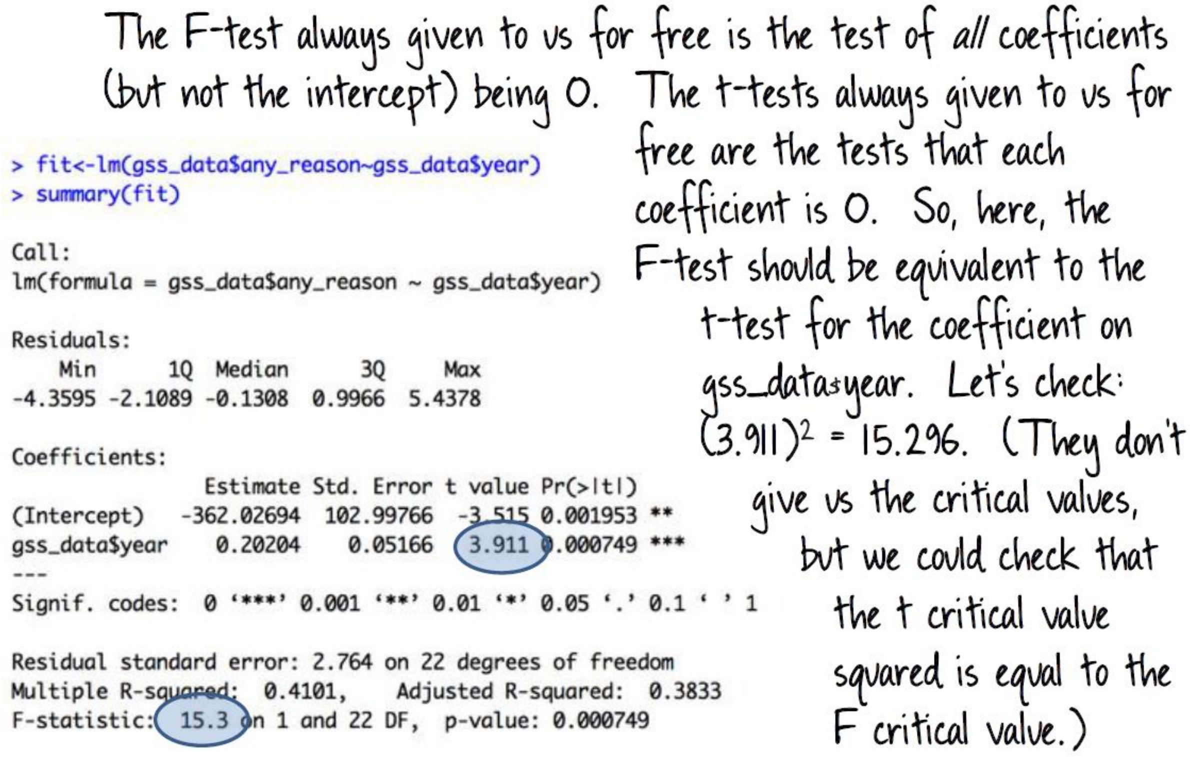
T-test versus F-test

Back to the question of when and how it's useful:

Well, for the hypothesis H_0: $\beta_j = c$ versus H_A: $\beta_j \neq c$, the F-test is equivalent to the t-test. (The t-test statistic and critical values are the square root of those for the F-test.)

So, you can use either, but it's easier to use the t-test for a single estimated coefficient if H_0: $\beta_j = 0$ since it's printed out right there for you.

One case where you *need* a t-test: if you want to carry out a one-sided test, like H_0: $\beta_j > 0$ versus H_A: $\beta_j < 0$.



The F-test always given to us for free is the test of *all* coefficients (but not the intercept) being 0. The t-tests always given to us for free are the tests that each coefficient is 0. So, here, the F-test should be equivalent to the t-test for the coefficient on gss_data$year. Let's check: $(3.911)^2 = 15.296$. (They don't give us the critical values, but we could check that the t critical value squared is equal to the F critical value.)

```
> fit<-lm(gss_data$any_reason~gss_data$year)
> summary(fit)

Call:
lm(formula = gss_data$any_reason ~ gss_data$year)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3595 -2.1089 -0.1308  0.9966  5.4378 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -362.02694  102.99766  -3.515 0.001953 ** 
gss_data$year    0.20204    0.05166   3.911 0.000749 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.764 on 22 degrees of freedom
Multiple R-squared:  0.4101,	Adjusted R-squared:  0.3833 
F-statistic:  15.3 on 1 and 22 DF,  p-value: 0.000749
```

In R

Lm linear regression model gives the t statistic next the coefficient with the p value (Proba > |t|), testing for coefficient is null

var(x) gives variance of vector x

covar(x, y)

dt()

pt (q,df=degrees of freedom, lower.tail=FALSE)

qt (mass, df=degrees of freedom, lower.tail=FALSE)

rt()

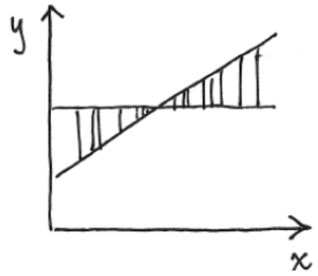
Linear regression Goodness of fit: R^2 (lecture 17)

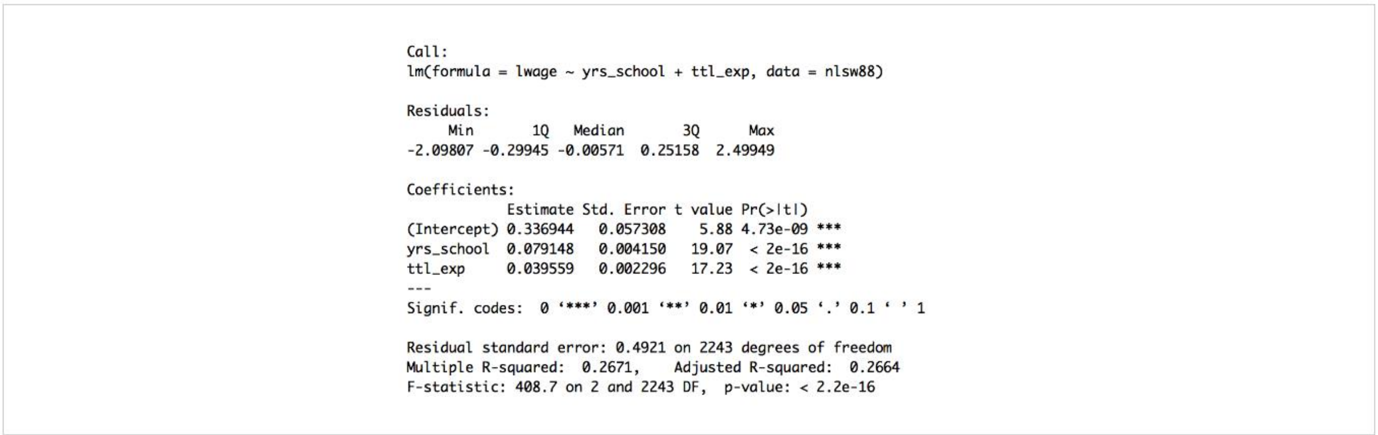
I guess we wanted a measure of fit that had larger values when the fit was better, or we explained more, so we defined

$R^2 = 1 - SSR/SST$.

It turns out that SST can be decomposed into two terms, SSR and the model sum of squares, SSM.

$$SSM = \sum_i(\hat{Y}_i - \bar{Y})^2$$

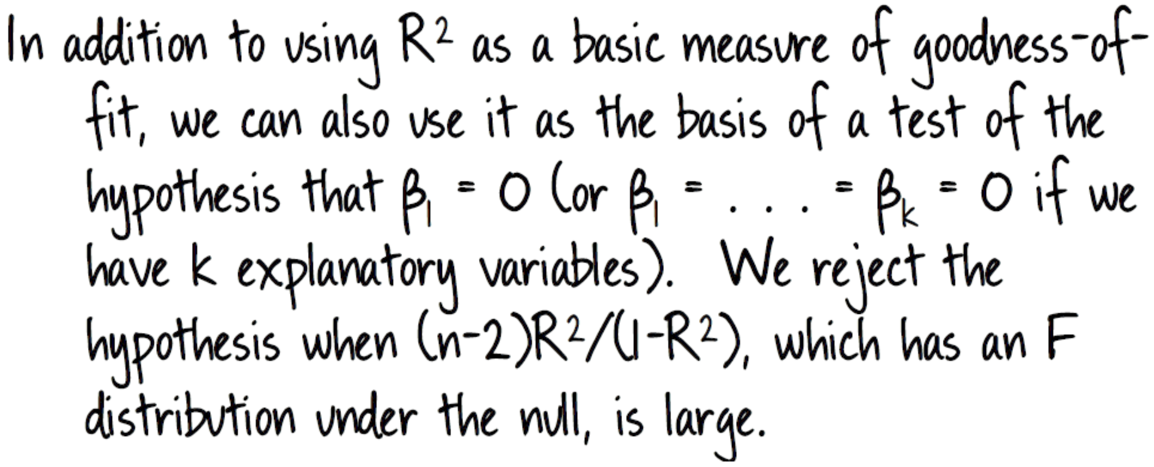




```
Call:
lm(formula = lwage ~ yrs_school + ttl_exp, data = nlsw88)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.09807 -0.29945 -0.00571  0.25158  2.49949 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.336944   0.057308    5.88 4.73e-09 ***
yrs_school  0.079148   0.004150   19.07  < 2e-16 ***
ttl_exp     0.039559   0.002296   17.23  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4921 on 2243 degrees of freedom
Multiple R-squared:  0.2671,	Adjusted R-squared:  0.2664 
F-statistic: 408.7 on 2 and 2243 DF,  p-value: < 2.2e-16
```

From there, we know that the R^2 is 0.2671. This implies that 26.71% of the total variance in the logarithm of the wage is explained by the years of schooling and the total experience.

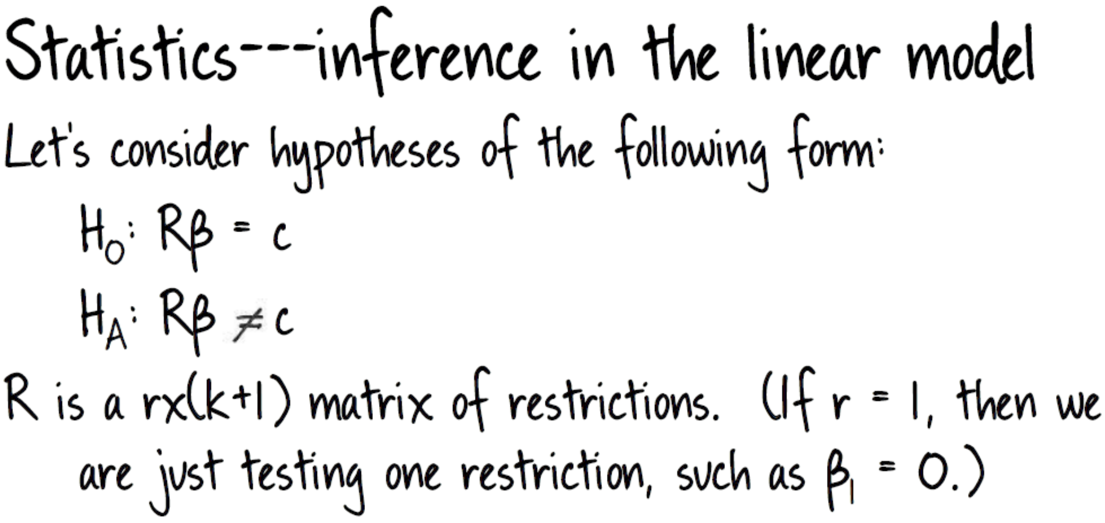
Test of hypothesis (regression coeff = 0) using R^2



In addition to using R^2 as a basic measure of goodness-of-fit, we can also use it as the basis of a test of the hypothesis that $\beta_1 = 0$ (or $\beta_1 = \ldots = \beta_k = 0$ if we have k explanatory variables). We reject the hypothesis when $(n-2)R^2/(1-R^2)$, which has an F distribution under the null, is large.

Hypothesis testing (inference) in the linear model (lecture 18)

- Cannot do one sided tests with matrix of restriction



Statistics---inference in the linear model

Let's consider hypotheses of the following form:

H_0: $R\beta = c$

H_A: $R\beta \neq c$

R is a rx(k+1) matrix of restrictions. (If r = 1, then we are just testing one restriction, such as $\beta_1 = 0$.)

Statistics---inference in the linear model

We have a super intuitive and cool way to test these hypotheses. (First, think of the null as describing a set of restrictions on the model.)

1. We estimate the unrestricted model.

2. We impose the restrictions of the null and estimate that model.

3. We compare the goodness-of-fit of the models. If the restrictions don't really affect the fit of the model much, then the null is probably true or close to true, so we do not want to reject it. If the restrictions really bind, then we do want to reject the null.

Estimating the unrestricted model should be simple---just run the regression. But how do we estimate the restricted model?

If the restriction is that certain βs = 0, then leave the regressors corresponding to those βs out of the restricted model.

If the restriction is that, say, two βs are equal, create a new regressor, which is the sum of the regressors corresponding to those βs and include that sum in the restricted model in place of the original regressors.

What if the restriction is that some $\beta = c$?

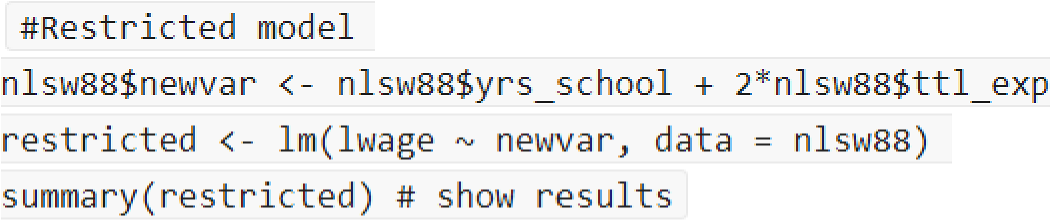
This is an F-test. (We've mentioned a special case of the F test before. This is a more general formulation.)

$T = ((SSR_R - SSR_U)/r)/(SSR_U/(n-(k+1))$

$T \sim F_{r,n-(k+1)}$ under the null and we reject the null for large values of the test statistic.

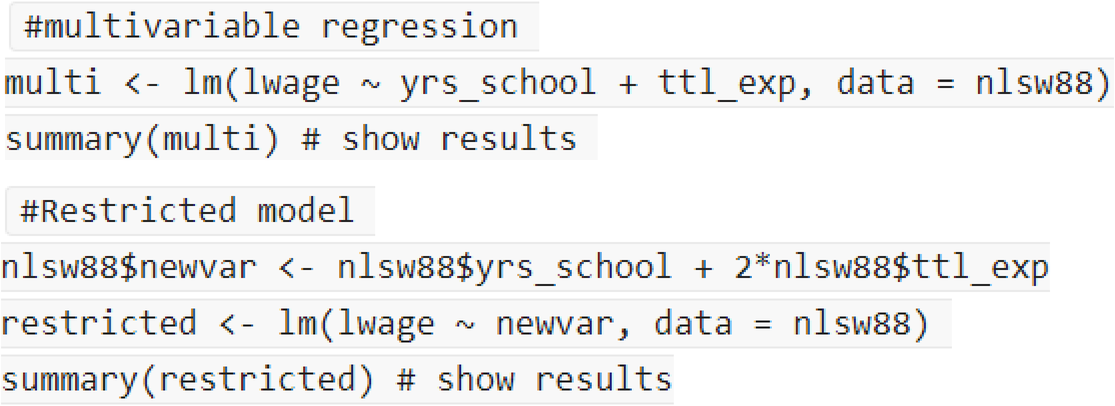
(Why an F distribution? Well, the reason goes back to one of the facts I told you about special distributions a couple of weeks ago. The ratio of two independent χ^2 random variables divided by their respective degrees of freedom are distributed F.)

Restricted model in R – New variable



```
#Restricted model
nlsw88$newvar <- nlsw88$yrs_school + 2*nlsw88$ttl_exp
restricted <- lm(lwage ~ newvar, data = nlsw88)
summary(restricted) # show results
```

Restricted model in R – Anova F test

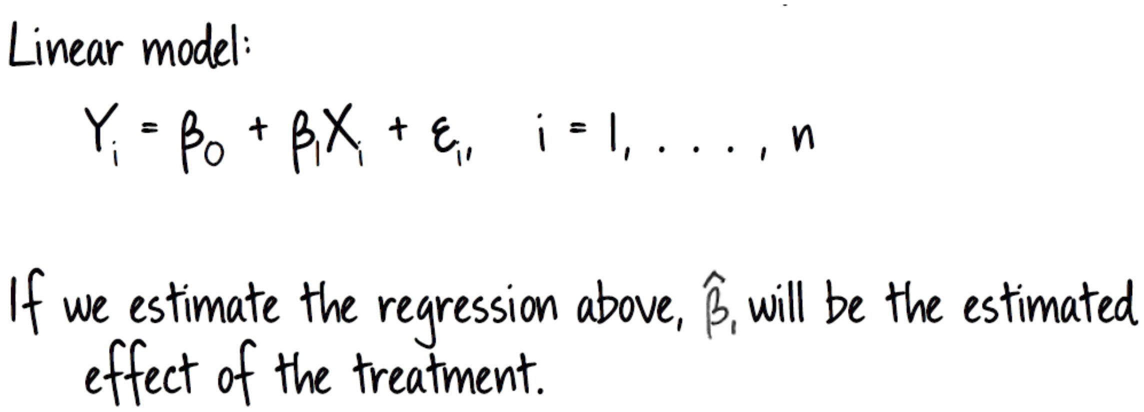


```
#multivariable regression
multi <- lm(lwage ~ yrs_school + ttl_exp, data = nlsw88)
summary(multi) # show results
#Restricted model
nlsw88$newvar <- nlsw88$yrs_school + 2*nlsw88$ttl_exp
restricted <- lm(lwage ~ newvar, data = nlsw88)
summary(restricted) # show results
```

```
anova(restricted, multi)
```

Interpretation of coefficients – simple model, 1 regressor

Each additional unit of X1 affects Y by B1



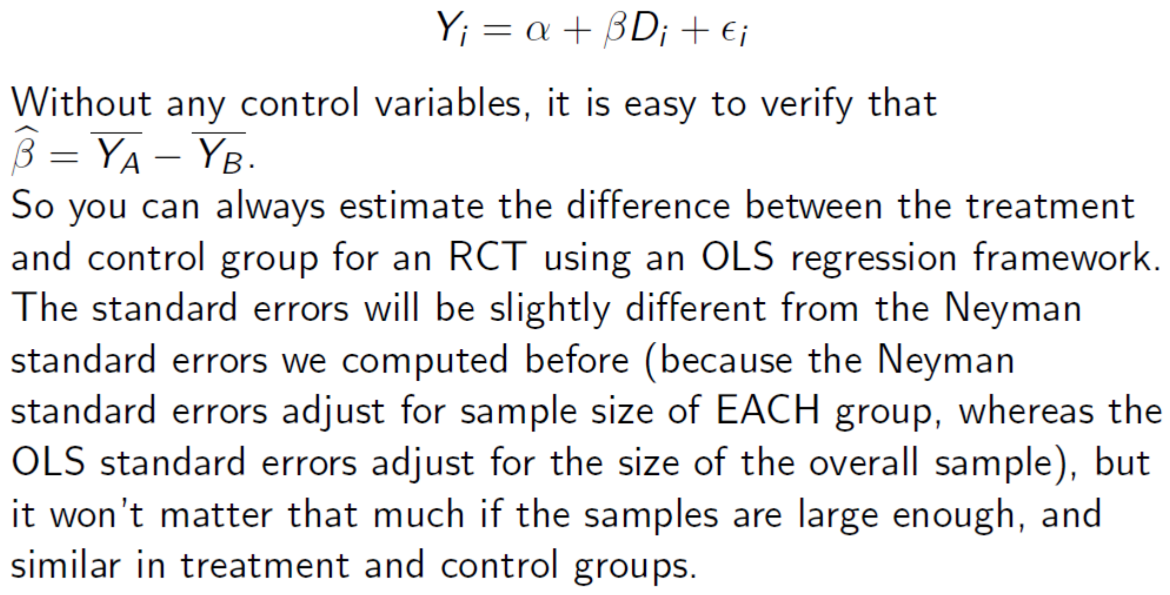
Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

If we estimate the regression above, $\hat{\beta}_1$ will be the estimated effect of the treatment.

Case of 1 dummy regressor

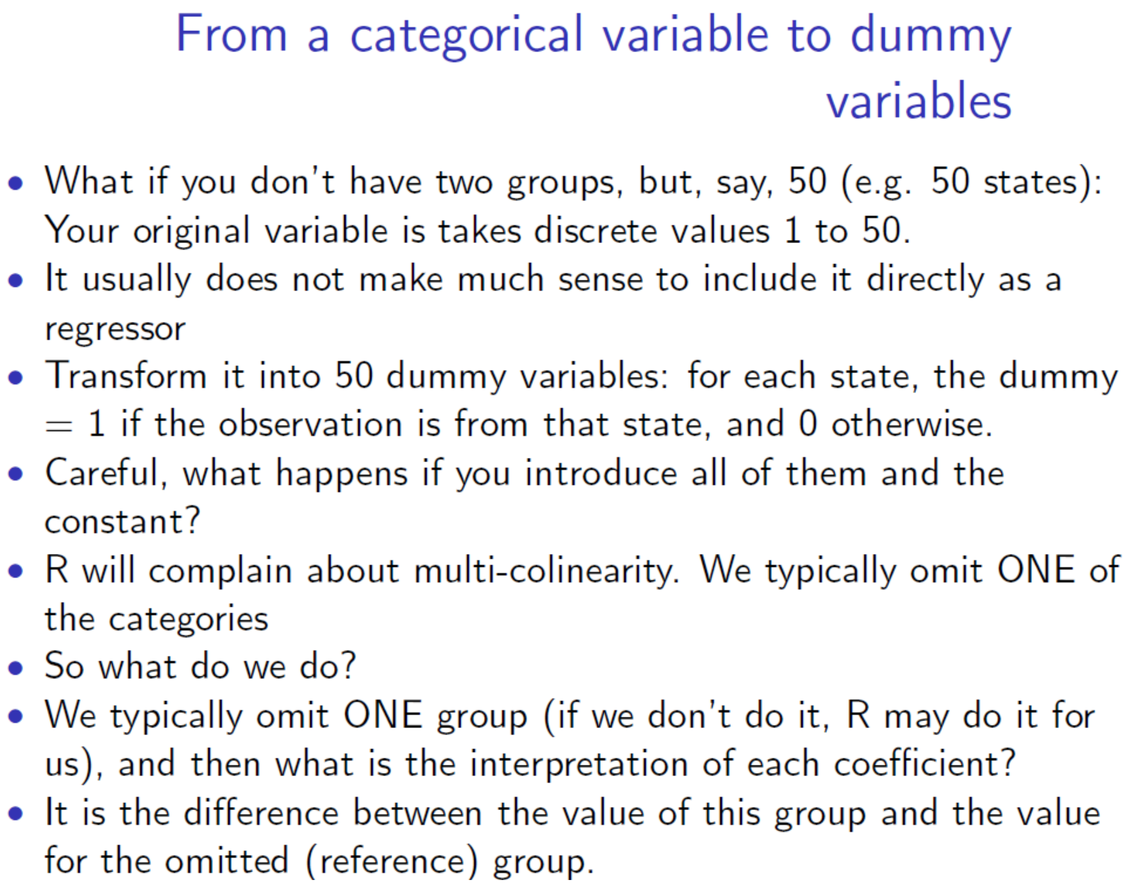
Beta = Avg(Y treated) – Avg(Y control)



$$Y_i = \alpha + \beta D_i + \epsilon_i$$

Without any control variables, it is easy to verify that $\widehat{\beta} = \overline{Y_A} - \overline{Y_B}$.
So you can always estimate the difference between the treatment and control group for an RCT using an OLS regression framework.
The standard errors will be slightly different from the Neyman standard errors we computed before (because the Neyman standard errors adjust for sample size of EACH group, whereas the OLS standard errors adjust for the size of the overall sample), but it won't matter that much if the samples are large enough, and similar in treatment and control groups.

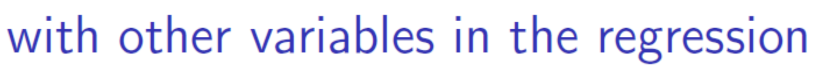
Categorical variables

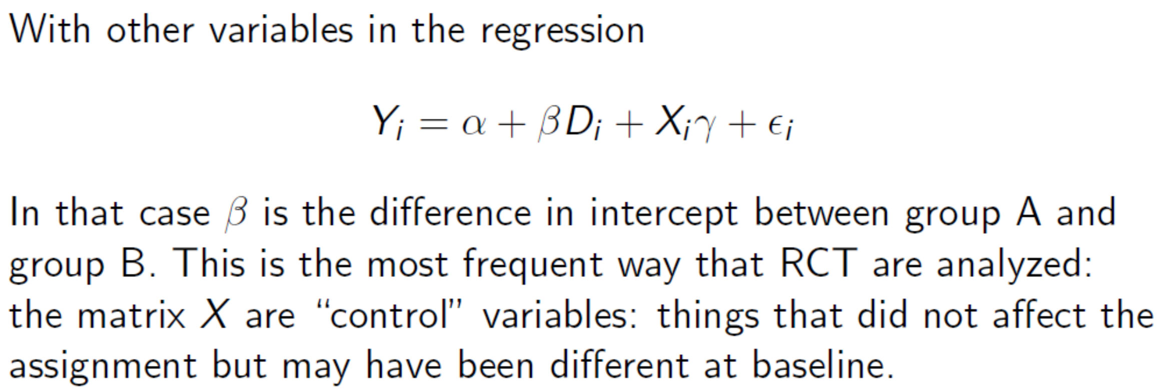


From a categorical variable to dummy variables

- What if you don't have two groups, but, say, 50 (e.g. 50 states): Your original variable is takes discrete values 1 to 50.
- It usually does not make much sense to include it directly as a regressor
- Transform it into 50 dummy variables: for each state, the dummy = 1 if the observation is from that state, and 0 otherwise.
- Careful, what happens if you introduce all of them and the constant?
- R will complain about multi-colinearity. We typically omit ONE of the categories
- So what do we do?
- We typically omit ONE group (if we don't do it, R may do it for us), and then what is the interpretation of each coefficient?
- It is the difference between the value of this group and the value for the omitted (reference) group.

Typical RCT


with other variables in the regression


With other variables in the regression

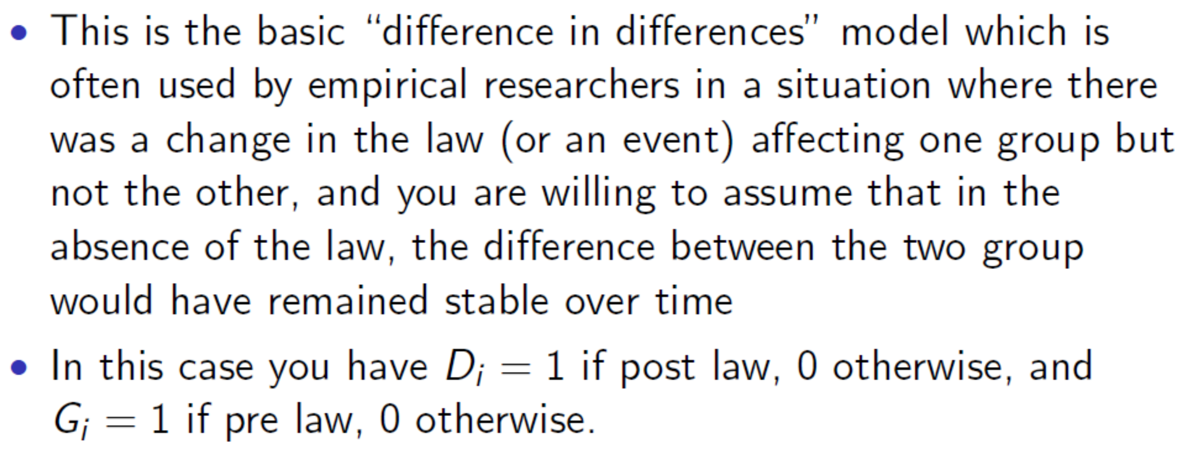
$$Y_i = \alpha + \beta D_i + X_i \gamma + \epsilon_i$$

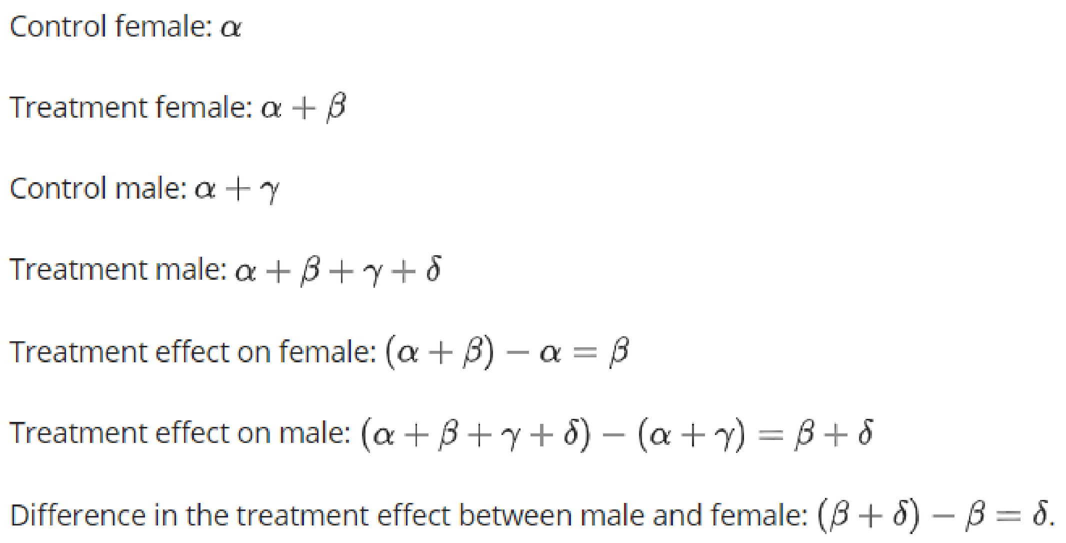
In that case β is the difference in intercept between group A and group B. This is the most frequent way that RCT are analyzed: the matrix X are “control” variables: things that did not affect the assignment but may have been different at baseline.

Interaction terms – Difference in differences

Y= alpha + Beta*(DummyDTreated) + Gamma*(DummyMaleGroup) + Delta* (D*Male) + error

Delta: interaction effect


- This is the basic “difference in differences” model which is often used by empirical researchers in a situation where there was a change in the law (or an event) affecting one group but not the other, and you are willing to assume that in the absence of the law, the difference between the two group would have remained stable over time
- In this case you have $D_i = 1$ if post law, 0 otherwise, and $G_i = 1$ if pre law, 0 otherwise.



Control female: α

Treatment female: $\alpha+\beta$

Control male: $\alpha+\gamma$

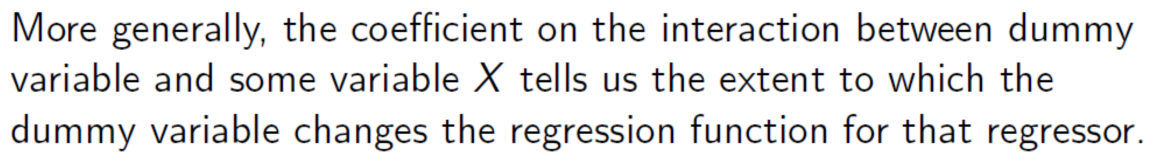
Treatment male: $\alpha+\beta+\gamma+\delta$

Treatment effect on female: $(\alpha+\beta)-\alpha=\beta$

Treatment effect on male: $(\alpha+\beta+\gamma+\delta)-(\alpha+\gamma)=\beta+\delta$

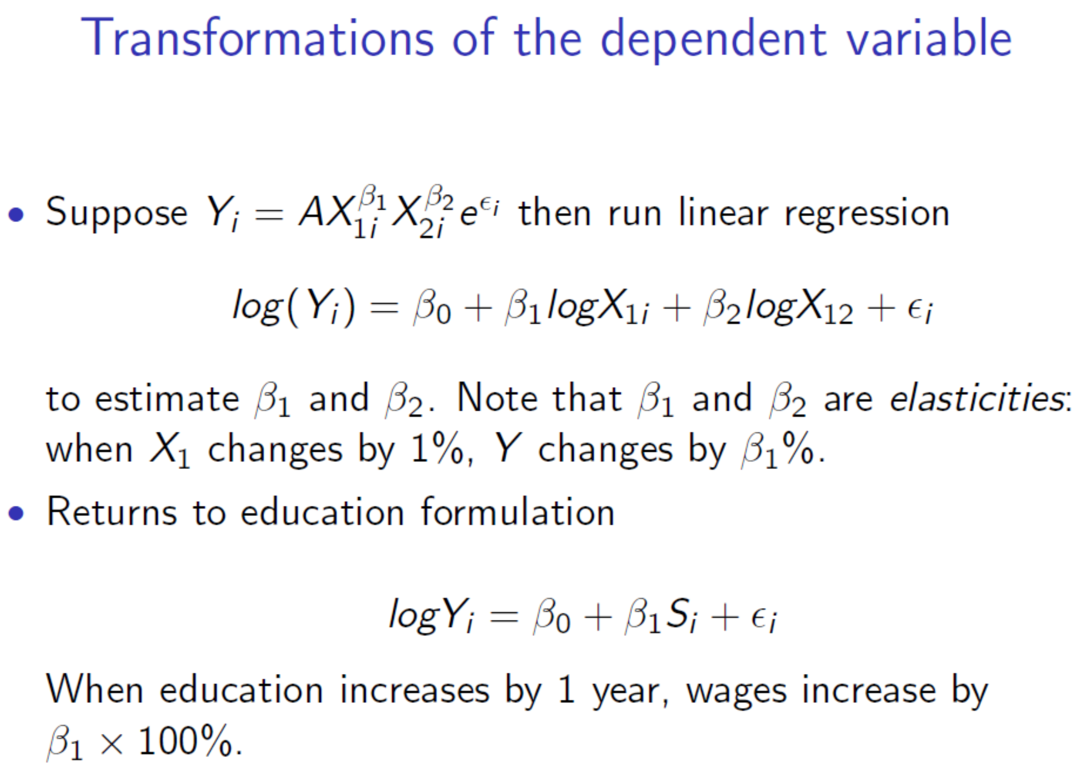
Difference in the treatment effect between male and female: $(\beta+\delta)-\beta=\delta$.

Interaction coefficient



More generally, the coefficient on the interaction between dummy variable and some variable X tells us the extent to which the dummy variable changes the regression function for that regressor.

Transformations of the dependant variable Y / ELASTICITIES



Transformations of the dependent variable

- Suppose $Y_i = AX_{1i}^{\beta_1}X_{2i}^{\beta_2}e^{\epsilon_i}$ then run linear regression

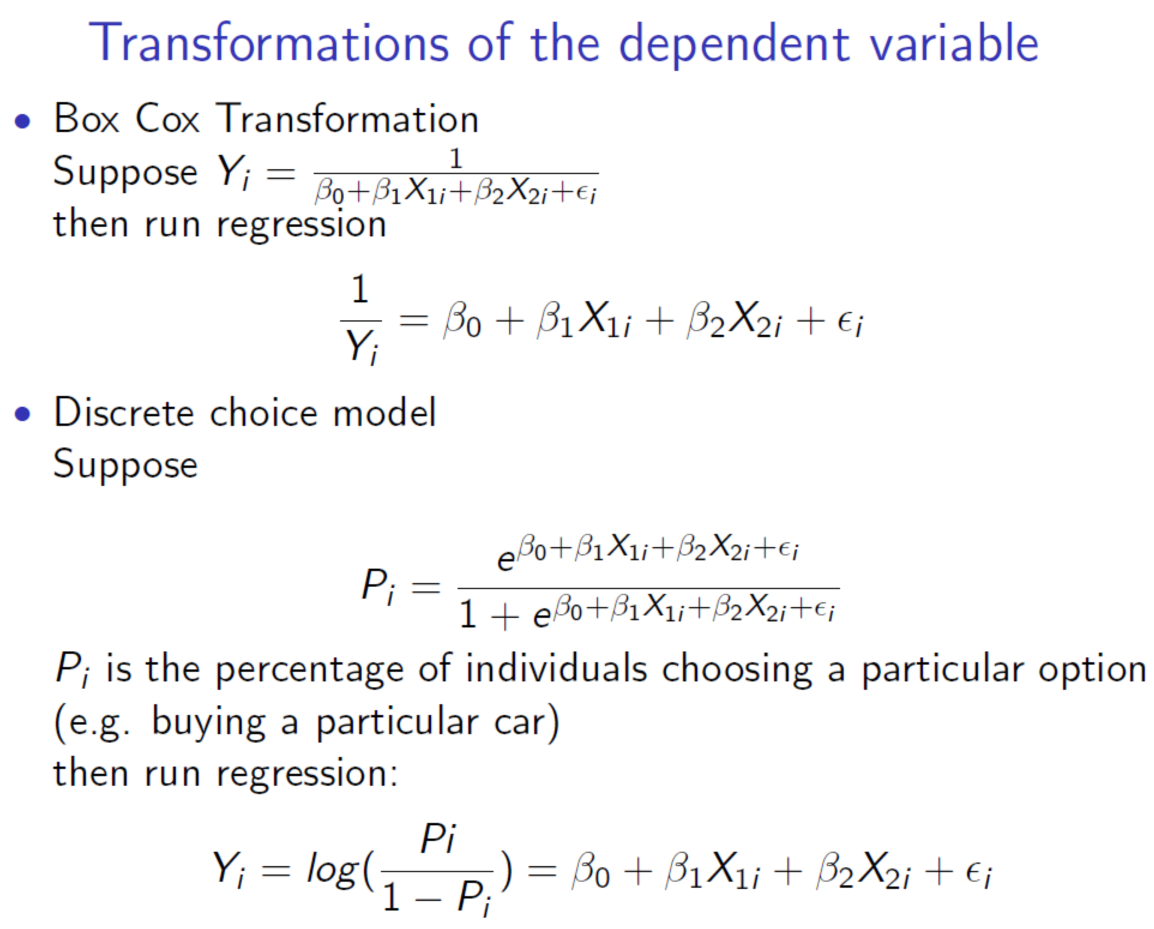
$$log(Y_i)=\beta_0+\beta_1 logX_{1i}+\beta_2 logX_{12}+\epsilon_i$$

to estimate β_1 and β_2. Note that β_1 and β_2 are *elasticities*: when X_1 changes by 1%, Y changes by β_1%.

- Returns to education formulation

$$logY_i=\beta_0+\beta_1 S_i+\epsilon_i$$

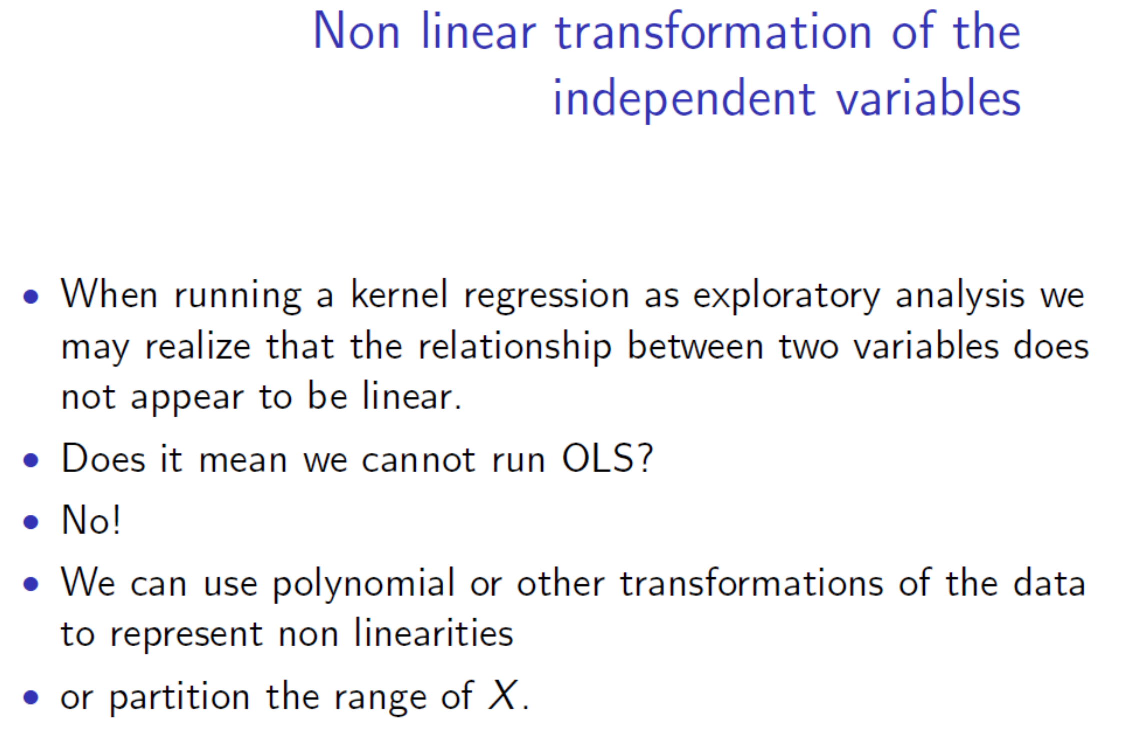
When education increases by 1 year, wages increase by $\beta_1 \times 100\%$.



Transformations of the dependent variable

- Box Cox Transformation
  Suppose $Y_i = \frac{1}{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\epsilon_i}$
  then run regression
  $$\frac{1}{Y_i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$
- Discrete choice model
  Suppose
  $$P_i = \frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\epsilon_i}}{1 + e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\epsilon_i}}$$
  P_i is the percentage of individuals choosing a particular option (e.g. buying a particular car)
  then run regression:
  $$Y_i = log(\frac{Pi}{1-P_i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

Non-linear transformations of Independent Variables

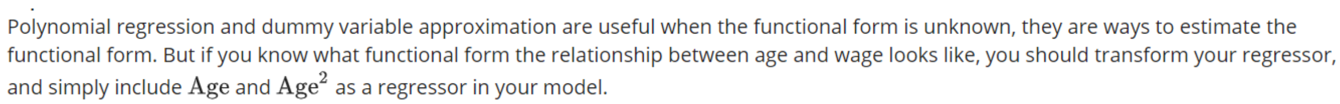


Non linear transformation of the independent variables

- When running a kernel regression as exploratory analysis we may realize that the relationship between two variables does not appear to be linear.
- Does it mean we cannot run OLS?
- No!
- We can use polynomial or other transformations of the data to represent non linearities
- or partition the range of X.

Unknown functional form

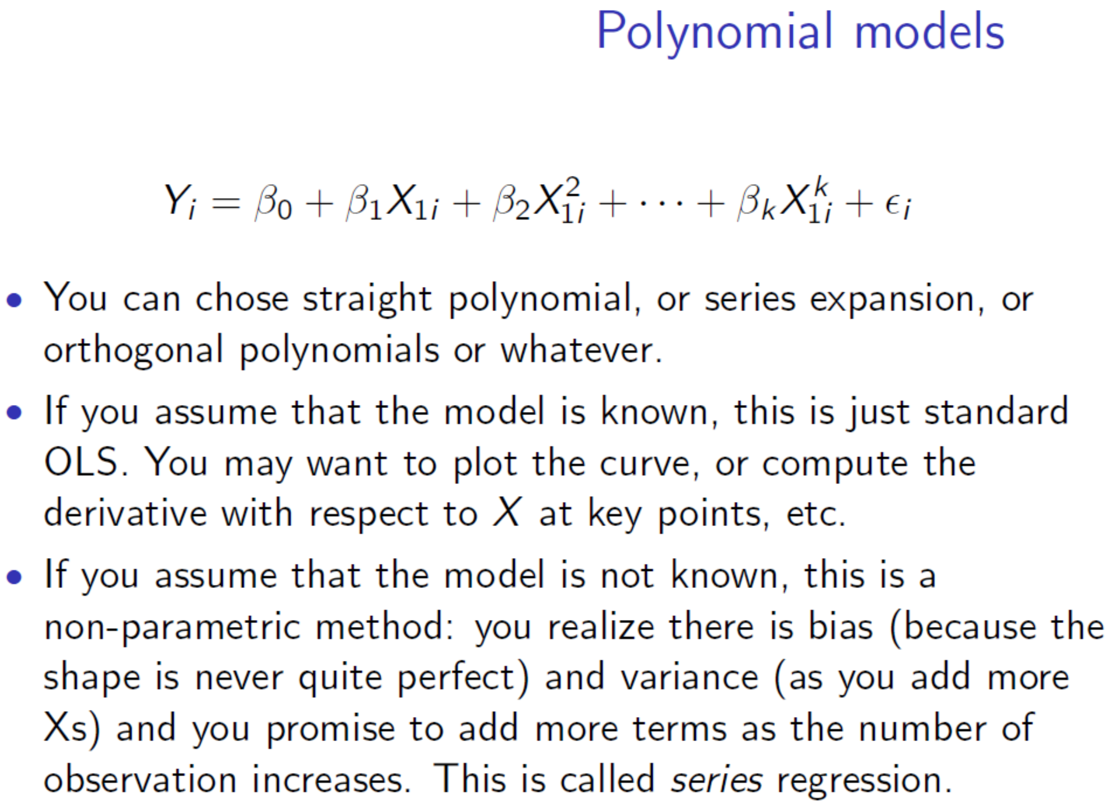
Functional form known vs unknown



Polynomial regression and dummy variable approximation are useful when the functional form is unknown, they are ways to estimate the functional form. But if you know what functional form the relationship between age and wage looks like, you should transform your regressor, and simply include Age and Age^2 as a regressor in your model.

If functional form unknown:

- Series regression (polynomials)
- Dummies approximation
- Local linear regression
- Kernel regression (y is the weighted average for the x's in that interval, the weights are given by a kernel function)



Polynomial models

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \cdots + \beta_k X_{1i}^k + \epsilon_i$$

- You can chose straight polynomial, or series expansion, or orthogonal polynomials or whatever.
- If you assume that the model is known, this is just standard OLS. You may want to plot the curve, or compute the derivative with respect to X at key points, etc.
- If you assume that the model is not known, this is a non-parametric method: you realize there is bias (because the shape is never quite perfect) and variance (as you add more Xs) and you promise to add more terms as the number of observation increases. This is called *series* regression.

Using dummies for approximation

- Partition the range of X is interval, $X_0, \dots X_J$
- Define the dummies as:

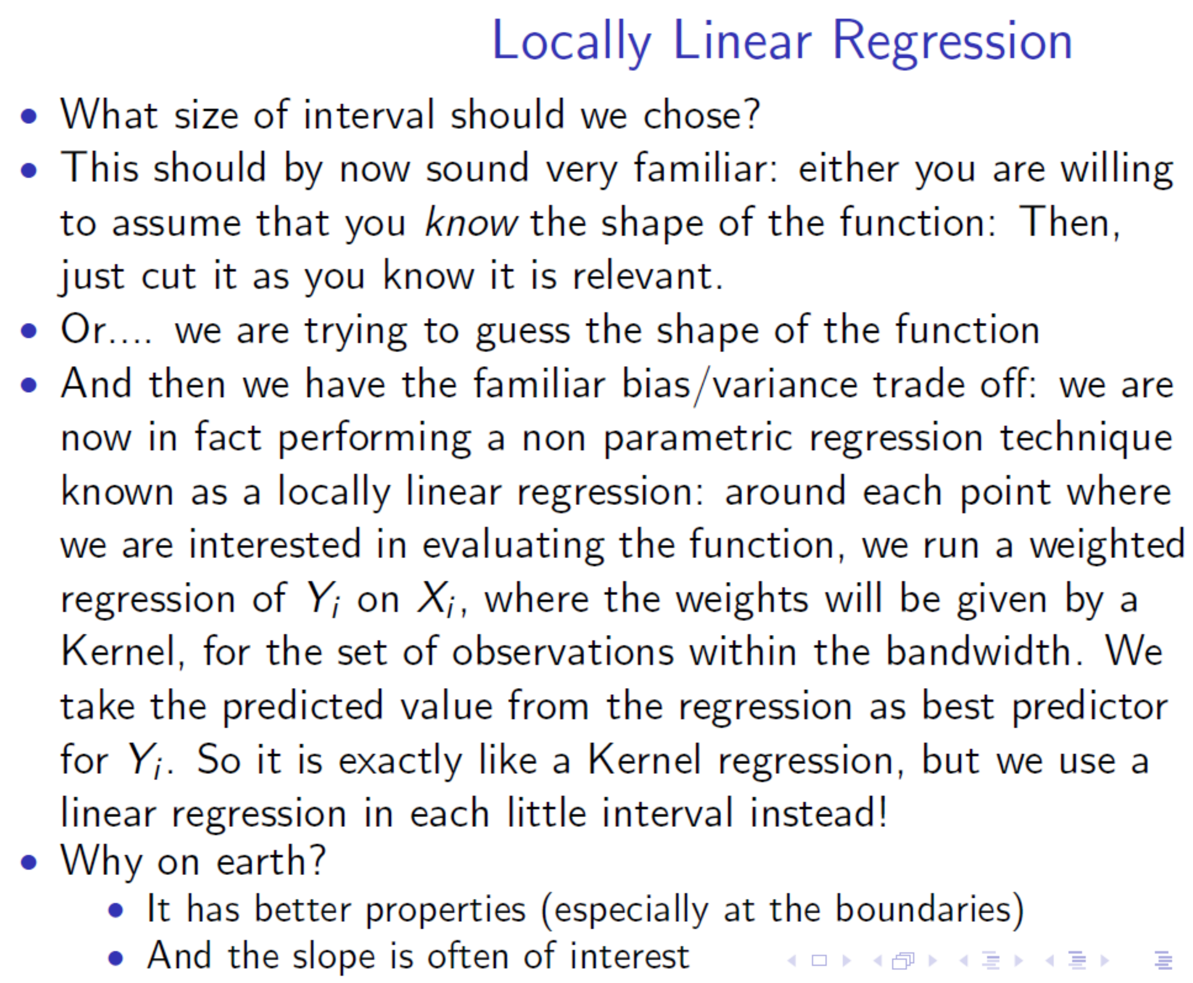
  $D_{1i} = I_{[X_0 \leq X_{1i} < X_1]}$
  $D_{2i} = I_{[X_1 \leq X_{1i} < X_2]}$
  $\vdots$
  $D_{ji} = I_{[X_{J-1} \leq X_{1i} < X_J]}$
- you can run regression:

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \cdots + \beta_J D_{ji} + \epsilon_i$$

  (note no intercept. why?)
- Define Piece wise linear variables as:

  $S_{1i} = I_{[X_0 \leq X_{1i} < X_1]}(X_{1i} - X_1)$ $S_{2i} = I_{[X_1 \leq X_{1i} < X_2]}(X_{1i} - X_2)$
- Run regression

$$Y_i = \beta_1 X_{1i} + \beta_2 S_{1i} + \cdots + \beta_J S_{j-1i} + \epsilon_i$$



Locally Linear Regression

- What size of interval should we chose?
- This should by now sound very familiar: either you are willing to assume that you *know* the shape of the function: Then, just cut it as you know it is relevant.
- Or.... we are trying to guess the shape of the function
- And then we have the familiar bias/variance trade off: we are now in fact performing a non parametric regression technique known as a locally linear regression: around each point where we are interested in evaluating the function, we run a weighted regression of Y_i on X_i, where the weights will be given by a Kernel, for the set of observations within the bandwidth. We take the predicted value from the regression as best predictor for Y_i. So it is exactly like a Kernel regression, but we use a linear regression in each little interval instead!
- Why on earth?
  - It has better properties (especially at the boundaries)
  - And the slope is often of interest

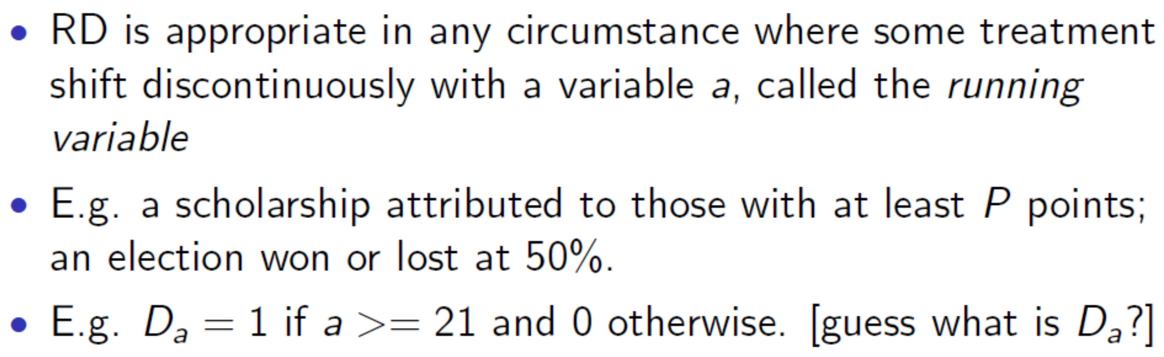
Comparison kernel regression, series regression, local linear regression

	Kernel Regression	**Series Regression**
Trade-off	For fixed N, a more flexible functional form (smaller band-width) increases variance and decreases bias.	For fixed N, a more flexible functional form (more terms in the polynomial) increases variances and decreases bias.
Promise	To make your bandwidth smaller as N increases.	To increase the number of terms in your polynomial as your sample size increases.
Cross Validation	The trade-off between bias and variance by using cross validation criterion, usually by minimizing mean squared error.	

Local linear regression performs better than kernel regressions at the boundaries, because they predict a line to the boundary whereas in kernel regression the predicted value is the weighted mean of observations within that interval.

Regression Continuity Design RD (lecture 20)

In statistics, econometrics, political science, epidemiology, and related disciplines, a **regression discontinuity design (RDD)** is a quasi-experimental **pretest-posttest** design that elicits the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned. By comparing observations lying closely on either side of the threshold, it is possible to estimate the average treatment effect in environments in which randomization is unfeasible.

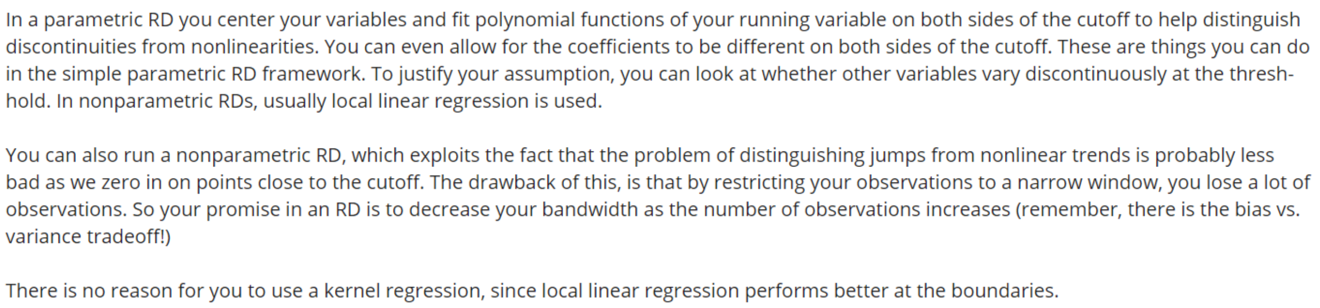


- RD is appropriate in any circumstance where some treatment shift discontinuously with a variable a, called the *running variable*
- E.g. a scholarship attributed to those with at least P points; an election won or lost at 50%.
- E.g. $D_a = 1$ if $a >= 21$ and 0 otherwise. [guess what is D_a?]

Running variable

The supposed cause

Ex: age of 21 just attained



In a parametric RD you center your variables and fit polynomial functions of your running variable on both sides of the cutoff to help distinguish discontinuities from nonlinearities. You can even allow for the coefficients to be different on both sides of the cutoff. These are things you can do in the simple parametric RD framework. To justify your assumption, you can look at whether other variables vary discontinuously at the threshold. In nonparametric RDs, usually local linear regression is used.

You can also run a nonparametric RD, which exploits the fact that the problem of distinguishing jumps from nonlinear trends is probably less bad as we zero in on points close to the cutoff. The drawback of this, is that by restricting your observations to a narrow window, you lose a lot of observations. So your promise in an RD is to decrease your bandwidth as the number of observations increases (remember, there is the bias vs. variance tradeoff!)

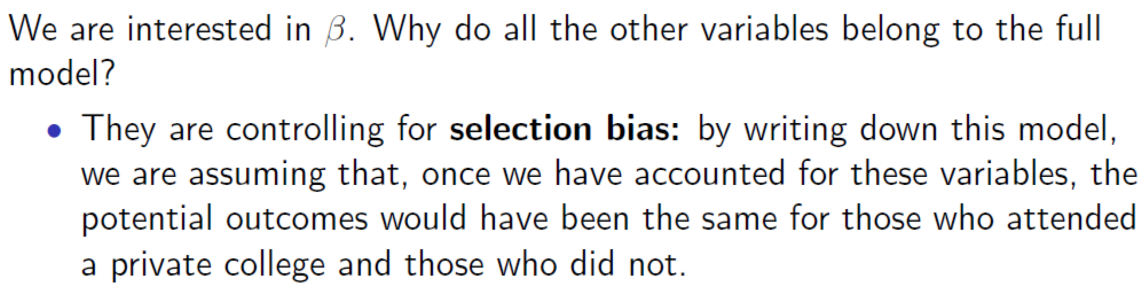
There is no reason for you to use a kernel regression, since local linear regression performs better at the boundaries.

Attention

Fit a polynomial on each side of the discontinuity: simplest design may take non-linearities for discontinuities.

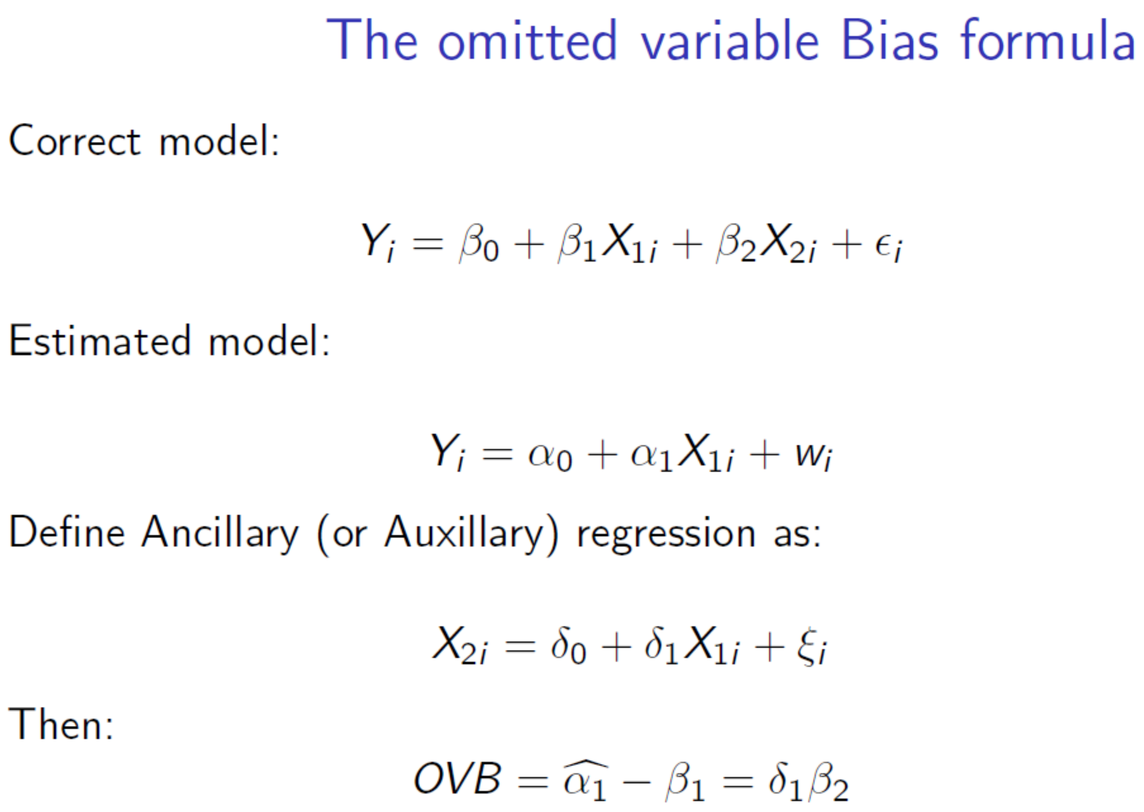
We can also solve the problem by narrowing the estimate to a band around the discontinuity (the bandwidth!). As usual, the risk is bias vs variance: if we promise to narrow the bandwidth as the number of observation increases, we now have a non –parametric RD!

Omitted Variable Bias OVB



We are interested in β. Why do all the other variables belong to the full model?

- They are controlling for **selection bias:** by writing down this model, we are assuming that, once we have accounted for these variables, the potential outcomes would have been the same for those who attended a private college and those who did not.



The omitted variable Bias formula

Correct model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

Estimated model:

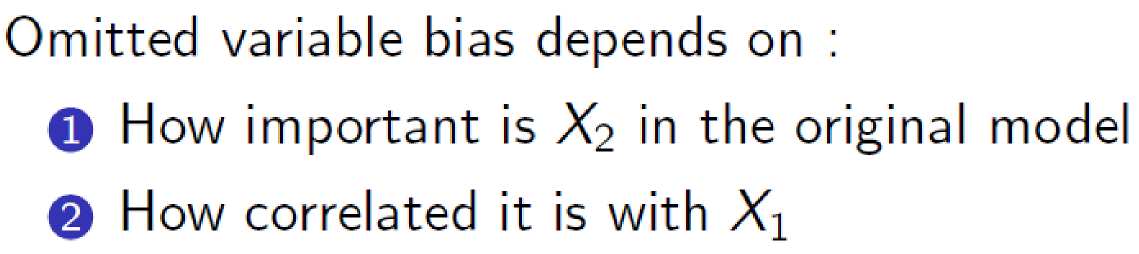
$$Y_i = \alpha_0 + \alpha_1 X_{1i} + w_i$$

Define Ancillary (or Auxillary) regression as:

$$X_{2i} = \delta_0 + \delta_1 X_{1i} + \xi_i$$

Then:

$$OVB = \widehat{\alpha_1} - \beta_1 = \delta_1 \beta_2$$



Omitted variable bias depends on :

1. How important is X_2 in the original model
2. How correlated it is with X_1

Bivariate derivation

It is worth spending some time with this formula because it is going to stay with you for your entire life as a data scientist!
Remember OLS bivariate formula:

$$\widehat{\alpha_1} = \frac{Cov(Y_i, X_{1i})}{V(X_{1i})}$$

substituting for Y_i we get:

$$\frac{Cov(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, X_{1i})}{V(X_{1i})}$$

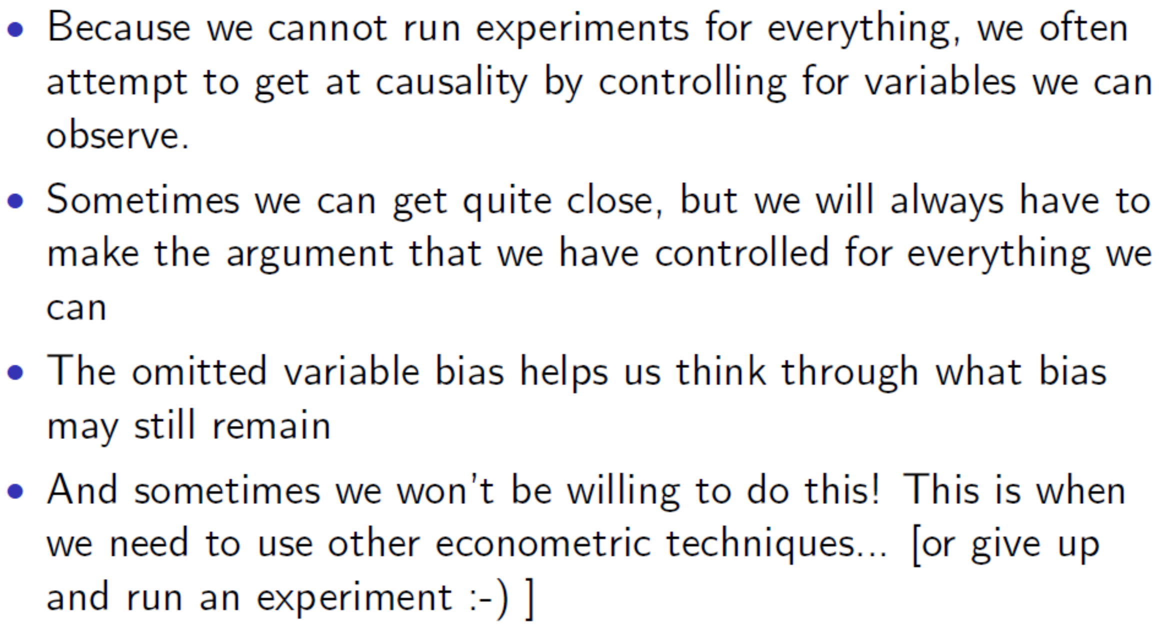
$$= \frac{\beta_1 V(X_{1i}) + \beta_2 Cov(X_{2i}, X_{1i}) + Cov(\epsilon_i, X_{1i})}{V(X_{1i})}$$

$$= \beta_1 + \delta_1 \beta_2$$

How do we use the OVB formula in general?

- Most of the times we don't have the variables we are not including... otherwise we would include them!
- So how is the OVB formula useful?
- It guides our economic thinking on whether the bias would be important
    - When we are running a regression, are we omitting variables that are likely to be important determinant of the *outcome*
    - And are they likely to be correlated with the *regressor of interest*

Conclusion OVB



- Because we cannot run experiments for everything, we often attempt to get at causality by controlling for variables we can observe.
- Sometimes we can get quite close, but we will always have to make the argument that we have controlled for everything we can
- The omitted variable bias helps us think through what bias may still remain
- And sometimes we won't be willing to do this! This is when we need to use other econometric techniques... [or give up and run an experiment :-)]

ENDOGENEITY AND INSTRUMENTAL VARIABLES (IV)

$$Y_i = \alpha + \beta A_i + \epsilon_i$$

Endogeneity when Y and A have a mutual effect one on the other.

Endogeneity when A is correlated with the error term.

IV is a response to:

- Endogeneity
- When A cannot be assigned randomly

Assign an instrument Z that will affect/ represent A.

Three conditions make Z a good instrument

1. It affects A_i : $E[A_i|Z_i = 1] - E[A_i|Z_i = 0]$
2. It is randomly assigned, or as good as randomly assigned, so that $E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$ can be interpreted as the causal effect of Z on Y [when we use RCT as an instrument this is guaranteed, otherwise it needs to be checked]
3. It has no direct effect on Y (exclusion restriction). This may or may not be true, has to be argued on a case by case basis, and cannot be tested.

$$\hat{\beta} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[A_i|Z_i = 1] - E[A_i|Z_i = 0]}$$

Equation 1 is the *first stage* relationship (the denominator). Equation 2 is the *reduced form* relationship (the numerator). Therefore, the Wald Estimate is the reduced form divided by the first stage. $\hat{\beta}$, given by the equation above, is the *Wald estimate* of the effect of SHS participation on Y_i. It is the simplest form of the instrumental variable estimator (Z_i is our instrument).

The interpretation of IV when the treatment effect is not constant

- In reality the effect of going to school on test scores is likely to be different for different children. [remember that when we first introduced causality we did not assume constant treatment effect]
- In that case the simple calculation we just did does not apply
- Yet, under a fairly mild assumption, the Wald estimate still has a causal interpretation, which is in fact quite intuitive: it captures the effect of the treatment on those who are compelled by the instrument to get treated: this is the *Local Average Treatment Effect*, or LATE.

(only the compliers, those who respond positively to the instrument, group participate to the calculous)

From the Wald Estimate to two state Least Squares

- Instead of computing differences in means and taking the ratio, we could have couched this in a regression framework.
- First stage , $\widehat{\pi_1}$ in the equation: $A_i = \pi_0 + \pi_1 Z_i + \upsilon_i$
- Reduced form : $\widehat{\gamma_1}$ in the equation $Y_i = \gamma_0 + \gamma_1 Z_i + \omega_i$
- Two stage least square: Run the first stage, and take the fitted values $\widehat{A}_i$,
- Then, in the second stage, run: $Y_i = \alpha + \beta \widehat{A}_i + \epsilon_i$

The two stage least squares and the Wald estimates are identical

$$
\begin{aligned}
\widehat{\beta} &= \frac{Cov(Y_i, \widehat{A}_i)}{Var(\widehat{A}_i)} \\
&= \frac{Cov(Y_i, \pi_0 + \pi_1 Z_i)}{Var(\pi_0 + \pi_1 Z_i)} \\
&= \frac{\pi_1 Cov(Y_i, Z_i)}{\pi_1^2 Var(Z_i)} = \frac{\gamma_1}{\pi_1}
\end{aligned}
$$

More generally: two stage least squares

Consider the model

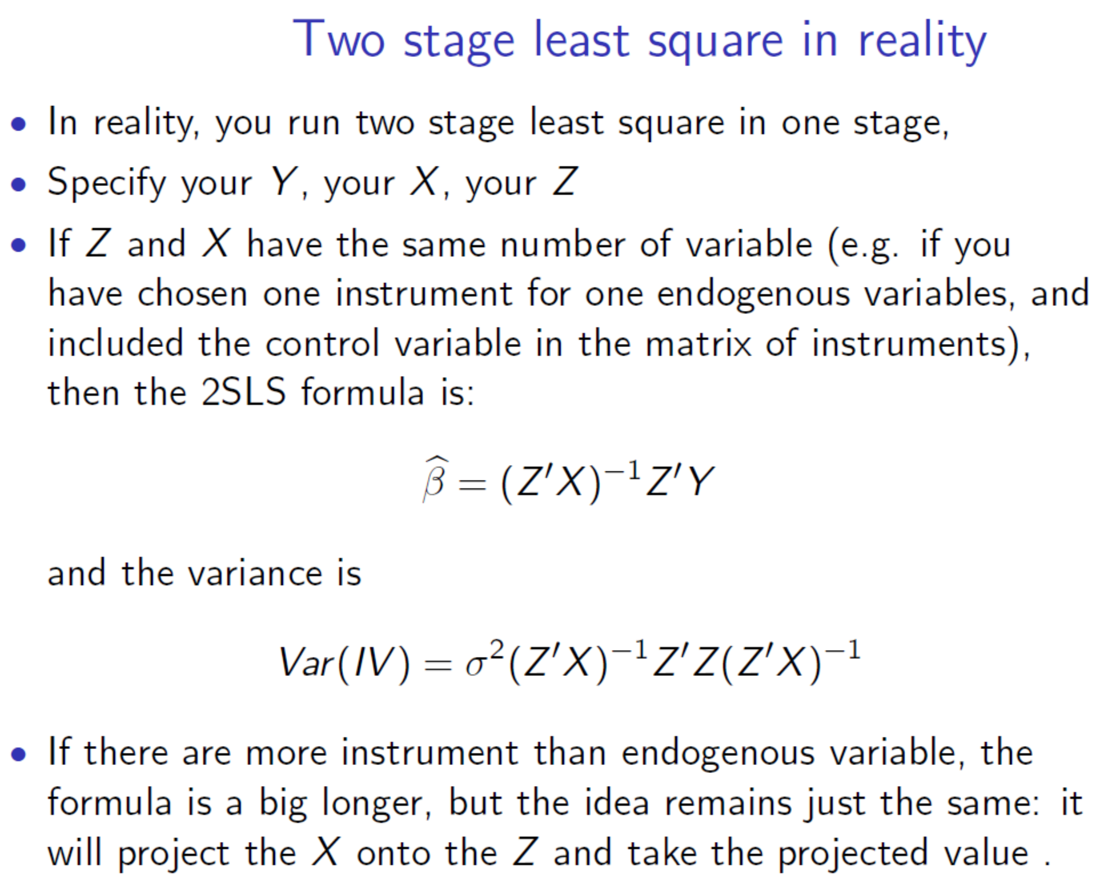
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

where X_{i1} a vector of endogenous variables, and X_{2i} some variables that you assume are not endogenous (control variables).
Look for an instrument. In this case you will need at least one instrument, you could have more. Denote Z the matrix $(Z_1, ...Z_k, X_2)$ [in other words the control variables, which do not need to be instrumented, are part of the matrix of instruments.

Intuitive steps:

- First stage: $X_{1i} = \pi_o + \pi_1 Z_1 + \pi_2 Z_2 + \cdots + \pi_k Z_k + X_{2i} + \omega_i$
- Second stage: $Y = \beta_0 + \beta_1 \widehat{X_{1i}} + X_{i2} + \epsilon_i$

In practice if you do that point estimates will be correct, but the standard errors and all the tests will be wrong (because you have estimated your first stage, rather than knowing it, and the standard errors must reflect this uncertainty).



Two stage least square in reality

- In reality, you run two stage least square in one stage,
- Specify your Y, your X, your Z
- If Z and X have the same number of variable (e.g. if you have chosen one instrument for one endogenous variables, and included the control variable in the matrix of instruments), then the 2SLS formula is:

$$\widehat{\beta} = (Z'X)^{-1}Z'Y$$

and the variance is

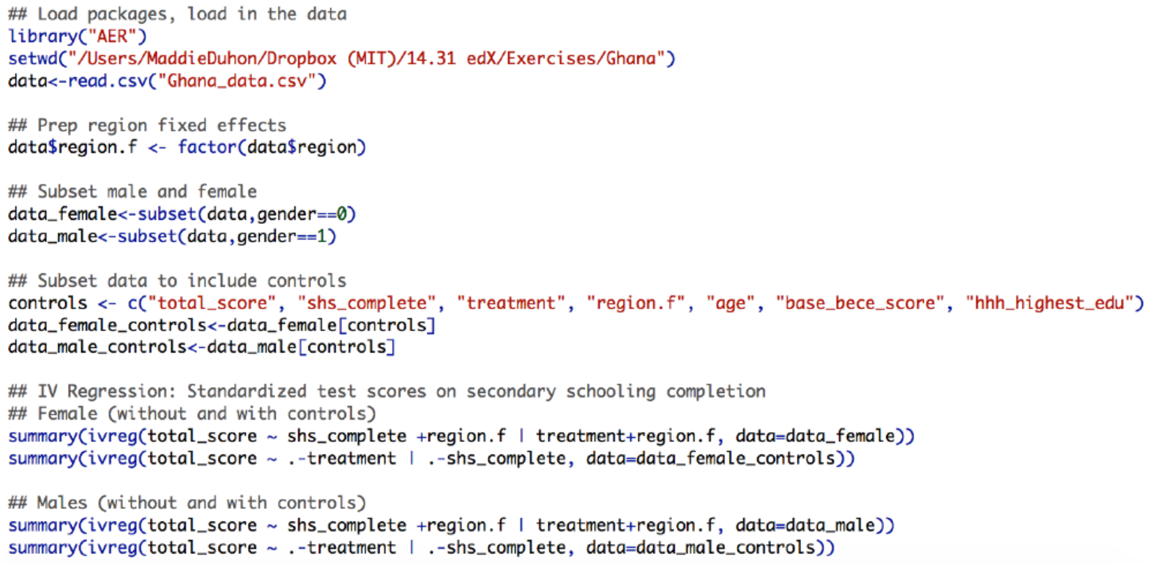
$$Var(IV) = \sigma^2(Z'X)^{-1}Z'Z(Z'X)^{-1}$$

- If there are more instrument than endogenous variable, the formula is a big longer, but the idea remains just the same: it will project the X onto the Z and take the projected value .

IV in R

(model original with endogeneous variable incl. control variables | control variables + instrument)

```
ivb1 <- ivreg(workedm ~ three + blackm + hispm + othracem |
blackm + hispm + othracem + samesex, data = census80)
IVb[1, 1] <- ivb1$coefficients[2]
pvalue <- summary(ivb1)
IVb[2, 1] <- pvalue$coefficients[2, 4]
 ivb2 <- ivreg(weeksm ~ three + blackm + hispm + othracem |
blackm + hispm + othracem + samesex, data = census80)
IVb[1, 2] <- ivb2$coefficients[2]
pvalue <- summary(ivb2)
IVb[2, 2] <- pvalue$coefficients[2, 4]
IVb
```



```
## Load packages, load in the data
library("AER")
setwd("/Users/MaddieDuhon/Dropbox (MIT)/14.31 edX/Exercises/Ghana")
data<-read.csv("Ghana_data.csv")

## Prep region fixed effects
data$region.f <- factor(data$region)

## Subset male and female
data_female<-subset(data,gender==0)
data_male<-subset(data,gender==1)

## Subset data to include controls
controls <- c("total_score", "shs_complete", "treatment", "region.f", "age", "base_bece_score", "hhh_highest_edu")
data_female_controls<-data_female[controls]
data_male_controls<-data_male[controls]

## IV Regression: Standardized test scores on secondary schooling completion
## Female (without and with controls)
summary(ivreg(total_score ~ shs_complete +region.f | treatment+region.f, data=data_female))
summary(ivreg(total_score ~ .-treatment | .-shs_complete, data=data_female_controls))

## Males (without and with controls)
summary(ivreg(total_score ~ shs_complete +region.f | treatment+region.f, data=data_male))
summary(ivreg(total_score ~ .-treatment | .-shs_complete, data=data_male_controls))
```

EXPERIMENT DESIGN