

Lecture 6-b Model Specification

Brooks (4th edition): Chapters 3 & 4

© R. Susmel, 2023 (for private use, not to be posted/shared online).

1

Review: OLS Estimation - Assumptions

- CLM Assumptions

(A1) DGP: $\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is correctly specified.

(A2) $E[\boldsymbol{\varepsilon} | \mathbf{X}] = 0$

(A3) $\text{Var}[\boldsymbol{\varepsilon} | \mathbf{X}] = \sigma^2 \mathbf{I}_T$

(A4) \mathbf{X} has full column rank $\rightarrow \text{rank}(\mathbf{X}) = k$, where $T \geq k$.

Q: What happens when (A1) is not correctly specified?

- First, we looked at (A1), in the context of linearity. Are we omitting a relevant regressor? Are we including an irrelevant variable? What happens when we impose restrictions in the DGP?
- Second, in (A1), we allow some non-linearities in its functional form.

Review: Specification – Omitted & Irrelevant X

- Omitting relevant variables: Suppose the correct model (DGP) is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \quad \text{–the “long regression,” with } \mathbf{X}_1 \text{ \& } \mathbf{X}_2.$$

But, we compute OLS omitting \mathbf{X}_2 , a true driver of \mathbf{y} . That is,

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \quad \text{–the “short regression.”}$$

Implication: Restricted estimator \mathbf{b}^* is **biased**, but **more efficient**.

- Irrelevant variables . Suppose the correct model is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \quad \text{–the “short regression,” with } \mathbf{X}_1$$

But, we estimate, ignoring the true restriction $\boldsymbol{\beta}_2 = 0$:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \quad \text{–the “long regression.”}$$

Implication: Estimator \mathbf{b} is **unbiased**, but **inefficient**.

Review: Trilogy of Tests – LR, Wald & LM

- Given that omitting explanatory variables is a big problem (bias estimation!), we use tests to check the specification of the model. We test $H_0: \boldsymbol{\beta}_J = 0$, where $\boldsymbol{\beta}_J$ is the vector of coefficients for the J variables we consider omitting.

We have three asymptotic tests that follow the same χ^2_J distribution:

- **Wald test**, W – estimates Unrestricted Model.
- **Likelihood Ratio test**, LR – estimates both Unrestricted and Restricted Models and assume a distribution (usually, normality).
- **Lagrange Multiplier test**, LM – estimates only Restricted Models.

We like LM tests because only the Restricted Model is estimated. If we reject $H_0: \boldsymbol{\beta}_J = 0$, then, we re-specify the model: We need to add the J explanatory variables.

Review: Trilogy of Tests – LR, Wald & LM

• Given that omitting explanatory variables is a big problem (bias estimation!), we use tests to check the specification of the model. We test $H_0: \beta_J = 0$, where β_J is the vector of coefficients for the J variables we consider omitting.

We have three asymptotic tests that follow the same χ_J^2 distribution:

- **Wald test**, W – estimates Unrestricted Model.
- **Likelihood Ratio test**, LR – estimates both Unrestricted and Restricted Models and assume a distribution (usually, normality).
- **Lagrange Multiplier test**, LM – estimates only Restricted Models.

We like LM tests because only the Restricted Model is estimated. If we reject $H_0: \beta_J = 0$, then, we re-specify the model: We need to add the J explanatory variables.

Review: Trilogy of Tests – Wald

• In our general framework, we test $H_0: R\beta = q$. In the particular case of testing for J omitted variables, $q = 0$. The $J \times k$ R matrix is a matrix of zeros, with ones for the omitted variables, for example, if we omit in the 3-factor FF model, SMB & HML, elements (1,3) and (2,4) will have a 1. In this example, the restriction to test will be:

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \beta_{Mkt} \\ \beta_{SMB} \\ \beta_{HML} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

We based the Wald test on the unrestricted OLS b :

$$F = W^*/J = (Rb)' \{R[s^2(X'X)^{-1}]R'\}^{-1} Rb$$

Distribution under H_0 if (A5) $F = W^*/J \sim F_{J, T-k}$

if not (A5) $J^*F \xrightarrow{d} \chi_J^2$

Review: Trilogy of Tests – LR

- The LR test is based on the (log) **Likelihood**. It requires two ML estimations:
 - The unrestricted estimation, producing $\hat{\theta}_{ML}$
 - The restricted estimation, producing $\hat{\theta}^R$.

Then, the LR test:

$$LR = 2[\log(L(\hat{\theta}_{ML})) - \log(L(\hat{\theta}^R))] \xrightarrow{d} \chi^2_f$$

Note: MLE requires **assuming a distribution**, usually, a normal.

Technical note: The LR has a very good property: It is a *consistent test*. An asymptotic test which rejects H_0 with probability one when the H_1 is true is called a *consistent test*. That is, a consistent test has asymptotic power of 1.

Review: Trilogy of Tests – LM

- The LM test needs only one estimation: the restricted estimation, that is, imposing $H_0: \mathbf{R}\beta = \mathbf{q}$, producing $\hat{\theta}^R$.

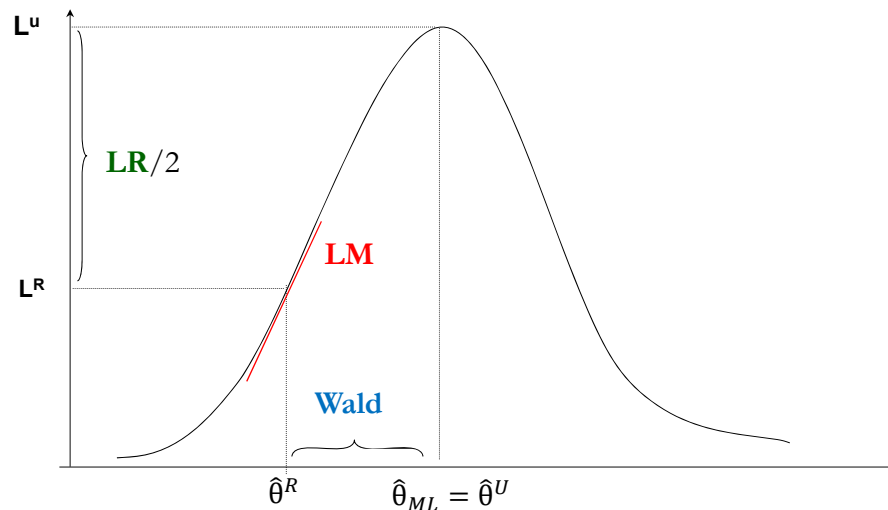
Then, if the restriction is true, then the slope of the objective function (say, the Likelihood) at $\hat{\theta}^R$ should be zero. The slope is called the Score, $S(\hat{\theta}^R)$.

- The LM test is based on a Wald test on $H_0: S(\hat{\theta}^R) = 0$.

$$LM = S(\hat{\theta}^R)' [Var(S(\hat{\theta}^R))]^{-1} S(\hat{\theta}^R) \xrightarrow{d} \chi^2_f$$

It turns out that there is a much simpler formulation for the LM test, based on the residuals of the restricted model. We will present this version of the test next.

Review: Trilogy of Tests – LR, Wald & LM



Remark: Asymptotically equivalent, but, for small T , in general, $W > LR > LM$.

Model Specification with LM Tests

- The popular version of the LM test involves the following steps:

(1) Run restricted model ($\mathbf{y} = \mathbf{X} \boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$). Get restricted residuals, \mathbf{e}_R .

(2) (Auxiliary Regression). Run the regression of \mathbf{e}_R on all the omitted J variables, \mathbf{Z} , and the k included variables, \mathbf{X} . In our case:

$$e_{R,i} = \alpha_0 + \alpha_1 x_{1,i} + \dots + \alpha_k x_{k,i} + \gamma_1 z_{1,i} + \dots + \gamma_J z_{J,i} + v_i$$

\Rightarrow Keep the R^2 from this regression, R_{eR}^2 .

(3) Compute LM-statistic:

$$LM = T * R_{eR}^2 \xrightarrow{d} \chi_J^2.$$

- Here, we use the LM test to check (A1). But, the LM test is very general. It can be used in many settings, for example, to test for nonlinearities, autocorrelation, heteroscedasticity, etc.

10

Model Specification with LM Tests

Example: We use an LM test to check if the standard CAPM for IBM returns omits **SMB** and **HML**. ($J = 2$)

```
fit_ibm_capm <- lm (ibm_x ~ Mkt_RF)           # Restricted Model
resid_r <- fit_ibm_capm$residuals             # extract residuals from R model
fit_lm <- lm (resid_r ~ Mkt_RF + SMB + HML)    # auxiliary regression
> summary(fit_lm)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0007021	0.0024875	0.282	0.7779
Mkt_RF	0.0125253	0.0567221	0.221	0.8253
SMB	-0.2124596	0.0841119	-2.526	0.0118 *
HML	-0.1715002	0.0846817	-2.025	0.0433 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05848 on 565 degrees of freedom

Multiple R-squared: **0.01649**, Adjusted R-squared: 0.01127

F-statistic: 3.158 on 3 and 565 DF, p-value: 0.02438

11

Model Specification with LM Tests

Example (continuation):

```
R2_r <- summary(fit_lm)$r.squared           # extracting R^2 from fit_lm
> R2_r
[1] 0.01649104

LM_test <- R2_r * T
> LM_test
[1] 9.383402                                ⇒ LM_test > qchisq (.95,df=2) ⇒ Reject H0.

> qchisq(.95, df = 2)                     # chi-squared (df=2) value at 5% level
[1] 5.991465

p_val <- 1 - pchisq(LM_test, df = 2)       # p-value of LM_test
> p_val
[1] 0.009171071                            ⇒ p-value is small ⇒ Reject H0.
```

Conclusion: We need to respecify the CAPM. Given the results of the LM test we need to add **SMB** and **HML**.

12

Functional Form: Linearity in Parameters

- So far, our models have been linear in variables and parameters:

$$y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon.$$

\Rightarrow OLS estimates all parameters: $\beta_1, \beta_2, \beta_3,$ & β_4 .

- But OLS can handle non-linear models in variables, as long as linearity in parameters is preserved –i.e., *intrinsic linear model*:

$$y = \beta_1 + \beta_2 X_2^2 + \beta_3 \sqrt{X_3} + \beta_4 \log X_4 + \varepsilon$$

Define: $Z_2 = X_2^2$, $Z_3 = \sqrt{X_3}$, & $Z_4 = \log X_4$

Then, the non-linear model becomes a linear model:

$$y = \beta_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \varepsilon$$

\Rightarrow OLS can be used to estimate all $\beta_1, \beta_2, \beta_3,$ & β_4 .

13

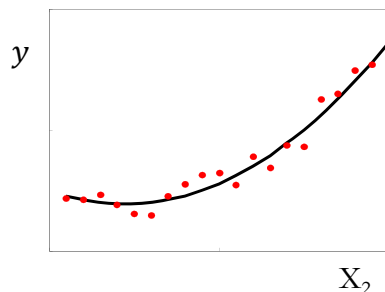
7

Functional Form: Linearity in Parameters

- Suppose we have:

$$y = \beta_1 + \beta_2 X_2 + \beta_3 X_2^2 + \varepsilon$$

This model allows for a quadratic relation between y and X_2 :



- Let $X_3 = X_2^2$, then, the model is intrinsic linear:

$$y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

14

Functional Form: Linearity in Parameters

Example: We do a Wald test to check if a measure of market risk $(r_{m,t} - r_f)^2$ is significant in the 3 FF factor model for IBM returns.

$$(r_t - r_f) = \beta_0 + \beta_1 (r_{m,t} - r_f) + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 (r_{m,t} - r_f)^2 + \varepsilon_t$$

We can do OLS, by redefining the variables: $X_1 = (r_{m,t} - r_f)$; $X_2 = SMB_t$; $X_3 = HML_t$; $X_4 = (r_{m,t} - r_f)^2$. Then,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

```
Mkt_RF2 <- Mkt_RF^2
```

```
fit_ibm_ff3_2 <- lm(ibm_x ~ Mkt_RF + SMB + HML + Mkt_RF2)
```

```
summary(fit_ibm_ff3_2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-0.004765	0.002854	-1.670	0.0955	.
Mkt_RF	0.906527	0.057281	15.826	<2e-16	***
SMB	-0.215128	0.084965	-2.532	0.0116	*
HML	-0.173160	0.085054	-2.036	0.0422	*
Mkt_RF2	-0.143191	0.617314	-0.232	0.8167	\Rightarrow Not significant!

15

Functional Form: Linearity in Parameters

Example (continuation): Now, we also check with an LM test if all variables squares $((r_{m,t} - r_f)^2, SMB^2, \text{ and } HML^2)$ are omitted from the 3-factor FF model for IBM returns.

```
Mkt_RF2 <- Mkt_RF^2
```

```
SMB2 <- SMB^2
```

```
HML2 <- HML^2
```

```
fit_ibm_ff3 <- lm(ibm_x ~ Mkt_RF + SMB + HML) # Restricted Model
```

```
resid_r <- fit_ibm_ff3$residuals # Extract residuals from R
```

```
fit_lm <- lm(resid_r ~ Mkt_RF + SMB + HML + Mkt_RF2 + SMB2 + HML2)
```

```
R2_r <- summary(fit_lm)$r.squared
```

```
LM_test <- R2_r * T
```

```
> LM_test
```

```
[1] 2.453822
```

```
p_val <- 1 - pchisq(LM_test, df = 3) # p-value of LM_test
```

```
> p_val
```

```
[1] 0.4836944  $\Rightarrow$  p-value is higher than standard levels  $\Rightarrow$  Cannot Reject  $H_{016}$ 
```


Functional Form: Linearity in Parameters

- Nonlinear in parameters:

$$y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_2 \beta_3 X_4 + \varepsilon$$

This model is nonlinear in parameters since the coefficient of X_4 is the product of the coefficients of X_2 and X_3 .

- Some nonlinearities in parameters can be linearized by appropriate transformations, but not this one. This is not an intrinsic linear model. Different estimation techniques should be used in these cases.

17

Functional Form: Linearity in Parameters

- Intrinsic linear models can be estimated using OLS. Sometimes, transformations are needed. Suppose we start with a power function:

$$y = \beta_1 X^{\beta_2} \varepsilon$$

- The errors enter in multiplicative form. Then, using logs:

$$\log y = \log \beta_1 X^{\beta_2} \varepsilon = \log \beta_1 + \beta_2 \log X + \log \varepsilon,$$

Define:

$$y' = \log y$$

$$X' = \log X$$

$$\beta'_1 = \log \beta_1$$

$$\varepsilon' = \log \varepsilon$$

Then, we have an intrinsic linear model:

$$y' = \beta'_1 + \beta_2 X' + \varepsilon',$$

18

Functional Form: Linearity in Parameters

- Similar intrinsic linear model can be obtained if:

$$\mathbf{y} = e^{\beta_1 + \beta_2 \mathbf{X} + \varepsilon}$$
$$\Rightarrow \log \mathbf{y} = \beta_1 + \beta_2 \mathbf{X} + \varepsilon$$

Define:

$$\mathbf{y}' = \log \mathbf{y}$$

Then, we have an intrinsic linear model:

$$\mathbf{y}' = \beta_1 + \beta_2 \mathbf{X} + \varepsilon$$

19

Functional Form: Linearity in Parameters

- Not all models are intrinsic linear. For example:

$$\mathbf{y} = \beta_1 \mathbf{X}^{\beta_2} + \varepsilon$$
$$\log \mathbf{y} = \log(\beta_1 \mathbf{X}^{\beta_2} + \varepsilon)$$

We cannot linearize the model by taking logarithms. There is no way of simplifying $\log(\beta_1 \mathbf{X}^{\beta_2} + \varepsilon)$.

- We will have to use some nonlinear estimation technique (ML can estimate this model, once we assume a distribution for ε).

20

Functional Form: Ramsey's RESET Test

- To test the specification of the functional form, Ramsey designed a simple test. We start with the fitted values from our (A1) model:

$$\hat{y} = \mathbf{X}\mathbf{b}. \quad (\text{for example, } \hat{y} = b_1X_1 + b_2X_2)$$

Then, we add \hat{y}^2 to the regression specification:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \hat{y}^2 \boldsymbol{\gamma} + \boldsymbol{\varepsilon} \quad (\hat{y}^2 = (b_1X_1)^2 + (b_2X_2)^2 + 2b_1b_2X_2X_1)$$

- If \hat{y}^2 is added to the regression specification, it should pick up quadratic and interactive nonlinearity, if present, without necessarily being highly correlated with any of the \mathbf{X} variables.
- We test H_0 (linear functional form): $\boldsymbol{\gamma} = 0$
 H_1 (non linear functional form): $\boldsymbol{\gamma} \neq 0$

21

Functional Form: Ramsey's RESET Test

- We test H_0 (linear functional form): $\boldsymbol{\gamma} = 0$
 H_1 (non linear functional form): $\boldsymbol{\gamma} \neq 0$
 $\Rightarrow t\text{-test on the OLS estimator of } \boldsymbol{\gamma}.$
- If the *t-statistic* for \hat{y}^2 is significant \Rightarrow evidence of nonlinearity.
- The RESET test is intended to detect nonlinearity, but not be specific about the most appropriate nonlinear model (no specific functional form is specified in H_1).

James B. Ramsey, England



Functional Form: Ramsey's RESET Test

Example: We want to test the functional form of the 3 FF Factor Model for IBM returns, using monthly data 1973-2020.

```
fit_ibm_ff3 <- lm(ibm_x ~ Mkt_RF + SMB + HML)
y_hat <- fitted(fit_ibm_ff3)
y_hat2 <- y_hat^2
fit_ramsey <- lm(ibm_x ~ Mkt_RF + SMB + HML + y_hat2)
> summary(fit_ramsey)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-0.004547	0.002871	-1.584	0.1137	
Mkt_RF	0.903783	0.058003	15.582	<2e-16 ***	
SMB	-0.217268	0.085128	-2.552	0.0110 *	
HML	-0.173276	0.084875	-2.042	0.0417 *	
y_hat2	-0.289197	0.763526	-0.379	0.7050	⇒ Not significant!

23

Functional Form: Ramsey's RESET Test

Example (continuation): Using R package, *lmtest*. (Install it first and, then call the library).

Note: The test reported is an F -test $\sim F_{1,T-k}$, which is equal to $(t_{T-k})^2$. The p -values should be the same.

```
library(lmtest)
> resettest(fit_ibm_ff3, power=2, type="fitted")
```

RESET test

data: y ~ Mkt_RF + SMB + HML

RESET = **0.14346**, df1 = 1, df2 = 564, p-value = **0.705** ⇒ cannot reject H_0 . Check: $(-0.379)^2 = 0.1434$

Conclusion: Given the result of the RESET test, we do not need to respecify the 3-factor FF model with quadratic and interactive terms.

24

Qualitative Variables and Functional Form

- We want to model CEO compensation as a function of education. We have data on annual total CEO compensation (*Comp*), annual returns, annual sales, CEO's age, and CEO's **last degree** (education). We have qualitative data.
- We can estimate CEO compensation regressions for each last degree – i.e., BA/BS; MS/MA/MBA; Doctoral. We have three regressions:

$$\text{Undergrad degree} \quad \text{Comp}_i = \beta_{0-u} + \beta_{1-u}' \mathbf{z}_i + \varepsilon_{u,i}$$

$$\text{Masters degree} \quad \text{Comp}_i = \beta_{0-m} + \beta_{1-m}' \mathbf{z}_i + \varepsilon_{m,i}$$

$$\text{Doctoral degree} \quad \text{Comp}_i = \beta_{0-d} + \beta_{1-d}' \mathbf{z}_i + \varepsilon_{d,i}$$

where the \mathbf{z}_i is a vector of the CEO i 's age and previous experience, and his/her firm's *annual* returns and annual sales.

Potential problem: We have 3 small samples –i.e, lose power & precision

Qualitative Variables and Functional Form

- Alternatively, we can combine the 3 regressions in one, using the whole sample. We use a *dummy variable (indicator variable)* that points whether an observation belongs to a category or class or not. For example:

$$D_{C,i} = \begin{cases} 1 & \text{if observation } i \text{ belongs to category C (say, male.)} \\ 0 & \text{otherwise.} \end{cases}$$

- For CEO's education, we define two dummy variables:

$$D_{m,i} = \begin{cases} 1 & \text{if CEO } i \text{'s has at least a Masters degree} \\ 0 & \text{otherwise.} \end{cases}$$

$$D_{d,i} = \begin{cases} 1 & \text{if CEO } i \text{'s has a Doctoral degree} \\ 0 & \text{otherwise.} \end{cases}$$

Then, we introduce the dummy/indicator variables in the model:

$$\text{Comp}_i = \beta_0 + \beta_1' \mathbf{z}_i + \beta_2 D_{m,i} + \beta_3 D_{d,i} + \gamma_1' \mathbf{z}_i D_{m,i} + \gamma_2' \mathbf{z}_i D_{d,i} + \varepsilon_{i_{26}}$$

Qualitative Variables and Functional Form

Our CEO Compensation model becomes:

$$Comp_i = \beta_0 + \beta_1'z_i + \beta_2 D_{m,i} + \beta_3 D_{d,i} + \gamma_1'z_i D_{m,i} + \gamma_2'z_i D_{d,i} + \varepsilon_i$$

- This model uses all the sample to estimate the parameters. It is flexible:

- Model for undergrads only ($D_{m,i} = 0$ & $D_{d,i} = 0$):

$$Comp_i = \beta_0 + \beta_1'z_i + \varepsilon_i$$

- Model for Masters degree only ($D_{m,i} = 1$ & $D_{d,i} = 0$):

$$Comp_i = (\beta_0 + \beta_2) + (\beta_1 + \gamma_1)'z_i + \varepsilon_i$$

- Model for Doctoral degree only ($D_{m,i} = 1$ & $D_{d,i} = 1$):

$$Comp_i = (\beta_0 + \beta_2 + \beta_3) + (\beta_1 + \gamma_1 + \gamma_2)'z_i + \varepsilon_i$$

27

Qualitative Variables and Functional Form

- Three models, encompassed by one regression:

$$Comp_i = \beta_0 + \beta_1'z_i + \varepsilon_i \quad \text{Undergrad degree}$$

$$Comp_i = (\beta_0 + \beta_2) + (\beta_1 + \gamma_1)'z_i + \varepsilon_i \quad \text{Masters degree}$$

$$Comp_i = (\beta_0 + \beta_2 + \beta_3) + (\beta_1 + \gamma_1 + \gamma_2)'z_i + \varepsilon_i \quad \text{Doctoral degree}$$

- The parameters for the different categories are:

- Constant:

Constant for undergrad degree: β_0

Constant for Masters degree: $\beta_0 + \beta_2$

Constant for Doctoral degree: $\beta_0 + \beta_2 + \beta_3$

- Slopes:

Slopes for Masters degree: $\beta_1 + \gamma_1$

Slopes for Doctoral degree: $\beta_1 + \gamma_1 + \gamma_2$

28

Qualitative Variables and Functional Form

- We can test the effect of education on CEO compensation:
 - (1) H_0 : No effect of grad degree: $\beta_3 = \beta_2 = 0$ & $\gamma_1 = \gamma_2 = \mathbf{0} \Rightarrow F\text{-test.}$
 - (2) H_0 : No effect of Masters degree on constant: $\beta_2 = 0 \Rightarrow t\text{-test.}$
 - (3) H_0 : No effect of doctoral degree: $\beta_3 = 0$ & $\gamma_2 = \mathbf{0} \Rightarrow F\text{-test.}$
 - (4) H_0 : No effect of Dr degree on marginal effect: $\gamma_2 = \mathbf{0} \Rightarrow t\text{-test.}$
- We may have more than one qualitative category (last degree above) in our data that we may want to introduce in our model.

Example: Suppose we also have data for CEO graduate school. Now, we can create another qualitative category, “quality of school”, defined as Top 20 school, to test if a Top 20 school provides “more value.” To do this, we use D_{TOP} to define if any schooling is in the Top 20.

$$D_{TOP,i} = \begin{cases} 1 & \text{if CEO } i\text{'s school is a Top 20 school} \\ 0 & \text{otherwise.} \end{cases}$$

29

Qualitative Variables and Functional Form

Example (continuation):

The model becomes:

$$Comp_i = \beta_0 + \beta_1'z_i + \beta_2 D_{m,i} + \beta_3 D_{d,i} + \beta_4 D_{TOP,i} + \gamma_1'z_i D_{m,i} + \gamma_2'z_i D_{d,i} + \gamma_3'z_i D_{TOP,i} + \varepsilon_i$$

In this setting, we can test the effect of a Top20 education on CEO compensation:

- (1) H_0 : No effect of Top20 degree: $\beta_4 = 0$ and $\gamma_3 = \mathbf{0} \Rightarrow F\text{-test.}$

- The omitted category is the *reference* or *control category*. In our first example, with only educational degrees, the reference category is undergraduate degree. In the second example, with educational degrees and quality of school (Top20 dummy), the reference category is undergraduate degree with no Top 20 education.

30

Qualitative Variables and Functional Form

- *Dummy trap.*

If there is a constant, the numbers of dummy variables per qualitative variable should be equal to the number of categories minus 1. If you put the number of dummies variables equals the number of categories, you will create perfect multicollinearity –i.e., you fell on the **dummy trap**.

31

Dummy Variables as Seasonal Factors

- A popular use of dummy variables is in estimating seasonal effects. We may be interested in studying the January effect in stock returns or if the returns of oil companies (say, Exxon or BP) are affected by the seasons, since in the winter people drive less and in the summer more.

In this case, we define dummy/indicator variables for Summer, Fall and Winter (the base case is, thus, Spring):

$$\begin{aligned}
 D_{Sum,i} &= 1 && \text{if observation } i \text{ occurs in Summer} \\
 &= 0 && \text{otherwise.} \\
 D_{Fall,i} &= 1 && \text{if observation } i \text{ occurs in Fall} \\
 &= 0 && \text{otherwise.} \\
 D_{Win,i} &= 1 && \text{if observation } i \text{ occurs in Winter} \\
 &= 0 && \text{otherwise.}
 \end{aligned}$$

Then, letting \mathbf{Z} be the vector of the three FF factors, we have:

$$(r_i - r_f) = \beta_0 + \beta_1' \mathbf{z}_i + \beta_2 D_{Sum,i} + \beta_3 D_{Fall,i} + \beta_4 D_{Win,i} + \varepsilon_i \quad 32$$

Dummy Variables as Seasonal Factors

Example: In the context of the 3-factor FF model, we test if Exxon's excess returns (XOM) are affected by seasonal (quarters) factors:

$$(r_{XOM,i} - r_f) = \beta_0 + \beta_1'z_i + \beta_2 D_{Sum,i} + \beta_3 D_{Fall,i} + \beta_4 D_{Win,i} + \varepsilon_i$$

```
x_xom <- SFX_da$XOM # Extract XOM prices
T <- length(x_xom)
lr_xom <- log(x_xom[-1]/x_xom[-T])
xom_x <- lr_xom - RF

T <- length(xom_x)
Summ <- rep(c(0,0,0,0,0,0,1,1,1,0,0,0), round(T/12)) # Create Summer dummy
Fall <- rep(c(0,0,0,0,0,0,0,0,0,1,1,1), round(T/12)) # Create Fall dummy
Wint <- rep(c(1,1,1,0,0,0,0,0,0,0,0,0), round(T/12)) # Create Winter dummy
T1 <- T+1
Fall_1 <- Fall[2:T1] # Adjust sample (starts in Feb)
Wint_1 <- Wint[2:T1]
Summ_1 <- Summ[2:T1]

fit_xom_s <- lm(xom_x ~ Mkt_RF + SMB + HML + Fall_1 + Wint_1 + Summ_1)
```

33

Dummy Variables as Seasonal Factors

Example (continuation):

```
fit_xom_s <- lm(xom_x ~ Mkt_RF + SMB + HML + Fall_1 + Wint_1 + Summ_1)
> summary(fit_xom_s)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.002445	0.003485	0.702	0.4832	⇒ constant for reference category (Spring) ≈ 0.
Mkt_RF	0.761816	0.040602	18.763	< 2e-16 ***	
SMB	-0.261925	0.060575	-4.324	1.81e-05 ***	
HML	0.370623	0.060049	6.172	1.29e-09 ***	
Fall_1	-0.006609	0.004947	-1.336	0.1822	
Wint_1	-0.011283	0.004928	-2.290	0.0224 *	⇒ significant. Reject H ₀ : No Winter effect.
Summ_1	-0.007100	0.004944	-1.436	0.1515	

Interpretation: In the Winter quarter, Exxon excess returns decrease, relative to the Spring, by **1.13%**. But since Spring's (& Fall's & Winter's) effect is non-significant, the decrease is in absolute terms.

34

Dummy Variables as Seasonal Factors

Example (continuation): We can test if all quarters *jointly* matter. That is, $H_0: \beta_2 = \beta_3 = \beta_4 = 0$.

We do an F-test:

```
fit_u <- lm(xom_x ~ Mkt_RF + SMB + HML + Fall_1 + Wint_1 + Summ_1)
fit_r <- lm(xom_x ~ Mkt_RF + SMB + HML)
resid_u <- fit_u$residuals
RSS_u <- sum((resid_u)^2)
resid_r <- fit_r$residuals
RSS_r <- sum((resid_r)^2)
f_test <- ((RSS_r - RSS_u)/2)/(RSS_u/(T-4))
> f_test
[1] 2.706574
>
p_val <- 1 - pf(f_test,df1=3, df2=T-3)          # p-value of F-test
> p_val
[1] 0.05504357
```

Conclusion: p-value is “marginal.” At 5% level, cannot reject H_0 : No joint seas effect. 35

Dummy Variables as Seasonal Factors

Example (continuation): Now, we are also interested in checking if the slopes –i.e., *marginal effects*– are affected by the Winter quarter. We fit:

$$(r_{XOM,i} - r_f) = \beta_0 + \beta_1'z_i + \beta_2 D_{Sum,i} + \beta_3 D_{Fall,i} + \beta_4 D_{Win,i} + \gamma_1'z_i D_{Win,i} + \varepsilon_i$$

```
Mkt_W <- Mkt_RF*Wint_1
SMB_W <- SMB*Wint_1
HML_W <- HML*Wint_1
fit_xom_s2 <- lm(xom_x ~ Mkt_RF + SMB + HML + Fall_1 + Wint_1 + Summ_1 + Mkt_W + SMB_W + HML_W)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.003127	0.003478	0.899	0.368962	
Mkt_RF	0.695762	0.048202	14.434	< 2e-16 ***	
SMB	-0.291199	0.075197	-3.872	0.000120 ***	
HML	0.270262	0.077416	3.491	0.000519 ***	
Mkt_W	0.208912	0.091972	2.271	0.023497 *	⇒ significant effect on Mkt slope
SMB_W	0.064753	0.126138	0.513	0.607911	
HML_W	0.198753	0.124261	1.599	0.110278	
Fall_1	-0.006795	0.004934	-1.377	0.169038	
Wint_1	-0.013747	0.005000	-2.750	0.006159 **	⇒ significant effect on constant.
Summ_1	-0.007492	0.004928	-1.520	0.129012	

36

Dummy Variables as Seasonal Factors

Example (continuation):

Interpretation: The only factor interacting significantly with Winter is the Market factor. Then, we have two significantly different slopes:

In the Winter, the Market slope is: $0.695762 + 0.208912 = 0.903674$

In all other quarters, the Market is: 0.695762

It looks like in the Winter, XOM behaves closer to the Market, while in all other quarters, it is significantly less risky than the market.

- Again, a joint interacting Winter effect is not significant:

```
> f_test
```

```
[1] 3.921696
```

```
p_val <- 1 - pf(f_test, df1=3, df2=T-7)
```

```
# p-value of F-test
```

```
> p_val
```

```
[1] 0.0007923967
```

```
⇒ p-value < .05, then, we reject H0 (joint Winter interactive effect): γ1 = 0.
```

37

Dummy Variables: Is There a January Effect?

Example: We want to test the January effect on IBM stock returns, where because of tax reasons/window dressing, stocks go down in December and recover in January. The test can be done by adding a dummy variable to the 3-factor FF model:

$$D_{J,t} = \begin{cases} 1 & \text{if observation } t \text{ occurs in January} \\ 0 & \text{otherwise.} \end{cases}$$

Then, we estimate the expanded model:

$$(r_{i,t} - r_f) = \beta_0 + \beta_1 (r_{m,t} - r_f) + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 D_{J,t} + \varepsilon_{i,t}$$

We test H₀(No January effect): $\beta_4 = 0 \Rightarrow t\text{-test}$.

Alternatively, we can estimate do an LM test on the residuals of the 3-factor FF model and check if $D_{J,t}$ is significant.

```
T <- length(ibm_x)
```

```
Jan <- rep(c(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), (round(T)/12+1))
```

```
# Create January dummy
```

```
T2 <- T+1
```

38

Dummy Variables: Is There a January Effect?

Example (continuation):

```
Jan_1 <- Jan[2:T2]                                # Adjust sample
fit_ibm_ff <- lm (ibm_x ~ Mkt_RF + SMB + HML)        # Restricted Regression
resid_r <- fit_ibm_ff$residuals                     # Keep residuals (eR)
fit_Jan <- lm (resid_r ~ Mkt_RF + SMB + HML + Jan_1) # Auxiliary Regression
> summary(fit_Jan)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.002111	0.002561	-0.824	0.41027
Mkt_RF	-0.005198	0.056405	-0.092	0.92661
SMB	-0.026306	0.084063	-0.313	0.75445
HML	-0.014914	0.083606	-0.178	0.85848
Jan_1	0.026966	0.008906	3.028	0.00258 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.058 on 565 degrees of freedom
Multiple R-squared: **0.01597**, Adjusted R-squared: 0.009
F-statistic: 2.292 on 4 and 565 DF, p-value: 0.05841

39

Dummy Variables: Is There a January Effect?

Example (continuation):

```
R2_r <- summary(fit_Jan)$r.squared                # Keep R^2 from Auxiliary Regression
> R2_r
[1] 0.01596528

LM_test <- R2_r * T
> LM_test
[1] 9.084247

p_val <- 1 - pchisq(LM_test, df = 1)              # p-value of LM_test
> p_val
[1] 0.002578207                                   ⇒ p-value is small ⇒ Reject H0.
```

Given this result, we modify the 3-factor FF and add the January Dummy to the FF model:

```
fit_ibm_new <- lm (ibm_x ~ Mkt_RF + SMB + HML + Jan_1)
summary(fit_ibm_new)
```

40

Dummy Variables: Is There a January Effect?

Example (continuation):

```
> summary(fit_ibm_new)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.007302	0.002561	-2.851	0.00452 **
Mkt_RF	0.905182	0.056405	16.048	< 2e-16 ***
SMB	-0.247691	0.084063	-2.946	0.00335 **
HML	-0.154093	0.083606	-1.843	0.06584 .
Jan_1	0.026966	0.008906	3.028	0.00258 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.058 on 565 degrees of freedom

Multiple R-squared: 0.3499, Adjusted R-squared: 0.3453

F-statistic: 76.01 on 4 and 565 DF, p-value: < 2.2e-16

Interpretation: We have two constants (excess return, Jensen's alpha):

Feb - Dec: -0.7302% (significant).

January: -0.7302% + 2.6966% = 1.9664% (significant).

41

Dummy Variables: Is There a January Effect?

Example (continuation):

Interpretation: We have two constants (excess return, Jensen's alpha):

Feb - Dec: -0.7302% (significant).

January: -0.7302% + 2.6966% = 1.9664% (significant).

When the January dummy was not in the model, we had: -0.005191, which is close to an average of the constants ($= -0.007302 * 11 + 0.019664 / 12 = -0.00505$).

Interpretation: During January IBM has an additional 2.6966% excess returns. This is a big number. Today, the evidence for the January effect is much weaker than in this case.

- Note that in the FF model we expect the constant to be very small (≈ 0). In this case, it is not zero. Maybe we have a misspecified (A1).

42

Dummy Variable for One Observation

- We can use a dummy variable to isolate a single observation.

$$D_j = \begin{cases} 1 & \text{for observation } j. \\ 0 & \text{otherwise.} \end{cases}$$

- Define \mathbf{d} to be the dummy variable in question.

$$\mathbf{Z} = \text{all other regressors. } \mathbf{X} = [\mathbf{Z}, \mathbf{D}_j]$$

- Multiple regression of \mathbf{y} on \mathbf{X} . We know that

$$\begin{aligned} \mathbf{X}'\mathbf{e} &= \mathbf{0} & \text{where } \mathbf{e} &= \text{the column vector of residuals.} \\ \Rightarrow D_j'\mathbf{e} &= 0 & \Rightarrow e_j &= 0 \text{ (perfect fit for observation } j\text{).} \end{aligned}$$

- This approach can be used to deal with (eliminate) *outliers*.

43

Dummy Variable for One Observation

Example: In Dec 1992, IBM reported record losses and gave a very bleak picture of its future. The stock tumbled **-30.64%** that month. We check the effect of that extreme observation, a potential outlier, on the 3-factor FF model + January dummy:

```
dec_1992 <- rep(0,T) # Define Dec 1992 dummy
dec_1992[239] <- 1 # Define Dec 1992 dummy (=1 if Dec 1992)
fit_d92 <- lm(ibm_x ~ Mkt_RF + SMB + HML + Jan_1 + dec_1992)
> summary(fit_d92)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-0.006772	0.002502	-2.707	0.00699	**
Mkt_RF	0.908775	0.055054	16.507	< 2e-16	***
SMB	-0.239213	0.082059	-2.915	0.00370	**
HML	-0.138629	0.081647	-1.698	0.09008	.
Jan_1	0.026163	0.008694	3.009	0.00273	**
dec_1992	-0.306202	0.056710	-5.399	9.86e-08	***

(same value of observation)

Note: Potential “Outlier” has no major effect on coefficients.

44

Functional Form: Structural Change (Again)

- We want to test if an event at that time T_{SB} affected our model, creating a “before” and an “after” in the parameters: That is,

$$\begin{aligned} y_i &= \beta_0^1 + \beta_1^1 X_{1,i} + \beta_2^1 X_{2,i} + \beta_3^1 X_{3,i} + \varepsilon_i & \text{for } i \leq T_{SB} \\ y_i &= \beta_0^2 + \beta_1^2 X_{1,i} + \beta_2^2 X_{2,i} + \beta_3^2 X_{3,i} + \varepsilon_i & \text{for } i > T_{SB} \end{aligned}$$

The event caused *structural change* in the model.

- A Chow test, an F-test, tests if one model applies to both regimes:

$$y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i \quad \text{for all } i$$

- We test H_0 (No *structural change*): $\beta_0^1 = \beta_0^2 = \beta_0$
 $\beta_1^1 = \beta_1^2 = \beta_1$
 $\beta_2^1 = \beta_2^2 = \beta_2$
 $\beta_3^1 = \beta_3^2 = \beta_3$

H_1 (*structural change*): For at least one k ($= 0, 1, 2, 3$): $\beta_k^1 \neq \beta_k^2$

Functional Form: Structural Change

- We structure the Chow test to test H_0 (No *structural change*), as usual.

- Steps for Chow (Structural Change) Test:

(1) Run OLS with all the data, with no distinction between regimes. (Restricted or pooled model). Keep RSS_R .

(2) Run two separate OLS, one for each regime (Unrestricted model):

Before Date T_{SB} . Keep RSS_1 .

After Date T_{SB} . Keep RSS_2 . $\Rightarrow RSS_U = RSS_1 + RSS_2$.

(3) Run a standard F-test (testing Restricted vs. Unrestricted models):

$$F = \frac{(RSS_R - RSS_U)/(k_U - k_R)}{(RSS_U)/(T - k_U)} = \frac{(RSS_R - [RSS_1 + RSS_2])/k}{(RSS_1 + RSS_2)/(T - 2k)}$$

46

Functional Form: Structural Change

- Before, when we presented the Chow test, we use the F-distribution, which will be appropriate under **(A5)**.
- In general, we rely on the asymptotic distribution –i.e., we do not rely on **(A5)**. Then, under H_0 , (& if the number of observations pre- and post-break are large), then

$$J * F \xrightarrow{d} \chi_J^2 \quad (\text{sometimes written as } F \xrightarrow{d} \chi_J^2/J).$$

- Note that it is also possible to do a Wald test to test H_0 :

47

Functional Form: Structural Change

Example: 3 Factor Fama-French Model for SLB

Q: Did the financial crisis (Sep 2008, $T_{SB} = 429$) affect the structure of the FF Model? Sample: January 1973 – December 2023 ($T = 611$).

Pooled RSS = **3.5290**

Jan 1973 – Sep 2008 RSS = $RSS_1 = 2.0010$ ($T = 428$)

Oct 2008 – Dec 2023 RSS = $RSS_2 = 1.1213$ ($T = 183$)

$$F = \frac{[RSS_R - (RSS_1 + RSS_2)]/J}{(RSS_1 + RSS_2)/(T - k)} = \frac{[3.5290 - (2.0010 + 1.1213)]/4}{(2.0010 + 1.1213)/(611 - 2*4)} = 19.6356$$

\Rightarrow Since $F_{4,611,05} = 2.39$, we reject H_0

	Constant	Mkt – rf	SMB	HML	RSS	T
1973-2020	-0.0073*	1.2138*	0.0123	0.4182*	3.5290	611
1973-2001	0.0013	0.9038*	-0.2394*	-0.3477*	2.0010	428
2002 – 2023	-0.0141*	1.3129*	0.3703	1.1496*	1.1213	183

48

Functional Form: Structural Change

Example (continuation): The R package *sctrucchange* estimates the Chow test. (As usual, you need to install package first.)

```
>x_slb <- SFX_da$SLB
>lr_slb <- log(x_slb[-1]/x_slb[-T])
>slb_x <- lr_slb - RF
>library(sctrucchange)
> t_s <- 428
> sctest(slb_x ~ Mkt_RF + SMB + HML, type = "Chow", point = t_s)
```

Chow test

data: slb_x ~ Mkt_RF + SMB + HML

F = **19.636**, p-value = **3.331e-15**

49

Functional Form: Structural Change

Example: We test if the Oct 1973 oil shock in quarterly GDP growth rates had an structural change on the GDP growth rate model.

We model the GDP growth rate with an **AR(1) model**, that is, GDP growth rate depends only on its own lagged growth rate:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$$

```
GDP_da <- read.csv("http://www.bauer.uh.edu/rsusmel/4397/GDP_q.csv", head=TRUE,
sep=",")
x_date <- GDP_da$DATE
x_gdp <- GDP_da$GDP
x_dummy <- GDP_da$D73
T <- length(x_gdp)
t_s <- 108                                # TSB = Oct 1973

lr_gdp <- log(x_gdp[-1]/x_gdp[-T])
T <- length(lr_gdp)
lr_gdp0 <- lr_gdp[-1]
lr_gdp1 <- lr_gdp[-T]
t_s <- t_s -1                             # Adjust t_s (we lost the first observation)
```

50

Functional Form: Structural Change

Example (continuation):

```

y <- lr_gdp0
x1 <- lr_gdp1
T <- length(y)
x0 <- matrix(1,T,1)
x <- cbind(x0,x1)
k <- ncol(x)

# Restricted Model (Pooling all data)
fit_ar1 <- lm(lr_gdp0 ~ lr_gdp1)
e_R <- fit_ar1$residuals
RSS_R <- sum(e_R^2)

# Fitting AR(1) (Restricted) Model
# regression residuals, e
# RSS Restricted

> summary(fit_ar1)

Coefficients:
              Estimate Std. Error t value Pr(> |t|)
(Intercept)  0.011406   0.001118   10.200  < 2e-16 ***
lr_gdp1      0.262234   0.055543    4.721  3.59e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01248 on 302 degrees of freedom

```

51

Functional Form: Structural Change

Example (continuation):

```

# Unrestricted Model (Two regimes)

y_1 <- y[1:t_s]
x_u1 <- x[1:t_s]
fit_ar1_1 <- lm(y_1 ~ x_u1 - 1)
e1 <- fit_ar1_1$residuals
RSS1 <- sum(e1^2)

# AR(1) Regime 1
# Regime 1 regression residuals, e
# RSS Regime 1

kk = t_s+1
# Starting date for Regime 2

y_2 <- y[kk:T]
x_u2 <- x[kk:T]
fit_ar1_2 <- lm(y_2 ~ x_u2 - 1)
e2 <- fit_ar1_2$residuals
RSS2 <- sum(e2^2)

# AR(1) Regime 2
# Regime 2 regression residuals, e
# RSS Regime 2

F <- ((RSS_R - (RSS1+RSS2))/k)/((RSS1+RSS2)/(T - 2*k))
> F
[1] 4.391997
p_val <- 1 - pf(F, df1 = 2, df2 = T - 2*k) # p-value of F_test
> p_val
[1] 0.0131817

```

⇒ small p-values: Reject H_0 (No structural change).

52

Structural Change: Specification with Dummies

- Under the H_0 (No *structural change*), we pool the data into one model. That is, the parameters are the same under both regimes. We fit the same model for all t , for example:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$$

- If the Chow test rejects H_0 , we need to reformulate the model. A typical reformulation includes a dummy variable ($D_{SB,t}$). For example, with vector \mathbf{x}_t of explanatory variables:

$$y_t = \beta_0 + \beta_1' \mathbf{x}_t + \beta_2 D_{SB,t} + \gamma_1' \mathbf{x}_t D_{SB,t} + \varepsilon_t$$

where

$$\begin{aligned} D_{SB,t} &= 1 && \text{if observation } t \text{ occurred after } T_{SB} \\ &= 0 && \text{otherwise.} \end{aligned}$$

53

Structural Change: Specification with Dummies

Example: We are interested in modelling the effect of the Oct 1973 oil shock in GDP growth rates. We include a dummy variable in the AR(1) model, say D_{73} :

$$\begin{aligned} D_{73,t} &= 1 \text{ if observation } t \text{ occurred after October 1973} \\ &= 0 \text{ otherwise.} \end{aligned}$$

Then,
$$y_t = \beta_0 + \beta_1' \mathbf{x}_t + \beta_2 D_{73,t} + \gamma_1' \mathbf{x}_t D_{73,t} + \varepsilon_t$$

In the model, the oil shock affects the constant and the slopes.

	Constant	Slopes:
Before oil shock ($D_{73} = 0$):	β_0	β_1
After oil shock ($D_{73} = 1$):	$\beta_0 + \beta_2$	$\beta_1 + \gamma_1$

- We estimate the above model and perform an F-test to test if H_0 (No *structural change*): $\beta_2 = 0$ & $\gamma_1 = 0$.

54

Structural Change: Specification with Dummies

Example: We add an Oct 1973 dummy in the **AR(1) GDP model**.

```
T1 <- T - t_s                                # Number of Observations after SB
D73_0 <- rep(0,t_s)                          # Dummy_t = 0 if t <= t_s
D73_1 <- rep(1,T1)                          # Dummy_t = 1 of t > t_s
D73 <- c(D73_0,D73_1)                      # SB Dummy variable t_s <- 108
lr_gdp1_D73 <- lr_gdp1 * D73                # interactive dummy (effect on slope)
fit_ar1_d_2 <- lm(lr_gdp0 ~ lr_gdp1 + D73 + lr_gdp1_D73)
summary(fit_ar1_d_2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.009139	0.001939	4.712	3.75e-06 ***	
lr_gdp1	0.457011	0.090716	5.038	8.15e-07 ***	
D73	0.003499	0.002362	1.482	0.13947	⇒ no significant effect on constant
lr_gdp1_D73	-0.316005	0.114197	-2.767	0.00601 **	⇒ significant effect of oil shock on slope.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Conclusion: After the oil shock the slope significantly changed from **0.457011** to **0.141006** (= **0.457011** + **(-0.316005)**).

55

Structural Change: Wald Test

- It is also possible to do a Wald test to test H_0 , using only the unrestricted estimators. Steps:

1) Run two separate OLS, one for each regime (Unrestricted model):

Before Date T_{SB} : Keep \mathbf{b}_1 & $\text{Var}[\mathbf{b}_1]$

After Date T_{SB} : Keep \mathbf{b}_2 & $\text{Var}[\mathbf{b}_2]$

2) Compute the Wald test:

$$W = (\mathbf{b}_1 - \mathbf{b}_2)' \{\text{Var}[\mathbf{b}_1 - \mathbf{b}_2]\}^{-1} (\mathbf{b}_1 - \mathbf{b}_2)$$

where $\text{Var}[\mathbf{b}_1 - \mathbf{b}_2]$ is computed using $\text{Var}[\mathbf{b}_1]$ and $\text{Var}[\mathbf{b}_2]$.

Under H_0 (& if the number of observations pre- and post-break are large), the Wald test follows: $W \xrightarrow{d} \chi^2_f$

56

Structural Change: Test with Unknown Break

- The previous examples compute the Chow test assuming that we know exactly when the break occurred –say, October 73 or Dec 2001. That is, the results are *conditional* on the assumed breaking point.
- In general, breaking points are unknown, we need to estimate them.
- One quick approach is to do a rolling Chow test –that is we run the Chow test for all dates in the sample– and pick the date that maximizes the F-tests.
- This test was proposed by Quandt (1958):

$$QLR_T = \max_{\tau \in \{\tau_{min}, \dots, \tau_{max}\}} F_T(\tau)$$

The max (supremum) is taken over all potential breaks in (τ_{min}, τ_{max}) . For example, $\tau_{min} = T * .15$; $\tau_{max} = T * .85$; that is we trim 30% of the observations ($\pi_0 = 15\%$ in each side) to run the test.

3

Structural Change: Test with Unknown Break

- It is also possible to run the Wald test version of the Chow test for all possible dates, again, selecting the date that maximizes

$$QLR_T = \max_{\tau \in \{\tau_{min}, \dots, \tau_{max}\}} W_T(\tau)$$

- The first QLR_T is called the **SupF** test, the second the **SupW**.

Technical Problem: With this approach, the technical conditions under which the asymptotic distribution is derived are not met in this setting.

- Andrews (1993) showed that under appropriate conditions, the QLR statistic, also known as Sup-test (F, W, LR) statistic, has a *non-standard limiting distribution* (“non-standard” = no existing table; needs a new one).
- The distribution depends on the number of parameters of the model, k , which are tested for stability, and trimming value, π_0 .

3

Structural Change: Test with Unknown Break

- Andrews (1993) tabulated the non-standard distribution of the **SupW** for different k , α , and trimming values (π_0).

Note: It is usual to test the **SupF**, using the critical values of **SupW**, by dividing the **SupW** critical values by k . In the next slide, Andrews (1993) table. (Andrews (2003) issued a slightly corrected Table.)

For example, for $k = 2$ & 4, (& $\pi_0 = \tau_{min}/T = (1 - \tau_{max}/T) = .15$), using $\alpha = .05$, the SupW critical values are **11.79** & **16.45**, respectively. Then, for the SupF critical values, we get **5.89** ($= 11.79/2$) and **4.11** ($= 16.45/2$), respectively.

3

Structural Change: Test with Unknown Break

Critical values of the QLR test Distribution, taken from Andrews (1993). Note: $p = \#$ of parameters (k), $\pi_0 =$ trimming value. (Ignore λ .)

840

DONALD W. K. ANDREWS

TABLE I
ASYMPTOTIC CRITICAL VALUES

π_0	λ	$p=1$			$p=2$			$p=3$			$p=4$			$p=5$		
		10%	5%	1%	10%	5%	1%	10%	5%	1%	10%	5%	1%	10%	5%	1%
.50	1.00	2.71	3.84	6.63	4.61	5.99	9.21	6.25	7.81	11.34	7.78	9.49	13.28	9.24	11.07	15.09
.49	1.08	3.47	4.73	7.82	5.42	6.86	10.30	7.19	8.83	12.58	8.93	10.63	14.64	10.39	12.28	16.34
.48	1.17	3.79	5.10	8.26	5.80	7.31	10.71	7.64	9.29	13.05	9.42	11.17	15.17	10.96	12.88	16.83
.47	1.27	4.02	5.38	8.65	6.12	7.67	11.01	7.98	9.62	13.39	9.82	11.63	15.91	11.40	13.27	17.32
.45	1.49	4.38	5.91	9.00	6.60	8.11	11.77	8.50	10.15	14.23	10.35	12.27	16.64	12.05	14.00	18.06
.40	2.25	5.10	6.57	9.82	7.45	9.02	12.91	9.46	11.17	14.88	11.39	13.32	17.66	13.09	15.16	19.23
.35	3.45	5.59	7.05	10.53	8.06	9.67	13.53	10.16	12.05	15.71	12.10	14.12	18.54	13.86	15.93	19.99
.30	5.44	6.05	7.51	10.91	8.57	10.19	14.16	10.76	12.58	16.24	12.80	14.79	19.10	14.58	16.48	20.67
.25	9.00	6.46	7.93	11.48	9.10	10.75	14.47	11.29	13.16	16.60	13.36	15.34	19.78	15.17	17.25	21.39
.20	16.00	6.80	8.45	11.69	9.59	11.26	14.69	11.80	13.69	17.28	13.82	15.84	20.24	15.63	17.88	21.90
.15	32.11	7.17	8.85	12.35	10.01	11.79	15.51	12.27	14.15	17.68	14.31	16.45	20.71	16.20	18.35	22.49
.10	81.00	7.63	9.31	12.69	10.50	12.27	16.04	12.81	14.62	18.28	14.94	16.98	21.04	16.87	18.93	23.34
.05	361.00	8.19	9.84	13.01	11.20	12.93	16.44	13.47	15.15	19.06	15.62	17.56	21.54	17.69	19.61	24.18

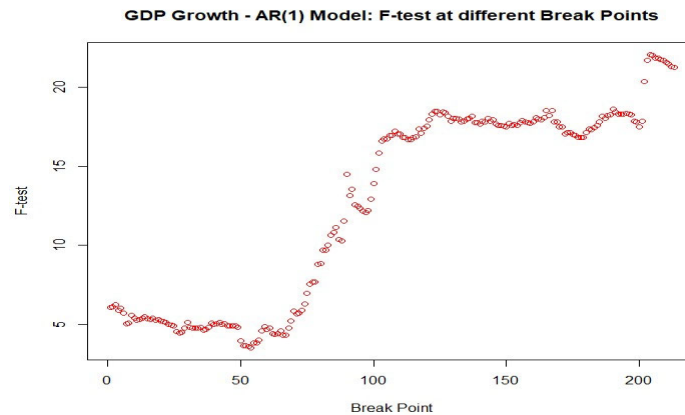
π_0	λ	$p=6$			$p=7$			$p=8$			$p=9$			$p=10$		
		10%	5%	1%	10%	5%	1%	10%	5%	1%	10%	5%	1%	10%	5%	1%
.50	1.00	10.64	12.59	16.81	12.02	14.07	18.48	13.36	15.51	20.09	14.68	16.92	21.67	15.99	18.31	23.21
.49	1.08	11.81	13.74	18.32	13.27	15.52	19.93	13.29	15.63	20.53	16.17	18.56	23.05	17.35	19.79	24.62
.48	1.17	12.42	14.45	19.12	13.92	16.14	20.64	13.89	16.31	21.14	16.82	19.25	23.83	18.08	20.35	25.75
.47	1.27	12.90	14.86	19.64	14.32	16.63	21.14	14.43	16.74	21.72	17.26	19.74	24.80	18.67	20.92	26.43
.45	1.49	13.53	15.59	20.45	14.97	17.38	22.32	15.05	17.53	22.28	18.10	20.59	25.52	19.39	21.78	27.30
.40	2.25	14.71	16.91	21.60	16.23	18.41	23.35	16.26	18.73	23.63	19.56	22.12	26.86	20.74	23.15	28.86
.35	3.45	15.56	17.75	22.33	17.09	19.34	24.10	17.06	19.46	24.64	20.49	22.93	27.77	21.87	24.17	29.76
.30	5.44	16.32	18.46	23.06	17.74	20.01	24.86	17.90	20.36	25.64	21.27	23.65	28.50	22.73	25.05	30.74
.25	9.00	17.00	19.07	23.65	18.38	20.63	25.11	18.61	20.95	26.10	21.93	24.31	29.23	23.32	25.80	31.32
.20	16.00	17.56	19.64	24.27	19.04	21.07	25.72	19.17	21.47	26.76	22.54	24.91	29.92	24.00	26.42	31.98
.15	32.11	18.12	20.26	24.79	19.69	21.84	26.23	19.82	22.13	27.25	23.15	25.47	30.52	24.62	27.03	32.33
.10	81.00	18.78	20.82	25.21	20.32	22.51	26.91	20.45	22.87	27.69	23.77	26.16	31.15	25.39	27.87	32.95
.05	361.00	19.49	21.56	25.96	21.02	23.22	27.53	21.23	23.60	28.77	24.64	26.94	31.61	26.24	28.63	33.86

Critical value
for test for
 $k=2$, $\pi_0 = .15$
and $\alpha = .05$.

Critical value
for test for
 $k=4$, $\pi_0 = .15$
and $\alpha = .05$.

Structural Change: Test with Unknown Break

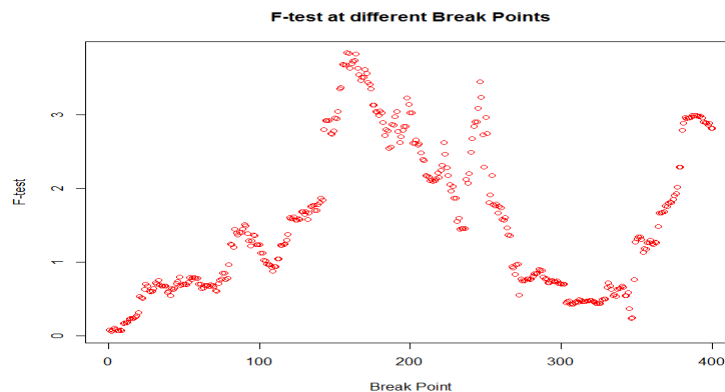
Example (continuation): We search for breaking points for GDP growth rate in AR(1) model. Below, we plot all F-tests starting at T^*15 :



- Maximum F is **22.08** occurs in Jan 2009 (observation #250). Then, $\overline{QLR} = 22.08 > 5.89 \Rightarrow$ Reject H_0 at 5% level & break is not Oct 73!.

Structural Change: Test with Unknown Break

Example: We search for breaking points for IBM returns in the 3-factor FF model. Below, we plot all F-tests starting at T^*15 :



- Maximum F is **3.83** occurs in May 1993 (observation #243). Then, $\overline{QLR} = 3.83 < 4.11 \Rightarrow$ Cannot reject H_0 at 5% level.

Structural Change: Test with Unknown Break (R)

- Chow Test for different breaking points, starting at T1.

```

y <- ibm_x;
x1 <- Mkt_RF
x2 <- SMB
x3 <- HML
T <- length(x1)
x0 <- matrix(1,T,1)
x <- cbind(x0,x1,x2,x3)
k <- ncol(x)
b <- solve(t(x)%*%x)%*%t(x)%*%y          # b = (X'X)-1 X' y (OLS regression)
e <- y - x%*%b                            # regression residuals, e
RSS_R <- as.numeric(t(e)%*%e)              # RSS for Restricted (no structural change)

T1 <- round(T * 1/5)                       # Trim .20 of data
t <- T1                                    # t will be the counter for loop. Starts at T1.
T2 <- round(T * 4/5)                       # Trim .20 of data
T_sam <- T2 - T1
All_F <- matrix(0,T_sam,1)                 # Matrix to accumulate the (T2-T1) F-tests

while (t <= T2) {                          # Start while loop with counter t
  y_1 <- y[1:t]
  x_u1 <- x[1:t]
```

3

Structural Change: Test with Unknown Break (R)

```

b_1 <- solve(t(x_u1)%*%x_u1)%*%t(x_u1)%*%y_1  # b = (X'X)-1 X' y (OLS regression)
e1 <- y_1 - x_u1%*%b_1                        # regression residuals, e
RSS1 <- as.numeric(t(e1)%*%e1)                # RSS for regime 1

kk = t+1
y_2 <- y[kk:T]
x_u2 <- x[kk:T]

b_2 <- solve(t(x_u2)%*%x_u2)%*%t(x_u2)%*%y_2  # b = (X'X)-1 X' y (OLS regression)
e2 <- y_2 - x_u2%*%b_2                        # regression residuals, e
RSS2 <- as.numeric(t(e2)%*%e2)                # RSS for regime 2

F <- ((RSS_R - (RSS1+RSS2))/k)/((RSS1+RSS2)/(T - 2*k))
kt <- t - T1 + 1                             # kt is an index that start at 1
All_F[kt] <- F                               # add F-test to All_F according to kt
t = t+1
}

plot(All_F, col="red",ylab ="F-test", xlab ="Break Point")
title("F-test at different Break Points")
F_max <- max(All_F)                          # Find the maximum F-test (QLR)
```

3

Structural Change Tests: Remarks

- The results are *conditional* on the breaking point –say, **October 73** or **Dec 2001**.
- The breaking point is usually unknown. It needs to be estimated.
- It can deal only with one structural break –i.e., two categories!
- The number of breaks is also unknown. They need to be estimated.
- Characteristics of the data (heteroscedasticity –for example, regimes in the variance- and unit roots (high persistence) complicate the test.
- In general, only asymptotic (consistent) results are available.
- There are many modern tests that take care of these issues, but usually also with *non-standard* distributions.

3