Erasmus School of Economics

MOOC Econometrics

Lecture 3.1 on Model Specification: Motivation

Michel van der Wel

Erasmus University Rotterdam

Introduction

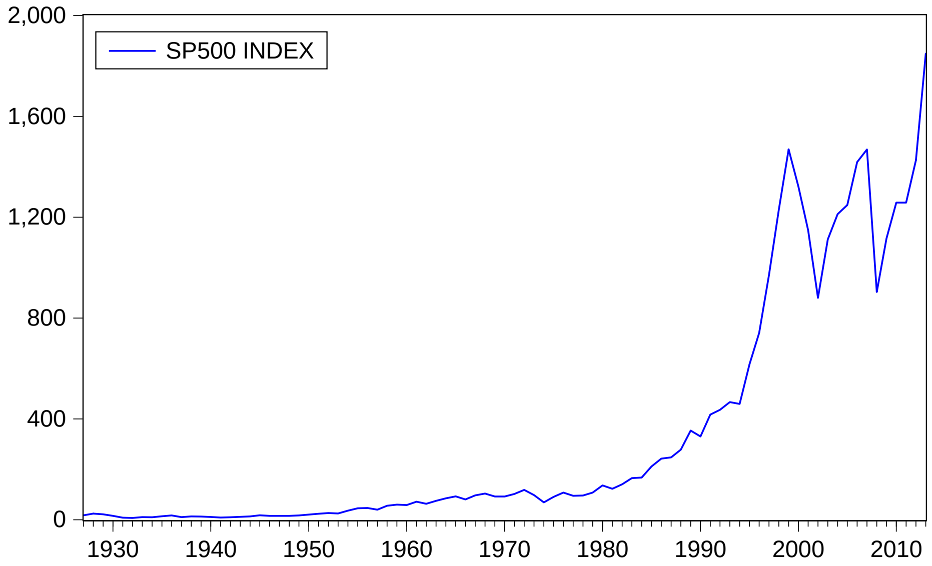
- Question 1: Do we include all explanatory variables or only a few?
- Question 2: Should we transform the variables?
- Question 3: How to evaluate a model?

Test

Suppose that all explanatory variables in a dataset are relevant for the dependent variable. Should we include all?

Answer: Not necessarily. Lecture 3.2 will explain the why and how.

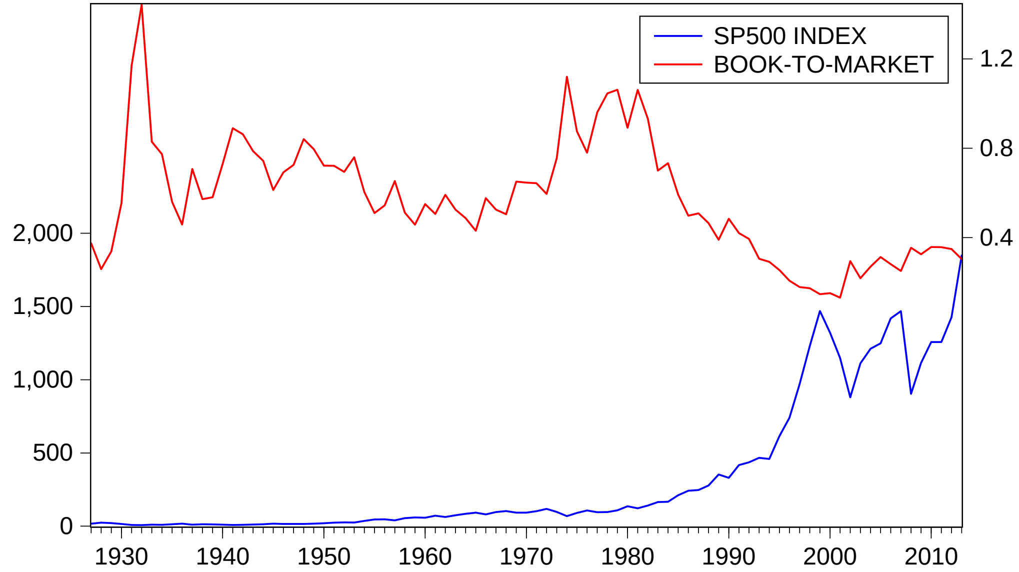
Example: Stock market index



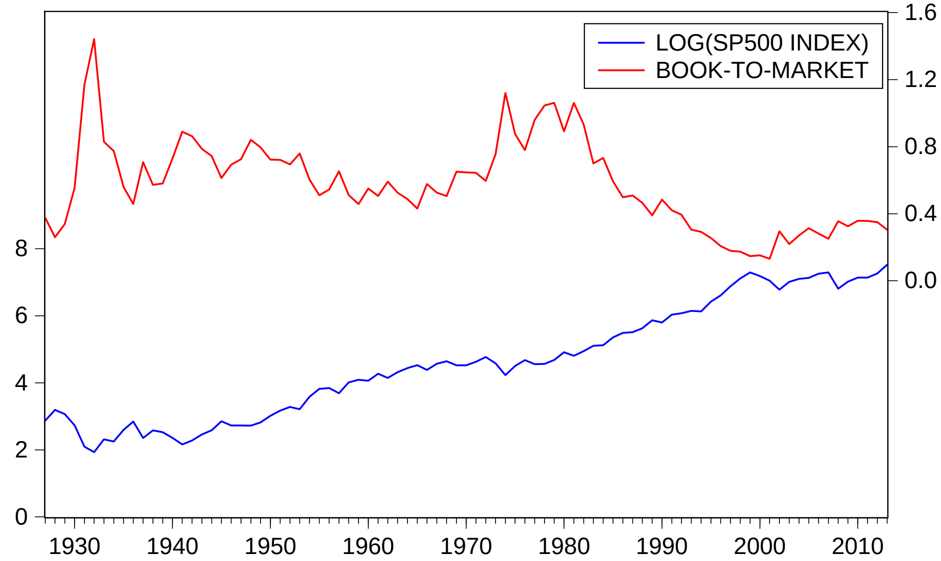
Explanatory variables

- Stock characteristics: Dividends, earnings, volatility, book value, issuing activity
- Interest-rate related: Treasury bill rates, long term yields, corporate bond returns
- Macroeconomic: Inflation, investment, consumption

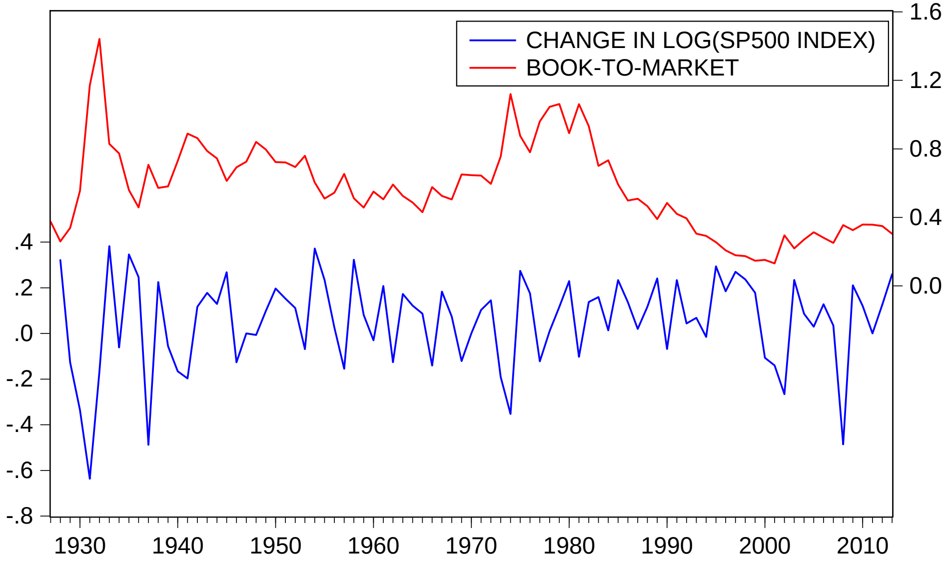
Stock index and book-to-market ratio



Log stock index and book-to-market ratio



Change in log stock index, and book-to-market ratio



Regression output

Dependent variable: change in log(SP500 index)
Sample size: 86

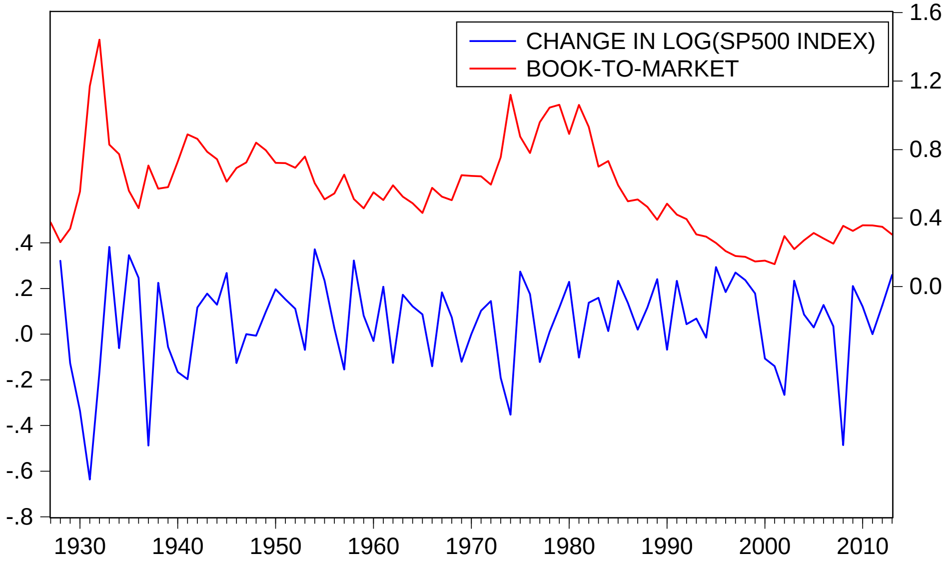
	Coefficient b_j	Standard error $SE(b_j)$	t-Statistic t_j	p-value $H_0 : \beta_j = 0$
Constant	0.177	0.050	3.543	0.001
Book-to-market	-0.213	0.079	-2.702	0.008
R-squared	0.080			
SE of regression	0.191			

Test

What is the interpretation of the negative sign of Book-to-market?

Answer: High book-to-market usually coincides with periods when market-value decreased.

Change in log stock index, and book-to-market ratio



TRAINING EXERCISE 3.1

- Train yourself by making the training exercise (see the website).
- After making this exercise, check your answers by studying the webcast solution (also available on the website).

Erasmus School of Economics

MOOC Econometrics

Lecture 3.2 on Model Specification: Specification

Michel van der Wel

Erasmus University Rotterdam

Bias-efficiency trade-off

Setting:

$$y_i = x_i'\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

or

$$y = X\beta + \varepsilon$$

in matrix form.

Which variables should we include in X?

- Too few variables $\rightarrow$ Bias.
- Too many variables $\rightarrow$ Efficiency loss. (Even if all variables really matter!)

Consequences of omitting variables

DGP: $y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad \rightarrow \quad b_1$ and b_2

Model: $y = X_1\beta_1 + \tilde{\varepsilon} \quad \rightarrow \quad b_R$

Test

Express $E(b_R)$ as function of β_1 and β_2.

Answer:

$$\begin{aligned} E(b_R) &= E((X_1'X_1)^{-1}X_1'y), \\ &= E((X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon)), \\ &= E((X_1'X_1)^{-1}X_1'X_1\beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon)), \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + 0. \end{aligned}$$

Consequences of omitting variables

DGP: $y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad \rightarrow \quad b_1$ and b_2

Model: $y = X_1\beta_1 + \tilde{\varepsilon} \quad \rightarrow \quad b_R$

It holds:

- $E(b_R) = \beta_1 + \underbrace{(X_1'X_1)^{-1}X_1'X_2}_{P}\beta_2 = \beta_1 + P\beta_2$

  $\rightarrow$ Bias if $\beta_2 \neq 0$ (omitted variable bias).

- $Var(b_R) = Var(b_1) - PVar(b_2)P'$

  $\rightarrow$ Variance of b_R is smaller than that of b_1 (even if $\beta_2 = 0$!).

Decision metrics

Possible decision metrics:

- Information criteria
- Out-of-sample prediction

Information criteria

Commonly used information criteria:

- Akaike: $AIC = \log(s^2) + \frac{2k}{n}$
- Bayes: $BIC = \log(s^2) + \frac{k \log n}{n}$

with s the standard error of the regression and k the number of variables.

Test

Which information criterion imposes the strongest penalty on the number of variables?

Answer: Penalty is $2/n$ for AIC and $\log(n)/n$ for BIC; BIC imposes stronger penalty if $\log(n) > 2$, $n \geq 8$.

Out-of-sample prediction

Commonly used out-of-sample prediction metrics:

- $RMSE = \left(\frac{1}{n_f} \sum_{i=1}^{n_f} (y_i - \hat{y}_i)^2\right)^{1/2}$
- $MAE = \frac{1}{n_f} \sum_{i=1}^{n_f} |y_i - \hat{y}_i|$

with n_f the number of observations "saved" for out-of-sample evaluation and $\hat{y}_i$ the i-th predicted value of the dependent variable.

Iterative selection methods

Commonly used methods to select explanatory variables:

- t-test and F-test
- Information criteria
- Out-of-sample predictions

Also iterative methods (based on tests) are commonly used:

- General-to-specific / backward elimination
- Specific-to-general / forward selection

TRAINING EXERCISE 3.2

- Train yourself by making the training exercise (see the website).
- After making this exercise, check your answers by studying the webcast solution (also available on the website).

Erasmus School of Economics

MOOC Econometrics

Lecture 3.3 on Model Specification: Transformation

Michel van der Wel

Erasmus University Rotterdam

Data transformation

Setting:

$$y_i = x_i'\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

or

$$y = X\beta + \varepsilon$$

in matrix form.

What is the most appropriate form of the data (y and X)?

- All variables should be incorporated in a compatible manner.
- If this is not the case, data can be transformed.

Taking logarithms

Use for:

- Exponential growth.

Taking differences

Use for:

- Trending patterns

  $\rightarrow$ Statistical assumptions may not hold.

First difference:

$$\Delta y_i = y_i - y_{i-1}.$$

Test

What is the result if you take the difference of $y_i = i$?

Answer: For $y_i = i$ the difference is:

$$\Delta y_i = y_i - y_{i-1} = i - (i-1) = 1.$$

Non-linear effects

$$y_i = x_i'\beta + \varepsilon_i = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

has linear set-up with fixed marginal effects ($dy_i/dx_{ji} = \beta_j$).

Extension with interaction and quadratic terms:

$$y_i = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \sum_{j=2}^{k} \gamma_{jj} x_{ji}^2 + \sum_{j=2}^{k} \sum_{h=j+1}^{k} \gamma_{jh} x_{ji} x_{hi} + \varepsilon_i.$$

Advantages:

- Get non-linear functional form.
- May provide economically meaningful specification.

Example (from lecture 2.2)

- $\log(\text{Wage})_i =$

  $\beta_1 + \beta_2 \text{Female}_i + \beta_3 \text{Age}_i + \beta_4 \text{Educ}_i + \beta_5 \text{Parttime}_i + \varepsilon_i.$

- $\log(\text{Wage})_i =$

  $\beta_1 + \beta_2 \text{Female}_i + \beta_3 \text{Age}_i + \beta_4 \text{Educ}_i + \beta_5 \text{Parttime}_i +$

  $\gamma_1 \text{Female}_i \text{Educ}_i + \gamma_2 \text{Age}_i^2 + \varepsilon_i.$

Now:

- Wage differential may depend on education ($\beta_2 + \gamma_1 \text{Educ}_i$).
- Age has non-linear effect for wage ($\beta_3 + 2\gamma_2 \text{Age}_i$).

Dummy variables

Quarterly data

$$y_i = \alpha_i + \sum_{j=2}^{k} \beta_j x_{ji} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where α_i is the quarter-specific mean level.

Use dummy variables:

- D_{hi} for $h = 1, 2, 3, 4$, with $D_{hi} = 1$ if observation i is in quarter h (and $D_{hi} = 0$ otherwise).
- Then

$$y_i = \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \sum_{j=2}^{k} \beta_j x_{ji} + \varepsilon_i.$$

Dummy variables

$$y_i = \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \sum_{j=2}^{k} \beta_j x_{ji} + \varepsilon_i.$$

Test

Can we add a constant term to the above specification that has a dummy for each quarter?

Answer: Only if one of the dummy parameters is set to 0.

If we omit D_{1i}, so $\alpha_1 = 0$, we get

$$y_i = \alpha_1 + \gamma_2 D_{2i} + \gamma_3 D_{3i} + \gamma_4 D_{4i} + \sum_{j=2}^{k} \beta_j x_{ji} + \varepsilon_i,$$

where $\gamma_h = \alpha_h - \alpha_1$ for $h = 2, 3, 4$.

TRAINING EXERCISE 3.3

- Train yourself by making the training exercise (see the website).
- After making this exercise, check your answers by studying the webcast solution (also available on the website).

Erasmus School of Economics

MOOC Econometrics

Lecture 3.4 on Model Specification: Evaluation

Michel van der Wel

Erasmus University Rotterdam

RESET

Extend linear model

$$y_i = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \varepsilon_i,$$

to non-linear model

$$y_i = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \sum_{j=2}^{k} \gamma_{jj} x_{ji}^2 + \sum_{j=2}^{k} \sum_{h=j+1}^{k} \gamma_{jh} x_{ji} x_{hi} + \varepsilon_i.$$

Test for linearity by testing significance of γ coefficients.

Challenge: Nonlinear model contains many parameters.

RESET

Instead, add fitted y-values $\hat{y} = Xb = X(X'X)^{-1}X'y$ to the model:

$$y_i = x_i'\beta + \sum_{j=1}^{p} \gamma_j (\hat{y}_i)^{j+1} + \varepsilon_i,$$

and test for joint significance of γ's. Under null of correct specification, H_0: $\gamma_j = 0$ for all j, test distribution approximately $F(p, n-k-p)$.

Test

For $p = 1$, compute the number of extra parameters in the alternative specification as compared to the total number of parameters in the RESET specification.

Answer: Above model with $p = 1$ has $k + 1$ parameters. Model with squares and cross-terms has $k + (k-1) + \frac{1}{2}(k-2)(k-1)$ coefficients. For example, if $k = 6$, then this is 7 compared to 21.

Chow break test

In case of a possible break, split the sample and test for constancy of parameters.

$$\begin{aligned} y_1 &= X_1\beta_1 + \varepsilon_1 && (n_1 \text{ observations}) \\ y_2 &= X_2\beta_2 + \varepsilon_2 && (n_2 = n - n_1 \text{ observations}) \end{aligned}$$

Combine:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

Test H_0: $\beta_1 = \beta_2$ against this unrestricted set-up.

Chow break test

F-test for null hypothesis of no break:

$$F = \frac{(e_R'e_R - e_U'e_U)/k}{e_U'e_U/(n-2k)}.$$

Here:

- Have $e_U = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$, with e_j the OLS residuals of each group.
- Thus $e_U'e_U = e_1'e_1 + e_2'e_2 \equiv S_1 + S_2$.
- Get $F = \frac{(S_0 - S_1 - S_2)/k}{(S_1+S_2)/(n-2k)}$, with $S_0 = e_R'e_R$.

Chow forecast test

A variation on the Chow break test is based on

$$y_i = x_i'\beta + \sum_{j=n_1+1}^{n_1+n_2} \gamma_j D_{ji} + \varepsilon_i,$$

test H_0: $\gamma_j = 0$ for all j.

Test

What is the number of parameters in the above specification?

Answer: The model contains the usual k variables and n_2 dummy-variables (one for each observation in group 2), so in total $k + n_2$ parameters.

- Perfect fit in second sample, thus $e_2 = 0$.
- Thus $F = \frac{(S_0 - S_1)/n_2}{S_1/(n_1-k)}$.

Test for normality of error terms

- Model misspecification may appear in the error terms.
- Normality of ε can be tested by distribution of residuals.
- Jarque-Bera test evaluates skewness S and kurtosis K:

$$JB = \left(\sqrt{\frac{n}{6}}S\right)^2 + \left(\sqrt{\frac{n}{24}}(K-3)\right)^2,$$

which approximately has $\chi^2(2)$ distribution if H_0: $\varepsilon_i \sim NID(0, \sigma^2)$ holds true.

TRAINING EXERCISE 3.4

- Train yourself by making the training exercise (see the website).
- After making this exercise, check your answers by studying the webcast solution (also available on the website).

Erasmus
School of
Economics

MOOC Econometrics

Lecture 3.5 on Model Specification: Application

Michel van der Wel

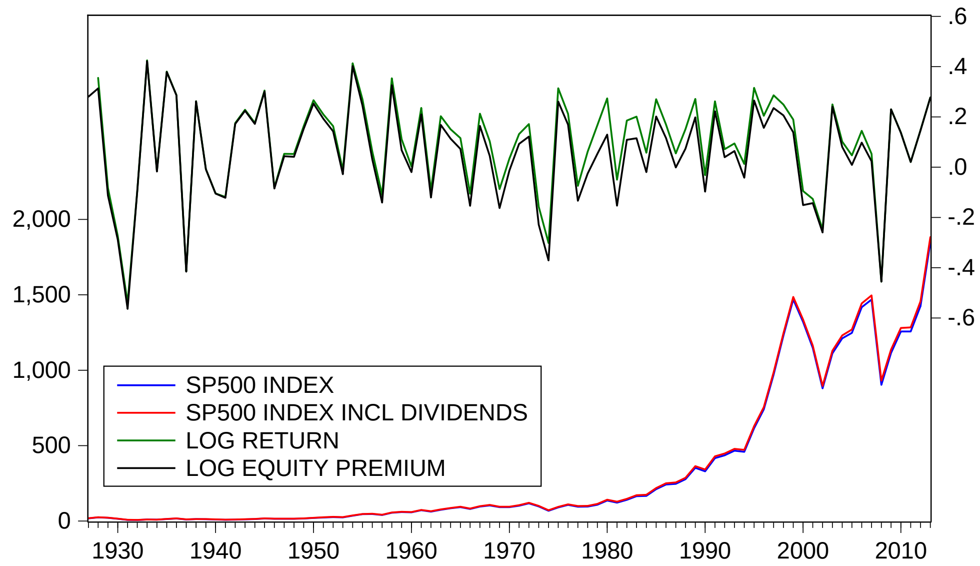
Erasmus University Rotterdam

Setting

Application:

- Model/forecast S&P500 stock index
  - Should we transform this index series?
- Large set of explanatory variables
  - Which to select?
- Choice of model
  - How to evaluate candidate models and how to compare them?
  - Is relationship stable over time?

Transformation



Variable selection

Dependent variable: Log of equity premium; Sample size: 87					
	Coefficients (p-values) by specification				
	(A)	(B)	(C)	(D)	(E)
Constant	0.166 (0.001)	0.062 (0.015)	−0.266 (0.076)	−0.027 (0.848)	0.065 (0.015)
Book-to-market	−0.185 (0.019)				
Issued Stock		−0.147 (0.850)			
Dividend/Price			−0.097 (0.029)		
Earnings/price				−0.032 (0.532)	
Inflation					−0.166 (0.746)
R-squared	0.063	0.000	0.055	0.005	0.001

General to specific

Dependent variable: Log of equity premium; Sample size: 87					
	Coefficients (p-values) by specification				
	(1)	(2)	(3)	(4)	(5)
Constant	0.234 (0.544)	0.205 (0.554)	0.215 (0.537)	0.489 (0.038)	0.166 (0.001)
Book-to-market	−0.176 (0.257)	−0.166 (0.249)	−0.191 (0.178)	−0.290 (0.008)	−0.185 (0.019)
Issued Stock	−0.146 (0.859)				
Dividend/Price	−0.120 (0.226)	−0.126 (0.174)	−0.090 (0.286)		
Earnings/price	0.167 (0.052)	0.167 (0.051)	0.127 (0.088)	0.097 (0.159)	
Inflation	−0.567 (0.337)	−0.564 (0.336)			
R-squared	0.108	0.108	0.098	0.085	0.063

Stability

$$\log(\text{EqPr})_i = \beta_1 + \beta_2 \text{BTM}_i + \beta_3 \text{BTM}_i \times D_i^{\text{War}} + \beta_4 \text{BTM}_i \times D_i^{\text{Oil}} + \varepsilon_i,$$

Dependent variable: Log of equity premium; Sample size: 87		
	Coefficients	p-values
Constant	0.160	0.002
Book-to-market	-0.175	0.036
Book-to-market × War-dummy	0.078	0.440
Book-to-market × Oil-dummy	-0.133	0.287
R-squared	0.085	

Test

What is the coefficient on Book-to-market during the war years?

Answer: $-0.175 + 0.078 = -0.097$.

Model comparison

	Full model	Book-to-market
R^2	0.108	0.063
AIC	-0.444	-0.486
BIC	-0.273	-0.430

Model evaluation

Book-to-market		
	Test statistic	p-value
RESET ($p = 1$)	3.446	0.067
Chow Break	2.269	0.110
Chow Forecast	0.765	0.794
Jarque-Bera	7.155	0.028

Note: As break-point 1980 is chosen.

Test

Will the p-values increase if the full model is considered?

Answer: Not possible to say.

TRAINING EXERCISE 3.5

- Train yourself by making the training exercise (see the website).
- After making this exercise, check your answers by studying the webcast solution (also available on the website).