Econometrics: Model Specification

Diego López Tamayo * Based on MOOC by Erasmus University Rotterdam

Contents

Model Specification **2**
How to specify? . . . 9
Estimation bias . . . 9
Efficiency loss . . . 9
Bias-variance trade-off . . . 10
Information Criteria . . . 11
Out of sample prediction . . . 11
Iterative selection methods . . . 11
Data Transformation . . . 12
Log and first difference . . . 12
Non-linearity . . . 13
Examples with SP500 dataset: . . . 14
Evaluation of models . . . 16
RESET . . . 17
Chow Break Test . . . 17
Chow forecast test . . . 18
Jarque-Bera . . . 19
Proofs for Chow tests . . . 19
Application on SP500 . . . 20
Variable transformation . . . 21
Testing Specification . . . 24
General-to-specific . . . 26
Stability . . . 26
Testing Information criteria . . . 28
RESET . . . 29
Chow Break . . . 29
Chow Forecast . . . 30
Normality test . . . 30

"There are two things you are better off not watching in the making: sausages and econometric estimates." -Edward Leamer

*El Colegio de México, diego.lopez@colmex.mx

Model Specification

```
dataset3 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/dataset3.csv")
dataset3 <- dataset3 %>% mutate(year_orig = Year)
dataset3$Year <- as.Date(ISOdate(dataset3$Year, 12, 31))
```

Datset:

This is a stock market data set for the United States for 1927-2013 (yearly data). The source of the data is the updated version of the Goyal and Welch (2008)1 data. The data are available from the website of Prof Amit Goyal

The variables are:

- **Year**
- **Index**: The S&P500 index
- **Dividends**: Dividends on the index ("D12" in the Goyal and Welch [GW] file)
- **Riskfree**: Riskfree rate ("Rfree" in GW)
- **LogEqPrem**: Log of the equity premium (calculated following GW) Calculated as: $\frac{(Index+D12)}{Index(-1)} - log(1+Rfree)$, where $x(-1)$ denotes value from previous period, log is the natural logarithm, $D12$ dividends and $Rfree$ the riskfree rate.
- **BookMarket**: Book to market ratio ("b/m" in GW)
- **NTIS**: Equity issued ("ntis" in GW)
- **DivPrice**: Dividend to price ratio (calculated following GW) Calculated as: $log(D12) - log(Index)$, where $D12$ are dividends.
- **EarnPrice**: Earnings to price ratio (calculated following GW) Calculated as: $log(E12)/log(Index)$, where $E12$ are earnings.
- **Inflation**: Inflation rate ("infl" in GW)

Suppose we have a data set of a stock price index with a large number of variables which of which we suspect they may explain movements in the stock index.

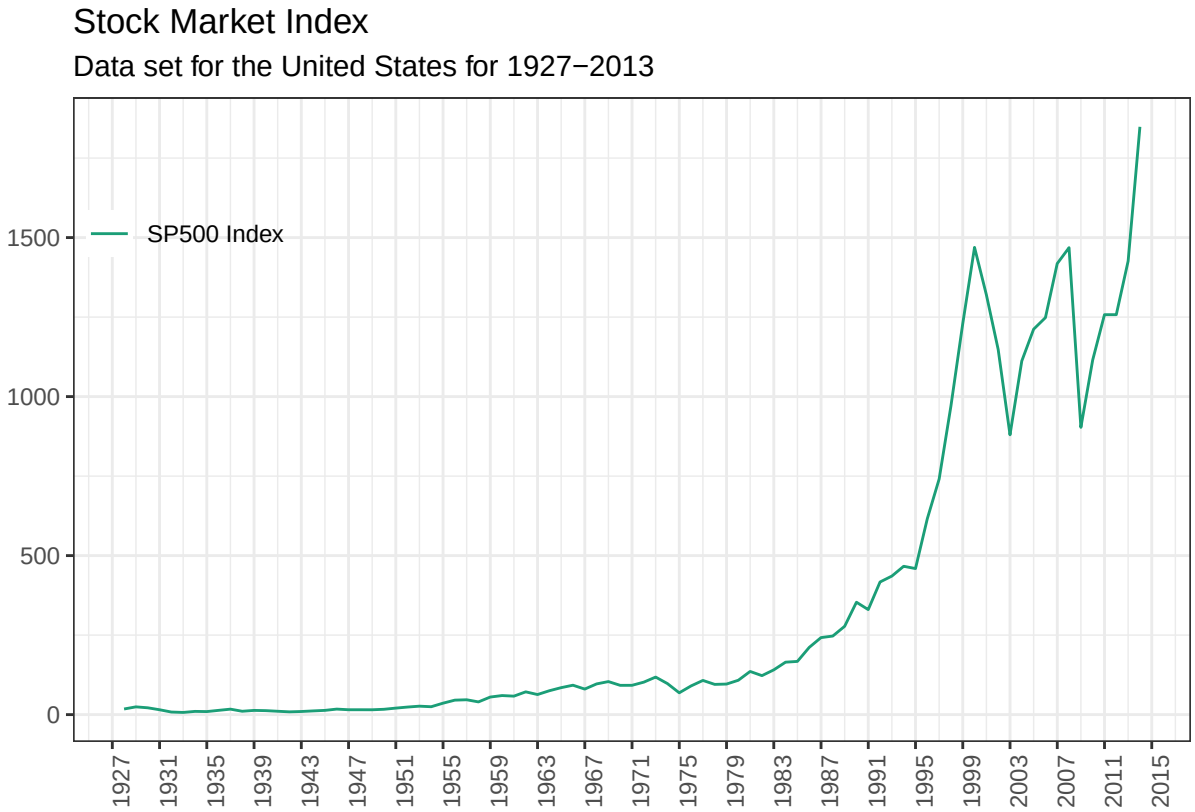
There are a number of questions that we need to address before we can actually formulate a model for a stock price index as function of the explanatory variables.

- Do we include all explanatory variables or only a few? And if we don't include all variables, how can we select which of the variables to include? Counterintuitive as it may seem, we do not always include all variables.
- Do we take the data as they are, or transform the variables?
- Once we have a model, how can we evaluate whether the model is appropriate in some sense?

These questions are, of course, relevant in any kind of application, not just the stock market setting which is the focus of this section.

We will illustrate these questions by looking at an example.

```
dataset3 %>% ggplot(aes(x=Year)) +
  geom_line(aes(y=Index, col = "SP500 Index")) +
  labs(x = "", y = "", title = "Stock Market Index",
       subtitle = ("Data set for the United States for 1927-2013")) +
    scale_x_date(date_breaks = "4 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.1, .8),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")
```



This figure shows the annual evolution of the S&P 500 stock price index over the years 1927 up to 2013. There's an exponential growth visible in the figure. Some interesting episodes stand out. For example, the .com bubble at the end of the 1990s and its burst in the early 2000s, and the financial crisis starting 2007, 2008. Of course, there were more crises, but those stand out less clearly in this figure, precisely because of the exponential growth.

We're interested in modeling this series, and have a number of explanatory variables. Modeling and forecasting of stock prices is not easy, and many variables have been examined:

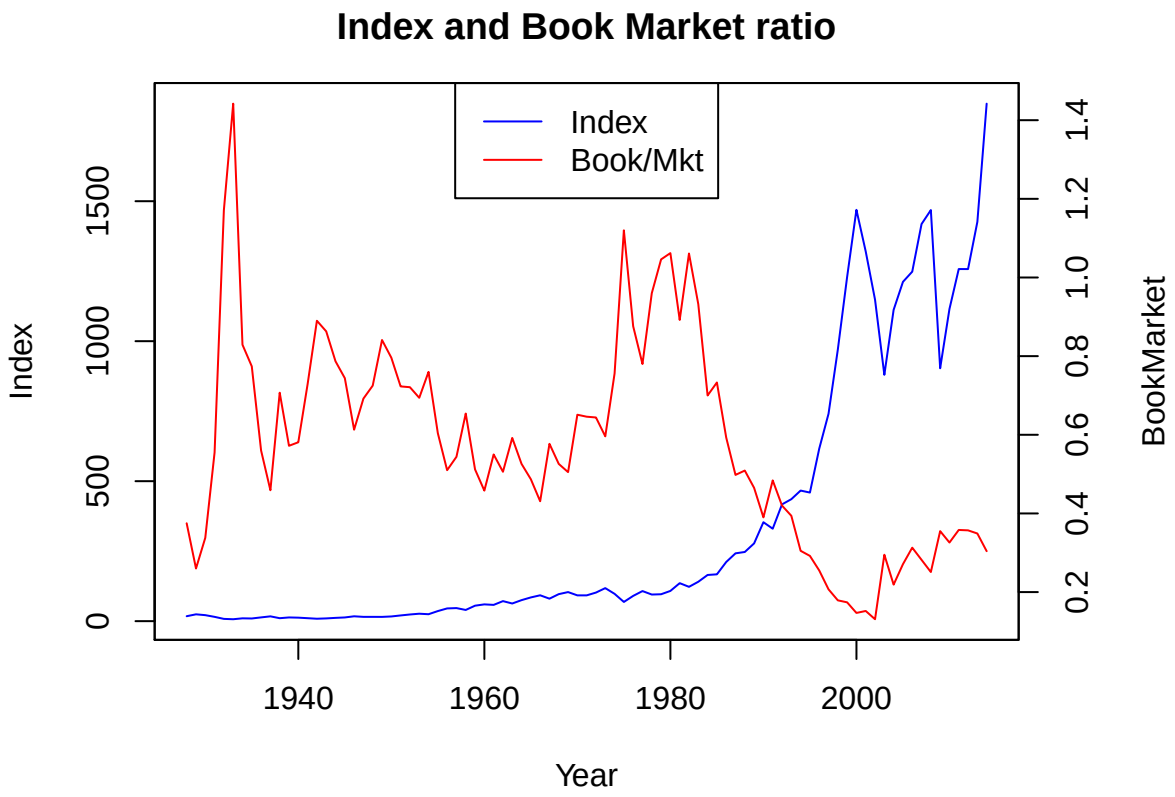
- Stock characteristics: Dividends, earnings, volatility, book value, issuing activity.
- Interest-rate related: Treasury bill rates, long term yields, corporate bond returns.
- Macroeconomic: Inflation, investment, consumption.

This list is not exhaustive, and it is hopefully already obvious that it is quite a challenge to select the important variables, if any. The first question we turn to is precisely on how to make this decision. Do we simply select all variables or just a few, and if a few, which ones?

Let's take one of the explanatory variables, which is the **book-to-market ratio**. This is the book value of the firms relative to the market value. The picture on the left plots the index together with this variable, with the index in blue on the left axis and the book-to-market ratio in red on the right axis.

```
par(mar = c(5, 5, 3, 5))
plot(dataset3$Year,dataset3$Index, type ="l", xlab = "Year", ylab = "Index", col="blue",main = "Index a
par(new = TRUE)
plot(dataset3$Year,dataset3$BookMarket, type ="l", xaxt = "n", yaxt = "n",
     ylab = "", xlab = "", col="red", lty = 1)
axis(side = 4)
mtext("BookMarket", side = 4, line = 3)
legend("top", c("Index", "Book/Mkt"),
```

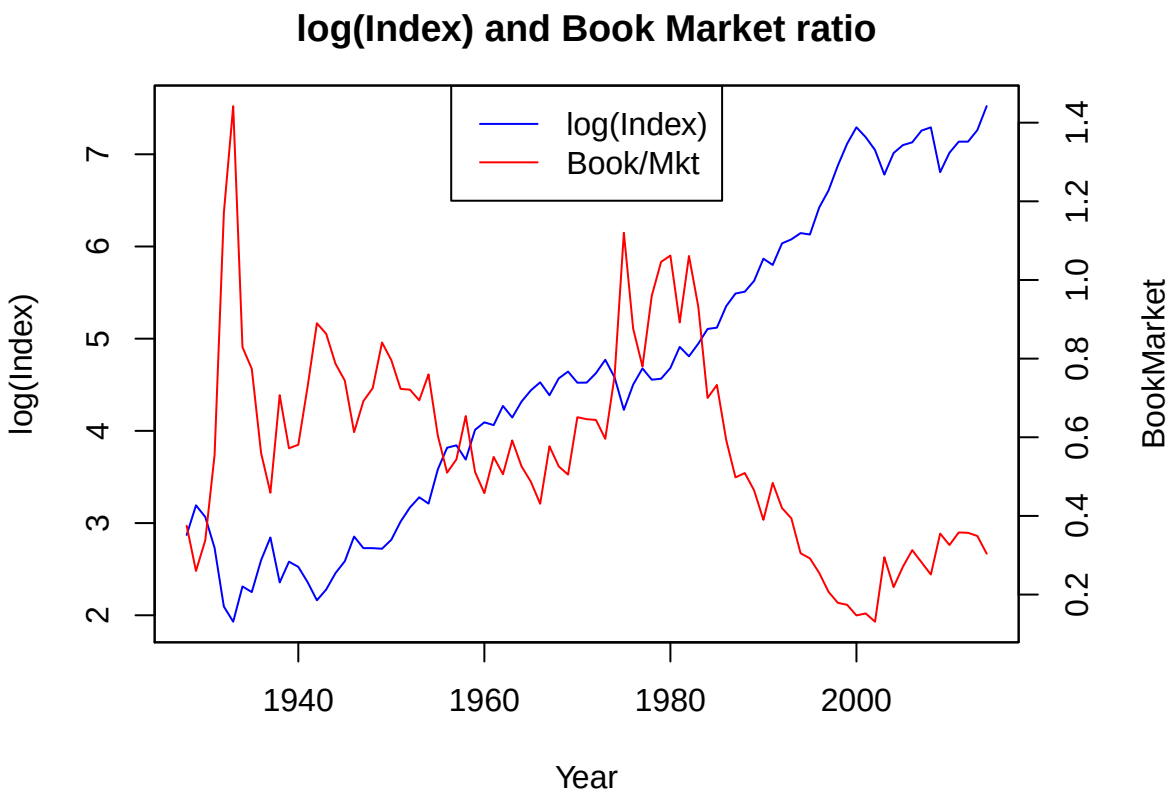
```
       col = c("blue", "red"), lty = c(1, 1))
```



It is obvious the two variables behave differently. The index grows exponentially, while the book-to-market ratio stays relatively stable over time. We can transform the series in order to get a more similar behavior. For example, to undo the exponential growth, we can take the log of the index.

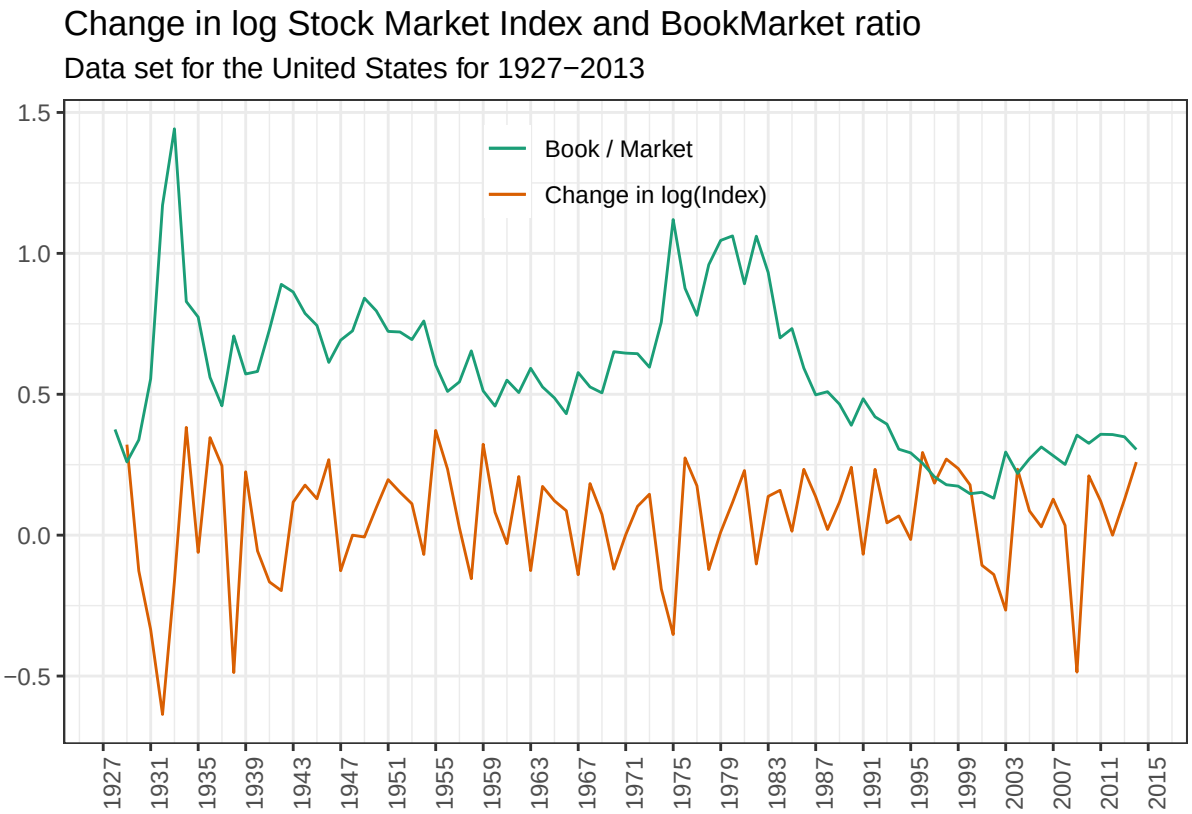
This figure plots the log of the index together with the book-to-market ratio, and just by looking at the picture, it seems we got the variables a bit more on the same scale. Taking the log of a series is a very common transformation.

```
par(mar = c(5, 5, 3, 5))
plot(dataset3$Year,log(dataset3$Index), type ="l", xlab = "Year", ylab = "log(Index)", col="blue",main =
par(new = TRUE)
plot(dataset3$Year,dataset3$BookMarket, type ="l", xaxt = "n", yaxt = "n",
     ylab = "", xlab = "", col="red", lty = 1)
axis(side = 4)
mtext("BookMarket", side = 4, line = 3)
legend("top", c("log(Index)", "Book/Mkt"),
       col = c("blue", "red"), lty = c(1, 1))
```



It turns out that in our current application, we still need another transformation. We do not considered the log of the series directly, but the change in the log of the index from one period to the next. This figure plots the **change of the log of the index** against the book to market ratio, and indeed now the variables move on the same scale.

```
dataset3 <- dataset3 %>% mutate(log_sp500=log(Index),dif_log_sp500=c(NA,diff(log(Index))))
dataset3 %>% ggplot(aes(x=Year)) +
  geom_line(aes(y=dif_log_sp500, col = "Change in log(Index)")) +
  geom_line(aes(y=BookMarket, col = "Book / Market")) +
  labs(x = "", y = "", title = "Change in log Stock Market Index and BookMarket ratio",
       subtitle = ("Data set for the United States for 1927-2013")) +
    scale_x_date(date_breaks = "4 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.5, .90),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")
```



After the 1980s, the book-to-market flattens out a bit, and goes to a lower level. It is not clear the relationship between the stock index and book-to-market ratio is stable before the 1980s and or after. Later, we talk about methods to test whether there's a break in the relationship and also discuss tests that can inform us whether the model is actually good enough.

We can regress the change of the log index on a constant and book-to-market to study this relation in more detail.

```
lm1 <- lm(dif_log_sp500 ~ BookMarket , data = dataset3)
# We add the residuals and fitted values into the dataset
dataset3 <- dataset3 %>% mutate(lm1.res = c(NA,resid(lm1)))
dataset3 <- dataset3 %>% mutate(lm1.pred = c(NA,predict(lm1)))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Jul 09, 2020 - 22:56:30

$$\Delta log(SP500index) = 0.177 - 0.213\ddot{O}BookMarket + e.$$

It turns out book-to-market is significant in explaining the change in the log of the stock index. It's significant at a 1% level, and the r-squared of this regression is 8%.Since book-to-market is defined as book value divided by market value, a high book-to-market period typically coincides with a period when the market value is low and has decreased. So when stock market values are low and have decreased, the stock market index has decreased. This is precisely what the coefficient tells us.

Perhaps, you already expected the significant explanatory power when modeling stock index movements with a variable that depends on the market value, but it turns out that book-to-market is also important when we forecast the stock market.

We took a transformation to get at the significant explanatory power for the stock market and this was rather ad hoc. More detailed considerations for transforming variables and related concepts, such as non-linear

Tabla 1: Regression Results

	Dependent variable:
	dif_log_sp500
BookMarket	-0.213^{***}
	(0.079)
Constant	0.177^{***}
	(0.050)
Observations	86
R^2	0.080
Adjusted R^2	0.069
Residual Std. Error	0.191 (df = 84)
F Statistic	7.295^{***} (df = 1; 84)
Note:	*p<0.1; **p<0.05; ***p<0.01

effects, will be treated later.

What would happen if we regress regress the S&P500 index (without any kind of transformation) on a constant and the book-to-market ratio.

```
lm2 <- lm(Index ~ BookMarket , data = dataset3)
# We add the residuals and fitted values into the dataset
dataset3 <- dataset3 %>% mutate(lm2.res = resid(lm2))
dataset3 <- dataset3 %>% mutate(lm2.pred = predict(lm2))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Jul 09, 2020 - 22:56:30

Tabla 2: Regression Results

	Dependent variable:
	Index
BookMarket	$-1{,}217.758^{***}$
	(150.794)
Constant	$1{,}035.403^{***}$
	(95.016)
Observations	87
R^2	0.434
Adjusted R^2	0.427
Residual Std. Error	366.477 (df = 85)
F Statistic	65.216^{***} (df = 1; 85)
Note:	*p<0.1; **p<0.05; ***p<0.01

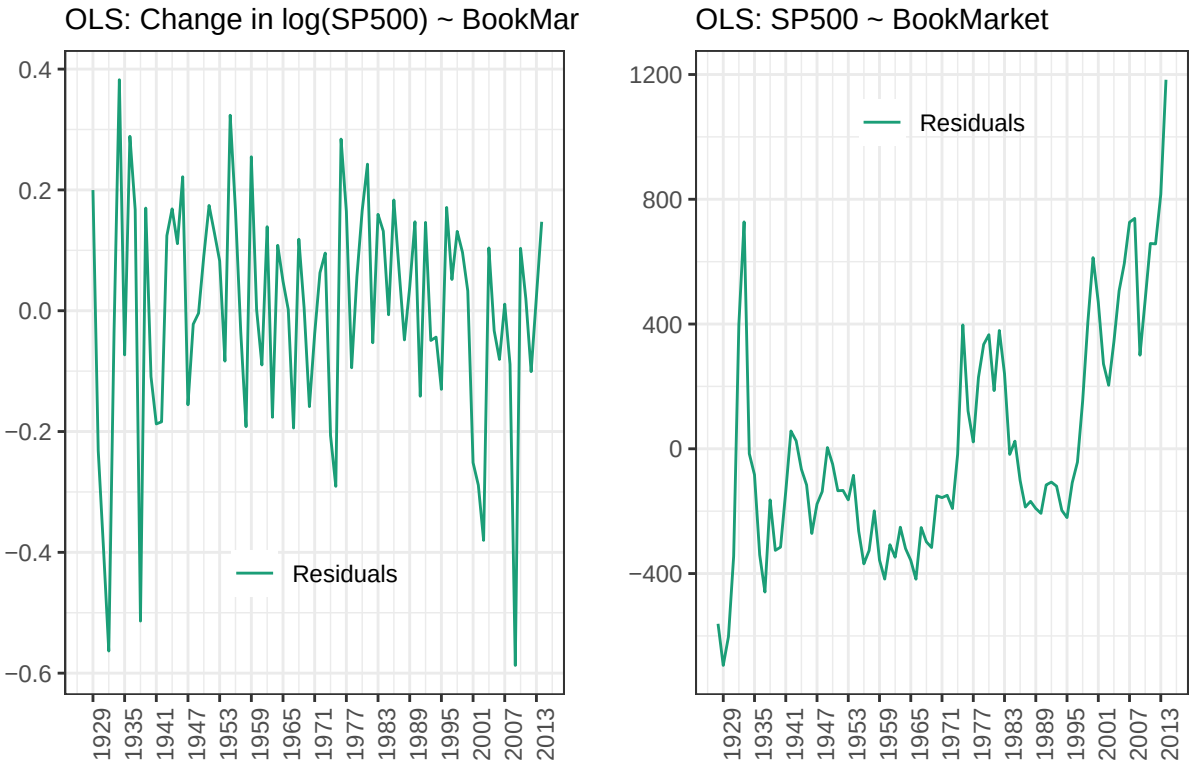
$$SP500index = 1035.35 - 1217.68 \cdot BookMarket + e.$$

The effect of BookMarket is still significant. We make a plot of the residuals e from both models:

```
plot_a <- ggplot(data=dataset3, aes(x=Year)) +
  geom_line(aes(y=lm1.res, col = "Residuals")) +
#  geom_line(aes(y=lm1.pred, col = "Fitted")) +
#  geom_line(aes(y=dif_log_sp500, col = "Actual")) +
  labs(x = "", y = "", title = "",
       subtitle = ("OLS: Change in log(SP500) ~ BookMarket")) +
    scale_x_date(date_breaks = "6 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.5, .20),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")

plot_b <- ggplot(data=dataset3, aes(x=Year)) +
  geom_line(aes(y=lm2.res, col = "Residuals")) +
#  geom_line(aes(y=lm2.pred, col = "Fitted")) +
#  geom_line(aes(y=Index, col = "Actual")) +
  labs(x = "", y = "", title = "",
       subtitle = ("OLS: SP500 ~ BookMarket")) +
    scale_x_date(date_breaks = "6 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.5, .90),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")

grid.arrange(plot_a, plot_b, nrow = 1)
```



The residuals in both regressions are clearly not the same. The most obvios difference is that in the Not transformed SP500 model, the residuals have a pattern and strong persistence (violating the assumption A7 $e \sim N(0,1)$)

How to specify?

We start with the familiar model where the dependent variable y is explained by a set of variables collected in X. Here, y can be a stock index return and X a number of variables that may explain movements in the stock index.

$$y = X\beta + \epsilon$$

The question we'll address now is which variables to include in the matrix X. It turns out that there's a tough trade off that we face.

If one considers a model with a small number of variables, there is the risk that relevant variables are missed, and thus actually too few variables are included. This will lead to an **estimation bias**.

If one, however, considers a model with too many variables, there's an **efficiency loss.**

- To few variables → Bias
- To many variables → Efficiency loss (more variance) even if all variables matter.

Estimation bias

We compare two models. Suppose that the data-generating process, DGP in short, contains two group of explanatory variables, X1 and X2.

$$\text{DGP} : y = X_1\beta 1 + X_2\beta 2 + \epsilon \rightarrow b_1, b_2$$
$$\text{Estimated Model} : y = X_1\beta 1 + \tilde{\epsilon} \rightarrow b_R$$

We contrast this with the actual estimated model which only contains X1. In this model we denote the estimator of β_1 by b_r, where r stands for restricted as we've restricted $\beta_2 = 0$ to zero. Also a tilde is added to the disturbance term, to indicate that it is different from the one in the DGP $\epsilon \neq \tilde{\epsilon}$. The estimators of β_1 and β_2 in the DGP are then ordered by b_1, b_2.

We can express $E(b_R)$ as a function of β_1 and β_2.

$$E(b_R) = E((X_1'X_1)^{-1}X_1y) = E((X_1'X_1)^{-1}X_1(X_1\beta 1 + X_2\beta 2 + \epsilon))$$
$$E(b_R) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 = \beta_1 + P\beta_2$$

With $P = (X_1'X_1)^{-1}X_1'X_2$.

This gives us the first result. The restricted estimator will be *biased* unless $\beta_2 = 0$ is zero, or X_1 and X_2 are completely orthogonal, such that the product and thus P is zero. We refer to this bias as the omitted variable bias.

Efficiency loss

Now we turn to the efficiency part. Efficiency concerns the variance of our estimators. We prefer estimators that have no or small bias with low variance. An estimator with the lowest possible variance is called efficient.

We can use the fact that $b_R = b_1 + Pb_2$ and $Cov(b_2, b_R) = 0$. Notice that we can rewrite $b_1 = b_R - Pb_2$ and

$$Var(b_1) = Var(b_R - Pb_2) = Var(b_R) + Var(Pb_2) - 2Cov(b_R, Pb_R)$$

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

Solving for $Var(b_R)$

$$Var(b_R) = var(b_1) - Pvar(b_2)P'$$

The variance of the restricted estimator, b_R, is equal to the variance of the unrestricted estimator, b_1 minus a positive semi-definite term, such that the variance of b_1 is always larger than that of b_R.

While the benefit of adding variables is bias reduction, a cost is thus increased variance.

Bias-variance trade-off

One way to get more insight into the bias-efficiency trade-off (also referred to as the bias-variance trade-off) is to combine bias and efficiency in the Mean Squared Error (MSE). The mean squared error is defined as:

$$MSE(b) = E((b-\beta)(b-\beta)')$$

with b a certain estimator of the unknown parameter β.

$$MSE(b) = E(bb' - b\beta' - \beta b' + \beta\beta') = E(bb') - E(b)\beta' - \beta E(b') + \beta\beta'$$

Notice that $Var(b) = E(bb') - E(b)E(b)'$ from the definition of variance. So $E(bb') = Var(b) + E(b)E(b)'$

$$MSE(b) = Var(b) + E(b)E(b)' - E(b)\beta' - \beta E(b') + \beta\beta'$$

We can add and subtract $E(b)E(b)'$ from the MSE expression and rewrite to get:

$$MSE(b) = Var(b) + E(b-\beta)E(b-\beta)'$$

Using this result in the context of b_1 and b_R, since $MSE(b_1) = Var(b_1) + E(b_1-\beta_1)E(b-\beta_1)' = Var(b_1)$ since $E(b_1) = \beta_1$ because we use the correct model and there's no bias- For $MSE(b_R) = Var(b_R) + E(b_R - \beta_R)E(b-\beta_R)'$ we cannot simplify any further. It can be shown that:

$$MSE(b_1) - MSE(b_R) = P(Var(b_2) - \beta_2\beta_2')P'$$

The resctricted estimator b_R is better when $MSE(b_1) - MSE(b_R) > 0$ there are two cases:

1. $\beta_2 = 0$, the restricted model is better as $P(Var(b_2))P' > 0$ so when the second group of regresson is not relevant, the MSE would tell us to ignore them and use b_R
2. $\beta_2 \neq 0$, the restricted model is if $Var(b_2) - \beta_2\beta_2'$ is PSD, thus when the variance of estimator of b_2 is large relative to it's influence.

The next step is to translate this finding into some measures, or Metrics, that we can use to find a good trade off between bias and efficiency. We turn to two commonly used decision metrics, **information criteria and out-of-sample prediction.**

Information Criteria

Often there's a preference for small models in the sense that a limited number of variables are included. When adding variables, at a certain stage the added benefit of yet another variable will be relatively small, and it is good to stop adding variables to the model. Information criteria capture this idea. They study the goodness of fit of a model, here captured with the standard error of the regression s, but impose a penalty on the number of parameters k. Two commonly used information criteria are the Akaike information criterion, abbreviated with AIC, and the Bayesian information criterion, abbreviated with BIC.

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

$$\text{Bayes}: \ BIC = log(s^2) + \frac{klog(n)}{n}$$

For both the AIC and BIC the value is equal to the log of the squared standard error of the regression plus a term that is a function of k, the number of variables in the model. The two information criteria differ in the penalty they impose on the number of parameters. When comparing models, a **lower value of the information criteria is preferred** as we aim for a low standard error of the regression.

The penalty on the number of parameters k is 2/n for the AIC and this is log(n) over n for BIC. When $log(n) > 2$, the BIC imposes a stronger penalty. Thus for eight or more observations, BIC imposes a stronger penalty than AIC.

Out of sample prediction

The information criteria are based on so-called **in-sample results**: using all observations in a sample. Often we're also interested in the predictive performance of our model. This can be in a time series sense, that we want to forecast a stock price to earn some money, but also if you have data on household consumption and want to predict whether they will buy a certain product or not.

In such cases, the full sample can be split in an in-sample part, often referred to as the training sample, and an out-of-sample part. The observations in the second out of sample part are kept out of the main analysis, for example when estimating beta, and they're only used to examine the predictive ability of the model.

Two commonly used out of sample criteria are the **root mean squared error, RMSE, and the mean absolute error, MAE**.

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

$$MAE = \frac{1}{n_f}\sum_{i=1}^{n_f} \mid y_i - \hat{y}_i \mid$$

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

Both criteria consider the difference between the actual observation y_i and the predicted value, $\hat{y}_i$, but they differ slightly in how the prediction errors are averaged. In both cases, a **lower value means a better model.**

Iterative selection methods

Now let us return to the problem that a researcher faces, how to decide which variables to include in X. If you consider removing a group of regressors, you can use an F-test for a joint significance of the second group of coefficients, or simply a t-test if you wish to remove only a single variable. However, be aware that **these tests are only concerned with the significance and do not incorporate the bias efficiency**

trade off. If you already have a set of candidate models that differ in the number of parameters, information criteria can be of use. These take into account that small models are preferred if more complex models do not perform sufficiently better.

Here you can also consider using out-of-sample prediction. If the goal is prediction and there are a number of candidate models, you may as well pick the one that has the most predictive power, and provides the lowest root mean squared error or mean absolute error.

Very often we're, however, not fortunate enough to start with two groups of regressors, X1 and X2, or with a candidate set of models, and we need to get just one model first. In this case, iterative selection methods can be of great help. These come in two variants:

- **General to specific:** you start with the most general model, including as many variables as are at hand. Then check whether one or more variables can be removed from the model. This can be based on individual t-tests, or a joint F-test in case of multiple variables. In case you remove one variable at a time, the variable with the lowest absolute t-value is removed from the model. The model is estimated again without that variable, and the procedure is repeated. The procedure continues until all remaining variables are significant.
- **Specific to general.:** follows the same logic, but starts with a very small model, sometimes even only consisting of the constant term. Variables get added one at a time, choosing the one that has the largest absolute t-statistic. This procedure is repeated until no significant variables can be added anymore.

Both procedures have pros and cons. The specific-to-general approach starts small, which is appealing. However, many variations need to be tried at the initial steps. Also, it can easily happen that important variables are missing in initial phases so that initial tests are performed in mis-specified models.

Data Transformation

We start again with the model where we explain a dependent variable y, with one or more explanatory variables collected in a vector x. A relevant question is, what is the most appropriate form of the data? An important consideration is that the variables should be incorporated in a compatible manner.

$$y = X\beta + \epsilon$$

If our y variable is a level, such as the number of unemployed individuals, it makes more sense to relay that to X variables that also capture levels, such as the level of production. Similarly if our y variable is some growth rate then it makes most sense to relate that to an X variable that also considers a growth rate. It makes less sense to explain the growth rate of unemployment with the level of production.

If variables are not similar in nature one should consider transforming data. We discussed two very common transformations.

Log and first difference

The first transformation is taking a logarithm of a series. A case where this is a sensible transformation is when there is some exponential growth. In case of exponential growth, such as commonly found in the level of macroeconomic and financial quantities, the properties of the series are not stable. The logarithmic transformation then brings back stability in the sense that the explosive behavior is removed.

The second transformation is taking the difference of a variable relative to its previous observed value. This transformation makes most sense when data capture observations for a variable at different points in time, and are thus ordered. Sometimes such a data set, often referred to as a time series data set, shows a **trend.** When there is such a trending pattern, it may affect the stability properties of the series, which causes statistical assumption to not hold.

Fortunately, the stability is oftentimes easily restored by taking the difference.

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

Non-linearity

So far, we've considered non-linear transformation on the variables. Let us study non-linearity a bit further.

$$y_i = x_i'\beta + \epsilon = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \epsilon_i$$

At the top here is the usual setting, where the dependence of y on a constant and k-1 other explanatory variables is written separately. The marginal effects are constant and simply equal to the beta parameters. $\frac{\partial y_i}{\partial x_{ji}} = \beta_j$

We can extend this setting to get nonlinear effects. For example, we can consider the square of all the explanatory variables. Also, we can consider cross-products of the explanatory variables, which we often refer to as **interaction terms**. Taking both together in our usual linear model, we get a set-up such as on the middle of the slide.

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

There are two reasons to consider this structure. First, it allows for a non-linear functional form, here quadratic. We can ask to extend this further by adding cubic or even higher order terms, which allows for very rich non-linear relationships. The nice thing is that for all sorts of variations, the relationship from X to y is non-linear, but the setup remains linear in the unknown parameters beta.

Taking the square of a series, or cross product of two series, does not depend on parameters, and enters linearly. Thus, ordinary least squares can still be used. A second reason for such a set-up is that, even though the structure itself may seem somewhat contrived, it may actually provide a meaningful economic specification.

As an example of this, let us go back to the second series of lectures, where attention was paid to wage regressions. One of the specifications considered is repeated here, where the log(Wage) is explained by a constant, a dummy whether the i-th observation is female or not, the age, education level, and dummy for part-time work.

$$log(Wage)_i = \beta_1 + \beta_2 Female_i + \beta_3 Age_i + \beta_4 Educ_i + \beta_5 Parttime_i + \epsilon_i$$

We extend this model with quadratic and interaction terms.

$$log(Wage)_i = \beta_1 + \beta_2 Female_i + \beta_3 Age_i + \beta_4 Educ_i + \beta_5 Parttime_i + \gamma_1 Female_i Educ_i + \gamma_2 Age_i^2 + \epsilon_i$$

In this specification, there's an interaction term for the gender dummy and education level measured by γ_1, and a quadratic term for age measured by γ_2. This small extension allows for two extra effects. First, because of the new interaction term, the partial wage differential is allowed to depend on education.

- The gender effect is now $\frac{\partial log(Wage)}{\partial Female_i} = \beta_2 + \gamma_1 \cdot Educ_i$

This allows for the possibility that the wage differential as compared to men is different for higher-educated woman or lower-educated woman. In fact, in this setting such a hypothesis can simply be tested by studying the significance of gamma1.

- The effect of an increase of age is $\frac{\partial log(Wage)}{\partial Age_i} = \beta_3 + 2\gamma_2 \cdot Age_i$

The squared term of age allows for a non-linear effect of age. This allows for the possibility that the wage increases more during relatively young age when climbing the career ladder and less for older age.

Naturally we could have added other squared and other interaction terms as well in this specification. In fact, it is possible to start again with a very general set up with all squares and interaction terms and use model selection of section Model Specification to get to a more specific model.

We can also use dummy variables to get a somewhat richer model structure and add non-linearities. The mean level of data that are measured quarterly may differ across each of the four quarters. This can be captured by replacing the constant term by the quarter specific mean level $alpha_i$. We can easily formulate this in our usually framework by use of dummy variables. These dummy variables take the value 1, if a certain condition holds and 0 if that is not the case.

$$y_i = \alpha_i + \sum_{j=2}^{k} \beta_j x_{ji} + \epsilon_i$$

Where α_i is the quarter-specific mean level. In this application we define dummy D_{hi} for each quarter, where h is 1 through 4 are the quarters. D_{hi} for $h = 1, 2, 3, 4$ with $D_{hi} = 1$ if observation i is in quarter h and $D_{hi} = 0$ otherwise.

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

With this notation, we obtain an equation much like before. We simply add the dummies to our X matrix, and use linear regression to get estimates of the quarter-specific constants alpha, as well of the parameters beta of the explanatory variables.

Can we add a constant term to this specification with dummies for each quarter? No, if we would add a constant and four quarterly dummies to our X matrix there would be linear dependence among the columns of X. Adding to four dummy variables gives exactly the intercept. So $(X'X)$ cannot be inverted. We can solve this, however, by simply taking out one of the dummies.

If we omit the first quarterly dummy DH_{1i} so that $\alpha_1 = 0$, this what the model becomes:

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

The model is equivalent to the model at the top of the slide, but the dummy coefficients have a different interpretation. As before, α_1 measures the mean level for the first quarter. In this specification the mean level of the second quarter is however given by $\gamma_2 + \alpha_1$. In the specification of the previous slide the mean level of the second quarter was given by alpha2.

We can thus easily relate the gammas and alphas to each other through the relationship that $\gamma_2 = \alpha_2 - \alpha_1$. Similar results hold for the third and the fourth quarter. $\gamma_h = \alpha_h - \alpha_1$ for $h = 1, 2, 3, 4$

Examples with SP500 dataset:

We have previously specified the model:

$$\Delta log(SP500index) = \beta_1 + \beta_2 BookMarket + \epsilon$$

Where we apllied two transformations to the SP500 variable, the log() and the first difference. These two transformations combined provide the interpretation of being an **(approximate) growth rate**.

Notice that traditionally a growth rate is calculated $\frac{y_i - y_{i-1}}{y_{i-1}} = \frac{\Delta y_i}{y_{i-1}}$

Remember the following rules for logarithms:

- $log(a) - log(b) = log(\frac{a}{b})$
- $log(\frac{a}{b}) = log(\frac{a}{b} + 1 - 1) = log(\frac{a}{b} + 1 - \frac{b}{b}) = log(1 + \frac{a-b}{b})$
- $log(1 + x) \approx x$ for small $x \to 0$

So the first difference can be seen as:

$$\Delta log(y_i) = log(y_i) - log(y_{i-1}) = log(\frac{y_i}{y_{i-1}}) = log(1 + \frac{y_i - y_{i-1}}{y_{i-1}})$$

$$\Delta log(y_i) = log(1 + \frac{\Delta y_i}{y_{i-1}}) \approx \frac{\Delta y_i}{y_{i-1}}$$

We now regress the change in the log of the S&P500 index on a constant, the book-to-market ratio, and the square of the book-to-market ratio.

$$\Delta log(SP500index) = \beta_1 + \beta_2 BookMarket + \beta_3 BookMarket^2 + \epsilon$$

To add the second order term we need to use the I() function in the model specification around our newly created predictor.

```
lm3 <- lm(dif_log_sp500 ~ BookMarket + I(BookMarket^2), data = dataset3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Jul 09, 2020 - 22:56:31

Tabla 3: Regression Results

	Dependent variable:
	dif_log_sp500
BookMarket	0.238
	(0.287)
I(BookMarket^2)	−0.347
	(0.213)
Constant	0.056
	(0.089)
Observations	86
R^2	0.109
Adjusted R^2	0.087
Residual Std. Error	0.189 (df = 83)
F Statistic	5.053^{***} (df = 2; 83)
Note:	*p<0.1; **p<0.05; ***p<0.01

We can see from the p-value of $I(BookMarket^2)$ that the coefficient is insignificant, thus the relationship is not quadratic.

Now we define a dummy that is 1 for 1980 and all following years using ifelse() base function within diplyr.

```
dataset3 <- dataset3 %>% mutate(D1980 = ifelse(year_orig >= 1980, 1, 0))
```

We now regress the change in the log of the S&P500 index on a constant, the book-to-market ratio, and an interaction between the book-to-market ratio and the just-defined dummy.

$$\Delta log(SP500index) = \beta_1 + \beta_2 BookMarket + \beta_3 BookMarket \cdot D1980 + \epsilon$$

```
lm4 <- lm(dif_log_sp500 ~ BookMarket + I(BookMarket*D1980), data = dataset3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Jul 09, 2020 - 22:56:32

Tabla 4: Regression Results

	Dependent variable:
	dif_log_sp500
BookMarket	-0.208^{**}
	(0.080)
I(BookMarket ∗D1980)	0.049
	(0.086)
Constant	0.166^{***}
	(0.054)
Observations	86
R^2	0.083
Adjusted R^2	0.061
Residual Std. Error	0.192 (df = 83)
F Statistic	3.776^{**} (df = 2; 83)
Note:	*p<0.1; **p<0.05; ***p<0.01

Is the relationship between the index and book-to-market stable over the pre and post 1980 period?

We can se from the result $\beta_3 = 0.048$ and is not statistically significant, therefore the relationship might be stable over the pre-post 1980 periods.

Evaluation of models

In this section, you will learn how to evaluate whether a model is actually a good model. Suppose you use the techniques from lectures How to specify? and Data Transformation, and obtained estimates for the parameters of some model. How to know whether the model is satisfactory? We will turn to a number of tests you can use to evaluate the model. In the last section, we started with a linear model and extended this to a non-linear model by adding square and interaction terms.

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

Suppose you want to test whether the linear model is good enough or that these extra terms should be added. A simple idea is to study the joint significance of the gamma coefficients on the squared and interaction

terms. The key challenge here is that this model contains many parameters. Here we have been even fairly modest by only considering squares, but of course more powers can be added which multiplies the number of parameters.

RESET

Fortunately, there is an easy way to reduce the number of parameters. We simply include powers of fitted y values based on the linear model, instead of the square and interaction terms.

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

$$y_i = x_i'\beta + \sum_{j=1}^{p} \gamma_j \hat{(y_i)}^{j+1} + \epsilon_i$$

$$\text{Correct linear specification : } H_0 : \gamma_j = 0 \forall j \text{ F-test(p,n-k-p)}$$

The test for non-linearity is then on the **joint significance of the gammas in this model**. The test here is written general, with p powers and thus p gamma coefficients. Under the null of a correct linear specification, the gammas are 0, and the test is an F-test. The number of restrictions are p, and the total number of parameters in the unrestricted model k+p, such that the degrees of freedom are p and n-k-p. **The F distribution is however approximate**, as the y hat is not a usual fixed regressor.

The test is called **RESET which stands for Regression Specification Error Test**. Strictly speaking the null is that of correct specification, which is more general than simply the null of linearity. For this reason the test is a general mis-specification test which the name RESET also alludes too.

- Notice for $p = 1$, we only have the k usual parameters plus one p extra. So in total, $(k + 1)$ parameters are to be estimated.
- In contrast, in the previous model with squares and cross-terms we would get the usual k β parameters, the k-1 squared terms, (note the square of an intercept is simply the intercept) and a number of interactions.So in total $k + (k - 1) + \frac{1}{2}(k - 2)(k - 1)$ parameters are to be estimated.

Chow Break Test

Now we turn to two tests that are both based on the idea that there is some possible break in the sample, with which the full sample can be split in two groups, one before and one after the break.

We write a model for the first and a model for the second group. We write n_1 for the number of observations in the first group, and n_2 for the number of observations in the second group.

$$y_1 = X_1\beta_1 + \epsilon_1 : n_1 \text{ observations}$$

$$y_2 = X_2\beta_2 + \epsilon_2 : n_2 = n - n_1 \text{ observations}$$

In both groups, we have similar models, and the only difference is that the parameter beta changes from $beta_1$ to $beta_2$. These two models can be written in one framework using vector and matrix notation. We stack the y_1 and y_2 vectors, make a block structure of X_1 and X_2 to get a new larger X matrix, and also stack the β and disturbance vectors.

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

$$\text{No Break : } H_0 : \beta_1 = \beta_2 \text{ such that } \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \beta + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

The idea of the Chow break test is that we test a restricted set-up, where $\beta_1 = \beta_2$, against this unrestricted setup. Under the null of no break, this is an F-test as follows:

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

As usual, e denote residuals, and the subscript R stands for the residuals from the restricted model, and U for the residuals of the unrestricted model. The degrees of freedom are k, the number of imposed restrictions in the restricted model, and $n-2k$, which is the number of observations minus the total number of parameters in the unrestricted model.

In this particular case it turns out that the unrestricted residuals can be split into two groups, the residuals from the first group and the residuals from the second group. In fact, the residuals from the first group are based on only data for the first group and similarly for the second group.

So we have $e_U = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$ thus $e_U' e_U = e_1' e_1 + e_2' e_2 = S_1 + S_2$. See the proof at the end of the section in Proofs for Chow tests

We can then express the F test as follows, where we've written S_0 for the sum of squared residuals in the restricted model $S_0 = e_R' e_R$:

$$F = \frac{(S_0 - S_1 - S_2)/k}{S_1 + S_2/(n-2k)} \sim F_{(k,n-2k)}$$

The Chow break test assumes that only the parameter vector beta changes across the two samples, but the rest of the model structure remains the same.

Chow forecast test

The second break test is a variant of the Chow break test and relaxes this assumption. The test equation is as follows:

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

$$\text{Constant structure : } H_0 : \gamma_j = 0 \forall j$$

The sum runs over n_2 elements. The dummy $D_{ji} = 1$ if $i = j$ and 0 else. There is thus exactly one dummy for each of the n_2 observations in group 2. In total there are the usual k parameters in the vector beta plus n_2 gamma parameters that we have to estimate.

Because of these dummies, the fit in the second sample will be perfect. The residuals for all observations i in the second group are equal to 0 as any deviation of $x_i'\beta$ from y_i is already captured with γ_i. Thus $e_2 = 0$

Compared to the Chow break test, the S_2 term drops out as the second sum of squared residuals is equal to 0 $e_2 = 0$. The number of restrictions imposed in the restricted model is n_2, as all the gammas are set equal to 0 $\gamma_j = 0 \forall j$. The degrees of freedom in the denominator is equal to the total number of observations $n = n_1 + n_2$ minus the total number of parameters in the unrestricted model, which is $n_2 + k$ Thus the denominator degrees of freedom is $n_1 - k$.

The F-test for the joint significance of all the gammas then simplifies to the following expression. $H_0 : \gamma = 0$

$$F = \frac{(S_0 - S_1)/n_2}{S_1/(n_1 - k)} \sim F_{(n_2,n_1-k)}$$

If the **test statistic is large**, the second group of observations does not fit the pattern from the first group of observations well and we **reject the null of constant module structure.**

The interpretation is that the test examines whether the relationship in the first sample can be used to forecast the relationship in the second sample, hence the name of the Chow forecast test.

We always do our very best to specify a good model, but of course this is not always easy. We should always perform checks on the chosen model specification, for example, by studying the residuals.

Jarque-Bera

We often assume, for example, in the t and the F-tests that the disturbances are normally distributed. We can test the validity of this assumption by studying the distribution of the residuals. Ideally, this distribution should resemble the nice bell shaped curve of the normal distribution, which is symmetric and does not have thick tails.

The test for normality is based on the third and fourth moments, which are skewness S and kurtosis K that were discussed in the building blocks. If the skewness and kurtosis of the residuals differ too much from those of the normal distribution, which are zero and three respectively, we reject the null that the disturbances are normally distributed.

The Jarque-Bera test is based on this idea:

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

$$\text{If null holds : } H_0 : \epsilon_i \sim NID(0,\sigma^2) \Rightarrow JB \sim \chi^2(2)$$

If normality is rejected, further inspection of the model is typically required.

Proofs for Chow tests

1. **Chow Break Test**. Prove that the vector of residuals from the unrestricted model $e_U = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$ thus $e_U' e_U = e_1' e_1 + e_2' e_2 = S_1 + S_2$. . Show that this is equivalent to calculating the sum of squared residuals for a regression for only the first sample, thus a regression of $y_1 \sim X_1$ plus the sum of squared residuals for a regression for only the second sample, thus a regression of $y_2 \sim X_2$

$$e_U = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} y_1 - X_1 b_1 \\ y_2 - X_2 b_2 \end{pmatrix}$$

$$e_U = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} y_1 - X_1 b_1 \\ y_2 - X_2 b_2 \end{pmatrix}$$

The vector of coefficients can be rewritten as:

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \left(\begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}' \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}' \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

We can use the fact that $\begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix}^{-1} = \begin{pmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{pmatrix}$ recall that $X_i'X_i$ is symmetric and the properties of the inverse of diagonal matrices.

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1'y_1 \\ X_2'y_2 \end{pmatrix} = \begin{pmatrix} (X_1'X_1)^{-1}X_1'y_1 \\ (X_2'X_2)^{-1}X_2'y_2 \end{pmatrix}$$

We can see that $b_i = (X_i'X_i)^{-1}X_i'y_i$ the coefficients are identical to the ones obtained in the regression respectively on group 1 and group 2 observations.

Is because of this that we can express $e_U = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$ and $e_U' e_U = e_1' e_1 + e_2' e_2 = S_1 + S_2$ equal the sum squared residuals of both models.

2. **Chow Forecast Test**. First we write the Chow forecast model $y_i = x_i'\beta + \sum_{j=n_1+1}^{n_1+n_2} \gamma_j D_{ji} + \epsilon_i$ in matrix form:

For the first n_1 observations we have:

- $y_1 = X_1\beta_1 + \epsilon_1$ **(Model 1)**

and for the last n_2 observations we have

- $y_2 = X_2\beta_2 + D\gamma + \epsilon_2$ **(Model 2)** with $D = I_{n_2}$, $(n_2 x n_2)$ identity matrix.

We combine these two to derive:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1\beta_1 + \epsilon_1 \\ X_2\beta_2 + D\gamma + \epsilon_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ X_2 & D \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

The test for the null $H_0 : \gamma = 0$ has the following F test:

$$F = \frac{(e_R'e_R - e_U'e_U)/n_2}{e_U'e_U/(n_1 + n_2 - (k - n_2))} = \frac{(e_R'e_R - e_U'e_U)/n_2}{e_U'e_U/(n_1 - k)} \sim F_{(k,n-2k)}$$

We denote the sum of square residuals when restricted model **(Model 1)** is applied to all observations as $S_0 = e_R'e_R$. Under the unrestricted alternative **(Model 2)** the sum of square residuals is obtained by the following minimization problem:

$$Min_{\beta,\gamma} \begin{pmatrix} y_1 - X_1\beta_1 \\ y_2 - X_2\beta_2 - D\gamma \end{pmatrix}' \begin{pmatrix} y_1 - X_1\beta_1 \\ y_2 - X_2\beta_2 - D\gamma \end{pmatrix}$$

That yields: $\hat{\beta} = (X_1'X_1)^{-1}Xy$, $\hat{\gamma} = y_2 - \alpha_2\hat{\beta}$.

This results that the residuals for the first n_1 observations $e_1 = y_1 - X_1\hat{\beta}$ and for last n_2 observations $e_2 = y_2 - X_2\hat{\beta} - D\hat{\gamma} = 0$. Therefore $e_U'e_U = e_1'e_1 = S_1$ and we have the following F-test: $F = \frac{(S_0 - S_1)/n_2}{S_1/(n_1-k)} \sim F_{(n_2,n_1-k)}$

Application on SP500

Datset:

This is a stock market data set for the United States for 1927-2013 (yearly data). The source of the data is the updated version of the Goyal and Welch (2008)1 data. The data are available from the website of Prof Amit Goyal

The variables are:

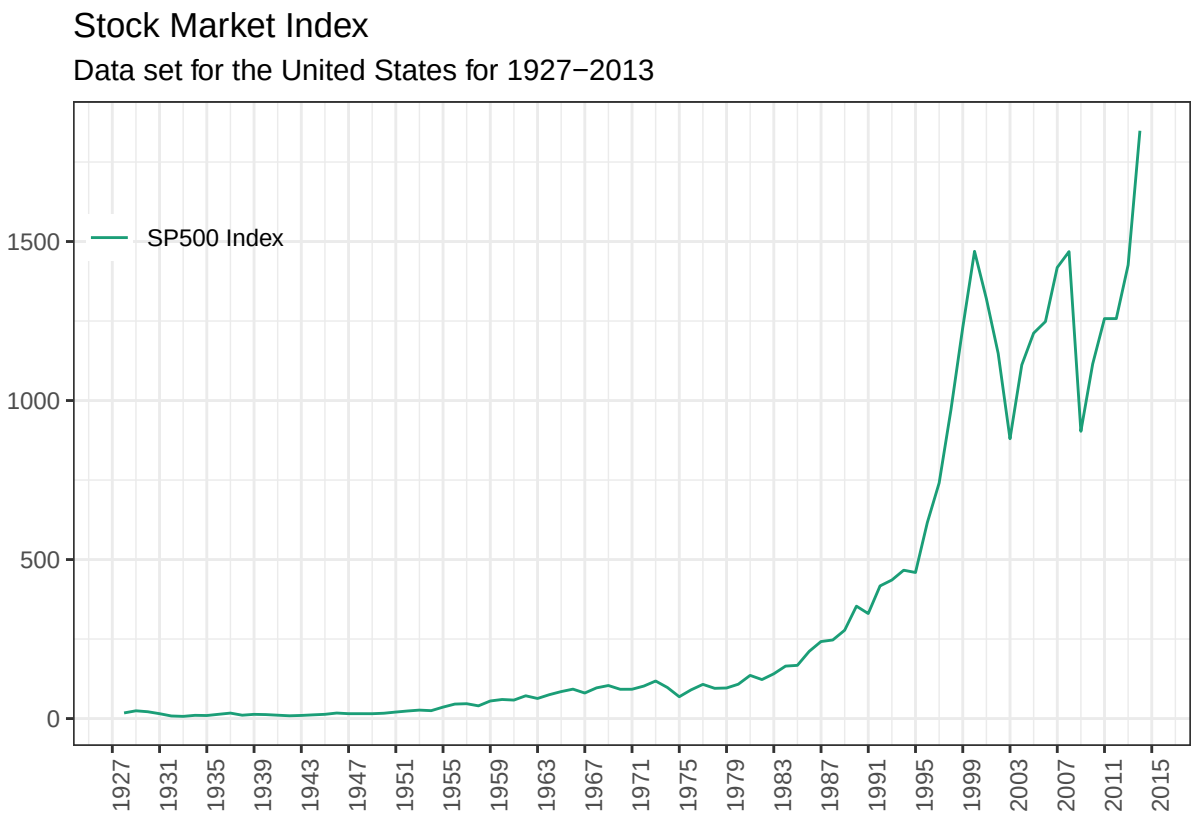
- **Year**
- **Index**: The S&P500 index
- **Dividends**: Dividends on the index ("D12" in the Goyal and Welch [GW] file)
- **Riskfree**: Riskfree rate ("Rfree" in GW)
- **LogEqPrem**: Log of the equity premium (calculated following GW) Calculated as: $\frac{(Index+D12)}{Index(-1)} - log(1 + Rfree)$, where $x(-1)$ denotes value from previous period, log is the natural logarithm, $D12$ dividends and $Rfree$ the riskfree rate.
- **BookMarket**: Book to market ratio ("b/m" in GW)
- **NTIS**: Equity issued ("ntis" in GW)
- **DivPrice**: Dividend to price ratio (calculated following GW) Calculated as: $log(D12) - log(Index)$, where $D12$ are dividends.
- **EarnPrice**: Earnings to price ratio (calculated following GW) Calculated as: $log(E12)/log(Index)$, where $E12$ are earnings.
- **Inflation**: Inflation rate ("infl" in GW)

The application that we consider is how to model the stock market index. A first question we turn to is whether the index series should be transformed. Then, we consider a number of explanatory variables and decide which ones we actually include in our model. Finally, we compare a set of candidate models and study whether the relationship is stable over time.

Variable transformation

Let's start with the transformation. Here is the S&P 500 index again, annual data of the period 1927 through 2013.

```
dataset3 %>% ggplot(aes(x=Year)) +
  geom_line(aes(y=Index, col = "SP500 Index")) +
  labs(x = "", y = "", title = "Stock Market Index",
       subtitle = ("Data set for the United States for 1927-2013")) +
    scale_x_date(date_breaks = "4 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.1, .8),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")
```

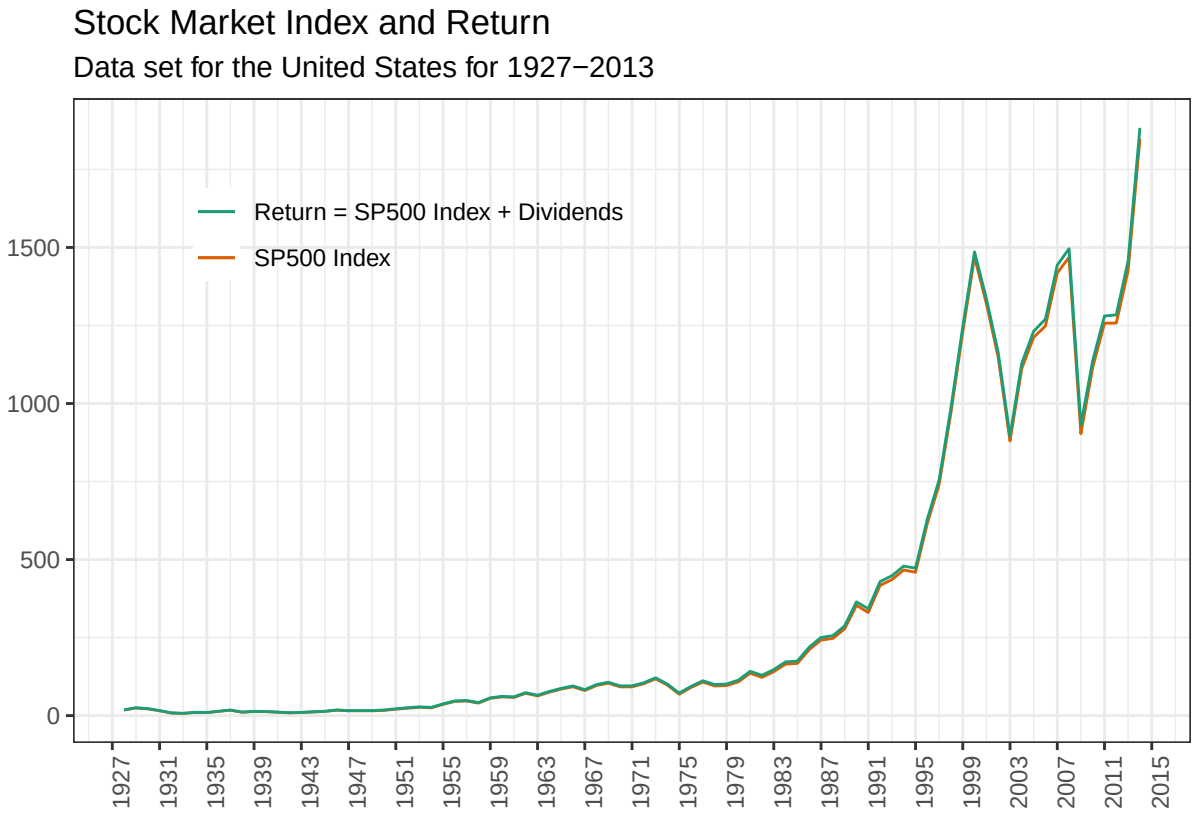


Now, we of course also have to think carefully about the economic setting. Rather than modeling the stock index directly, or some appropriate transformed version, we consider **how much the stock market index earns in total**, on top of simply putting money in a risk-free asset. This difference tells us how high the total reward is for taking on the risk of the stock market.

First, to consider all gains from holding stock, we add dividends to the stock market index, as these also form an important part of the income of holding stocks. The green line gives the index including dividends, and here, we see some instability.

$$\text{Index} + \text{Dividends} = Index_i + D12_i$$

```
dataset3 %>% ggplot(aes(x=Year)) +
  geom_line(aes(y=Index, col = "SP500 Index")) +
  geom_line(aes(y=Index+Dividends, col = "Return = SP500 Index + Dividends")) +
  labs(x = "", y = "", title = "Stock Market Index and Return",
       subtitle = ("Data set for the United States for 1927-2013")) +
    scale_x_date(date_breaks = "4 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.3, .8),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")
```



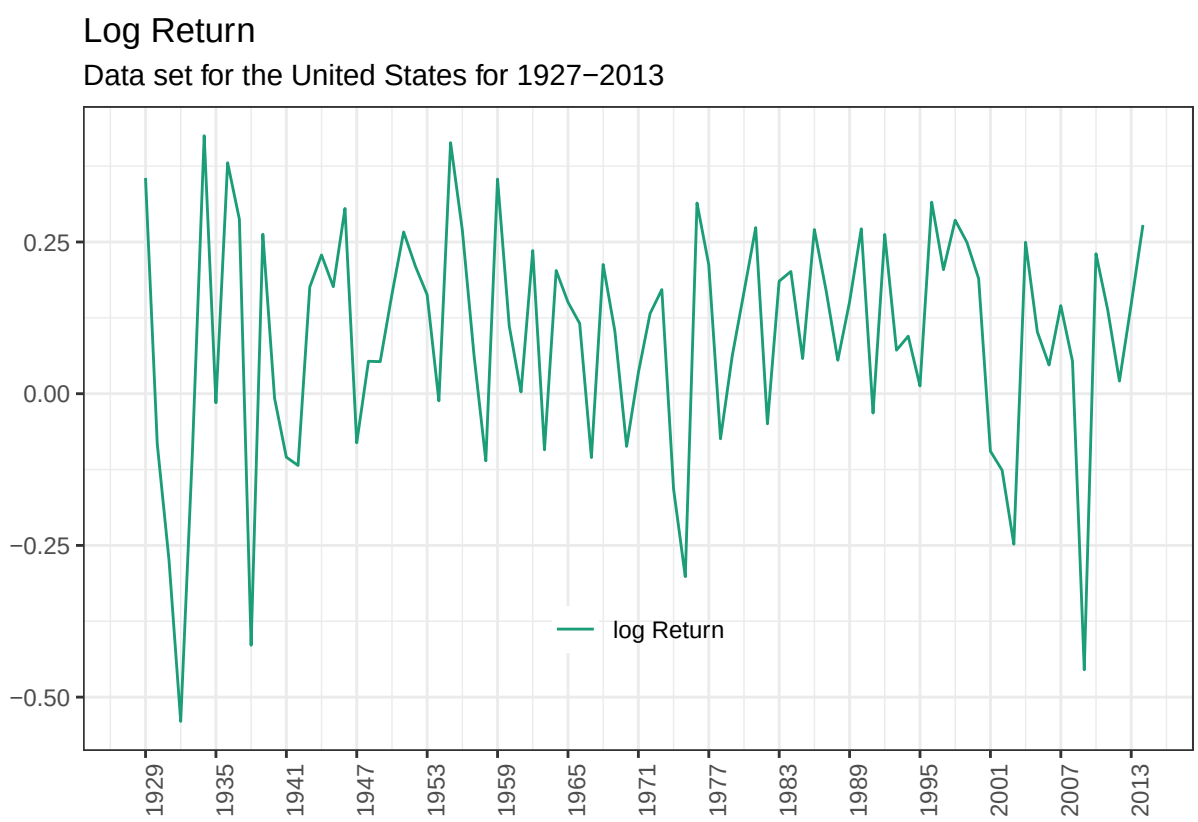
We take the log to undo the exponential growth, and consider the difference of the log index to take out the trend in the log index. It turns out that the combination of these transformations gives a series that is **approximately equal to a growth rate**. The green line plus the series, with the values on the right axis.

$$\text{log Return} = log((Index_i + D12_i)/Index_{i-1})$$

```
dataset3 <- dataset3 %>% mutate(log_Ret=log((Index+Dividends)/lag(Index)))
plot_a <- ggplot(data=dataset3, aes(x=Year)) +
  geom_line(aes(y=log_Ret, col = "log Return")) +
  labs(x = "", y = "", title = "Log Return",
       subtitle = ("Data set for the United States for 1927-2013")) +
    scale_x_date(date_breaks = "6 year", date_labels = "%Y") +
  theme_bw() +
```

```
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.5, .20),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")

plot_a
```

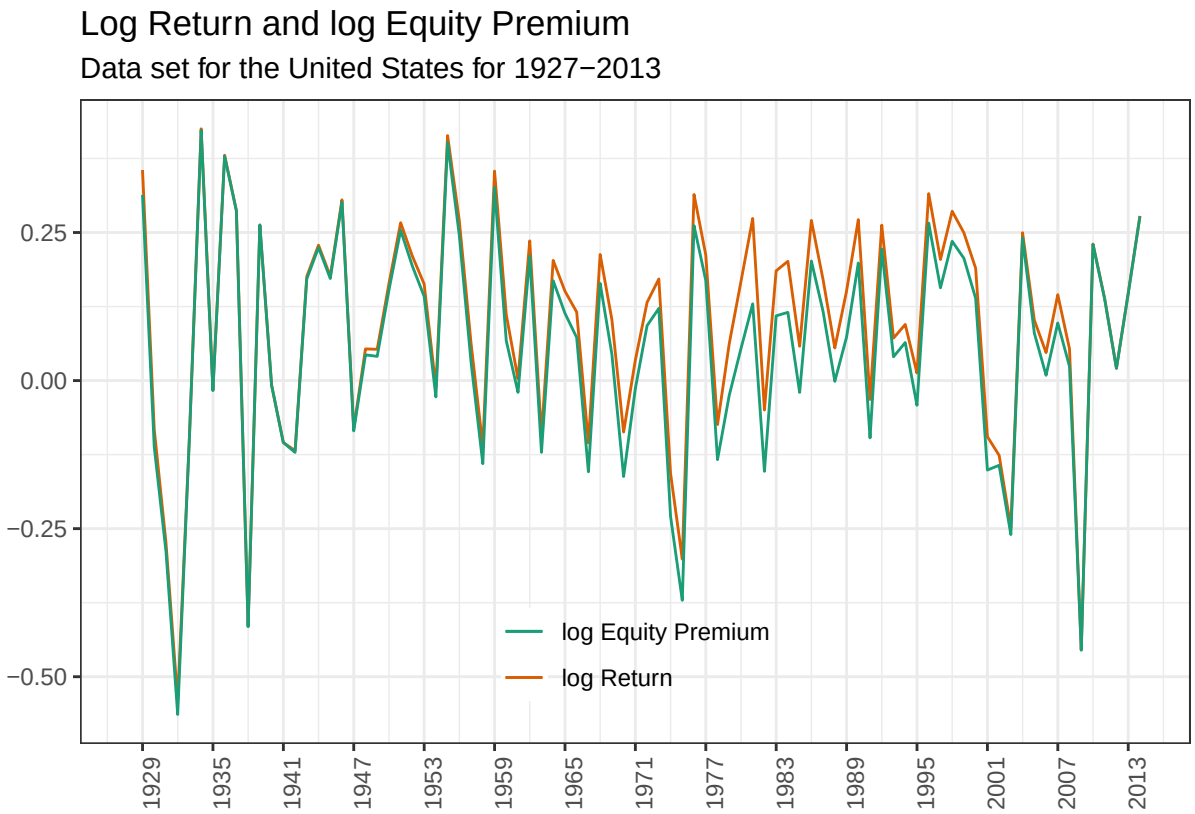


In terms of econometrics, this is already a series we can work with. For the economic setting, we also subtract the risk-free rate, for which we take the treasury bill rate, the return on short-term government bonds.

$$\text{log Equity Premium} = log((Index_i + D12_i)/Index_{i-1}) - log(1 + Rfree)$$

```
dataset3 <- dataset3 %>% mutate(log_Equity_Prem = log_Ret - log(1+Riskfree))
plot_b <- ggplot(data=dataset3, aes(x=Year)) +
  geom_line(aes(y=log_Ret, col = "log Return")) +
  geom_line(aes(y=log_Equity_Prem, col = "log Equity Premium")) +
  labs(x = "", y = "", title = "Log Return and log Equity Premium",
       subtitle = ("Data set for the United States for 1927-2013")) +
    scale_x_date(date_breaks = "6 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.5, .15),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")

plot_b
```



The green line is the **log Equity Premium**, which is the series we actually model in our analyses, and represents the extra reward for investing in stock, relative to putting money in safe assets.

Testing Specification

In the analysis, we consider only five of the many explanatory variables available. These five are:

1. Book-to-market ratio
2. A net equity expansion variable issued stock that measures how much stock is issued
3. Dividends relative to prices
4. Earnings relative to prices
5. Inflation.

If you have no feeling for the field of finance and do not understand the precise motivation for these variables, that is fine. Just treat them as X's we use to model a certain y, and focus on the approach.

Just to get going, and as we only have five explanatory variables, we run five separate simple regressions. In each of these regressions, we regress the log equity premium on one of the variables. Each of the columns in the table, labeled from 1 to 5, provides the output for one of these simple regressions.

```
lm1 <- lm(LogEqPrem ~ BookMarket, data = dataset3)
lm2 <- lm(LogEqPrem ~ NTIS, data = dataset3)
lm3 <- lm(LogEqPrem ~ DivPrice, data = dataset3)
lm4 <- lm(LogEqPrem ~ EarnPrice, data = dataset3)
lm5 <- lm(LogEqPrem ~ Inflation, data = dataset3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Jul 09, 2020 - 22:56:35

Tabla 5: Regression Results

	Dependent variable:				
	LogEqPrem				
	(1)	(2)	(3)	(4)	(5)
Book to Market	-0.185^{**}				
	(0.077)				
Issued Stock		-0.148			
		(0.771)			
Dividend/Price			-0.097^{**}		
			(0.044)		
Earnings/Price				-0.032	
				(0.051)	
Inlfation					-0.167
					(0.511)
Constant	0.166^{***}	0.062^{**}	-0.266^{*}	-0.027	0.065^{**}
	(0.049)	(0.025)	(0.148)	(0.140)	(0.026)
Observations	87	87	87	87	87
R^2	0.063	0.0004	0.055	0.005	0.001
Adjusted R^2	0.052	-0.011	0.044	-0.007	-0.011
Residual Std. Error (df = 85)	0.188	0.194	0.188	0.193	0.194
F Statistic (df = 1; 85)	5.763^{**}	0.037	4.938^{**}	0.395	0.106
Note:					*p<0.1; **p<0.05; ***p<0.01

From this table, you can see that both the book-to-market and the dividend/price ratio are significant at the 5% level for the log equity premium. Also the R-squareds are highest for these two regressions.

General-to-specific

To develop a model for the log equity premium, we apply the general-to-specific approach. In column one, the output is given for the regression of the log equity premium on all variables. For all variables, we inspect whether they are significant, and if there are insignificant variables, we eliminate the variable with the highest p-value. In this case, the stock issued has the highest p-value, so we drop it and run a second regression using all variables except for this variable.

We follow the same logic, and again drop the non-significant variable with the highest p-value. In this case, this is inflation. Note, the constant is also insignificant, with an even higher p-value, but we do prefer to keep this in the model. A reason for this is that the variables are not demeaned, and we need to ensure that the disturbance term has mean zero.

In the third regression the dividend/price ratio is the non-significant variable with the highest p-value. This is quite interesting, because it did give us significance in the simple regression setting. Apparently, the dividend/price ratio has limited explanatory power for the log equity premium, when controlling for book-to-market and earnings/price effects.

The fourth regression considers only a constant, book-to-market and the earnings/price ratio. It turns out this latter variable is insignificant and also has to be dropped in the general-to-specific approach.

Our final model is a simple regression model with only book-to-market.

```
lm1 <- lm(LogEqPrem ~ BookMarket + NTIS + DivPrice + EarnPrice + Inflation, data = dataset3)
lm2 <- lm(LogEqPrem ~ BookMarket + DivPrice + EarnPrice + Inflation, data = dataset3)
lm3 <- lm(LogEqPrem ~ BookMarket + DivPrice + EarnPrice, data = dataset3)
lm4 <- lm(LogEqPrem ~ BookMarket + EarnPrice, data = dataset3)
lm5 <- lm(LogEqPrem ~ BookMarket, data = dataset3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Jul 09, 2020 - 22:56:36

Stability

Now we evaluate this model in various ways. First, we check the stability of this relationship.

As an example, we consider the stability during two important periods, the Second World War during 1939 up to 1945, and the oil crisis during 1973 up to 1975.

$$log(EqPr)_i = \beta_1 + \beta_2 BTM_i + \beta_3 BTM_i \cdot DummyWar_i + \beta_4 BTM_i \cdot DummyOil_i + \epsilon_i$$

```
dataset3 <- dataset3%>% mutate(war_dummy = ifelse(year_orig >= 1939 & year_orig <= 1945, 1, 0),oil_dummy
```

The table shows the results, including two extra coefficients for the interaction of the book-to-market value with the war-dummy, and the interaction with the oil-dummy.

```
lm6 <- lm(LogEqPrem ~ BookMarket + I(BookMarket*war_dummy) + I(BookMarket*oil_dummy), data = dataset3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Jul 09, 2020 - 22:56:36

The p-value of both the war and oil-dummy interaction term are not significant, so the relationship does not differ significantly during these periods.

Tabla 6: Regression Results

	Dependent variable:				
	LogEqPrem				
	(1)	(2)	(3)	(4)	
Book to Market	−0.177	−0.166	−0.191	−0.290***	-
	(0.154)	(0.143)	(0.141)	(0.107)	
Issued Stock	−0.150				
	(0.818)				
Dividend/Price	−0.120	−0.126	−0.090		
	(0.098)	(0.092)	(0.084)		
Earnings/Price	0.167*	0.167*	0.128*	0.097	
	(0.085)	(0.084)	(0.074)	(0.068)	
Inlfation	−0.569	−0.567			
	(0.587)	(0.583)			
Constant	0.235	0.205	0.214	0.490**	0
	(0.385)	(0.346)	(0.346)	(0.233)	
Observations	87	87	87	87	
R^2	0.109	0.108	0.098	0.086	
Adjusted R^2	0.054	0.065	0.065	0.064	
Residual Std. Error	0.188 (df = 81)	0.186 (df = 82)	0.186 (df = 83)	0.187 (df = 84)	0.18
F Statistic	1.977* (df = 5; 81)	2.492** (df = 4; 82)	3.009** (df = 3; 83)	3.927** (df = 2; 84)	5.763*
Note:					*p<0.1; **p<0.0

Tabla 7: Regression Results

	Dependent variable:
	LogEqPrem
BookMarket	−0.175**
	(0.082)
I(BookMarket ∗war_dummy)	0.078
	(0.101)
I(BookMarket ∗oil_dummy)	−0.133
	(0.124)
Constant	0.160***
	(0.050)
Observations	87
R^2	0.085
Adjusted R^2	0.052
Residual Std. Error	0.188 (df = 83)
F Statistic	2.580* (df = 3; 83)
Note:	*p<0.1; **p<0.05; ***p<0.01

Testing Information criteria

We can compare the model we obtained (lm5) to the full model (lm1), including all considered variables. While the R squared is higher for the full model, our analyses show that most of the other explanatory variables did not carry significant explanatory power.

Recall our two information criteria:$AIC = log(s^2) + \frac{2k}{n}$ and $BIC = log(s^2) + \frac{klog(n)}{n}$ where s^2 is the standard error of the regression s and k the number of parameters.

```
Rsqr_lm1 <- summary(lm1)$r.squared
Rsqr_lm5 <- summary(lm5)$r.squared
#AIC_lm1 <- AIC(lm1)
#BIC_lm1 <- BIC(lm1)
#AIC_lm5 <- AIC(lm5)
#BIC_lm5 <- BIC(lm5)
AIC_lm1 <- log(sqrt(deviance(lm1)/df.residual(lm1))^2) + (2*6)/nobs(lm1)
BIC_lm1 <- log(sqrt(deviance(lm1)/df.residual(lm1))^2) + (6*log(nobs(lm1)))/nobs(lm1)
AIC_lm5 <- log(sqrt(deviance(lm5)/df.residual(lm5))^2) + (2*2)/nobs(lm5)
BIC_lm5 <- log(sqrt(deviance(lm5)/df.residual(lm5))^2) + (2*log(nobs(lm5)))/nobs(lm5)

info <- matrix(c(Rsqr_lm1,Rsqr_lm5,AIC_lm1,AIC_lm5,BIC_lm1,BIC_lm5),nrow = 3, byrow = T)
colnames(info) <- c("Full Model","Book Market")
rownames(info) <- c("Rsqr","AIC","BIC")
kable(info,booktabs = TRUE, digits = 3) %>%
  kable_styling()
```

	Full Model	Book Market
Rsqr	0.109	0.063
AIC	-3.210	-3.301
BIC	-3.040	-3.244

The Akaike and Bayesian information criteria confirm this. The lowest AIC and BIC values are indeed obtained for the book-to-market model, confirming this is the preferred approach.

RESET

We use the function resettest

```
# resettest(formula, power = 2:3, type = c("fitted", "regressor","princomp"), data = list())
reset_lm5 <- resettest(lm5,power = 2, type = "fitted")
reset_lm5
```

```
##
##  RESET test
##
## data:  lm5
## RESET = 3.4563, df1 = 1, df2 = 84, p-value = 0.06651
```

The null hyp is not rejected H_0 : the model is a linear regression model.

Otherwise we could follow the following process:

1. Regress the log equity premium on a constant and the book-to-market ratio. $e_0'e_0 = 2.992$
2. Store the fitted log equity premium based on the output from this regression.
3. Regress the log equity premium on a constant, the book-to-market ratio, and the square of the fitted log equity premium that was stored in the previous step. $e_1'e_1 = 2.992$
4. The RESET test statistic is the statistic of an F-test on the fitted log equity premium parameter. $F = \frac{(e_0'e_0 - e_1'e_1)/g}{(e_1'e_1)/(n-k)} = 3.4563$

Chow Break

We can follow the following process:

1. Regress the log equity premium on a constant and the book-to-market ratio and store the sum of squared residuals. $S_0 = 2.992$
2. Then perform the same regression for both the subsample of observations over 1927-1979 $S_1 = 1.981$, and the subsample of observations over 1980-2013, $S_12 = 0.855$
3. For both regressions, store the sum of squared residuals.
4. Use these sum of squared residuals to calculate the Chow break statistic. $F = \frac{(S_0 - S_1)/n_2}{S_1/(n_1 - k)}$

```
# For Chow test we split our dataset:
dataset <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/dataset3.csv")
dataset_1 <- dataset %>% filter(Year < 1980) # We have 53 obs
dataset_2 <- dataset %>% filter(Year >= 1980) # We have 34 obs
y1 <- unlist(dataset_1 %>% dplyr::select(LogEqPrem),use.names = FALSE)
y2 <- unlist(dataset_2 %>% dplyr::select(LogEqPrem),use.names = FALSE)
x1 <- unlist(dataset_1 %>% dplyr::select(BookMarket),use.names = FALSE)
x2 <- unlist(dataset_2 %>% dplyr::select(BookMarket),use.names = FALSE)
```

Now we use the function chow.test

```
chowbreak <- chow.test(y1,x1,y2,x2)
chowbreak
```

```
##    F value      d.f.1      d.f.2    P value
##  2.2708835  2.0000000 83.0000000  0.1095987
```

Again, the Null Hyp is not rejected, the model parameters do not suffer from structural break.

Chow Forecast

For the Chow's forecast test there is no available library in R. So we calculate the F-test as follows:

1. Estimate the OLS vector from the first n_1 observations, obtaining $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y_1$, the vector of residuals $\hat{e}_1 = y_1 - X_1\beta_1$ and $S_1 = \hat{e}_1'\hat{e}_1$

```
x1 <- matrix(c(rep(1,length(x1)),x1),nrow = length(x1), byrow = F)
x2 <- matrix(c(rep(1,length(x2)),x2),nrow = length(x2), byrow = F)
y1 <- matrix(y1,nrow = length(y1), byrow = F)
y2 <- matrix(y2,nrow = length(y2), byrow = F)
beta_1 <- (solve(t(x1)%*%x1))%*%(t(x1)%*%y1)
e_1 <- y1-x1%*%beta_1
sse_1 <- t(e_1)%*%e_1
```

2. Fit the same e regression to all $N = n_1 + n_2$ observations and obtain the restricted $S_0 = \hat{e}'\hat{e}$.

```
y <- rbind(y1,y2)
x <- rbind(x1,x2)
beta_0 <- (solve(t(x)%*%x))%*%(t(x)%*%y)
e <- y-x%*%beta_0
sse_0 <- t(e)%*%e
```

3. Employt the F-test : $F = \frac{(S_0 - S_1)/n_2}{S_1/(n_1-k)} \sim F_{(n_2,n_1-k)}$ with $n_1 = 53$ and $n_2 = 34$.

```
chowfore <- ((sse_0-sse_1)/34)/(sse_1/(53-2))
# We check if the statistic falls within the interval at 99%
#(T We accept H0: Constant module structure) (F We reject H0)
c_R=qf(0.95,34,(53-2))
Ho_R=chowfore<c_R
chowfore
```

```
##          [,1]
## [1,] 0.765064
```

```
Ho_R
```

```
##      [,1]
## [1,] TRUE
```

We do not reject the Null Hyp H_0 : There is no structural change in the prediction parameters.

Normality test

We apply the Jarque Bera on the residuals of lm_5:

```
jb <- jarque.test(resid(lm5))
jb
```

```
##
##  Jarque-Bera Normality Test
##
```

```
## data:  resid(lm5)
## JB = 7.1616, p-value = 0.02785
## alternative hypothesis: greater
```

```
tests <- matrix(c(reset_lm5$statistic,reset_lm5$p.value,chowbreak[1],chowbreak[4],chowfore,chowfore,jb$
colnames(tests) <- c("Test Statistic","p-value")
rownames(tests) <- c("Reset(p=1)","Chow Break","Chow Forecast","Jarque-Bera")
kable(tests,booktabs = TRUE, digits = 3) %>%
  kable_styling() %>%
  footnote(general = "As break-point 1980 is chosen.")
```

	Test Statistic	p-value
Reset(p=1)	3.456	0.067
Chow Break	2.271	0.110
Chow Forecast	0.765	0.765
Jarque-Bera	7.162	0.028

Note:
As break-point 1980 is chosen.

The model does fairly well. Reset with p=1 does not reject the null of correct specification, and both Chow tests do not reject the null of no breaks. Only Jarque-Bera seem somewhat doubtful, as at 5%, we reject normality of the residuals. This may hint to some remaining specification problems.

Will the p-values of these tests increase if the full model is considered?
This is actually a tough question and the answer is not trivial. If the book-to-market model is correct and actually generated the data, we would add insignificant variables to the model. The p-values should be similar but may differ slightly, simply because of the added variance in the results. If the full model is correct, the p-values should increase when considering the full model. If neither the full nor the book-to-market model is correct, it is not clear what will happen to the p-values. This is the challenge you face when doing applied work. It is never certain how the data are actually generated. What we have are tests to rely on and help inform us if what we are doing makes sense.

Take a look at the other materials available in my website