Introduction to Time Series Analysis of Macroeconomic- and Financial-Data

Felix Pretis

Programme for Economic Modelling
Oxford Martin School, University of Oxford

Lecture 2: Testing & Dependence over time

Yesterday:

- Econometric Models
- Ordinary Least-Squares (OLS) Regression
  - Line of best fit
  - Hypothesis testing
  - Goodness of fit
  - More than one explanatory variable

Today:

- Is the model well-specified?
- Time-series: dependence over time!

- Extend model to include **two explanatory variables**: X_t and Z_t.
- Triplets of data, (Y_t, X_t, Z_t): Three economic variables occurring simultaneously.
- Assumptions:
  - **(i)** (Y_t, X_t, Z_t) **independent** across t.
  - **(ii)** Identical conditional distribution:
    $(Y_t|X_t, Z_t) \sim (\beta_1 + \beta_2 X_t + \beta_3 Z_t, \sigma^2)$.
  - **(iii)** X_t and Z_t **exogenously** determined for Y_t.
  - **(iv)** A parameter space exists: $\beta_1, \beta_2, \beta_3, \sigma^2 \in \mathbb{R}^{\nVdash} \times \mathbb{R}^+$,
- Gives model:

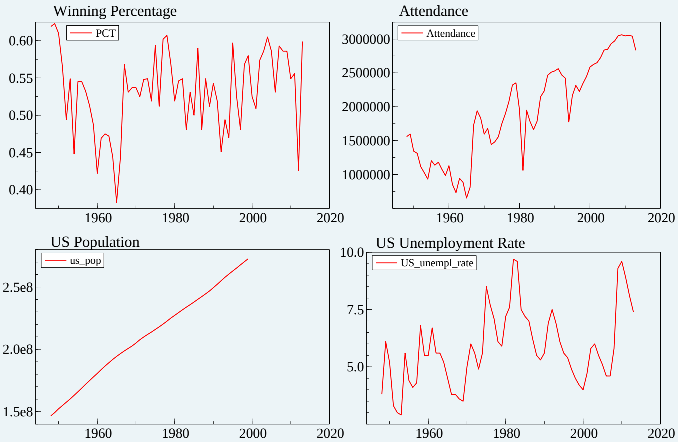
$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Z_t + \epsilon_t, \epsilon_t \sim \mathbb{N}0, \sigma^2 \tag{1}$$

- Parameter interpretation as before except **ceteris paribus**:

$$\frac{\partial \mathbb{E} Y_t | X_t, Z_t}{\partial X_t} = \beta_2, \tag{2}$$

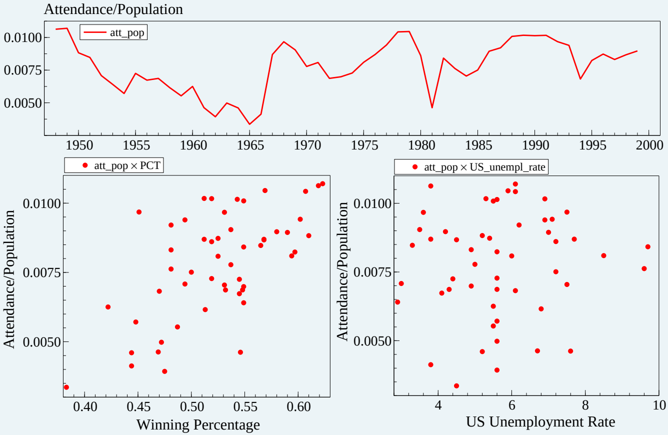
Baseball Attendance

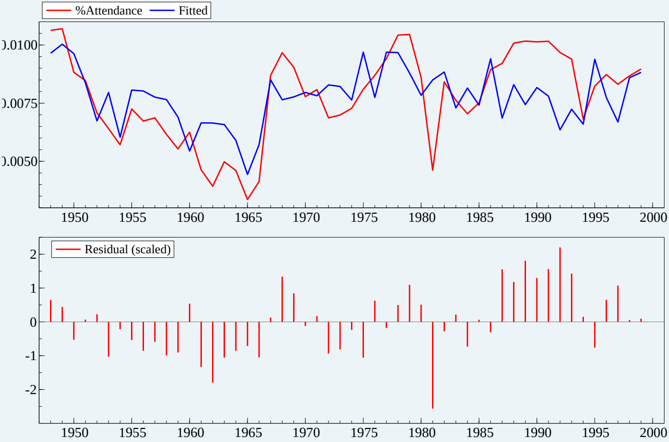
- Red Sox Game Attendance 1948-1999
- Winning Percentage (PCT)
- US Populations - to scale Attendance (att_pop)
- US Unemployment Rate



OXFORD MARTIN SCHOOL
UNIVERSITY OF OXFORD
Institute for New Economic Thinking AT THE OXFORD MARTIN SCHOOL

Model: $\%\text{Attendance}_t = \beta_1 + \beta_2 \text{PCT}_t + \beta_3 \text{Unemp.}_t + \epsilon_t$




%Attendance
Fitted
).0100
).0075
).0050
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
Residual (scaled)
2
1
0
-1
-2

```
EQ( 1) Modelling att_pop by OLS
       The dataset is: red_sox_1948.in7
       The estimation sample is: 1948 - 1999

                  Coefficient  Std.Error  t-value  t-prob Part.R^2
Constant          -0.00475283   0.002263    -2.10  0.0409   0.0826
PCT                 0.0224757   0.003977     5.65  0.0000   0.3946
US_unempl_rate    0.000129058  0.0001368    0.943  0.3502   0.0178

sigma              0.00151203  RSS              0.000112024814
R^2                  0.398903  F(2,49) =     16.26 [0.000]**
Adj.R^2              0.374369  log-likelihood          265.464
no. of observations        52  no. of parameters             3
mean(att_pop)      0.00779052  se(att_pop)          0.00191161
When the log-likelihood constant is NOT included:
AIC                  -12.9326  SC                     -12.8201
HQ                   -12.8895  FPE                 2.41812e-006
When the log-likelihood constant is included:
AIC                  -10.0948  SC                     -9.98220
HQ                   -10.0516  FPE                 4.13002e-005

AR 1-2 test:      F(2,47)   =   8.0881 [0.0010]**
ARCH 1-1 test:    F(1,50)   =   1.2062 [0.2773]
Normality test:   Chi^2(2)  =  0.12384 [0.9400]
Hetero test:      F(4,47)   =   1.5738 [0.1969]
Hetero-X test:    F(5,46)   =   1.2577 [0.2983]
RESET23 test:     F(2,47)   =  0.67768 [0.5127]
```

t-testing Procedure to Check Variable Significance

- Create **null hypothesis**: $H0 : \beta_2 = 0$.
  - Variable is insignificant: Regression coefficient has zero mean.
- Divide estimator ($\hat{\beta}_2$ here) by estimated standard error ($\sqrt{V[\hat{\beta}_2]}$ here).
  - This calculates the t-statistic.
- Check t-statistic against critical values from t distribution (t_T here):
  - If 5% significance level then check against tables or take roughly ± 2.[1]
- If t-statistic greater than ± 2 then we **reject null hypothesis**.
  - Otherwise we **do not reject null hypothesis**.

[1] Recall test is two-tailed: We do not care whether significance is positive or negative!

Institute for New Economic Thinking at the Oxford Martin School

- In regression output, the `t-prob` column contains **p-values**.
- In output of other tests (details later), p-values are in [square brackets].
- p-value is probability of rejecting null-hypothesis if null hypothesis is true
- probability of observing test-statistic under the null hypothesis
  - if "sufficiently unlikely" (small), reject H_0
  - Calculated by imposing null hypothesis, using data and assumed distribution.
- If p-value very small, null hypothesis is statistically unlikely:
  - If p-value below significance level α (e.g. 0.05) we **reject null**.
- p-value **probability of falsely rejecting null hypothesis**.

Baseball Model:

$$\text{Att\_pop}_t = \underset{(0.00226)}{-0.004753} + \underset{(0.00398)}{0.02248}\ \text{PCT}_t + \underset{(0.000137)}{0.0001291}\ \text{Unemp}_t$$

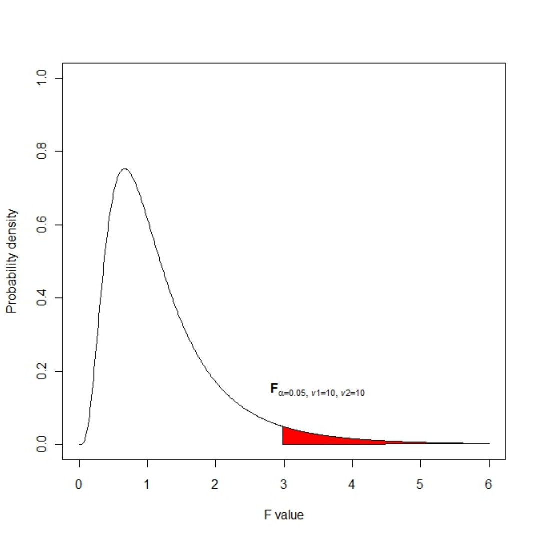
- In multivariate regression, F-test of joint significance different from t-statistics.
- t-testing is for individual regression coefficients:
  - Testing the (null) hypothesis that each one individually is insignificant.
- F-testing is for combinations of regression coefficients:
  - Testing the (null) hypothesis that together they are insignificant.
- Unemployment individually insignficant, variables jointly significant?

```
F(2,49) =     16.26 [0.000]**
```

- We test joint hypotheses using the F-test:
    - Calculate test statistic and find p-value: Probability of observing test statistic under null
- We find test statistic and distribution:

$$F = \frac{(RSS_R - RSS_U)/(K_U - K_R)}{RSS_U/(T - K_U)} \sim F_{K_U - K_R, T - K_U}. \quad (3)$$

- Baseball model: $H_0 : \beta_2 = 0.025, \beta_3 = 0.$
    - We *could* impose null by creating $Y_t - X_t$ variable but much easier: Use PcGive:
        - *Test* then *General Restrictions*.
        - Each variable denoted by ampersand (&) and number (See key beneath).
        - Each restriction is line of code; **must** be ended with semi-colon `;`.


Probability density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
F value
F α=0.05, ν1=10, ν2=10

- Common use of F test: Overall significance of model.
  - More relevant when we add **even more regressors**: Thursday and Friday.
- F-test statistic:

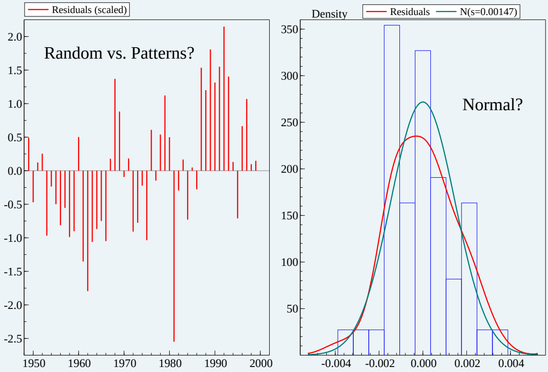
$$F = \frac{(RSS_R - RSS_U)/(K_U - K_R)}{RSS_U/(T - K_U)} = \frac{(TSS - RSS)/(K - 1)}{RSS/(T - K)}. \tag{4}$$

- $ESS = TSS - RSS$: How much do variables explain of variation?
  - Above constant: Base model is still one-variable model $Y_t = \beta_1 + \epsilon_t$.
- Form of F yields distribution: $F_{K-1,T-K}$ and hence **critical values**.
  - Null hypothesis: Explanatory variables are **jointly statistically insignificant**.
  - E.g. $H_0 : \beta_2 = \beta_3 = 0$.

- Our assumptions imply ϵ_t has **independent** (i) **identical** (ii) distribution.
  - Often written as $\epsilon_t \sim \mathrm{iid}(0, \sigma^2)$ (distribution need not be Normal).

$$Y_t = \underbrace{\beta_1 + \beta_2 X_t + \beta_3 Z_t}_{\text{Our model: What we know}} + \underbrace{\epsilon_t, \quad \epsilon_t \sim \mathrm{iid}(0, \sigma^2)}_{\text{Errors: What we don't know}}. \qquad (5)$$

- **Important** because want errors (mistakes) from our model to be:
  - Unrelated to each other.
  - Always of a similar size.
- This is **central principle of all economic and econometric modelling**.
- If not then our model does not include some important information!
  - More importantly, $\hat{\beta}_2$, $\hat{\beta}_3$ may be **biased** and **inefficient**.



AR 1-2 test:	F(2,47)	=	8.0881	[0.0010]**
ARCH 1-1 test:	F(1,50)	=	1.2062	[0.2773]
Normality test:	Chi^2(2)	=	0.12384	[0.9400]
Hetero test:	F(4,47)	=	1.5738	[0.1969]
Hetero-X test:	F(5,46)	=	1.2577	[0.2983]
RESET23 test:	F(2,47)	=	0.67768	[0.5127]

- Estimated regression model:

$$Y_t = \hat{\beta}_1 + \hat{\beta}_2 X_t + \hat{\beta}_3 Z_t + \hat{\epsilon}_t. \quad (6)$$

- Everything so far depends on iid assumption holding:
  - Unbiasedness and efficiency of estimators.
  - Accuracy of standard errors and all other test statistics.
- Must test whether assumptions hold: What do $\hat{\epsilon}_t$ look like?
- Tests called **misspecification tests** or **diagnostic tests**.

Assume $\epsilon_t \sim iid(0, \sigma^2) \Rightarrow$

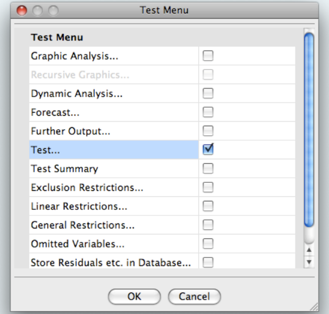
Model:
• $X_{1,t}, \ldots, X_{K,t}$.
• Functional form.
• Data transformations.

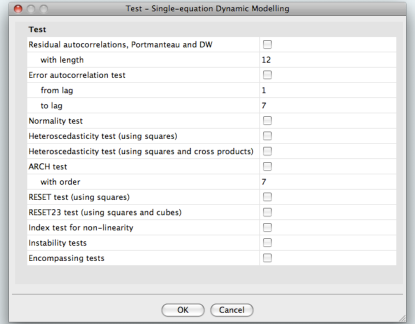
$\Longleftrightarrow$ **Check**

1. Identical distribution:
   - Is $\epsilon_t \sim (0, \sigma^2)$?
   - Mean must be zero (if constant included), but variance can change.
   - **Heteroskedasticity**: $V\epsilon_t = \sigma_t^2$.
2. Independent distribution:
   - Is $\mathrm{Corr}[\epsilon_s, \epsilon_t] = \mathrm{E}[\epsilon_s \epsilon_t] = 0$, for all $s \neq t$? If not, **autocorrelation**.
   - Variance not autocorrelated (**ARCH**)
3. Normal distribution:
   - Is $\epsilon_t \sim N0, \sigma^2$?
   - Residuals may have very different distribution.
4. Functional Form:
   - Model specification test

$\Longrightarrow$ will consider in turn!

- OxMetrics has very developed ways (tests) for detecting misspecification.

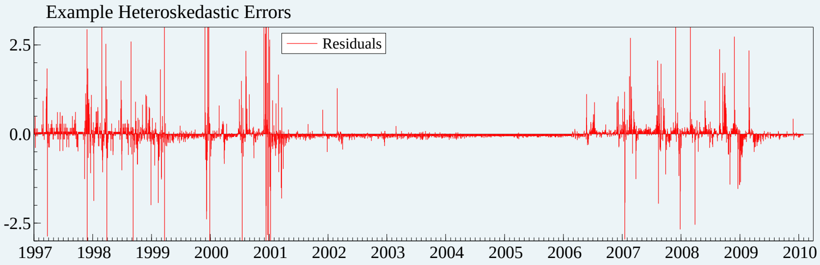


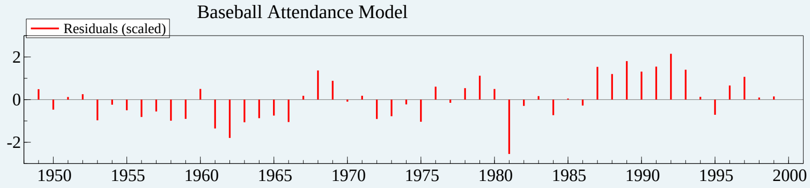


- When conducting testing in OxMetrics, will come across many different test.
  - t-test: Tests on individual regression coefficients.
  - F-test: Tests on combinations of regression coefficients.
- Additionally three types of test based on Likelihood:
  - **Maximum likelihood** is alternative, more flexible form of estimation.
  - Choose parameters (β_1, β_2) to maximise likelihood of having observed our data.
  - Likelihood function is probability density as function of parameters.
- Tests are based on different aspects of likelihood function:
  - **Likelihood Ratio (LR) test**: How likely is null hypothesis?
  - **Wald test**: Computes only under alternative hypothesis.
  - **Lagrange Multipler (LM) test**: Evaluates only under null hypothesis.
- Type of test not too important; **p-values are the fundamental concept**.

- Assumption of $\epsilon_t \sim (0, \sigma^2)$ violated.
    - Variance changes through sample: $\epsilon_t \sim (0, \sigma_t^2)$.
- Examples:
    - Stock market more volatile in financial crisis.
    - Inflation more volatile when higher.
- Loss of precision:
    - **Efficiency** of estimator relies on constant σ^2 (homoskedasticity).
    - Standard errors larger than they should be.
    - Possible mistake: May conclude variable insignificant that is actually significant.
- On average we still get right answer, but more possibility **we make mistakes**.

- Detection: Use residuals $\hat{\epsilon}_t$ in place of errors ϵ_t.
- Graphic inspection: Does the variance 'look' constant?
  - Residuals (scaled) are standard graphics output in OxMetrics





- White test of heteroskedasticity: regress $\hat{\epsilon}_t^2$ on $X_{1,t}, \ldots, X_{K,t}$, and $X_{1,t}^2, \ldots, X_{K,t}^2$:
  - Does residual variance depend on regressors included in model, or their squares?
  - Square of variable rough approximation to variance.
- Model:

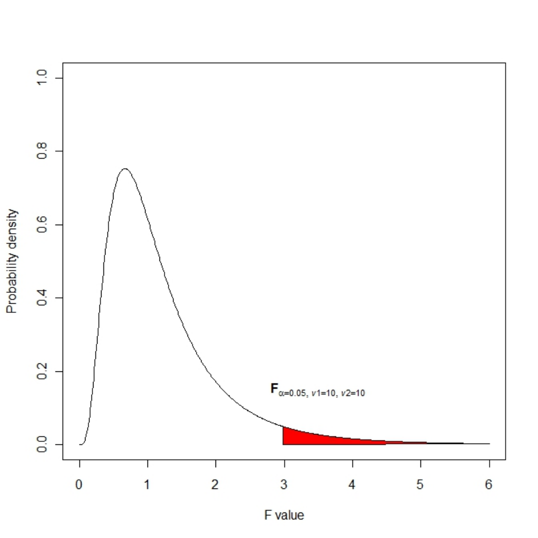
$$\hat{\epsilon}_t^2 = \alpha_0 + \alpha_1 X_{1,t} + \cdots + \alpha_K X_{K,t} + \alpha_{K+1} X_{1,t}^2 + \cdots + \alpha_{2K} X_{K,t}^2 + \nu_t.$$

- Null hypothesis:

$$H0 : \alpha_1 = \cdots = \alpha_{2K} = 0, \tag{7}$$

  - Absence of heteroskedasticity, or homoskedasticity (constant variance).
- F-test of joint significance:

$$F_{Het} = \frac{R_{Het}^2 / m}{\left(1 - R_{Het}^2\right) / (T - m)} \sim F_{m,T-m}. \tag{8}$$


Probability density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
F value
$F_{\alpha=0.05,\ v1=10,\ v2=10}$

Formal Test in Attendance Model:

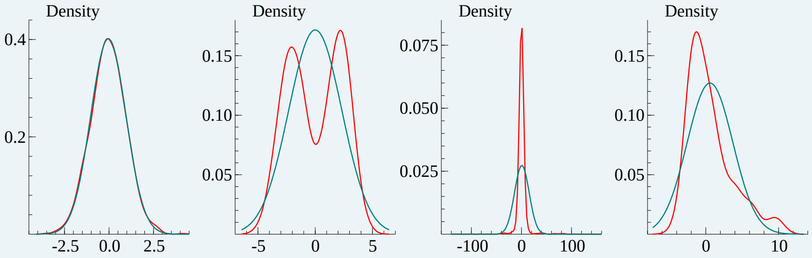
- Hetero test : $\mathrm{F}(4, 47) = 1.5738[0.1969]$

Causes:

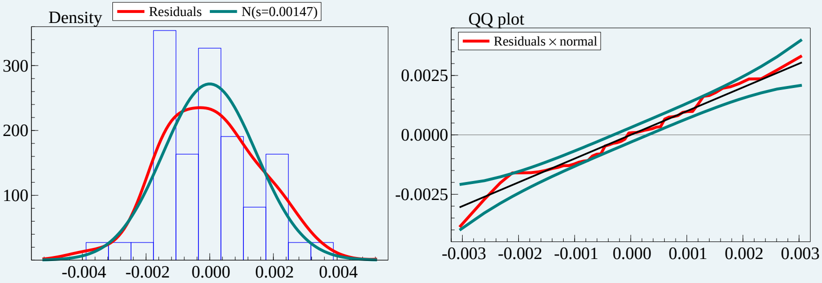
- Omitted Variables:
  - Systematic variation in residuals.
  - Unincluded variable may explain change in variation.
    - E.g. Age may affect how volatile income is across individuals.
- Wrong functional form/data transformation:
  - Logarithmic transformation can stabilise variance in series.
- Non-constant parameters:
  - We are assuming β_1, β_2 stay same throughout sample.
  - If they don't then result may be heteroskedasticity.

- Active:
  - Add the regressors you think cause heteroskedasticity.
- Passive: Calculate Heteroskedasticity Consistent Standard Errors (HCSE).
  - Heteroskedasticity affects standard errors: Become bigger.
  - These standard errors are robust to heteroskedasticity.

- Do not assume Normal distribution but Normal distribution is 'nice' distribution.
- Assume $\epsilon_t \sim \mathrm{iid}(0, \sigma^2)$, so:
  - Distribution unchanging over time.
  - Distribution symmetric around zero.



- Normality not assumed for OLS so provided iid assumption holds inference unaffected:
  - OLS estimator unbiased, consistent, and efficient.
- **But:** Non-Normality may reveal information about dataset:
  - Inappropriate functional form and/or data transformation.
- Informally "test" by plotting data series:



- Skewness and kurtosis of distribution are like mean and variance:
  - They describe 'shape' of distribution:
    - Skewness: How symmetric is the distribution?
    - Kurtosis: How flat/spiky is distribution?
- Test statistics:

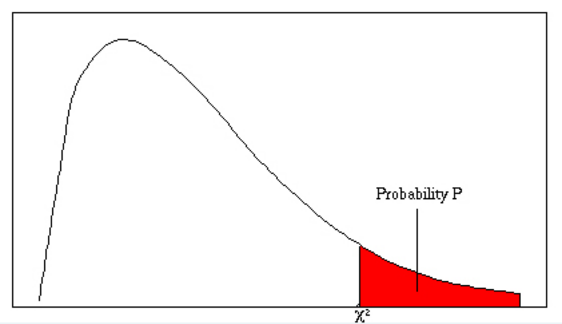
$$\chi^2_{\text{skewness}} = T\frac{\hat{\kappa}_3^2}{6} \sim \chi^2_1, \tag{9}$$

$$\chi^2_{\text{kurtosis}} = T\frac{\hat{\kappa}_4^2}{24} \sim \chi^2_1, \tag{10}$$

$$\chi^2_{\text{normality}} = \chi^2_{\text{skewness}} + \chi^2_{\text{kurtosis}} \sim \chi^2_2. \tag{11}$$

- PcGive reports this test. Null hypothesis is Normal residuals.
  - Hence if test rejected (we see stars) then our model is **misspecified.**
  - Test in Baseball model: Normality test: $\mathrm{Chi}^2(2) = 0.12384[0.9400]$

Under null-hypothesis:



- Example: Normality test: $\mathrm{Chi}^2(2) = 0.12384[0.9400]$
- We should investigate non-Normalities:
  1. Non-Normality may be harmless for inference.
  2. Non-Normality may reveal important information about model specification.

1. Non-Normality may be harmless for inference:
   - iid assumption translates into **non-skewed distribution**.
   - Hence if test fail caused by **excess kurtosis** then model is fine.
2. May reveal information about model to help diagnose other problems found.
   - Inappropriate data transformation.
   - Structural breaks (see recursive testing later)

- Additional information via Test menu and `Test...`

- Economic theory often predicts which variables matter.
  - Less often predicts mathematical form of dependence.
- Test for functional form: RESET test (Ramsey, 1969).[2]
  - Include squares, cubes of fitted values.
  - Null hypothesis of correct functional form: Additional variables do not matter.
- Auxiliary regression:

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \psi_1 \hat{Y}_i^2 + \psi_2 \hat{Y}_i^3 + \nu_i.$$

- Null hypothesis:

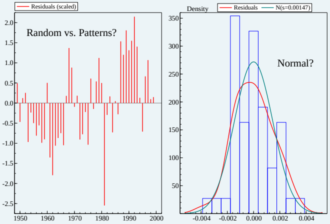
$$H0 : \psi_1 = \psi_2 = 0. \quad (12)$$

- F-test statistic:

$$F_{RESET} = \frac{(RSS_R - RSS_U)/2}{RSS_U/(T-K-2)} \sim F_{2,T-K-2}. \quad (13)$$

- `RESET23 test:        F(2,47)   =   0.67768 [0.5127]`

[2] **RE**gression **S**pecification **E**rror **T**est.




```
AR 1-2 test:        F(2,47)   =   8.0881 [0.0010]**
ARCH 1-1 test:      F(1,50)   =   1.2062 [0.2773]
Normality test:     Chi^2(2)  =  0.12384 [0.9400]
Hetero test:        F(4,47)   =   1.5738 [0.1969]
Hetero-X test:      F(5,46)   =   1.2577 [0.2983]
RESET23 test:       F(2,47)   =  0.67768 [0.5127]
```

Common Model Specification: Log Model

- So far:

$$Y_t = \beta_1 + \beta_2 X_t + \epsilon_t \tag{14}$$

  - $\beta_2 = \frac{dY_t}{dX_t}$
- Log-log Model: interpret as percentage changes/elasticities

$$\log(Y)_t = \beta_1 + \beta_2 \log(X_t) + \epsilon_t \tag{15}$$

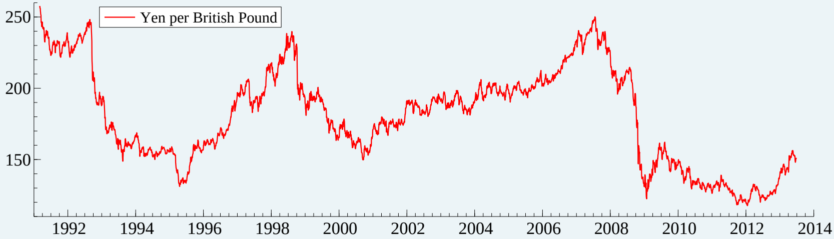
- $\beta_2 = \frac{d\log(Y)_t}{d\log(X)_t} \approx \frac{\frac{dY_t}{Y_t}}{\frac{dX_t}{X_t}} = \frac{dY_t}{dX_t}\frac{X_t}{Y_t}$ = elasticity of Y_t with respect to X_t
- Percentage change! For a 1% change in X_t expect a β_2% change in Y_t.

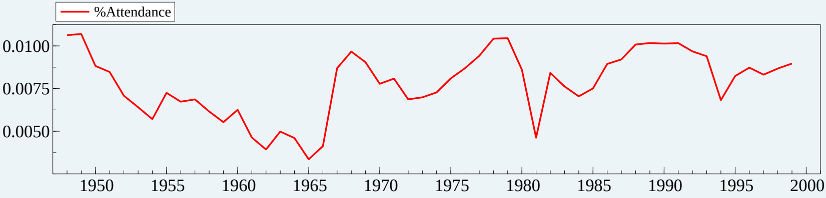
- So far: Test theory using econometric model.
  - E.g. Baseball attendance: positive relationship between winning percentage and attendance.
- **Model**: %Attendance$_t = \beta_1 + \beta_2 PCT_t + \beta_3 Unemp._t + \epsilon_t$
- So far: All variables dated t: This is a **static** regression.
  - But where variables are today depends on where they were previously.
  - How can we predict variables if all our dates are t?
- Today: **Dynamic** models.

- May be that variable is correlated with itself at a previous point in time.
  - E.g.: Exchange rate, interest rates, inflation (low and stable prices), unemployment, baseball attendance.
- Correlation through time is concept called **time dependence**.
  - We recognise how economic variable today is dependent on where it was yesterday.
- Expect persistence in many economic variables; especially prices.
  - Supply reflects production capacity which does not drastically change.
  - Demand reflects frequent purchases if a necessity good.

- Y_{t-r} is **lagged variable**:
  - The same variable at previous point in time.
- Y_{t-1} is first lag: First previous time period.
- Y_{t-2} is second lag, Y_{t-3} third, so on...
- Use lags to learn about:
  - Persistence: How important is a variable's history?
  - Effects of explanatory variables distributed over time (e.g. advertising).

- E.g. Spot exchange rate: Demand and supply of UK pounds for Japanese Yen (Top), % Attendance (bottom).





- Want to understand persistence:
  - Tells us much about economic variables.
    - E.g. price efficiency, partial adjustments, interest rate smoothing.
  - If we don't model it properly, can cause big mistakes.
- **Autoregressive models**:
  - Regression model of variable Y_t on itself in previous time period Y_{t-1}.

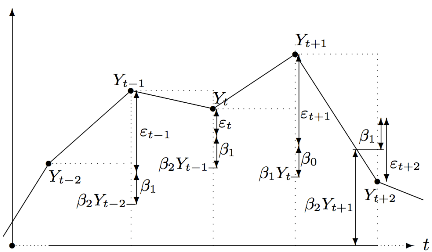
- Autoregressive model has three elements:
  - **(1)** Where Y_t was the last time period.
  - **(2)** The unexpected event ϵ_t.
  - **(3)** Constant term allowing mean of Y_t to be non-zero.

$$Y_t = \underbrace{\alpha_0}_{(3)} + \underbrace{\alpha_1 Y_{t-1}}_{(1)} + \underbrace{\epsilon_t}_{(2)}, \quad \epsilon_t \sim \mathbb{N}[0, \sigma^2]. \tag{16}$$

Notation: normally use α for autoregressive, but equivalent to:

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathbb{N}[0, \sigma^2]. \tag{17}$$

(a) The effect of shocks



- AR(1) model allows us to determine many things about theory:
  - α_1: How quickly equilibrium re-established.
  - α_0 and α_1: Whether equilibrium is zero or otherwise.
  - σ^2: How much variation there is in Y_t around equilibrium.
    - How big are the unexpected events?
- What is equilibrium value? Again expectations:

$$EY_t = \alpha_0 + \alpha_1 EY_{t-1}. \tag{18}$$

- Since $EY_t = EY_{t-1}$ we find that $\mu_Y = EY = \alpha_0/(1-\alpha_1)$.
  - We define μ_Y to be the equilibrium value, or **unconditional mean** of Y_t.

- We learn about the persistence of deviations from equilibrium from α_1.
- To see why note that $\mu_Y = \alpha_0/(1-\alpha_1)$ implies $\alpha_0 = \mu_Y(1-\alpha_1)$ so that:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \epsilon_t \quad \Longrightarrow \quad Y_t - \mu_Y = \alpha_1(Y_{t-1} - \mu_Y) + \epsilon_t. \tag{19}$$

- We have **de-meaned** Y_t: We only care about α_1 and deviations from equilibrium.
- If assume **no more shocks happen** can see how quickly impact of shock disappears.
- $Y_t - \mu_Y = \alpha_1(Y_{t-1} - \mu_Y)$ and $Y_{t-1} - \mu_Y = \alpha_1(Y_{t-2} - \mu_Y)$ so:

$$Y_t - \mu_Y = \alpha_1^2(Y_{t-2} - \mu_Y). \tag{20}$$

- We can carry on doing this:

$$Y_t - \mu_Y = \alpha_1^3(Y_{t-3} - \mu_Y) \quad \ldots \quad Y_t - \mu_Y = \alpha_1^k(Y_{t-k} - \mu_Y) \tag{21}$$

- It so happens that:

$$\text{Corr}\left[Y_t, Y_{t-k}\right] = \frac{\text{Cov}(Y_t, Y_{t-k})}{\sqrt{V(Y_t)}\sqrt{V(Y_{t-k})}} = \frac{\alpha_1^k \sigma_Y^2}{\sigma_Y \times \sigma_Y} = \alpha_1^k. \tag{22}$$

  - Have measured autocorrelation, or correlation through time, of Y_t from α_1!
- The bigger is α_1 and hence nearer to 1, the more persistent is the series:
  - If $\alpha_1 = 0.9$ then $\alpha_1^2 = 0.81$ and $\alpha_1^{10} = 0.35$.
  - If $\alpha_1 = 0.2$ then $\alpha_1^2 = 0.04$ and $\alpha_1^{10} \approx 0$.

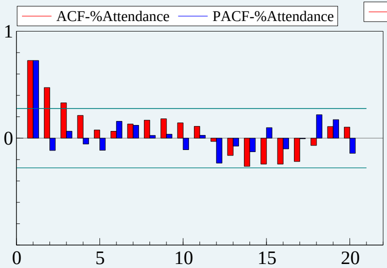
- May need more than one lag to explain dynamics of variable:
  - If we model p lags, we have AR(p) model.
- E.g. AR(2): $Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \epsilon_t$.
- Estimators like in **multivariate regression**:
  - $\hat{\alpha}_2$ asks Y_{t-1} to **be still**! It controls for first lag to get **only second lag effect**.

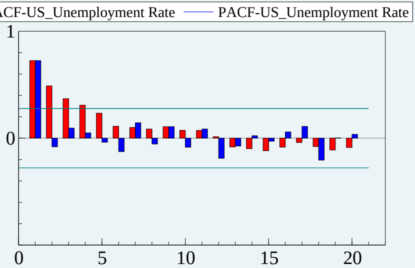
$$\hat{\alpha}_2 = \frac{\sum_{t=2}^{T} Y_{t-2}(Y_t | Y_{t-1})}{\sum_{t=2}^{T} Y_{t-2}(Y_{t-2} | Y_{t-1})}.$$

- Unconditional mean, variance and covariance affected. E.g. unconditional mean:

$$\mu_Y = \frac{\alpha_0}{1 - \alpha_1 - \alpha_2}. \tag{23}$$

- Common method for learning about autocorrelation is graphically.
  - Autocorrelation function (ACF): $\mathrm{Corr}[Y_t, Y_{t-p}]$, $p = 1, 2, \ldots, 20$.
  - Partial ACF (PACF): $\mathrm{Corr}[Y_t, Y_{t-p} | Y_{t-1}, \ldots, Y_{t-p+1}]$, $p = 1, 2, \ldots, 20$.
    - Number of significant PACF lags = number of lags needed in model.



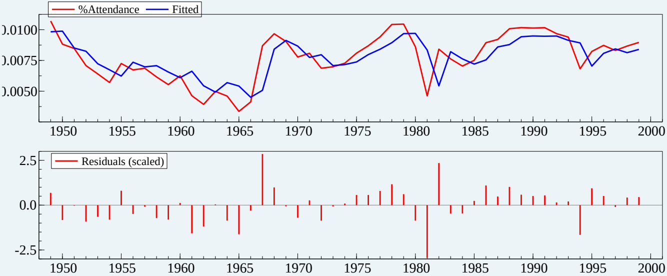


- **Autoregressive model for Baseball Attendance**:

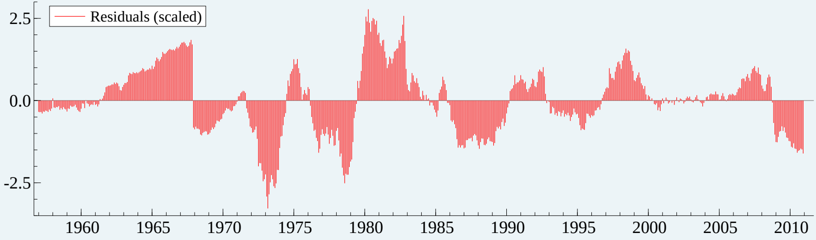
$$\%\text{Attendance}_t = \mu_0 + \alpha_1 \%\text{Attendance}_{t-1} + \epsilon_t \qquad (24)$$

- **Estimated Model**:

$$\%\text{Attendance}_t = \underset{(0.000747)}{0.002052} + \underset{(0.0935)}{0.7316} \ \%\text{Attendance}_{t-1}$$



- Earlier: Misspecification testing: Check iid assumption.
  - $\hat{\epsilon}_t$ all approximately same size?
  - $\hat{\epsilon}_t$ all unrelated to each other?
- Latter failure in time series is **residual autocorrelation**: $\text{Corr}(\epsilon_t, \epsilon_{t-1}) \neq 0$.
- Omitted variable bias: We omit lags that matter for explaining Y_t.



- Test using auxiliary regression:

$$\hat{\epsilon}_t = \gamma_0 + \gamma_1 Y_{t-1} + \phi_1 \hat{\epsilon}_{t-1} + \nu_t, \quad (25)$$

- Null hypothesis:

$$H0 : \gamma_1 = \phi_1 = 0, \quad (26)$$

  - Absence of autocorrelation in errors of model (model well specified).
- F-test statistic:

$$F_{AR} = \frac{R^2_{AR}/r}{\left(1 - R^2_{AR}\right) / (T - r)} \sim F_{r,T-r}. \quad (27)$$

  - Can include up to r lags when testing. Can vary in Test menu and
- Remedy autocorrelation by including time-dependent information:
  - More lags or extra variables (with lags).

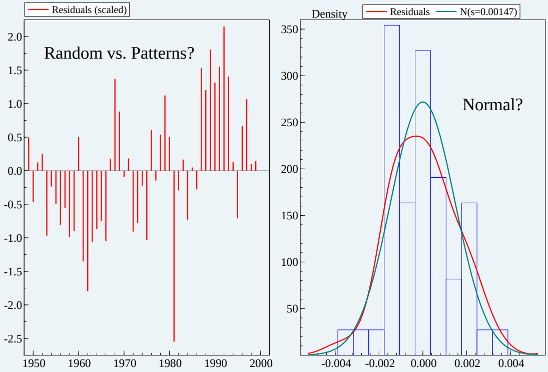
- ARCH: **A**utoregressive **C**onditional **H**eteroskedasticity.
- Heteroskedasticity: $\epsilon_t \sim (0, \sigma_t^2)$, variance changes over time.
- Autoregressive: Variable correlated with itself.
- ARCH: $\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + u_t$.
    - Variance is persistent: Spells of high and low volatility.
    - Common in financial markets.
- Impact on regression model same as heteroskedasiticity: Efficiency.
- Test using AR(1) on squared residuals: $\hat{\epsilon}_t^2 = \alpha_0 + \alpha_1 \hat{\epsilon}_{t-1}^2 + \cdots + \alpha_r \hat{\epsilon}_{t-r}^2 + \nu_t$.
    - F-test: $H0 : \alpha_1 = \cdots = \alpha_r = 0$.

- Recall our model: **Model**:
  $\%\text{Attendance}_t = \beta_1 + \beta_2 PCT_t + \beta_3 Unemp._t + \epsilon_t$
- Misspecification Tests

```
AR 1-2 test:        F(2,47)    =   8.0881 [0.0010]**
ARCH 1-1 test:      F(1,50)    =   1.2062 [0.2773]
Normality test:     Chi^2(2)   =  0.12384 [0.9400]
Hetero test:        F(4,47)    =   1.5738 [0.1969]
Hetero-X test:      F(5,46)    =   1.2577 [0.2983]
RESET23 test:       F(2,47)    =  0.67768 [0.5127]
```

$\rightarrow$ Residual Autocorrelation!

Baseball Attendance: Specification




```
AR 1-2 test:        F(2,47)    =   8.0881 [0.0010]**
ARCH 1-1 test:      F(1,50)    =   1.2062 [0.2773]
Normality test:     Chi^2(2)   =  0.12384 [0.9400]
Hetero test:        F(4,47)    =   1.5738 [0.1969]
Hetero-X test:      F(5,46)    =   1.2577 [0.2983]
RESET23 test:       F(2,47)    =  0.67768 [0.5127]
```

Add extra lags of dependent variable:

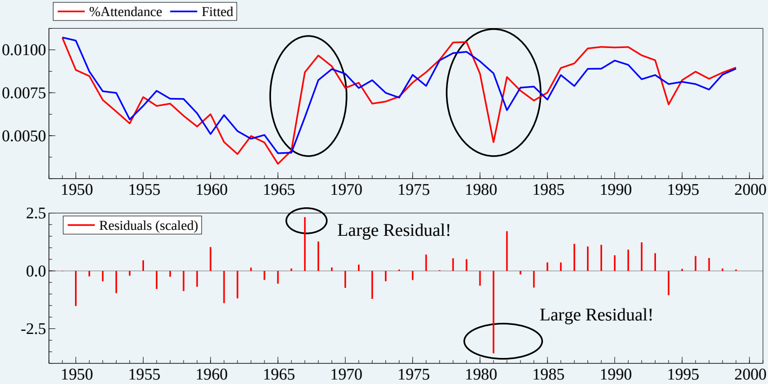
$$\%\text{Attend.}_t = \alpha_1 \%\text{Attend.}_{t-1} + \beta_1 + \beta_2 PCT_t + \beta_3 Unemp._t + \epsilon_t$$

$$\%\text{Attend.} = \underset{(0.0913)}{0.5815}\ \%\text{Attend.}_{t-1} - \underset{(0.00171)}{0.003985} + \underset{(0.00335)}{0.01307}\ \text{PCT}_t + \underset{(0.000104)}{0.00006}\ \text{Unempl.}_t$$

```
AR 1-2 test:        F(2,45)    =  0.60392 [0.5510]
ARCH 1-1 test:      F(1,49)    =  0.67357 [0.4158]
Normality test:     Chi^2(2)   =   11.261 [0.0036]**
Hetero test:        F(6,44)    =  0.53186 [0.7810]
Hetero-X test:      F(9,41)    =  0.43033 [0.9109]
RESET23 test:       F(2,45)    =  0.10631 [0.8994]
```

What now??

Look at the model fit and residuals:



May distort normality tests! What happened? More on Thursday!

Computer Lab Session 2:

- Testing joint-hypotheses
  - F-test
- Interpreting Misspecification Tests
- Investigating Time Series Properties
  - Time Series plots

Fulton Fish Market: Price, Quantity, Weather

- Load ”fish.in7”
- Series
  - Model for qty = log(Quantity)
  - Weather: Stormy, Rainy, Cold
- Graph the series! (Important first step!)

Construct Auto-regressive models for log(Quantity) sold:

- Determine lag length: (Partial) Auto-correlation function (max 10 lags)
- Estimate an AR(1), AR(2) models
- What is the long-run equilibrium?
- Interpret mis-specification tests

```
EQ(15) Modelling qty by OLS
       The estimation sample is: 2 - 111

                  Coefficient  Std.Error  t-value  t-prob Part.R^2
qty_1                0.203549    0.09406     2.16  0.0327   0.0416
Constant              6.78385     0.8049     8.43  0.0000   0.3968

sigma                0.731432  RSS               57.7792493
R^2                 0.0415598  F(1,108) =     4.683 [0.033]*
Adj.R^2             0.0326853  log-likelihood      -120.671
no. of observations       110  no. of parameters          2
mean(qty)             8.51915  se(qty)             0.743687


AR 1-2 test:      F(2,106)  =   1.9872 [0.1422]
ARCH 1-1 test:    F(1,108)  =   2.0874 [0.1514]
Normality test:   Chi^2(2)  =   6.9103 [0.0316]*
Hetero test:      F(2,107)  =   3.6890 [0.0282]*
Hetero-X test:    F(2,107)  =   3.6890 [0.0282]*
RESET23 test:     F(2,106)  =  0.69995 [0.4989]
```

Interested in **effects of weather** on quantity sold:

- Estimate auto-regressive model with weather variables added in
- Include: Stormy, Rainy, Cold
- Which variables are individually significant?
- Which variables are jointly significant?

EQ(17) Modelling qty by OLS

	Coefficient	Std.Error	t-value	t-prob	Part.R^2
qty_1	0.184254	0.09336	1.97	0.0511	0.0358
Constant	7.06086	0.8097	8.72	0.0000	0.4200
stormy	-0.342175	0.1681	-2.04	0.0443	0.0380
rainy	0.0824118	0.1918	0.430	0.6683	0.0018
cold	-0.0566163	0.1524	-0.372	0.7109	0.0013

sigma	0.721793	RSS	54.7034867
R^2	0.0925804	F(4,105) =	2.678 [0.036]*
Adj.R^2	0.0580121	log-likelihood	-117.663
no. of observations	110	no. of parameters	5
mean(qty)	8.51915	se(qty)	0.743687

AR 1-2 test:	F(2,103)	=	0.82520	[0.4410]
ARCH 1-1 test:	F(1,108)	=	1.7838	[0.1845]
Normality test:	Chi^2(2)	=	8.7179	[0.0128]*
Hetero test:	F(5,104)	=	1.3869	[0.2352]
Hetero-X test:	F(5,104)	=	1.3869	[0.2352]
RESET23 test:	F(2,103)	=	0.65843	[0.5198]

F-Test for joint significance of all 3 weather variables

- *Test* - *Exclusion Restrictions*
- Select all 3 variables

```
Test for excluding:
[0] = stormy
[1] = rainy
[2] = cold
Subset F(3,105)  =   1.9679 [0.1233]
```

Modelling share price (Apple) using Google search data:

Can share price be modelled using consumer interest (measured through google search term: ”iphone”)

- load data ”apple_google.in7”
- Build a model for Apple closing price
- What lag length should you choose?
- Is the google search term a useful predictor? What about lagged search terms?
- What about the effect of google searches on the trading volume?