Sections 6.3-6.4

- Seasonal variation: The effect of the time of year on the time series
  - Constant seasonal variation: The magnitude of the seasonal swing does not depend on time
  - Increasing seasonal variation: The magnitude of the seasonal swing varies with time
- It is common practice to transform an increasing seasonal variation into a constant seasonal variation for the model
  - Use a power transformation $y_t^* = y_t^\lambda$ with $0 < \lambda < 1$
  - Common transformations:
    * Square root transformation
    * Logarithmic transformation
- A constant seasonal trend can be added into a time series model

$$y_t = TR_t + SN_t + \epsilon_t$$

- If a non-constant seasonal trend is added to a time series model, the error terms will not have constant variance
- A constant seasonal trend can be modeled by
  1. Dummy variables
     - If there are L seasons, create $L-1$ dummy variables such that $D_{j,t}$ is 1 if time period t is in the j^{th} season and 0 otherwise, then

$$SN_t = \beta_1 D_{1,t} + \beta_2 D_{2,t} + \ldots + \beta_{L-1} D_{L-1,t}$$

     - If the seasons of interest are the four business quarters then

$$SN_t = \beta_1 Q_{1,t} + \beta_2 Q_{2,t} + \beta_3 Q_{3,t}$$

     - Each dummy variable relates the seasonality of the season it represents with respect to the baseline season
  2. Trigonometric functions
     (a) For very regular constant seasonal variation:

$$SN_t = \beta_1 \sin\left(\frac{2\pi t}{L}\right) + \beta_2 \cos\left(\frac{2\pi t}{L}\right)$$

     (b) For a more complicated pattern:

$$SN_t = \beta_1 \sin\left(\frac{2\pi t}{L}\right) + \beta_2 \cos\left(\frac{2\pi t}{L}\right) + \sin\left(\frac{4\pi t}{L}\right) + \beta_2 \cos\left(\frac{4\pi t}{L}\right)$$

- Dummy variables are generally preferred to trigonometric functions because they treat each season individually instead of treating the entire period as a unit

Sections 6.5-6.6

- There are times when the data can be described better by a model that is not linear
- Growth curve model

$$y_t = \beta_0(\beta_1^t)\epsilon_t$$

  - The growth curve will be increasing if $\beta_1 > 1$ and decreasing if $0 < \beta_1 < 1$
  - The further β_1 is from 1, the more extreme the curvature will be
  - Since the error term is multiplied, the points will be randomly fluctuating around this trend but fanning out at higher values and fanning in at lower values
  - To be able to predict, a logarithmic transformation is used on the entire model when $\beta_0 > 0$ and $\beta_1 > 0$

$$\begin{aligned} \log(y_t) &= \log(\beta_0) + \log(\beta_1)t + \log(\epsilon_t) \\ y_t^* &= \alpha_0 + \alpha_1 t + u_t \end{aligned}$$

  - The coefficients α_j must be untransformed to find the coefficients β_j
  - The predicted value and associated intervals must be untransformed to yield the appropriate value for y_t
- Growth rate: The amount higher we expect y_{t+1} to be over y_t, which is found by $100(\beta_1 - 1)\%$

$$\begin{aligned} y_t &= \beta_0(\beta_1^t)\epsilon_t \\ &= \left(\beta_0(\beta_1^{t-1})\right)\beta_1\epsilon_t \\ &= y_{t-1}\beta_1\epsilon_t \end{aligned}$$

- Accounting for autocorrelation will shorten prediction intervals and yield better predictive power
- The first-order autoregressive model is

$$\epsilon_t = \phi_1\epsilon_{t-1} + a_t$$

- The p-order autoregressive model is

$$\epsilon_t = \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \ldots + \phi_p\epsilon_{t-p} + a_t$$

- When there is first-order autocorrelation, the modified time series model becomes

$$y_t = TR_t + SN_t + \phi_1\epsilon_{t-1} + a_t$$

- In order to predict $y_{T+\tau}$, where T is the last time period observed and τ is a number of time periods into the future, the prediction equation is

$$\hat{y}_{T+\tau} = \widehat{TR}_{T+\tau} + \widehat{SN}_{T+\tau} + \hat{\phi}_1\hat{\epsilon}_{T+\tau-1}$$

- If $\tau = 1$, to predict the first point after the set of observed data, $\hat{\epsilon}_{T+\tau-1}$ is estimated by

$$\hat{\epsilon}_{T+\tau-1} = y_T - \widehat{TR}_T - \widehat{SN}_T$$

- If $\tau > 1$, to predict points after the first point following the observed data, $y_{T+\tau-1}$ must also be predicted and $\hat{\epsilon}_{T+\tau-1}$ is estimated by

$$\hat{\epsilon}_{T+\tau-1} = \hat{y}_{T+\tau-1} - \widehat{TR}_{T+\tau-1} - \widehat{SN}_{T+\tau-1}$$

- The $100(1-\alpha)\%$ prediction intervals depend on τ
  - If $\tau = 1$

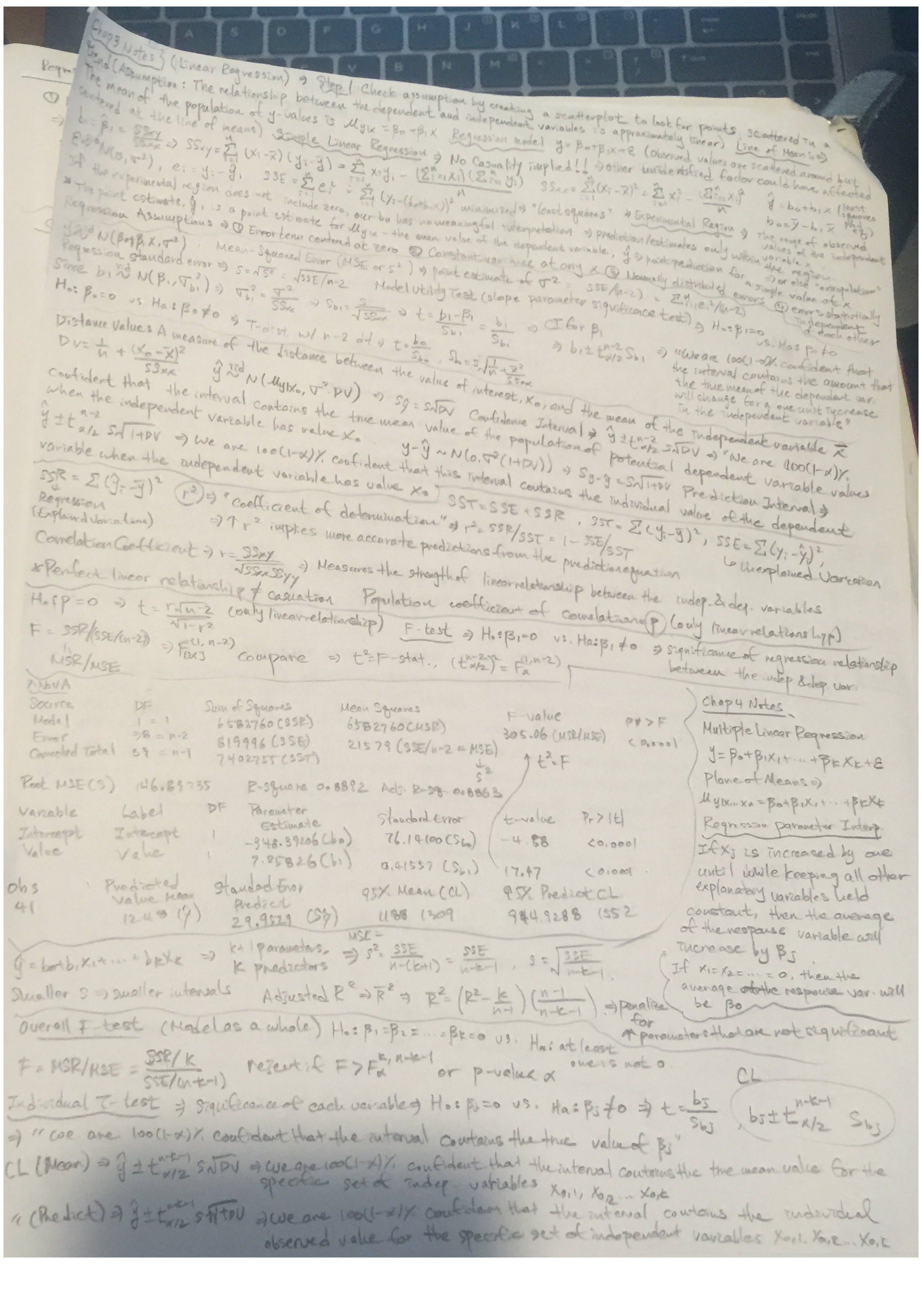
$$\left[\hat{y}_{T+1} \pm z_{[\alpha/2]}s\right]$$

  - If $\tau = 2$

$$\left[\hat{y}_{T+2} \pm z_{[\alpha/2]}s\sqrt{1 + (\hat{\phi}_1)^2}\right]$$

  - If $\tau \geq 3$

$$\left[\hat{y}_{T+\tau} \pm z_{[\alpha/2]}s\sqrt{1 + (\hat{\phi}_1)^2 + \ldots + (\hat{\phi}_1)^{2(\tau-1)}}\right]$$


Chap3 Notes (Linear Regression) ⇒ Step 1 Check assumption by creating a scatterplot to look for points scattered in a band (Assumption: The relationship between the dependent and independent variables is approximately linear) Line of Means ⇒ The mean of the population of y-values is $\mu_{y|x} = \beta_0 + \beta_1 x$ Regression model $y = \beta_0 + \beta_1 x + \varepsilon$ (observed values are scattered around but centered at the line of means) Simple Linear Regression ⇒ No Causality implied!! ⇒ other unidentified factor could have affected
$b_1 = \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$ ⇒ $SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$, $SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$, $\hat{y} = b_0 + b_1 x$ (least squares)
$\varepsilon_i \sim N(0, \sigma^2)$, $e_i = y_i - \hat{y}_i$, $SSE = \sum e_i^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$ minimized ⇒ "least squares", $b_0 = \bar{y} - b_1\bar{x}$
If the experimental region does not include zero, our b_0 has no meaningful interpretation. * Experimental Region ⇒ The range of observed values of the independent variable x
* The point estimate, $\hat{y}$, is a point estimate for $\mu_{y|x}$ - the mean value of the dependent variable, $\hat{y}$ ⇒ point prediction for a single value of x ⇒ predictions/estimates only within the region ⇒ or else "extrapolation"
Regression Assumptions ⇒ ① Error term centered at zero ② Constant variance at any x ③ Normally distributed errors ④ errors statistically independent of each other
$y \overset{iid}{\sim} N(\beta_0 + \beta_1 x, \sigma^2)$, Mean-Squared Error (MSE or s^2) ⇒ point estimate of σ^2 = $SSE/(n-2)$ = $\sum_{i=1}^{n} e_i^2/(n-2)$
Regression standard error ⇒ $s = \sqrt{s^2} = \sqrt{SSE/n-2}$
Since $b_1 \overset{iid}{\sim} N(\beta_1, \sigma_{b_1}^2)$ ⇒ $\sigma_{b_1}^2 = \frac{\sigma^2}{SS_{xx}}$ ⇒ $s_{b_1} = \frac{s}{\sqrt{SS_{xx}}}$ Model Utility Test (slope parameter significance test) ⇒ $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$
⇒ $t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1}{s_{b_1}}$ ⇒ CI for β_1 ⇒ $b_1 \pm t_{\alpha/2}^{n-2} s_{b_1}$ ⇒ "We are $100(1-\alpha)\%$ confident that the interval contains the amount that the true mean of the dependent var. will change for a one unit increase in the independent variable"
$H_0: \beta_0 = 0$ vs. $H_a: \beta_0 \neq 0$ ⇒ T-dist. w/ n-2 df ⇒ $t = \frac{b_0}{s_{b_0}}$, $s_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}$
Distance Value: A measure of the distance between the value of interest, x_0, and the mean of the independent variable $\bar{x}$
$DV = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}$, $\hat{y} \overset{iid}{\sim} N(\mu_{y|x_0}, \sigma^2 \cdot DV)$ ⇒ $s_{\hat{y}} = s\sqrt{DV}$ Confidence Interval ⇒ $\hat{y} \pm t_{\alpha/2}^{n-2} s\sqrt{DV}$ ⇒ "We are $100(1-\alpha)\%$ confident that the interval contains the true mean value of the population of potential dependent variable values when the independent variable has value x_0"
$y - \hat{y} \sim N(0, \sigma^2(1+DV))$ ⇒ $s_{y-\hat{y}} = s\sqrt{1+DV}$ Prediction Interval ⇒ $\hat{y} \pm t_{\alpha/2}^{n-2} s\sqrt{1+DV}$ ⇒ We are $100(1-\alpha)\%$ confident that this interval contains the individual value of the dependent variable when the independent variable has value x_0
$SSR = \sum(\hat{y}_i - \bar{y})^2$ Regression (Explained Variation) r^2 ⇒ "Coefficient of determination" ⇒ $r^2 = SSR/SST = 1 - SSE/SST$, $SST = SSE + SSR$, $SST = \sum(y_i - \bar{y})^2$, $SSE = \sum(y_i - \hat{y}_i)^2$ Unexplained Variation
⇒ ↑ r^2 implies more accurate predictions from the prediction equation
Correlation Coefficient ⇒ $r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$ ⇒ Measures the strength of linear relationship between the indep. & dep. variables
* Perfect linear relationship ≠ causation Population coefficient of correlation ρ (only linear relationship)
$H_0: \rho = 0$ ⇒ $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ (only linear relationship) F-test ⇒ $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$ ⇒ significance of regression relationship between the indep & dep var.
$F = SSR/(SSE/(n-2))$ ⇒ $F_{[\alpha]}^{(1, n-2)}$ compare ⇒ t^2 = F-stat., $(t_{\alpha/2}^{n-2})^2 = F_\alpha^{(1, n-2)}$
MSR/MSE
ANOVA

Source	DF	Sum of Squares	Mean Squares	F value	Pr > F
Model	1 = 1	6582760 (SSR)	6582760 (MSR)	305.06 (MSR/MSE)	< 0.0001
Error	38 = n-2	819996 (SSE)	21579 (SSE/n-2 = MSE) s^2		
Corrected Total	39 = n-1	7402755 (SST)			

$t^2 = F$
Root MSE (S) 146.89735 R-Square 0.8892 Adj. R-Sq 0.8863

Variable	Label	DF	Parameter Estimate	Standard Error	t-value	Pr > |t|
Intercept	Intercept	1	-348.39206 (b_0)	76.14100 (s_{b_0})	-4.58	<0.0001
Value	Value	1	7.25826 (b_1)	0.41557 (s_{b_1})	17.47	<0.0001

Obs	Predicted Value Mean	Standard Error Predict	95% Mean (CL)		95% Predict CL	
41	1248 ($\hat{y}$)	29.9529 ($s_{\hat{y}}$)	1188	1309	944.9288	1552

Chap4 Notes
Multiple Linear Regression
$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$
Plane of Means ⇒ $\mu_{y|x_1 \ldots x_k} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
Regression parameter Interp.
If x_j is increased by one unit while keeping all other explanatory variables held constant, then the average of the response variable will increase by β_j.
If $x_1 = x_2 = \ldots = 0$, then the average ~~of~~ the response var. will be β_0
$\hat{y} = b_0 + b_1 x_1 + \ldots + b_k x_k$ ⇒ k+1 parameters, k predictors ⇒ MSE = $s^2 = \frac{SSE}{n-(k+1)} = \frac{SSE}{n-k-1}$, $s = \sqrt{\frac{SSE}{n-k-1}}$
Smaller s ⇒ smaller intervals Adjusted R^2 ⇒ $\bar{R}^2$ ⇒ $\bar{R}^2 = \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-k-1}\right)$ ⇒ Penalize for parameters that are not significant
Overall F-test (Model as a whole) $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$ vs. H_a: at least one is not 0
$F = MSR/MSE = \frac{SSR/k}{SSE/(n-k-1)}$ reject if $F > F_\alpha^{k, n-k-1}$ or p-value < α
Individual T-test ⇒ significance of each variable ⇒ $H_0: \beta_j = 0$ vs. $H_a: \beta_j \neq 0$ ⇒ $t = \frac{b_j}{s_{b_j}}$, CL $b_j \pm t_{\alpha/2}^{n-k-1} s_{b_j}$
⇒ "We are $100(1-\alpha)\%$ confident that the interval contains the true value of β_j"
CL (Mean) ⇒ $\hat{y} \pm t_{\alpha/2}^{n-k-1} s\sqrt{DV}$ ⇒ We are $100(1-\alpha)\%$ confident that the interval contains the true mean value for the specific set of indep. variables $x_{0,1}, x_{0,2} \ldots x_{0,k}$
" (Predict) ⇒ $\hat{y} \pm t_{\alpha/2}^{n-k-1} s\sqrt{1+DV}$ ⇒ We are $100(1-\alpha)\%$ confident that the interval contains the individual observed value for the specific set of independent variables $x_{0,1}, x_{0,2} \ldots x_{0,k}$

Quadratic $y=\beta_0+\beta_1x+\beta_2x^2+e$ ⇒ use if a scatterplot shows a parabolic relationship.
β_0 ⇒ y-intercept β_1 ⇒ shift parameter ⇒ horizontal shift of the center of the parabola from 0.
If $\beta_1>0$ ⇒ center of the parabola is left of 0. → $\beta_1<0$ ⇒ right of 0 ... ⇒ $\beta_1\uparrow$ ⇒ center moves left
β_2 ⇒ Rate of curvature ⇒ width & direction of parabola $\beta_2>0$ ⇒ opens upward → $\beta_2\uparrow$ ⇒ parabola narrows
If quadratic term is significant ⇒ keep the linear term in the model, even if not significant

Interactive Term ⇒ Two or more explanatory var ⇒ have combined effect on dep. var ⇒ interactive term!
⇒ Plot the response variable against one of the explanatory variables for the different levels of the other explanatory variables. If an interaction term is significant, the two independent var should be included, no matter their p-values (interactive term between linear & quad terms is also possible)

Dummy Var ⇒ **Qualitative Var** ⇒ $y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon$ $x_2=\begin{cases}1 & \text{if category A}\\ 0 & \text{otherwise}\end{cases}$ $\mu_{y|x_2=0}=\beta_0+\beta_1x_1$, $\mu_{y|x_2=1}$ ⇒ $(\beta_0+\beta_2)+\beta_1x_1$ parallel
$x_2=0$ ⇒ baseline category if qualitative var. has m levels (m categories) ⇒ m-1 dummy var added

	x_2	x_3	
1	0	0	→ **Baseline**
2	1	0	
3	0	1	

$\mu_{y|x_2=0,x_3=0}=\beta_0+\beta_1x_1$
$\mu_{y|x_2=1,x_3=0}=(\beta_0+\beta_2)+\beta_1x_1$
$\mu_{y|x_2=0,x_3=1}=(\beta_0+\beta_3)+\beta_1x_1$.
⇒ Interactive terms may also be included

Partial F-test $H_0: \beta_{k_1}=\beta_{k_2}=\dots=\beta_{k}=0$
H_a: at least one of them is not zero
Full Regression Model ⇒ includes all k variables
Reduced ⇒ includes all variables except those being tested ⇒ so it has g variables in this model
Test-Stat ⇒ $F=\dfrac{(SSE_R-SSE_F)/(k-g)}{SSE_F/(n-k-1)}$ Reject if ⇒ $F>F_\alpha^{k-g,\,n-k-1}$
Ex) $H_0: \beta_2=\beta_3=0$ vs. H_a: at least one not zero Full model ⇒ $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon$ (k=3)
Reduced ⇒ $y=\beta_0+\beta_1x_1+\varepsilon$ (g=1)
$F=\dfrac{(SSE_R-SSE_F)/(2)}{SSE_F/(n-4)}$

Chap 5 Multicollinearity ⇒ Explanatory variables are related to or dependent upon each other
Correlation Matrix ⇒ Matrix that contains coefficient of correlation for the row, column variable.
VIF ⇒ Variance Inflation Factor ⇒ A measure of multicollinearity for each independent variable in the regression model $VIF_j=\frac{1}{1-R_j^2}$ where R_j^2 is the multiple coefficient of correlation for the regression model that uses x_j as the response variable and all of the remaining explanatory variables as the explanatory variables ⇒ Multicollinearity detected if VIF_j substantially greater than 1 (greater than 10 usually) ⇒ not accurate ⇒ unreliable model
⇒ Multicollinearity inflates variances of parameter estimates

Model Selection Criteria
⇒ Higher adjusted R^2, $\bar{R}^2$ ⇒ Low standard errors, Low Mallow's C_p (close to k+1)
$C_p=\frac{SSE_k}{s_p^2}-n+2k+2$ (s_p^2 is the MSE of the model including all p potential var and SSE_k is the SSE of the model including k of the p explanatory variables
⇒ significant variable added ⇒ decrease C_p
⇒ ideal value is k+1
⇒ substantially greater than k+1 is biased, not useful
C_p close to k+1 preferred to a small C_p significantly larger than k+1

Regression Analysis

① Exploratory Data Analysis
⇒ Scatterplots of y vs. all predictors
⇒ Correlation y vs. x
⇒ Multicollinearity
⇒ Outliers

② Decide on your hypothesized model
③ get the best model
→ model significance?
④ Assumptions
⑤ Transform if needed
⑥ Influential?

⑦ Prediction eq
⑧ Assess R^2 adj, S, F
⑨ Prediction

Ch 6 Time Series Regression
$y_t=TR_t+SN_t+\varepsilon_t$
$y_t=\beta_0+\beta_1t+\varepsilon_t$
increasing seasonal var.
$y_t=\ln(y_t)=y_t^*$
$y_t^*=\beta_0+\beta_1M_1+\dots+\beta_{11}M_{11}+\varepsilon_t$ (L=12)

Growth Curve
$y_t=\beta_0\beta_1^t\varepsilon_t$
$100(\beta_1-1)\%$ ⇒ growth rate

Ch 9-11
Step 1 Identify
→ stationary
→ look at SAC
→ first reg. diff
seasonal reg. diff
first seasonal diff
first reg. seasonal diff
→ Behavior of SAC/SPAC
⇒ what model?
⇒ Seasonal/Nonseasonal/combined?

Step 2 Estimation
$\hat{\delta}=\hat{\mu}(1-\hat{\phi}_1-\hat{\phi}_2\dots)$
$z_t=y_t-y_{t-1}$

Step 3 Diagnostics
parsimonious?
correlation?
Stationary?/Invertible?
Improve RSAC/RSPAC?
Random shock check

Chap 6 Notes Time Series Regression ⇒ ① The true value of the time series at time t is μ_t ② Error terms represent the average fluctuation around the time series model at t ③ Parameters will stay constant over time

3 Common Trends ① No trend ⇒ $TR_t=\beta_0$ (no long run growth/decline) ② Linear Trend ⇒ $TR_t=\beta_0+\beta_1t$ (constant long-run growth/decline)
③ Quadratic trend ⇒ $\beta_0+\beta_1t+\beta_2t^2$ ⇒ changing long-run growth/decline increasing/decreasing rate growth/decline
Check assumptions using residual vs. time and normal prob. plot
Autocorrelation ⇒ indep. assumption violated ⇒ Positive autocorrelation ⇒ positive (negative) error term tends to be followed by a positive (negative) error term, Negative autocorrelation ⇒ positive (negative) error term