Statistics 5350/7110
Forecasting

Lecture 20
Regression with ARIMA Errors

Professor Robert Stine

Preliminaries

- Questions

- Assignments
  - Next coming Thursday

- Quick review
  - Examples of
    - Model drift
    - Controlling outliers
  - Examples of SARIMA models — Lecture_19.Rmd
    - Births
    - CO_2

Today's Topics

Text, §5.4

- Autocorrelation in regression errors
  - Violating the key MRM assumption
  - Consequences
  - Durbin-Watson test

- Regression with seasonal terms
  - Dummy variable model compared to SARIMA process
  - Rigid vs fluctuating seasonal patterns

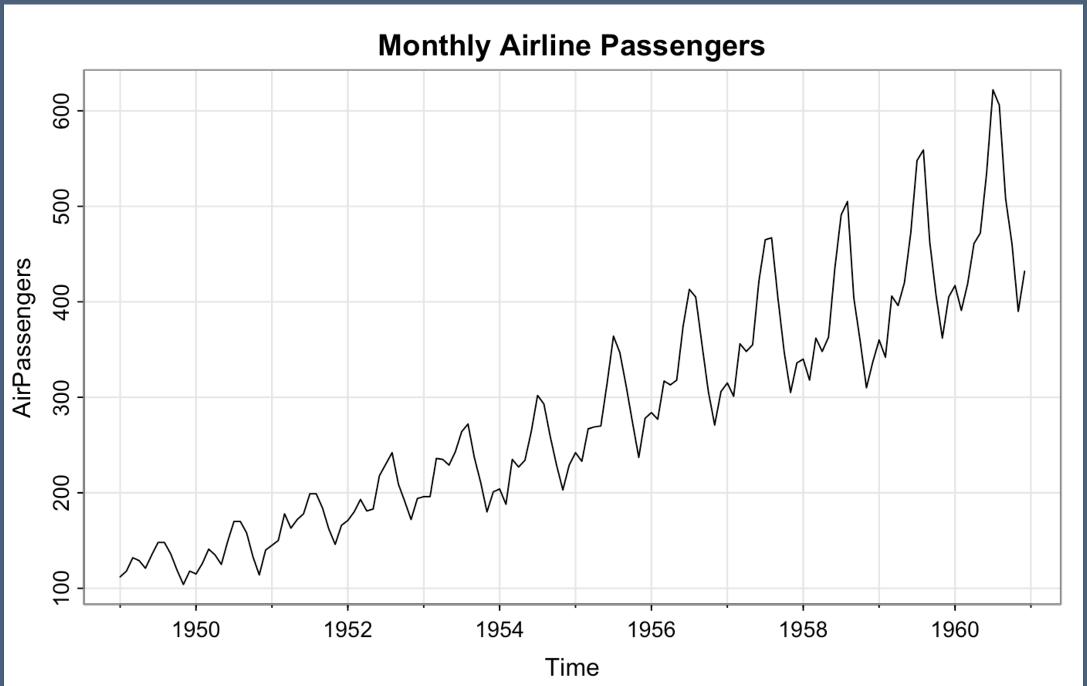
- Calendar effects in monthly data

Seasonal Regression

Regression with autocorrelated errors

Seasonal Regression Models

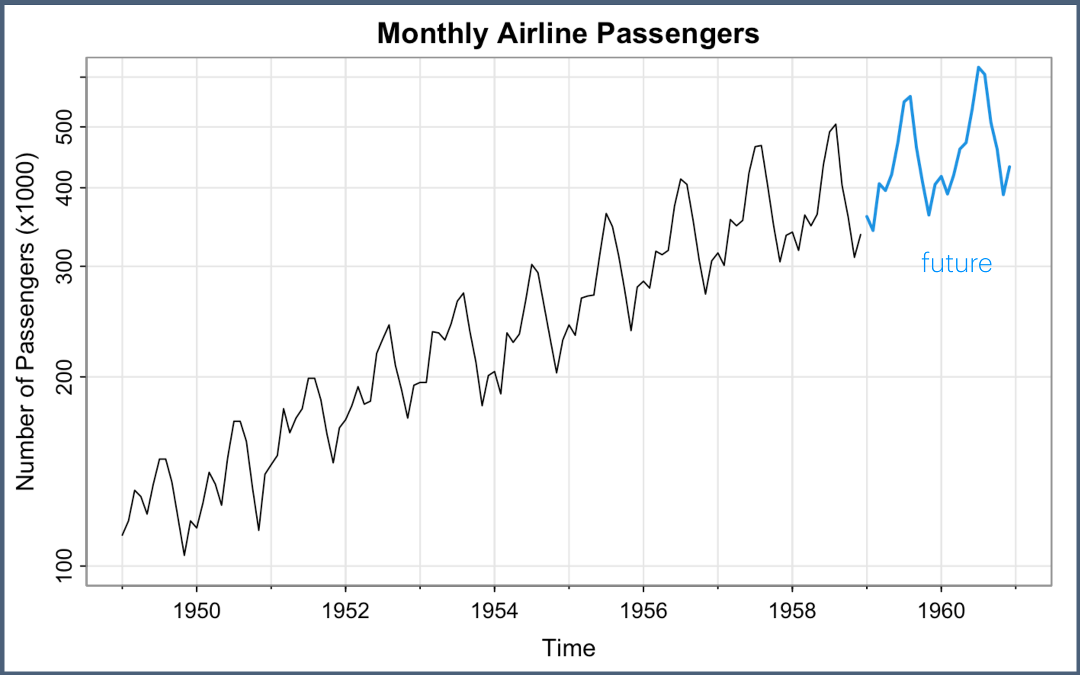
- Classical example: monthly airline passengers
  - Appears in many texts, originally Box & Jenkins (1976)
  - Data included in R (AirPassengers)
  - Combines strong growth with seasonal variation



Useful to transform the counts before continuing to “uncouple” mean level from variance.

Seasonal Regression Models

- Classical example: monthly airline passengers
  - Appears in many texts, originally Box & Jenkins (1976)
  - Data included in R (AirPassengers)
  - Combines strong growth with seasonal variation



See a log scale with original units rather than displaying log(x) units

Regression with Seasonal Terms

- Dummy variable representation
  - Assuming a linear trend, then

$$Y_t = \beta_0 + \beta_1\, t + \sum_{j=2}^{12} \beta_j M_j(t) + w_t, \quad \text{where} \quad M_j \in \{0,1\}$$

  - Dummy coding for months

$$M_j(t) = \begin{cases} 1 & \text{if } (t \bmod 12) = j, \\ 0 & \text{otherwise.} \end{cases}$$

- Interpreting estimated coefficients
  - Trend coefficient estimates annual growth rate (≈ 12%)
  - Intercept for baseline (omitted) month (January)
  - Coefficients of dummy variables shift the fit

    Lower in February

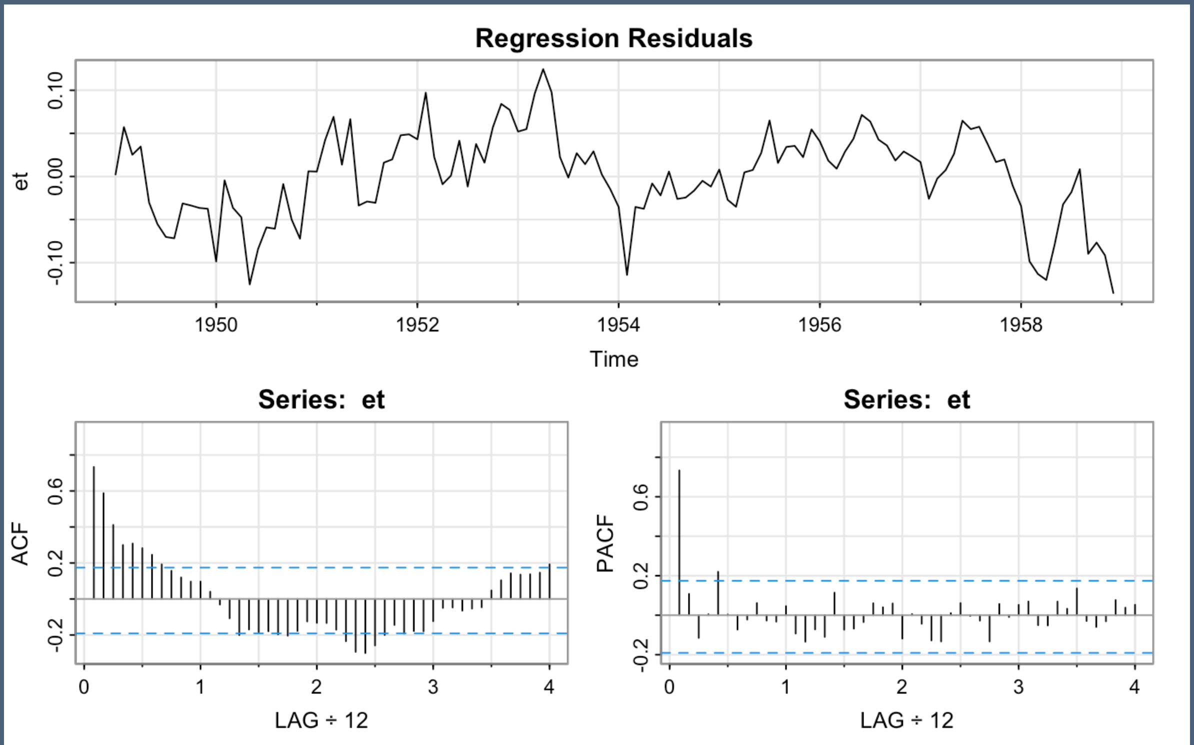
    Peak in July - August

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.437e+02  3.401e+00 -71.660  < 2e-16 ***
trend        1.275e-01  1.741e-03  73.217  < 2e-16 ***
monthFeb    -1.344e-02  2.450e-02  -0.549  0.58446    
monthMar     1.202e-01  2.450e-02   4.907 3.32e-06 ***
monthApr     7.710e-02  2.450e-02   3.147  0.00214 ** 
monthMay     6.747e-02  2.450e-02   2.754  0.00693 ** 
monthJun     1.913e-01  2.451e-02   7.806 4.23e-12 ***
monthJul     2.875e-01  2.451e-02  11.729  < 2e-16 ***
monthAug     2.784e-01  2.452e-02  11.356  < 2e-16 ***
monthSep     1.428e-01  2.452e-02   5.823 6.13e-08 ***
monthOct     1.081e-03  2.453e-02   0.044  0.96493    
monthNov    -1.415e-01  2.454e-02  -5.765 7.97e-08 ***
monthDec    -2.483e-02  2.455e-02  -1.012  0.31406    
---
```

```
Residual standard error: 0.05478 on 107 degrees of freedom
Multiple R-squared:  0.9824,	Adjusted R-squared:  0.9804 
F-statistic: 498.2 on 12 and 107 DF,  p-value: < 2.2e-16
```

Residual Autocorrelation

- Correlation evident in residuals from the regression
  - AR(1) pattern in ACF/PACF



- Test statistic
  - Durbin-Watson test statistic related to corr(e_t, e_{t-1}):

$$D = 2\,(1 - \hat{\rho})$$

  - Recognizes that data are residuals rather than direct observations.
  - Autocorrelation 0.73 is highly significant

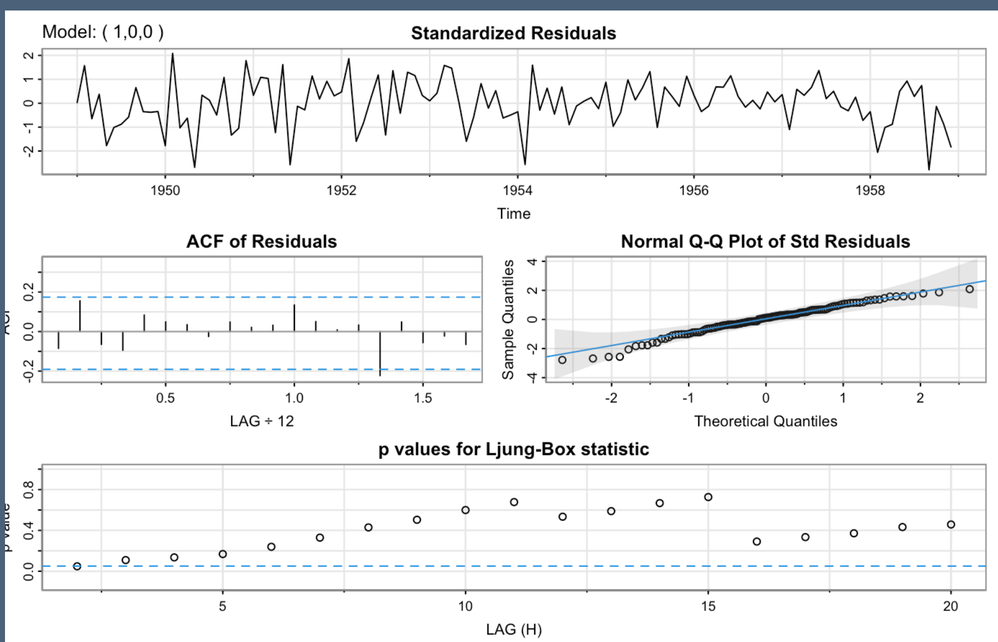
Autocorrelation in Regression

- Serious problem
  - Don’t ignore it!
- Biases test statistics
  - Unbiased estimated coefficients, but standard errors are biased.
  - Correlation implies less information (using the “wrong n”)
  - Positive autocorrelation biases t-statistics of estimated slopes
  - Size of bias depends on autocorrelation in explanatory variables
- Opportunity for better predictions
  - Incorporate structure into model
  - RMSE in regression is 0.055… We can do better
- Remedies
  - Ideally, identify omitted explanatory variables
  - Practically, model residuals as ARMA process

See Maddala, Chap 6

Revised Model

- Incorporate residual correlation
  - Model residuals as AR(1)
  - Similar estimated coefficients, smaller RMSE (≈ 0.034 compared to 0.055)
  - Fit model using `sarima` with regression components
  - Diagnostics look much better



Coefficients:

	Estimate	SE	t.value
ar1	0.7771	0.0758	10.2579
intercept	-238.9484	8.7972	-27.1619
trend	0.1250	0.0045	27.6669
monthFeb	-0.0110	0.0114	-0.9633
monthMar	0.1246	0.0152	8.2019
monthApr	0.0831	0.0175	4.7598
monthMay	0.0747	0.0188	3.9706
monthJun	0.1996	0.0195	10.2140
monthJul	0.2966	0.0197	15.0271
monthAug	0.2883	0.0195	14.8158
monthSep	0.1532	0.0187	8.2069
monthOct	0.0120	0.0173	0.6954
monthNov	-0.1301	0.0152	-8.5835
monthDec	-0.0130	0.0118	-1.1065

OLS

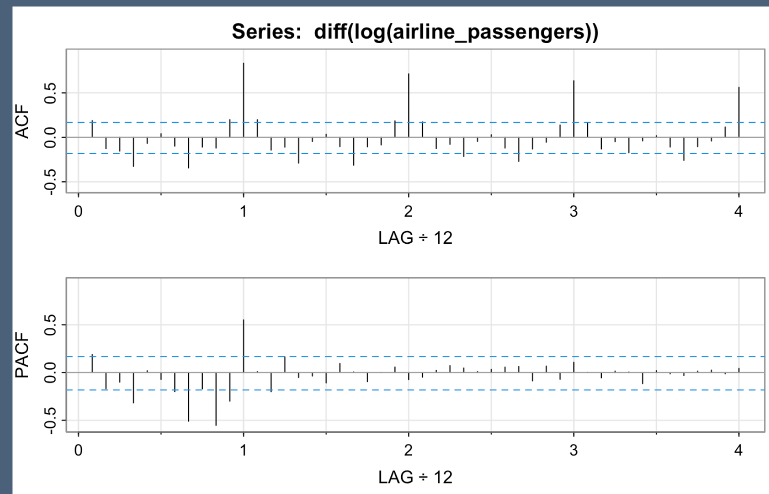
	Estimate
(Intercept)	-243.711
trend	0.127
monthFeb	-0.013
monthMar	0.120
monthApr	0.077
monthMay	0.067
monthJun	0.191
monthJul	0.288
monthAug	0.278
monthSep	0.143
monthOct	0.001
monthNov	-0.141
monthDec	-0.025

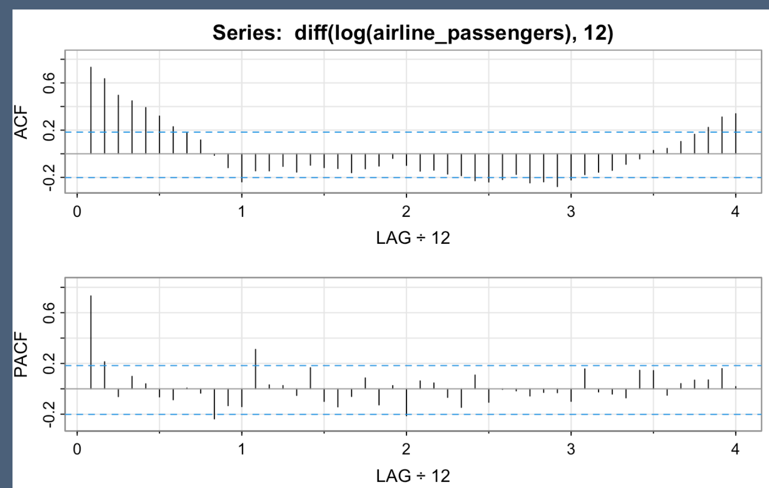
Seasonal ARIMA

Replace monthly regressors and trend with differences

Seasonal Time Series

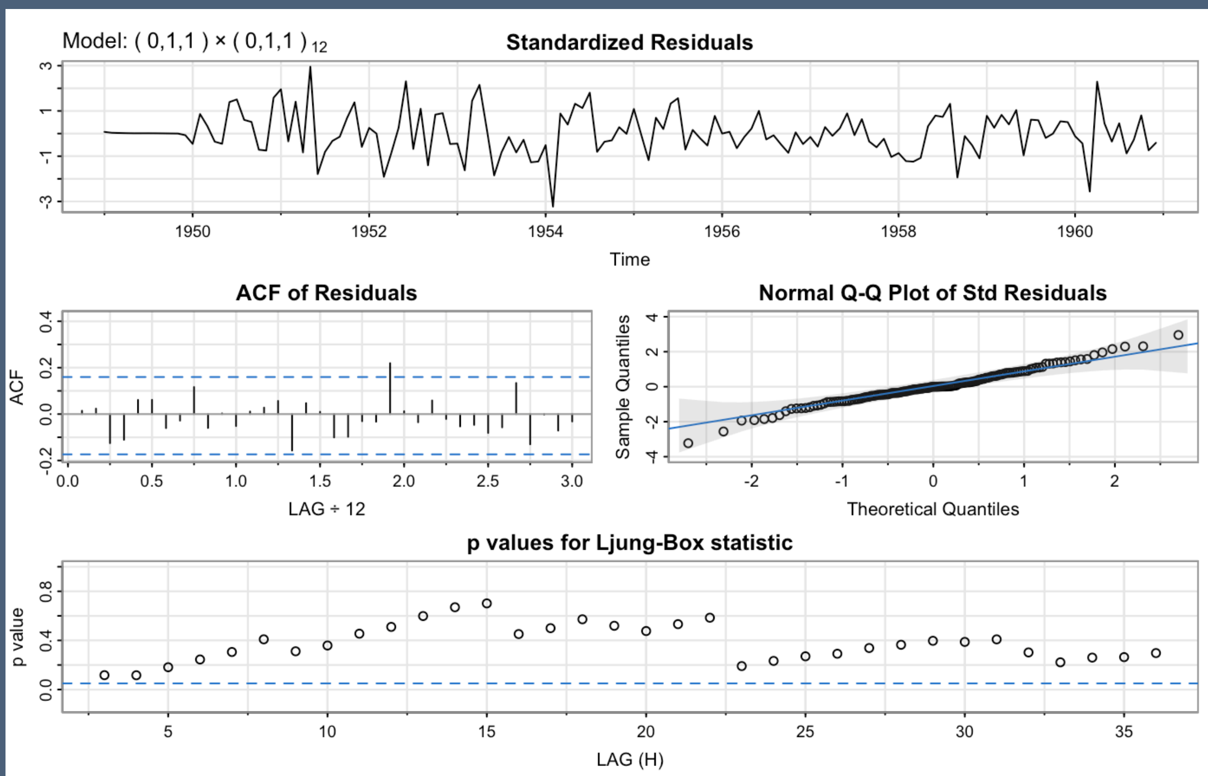
- Alternative approach
  - Rather than fit a trend and dummy variables, fit a seasonal time series model with differencing
- Model selection
  - Difference at lags 1 and 12
  - Strong seasonal structure (AR), other less clear
- Iterative modeling
  - Try several models, starting with SARIMA$(1,1,1)(1,1,1)_{12}$, then revise
  - Model selection tools don't recognize multiplicative structure of SARIMA style models.





Seasonal Time Series

- Parsimonious model
  - Try several, end up with SARIMA(0,1,1)(0,1,1)$_{12}$
  - Residual diagnostics look okay, perhaps changing variance




```
Coefficients:
     Estimate     SE t.value p.value
ma1   -0.3424 0.1009 -3.3925   0.001
sma1  -0.5405 0.0877 -6.1626   0.000

sigma^2 estimated as 0.001402458 on 105
```

sqrt(0.0014) ≈ 0.0037

Why 105?

Details of SARIMA Model

- Multiplicative
  - Model is

$$(1-B)(1-B^{12})\,X_t = (1-0.34B)(1-0.54B^{12})\,w_t$$

  - Expanded becomes MA(13) for differences

$$(1-B)(1-B^{12})\,X_t = (1-0.34B-0.54B^{12}+0.18B^{13})\,w_t$$

```
Coefficients:
     Estimate     SE t.value p.value
ma1   -0.3424 0.1009 -3.3925   0.001
sma1  -0.5405 0.0877 -6.1626   0.000

sigma^2 estimated as 0.001402458 on 105
```

- Fit model directly
  - Estimates from multiplicative model are similar
  - Loss of efficiency (larger SEs)
  - Numerous non-significant estimates

```
Coefficients:
     Estimate     SE t.value
ma1   -0.3476 0.1487 -2.3373
ma2    0.0701 0.1331  0.5268
ma3   -0.1786 0.1338 -1.3347
ma4   -0.1973 0.1241 -1.5900
ma5    0.1105 0.1114  0.9920
ma6   -0.0244 0.1201 -0.2029
ma7    0.0186 0.1313  0.1419
ma8   -0.0978 0.1358 -0.7199
ma9    0.0207 0.1204  0.1722
ma10   0.0120 0.1068  0.1122
ma11   0.1003 0.1089  0.9214
ma12  -0.6465 0.1127 -5.7366
ma13   0.1602 0.1320  1.2135
```

```
sigma^2 estimated as 0.001196359 on 94
```

- Constrained
  - Force 0 at same locations as ARIMA

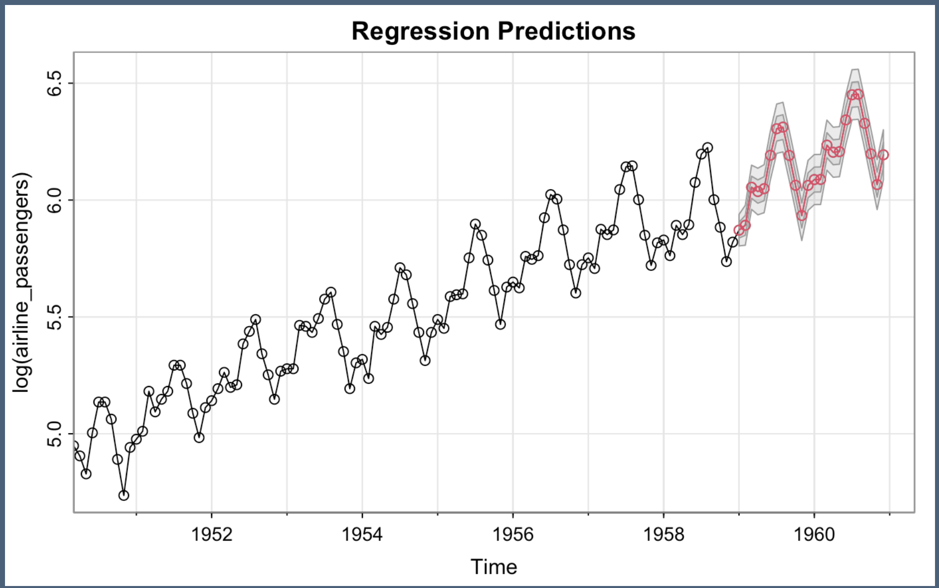
```
Coefficients:
     Estimate     SE t.value p.value
ma1   -0.3419 0.0945 -3.6168  0.0005
ma12  -0.6032 0.1058 -5.7011  0.0000
ma13   0.3248 0.1284  2.5289  0.0129

sigma^2 estimated as 0.001359599 on 104
```

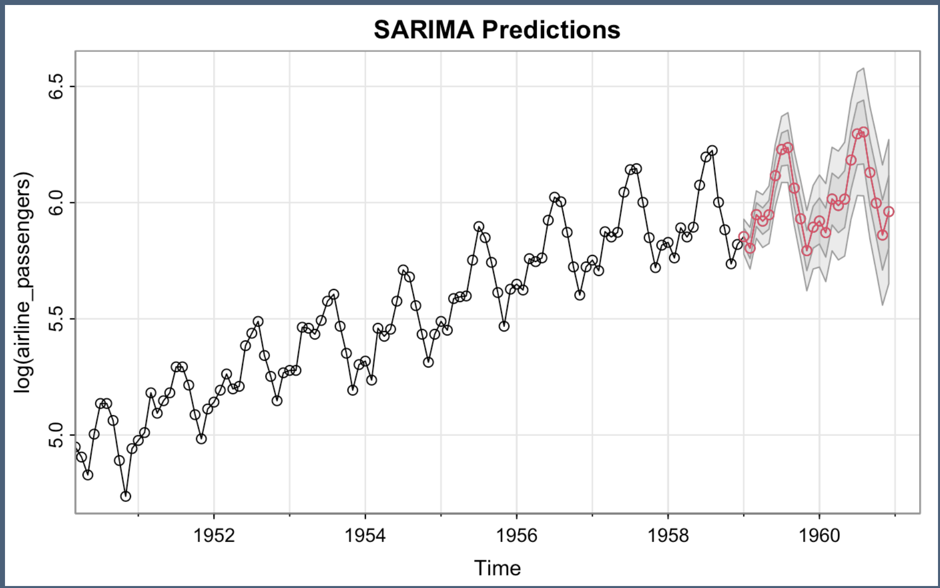
Forecast Comparison

Forecasts from Seasonal Models

- Forecast from two seasonal differences model
  - Regression with fixed seasonal coefficients (dummy variables)
  - SARIMA with differences
- How do the predictions differ?



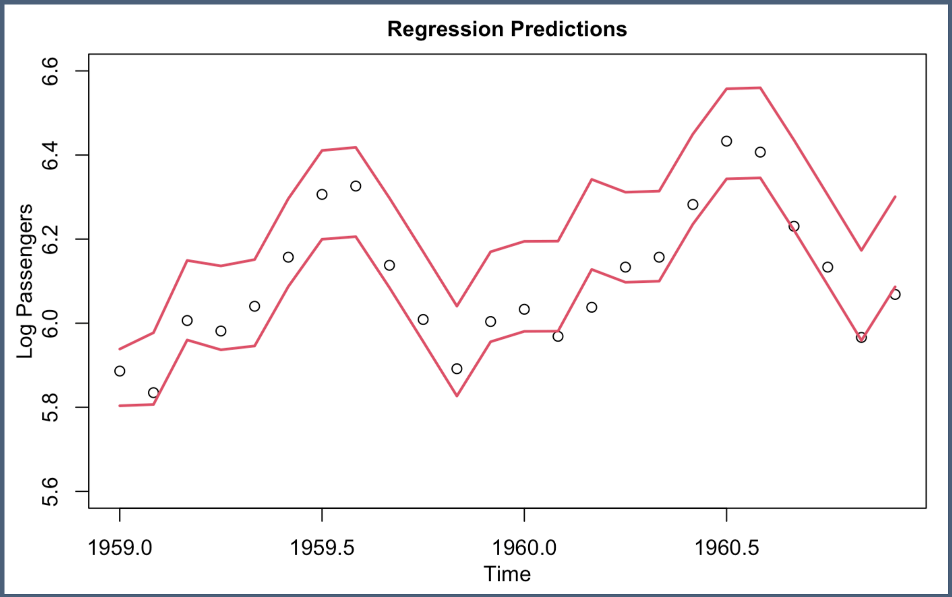
Regression model with explicit trend and seasonal dummies

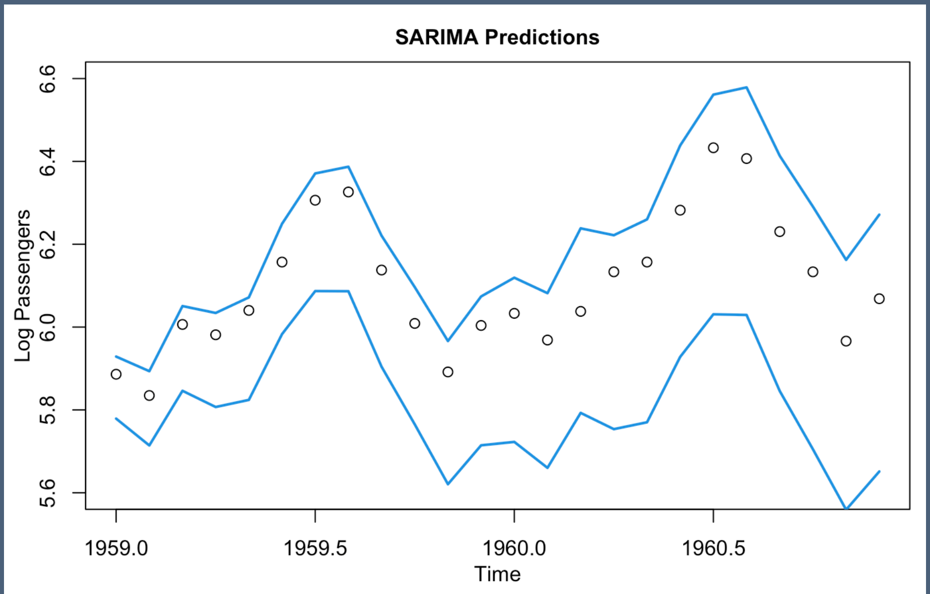


Seasonal arima model with differencing at d=1, D=12

Closer Comparison

- Forecasts on common scales
  - Interval is forecast ± two standard errors
  - Regression bounds are tighter, fixed length (after initial few)
  - SARIMA intervals open to become far longer as extrapolate

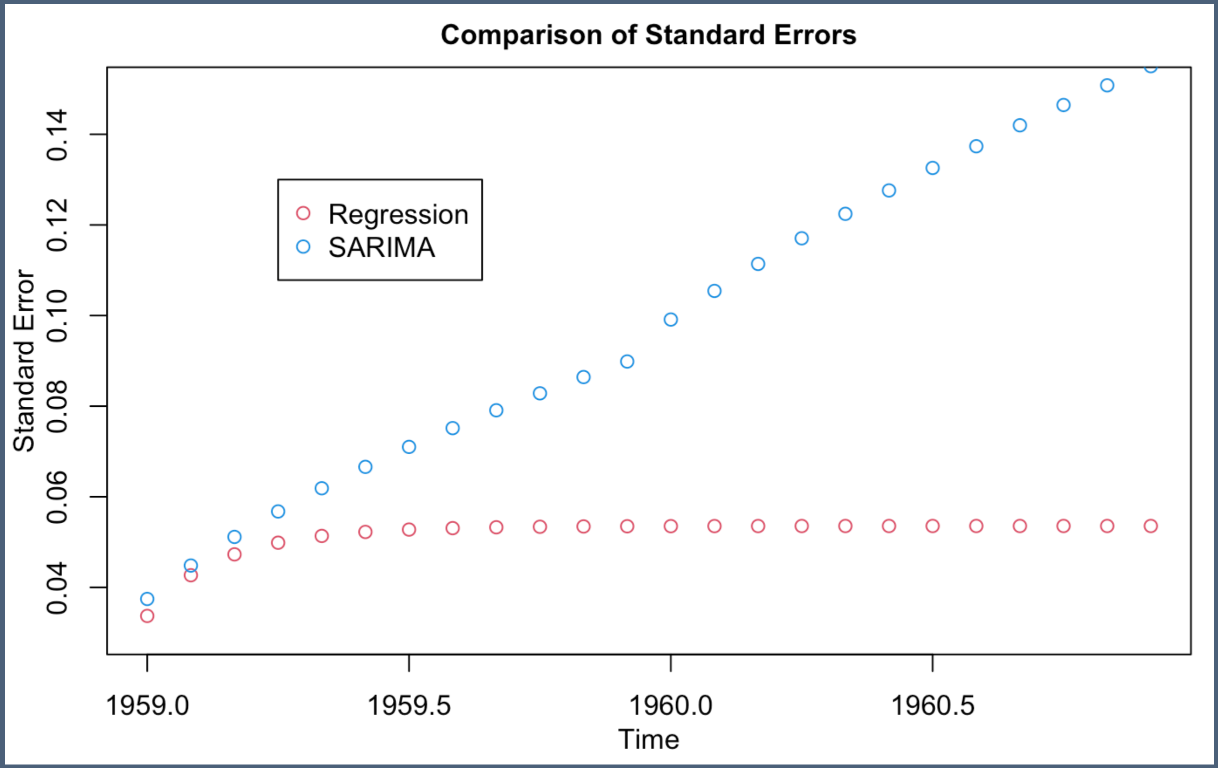




Comparison of Standard Errors

- Regression
  - SE constant after AR(1) component used to model residual correlation
  - Too optimistic that linear trend continues

- ARIMA
  - Integrating differences leads to steadily increasing SE
  - Adjusts trend estimate based on intervening year

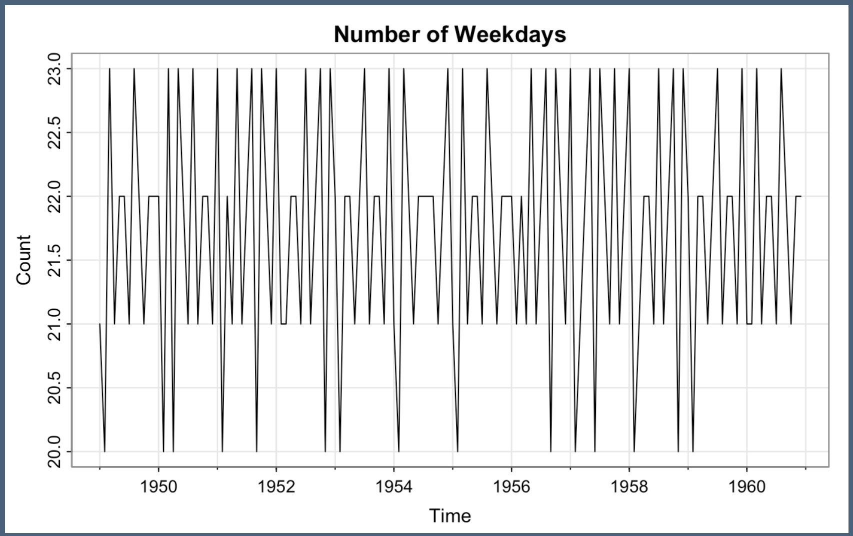


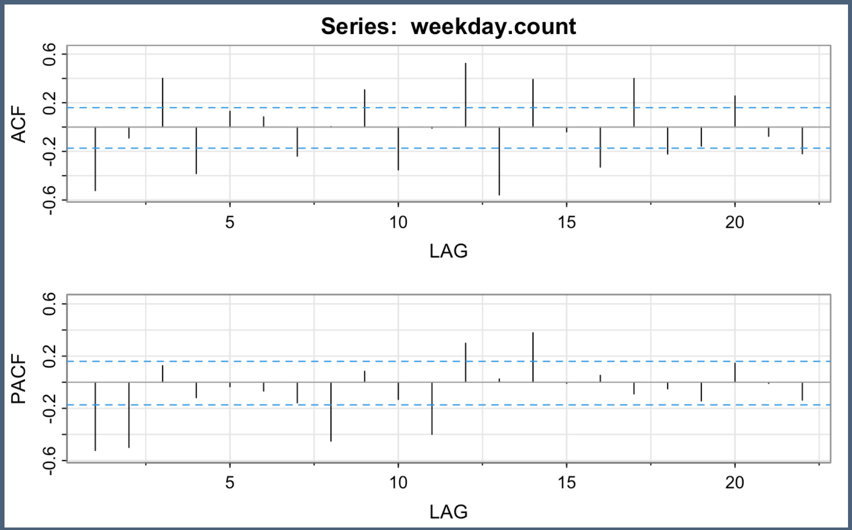
Calendar Effects

Improve both types of models

Calendar Effects in Monthly Time Series

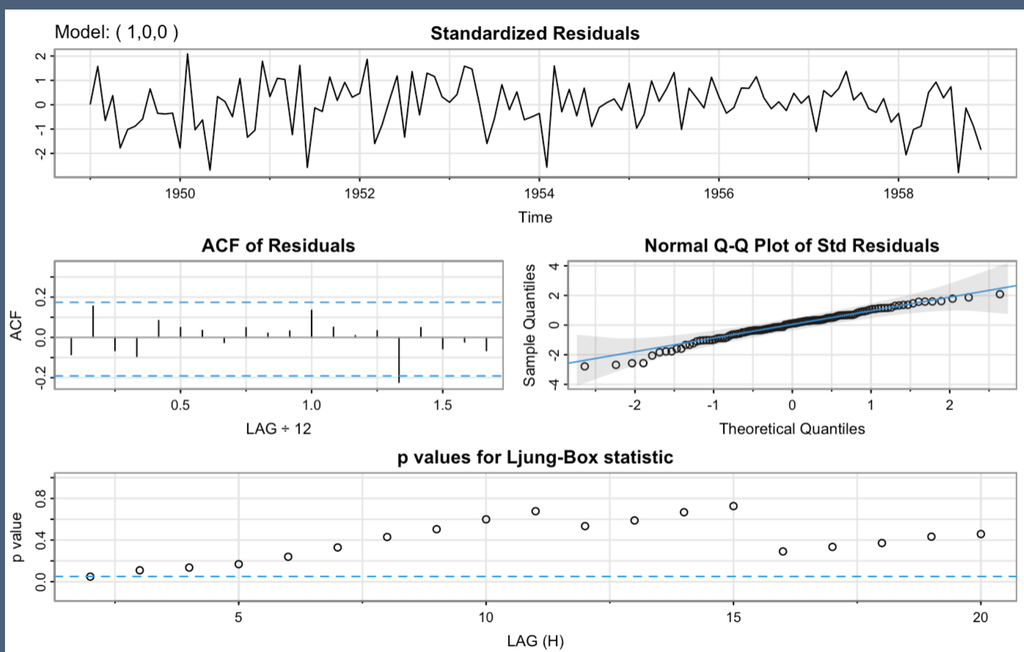
- Specific to monthly time series
  - Some months are longer than others (fewer days for production in February than March)
  - Effects most evident in aggregated time series (stock variables)
- Example
  - Varying number of business days in March, say, from year to year.
  - If omit this feature from regression, its structure remains in the residuals





Expanded Regression

- Another predictor
  - Add the weekday count to the prior regression
  - Significant improvement, slight change in other coefficients (except for February, April)



	Estimate	SE	t.value
ar1	0.7972	0.0765	10.4185
intercept	-240.3132	9.0552	-26.5388
weekend.count	0.0080	0.0024	3.3426
month.count	0.0381	0.0197	1.9376
trend	0.1251	0.0046	26.9508
monthFeb	0.1027	0.0563	1.8257
monthMar	0.1242	0.0144	8.6546
monthApr	0.1231	0.0258	4.7659
monthMay	0.0749	0.0179	4.1870
monthJun	0.2395	0.0272	8.8029
monthJul	0.2959	0.0188	15.7389
monthAug	0.2884	0.0185	15.5781
monthSep	0.1930	0.0266	7.2646
monthOct	0.0119	0.0164	0.7297
monthNov	-0.0905	0.0243	-3.7185
monthDec	-0.0141	0.0110	-1.2809

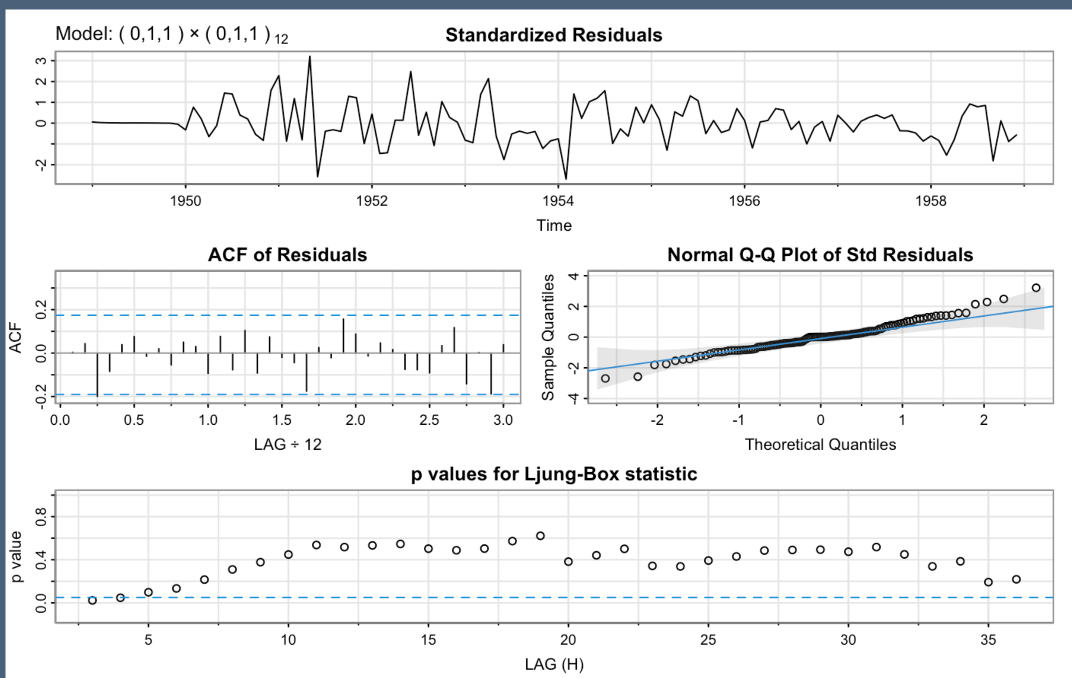
prior model

0.1250
-0.0110
0.1246
0.0831
0.0747
0.1996
0.2966
0.2883
0.1532
0.0120
-0.1301
-0.0130

```
sigma^2 estimated as 0.001011807 on 104
```

Augment SARIMA

- Use calendar features as well
  - Significant effects
  - Some residual correlation remains
  - Error variance estimate larger than obtained by regression




```
Coefficients:
              Estimate     SE t.value p.value
ma1            -0.2744 0.1022 -2.6852  0.0085
sma1           -0.5262 0.0819 -6.4240  0.0000
weekend.count   0.0076 0.0025  3.0048  0.0033
month.count     0.0473 0.0178  2.6622  0.0090

sigma^2 estimated as 0.001230245 on 103 degrees
```

What’s next?

- More modeling examples
  - Examples here have non-stochastic explanatory variables
    - linear trend
    - dummies
    - other calendar features
  - Two or more stochastic time series as predictors