Section 4.1: Time Series I

Jared S. Murray
The University of Texas at Austin
McCombs School of Business

Time Series Data and Dependence

Time-series data are simply a collection of observations gathered over time. For example, suppose $y_1 \dots y_T$ are

- Annual GDP.
- Quarterly production levels
- Weekly sales.
- Daily temperature.
- 5 minute stock returns.

In each case, we might expect what happens at time t to be correlated with what happens at time $t-1$.

Time Series Data and Dependence

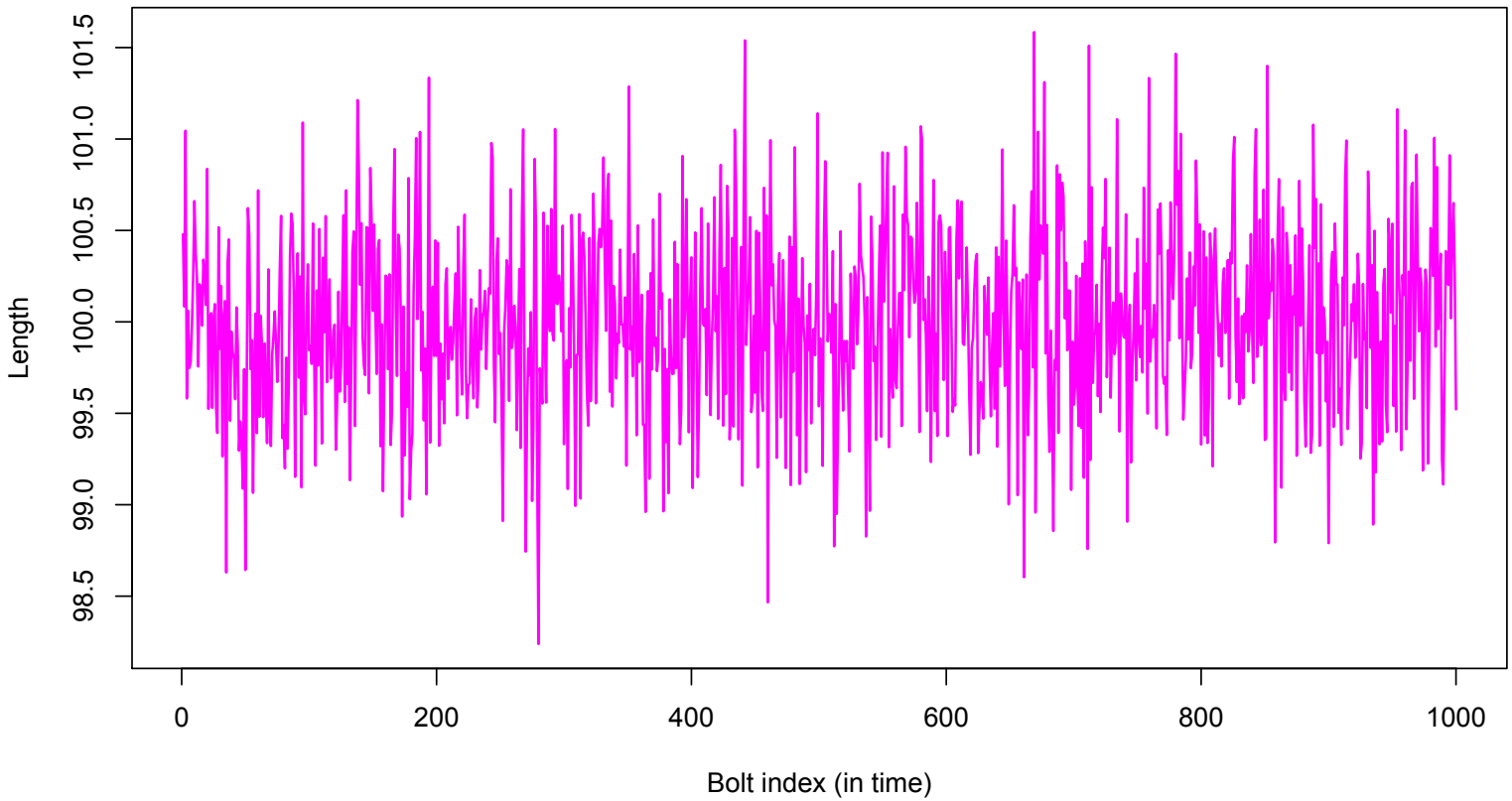
Suppose we measure temperatures daily for several years.

Which would work better as an estimate for today's temp:

- The average of the temperatures from the previous year?
- The temperature on the previous day?

Example: Length of a bolt...

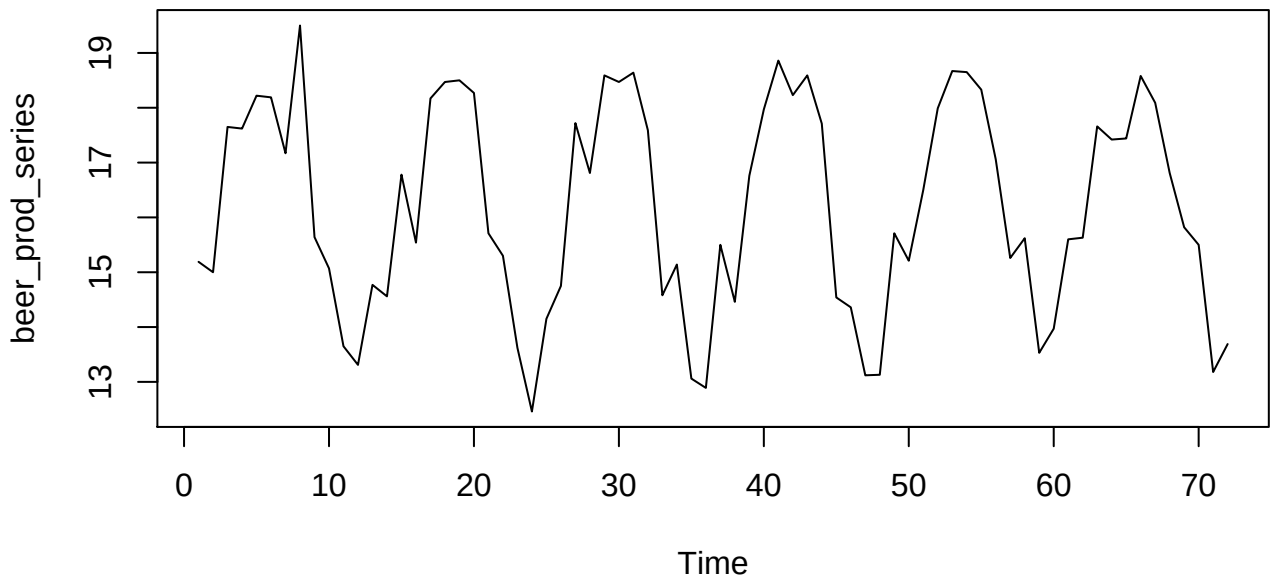
Suppose you have to check the performance of a machine making bolts... in order to do so you want to predict the length of the next bolt produced...



What is your best guess for the next part?

Example: Beer Production

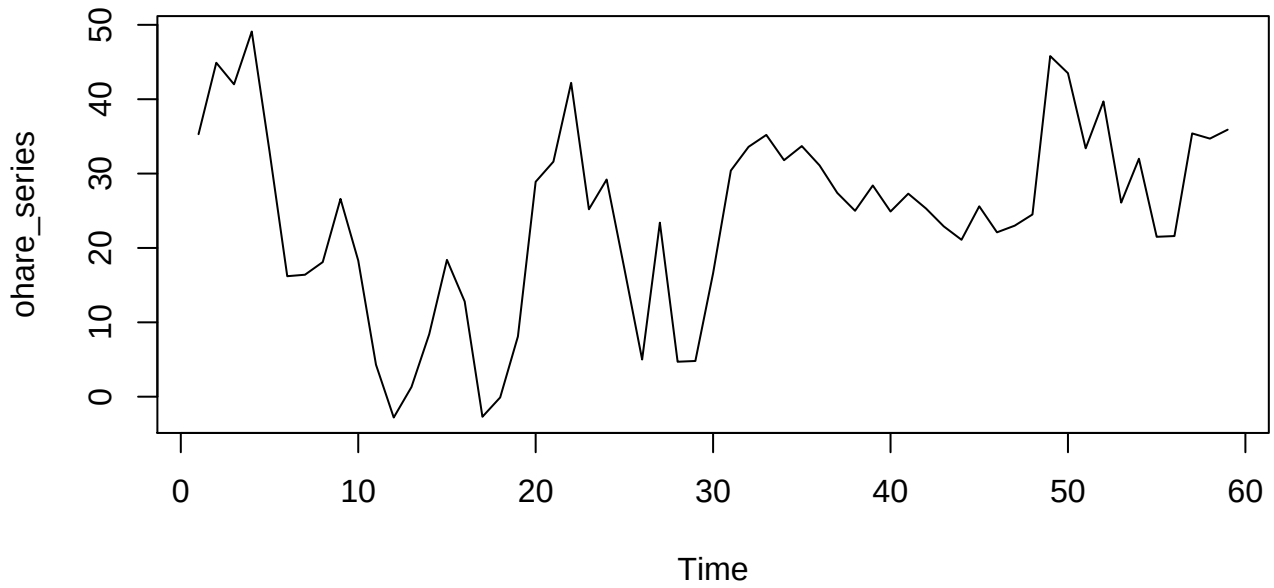
Now, say you want to predict the monthly U.S. beer production (in millions of barrels).



What about now, what is your best guess for the production in the next month?

Examples: Temperatures

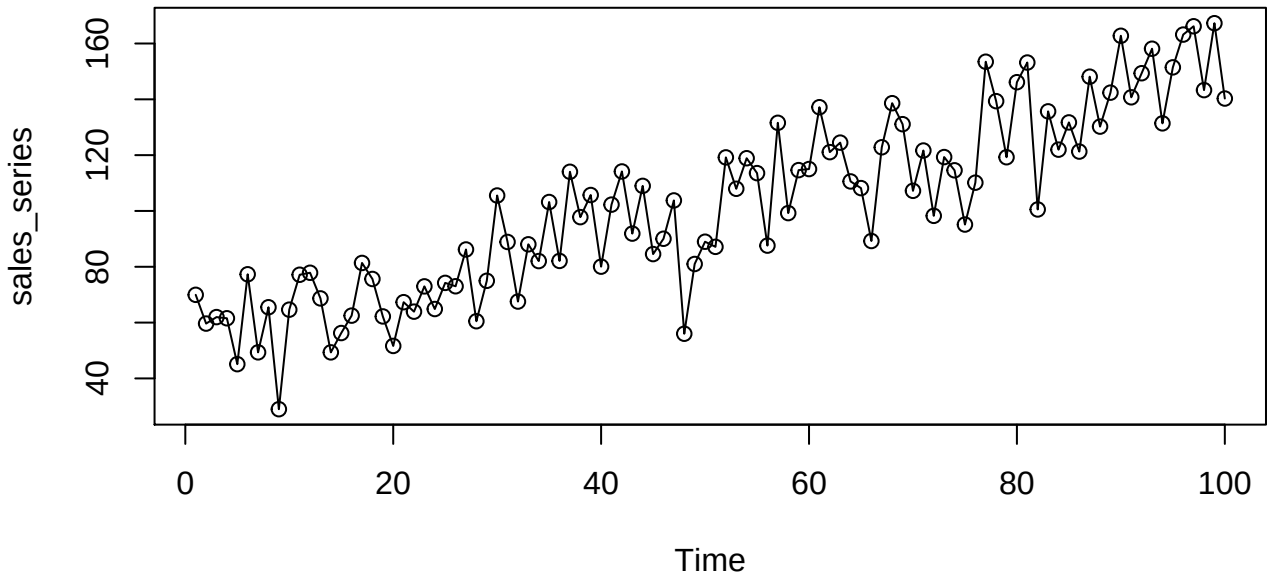
Now you need to predict the temperature on March 1 at O'Hare using data from Jan-Feb.



Is this one harder? Our goal in this section is to use regression models to help answer these questions...

Fitting a Trend

Here's a time series plot of monthly sales of a company...



What would be a reasonable prediction for Sales 5 months from now?

Fitting a Trend

The sales numbers are "trending" upwards... What model could capture this trend?

$$S_t = \beta_0 + \beta_1 t + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma^2)$$

This is a regression of Sales (y variable) on "time" (x variable).
This allows for shifts in the mean of Sales as a function of time.

Fitting a Trend

The data for this regression looks like:

months(t)	Sales
1	69.95
2	59.64
3	61.96
4	61.55
5	45.10
6	77.31
7	49.33
8	65.49
...	...
100	140.27

Fitting a Trend

$$S_t = \beta_0 + \beta_1 t + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma^2)$$

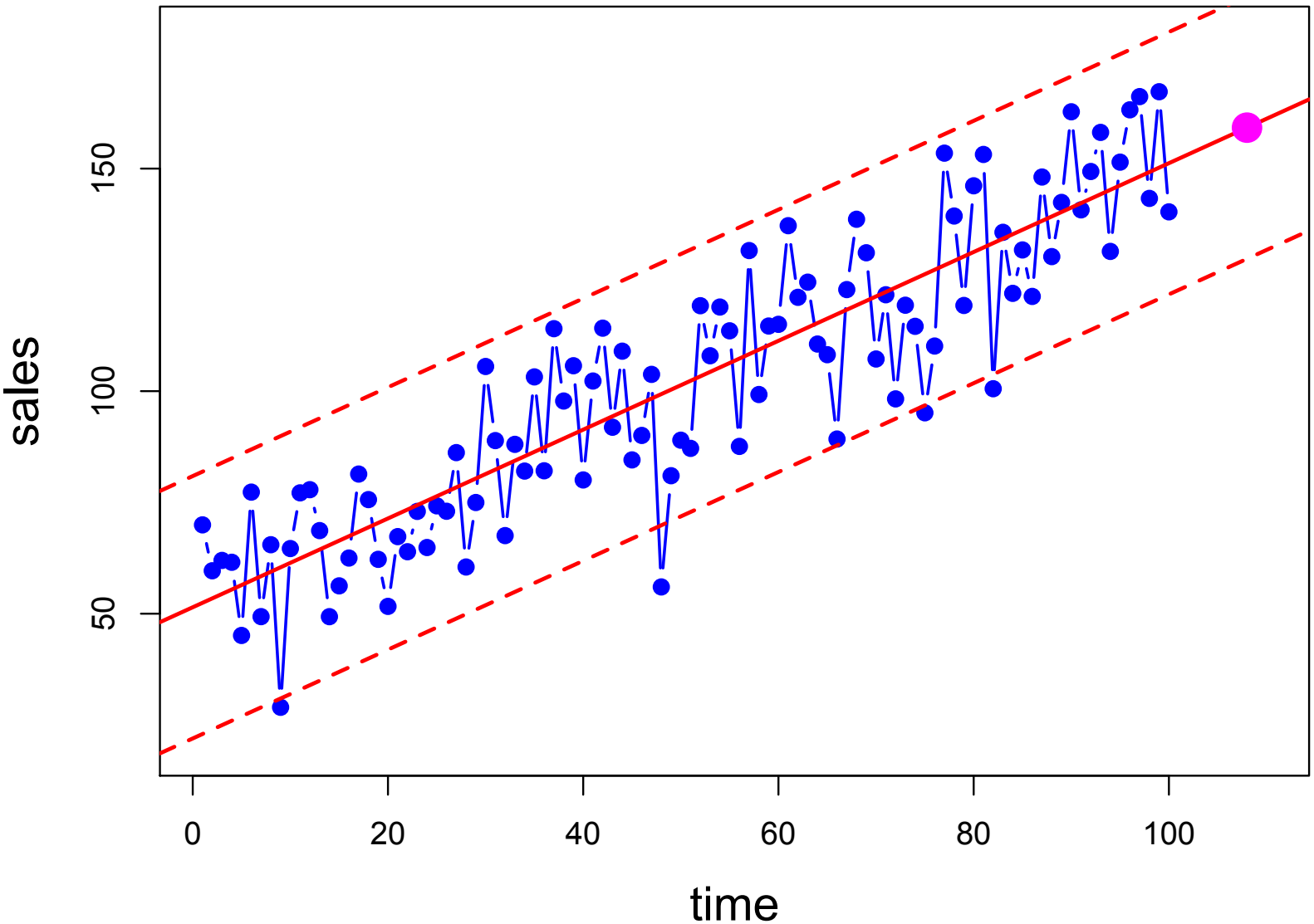
```
library(forecast)
sales_fit = tslm(sales_series~trend)
print(sales_fit)

##
## Call:
## tslm(formula = sales_series ~ trend)
##
## Coefficients:
## (Intercept)        trend
##     51.4419       0.9978
```

$$\hat{S}_t = 51.44 + 0.998t$$

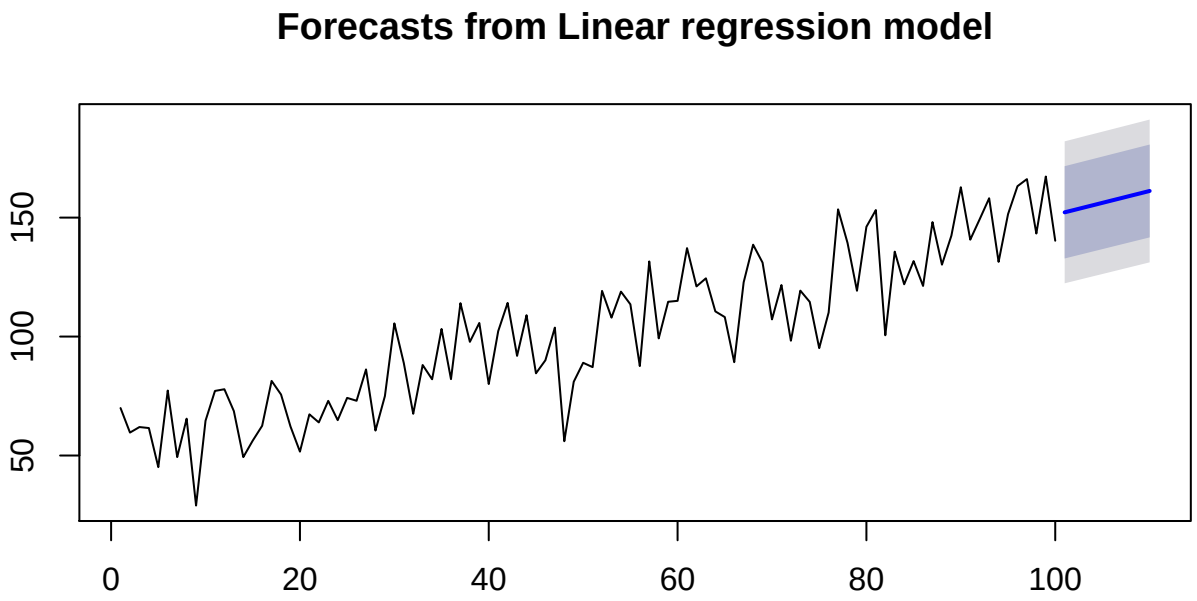
Fitting a Trend

Plug-in prediction...



Fitting a Trend

```
sales_pred = forecast(sales_fit, h=10)
plot(sales_pred)
```

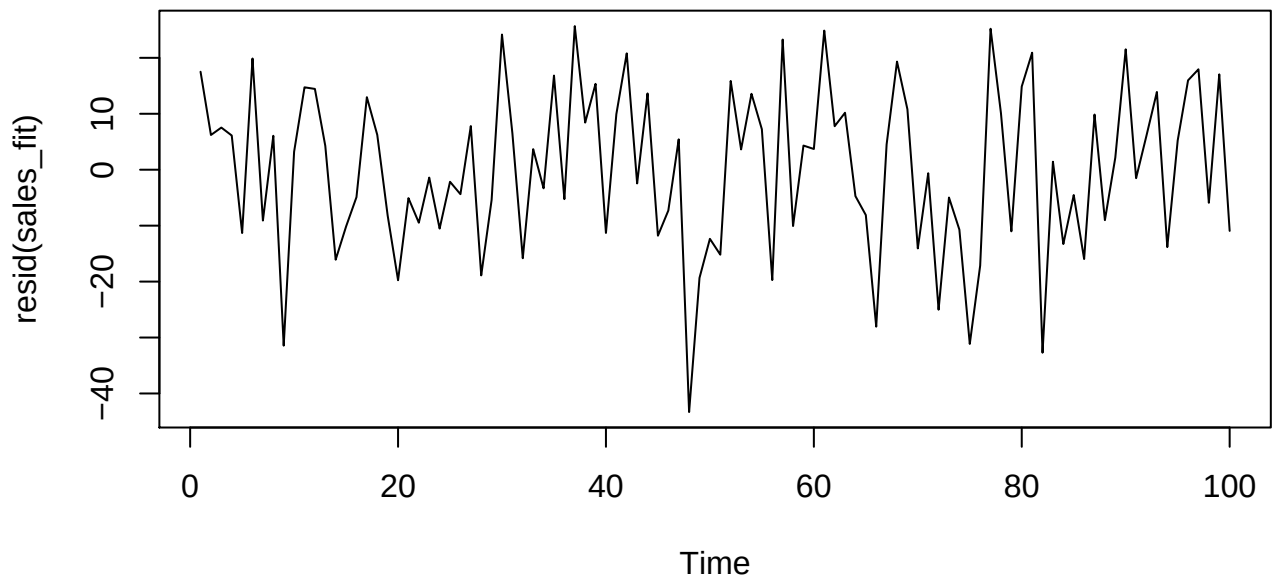



```
print(sales_pred)

##     Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 101       152.2150 132.8183 171.6117 122.3819 182.0481
## 102       153.2128 133.8047 172.6209 123.3621 183.0634
```

Residuals

How should our residuals look? If our model is correct, the trend should have captured the time series structure is sales and what is left, should not be associated with time... i.e., it should be iid normal.



Great!

Time Series Regression... Hotel Occupancy Case

In a recent legal case, a Chicago downtown hotel claimed that it had suffered a loss of business due to what was considered an illegal action by a group of hotels that decided to leave the plaintiff out of a hotel directory.

In order to estimate the loss business, the hotel had to predict what its level of business (in terms of occupancy rate) would have been in the absence of the alleged illegal action.

In order to do this, experts testifying on behalf of the hotel use data collected before the period in question and fit a relationship between the hotel's occupancy rate and overall occupancy rate in the city of Chicago. This relationship would then be used to predict occupancy rate during the period in question.

Example: Hotel Occupancy Case

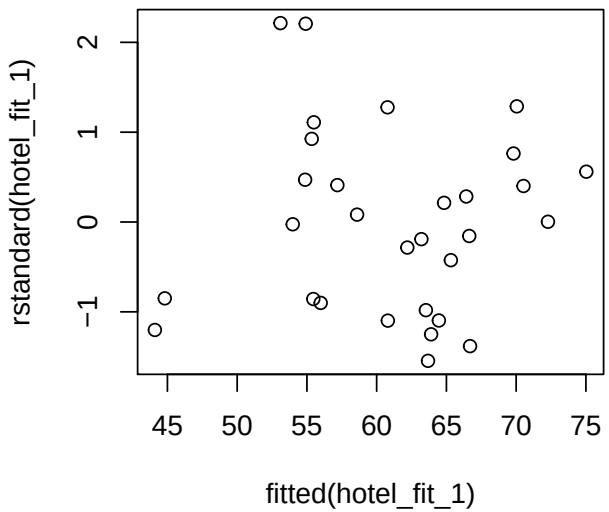
$$Hotel_t = \beta_0 + \beta_1 Chicago + \epsilon_t$$

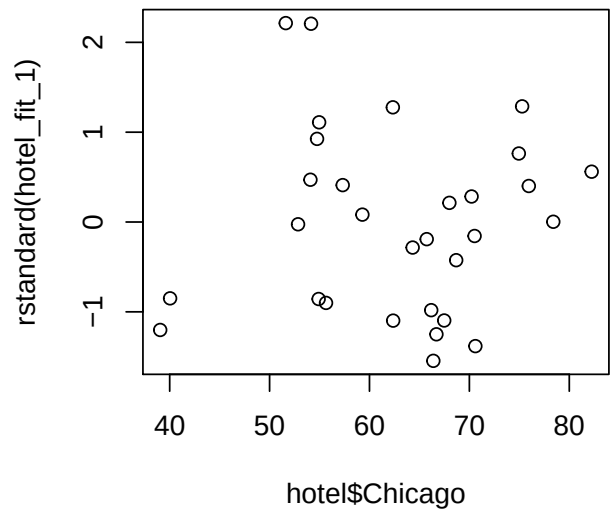
```
##
## Call:
## lm(formula = Hotel ~ Chicago, data = hotel)
##
## Coefficients:
## (Intercept)      Chicago
##     16.1357       0.7161
```

- In the month after the omission from the directory the Chicago occupancy rate was 66%. The plaintiff claims that its occupancy rate should have been 16 + 0.71*66 = 62%.
- It was actually 55%!! The difference added up to a big loss!!

Example: Hotel Occupancy Case

A statistician was hired by the directory to access the regression methodology used to justify the claim. As we should know by now, the first thing he looked at was the residual plot...





Looks fine. However...

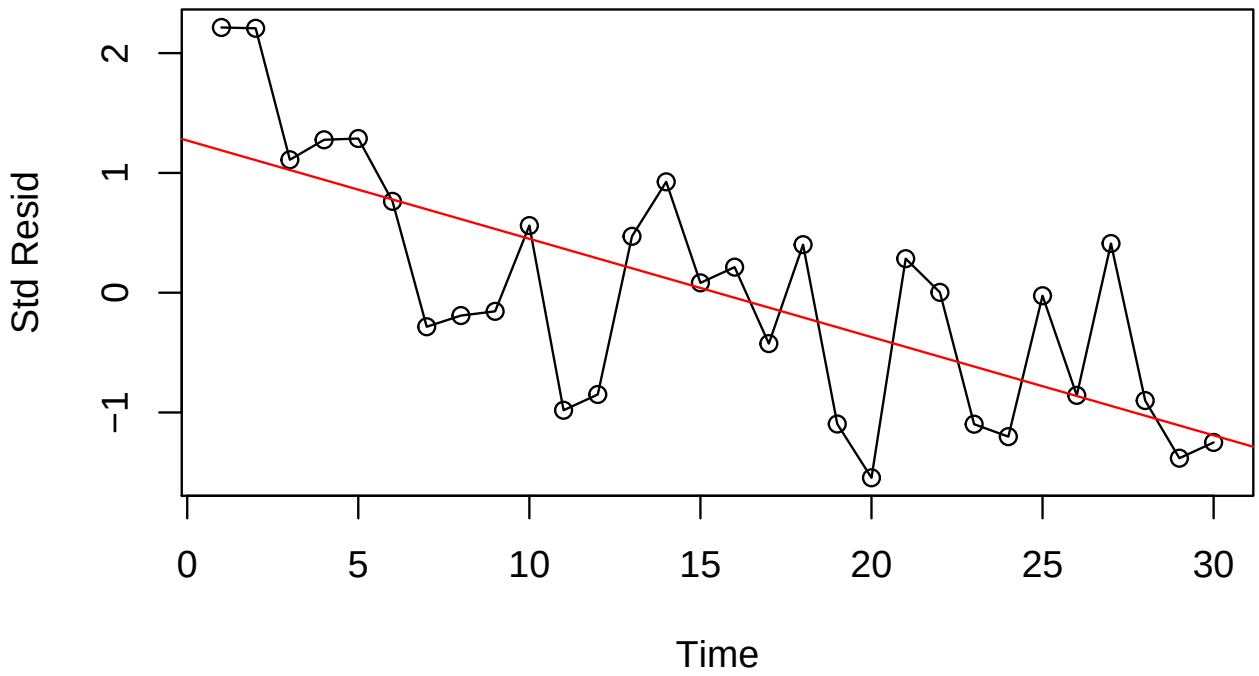
Example: Hotel Occupancy Case

... this is a *time series regression*, as we are regressing one time series on another.

In this case, we should also check whether or not the residuals show some temporal pattern.

If our model is correct the residuals should look iid normal over time.

Example: Hotel Occupancy Case



Does this look like independent normal noise to you? Can you guess what the red line represents?

Example: Hotel Occupancy Case

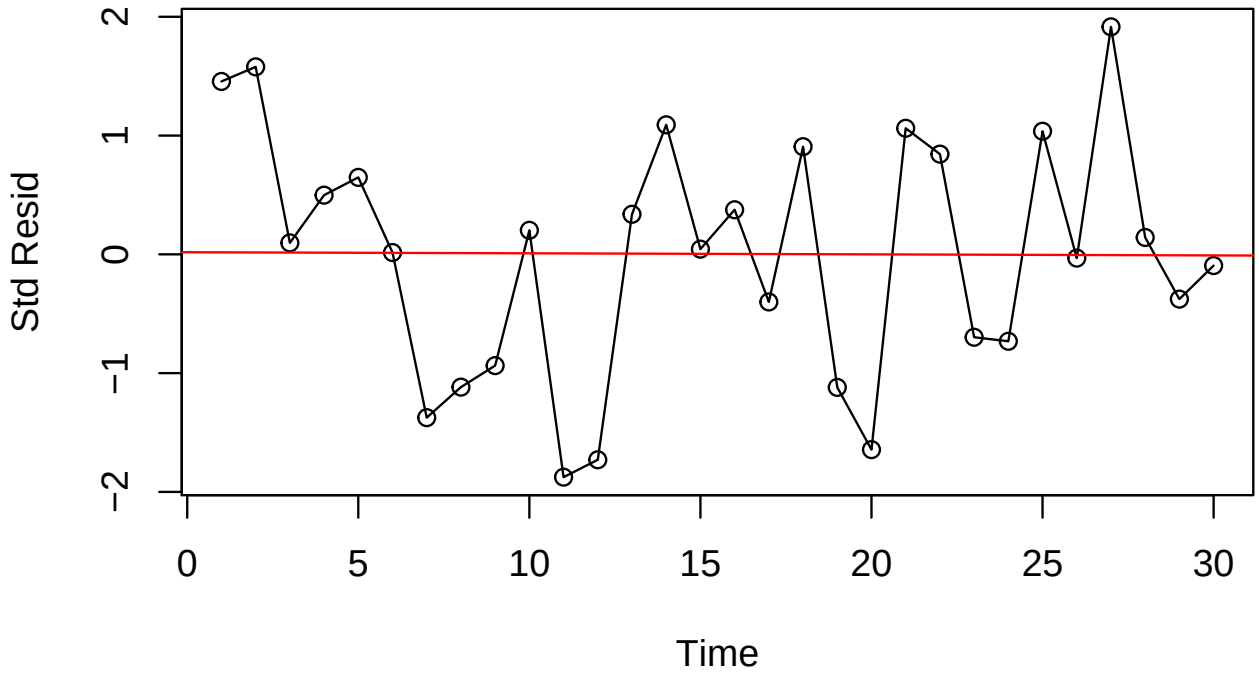
It looks like part of hotel occupancy (y) not explained by the Chicago downtown occupancy (x) – i.e., the SLR residuals – is moving down over time. We can try to control for that by adding a trend to our model...

$$Hotel_t = \beta_0 + \beta_1 Chicago + \beta_2 t + \epsilon_t$$

```
hotel_ts = ts(hotel)
hotel_fit_2 = tslm(Hotel~Chicago + trend, data=hotel_ts)
coef(hotel_fit_2)

## (Intercept)     Chicago       trend
##  26.6939111   0.6952379  -0.5964767
```

Example: Hotel Occupancy Case



Much better!! What is the slope of the red line?

Example: Hotel Occupancy Case

Okay, what happened?!

Well, once we account for the downward trend in the occupancy of the plaintiff, the prediction for the occupancy rate is

$$26 + 0.69 * 66 - 0.59 * 31 = 53.25\%$$

What do we conclude?

Example: Hotel Occupancy Case

Take away lessons...

- When regressing a time series on another, always check the residuals as a time series
- What does that mean... plot the residuals over time. If all is well, you should see no patterns, i.e., they should behave like iid normal samples.

Example: Hotel Occupancy Case

Question

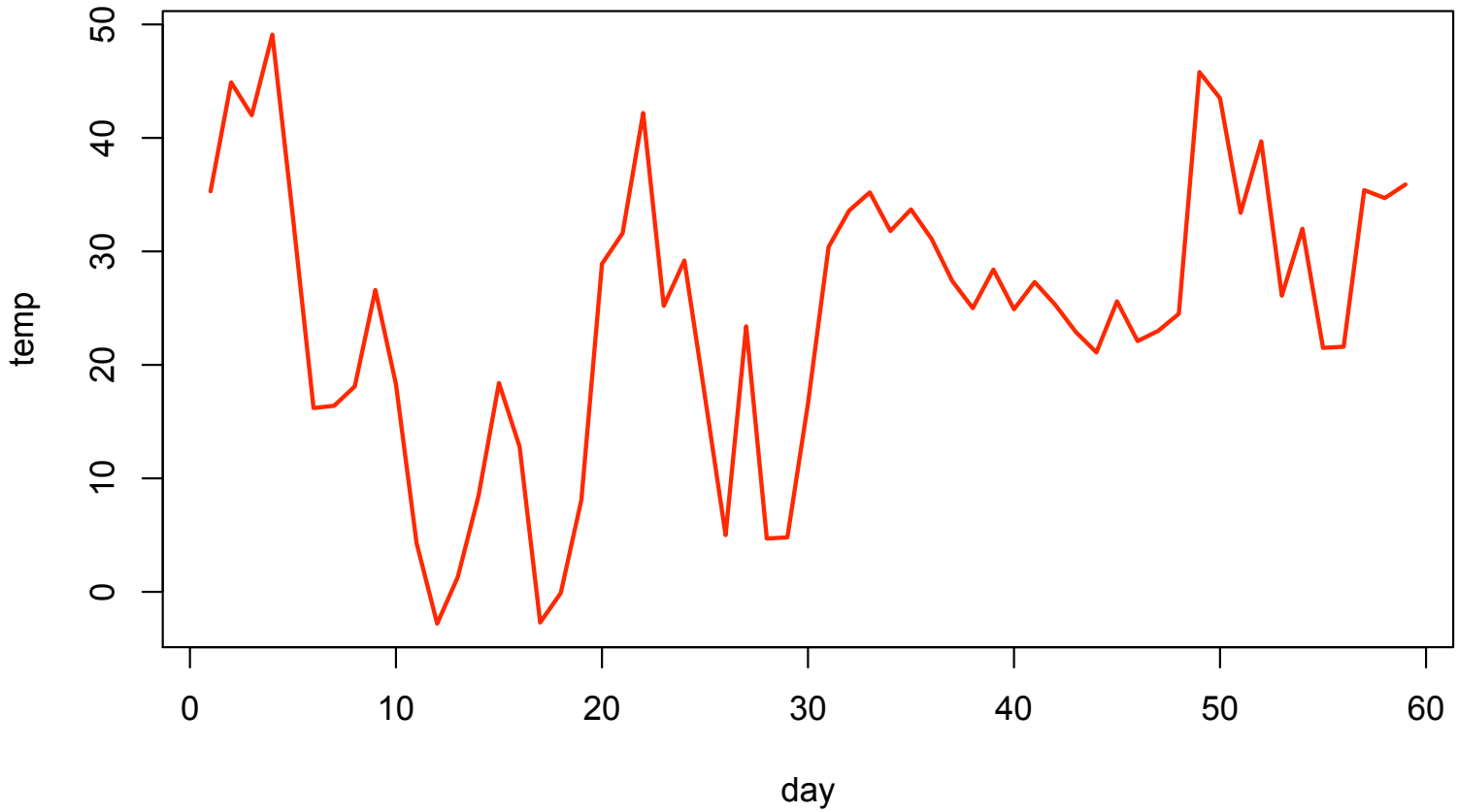
- What if we were interested in predicting the hotel occupancy ten years from now? We would compute

$$26 + 0.69 * 66 - 0.59 * 150 = -16.96\%$$

- Would you trust this prediction? Could you defend it in court?
- Remember: always be careful with extrapolating relationships!

Examples: Temperatures

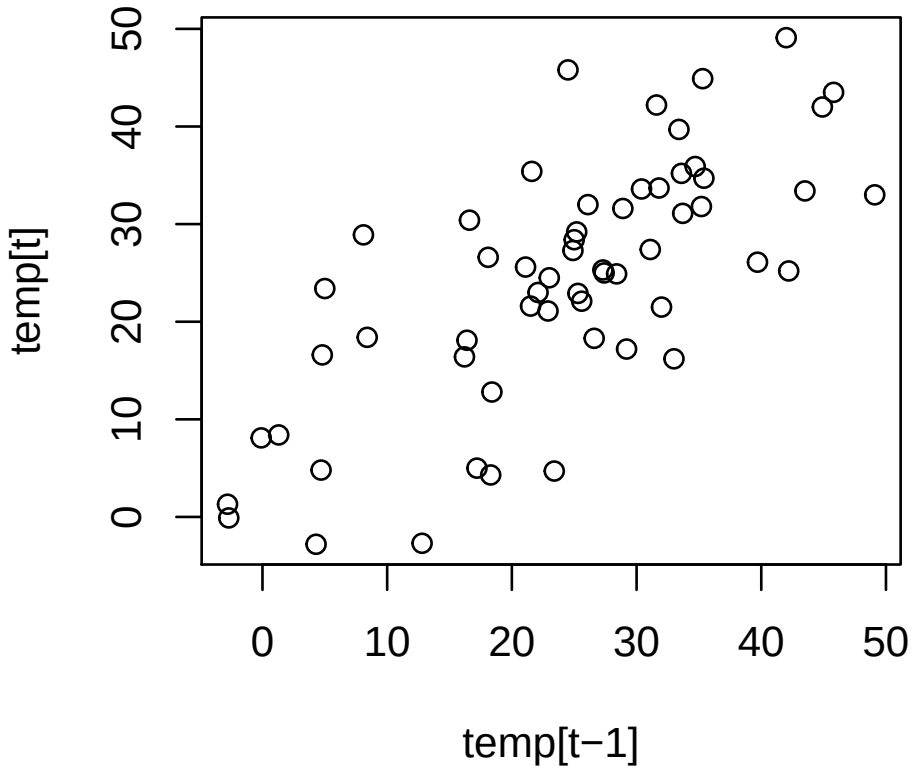
Now you need to predict tomorrow's temperature at O'Hare from (Jan-Feb).



Does this look iid? If it is iid, tomorrow's temperatures should not depend on today's... does that make sense?

Checking for Dependence

To see if Y_{t-1} would be useful for predicting Y_t, we can plot them together and see if there is a relationship.



Here $Cor(Y_t, Y_{t-1}) = 0.72$. Correlation between Y_t and Y_{t-1} is called autocorrelation.

Checking for Dependence

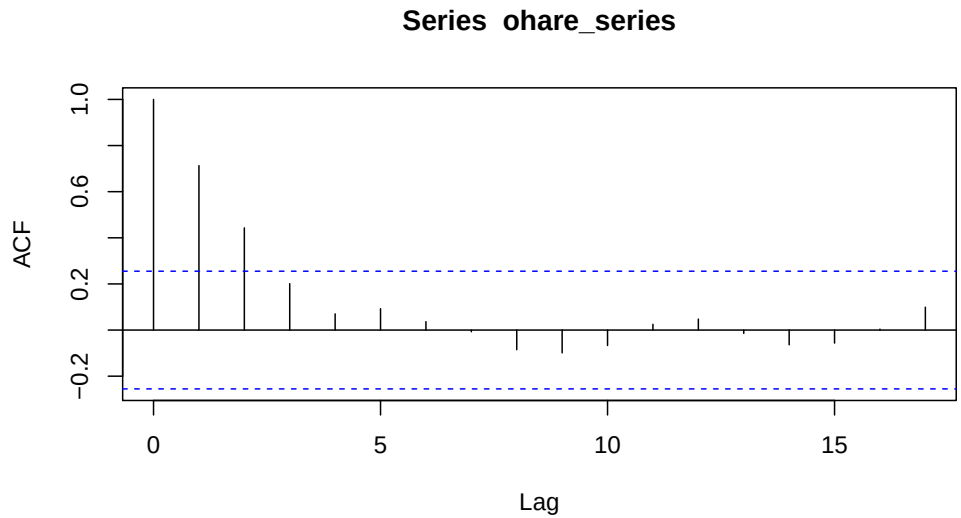
We created a "lagged" variable $temp_{t-1}$... the data looks like this:

t	temp(t)	temp(t-1)
1	42	35
2	41	42
3	50	41
4	19	50
5	19	19
6	20	19
...	...	

Checking for Dependence

We could plot Y_t against Y_{t-h} to see h-period lagged relationships. As a shortcut we could make a plot of $Cor(y_t, y_{t-h})$ as a funciton of the lag h. This is the **autocorrelation function**:

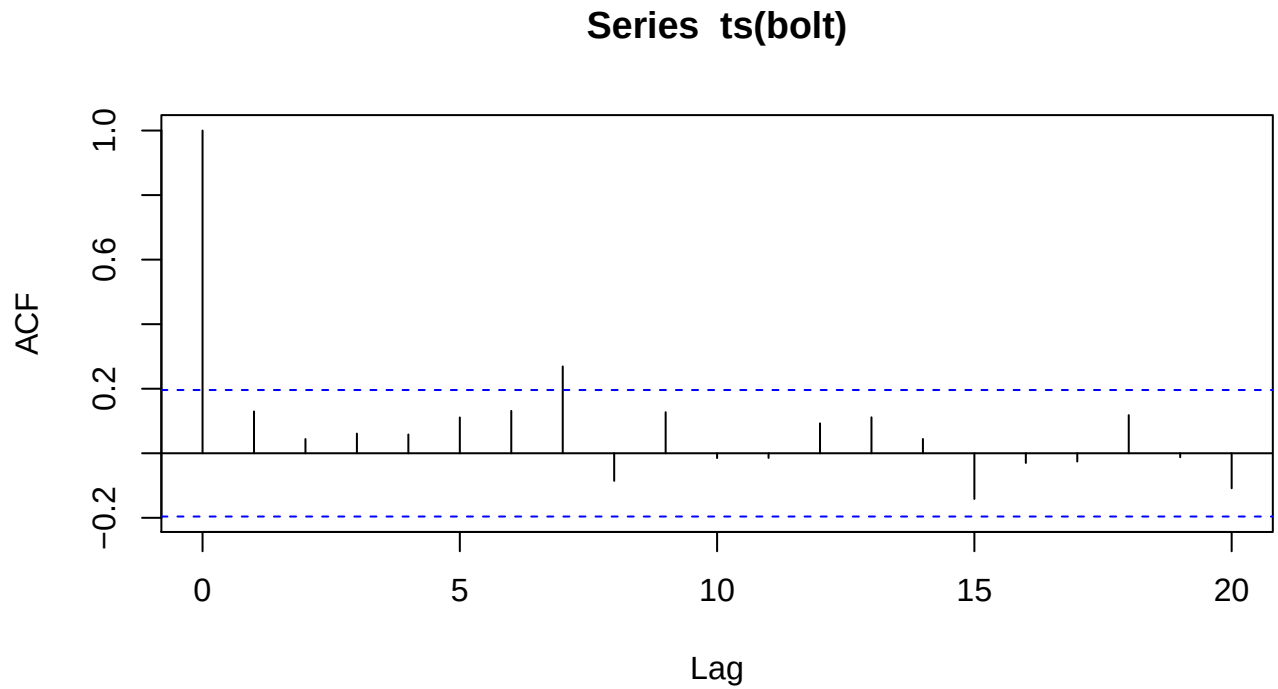
```
acf(ohare_series)
```



- It appears that the correlation is getting weaker with increasing L.
- How could we test for this dependence?

Checking for Dependence

Back to the "length of a bolt" example. When things are not related in time we should see...



The AR(1) Model

A simple way to model dependence over time in with the autoregressive model of order 1...

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$$

- What is the mean of Y_t for a given value of Y_{t-1}?
- If the model successfully captures the dependence structure in the data then the residuals should look iid.
- Remember: if our data is collected in time, we should always check for dependence in the residuals...

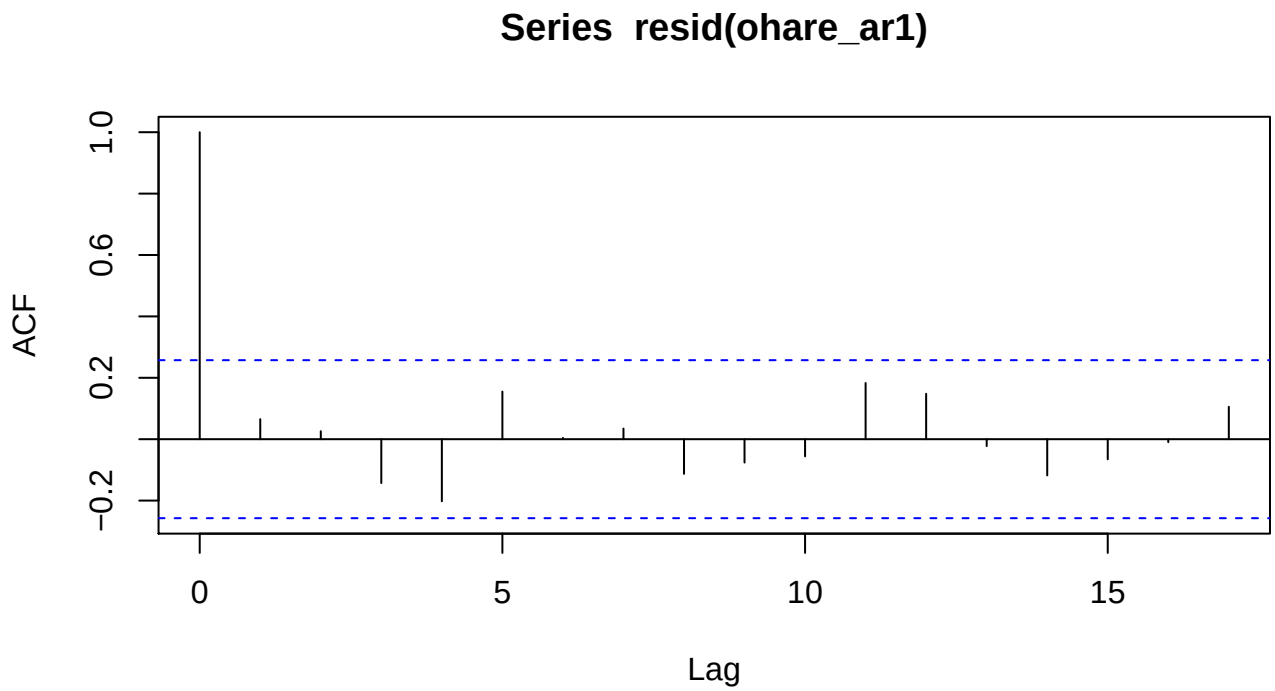
The AR(1) Model

Again, regression is our friend here...

```
##
## Call:
## tslm(formula = y ~ lag1, data = ohare_comb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9308  -4.8319   0.1644   4.2484  21.3736
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.70580    2.51661   2.665   0.0101 *
## lag1         0.72329    0.09242   7.826  1.5e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.79 on 56 degrees of freedom
## Multiple R-squared:  0.5224,Adjusted R-squared:  0.5138
## F-statistic: 61.24 on 1 and 56 DF,  p-value: 1.497e-10
```

The AR(1) Model

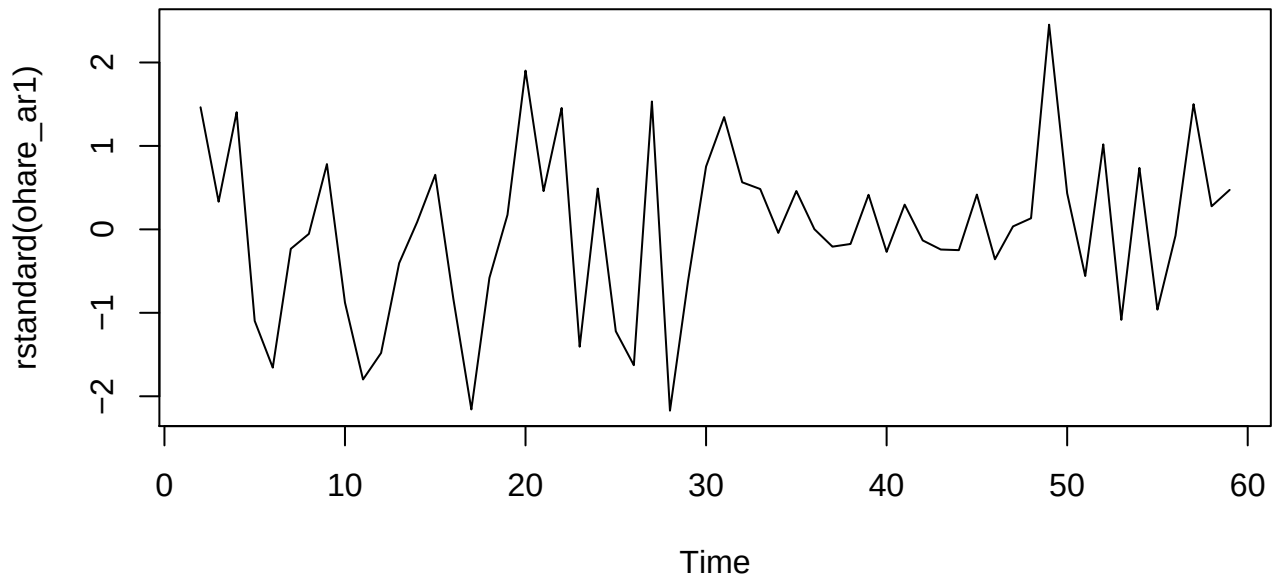
If the AR(1) model fits well, there should be no more time dependence in the residuals...



Good!

The AR(1) Model

We can also check residuals vs. time...



Again, looks pretty good...

Forecasting with the AR(1) Model

Forecasting the next observation Y_{T+1} from observations $Y_1, Y_2, \ldots Y_T$ is straightforward:

$$E(Y_{T+1} \mid Y_1, Y_2, \ldots Y_T) = \beta_0 + \beta_1 Y_T$$

For prediction intervals we know that

$$Y_{T+1} \sim N(\beta_0 + \beta_1 Y_T, \sigma^2)$$

Just like SLR (this is SLR!), we can use plug-ins b_0, b_1, and s for β_0, β_1 and σ.

What about 2 steps ahead?

Forecasting with the AR(1) Model

We can write:

$$\begin{aligned} Y_{T+2} &= \beta_0 + \beta_1 Y_{T+1} + \epsilon_{T+2} \\ &= \beta_0 + \beta_1(\beta_0 + \beta_1 * Y_T + \epsilon_{T+1}) + \epsilon_{T+2} \\ &= (1+\beta_1)\beta_0 + \beta_1^2 * Y_T + \beta_1\epsilon_{T+1} + \epsilon_{T+2} \end{aligned}$$

Remember, all the ϵ's are independent normally distributed variables with variance σ^2 So:

$$E(Y_{T+2} \mid Y_1, Y_2, \ldots Y_T) = (1+\beta_1)\beta_0 + \beta_1^2 Y_T$$

$$Var(Y_{T+2} \mid Y_1, Y_2, \ldots Y_T) = (1+\beta_1^2)\sigma^2$$

$$(Y_{T+2} \mid Y_1, Y_2, \ldots Y_T) \sim N([1+\beta_1]\beta_0 + \beta_1^2 Y_T, [1+\beta_1^2]\sigma^2)$$

We can still use plug-ins b_0, b_1, and s for β_0, β_1 and σ.

Forecasting with the AR(1) Model

For forecasting h steps ahead,

$$E(Y_{T+h} \mid Y_1, Y_2, \ldots Y_T) = \left(1 + \sum_{\ell=1}^{h-1} \beta_1^{\ell}\right) \beta_0 + \beta_1^h Y_T$$

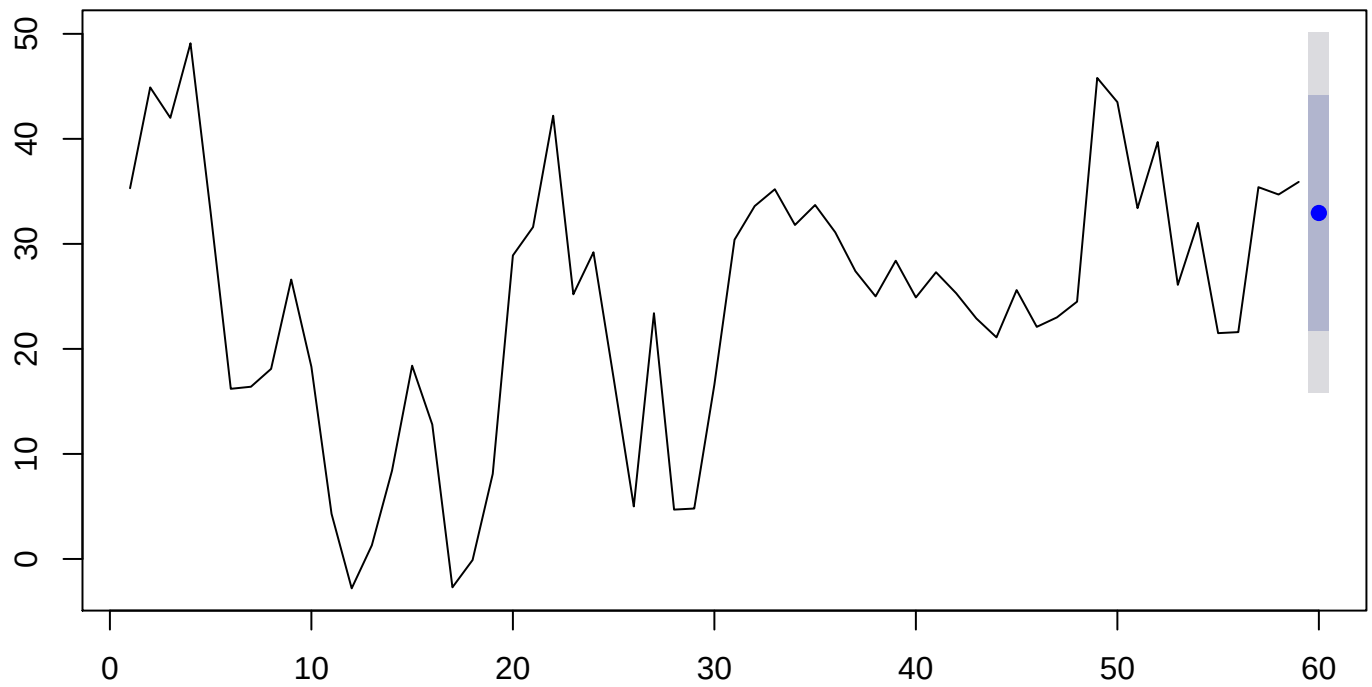
$$Var(Y_{T+h} \mid Y_1, Y_2, \ldots Y_T) = \left(1 + \sum_{\ell=1}^{h-1} \beta_1^{2\ell}\right) \sigma^2$$

and the conditional distribution of Y_{T+h} is normal.

Usually, $|\beta_1| < 1$. What happens to forecasts when h is large?

Forecasting with the AR(1) Model

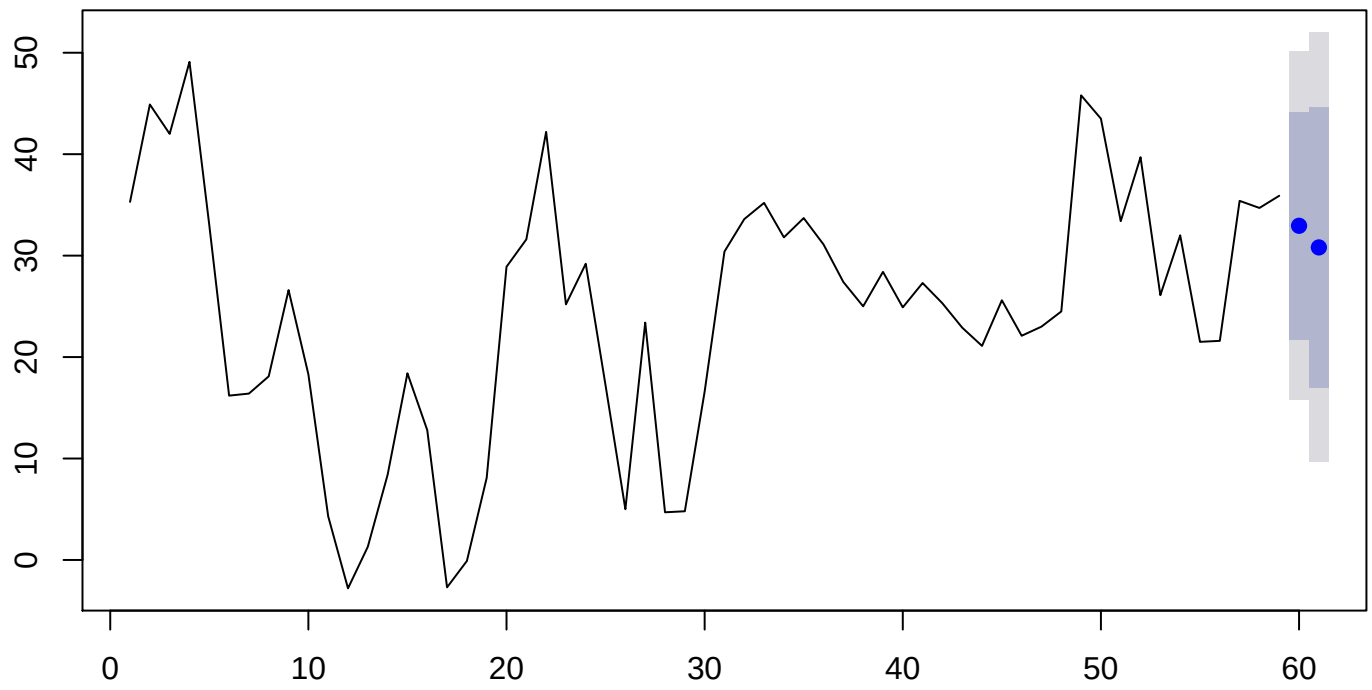
Let's look at the O'Hare data. Forecasting 1 day ahead:



(The gray bars are 80 and 95% prediction intervals)

Forecasting with the AR(1) Model

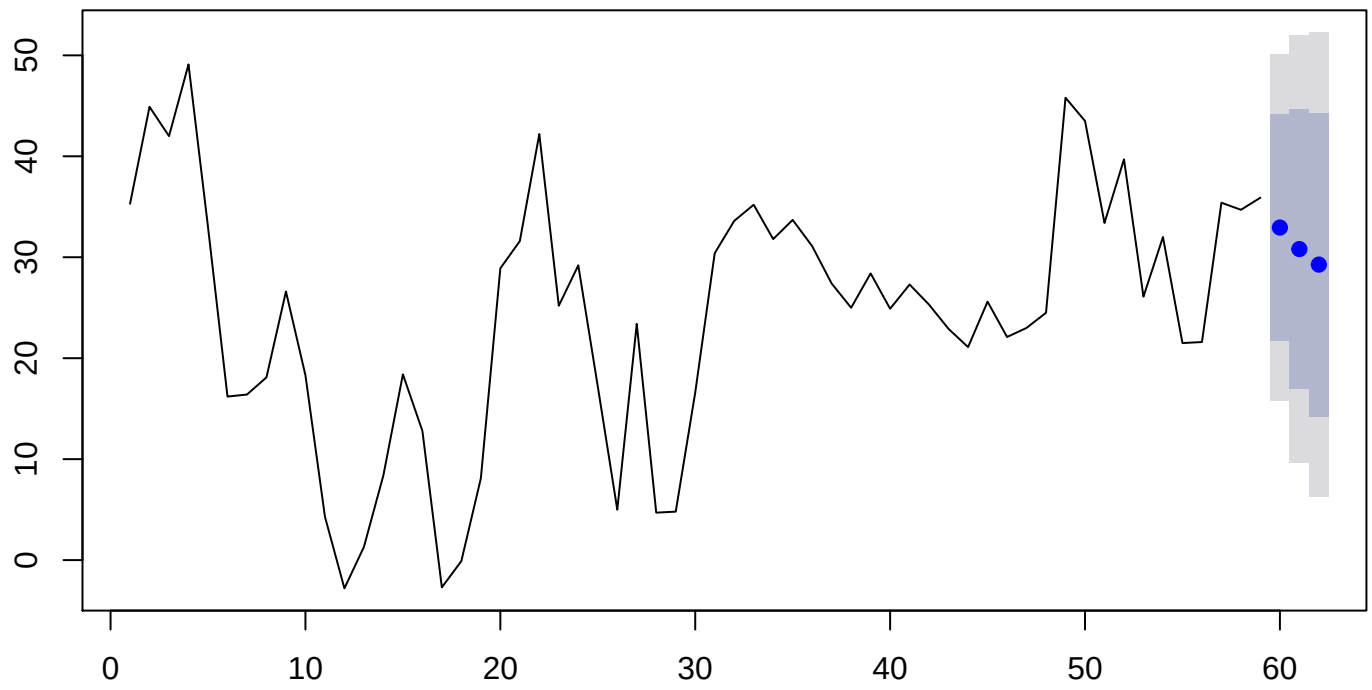
Forecasting 2 days ahead:



(The gray bars are 80 and 95% prediction intervals)

Forecasting with the AR(1) Model

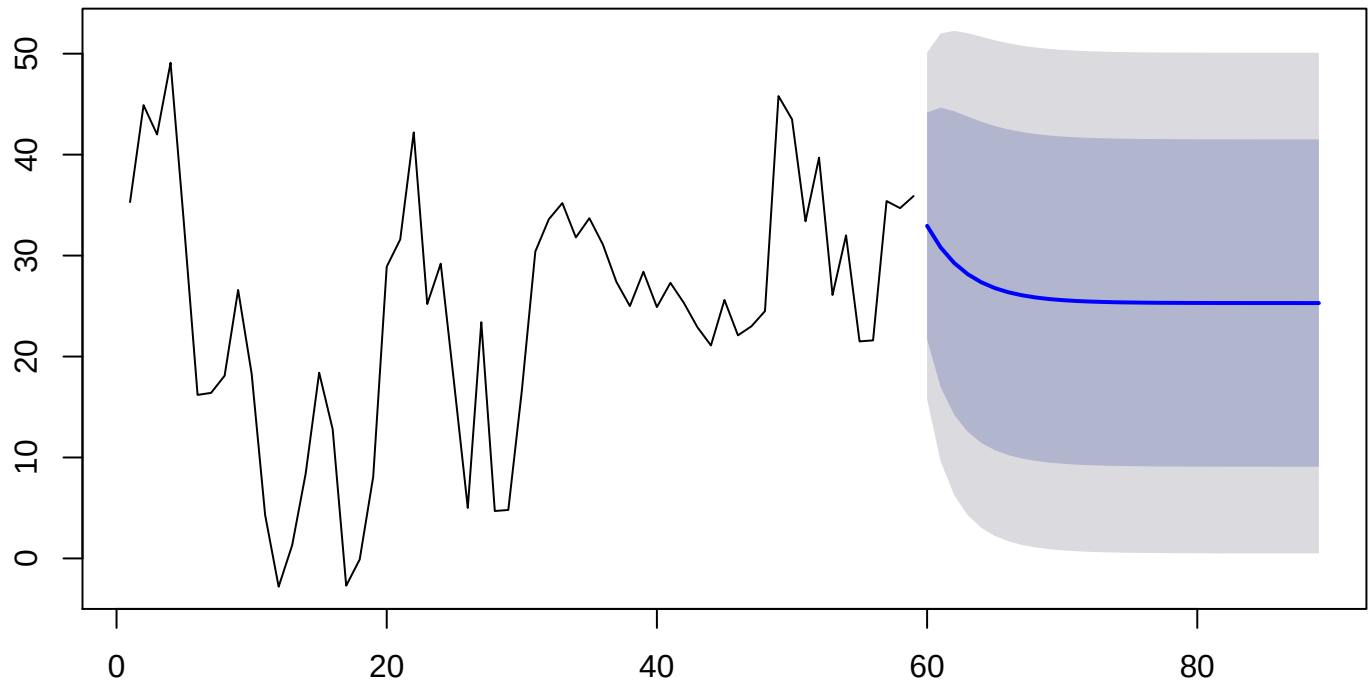
Forecasting 3 days ahead:



(The gray bars are 80 and 95% prediction intervals)

Forecasting with the AR(1) Model

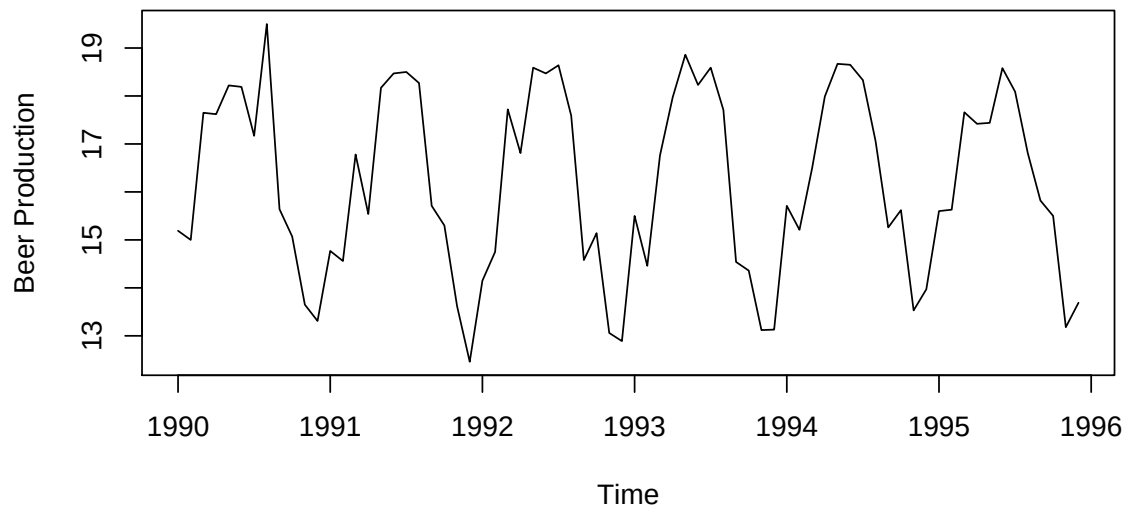
Forecasting 30 days ahead:



Do you trust this model to make long-term forecasts?

The Seasonal Model

- Many time-series data exhibit some sort of seasonality
- The simplest solution is to add a set of dummy variables to deal with the "seasonal effects"



Y_t = monthly U.S. beer production (in millions of barrels).

The Seasonal Model

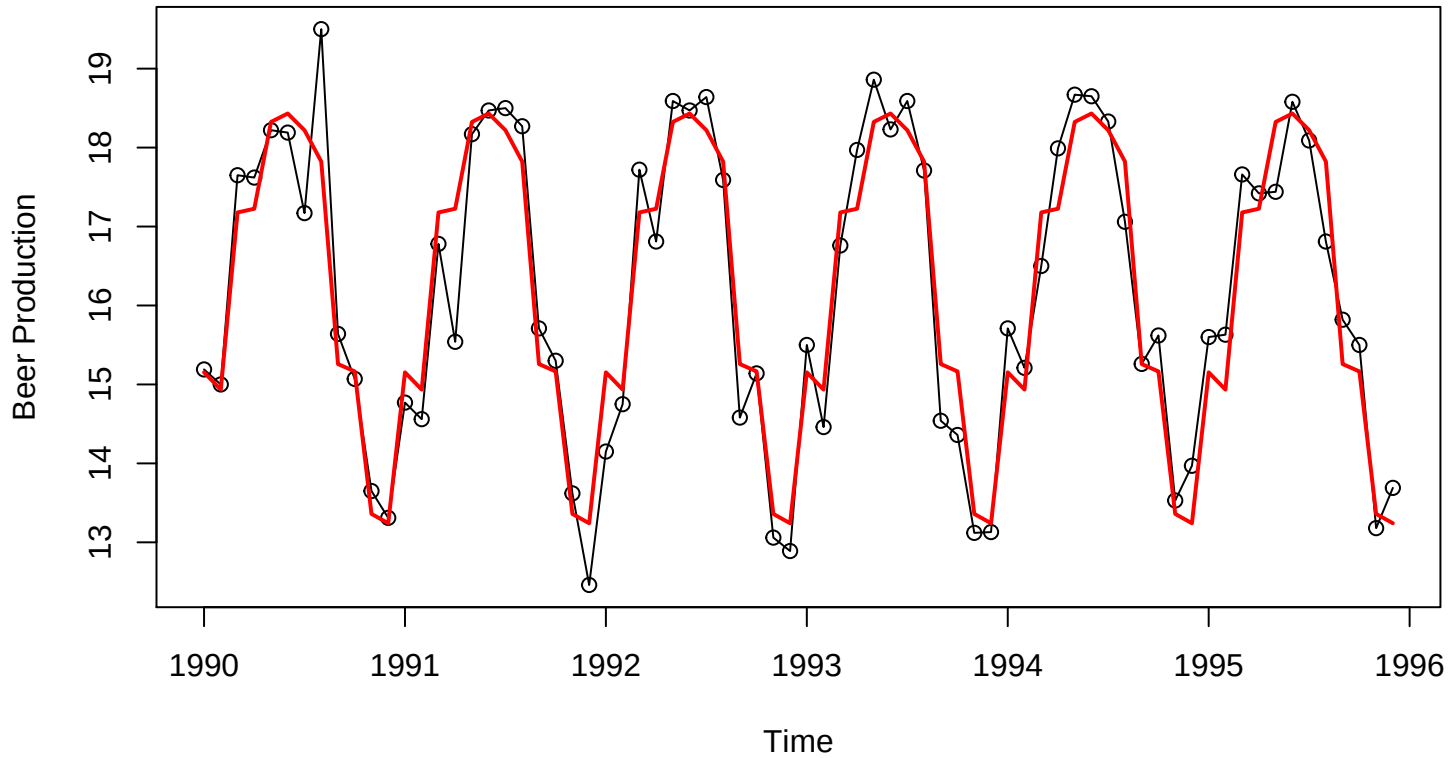
```
beer_series = ts(beer$X1, start=c(1990, 1), frequency=12)
```

```
beer_fit = tslm(beer_series~season)
print(beer_fit)
```

```
##
## Call:
## tslm(formula = beer_series ~ season)
##
## Coefficients:
## (Intercept)      season2      season3      season4      season5
##    15.15333     -0.21833      2.02500      2.07167      3.17167
##     season6      season7      season8      season9     season10
##     3.27833      3.06667      2.67000      0.10500      0.01167
##    season11     season12
##    -1.79333     -1.91167
```

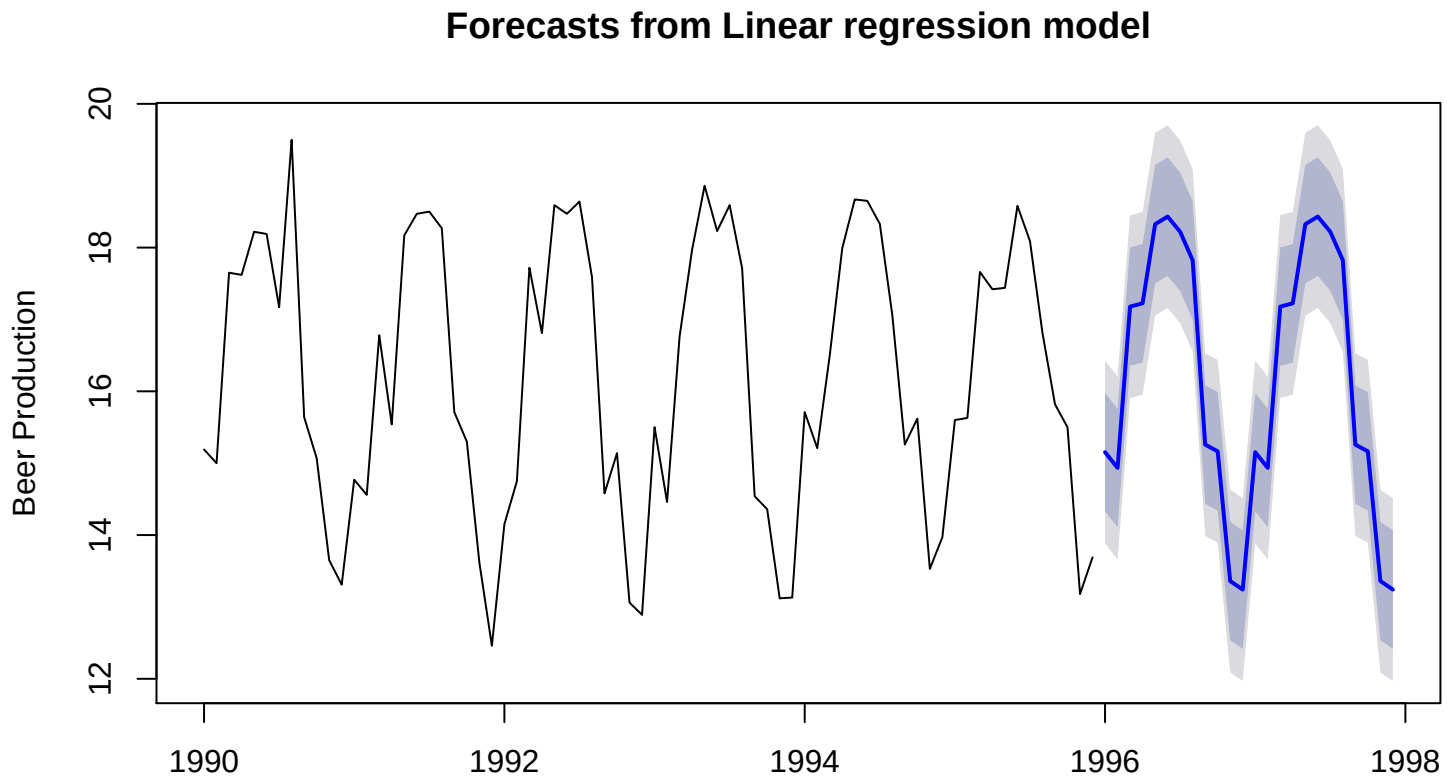
The Seasonal Model

The fitted model is in red:



What would our future predictions look like?

The Seasonal Model



Summary

We've looked at modeling and forecasting time series with

- Trends
- Seasonality
- General serial dependence (using lags)

Fundamentally, these are just multiple regression models with special covariates!

Often a proper time series analysis will involve all these pieces...