Hotels Case

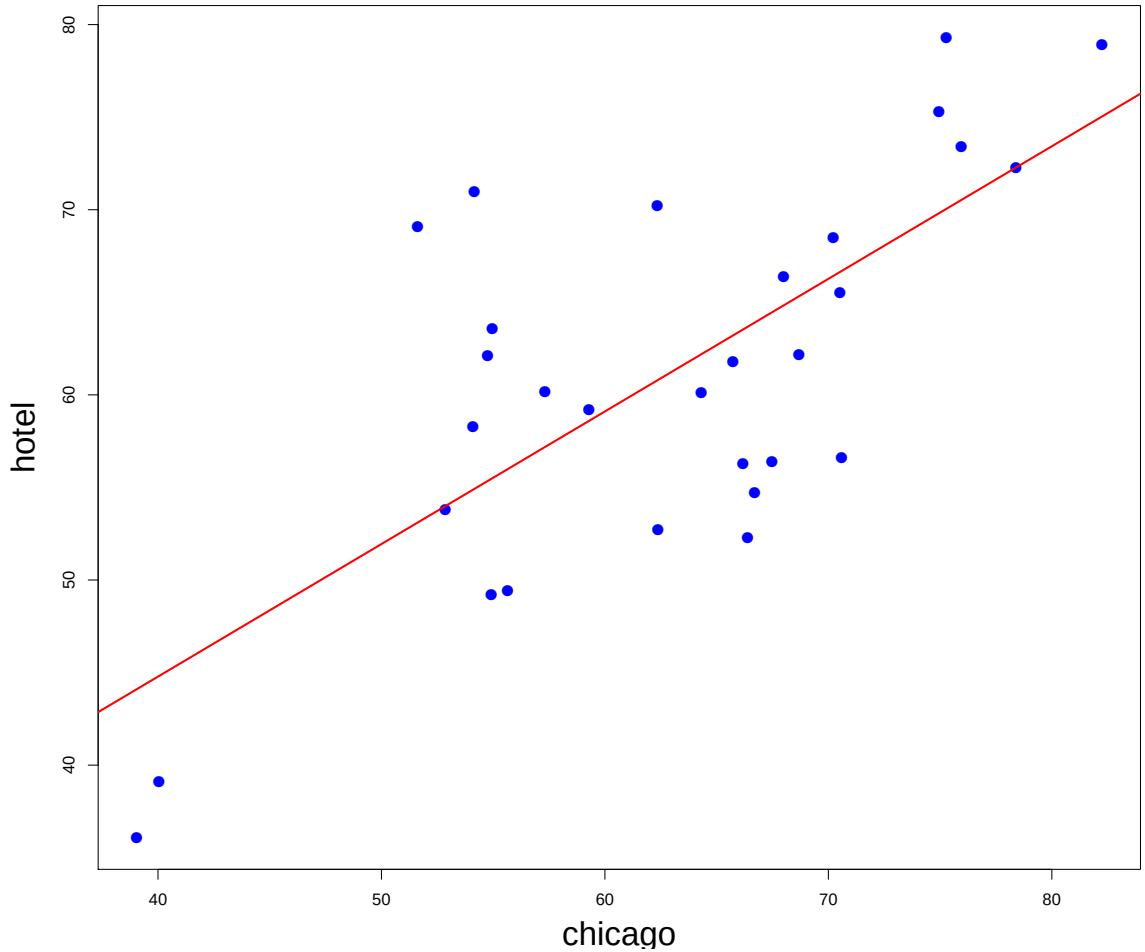
The Hotel Lawsuit

A downtown Chicago hotel paid to be included in a directory. It was then not included. The hotel sued the publisher of the directory to recover the lost income due to lost business.

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

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

Here is the data with the regression line drawn.



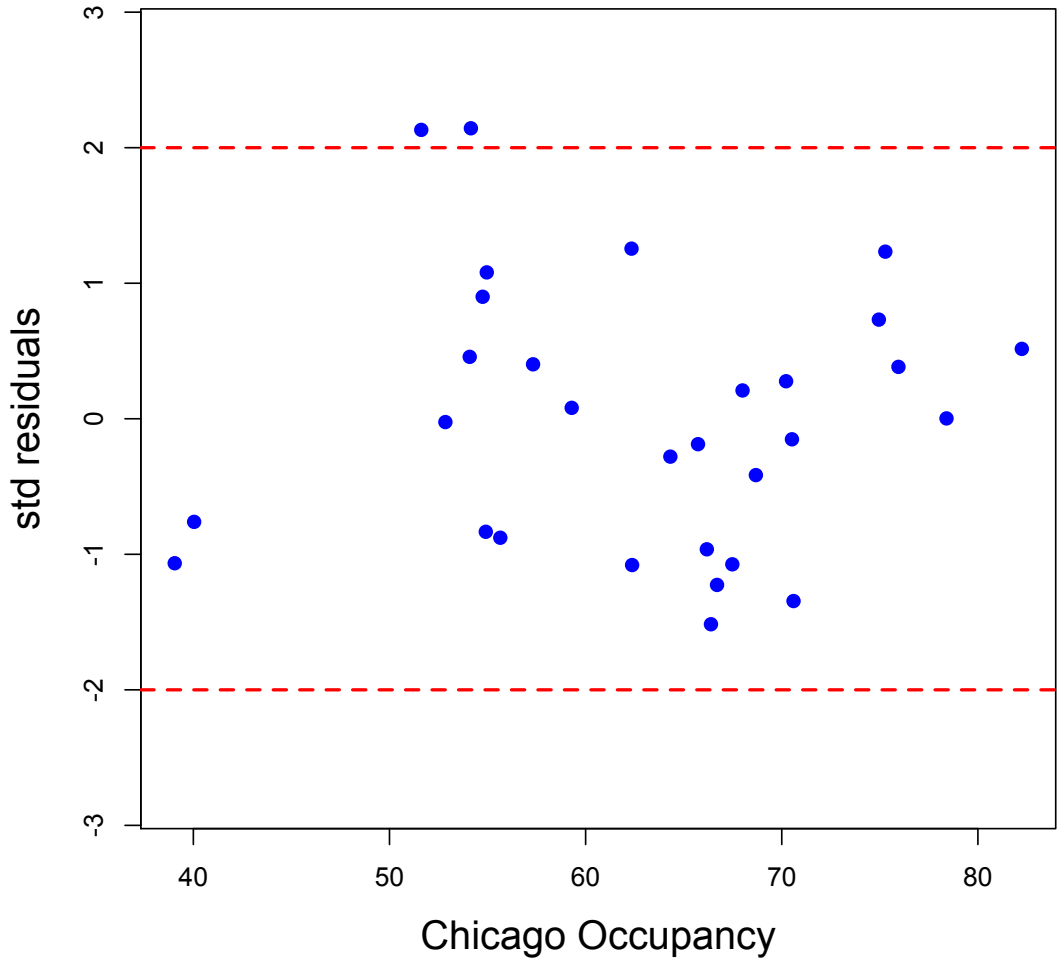
Does this look reasonable ??

$$Hotel_t = \alpha + \beta\, Chicago_t + \epsilon_t$$

```
Call:
lm(formula = hotel ~ ., data = hd)

Residuals:
     Min       1Q   Median       3Q      Max
-11.3797  -6.5059  -0.0818   3.7594  16.0858

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.1357     8.5189   1.894   0.0686 .
chicago       0.7161     0.1338   5.352 1.06e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.506 on 28 degrees of freedom
Multiple R-squared: 0.5057,Adjusted R-squared: 0.488
F-statistic: 28.64 on 1 and 28 DF,  p-value: 1.061e-05
```

- 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!!

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

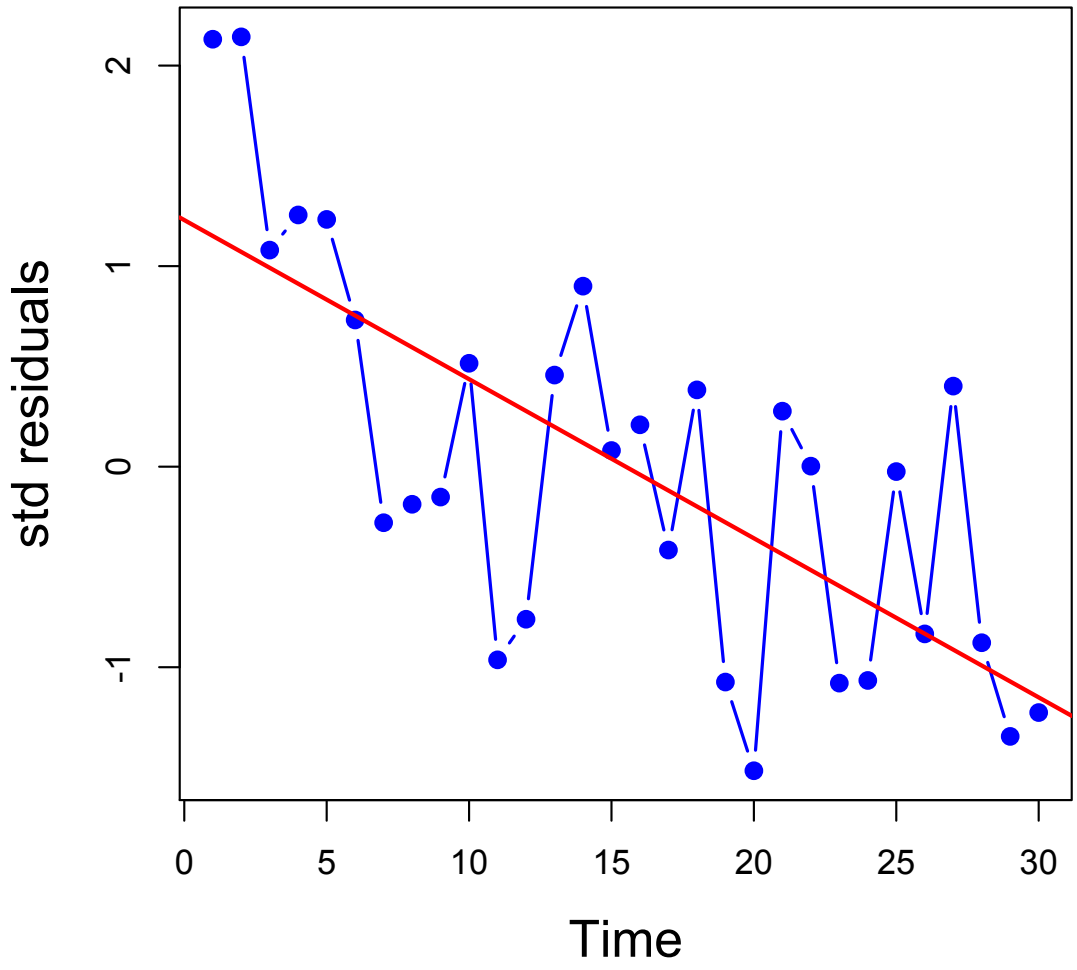


Looks fine. However...

... 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*.

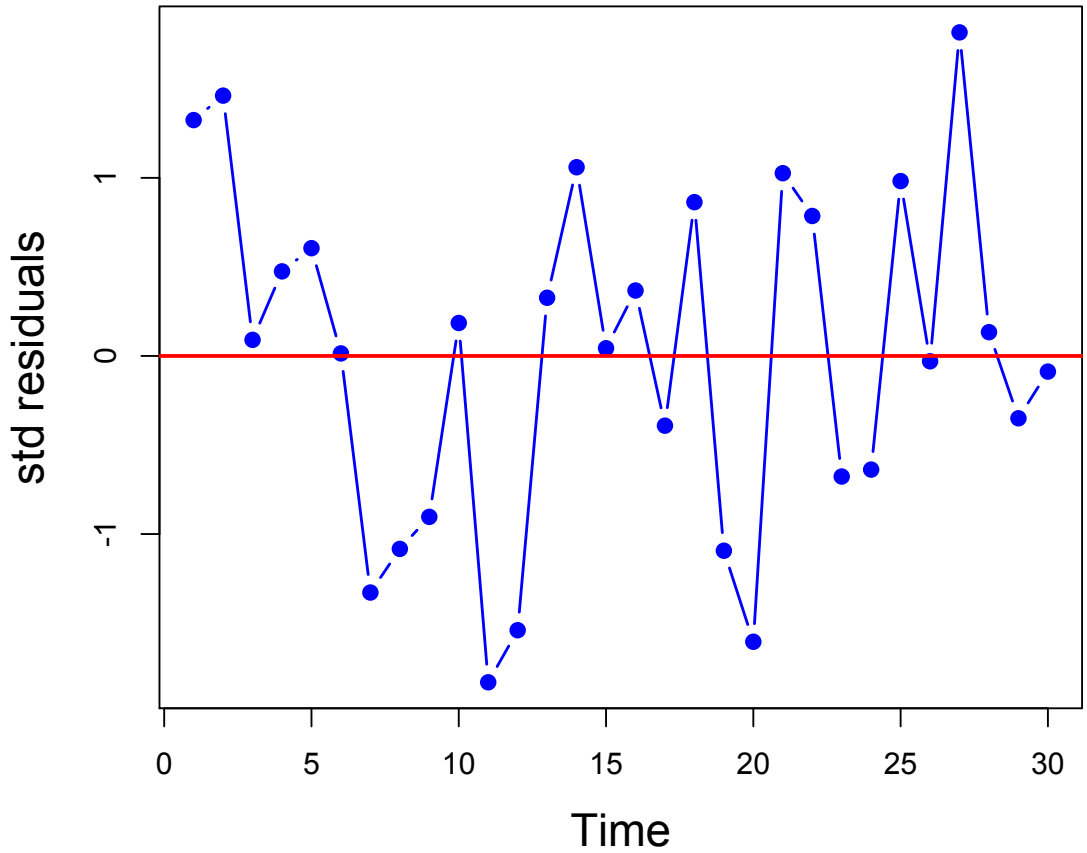


Does this look iid to you?

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

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

```
Call:
lm(formula = hotel ~ ., data = hd)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8435 -3.5846  0.3589  3.9792  9.7599

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.69391    6.41884   4.159  0.00029 ***
chicago      0.69524    0.09585   7.253 8.41e-08 ***
t           -0.59648    0.11340  -5.260 1.52e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.372 on 27 degrees of freedom
Multiple R-squared: 0.7558,Adjusted R-squared: 0.7378
F-statistic: 41.79 on 2 and 27 DF,  p-value: 5.415e-09
```



Much better!!

Okay, what happened?!

The actual occupancy rate for the hotel in the month after the omission was 55.

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?

Question

- What if we were interested in predicting the hotel occupancy ten years from now?? Suppose the occupancy rate for downtown Chicago hotels is 65, we put in $t = 30 + 10(12) = 150$:
  We should compute

$$26 + 0.69 * 66 - 0.59 * 150 = -17.65.$$

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

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 draws.