Forecasting and Time Series Analysis

San Cannon

Rockhurst University

Week 3

Regression time

Remember linear regression?

Say you wanted to understand what affects the mileage of automobiles. There are possibly lots of things but we suspect that the power in the engine might be related.

Start with a hypothesized linear model:

$y = \beta_0 + \beta_1 * x + \epsilon$

And fit a linear regression line:

$\hat{y} = b_0 + b_1 * x$

Where

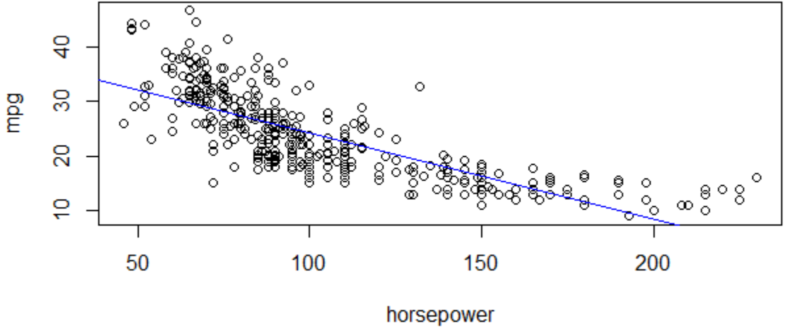
- Target variable y: Miles per gallon
- Single independent variable x: horsepower
- Fit the model using lm() function in R

What do you think the relationship looks like?

Regression results

```
Call:
lm(formula = mpg ~ horsepower, data = autos)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

Obligatory picture



Blue line is **mpg = 39.93 - 0.158 * horsepower**

What do these numbers tell us?

Interpreting the numbers:

- The intercept estimate b_0 is the expected value of y when x is zero.
- The slope estimate b_1 is the expected change in y for a one unit change in x
- R^2 is the amount of variation in y captured by the variation in x

Hypothesis testing: what is the null hypothesis of interest for linear regression?

$$
\begin{aligned}
H_0 : &\quad \beta_1 = 0 \\
H_a : &\quad \beta_1 \neq 0
\end{aligned}
$$

Interpretation here?

Forecasting using linear model: without time

The regression line is used for forecasting. For each value of x, we can forecast a corresponding value of y using $\hat{y} = b_0 + b_1 * x$

Note - there are no time subscripts here. We are trying to predict what value for y we would see for any given value of x

For our mpg problem, we would create our train/test split and then project based on the coefficients estimated for the training data set.

Then we would typically investigate the prediction accuracy.

Fitted values and residuals

The forecast values of y obtained from the observed x values are called fitted values. We write these as $\hat{y}_i = b_0 + b_1 * x_i$ for $i = 1, \ldots, N$.

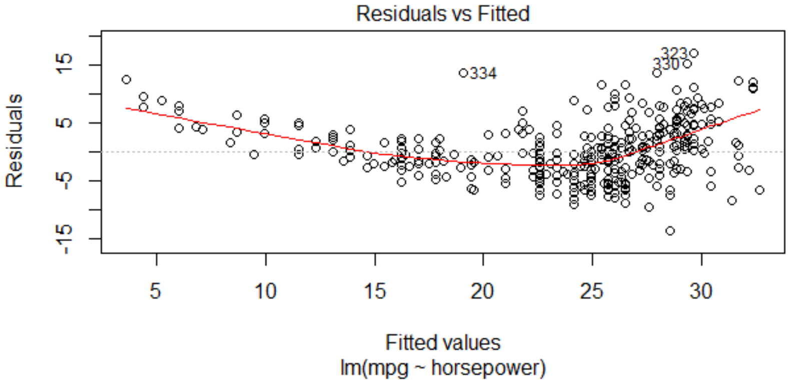
Each $\hat{y}_i$ is the point on the regression line corresponding to observation x_i

The difference between the observed y values and the corresponding fitted values are the residuals: $e_i = y_i \hat{y}_i = y_i b_0 - b_1 * X_i$

The residuals have some useful properties including the following two: $\sum_{i=1}^{N} e_i = 0$ and $\sum_{i=1}^{N} x_i * e_i = 0$

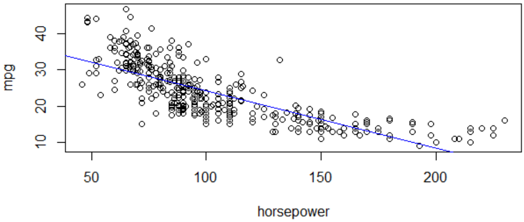
As a result of these properties, it is clear that the average of the residuals is zero, and that the correlation between the residuals and the observations for predictor variable is also zero.

What do the residuals look like?



What about non-linear regression?

We are assuming a linear relationship between variables. What if that isn't true? Do we think this relationship is linear?



So what should we do?
We need to figure out a way to express either the data or the relationship in a way that is linear.

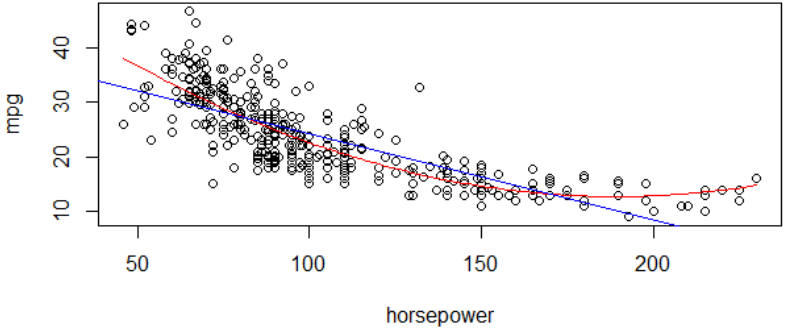
Alternative specifications

For our mpg example, we can add a quadratic term to try to match the curvature.[1]

```
Call:
lm(formula = mpg ~ horsepower + I(horsepower^2), data = autos)

Residuals:
     Min       1Q   Median       3Q      Max
-14.7135  -2.5943  -0.0859   2.2868  15.8961

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     56.9000997  1.8004268   31.60   <2e-16 ***
horsepower      -0.4661896  0.0311246  -14.98   <2e-16 ***
I(horsepower^2)  0.0012305  0.0001221   10.08   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.374 on 389 degrees of freedom
Multiple R-squared:  0.6876,Adjusted R-squared:  0.686
F-statistic:   428 on 2 and 389 DF,  p-value: < 2.2e-16
```

[1] Simple explanation for including polynomials:
http://www.theanalysisfactor.com/regression-modelshow-do-you-know-you-need-a-polynomial/

Picture please



Other options

Sometimes we need to change the data to get things to be linear. Remember what we said about multiplicative decomposition?

Take logs to make the relationship linear. (Taking the log of one variable helps with skew as well.)

How do we interpret the coefficients for various types of equations?

Model	Dependent Variable	Independent Variable	Interpretation of the Coefficient
Level - Level	Y	X	$\Delta y = \Delta x * b$
Level – Log	Y	ln(x)	$\Delta y = \% \Delta x * b$
Log - Level	ln(y)	x	$\% \Delta y = \Delta x * b$ (...ish)
Log - log	ln(y)	ln(x)	$\% \Delta y = \% \Delta x * b$

Remember the assumptions of the linear model

- No autocorrelation: Error values ϵ are statistically independent
- Normality of error distribution: Error values are normally distributed for any given value of x
- Homoskedasticity: The probability distribution of the errors has constant variance
- Linearity and additive: The underlying relationship between the x variable and the y variable is linear

So that's why we waded through all the stuff about autocorrelation - because it violates our assumptions.

Understanding the properties of a time series is important for understanding how to use them with basic statistics.

Modeling data using linear regression

Sometimes we want to understand a particular data series better so we use linear regression techniques to do that.

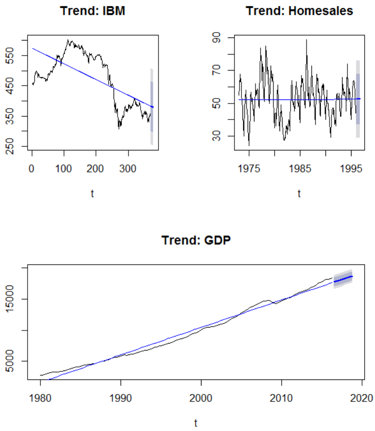
We've taken a quick look at seasonality (and will come back to it shortly), now let's do the same for trend.

It's a common feature of time series. We can add a new model to our simple forecast arsenal by strictly estimating the linear trend for a series:

$$\hat{y}_t = \beta_0 + \beta_1 t + \epsilon_t.$$

How does that work? Let's look at some of our recent examples.

Where do we find trend?



Random Walk

Now we'll take a closer look at some special cases - the random walk and random walk with drift (our drift model from before)

A random walk is defined as a process where the current value of a variable is composed of the past value plus an error term defined as a white noise (a normal variable with zero mean and variance one). Algebraically a random walk is represented as follows: $y_t = y_{t-1} + \epsilon_t$

Adding a drift component to the random walk model means that the drift acts like a trend, and the process has the following form:
$y_t = y_{t-1} + \alpha + \epsilon_t$

For $\alpha > 0$ the process will show an upward trend

A word about spurious regressions

A quick word about spurious regressions. Because there are so many time series have trend components, it is easy to confuse a common time trend with a meaningful relationship.

The most absurd examples can be found here: Spurious Correlations

We'll learn how to deal with such things next week.

Multiple regression

Everything we've done so far is based on one time series. We have more information we can use but let's remember some fundamentals about multiple regression.

Everything we said about linear regression before still holds but now we have multiple $x's$ on the right hand side. We not only have to think about the relationship between each x and y but also any relationship between the $x's$.

We model the relationship as: $y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots + \beta_k * x_k + \epsilon$
and we estimate: $\hat{y} = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_k * x_k$

where the coefficients $b_1, b_2, \ldots, b_k$ estimate the effect of each predictor after taking account of the effect of all other predictors in the model.

Let's start with an example

Adding on to our car example: what else might affect gas mileage?

```
Call:
lm(formula = mpg ~ horsepower + I(horsepower^2) + cylinders +
    weight + year, data = autos)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6151 -2.0313 -0.1192  1.8801 12.9871

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.654e+00  3.955e+00  -0.924    0.356
horsepower      -2.480e-01  2.791e-02  -8.886   <2e-16 ***
I(horsepower^2)  8.658e-04  9.373e-05   9.237   <2e-16 ***
cylinders        6.592e-02  2.211e-01   0.298    0.766
weight          -5.197e-03  4.901e-04 -10.605   <2e-16 ***
year             7.554e-01  4.742e-02  15.931   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.111 on 386 degrees of freedom
Multiple R-squared:  0.8431,Adjusted R-squared:  0.8411
F-statistic: 414.9 on 5 and 386 DF,  p-value: < 2.2e-16
```

How do we interpret these results?

How do we use this with time?

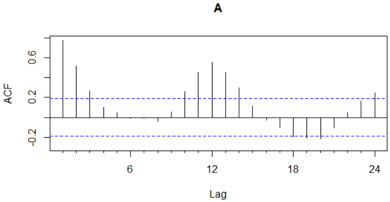
Dummy variables

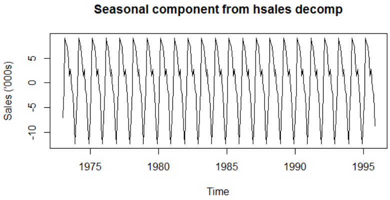
We can address issues such as seasonality by using dummy variables as part of a multiple regression model. How might we test for the seasonal pattern in our home sales data?

```
Call:
tslm(formula = hsales ~ season)

Residuals:
     Min       1Q   Median       3Q      Max
-28.2609  -6.0652   0.8696   6.6087  27.6087

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   45.348      2.108  21.517  < 2e-16 ***
season2        5.783      2.980   1.940 0.053430 .
season3       16.043      2.980   5.383 1.62e-07 ***
season4       14.913      2.980   5.004 1.03e-06 ***
season5       14.261      2.980   4.785 2.86e-06 ***
season6       11.652      2.980   3.909 0.000118 ***
season7        8.435      2.980   2.830 0.005014 **
season8        9.783      2.980   3.282 0.001169 **
season9        5.478      2.980   1.838 0.067184 .
season10       4.391      2.980   1.473 0.141852
season11      -2.217      2.980  -0.744 0.457559
season12      -5.802      3.014  -1.925 0.055304 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.11 on 263 degrees of freedom
Multiple R-squared:  0.312,Adjusted R-squared:  0.2833
F-statistic: 10.84 on 11 and 263 DF,  p-value: < 2.2e-16
```

Interpreting dummy variables

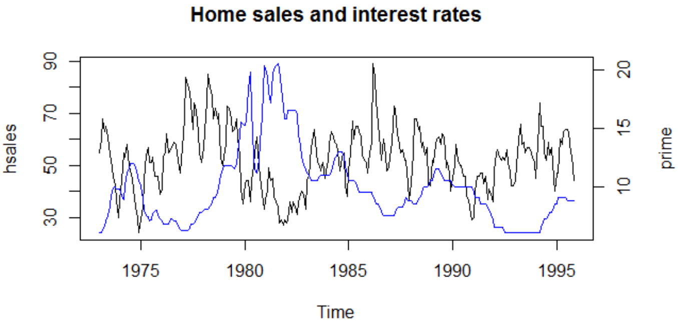
How do these coefficients relate to all the other info we've gathered on home sales data?





Putting it all together

We can model the trend and seasonal information in order to be able to find relationships with other variables that are not due to time. Let's look at the relationship between home sales and interest rates.



Simple model

Regression results:

```
Call:
tslm(formula = hdata[, 1] ~ hdata[, 2])

Residuals:
    Min      1Q  Median      3Q     Max 
-27.263  -7.292  -0.612   7.097  35.406 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   68.747      2.031  33.841  < 2e-16 ***
hdata[, 2]    -1.665      0.195  -8.538 9.65e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.63 on 273 degrees of freedom
Multiple R-squared:  0.2108,Adjusted R-squared:  0.2079 
F-statistic:  72.9 on 1 and 273 DF,  p-value: 9.651e-16
```


```
Call:
tslm(formula = hdata[, 1] ~ hdata[, 2])

Residuals:
    Min      1Q  Median      3Q     Max 
-27.263  -7.292  -0.612   7.097  35.406 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   68.747      2.031  33.841  < 2e-16 ***
hdata[, 2]    -1.665      0.195  -8.538 9.65e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.63 on 273 degrees of freedom
```

Simple model plus seasonals

Capturing seasonality:

```
Call:
tslm(formula = hsales ~ prime + season)

Residuals:
     Min       1Q   Median       3Q      Max
-20.3458  -6.2804  -0.0081   5.4044  26.5622

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.3375     2.3449  26.158  < 2e-16 ***
prime        -1.6307     0.1563 -10.434  < 2e-16 ***
season2       5.6770     2.5099   2.262 0.024529 *
season3      15.9400     2.5099   6.351 9.41e-10 ***
season4      15.0010     2.5099   5.977 7.41e-09 ***
season5      14.4856     2.5100   5.771 2.22e-08 ***
season6      11.7174     2.5099   4.668 4.85e-06 ***
season7       8.4929     2.5099   3.384 0.000824 ***
season8       9.8379     2.5099   3.920 0.000113 ***
season9       5.7151     2.5100   2.277 0.023598 *
season10      4.6813     2.5101   1.865 0.063298 .
season11     -1.9579     2.5100  -0.780 0.436079
season12     -5.2935     2.5387  -2.085 0.038030 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.512 on 262 degrees of freedom
Multiple R-squared:  0.514,Adjusted R-squared:  0.4917
F-statistic: 23.09 on 12 and 262 DF,  p-value: < 2.2e-16
```

Is this an improvment?

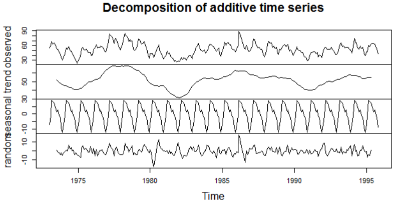
Other effects/dummy variables

- It is often necessary to model interventions that may have affected the variable to be forecast
- When the effect lasts only for one period, we use a spike variable. This is a dummy variable taking value one in the period of the intervention and zero elsewhere.
- Other interventions have an immediate and permanent effect. If an intervention causes a level shift (i.e., the value of the series changes suddenly and permanently from the time of intervention), then we use a step variable. A step variable takes value zero before the intervention and one from the time of intervention onwards.
- Another form of permanent effect is a change of slope. Here the intervention is handled using a piecewise linear trend as discussed earlier (where is the time of intervention).

We still need to check the residuals

We have an idea now how to model some of time series issues that we want to account for. But we still need to see how well we are doing. Hint: throwing seasonal dummies into a regression doesn't explain everything....

Remember our decomposition of home sales:



There is still a great deal of variability in the "random" or "residual" category and there might be information there that we can/should capture.

One last statistical test

Statisticians love test statistics!

We've looked at accuracy measures (ME, MAE,RMSE, MAPE,...)
We've looked at ACF confidence intervals
We've looked and Ljung Box (and Box Pierce) statistics
Now one more: Durbin-Watson test for autocorrelation. This one is just to see if we have first order autocorrlation: y_t is correlated with y_{t-1}
We'll look at some examples in the code.

Summary

Tonight we covered:

- Reviewed linear regression
- Reviewed capturing non-linearities
- Introduced regression with time - including trend
- Took a quick look at random walks and spurious regressions
- Reviewed multiple regression
- Introduced multiple regression with time - including seasonality
- Reviewed residual effects

Now on to the R code...