Extensions and applications of linear regression

STAT315 April 3, 9, 10

Probable contents:

- Random versus fixed effects.
- Interaction terms.
- Additive models.
- Multivariate regression.

Random versus fixed effects

Sauer et al. analysis of the Breeding Bird Survey. 30 species of bird, with counts for year 1 and year 2. Aim: to determine which species are declining, becoming more abundant, or are stable.

(year 2 count) = trend x (year 1 count)

Model:

log(trend estimate for ith species) ~ $N(\mu, \sigma^2)$

Why?
Why would we assume that the trend estimates are a random sample from a normal distribution?

Regression example:

Working out the sale price for a second-hand car, based on:

- Mileage
- Age
- Make & model
- Etc.

It's clear that make and model are important and we shouldn't just throw this information away. But if we use the model of car as a categorical variable, we will have way too many categories and will be over-fitting (some models, like the Honda Civic, will appear often in our sample. But others, like the Lambourghini Countach, are likely to be very rare.) And what do we do if a particular type of car doesn't appear at all?

$$y = \beta_1(age) + \beta_2(mileage) + \beta_0[make] + \cdots + \varepsilon$$

Where $\beta_0[make] \sim \mathrm{N}(\mu, \sigma^2)$ for some unknown mean and standard deviation.

The $\beta_0[make]$ are numbers called **random effects**. The other β_i are called **fixed effects**.

A model like this is called a **multilevel model** or **hierarchical model** or **mixed model**.

What are the advantages of using random effects?

- More parsimonious; fewer parameters to estimate.

- Reflects the fact that the cars in our sample really *did* come from a larger population.

- Allows us to make predictions for a make/model we have never seen before: draw from the fitted $\mathrm{N}(\mu, \sigma^2)$ distribution.

- Can be shown to matter when real data generating process is of this form.

Downsides:

- Models are much harder to fit. If the model is complicated enough, may be practically impossible.
- Nobody can agree how to quantify uncertainty in the fitted regression coefficients.
- Nobody can agree how to make predictions from the fitted model.

Hopefully these problems will be overcome.

Reference: Gelman and Hill: ***Data Analysis Using Regression and Multilevel/Hierarchical Models***

Discussion: Random effects or fixed effects?

Scenario 1: You are interested in the effect of various soil variables on the growth of pine trees. You have measured the variables for eight pine trees and want to take account of variation between trees in your regression.

Scenario 2: You are studying the relationship between education and crime rate in the US and want to include the state as a categorical variable in your regression (i.e. "control" for differences between states.)

Interaction terms

Going back to the standard linear model, we often find that just doing a linear regression isn't flexible enough; the model has *high bias* in the parlance of an earlier lecture.

(re-draw bv tradeoff)

An **interaction** is a term of the form $x_i\ x_j$ where x_i and x_j are variables in the model. You can also have higher-order interactions.

Including interactions gives you a more flexible model (at the cost of …. ?)

An important principle: don't include the product $x_i\ x_j$ unless x_i and x_j are also in the model!

(Motivation: expand $(x_i - a)\ (x_j - b)\ \ldots$; inference should be invariant when you shift the variables.)

How can we test whether an interaction term is significant/should be added to the model?

- We could use cross-validation.
- We could use ANOVA.

Example: Boston housing data.

```
model <- lm(medv ~., data=Boston)
model2 <- lm(medv ~. + crim*lstat, data=Boston)
```

```
> anova(model, model2)
Analysis of Variance Table

   Res.Df RSS     Df   Sum of Sq  F      Pr(>F)
1  492    130.97
2  491    129.62  1    1.3515     5.1195 0.0241 *
```

(Go through calculation here as well)

10 x 10 fold cross-validation: no significant difference detected between cv scores of the two models (t-test, p=0.99).

Discussion Question:

Why do the two methods give different conclusions?

Which is correct?

How do you know when you should add an interaction?

Is it a good idea to keep trying all possible interactions until you find a significant one? Why or why not?

How else can you detect interactions?

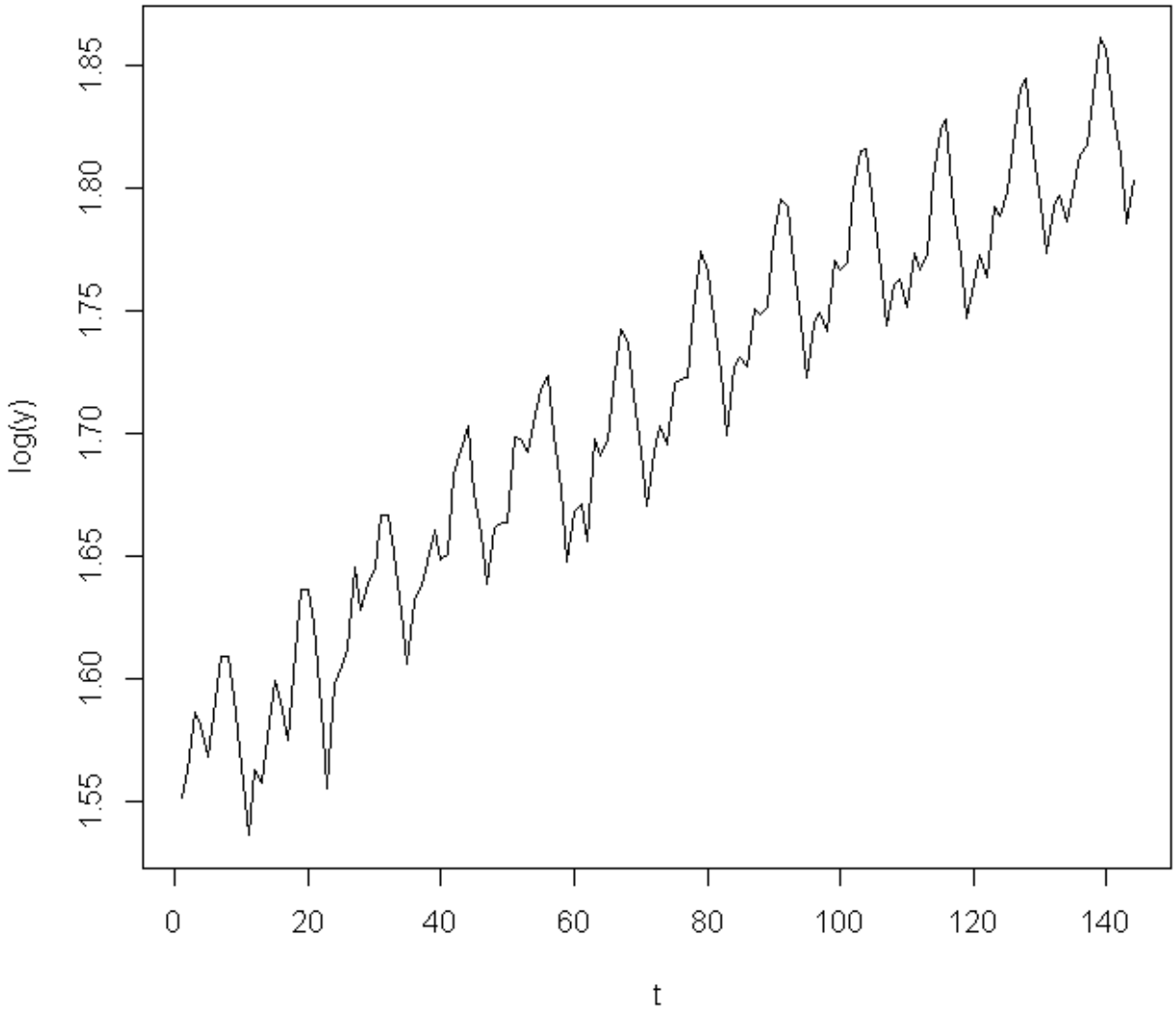
Additive Models

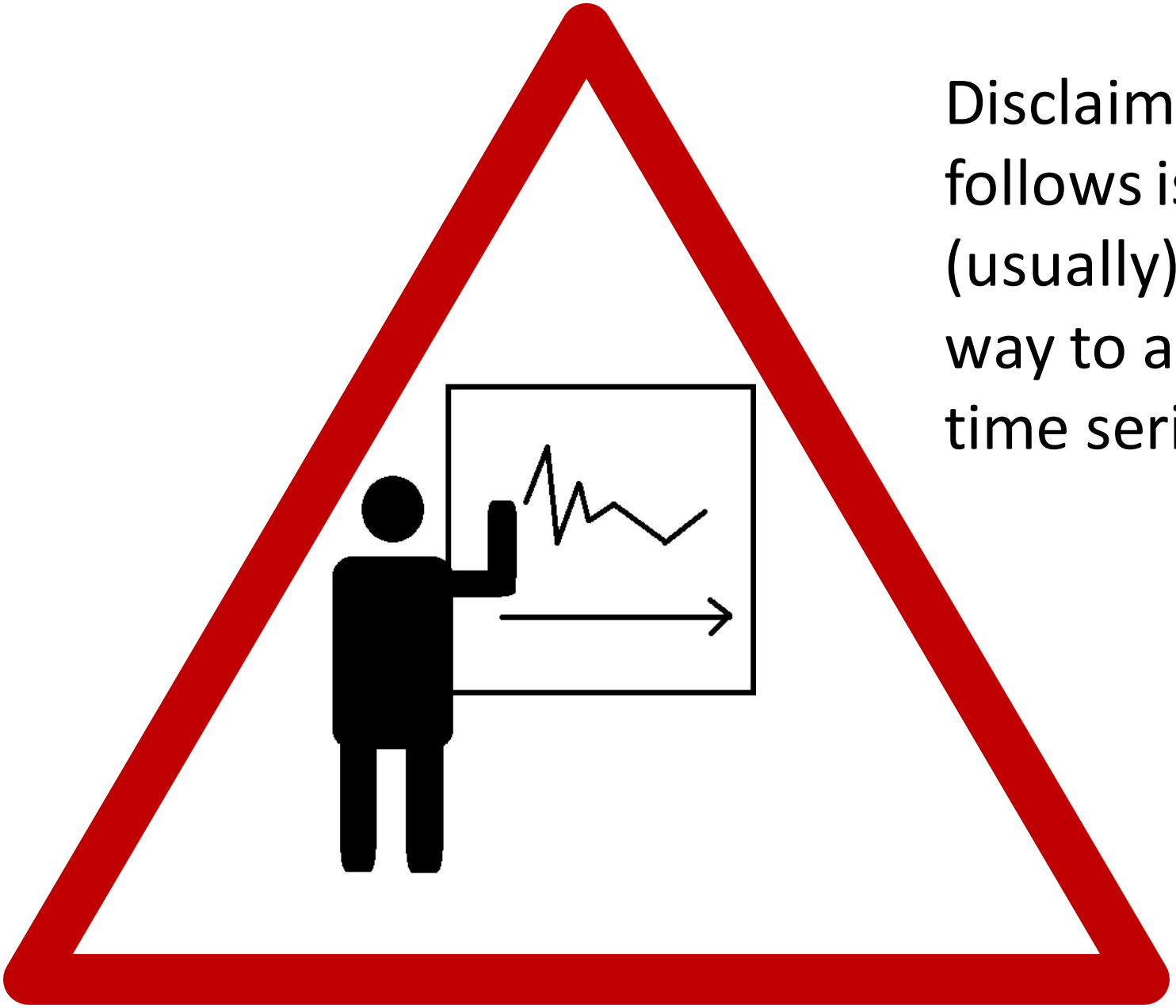
An **additive model** has the form

$$y = \beta_1 f_1(x_1) + \beta_2 f_2(x_2) + \ \ldots + \beta_p f_p(x_p)$$

where x is a vector of covariates, and the f's are functions. Notice that there is no need for a constant term since it could be part of one of the f's. Similarly, the β_i could be omitted. An additive model is just another way of saying "a model which is linear in some transformed versions of the predictors". Fitting an additive model is the same as transforming the x-variables.

Example: log(AirPassengers) measured monthly from 1949 to 1960. t=1 corresponds to January 1949.

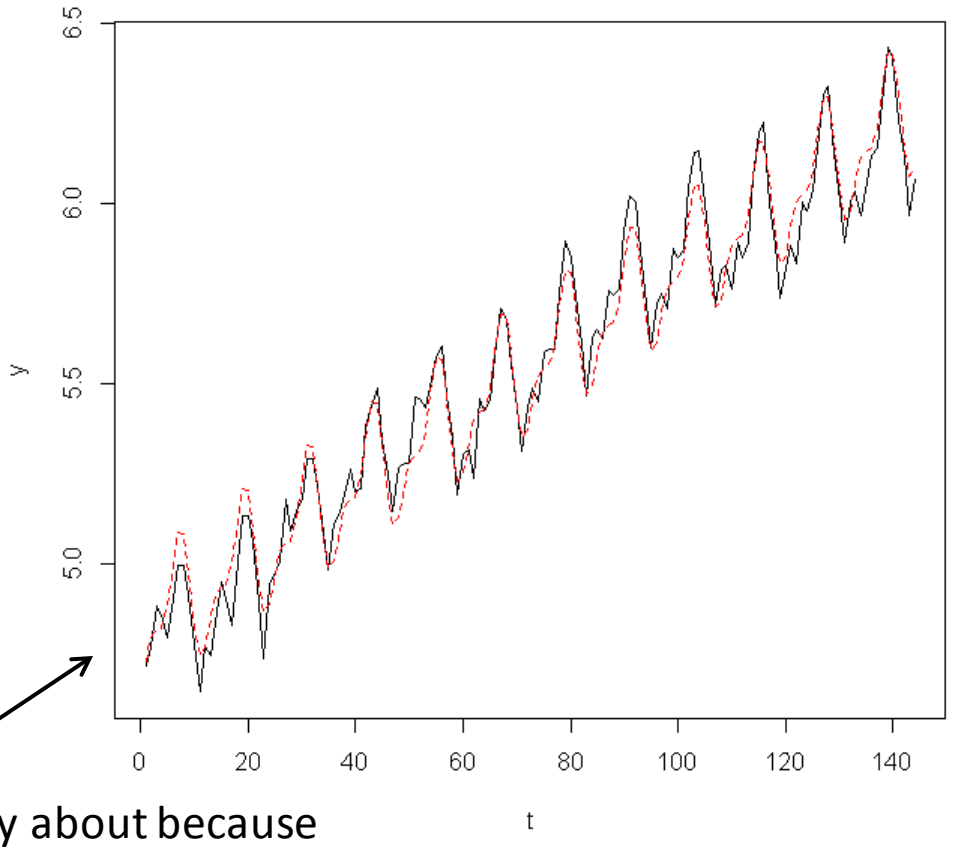




Disclaimer: What follows is **not** (usually) the right way to analyse time series data!

Seasonal Dummy (1 2 3 4 5 … 12 1 2 3 4 5 6 etc…)

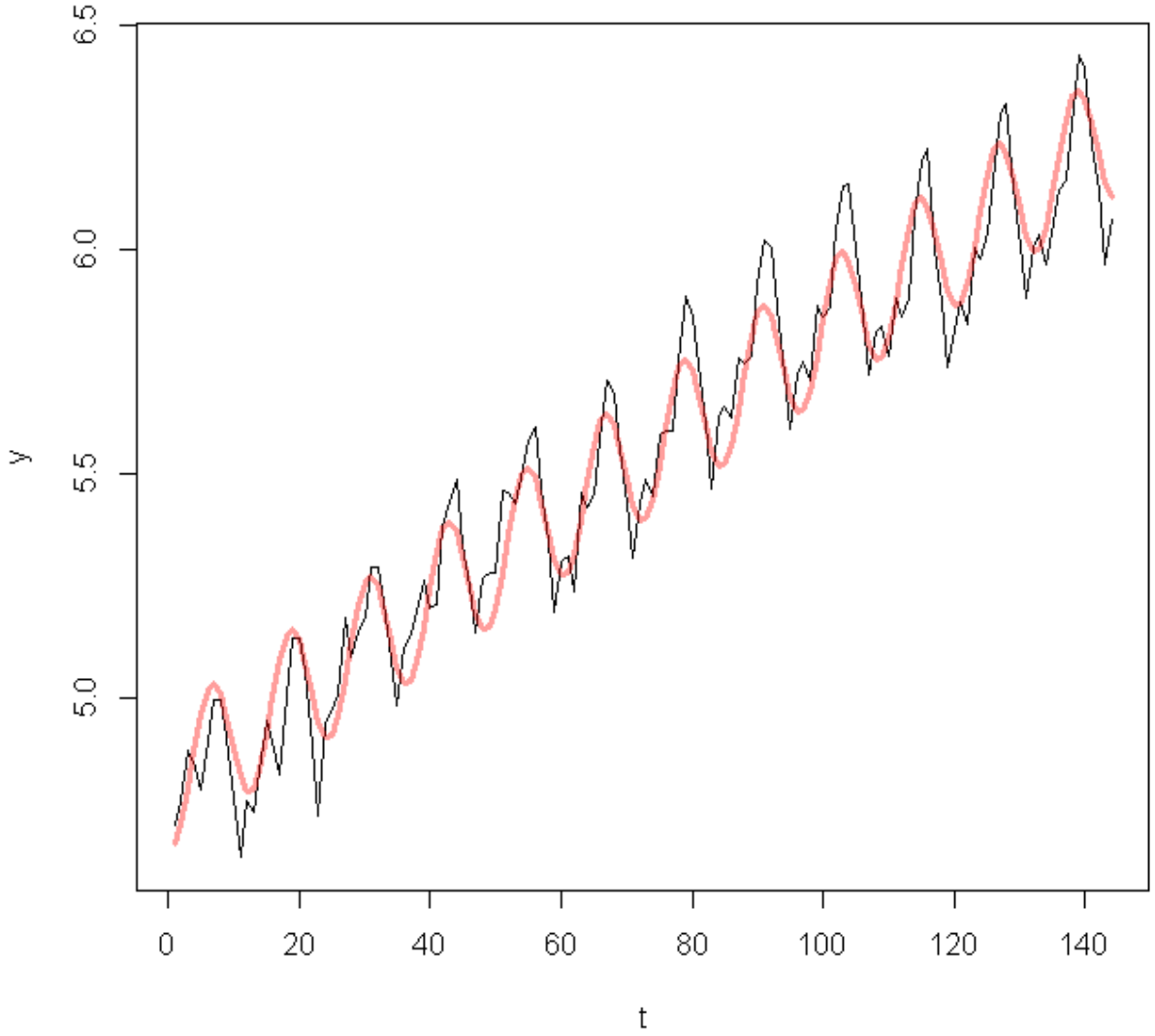
```
y <- log(AirPassengers)
n <- length(AirPassengers)
t <- 1:n
month <- rep(1:12, 50)[1:n]
model1 <- lm(y ~ t + factor(month))
#13 x-variables
plot(t, y, "l")
lines(t, model$fitted, col="red",
lty=2)
```



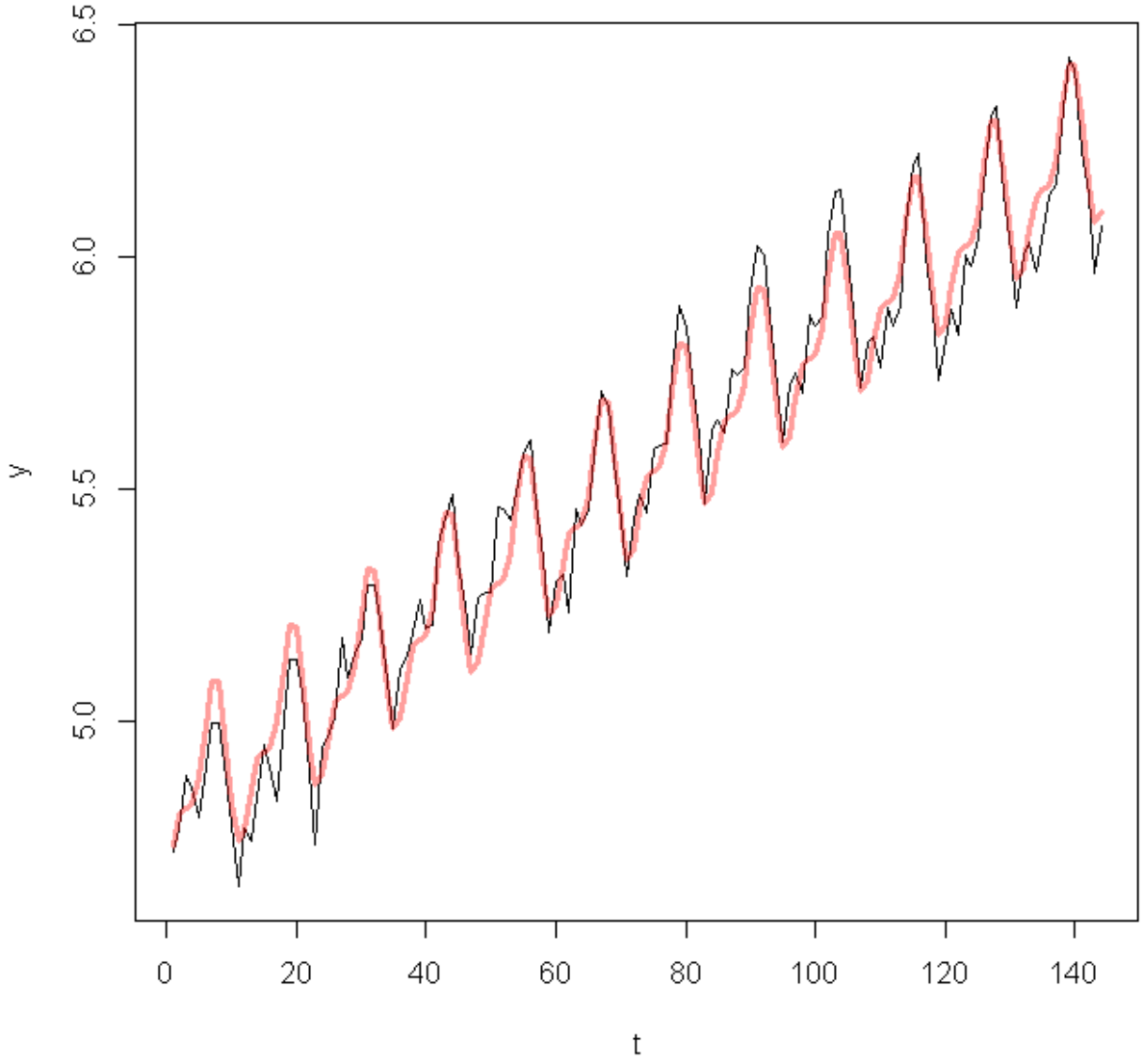
There are other things we should worry about because this is a time series problem (why would we be analysing these data anyway?), but this fit looks OK.

13 is a lot. Can we be more parsimonious?

```
# Try sin and cos terms
# Frequency 1/12:
s1 <- sin(2*pi*t/12)
c1 <- cos(2*pi*t/12)
model.trig1 <- lm(y ~ t + s1 + c1)
```



```
s2 <- sin(2* 2*pi*t/12)
c2 <- cos(2* 2*pi*t/12)
model.trig2 <- lm(y ~ t + s1 + c1 + s2 + c2)
```



```
> anova(model.trig1, model.trig2)
Analysis of Variance Table

Model 1: y ~ t + s1 + c1
Model 2: y ~ t + s1 + c1 + s2 + c2
  Res.Df     RSS Df Sum of Sq     F    Pr(>F)
1    140 1.12161
2    138 0.63864  2   0.48296 52.18 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

What if we want to compare the 13-parameter seasonal model with the 6-parameter model with sine and cosine terms of frequency 1/12 and 1/24? How can we compare these models?

Suddenly we can’t just use cross-validation.

Instead of finding your own transformations, you can use smoothing splines or loess-type smoothers to transform the predictors. This gives a **generalized additive model** or **GAM**.

In SAS: **`proc gam`**

Similar to **`proc reg`**

"there is hardly ever any reason to prefer linear models to additive ones, and the continued thoughtless use of linear regression is a scandal."

- Cosma Shalizi

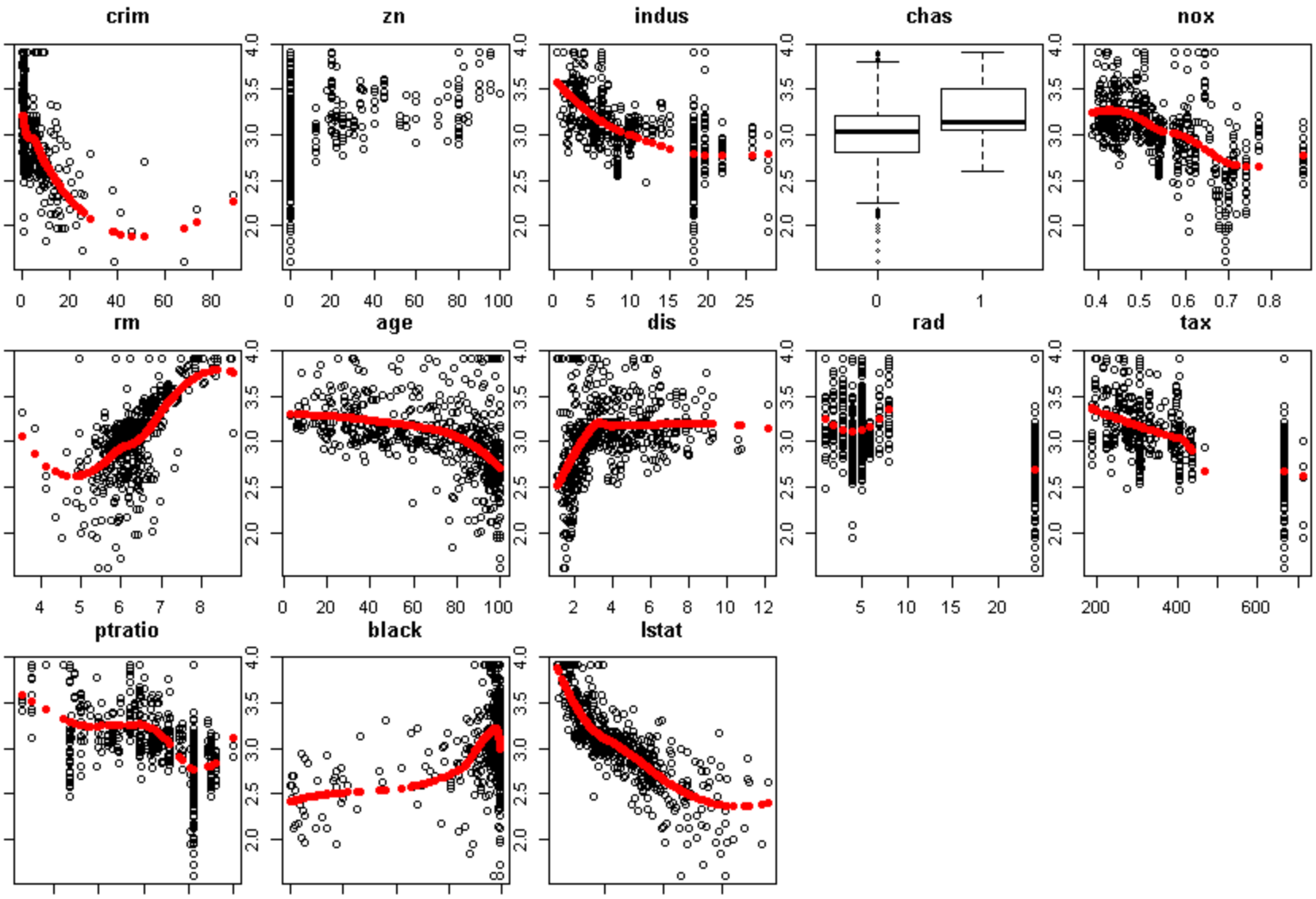
GAMs can be difficult to fit, but you can cheat and smooth the variables yourself.

On the Boston housing price data:

CV error of linear model `log(medv) ~.` was 19 ± 0.4.

CV error of GAM (using R's `loess`) is 17 ± 0.07.

Many of the plots on the next page look linear, but GAM sometimes predicts better at the edges. You could choose your own transformations instead of using loess.


crim
zn
indus
chas
nox
rm
age
dis
rad
tax
ptratio
black
lstat

Multivariate regression

Multivariate regression is where you have several predictor (x) variables and several response (y) variables. Each y-variable has its own regression on the x's.

We now have a matrix problem

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(demonstrate at board)

The matrix $\boldsymbol{X}$ is the same as in the one-variable problem. The least-squares solution is the same.

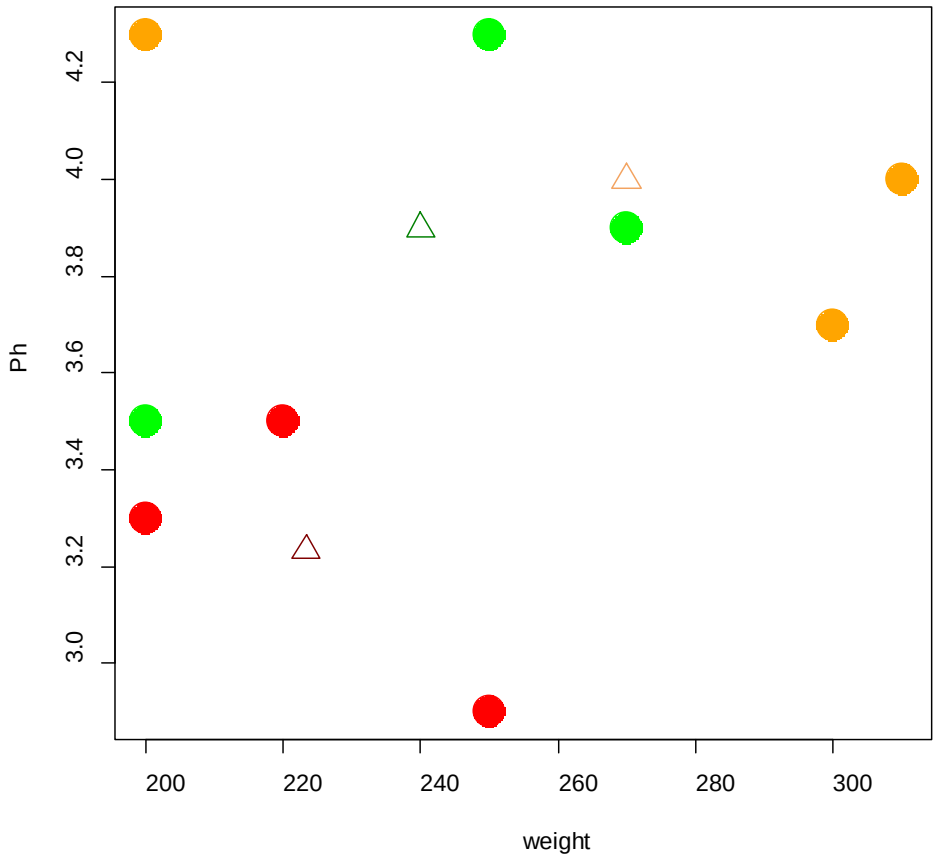
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

But now $\widehat{\boldsymbol{\beta}}$ is a matrix. The advantage of doing several regressions at once is that we can consider the y's simultaneously.

For example, what is the correlation structure between them? *(Usually if the y's are not correlated, there is no point in doing this kind of analysis. Also, if the y's are highly correlated, there is no point; it's like having just one y.)*

When the x's are categorical, this is called one-way MANOVA (multivariate analysis of variance)

	weight	Ph	fruit
1	200	3.3	apple
2	250	2.9	apple
3	220	3.5	apple
4	300	3.7	orange
5	310	4.0	orange
6	200	4.3	orange
7	200	3.5	pear
8	270	3.9	pear
9	250	4.3	pear



Q. Evidence of a difference in (weight, Ph) among different kinds of fruit?

```
lm(cbind(weight, Ph) ~ fruit, data=dat) -> model

> summary(manova(model))
          Df  Pillai approx F num Df den Df Pr(>F)
fruit      2 0.74359   1.7755      4     12 0.1986
Residuals  6
```

Test of H0 : all regression coefficients are zero except for constant term

vs.

H1: H0 is false.

Very similar to ANOVA.

Conclusion here: no evidence of a difference in mean (weight, Ph) between apples and pears and between apples and oranges.

Important question: why is this “better” than just doing a separate regression for each y-variable?

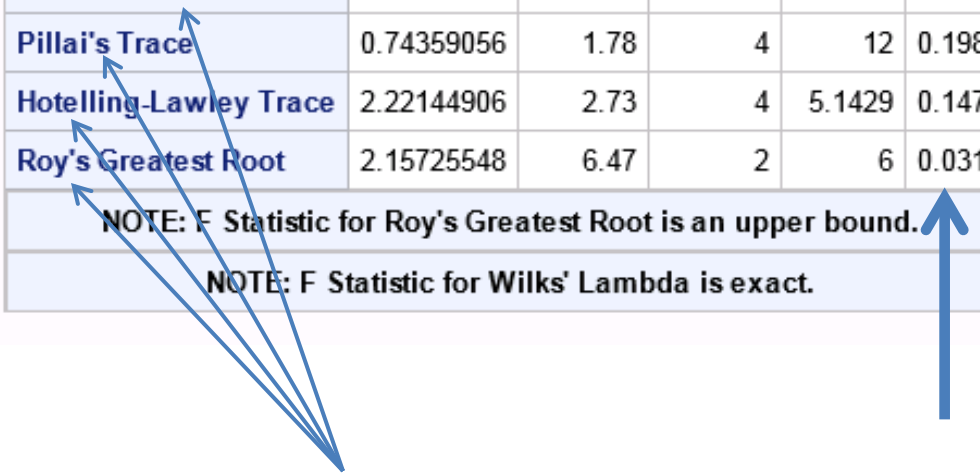
SAS version:

```
data fruit;
  input weight Ph apple orange pear;
  cards;
  200 3.3 1 0 0
  250 2.9 1 0 0
  220 3.5 1 0 0
  300 3.7 0 1 0
  310 4.0 0 1 0
  200 4.3 0 1 0
  200 3.5 0 0 1
  270 3.9 0 0 1
  250 4.3 0 0 1
run;

proc print;
 run;

proc reg data = fruit;
   model weight Ph = apple orange;
   M1: mtest;
run;
```

The SAS System

The REG Procedure
Model: MODEL1
Multivariate Test: M1

Multivariate Statistics and F Approximations					
S=2 M=-0.5 N=1.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.29762516	2.08	4	10	0.1581
Pillai's Trace	0.74359056	1.78	4	12	0.1986
Hotelling-Lawley Trace	2.22144906	2.73	4	5.1429	0.1473
Roy's Greatest Root	2.15725548	6.47	2	6	0.0318
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					



TMI !

You can also do these tests with continuous predictors (this is harder to do in R than in SAS). If x_i is the ith predictor, you can test

H0 : $\beta_i = 0$ for all the y's.

H1: some β_i is not zero.

e.g. in SAS, add `M2: mtest apple;`

In R, use the `anova` function. Note: the order in which you specify the variables in `lm` makes a difference to the results in R!

Example: Rohwer data. Three test scores SAT, PPVT, Raven; several variables give the scores of students on earlier test. Searching for an association between earlier scores and SAT+PPVT+Raven scores.

```
data Rohwer;
infile
"H:\SASfiles\Rohwer.tx
t";
input SAT PPVT Raven n
s ns na ss;
run;

proc print;
run;

proc reg data=Rohwer;
 model SAT PPVT Raven
= n s ns na ss;
 M1: mtest;
 M2: mtest n;
 M3: mtest n, s;
run;
```

Results Viewer - sashtml

The REG Procedure
Model: MODEL1
Multivariate Test: M1

Multivariate Statistics and F Approximations					
S=3 M=0.5 N=29.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.44010705	3.90	15	168.8	<.0001
Pillai's Trace	0.65935355	3.55	15	189	<.0001
Hotelling-Lawley Trace	1.05562391	4.23	15	110.09	<.0001
Roy's Greatest Root	0.81598982	10.28	5	63	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

The SAS System

The REG Procedure
Model: MODEL1
Multivariate Test: M2

Multivariate Statistics and Exact F Statistics					
S=1 M=0.5 N=29.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.94003599	1.30	3	61	0.2836
Pillai's Trace	0.05996401	1.30	3	61	0.2836
Hotelling-Lawley Trace	0.06378905	1.30	3	61	0.2836

Ironically, although the MANOVA-type tests are supposed to avoid multiple comparisons, they can offer pretty good scope for data dredging if not used with care.

However, in some research areas these techniques are extremely popular.

We haven't discussed the assumptions underlying the tests and what is actually being calculated. It would be better to take a specialist course if you wish to use these methods on real data. One important fact is that MANOVA assumes multivariate normality (there is no normality assumption like this in linear regression.)

Example writeup

A one-way multivariate analysis of variance (MANOVA) was conducted to determine the effect of the three types of study strategies (thinking, writing and talking) on two dependent variables (recall and application test scores). A nonsignificant Box's *M*, indicated a lack of evidence that the homogeneity of variance-covariance matrix assumption was violated. No univariate or multivariate outliers were evident and MANOVA was considered to be an appropriate analysis technique.

Significant differences were found among the three study strategies on the dependent measures, Wilks' = .42, $F(4,52) = 7.03$, $p < .001$. The multivariate Wilks' was quite strong at .35. Table 1 presents the means and standard deviations of the dependent variables for the three strategies.

Univariate analyses of variance (ANOVAs) for each dependent variable were conducted as follow-up tests to the MANOVA. Using the Bonferroni method for controlling Type I error rates for multiple comparisons, each ANOVA was tested at the .025 level. The ANOVA of the recall scores was significant, $F(2,27) = [..]$ while the ANOVA based on the application scores was nonsignificant, [..]

Post hoc analysis for the recall scores consisted of conducting pairwise comparisons to determine which study strategy affected performance most strongly. Each pairwise comparison was tested at the .025/3, or .008, significance level. The writing group produced significantly superior performance on the recall questions in comparison with either of the other two groups. The thinking and talking groups did not differ significantly from each other.

	Recall		Application	
Strategy	*M*	*SD*	*M*	*SD*
Thinking	3.30	0.68	3.20	1.23
Writing	5.80	1.03	5.00	1.76
Talking	4.20	1.14	4.40	1.17