Data, Environment and Society: Lecture 11: Multiple Regression

Instructor: Duncan Callaway
GSI: Seigi Karasaki

September 27, 2018

Announcements

Today

- First: slides, covering multiple regression and (one form of) model selection. Slides in GitHub
- Second: Start working with NO2 data in Jupyter notebook
- Third: Group discussion for Alstone et al

Reading

- Next tuesday: Novotny *et al*, see questions in GitHub folder for lecture 12 reading.
- Next week: ISLR Ch 3.3.

Survey posted! Please respond

Mid term

Not intended to be hard.

- Some basic theory, formula recall and application of mathematical concepts
  - Anything on slides or in labs
  - ISLR will reinforce, but I won't test things from ISLR not covered in lecture of lab.
- Principles of EDA and visualization
- Basic questions about working in Python
  - Setting up libraries
  - Accessing information from data frames
  - Etc.

Final project

- You can work with your own data
- But we will also suggest data sets
- Working in groups up to three ok but not required (you can self-organize)
- We will give you basic guardrails on what to do
  - Pose a coherent question that can be addressed using the skills we are learning
  - EDA and visualization requirements
  - Carry out multiple prediction exercises using the tools we are learning.
  - Critique the performance of your models
  - Interpret your results within the confines of what your models are capable of.

What if the confidence interval contains zero?

For example, if

$$-10.3 < \beta_1 < 24.8?$$

...where the upper and lower bounds comprise the 95% confidence interval.

What if the confidence interval contains zero?

For example, if

$$-10.3 < \beta_1 < 24.8?$$

...where the upper and lower bounds comprise the 95% confidence interval.

This implies there is more than a remote chance that there is no significant relationship between the dependent and independent variables.

p-values

p-values measure the probability that the estimated coefficients arose by chance from a data generating process that actually has *no* relationship between the inputs and outputs.

$p = 0.05$ implies a 5% chance that the true parameter value is *zero*.

If $p \ll 0.05$, then the parameter is strongly inside the 95% confidence interval.

If $p > 0.05$, then the parameter is outside the 95% confidence interval.

A small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance.

p-hacking?

What's wrong with these practices:

- Stop collecting data once $p < 0.05$
- Analyze many independent variables, but only report those for which $p < 0.05$
- Collect and analyze many data samples, but only report those with $p < 0.05$
- Exclude participants to get $p < 0.05$.
- Transform the data to get $p < 0.05$.

(credit to Leif Nelson, UCB Haas)

The trouble with p-hacking...

...is that by looking for the data set and the models that give low p-values, you could just be looking for those 5% “chances” where the real relationship is non-existent.

The trouble with p-hacking...

...is that by looking for the data set and the models that give low p-values, you could just be looking for those 5% “chances” where the real relationship is non-existent.

Some estimates suggest that this practice leads to false positive rates of 61%!

Model accuracy: R^2

TSS = total sum of squares

RSS = residual sum of squares

$$R^2 = \frac{TSS - RSS}{TSS}$$

Model accuracy: R^2

TSS = total sum of squares

RSS = residual sum of squares

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Model accuracy: R^2

TSS = total sum of squares

RSS = residual sum of squares

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

R^2 measures the fraction of variation in the dependent variable that is captured by the model.

Model accuracy: R²

TSS = total sum of squares

RSS = residual sum of squares

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

R^2 measures the fraction of variation in the dependent variable that is captured by the model.

It's good for capturing predictive power, but not for evaluating the significance of the model.

Multivariate regression

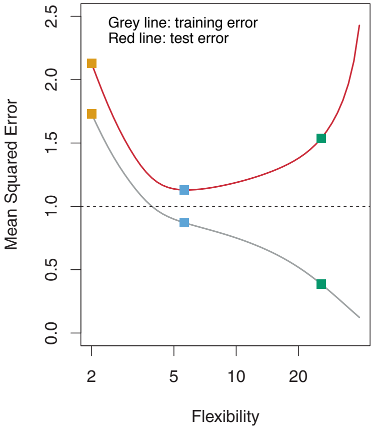
This is exactly the same process as single (independent) variable regression. Parameters solutions can be found by

- Gradient search
- Normal equations
- Setting partial derivatives of MSE to zero and solving – but now for $\beta_0, \beta_1, \beta_2, \ldots, \beta_d$ (d is the number of features, a.k.a. independent variables).

The mechanics of finding parameters is easy. The real challenge is: Which features to include?

Model selection

The challenge: Don't include variables in your model that lead to over-fit.



With multiple regression, increasing the number of variables increases the flexibility of the model.

Model selection methods

Two basic methods:

- Computationally heavy and theoretically robust:
  - repeated sampling of train and test data sets
  - build and test models with each sampled set
  - choose the model form that minimizes test error, on average.
  - the figure on the previous slide is an example of this approach.
- Easy to implement (no need for significant computing):
  - Use the full data set
  - Fit each candidate model once
  - Choose the model that minimizes an “adjusted” measure of R2 or mean squared error.

An easy-to-implement method

Akaike information criterion (AIC):

1. Construct all the models you have time for using *all* the data to train the models.
2. Then, choose the model with the lowest AIC, where

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}(\text{RSS} + 2d\hat{\sigma}^2)$$

An easy-to-implement method

Akaike information criterion (AIC):

1. Construct all the models you have time for using *all* the data to train the models.
2. Then, choose the model with the lowest AIC, where

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}(\text{RSS} + 2d\hat{\sigma}^2) = \frac{1}{\hat{\sigma}^2}\left(\frac{\text{RSS}}{n}\right) + \frac{2d}{n}$$

As you can see, AIC "penalizes" models with a high value of d.

What the heck is AIC?

It actually has a rigorous theoretical underpinning. Understanding the derivation requires background in information theory and more time than we have here.

What the heck is AIC?

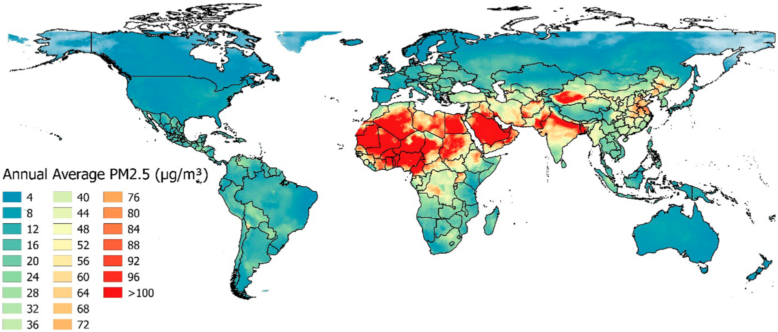
It actually has a rigorous theoretical underpinning. Understanding the derivation requires background in information theory and more time than we have here.

But:

- It gives unbiased estimate of the MSE you'd get if you *did* use a test data set (as long as the errors are Gaussian)
- It's ok to just work with the intuition that choosing models that minimize AIC is analogous to
  - choosing models that minimize MSE ...
  - plus a penalty for the number of features.

Prediction application: Land use regression

- Suppose we'd like to know pollutant concentrations at a fine spatial resolution
- We only have pollutant measurements at low resolution (coarse spatial scale)
- But we have other measurements at finer spatial resolution
- This is an ideal job for forecasting.
- But rather than forecast in *time* we will forecast in *space*.



(From Shaddick *et al* ES&T 2018)

Nitrogen dioxide

NO_2:

- Direct product of fossil fuel combustion
- Used as an indicator for larger group of nitrogen oxides.
- Health impact: Contributes to development of, and aggravates, asthma
- Environmental impact: Haze, acid rain, nutrient pollution in coastal waters

EPA Regulates NO2:

Nitrogen Dioxide (NO_2)	primary	1 hour	100 ppb	98th percentile of 1-hour daily maximum concentrations, averaged over 3 years
	primary and secondary	1 year	53 ppb [(2)]	Annual Mean

Novotny *et al* setup

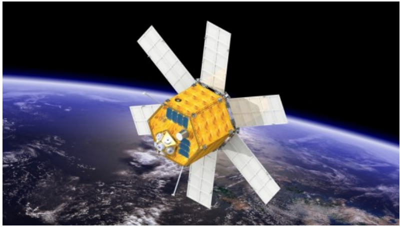
- NO_2 concentrations are known where monitors are present.
- But we don't have monitors everywhere
- Can we *predict* concentrations where monitors are absent?

Novotny *et al* setup

- NO_2 concentrations are known where monitors are present.
- But we don't have monitors everywhere
- Can we *predict* concentrations where monitors are absent?

"Remote sensing" data from satellites can be useful:

- Aurora satellite "Ozone Monitoring Instrument" provides tropospheric NO2 column abundance (units: ppb; Called "WRF+DOMINO" in data set we'll work with).

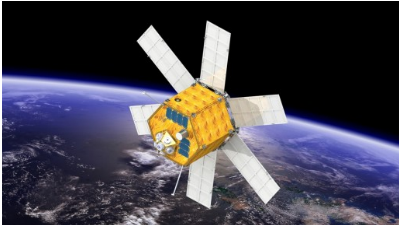


Novotny *et al* setup

- NO_2 concentrations are known where monitors are present.
- But we don't have monitors everywhere
- Can we *predict* concentrations where monitors are absent?

"Remote sensing" data from satellites can be useful:

- Aurora satellite "Ozone Monitoring Instrument" provides tropospheric NO2 column abundance (units: ppb; Called "WRF+DOMINO" in data set we'll work with).



But!

- Measurements are for entire column of air above a location, not ground-level
- Spatial resolution is low

Land use regression for NO_2

Dependent variable: Hourly NO_2 concentrations from EPA sensors.

Independent variables to consider:

parameter	units	spatial resolution	buffer[a] or point estimate
impervious surface	%	30 m (United States only[32]); 1000 m (global[29])	buffer
tree canopy	%	30 m (United States only[33]); 500 m (global[30])	buffer
population	no.	Census block (United States only[34]); 1 km (global[31])	buffer
major road length[35]	km	NA	buffer
minor road length[35]	km	NA	buffer
total road length[35]	km	NA	buffer
elevation[36]	km	90 m	point
distance to coast	km	NA	point
OMI NO_2[25,26]	ppb	13×24 km^2 at nadir	point

Novotny *et al* Table 1.

Let's run some linear regression models with these data. Move over to data hub.

Reading questions

- Review Figure 1 in detail. From a visualization perspective, what features of the figure do you appreciate? Do you think it could be improved upon?
- In the section **Electricity and human development**, the authors state that their figure is 'consistent with an aggregate view of household-level diminishing returns on energy consumption....' Discuss what the authors mean by this statement. Do you agree?
- In the section **The electricity continuum**, the authors argue that 'By overcoming access barriers, often through market-based structures, these systems provide incremental and often substantial increases in access to services, compared with the status quo.' Contrast this statement to the premise of Lee *et al* (which we read last week). Is Alstone *et al*'s view consistent or in conflict with Lee *et al*?

Supplemental

Important Questions for Multiple Linear Regression

- Is at least one of the predictors X1 , X2 , . . . , Xp useful in predicting the response?
  - cover only briefly
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
  - Review variable selection.
  - Cue attention to Marshall et al approach.
- How well does the model fit the data?
  - Return to question of Rsq
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?
  - Prediction intervals (contrast to confidence intervals)

Basic sketch of the Novotny et al paper:

"Stepwise multivariate regression":

1. The independent variable most correlated with the dependent variable is added to the model first
2. Of the remaining variables, the one most correlated with model residuals is selected as the next independent variable
3. Step 2 repeats on each new model with the remaining variables.
   - Variables are not kept in the model if $p > 0.05$
   - Variables are not kept in the model if they are collinear with others (we'll discuss this Tuesday).

Basic sketch for Novotny, ctd

Choosing the model:

- Model-building using a random sample of 90% of the monitoring data
- Tested the model's ability to predict the remaining 10%.
- Create 500 different random samples of training data – calculate R2, error, and bias for all 500 iterations.

Question: Why build the model with 90% of the data only?

Basic sketch for Novotny, ctd

Choosing the model:

- Model-building using a random sample of 90% of the monitoring data
- Tested the model's ability to predict the remaining 10%.
- Create 500 different random samples of training data – calculate R2, error, and bias for all 500 iterations.

Question: Why build the model with 90% of the data only?

Answer: to avoid choosing a model that over-fits the data.

In HW9, we'll go through a process similar (and arguably superior) to this one.

In HW6 – next week – we'll use a different method, known as AIC. We'll go over that today.