Machine learning

Multiple linear regression

Joshua Loftus

From the stars to "Poor Law Statistics"

- Almost a century after Gauss
- Scientists correlating/regressing anything
- Problem: what does it mean?

e.g. Francis Galton correlated numeric traits between generations of organisms...

But *why*? "Nature versus nurture" debate never ends?

e.g. Udny Yule and others correlated poverty ("pauperism") with welfare ("out-relief")...

But *why*? The "welfare trap" debate never ends?

Origin of multiple regression

- Udny Yule (1871-1951)
- Studied this poverty question
- First paper using multiple regression in 1897
- Association between poverty and welfare while "controlling for" age

Yule, in 1897:

> Instead of speaking of "causal relation," ... we will use the terms "correlation," ...

- Variables, roughly:
  - $Y =$ prevalence of poverty
  - $X_1 =$ generosity of welfare policy
  - $X_2 =$ age
- Positive correlations:
  - $\text{cor}(Y, X_1) > 0$
  - $\text{cor}(X_2, X_1) > 0$

Do more people enter/stay in poverty if welfare is more generous?

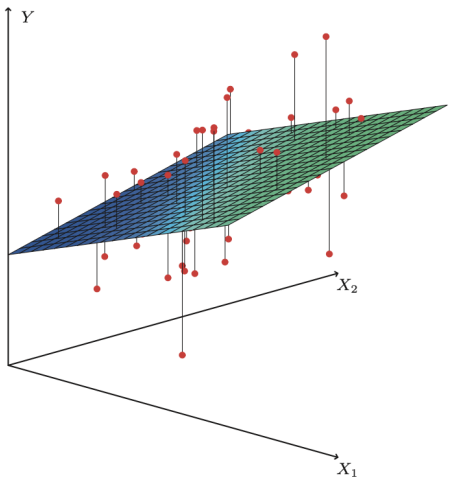
Or is this association "due to" age?

Yule, in 1897:

> The investigation of **causal relations** between economic phenomena presents many problems of peculiar difficulty, and offers many opportunities for fallacious conclusions.
>
> Since the statistician can seldom or never make experiments for himself, he has to accept the data of daily experience, and discuss as best he can the relations of a whole group of changes; he **cannot, like the physicist, narrow down the issue to the effect of one variation at a time. The problems of statistics are in this sense far more complex than the problems of physics**.

When $p > 1$

- Instead of a regression line, we fit a regression (hyper)plane
- Among all possible such planes, find the one minimizing sum of squared errors (represented by vertical lines in ISLR Fig 3.4)
- How to find the coefficients? Calculus?



Regression estimates when $p = 3$

836 YULE—*On the Theory of Correlation.* [Dec.

ndicate briefly the results for the case of four variables, which will serve to illustrate the very rapid growth in the complexity of formulæ and arithmetic as the number of variables increases.

If x_1, x_2, x_3, x_4 be associated deviations of the four variables from their respective means, the characteristic equation will be of the form

$$x_1 = b_{12}x_2 + b_{13}x_3 + b_{14}x_4 \quad \ldots \quad \ldots \quad (14).$$

The normal equations for the b's are in our previous notation—

$$\left.\begin{aligned} r_{12}\sigma_1 &= b_{12}\sigma_2 + b_{13}r_{23}\sigma_3 + b_{14}r_{24}\sigma_4 \\ r_{13}\sigma_1 &= b_{12}r_{23}\sigma_2 + b_{13}\sigma_3 + b_{14}r_{34}\sigma_4 \\ r_{14}\sigma_1 &= b_{12}r_{24}\sigma_2 + b_{13}r_{34}\sigma_3 + b_{14}\sigma_4 \end{aligned}\right\} \ldots \quad \ldots \quad (15).$$

Hence

$$b_{12} = \frac{r_{12}(1 - r_{34}^2) + r_{12}(r_{34}r_{24} - r_{23}) + r_{14}(r_{23}r_{34} - r_{24})}{(1 - r_{34}^2) + r_{23}(r_{34}r_{24} - r_{23}) + r_{24}(r_{23}r_{34} - r_{24})} \quad (16),$$

Progress(?) in regression

p variables $\rightarrow \binom{p}{2}$ "product sums" to compute by hand...

Yule:

> ... if we wished to discuss the causality [note: correlation?! -JL] of pauperism on the basis of as many as eight variables, the **work** involved would be something like twenty-eight times as much as that necessary for the example taken on pp. 824-831. The **labour** would, in fact, be almost impossible for a single individual.

- 1958: Ted Anderson *An Introduction to Multivariate Analysis*
- 1960's: **Electric** desktop calculators made it easier
- Present: linear algebra notation and computers (R, etc)

Notation

Writing the same thing in various ways

- For observation i:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

or using the inner product (of column vectors)

$$y_i = x_i^T \beta + \varepsilon_i$$

- Random variable version:

$$Y = X^T \beta + \varepsilon$$

Notation, continued

- For all n observations

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

Note: column of 1's for intercept term. Sometimes omitted by assuming $\mathbf{y}$ is already "centered"

Notational conventions

We'll use common conventions in this course

- Bold for vectors, bold and upper case for matrices
- Otherwise upper case denotes random variable
- Error terms $\varepsilon = y - \mathbf{x}^T\beta$ never truly observed
- Residuals $r = y - \mathbf{x}^T\hat{\beta}$ used as a proxy for errors
- Greek letters like $\beta, \theta, \sigma, \Sigma$ usually *unknown parameters*
- Greek letters with hats like $\hat{\beta}$ are estimates computed from data
- Roman letters that usually denote functions with hats, like $\hat{f}$ are also estimates
- Other Roman letters with hats like $\hat{y}$ are predictions

Matrices and vectors in R

```
# byrow = FALSE default
x <- matrix(1:9, nrow = 3, ncol = 3)
x
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
```

```
beta <- rep(1,3)
beta
```

```
## [1] 1 1 1
```

Multiplication: %*% (yes, really)

- Beware "Error: non-conformable arguments"
- Always remember to check dimensions
- If dimension of one object divides dimension of another, R may "conveniently" (unintuitively) repeat the smaller one

```
dim(x)
```

```
## [1] 3 3
```

```
dim(beta) # frustrating
```

```
## NULL
```

```
x %*% beta
```

```
##      [,1]
## [1,]   12
## [2,]   15
## [3,]   18
```


Transpose and symmetry

Recall: even if a matrix $\mathbf{A}$ is not square, both $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ are square and symmetric (often one is invertible)

```
A <- matrix(c(rep(1,4), 0,0), nrow = 3, byrow = FALSE)
A
```

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    1    0
## [3,]    1    0
```

```
t(A) %*% A
```

```
##      [,1] [,2]
## [1,]    3    1
## [2,]    1    1
```

Note: this matrix is even invertible! (But $\mathbf{A}\mathbf{A}^T$ is not)

Pseudoinversion

```
# ginv() function in MASS library
Ainv <- MASS::ginv(A)
Ainv
```

```
##              [,1] [,2] [,3]
## [1,] 1.578934e-16  0.5  0.5
## [2,] 1.000000e+00 -0.5 -0.5
```

```
Ainv %*% A
```

```
##      [,1]         [,2]
## [1,]    1 1.578934e-16
## [2,]    0 1.000000e+00
```

The 2×2 identity matrix

Pseudoinversion

Why does this work?

Let $\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$, then

$$\mathbf{A}^\dagger\mathbf{A} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{A} = (\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{A}) = \mathbf{I}$$

`ginv` in the library `MASS` apparently computes the left or right pseudoinverse, whichever one works

```
A %*% Ainv # not a right inverse
```

```
##              [,1]         [,2]         [,3]
## [1,] 1.000000e+00 5.551115e-17 5.551115e-17
## [2,] 1.578934e-16 5.000000e-01 5.000000e-01
## [3,] 1.578934e-16 5.000000e-01 5.000000e-01
```

Least-squares solutions in matrix notation

Instead of those long expressions that Yule found were already very complicated with $p = 3$, we can always write very simply:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^\dagger\mathbf{y}$$

This assumes $\mathbf{X}^T\mathbf{X}$ to be invertible, i.e. the *columns* of $\mathbf{X}$ have full rank (columns = variables)

- That's often true if $n > p$, unless some problem like one variable is a copy of another
- Impossible if $p > n$. "High-dimensional" regression requires special methods, covered soon in this course!

Linear algebra and geometric intuition

Predictions from the linear model:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}y$$

if we define

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

COOL FACTS about $\mathbf{H}$:

- $\mathbf{H}$ is a projection: $\mathbf{H}^2 = \mathbf{H}$
- For any n-vector $\mathbf{v}$, the n-vector $\mathbf{H}\mathbf{v}$ is the orthogonal projection of $\mathbf{v}$ onto the column space of $\mathbf{X}$
- Of all linear combinations of columns of $\mathbf{X}$, $\mathbf{H}\mathbf{v}$ is the one closest (in Euclidean distance) to $\mathbf{v}$.

Exercise: do the calculus

We have the loss function

$$L(\mathbf{X}, \mathbf{y}, \beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

(just a different way of writing sum of squared errors)

- Consider each coordinate separately and take univariate partial derivatives
- Use vector calculus and compute the gradient
- (Or even use matrix calculus identities)

Reach the same conclusion: at a stationary point of L,

$$\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{y}$$

Categorical predictors

This is an interesting/practically important special case

```
x <- as.factor(
  sample(c("A", "B", "C"),
         10,
         replace = TRUE))
x
```

```
##  [1] A C A B A C C B B C
## Levels: A B C
```

Categorical predictor with 3 levels, what does the **design matrix** (common terminology in regression) $\mathbf{X}$ look like?

```
model.matrix(~x)
```

```
##    (Intercept) xB xC
## 1            1  0  0
## 2            1  0  1
## 3            1  0  0
## 4            1  1  0
## 5            1  0  0
## 6            1  0  1
## 7            1  0  1
## 8            1  1  0
## 9            1  1  0
## 10           1  0  1
## attr(,"assign")
## [1] 0 1 1
## attr(,"contrasts")
## attr(,"contrasts")$x
## [1] "contr.treatment"
```

```r
X <- model.matrix(~x-1) # take out intercept
head(X)
```

```
##   xA xB xC
## 1  1  0  0
## 2  0  0  1
## 3  1  0  0
## 4  0  1  0
## 5  1  0  0
## 6  0  0  1
```

```r
round(head(X %*% MASS::ginv(X)), 3) # hat matrix
```

```
##    [,1] [,2]  [,3]  [,4]  [,5] [,6] [,7]  [,8]  [,9] [,10]
## 1 0.333 0.00 0.333 0.000 0.333 0.00 0.00 0.000 0.000  0.00
## 2 0.000 0.25 0.000 0.000 0.000 0.25 0.25 0.000 0.000  0.25
## 3 0.333 0.00 0.333 0.000 0.333 0.00 0.00 0.000 0.000  0.00
## 4 0.000 0.00 0.000 0.333 0.000 0.00 0.00 0.333 0.333  0.00
## 5 0.333 0.00 0.333 0.000 0.333 0.00 0.00 0.000 0.000  0.00
## 6 0.000 0.25 0.000 0.000 0.000 0.25 0.25 0.000 0.000  0.25
```

```r
which(x == "C") # predicting within-group averages!
```

```
## [1]  2  6  7 10
```


Differences from simple regression

- **Interpreting coefficients**: "*ceteris paribus*" -- all other things being equal

The status quo is *ridiculous*, but I must tell you... (I will also tell you about better ways)

- **Diagnostic plots**: can't see higher dimensional relationships

Plot residuals vs fitted values, and/or various pair plots

`GGally` package 👍

- **Inference**: testing multiple coefficients

See ISLR's discussion of F-tests, beginning of Section 3.2.2

Interpreting coefficients

People *want* these two things to be true:

1. $$\frac{\partial}{\partial x_j}\mathbb{E}[\mathbf{y}|\mathbf{X}] = \beta_j \approx \hat{\beta}_j$$

2. β_j is a causal parameter, i.e. **intervening** to increase x_j by 1 unit would result in conditional average of y changing by β_j units

Both of these can be *importantly wrong*! Always remember:

- Think about *general* conditional expectation functions vs the **linear model assumption** (strength of ML!)
- Think about **relationships between predictors** (causal or associations)
- Consider **unobserved variables** not in the dataset

Non-linear example

Suppose there is one predictor x, and a (global) non-linear model fits the CEF:

$$\mathbb{E}[\mathbf{y}|X = x] = \beta_0 + \beta_1 x + \beta_2 x^2$$

We don't know the β's but we have some data, and we use multiple linear regression to fit the coefficients

```
x2 <- x^2
lm(y ~ x + x2)
```

But, there's a **problem**:

$$\frac{\partial}{\partial x}\mathbb{E}[\mathbf{y}|x] = \beta_1 + 2\beta_2 x \neq \beta_1 \approx \hat{\beta}_1$$

What went wrong?

In this simple example we know the problem is that x_2 is actually a function of x. **Solution**: interpret $\frac{\partial}{\partial x}$ locally as a function of x, not as a global constant

Sometimes simplifying assumptions are **importantly wrong**, and we must reject simple interpretations

Machine learning provides tools for fitting more complex models, like non-linear models

Interpreting causality

- In the real world when is it true that one predictor variable *does not depend* on any of the other ones?
- Or when is it true that there are no important **unobserved confounders**, variables that are related to both the predictor and the outcome?

Consider Yule's regression analysis of poverty

Other important but unobserved variables?

Reverse causation? Simultaneity (feedback loop)?

In the real world

ceteris is never *paribus*

Fortunately, we will also study "causal inference" - a field with methods specialized for interpreting coefficients the way *people generally want to*

Remember Yule!

[We] cannot, like the physicist, narrow down the issue to the effect of ***one variation at a time***

Collinearity between predictors

Another important difference from simple regression

Related to problems with interpreting coefficients

For a set of predictors, the more closely they are mutually linearly dependent the more difficult it is to estimate their separate coefficients

If the problem is bad enough, can result in numerical instability

Assessing if it's a problem

- For each predictor x_j, treat it as an outcome in a regression model using all the other predictors
- R^2 of this model: closer to 1, worse collinearity

Concluding points

One of the most commonly used methods, even with more complex ML often compare to regression as a "baseline"

Perhaps the most complex method that is still considered relatively interpretable. But interpretation is actually trickier than most understand! *Ceteris paribus* and causality...

Always remember bias, even if sample is large our estimates could be far from truth

Could be estimating the wrong thing, using a model that's importantly wrong, asking the wrong question, analyzing/collecting the wrong data, including wrong predictors, etc

the end

Wisdom from one of the great early statistical explorers

Udny Yule:

> Measurement does not necessarily mean progress. Failing the possibility of measuring that which you desire, the lust for measurement may, for example, merely result in your measuring something else - and perhaps forgetting the difference - or in your ignoring some things because they cannot be measured.