Multiple linear regression: estimation and model fitting

Patrick Breheny

January 25

Introduction

- The goal of today's class is to set up a multiple regression model in terms of matrices and then solve for the regression coefficients, using the results we introduced last Thursday
- Our data set for today consists of daily measurement of air quality (in terms of ozone concentration) taken in New York during the summer of 1973

Ozone

- While the ozone layer in the upper atmosphere is beneficial and protects us from ultraviolet light, in the lower atmosphere it is a pollutant that has been linked to a number of respiratory diseases as well as heart attacks and premature death
- The EPA's national air quality standard for ozone concentration is 75 parts per billion (ppb); in Europe, the standard is 60 ppb; and according to some studies, at-risk individuals may be adversely affected by ozone levels as low as 40 ppb
- Ozone concentrations, however, are not constant, and fluctuate quite a bit from day to day, depending on many factors

Ozone data set

The file ozone.txt contains the following variablesozone.txt

- `Ozone`: Ozone concentration (in ppb)
- `Solar`: Solar radiation (in Langleys)
- `Wind`: Average wind speed (in miles/hour)
- `Temp`: Daily high temperature (in Fahrenheit)
- `Day`: Day of the year, with January $1 = 1$, February $1 = 32$, etc.

We will be considering ozone concentration to be the outcome variable and the rest as explanatory variables

Scatterplot matrices

- A useful way of visualizing multivariate relationships is with *scatterplot matrices*:
  - In `R`,

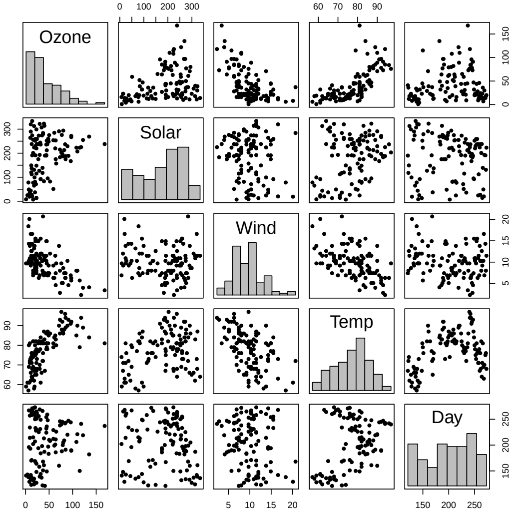
    ```
    pairs(ozone)
    ```

  - In SAS,

    ```
    PROC SGSCATTER;
        MATRIX Ozone Solar Wind Temp Day;
    RUN;
    ```

- Both SAS (using the `DIAGONAL` option) and `R` (using the `diag.panel` option) come with options allowing you to also plot things like histograms and kernel density estimates along the diagonal

Scatterplot matrix for the ozone data



The regression model

- Let Y_i represent the ozone concentration on day i, x_{i1} represent solar radiation on day i, x_{i2} represent wind speed on day i, and so on
- A linear regression model for ozone concentration could then be written as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

- Or equivalently, letting $\mathbf{x}_i^T = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4})$,

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

- Or still equivalently, letting $\mathbf{X}$ be the 111×5 matrix with rows $\mathbf{x}_i^T$,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The design matrix

- The quantity $\mathbf{X}$ is referred to as the *design matrix*
- Note that each column of $\mathbf{X}$ corresponds to a different explanatory variable, and each row of $\mathbf{X}$ corresponds to a different observation
- Thus, if there are p explanatory variables (counting the intercept) and n observations, $\mathbf{X}$ will have dimension $n \times p$
- Remarks:
  - The name "design matrix" is used regardless of whether $\mathbf{X}$ was actually chosen by design (as it might be in, say, a controlled experiment) or not
  - While certainly important to the scientific conclusions, whether $\mathbf{X}$ is chosen by design or not has no impact on the statistical modeling, as $\mathbf{X}$ is considered to be fixed (*i.e.*, not random) in the regression setting

Solving for $\hat{\boldsymbol{\beta}}$

- Just as in simple linear regression, a reasonable way to estimate the regression coefficients is to choose the ones which minimize the residual sum of squares:

$$\begin{aligned} RSS &= \sum r_i^2 \\ &= \mathbf{r}^T\mathbf{r} \end{aligned}$$

  where $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is an $n \times 1$ vector with elements $r_i = y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{i4}\beta_4$

- **Proposition:** Suppose $(\mathbf{X}^T\mathbf{X})^{-1}$ exists. Then the value $\hat{\boldsymbol{\beta}}$ for the regression coefficients that minimizes the residual sum of squares is given by

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

Fitting linear regression models in SAS/R

- There is essentially no difference between simple and multiple linear regression as far as the syntax in SAS/R is concerned:
  - In R,

    ```
    lm(Ozone~Solar+Temp+Wind+Day)
    ```
  - In SAS,

    ```
    PROC REG;
        MODEL Ozone = Solar Wind Temp Day;
    RUN;
    ```

Manual fitting

We can also solve for these coefficients manually using the results we derived earlier:

```
y <- Ozone
X <- cbind(1,as.matrix(ozone[,-1]))
beta <- solve(t(X) %*% X) %*% t(X) %*% y
```

Notes:

- `cbind`: binds objects together by column (there is also an `rbind` column for rows)
- `t`: Take the transpose of a matrix
- `solve`: Take the inverse of a matrix
- `%*%`: Matrix multiplication

Comparison to univariate solutions

- Below is a table comparing the estimates obtained from simple linear regression and multiple regression

	Multiple regression	Simple regression
Solar	0.05	0.13
Wind	-3.32	-5.73
Temp	1.83	2.44
Day	-0.08	0.10

- Keep in mind the interpretation:
  - As wind speed goes up by 1 mile/hour, ozone levels go down by 5.7 ppb
  - As wind speed goes up by 1 mile/hour, *but solar radiation, temperature, and day of the year stay the same*, ozone levels go down by 3.3 ppb

Comparison to univariate solutions (cont'd)

Remarks:

- When unadjusted for confounding correlations (especially between wind and temperature), simple linear regression systematically overestimates the effect of the explanatory variables
- This is the pattern observed in this particular data set, but it is not a rule: systematic underestimation can also occur

Standardized regression coefficients

- Note that regression coefficients are dependent on the scale of measurement
- For example, if we measured temperature in Celsius instead of Fahrenheit, its regression coefficient would be 3.29 instead of 1.83
- Thus, looking directly at regression coefficients has the potential to be misleading in terms of judging relative importance

Standardized regression coefficients

- In the univariate case, correlation is often used to put measurements on a common scale
- As we saw last week,

$$r = \hat{\beta}\frac{s_x}{s_y}$$

- Applying this logic to multiple regression yields

	Original		Standardized	
	Multiple	Simple	Multiple	Simple
Solar	0.05	0.13	0.14	0.35
Wind	-3.32	-5.73	-0.35	-0.61
Temp	1.83	2.44	0.52	0.70
Day	-0.08	0.10	-0.11	0.14

Standardized regression coefficients in R/SAS

- Note that this is equivalent to standardizing all the variables, then performing the regression:

```
lm(Ozone~0+Solar+Wind+Temp+Day,
   data=as.data.frame(scale(ozone)))
```

- In SAS, one can obtain the standardized regression coefficients with the STB option:

```
PROC REG;
    MODEL Ozone = Solar Wind Temp Day / STB;
RUN;
```

The hat matrix

- Let $\hat{\mathbf{y}}$ denote the fitted values of $\mathbf{y}$ arising from the multiple regression model
- Note that

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}}\\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\\ &= \mathbf{H}\mathbf{y}\end{aligned}$$

  where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- The matrix $\mathbf{H}$ is often called the *hat matrix*, because it "puts the hat on $\mathbf{y}$"
- It is also called the *projection matrix* because it projects $\mathbf{y}$ onto the *column space* of $\mathbf{X}$ (the vector space spanned by the columns of $\mathbf{X}$)

Predictions

- An actual prediction of the ozone level for a new day not used in the fitting of the model would be

$$\hat{\mu} = \mathbf{x}^T \hat{\boldsymbol{\beta}}$$

  where $\mathbf{x}$ is the vector containing that day's wind speed, temperature, solar radiation, and day (along with a 1 for the intercept)
- So, for example, the predicted ozone level on day 274, if that day was 70 degrees with 180 Langleys of solar radiation and 15 mph wind speed, would be 5.16 ppb
- If that day instead had a temperature of 90 and a wind speed of 5 mph, our predicted ozone level would be 74.9

The Hessian matrix

- Thus far, we have been somewhat loose in claiming that setting the derivative equal to zero is equivalent to minimizing the residual sum of squares
- To actually prove this, we need the multivariate version of the "second derivative test"
- Let $\nabla^2_{\mathbf{x}} f(\mathbf{x})$ denote the second derivative of a function $f(\mathbf{x})$:

$$\nabla^2_{\mathbf{x}} f(\mathbf{x}) \equiv \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = \frac{\partial}{\partial \mathbf{x}} \left(\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) \right)$$

  This matrix is referred to as the *Hessian* of the function f (after the German mathematician Ludwig Hesse)

The multivariate second derivative test

- **Theorem:** Suppose f is a scalar-valued function of a vector $\mathbf{x}$ and that $\nabla^2_{\mathbf{x}} f(\mathbf{x})$ is positive definite. If $\mathbf{x}_0$ is a point satisfying $\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}_0) = \mathbf{0}$, then $\mathbf{x}_0$ is a unique global minimum of $f(\mathbf{x})$.
- In the linear regression case,

$$\nabla^2_{\boldsymbol{\beta}} RSS = 2\mathbf{X}^T\mathbf{X}$$

- This quantity is positive definite if and only if the rank of $\mathbf{X}$ is equal to p (*i.e.*, $\mathbf{X}$ has *full column rank*)

Summary

To summarize, there are two possibilities:

- $\mathbf{X}$ is full rank, $\mathbf{X}^T\mathbf{X}$ is positive definite and invertible, and there is exactly one unique value of $\boldsymbol{\beta}$ which minimizes the RSS
- The rank of $\mathbf{X}$ is less than p, $\mathbf{X}^T\mathbf{X}$ is positive semidefinite, not invertible, and there are an infinite number of solutions which minimize the RSS

Example: Non-full-rank design

- For example, suppose we define a new variable in our ozone data which is a linear combination of the others, and try to fit the model:

```
Extra <- 3*Solar - 2*Wind + 0.5*Temp
lm(Ozone~Solar+Wind+Temp+Day+Extra)
XX <- cbind(1,Solar,Wind,Temp,Day,Extra)
solve(crossprod(XX))
```

- No unique solution for $\hat{\boldsymbol{\beta}}$ exists; multiple values produce the exact same fitted values $\hat{\boldsymbol{\mu}}$ and thus the exact same RSS
- Remark: although $\hat{\boldsymbol{\beta}}$ is not unique, the fitted values $\hat{\boldsymbol{\mu}}$ are unique, as are all predictions of future observations