# Multivariable regression

Brian Caffo, Roger Peng and Jeff Leek
Johns Hopkins Bloomberg School of Public Health

# Multivariable regression analyses

- If I were to present evidence of a relationship between breath mint useage (mints per day, X) and pulmonary function (measured in FEV), you would be skeptical.
  - Likely, you would say, 'smokers tend to use more breath mints than non smokers, smoking is related to a loss in pulmonary function. That's probably the culprit.'
  - If asked what would convince you, you would likely say, 'If non-smoking breath mint users had lower lung function than non-smoking non-breath mint users and, similarly, if smoking breath mint users had lower lung function than smoking non-breath mint users, I'd be more inclined to believe you'.
- In other words, to even consider my results, I would have to demonstrate that they hold while holding smoking status fixed.

# Multivariable regression analyses

- An insurance company is interested in how last year's claims can predict a person's time in the hospital this year.
  - They want to use an enormous amount of data contained in claims to predict a single number. Simple linear regression is not equipped to handle more than one predictor.
- How can one generalize SLR to incoporate lots of regressors for the purpose of prediction?
- What are the consequences of adding lots of regressors?
  - Surely there must be consequences to throwing variables in that aren't related to Y?
  - Surely there must be consequences to omitting variables that are?

# The linear model

- The general linear model extends simple linear regression (SLR) by adding terms linearly into the model.

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i = \sum_{k=1}^{p} X_{ik} \beta_j + \epsilon_i$$

- Here $X_{1i} = 1$ typically, so that an intercept is included.
- Least squares (and hence ML estimates under iid Gaussianity of the errors) minimizes

$$\sum_{i=1}^{n} \left( Y_i - \sum_{k=1}^{p} X_{ki} \beta_j \right)^2$$

- Note, the important linearity is linearity in the coefficients. Thus

$$Y_i = \beta_1 X_{1i}^2 + \beta_2 X_{2i}^2 + \ldots + \beta_p X_{pi}^2 + \epsilon_i$$

  is still a linear model. (We've just squared the elements of the predictor variables.)

# How to get estimates

- The real way requires linear algebra. We'll go over an intuitive development instead.
- Recall that the LS estimate for regression through the origin, $E[Y_i] = X_{1i}\beta_1$, was $\sum X_i Y_i / \sum X_i^2$.
- Let's consider two regressors, $E[Y_i] = X_{1i}\beta_1 + X_{2i}\beta_2 = \mu_i$.
- Also, recall, that if $\hat \mu_i$ satisfies

$$\sum_{i=1} (Y_i - \hat \mu_i) (\hat \mu_i - \mu_i) = 0$$

  for all possible values of $\mu_i$, then we've found the LS estimates.

$$\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = \sum_{i=1}^{n}(Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})\left\{ X_{1i}(\hat{\beta}_1 - \beta_1) + X_{2i}(\hat{\beta}_2 - \beta_2)\right\}$$

- Thus we need
  1. $\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})X_{1i} = 0$
  2. $\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})X_{2i} = 0$

- Hold $\hat{\beta}_1$ fixed in 2. and solve and we get that

$$\hat{\beta}_2 = \frac{\sum_{i=1}(Y_i - X_{1i}\hat{\beta}_1)X_{2i}}{\sum_{i=1}^{n} X_{2i}^2}$$

- Plugging this into 1. we get that

$$0 = \sum_{i=1}^{n}\left\{ Y_i - \frac{\sum_j X_{2j} Y_j}{\sum_j X_{2j}^2} X_{2i} + \hat{\beta}_1 \left( X_{1i} - \frac{\sum_j X_{2j} X_{1j}}{\sum_j X_{2j}^2} X_{2i} \right)\right\} X_{1i}$$

# Continued

- Re writing this we get

$$0 = \sum_{i=1}^{n} \left\{ e_{i,Y|X_2} - \hat{\beta}_1 e_{i,X_1|X_2} \right\} X_{1i}$$

where $e_{i,a|b} = a_i - \frac{\sum_{j=1}^{n} a_j b_j}{\sum_{i=1}^{n} b_j^2} b_i$ is the residual when regressing $b$ from $a$ without an intercept.

- We get the solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} e_{i,Y|X_2} e_{i,X_1|X_2}}{\sum_{i=1}^{n} e_{i,X_1|X_2} X_1}$$

- But note that

$$\sum_{i=1}^{n} e_{i,X_1|X_2}^2 = \sum_{i=1}^{n} e_{i,X_1|X_2}\left(X_{1i} - \frac{\sum_j X_{2j}X_{1j}}{\sum_j X_{2j}^2} X_{2i}\right)$$

$$= \sum_{i=1}^{n} e_{i,X_1|X_2} X_{1i} - \frac{\sum_j X_{2j}X_{1j}}{\sum_j X_{2j}^2} \sum_{i=1}^{n} e_{i,X_1|X_2} X_{2i}$$

But $\sum_{i=1}^{n} e_{i,X_1|X_2} X_{2i} = 0$. So we get that

$$\sum_{i=1}^{n} e_{i,X_1|X_2}^2 = \sum_{i=1}^{n} e_{i,X_1|X_2} X_{1i}$$

Thus we get that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} e_{i,Y|X_2} e_{i,X_1|X_2}}{\sum_{i=1}^{n} e_{i,X_1|X_2}^2}$$

# Summing up fitting with two regressors

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n e_{i,Y|X_2} e_{i,X_1|X_2}}{\sum_{i=1}^n e_{i,X_1|X_2}^2}$$

- That is, the regression estimate for $\beta_1$ is the regression through the origin estimate having regressed $X_2$ out of both the response and the predictor.
- (Similarly, the regression estimate for $\beta_2$ is the regression through the origin estimate having regressed $X_1$ out of both the response and the predictor.)
- More generally, multivariate regression estimates are exactly those having removed the linear relationship of the other variables from both the regressor and response.

# Example with two variables, simple linear regression

- $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i}$ where $X_{2i} = 1$ is an intercept term.
- Then $\frac{\sum_j X_{2j} X_{1j}}{\sum_j X_{2j}^2} X_{2i} = \frac{\sum_j X_{1j}}{n} = \bar{X}_1$.
- $e_{i,X_1|X_2} = X_{1i} - \bar{X}_1$.
- Simiarly $e_{i,Y|X_2} = Y_i - \bar{Y}$.
- Thus

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n e_{i,Y|X_2} e_{i,X_1|X_2}}{\sum_{i=1}^n e_{i,X_1|X_2}^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \mathrm{Cor}(X, Y) \frac{\mathrm{Sd}(Y)}{\mathrm{Sd}(X)}$$

# The general case

- The equations

$$\sum_{i=1}^{n} (Y_i - X_{1i}\hat\beta_1 - \ldots - X_{ip}\hat\beta_p)X_k = 0$$

  for $k = 1, \ldots, p$ yields $p$ equations with $p$ unknowns.

- Solving them yields the least squares estimates. (With obtaining a good, fast, general solution requiring some knowledge of linear algebra.)
- The least squares estimate for the coefficient of a multivariate regression model is exactly regression through the origin with the linear relationships with the other regressors removed from both the regressor and outcome by taking residuals.
- In this sense, multivariate regression "adjusts" a coefficient for the linear impact of the other variables.

# Fitting LS equations

Just so I don't leave you hanging, let's show a way to get estimates. Recall the equations:

$$\sum_{i=1}^{n}(Y_i - X_{1i}\hat{\beta}_1 - \ldots - X_{ip}\hat{\beta}_p)X_k = 0$$

If I hold $\hat{\beta}_1, \ldots, \hat{\beta}_{p-1}$ fixed then we get that

$$\hat{\beta}_p = \frac{\sum_{i=1}^{n}(Y_i - X_{1i}\hat{\beta}_1 - \ldots - X_{i,p-1}\hat{\beta}_{p-1})X_{ip}}{\sum_{i=1}^{n} X_{ip}^2}$$

Plugging this back into the equations, we wind up with

$$\sum_{i=1}^{n}(e_{i,Y|X_p} - e_{i,X_1|X_p}\hat{\beta}_1 - \ldots - e_{i,X_{p-1}|X_p}\hat{\beta}_{p-1})X_k = 0$$

# We can tidy it up a bit more, though

Note that

$$X_k = e_{i,X_k|X_p} + \frac{\sum_{i=1}^n X_{ik} X_{ip}}{\sum_{i=1}^n X_{ip}{}^2} X_p$$

and $\sum_{i=1}^n e_{i,X_j|X_p} X_{ip} = 0$. Thus

$$\sum_{i=1}^n (e_{i,Y|X_p} - e_{i,X_1|X_p} \hat{\beta}_1 - \ldots - e_{i,X_{p-1}|X_p} \hat{\beta}_{p-1}) X_k = 0$$

is equal to

$$\sum_{i=1}^n (e_{i,Y|X_p} - e_{i,X_1|X_p} \hat{\beta}_1 - \ldots - e_{i,X_{p-1}|X_p} \hat{\beta}_{p-1}) e_{i,X_k|X_p} = 0$$

# To sum up

- We've reduced $p$ LS equations and $p$ unknowns to $p - 1$ LS equations and $p - 1$ unknowns.
  - Every variable has been replaced by its residual with $X_p$.
  - This process can then be iterated until only Y and one variable remains.
- Think of it as follows. If we want an adjusted relationship between y and x, keep taking residuals over confounders and do regression through the origin.
  - The order that you do the confounders doesn't matter.
  - (It can't because our choice of doing $p$ first was arbitrary.)
- This isn't a terribly efficient way to get estimates. But, it's nice conceputally, as it shows how regression estimates are adjusted for the linear relationship with other variables.

# Demonstration that it works using an example

## Linear model with two variables and an intercept

```
n <- 100; x <- rnorm(n); x2 <- rnorm(n); x3 <- rnorm(n)
y <- x + x2 + x3 + rnorm(n, sd = .1)
e <- function(a, b) a -  sum( a * b ) / sum( b ^ 2) * b
ey <- e(e(y, x2), e(x3, x2))
ex <- e(e(x, x2), e(x3, x2))
sum(ey * ex) / sum(ex ^ 2)
```

```
[1] 1.004
```

```
coef(lm(y ~ x + x2 + x3 - 1)) #the -1 removes the intercept term
```

```
     x     x2     x3 
1.0040 0.9899 1.0078 
```

# Showing that order doesn't matter

```
ey <- e(e(y, x3), e(x2, x3))
ex <- e(e(x, x3), e(x2, x3))
sum(ey * ex) / sum(ex ^ 2)
```

```
[1] 1.004
```

```
coef(lm(y ~ x + x2 + x3 - 1)) #the -1 removes the intercept term
```

```
     x     x2     x3 
1.0040 0.9899 1.0078 
```

# Residuals again

```
ey <- resid(lm(y ~ x2 + x3 - 1))
ex <- resid(lm(x ~ x2 + x3 - 1))
sum(ey * ex) / sum(ex ^ 2)
```

```
[1] 1.004
```

```
coef(lm(y ~ x + x2 + x3 - 1)) #the -1 removes the intercept term
```

```
     x     x2     x3 
1.0040 0.9899 1.0078 
```

# Interpretation of the coeficient

$$E[Y|X_1 = x_1, \dots, X_p = x_p] = \sum_{k=1}^{p} x_k \beta_k$$

So that

$$E[Y|X_1 = x_1 + 1, \dots, X_p = x_p] - E[Y|X_1 = x_1, \dots, X_p = x_p]$$

$$= (x_1 + 1)\beta_1 + \sum_{k=2}^{p} x_k + \sum_{k=1}^{p} x_k \beta_k = \beta_1$$

So that the interpretation of a multivariate regression coefficient is the expected change in the response per unit change in the regressor, holding all of the other regressors fixed.

In the next lecture, we'll do examples and go over context-specific interpretations.

# Fitted values, residuals and residual variation

All of our SLR quantities can be extended to linear models

- Model $Y_i = \sum_{k=1}^p X_{ik} \beta_k + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$
- Fitted responses $\hat Y_i = \sum_{k=1}^p X_{ik} \hat \beta_k$
- Residuals $e_i = Y_i - \hat Y_i$
- Variance estimate $\hat \sigma^2 = \frac{1}{n-p} \sum_{i=1}^n e_i^2$
- To get predicted responses at new values, $x_1, \ldots, x_p$, simply plug them into the linear model $\sum_{k=1}^p x_k \hat \beta_k$
- Coefficients have standard errors, $\hat \sigma_{\hat \beta_k}$, and $\frac{\hat \beta_k - \beta_k}{\hat \sigma_{\hat \beta_k}}$ follows a $T$ distribution with $n-p$ degrees of freedom.
- Predicted responses have standard errors and we can calculate predicted and expected response intervals.

# Linear models

- Linear models are the single most important applied statistical and machine learning techniqe, *by far*.
- Some amazing things that you can accomplish with linear models
  - Decompose a signal into its harmonics.
  - Flexibly fit complicated functions.
  - Fit factor variables as predictors.
  - Uncover complex multivariate relationships with the response.
  - Build accurate prediction models.