Multiple Linear Regression (MLR) Introduction: Motivation, Assumptions, Example

BIOS 6611

CU Anschutz

Week 10

Multiple Linear Regression (MLR) Introduction

Multiple Linear Regression (MLR) Introduction

Multiple linear regression (MLR) can be used to summarize the relationship between a continuous response variable, Y, and *multiple* explanatory predictor variables, $X_1, X_2, \ldots, X_k$, using linear relationships.

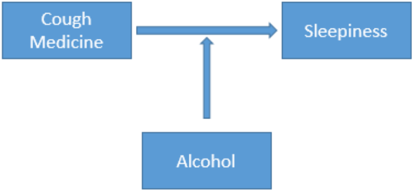
Reasons to include multiple predictors:

- To address the scientific question
- To adjust for confounding
- To gain precision

Addressing the scientific question

Address the scientific question. The scientific question may dictate inclusion of predictors:

- *Predictor(s) of interest:* The scientific factor(s) under investigation may need to be modeled by multiple predictors (e.g., dummy variables, polynomials). Or, there may be more than one predictor of interest.
- *Effect modifiers:* The magnitude of the effect of the predictor of interest may vary depending on levels of an effect modifier.

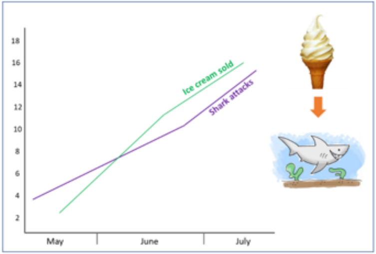


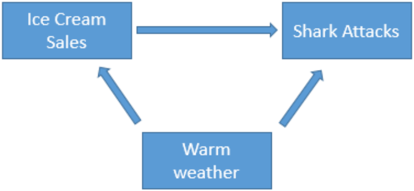
- *Confounders:*...

Addressing the scientific question (cont.)

Address the scientific question. The scientific question may dictate inclusion of predictors:

- ...
- *Confounders:* Confounders are a variable that effect both the predictor of interest and the outcome variable.





Precision

Precision. Adjusting for an additional covariate(s) can change the standard error of the slope estimate corresponding to the predictor of interest.

- The standard error decreases when smaller within group variance.
- The standard error increases when there is a correlation between predictor of interest and other covariates in the model.

The MLR Model

The MLR Model

As in SLR, assume $Y_i|X_1, \ldots, X_k \sim N(\mu_{Y|\mathbf{X}}, \sigma^2_{Y|\mathbf{X}})$, but now we assume underlying center changes linearly with several other factors:

$$\mu_{Y|\mathbf{X}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Or, equivalently,

$$Y_i|\mathbf{X}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i1} + \ldots + \beta_k X_{ki} + \epsilon_i$$

where ϵ_i represents the random error and $\epsilon_i \sim N(0, \sigma^2_{Y|\mathbf{X}})$.

Interpretation of coefficients:

$$\mu_{Y|\mathbf{X}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Interpretation of coefficients:

- Intercept: β_0 is the expected value of Y **when all other predictors**, $X_1, \ldots, X_k$, are equal to 0.
- Slope: β_j is the expected change in Y associated with a one-unit change in X_j, assuming all other predictors are held constant.

For example, say we are interested in the change in Y for a one unit increase in X_1, assuming all other predictors are held constant. Then

$$\begin{aligned}
\mu_{Y|X_1=x+1} - \mu_{Y|X_1=x} &= (\beta_0 + \beta_1(x+1) + \beta_2 c_2 + \ldots + \beta_k c_k) \\
&\quad -(\beta_0 + \beta_1 x + \beta_2 c_2 + \ldots + \beta_k c_k) \\
&= \beta_1
\end{aligned}$$

Least Squares Estimation (LSE) for Multiple Linear Regression

- As in SLR, use data $(Y_i, X_{1i}, \ldots, X_{ki}; i = 1, \ldots, n)$ to estimate the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$.
- The model that represents the fitted values is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$$

- Differences between fitted values and observed values are called *residuals*

$$\hat{e}_i \quad = \quad Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{ki}$$

- Coefficient estimates are chosen to minimize the residual sum of squares (also called error sum of squares).

$$RSS = SSE = \sum_{i=1}^{n} \hat{e}_i^2$$

MLR Assumptions

MLR Assumptions

The assumptions for multiple linear regression are the same as for simple linear regression:

- **Existence:** For each combination of values of the predictors $(X_1, X_2, \ldots, X_k)$, Y is a random variable with a certain probability distribution having finite mean and variance.
- **Linearity:** The mean value of Y for each specific combination of values of $X_1, X_2, \ldots, X_k$ is a linear function of $X_1, X_2, \ldots, X_k$
- **Independence:** Y_i are statistically independent
- **Homoscedasticity:** The variance of Y, $\sigma^2_{Y|\mathbf{X}}$ is the same for any fixed combination of $X_1, X_2, \ldots, X_k$.
- **Normality:** For any fixed combination of $X_1, X_2, \ldots, X_k$, the residuals are normally distributed. (This assumption is primarily used for hypothesis testing and CIs.)

MLR Example

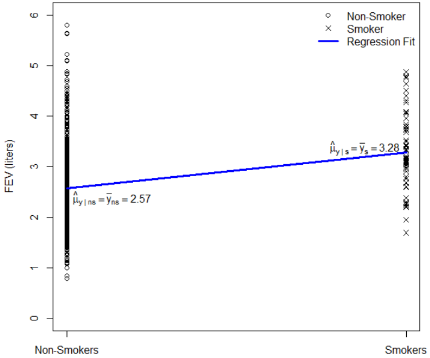
MLR Example: Starting with SLR

In the Rosner FEV data set, let's say we are interested in the effect of smoking on FEV. We could naively fit a SLR model:

```
fev <- read.csv('FEV_rosner.csv')
slr <- glm( fev ~ smoke, data=fev )
summary(slr)
```

```
##
## Call:
## glm(formula = fev ~ smoke, data = fev)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7751  -0.6339  -0.1021   0.4804   3.2269
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.56614    0.03466  74.037  < 2e-16 ***
## smokesmoker   0.71072    0.10994   6.464 1.99e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Starting with SLR



Least squares regression fitted model: $\hat{FEV} = 2.57 + 0.71 \times \text{smoker}$

Interpretation: There is an expected FEV increase of 0.71 for smokers compared to non-smokers. Therefore, smokers have better lung function than non-smokers (p<0.0001). *(Does this make clinical sense?)*

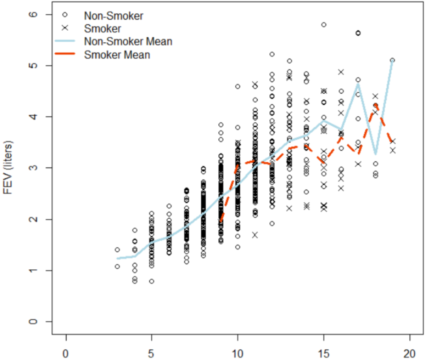
Example: Control for age

We realize that smokers tend to be older than non-smokers, and that older children tend to have higher FEV than young children. (Thus, age has potential to be a *confounder*.) We decide to control for age in our analysis.

New question: For a group of children at a given age, do smokers have lower FEV compared to non-smokers?

- Option 1: Perform a stratified analysis and compare smokers to non-smokers within age strata. Note: requires that we break up age into strata, losing some information.
- Option 2: With MLR, get a single estimate of the average effect of smoking on FEV, adjusting for differences due to age.

Option 1: Stratified analysis



Stratified Analysis:

Age Group	Smokers	Non-smokers	FEV Smokers	FEV Non-smokers	Smoke-NonSmoke difference	T statistic	p-value
3-8	0	215	-	1.86 (0.42)	-	-	-
9-10	6	169	2.88 (0.60)	2.54 (0.51)	0.34	-1.57	.118
11-12	16	131	3.11 (0.67)	3.11 (0.64)	0.00	0.01	.993
13-14	20	48	3.40 (0.83)	3.57 (0.68)	-0.17	0.87	.389
15-16	17	15	3.30 (0.82)	3.85 (0.81)	-0.55	1.92	.065
17-19	6	11	3.64 (0.50)	4.19 (1.03)	-0.55	1.21	.244

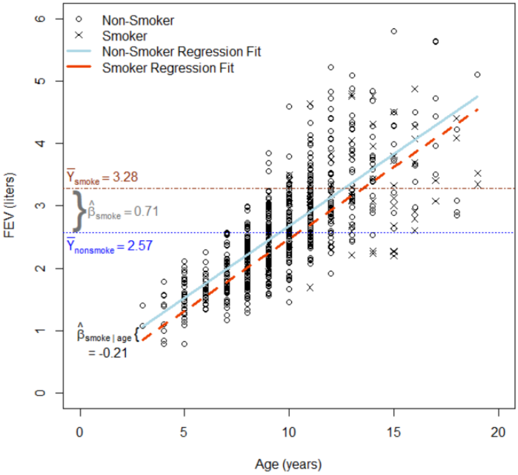
Option 2: Multiple Linear Regression

```r
mlr <- glm( fev ~ smoke + age, data=fev )
summary(mlr)
```

```
##
## Call:
## glm(formula = fev ~ smoke + age, data = fev)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6653  -0.3564  -0.0508   0.3494   2.0894
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.367373   0.081436   4.511 7.65e-06 ***
## smokesmoker -0.208995   0.080745  -2.588  0.00986 **
## age          0.230605   0.008184  28.176  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.3192958)
##
```

MLR Example

Least squares regression line: $\hat{FEV} = 0.37 + 0.23 \times \text{age} - 0.21 \times \text{smoker}$.



Interpretation: On average, FEV is 0.21 liters lower in smokers compared to non-smokers when age is held constant. Thus, we conclude smokers have worse lung function compared to non-smokers of the same age (p=0.0099).

Preview: Interaction terms

Note that the effect of age on FEV is assumed to be the same for smokers and non-smokers.

Likewise, the effect of smoking on FEV is assumed to be the same for every age.

We can relax these restrictions by including *interaction terms*, which we will learn about in a future lecture.

Adjusted R^2

R^2 refresher

Recall the **coefficient of determination**, or "R-squared":

$$R^2 = \frac{SS_{Total} - SS_{Error}}{SS_{Total}} = \frac{SS_{Model}}{SS_{Total}}$$

which gives the proportion of variance of Y that can be explained by $X_1, \ldots, X_k$.

When more explanatory variables are added to the model, R^2 automatically increases. Therefore, we cannot use R^2 to compare models with differing numbers of predictors.

Adjusted R^2

The **adjusted** R^2 accounts for this phenomenon. It is a modification of R^2 that adjusts for the number of explanatory terms in the model (k) relative to the number of data points (n).

$$
\begin{aligned}
R^2_{adj} &= 1 - (1 - R^2)\frac{n-1}{n-k-1} \\
&= 1 - \frac{SS_{Error}/(n-k-1)}{SS_{Total}/(n-1)}
\end{aligned}
$$

R^2_{adj} can be negative, and will always be less than or equal to R^2. It will only increase when the increase in R^2 is more than one would expect to see by chance. R^2_{adj} is more appropriate when evaluating model fit and when comparing alternative models with differing number of predictors.

MLR Introduction Summary

In summary:

- MLR allows us to investigate multiple predictors of interest, to control for confounders and effect modifiers, and to gain precision.
- The MLR model has the same assumptions of the simple linear regression: existence, linearity, independence, homoscedasticity
- Coefficient estimates are obtained my minimizing the residual sum of squares
- We looked at an example where controlling for age, a confounder, changed our conclusion about effect of the predictor of interest, smoking status, on the FEV.
- The adjusted R^2 gives us information about the amount of variability explained by the model, while accounting for the addition of more explanatory variables.