# Multiple Linear Regression

Sumanta Basu

# Multiple Linear Regression

# Why MLR?

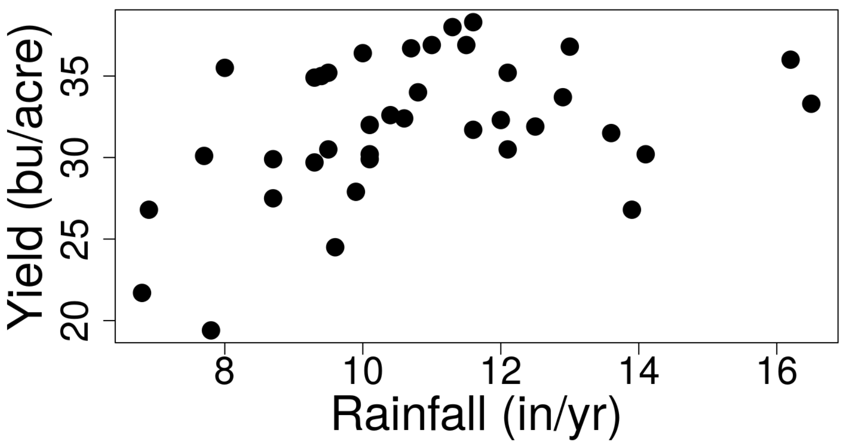
- the SLR model

$$Y = \mu(x) + \epsilon$$

  is useful when response only depends on one predictor and that association is linear.

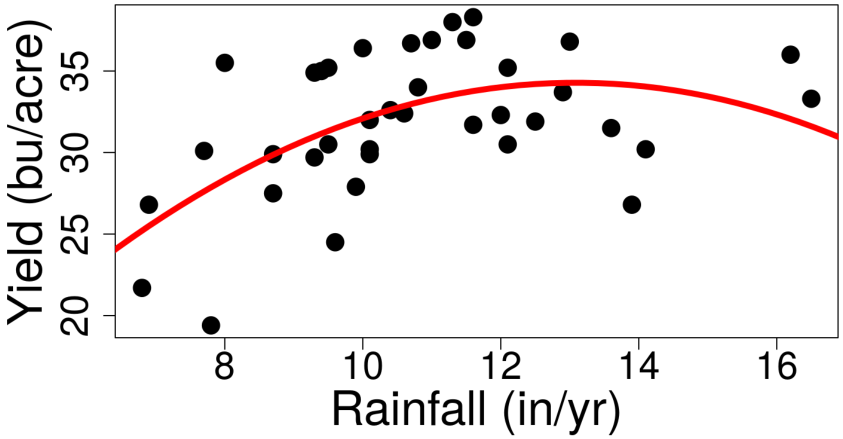
- in many situations this doesn't hold
- MLR is an incredibly flexible framework
- we'll start with some examples

# Example: Corn yield




- is trend linear?

# Example: Corn yield




- MLR lets us fit nonlinear associations and test them against a “null” linear association

# Example: Corn yield

$$\mu(\texttt{rain}) = \beta_0 + \beta_1 \cdot \texttt{rain} + \beta_2 \cdot \texttt{rain}^2$$

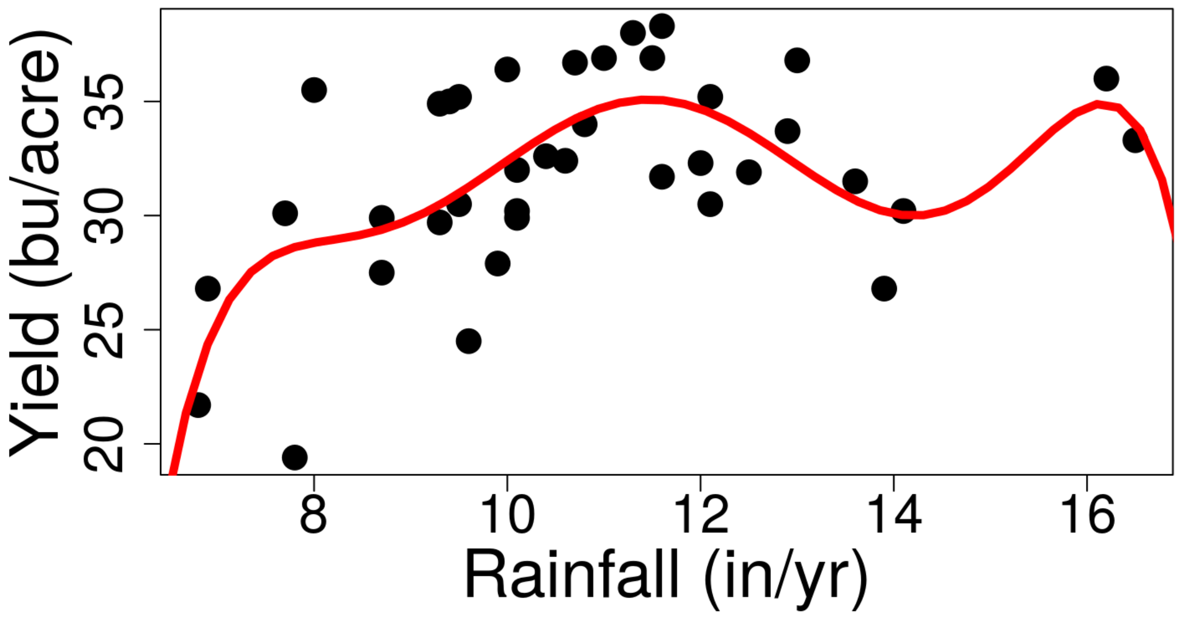
Can test $H_0 : \beta_2 = 0$ versus $H_A : \beta_2 \neq 0$.

# Example: Corn yield

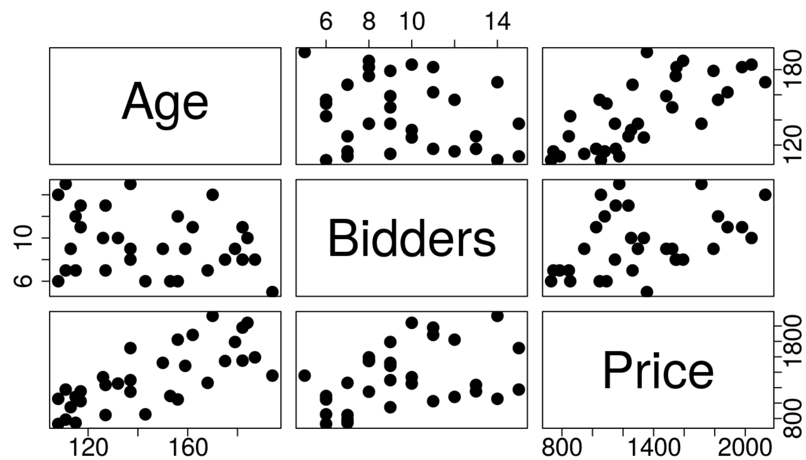
Could also do something much more complicated, e.g.,

$$\mu(\texttt{rain}) = \beta_0 + \beta_1 \cdot \texttt{rain} + \beta_2 \cdot \texttt{rain}^2$$
$$+ \beta_3 \cdot \texttt{rain}^3 + \beta_4 \cdot \texttt{rain}^4 + \beta_5 \cdot \texttt{rain}^5 + \beta_6 \cdot \texttt{rain}^6$$

but this usually ill-advised:



# Example: Auction for Grandfather Clocks




what is expected selling price given clock's age and number of bidders?

$$\mu(bidders, age) = ?$$

- both appear associated with Price
- but they are not too correlated with each other, so they may "get at" different aspects of Price

# Example: Auction for Clocks

Natural extension of SLR to situations when we have multiple predictors:

Expected selling price based on number of bidders and age of clock

$$\mu(bidders, age) = \beta_0 + \beta_1 \cdot \texttt{bidders} + \beta_2 \cdot \texttt{age}$$

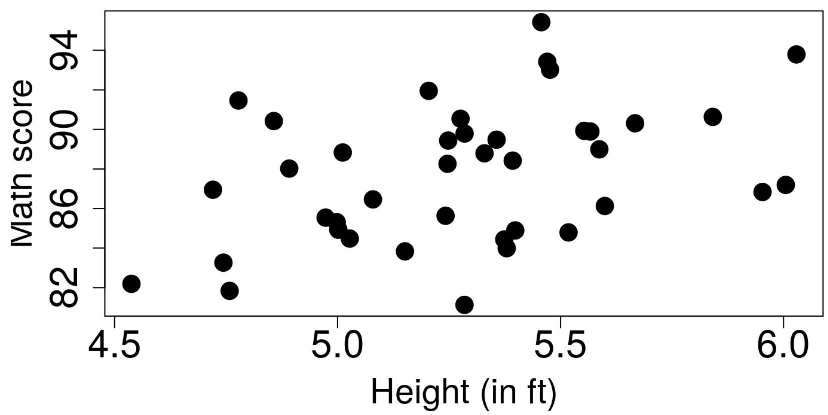
*Important to keep in mind the units*

- observe that

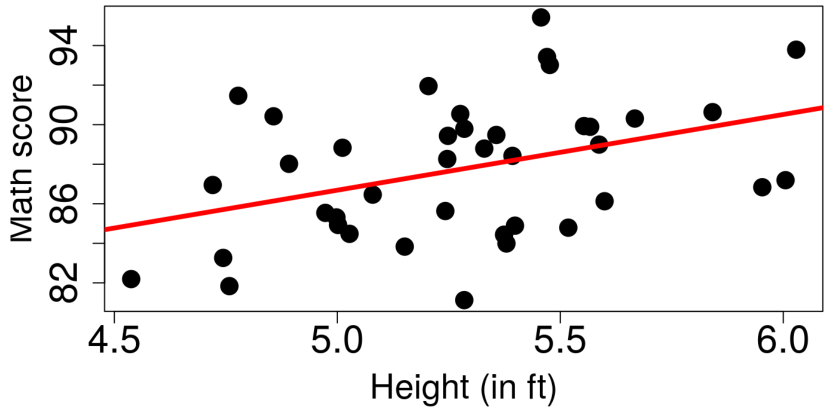
$$\mu(bidders + 1, age) - \mu(bidders, age) = \beta_1$$

- so $\beta_1$ = effect of increasing `bidders` by 1 while holding age of clock fixed
- that is, among subpopulations of clocks of the same age, this is effect of `bidders`.

# Example: Math ability vs. height

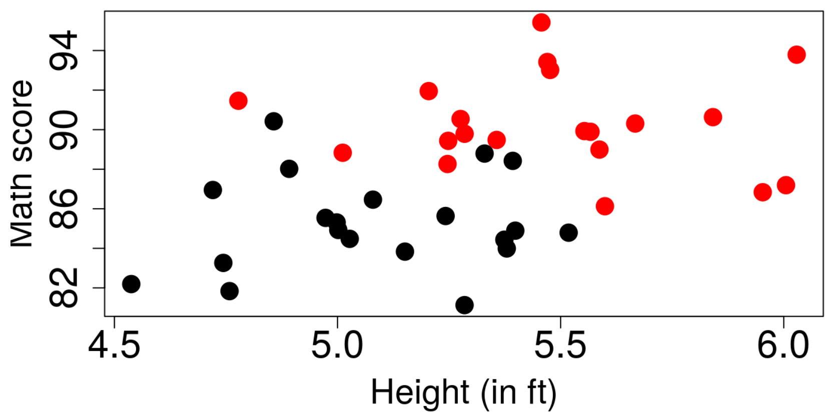


# Example: Math ability vs. height



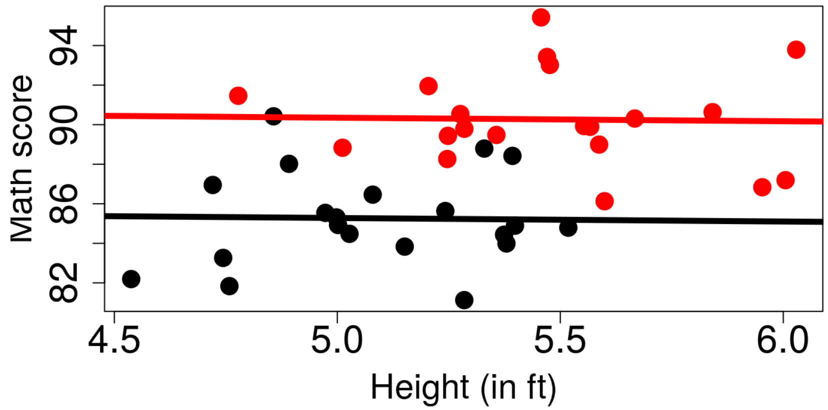

- $\hat{\beta} \approx 3.8$
- $H_0 : \beta = 0$ has p-value $0.01$
- an increase in height by 1 foot is associated with an increase in expected math score of 3.8 points.
- **Correlation does not imply causation**
- to make causal statements requires a randomized, controlled experiment

# Example: Math ability vs. height




- suppose researcher also recorded **age** of student
- red is “older”; black is “younger”
- fitting separate SLR’s for each age group is not great - can’t pool information across both groups; difficult to make comparisons between groups; gets worse if there were 10 groups!

# Example: Math ability vs. height




- MLR allows us to fit a **single model** with a separate intercept for each age group and a shared slope for height

$$\mu(h, a) = \begin{cases} \beta_0 + \beta \cdot h & \text{if } a = old \\ \beta_0 + \beta_{young} + \beta \cdot h & \text{if } a = youn \end{cases}$$

- in this parameterization, $\beta_{young}$ represents the difference between intercepts

- can test $\beta_{young} = 0$ to see if lines have same intercept

# In R…

```
fit = lm(math ~ height + age)
```

- show result of `summary(fit)` in R.
- with age in the model, we fail to reject $H_0 : \beta = 0$
- no evidence of an association between math ability and height "within subpopulations of fixed age"
- "conditional on age" or "controlling for age"

# Multiple Linear Regression Model

Suppose we measure $n$ observations of a response variable and $p$ predictors. That is, for $i = 1, \dots, n$, we observe

$$(y_i, x_{i1}, \dots, x_{ip}).$$

> MLR models this data as a *realization* of
>
> $$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$
>
> where $\epsilon_i \sim N(0, \sigma)$ are independent.

- if $p = 1$, this is SLR.

# Previous examples as special cases

- **Grandfather clocks:** For $i$ th clock,
  - $y_i$ = Price
  - $x_{i1}$ = Bidders; $x_{i2}$ = Age
- **Corn yield:** For $i$ th year,
  - $y_i$ = Yield
  - $x_{i1}$ = Rainfall; $x_{i2}$ = Rainfall$^2$
- **Math ability:** For $i$ th student,
  - $y_i$ = Math score
  - $x_{i1}$ = *dummy variable* indicating whether Age = "young"
  - $x_{i2}$ = Height

# Dummy variable "trick"

Age is a categorical variable. It is *coded* as a 0 or 1:

$$x_{i1} = \begin{cases} 1 & \text{if } i \text{ th student's Age} = \text{"young"} \\ 0 & \text{otherwise} \end{cases}$$

Notice

$$\mu(x_{i1}, x_{i2}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

becomes

$$\begin{cases} \beta_0 + \beta_1 + \beta_2 x_{i2} & \text{if } i \text{ th student's Age} = \text{"young"} \\ \beta_0 + \beta_2 x_{i2} & \text{otherwise} \end{cases}$$

- thus these are both straight lines, but the intercept differs by $\beta_1$ between the old and young subpopulations

# Interpreting MLR coefficients

$$\mu(x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- note that $\beta_1 = \mu(x_1 + 1, x_2, \dots, x_p) - \mu(x_1, x_2, \dots, x_p)$
- $\beta_j$ is the mean increase of the response associated with a unit increase of the $j$ th predictor if one could hold all other predictors fixed
- **only if** data is from a randomized experiment, can one say that a unit increase in $j$ th predictor *causes* an increase in the average
- For observational studies, no causality implied.

# Interpretation in observational studies

$$\beta_1 = \mu(x_1 + 1, x_2, \ldots, x_p) - \mu(x_1, x_2, \ldots, x_p)$$

- We are imagining two subpopulations that are identical in all predictors except that $x_1$ differs by 1.
- In some cases, such a situation wouldn't make sense
  - for example, suppose $x_1$ and $x_2$ are necessarily correlated (extreme example: height in feet and cm). Imagining $x_2$ staying fixed while $x_1$ varying might not make sense
  - furthermore, in such a case the data doesn't let us see what would happen in this situation and thus our estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$ would have high variance

# Fitting MLR

# Least squares

Click here for image.

- instead of best fitting line, we now seek "**best fitting hyperplane**"
- still minimize sum-of-squared errors $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, but $\hat{y}_i$ now depends on $p + 1$ parameters: $\beta_0, \ldots, \beta_p$.
- image credit: "Introduction to Statistical Learning" (2003) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

# In R…

Grandfather clock data: $n = 32$ clocks; $p = 2$ predictors.

```
##   Age Bidders Price
## 1 127      13  1235
## 2 115      12  1080
## 3 127       7   845
## 4 150       9  1522
## 5 156       6  1047
## 6 182      11  1979
```

# MLR in R...

```
fit = lm(Price ~ Age + Bidders, data=auction)
fit
```

```
##
## Call:
## lm(formula = Price ~ Age + Bidders, data = auction)
##
## Coefficients:
## (Intercept)          Age      Bidders
##    -1336.72        12.74        85.82
```

- more information available with `summary(fit)`

# Inference

# Inferential goals

MLR has all the same goals as in SLR:

- Test whether **coefficient** $\beta_j = 0$ or get confidence interval
- Confidence interval for **mean response** given a particular set of predictor values
- **Prediction** interval for a new response with a given set of predictor values

But MLR has some additional ones:

- test whether a group of coefficients are zero (i.e., test whether a submodel is sufficient)

# Testing a single coefficient (t test)

Consider testing $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$

Under the null, we have

$$Y_i = \beta_0 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

and (it turns out that)

$$\hat{\beta}_1 \sim N(0, \sigma_{\hat{\beta}_1}).$$

Therefore,

> Under $H_0 : \beta_1 = 0$, we have $\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-p-1}$.

# In R...

```r
summary(fit)
```

```
##
## Call:
## lm(formula = Price ~ Age + Bidders, data = auction)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -207.2 -117.8   16.5  102.7  213.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1336.7221   173.3561  -7.711 1.67e-08 ***
## Age            12.7362     0.9024  14.114 1.60e-14 ***
## Bidders        85.8151     8.7058   9.857 9.14e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 133.1 on 29 degrees of freedom
## Multiple R-squared:  0.8927, Adjusted R-squared:  0.8853
## F-statistic: 120.7 on 2 and 29 DF,  p-value: 8.769e-15
```

# In R…

- each row corresponds to a parameter ($\beta_0$, $\beta_1$, and $\beta_2$)
- first column gives estimated coefficient, $\hat{\beta}_j$
- second column gives $\hat{\sigma}_{\hat{\beta}_j}$, estimated standard error of $\hat{\beta}_j$
- third column gives t statistic for testing $H_0 : \beta_j = 0$ vs. $H_A : \beta_j \neq 0$ while leaving all other variables in model
- fourth column gives p-value associated with this test

# A puzzle, predicting `sqrt(income)`

Predict `sqrt(income)` of a profession based on `education` and `prestige`.

```
library(car)
fit <- lm(sqrt(income) ~ education + prestige, data=Prestige)
```

# Isn’t education related to income??

```
summary(fit)
```

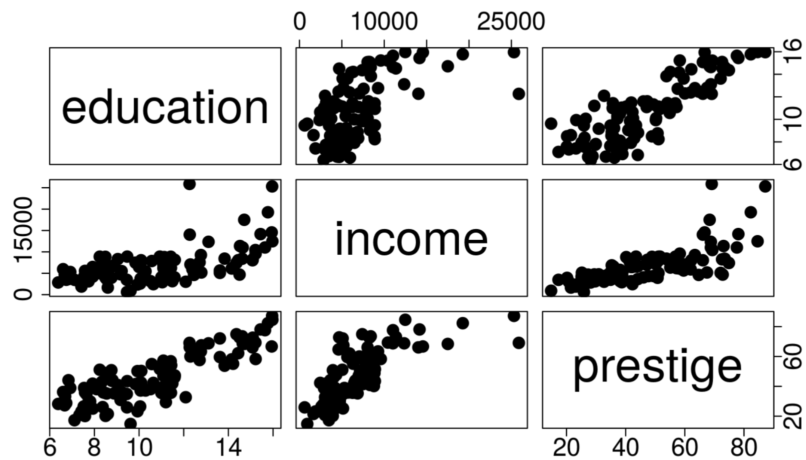
```
##
## Call:
## lm(formula = sqrt(income) ~ education + prestige, data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.595  -9.645   2.396   8.695  56.832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.7713     6.3477   6.108 1.99e-08 ***
## education    -1.5683     1.0449  -1.501    0.137
## prestige      1.2228     0.1657   7.379 5.01e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.09 on 99 degrees of freedom
## Multiple R-squared:  0.5798, Adjusted R-squared:  0.5713
## F-statistic: 68.31 on 2 and 99 DF,  p-value: < 2.2e-16
```

# What happened?

```
fit2 <- lm(sqrt(income) ~ education, data=Prestige)
summary(fit2)
```

```
##
## Call:
## lm(formula = sqrt(income) ~ education, data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.106 -11.215   0.199  11.449  74.082
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.6483     7.5483   3.398 0.000976 ***
## education     4.9869     0.6815   7.317 6.46e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.69 on 100 degrees of freedom
## Multiple R-squared:  0.3487, Adjusted R-squared:  0.3422
## F-statistic: 53.55 on 1 and 100 DF,  p-value: 6.459e-11
```

# Remember what is being tested




- `fit`'s p-value for `education` compares the model with `education` and `prestige` to the one with just `prestige`
- if you know the profession's prestige, knowing education doesn't add anything useful
- `fit2`'s p-value compares model with `education` versus no predictors.

# Testing multiple variables at once

Sometimes we wish to test *simultaneously* whether a **group of predictors** should be excluded from a model.

E.g., should all environmental variables be together excluded? (leaving only genetic ones)

# Testing a submodel

Consider $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

versus $H_A$ : not all of $\beta_1, \ldots, \beta_k$ are zero.

- In the absence of evidence against $H_0$, we would favor it in that it is a simpler model.

> Increasing the number of predictors will always reduce the residual sum-of-squares. The F-test tests whether it drops significantly more than we would expect under the null.

# In R...

```
# (note: n = 100 here)
fit = lm(y ~ x1 + x2 + x3 + x4)
fit.sub = lm(y ~ x1 + x4)
anova(fit.sub, fit)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x4
## Model 2: y ~ x1 + x2 + x3 + x4
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     97 3.7291
## 2     95 3.6450  2   0.08413 1.0963 0.3383
```

- There is insufficient evidence that we should favor the more complicated model that includes $x_2$ and $x_3$. We therefore stick with the simpler 2-predictor model.

# ANOVA as MLR

# Does mean age depend on favorite color?

color is a categorical variable with $K > 2$ levels.

age is numerical

ANOVA answers this question.

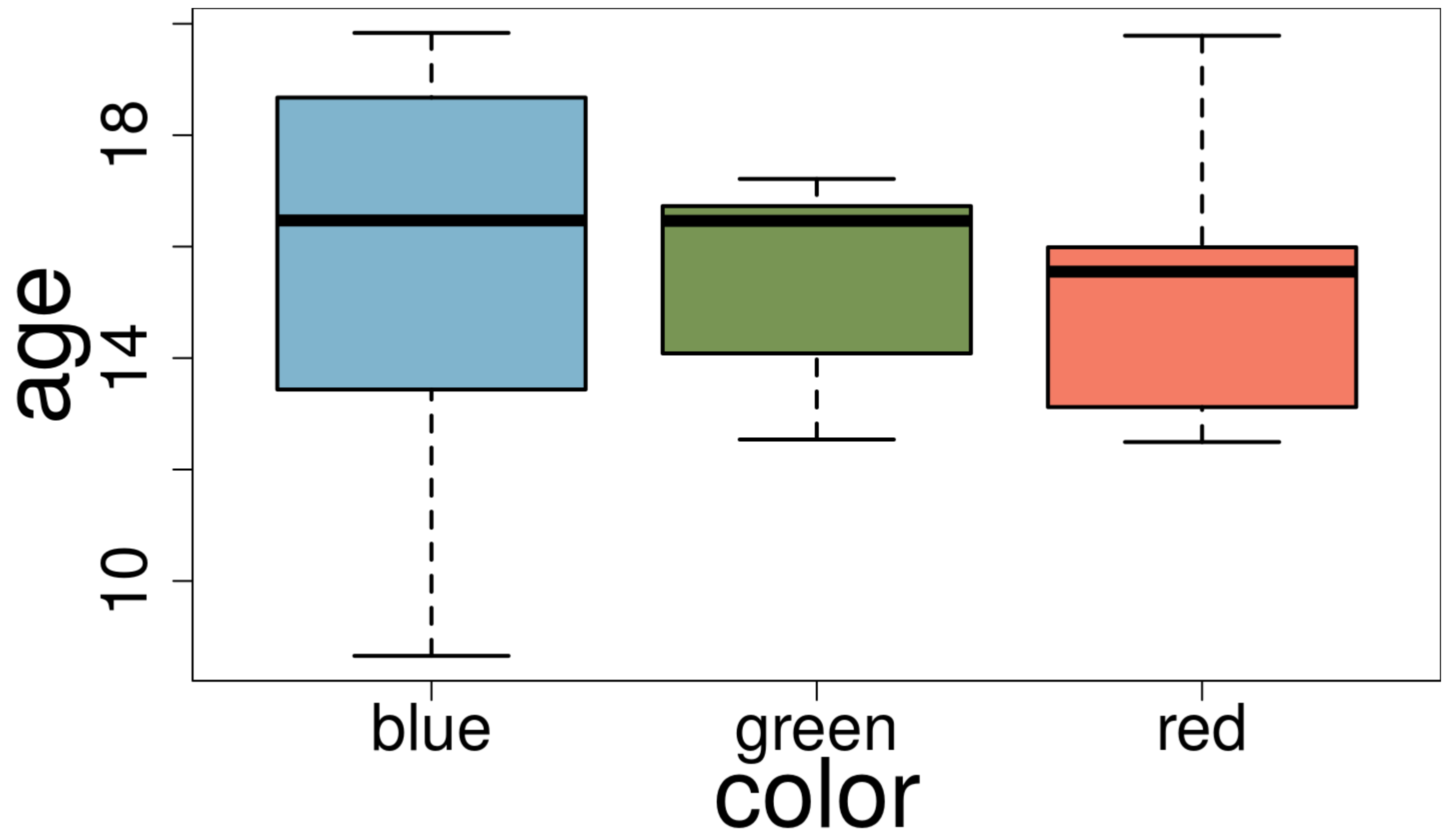
$H_0 : \mu_{red} = \mu_{blue} = \mu_{green}$

Recall R command

```
fit = lm(age ~ color)
anova(fit)
```

- why are we using `lm`?

# Fictitious data



# What R is doing in lm

Internally, R uses dummy coding to represent categorical variable with $K$ levels as $K-1$ separate binary predictors:

$$x_{i1} = \begin{cases} 1 & \text{person } i\text{'s color} = \text{green} \\ 0 & \text{person } i\text{'s color} \neq \text{green} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{person } i\text{'s color} = \text{red} \\ 0 & \text{person } i\text{'s color} \neq \text{red} \end{cases}$$

MLR:

$$E[y_i | color_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

$$x_{i1} = \begin{cases} 1 & \text{person } i\text{'s color} = \text{green} \\ 0 & \text{person } i\text{'s color} \neq \text{green} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{person } i\text{'s color} = \text{red} \\ 0 & \text{person } i\text{'s color} \neq \text{red} \end{cases}$$

$$E[y_i \mid color_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

Three cases to consider:

$$\begin{aligned} E[y_i \mid color_i = blue] &= \beta_0 \\ E[y_i \mid color_i = green] &= \beta_0 + \beta_1 \\ E[y_i \mid color_i = red] &= \beta_0 + \beta_2 \end{aligned}$$

- each group gets its own mean… this is exactly ANOVA. Just parameterized in terms of $\beta$'s instead of $\mu$'s!

# A look inside R

The factor `color`

```
## [1] red   green green blue  red
## Levels: blue green red
```

is coded as two dummy predictors

```
##      colorgreen colorred
## [1,]          0        1
## [2,]          1        0
## [3,]          1        0
## [4,]          0        0
## [5,]          0        1
```

# Which explains lm output

```
fit = lm(age ~ color)
fit
```

```
##
## Call:
## lm(formula = age ~ color)
##
## Coefficients:
## (Intercept)   colorgreen     colorred
##   15.414369    -0.008962    -0.026559
```

- Recall that $\mu_{blue} = \beta_0 \qquad \mu_{green} = \beta_0 + \beta_1 \qquad \mu_{red} = \beta_0 + \beta_2.$
- How would you test $H_0 : \mu_{blue} = \mu_{green} = \mu_{red}$?

# Checking assumptions

# Assumptions of MLR

> MLR models this data as a *realization* of
>
> $$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$
>
> where $\epsilon_i \sim N(0, \sigma)$ are independent.

Need:

- correctly specified dependence on the predictors
  - linear?
  - missing any?
- errors independent? same variance? normal?

# Relevant side note

> “All models are wrong, but some are useful.” –George Box (1979)

# Checking assumptions

- **independence**: depends on how data were collected
- **equal variance** (homoscedasticity): plot residuals versus predicted values. Point spread should look similar across range of predicted values
- **outliers** and influential points (Cook's distance)
- **normality**: QQ-plot of residuals

# Multicollinearity

- ideal situation: predictors are statistically uncorrelated
- **multicollinearity** means one or more of the predictors can be well-predicted by a linear function of the other predictors
- MLR works best when there is little multicollinearity
- can lead to numerical instability in estimates
- can lead to very large standard errors of coefficient estimates
- creates challenges in properly interpreting relative importance of predictors in model

# Simple tools for detecting multicollinearity

- **Plotting** predictors using `pairs(dat)` in R can help
- entries close to $\pm 1$ in **correlation matrix** for the predictor variables, `cor(dat)` in R.
- checking for large **variance inflation factors**, `vif(fit)` in R, using package `car`

# Summing up MLR

# Multiple Linear Regression Model

Suppose we measure $n$ observations of a response variable and $p$ predictors. That is, for $i = 1, \ldots, n$, we observe

$$(y_i, x_{i1}, \ldots, x_{ip}).$$

> MLR models this data as a *realization* of
>
> $$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$
>
> where $\epsilon_i \sim N(0, \sigma)$ are independent.

- if $p = 1$, this is SLR.

# Previous examples as special cases

- **Grandfather clocks:** For $i$ th clock,
  - $y_i$ = Price
  - $x_{i1}$ = Bidders; $x_{i2}$ = Age
- **Corn yield:** For $i$ th year,
  - $y_i$ = Yield
  - $x_{i1}$ = Rainfall; $x_{i2}$ = Rainfall$^2$
- **Math ability:** For $i$ th student,
  - $y_i$ = Math score
  - $x_{i1}$ = *dummy variable* indicating whether Age = “young”
  - $x_{i2}$ = Height

# Dummy variable "trick"

Age is a categorical variable. It is *coded* as a 0 or 1:

$$x_{i1} = \begin{cases} 1 & \text{if } i \text{ th student's Age = "young"} \\ 0 & \text{otherwise} \end{cases}$$

Notice

$$\mu(x_{i1}, x_{i2}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

becomes

$$\begin{cases} \beta_0 + \beta_1 + \beta_2 x_{i2} & \text{if } i \text{ th student's Age = "young"} \\ \beta_0 + \beta_2 x_{i2} & \text{otherwise} \end{cases}$$

- thus these are both straight lines, but the intercept differs by $\beta_1$ between the old and young subpopulations

# Interpreting MLR coefficients

$$\mu(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- note that $\beta_1 = \mu(x_1 + 1, x_2, \ldots, x_p) - \mu(x_1, x_2, \ldots, x_p)$
- $\beta_j$ is the mean increase of the response associated with a unit increase of the $j$ th predictor if one could hold all other predictors fixed
- **only if** data is from a randomized experiment, can one say that a unit increase in $j$ th predictor *causes* an increase in the average
- For observational studies, no causality implied.

# Interpretation in observational studies

$$\beta_1 = \mu(x_1 + 1, x_2, \dots, x_p) - \mu(x_1, x_2, \dots, x_p)$$

- We are imagining two subpopulations that are identical in all predictors except that $x_1$ differs by 1.
- In some cases, such a situation wouldn't make sense
  - for example, suppose $x_1$ and $x_2$ are necessarily correlated (extreme example: height in feet and cm). Imagining $x_2$ staying fixed while $x_1$ varying might not make sense
  - furthermore, in such a case the data doesn't let us see what would happen in this situation and thus our estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$ would have high variance