Confounding in Regression Analysis

CRP 241 Tutorial

Duke University Clinical Research Training Program

2025-11-07

Table of contents

Introduction

Learning Objectives

After completing this tutorial, you will be able to:

- **Identify** confounding variables in research studies
- **Explain** how confounding distorts exposure-outcome relationships
- **Apply** multiple linear regression to adjust for confounders
- **Interpret** the difference between unadjusted and adjusted associations
- **Understand** what regression is doing "behind the scenes" when it adjusts

0.1 What is Confounding?

Confounding occurs when a third variable is associated with **both** the exposure and the outcome, creating a spurious or distorted association between them. Think of a confounder as creating a "false connection" or hiding the "true connection" between two variables of interest.

! Key Concept: Confounding Criteria

For a variable to be a confounder, it must meet **both** criteria:

1. **Associated with the exposure** (e.g., different distribution across exposure groups)
2. **Associated with the outcome** (e.g., correlated with or predictive of the outcome)

0.2 Why Regression Matters

While t-tests can compare two groups, they **cannot** adjust for confounders. Regression allows us to:

- Account for multiple variables simultaneously
- Estimate the "true" exposure-outcome relationship after removing confounding
- Compare individuals who are similar on the confounder (like matching)

Clinical Analogy

Adjusting for confounders is like comparing apples to apples. Without adjustment, we might compare apples to oranges and draw incorrect conclusions.
For example, comparing elderly patients to young patients without adjusting for age differences could lead to biased conclusions about a treatment effect.

0.3 Quick Reference Guide

Table 1: Comparison of Statistical Methods

Model Type	**What It Shows**	**Can Adjust?**
t-test	Difference between 2 groups	No
Simple Linear Regression	Same as t-test for 2 groups	No
Multiple Linear Regression	Adjusted differences	Yes

1 Example 1: FEV1 and Genetic Variation

1.1 Study Background

A study investigated whether a genetic variant affects lung function (FEV1) in patients with COPD:

- **Sample:** 100 patients randomly selected from a clinical practice
- **Exposure:** Genotype (Wild Type vs. Mutant)
- **Outcome:** FEV1 (forced expiratory volume in 1 second, measured in liters)
- **Potential Confounder:** Sex at birth

> ℹ Clinical Question
>
> Does this genetic variant affect lung function? Or could any observed difference be explained by sex differences between genotype groups (since males and females have different baseline lung capacities)?

1.2 Data Dictionary

Table 2: FEV1 Study Variables

Variable	Description	Coding
`FEV1`	Forced expiratory volume in 1 second	Liters (continuous)
`GENO`	Patient genotype	0 = Wild Type, 1 = Mutant
`SEX`	Sex at birth	0 = Male, 1 = Female

1.3 Loading and Examining the Data

First, we'll load the dataset and examine its structure:

```r
# Load the FEV1 genotype dataset
load(url("https://www.duke.edu/~sgrambow/crp241data/fev1_geno.RData"))

# Examine the structure - shows variable types and first few values
str(fgdata)
```

```
'data.frame':	200 obs. of  3 variables:
 $ FEV1: num  2.66 3.62 4.8 1.61 3.19 ...
 $ GENO: int  0 0 0 0 1 1 0 0 0 1 ...
 $ SEX : int  0 1 0 1 0 1 0 1 0 1 ...
```

What to Look For

The `str()` output shows:

- 100 observations (patients)
- 3 variables (FEV1, GENO, SEX)
- Variable types (numeric or integer)
- Coding values (e.g., 0s and 1s for categorical variables)

Let's get summary statistics for all variables:

```
# Get summary statistics - shows min, quartiles, mean, median, max
summary(fgdata)
```

```
      FEV1            GENO           SEX     
 Min.   :1.283   Min.   :0.00   Min.   :0.0  
 1st Qu.:2.706   1st Qu.:0.00   1st Qu.:0.0  
 Median :3.387   Median :0.00   Median :0.5  
 Mean   :3.458   Mean   :0.44   Mean   :0.5  
 3rd Qu.:4.137   3rd Qu.:1.00   3rd Qu.:1.0  
 Max.   :6.280   Max.   :1.00   Max.   :1.0  
```

Interpreting the Summary

- **FEV1:** Range and distribution of lung function values
- **GENO & SEX:** The mean tells you the proportion coded as 1
  - Example: mean = 0.5 indicates 50% mutant (or 50% female)

1.4 Creating Labeled Variables

Numeric codes (0/1) work for analysis but are hard to interpret. Let's create labeled versions:

```
# Create labeled versions for easier interpretation in tables and plots
fgdata$fSEX <- factor(fgdata$SEX, labels = c('Male', 'Female'))
fgdata$fGENO <- factor(fgdata$GENO, labels = c('Wild Type', 'Mutant'))

# Verify the structure with new factor variables
str(fgdata)
```

```
'data.frame':	200 obs. of  5 variables:
 $ FEV1 : num  2.66 3.62 4.8 1.61 3.19 ...
 $ GENO : int  0 0 0 0 1 1 0 0 0 1 ...
 $ SEX  : int  0 1 0 1 0 1 0 1 0 1 ...
 $ fSEX : Factor w/ 2 levels "Male","Female": 1 2 1 2 1 2 1 2 1 2 ...
 $ fGENO: Factor w/ 2 levels "Wild Type","Mutant": 1 1 1 1 2 2 1 1 1 2 ...
```

Let's verify the labels match the numeric codes:

```r
# Cross-tabulation to verify Sex labels (rows) match numeric codes (columns)
table(fgdata$fSEX, fgdata$SEX)
```

```
         0   1
  Male   100   0
  Female   0 100
```

```r
# Cross-tabulation to verify Genotype labels match numeric codes
table(fgdata$fGENO, fgdata$GENO)
```

```
             0   1
  Wild Type 112   0
  Mutant      0  88
```

1.5 Checking for Confounding: Criterion 1

> ❗ Criterion 1: Is sex associated with genotype?
>
> If the distribution of males and females differs between genotype groups, then sex is associated with genotype (the exposure).

```r
# Cross-tabulation of sex by genotype (counts)
table(fgdata$fSEX, fgdata$fGENO)
```

```
        Wild Type Mutant
  Male         70     30
  Female       42     58
```

```r
# Convert to proportions within each genotype group
prop.table(table(fgdata$fSEX, fgdata$fGENO), 2)
```

```
         Wild Type    Mutant
  Male   0.6250000 0.3409091
  Female 0.3750000 0.6590909
```

Interpretation

Compare the proportion of females in Wild Type vs. Mutant groups:

- If proportions are similar (e.g., ~50% female in both), sex is **NOT** associated with genotype
- If proportions differ substantially, sex **IS** associated with genotype

Example: If we see 38% female in Wild Type vs. 66% female in Mutant, this suggests sex is associated with genotype **Criterion 1 met**

1.6 Checking for Confounding: Criterion 2

Criterion 2: Is sex associated with FEV1?

If FEV1 values differ between males and females, then sex is associated with the outcome.

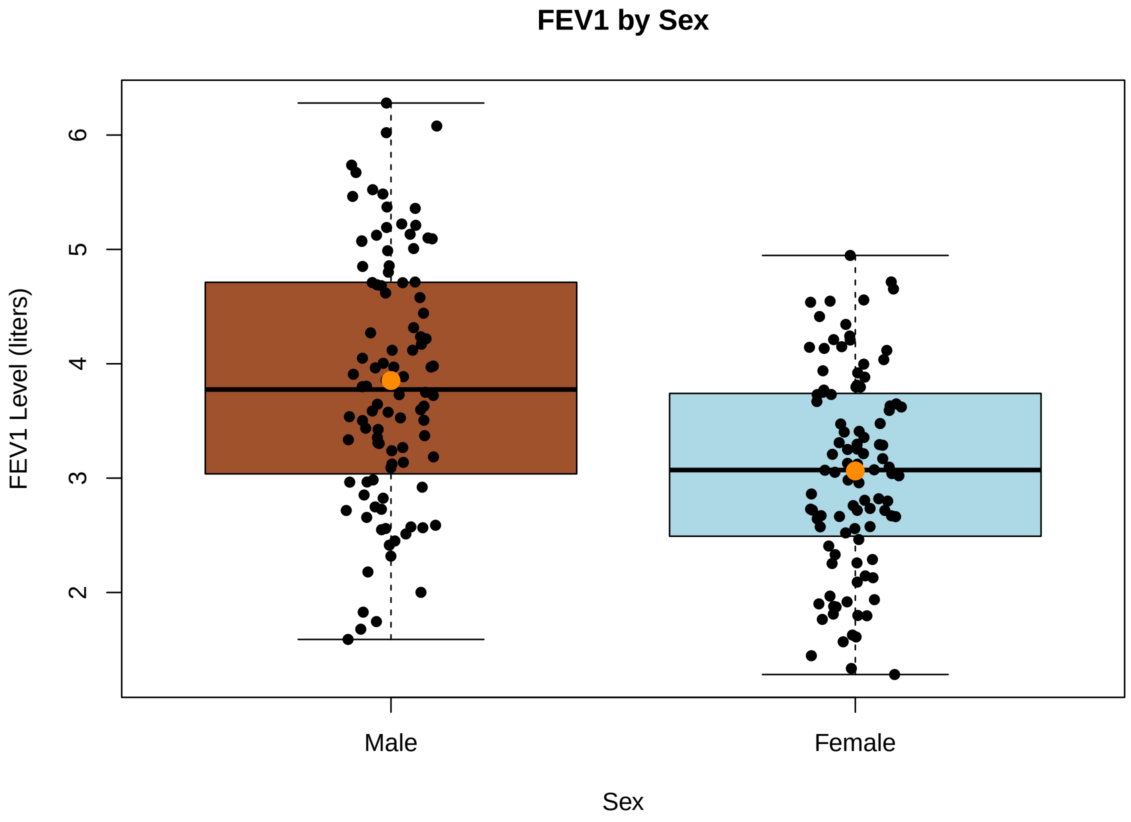
Let's visualize FEV1 distribution by sex:

```r
# Create enhanced boxplot with individual data points
boxplot(fgdata$FEV1 ~ fgdata$fSEX,
        main = 'FEV1 by Sex',
        ylab = 'FEV1 Level (liters)',
        xlab = 'Sex',
        col = c('sienna', 'lightblue'),
        range = 0)

# Overlay individual patient data points (jittered to avoid overlap)
stripchart(fgdata$FEV1 ~ fgdata$fSEX,
           method = "jitter",
           pch = 16,
           vertical = TRUE,
           add = TRUE)
```

```
# Calculate and overlay mean values
males <- subset(fgdata, fSEX == 'Male')
females <- subset(fgdata, fSEX == 'Female')
sex.means <- c(mean(males$FEV1), mean(females$FEV1))

points(sex.means, cex = 1.7, pch = 16, col = "dark orange")
```



ℹ Interpretation

Look at the boxplot:

- Are the boxes clearly separated or do they overlap substantially?
- Are the mean values (large orange dots) noticeably different?
- Clinical context: Males typically have larger lung capacity than females

If FEV1 differs between males and females, then sex **IS** associated with FEV1 **Criterion 2 met**

> **Conclusion:** If BOTH criteria are met, sex is likely a confounder!

1.7 Analysis 1: Two-Sample t-Test (Unadjusted)

Let's first compare FEV1 between genotypes using a traditional t-test. This gives us the **unadjusted (crude)** association, which does **NOT** account for sex differences.

```r
# Create genotype subsets for calculating means
wild <- subset(fgdata, fGENO == 'Wild Type')
mutant <- subset(fgdata, fGENO == 'Mutant')

mean.wild <- mean(wild$FEV1)
mean.mutant <- mean(mutant$FEV1)
```

```r
# Two-sample t-test with equal variances
t.test(fgdata$FEV1 ~ fgdata$GENO, var.equal = TRUE)
```

```

	Two Sample t-test

data:  fgdata$FEV1 by fgdata$GENO
t = 2.7962, df = 198, p-value = 0.005681
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
 0.1228584 0.7108077
sample estimates:
mean in group 0 mean in group 1
       3.641340        3.224507
```

```r
# Calculate difference in means manually
mean.wild - mean.mutant
```

```
[1] 0.416833
```

> **Interpreting the t-Test**
>
> Look for:
>
> - **Mean in each group:** Wild Type vs. Mutant
> - **Difference in means:** How much higher/lower is one group?
> - **95% confidence interval:** Uncertainty around the difference

- **p-value:** Is the difference statistically significant ($p < 0.05$)?

Note: This is the unadjusted difference - it may be confounded by sex!

1.8 Analysis 2: Simple Linear Regression (Unadjusted)

Now let's analyze the same comparison using simple linear regression. This demonstrates that **t-tests and simple linear regression are equivalent** when comparing two groups!

```r
# Fit simple linear regression: FEV1 ~ GENO
ufit <- lm(FEV1 ~ GENO, data = fgdata)

# Display regression results
summary(ufit)
```

```

Call:
lm(formula = FEV1 ~ GENO, data = fgdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.35865 -0.68975 -0.02921  0.59003  2.63869 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.64134    0.09888  36.824  < 2e-16 ***
GENO        -0.41683    0.14907  -2.796  0.00568 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.046 on 198 degrees of freedom
Multiple R-squared:  0.03799,	Adjusted R-squared:  0.03313 
F-statistic: 7.819 on 1 and 198 DF,  p-value: 0.005681
```

```r
# Calculate 95% confidence intervals for coefficients
confint(ufit)
```

```
                 2.5 %     97.5 %
(Intercept)  3.4463389  3.8363403
GENO        -0.7108077 -0.1228584
```

```r
# ANOVA table - tests overall model significance
summary(aov(ufit))
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)   
GENO          1   8.56   8.562   7.819 0.00568 **
Residuals   198 216.84   1.095                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# The regression coefficient has opposite sign from t-test
# because it compares Mutant (1) to Wild Type (0)
mean.mutant - mean.wild
```

```
[1] -0.416833
```

KEY TEACHING POINT: t-Test = Simple Linear Regression

The slope coefficient from regression equals the difference in means from the t-test (just with opposite sign due to comparison direction).
Regression advantage: We can extend regression to adjust for confounders, but we **cannot** do this with a t-test!

Interpreting Regression Output

Key values to examine:

- **Intercept:** Mean FEV1 for Wild Type (GENO = 0)
- **Slope (GENO):** Change in mean FEV1 for Mutant vs. Wild Type
  - Example: -0.417 means Mutants have 0.417 liters lower FEV1
- **p-value for GENO:** Is genotype significantly associated with FEV1?
- **R-squared:** Proportion of FEV1 variation explained by genotype
- **95% CI:** Uncertainty around the slope estimate

1.9 Analysis 3: Multiple Linear Regression (Adjusted)

Now for the key analysis: we'll **adjust for sex** by including it in the regression model alongside genotype. This controls for sex differences and reveals the "true" genotype effect.

❗ Critical Point

We **CANNOT** do this adjustment with a t-test. This is why multiple linear regression is so powerful for observational research!

```
# Fit multiple linear regression: FEV1 ~ GENO + SEX
afit <- lm(FEV1 ~ GENO + SEX, data = fgdata)

# Display adjusted regression results
summary(afit)
```

```

Call:
lm(formula = FEV1 ~ GENO + SEX, data = fgdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.23691 -0.69417 -0.01816  0.77985  2.36431 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.9157     0.1081  36.213  < 2e-16 ***
GENO         -0.2090     0.1467  -1.425    0.156    
SEX          -0.7317     0.1456  -5.025 1.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9877 on 197 degrees of freedom
Multiple R-squared:  0.1473,	Adjusted R-squared:  0.1386 
F-statistic: 17.02 on 2 and 197 DF,  p-value: 1.526e-07
```

```
# 95% confidence intervals for all coefficients
confint(afit)
```

```
                 2.5 %      97.5 %
(Intercept)  3.7024806  4.12896143
GENO        -0.4981893  0.08025266
SEX         -1.0188149 -0.44455282
```

```
# ANOVA table showing contribution of each variable
summary(aov(afit))
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)    
GENO          1   8.56   8.562   8.776  0.00343 ** 
SEX           1  24.64  24.639  25.254 1.12e-06 ***
Residuals   197 192.20   0.976                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting the Adjusted Model

Compare the **GENO coefficient** in the adjusted model to the unadjusted model:

- **Unadjusted (ufit):** Effect without considering sex
- **Adjusted (afit):** Effect after accounting for sex differences

Key questions:

1. Did the coefficient change? By how much?
2. Did the p-value change?
3. Which estimate is more trustworthy (adjusted or unadjusted)?

Interpretation:

- **Intercept:** Mean FEV1 for Wild Type males (GENO=0, SEX=0)
- **GENO coefficient:** Difference in FEV1 for Mutant vs. Wild Type, **holding sex constant** (this is the adjusted effect!)
- **SEX coefficient:** Difference in FEV1 for females vs. males, holding genotype constant

Clinical Interpretation

If the GENO coefficient changed substantially after adjustment:

- Sex was confounding the genotype-FEV1 relationship
- The adjusted coefficient is the “true” effect, removing distortion caused by sex differences between genotype groups
- We’re now comparing males to males and females to females (effectively)

1.10 Understanding Adjustment: Stratified Analysis

What is regression actually doing when it “adjusts”? Let’s look behind the scenes by manually calculating the genotype effect within each sex group.

1.10.1 Analysis Within Females Only

```r
# Create subsets by genotype within females
females.mutant <- subset(females, fGENO == 'Mutant')
females.wild <- subset(females, fGENO == 'Wild Type')

# Calculate means
mean.females.mutant <- mean(females.mutant$FEV1)
mean.females.wild <- mean(females.wild$FEV1)

# Display means
cat("Mean FEV1 for Mutant females:", mean.females.mutant, "\n")
```

```
Mean FEV1 for Mutant females: 3.006362
```

```r
cat("Mean FEV1 for Wild Type females:", mean.females.wild, "\n")
```

```
Mean FEV1 for Wild Type females: 3.140823
```

```r
# Calculate difference (genotype effect among females only)
mean.females.mutant - mean.females.wild
```

```
[1] -0.1344618
```

> **ℹ Note**
>
> This difference represents the genotype effect **among females only**, free from confounding by sex (since we're only looking at one sex).

1.10.2 Analysis Within Males Only

```r
# Create subsets by genotype within males
males.mutant <- subset(males, fGENO == 'Mutant')
males.wild <- subset(males, fGENO == 'Wild Type')

# Calculate means
mean.males.mutant <- mean(males.mutant$FEV1)
mean.males.wild <- mean(males.wild$FEV1)

# Display means
cat("Mean FEV1 for Mutant males:", mean.males.mutant, "\n")
```

```
Mean FEV1 for Mutant males: 3.646253
```

```r
cat("Mean FEV1 for Wild Type males:", mean.males.wild, "\n")
```

```
Mean FEV1 for Wild Type males: 3.941649
```

```r
# Calculate difference (genotype effect among males only)
mean.males.mutant - mean.males.wild
```

```
[1] -0.2953959
```

> Note
>
> This difference represents the genotype effect **among males only**, free from confounding by sex.

1.10.3 The Magic of Adjustment

```r
# Calculate the average of sex-specific differences
avg_effect <- ((mean.females.mutant - mean.females.wild) +
                (mean.males.mutant - mean.males.wild)) / 2

cat("Average of sex-specific effects:", avg_effect, "\n")
```

```
Average of sex-specific effects: -0.2149288
```

```r
cat("Compare to adjusted GENO coefficient from afit\n")
```

```
Compare to adjusted GENO coefficient from afit
```

❗ KEY INSIGHT: How Regression “Adjusts”

When regression adjusts for sex, it essentially:

1. Calculates the genotype effect within each sex group (as we just did)
2. Averages these sex-specific effects (weighted by sample size)
3. Reports this average as the adjusted coefficient

The result should be very similar to what we calculated manually! This is what “adjustment” means - comparing within strata and averaging the results.

1.10.4 Regression Within Strata

We can also fit separate regression models within each sex group:

```r
# Regression within females only
ffit <- lm(FEV1 ~ GENO, data = females)
summary(ffit)
```

```

Call:
lm(formula = FEV1 ~ GENO, data = females)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.85813 -0.63437  0.02404  0.72909  1.80653 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.1408     0.1354  23.202   <2e-16 ***
GENO         -0.1345     0.1777  -0.756    0.451    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8773 on 98 degrees of freedom
Multiple R-squared:  0.005805,	Adjusted R-squared:  -0.004339 
F-statistic: 0.5723 on 1 and 98 DF,  p-value: 0.4512
```

```r
confint(ffit)
```

```
                 2.5 %    97.5 %
```

```
(Intercept)  2.8721889 3.4094579
GENO        -0.4871961 0.2182726
```

> ℹ Note
>
> The slope for GENO should match our manual calculation: `mean.females.mutant - mean.females.wild`

```
# Regression within males only
mfit <- lm(FEV1 ~ GENO, data = males)
summary(mfit)
```

```

Call:
lm(formula = FEV1 ~ GENO, data = males)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.26284 -0.70681 -0.09617  0.91070  2.33838 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.9416     0.1303  30.249   <2e-16 ***
GENO         -0.2954     0.2379  -1.242    0.217    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.09 on 98 degrees of freedom
Multiple R-squared:  0.01549,	Adjusted R-squared:  0.005442 
F-statistic: 1.542 on 1 and 98 DF,  p-value: 0.2173
```

```
confint(mfit)
```

```
                 2.5 %    97.5 %
(Intercept)  3.6830593 4.2002394
GENO        -0.7675146 0.1767227
```

> ℹ Note
>
> The slope for GENO should match our manual calculation: `mean.males.mutant - mean.males.wild`

KEY TEACHING POINT

The sex-specific slopes from these stratified regressions (ffit and mfit) are averaged (conceptually) in the adjusted multiple regression model (afit).
This is how regression adjusts for confounders: it estimates the exposure-outcome relationship within levels of the confounder, then combines them into a single adjusted estimate.

---

2 Example 2: Lead Exposure and Neurological Function

2.1 Study Background

A study examined the effects of lead exposure on children's neurological development:

- **Sample:** 102 children living near a lead smelter in El Paso, Texas
- **Exposure:** Blood lead levels (Control: <40 g/ml; Exposed: 40 g/ml)
- **Outcome:** Finger-wrist tapping test (measure of fine motor coordination)
- **Potential Confounder:** Age in years

Clinical Question

Does lead exposure affect neurological function (tapping test score)? Or could any observed difference be explained by age differences between groups (since neurological development improves with age)?

2.2 Data Dictionary

Table 3: Lead Study Variables

Variable	Description	Coding
`maxfwt`	Finger-wrist tapping test score	Number of taps in 10 seconds (continuous); higher = better
`Group`	Exposure group	1 = Control, 2 = Exposed
`ageyrs`	Age of child	Years (decimal format, e.g., 8.5)

2.3 Loading and Cleaning the Data

```r
# Load the lead exposure dataset
load(url("https://www.duke.edu/~sgrambow/crp241data/lead.RData"))
```

⚠ Important Data Cleaning Step

In this dataset, missing values for `maxfwt` were coded as 99 (a common convention in older studies). We need to convert these to `NA` so R recognizes them as missing and handles them correctly.
Why this matters: If we leave 99 as a number, R will treat it as a real score (99 taps is impossibly high!), severely biasing our results.

```r
# Convert 99 to NA for missing tapping test scores
lead$maxfwt[lead$maxfwt == 99] <- NA

# Check how many values were converted
sum(is.na(lead$maxfwt))
```

```
[1] 25
```

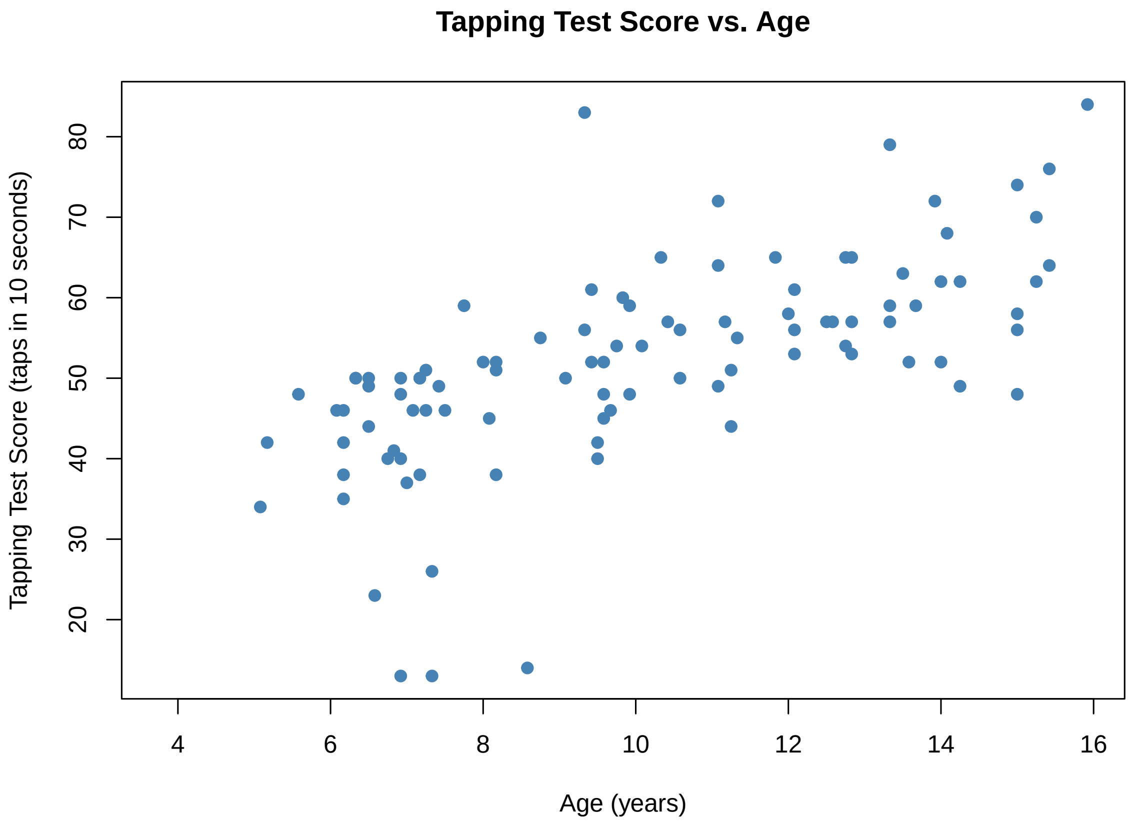
2.4 Question 1: Is Age a Confounder?

For age to be a confounder, it must be associated with **both** the exposure (Group) and outcome (maxfwt).

2.4.1 Criterion 1: Age Associated with Outcome?

Let's examine whether tapping test scores vary with age:

```r
# Scatterplot of age vs. tapping score
plot(lead$ageyrs, lead$maxfwt,
     main = "Tapping Test Score vs. Age",
     xlab = "Age (years)",
     ylab = "Tapping Test Score (taps in 10 seconds)",
     pch = 19,
     col = "steelblue")
```



```r
# Pearson's correlation test
cor.test(lead$ageyrs, lead$maxfwt)
```

```
	Pearson's product-moment correlation

data:  lead$ageyrs and lead$maxfwt
t = 8.3903, df = 97, p-value = 3.947e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5173061 0.7499017
sample estimates:
      cor 
0.6484923 
```

Interpreting the Correlation

Look for:

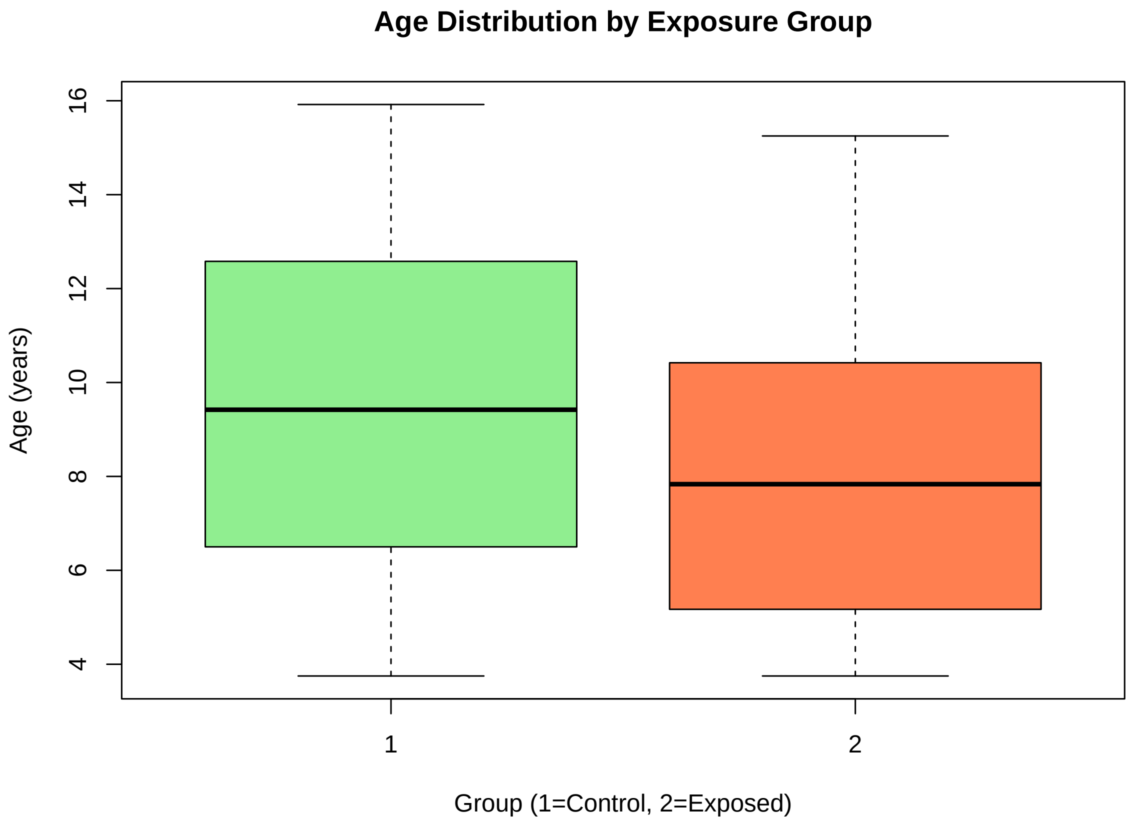
- **Correlation coefficient (r):** Strength and direction
  - Close to 0: weak relationship
  - Close to ±1: strong relationship
  - Positive: older children score higher
  - Negative: older children score lower
- **95% confidence interval:** Uncertainty around r
- **p-value:** Is the correlation statistically significant?

If $p < 0.05$ and r 0, age **IS** associated with maxfwt **Criterion 1 met**

2.4.2 Criterion 2: Age Associated with Exposure?

Now let's check if exposed and control children differ in age:

```r
# Boxplot of age by exposure group
boxplot(lead$ageyrs ~ lead$Group,
        main = "Age Distribution by Exposure Group",
        xlab = "Group (1=Control, 2=Exposed)",
        ylab = "Age (years)",
        col = c("lightgreen", "coral"))
```



```
# Summary statistics by group
by(lead$ageyrs, lead$Group, summary)
```

```
lead$Group: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.750   6.500   9.420   9.327  12.560  15.920 
------------------------------------------------------------ 
lead$Group: 2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.750   5.272   7.835   8.270  10.335  15.250 
```

```
# Two-sample t-test
t.test(lead$ageyrs ~ lead$Group, var.equal = TRUE)
```

```

    Two Sample t-test
```

```
data:  lead$ageyrs by lead$Group
t = 1.6188, df = 122, p-value = 0.1081
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to
95 percent confidence interval:
 -0.2356665  2.3507167
sample estimates:
mean in group 1 mean in group 2 
       9.327308        8.269783 
```

```
# Calculate difference in means
9.327 - 8.270
```

```
[1] 1.057
```

```
# Alternative: Simple linear regression (equivalent to t-test)
summary(lm(lead$ageyrs ~ lead$Group, data = lead))
```

```

Call:
lm(formula = lead$ageyrs ~ lead$Group, data = lead)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5773 -2.8698 -0.0485  2.9202  6.9802 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.3848     0.9496  10.936   <2e-16 ***
lead$Group   -1.0575     0.6533  -1.619    0.108    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.514 on 122 degrees of freedom
Multiple R-squared:  0.02103,	Adjusted R-squared:  0.013 
F-statistic: 2.621 on 1 and 122 DF,  p-value: 0.1081
```

ℹ Interpretation

Look at the results:

- **Mean age by group:** Control ~9.3 years, Exposed ~8.3 years

- **Difference:** Control children are about 1 year older
- **p-value:** Is this difference statistically significant?

If mean ages differ significantly ($p < 0.05$), age **IS** associated with exposure group
Criterion 2 met
Conclusion: If BOTH criteria are met, age is likely a confounder!

2.5 Question 2: Unadjusted Association

Let's estimate the crude (unadjusted) association between lead exposure and tapping test score. This does **NOT** account for age differences.

```r
# Simple linear regression: maxfwt ~ Group
ufit <- lm(maxfwt ~ Group, data = lead)
summary(ufit)
```

```

Call:
lm(formula = maxfwt ~ Group, data = lead)

Residuals:
    Min      1Q  Median      3Q     Max 
-41.438  -5.933   0.562   7.067  35.571 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   61.446      3.758  16.351  < 2e-16 ***
Group         -7.009      2.618  -2.677  0.00872 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.45 on 97 degrees of freedom
  (25 observations deleted due to missingness)
Multiple R-squared:  0.06881,	Adjusted R-squared:  0.05921 
F-statistic: 7.168 on 1 and 97 DF,  p-value: 0.008718
```

Interpretation

Key values:

- **Coefficient for Group:** Difference in mean tapping score between exposed and control children

  - Negative value: Exposed children score lower
  - Example: -7.009 means exposed children have 7 fewer taps on average

- **p-value:** Is this difference statistically significant?
- **R-squared:** Proportion of variation in tapping scores explained by exposure alone

Remember: This is unadjusted - it may be biased by age confounding!

2.6 Question 3: Adjusted Association

Now let's adjust for age to get the "true" association:

```r
# Multiple linear regression: maxfwt ~ Group + ageyrs
afit <- lm(maxfwt ~ Group + ageyrs, data = lead)
summary(afit)
```

```

Call:
lm(formula = maxfwt ~ Group + ageyrs, data = lead)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.380  -4.301   0.977   5.495  36.150 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  31.7367     4.6345   6.848 7.10e-10 ***
Group        -4.8489     2.0342  -2.384   0.0191 *  
ageyrs        2.6592     0.3239   8.210 1.02e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.595 on 96 degrees of freedom
  (25 observations deleted due to missingness)
Multiple R-squared:  0.4529,	Adjusted R-squared:  0.4415 
F-statistic: 39.74 on 2 and 96 DF,  p-value: 2.669e-13
```

Interpretation

Compare the **Group coefficient** in adjusted vs. unadjusted models:

- **Unadjusted:** ~-7.009 taps

- **Adjusted:** ~-4.85 taps (or whatever your data shows)

Key questions:

1. Did the coefficient change substantially?
2. Is it still statistically significant?
3. What does this tell us about confounding?

Interpretation of coefficients:

- **Group (adjusted):** Difference in tapping score between exposed vs. control children **of the same age**
- **ageyrs:** Change in tapping score per 1-year increase in age, holding exposure constant (useful for understanding the confounder's effect)

2.7 Question 4: Impact of Adjustment

Let's create a comparison table to see the impact of adjusting for age:

```r
# Extract coefficients for comparison
unadj_coef <- coef(summary(ufit))["Group", "Estimate"]
unadj_pval <- coef(summary(ufit))["Group", "Pr(>|t|)"]

adj_coef <- coef(summary(afit))["Group", "Estimate"]
adj_pval <- coef(summary(afit))["Group", "Pr(>|t|)"]

# Create comparison data frame
comparison <- data.frame(
  Model = c("Unadjusted", "Adjusted for Age"),
  Coefficient = c(unadj_coef, adj_coef),
  P_value = c(unadj_pval, adj_pval),
  Change = c("-",
             sprintf("%.1f%%", abs((adj_coef - unadj_coef) / unadj_coef * 100)))
)

knitr::kable(comparison,
             digits = 3,
             col.names = c("Model", "Group Coefficient", "p-value",
                           "% Change from Unadjusted"),
             caption = "Impact of Adjusting for Age Confounding")
```

Table 4: Impact of Adjusting for Age Confounding

Model	Group Coefficient	p-value	% Change from Unadjusted
Unadjusted	-7.009	0.009	-
Adjusted for Age	-4.849	0.019	30.8%

❗ What Does This Mean?

Observed pattern: The coefficient moved toward zero (was attenuated) after adjusting for age.
Explanation:

1. The unadjusted model **overestimated** the impact of lead exposure
2. Control children were ~1 year older than exposed children on average
3. Older children naturally perform better on the tapping test (developmental maturation)
4. Part of the observed difference was due to **age**, not lead exposure
5. After adjusting for age (comparing children of the **same age**), the true effect of lead is smaller

Clinical interpretation:
Age was a confounder that **exaggerated** the apparent effect of lead exposure. The adjusted analysis gives a more accurate estimate of lead's impact on neurological function. However, even after adjustment, exposed children still perform worse, suggesting a **real adverse effect** of lead exposure on neurological development.

2.8 Question 5: Missing Data Handling

How R Handles Missing Values

After we converted 99 to `NA`, R automatically uses **complete case analysis** (also called "listwise deletion"):

- Any observation missing ANY variable in the model is **excluded** from the analysis
- Example: A child missing `maxfwt` is dropped from both unadjusted and adjusted models
- Example: A child missing `ageyrs` is included in the unadjusted model (which doesn't use age) but **excluded** from the adjusted model

```
# Check how many observations were used in each model
cat("Unadjusted model N:", nobs(ufit), "\n")
```

```
Unadjusted model N: 99
```

```
cat("Adjusted model N:", nobs(afit), "\n")
```

```
Adjusted model N: 99
```

```
cat("Difference:", nobs(ufit) - nobs(afit),
    "observations excluded when age added\n")
```

```
Difference: 0 observations excluded when age added
```

Why This Matters

Complete case analysis is:

- Simple and the default in most software
- Unbiased IF data are "missing completely at random" (MCAR)
- Potentially biased if missingness is "informative"

Example of informative missingness:
If children with very poor neurological function were less able to complete the tapping test (resulting in missing values), excluding them would **underestimate** the true impact of lead exposure.
Best practices:

1. Examine patterns of missing data before analysis
2. Report the number of observations excluded
3. Consider: Why might data be missing? Is missingness related to variables in the analysis?
4. For substantial missingness, consider advanced methods (e.g., multiple imputation)

Summary and Key Takeaways

2.1 Confounding Basics

Key Points

- A **confounder** is associated with both the exposure and the outcome
- Confounding **distorts** the true relationship between exposure and outcome
- Always check: Is the potential confounder associated with **both**?
- Use clinical and biological knowledge to identify plausible confounders

2.2 Why Regression is Powerful

Key Points

- **t-tests** can compare two groups but cannot adjust for confounders
- **Simple linear regression** gives the same answer as a t-test for two groups
- **Multiple linear regression** can adjust for confounders by including them alongside the exposure
- This is a major advantage for observational research!

2.3 How Adjustment Works

Key Points

- "Adjusting" means estimating the exposure effect **within levels** of the confounder
- Example: Comparing males to males and females to females separately
- Regression **averages** these stratum-specific effects
- The adjusted coefficient is the "true" effect, free from confounding

2.4 Interpreting Results

Key Points

Unadjusted model:

- Shows crude association
- May be biased by confounding
- Easier to calculate and communicate

Adjusted model:

- Accounts for confounders

- More accurate estimate
- Essential for causal inference in observational studies

Compare them: Did adjustment change the estimate substantially? If yes, confounding was present and adjustment was necessary!

2.5 Clinical Examples Recap

Example 1: FEV1 and Genetic Variation

Confounder: Sex

- Sex was associated with genotype (different % females in each group)
- Sex was associated with FEV1 (males vs. females have different lung capacity)
- Adjustment revealed the "true" genotype effect after accounting for sex differences

2.6 Example 2: Lead Exposure and Neurological Function

Confounder: Age

- Age was associated with exposure group (control children were older)
- Age was associated with tapping scores (older children score higher)
- Adjustment attenuated the lead effect, showing part of the crude association was due to age, not lead alone

Next Steps

Practice These Skills

1. **Identify** potential confounders in your own research using clinical knowledge
2. **Always compare** unadjusted and adjusted models to assess confounding
3. **Create** stratified analyses to understand how adjustment works
4. **Think carefully** about which variables to adjust for:
   - Must be associated with both exposure and outcome
   - Should not be on the causal pathway (mediators)
   - Consider directed acyclic graphs (DAGs) for complex situations
5. **Report** both unadjusted and adjusted estimates in your papers
6. **Consider** missing data patterns and their potential impact

---

Additional Resources

2.1 Recommended Reading

- Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern Epidemiology* (3rd ed.). Chapter on confounding.
- Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What If.* Free at https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

2.2 R Resources

- R for Data Science: https://r4ds.had.co.nz/
- Statistical Modeling with R: Multiple regression chapters
- Quarto documentation: https://quarto.org/

Session Information

This document was created using Quarto and R. Here's the session information for reproducibility:

```r
sessionInfo()
```

```
R version 4.4.2 (2024-10-31)
Platform: aarch64-apple-darwin20
Running under: macOS Sequoia 15.7.1

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: America/New_York
tzcode source: internal

attached base packages:
```

```
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] ellmer_0.1.1

loaded via a namespace (and not attached):
 [1] digest_0.6.37     coro_1.1.0        R6_2.5.1          fastmap_1.2.0
 [5] xfun_0.50         magrittr_2.0.3    rappdirs_0.3.3    glue_1.8.0
 [9] knitr_1.49        htmltools_0.5.8.1 rmarkdown_2.29    lifecycle_1.0.4
[13] cli_3.6.3         S7_0.2.0          compiler_4.4.2    tools_4.4.2
[17] evaluate_1.0.3    yaml_2.3.10       httr2_1.1.0       jsonlite_1.8.9
[21] rlang_1.1.5
```