Tidy Tuesday: From Biased to Unbiased - Addressing OMV in Sports Expenditure Modeling

Shafqat Shafiq

2025-03-08

1 Introduction:

We are going to use data on U.S. Collegiate Sports to understand what affects the expenditures ~ How much a college spends on certain sports teams.

We will look at things like genders (are there differences between men's and women's sports), what sports teams have the most money spent on them, and so on.

Instead of it being more like a predictive modeling exercise, we'll focus on the following:

- How to build models
- How to handle the uncertainty around the coefficients

Our modeling goal is to understand *what affects expenditures in Collegiate Sports in the US.*

2 Explore Data:

```
# Load dataset
sports_raw <- read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/

# Explore dataset
sports_raw %>% slice_sample(n = 20)
```

```
# A tibble: 20 x 28
    year unitid institution_name  city_txt state_cd zip_text classification_code
   <dbl>  <dbl> <chr>             <chr>    <chr>    <chr>                  <dbl>
 1  2019 218399 Morris College    Sumter   SC       29150                      9
 2  2016 486840 Kennesaw State U~ Kennesaw GA       30144                      2
 3  2016 141486 Chaminade Univer~ Honolulu HI       96816                      5
 4  2019 221397 Roane State Comm~ Harriman TN       37748                     12
 5  2017 220312 Hiwassee College  Madison~ TN       37354                     15
 6  2015 216542 University of Va~ Phoenix~ PA       1946023~                   7
 7  2015 181020 Doane University~ Crete    NE       68333                     10
 8  2015 215770 Saint Joseph's U~ Philade~ PA       1913113~                   3
 9  2019 150534 University of Ev~ Evansvi~ IN       47722                      3
10  2015 155292 Kansas City Kans~ Kansas ~ KS       66112                     12
11  2019 221519 The University o~ Sewanee  TN       37383                      6
12  2017 154518 Waldorf Universi~ Forest ~ IA       50436                     10
13  2019 179265 St. Louis Colleg~ Saint L~ MO       63110                      9
14  2019 199102 North Carolina A~ Greensb~ NC       27411                      2
15  2018 168227 Wentworth Instit~ Boston   MA       02115                      7
16  2019 147244 Millikin Univers~ Decatur  IL       62522                      6
17  2019 186469 Salem Community ~ Carneys~ NJ       0806927~                  14
18  2019 141185 Toccoa Falls Col~ Toccoa ~ GA       30598                     16
19  2016 237950 West Virginia Un~ Beckley  WV       25801                     10
20  2015 195526 Skidmore College  Saratog~ NY       12866                      7
# i 21 more variables: classification_name <chr>, classification_other <chr>,
#   ef_male_count <dbl>, ef_female_count <dbl>, ef_total_count <dbl>,
#   sector_cd <dbl>, sector_name <chr>, sportscode <dbl>, partic_men <dbl>,
#   partic_women <dbl>, partic_coed_men <dbl>, partic_coed_women <dbl>,
#   sum_partic_men <dbl>, sum_partic_women <dbl>, rev_men <dbl>,
#   rev_women <dbl>, total_rev_menwomen <dbl>, exp_men <dbl>, exp_women <dbl>,
#   total_exp_menwomen <dbl>, sports <chr>
```

```
# Use glimpse
sports_raw %>% glimpse()
```

```
Rows: 132,327
Columns: 28
$ year                  <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2~
$ unitid                <dbl> 100654, 100654, 100654, 100654, 100654, 100654, 1~
$ institution_name      <chr> "Alabama A & M University", "Alabama A & M Univer~
$ city_txt              <chr> "Normal", "Normal", "Normal", "Normal", "Normal",~
$ state_cd              <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "~
$ zip_text              <chr> "35762", "35762", "35762", "35762", "35762", "357~
$ classification_code  <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1~
$ classification_name  <chr> "NCAA Division I-FCS", "NCAA Division I-FCS", "NC~
$ classification_other <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ ef_male_count        <dbl> 1923, 1923, 1923, 1923, 1923, 1923, 1923, 1923, 1~
$ ef_female_count      <dbl> 2300, 2300, 2300, 2300, 2300, 2300, 2300, 2300, 2~
$ ef_total_count       <dbl> 4223, 4223, 4223, 4223, 4223, 4223, 4223, 4223, 4~
$ sector_cd            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ sector_name          <chr> "Public, 4-year or above", "Public, 4-year or abo~
$ sportscode           <dbl> 1, 2, 3, 7, 8, 15, 16, 22, 26, 33, 1, 2, 3, 8, 12~
$ partic_men           <dbl> 31, 19, 61, 99, 9, NA, NA, 7, NA, NA, 32, 13, NA,~
$ partic_women         <dbl> NA, 16, 46, NA, NA, 21, 25, 10, 16, 9, NA, 20, 68~
$ partic_coed_men      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ partic_coed_women    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ sum_partic_men       <dbl> 31, 19, 61, 99, 9, 0, 0, 7, 0, 0, 32, 13, 0, 10, ~
$ sum_partic_women     <dbl> 0, 16, 46, 0, 0, 21, 25, 10, 16, 9, 0, 20, 68, 7,~
$ rev_men              <dbl> 345592, 1211095, 183333, 2808949, 78270, NA, NA, ~
$ rev_women            <dbl> NA, 748833, 315574, NA, NA, 410717, 298164, 13114~
$ total_rev_menwomen   <dbl> 345592, 1959928, 498907, 2808949, 78270, 410717, ~
$ exp_men              <dbl> 397818, 817868, 246949, 3059353, 83913, NA, NA, 9~
$ exp_women            <dbl> NA, 742460, 251184, NA, NA, 432648, 340259, 11388~
$ total_exp_menwomen   <dbl> 397818, 1560328, 498133, 3059353, 83913, 432648, ~
$ sports               <chr> "Baseball", "Basketball", "All Track Combined", "~
```

The data is in **long** format and contains 28 variables including `year`, `institution_name`, `sports`, `exp_men`, `exp_women`, and so on.

Let's look at the unique `year` and `sports` values/categories that are present in the dataset.

```
unique(sports_raw$year) # 2015, 2016, 2017, 2018, 2019
```

```
[1] 2015 2016 2017 2018 2019
```

```
unique(sports_raw$sports) # 38 different sports teams
```

```
 [1] "Baseball"                     "Basketball"
 [3] "All Track Combined"           "Football"
 [5] "Golf"                         "Soccer"
 [7] "Softball"                     "Tennis"
 [9] "Volleyball"                   "Bowling"
[11] "Rifle"                        "Beach Volleyball"
[13] "Ice Hockey"                   "Lacrosse"
```

```
[15] "Gymnastics"                  "Rowing"
[17] "Swimming and Diving"         "Track and Field, X-Country"
[19] "Equestrian"                  "Track and Field, Indoor"
[21] "Track and Field, Outdoor"    "Wrestling"
[23] "Other Sports"                "Rodeo"
[25] "Skiing"                      "Swimming"
[27] "Water Polo"                  "Archery"
[29] "Field Hockey"                "Fencing"
[31] "Sailing"                     "Badminton"
[33] "Squash"                      "Diving"
[35] "Synchronized Swimming"       "Table Tennis"
[37] "Weight Lifting"              "Team Handball"
```

There are too many sports teams (almost 40!). Let's combine a few of them to make our dataset more pallatable and concise.

Let's combine all sports containing swimming and/or diving into swimming and all sports containing track into track.

```r
sports_parsed <-
sports_raw %>% mutate(
  sports = case_when(
    str_detect(sports, "Swimming") ~ "Swimming and Diving",
    str_detect(sports, "Diving") ~ "Swimming and Diving",
    str_detect(sports, "Track") ~ "Track",
    TRUE ~ sports
  )
)
```

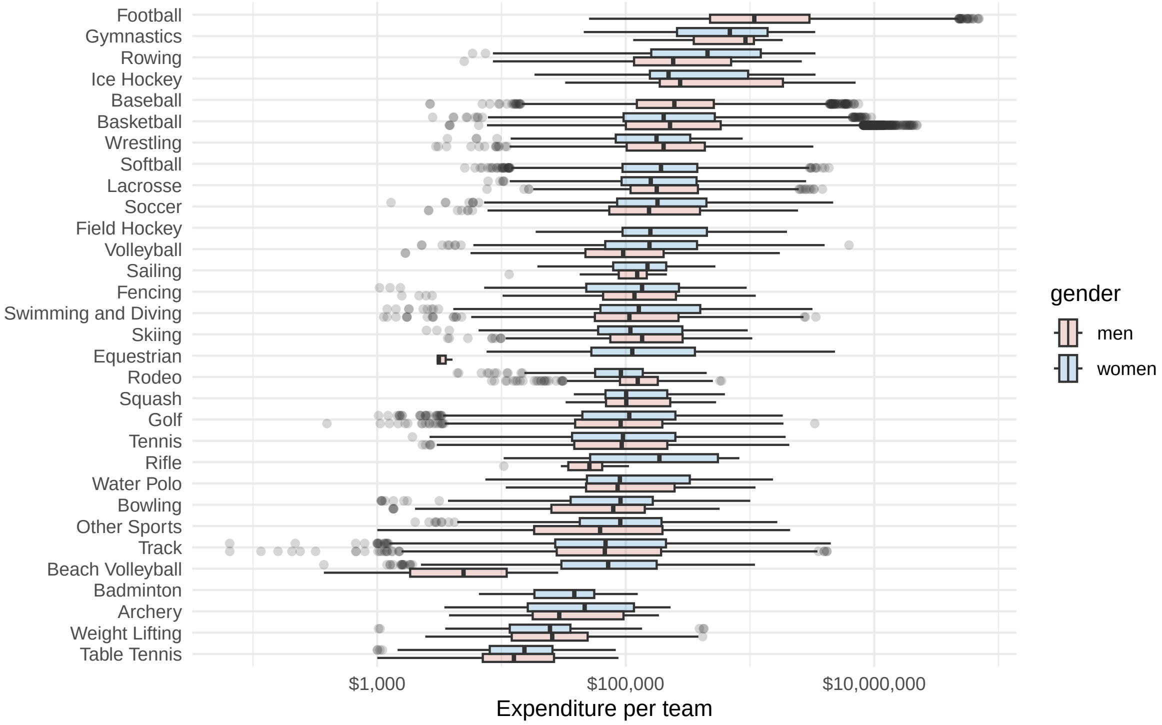
Recall that the data is in a **long** format meaning that each row corresponds to a unique combination of `sport`, `year`, `male participation`, `female participation` and `school` combination. This particular dataset is often used to look at the status of women's sports in colleges.

Let's change the structure of this data a little bit, by making it pseudo-wider. That is, make it one row per `year`, `college` and `sport`, and by combining the `male/female participations` into one column.

```r
# Transforming Wide Data to Long Format using bind_rows

sports <- bind_rows(
  # 1. Select and transform data for male participants
  sports_parsed %>%
    select(year, institution_name, sports,
           participants = partic_men, # Rename male participants
           revenue = rev_men,         # Rename male revenue
           expenditure = exp_men) %>%  # Rename male expenditure
    mutate(gender = "men"),          # Add a 'gender' column for males

  # 2. Select and transform data for female participants
  sports_parsed %>%
    select(year, institution_name, sports,
           participants = partic_women, # Rename female participants
           revenue = rev_women,         # Rename female revenue
           expenditure = exp_women) %>%  # Rename female expenditure
    mutate(gender = "women")          # Add a 'gender' column for females
) %>% na.omit()

# Explanation:
# This code takes a 'wide' format dataset (sports_parsed) where male and female
# data are in separate columns and converts it to a 'long' format.
#
# Steps:
# 1. For male data:
#    - Selects relevant columns and renames them to generic names (participants,
#      revenue, expenditure).
#    - Adds a 'gender' column with the value "men".
# 2. For female data:
#    - Selects relevant columns and renames them to generic names.
#    - Adds a 'gender' column with the value "women".
# 3. Combines the male and female data frames using bind_rows(), stacking them
#    on top of each other.
#
# Result:
# The resulting 'sports' data frame is in a 'long' format, where each row represents
# a single observation (a sports program at an institution), and the 'gender'
# column distinguishes between male and female data. This format is more suitable
# for analysis and visualization.
```

Let's now make boxplots that show how expenditure is distributed across these sports teams. That is, let's look at how much is actually spent on them.

```
library(ggsci)

sports %>%
  mutate(sports = fct_reorder(sports, expenditure)) %>%
  ggplot(aes(y = sports, x = expenditure, fill = gender)) +
  geom_boxplot(alpha = 0.2, position = position_dodge(preserve = "single")) +
  scale_x_log10(labels = scales::dollar) +
  scale_fill_nejm() + # Or scale_fill_jama(), scale_fill_aaas(), etc.
  labs(y = NULL, color = NULL, x = "Expenditure per team")
```



We can immediately notice some massive differences in expenditures between the different teams, especially ones that are single gendered sports such as Football and Equestrian.

The most conspicuous difference is in Beach Volleyball where women's team gets a lot more funding compared to the men's team.

But does gender and sports actually have an impact on how much is spent on a team? Let's build a linear model to investigate the (possible) causal relationship.

3 Build Linear Models:

We first look to create a Linear model that **does NOT** take into account differences into account `sports` and only really looks at `expenditure` as a function of `gender` and the number of `participants`.

In short, we'll try to answer the following:

How much does expenditure per team change based on:

1. How many participants there are
2. The gender of the (sports) team

The second model we'll create will **take into account** `sports` as an explanatory variable, along with `gender` and `expenditure`

3.1 Omitted Variable Bias (OMV) in Sports Expenditure Analysis:

In our analysis of sports expenditures, we are examining how `gender` and `participants` influence `expenditure`. However, if we construct a model that **excludes** the categorical variable `sports`, we risk encountering Omitted Variable Bias (OMV).

- Why OMV Might Occur:
  - It's likely that different sports have inherently different expenditure levels (e.g., football vs. tennis).
  - Furthermore, the distribution of `gender` might vary across sports (e.g., more male participants in football, more female participants in volleyball).
  - Thus, `sports` is correlated with both our independent variables (`gender`, `participants`) and our dependent variable (`expenditure`).
- Consequences of OMV:
  - If we omit `sports`, the estimated coefficients for `gender` and `participants` will be biased.
  - This means we might incorrectly attribute the effect of `sports` to `gender` or `participants`. For example, we might over- or underestimate the effect of `gender` on `expenditure` if we don't account for the fact that certain sports, which have different expenditure levels, also have different gender distributions.
- Mitigation:
  - To address OMV, we include `sports` as a categorical variable in our second model.
  - This allows us to control for the effect of `sports` and obtain more accurate estimates of the relationships between `gender`, `participants`, and `expenditure`.

  - By comparing the models, we can see the degree to which OMV effected the first model.
- Importance:
  - Recognizing and mitigating OMV is crucial for drawing valid conclusions in our analysis.

  - It ensures that our model accurately reflects the underlying relationships in the data.

3.2 The Linear Models:

```
# Without sports
ignore_sports <- lm(expenditure ~ participants + gender, data = sports)

# With sports
account_sports <- lm(expenditure ~ participants + gender + sports, data = sports)
```

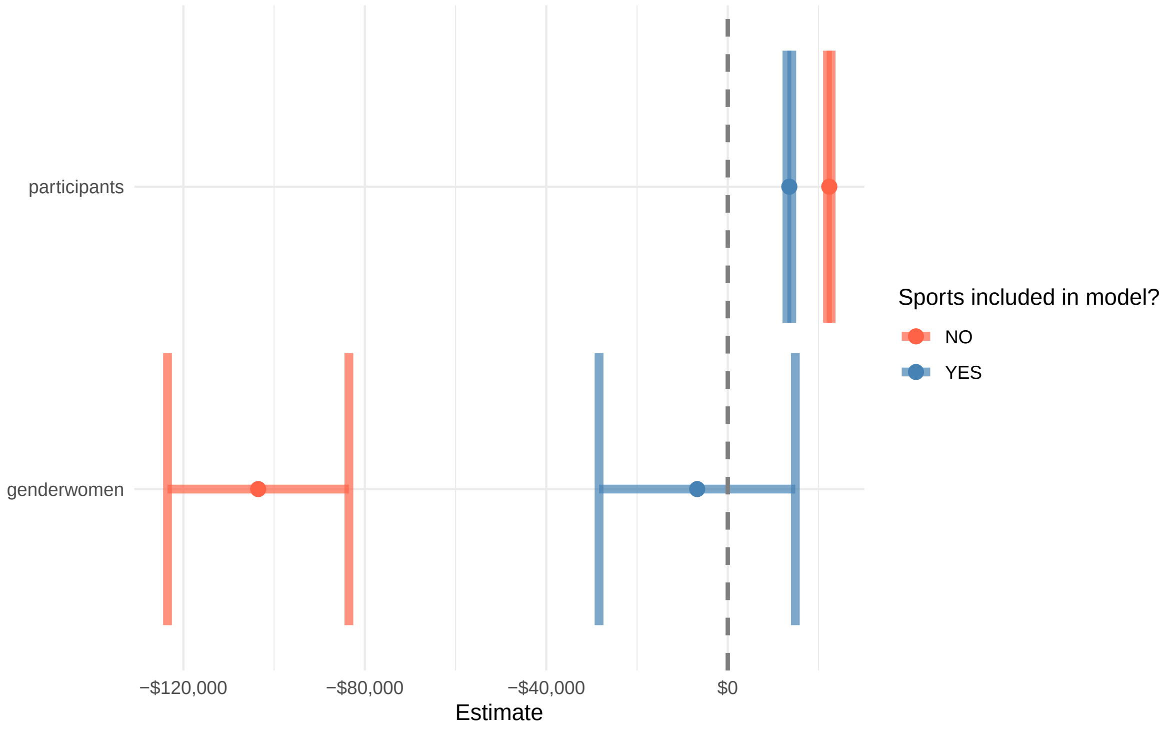
Let's look at the summary and plot using `tidy()`. `tidy()` will return a tibble of the results. In order to understand whether there is a significant difference in the coefficients of `gender` and `participants` before and after accounting for `sports`, we can once again make use of `bind_rows()` function.

We first create a new column called `ignore_sports` which takes on either `Yes`/`No`. We then attach it to the tidy results for `ignore_sports` and `account_sports` respectively. After that, we remove all the terms unique to either model, and only keep the terms that are common between both `ignore_sports` and `account_sports` (specifically, `gender` and `participants`). Once we have done that, we can then visualize and see if there's a significant difference in the coefficients for `gender` and `participants`. If there is a visually and statistically significant difference, it means that `sports` is an important variable and omitting it will lead to **Omitted Variable Bias** in the model.

```
library(broom)

results <-
  bind_rows(
    tidy(ignore_sports) %>%
      mutate(account_sports = "NO"),

    tidy(account_sports) %>%
      mutate(account_sports = "YES")
)
```

Let's now make a plot of the common coefficients between the 2 models; i.e. get rid of terms not common between the 2 models. By looking at the data, we should get rid of all terms that start with `sports`. We can also get rid of the intercept (but do it at your own peril!).

```
# Filter intercept and terms that start with "sports"
results_modified <-
results %>%
  filter(!str_detect(term, "sports"),
         term != "(Intercept)")


# Visualize, where term = coefficient name

results_modified %>%
  ggplot(aes(x = estimate, y = term, color = account_sports)) +
  geom_point(size = 3) +
  geom_errorbar(aes(
    xmin = estimate - (1.96 * std.error),
    xmax = estimate + (1.96 * std.error)
  ), size = 2, alpha = 0.7) + # Reduced errorbar line thickness
  scale_x_continuous(labels = scales::dollar) +
  geom_vline(xintercept = 0,
             lty = 2, linewidth = 1, color = "gray50") + # Reduced vline thickness
  labs(x = "Estimate", y = NULL, color = "Sports included in model?") +
  scale_color_manual(values = c("YES" = "steelblue", "NO" = "tomato")) # Custom color palett
```



3.3 Impact of Omitted Variable Bias (OMV) on Gender Coefficient

We observe a significant reduction in the `genderwomen` coefficient when the `sports` variable was included in the model. Without `sports`, the coefficient was approximately `-$103,506`, indicating a substantial negative effect of being a female team on expenditures. After accounting for `sports`, the coefficient decreased to approximately `-$6,739`, and became statistically insignificant.

This drastic change demonstrates OMV. `sports` acts as a confounding variable, correlated with both `gender` and `expenditure`. Omitting `sports` led to an overestimation of the gender effect, as the model attributed the influence of `sports` to `gender`.

Therefore, including `sports` provides a more accurate and unbiased estimate of the `genderwomen` effect, highlighting the importance of addressing OMV in regression modeling.

Basically, if we ignore `sports` as a variable, then teams spend almost `$100,000` less on female sports teams, on average. But if we account for `sports` the `genderwomen` coefficient becomes almost 0 (on the log scale), which means there's no statistically significant difference between men's and women's sports teams expenses, when account for `sports`.

4 Bootstrap Confidence Intervals:

In the previous section we performed regression analysis and made inferences using the standard OLS intervals. It is worth noting however that sometimes we don't actually want to do that. In most applied cases, OLS assumptions are violated and therefore may provide incorrect intervals.

- Incorrect Standard Errors/Intervals lead to incorrect test statistics
- Incorrect test statistics lead to incorrect p-values
- Underestimated Standard Errors lead to smaller p-vales and more **False Positives**
- Overestimated Standard Errors lead to larger p-values and more **False Negatives**

Recall that **p-value** is the probability of observing a test statistic *as extreme or more* than the computed test statistic, while assuming the Null hypothesis being true.

- **False Positives (Type I Error):**
  - Reject Null when it's actually true.
  - Small p-vale (less than `alpha`) leads to more frequent rejections.
  - `alpha` is the probability of false positive.
- **False Negatives (Type II Error):**
  - Fail to reject Null when it's actually false.
  - Large p-value (greater than `alpha`) leads to more frequent acceptance of Null.

Bootstrap Intervals ignore the estimate standard errors (from OLS). Instead, we fit the model a whole bunch of times using resamples of the data, and observe how much the coefficients change over each resample/bootstrap fit.

```r
library(infer)
library(future)

# Set up parallel processing with 20 workers
plan(multisession, workers = 20)

set.seed(123)

# Create Bootstrap Intervals using reg_intervals
```

```
ignore_sports_intervals <- reg_intervals(
  expenditure ~ gender + participants, data = sports,
  times = 500
)

account_sports_intervals <- reg_intervals(
  expenditure ~ gender + sports + participants, data = sports,
  times = 500
)

# Optional: Reset the plan if needed
# plan(sequential)
```

Let's check what the Bootstrap estimates are for each sport.

```
account_sports_intervals %>%
  filter(str_detect(term, "sports")) %>%
  arrange(desc(.estimate))
```

```
# A tibble: 30 x 6
   term                 .lower .estimate   .upper .alpha .method
   <chr>                 <dbl>     <dbl>    <dbl>  <dbl> <chr>
 1 sportsFootball     2644553.  2836230. 3045354.   0.05 student-t
 2 sportsGymnastics    639676.   704364.  763908.   0.05 student-t
 3 sportsIce Hockey    508854.   581918.  644952.   0.05 student-t
 4 sportsBasketball    549559.   575211.  603436.   0.05 student-t
 5 sportsEquestrian    114303.   206443.  298579.   0.05 student-t
 6 sportsRifle          85377.   156425.  206370.   0.05 student-t
 7 sportsVolleyball    130314.   146592.  163526.   0.05 student-t
 8 sportsSkiing         99015.   126090.  153753.   0.05 student-t
 9 sportsField Hockey   90133.   118288.  144998.   0.05 student-t
10 sportsRowing         77110.   117921.  161518.   0.05 student-t
# i 20 more rows
```

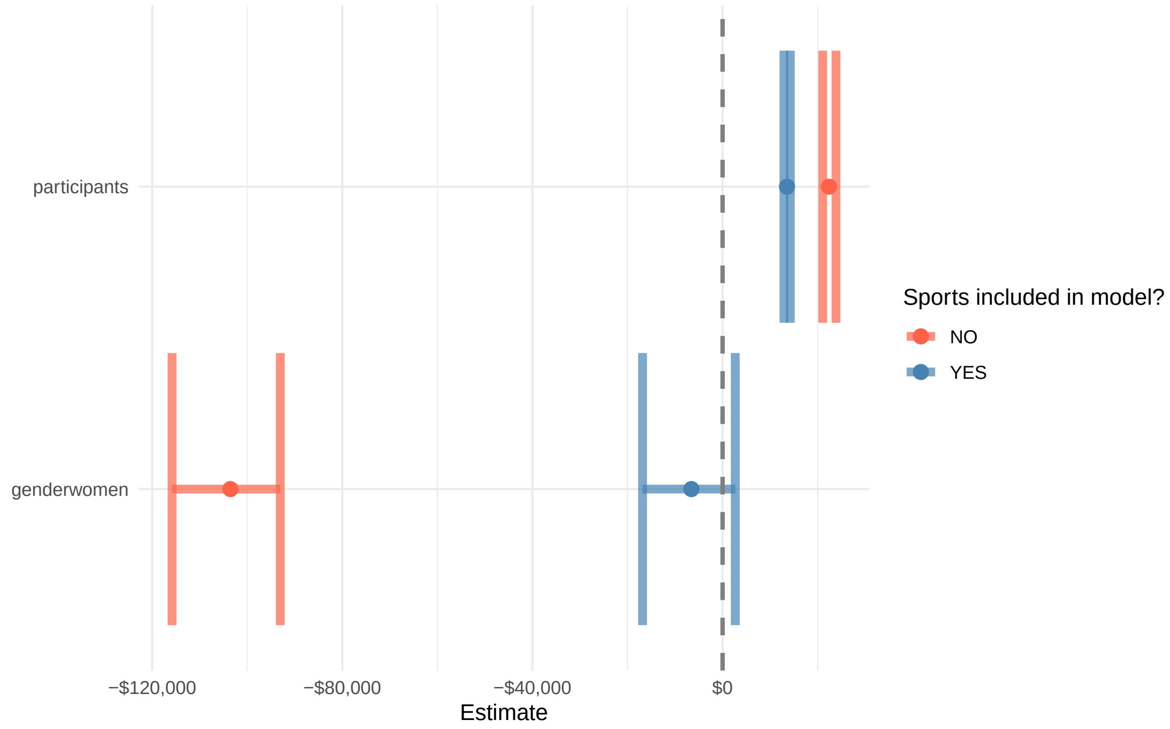
Let's once again, keep only the common terms, bind the rows and visualize the two dataframes.

```
bootstrap_results <-
bind_rows(
ignore_sports_intervals %>%
  mutate(account_sports = "NO"),
account_sports_intervals %>%
  mutate(account_sports = "YES")
  ) %>% filter(!str_detect(term, "sports"))

bootstrap_results
```

```
# A tibble: 4 x 7
  term            .lower .estimate  .upper .alpha .method   account_sports
  <chr>            <dbl>     <dbl>   <dbl>  <dbl> <chr>     <chr>
1 genderwomen  -115821.  -103542. -93056.   0.05 student-t NO
2 participants   21063.    22369.  23857.   0.05 student-t NO
3 genderwomen   -16838.    -6566.   2669.   0.05 student-t YES
4 participants   12907.    13557.  14257.   0.05 student-t YES
```

Let’s visualize these results.

```
ggplot(bootstrap_results, aes(x = .estimate, y = term, color = account_sports)) +
  geom_point(size = 3) +
  geom_errorbar(aes(
    xmin = .lower,
    xmax = .upper
  ), size = 2, alpha = 0.7) +
  scale_x_continuous(labels = scales::dollar) +
  geom_vline(xintercept = 0, lty = 2, linewidth = 1, color = "gray50") +
  labs(x = "Estimate", y = NULL, color = "Sports included in model?") +
  scale_color_manual(values = c("YES" = "steelblue", "NO" = "tomato"))
```



While there aren't any visibly significant shifts in the estimates, the confidence intervals have become tighter! This means we are much more confident about the uncertainty around the coefficients now, since tighter confidence intervals suggest more precise estimates of uncertainty.

5 Summary of Sports Expenditure Analysis:

5.1 Initial Findings:

The inclusion of the `sports` variable resulted in significant changes to the point estimates of both `genderwomen` and `participants`. The changes in the estimates demonstrate that omitting `sports` introduces bias, and therefore it should be included to ensure the model's validity.

5.2 Bootstrap Analysis:

A bootstrap analysis was conducted to assess the uncertainty around the coefficients. The bootstrap analysis confirmed the significant impact of including `sports` on the point estimates, and also provided tighter confidence intervals. This refined our understanding of the uncertainty around these coefficients, indicating more precise estimates. The tighter bounds reaffirmed the importance of including `sports` to avoid bias, and increased our confidence in the magnitude of the effects.

6 Analysis of Sports Expenditure and the Importance of Omitted Variable Bias Checks:

6.1 Initial Model and Omitted Variable Concerns

We began by examining the relationship between `gender` and `participants` on `expenditure` within a dataset named `sports`. Initially, we fit a model excluding the `sports` variable. However, we suspected that `sports` might be a relevant predictor and a potential confounder, influencing both the independent variables and the outcome. Therefore, we fit a second model including `sports`. The initial comparison of point estimates from these models revealed significant changes in the coefficients for `genderwomen` and `participants` when `sports` was included. This highlighted the potential for omitted variable bias (OVB) in the model without `sports`.

6.2 Bootstrap Analysis and Refinement

To further assess the uncertainty and robustness of our findings, we conducted a bootstrap analysis. This resampling-based approach allowed us to generate confidence intervals for our coefficients, providing a more reliable measure of their variability than standard errors alone. The bootstrap analysis reaffirmed the substantial impact of including `sports` on the point estimates, validating our concerns about OVB. Furthermore, the bootstrap provided tighter confidence intervals, indicating more precise estimates of uncertainty around the coefficients. This increased our confidence in the magnitude and significance of the observed effects.

6.3 Importance of Omitted Variable Checks and Robust Analysis

This exercise demonstrates the critical importance of conducting omitted variable bias checks in regression analysis. Failing to consider potential confounders like `sports` can lead to biased estimates and misleading conclusions about the relationships between variables. The significant changes in coefficients observed when `sports` was included underscore this point. Additionally, the bootstrap analysis highlights the value of robust statistical methods in validating and refining our understanding of model parameters. By providing tighter confidence intervals, the bootstrap increased our confidence in the precision of our estimates. These checks are essential for ensuring the validity and reliability of our findings, ultimately leading to more informed and accurate interpretations of the data.