# Homework 4

## Background: California Department of Developmental Services

From Taylor, S. A., & Mickel, A. E. (2014). Simpson's Paradox: A Data Set and Discrimination Case Study Exercise. Journal of Statistics Education, 22(1):

> Most states in the USA provide services and support to individuals with developmental disabilities (e.g., intellectual disability, cerebral palsy, autism, etc.) and their families. The agency through which the State of California serves the developmentally-disabled population is the California Department of Developmental Services (DDS) … One of the responsibilities of DDS is to allocate funds that support over 250,000 developmentally-disabled residents. A number of years ago, an allegation of discrimination was made and supported by a univariate analysis that examined average annual expenditures on consumers by ethnicity. The analysis revealed that the average annual expenditures on Hispanic consumers was approximately one-third of the average expenditures on White non-Hispanic consumers. This finding was the catalyst for further investigation; subsequently, state legislators and department managers sought consulting services from a statistician.

In this assignment, you'll analyze the deidentified DDS data published with this article to answer the question: *is there evidence of ethnic or gender discrimination in allocation of DDS funds?* This will involve practicing the following:

- exploratory data visualization
- regression analysis
- model visualization

**Aside**: The JSE article focuses on what's known as Simpson's paradox, an arithmetic phenomenon in which aggregate trends across multiple groups show the *opposite* of within-group trends. We won't emphasize tis topic in this assignment, although we will discuss it in class. As you work through the assignment, think how this data might highlight Simpson's paradox?

## DDS data

The data for this assignment are already tidy, so in this section you'll just familiarize yourself with basic characteristics. The first few rows of the data are shown below:

```
dds <- read_csv('data/california-dds.csv')
print(dds)
```

Take a moment to open and read the data documentation (*data > california-dds-documentation.md*).

#### Question 1: Data description

Write a short paragraph answering the following questions based on the data documentation.

- (i) Why were the data collected? What is the purpose of this dataset?
- (ii) What are the observational units?
- (iii) What is the population of interest?
- (iv) How was the sample obtained (*e.g.* random sampling, adminsitrative data, convenience sampling, etc.)?
- (v) Can inferences about the population be drawn from the sample?

In addition, make a table summarizing the variables measured. Use the format below.

| Name | Variable description | Type | Units of measurement |
|---|---|---|---|
| ID | Unique consumer identifier | Numeric | None |

*Type your answer here, replacing this text.*

## Exploratory analysis

Here you'll use graphical and descriptive techniques to explore the allegation of discriminatory allocation of benefits.

#### Question 2: Alleged discrimination

Construct a data frame showing median expenditures by ethnicity that also shows the sample size for each ethnic group in the data.

```r
median_expend_by_eth <- ...
# print
print(median_expend_by_eth, n=15)
```

### Question 3: Plot median expenditures

Construct a point-and-line plot of median expenditure (y) against ethnicity (x), with:

- ethnicities sorted by descending median expenditure;
- the median expenditure axis shown on the log scale;
- the y-axis labeled 'Median expenditure'; and
- no x-axis label (since the ethnicity group names are used to label the axis ticks, the label 'Ethnicity' is redundant).

A point-and-line plot is a plot using both `geom_point` and `geom_line` with the same x and y aesthetics. Store the result as `fig_1` and display the plot.

```r
# Create visualization
fig_1 <- ...

# Display plot
print(fig_1)
```

### Question 4: Age and expenditure

How does expenditure differ by age? Construct a scatterplot of expenditure (y) against age (x). Store the plot as `fig_2`. In one or two sentences, comment on the plot – what is the main pattern it reveals? *Hint* consider transforming the y-axis to better visualize the results.

*Type your answer here, replacing this text.*

```r
# Create scatterplot
fig_2 <- ...

# Display plot
print(fig_2)
```

Precisely because recipients have different needs at different ages that translate to jumps in expenditure, age has been discretized into age cohorts defined based on need level. Going forward, we'll work with these age cohorts – by treating age as discrete, we won't need to attempt to model the discontinuities in the relationship between age and expenditure.

The cohort labels are stored as `Age Cohort` in the dataset. There are six cohorts; the cell below puts them in the proper order, and prints the category levels.

```
# Convert columns to factors (R's categorical type)
dds_cat <- dds %>%
  mutate(across(c('Age Cohort', 'Ethnicity', 'Gender'), as.factor))

# Reorder Age Cohort levels
dds_cat$`Age Cohort` <- factor(
  dds_cat$`Age Cohort`,
  levels = levels(dds_cat$`Age Cohort`)[c(1, 6, 2, 3, 4, 5)],
)

# Display age cohort levels
levels(dds_cat$`Age Cohort`)
```

Here is an explanation of how the cohort age boundaries were chosen:

> The 0-5 cohort (preschool age) has the fewest needs and requires the least amount of funding. For the 6-12 cohort (elementary school age) and 13-17 (high school age), a number of needed services are provided by schools. The 18-21 cohort is typically in a transition phase as the consumers begin moving out from their parents' homes into community centers or living on their own. The majority of those in the 22-50 cohort no longer live with their parents but may still receive some support from their family. Those in the 51+ cohort have the most needs and require the most amount of funding because they are living on their own or in community centers and often have no living parents.

### Question 5: age structure of the sample

Here you'll explore the age composition of each ethnic group in the sample.

- (i) Group the data by ethnic group and tabulate the sample sizes for each group. Use `dds_cat` so that the order of age cohorts is preserved. Store the result as `samp_sizes`.

- (ii) Visualize the age structure of each ethnic group in the sample. Construct a point-and-line plot of the sample size (y) against age cohort (x) by ethnicity (color or linetype). Make sure the ordering of age cohorts is preserved on the x axisStore the plot as `fig_3` and display.

Comment on the figure. Are there differences in age composition by ethnic group among the individuals sampled?

*Type your answer here, replacing this text.*

```
# Compute sample sizes for each age/ethnic group
samp_sizes <- ...

# Create the plot
fig_3 <- ...

# Display plot
print(fig_3)
```

Age structure among ethnic groups might be related to the observed differences in median expenditure, because we know that:

- (i) among the individuals in the sample, age distributions differed by ethnic group
- (ii) age is related to benefit expenditure

To see this, think through an example.

### Question 6: potential confounding

Look at the age distribution for `Multi Race` and consider the age-expenditure relationship. Can you explain why the median expenditure for this group might be lower than the others? Answer in 1-2 sentences.

*Type your answer here, replacing this text.*

### Question 7: correcting for age

Hopefully, the last few prompts convinced you that the apparent discrimination *could* simply be an artefact of differing age structure. We can confirm this by plotting median expenditure against ethnicity, as in Q3, but now also correcting for age cohort.

Construct a point-and-line chart based on `dds_cat` with:

- ethnicity on the x axis

- no x axis label
- median expenditure on the y axis
- the y axis displayed on the log scale
- age cohort mapped to color and sorted in order of age
- lines connecting points that display the median expenditure for each ethnicity and cohort, with one line per age cohort

Store the result as `fig_4` and display the graphic.

```
# Calculate medians and create plot
fig_4 <- ...

# Display plot
print(fig_4)
```

## Statistical Inference

Our exploratory analysis revealed two key insights: 1. There appear to be substantial differences in expenditures between racial/ethnic groups 2. These differences might be explained by different age distributions across groups

However, exploratory analysis alone cannot tell us whether these patterns reflect real population-level differences or just random variation in our sample. While our visualization in Question 7 suggests that age differences might explain the apparent ethnic disparities, we need more rigorous statistical methods to properly account for age when comparing groups.

We'll approach this analysis in two complementary ways:

1. First, we'll use bootstrapping and inverse probability weighting (IPW) to directly estimate and compare expenditures between minority and non-minority groups while accounting for age differences. This approach:
    - Gives us a clear single number: the age-adjusted difference in expenditures
    - Introduces the concept of sampling variability through bootstrapping
    - Directly adjusts for differences in age distributions between groups
2. Later, we'll use regression analysis to:
    - Simultaneously account for multiple factors (age, gender, and specific ethnic groups)
    - Estimate specific differences between individual ethnic groups correcting for the other multiple factors
    - Provide an alternative framework for assessing sampling variability

Using both approaches provides a more complete understanding and serves as a check on our conclusions - if both methods suggest similar findings, we can be more confident in our results.

### Question 8: Chracterizing sampling variance via bootstrapping

We'll start by exploring whether there are differences in expenditures between minorities and non-minorities. Let's define a "non-minority" as "White not Hispanic" and "minority" as all other race/ethnicities.

i) Add the corresponding variable named `Minority` to `dds_cat`.

ii) Write a function, `compute_mean_diff` that reads in a data frame and computes the mean difference in observed expenditures (in dollars) between minorities and non minorities.

iii). Use `bootstraps` to run `compute_mean_diff` on each boostrapped dataframe. Use 1000 bootstrap replicates and plot a histogram of the boostrap distribution of mean expenditure differences.

Based on the bootstrap distribution, what are the reasonable a reasonable range for the mean difference in expenditures between minorities and non-minorities **in the population**? Based on the previous questions, why can we **not** take this as evidence of unfair racial discrimination?

```
## Add `Minority` to dds_cat
dds_cat <- ...

compute_mean_diff <- \(x) {
  ## fill in function
...
}

## compute the bootstrap replicates of the observed df
obs_diff_bootstraps <-

## make the plot
obs_diff_bootrap_plot <- ...
```

### Question 9: Using IPW to adjust for age differences.

In this question, we'll use a technique called inverse propensity weighting (IPW) to adjust for the effect of age. We previously used IPW in an earlier lab to account for sampling bias

in the hawk length example. Here, we're going to use it to correct for the difference in age distributions in each racial/ethnic group.

To do so, let $Z$ be an indicator for whether the individual is a minority (1 if minority, 0 if not), let $Y$ be the expenditure and $X$ be age. The IPW formula tells us that we can estimate the mean age-adjusted difference in expenditures between minorities and non-minorities as:

$$\text{(adjusted difference)} = \sum_{i}^{n} \frac{Y_i Z_i}{e(X_i)} - \frac{Y_i(1-Z_i)}{(1-e(X_i))}$$

e(X) is called the *propensity score* and corresponds to $Pr(Z = 1|X)$, in this case, the probability that an individual is a minority given the age level.

i) Write a function, `compute_adjusted_mean_diff` that takes in a data frame and computes the adjusted mean difference in expenditures between minorities and non minorities, using the IPW formula above. Within `compute_adjusted_mean_diff` you should first make sure to compute e(X) by adding the variable `e_x` using `mutate(e_x = glm(Minority ~ Age, x, family = "binomial")$fitted.values)`. This line simply estimates `e(X)` via logistic regression. We'll cover logistic regression in more detail in weeks 9 and 10.

ii). Use `bootstraps` to run `compute_adjusted_mean_diff` on each boostrapped dataframe. Plot a histogram of the boostrap distribution of mean expenditure differences.

```
## A function to compute the IPW estimate
compute_adjusted_diff <- \(x) {
...
}

## Do the boostraping
bootstrap_adj_diff <- ...

## Make the plot
bootstrap_adj_diff_plot <- ...

bootstrap_adj_diff_plot
```

**Question 10: Interpreting the bootstrap IPW results.**

Based on the bootstrap distribution, what are the reasonable range of values that the age-corrected population difference in expenditures could take? How does this compare to the

range of unadjusted values below? Does there still seem to be a difference between expenditures after adjusting for age?

**Your answer here replacing this line**

## Regression analysis

As an alternative to IPW adjustment, we can adjust for confounders via regression. In this part, we'll use a linear model to estimate the differences in median expenditure in finer detail.

More specifically, you'll model the log of expenditures (response variable) as a function of gender, age cohort, and each race/ethnicity category:

$$\log\left(\text{expend}_i\right) = \beta_0 + \underbrace{\beta_1\left(\text{6-12}\right)_i + \cdots + \beta_5\left(\text{51+}\right)_i}_{\text{age cohort}} + \underbrace{\beta_6 \text{male}_i}_{\text{sex}} + \underbrace{\beta_7 \text{hispanic}_i + \cdots + \beta_{13}\text{other}_i}_{\text{ethnicity}} + \epsilon_i$$

In this model, *all* of the explanatory variables are categorical and encoded using indicators; in this case, the linear model coefficients capture means for each group.

Because this model is a little different than the examples you've seen so far in two respects – the response variable is log-transformed and all explanatory variables are categorical – some comments are provided below on these features.

### Data preprocessing

Each coefficient represents a difference in means from the 'baseline' group. By default, the baseline category of a factor variable is the first level of that factor. It should be that the first level of the `Age Cohort` factor is "0-5", which will be your baseline. For the ethnicity, use `fct_relevel` to set the most common category, "White not Hispanic" as the baseline category.

```
dds_lm <- dds_cat |> mutate(Ethnicity = fct_relevel(Ethnicity, "White not Hispanic", after=0)
```

For a white female recipient between ages 0 and 5, all indicators in the formula above are 0, so this is the baseline group and:

$$\mathbb{E}\left(\log(\text{expend}) \mid \text{male, white, 0-5}\right) = \beta_1$$

Similarly, the expected log expenditure for a hispanic male recipient between ages 0 and 5 is:

$$\mathbb{E}\left(\log(\text{expend}) \mid \text{male, hispanic, 0-5}\right) = \beta_0 + \beta_6 + \beta_7$$

So $\beta_6 + \beta_7$ is *the difference in mean log expenditure between hispanic male and white female recipients after accounting for age.* The other parameters have similar interpretations.

You should know that the parameters represent marginal differences in means between genders (holding age and ethnicity fixed), between ages (holding gender and ethnicity fixed), and between ethnicities (holding age and gender fixed).

**Comments about the log transformation**

The response in this model is the *log* of expenditures (this gives a better model for a variety of reasons). The statistical assumption then becomes that:

$$\log(\text{expend})_i \sim N\left(\mathbf{x}_i'\beta, \sigma^2\right)$$

If the log of a random variable $Y$ is normal, then $Y$ is known as a *lognormal* random variable; it can be shown mathematically that the exponentiated mean of $\log Y$ is the median of $Y$. As a consequence, according to our model:

$$\text{median}(\text{expend}_i) = \exp\left\{\mathbf{x}_i'\beta\right\}$$

You'll work on the log scale throughout to avoid complicating matters, but know that this model for the log of expenditures is *equivalently* a model of the median expenditures.

### Question 11: model fitting

Fit the linear regression of log expenditures on Age Cohort, Gender and Race/Ethnicity.

Use the `tidy` function on the resulting `lm` object to store the parameter estimates, standard errors and p-values as a tidy dataframe called `coef_tbl`. Display the result.

```
## run lm and store result
expenditures_fit <- ...

coef_tbl <- tidy(expenditures_fit)
print(coef_tbl, n=15)
```

Now look at both the estimates and standard errors for each level of each categorical variable; if some estimates are large for at least one level and the standard errors aren't too big, then estimated mean log expenditures differ according to the value of that variable when the other variables are held constant.

For example: the estimate for `Gender_Male` is about -0.04; that means that, if age and ethnicity are held fixed, the estimated mean log expenditure is 0.04 lower for male recepients than female recipients. If $\log(a) - \log(b) = 0.04$, then $\frac{a}{b} = e^{0.04} \approx 1.041$; so the estimated expenditures (not on the log scale) differ by a factor of about 1, *i.e.*, are about the same. Further, the standard error is 0.02, so the estimate is within 2SE of 0; the difference could well be zero. So the model suggests there is no difference in expenditure by gender.

### Question 12: interpretation

Do the parameter estimates suggest differences in expenditure by age or ethnicity?

First consider the estimates and standard errors for each level of age, and state whether any differences in mean log expenditure between levels appear significantly different from zero; if so, cite one example. Then do the same for the levels of ethnicity. Answer in 2-4 sentences. How many *times larger* do we expect the *expenditures* (not log expenditures) to be for a 55+ year old compared to a 0-5 year old? **Hint:** consider exponentiation.

*Type your answer here, replacing this text.*

### Question 13.

Now fit the linear regression of log expenditures on **Age** (not Age Cohort), Gender and Race/Ethnicity. Again, use the `tidy` function on the resulting `lm` object to store the parameter estimates, standard errors and p-values as a tidy dataframe called `coef_tbl2`. Display the table. Interpret the coefficient for `Age`. Does it appear significantly different from zero?

```
## Run the regression
expenditures_fit2 <- ...

coef_tbl <- tidy(expenditures_fit2)
print(coef_tbl, n=15)
```

### Question 14: model visualization

For females only, plot the log expenditures (y) against age (x) colored by gender and faceted by race/ethnicity. Include the linear fit by using `geom_smooth`.

```
...
```

Based on this plot, comment on the pros and cons of running a regression on the continuous `Age` versus the `Age Cohort` factor?

**Your answer here replacing this text**

### Question 15: uncertainty

In the original linear model using `Age Cohort`, which estimates have greatest statistical uncertainty and why? Identify the ethnic groups for which the uncertainty band is relatively wide. Why might uncertainty be higher for these groups? Answer in 2 sentences. (it may help to refer to the previous figure.)

*Type your answer here, replacing this text.*

### Question 16: summary

Write a one-paragraph summary of your analysis. Focus on answering the question, 'do the data provide evidence of ethnic or gender discrimination in allocation of DDS funds?'

Your summary should include the following:

- a description of the data indicating observations, variables, and sampling mechanism;
- a description of any important exploratory findings;
- a description of the method you used to analyze the data (don't worry about capturing every detail);
- a description of the findings of the analysis;
- an answer to the question.

*Type your answer here, replacing this text.*