Predictive Analytics Homework 1: Inference in Linear Models

Introduction

This homework assignment will give you practice interpreting this week's concepts. A few exercises require modifying code that is provided.

Completing this assignment will help you be able to:

- Explain the difference between a sample and a population.
- Explain how a linear regression model computes a prediction.

Instructions

Getting Started

Create your own Quarto solution file. If you haven't done that before, follow the instructions given in the "Creating Documents" tutorial in Unit 0. You can switch back and forth between Source and Visual view by clicking the buttons in the top-left of the editor window; here's an example of what a document might look like in Source view (you can copy-paste this as a template):

```
---
title: "Predictive Analytics Homework 1"
author: "My Name"
engine: knitr
execute:
  echo: true
  error: true
knitr:
  opts_chunk:
    message: false
```

~~~
format:
  html:
    embed-resources: true
    code-tools: true
    code-fold: true
---

```{r}
#| context: setup
#| include: false
library(tidyverse)
library(tidymodels)
library(mosaic)
library(broom)
theme_set(theme_bw())
```

## Exercise 1

```{r}
1+1 # This is just an example; remove this code chunk.
```



```{r}
gestation_no_missing <- mosaicData::Gestation %>%
  filter(!is.na(age), !is.na(smoke)) %>%
  mutate(ever_smoked = ifelse(smoke != "never", "smoked", "never smoked"))
```

```{r}
gestation_no_missing %>% summarize(mean(age))
```

Some flaws in my colleague's statement:

- Issue 1.
- Issue 2.
~~~

The following exercises refer to the `Gestation` dataset in the `mosaicData` package. You may use the command `help("Gestation", package = "mosaicData")` to view the documentation for this dataset.

We'll use a mildly filtered version of this dataset that omits cases where `age` or `smoke` are missing. We also simplify the `smoke` variable into a binary indicator of whether they `ever_smoked`

~~~
```{r}
gestation_no_missing <- mosaicData::Gestation %>%
  filter(!is.na(age), !is.na(smoke)) %>%
  mutate(ever_smoked = ifelse(smoke != "never", "smoked", "never smoked"))
```
~~~

Exercise 1 (5pt)

~~~
```{r}
gestation_no_missing %>% summarize(mean(age))
```
~~~

```
# A tibble: 1 x 1
  `mean(age)`
        <dbl>
1        27.2
```

A colleague sees this result and says, "Ah, so women give birth when they're around 27 years old."

Exercise: Explain at least two flaws in your colleague's statement.

Exercise 2 (5pt)

The following code correctly uses bootstrap resampling to compute a 95% confidence interval for the mean.

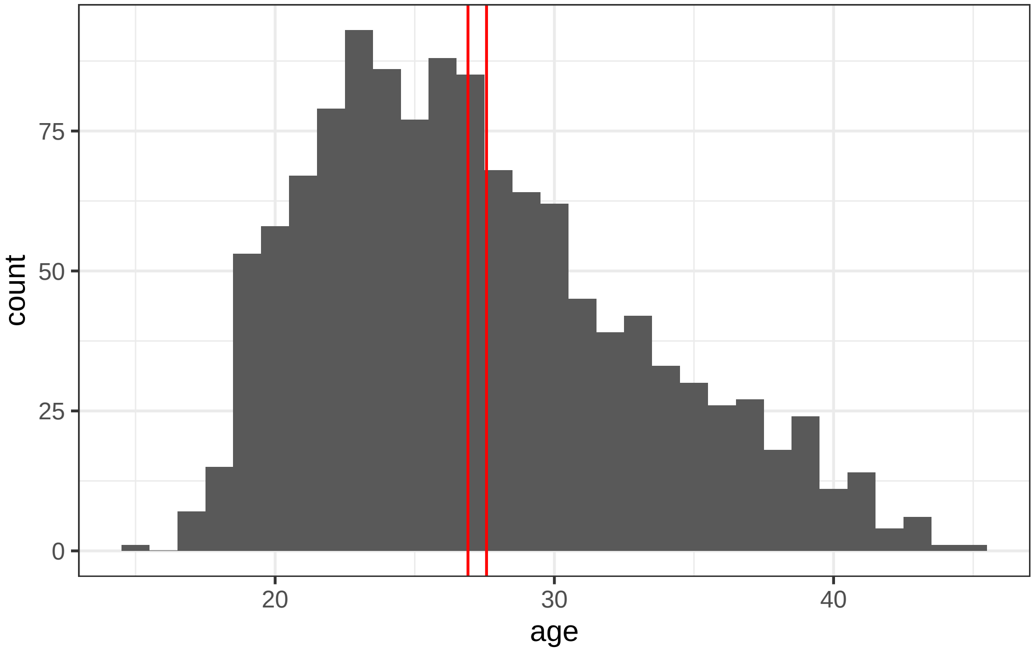
~~~
```{r cache=TRUE}
mosaic::set.rseed(123)
bootstrap <-
  mosaic::do(1000) * { # <1>
    gestation_no_missing %>% # <2>
      mosaic::resample() %>% # <2>
      summarize(mean_age = mean(age)) # <3>
  }
ci_stats <- mosaic::cdata(~ mean_age, data = bootstrap, p = 0.95) # <4>
ci_stats
```
~~~

① Repeat the following stuff 1000 times...
② Get a bootstrap resample from `gestation_no_missing`.
③ Summarize that sample by the mean of the `age` variable.
④ Compute a 95% confidence interval for the summary we computed.

```
        lower   upper central.p
2.5% 26.90592 27.5703      0.95
```

Now we plot a histogram of the data, and overlay the confidence interval that we just computed as two vertical bars.

~~~
```{r gestation-ci}
ggplot(gestation_no_missing, aes(x = age)) + geom_histogram(binwidth = 1.0) +
  geom_vline(xintercept = ci_stats$lower, color = "red") +
  geom_vline(xintercept = ci_stats$upper, color = "red")
```
~~~



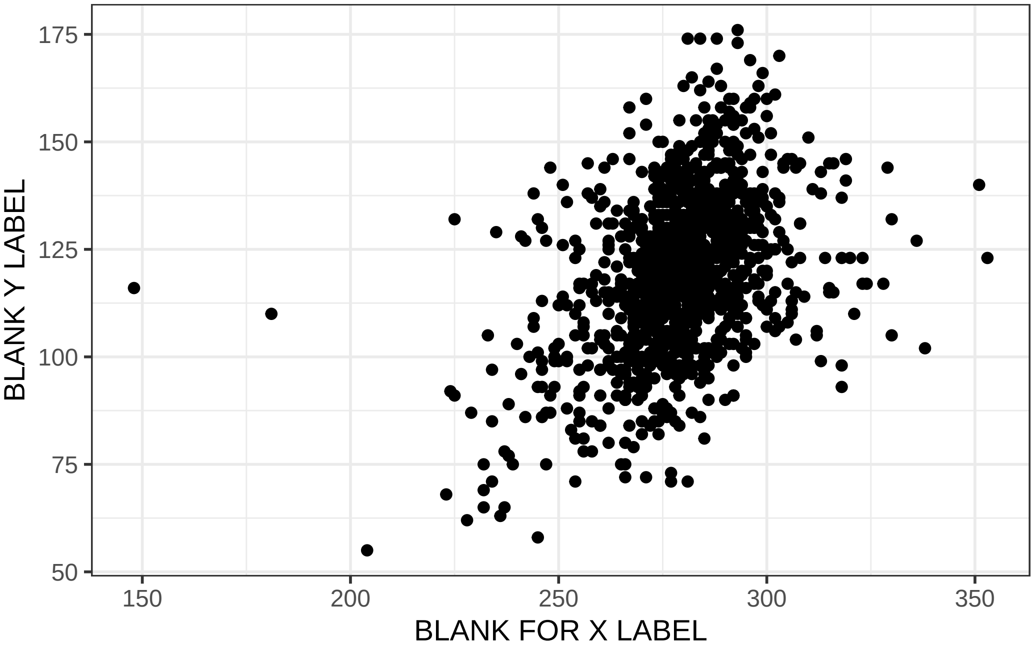
A colleague is puzzled: *why is most of the data outside of the confidence interval*?

Exercise: Explain to your hypothetical colleague why this situation actually makes sense.

Exercise 3 (12pt)

A. Here is a plot showing how birth weight (`wt`, in ounces) relates to gestation duration (`gestation`, in days). **Exercise**: The plot is unlabeled; give it meaningful labels by filling in the blanks. (2pt)

~~~
```{r}
#| label: "weight-vs-gestation"
Gestation %>%
  filter(!is.na(gestation)) %>%
  ggplot(aes(x = gestation, y = wt)) +
    geom_point() +
    labs(x = "BLANK FOR X LABEL", y = "BLANK Y LABEL")
```
~~~



B. Here is the code to fit a linear model to predict birth weight from gestation duration (using the full dataset). **Exercise**:

1. Write out (by hand, not using `equatiomatic` or the like) the equation that the fitted model uses to compute a prediction.
2. Use that equation to calculate (using a calculator, not code) the model's prediction for when the gestation duration is 250 days. (3pt)

~~~
```{r}
model <- lm(wt ~ gestation, data = Gestation)
model
```
~~~

```
Call:
lm(formula = wt ~ gestation, data = Gestation)

Coefficients:
(Intercept)    gestation
   -10.0642       0.4643
```

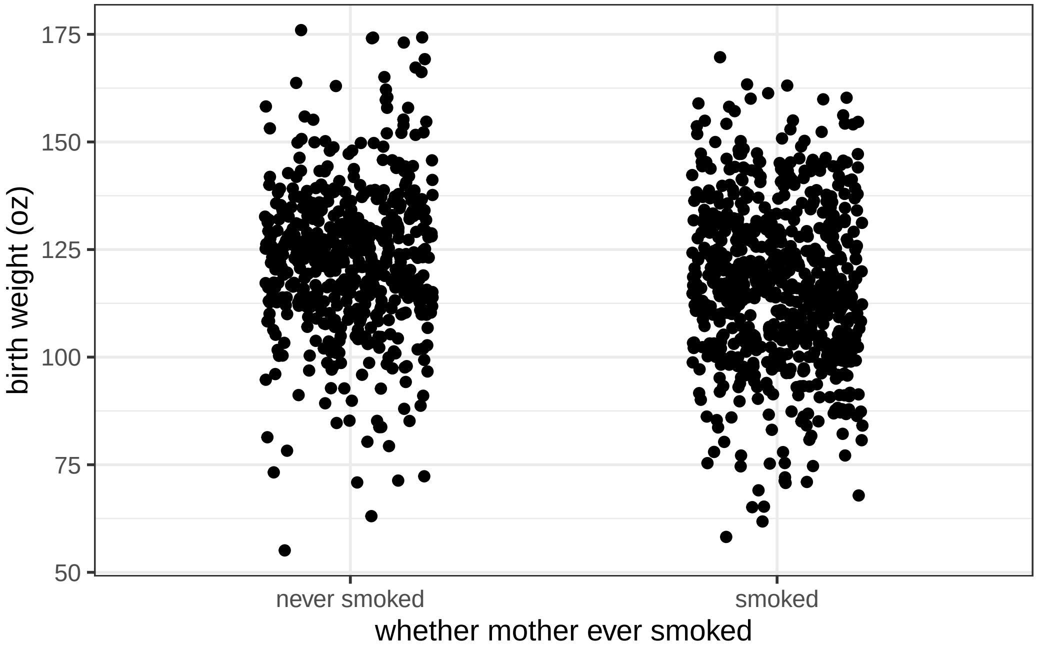
C. The following code reports a 95% confidence interval for the parameters of the model. **Exercise**: Write a concise but precise explanation of what the confidence interval for the `gestation` coefficient is telling us, to a colleague unfamiliar with confidence intervals. Make sure your explanation addresses whether the confidence interval gives evidence about whether this is a true relationship between `gestation` and some other variable, and if so, what that other variable is. (3pt)

~~~
```{r}
model %>%
  tidy(conf.int = TRUE) %>%
  select(term, estimate, conf.low, conf.high)
```
~~~

```
# A tibble: 2 x 4
  term        estimate conf.low conf.high
  <chr>          <dbl>    <dbl>     <dbl>
1 (Intercept)  -10.1    -26.4       6.26
2 gestation      0.464    0.406     0.523
```

D. Repeat parts A, B, and C, but instead for predicting birth weight from whether the mother `ever_smoked`. (The model is `wt ~ ever_smoked`. Note that the independent variable is *categorical.* For B, just write the equation; you don't need to give an example prediction.) When explaining the confidence interval, explain in particular what the confidence interval for the `ever_smoke` coefficient tells us about the relationship of smoking history and birth weight in this data: **Does maternal smoking correlate with birth weight, and if so, how large is the effect?** (4pt)

~~~
```{r weight-vs-smoke}
gestation_no_missing %>%
  filter(!is.na(age)) %>%
  ggplot(aes(y = wt, x = ever_smoked)) +
    geom_jitter(width = .2) +
    labs(x = "whether mother ever smoked", y = "birth weight (oz)")
```
~~~



Submission

To submit this assignment:

1. Use RStudio to Render your work to HTML (check that your header block includes the `embed-resources` and `code-tools` lines from the example above).
2. Check the box next to the HTML file in the `Files` pane.
3. Select **More > Export** from the menu at the top of the Files tab. You should get a file downloaded to your computer.
4. Open the file that just got downloaded and check that it displays correctly.
5. Upload the file to the Moodle assignment.