Homework 6

Brady Lamson

4/2/2022

Chapter 9 Problems

Problem 2

Calculate and interpret a 95% confidence interval for the mean age of mothers from the `Gestation` data set from the `mosaicData` package.

```
# Caclulate confidence interval given a certain number of standard deviations
# Note: 1.96 is a 95% CI, n_sds = 2 is an approximation of that.
ci_mean <- function(vector, n_sds = 2) {
    samp_mean <- vector %>% mean(na.rm = TRUE)
    samp_sd <- vector %>% sd(na.rm = TRUE)
    root_samp_size <- vector %>% length() %>% sqrt()

    upper <- (samp_mean + (n_sds * samp_sd / root_samp_size)) %>% round(digits = 3)
    lower <- (samp_mean - (n_sds * samp_sd / root_samp_size)) %>% round(digits = 3)

    glue::glue("
        The 95% confidence interval for the mean is [{lower}, {upper}]
    ")
}
```

```
# Read in data
gestation <- mosaicData::Gestation

# Calculate 95% CI
ci_mean(gestation$age, n_sds = 2)
```

```
##     The 95% confidence interval for the mean is [26.926, 27.584]
```

Based on the given sample, the population mean for the age of mothers is anywhere between 26.926 years old and 27.584 years old.

Problem 3

Use the bootstrap to generate and interpret a 95% confidence interval for the median age of mothers for the `Gestation` data set from the `mosaicData` package.

```
set.seed(100)

boot_ages <- gestation %>%
    dplyr::slice_sample(n = 500, replace = TRUE) %>%
    dplyr::select(age)

sample_med <- boot_ages$age %>% median()
se <- sd(boot_ages$age) / sqrt(length(boot_ages$age))

upper <- (sample_med + (2 * se)) %>% round(digits = 3)
lower <- (sample_med - (2 * se)) %>% round(digits = 3)


glue::glue("
    The 95% confidence interval for the median age is [{lower}, {upper}]
")
```

```
## The 95% confidence interval for the median age is [26.496, 27.504]
```

Based on the given sample, we can estimate that the median age of mothers is between 26.496 and 27.504.

Appendix E Problems

Problem 1

The statement *"Roughly 78% of the foster twins' IQ can be accurately predicted by the model"* is **FALSE**. R squared is not a measure of accuracy, it's a measure of variance. A proper statement would be *"approximately 78% of the observed variation can be explained by the model."*

Problem 3

```
# Create linear model and fit it to the data
model <- lm(wt ~ age, data = gestation)

# Get summary information on the model
model %>% broom::tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  117.       2.50       46.6  4.56e-274
## 2 age            0.106    0.0899      1.18 2.38e-  1
```

```
# Calculate 95% CI for the slope coefficient
confint(model, 'age', level = 0.95)
```

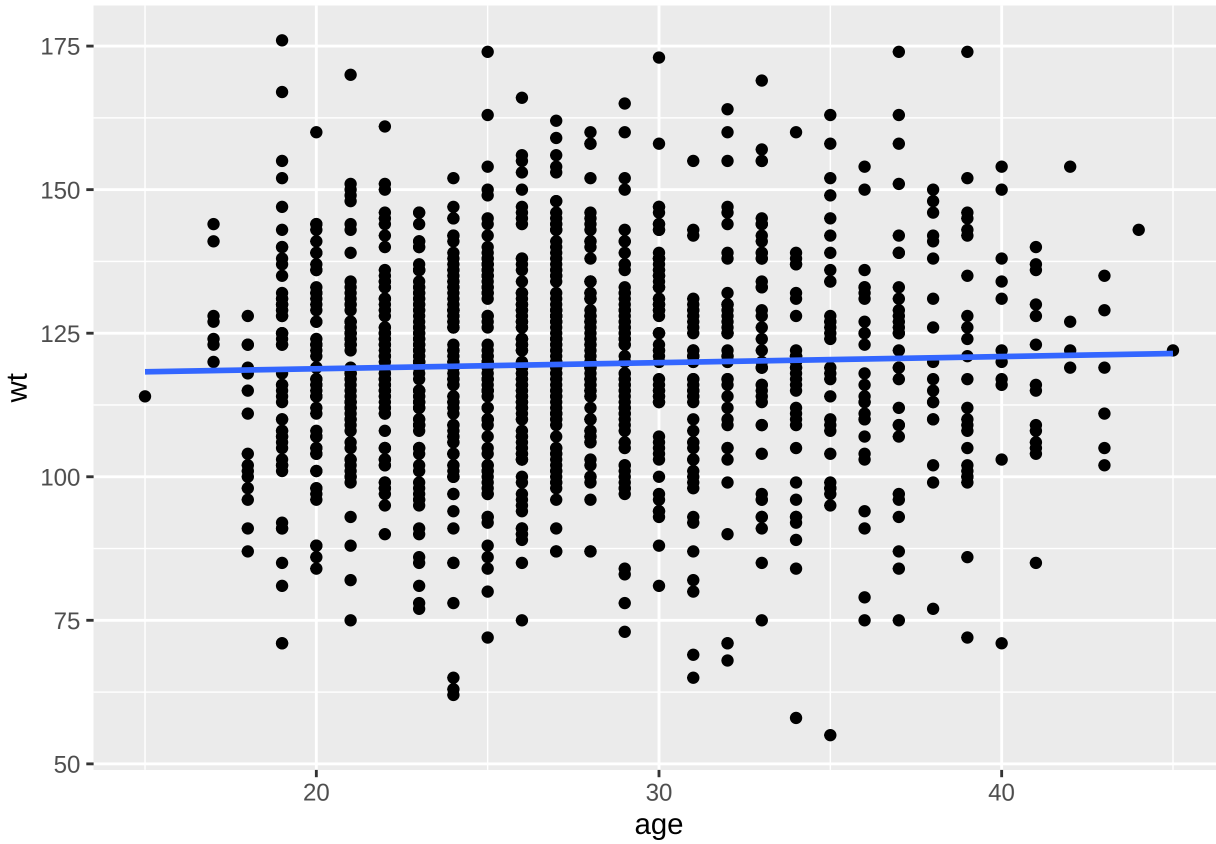
```
##           2.5 %   97.5 %
## age -0.07012632 0.282573
```

`broom` gives us the p-value for our slope coefficient. We get $2.375387 \cdot 10^{-1} = 0.238$.

`stats` has a built in way to calculate the confidence interval for the slope coefficient. The range it gives us here is $[-0.07, 0.28]$.

What we can glean from this information is that age is probably not a good predictor of weight for the specific group that we're looking at. The potential values for the slope coefficient is very low on either side of the range, and the p-value is very high. If there is a relationship, the data we have does not give us significant evidence to conclude it exists. We can examine this relationship visually with a scatterplot, and it shows the poor predictive nature of age firsthand.

```
gestation %>%
    ggplot(aes(x = age, y = wt)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE)
```



Note:

I thought this was due monday. My bad. I'm just going to turn in the partial work I did so far.