Lab 02: Ice duration and air temperature in Madison, WI

Inference for simple linear regression

Dav King

2022-09-16

Setup

Load packages and data:

```
library(tidyverse)
library(tidymodels)
library(knitr)
```

```
ice <- read_csv("data/wi-icecover.csv")
air <- read_csv("data/wi-air-temperature.csv")
```

Exercises

Exercise 1

```r
icecover_avg <- ice %>%
  group_by(lakeid, year) %>%
  summarize(avg_ice_duration = mean(ice_duration))
```

```
`summarise()` has grouped output by 'lakeid'. You can override using the
`.groups` argument.
```

There are 270 observations of 3 variables in `icecover_avg`.

Exercise 2

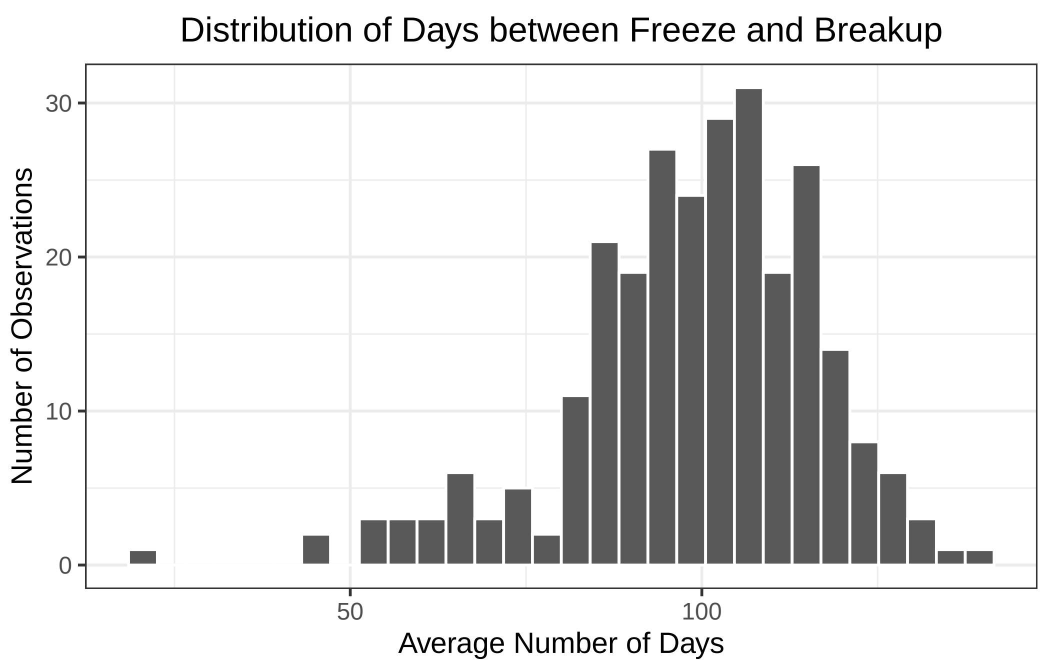
```r
airtemp_avg <- air %>%
  group_by(year) %>%
  summarize(avg_air_temp = mean(ave_air_temp_adjusted))

ice_air_joined <- inner_join(icecover_avg, airtemp_avg)
```

```
Joining, by = "year"
```

Exercise 3

```
ggplot(ice_air_joined, aes(x = avg_ice_duration)) +
  geom_histogram(color = "white") +
  theme_bw() +
  labs(x = "Average Number of Days", y = "Number of Observations",
       title = "Distribution of Days between Freeze and Breakup") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
fivenum(ice_air_joined$avg_ice_duration)
```

```
[1]  21  91 101 111 140
```

```
mean(ice_air_joined$avg_ice_duration)
```

```
[1] 99.04851
```

```r
sd(ice_air_joined$avg_ice_duration)
```

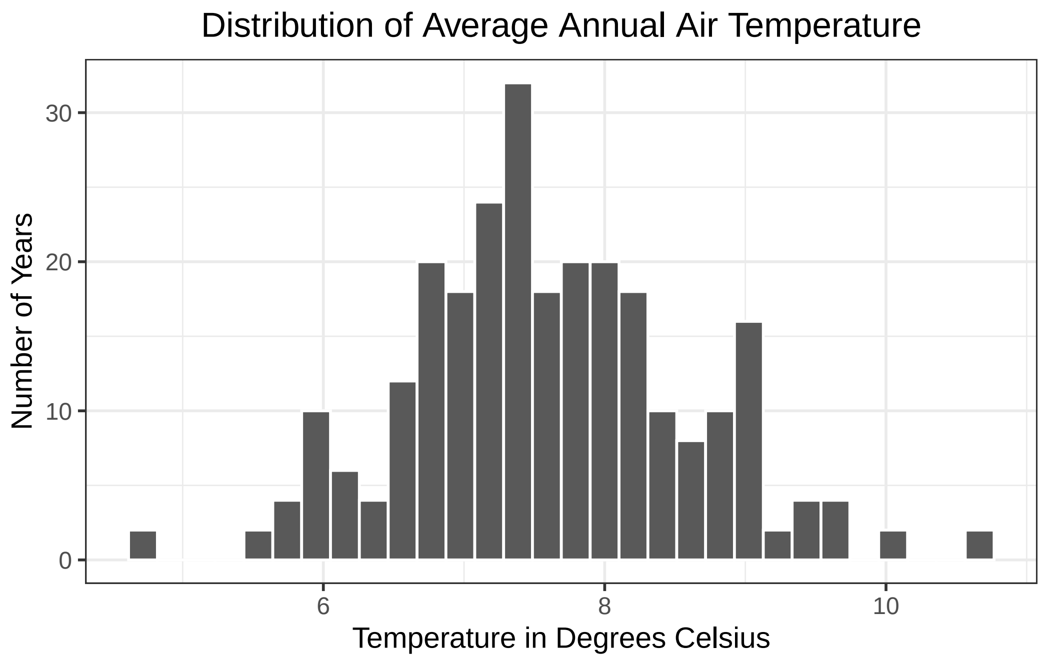
```
[1] 17.42618
```

These data are roughly unimodal and normally distributed, with a center around 100 (mean = 99.049, sd = 17.426, median = 101) and a spread from 21 to 140, with first and third quartiles at 91 and 111, respectively. There appears to be a possible low-end outlier at 21 - otherwise, the data all seem to be clustered closely with one another and there are not serious concerns about outliers.

Exercise 4

```r
ggplot(ice_air_joined, aes(x = avg_air_temp)) +
  geom_histogram(color = "white") +
  theme_bw() +
  labs(title = "Distribution of Average Annual Air Temperature",
       x = "Temperature in Degrees Celsius", y = "Number of Years") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
fivenum(ice_air_joined$avg_air_temp)
```

```
[1]  4.811507  6.977534  7.485764  8.154918 10.759836
```

```r
mean(ice_air_joined$avg_air_temp)
```

```
[1] 7.595385
```

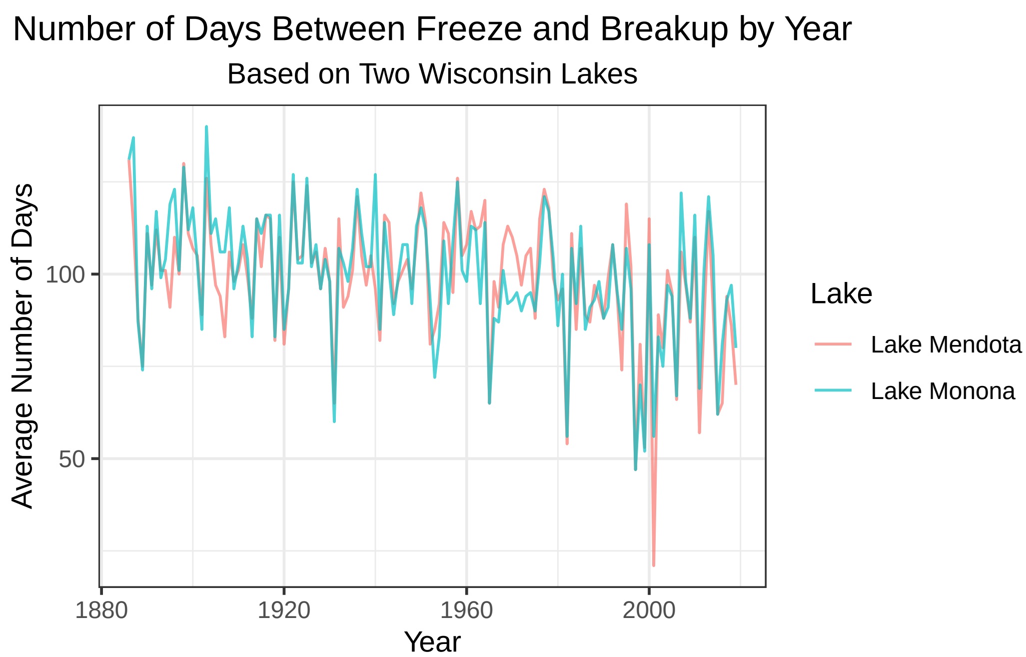
```r
sd(ice_air_joined$avg_air_temp)
```

```
[1] 0.9801526
```

These data are roughly unimodal and normally distributed, with a center around 7.5 (mean = 7.595, sd = 0.98, median = 7.486) and a spread from 4.811 to 10.760, with first and third quartiles at 6.978 and 8.155, respectively. There appear to be a few outliers below an annual average of 5 degrees or above an average of 10; otherwise, there are no notable outliers in the data.

Exercise 5

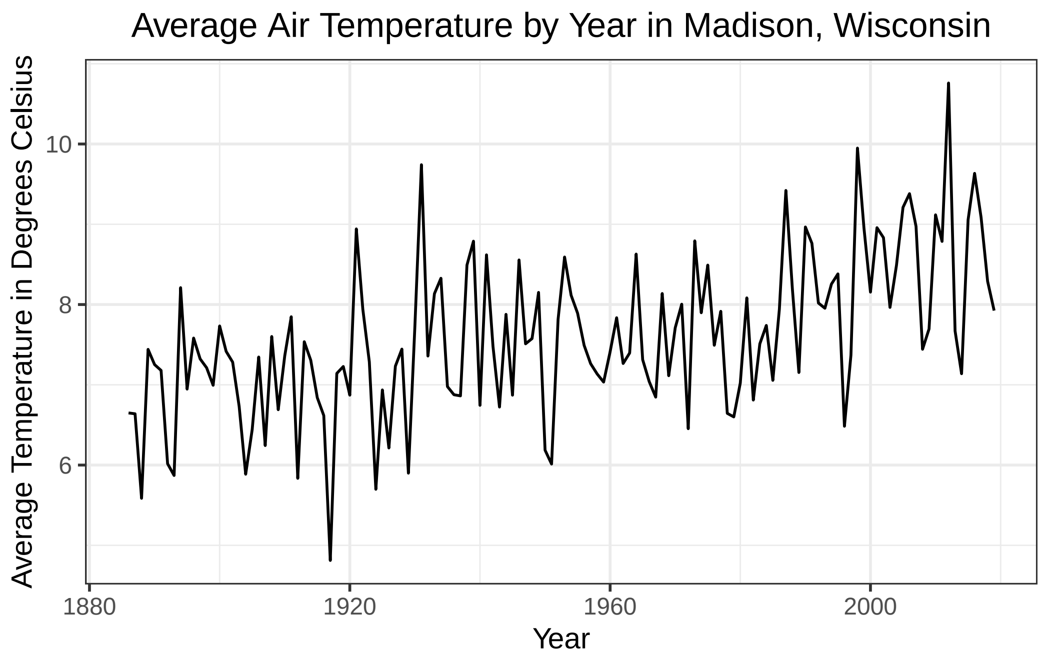
```
ggplot(ice_air_joined, aes(x = year, y = avg_ice_duration, color = lakeid)) +
  geom_line(alpha = 0.7) +
  labs(x = "Year", y = "Average Number of Days", color = "Lake",
       title = "Number of Days Between Freeze and Breakup by Year",
       subtitle = "Based on Two Wisconsin Lakes") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```



Over time, the average number of days seems to have been trending slightly lower in both of the two lakes. However, there is also a lot of variability from year to year, and the trend can only be seen over a long period of time.

Exercise 6

```
ggplot(ice_air_joined, aes(x = year, y = avg_air_temp)) +
  geom_line() +
  labs(x = "Year", y = "Average Temperature in Degrees Celsius",
       title = "Average Air Temperature by Year in Madison, Wisconsin") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



Over time, the average air temperature seems to be increasing over time. However, it still fluctuates from year to year - the trend can only be seen when looking at a long period of time.

Exercise 7

$$avg\_ice\_duration = \beta_0 + \beta_1 \times avg\_air\_temp$$

Exercise 8

```r
airModel <- linear_reg() %>%
  set_engine("lm") %>%
  fit(avg_ice_duration ~ avg_air_temp, data = ice_air_joined)

airModel %>%
  tidy() %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	149.454	7.745	19.297	0
avg_air_temp	-6.636	1.011	-6.562	0

$$avg\_ice\hat{\_}duration = 149.454 - 6.636 \times avg\_air\_temp$$

The slope of -6.636 means that, for every one degree increase in annual temperature, we would expect to see a decrease of 6.636 days between freeze and breakup for the ice on the lakes. The intercept does have a theoretically meaningful interpretation, since it is possible (though highly unlikely) to have a year with an average temperature of 0; in such a year, we would expect an average of 149.454 days between freeze and breakup.

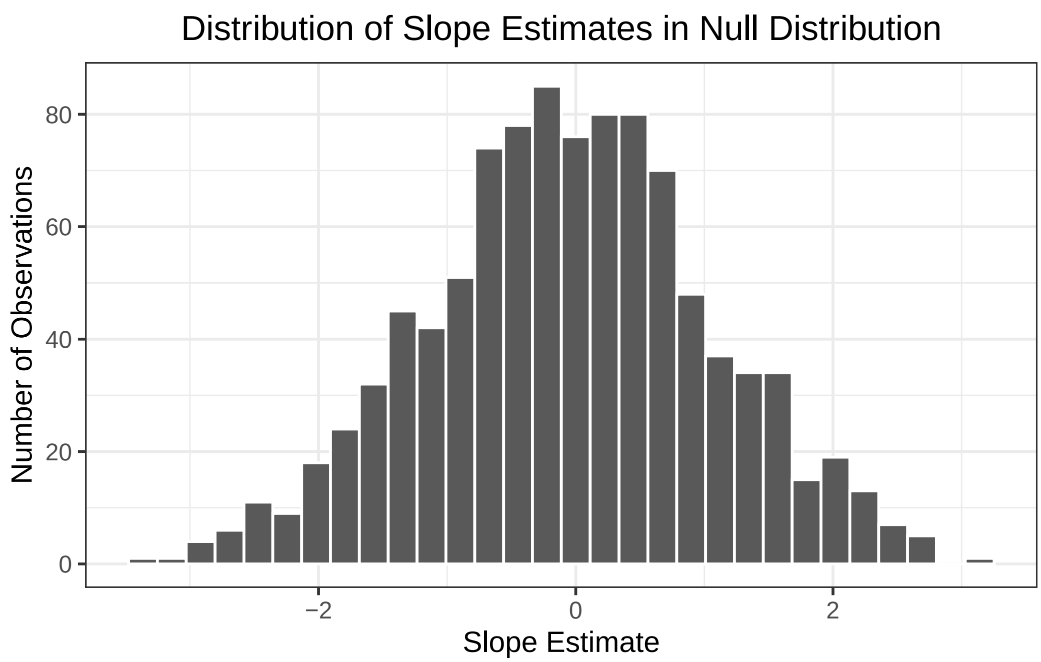
Exercise 9

$H_0 : \beta_1 = 0$, or the slope of the relationship between average air temperature and the number of days between freeze and breakup in a given year is equal to zero.

$H_a : \beta_1 \neq 0$, or the slope described above is not equal to zero.

```r
set.seed(9)

perm_fits <- ice_air_joined %>%
  specify(avg_ice_duration ~ avg_air_temp) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  fit()

perm_fits %>%
  filter(term == "avg_air_temp") %>%
  ggplot(aes(x = estimate)) +
  geom_histogram(color = "white") +
  theme_bw() +
  labs(x = "Slope Estimate", y = "Number of Observations",
       title = "Distribution of Slope Estimates in Null Distribution") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
obs <- ice_air_joined %>%
  specify(avg_ice_duration ~ avg_air_temp) %>%
  fit()

get_p_value(
  perm_fits,
  obs_stat = obs,
  direction = "two-sided"
)
```

```
Warning: Please be cautious in reporting a p-value of 0. This result is an
approximation based on the number of `reps` chosen in the `generate()` step. See
`?get_p_value()` for more information.

Warning: Please be cautious in reporting a p-value of 0. This result is an
approximation based on the number of `reps` chosen in the `generate()` step. See
`?get_p_value()` for more information.
```

```
# A tibble: 2 x 2
  term         p_value
  <chr>          <dbl>
```

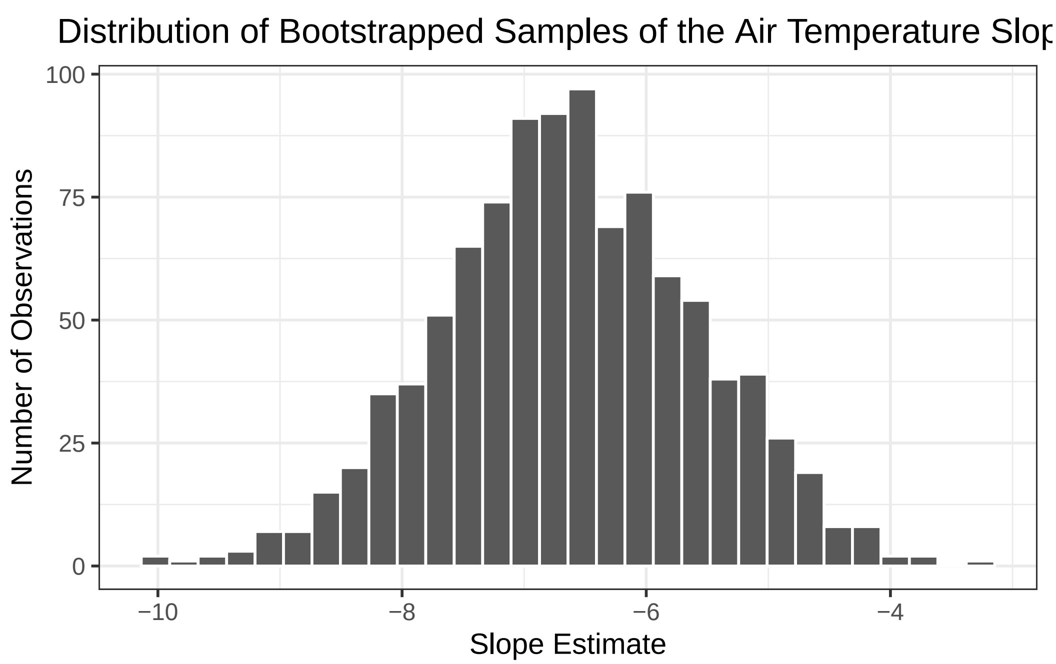
```
1 avg_air_temp       0
2 intercept          0
```

The p-value of getting a statistic like ours if there is no relationship between the two variables is ~0. Thus, we can reject H_0. We have sufficient evidence to suggest that there is a significant relationship between average air temperature and the number of days between freeze and breakup in any given year.

Exercise 10

```r
set.seed(10)

nullCI <- ice_air_joined %>%
  specify(avg_ice_duration ~ avg_air_temp) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  fit()

nullCI %>%
  filter(term == "avg_air_temp") %>%
  ggplot(aes(x = estimate)) +
  geom_histogram(color = "white") +
  theme_bw() +
  labs(x = "Slope Estimate", y = "Number of Observations",
   title = "Distribution of Bootstrapped Samples of the Air Temperature Slope")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
get_confidence_interval(
  nullCI,
  point_estimate = obs,
  level = .95,
  type = "percentile"
)
```

```
# A tibble: 2 x 3
  term         lower_ci upper_ci
  <chr>           <dbl>    <dbl>
1 avg_air_temp    -8.62    -4.64
2 intercept      134.     165.
```

The confidence interval says that we can be 95% confident that the true population slope of the relationship between average air temperature and average ice duration lies on the interval (-8.62, -4.64). This is consistent with the results of the hypothesis test because the confidence interval does not contain zero - if it did, we would not be able to reject the null hypothesis that there is no relationship between the two variables.