AE 04: Simulation-based hypothesis testing

Houses in Duke Forest

Sep 14, 2022

```r
library(tidyverse)
library(tidymodels)
library(openintro)
```

```
Warning: package 'openintro' was built under R version 4.1.3
```

```
Warning: package 'airports' was built under R version 4.1.3
```

```
Warning: package 'cherryblossom' was built under R version 4.1.3
```

```
Warning: package 'usdata' was built under R version 4.1.3
```

```r
library(knitr)
```

Data

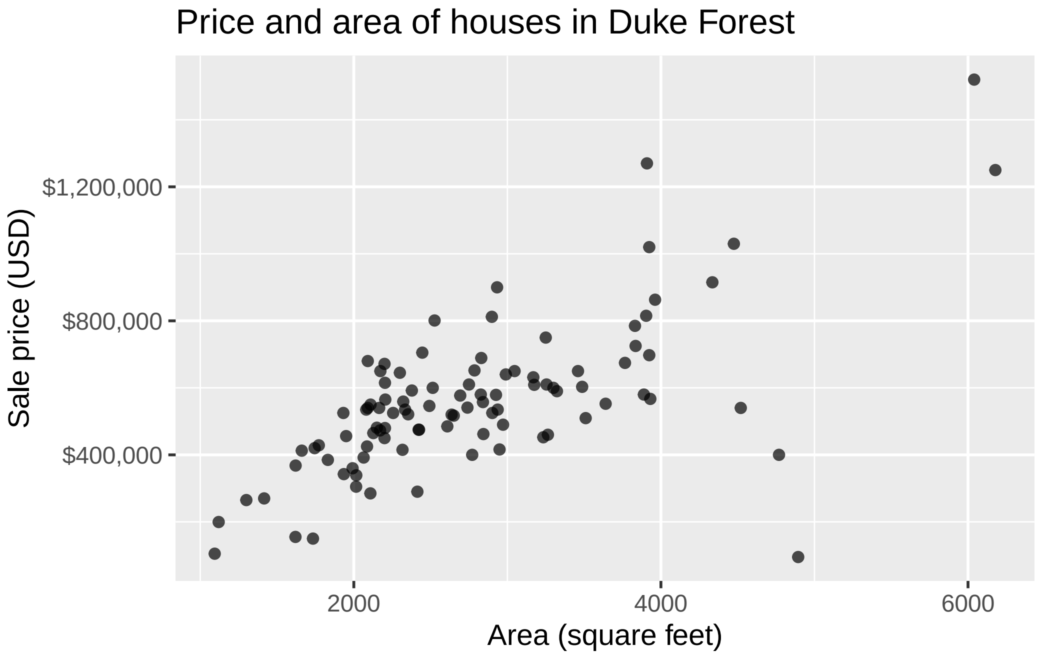
Today's dataset is about houses that were sold in the Duke Forest neighborhood of Durham, NC around November 2020. The data were originally scraped from Zillow, and can be found in the `duke_forest` data set in the **openintro** R package.

```r
glimpse(duke_forest)
```

```
Rows: 98
Columns: 13
$ address    <chr> "1 Learned Pl, Durham, NC 27705", "1616 Pinecrest Rd, Durha~
$ price      <dbl> 1520000, 1030000, 420000, 680000, 428500, 456000, 1270000, ~
$ bed        <dbl> 3, 5, 2, 4, 4, 3, 5, 4, 4, 3, 4, 4, 3, 5, 4, 5, 3, 4, 4, 3,~
$ bath       <dbl> 4.0, 4.0, 3.0, 3.0, 3.0, 3.0, 5.0, 3.0, 5.0, 2.0, 3.0, 3.0,~
$ area       <dbl> 6040, 4475, 1745, 2091, 1772, 1950, 3909, 2841, 3924, 2173,~
$ type       <chr> "Single Family", "Single Family", "Single Family", "Single ~
$ year_built <dbl> 1972, 1969, 1959, 1961, 2020, 2014, 1968, 1973, 1972, 1964,~
$ heating    <chr> "Other, Gas", "Forced air, Gas", "Forced air, Gas", "Heat p~
$ cooling    <fct> central, central, central, central, central, central, centr~
$ parking    <chr> "0 spaces", "Carport, Covered", "Garage - Attached, Covered~
$ lot        <dbl> 0.97, 1.38, 0.51, 0.84, 0.16, 0.45, 0.94, 0.79, 0.53, 0.73,~
$ hoa        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ url        <chr> "https://www.zillow.com/homedetails/1-Learned-Pl-Durham-NC-~
```

Exploratory data analysis

```r
ggplot(duke_forest, aes(x = area, y = price)) +
  geom_point(alpha = 0.7) +
  labs(
    x = "Area (square feet)",
    y = "Sale price (USD)",
    title = "Price and area of houses in Duke Forest"
  ) +
  scale_y_continuous(labels = label_dollar())
```



Model

```r
df_fit <- linear_reg() |>
  set_engine("lm") |>
  fit(price ~ area, data = duke_forest)

tidy(df_fit) |>
  kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	116652.33	53302.46	2.19	0.03
area	159.48	18.17	8.78	0.00

Hypothesis test for the slope

> Tip
>
> For code chunks with fill-in-the-blank code, change code chunk option to `#| eval: true` once you've filled in the code.

State the null and alternative hypotheses

$$H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0$$

Generate null distribution using permutation

Fill in the code, then set `eval: true` .

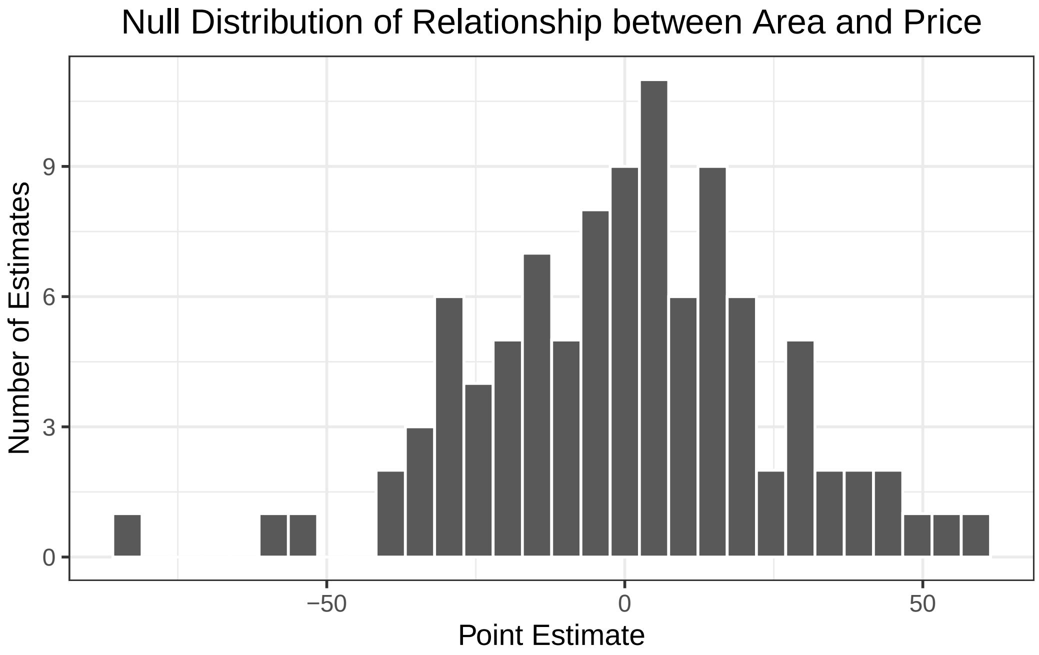
```r
n = 100
set.seed(09142022)

null_dist <- duke_forest |>
  specify(price ~ area) |>
  hypothesize(null = "independence") |>
  generate(reps = n, type = "permute") |>
  fit()
```

Visualize distribution

```r
null_dist %>%
  filter(term == "area") %>%
  ggplot(aes(x = estimate)) +
  geom_histogram(color = "white") +
  theme_bw() +
  labs(x = "Point Estimate", y = "Number of Estimates",
       title = "Null Distribution of Relationship between Area and Price") +
  theme(plot.title = element_text(hjust = 0.5))
```

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



Calculate the p-value.

```
# get observed fit
observed_fit <- duke_forest |>
  specify(price ~ area) |>
  fit()

# calculate p-value
get_p_value(
  null_dist,
  obs_stat = observed_fit,
  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 area            0
2 intercept       0
```

- What does the warning message mean?

It means that you should take caution in the reporting of your results - while the p-value of 0 is a rounded approximation, the probability of us getting a statistic like this given that no such relationship exists is not actually zero (but rather something very close to it).

State conclusion

A test comparing our observed statistic to the null distribution of relationships between area and price yielded a p-value of ~0. Thus, we reject H_0. We have sufficient evidence to say that there is a significant relationship between the area of a house and its price.

! Important

To submit the AE:

- Render the document to produce the PDF with all of your work from today's class.
- Push all your work to your `ae-04-` repo on GitHub. (You do not submit AEs on Gradescope).