SMML Project 1

Kevin Linares (klinares@umd.edu) and Jamila Sani (jsani@umich.edu)

09 September, 2024

1. Use hprice in faraway package

- The data includes 324 observations coming from 36 US metropolitan statistical areas (MSAs) over 9 years from 1986-1994: 36 × 9 = 324
- Assume that the MSAs in the data are a simple random sample of the population of MSAs in the US. See https://www2.census.gov/geo/maps/metroarea/us_wall/Mar2020/CBSA_WallMap_Mar2020.pdf for MSAs.
- Refer to the R manual for faraway for the background information about this dataset as well as variable definitions.
- The housing sale price is the outcome variable of interest. Since the data set has a natural log transformed price variable, narsp, we recode this to create **homeprice** by transforming **narsp** back to the dollar unit for an easier interpretation as follows:

```
data("hprice") # read in data from pacakge

hprice |> summary() # print summary statistics for each variable
```

```
##      narsp            ypc            perypc          regtest      rcdum
##  Min.   :3.920   Min.   :12535   Min.   :-2.054   Min.   :13.00   0:279
##  1st Qu.:4.264   1st Qu.:16609   1st Qu.: 3.535   1st Qu.:18.00   1: 45
##  Median :4.412   Median :18454   Median : 3.964   Median :20.00
##  Mean   :4.484   Mean   :18769   Mean   : 4.268   Mean   :20.42
##  3rd Qu.:4.575   3rd Qu.:20323   3rd Qu.: 5.711   3rd Qu.:22.00
##  Max.   :5.563   Max.   :33383   Max.   : 8.788   Max.   :29.00
##
##  ajwtr        msa           time
##  0:189   1      :  9   Min.   :1
##  1:135   2      :  9   1st Qu.:3
##          3      :  9   Median :5
##          4      :  9   Mean   :5
##          5      :  9   3rd Qu.:7
##          6      :  9   Max.   :9
##          (Other):270
```

```
?faraway::hprice# view variable information
```

1. What are the mean and the variance of homeprice? What do they mean?

- The average house sale price from 1986-1994 was $94,411, with a very high variance due to a few houses being sold for a lot more than the average.

```
hprice <- hprice |>
  as_tibble() |>
  mutate(homeprice = exp(narsp)*1000)
```

```r
hprice |>
  summarise(mean_homeprice = mean(homeprice),
            var_homeprice = var(homeprice))
```

```
## # A tibble: 1 x 2
##   mean_homeprice var_homeprice
##            <dbl>         <dbl>
## 1         94411.   1583110349.
```

2. Construct a 95% confidence interval of the average homeprice. What does the confidence interval imply?

```r
CI_output <- hprice |>
  summarise(t_score = qt(p=.05/2, df=n()-1, lower.tail=F),
    se = sd(homeprice) / sqrt(n()), # calculate standard error
    lower = round(mean(homeprice) - t_score*se, 2),
    upper = round(mean(homeprice) + t_score*se), 2)

CI_output
```

```
## # A tibble: 1 x 5
##   t_score    se  lower upper   `2`
##     <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1    1.97 2210. 90063. 98760     2
```

- The confidence interval of house prices (mean $94,411 [95% CI 90062.7, 98760]) implies that if we were to take repeated samples from the same population with n sample sizes, we would expect the true price to be contained in this interval 95% of the time.

3. Estimate the average homeprice by whether the MAS was adjacent to a coastline, noted in ajwtr, and the standard errors.

- Houses near the coast line had an average sale price of $111,243 (se = $4,656) vs non-coastal houses $82,389 (se = $1,229).

```r
hprice <- hprice |>
  mutate(ajwtr = ifelse(ajwtr == 1, "Coastal", "Non-Coastal")) # give labels to the numeric values


hprice |>
  group_by(ajwtr) |>
  reframe(se = sd(homeprice) / sqrt(n()), # reframe ungroups the previous function, returns summary
          sample_mean = mean(homeprice))
```

```
## # A tibble: 2 x 3
##   ajwtr          se sample_mean
##   <chr>       <dbl>       <dbl>
## 1 Coastal     4656.     111243.
## 2 Non-Coastal 1229.      82389.
```

4. Test the difference in homeprice between coastline MSAs and non-coastline MSAs. Clearly state the formula for the hypothesis, the test method and your rationale for selecting the method. What do you conclude about the hypothesis?

- We will conduct a two sample right-tailed t-test with unknown σ^2, and we hypothesize that home prices are higher in coastal areas than non-coastal forming the basis of our alternative hypothesis. Conroy and Molisch 2009 found that coastal houses are twice the cost of comparable homes 9 miles away from the coastal line, suggesting that waterfront houses are more appealing to home buyers thus more expensive.
  - Our Null hypothesis is that non-coastal house prices are higher than coastal.

$$H_0 : \mu_{ajwtr=1} - \mu_{ajwtr=0} =\leq 0$$
$$H_A : \mu_{ajwtr=1} - \mu_{ajwtr=0} => 0$$

- First, we examine the assumption of equal variances.
  - Our f-test's p-value is less than .05, therefore we cannot reject the Null hypothesis, and the variances between coastal and non-coastal metropolitan areas are not equal. We can also use the confidence intervals to test the ratio of the two variances, and we can see that 1 is not in the $95% CI; therefore, the variances are not equal and we can account for this in our two sample t-test.

```r
var.test(homeprice ~ ajwtr, hprice, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  homeprice by ajwtr
## F = 10.257, num df = 134, denom df = 188, p-value < 0.00000000000000022
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   7.520273 14.107151
## sample estimates:
## ratio of variances
##           10.25681
```

- We reject the null hypothesis in favor of the alternative hypothesis that coastal houses probably have higher prices than non-coastal houses.
  - In our one sided right tailed test, the upper confidence interval bound is a naturally Inf.

```r
t.test(homeprice ~ ajwtr, var.equal = FALSE, hprice, alt="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  homeprice by ajwtr
## t = 5.9922, df = 152.79, p-value = 0.000000007148
## alternative hypothesis: true difference in means between group Coastal and group Non-Coastal is grea
## 95 percent confidence interval:
##  20885.34      Inf
## sample estimates:
##     mean in group Coastal mean in group Non-Coastal
##                 111242.96                  82388.89
```

5. Estimate the Pearson correlation coefficient between homeprice and per capita income of the MSA of a given year, noted in ypc.

- Our correlation of $r = .74$ indicates a strong association between home prices and per capita income, and is positive meaning that as one increases the other variable increases as well.

```r
cor(hprice$homeprice, hprice$ypc)
```

```
## [1] 0.7437474
```

6. Test whether the correlation coefficient between homeprice and ypc is 0 or not. Clearly state the hypothesis including the formula. What do you conclude?

- We will perform a Pearson's correlation coefficient to test our alternative hypothesis that these two variables are associated. Our Null hypothesis is that the correlation =0.

$$H_0 : r = 0$$
$$H_A : r \neq 0$$

- We reject the null hypothesis in our Pearson's correlation coefficient in favor of our alternative hypothesis suggesting that there is evidence that home prices and per capita income have a positive linear relationship of $.74(95\%[.69, .79])$.

```r
cor.test(hprice$homeprice, hprice$ypc)
```

```
##
##  Pearson's product-moment correlation
##
## data:  hprice$homeprice and hprice$ypc
## t = 19.965, df = 322, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6907661 0.7887854
## sample estimates:
##       cor
## 0.7437474
```

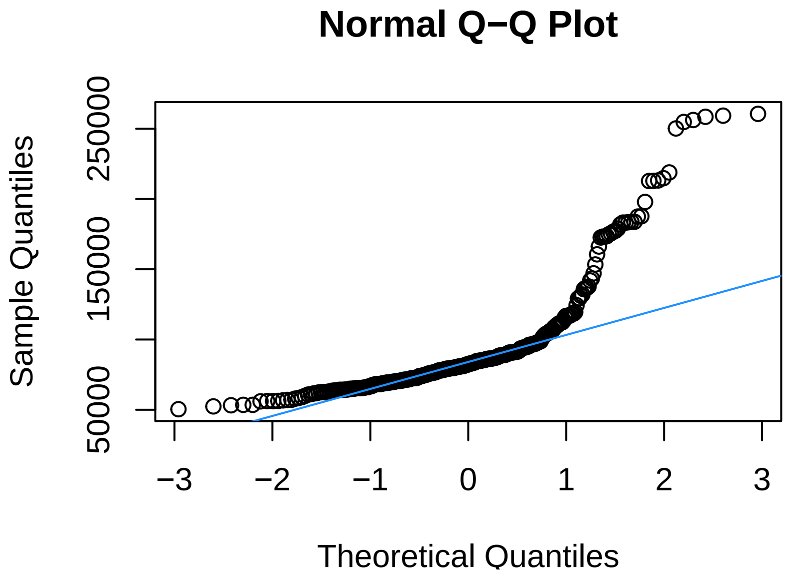
7. Can you say that per capita income has an effect on the home sales price using the results from #6)? Why or why not?

- We cannot conclude that one variable has an effect, nor in the causal sense, on the other at this point until we conduct an OLS regression with home sales regressed on per capita income.

8. Test the normality of homeprice. Would this test result change your responses to #1) to 7)? Why or why not?

- Home prices appears to be skewed to the right according to our qq plots. Additionally, we used the Shapiro-Wilk normality test and rejected the null hypothesis that the data are normally distributed in favor of the alternative hypothesis that they are probably not normal.
- We can conclude that reporting the mean is not robust to the skewness found in the data and would want to reconsider our t-test to include a Mann Whitney U test used to test sample distributions that are not normally distributed.

```r
qqnorm(hprice$homeprice)
qqline(hprice$homeprice, col="dodgerblue") # add a blue line
```



```
shapiro.test(hprice$homeprice)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  hprice$homeprice
## W = 0.7327, p-value < 0.00000000000000022
```