# Project 1

Sol and Keqing lu

2023-09-05

- Note: for each question, be sure to include both R code and output that is pertinent to your answer(s).
- Use hprice in the faraway package.
- The data include 324 observations from 36 US metropolitan statistical areas (MSAs) over 9 years from 1986-1994: 36 x 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 (on Canvas - see Module 1) for background information about this dataset as well as variable definitions.
- The housing sale price is the outcome variable of interest. Because the dataset has a natural log transformed price variable, narsp, we will recode this to create a variable called "homeprice" by transforming narsp back to the dollar unit for an easier interpretation as follows:

```
hprice$homeprice <- exp(hprice$narsp)*1000
```

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

```
mean_homeprice <- mean(hprice$homeprice)
variance_homeprice <- var(hprice$homeprice)
mean_homeprice
```

```
## [1] 94411.42
```

```
variance_homeprice
```

```
## [1] 1583110349
```

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

```
confidence_interval <- qt(0.975, df=length(hprice$homeprice)-1) * (sd(hprice$homeprice) / sqrt(length(hp
lower_bound <- mean_homeprice - confidence_interval
upper_bound <- mean_homeprice + confidence_interval
c(lower_bound, upper_bound)
```

```
## [1] 90062.70 98760.14
```

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

```
coastline_means <- tapply(hprice$homeprice, hprice$ajwtr, mean)
coastline_se <- tapply(hprice$homeprice, hprice$ajwtr, function(x) sd(x) / sqrt(length(x)))
coastline_means
```

```
##         0         1
##  82388.89 111242.96
```

```
coastline_se
```

```
##        0        1
## 1228.660 4655.883
```

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

```
#use t test to determine if mean homeprice in coastline and non-coastline MSAs is the same
#null- mean homeprice of coastal cities= mean homeprice of non-coastal cities

#test equal variance
#null- Variance of homeprice in coastal cities= variance of homeprice in non costal cities
var.test(homeprice ~ ajwtr, hprice, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  homeprice by ajwtr
## F = 0.097496, num df = 188, denom df = 134, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.07088604 0.13297389
## sample estimates:
## ratio of variances
##         0.09749617
```

```
#reject null

#use t-test
#null- mean homeprice of costal cities= mean homeprice of non-coastal cities
t_test_result <- t.test(homeprice ~ ajwtr, data = hprice)
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  homeprice by ajwtr
## t = -5.9922, df = 152.79, p-value = 1.43e-08
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
```

```
## 95 percent confidence interval:
##  -38367.19 -19340.96
## sample estimates:
## mean in group 0 mean in group 1
##        82388.89       111242.96
```

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

```
correlation <- cor(hprice$homeprice, hprice$ypc)
correlation
```

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

6. Test whether the correlation coefficient between homeprice and ypc is 0, or not. Clearly state the null hypothesis being tested and include the formula.

```
cor_test_result <- cor.test(hprice$homeprice, hprice$ypc)
cor_test_result
```

```
##
##  Pearson's product-moment correlation
##
## data:  hprice$homeprice and hprice$ypc
## t = 19.965, df = 322, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6907661 0.7887854
## sample estimates:
##       cor
## 0.7437474
```

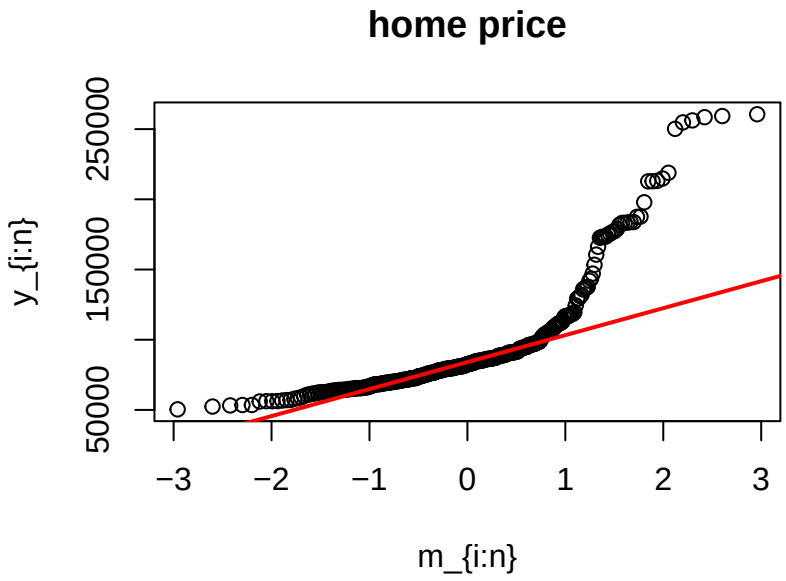
7. Can you say that per capita income has an effect on the home sales price using the results from #6? Why or why not?

```
if (cor_test_result$p.value < 0.05) {
  cat("Yes, per capita income has a significant effect on home sales price.")
} else {
  cat("No, per capita income doesn't have a significant effect on home sales price.")
}
```

```
## Yes, per capita income has a significant effect on home sales price.
```

8. Test the normality of homeprice. Would this change your responses to questions 1-7? Why or why not?

```
library(ggplot2)
qqnorm(hprice$homeprice, main="home price", ylab="y_{i:n}", xlab="m_{i:n}")
qqline(hprice$homeprice, col="red",lwd=2)
```




#homeprice does not follow a normal distribution, therefore answers 1-7 are not accurate. The sample means are innacurate to the population mean