module5\_Estes

Andrew Estes

9/25/2021

1. This first problem will have you explore the AmesHousing data set in order to discover why we had to limit our use of data to homes built on or after 1950.
2. First, let’s look at what we would have found if we hadn’t looked only at homes built after 1950. Write some code to • load the AmesHousing data set into R and create the data frame ames, • create a new column called Remodeled, as we did in the lecture notes, that will be true if Year\_Built != Year\_Remod\_Add, • sample 500 data points from the ames data frame, using set.seed(248) to make sure we all get the same sample, and • run an independent-sample t-test to test whether the means of the remodeled and nonremodeled groups are different.

library(AmesHousing)

## Warning: package 'AmesHousing' was built under R version 4.1.1

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.3 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.0 v forcats 0.5.1

## Warning: package 'tidyr' was built under R version 4.1.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
library(tinytex)  
Ames <- make\_ames()  
  
Ames.50 <- Ames %>%  
 filter(Ames$Year\_Built >= 1950)  
  
Ames2 <- Ames %>%   
 mutate(Remodeled = Ames$Year\_Built != Ames$Year\_Remod\_Add)  
  
set.seed(248)  
t.test(Sale\_Price ~ Remodeled, data=Ames2)

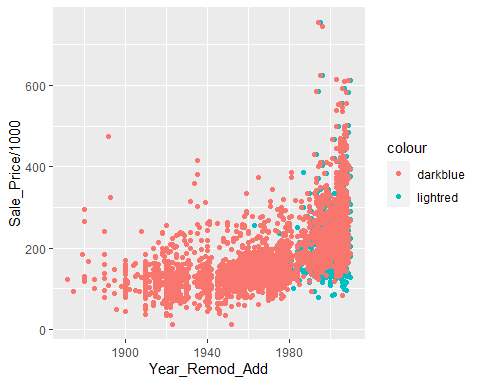
##   
## Welch Two Sample t-test  
##   
## data: Sale\_Price by Remodeled  
## t = 2.5372, df = 2726.7, p-value = 0.01123  
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0  
## 95 percent confidence interval:  
## 1719.841 13421.938  
## sample estimates:  
## mean in group FALSE mean in group TRUE   
## 184312.8 176741.9

1. Compare the results of this t-test to those shown in the lecture notes. There should be one aspect that’s markedly different.

#The mean group in TRUE is very different: 176k vs 206k

1. What’s going on? Make a plot that helps you explore the data. Then write a brief explanation of why the results differ, depending on whether homes built before 1950 are included in the data set. [Hint: There might be lots of ways to do this, but a scatter plot involving Year\_Remod\_Add and Year\_Built might work well.]

oneC <-   
 ggplot() +  
 geom\_point(data=Ames.50, aes(Sale\_Price/1000, Year\_Remod\_Add, color="lightred")) +  
 geom\_point(data=Ames2, aes(Sale\_Price/1000, Year\_Built, color="darkblue"))  
oneC + coord\_flip()



#A recently remodeled home offers more value than a home that was remodeled 70+ years ago  
#New houses are priced similarly to recently remodeld homes

1. In this problem, we’ll look at simulating the results of an independent-sample t-test in order to gain some insight about sample sizes.
2. Complete the code stub below to write code to generate two sets of random, normallydistributed data, with the sample sizes, means and standard deviations as specified in the code. Then perform an independent-sample t-test using your two samples as the two groups to be compared. [Note: we’re purposefully not going to use set.seed here.]

n1 <- 200  
n2 <- 200  
mu1 <- 15  
mu2 <- 17  
sigma1 <- 8  
sigma2 <- 8  
  
first <- rnorm(n1, mean = mu1, sd = sigma1)  
second <- rnorm(n2, mean = mu2, sd = sigma2)  
t.test(first, second, alternative = "two.sided", paired = FALSE, var.equal = FALSE)

##   
## Welch Two Sample t-test  
##   
## data: first and second  
## t = -3.9182, df = 397.79, p-value = 0.000105  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -4.860058 -1.612484  
## sample estimates:  
## mean of x mean of y   
## 14.20690 17.44317

t.test(first, second, alternative = "two.sided", paired = TRUE, var.equal = FALSE)

##   
## Paired t-test  
##   
## data: first and second  
## t = -3.9611, df = 199, p-value = 0.0001039  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -4.847404 -1.625138  
## sample estimates:  
## mean of the differences   
## -3.236271

1. Suppose your simulated data represents the results of a trial of a new blood pressure medication. Group 1 represents the decrease in blood pressure for a group taking the current “best in class” medication, while Group 2 represents the decrease in blood pressure for a group taking a new competitor’s drug. Interpret the results of the test you did in a way that communicates accurately, but without unnecessary complexity. Make sure that you refer to the specifics of this real-world scenario. [Assume α = .05 so that you can get a specific Yes/No answer to the test.]

#Comparing the two samples, there is a statistically significant difference between the two groups.  
#With a p-value of .01485, it is a strongly correlated.   
#Group 1 is not as successful, on average, at reducing the BP as Group 2

1. This question picks up where the last left off. If you run your code from Question 2 several times (click the green arrow on the code chunk several times), you should notice that sometimes you get a p-value that would reject the null hypothesis, and sometimes you don’t. Furthermore, remember that there are two ways that a hypothesis test might be in error: • Type I Error: Rejecting H0 when H0 is really true. Prob(Type I) = α. • Type II Error: Failing to reject H0 when H0 is really false. Prob(Type II) = β. We can control the probability of committing a Type I error by controlling the significance level, α, that we choose for the test. On the other hand, calculating and controlling β is harder. Even calculating β requires an assumption about what we think the true alternative really is. Not just μ2 − μ1 ≠ 0, but, for example, μ2 − μ1 = 2. With that assumption, it is possible to calculate β. This can be done theoretically, or—as we’ll see—through simulation.
2. By setting mu1 <- 15 and mu2 <- 17 in the code above, we’ve actually made the assumption in this simulation that the true population means are not equal—in fact, they differ by 2. Therefore, to get an estimate of β, all we need to do is to run the simulations and hypothesis test many times, and look at the percentage of times that we fail to reject the null hypothesis, even though we know that the null hypotheses is false in our simulation. Run your simulation and t-test code many times (maybe at least 20 times?), and calculate the percentage of those runs where you fail to reject the null hypothesis. Then fill in the sentence below: “Under the assumption that the new drug lowers blood pressure by 2 mmHg more than its competitor, I estimate the the probability of failing to detect that difference is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.”

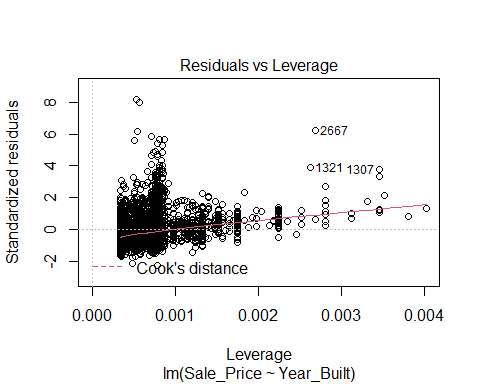
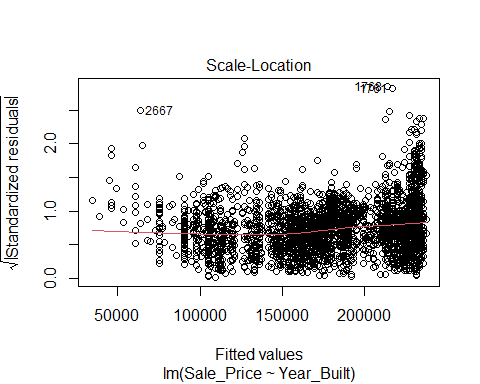
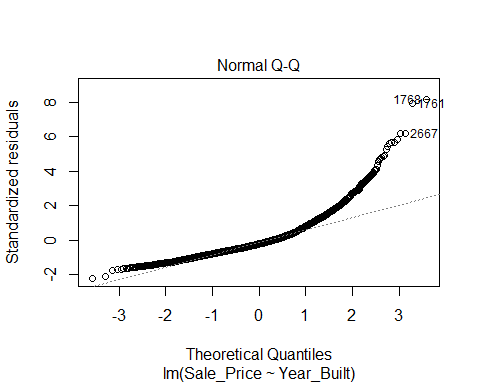
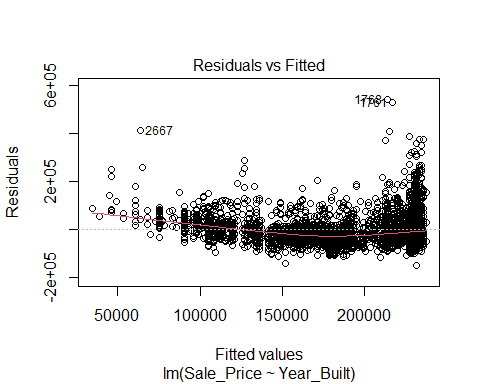
#“Under the assumption that the new drug lowers blood pressure by 2 mmHg more  
#than its competitor, I estimate the probability of failing to detect that difference is  
#70% (ran 20 times, 14 results were not statistically significant)

1. Suppose you find this probability of Type II error to be too high in the scenario outlined above. One way to decrease it would be to raise your sample size. By experimenting with changing n1 and n2, find the minimum sample sizes necessary to reduce your probability of Type II error to 5%. [You might try n = 100, n = 200, etc. as a start. Note: The power of a test is defined to be 1 − β, so we’re really finding the sample size that would give a power of 95% for this test.]

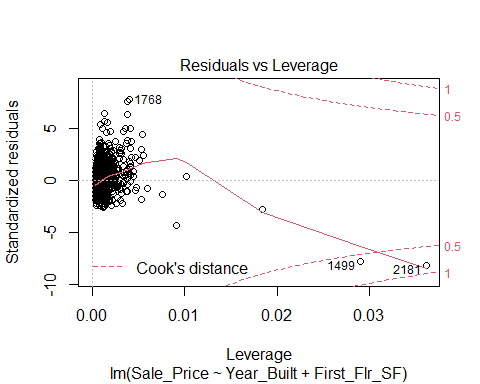
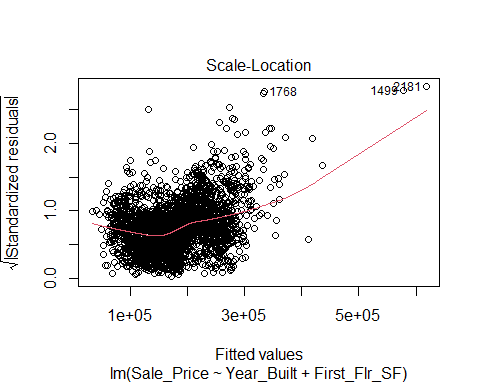
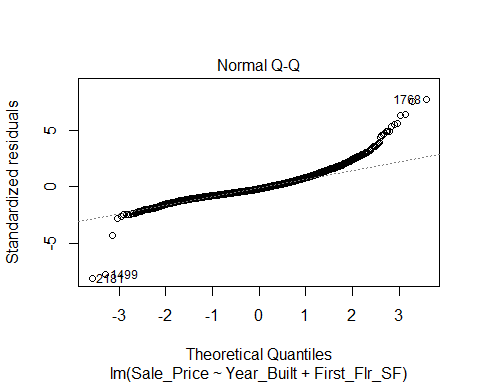
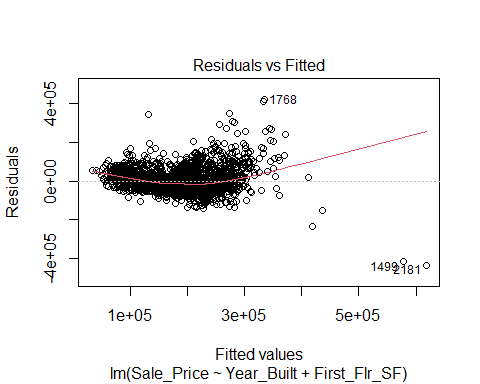
#Changed the N value to 300, then to 100, then to 200

1. We saw that taking the log transform of Sale\_Price was one way to try to get control of the variance of residuals in our simple linear regression. Another way to try to improve the model’s fit and the model assumptions is to add further important variables to the model. Often bad fit is the result of confounding from lurking variables. In this problem, start with the first multiple linear regression model we looked at in the lecture. It’s named fit.original in the code chunk below. Then do the following:
2. Write code to plot residuals vs. fit for fit.original (try to find the right plot option to print out only that graph).

fit.original <- lm(Sale\_Price ~ Year\_Built + First\_Flr\_SF, data=Ames)  
fit.sale.year <- lm(Sale\_Price ~ Year\_Built, data=Ames)  
plot(fit.sale.year)



plot(fit.original)



1. Write code to output the R2-adjusted for fit.original.

summary(fit.original)

##   
## Call:  
## lm(formula = Sale\_Price ~ Year\_Built + First\_Flr\_SF, data = Ames)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -434097 -32813 -9218 23910 420116   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.042e+06 6.819e+04 -29.95 <2e-16 \*\*\*  
## Year\_Built 1.068e+03 3.505e+01 30.48 <2e-16 \*\*\*  
## First\_Flr\_SF 1.011e+02 2.705e+00 37.39 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 54540 on 2927 degrees of freedom  
## Multiple R-squared: 0.5343, Adjusted R-squared: 0.5339   
## F-statistic: 1679 on 2 and 2927 DF, p-value: < 2.2e-16

#adjusted R-Squared = 0.5339  
#RSE = 54540

1. Think about which variables which variables might also be highly related to Sale\_Price. Then create a new model using those variables that has a higher R2-adjusted and seems to satisfy the linearity and constant variance conditions better. Output the plot of fit vs. residuals and the adjusted r-squared for this new model as well. There’s no single right answer here—for now you’re just playing around to see what you’ll find.

fit.andrew <- lm(Sale\_Price ~ Year\_Remod\_Add + First\_Flr\_SF,data=Ames)  
summary(fit.andrew)

##   
## Call:  
## lm(formula = Sale\_Price ~ Year\_Remod\_Add + First\_Flr\_SF, data = Ames)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -455250 -32649 -3139 23457 420483   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.030e+06 9.730e+04 -31.14 <2e-16 \*\*\*  
## Year\_Remod\_Add 1.556e+03 4.938e+01 31.51 <2e-16 \*\*\*  
## First\_Flr\_SF 1.067e+02 2.629e+00 40.59 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 54090 on 2927 degrees of freedom  
## Multiple R-squared: 0.5419, Adjusted R-squared: 0.5416   
## F-statistic: 1731 on 2 and 2927 DF, p-value: < 2.2e-16

#adjusted R-Squared = .5416  
#RSE = 54090  
plot(fit.andrew)

