EEP/IAS 118 - Introductory Applied Econometrics Problem Set 3

Problem Set 3, Spring 2024, Villas-Boas

Deadline: See Gradescope

DSP students. Given we must release solutions sufficiently before the MT, we are only permitting a shorter extension than usual. DSP deadline is March 10, Sunday, 11:59 AM (midday, not midnight).

Submit materials as **one pdf** on Gradescope. After uploading the pdf to Gradescope, please **assign appropriate pages to each question**. Questions that do not have assigned pages on Gradescope may not be graded. Codes and outputs not properly displayed will be marked as incorrect.

Before submitting, make sure that all code cells are run with all output fully visible, and **do not print the entire dataset in your submission**. If you viewed the data earlier, remove that line of code and re-run the code cell (as datasets get bigger this adds many pages to pdf submissions and increases the likelihood we miss your answer).

*Note: Coding Bootcamp Part 3 and Part 4 covers all necessary R methods.

Preamble

When writing R code, it's a good habit to start your notebooks or R scripts with a preamble, a section where you load all necessary packages, set paths or change the working directory, or declare other options.

Use the below code cell to load in packages you will use throughout the problem set (at least `haven`, `tidyverse`, and `ggplot2` this week).

In [43]:
```
library(haven)
library (tidyverse)
library(ggplot2)
```

Exercise 1: Relationship between Housing Prices and Characteristics of Locations

This exercise is to be completed using R. We will establish a simple linear relationship between **housing prices and characteristics of location** in the

sample. This is called a hedonic regression, relating price to characteristics. The idea is that if a characteristic is valued in a region, like good schools, demand for housing increases as people move there, and then housing prices increase, all else constant. Vice versa, if people do not value a characteristic, like crime.

When there is a clean-up in former hazardous waste locations, what happens to the prices of houses nearby? This problem set explores individuals' willingness to pay (WTP) for environmental quality by estimating whether increasing environmental quality impacts housing prices in the area. For this problem set, the change in environmental quality results from the cleanup of hazardous waste sites. Consider Superfund sites, areas designated by the US government as contaminated by hazardous waste that pose a hazard to environmental/human health. The EPA placed certain Superfund sites on the National Priorities List (npl=1), which meant that these sites were legally required to undergo remediation. If individuals value environmental quality, then housing prices should increase after nearby NPL sites are cleaned up. To determine the extent to which individuals value the cleanup, we can compare housing close to a hazardous waste site that was cleaned up to "comparable" homes near a similar waste site that was not cleaned up.

Data description

The data include observations on census tracts within 2 miles of a hazardous waste site (Census tracts are statistical subdivisions of a county defined by the Census Bureau to allow comparisons from census to census). This includes census tracts where the site in question was on the NPL, and thus was legally required to be cleaned up, as well as those that were not on the NPL. The shared dataset contains the following variables for each census tract in the year 2000:

Variable name	Definition
fips	Federal Information Processing Standards (FIPS) census tract identifier
npl	A binary indicator for whether the site in a given census tract was placed on the NPL!
price	Median housing value (in USD.) -- called housing price, henceforth
ba_or_better	% of the population that has a Bachelors degree or higher
unemprt	Unemployment rate
povrat	Poverty rate

1. (i) Read the data into R using `mydata <- read_dta("dataPset3_2024.dta")`. (ii) Create a new variable, by computing

the natural log of housing price, and call it *lprice*. (iii) How many observations are there in the data?

In [44]:
```
mydata <- read_dta("dataPset3_2024.dta")
mydata$lprice <- log(mydata$lprice)
t_obs <- nrow(mydata)

t_obs
print(" Total number of observations is 445")
```

```
445
[1] " Total number of observations is 445"
```

2. Report (a) the minimum, (b) the median, (c) the sample mean, and (d) the maximum, using `summary()` for the variables lprice, price, povrate, unemprt, and npl.

(Hint: see the section on summary() in Coding Bootcamp Part 1)

In [45]:
```
summary(mydata)
```

```
    fips                npl          ba_or_better        unemprt       
Length:445         Min.   :0.0000   Min.   :0.02228   Min.   :0.02242  
Class :character   1st Qu.:0.0000   1st Qu.:0.08063   1st Qu.:0.05174  
Mode  :character   Median :1.0000   Median :0.11526   Median :0.06784  
                   Mean   :0.6899   Mean   :0.13417   Mean   :0.07539  
                   3rd Qu.:1.0000   3rd Qu.:0.16911   3rd Qu.:0.09582  
                   Max.   :1.0000   Max.   :0.49853   Max.   :0.25319  
    povrat            price            lprice      
Min.   :0.01836   Min.   : 40552   Min.   :2.362  
1st Qu.:0.05460   1st Qu.: 83401   1st Qu.:2.428  
Median :0.09160   Median :117068   Median :2.457  
Mean   :0.10833   Mean   :130036   Mean   :2.456  
3rd Qu.:0.14311   3rd Qu.:159378   3rd Qu.:2.483  
Max.   :0.44887   Max.   :535170   Max.   :2.579  
```

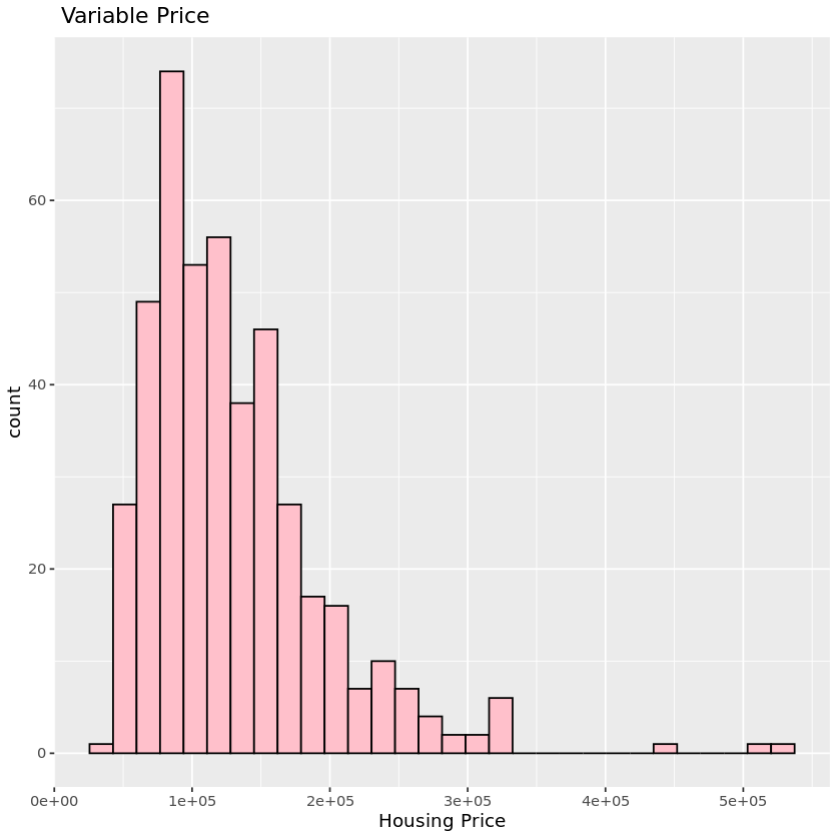
3. Create a histogram for the variable price.

Remember to title and label everything.

Hint: see the Histograms section of Coding Bootcamp Part 4

In [46]:
```
ggplot(data=mydata,aes(x=price)) + geom_histogram(fill = "pink", color = "bl
    labs (title = " Variable Price", x= "Housing Price")
```

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



4. Using the dataset, we now compare housing across the two groups (group **npl= 1** and **npl= 0**). How many census tracts have been cleaned up (that is, npl=1), and how many census tracts have npl=0?

In [47]:
```
summarise(group_by(mydata,npl), count =n())
# 138 have 0 and 307 have 1
```

A tibble: 2 × 2

npl	count
<dbl>	<int>
0	138
1	307

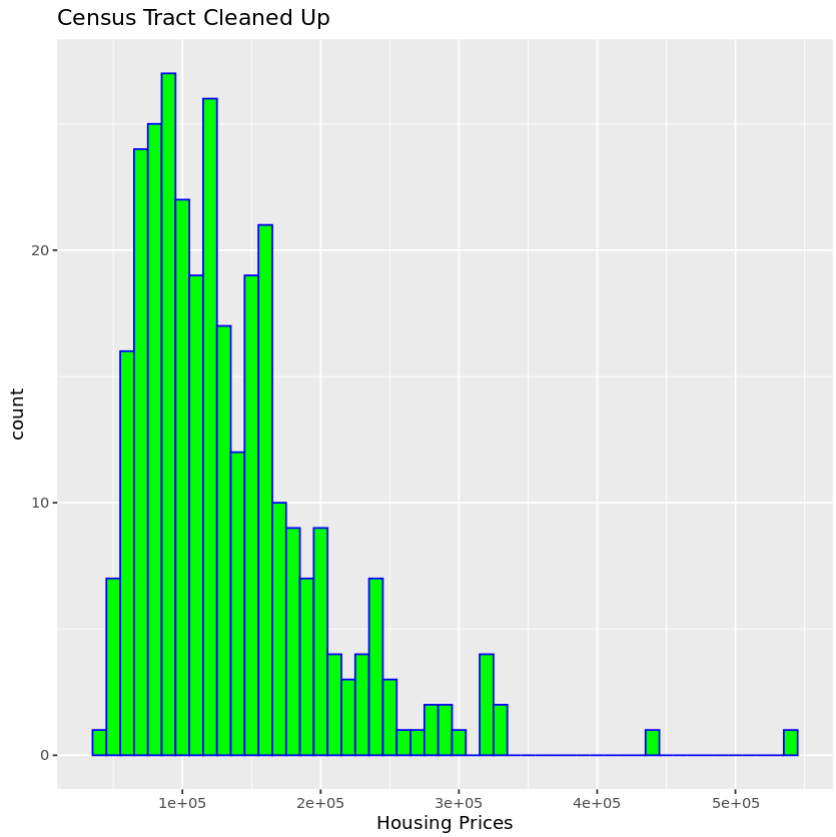
The number of census tracts that have been cleaned up is 307 and 138 have not been cleaned up.

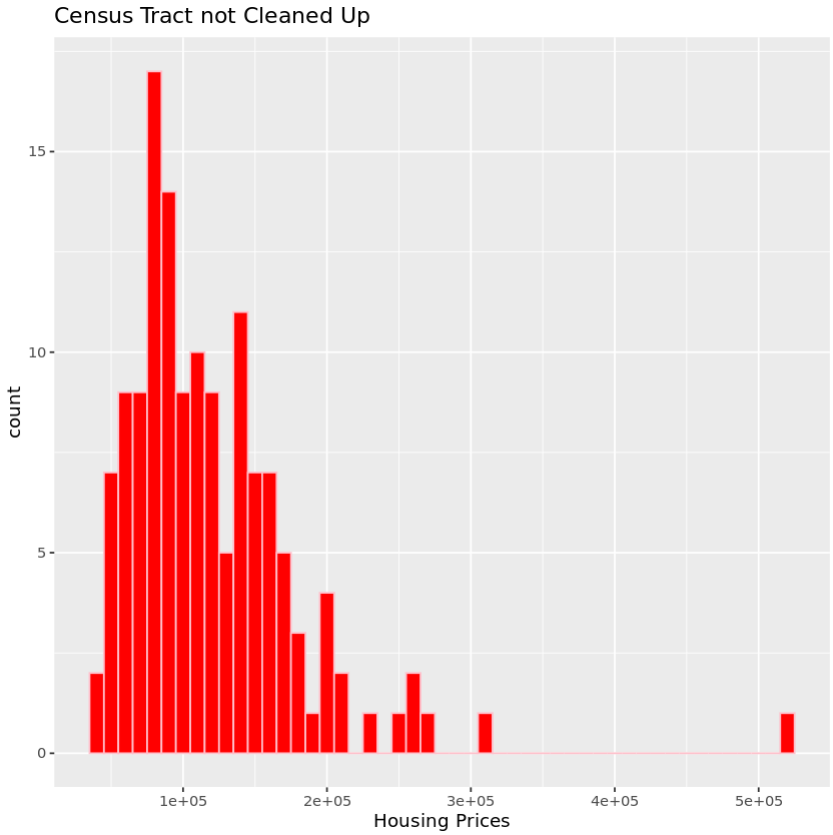
5. Draw separate histograms of the housing price in each group of census tracts with and without NPL.

Hint: see the Histograms section of Coding Bootcamp Part 4

In [48]:
```
# With NPL
npl_data <- subset(mydata, npl == 1)
ggplot(data = npl_data, aes(x=price)) +
  geom_histogram(binwidth=10000, color="blue", fill="green") +
```

```
    labs(title = "Census Tract Cleaned Up ",
      x = "Housing Prices", y = "count")
# Histogram for data with NPL



#Without NPL
withoutnpl_data <- subset(mydata, npl== 0)
ggplot(data = withoutnpl_data, aes(x=price)) +
  geom_histogram(binwidth=10000, color="pink", fill="red") +
  labs(title = "Census Tract not Cleaned Up",
       x = "Housing Prices", y = "count")
```

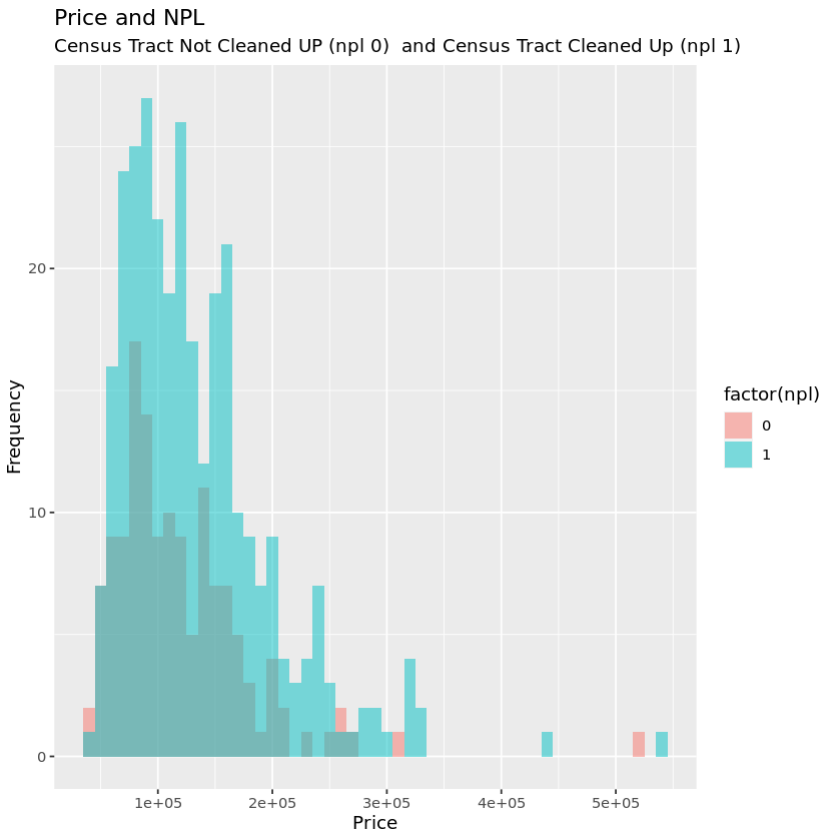




6. Overlap both histograms into the same graph and comment on differences (be precise - and explain why the differences intuitively make sense).

Hint: see the "Stacking/Multiple Histograms" section of Coding Bootcamp Part_4

In [49]:
```
ggplot(data = mydata, aes(x = price)) +
    geom_histogram(aes(fill = factor(npl),), position = "identity",

                   alpha = 0.5, binwidth = 10000) +

     labs(title = "Price and NPL",

          subtitle = "Census Tract Not Cleaned UP (npl 0)  and Census Tract

          x = "Price",

          y = "Frequency")
```



My first observation is that the histograms are right skewed. However NPL = 1 is less rightly skewed than NPL = 0. This makes sense because the homes that are on the priority list are taken care of first. Most of the houses that are not cleaned up makes the price of those houses lower. We can see that those houses that have been cleaned up have a higher price, as we can see them right skewed compared to NPL = 0, since those houses with NPL = 0 are not on the priority list to get cleaned up.

7. Construct a 99% confidence interval for the mean housing price for census tracts listed on the NPL (**npl == 1**), clearly writing out the estimates of the three components required for building the relevant CI. Give an interpretation of these results in a sentence.

Hint: use `mean()` *and* `sd()` *to get the necessary information to construct the CI*

In [50]:
```
#mean of housing price
mean(npl_data$price)

#99% confidence interval
sample.length <- length(npl_data$price)
sample.sd <- sd(npl_data$price)
sample.mean <- mean(npl_data$price)
sample.se <- sample.sd/sqrt(sample.length)

degrees.freedom = sample.length - 1
```

```
t.score = qt(p = .01/2, df = degrees.freedom, lower.tail = F)

margin.error <- t.score * sample.se

margin.error
lower.bound <- sample.mean - margin.error
upper.bound <- sample.mean + margin.error

print(c(lower.bound,upper.bound))
# Printed in order is mean of npl, marin error and lower and upper bound for
```

```
134371.818314841
9842.3819138578
 [1] 124529.4 144214.2
```

In [51]:
```
sample.mean
```

```
134371.818314841
```

We use the mean of 134,371.82 and the marin of error of 9,842.38 are used to make our interval. In other words, we are 99% confident that the population parameter of the true median housing price for NPL =1 is in between the interval of 124,529.44 and 144,214.20. Our 99% confidence interval for the mean housign price for census tract clean ups on the NPL have the bounds of 124,529, for the lower end and 144,214.20 for the upper bound.

8. Let D be the difference in population mean of prices between the **NPL=1** and **NPL=0** groups. State an estimator $\hat{D}$ for D and use the estimator to compute an estimate of D. Compute a standard error for $\hat{D}$. Derive a 90% confidence interval for D and interpret in one sentence.

In [52]:
```
mean_without_npl <- mean(withoutnpl_data$price)
sd_without_npl <- sd(withoutnpl_data$price)
Dhat <- sample.mean - mean_without_npl

npl_n <- nrow(npl_data)
without_npl_n <- nrow(withoutnpl_data)
se <- sqrt((sample.sd^2 / npl_n) + (sd_without_npl^2) / without_npl_n)

a = 0.1
c <- qnorm(1- a/2)
moe <- c * se
lower_bound <- Dhat - se
upper_bound <- Dhat + se
conf_int_90 <- c(lower_bound, upper_bound)

Dhat
se
conf_int_90
# printed in order is our D hat, standard error for d hat and our confidence
```

```
13981.1550449137
```

6468.54248856672
7512.61255634694 · 20449.6975334804

This means that there is a 90% probability that this random interval catches the true difference in the mean housing price value between the houses that have been cleaned up versus not cleaned up. In other words, the confidence interval of [7512.61255, 20449.6975] tells us explicitly that we are 90% confident that this is the difference parameter that is within the interval.

9. Using the data, test whether the average of the housing price for the **NPL=1** group is statistically different at the 1% significance level (α = 0.01) from average housing price in the **NPL=0** group (that is, in terms of the hypothesis, the null is equal, and the alternative is not equal).

Explicitly write out the 5-step procedure for hypothesis testing.

In [53]:
```
#called in order: Average Differnce, Standard Error, Total Value, Critical V

averag_diff <-Dhat
averag_diff #called

standard_e <- se
standard_e #called

t <- averag_diff/standard_e
t #called

a <- .01
df <-nrow(npl_data)+nrow(withoutnpl_data) -2

c_val <- qt(1-(a/2), df)
c_val #called
```

13981.1550449137
6468.54248856672
2.16140731387722
2.58697276325338

Step 1:

$H_0 : \mu_0 = \mu_1$
$H_1 : \mu_0 \neq \mu_1$

Step 2:

$t_{(307+138-2)} = t = \frac{\bar{y_0} - \bar{y_1}}{se(\bar{y_0} - \bar{y_1})} = \frac{13981.1550}{6468.5425}$ = 2.1614

Step 3:

We have a 1% significance level with $\alpha = 0.01$. This corresponds to the critical value for df = 443 is of 2.5870

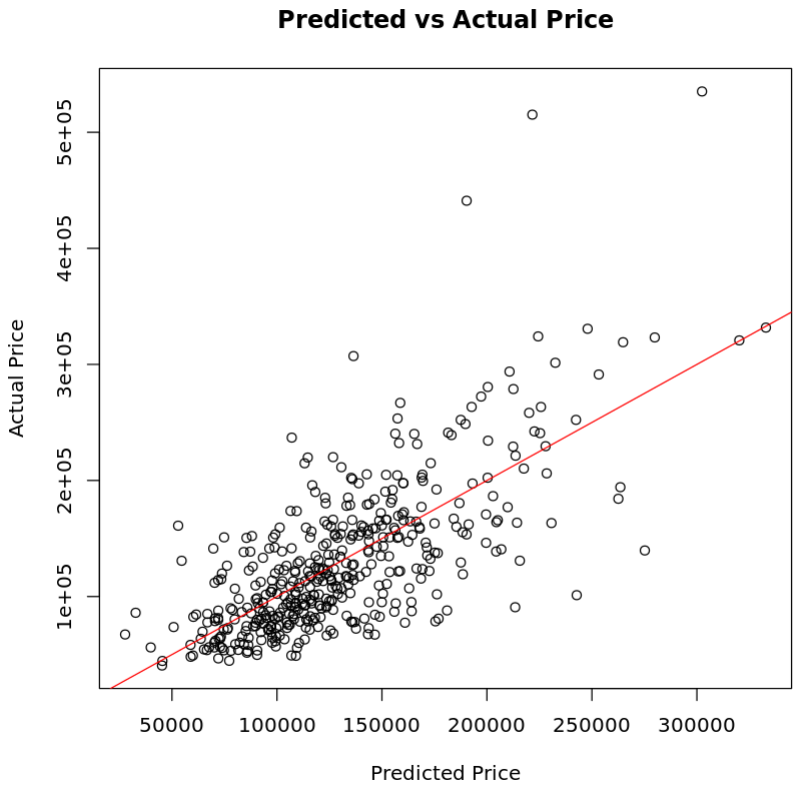
Step 4:
Our t-value of 2.1614 and our critical value(c_value in the code) of 2.5870, we can state that t < c. In other words, this means we fail to reject the null hypothesis at the 1% significance level because t < c. This means that we perfer the null hypothesis over the alternative.

Step 5:
It is fair to conclude that there is no statistical evidence that indicates a change in price between homes in areas where the NPL=1 compared to areas where NPL=0.

10. Regress **price** on a **constant**, **npl** , **ba_or_better**, **unemprt**, and **povrat**. Generate a series of the predicted values of price and plot those against the price data series: What do you see in terms of fit? (Start by thinking about what we would see in the graph if the model perfectly predicted price.)

For generating predicted values of price from a model, see R bootcamp 2.5. Use `lm_object$fitted.values`

In [54]:
```
lm_model <- lm(price ~ 1 + npl + ba_or_better + unemprt + povrat, data = myc
pred_values <- lm_model$fitted.values

plot(pred_values, mydata$price, xlab = "Predicted Price", ylab = "Actual Pri
abline(a = 0, b = 1, col = "red")
```



The model seems like it fits in our first set of observations, however the restof the observations do not seem to fit on the graph. The model does not have an overall perfect fit

11. What is the percent variation of housing prices that the model is explaining, and what percent is the model NOT explaining?

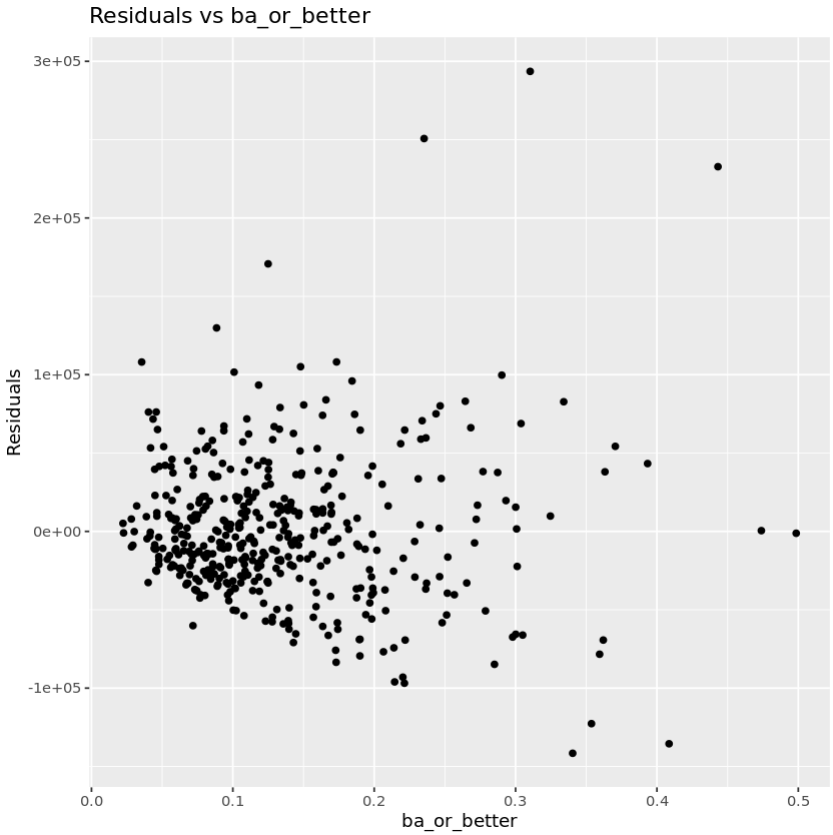
Using R^2 from our summary, we can see that 50.59% of houses is explained by the model and 49.41% is not explained by the model. We got this from our summary above.

In [55]:
```
summary(lm_model)
```

```
Call:
lm(formula = price ~ 1 + npl + ba_or_better + unemprt + povrat,
    data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-141581  -26283   -4923   19733  293546

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     75992      10184   7.462 4.59e-13 ***
npl             -2971       4813  -0.617    0.537
ba_or_better   531656      34182  15.553  < 2e-16 ***
unemprt         34365      85038   0.404    0.686
povrat        -164565      37740  -4.360 1.62e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46090 on 440 degrees of freedom
Multiple R-squared:  0.5059,    Adjusted R-squared:  0.5014
F-statistic: 112.6 on 4 and 440 DF,  p-value: < 2.2e-16
```

12. Compute the residuals series and plot the **residuals** on the vertical axis against **ba_or_better** in the x axis, using `ggplot()`. Is the constant variance assumption for the residuals valid or not for different levels of **ba_or_better**, when you look at the scatter plot of the estimated residuals?

In [56]:
```
residuals <- residuals(lm_model)
residuals_df <- data.frame(ba_or_better = mydata$ba_or_better, residuals = r
ggplot(residuals_df, aes(x = ba_or_better, y = residuals)) +
       geom_point() +
       labs(title = "Residuals vs ba_or_better", x = "ba_or_better", y = "Re
```



I noticed that as ba_or_better increases the spread of the residuals, it also increases meaning that our constant variance assumption is not cohiesive. Given the graph we made we can see that between ba_or_better and the residuals have are related in some way. One point to consider is that we cannot see a pattern between the residuals. The assumption is not valid.

13. Using the triplet Sign, Size, Significance (SSS), let's interpret two of the coefficients from the model in Question 10.

(a) What can you say of the effect of **povrat** on housing prices holding other factors constant?

(For this question, when interpreting "Size", rather than moving the explanatory variable of interest **povrat** by 1 unit, which is a shift in proportion of households in poverty by 100 percentage points, consider shifting it by more sensible 0.01 unit. "As poverty rate in the Census Tract increases by 1 percentage point, ...")

In [57]:
```
lm_model_2 <- lm(price ~ 1 + npl + ba_or_better + unemprt + povrat, data = m
summary(lm_model_2)
```

```
Call:
lm(formula = price ~ 1 + npl + ba_or_better + unemprt + povrat,
    data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-141581  -26283   -4923   19733  293546

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     75992      10184   7.462 4.59e-13 ***
npl             -2971       4813  -0.617    0.537
ba_or_better   531656      34182  15.553  < 2e-16 ***
unemprt         34365      85038   0.404    0.686
povrat        -164565      37740  -4.360 1.62e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46090 on 440 degrees of freedom
Multiple R-squared:  0.5059,    Adjusted R-squared:  0.5014
F-statistic: 112.6 on 4 and 440 DF,  p-value: < 2.2e-16
```

➡ Type your answer to *Exercise 1 Question 13 part (a)* here (replacing this text)

Sign: Negative

Size: -1,546.65 in dollars. We get this by -164565 * 0.01 because we need to increase it by .01 or 1% point. The housing price is expected to decrease 1,645.65.

Significance: We can see that P-value for povrat is 1.62e-05 and we can compare this to .05, which is obviously much less. It is safe to conclude that there is a relationship between poverty rate and housing prices that is statistically significant at the 5% level.

(b) What about the coefficient on **npl**? Use the (SSS) interpretation again.

Sign: Negative

Size: -2971. This would mean that a house on the census tract with an npl it would decrease by that value compared to one without a npl.

Significance:
We notice that the value of npl is .537 which is greater than the significance level of .05, which means that we are not able to reject the null hypothisis. This would mean that the relationship between housing prices and npl is not statistically significant. It is also important to note that while the coefficient brings to light that being in a Census Tract with an NPL site might be related with lower housing prices by $2,971, this does does mean it is statistically significant. By the same token we cannot be sure that the true effect is not 0.

14. First, using the data, estimate correlation between **povrat** and the **ba_or_better**.

Given what you found in Question 10's regression, and using the correlation between **povrat** and the **ba_or_better**, what do you expect would happen to the estimated coefficient on **ba_or_better** if you omit the poverty rate (**povrat**) from the regression considered in question 10?

Go through the Omitted Variable formula and explain briefly. (We are not asking you to run the linear regression on the "naive" model.)

In [58]:
```
corre_pov <-cor(mydata$povrat, mydata$ba_or_better)
corre_pov
```

-0.447898192945506

Based on our correlation variable of -0.4478981, we can see that there is a moderate negative correlation between poverate and ba_or_better. On the same hand we notice that as poverate increases then individuals with higher education decreases. If we were to consider OVB and omit povrat from the regression and povrate correlated with ba_or_better then the estimated coefficient of ba_or_better would be baised. To ommit povrat would mean that there would be an upward bias in the estimated coefficient of ba_or_better since povrat is negatively correlated and would not be taken into consideration in our model.

15. Now estimate the model in Question 10 but do not include **povrat**. What is the new estimate of the coefficient on **ba_or_better**, and do you confirm your answer in Question 14?

In [59]:
```
lm_model4 <- lm(price ~ 1 + npl + unemprt + ba_or_better, data = mydata)
summary(lm_model4)
```

```
Call:
lm(formula = price ~ 1 + npl + unemprt + ba_or_better, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-162422  -26587   -6289   21202  281807

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     64777      10053   6.444 3.06e-10 ***
npl             -1794       4902  -0.366    0.715
unemprt       -122744      78585  -1.562    0.119
ba_or_better   564599      34011  16.600  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47020 on 441 degrees of freedom
Multiple R-squared:  0.4845,    Adjusted R-squared:  0.481
F-statistic: 138.2 on 3 and 441 DF,  p-value: < 2.2e-16
```

The estimate we recently got on the coefficient on ba_or_better is higher. We can also confirm that our answer in q14 that the coefficient would increase. We see that ba_or_better goes from 531656 to 564599 clearly indicating that when povrate is omitted, the value of ba_or_better would be overvalued yet because the correlation isn't very strong, it's effect wont be as great. This does confirm my answer in question 14. To ommit povrat did create an upward bias.

In [60]:
```
summary(lm_model4)$r.squared
```

0.484521833874771

1**16.** What happens to the R squared (R^2) when you omit the **povrat** variable in the equation compared to the R squared in Question 10?

Answer: When we exclude povrate variable in the equation then R^2 would decrease from 0.5059 to 0.4845 indicating that omitting povrate decreases the fitness of our model.

In [61]:
```
summary(lm_model4)$r.squared
```

0.484521833874771

17 Run a linear model of the log of prices (**lprice)** on the same variables in question 10. Please interpret the estimated coefficient for **npl**. Use SSS interpretation again. What do you conclude?

In [62]:
```
lm_model_5 <- lm(lprice ~ 1 + npl + ba_or_better + unemprt + povrat, data =
summary(lm_model_5)
```

```
Call:
lm(formula = lprice ~ 1 + npl + ba_or_better + unemprt + povrat,
    data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.070432 -0.018734 -0.000961  0.018067  0.078915

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.439494   0.005972 408.510  < 2e-16 ***
npl          -0.001232   0.002822  -0.437    0.663
ba_or_better  0.265840   0.020044  13.263  < 2e-16 ***
unemprt      -0.022396   0.049866  -0.449    0.654
povrat       -0.150126   0.022131  -6.784 3.79e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02703 on 440 degrees of freedom
Multiple R-squared:  0.5088,    Adjusted R-squared:  0.5043
F-statistic: 113.9 on 4 and 440 DF,  p-value: < 2.2e-16
```

Sign: Negative

Size: -.001232. The negative sign indicates that to increase npl would mean that it would be associated with a decrease in the log of house prices, assuming all other variables in our model stay the same.

Significance: We notice that the p-value for npl is 0.663 which is greater than .05, which is our threshold for statistical significance. This means that our coefficient is not statistically significant. We also cannot say that being on the NPL list has a significant effect on the housing prices, when taken into consideration for this model. In conclusion that our model suggests that being on NPL list could be associated with a slightly higher house price. By the same token, the effect suggests thatit is not statistically significant.

Exercise 2: Survey Evidence about OVB

Assume we surveyed a *random sample* of 68 EEP118 Students.

	Percent Answering both OVB Questions after Lecture 7 Correctly	Total number of respondents
Overall	41.2%	68
Among Soccer Fans	44.7%	38
Among Non Soccer Fans	36.7%	30

Let p be the true but unknown proportion of the population (all EEP118 students) that would have answered both OVB questions correctly.

1. Use the survey results to estimate p. Also estimate the standard error of your estimate.

In [65]:
```
p = 0.412
n = 68
se = sqrt((p*(1-p))/n)
p
se
```

0.412
0.0596874210809457

We have an estimate of p is .412 and our standard error .059689 or 5.97%.

2. Construct a 95% confidence interval for p. Interpret.

In [66]:
```
val <- 1.96

lowerbound <- p - (val*se)
upperbound <- p + (val*se)

lowerbound
upperbound
```

0.295012654681346
0.528987345318654

We got with a 95% confidence interval for p has a lower bound of 0.2950 and upper bound of 0.5290. We used interval [29.5%, 52.9%] as the percent form.

3. Is there statistical evidence that more than 50% of respondents would have answered correctly? Use the 5 steps for hypothesis testing with a 5% significance level.

In [68]:
```
p_0 = 0.50
z_num = p - p_0
z_den = sqrt((p_0 * (1 - p_0))/n)
z = z_num/z_den
z_num
z_den
z
```

-0.088
0.0606339062590832
-1.45133318021742

-1.45 is the z vlue and it is not greater than the critical z value of 1.645 for a one-sided test at a 5% signicance level. -0.088 is the differnce of the sample

proportion and .0606 is the standard error. **Step 1:**
H_0 : p <= 0.50
H_1 : p > 0.50

Step 2:
Z = $\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ = $\frac{-0.088}{0.0606}$ = -1.4513

Step 3:
With α = 0.05 or significance level set at 5% we have a critical value of 1.96

Step 4:
We have a z value of -1.45 and absolute value of 1.45 < 1.96, which is our critical value, at the 5% significance level.We fail to reject the null hypothesis

Step 5: We fail to reject the null hypothesis. Therefore we can conclude that there isn't sufficient evidence to support the claim that more than 50% of respondents would have answered both OVB questions correctly.

4. Is there statistical evidence that answering correctly is more likely for respondents who are soccer fans versus those who are not, at the 5% significance level? Explain (to answer this question, use the 5 steps for hypothesis testing).

In [77]:
```
p_0 = 0.367
p_1 = 0.447
p_hat = 0.412
n_0 = 30
n_1 = 38

z_n = p_1 - p_0
z_d = sqrt(p_hat*(1-p_hat)*(1/n_0 + 1/n_1))
z = z_n/z_d

df = (n_0 + n_1 - 2)
a = 0.05
c <- qt(1 - (a / 2), df)
z_n
z_d
z
c
```

0.08
0.120209641438351
0.665504023161303
1.99656441895231

.08 is the sample between two sample proportions, .12 is the standard error between the two individual proportions, .66 z value, 1.99 critical value for the t

distribution, **Step 1:**
$H_0 : p_0 = p_1$
$H_! : p_0 \neq p_1$

Step 2: Z = $\frac{\hat{p_1} - \hat{p_0}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_0})}} = \frac{0.08}{0.1202} = 0.6655$

Step 3: With α = 0.05 and 66 degrees of freedom we have a critical value of 1.9966

Step 4: We have a critical value of 1.9966 and with the z value being less than that at 0.6655 we can state that we fail to reject the null hypothesis

Step 5: We fail to reject the null hypothesis. So we can state that there isn't sufficient evidence to suggest that soccer fans are more likely to to answer both questions correctly compared to non-soccer fans.

In [ ]: