1. a) The bias for an estimator W for the parameter θ is given by:

Bias[W] = E[W] – θ,

or if an estimator is biased, the expected value of the estimator is not equal to the value of the population parameter.

b) An estimator is biased if when it is estimated many times the average value of is not equal to the population parameter it is trying to estimate.

c) We care if an estimator is biased because a biased estimator means that the models we make are not accurate. We want to estimate the true relationship between explanatory variables and an outcome, but if we estimators we use to describe that relationship are biased, we will describe a different relationship than the true one, and that is not a very useful thing to describe.

d) A consistent estimator has a probability limit, which means that as the number of observations increases, the distribution collapses, and the spike is centered at the parameter the estimator predicts.

e) True

f) I would choose estimator B. Even though estimator B is biased, because it is consistent, when the number of observations grows large enough, it centers on the correct value of the population parameter. While the unbiased estimator is always centered on the correct value, it has a larger distribution, so when the number of observations grows large enough, the consistent estimator gives the true population value a larger percentage of the time.

2. a) Both figures B and C likely have heteroskedasticity. In figure B, the ui terms distribution is in the shape of a cone. There is not a constant distribution, instead the ui terms seem to be correlated with the X term, so its likely heteroskedastic. Similarly, in figure C, the ui terms have a larger distribution for small and large values of X than for medium values. So, again they seem to be correlated with the X terms.

b) OLS is still unbiased in the presence of heteroskedasticity.

c) In the presence of heteroskedasticity, OLS is no longer the most efficient linear unbiased estimator. Weighted least squares is more efficient in this scenario.

d) There are a few ways to live with heteroskedasticity. The first of these is to respecify the model. The true model may have the logs of some of the variables or a variable may be omitted. If this is corrected, the heteroskedasticity may also be corrected. A second option is to use weighted least squares instead of OLS as the estimator. This is the most efficient estimator with heteroskedasticity, but it can be infeasible to use this method. A final option is to ignore the inefficiency of OLS and just use unbiased estimators for the standard errors. We can use heteroskedasticity-robust standard errors which are less efficient than normal standard errors but are unbiased in the present of heteroskedasticity.

e) σ2­­3 must have a value of 15 for the model to be homoscedastic. If it is 15, then the variance is constant for all values of x.

f) H0: σ21 = σ23 Ha: σ21 != σ23

FGQ = (SSE3/(n\*-k))/(SSE1/(n\*-k))

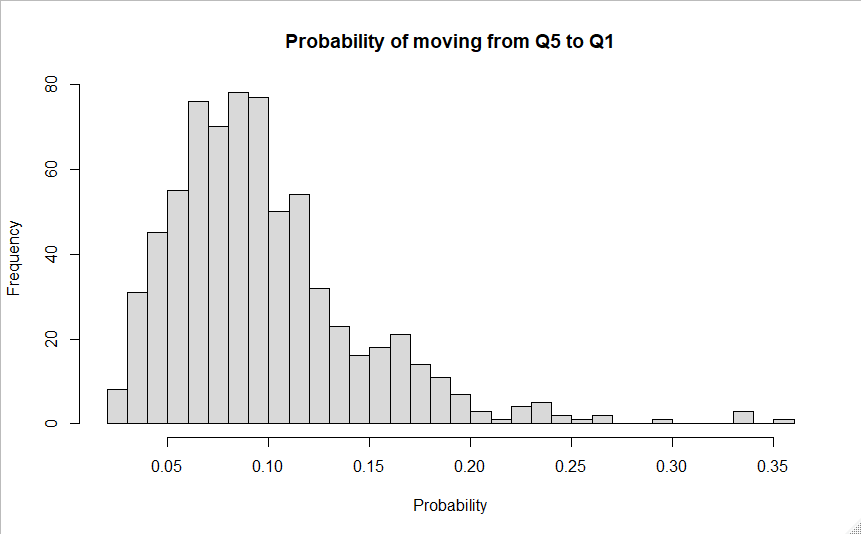
FGQ = (100/(100-2))/(70/(100-2))

FGQ = 10/7 = 1.426

Test this against an F98,98 distribution which produces a p-value of 0.0395.

So, at the 5% level, we reject the null hypothesis and conclude that there is statistically significant evidence that σ21 and σ23 differ. This means we have evidence of heteroskedasticity in this model.

1. b) There is lost of variation in the probability of an individual born to parents in the bottom 5th of income moving up to the top 5th. Most frequently, people have under a 20% chance of achieving this, but there are some that are as high as 35%. There is a much wider disparity between the mean value of approximately 9.8% and the maximum value of approximately 35.8% than between the mean and the minimum value of approximately 2.2%. This shows that most people have a very low probability of moving up.



Min. 1st Qu. Median Mean 3rd Qu. Max.

0.02210 0.06588 0.08889 0.09761 0.11715 0.35714

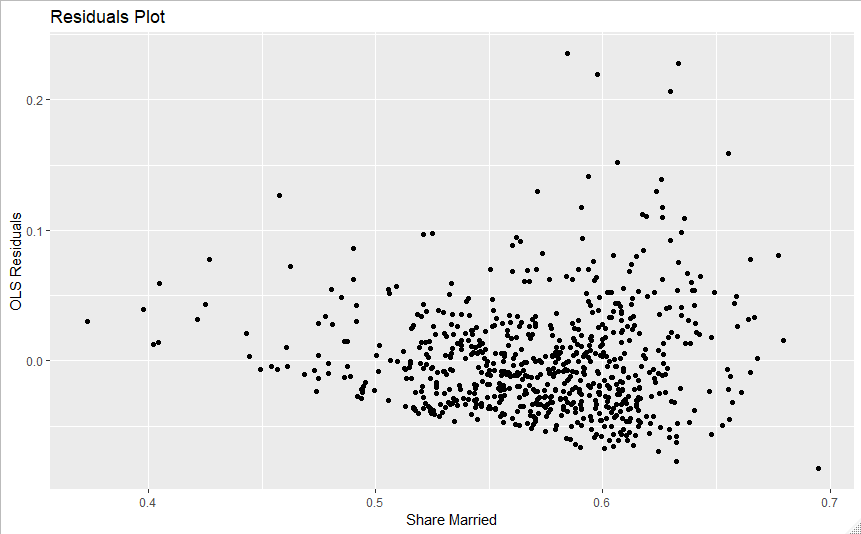
c)

|  |
| --- |
| term estimate std.error statistic p.value  <chr> <dbl> <dbl> <dbl> <dbl>  1 (Intercept) -0.198 0.0196 -10.1 1.34e-22  2 share\_married 0.517 0.0341 15.2 3.88e-45 |

The coefficient for share\_married is 0.517. This means that we would expect that if the share of people married in a community increases by 1 percentage point, we would expect that the percentage of people moving from the bottom to the top 5th of income to increase by 0.517%. The p-value for this estimate is 3.88\*10^-45, which is less than 5%, so at the 5% level this is statistically significant.

d) It does not make sense to interpret the intercept in this case. The intercept tells us what we would expect the percentage of people moving from the bottom to the top 5th of incomes if the proportion of people married in a community was 0%. The data set only contains proportions of people married from approximately 37% to 69%, so 0% is not reasonable. Also, the intercept says that we would expect the proportion of people jumping income levels to be -1.98%. This is not a possibility since it is negative, so the intercept does not make sense to interpret.

e) There is evidence of heteroskedasticity from this plot. The residuals do not have a constant variance, which indicates that the model is heteroskedastic.



f) H0: β1 = 0 Ha: β1 != 0 for the regression ei2 = β0 + β1 share\_married + vi where ei2 estimates ui2. Rejecting the null gives evidence of heteroskedasticity

B-P test statistic: LM^ = 15.77

For the chi-square distribution with one degree of freedom, the LM test statistic gives a p-value of 7.15\*10^-5.

The p-value is less than .05, so we reject the null hypothesis at the 5% level. There is a correlation between the residuals and the explanatory variable. This shows that we have statistically significant evidence for heteroskedasticity in the model.

g) H0: β1 = β2 = 0 Ha: β1 != 0 of β2 != 0 for the model ei2 = β0 + β1 share\_married + β2 share\_married^2 + vi.

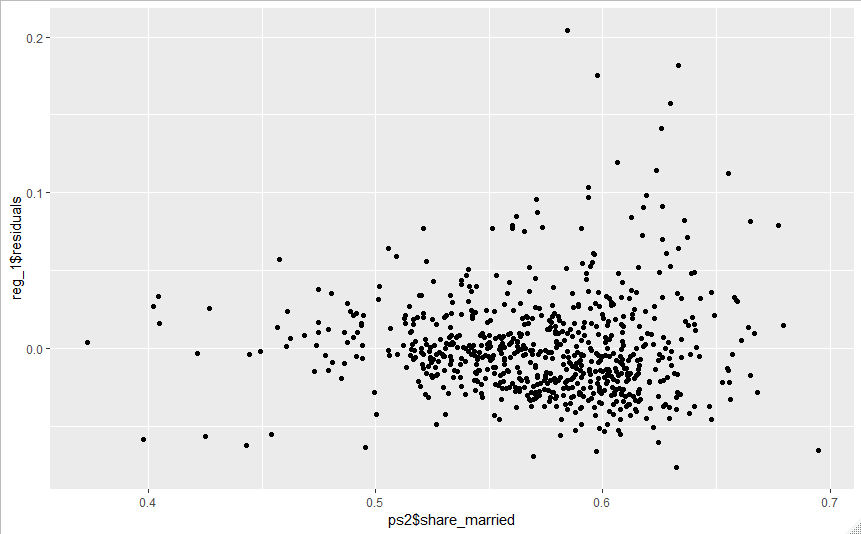
White test statistic: LM^ = 24.21

For the chi-square distribution with two degrees of freedom, this LM test statistic gives a p-value of 5.54\*10^-6

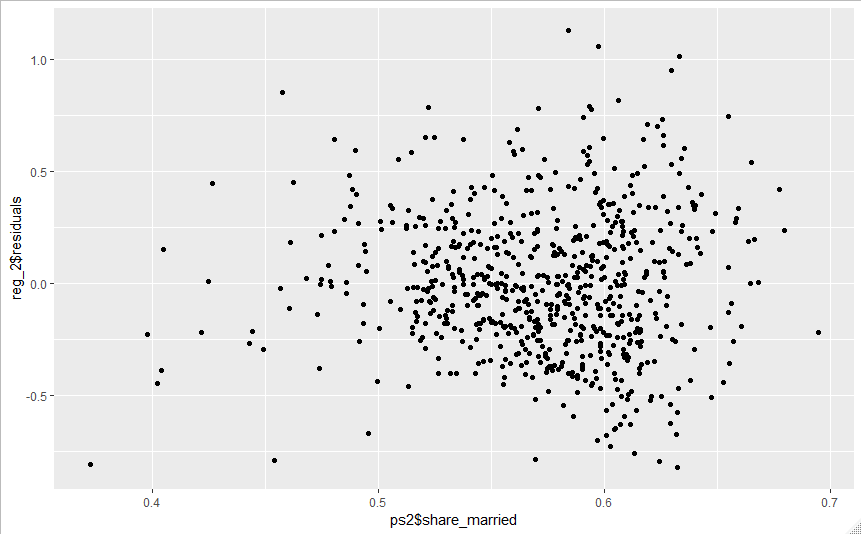
At the 5% significance level, we reject the null hypothesis and conclude that there is statistically significant evidence of heteroskedasticity.

h) The coefficients are the same across both estimates. This is the case because they both use the same estimator, OLS, to determine their value. Nothing changes if we change to regression to be robust for the coefficients. The standard errors do change. Using robust standard errors increases them when compared to the non-robust ones. This is the case because heteroskedasticity robust standard errors use a different estimator to get their values.

i) The first regression I ran was prob\_q5\_q1 against all the other explanatory variable in the data set, i\_urban, share\_black, share\_middleclass, share\_divorced, and share\_married. It didn’t seem to change the appearance of heteroskedasticity that much. The graph is pretty much the same as the original with only a little movement.



The second regression I ran was the log of prob\_q5\_q1 against share\_married, share\_married squared, share\_black and the interaction between share\_married and share\_black. This had a huge effect on the appearance of heteroskedasticity. The graph is drastically different and its not as obvious that there is any heteroskedasticity still in the model.



j) We should not interpret the results of the regression earlier as causal. There are likely omitted variables in this regression that affect the probability an individual moves from the bottom to the top 5th of income and are correlated with the share of a community that is married. If this is the case, the regression suffers from omitted variable bias, so the coefficient is either over or underestimating the true relationship. We can still learn interesting things from the model even if we don’t interpret it as causal. For example, we learn that there is a correlation between the share of a community that is married and the probability an individual moves up from the bottom to the top 5th of income. Even if this relationship is exaggerated or underexaggerated from omitted variable bias, it is still interesting to know this is true. Knowing what is correlated with this is an important first step in helping people move out of poverty.