# MGT581 Introduction to Econometrics - Problem Set 2

Monday, October 28

## General information

- This is the second of the three graded problem sets we will have this semester.
- It is due on at **midnight (12 pm) on Monday, November 11**
- Please submit both:
  - The .Rmd file with your code and answers.
  - The PDF file generated by knitting the .Rmd file.
- If the file fails to Knit:
  - Remove install.package("…") comands from your code (preferrably use the lib function provided in the template).
  - Knit to HTML and then convert the HTML file to PDF.

## Exercise 1: College Distance (6 points)

In this exercise, you will work with the CollegeDistance dataset. These data are taken from the HighSchool and Beyond survey conducted by the U.S. Department of Education in 1980, with a follow-up in 1986. The survey included students from approximately 1100 high schools.

**Note: In the United States, "College" is typically used to refer to a standard undergraduate program. In this context, a "four-year college" is a college that offers a standard undergraduate program that typically takes four years to complete.**

The dataset contains the following variables:

- `ed`: Years of education completed.
- `female`: Dummy variable, 1 if female, 0 if male.
- `black`: Dummy variable, 1 if Black, 0 otherwise.
- `Hispanic`: Dummy variable, 1 if Hispanic, 0 otherwise.
- `bytest`: Base year composite test score.
- `dadcoll`: Dummy variable, 1 if father is a college graduate, 0 otherwise.
- `momcoll`: Dummy variable, 1 if mother is a college graduate, 0 otherwise.
- `incomehi`: Dummy variable, 1 if family income is above $25,000, 0 otherwise.
- `ownhome`: Dummy variable, 1 if family owns a home, 0 otherwise.
- `urban`: Dummy variable, 1 if high school is in an urban area, 0 otherwise.
- `cue80`: County unemployment rate in 1980.
- `stwmfg80`: State hourly wage in manufacturing in 1980.
- `dist`: Distance from nearest 4-year college, measured in tens of miles.
- `tuition`: Average state 4-year college tuition in thousands of dollars.

Load and inspect the data

```r
# Load the data here
college_dist <- read_excel("CollegeDistance.xls")
head(college_dist)
```

```
## # A tibble: 6 x 14
##   female black hispanic bytest dadcoll momcoll ownhome urban cue80 stwmfg80
##    <dbl> <dbl>    <dbl>  <dbl>   <dbl>   <dbl>   <dbl> <dbl> <dbl>    <dbl>
## 1      0     0        0   39.2       1       0       1     1   6.2     8.09
## 2      1     0        0   48.9       0       0       1     1   6.2     8.09
## 3      0     0        0   48.7       0       0       1     1   6.2     8.09
## 4      0     1        0   40.4       0       0       1     1   6.2     8.09
## 5      1     0        0   40.5       0       0       0     1   5.6     8.09
## 6      0     0        0   54.7       0       0       1     1   5.6     8.09
## # i 4 more variables: dist <dbl>, tuition <dbl>, ed <dbl>, incomehi <dbl>
```

*Unless otherwise stated, use the 5% significance level throughout this exercise and run regressions with the lm_robust function from the estimatr package. Use HC2 standard errors for all statistical tests.*

**a.** Run a regression of years of completed education (`ed`) on the distance to the nearest four-year college (`dist`). Discuss how you expect the coefficient on `dist` to be signed, and why.

```r
# Running robust linear regression
model_a <- lm_robust(ed ~ dist, data = college_dist, se_type = "HC2")
summary(model_a)
```

```
##
## Call:
## lm_robust(formula = ed ~ dist, data = college_dist, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept) 13.95586    0.03783 368.952 0.000e+00 13.88170 14.03002 3794
## dist        -0.07337    0.01346  -5.451 5.324e-08 -0.09976 -0.04698 3794
##
## Multiple R-squared:  0.00745 ,   Adjusted R-squared:  0.007188
## F-statistic: 29.71 on 1 and 3794 DF,  p-value: 5.324e-08
```

Coefficient on distance is going to represent how does distance from nearest 4-year college (measured in tens of miles) affect the years of education completed for the person. Without looking at the product of regression, we would say that being far away from college would impact negatively the years of education completed, thus the coefficient would be negative - as distance grows, the years of education go down. Like we expected, we do indeed get a negative coefficient of $-0.07337$, meaning that with every growth in distance of 10 miles, the years of completed education go down by $-0.073375$.

**b.** Interpret the (slope) coefficient on `dist` (Hint: double check the units of measurement of the dependent and independent variables). What are the associated p-value and t-statistic? State the null hypothesis `H0` and the alternative hypothesis `H1` that correspond to the regression output you observe. What do you conclude on the significance of `dist`?

Like previously mentioned, the slope coefficient implies that for every 10 mile increase in distance from nearest 4-year college, the years of completed education decrease by $-0.073375$ years, when holding the other factors constant. We can see that the associated p-value is $p = 5.324e - 08$, and t-statistic is $t = -5.451$. We will use these values for hypothesis testing, and these are our hypothesis: H0:The distance from the nearest 4-year college has no effect on years of completed education → $\beta_{dist} = 0$. H1:The distance from the nearest 4-year college has an effect on years of completed education → $\beta_{dist} \neq 0$. Since our p-value is smaller than 5% significance level: $p = 5.324e - 08 < 0.05$ and our t-statistic is $|t| = | - 5.451| > 1.96$, we can reject the null hypothesis H0, and conclude distance does have the significant negative effect on years of completed education.

**c.** Interpret the intercept. Is it significant?

Intercept represent the number of years of completed education when the distance is equal (which is quite unlikely, but still good to know). Intercept here is 13.95586, meaning that when distance is zero, years of education are 13.95586. P-value of the intercept is $p = 0.000e{+}00$, which is so small that it is practically zero, meaning that it is definitely smaller than 0.05 (5% significance level), and thus is statistically significant. Even though it seems that this intercept doesn't have any practical use, it is good to provide baseline.

**d.** Can you interpret the coefficient on ed causally? If yes, why? If no, which biases are you concerned about?

No, we can't interpret the coefficient on ed causally, since other factors might influence both the distance to college and educational outcomes. Some of the possible factors are family income, parental education, neighborhood. For example, having bigger household income can result in living closer to central areas with more colleges, or having more resources to help a child with their education.

**e.** Add incomehi to the regression in question a) and interpret the coefficient. Compute a 99% confidence interval for the coefficient. Based on this confidence interval, can you reject the null hypothesis that the coefficient is equal to 0?

```r
# Regression with added control variable incomehi
model_e <- lm_robust(ed ~ dist + incomehi, data = college_dist, se_type = "HC2")
summary(model_e)
```

```
##
## Call:
## lm_robust(formula = ed ~ dist + incomehi, data = college_dist,
##     se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept) 13.68735    0.04208 325.267 0.000e+00  13.6049 13.76986 3793
## dist        -0.05821    0.01330  -4.376 1.243e-05  -0.0843 -0.03213 3793
## incomehi     0.84635    0.06506  13.008 7.059e-38   0.7188  0.97391 3793
##
## Multiple R-squared:  0.05163 ,   Adjusted R-squared:  0.05113
## F-statistic: 101.9 on 2 and 3793 DF,  p-value: < 2.2e-16
```

We would expect the coefficient for incomehi to be positive, as it makes sense that student whose family has higher income, also has more completed years of education. As expected, we get the positive value of 0.84635, meaning that when holding distance constant, students with families who earn more than $25000 are complete 0.84635 more years of education.

```r
# Calculating 99% confidence interval
confint_incomehi <- confint(model_e, "incomehi", level = 0.99)
confint_incomehi
```

```
##              0.5 %   99.5 %
## incomehi 0.6786687 1.014029
```

Hypothesis testing: `H0`:The income has no effect on years of completed education → $\beta_{incomehi} = 0$. `H1`:The income has an effect on years of completed education → $\beta_{incomehi} \neq 0$. If we look at the confidence interval, we can say that we are 99% confident that the effect of income on complete years of education is between 0.6786687 and 1.014029. Since our confidence interval doesn't include 0, that means that income is actually a significant factor, and we can reject `H0` at 1% significance level (1% level corresponds to 99% confidence interval).

**f.** Compare the coefficient on `dist` between models **a** and **e**. How did the point estimate change? Justify your answer in light of how you expect educational attainment, distance from a four-year college, and household income to be related. Based on your answer, what do you conclude on the adequacy of model **a**?

```r
c("Coefficient for distance for model a: ", coef(model_a)["dist"])
```

```
##
## "Coefficient for distance for model a: "
##                                     dist
##                     "-0.073372707129185"
```

```r
c("Coefficient for distance for model e: ", coef(model_e)["dist"])
```

```
##
## "Coefficient for distance for model e: "
##                                     dist
##                    "-0.0582140409612618"
```

```r
c("Coefficient for income for model e: ", coef(model_e)["incomehi"])
```

```
##                                                                     incomehi
## "Coefficient for income for model e: "                   "0.846348807484359"
```

We see that the coefficient for distance for model e is smaller than for model a. The reason for this decrease is the adding of the new control variable incomehi. This means that some of the negative relationship between distance and education can be explained by income. Part of the negative effect the distance had on education in model e was maybe due to some families having smaller income and not affording costs of attending college that is more far away and maybe even not attending at all, while the richer ones can afford extra costs. We can conclude that model a isn't very adequate. It overestimates the true effect the distance has on years of completed education, because it doesn't account for some important factors, like income. This coefficient then might be biased.

**g.** Consider again the regression in **e** and add `momcoll` and `dadcoll` as explanatory variables. Test the hypothesis that parental college attendance has no effect on education at the 1% significance level. What is the null hypothesis? What is the alternative hypothesis? What do you conclude?

```r
# Regression with added control variables momcoll and dadcoll
model_g <- lm_robust(ed ~ dist + incomehi + momcoll + dadcoll, data = college_dist, se_type = "HC2")
summary(model_g)
```

```
##
## Call:
```

```
## lm_robust(formula = ed ~ dist + incomehi + momcoll + dadcoll,
##     data = college_dist, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept) 13.50360    0.04235 318.842 0.000e+00 13.42057 13.58664 3791
## dist        -0.03863    0.01277  -3.024 2.507e-03 -0.06368 -0.01359 3791
## incomehi     0.45476    0.06809   6.678 2.768e-11  0.32126  0.58827 3791
## momcoll      0.59442    0.09105   6.529 7.515e-11  0.41591  0.77292 3791
## dadcoll      0.88722    0.08283  10.711 2.153e-26  0.72482  1.04962 3791
##
## Multiple R-squared:  0.1125 ,    Adjusted R-squared:  0.1116
## F-statistic: 129.4 on 4 and 3791 DF,  p-value: < 2.2e-16
```

If we want to test joint hypotheses that involve multiple coefficients we need to use F-test based on the F-statistic (with $q = 2$ restrictions in our case). Hypothesis testing: `H0`:The parental college attendance has no effect on years of completed education → $\beta_{momcoll} = \beta_{dadcoll} = 0$. `H1`:The parental college attendance has an effect on years of completed education → at least one of $\beta_{momcoll} \neq 0$ or $\beta_{dadcoll} \neq 0$.

```r
# We do an F-test
linearHypothesis(model_g, c("momcoll=0", "dadcoll=0"), test=c("F"), white.adjust = "HC2")
```

```
##
## Linear hypothesis test:
## momcoll = 0
## dadcoll = 0
##
## Model 1: restricted model
## Model 2: ed ~ dist + incomehi + momcoll + dadcoll
##
##   Res.Df Df      F    Pr(>F)
## 1   3793
## 2   3791  2 131.27 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From known table for F-statistic, we know that the critical value at a 1% significance level equals 4.61 for $q = 2$ restrictions in our case. Since our $F = 131.27 > 4.61$, we can safely reject our null hypothesis. Similarly, we see that our p-value is also smaller than 1%: $p = 2.2e - 16 < 0.01$. Thus, we can conclude that the college education of parents is a important factor for years of completed education of their children, since mother being a college graduate increases years by 0.59442, and father being a college graduate increases years by 0.88722.

**h.** Consider again the regression in **g**. Test the hypothesis that the coefficient on `momcoll` and the coefficient on `dadcoll` are equal.

The logic stays the same, we just change our hypotheses: `H0`:The coefficients for mom being a college graduate and dad being a college graduate are the same → $\beta_{momcoll} = \beta_{dadcoll}$. `H1`:The coefficients for mom being a college graduate and dad being a college graduate are different → $\beta_{momcoll} \neq \beta_{dadcoll}$.

```
# We do an F-test
linearHypothesis(model_g, c("momcoll=dadcoll"), test=c("F"), white.adjust = "HC2")
```

```
##
## Linear hypothesis test:
## momcoll - dadcoll = 0
##
## Model 1: restricted model
## Model 2: ed ~ dist + incomehi + momcoll + dadcoll
##
##   Res.Df Df     F  Pr(>F)
## 1   3792
## 2   3791  1 3.989 0.04587 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The significance level is 5% (as defined at the beginning). Since our p-value is smaller than this $p = 0.04587 < 0.05$, we can reject the null hypothesis (although it’s quite close, and would not be rejected at 1% level). We can also look at F. F for 5% significance level and 2 restrictions is $F = 3.00$. Since our F is $F = 3.989$, we can again reject the null hypothesis. Thus, we conclude that the effect of mom being a college graduate and dad being a college graduate on education of their child is different (their coefficients are not the same). If we consider significance level from point g, so 1%, then p-value will not be smaller as $p = 0.04587 > 0.01$. Additionally, if we look at at F, for significance level of 1% and 2 restrictions, our F needs to be greater than 4.61, and it isn’t:$F = 3.989 < 4.61$. Thus, we couldn’t reject the null hypothesis, and we don’t have enough evidence to say that the effect of mom being a college graduate and dad being a college graduate on education of their child is different.

**i.** Add the interaction between `dadcoll` and `momcoll` to the regression in the previous question. Interpret the coefficient on the interaction term, and discuss whether its sign is consistent with your expectations.

```
# Regression with added interaction between momcoll and dadcoll
model_i <- lm_robust(ed ~ dist + incomehi + momcoll + dadcoll + momcoll:dadcoll,
                     data = college_dist, se_type = "HC2")
summary(model_i)
```

```
##
## Call:
## lm_robust(formula = ed ~ dist + incomehi + momcoll + dadcoll +
##     momcoll:dadcoll, data = college_dist, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                 Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)     13.49073    0.04270 315.921 0.000e+00 13.40700 13.57445 3790
## dist            -0.03802    0.01275  -2.981 2.893e-03 -0.06303 -0.01301 3790
## incomehi         0.44731    0.06802   6.577 5.470e-11  0.31396  0.58066 3790
## momcoll          0.80940    0.13445   6.020 1.910e-09  0.54579  1.07300 3790
## dadcoll          0.99155    0.09455  10.487 2.198e-25  0.80618  1.17692 3790
## momcoll:dadcoll -0.41159    0.18068  -2.278 2.278e-02 -0.76582 -0.05736 3790
##
## Multiple R-squared:  0.1137 ,    Adjusted R-squared:  0.1126
## F-statistic: 103.6 on 5 and 3790 DF,  p-value: < 2.2e-16
```

From previous example, we know that mom being a college graduate and dad being a college graduate has positive effect on their child's education-number of years completed gets bigger. Right of the bat we would say that having both parents with college degree would probably increase the years of completed education of their child, thus coefficient of interaction would be positive. But here we see that coefficient for that interaction is $-0.41159$, so negative. This implies that the effect of having both parents as college graduates is just not equal to sum of their individual effects; their combined effect is just slightly less than that, so we have to subtract from that sum, and we have $0.80940 + 0.99155 - 0.41159 = 1.38936$ increase in years. So, having both parents being college graduates does increase the years of completed education of their child, just not by the same amount as sum of their individual effects. This seems like a reasonable explanation, but there is still a chance that having both parents as college graduates might have an effect on their child's career - maybe they encourage their child to excel in some craft, or maybe work for the family business, so that's why the coefficient is this negative.

**j.** Test the hypothesis that a mother's college attendance does NOT help to explain the child's education. (Hint: think carefully of what this statement means in the context of this particular regression).

What we have to keep in mind is that model also includes the interaction term between momcoll and dadcoll. This means that the effect of momcoll is not isolated, it is also influenced by dadcoll. So, we have to be aware of both the variable momcoll and interaction term momcoll:dadcoll-we test if mother's has an effect either independently or through interaction. Hypothesis testing: `H0`:Mother's college attendance does NOT help to explain the child's education → $\beta_{momcoll} = 0$ and $\beta_{momcoll:dadcoll} = 0$. `H1`:Mother's college attendance does help to explain the child's education → at least one of $\beta_{momcoll} \neq 0$ or $\beta_{momcoll:dadcoll} \neq 0$.

```
# We do an F-test
linearHypothesis(model_i, c("momcoll=0", "momcoll:dadcoll=0"), test=c("F"), white.adjust = "HC2")
```

```
##
## Linear hypothesis test:
## momcoll = 0
## momcoll:dadcoll = 0
##
## Model 1: restricted model
## Model 2: ed ~ dist + incomehi + momcoll + dadcoll + momcoll:dadcoll
##
##   Res.Df Df      F    Pr(>F)
## 1   3792
## 2   3790  2 23.305 8.717e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value is lower than significance level: $p = 8.717e-11 < 0.05$, we can safely reject the null hypothesis, that a mother's college attendance does NOT help to explain the child's education. Additionally we can check F: threshold for 5% would be 3.00, thus we have $F = 23.305 > 3.00$ and again, we are sure to reject the null hypothesis.

**k.** Add tuition to model **i**. Notice that the regression $R^2$ increases. A colleague tells you that, based on this observation, you can conclude that the coefficient on tuition must be statistically different from 0. Do you agree? Why?

```
# Regression with added control variable tuition
model_k <- lm_robust(ed ~ dist + incomehi + momcoll + dadcoll + momcoll:dadcoll + tuition,
                     data = college_dist, se_type = "HC2")
summary(model_k)
```

```
##
```

```
## Call:
## lm_robust(formula = ed ~ dist + incomehi + momcoll + dadcoll +
##     momcoll:dadcoll + tuition, data = college_dist, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                 Estimate Std. Error t value  Pr(>|t|) CI Lower  CI Upper   DF
## (Intercept)     13.37939    0.10159 131.694 0.000e+00 13.18020 13.578571 3789
## dist            -0.03514    0.01301  -2.701 6.950e-03 -0.06065 -0.009629 3789
## incomehi         0.44117    0.06838   6.452 1.244e-10  0.30711  0.575226 3789
## momcoll          0.81029    0.13481   6.011 2.021e-09  0.54599  1.074589 3789
## dadcoll          0.98975    0.09441  10.483 2.287e-25  0.80464  1.174850 3789
## tuition          0.11876    0.09916   1.198 2.311e-01 -0.07566  0.313170 3789
## momcoll:dadcoll -0.41241    0.18089  -2.280 2.267e-02 -0.76706 -0.057762 3789
##
## Multiple R-squared:  0.1141 ,    Adjusted R-squared:  0.1127
## F-statistic:  86.9 on 6 and 3789 DF,  p-value: < 2.2e-16
```

R-squared for the model without control variable tuition is $R^2 = 0.1137$, and the new one with this added variable has $R2 = 0.1141$. So, R-squared really did increase a bit. R-squared usually does increase when added a new variable, because this variable can potentially help with explaining the variation in education. We actually can't use $R^2$ to interpret statistical significance. $R^2$ is not good for causal analysis - it can increase when adding more variables, but that doesn't necessarily mean those variables are causing any changes in the outcome.In this case, $R^2$ did really increase, but when we look at the p-value for tuition: $p = 2.311e - 01 > 0.05$, we see that we can't reject the null hypothesis that the tuition coefficient is equal to zero. In conclusion, our colleague was lying to us.

## Exercise 2: Loan Application data (6 points)

In this exercise, you will work with a dataset on mortgage loan applications.

**Note: The purpose of a mortgage loan is to finance the purchase of a home or other real estate property. Lenders set the interest rate on a loan based on its perceived riskiness. If the loan is perceived as risky, the lender will charge a higher interest rate to "compensate" for potential losses in case the borrower defaults.**

The dataset contains the following variables:

- `action`: Type of action taken by the lender on the application. 1 if approved, 2 if denied, 3 if pending or withdrawn.
- `loan_amount`: Loan amount requested by the applicant in thousands of $.
- `income`: Applicant's income in thousands of $.
- `units`: Number of units in the property.
- `hispan`: Dummy variable, 1 if applicant is Hispanic, 0 otherwise.
- `black`: Dummy variable, 1 if applicant is Black, 0 otherwise.
- `male`: Dummy variable, 1 if applicant is male, 0 otherwise.
- `married`: Dummy variable, 1 if applicant is married, 0 otherwise.

Load and inspect the data.

```r
# Load the data here
loan_app <- read.csv("loan_applications.csv")
head(loan_app)
```

```
##   action loan_amount income units black hispan male married
## 1      3         128     74     1     0      0    1       1
## 2      1         276     90     1     0      0    1       0
## 3      3         158    666     1     0      0    1       1
## 4      1          40     48     1     0      0    1       1
## 5      1          70     33     1     1      0    1       1
## 6      1         126     55     1     0      0    1       1
```

*Unless otherwise stated, use the 5% significance level throughout this exercise and run regressions with the lm_robust function from the estimatr package.*

*For questions from **a** through **f**, use should use the lm() function (non lm_robust) and standard errors computed under the assumption of homoskedasticity*

**a.** Note that action is in integer format. What implicit assumption should hold for a regression of `loan_amount` on the integer action to be meaningful? Do you think this assumption is likely to hold in this case? How would you proceed, if you wanted to estimate the effect of action on `loan_amount`?

Usually when we have this type of regression, we would have one dummy variable and one continuous variable, so the cases are just: dummy variable holds (is 1), or doesn't (is 0). Since the action is in integer format, regression will interpret it as continuous predictor. This means that it will assume that difference between each level (1 and 2, and 2 and 3) is the same. For example, if moving from $action = 1$ to $action = 2$ reduces loan_amount by $20000, this would imply that moving from $action = 2$ to $action = 3$ would also decrease the amount by $20000. Is this actually what would happen? In this case, probably not. The integer values for the action here are just a way to represent distinct categories; there is no actual ordering. It doesn't make sense to assume these categories are spaced equally, and also, these categories don't imply any specific change in loan_amount - they are purely categorical values. So how to proceed? The rest of the exercise suggests we should use only action=1, which might be one of the solutions of this problem. We could also divide these integer values to 3 different dummy variables - approved, denied and pending/withdrawn, and then exclude one and treat it like a baseline (approved for example).

*Keep only the observations where action is equal to 1 (i.e. the loan was approved) for the rest of the exercise.*

```r
# First keep just action=1
loan_app <- loan_app[loan_app$action==1,]
head(loan_app)
```

```
##   action loan_amount income units black hispan male married
## 2      1         276     90     1     0      0    1       0
## 4      1          40     48     1     0      0    1       1
## 5      1          70     33     1     1      0    1       1
## 6      1         126     55     1     0      0    1       1
## 8      1         208     82     1     0      1    1       1
## 9      1         187    177     1     0      0    1       1
```

**b.** Produce a table with summary statistics for the dataset (excluding `action`). You should report the number of observations, the mean, the standard deviation, the minimum, the 25th percentile, the median, the 75th percentile, and the maximum for each variable. Format output with two decimal places. Comment on any relevant patterns you observe.

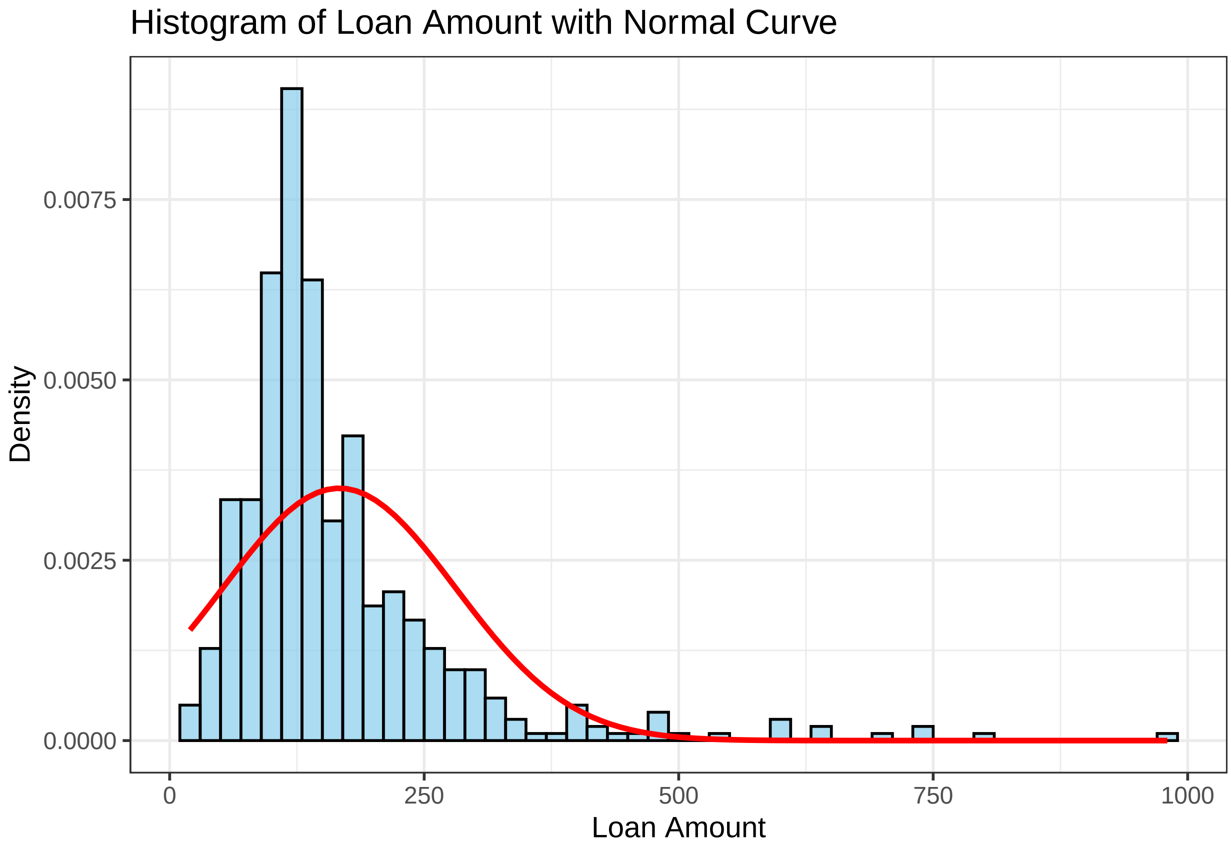
```r
# We don't need column action for this summary
loan_app_no_action <- loan_app[, names(loan_app) != "action"]

# We make a function for all the needed statistical values
compute_summary_stats <- function(x) {
  stats <- c(
    Observations = sum(!is.na(x)),
    Mean = round(mean(x, na.rm = TRUE), 2),
    Std_Dev = round(sd(x, na.rm = TRUE), 2),
    Min = round(min(x, na.rm = TRUE), 2),
    Q1 = round(quantile(x, 0.25, na.rm = TRUE), 2),
    Median = round(median(x, na.rm = TRUE), 2),
    Q3 = round(quantile(x, 0.75, na.rm = TRUE), 2),
    Max = round(max(x, na.rm = TRUE), 2)
  )
  return(stats)
}

# I'm transposing the result because for me it is more readable when one row=one variable
summary_stats <- data.frame(t(sapply(loan_app_no_action, compute_summary_stats)))

# I didn't like that for quantiles the names were Q1.25. and Q3.75., so I'm changing the names by hand
colnames(summary_stats) <- c("Observations", "Mean", "Std_Dev", "Min", "Q1", "Median", "Q3", "Max")


print(summary_stats)
```

```
##             Observations   Mean Std_Dev Min  Q1 Median  Q3 Max
## loan_amount          509 166.56  114.04  20 105    136 190 980
## income               509  91.49   85.27   5  48     67 103 870
## units                509   1.12    0.42   1   1      1   1   4
## black                509   0.08    0.28   0   0      0   0   1
## hispan               509   0.06    0.23   0   0      0   0   1
## male                 509   0.82    0.39   0   1      1   1   1
## married              509   0.69    0.46   0   0      1   1   1
```

We can quickly go through every variable. Loan_amount - it has mean of 166.56, with standard deviation of 114.04, so we can deduct that the amount is quite variable. We can also see this in minimum and maximum amounts - minimum is just 20, while maximum is 980. Another interesting thing are quantiles: Q1 is 105, and Q3 is 190, while the median is 136, which suggests that amounts are more concentrated at the lower end, and probably have some large amounts in the right end that are pulling the mean to a higher value. Income - everything here makes sense. What is interesting is the range of income - the lowest one is just $5000, while the highest is $870000. Units - we can deduct that most loan takers have 1 unit properties. Mean is 1.12, and both Q1 and Q3 are 1, meaning that the units values are clustered around value 1. Black - since mean is 0.08, this indicates that small amount of loan getters in this dataset is black. Hispan - similarly, we can deduct that small amount of loan getters is hispanic. Male - mean of 0.82 suggests that majority of people in this dataset are male. Married - mean of 0.69 suggests the big portion of poeple here are married.

**c.** Plot a histogram of `loan_amount`. What do you observe? What does the histogram suggest on the functional form a potential regression trying to explain loan_amount?

```r
# Plotting the histogram
ggplot(loan_app, aes(x = loan_amount)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 20, fill = "skyblue", color = "black",
alpha = 0.7) + stat_function(fun = dnorm, args = list(mean = mean(loan_app$loan_amount, na.rm = TRUE),
sd = sd(loan_app$loan_amount, na.rm = TRUE)), color = "red", linewidth = 1) +
  labs(title = "Histogram of Loan Amount with Normal Curve",
x = "Loan Amount", y = "Density") + theme_bw()
```



We see that most of the amounts fall between 0 and 350, with most of them being clustered around 125. The tail extends to the right, so we do have a smaller number of bigger amounts of loan, but most of them are concentrated on the lower end. We thus conclude that our distribution is right-skewed. The skewness of our histogram is implying that the loan amount does not follow the normal distribution. If it was normal, it would have symmetric, bell-shaped curve. Since normal distribution is often an assumption for linear regression problem, this might be a problem. What can be useful is to consider log transformation, to make the distribution more normal-looking - more symmetrical. This can help with the assumption for normal distribution regarding the regression.

**d.** Run a regression of `loan_amount` on income and interpret the slope coefficient. What is the associated p-value? What is the t-statistic? What are the null and alternative hypotheses that correspond to the regression output you observe? Can you reject the null hypothesis at the 5% level?

```
# Run regression of loan_amount and income
loan_model_d <- lm(loan_amount ~ income, data = loan_app)
summary(loan_model_d)
```

```
##
## Call:
## lm(formula = loan_amount ~ income, data = loan_app)
```

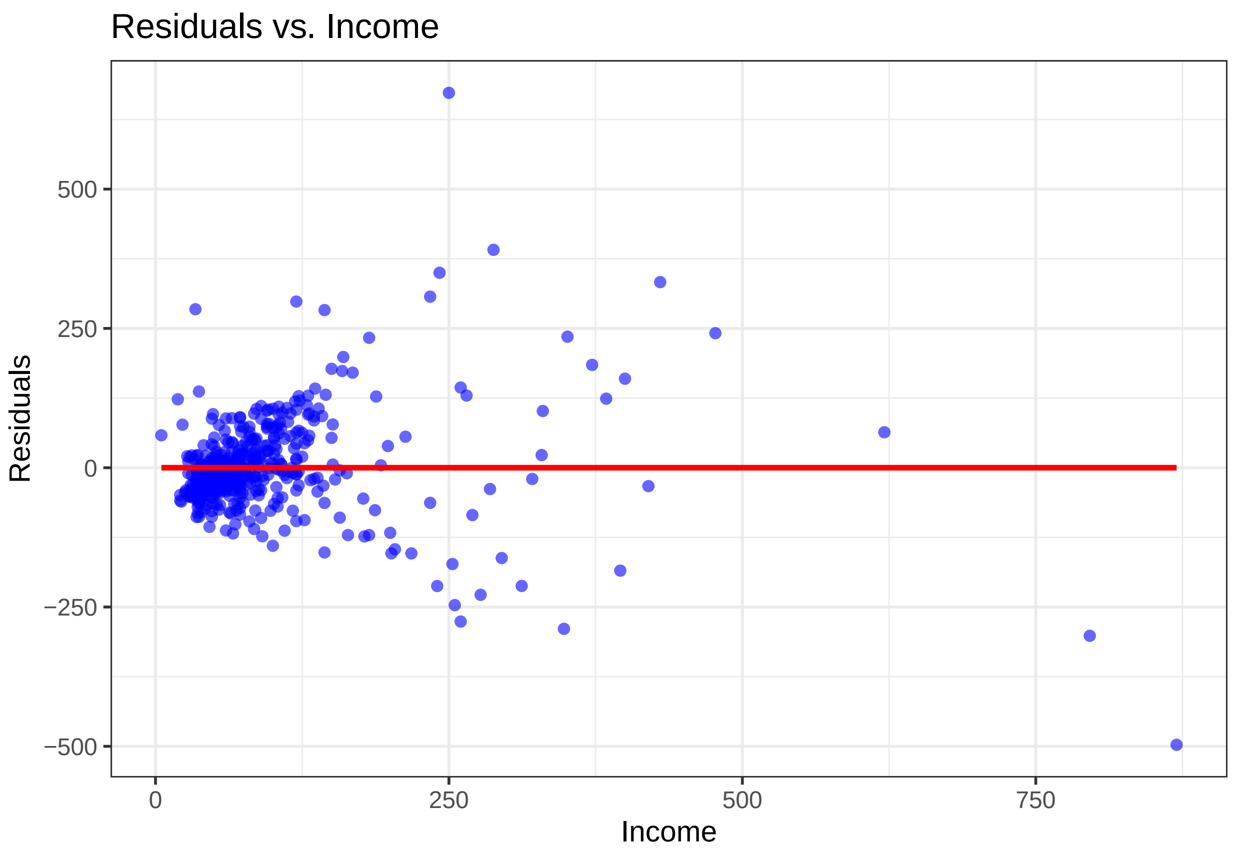
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -497.30  -36.21   -9.19   24.96  672.80
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 85.38320    5.55544   15.37   <2e-16 ***
## income       0.88726    0.04444   19.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.41 on 507 degrees of freedom
## Multiple R-squared:  0.4401, Adjusted R-squared:  0.439
## F-statistic: 398.6 on 1 and 507 DF,  p-value: < 2.2e-16
```

Since the slope is $\beta_{income} = 0.88726$, this means that with every \$1000 increase in house hold income, the taken loan amount is expected to increase by \$887. This does make sense, since it is more reasonable to give bigger amounts of loan to people who can actually afford to pay them off (this is the main thing banks look for). The associated p-value for income is $p < 2e - 16 < 0.05$ - it is practically zero, so definitely less than 5% significance level. This means that income truly is significant factor for loan amount. T-Statistic is $t = 19.96 > 1.96$, and we again conclude that income is significant, thus we can reject null hypothesis. Here is the written down hypothesis testing process: `H0`:The income has no effect on loan amount $\rightarrow \beta_{income} = 0$. `H1`:The income has an effect on loan amount $\rightarrow \beta_{income} \neq 0$. Since we saw that both p-value and t-statistic are smaller/bigger than needed thresholds, we can safely reject the null hypothesis.

**e.** Plot a scatterplot of the residuals of the regression in **d** against `income` together with a regression line (i.e. the regression of the model **e**). Do you find evidence supporting homoskedasticity or heteroskedasticity? Discuss.

```r
# First we extract the residuals from the model
residuals_model_d <- residuals(loan_model_d)
residuals_income_data <- data.frame(income = loan_app$income, residuals = residuals_model_d)

# Then we do the plot
ggplot(residuals_income_data, aes(x = income, y = residuals)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Residuals vs. Income", x = "Income", y = "Residuals") +
  theme_bw()
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Right of the bat, we see that most residuals are clustered on the lower end of income, and they have both positive and negative sign. As income increases, variance of residuals also grows - residuals are getting further from the zero. This means that the variability is not constant for all values of income, which is a sign supporting heteroskedasticity! Why is there a zero line? This line should reflect that on average, the model should have no bias, or in other words, the residuals should sum up to zero. For both homoskedasticity and heteroskedasticity the residuals still average around zero. The difference is that for homoskedasticity, the residuals are going to be consistently spaced around zero, while for heteroskedasticity, the spread will vary depending on the change of independent variable - which is our case.

**f.** Consider again the scatterplot in **e**. A colleague tells you that, since the slope of the regression line is zero, you do not have to worry about potential heteroskedasticity in the regression. Do you agree? Why or why not?

In linear regression model, residuals are expected to sum up to zero, thus the regression line in e) is zero. This is not an indicator of homoskedasticity, and doesn't tell us anything about the spread fo residuals. Since we clearly see that variance of residuals is not constant, and grows bigger with higher income, we conclude heteroskedasticity. It is pretty clear the residuals fan out, and if homoskedasticity would hold, they would stay evenly distributed around the line. So, our colleague is talking nonsense, and we shouldn't trust them.

*You decide that, to be on the safe side, you should run all the regressions in the remainder of the exercise using the lm_robust function and perform statistical tests with the HC2 standard errors.*

**g.** Run the regression in **d** again, this time using `lm_robust`. Your colleague points out that the coefficient on `income` did not change, and the $R^2$ is also the same as in **e**. Based on these observations, your colleague concludes that the heteroskedasticity was indeed not a problem. Do you agree? Why or why not?

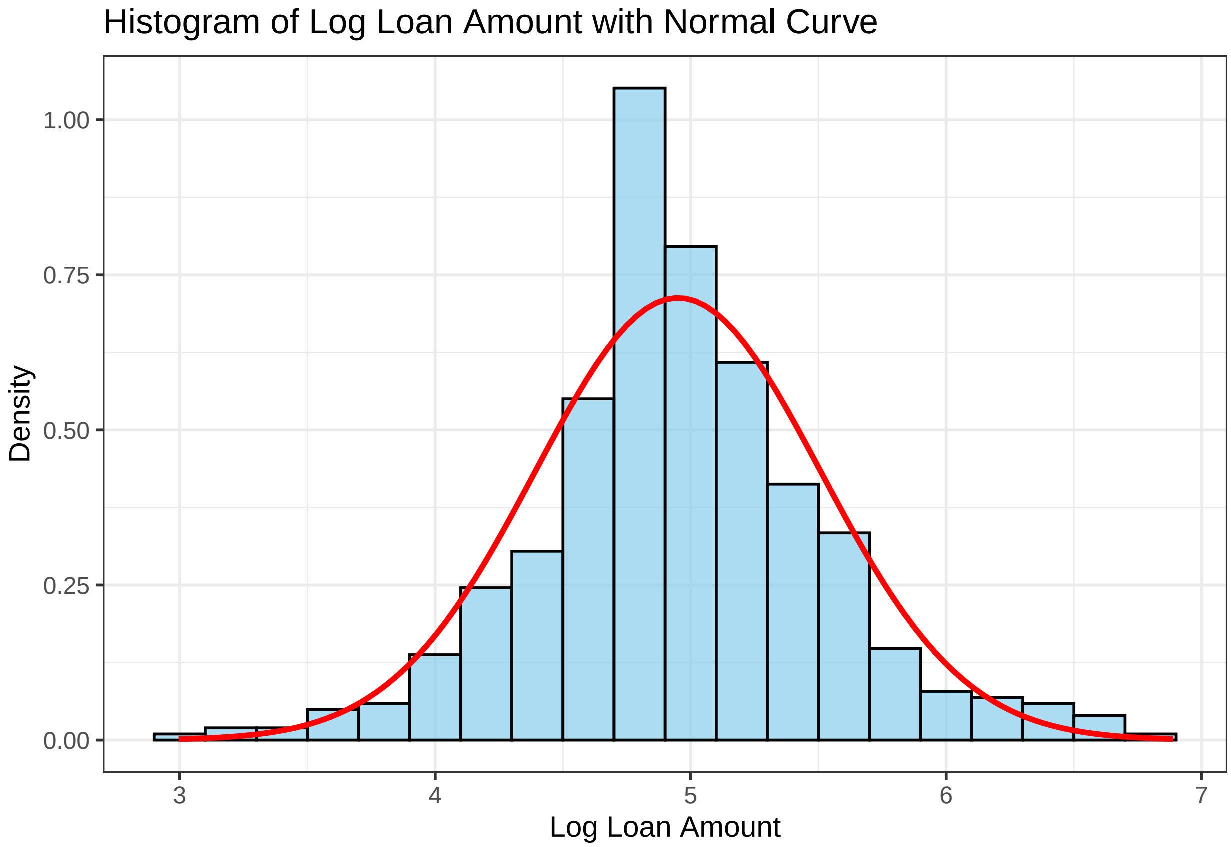
```r
# Run regression of loan_amount and income, now with lm_robust
loan_model_g <- lm_robust(loan_amount ~ income, data = loan_app, se_type = "HC2")
summary(loan_model_g)
```

```
##
## Call:
## lm_robust(formula = loan_amount ~ income, data = loan_app, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  85.3832    11.9344   7.154 2.959e-12  61.9363  108.830 507
## income        0.8873     0.1519   5.842 9.236e-09   0.5889    1.186 507
##
## Multiple R-squared:  0.4401 ,    Adjusted R-squared:  0.439
## F-statistic: 34.13 on 1 and 507 DF,  p-value: 9.236e-09
```

We do see that truly, coefficient on income $\beta_{income} = 0.8873$ and R-squared $R^2 = 85.3832$ stayed the same, like in the regression d. But what does this mean? Heteroskedasticity will not affect the coefficient, because it does not bias the estimates of the coefficients. These coefficients are calculated by minimizing sum of squared residuals, which doesn't depend on heteroskedasticity. But what does it affect? Heteroskedasticity will only affect variance-covariance matrix, and thus standard errors, which we can see in our regression - standard errors changed! Also, R-squared not changing is also normal. It measures the proportion of variance explained by the model, which can be totally unaffected by heteroskedasticity. So, we shouldn't trust our colleague, and shouldn't conclude that heteroskedasticity isn't a problem. We can clearly see this by looking at standard errors, also in p-values and t-values, and finally, it's our best bet to look at the plot of residuals, like we did.

**h.** Create a new variable, `ln_loan_amount`, which is the natural logarithm of `loan_amount`. Plot a histogram of `ln_loan_amount`. What do you observe now? Compare with the histogram of `loan_amount`.

```r
# Creating new variable
loan_app$ln_loan_amount <- log(loan_app$loan_amount)

# Then plot the histogram
ggplot(loan_app, aes(x = ln_loan_amount)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 0.2, fill = "skyblue", color = "black",
alpha = 0.7) + stat_function(fun = dnorm, args = list(mean = mean(loan_app$ln_loan_amount, na.rm = TRUE
sd = sd(loan_app$ln_loan_amount, na.rm = TRUE)), color = "red", linewidth = 1) +
  labs(title = "Histogram of Log Loan Amount with Normal Curve",
x = "Log Loan Amount", y = "Density") + theme_bw()
```



Original histogram was very right-skewed, with most values concentrating in the lower end of loan amount, and then histogram having a long tail extending to the right. The red normal curve didn't really fit nicely the data in histogram. Here we concluded that data deviates from normal distribution, which puts in danger our needed normality assumption. On the other hand, our new, log-transformed data histogram has more symmetric shape around the center. Log-transformation reduced skewness and compressed the range of loan amounts, and now, histogram pictures a nice bell-shaped distribution. We can conclude that log-transformed data better satisfies the normality distribution.

**i.** Add the natural log of `income` to the data. Call this new variable `ln_income`. Run a regression of `ln_amount` on on `ln_income`, `units`, `black`, `hispan`, `male`, and `married`. Interpret the coefficient on `ln_income`. What is the associated p-value? What is the t-statistic? What are the null and alternative hypotheses that correspond to the regression output you observe? Can you reject the null hypothesis at the 5% level?

```r
# First we add the new variable
loan_app$ln_income <- log(loan_app$income)

# Then run a regression on all needed variables
loan_model_i <-lm_robust(ln_loan_amount ~ ln_income + units + black + hispan + male + married,
                         data = loan_app, se_type = "HC2")
summary(loan_model_i)
```

```
##
## Call:
## lm_robust(formula = ln_loan_amount ~ ln_income + units + black +
##     hispan + male + married, data = loan_app, se_type = "HC2")
##
```

```
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  2.25044    0.20692 10.8757 6.992e-25  1.84389   2.6570 502
## ln_income    0.57490    0.05004 11.4893 2.757e-27  0.47659   0.6732 502
## units        0.12012    0.04782  2.5120 1.232e-02  0.02617   0.2141 502
## black        0.02824    0.05565  0.5075 6.120e-01 -0.08109   0.1376 502
## hispan       0.03345    0.05866  0.5703 5.687e-01 -0.08179   0.1487 502
## male         0.05467    0.05848  0.9350 3.503e-01 -0.06021   0.1696 502
## married      0.07242    0.04937  1.4670 1.430e-01 -0.02457   0.1694 502
##
## Multiple R-squared:  0.4268 ,    Adjusted R-squared:   0.42
## F-statistic: 36.61 on 6 and 502 DF,  p-value: < 2.2e-16
```

The coefficient on ln_income here is $\beta_{ln\_income} = 0.57490$, meaning that with every 1% increase in income, loan amount will increase by 0.57490% (since both loan amount and income are here in natural logarithm form). The p-value is $p = 2.757e-27 < 0.05$, so drastically smaller than 5% significance level. The t-statistic is $t = 11.4893 > 1.96$, so also differs drastically from it's respectable threshold. These two numbers prove that there is a strong evidence against null hypothesis, and we can safely reject the null hypothesis stating that this coefficient is not statistically significant. I other words, effect of ln_income on ln_loan_amount is indeed statistically significant. Here are written down hypotheses: `H0`:The natural logarithm of income has no effect on natural logarithm of loan amount $\rightarrow \beta_{ln\_income} = 0$. `H1`:The natural logarithm of income has an effect on natural logarithm of loan amount $\rightarrow \beta_{ln\_income} \neq 0$. Thus, we safely reject `H0`, and conduct that ln_income is important predictor of the ln_loan_amount.

**j.** Predict the `loan_amount` for a `black`, non-hispanic, `married` and `male` applicant. The applicant has an `income` of $100,000. The property comprises of 1 unit. (double check the units of measurement of the variables).

```
# First we load the data for applicant - be careful, we have to do ln for income!
applicant_data <- data.table(ln_income = log(100), units = 1, black = 1, hispan = 0, male = 1,
                             married = 1)

# Then we use the previously defined model
predicted_ln_loan_amount <- predict(loan_model_i, newdata = applicant_data)

# We can transform the loan amount to regular form, not to be in logarithm form
predicted_loan_amount <- exp(predicted_ln_loan_amount)

cat(paste("The predicted logarithm of loan amount is:", predicted_ln_loan_amount), "\n")
```

```
## The predicted logarithm of loan amount is: 5.17343093642261
```

```
cat(paste("The predicted loan amount is:", predicted_loan_amount), "\n")
```

```
## The predicted loan amount is: 176.519426667724
```

```
# I used cat(paste()) because c was adding .1 in output, which wasn't nicely readable
```

The predicted logarithm of the loan amount for an applicant with an income of $100,000, who is black, non-Hispanic, married, and male, with a property comprising of 1 unit, is approximately $176.519 thousand dollars.

**k.** Test the null hypothesis that the coefficient on `male` is the same as the coefficient on `married` using an F-test. State the null and alternative hypotheses. What are the F-statistic and the associated p-value? What do you conclude?

Let's start with forming the hypotheses: `H0`:The coefficient on `male` is the same as the coefficient on `married` $\rightarrow \beta_{male} = \beta_{married}$. `H1`:The coefficient on `male` is not the same as the coefficient on `married` $\rightarrow \beta_{male} \neq \beta_{married}$.

```r
# Do the F-test
linearHypothesis(loan_model_i, c("male = married"), test = "F", white.adjust = "HC2")
```

```
##
## Linear hypothesis test:
## male - married = 0
##
## Model 1: restricted model
## Model 2: ln_loan_amount ~ ln_income + units + black + hispan + male +
##     married
##
##   Res.Df Df      F Pr(>F)
## 1    503
## 2    502  1 0.0395 0.8426
```

From known table for F-statistic, we know that the critical value at a 5% significance level equals 3.84 for $q = 1$ restrictions in our case. Since our $F = 0.0395 < 3.84$, we can not reject our null hypothesis. Similarly, we see that our p-value is also bigger than 5%: $p = 0.8426 > 0.05$. Thus, we fail to reject the null hypothesis, meaning that there is no significant difference between the coefficients on `male` and `married`. The evidence does not support the claim that these two coefficients are different from each other.

**l.** How could test the linear restriction in point **k** using a t-test, rather than an F-test? Implement the test and verify that the p-value is the same. (Hint: you will need to run a slightly modified regression, reparametrizing the model, to accomplish this. Think of how you can algebraically manipulate the null hypothesis you wrote at the previous point to make it testable using a t-test.)

To make the restriction in point k t-test testable, we have to transform the model so that the null hypothesis involves only 1 coefficient. We will use one trick: instead of writing $\beta_{male} \cdot male + \beta_{married} \cdot married$, we will add and subtract $\beta_{married} \cdot male$, and then we have:

$$\beta_{male} \cdot male + \beta_{married} \cdot married + \beta_{married} \cdot male - \beta_{married} \cdot male = (\beta_{male} - \beta_{married}) \cdot male + \beta_{married} \cdot (male + married)$$

. We will have to create new variable $male + married$, which will be added to our new model. The initial hypothesis $\beta_{male} = \beta_{married}$ can be also written as: $\beta_{male} - \beta_{married} = 0$, and in our new model, that is the coefficient on `male`. Thus, our hypotheses for new model are: `H0`:The coefficient on `male` is equal to zero $\rightarrow \beta_{male} = 0$. `H1`:The coefficient on `male` is not equal to zero $\rightarrow \beta_{male} \neq 0$.

```r
# We add a new variable male + married
loan_app$male_plus_married <- loan_app$male + loan_app$married

# Run the regression
loan_model_l <-  lm_robust(ln_loan_amount ~ ln_income + units + black + hispan + male
  + male_plus_married , data = loan_app, se_type = "HC2")
summary(loan_model_l)
```

```
##
## Call:
```

```
## lm_robust(formula = ln_loan_amount ~ ln_income + units + black +
##     hispan + male + male_plus_married, data = loan_app, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                    Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)         2.25044    0.20692 10.8757 6.992e-25  1.84389   2.6570 502
## ln_income           0.57490    0.05004 11.4893 2.757e-27  0.47659   0.6732 502
## units               0.12012    0.04782  2.5120 1.232e-02  0.02617   0.2141 502
## black               0.02824    0.05565  0.5075 6.120e-01 -0.08109   0.1376 502
## hispan              0.03345    0.05866  0.5703 5.687e-01 -0.08179   0.1487 502
## male               -0.01775    0.08933 -0.1987 8.426e-01 -0.19325   0.1578 502
## male_plus_married  0.07242    0.04937  1.4670 1.430e-01 -0.02457   0.1694 502
##
## Multiple R-squared:  0.4268 ,    Adjusted R-squared:   0.42
## F-statistic: 36.61 on 6 and 502 DF,  p-value: < 2.2e-16
```

When we look at the p-value for the coefficient on `male` in our new model, it is actually the same as in model in question k: $p = 8.426e - 01$. So, we managed to test the linear restriction in point k using a t-test.

**m.** Include a discussion of the following threats to inference: (i) possible omitted variable bias, (ii) errors-in variables, (iii) misspecification of the functional form, and (iv) inconsistency of the OLS standard errors. For each threat, explain if it is a concern in this context and how you could address it.

(i) Omitted variable bias occurs when we don't include important relevant variable, which leads to biased and inconsistent estimates. Is this a problem here? Yes, it could be. Some potentially very important variables related to loan amount are left out. For example, credit score is a really important for getting a loan in USA. Only a good credit score will allow you to take a loan, and you need even better one to take a bigger amount. Similarly, debt status is also important - banks aren't going to approve big loans to applicant who already have debt and are struggling to pay it off (this can be seen through debt-income ratio). So yes, our model might have omitted variable bias. Of course, solution to this problem would be to explore the data and realize which variables are missing, and then to gather them and include them in our model.

(ii) Errors-in-variables occur when a variable/variables have measurement errors, which can lead to model giving biased coefficients and inconsistent standard errors. Overall, a lot of variables can be prone to this. Every variable that is hard to be measured can cause these errors, which can also happen in our model. For example, if we gathered data by people self-reporting it - people might report their income a bit off, or number of units. On the other hand, loan information is quite strict because it's coming from the banks, so we usually shouldn't be worried about those. The solution would be of course to improve the accuracy of the data by overlooking the gathering process. It would be great if we could get as many observations as possible, to reduce random errors.

(iii) Misspecification of the functional form is occurs when the relationship between dependent and independent variables is not correctly modeled. For example, this could happen if we model the relationship as linear, when in reality it's non-linear; or we miss out on some interaction terms or high-order terms. In theory, this could be a concern for our problem. We could have some interaction terms that we didn't include, or maybe the relationship between some variables is not linear. For example, relationship between income and loan amount could be non-linear - after some threshold, after which income doesn't really matter that much. Maybe banks have some upper limit on loan amounts, and couldn't approve high enough loan to be proportional with applicants income. Solving these problems may include testing different interaction terms, testing for non-linearity by doing log-transformation and including polynomial terms…

(iv) Inconsistency of the OLS standard errors occurs when standard errors aren't correctly estimated, which can happen when there is heteroskedasticity (like in our case!). This could lead to confidence intervals and significance tests being misleading. This is definitely a concern in our problem, because we concluded that heteroskedasticity indeed holds. Solution for this problem is use of robust standard errors, which we later did use.