CS130 Assignment 3 Causal Inference

Preliminaries

In [2]:

```
# Question 1
data <- read.csv('House_Rent_Dataset.csv', stringsAsFactors = TRUE)
head(data)

# extract number of houses in Hyderabad
cat('Houses in Hyderabad to rent:', nrow(data[data$City == 'Hyderabad',]))
```

A data.frame: 6 × 12

	Posted.On	BHK	Rent	Size	Floor	Area.Type	Area.Locality	City	Furnishing.Statu
	<fct>	<int>	<int>	<int>	<fct>	<fct>	<fct>	<fct>	<fct
1	2022-05-18	2	10000	1100	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnishe
2	2022-05-13	2	20000	800	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnishe
3	2022-05-16	2	17000	1000	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnishe
4	2022-07-04	2	10000	800	1 out of 2	Super Area	Dumdum Park	Kolkata	Unfurnishe
5	2022-05-09	2	7500	850	1 out of 2	Carpet Area	South Dum Dum	Kolkata	Unfurnishe
6	2022-04-29	2	7000	600	Ground out of 1	Super Area	Thakurpukur	Kolkata	Unfurnishe

```
Houses in Hyderabad to rent: 868
```

In [3]:

```
# Question 2
summary(data$Rent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1200   10000   16000   34993   33000 3500000 
```

In [4]:

```
# Question 3
summary(data$Point.of.Contact)
```

Contact Agent: 1529 **Contact Builder:** 1 **Contact Owner:** 3216

In [5]:

```
# Question 4

# removing Contact Builder (not treatment nor control group)
data <- subset(data, Point.of.Contact != "Contact Builder")

# change control and treatment to 0 and 1
data$Point.of.Contact <- ifelse(data$Point.of.Contact == "Contact Owner", 0, 1)
head(data)
```

A data.frame: 6 × 12

	Posted.On	BHK	Rent	Size	Floor	Area.Type	Area.Locality	City	Furnishing.Statu
	<fct>	<int>	<int>	<int>	<fct>	<fct>	<fct>	<fct>	<fct
1	2022-05-18	2	10000	1100	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnishe
2	2022-05-13	2	20000	800	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnishe
3	2022-05-16	2	17000	1000	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnishe
4	2022-07-04	2	10000	800	1 out of 2	Super Area	Dumdum Park	Kolkata	Unfurnishe
5	2022-05-09	2	7500	850	1 out of 2	Carpet Area	South Dum Dum	Kolkata	Unfurnishe
6	2022-04-29	2	7000	600	Ground out of 1	Super Area	Thakurpukur	Kolkata	Unfurnishe

In [6]:

```
# Question 5
data$Furnishing.Status <- ifelse(data$Furnishing.Status == "Furnished", 2,
                                 ifelse(data$Furnishing.Status == "Semi-Furnished", 1, 0)
head(data)
```

A data.frame: 6 × 12

	Posted.On	BHK	Rent	Size	Floor	Area.Type	Area.Locality	City	Furnishing.Statu
	<fct>	<int>	<int>	<int>	<fct>	<fct>	<fct>	<fct>	<db
1	2022-05-18	2	10000	1100	Ground out of 2	Super Area	Bandel	Kolkata	
2	2022-05-13	2	20000	800	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	
3	2022-05-16	2	17000	1000	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata	
4	2022-07-04	2	10000	800	1 out of 2	Super Area	Dumdum Park	Kolkata	
5	2022-05-09	2	7500	850	1 out of 2	Carpet Area	South Dum Dum	Kolkata	
6	2022-04-29	2	7000	600	Ground out of 1	Super Area	Thakurpukur	Kolkata	

In [7]:

```
set.seed(130)
```

Analysis

Question 1

In [8]:

```
X <- cbind(data$BHK, data$Size, data$Furnishing.Status, data$Bathroom)

lm1 <- lm(Rent ~ X + Point.of.Contact, data = data)
summary(lm1)
```

```
Call:
lm(formula = Rent ~ X + Point.of.Contact, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-144200  -17909   -5617   13849 3381277 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -41441.877   2825.102 -14.669  < 2e-16 ***
X1                  -18.486   2039.183  -0.009 0.992767    
X2                   24.929      2.421  10.299  < 2e-16 ***
X3                 5016.855   1459.151   3.438 0.000591 ***
X4                18622.872   2052.565   9.073  < 2e-16 ***
Point.of.Contact  37013.050   2224.058  16.642  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 67310 on 4739 degrees of freedom
Multiple R-squared:  0.2583,	Adjusted R-squared:  0.2575 
F-statistic: 330.1 on 5 and 4739 DF,  p-value: < 2.2e-16
```

The coefficients of the linear regression can be interpreted as follows:

- Intercept: when the value of all predictors are 0 (i.e. no bedroom/hall/kitchen, 0 size, unfurnished, no bathrooms, and contact owner), we would expect the rent to be -41,441.9.
- *BHK*: a one unit increase in the number of bedrooms, halls and kitchens is associated with a decrease in rent of -18.5.
- *Size*: a one unit increase in the size of a rental unit is associated with an increase in rent of 24.9.
- *Furnishing.Status*: a one unit increase in furnishing status (i.e. unfurnished to semi-furnished, or semi-furnished to furnished) is associated with an increase in rent of 5016.9.
- *Bathroom*: a one unit increase in the number of bathrooms is associated with an increase in rent of 18,622.9.
- *Point.of.Contact*: a one unit increase in point of contact (from control/contact owner to treatment/contact agent) is associated with an increase in rent of 37,013.1.

The p-values for all of these coefficients (the intercept, size, furnishing, bathrooms and point of contact) are all statistically significant (all less than 0.001) except BHK, which is not statistically significant (0.99). This means that we can be fairly certain that size, furnishing, bathrooms and point of contact are all predictors of rent with these coefficients, but for BHK, we can't be sure.

The adjusted R^2 value of 0.2575 means that 25.75% of the variation in the rent can be explained by the predictors (BHK, size, furnishing, bathrooms, and point of contact). The remaining 74.25% of variation in rent is explained by other factors not included in this regression.

Question 2

In [22]:

```
library(rgenoud)
library(Matching)
```

In [10]:

```
# genetic matching without Y
genout <- GenMatch(X = X, Tr = data$Point.of.Contact, pop.size = 200)
matchout.gen <- Match(X = X, Tr = data$Point.of.Contact, Weight.matrix=genout)
mout <- MatchBalance(Point.of.Contact ~ BHK + Size + Furnishing.Status + Bathroom,
                     data = data, match.out = matchout.gen, nboots = 500)
```

```


Fri Apr 07 15:38:42 2023
Domains:
 0.000000e+00   <=  X1   <=    1.000000e+03
 0.000000e+00   <=  X2   <=    1.000000e+03
 0.000000e+00   <=  X3   <=    1.000000e+03
 0.000000e+00   <=  X4   <=    1.000000e+03

Data Type: Floating Point
Operators (code number, name, population)
        (1) Cloning...........................  22
        (2) Uniform Mutation..................  25
        (3) Boundary Mutation.................  25
        (4) Non-Uniform Mutation..............  25
        (5) Polytope Crossover................  25
        (6) Simple Crossover..................  26
        (7) Whole Non-Uniform Mutation........  25
        (8) Heuristic Crossover...............  26
```

a. Before matching, the mean treatment value of BHK was 2.36, and the mean control was 1.95 (a difference of 0.41). After matching, the mean treatment value of BHK was 2.36, and the mean control was 2.37 (a difference of only 0.01). This means that there is more balance in the BHK variable after matching than there was before. Here, we only consider the difference in means before and after matching. To better analyse the differences in BHK before and after matching, we should ideally also do a KS test in which we find the difference between the control and treatment BHK distributions over their entire support.

In [11]:

```
mout$AMsmallestVarName
mout$AMsmallest.p.value
```

'Furnishing.Status'

0.317310585540041

b. The smallest p-value after matching is 0.317 for Furnishing.Status. This means that the 'worst-balanced' variable after matching is furnishing.

c. Ideally, we want large p-values after matching. A small p-value indicates that the difference between the control and treatment distributions is statistically significant, and a large p-value indicates that the difference is not statistically significant. Since we want the distributions of control and treatment values for each variable to be as similar as possible, we do not want the difference between them to be statistically significant, and thus, we desire a large p-value.

Question 3

In [12]:

```
# genetic matching with Y
genout <- GenMatch(X = X, Tr = data$Point.of.Contact, pop.size = 200)
matchout.gen <- Match(X = X, Tr = data$Point.of.Contact, Y = data$Rent,
                      Weight.matrix=genout)
mout <- MatchBalance(Point.of.Contact ~ BHK + Size + Furnishing.Status + Bathroom,
                     data = data, match.out = matchout.gen, nboots = 500)
```

```


Fri Apr 07 15:43:15 2023
Domains:
 0.000000e+00   <=  X1   <=    1.000000e+03
 0.000000e+00   <=  X2   <=    1.000000e+03
 0.000000e+00   <=  X3   <=    1.000000e+03
 0.000000e+00   <=  X4   <=    1.000000e+03

Data Type: Floating Point
Operators (code number, name, population)
        (1) Cloning...........................  22
        (2) Uniform Mutation..................  25
        (3) Boundary Mutation.................  25
        (4) Non-Uniform Mutation..............  25
        (5) Polytope Crossover................  25
        (6) Simple Crossover..................  26
        (7) Whole Non-Uniform Mutation........  25
        (8) Heuristic Crossover...............  26
```

In [13]:

```
summary(matchout.gen)
```

```
Estimate...  46351
AI SE......  4309.1
T-stat.....  10.757
p.val......  < 2.22e-16

Original number of observations..............  4745
Original number of treated obs...............  1529
Matched number of observations...............  1529
Matched number of observations  (unweighted).  15054
```

a. The estimated treatment effect is 47,046.19. This means that we would predict rent to be 47,046.17 more for a property where the point of contact is an agent as opposed to the owner (given that all other predictors are the same between the two properties).

b. In this case, all the treatment examples are matched (1529 are matched out of 1529). Because we have not set a caliper (a limit on the distance between a treatment sample and a matched control sample), there may be some treatment samples that are matched to control samples far away.

Question 4

In [14]:

```
# create new dataset from treated/non-treated matches
data_matched <- data[c(matchout.gen$index.treated, matchout.gen$index.control), ]
head(data_matched)
matched_weights <- c(matchout.gen$weights, matchout.gen$weights)

lm2 <- lm(Rent ~ BHK + Size + Furnishing.Status + Bathroom + Point.of.Contact,
          data = data_matched, weights = matched_weights)
summary(lm2)
```

A data.frame: 6 × 12

	Posted.On	BHK	Rent	Size	Floor	Area.Type	Area.Locality	City	Furnishing.Sta
	<fct>	<int>	<int>	<int>	<fct>	<fct>	<fct>	<fct>	<d
7	2022-06-21	2	10000	700	Ground out of 4	Super Area	Malancha	Kolkata	
7.1	2022-06-21	2	10000	700	Ground out of 4	Super Area	Malancha	Kolkata	
7.2	2022-06-21	2	10000	700	Ground out of 4	Super Area	Malancha	Kolkata	
7.3	2022-06-21	2	10000	700	Ground out of 4	Super Area	Malancha	Kolkata	
7.4	2022-06-21	2	10000	700	Ground out of 4	Super Area	Malancha	Kolkata	
7.5	2022-06-21	2	10000	700	Ground out of 4	Super Area	Malancha	Kolkata	

```
Call:
lm(formula = Rent ~ BHK + Size + Furnishing.Status + Bathroom + 
    Point.of.Contact, data = data_matched, weights = matched_weights)

Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-224416   -4681    -294    3294 3380434 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -5.741e+04  1.540e+03 -37.285   <2e-16 ***
BHK               -9.261e+03  1.002e+03  -9.241   <2e-16 ***
Size               2.478e+01  9.918e-01  24.980   <2e-16 ***
Furnishing.Status  8.630e+03  7.016e+02  12.300   <2e-16 ***
Bathroom           2.929e+04  9.465e+02  30.942   <2e-16 ***
Point.of.Contact   4.633e+04  9.637e+02  48.072   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26650 on 30102 degrees of freedom
Multiple R-squared:  0.2388,	Adjusted R-squared:  0.2387 
F-statistic:  1889 on 5 and 30102 DF,  p-value: < 2.2e-16
```

The coefficients of the linear regression can be interpreted as follows:

- Intercept: when the value of all predictors are 0 (i.e. no bedroom/hall/kitchen, 0 size, unfurnished, no bathrooms, and contact owner), we would expect the rent to be -57,410. This value has become more negative after matching (decrease of 15,968). This means that our prediction of rent in the case that all other predictors are 0 is now more negative.
- BHK: a one unit increase in the number of bedrooms, halls and kitchens is associated with a decrease in rent of -9,261. This value has become several orders of magnitude more negative after matching (decrease of 9,243). This means that a one unit increase in BHK is now associated with a much greater decrease in rent than before.
- Size: a one unit increase in the size of a rental unit is associated with an increase in rent of 24.78. This value has decreased very slightly after matching (by 0.1). This means that a one unit increase in size is now associated with a (very slightly) smaller increase in rent than before.
- Furnishing.Status: a one unit increase in furnishing status (i.e. unfurnished to semi-furnished, or semi-furnished to furnished) is associated with an increase in rent of 8,630. This value has increased after matching (by 3,613). This means that a one unit increase in furnishing status is now associated with a greater increase in rent than before.
- Bathroom: a one unit increase in the number of bathrooms is associated with an increase in rent of 29,290. This value has increased after matching (by 10,667). This means that a one unit increase in the number of bathrooms is now associated with a much greater increase in rent than before.
- Point.of.Contact: a one unit increase in point of contact (from control/contact owner to treatment/contact agent) is associated with an increase in rent of 46,330. This value has increased after matching (by 9,317). This means that the estimated treatment effect is now greater than it was before matching.

The p-values for all of these coefficients (the intercept, BHK, size, furnishing, bathrooms and point of contact) are all highly statistically significant (all pretty much 0). This means that we can be very certain that BHK, size, furnishing, bathrooms and point of contact are all predictors of rent with the above coefficients.

The adjusted R^2 value of 0.2387 means that 23.87% of the variation in the rent can be explained by the predictors (BHK, size, furnishing, bathrooms, and point of contact). The remaining 76.13% of variation in rent is explained by other factors not included in this regression.

Question 5

In [23]:

```
library(sensemakr)
```

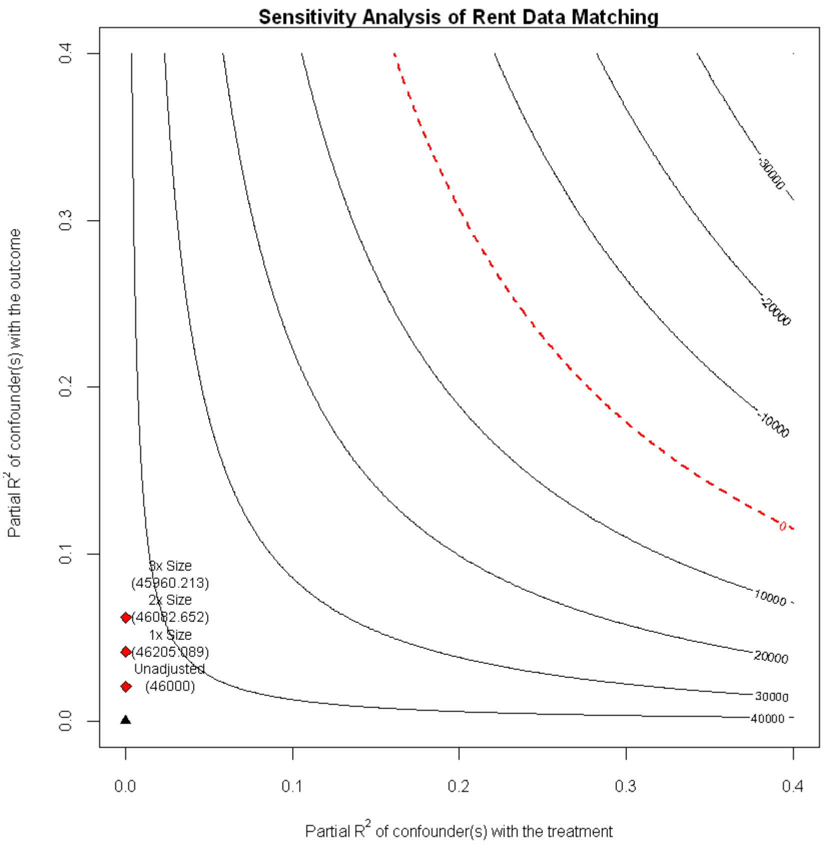
In [16]:

```
# sensitivity analysis using size as a benchmark
sensitivity <- sensemakr(model = lm2, treatment = 'Point.of.Contact',
                                benchmark_covariates = 'Size', kd = 1:3)
print(sensitivity)
```

```
Sensitivity Analysis to Unobserved Confounding

Model Formula: Rent ~ BHK + Size + Furnishing.Status + Bathroom + Point.o
f.Contact

Null hypothesis: q = 1 and reduce = TRUE

Unadjusted Estimates of ' Point.of.Contact ':
  Coef. estimate: 46327.52
  Standard Error: 963.7053
  t-value: 48.07229

Sensitivity Statistics:
  Partial R2 of treatment with outcome: 0.0713
  Robustness Value, q = 1 : 0.24134
  Robustness Value, q = 1 alpha = 0.05 : 0.2328

For more information, check summary.
```

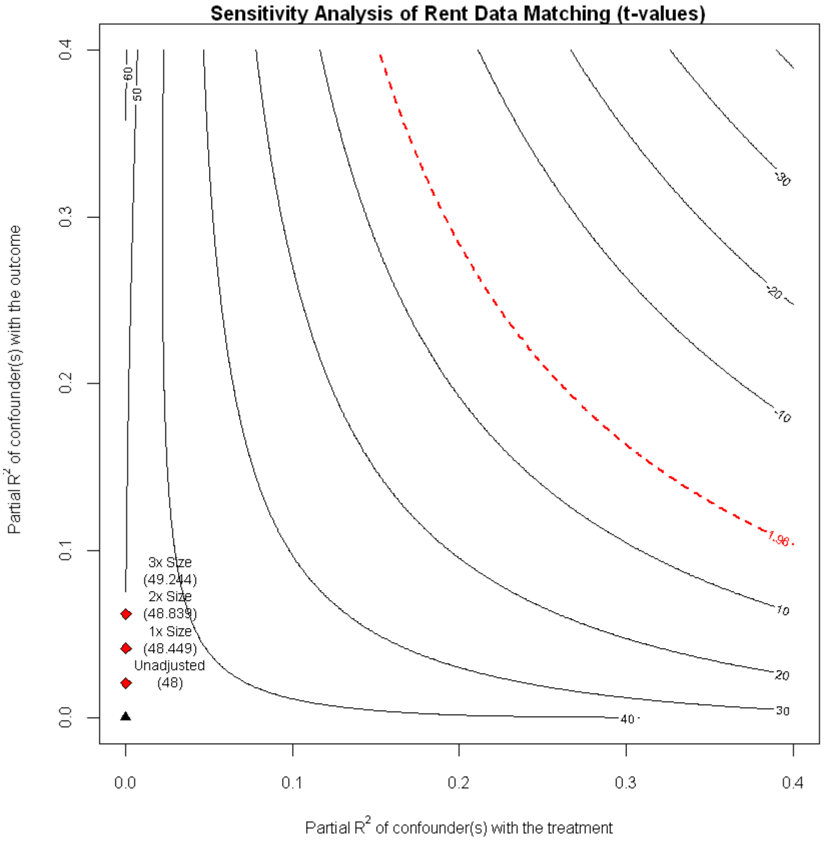
In [17]:

```
plot(sensitivity, main = 'Sensitivity Analysis of Rent Data Matching')
```



In [18]:

```
plot(sensitivity, main = 'Sensitivity Analysis of Rent Data Matching (t-values)', sensit
```



The robustness value for setting the *Point.of.Contact* to zero ($RV_{q=1}$) is 24.1%. This indicates that unobserved confounders that account for 24.1% of the residual variance in both the *Point.of.Contact* and *Rent* outcome variables are strong enough to fully explain the observed effect. Conversely, unobserved confounders that account for less than 24.1% of the residual variance in both *Rent* and *Point.of.Contact* variables are not strong enough to explain away the observed effect. This indicates that our result is robust, since we require unobserved confounders to be so highly correlated with both *Point.of.Contact* and *Rent* for the treatment effect to be otherwise explained.

The value of robustness when testing the null hypothesis that the coefficient of the variable denoted by *Point.of.Contact* is zero ($RV_{q=1,\alpha=0.05}$) is 23.3%, which is slightly lower than the previous value mentioned. This indicates that unobserved factors that account for 23.3% of the residual variance in both *Point.of.Contact* treatment and the *Rent* outcome are strong enough to bring the lower limit of the confidence interval to zero (at the chosen significance level of 5%). Conversely, factors that account for less than 23.3% of the residual variance in both the *Point.of.Contact* treatment and the *Rent* outcome are not strong enough to achieve this. This also indicates that our result is robust, since we require unobserved confounders to be highly correlated with both *Point.of.Contact* and *Rent* for the treatment effect to be otherwise explained at the 5% level of significance.

Moreover, the partial R^2 coefficient between *Point.of.Contact* and *Rent* suggests that in an extreme scenario where we assume that unobserved factors account for all the omitted variation in the *Rent* outcome, these factors should explain at least 7.13% of the residual variance in *Point.of.Contact* in order to fully explain the observed effect. This also supports the fact that our results are robust, since we would require an unobserved confounder to not only explain the remaining 76.13% of variation in *Rent* (as above), but also 7.13% of variation in *Point.of.Contact*.

From the first plot, we see that, even if we have a variable (or aggregate of multiple unobserved confounders) that is thrice as correlated with the *Rent* and thrice as correlated with *Point.of.Contact* as *Size*, there is still minimal impact on the treatment effect. The lowest possible treatment effect in this case is 45,960, only 368 less than our estimated treatment effect of 46,328. Given that our treatment effect is so large, a change of 368 is not very significant. This indicates that our result is very robust as, even with strong unobserved confounders, our treatment effect changes very minimally.

From the second plot, we see that the t-value does not change very much in the case that we have unobserved confounders that are thrice as correlated with the *Rent* and thrice as correlated with *Point.of.Contact* as *Size*. It only changes from 48.4 to 49.2 (0.8 increase). These values are all well below above the statistically-significant boundary of 1.96. Thus, we are very certain about the treatment effects in the first plot.

Extra Credit

In [24]:

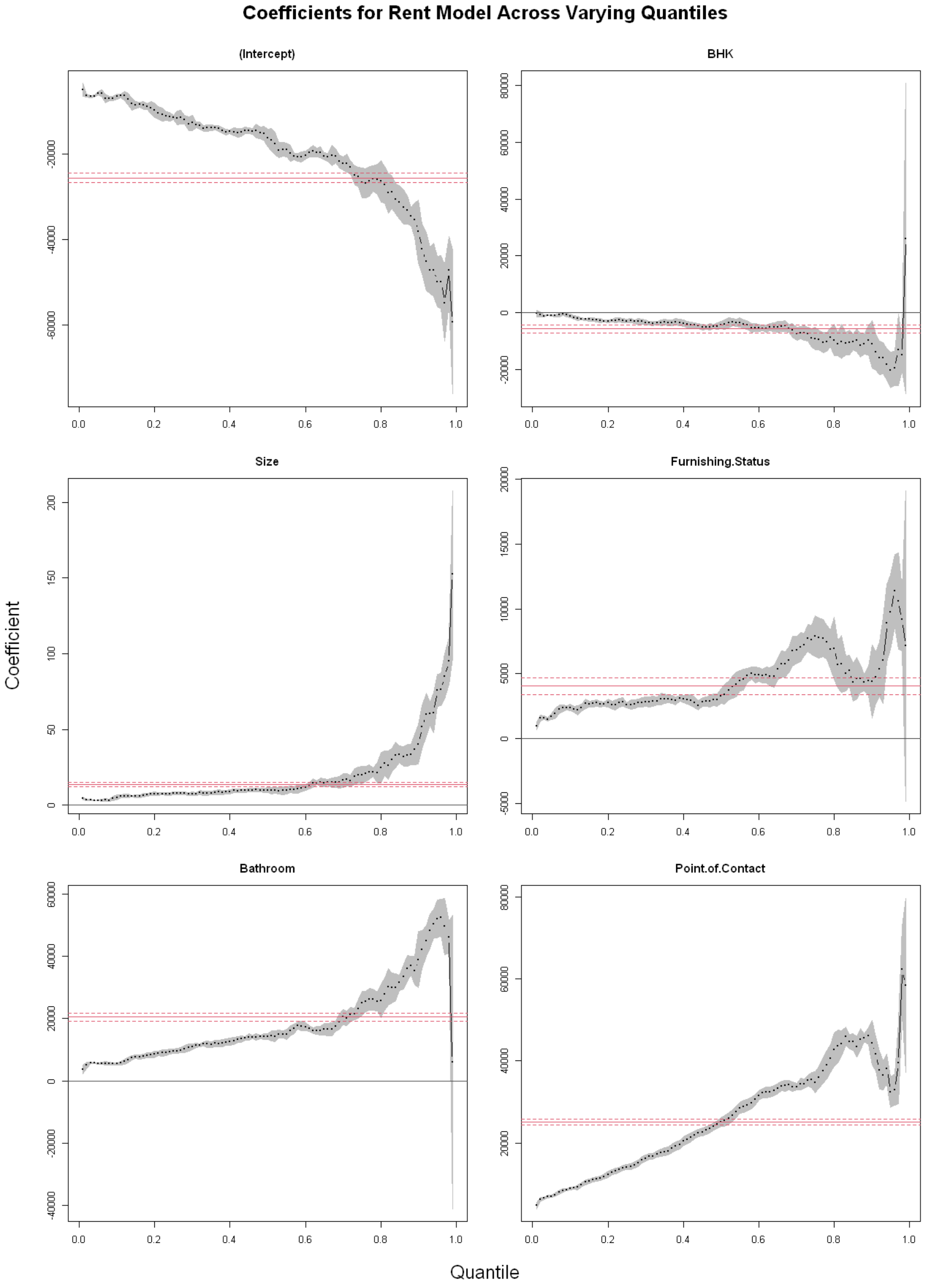
```
library(quantreg)
```

In [20]:

```
qreg <- suppressWarnings(summary(rq(Rent ~ BHK + Size + Furnishing.Status +
                                    Bathroom + Point.of.Contact, tau = 1:99/100,
                                    weights = matched_weights, data = data_matched)))
```

In [21]:

```
# increase plot size
options(repr.plot.width = 11, repr.plot.height = 15)

# position subplots
par(mfrow = c(3, 2), mar = c(4, 4, 2, 1), oma = c(4, 4, 4, 1))
plot(qreg)

# add title and axis labels
title(main = 'Coefficients for Rent Model Across Varying Quantiles',
      outer = TRUE, cex.main = 2)
mtext('Quantile', side = 1, line = 1, outer = TRUE, cex = 1.3)
mtext('Coefficient', side = 2, line = 1, outer = TRUE, cex = 1.3)
```



We see very different results when we specify specific quantiles. For all five predictors, the least-squares coefficients and intercept (seen in the plot as a solid red line, with uncertainty marked as a dashed red line) do not capture the information across different quantiles; for all variables, the coefficient values lie outside the red boundaries across the majority of quantiles. For example, for the point of contact variable, the least-squares range only agrees with the quantiles at ~0.48-0.52. For quantiles smaller than 0.48 and those larger than 0.52, the coefficient varies significantly from ~500 at quantile 0.01 to ~6000 at quantile 0.99. This range shows that the treatment effect varies across different quantiles in the data.

This analysis gives us a more holistic view of the regression.

AI tools note: I did not use any AI tools to complete this assignment.