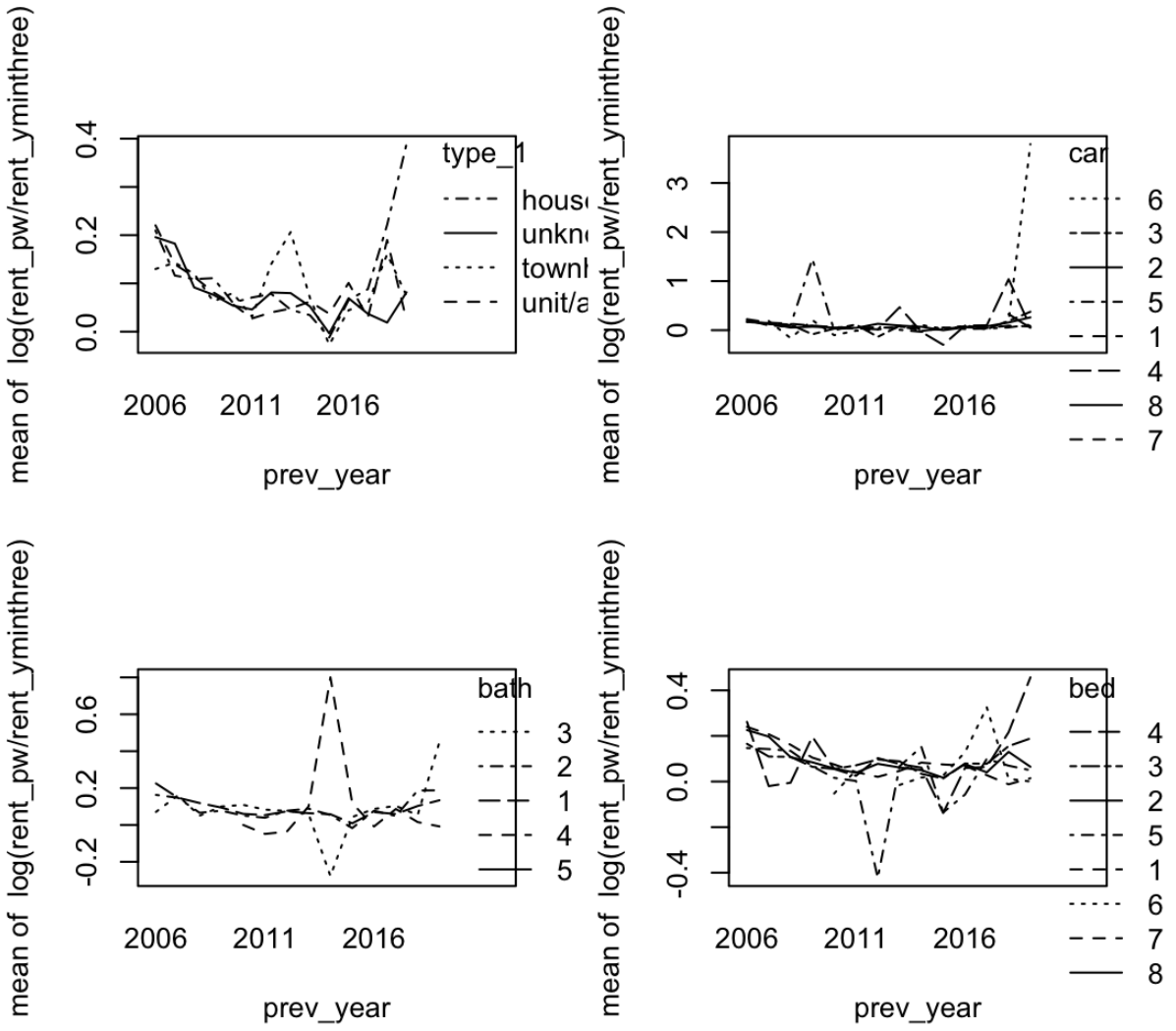
To start with, the response variable is essentially this year's listed rent that is dependent on rent listed three years ago. To deal with this dependencies, this problem can be modelled using a log link via $\mathbf{\log(\lambda_i/t_i) = x_i^T\beta \Rightarrow \log(\lambda_i) = \log(t_i) + x_i^T\beta}$ where $\lambda_i$ in this case represent this year's listed rent, $t_i$ represents rent listed three years ago, $x_i$ are known predictors and $\beta$ unknown parameters. This is called a rate model. In an R model description we can fix the coefficient of a variable to 1 by enclosing it in the offset function, viz y ~ _offset(log(t)) + x1 + x2 + · · · [source: Modern Applied Statistics tutorial sheet week 5 question 2].

Next, we want to plot the interaction between the variables



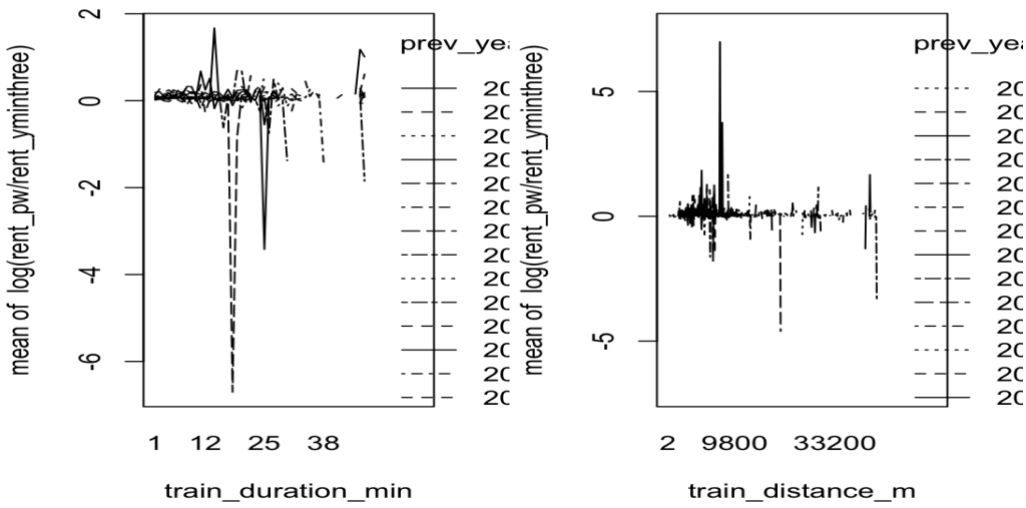
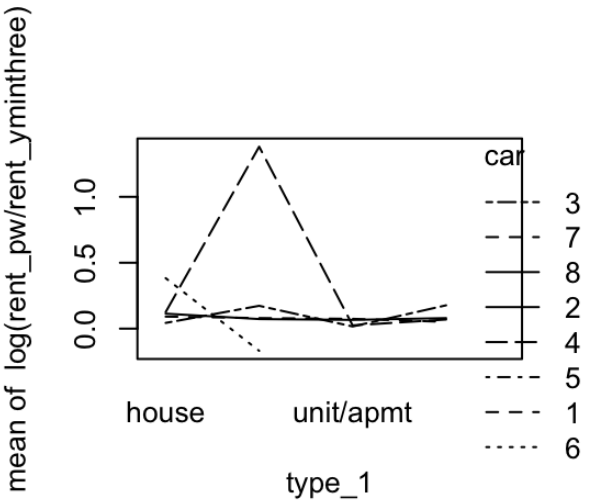
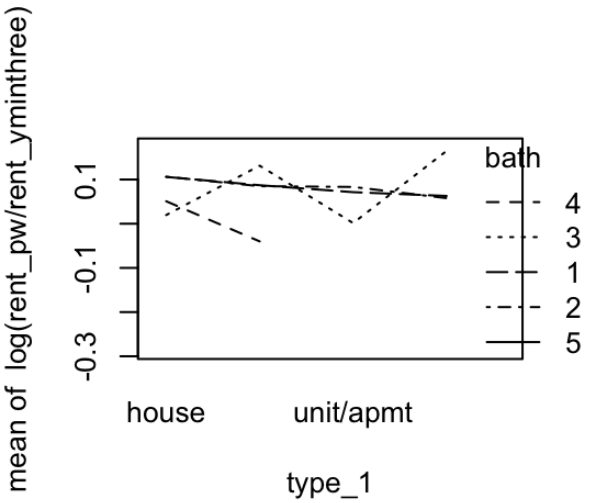
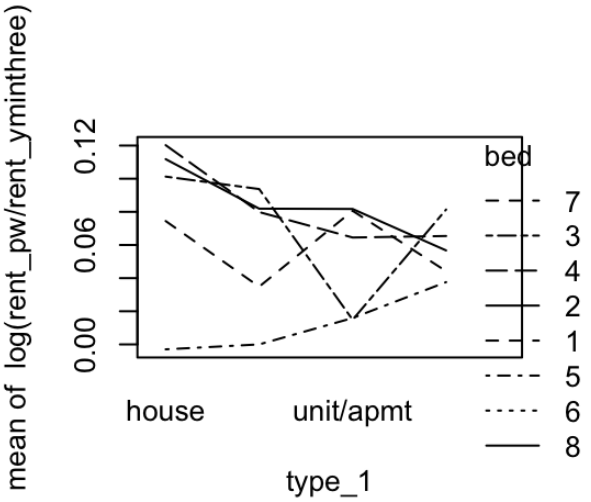
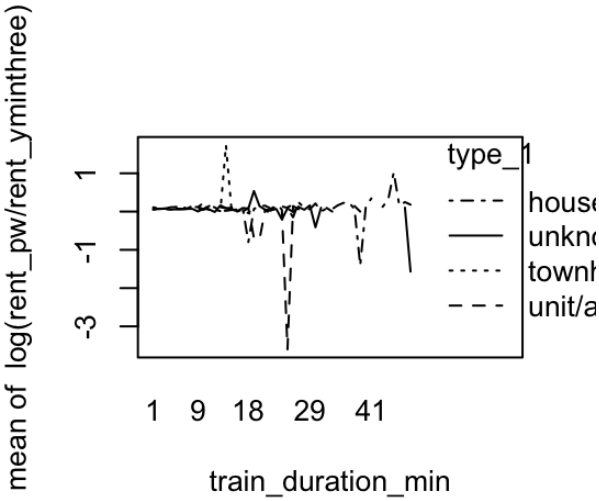



Figure 1 : Beginning Year interactions with other variables


mean of log(rent_pw/rent_yminthree)
0.0
0.5
1.0
car
3
7
8
2
4
5
1
6
house
unit/apmt
type_1


mean of log(rent_pw/rent_yminthree)
-0.3
-0.1
0.1
bath
4
3
1
2
5
house
unit/apmt
type_1


mean of log(rent_pw/rent_yminthree)
0.00
0.06
0.12
bed
7
3
4
2
1
5
6
8
house
unit/apmt
type_1


mean of log(rent_pw/rent_yminthree)
-3
-1
1
type_1
house
unkno
townh
unit/a
1
9
18
29
41
train_duration_min

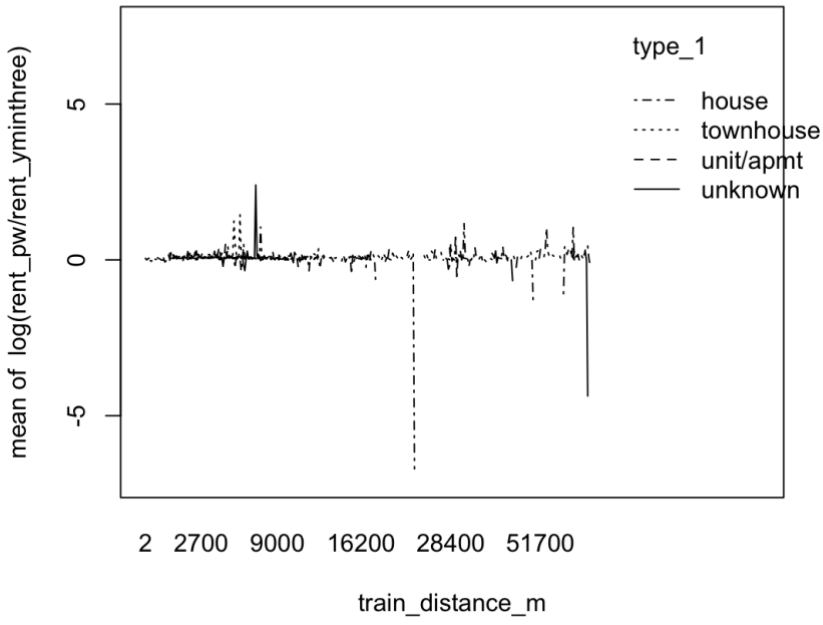

mean of log(rent_pw/rent_yminthree)
-5
0
5
type_1
house
townhouse
unit/apmt
unknown
2
2700
9000
16200
28400
51700
train_distance_m

Figure 2 : Property type interactions with other variables

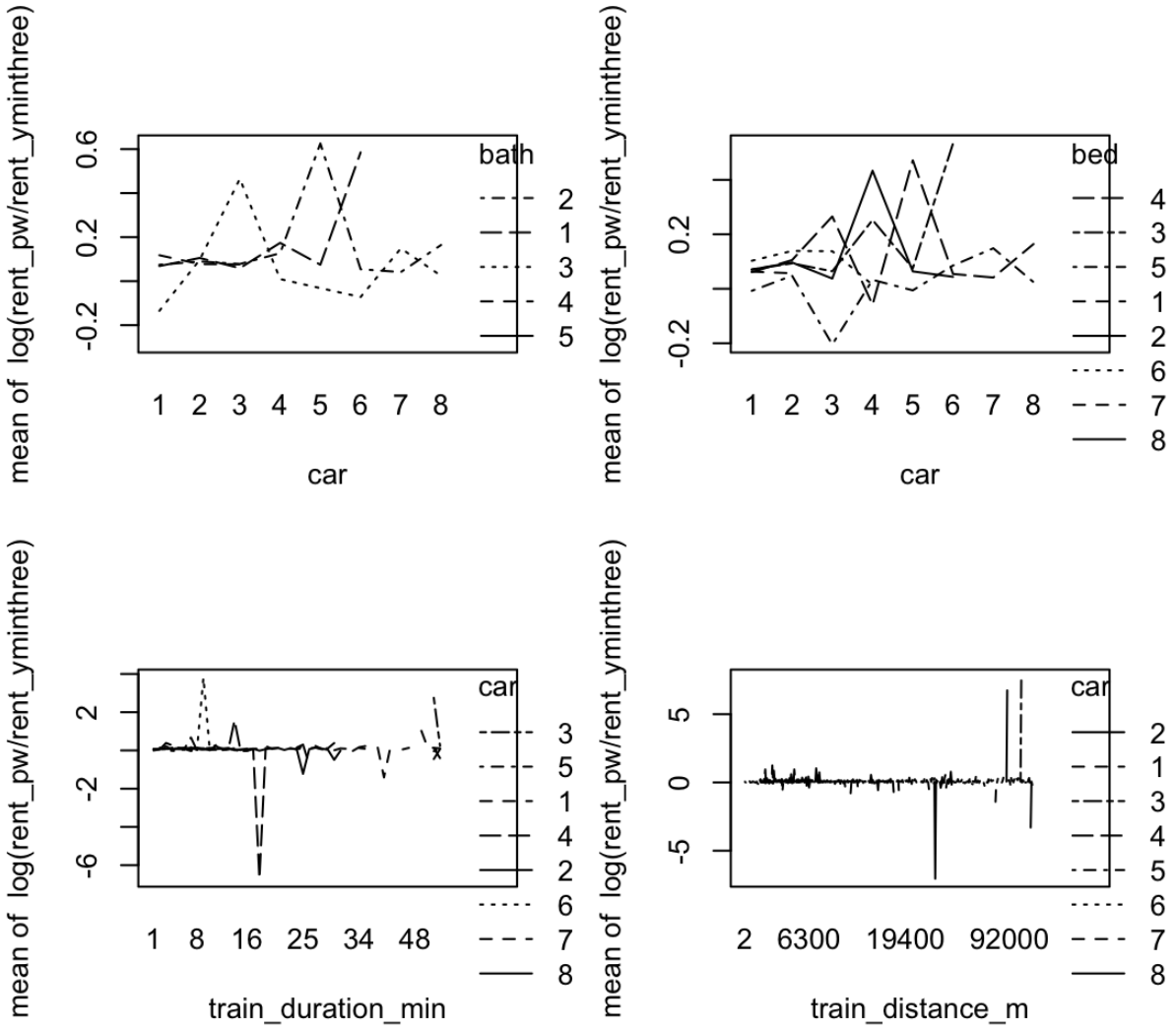



Figure 3 : Number of garage / car space interactions with other variables

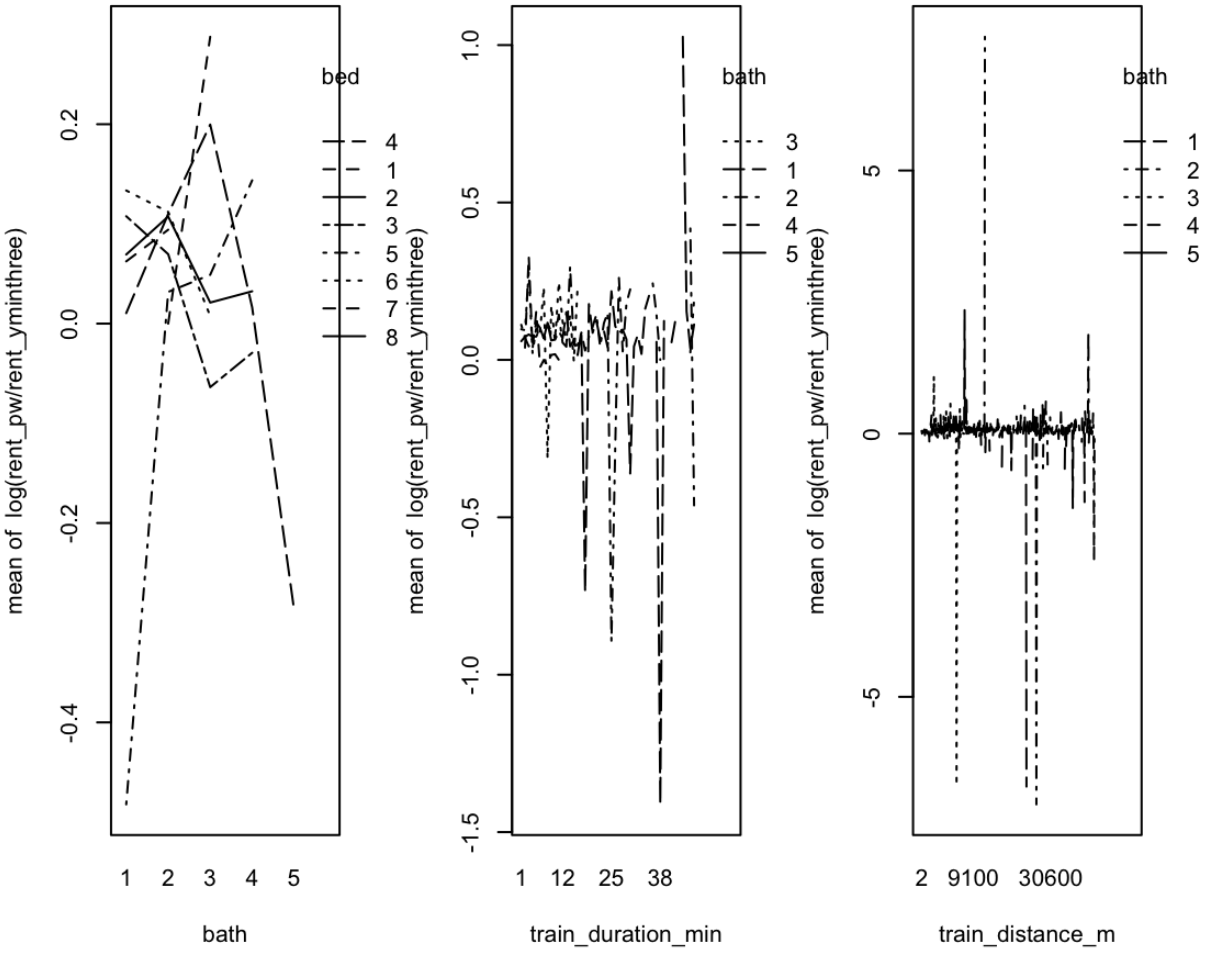



Figure 4 : Number of Bathroom interactions with other variables

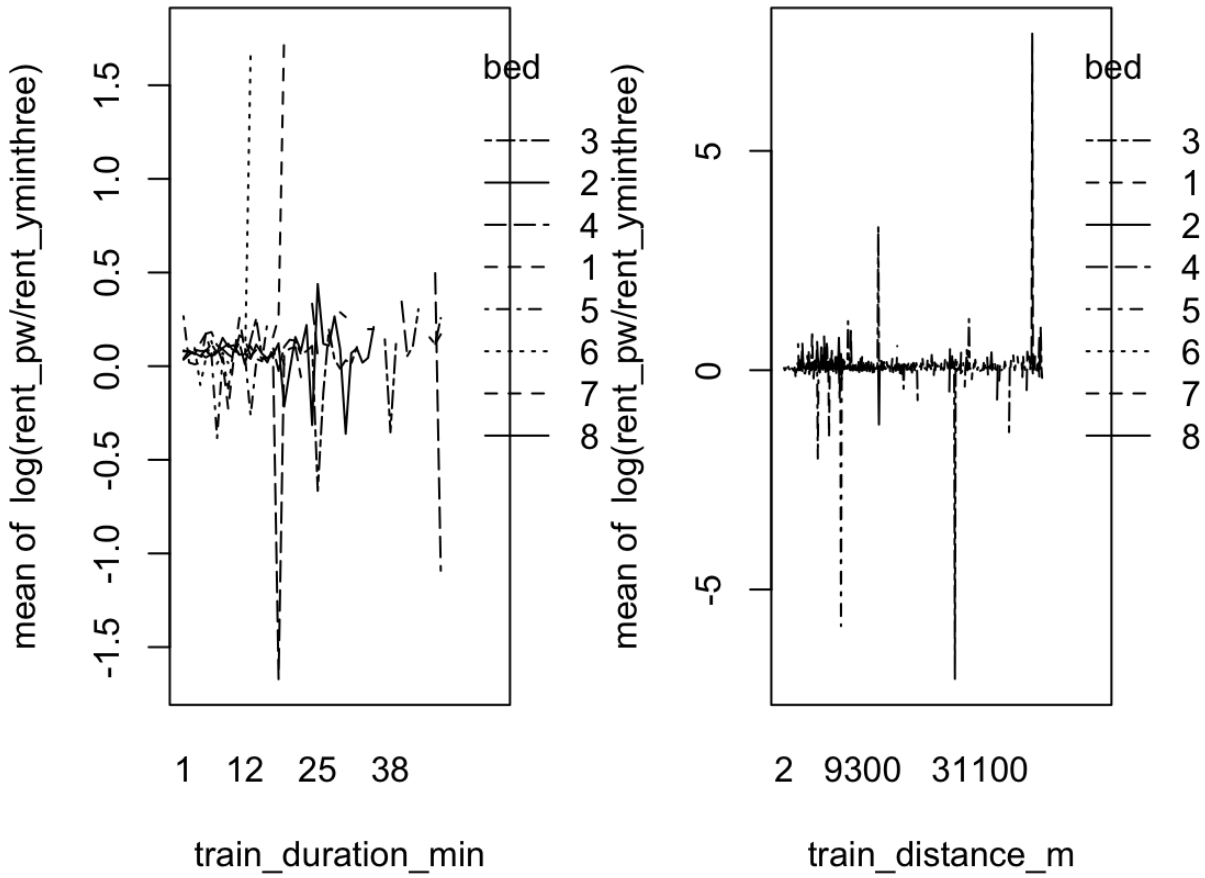



Figure 5 : Number of bedroom interactions with other variables

From these figures, the plots seem to show that there are indeed big interactions between these variables. Therefore, it is worth including these interactions to the base model so that the significance of these interaction can be tested. To test it, stepwise selection would be used to select the significant features from the base model.

```
> model <- lm(rent_pw ~ offset(log(rent_yminthree)) +
+                 ( prev_year+car + bath + bed + train_duration_min + train_distance_m)^2
+             ,data=data);
> mod2 <- step(model, scope=~.,trace = FALSE)
> summary(mod2)

Call:
lm(formula = rent_pw ~ prev_year + car + bath + bed + train_duration_min +
    train_distance_m + prev_year:car + prev_year:bath + prev_year:bed +
    prev_year:train_distance_m + car:bed + bed:train_duration_min +
    bed:train_distance_m + train_duration_min:train_distance_m +
    offset(log(rent_yminthree)), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-148570   -3417   -1248     -81 1339698

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1.756e+06  3.976e+05   4.416 1.01e-05 ***
prev_year                         -8.751e+02  1.974e+02  -4.432 9.37e-06 ***
car                               -1.169e+06  1.946e+05  -6.010 1.87e-09 ***
bath                              -4.875e+05  2.811e+05  -1.734  0.08287 .
bed                               -3.302e+05  1.749e+05  -1.888  0.05901 .
train_duration_min                 1.063e+03  3.340e+02   3.182  0.00146 **
train_distance_m                   8.667e+01  1.233e+01   7.026 2.17e-12 ***
prev_year:car                      5.828e+02  9.663e+01   6.031 1.64e-09 ***
prev_year:bath                     2.423e+02  1.396e+02   1.736  0.08259 .
prev_year:bed                      1.648e+02  8.684e+01   1.898  0.05774 .
prev_year:train_distance_m        -4.385e-02  6.115e-03  -7.172 7.59e-13 ***
car:bed                           -7.297e+02  3.408e+02  -2.141  0.03227 *
bed:train_duration_min            -2.887e+02  1.183e+02  -2.440  0.01469 *
bed:train_distance_m               4.545e-01  7.913e-02   5.744 9.36e-09 ***
train_duration_min:train_distance_m 3.507e-03  5.939e-04   5.906 3.54e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37330 on 30268 degrees of freedom
Multiple R-squared:  0.01727,   Adjusted R-squared:  0.01681
F-statistic: 37.98 on 14 and 30268 DF,  p-value: < 2.2e-16
```

Figure 6: Features selected using stepwise function

It is clear from figure 6 that all of the features scrapped from the web are indeed significant. On top of that a handful of the interactions between these features are useful too. These findings are also supported by the ANOVA of the original model shown on figure 7.

```
> anova(model, test="Chi")
Analysis of Variance Table

Response: rent_pw
                                    Df     Sum Sq    Mean Sq F value    Pr(>F)
prev_year                            1 1.3274e+11 1.3274e+11 95.2367 < 2.2e-16 ***
car                                  1 9.5720e+10 9.5720e+10 68.6777 < 2.2e-16 ***
bath                                 1 8.2734e+09 8.2734e+09  5.9360  0.014840 *
bed                                  1 3.8242e+09 3.8242e+09  2.7438  0.097641 .
train_duration_min                   1 7.2400e+10 7.2400e+10 51.9462 5.838e-13 ***
train_distance_m                     1 1.1091e+10 1.1091e+10  7.9578  0.004791 **
prev_year:car                        1 8.8090e+10 8.8090e+10 63.2035 1.929e-15 ***
prev_year:bath                       1 9.7073e+09 9.7073e+09  6.9648  0.008317 **
prev_year:bed                        1 1.8743e+09 1.8743e+09  1.3448  0.246202
prev_year:train_duration_min         1 9.4187e+10 9.4187e+10 67.5779 < 2.2e-16 ***
prev_year:train_distance_m           1 2.4222e+10 2.4222e+10 17.3792 3.070e-05 ***
car:bath                             1 5.1182e+08 5.1182e+08  0.3672  0.544527
car:bed                              1 1.6718e+09 1.6718e+09  1.1995  0.273424
car:train_duration_min               1 1.1888e+09 1.1888e+09  0.8530  0.355723
car:train_distance_m                 1 1.1688e+08 1.1688e+08  0.0839  0.772135
bath:bed                             1 5.0101e+07 5.0101e+07  0.0359  0.849626
bath:train_duration_min              1 6.7837e+10 6.7837e+10 48.6725 3.087e-12 ***
bath:train_distance_m                1 1.4607e+10 1.4607e+10 10.4806  0.001208 **
bed:train_duration_min               1 3.8397e+10 3.8397e+10 27.5496 1.542e-07 ***
bed:train_distance_m                 1 3.1846e+10 3.1846e+10 22.8490 1.761e-06 ***
train_duration_min:train_distance_m  1 4.8366e+10 4.8366e+10 34.7018 3.883e-09 ***
Residuals                        30261 4.2176e+13 1.3938e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7: ANOVA of the original model

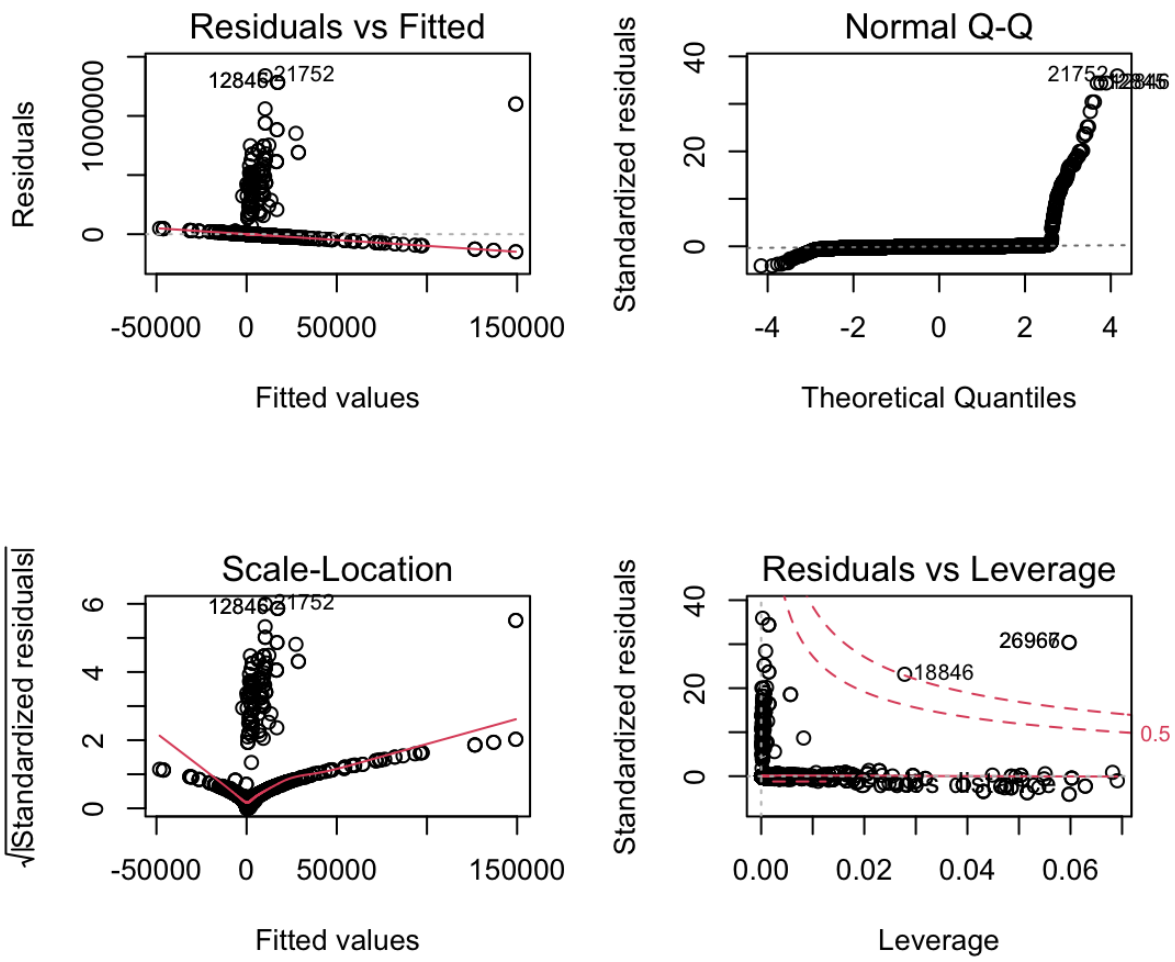



Figure 8: Diagnostic Plots

Now, since our model is a linear model, to see its performance it makes a lot of sense for us to see whether this dataset fits the linear model assumption. From the diagnostic plots on figure 8, the Residual vs Fitted shows that the red line approximately follow the zero line, this shows linearity in the data. The QQ plot shows that the data mostly follow a straight line that shows normality of residuals. However, it is still heavy on the right tail. Now the Scale-Location shows that there are poor Homogeneity of variance. It seems like homogeneity of variance heavily centred at the zero fitted values. Lastly the Residuals vs Leverage shows that there are only two outliers on the data out of thousands.

Now it seems like there are a handful of properties that should not have any change in price after three years according to our model but received an increase in price in reality. This might be because there are features that are still unknown that can explain the increase in price, like inflation or distance from the property to nearest shopping mall. Or it might because we assume that when the apartment that is unlisted, it was actually rented. When in reality, the asked-rent was too high for too long that these apartment are perceived to be unrentable, therefore are taken out from the market. The contrast does not have the same

effect because if the rent was too low, it is most likely to be the first ones that are leased. Hence, it is now clear that by definition, those points should not be included in the data set in the first place and thus our model should work seamlessly and following most of the assumption of the linear model.

R Code :

```r
#written by : Stefan Solagratio Simanjuntak/1039092

library(faraway)

data <- read.csv(
  file = 'generic-real-estate-consulting-project-group-0/notebooks/final.csv')


par(mfrow=c(2,2))

with(data, interaction.plot(prev_year, type_1, log(rent_pw/rent_yminthree)));
with(data, interaction.plot(prev_year, car, log(rent_pw/rent_yminthree)));
with(data, interaction.plot(prev_year, bath, log(rent_pw/rent_yminthree)));
with(data, interaction.plot(prev_year, bed, log(rent_pw/rent_yminthree)));

par(mfrow=c(1,2))
with(data, interaction.plot(train_duration_min, prev_year, log(rent_pw/rent_yminthree)));
with(data, interaction.plot(train_distance_m, prev_year, log(rent_pw/rent_yminthree)));

par(mfrow=c(2,2))
with(data, interaction.plot(type_1, car, log(rent_pw/rent_yminthree)));
with(data, interaction.plot(type_1, bath, log(rent_pw/rent_yminthree)));
with(data, interaction.plot(type_1, bed, log(rent_pw/rent_yminthree)));
with(data, interaction.plot(train_duration_min,type_1, log(rent_pw/rent_yminthree)));

par(mfrow=c(1,1))
with(data, interaction.plot(train_distance_m,type_1, log(rent_pw/rent_yminthree)));

par(mfrow=c(2,2))
with(data, interaction.plot(car, bath, log(rent_pw/rent_yminthree)));
with(data, interaction.plot(car, bed, log(rent_pw/rent_yminthree)));
with(data, interaction.plot(train_duration_min,car, log(rent_pw/rent_yminthree)));
with(data, interaction.plot(train_distance_m,car, log(rent_pw/rent_yminthree)));

par(mfrow=c(1,3))
with(data, interaction.plot(bath, bed, log(rent_pw/rent_yminthree)));
with(data, interaction.plot(train_duration_min,bath, log(rent_pw/rent_yminthree)));
with(data, interaction.plot(train_distance_m,bath, log(rent_pw/rent_yminthree)));

par(mfrow=c(1,2))
with(data, interaction.plot(train_duration_min,bed, log(rent_pw/rent_yminthree)));
with(data, interaction.plot(train_distance_m,bed, log(rent_pw/rent_yminthree)));

```

```r
model <- lm(rent_pw ~ offset(log(rent_yminthree)) +
              ( prev_year+car + bath + bed + train_duration_min + train_distance_m)^2
                 ,data=data);
mod2 <- step(model, scope=~.,trace = FALSE)
summary(mod2)

anova(model, test="Chi")

par(mfrow = c(2, 2))
plot(mod2)
```

6:66 (Top Level)