Labbrapport i Statistik

Laboration 1

732G53

Författare 1 Anton Roshamn
Författare 2 Nils af Petersens

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2024-09-10

Innehåll

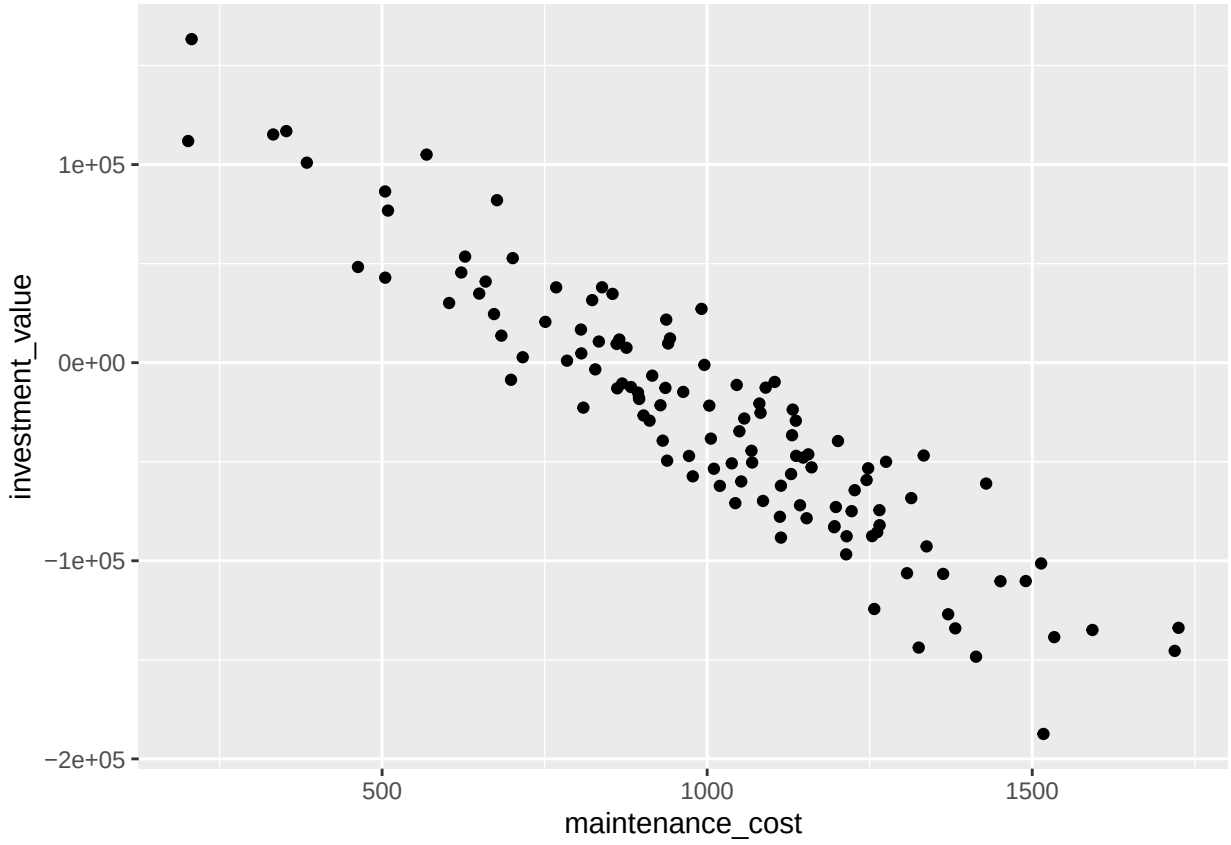
1 Uppgifter

1.1 Uppgift a

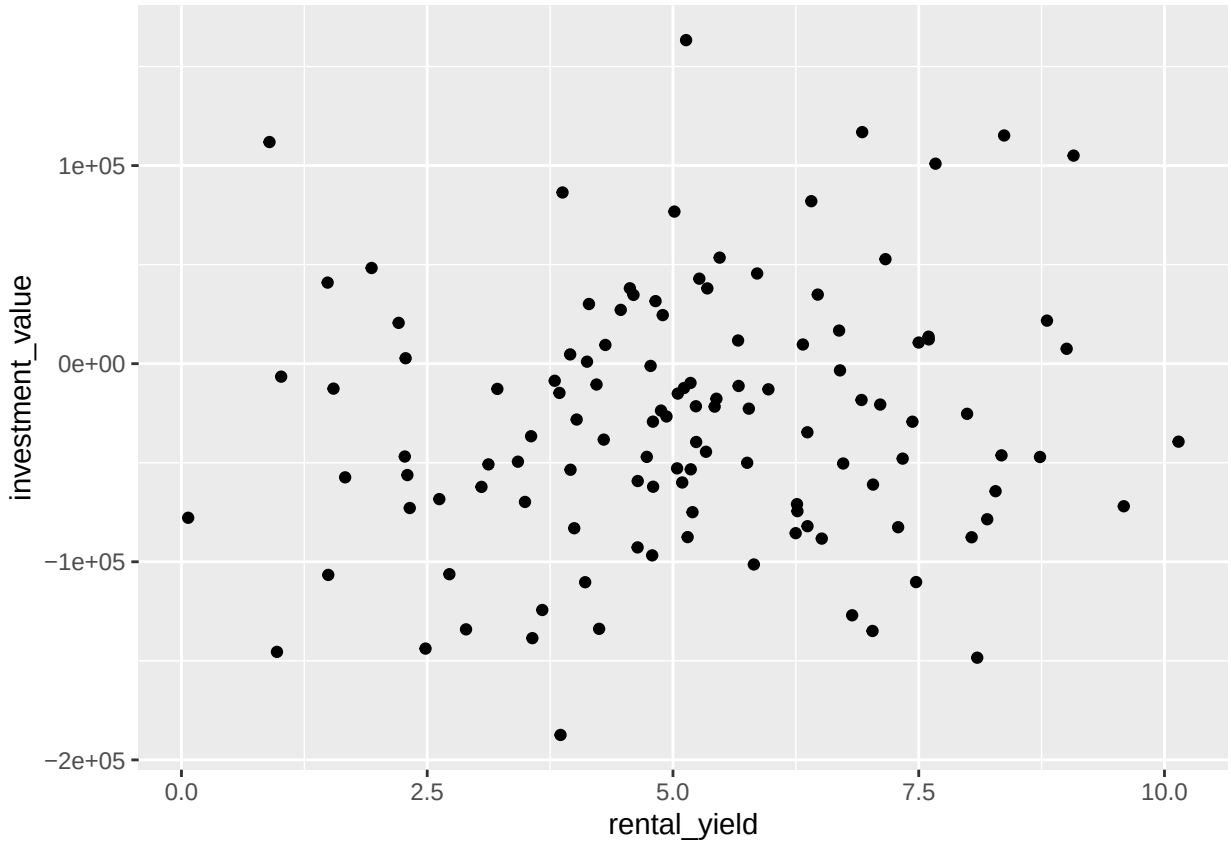
1.1.1 a.i)

```
##           Variabelnamn Variabeltyp        Skala
## 1     investment_value    Numerisk    Kvotskala
## 2      location_rating    Numerisk      Ordinär
## 3 construction_quality    Numerisk      Ordinär
## 4         rental_yield    Numerisk    Kvotskala
## 5     maintenance_cost    Numerisk    Kvotskala
## 6        property_type  Kategorisk Nominalskala
## 7  economic_conditions  Kategorisk Nominalskala
```

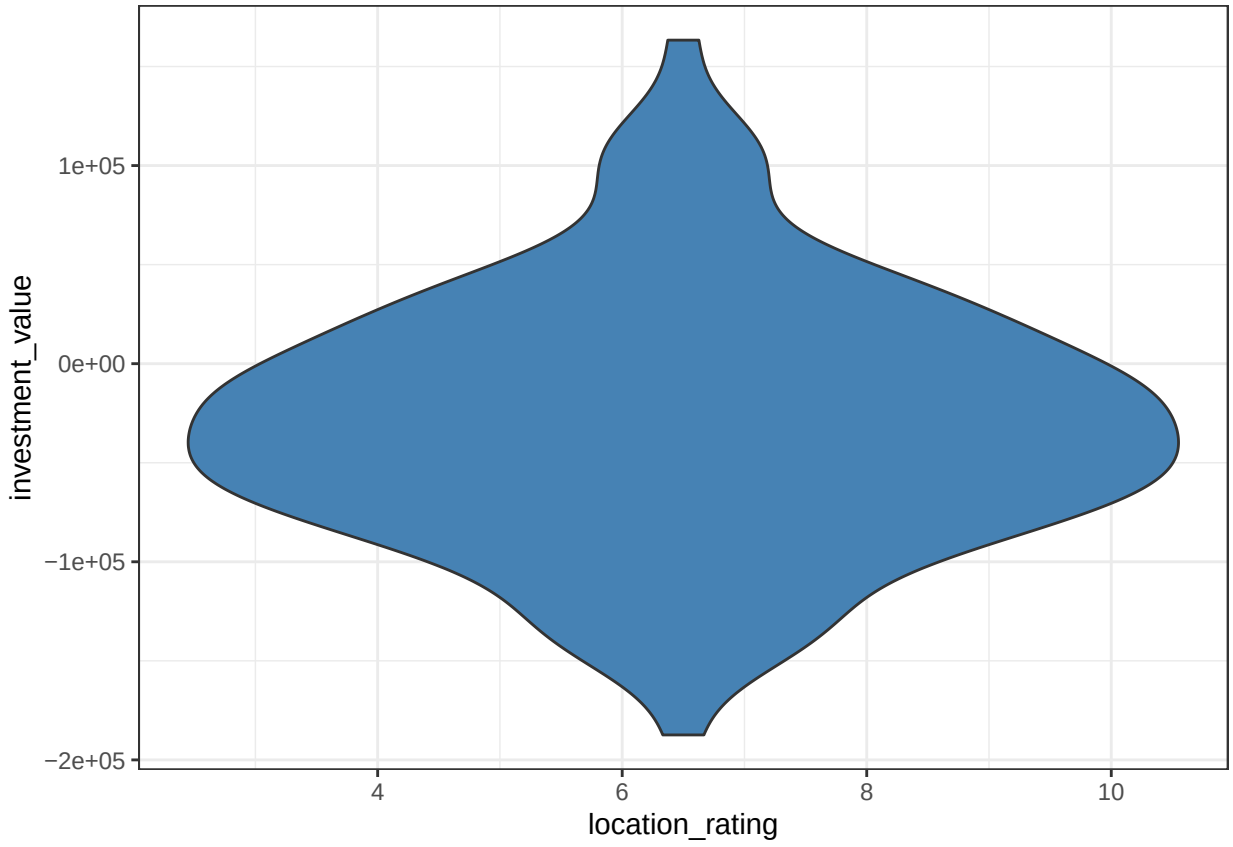
1.1.2 a.ii)



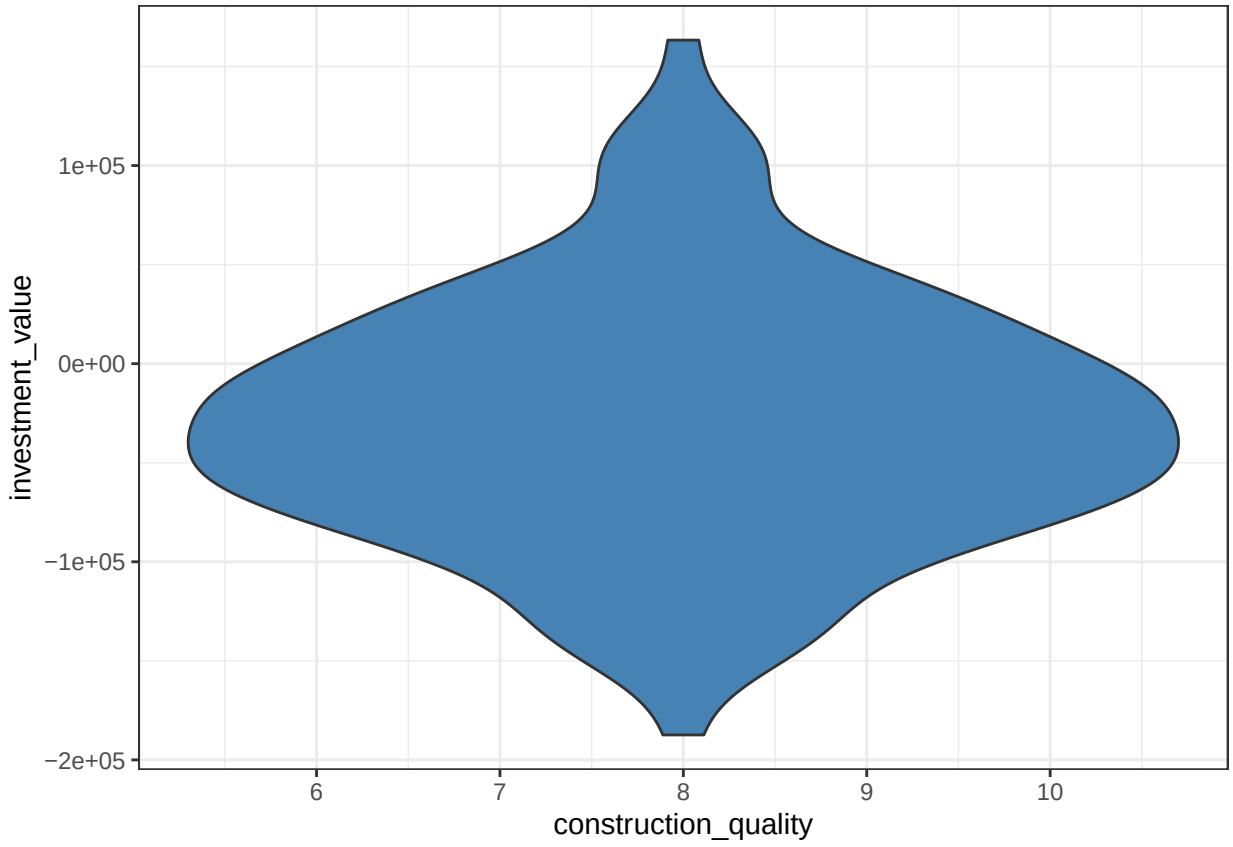
Figur 1: Samband mellan investment value och maintenance cost



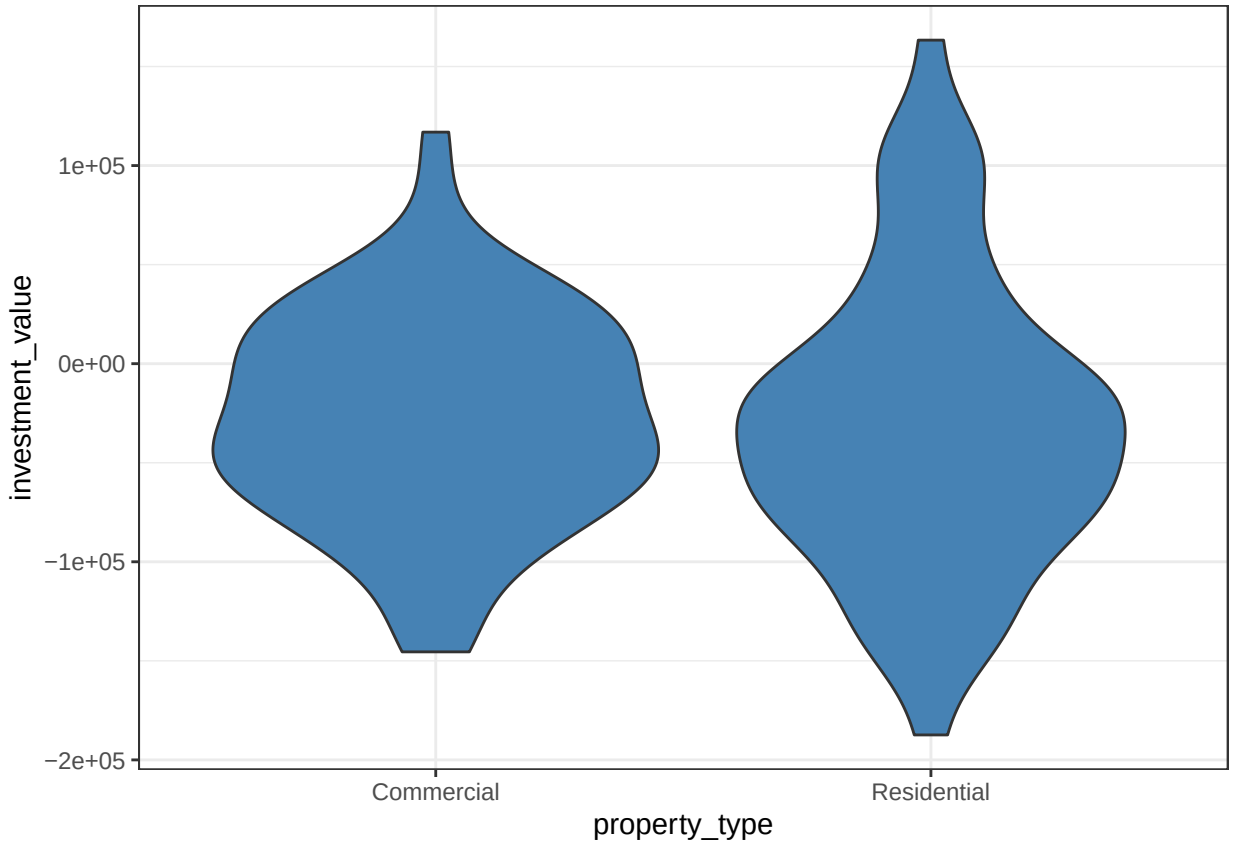
Figur 2: Samband mellan investment value och rental yield



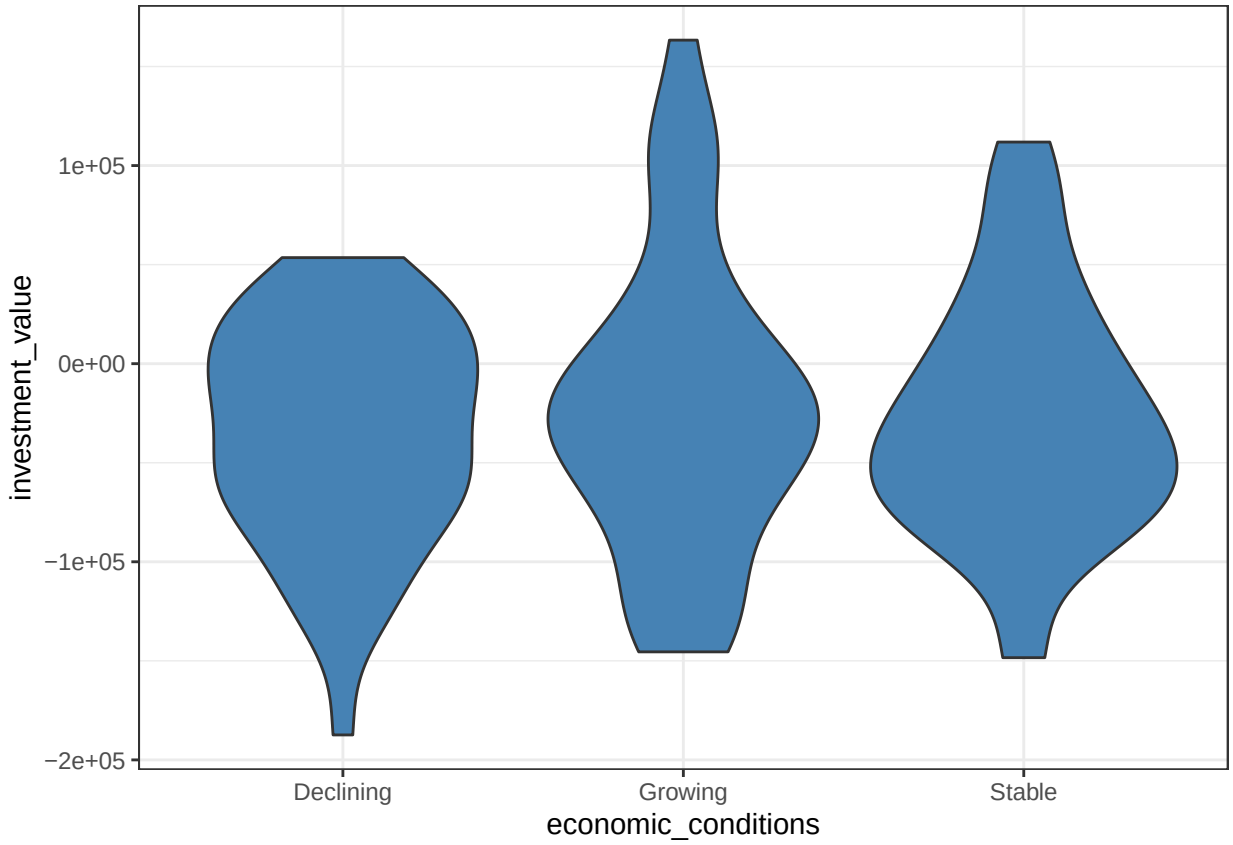
Figur 3: Samband mellan investment value och location rating



Figur 4: Samband mellan investment value och construction quality



Figur 5: Samband mellan investment value och property type



Figur 6: Samband mellan investment value och economic conditions

1.1.3 a.iii)

Vi inkluderar samtliga variabler i regressionsmodellen då vi från figurerna kan utläsa samband mellan investment value och samtliga övriga variabler. Ett statistiskt test kommer utföras senare i rapporten för att testa signifikansen av dessa samband och avgöra ifall vissa variabler ska exkluderas från den slutgiltiga modellen.

1.2 b

1.2.1 b.i)

```
linreg <- lm(investment_value ~ ., data = data_clean)
linreg
```

```
##
## Call:
## lm(formula = investment_value ~ ., data = data_clean)
##
## Coefficients:
##                 (Intercept)             location_rating
##                     58231.8                      9914.2
##       construction_quality               rental_yield
##                      5190.5                      1950.6
##           maintenance_cost    property_typeResidential
##                      -199.7                    -19738.2
## economic_conditionsGrowing   economic_conditionsStable
##                     20522.0                     11475.9
```

1.2.2 b.ii)

```
conf_intervall <- confint(linreg, level = 0.95)
conf_intervall["location_rating", ]
```

```
##     2.5 %    97.5 %
##  9417.112 10411.331
```

```
conf_intervall["economic_conditionsStable", ]
```

```
##     2.5 %    97.5 %
##  9232.694 13719.089
```

1.2.3 b.iii)

```
anovi <- anova(linreg)
summary(anovi)
```

```
##        Df            Sum Sq             Mean Sq            F value
##  Min.   :  1.0   Min.   :2.875e+09   Min.   :2.567e+07   Min.   :  137.4
##  1st Qu.:  1.0   1st Qu.:5.263e+09   1st Qu.:3.763e+09   1st Qu.:  185.1
##  Median :  1.0   Median :8.001e+09   Median :7.001e+09   Median :  323.5
##  Mean   : 17.0   Mean   :6.934e+10   Mean   :6.836e+10   Mean   : 3107.1
##  3rd Qu.:  1.5   3rd Qu.:2.798e+10   3rd Qu.:2.798e+10   3rd Qu.: 1448.1
##  Max.   :112.0   Max.   :4.080e+11   Max.   :4.080e+11   Max.   :15896.5
##                                                          NA's   :1
##      Pr(>F)
##  Min.   :0
##  1st Qu.:0
##  Median :0
```

```
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
##  NA's   :1
```

```r
krit <- qf(0.95, df1 = anovi[1,"Df"], df2 = anovi[2, "Df"])
cat("kritiskt varde for f-testet", krit, "\n")
```

```
## kritiskt varde for f-testet 161.4476
```

```r
# Räkna fram det kritiska F-värdet (med 95% konfidensnivå)
krit_fvarde <- qf(0.95, df1 = anovi[1, "Df"], df2 = anovi[2, "Df"])
cat("Det kritiska F-värdet är:", krit_fvarde, "\n")
```

```
## Det kritiska F-värdet är: 161.4476
```

```r
# Kontrollera om p-värdet indikerar att vi bör förkasta nollhypotesen
if (anovi$`Pr(>F)`[1] < 0.05) {
  cat("Nollhypotesen förkastas: Någon av de oberoende variablerna påverkar den beroende variabeln.\n")
} else {
  cat("Nollhypotesen kan inte förkastas: Inga signifikanta effekter upptäcktes.\n")
}
```

```
## Nollhypotesen förkastas: Någon av de oberoende variablerna påverkar den beroende variabeln.
```

```r
drop1(linreg, test = "F", direction = "backward")
```

```
## Single term deletions
##
## Model:
## investment_value ~ location_rating + construction_quality + rental_yield +
##     maintenance_cost + property_type + economic_conditions
##                       Df  Sum of Sq        RSS    AIC   F value    Pr(>F)
## <none>                               2.8748e+09 2055.0
## location_rating        1 4.0081e+10 4.2956e+10 2377.5  1561.512 < 2.2e-16 ***
## construction_quality   1 3.3500e+09 6.2248e+09 2145.7   130.511 < 2.2e-16 ***
## rental_yield           1 1.8317e+09 4.7066e+09 2112.2    71.362  1.22e-13 ***
## maintenance_cost       1 4.1749e+11 4.2036e+11 2651.2 16264.832 < 2.2e-16 ***
## property_type          1 1.0649e+10 1.3523e+10 2238.8   414.857 < 2.2e-16 ***
## economic_conditions    2 8.0007e+09 1.0876e+10 2210.7   155.850 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.2.4 b.iv)

```r
linreg_mini <- lm(investment_value ~ location_rating + construction_quality + rental_yield + maintenance
anova_test <- anova(linreg_mini, linreg)
print(anova_test)
```

```
## Analysis of Variance Table
##
## Model 1: investment_value ~ location_rating + construction_quality + rental_yield +
##     maintenance_cost
```

```
## Model 2: investment_value ~ location_rating + construction_quality + rental_yield +
##     maintenance_cost + property_type + economic_conditions
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    115 1.7877e+10
## 2    112 2.8748e+09  3 1.5002e+10 194.82 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Resultatet visar ett väldigt lågt p-värde och ett högt F-värde vilket innebär att vi förkastar H0, de k

# Utför bakåtselektion med F-test
reviderad_modell <- step(linreg, direction = "backward")
```

```
## Start:  AIC=2055.01
## investment_value ~ location_rating + construction_quality + rental_yield +
##     maintenance_cost + property_type + economic_conditions
##
##                         Df  Sum of Sq        RSS    AIC
## <none>                                2.8748e+09 2055.0
## - rental_yield           1 1.8317e+09 4.7066e+09 2112.2
## - construction_quality   1 3.3500e+09 6.2248e+09 2145.7
## - economic_conditions    2 8.0007e+09 1.0876e+10 2210.7
## - property_type          1 1.0649e+10 1.3523e+10 2238.8
## - location_rating        1 4.0081e+10 4.2956e+10 2377.5
## - maintenance_cost       1 4.1749e+11 4.2036e+11 2651.2
```

```
summary(reviderad_modell)
```

```
##
## Call:
## lm(formula = investment_value ~ location_rating + construction_quality +
##     rental_yield + maintenance_cost + property_type + economic_conditions,
##     data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14427.1  -3160.5    -54.5   3091.7  16261.9
##
## Coefficients:
##                               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                  58231.843   4446.661   13.096  < 2e-16 ***
## location_rating               9914.222    250.892   39.516  < 2e-16 ***
## construction_quality          5190.500    454.344   11.424  < 2e-16 ***
## rental_yield                  1950.579    230.903    8.448 1.22e-13 ***
## maintenance_cost              -199.723      1.566 -127.534  < 2e-16 ***
## property_typeResidential   -19738.228    969.078  -20.368  < 2e-16 ***
## economic_conditionsGrowing  20521.962   1171.036   17.525  < 2e-16 ***
## economic_conditionsStable   11475.892   1132.144   10.136  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 5066 on 112 degrees of freedom
## Multiple R-squared:  0.9941, Adjusted R-squared:  0.9937
## F-statistic:  2686 on 7 and 112 DF,  p-value: < 2.2e-16
```

```
# Resultatet visar att alla variabler har en signifikant påverkan på investment value, därmed plockar vi
```

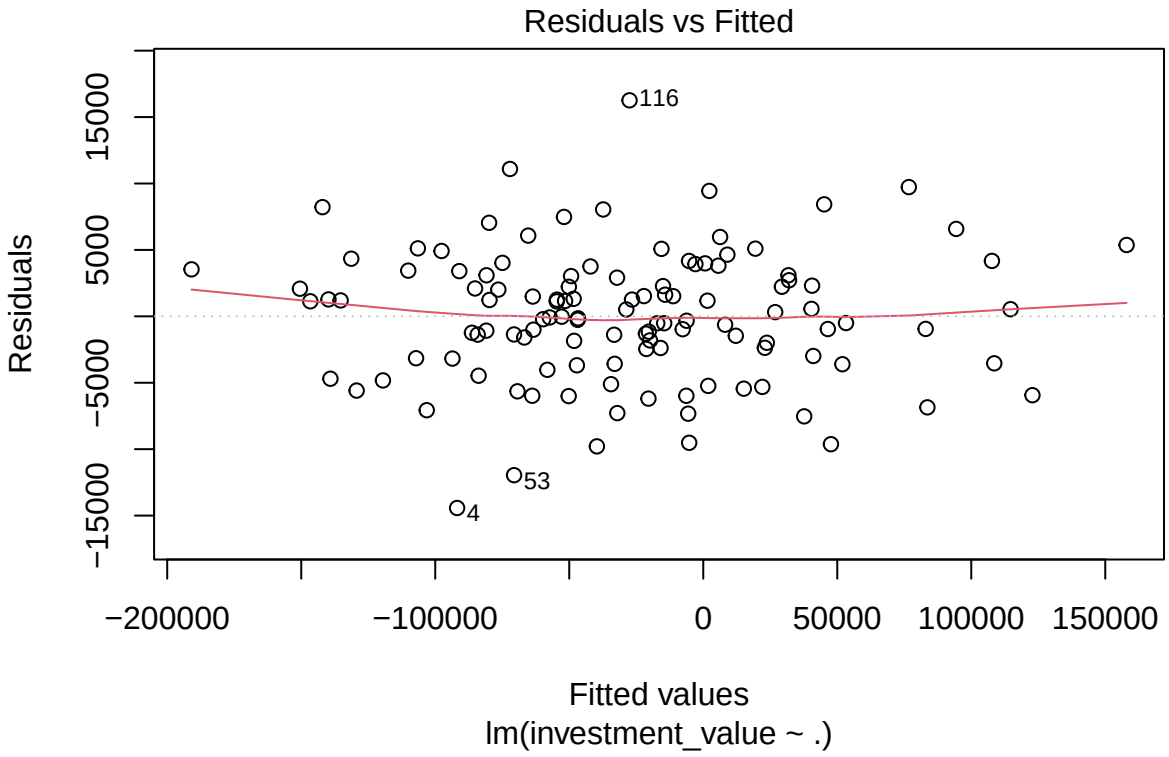
Och en ANOVA-analys.

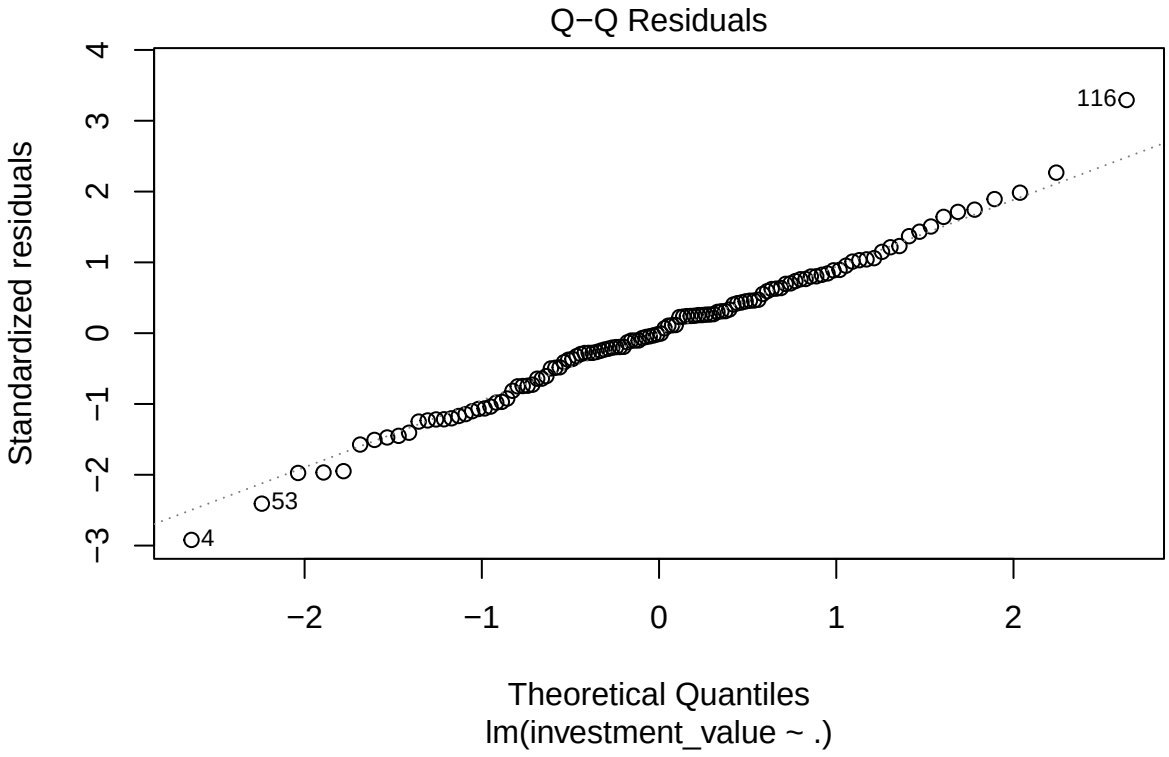
Signifikanta samband mellan investment value och samtliga övriga variabler har visats, därmed inkluderar vi variablerna i den slutgiltiga modellen inför residualanalys.

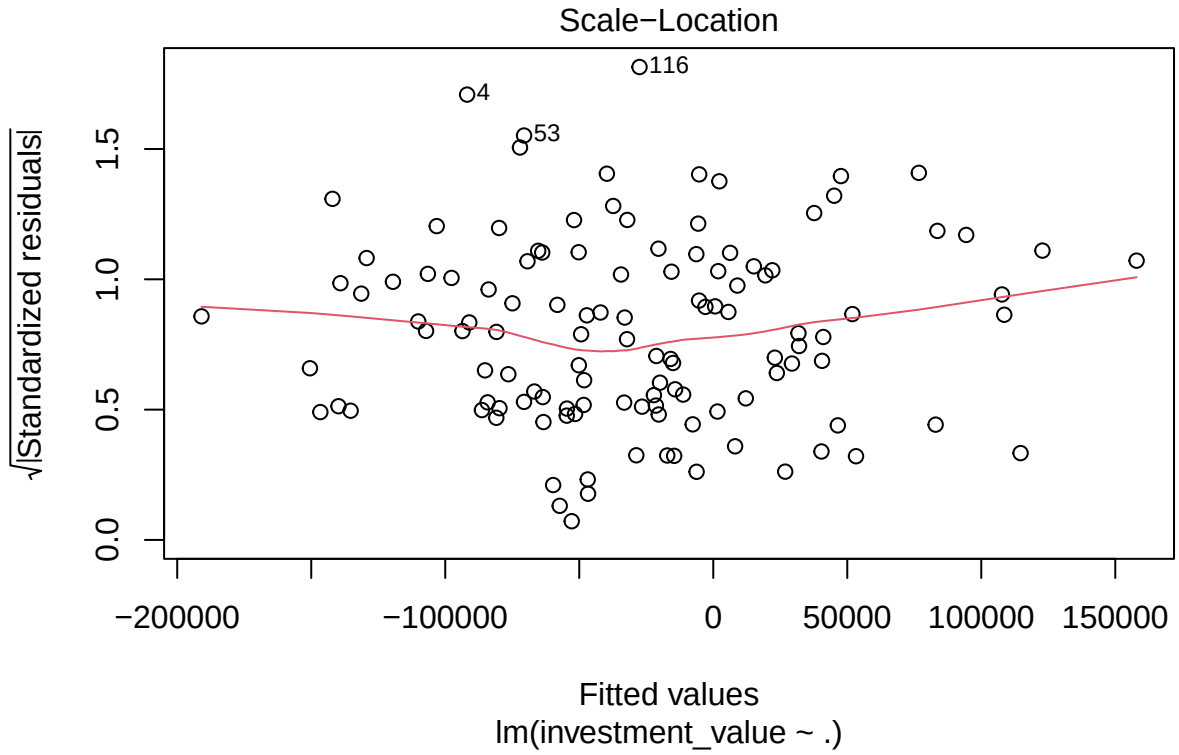
1.3 c

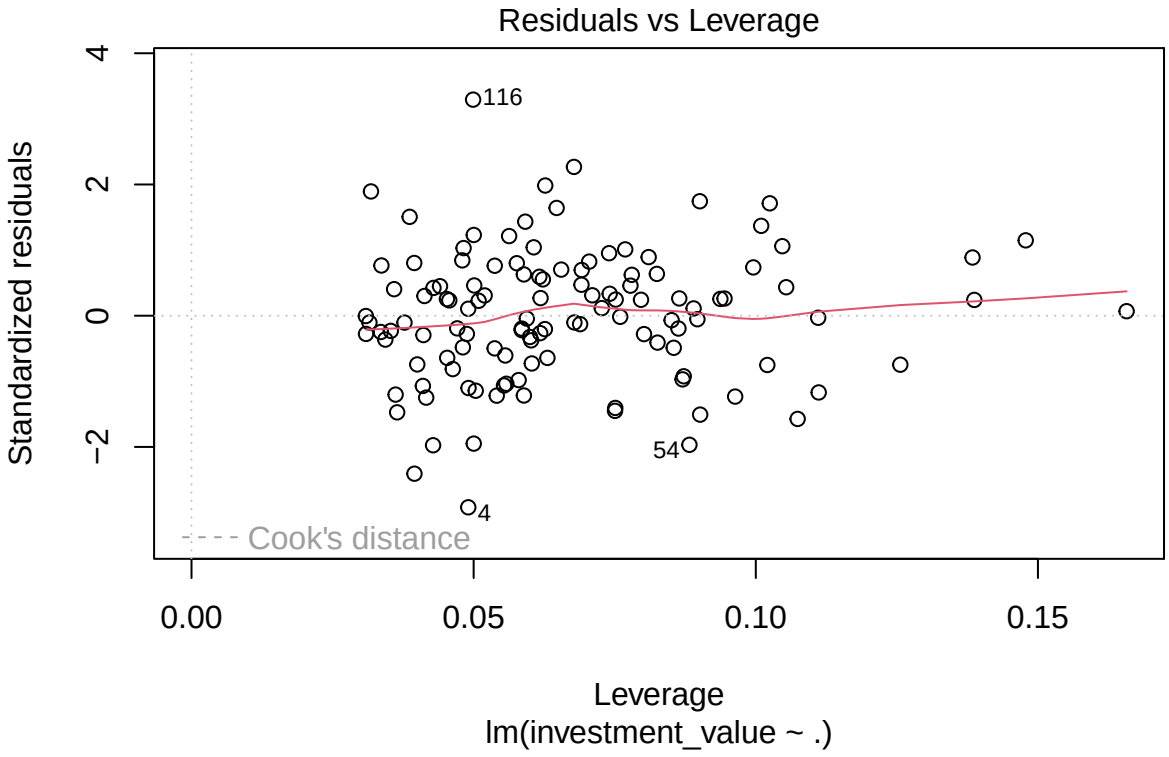
##c.i)

```
plot(linreg)
```

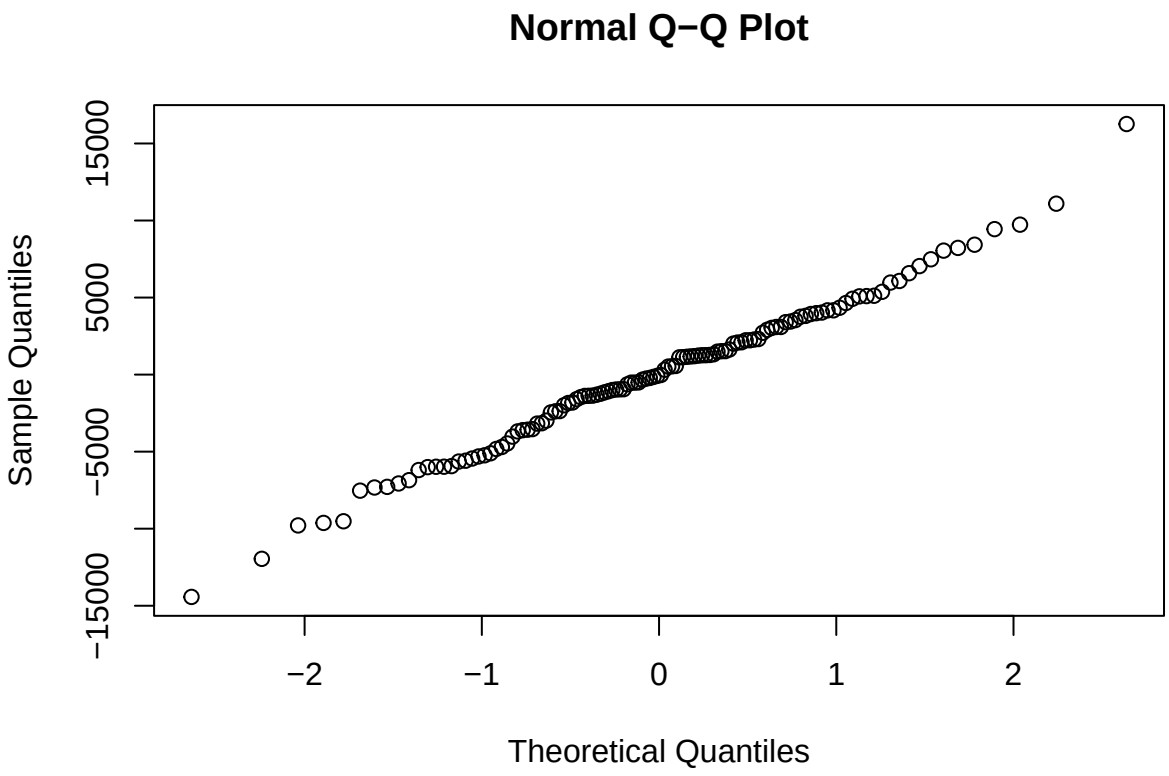



Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(investment_value ~ .)
116
53
4
−3
−2
−1
0
1
2
3
4


Scale−Location
116
4
53
√|Standardized residuals|
0.0
0.5
1.0
1.5
−200000
−100000
0
50000
100000
150000
Fitted values
lm(investment_value ~ .)




```
qq <- qqnorm(linreg$residuals)
```



##c.ii)

Vi identifierade 3 extremvärden.

##d) Vi anser att modellen är bra, samtliga variabler som inkluderades i den slutgiltiga modellen visade ha ett signifikant samband med investment value, alltså är det giltigt att använda dessa som prediktorer för utfallet i variabeln investment value. Detta innebär att samtliga variabler förklarar någon mån av signifikant varians i investment value.