

Multiple Regression

Reading Data

```
data=(read.csv("rent99.raw", sep=" "))
attach(data) # able to use location instead of data$location
head(data) # prints data header
```

```
##      rent   rentsqm area yearc location bath kitchen cheating district
## 1 120.9744  3.456410   35  1939         1    0         0         0      1112
## 2 436.9743  4.201676  104  1939         1    1         0         1      1112
## 3 355.7436 12.267021   29  1971         2    0         0         1      2114
## 4 282.9231  7.254436   39  1972         2    0         0         1      2148
## 5 807.2308  8.321964   97  1985         1    0         0         1      2222
## 6 482.8205  7.787426   62  1962         1    0         0         1      2222
```

```
data$location=as.factor(data$location) # treats location as a categorical variable (no linear correlation)
levels(data$location)=c("avg","good","top") # names the categorical variable's values
```

basic description and scatter plot

```
library(dplyr)
```

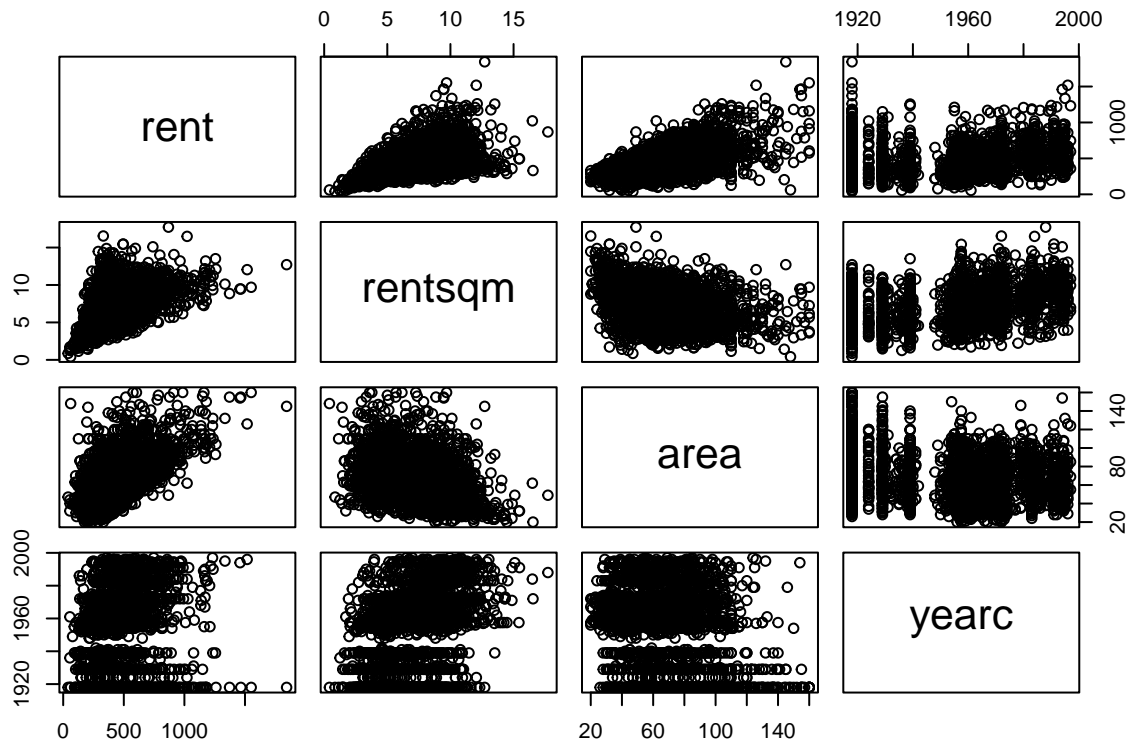
```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##   filter, lag
##
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
library(tableone)
CreateTableOne(factorVars=c("location", "bath", "kitchen", "cheating"), # makes a description table, with
                    data=data%>%select(-district)) # syntactic "sugar" for select(data, -district) - we're t
```

```
##
##              Overall
##  n              3082
##  rent (mean (SD))  459.44 (195.66)
##  rentsqm (mean (SD))  7.11 (2.44)
##  area (mean (SD))    67.37 (23.72)
##  yearc (mean (SD))   1956.31 (22.31)
##  location (%)
##    avg              1794 (58.2)
##    good             1210 (39.3)
##    top               78 ( 2.5)
##  bath = 1 (%)       191 ( 6.2)
```

```
## kitchen = 1 (%)          131 ( 4.3)
## cheating = 1 (%)        2761 (89.6)
```

```
plot(data[,1:4]) # plots all rows and first 4 columns of data (the 12 scatterplots)
```



Linear regression models

multiple regression of rent onto area and yearc

```
fit=lm(rent~area+I(yearc-1956),data=data) # I() - idiot function - treat arithmetics literally; fits a
summary(fit)
```

```
##
## Call:
## lm(formula = rent ~ area + I(yearc - 1956), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -734.76  -94.75  -10.87   82.55 1063.17
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    97.4225     8.3079   11.73  <2e-16 ***
## area           5.3618     0.1165   46.01  <2e-16 ***
## I(yearc - 1956)  2.4913     0.1239   20.11  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 149.3 on 3079 degrees of freedom
## Multiple R-squared:  0.4181, Adjusted R-squared:  0.4177
```

```
## F-statistic: 1106 on 2 and 3079 DF, p-value: < 2.2e-16
```

polynomial regression / quadratic effects

```
fit.2=lm(rent~area+I(yearc-1956)+I((yearc-1956)^2),data=data)
summary(fit.2)
```

```
##
## Call:
## lm(formula = rent ~ area + I(yearc - 1956) + I((yearc - 1956)^2),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -757.99  -88.89   -8.39   83.52 1039.27
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    81.607327   8.237114   9.907  <2e-16 ***
## area           5.136457   0.115594  44.435  <2e-16 ***
## I(yearc - 1956)  2.942822   0.127113  23.151  <2e-16 ***
## I((yearc - 1956)^2) 0.062017   0.005255  11.802  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 146.1 on 3078 degrees of freedom
## Multiple R-squared:  0.4433, Adjusted R-squared:  0.4427
## F-statistic: 816.9 on 3 and 3078 DF, p-value: < 2.2e-16
```

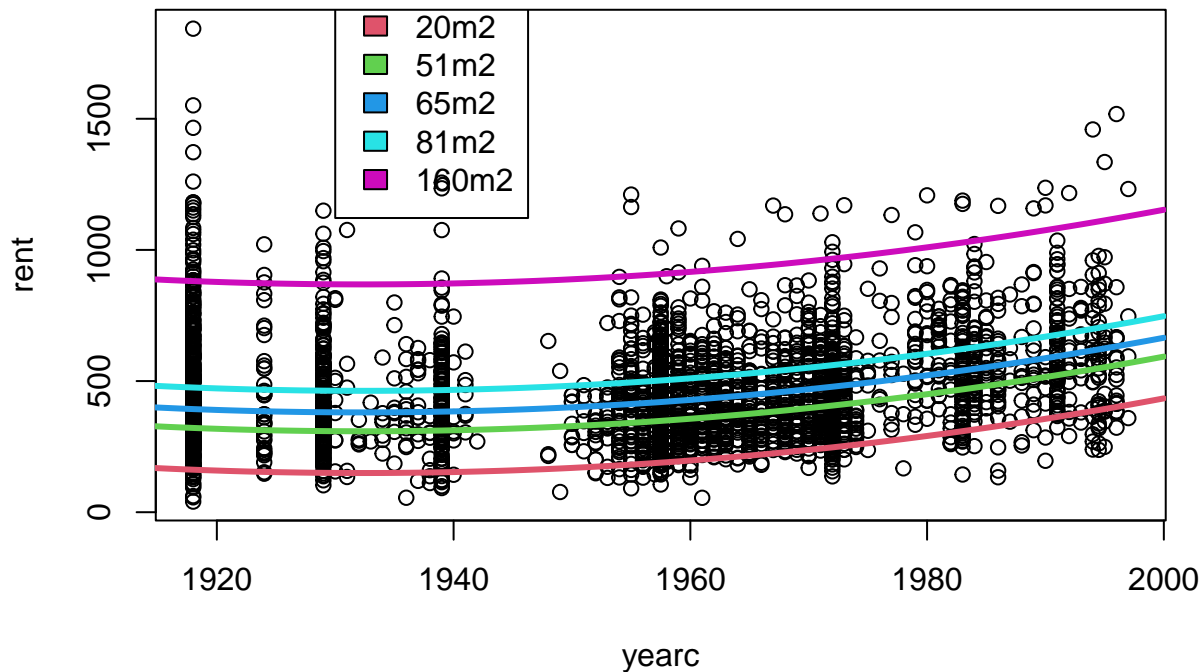
plotting model estimates with basic plotting tools

```
summary(data$area)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    20.00   51.00   65.00   67.37   81.00  160.00
```

```
with(data,plot(rent~yearc)) # with() - syntactic sugar for plot(data$rent~data$yearc)
```

```
H=predict(fit.2, newdata= # uses fit.2 lm to predict rent for new data frame (90 x 5 new data rows)
             data.frame(expand.grid(yearc=1911:2000, area=c(20,51,65,81,160))))
for (i in 0:4){ # plots a line for each area (5 of them)
  lines(1911:2000,H[i*90+1:90], lwd=3, col=i+2) # predicted rent vs yearc
} # lwd - line thickness, col - color
legend(1930,2000, legend=c("20m2","51m2","65m2","81m2","160m2"), fill=c(2:6))
```



R^2 depends on range of X

previous model

```
fit=lm(rent~area+I(yearc-1956),data=data)
S=summary(fit)
S
```

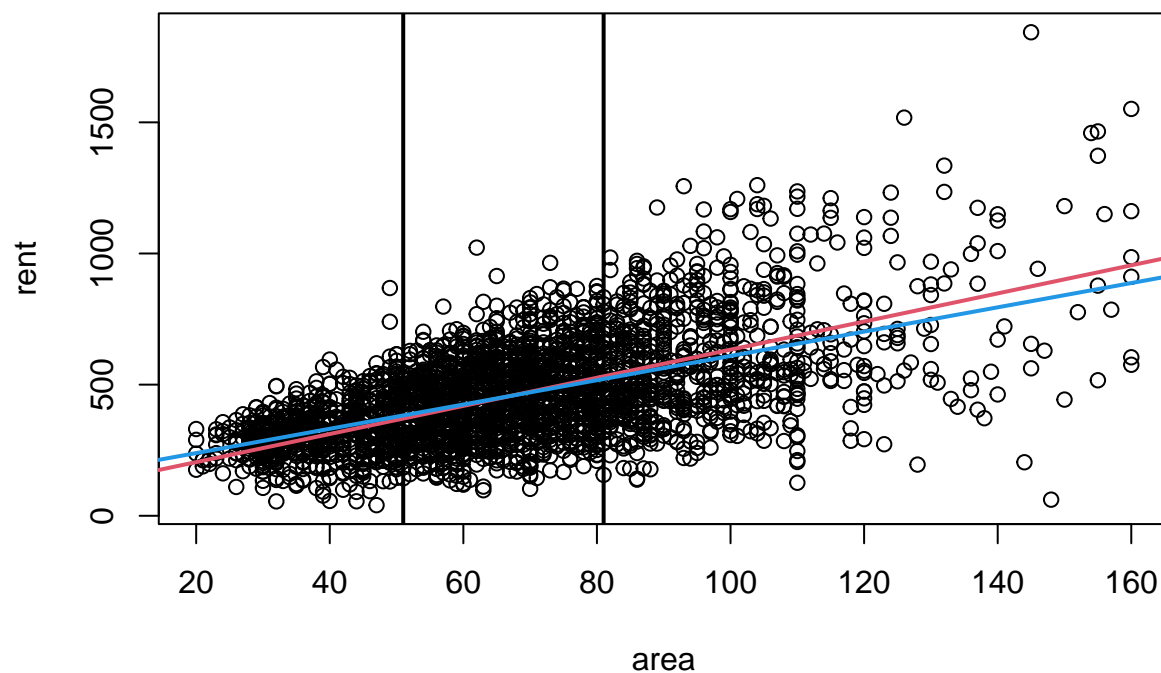
```
##
## Call:
## lm(formula = rent ~ area + I(yearc - 1956), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -734.76  -94.75  -10.87   82.55 1063.17
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    97.4225     8.3079   11.73  <2e-16 ***
## area             5.3618     0.1165   46.01  <2e-16 ***
## I(yearc - 1956)  2.4913     0.1239   20.11  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 149.3 on 3079 degrees of freedom
## Multiple R-squared:  0.4181, Adjusted R-squared:  0.4177
## F-statistic: 1106 on 2 and 3079 DF, p-value: < 2.2e-16
```

regression on subset with medium areas only

```
fit.3=lm(rent~area+I(yearc-1956) ,data=data%% filter(area>51, area<81)) # only picking data with medium area
S.3=summary(fit.3)
S.3
```

```
##
## Call:
## lm(formula = rent ~ area + I(yearc - 1956), data = data %% filter(area >
##      51, area < 81))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -373.56  -96.17   -8.46   90.30  528.87
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    146.5575     27.6437   5.302 1.32e-07 ***
## area              4.6265      0.4173  11.086 < 2e-16 ***
## I(yearc - 1956)  2.1478      0.1602  13.411 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 135 on 1520 degrees of freedom
## Multiple R-squared:  0.1626, Adjusted R-squared:  0.1614
## F-statistic: 147.5 on 2 and 1520 DF,  p-value: < 2.2e-16
```

```
with(data,plot(area, rent))
abline(v=c(51,81), lwd=2, col=1) # two vertical lines
abline(c(97.42, 5.36), lwd=2, col=2) # linear function w/ intercept 97, slope 5
abline(c(146.56, 4.63), lwd=2, col=4)
```



We can observe that limiting data subset has a high impact on the model and how well it will fit.

residual SE = sigma

```
round(S$sigma,2)
```

```
## [1] 149.3
```

```
round(S.3$sigma,2)
```

```
## [1] 134.99
```

R^2

```
round(S$r.squared,2)
```

```
## [1] 0.42
```

```
round(S.3$r.squared,2)
```

```
## [1] 0.16
```

estimated beta's

```
round(coef(fit),2)
```

```
##      (Intercept)      area I(yearc - 1956)
##           97.42           5.36           2.49
```

```
round(coef(fit.3),2)
```

```
##      (Intercept)      area I(yearc - 1956)
##          146.56           4.63           2.15
```