Notes: MS 204 Chapter 6.2

Overview

- Statistical significance
- Model selection
- Model validation

Multiple linear regression: Boston city home prices

Ex: $X_1 = \text{crim}$, $X_2 = rm$, $\ldots$, $Y = \text{medv}$

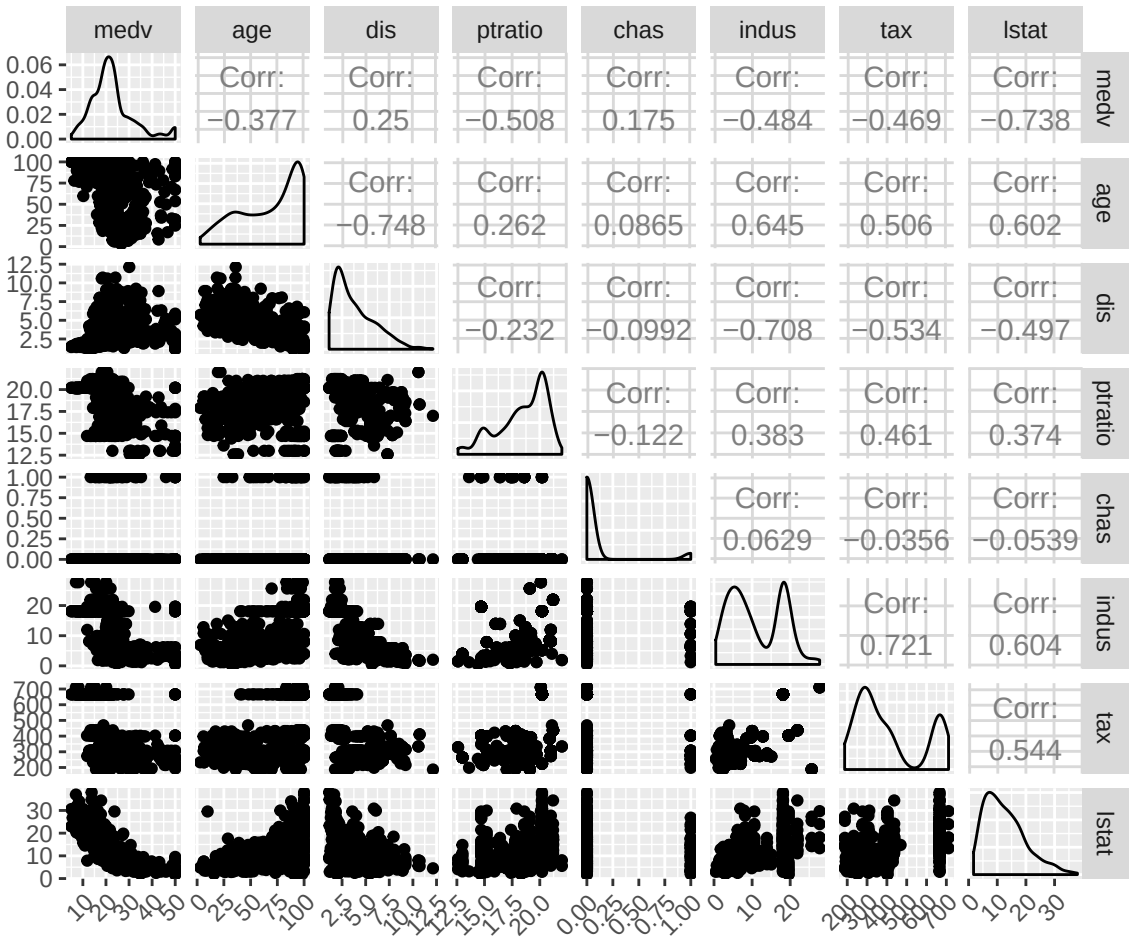
```
library(mosaic); library(tidyverse); library(MASS)
data(Boston)
dim(Boston)
```

```
## [1] 506  14
```

```
Boston.reg <- Boston %>%
  dplyr::select(medv, age, dis, ptratio, chas, indus, tax, lstat)
head(Boston.reg)
```

```
##   medv  age    dis ptratio chas indus tax lstat
## 1 24.0 65.2 4.0900    15.3    0  2.31 296  4.98
## 2 21.6 78.9 4.9671    17.8    0  7.07 242  9.14
## 3 34.7 61.1 4.9671    17.8    0  7.07 242  4.03
## 4 33.4 45.8 6.0622    18.7    0  2.18 222  2.94
## 5 36.2 54.2 6.0622    18.7    0  2.18 222  5.33
## 6 28.7 58.7 6.0622    18.7    0  2.18 222  5.21
```

```
library(GGally)
ggpairs(Boston.reg) + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Describe the overall association between variables.

Stepwise selection (approach, weaknesses, alternatives)

Full model

```
fit.full <- lm(medv ~ age + dis + ptratio + chas + indus + tax + lstat, data = Boston)
msummary(fit.full)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 58.4021721  2.6119146  22.360  < 2e-16 ***
## age          0.0020650  0.0143719   0.144 0.885807
## dis         -1.0708948  0.1966344  -5.446 8.10e-08 ***
## ptratio     -1.0142137  0.1303800  -7.779 4.25e-14 ***
## chas         3.4915589  0.9878544   3.534 0.000447 ***
## indus       -0.2320387  0.0643231  -3.607 0.000341 ***
## tax         -0.0002464  0.0022170  -0.111 0.911539
## lstat       -0.8524239  0.0475568 -17.924  < 2e-16 ***
##
## Residual standard error: 5.491 on 498 degrees of freedom
## Multiple R-squared:  0.6485, Adjusted R-squared:  0.6436
## F-statistic: 131.3 on 7 and 498 DF,  p-value: < 2.2e-16
```

```
fit.red1 <- lm(medv ~ age + dis + ptratio + chas + indus + lstat, data = Boston)
msummary(fit.red1)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 58.41054    2.60825  22.395  < 2e-16 ***
## age          0.00206    0.01436   0.143 0.885954
## dis         -1.06990    0.19624  -5.452 7.84e-08 ***
## ptratio     -1.01789    0.12599  -8.079 4.95e-15 ***
## chas         3.49901    0.98460   3.554 0.000416 ***
## indus       -0.23533    0.05706  -4.124 4.36e-05 ***
## lstat       -0.85309    0.04713 -18.102  < 2e-16 ***
##
## Residual standard error: 5.485 on 499 degrees of freedom
## Multiple R-squared:  0.6485, Adjusted R-squared:  0.6443
## F-statistic: 153.5 on 6 and 499 DF,  p-value: < 2.2e-16
```

```r
fit.red2 <- lm(medv ~  dis + ptratio + chas + indus + lstat, data = Boston)
msummary(fit.red2)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 58.56949    2.35903  24.828  < 2e-16 ***
## dis         -1.08478    0.16642  -6.518 1.74e-10 ***
## ptratio     -1.01785    0.12587  -8.087 4.66e-15 ***
## chas         3.50877    0.98129   3.576 0.000383 ***
## indus       -0.23458    0.05677  -4.132 4.21e-05 ***
## lstat       -0.85080    0.04430 -19.205  < 2e-16 ***
##
## Residual standard error: 5.48 on 500 degrees of freedom
## Multiple R-squared:  0.6485, Adjusted R-squared:  0.645
## F-statistic: 184.5 on 5 and 500 DF,  p-value: < 2.2e-16
```

Final model:

Coefficient for `chas`

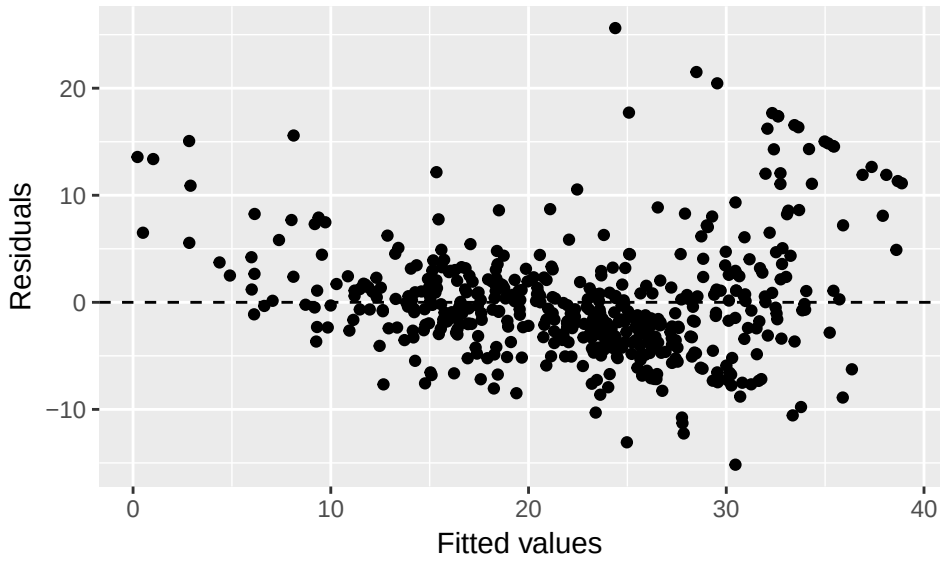
Coefficient for `dis`

Model validation

1. Linearity
2. Nearly normal residuals
3. Constant variability

Linearity

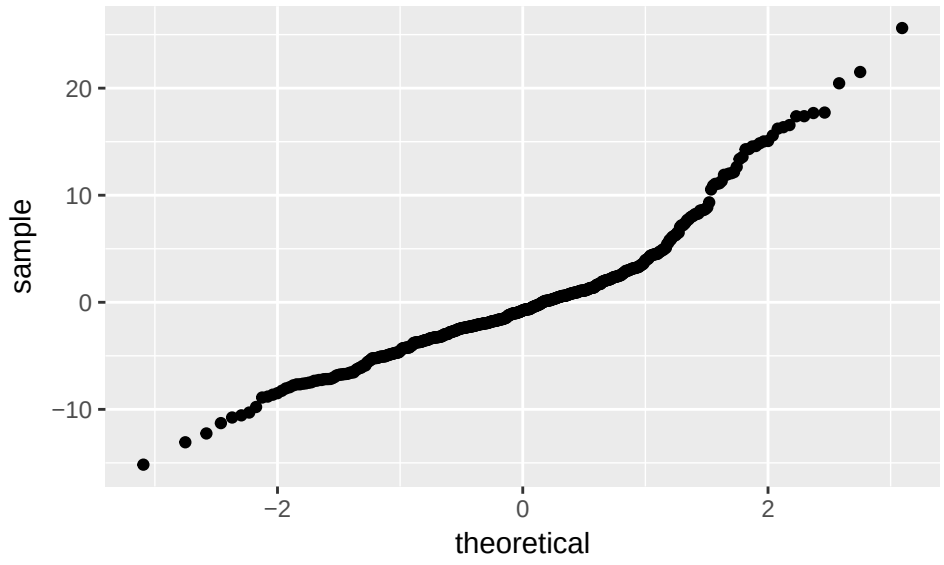
```r
qplot(x = .fitted, y = .resid, data = fit.red2) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



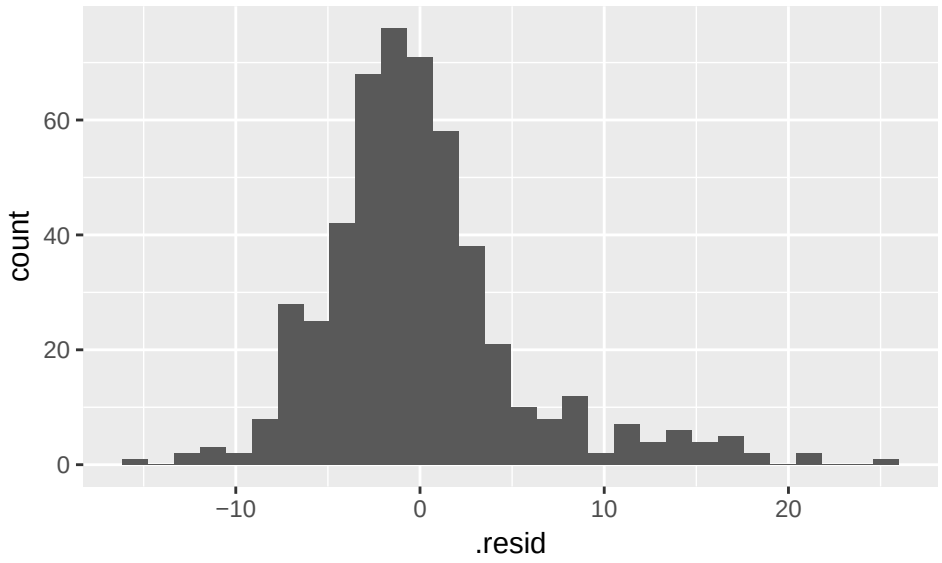
How else to assess linearity?

Nearly normal residuals

```
qplot(sample = .resid, data = fit.red2, geom = "qq")
```




```
qplot(x = .resid, data = fit.red2, geom = "histogram")
```



Constant variability

```
qplot(x = .fitted, y = .resid, data = fit.red2) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```