11) Other Considerations in the Regression Model

Vitor Kamada

January 2018

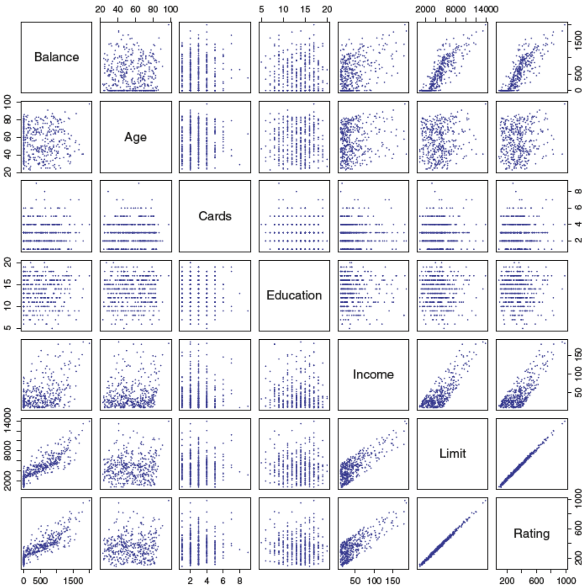
Reference

Tables, Graphics, and Figures from

An Introduction to Statistical Learning

James et al. (2017): Chapters: 3.3, 3.6.4, 3.6.5, 3.6.6

Credit Data Set



Predictors with Only Two Levels

$$Balance_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\beta_0 + \beta_1 + \epsilon_i \text{ if Female}$$

$$\beta_0 + \epsilon_i \text{ if Male}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

Predictors with More than Two Levels

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

$$\beta_0 + \beta_1 + \epsilon_i \text{ if Asian}$$

$$\beta_0 + \beta_2 + \epsilon_i \text{ if Caucasian}$$

$$\beta_0 + \epsilon_i \text{ if African American}$$

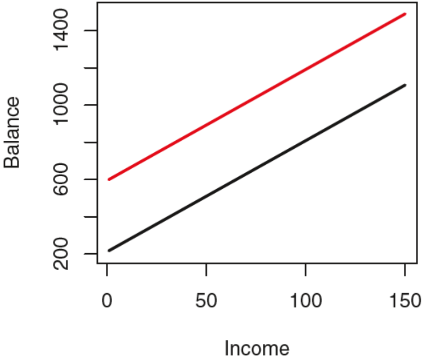
	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
ethnicity[Asian]	-18.69	65.02	-0.287	0.7740
ethnicity[Caucasian]	-12.50	56.68	-0.221	0.8260

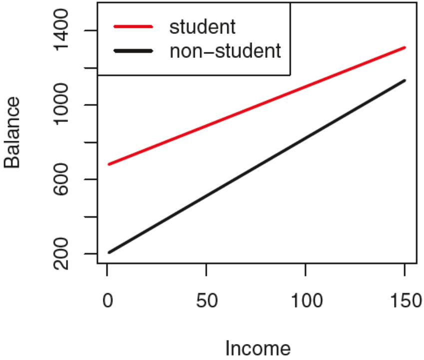
No Interaction vs Interaction

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$

$$\beta_0 + \beta_2 + (\beta_1 + \beta_3) x_{i1} \text{ if student}$$

$$\beta_0 + \beta_1 x_{i1} \text{ if not student}$$





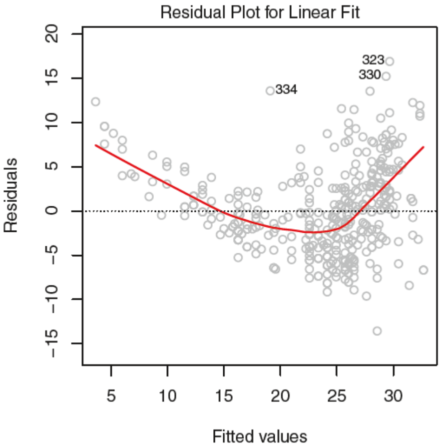
lm.fit=lm(Sales~.+Income:Advertising+Price:Age, data=Carseats); contrasts(ShelveLoc)

```
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        6.5755654  1.0087470   6.519 2.22e-10 ***
CompPrice          0.0929371  0.0041183  22.567  < 2e-16 ***
Income             0.0108940  0.0026044   4.183 3.57e-05 ***
Advertising        0.0702462  0.0226091   3.107 0.002030 ** 
Population         0.0001592  0.0003679   0.433 0.665330    
Price             -0.1008064  0.0074399 -13.549  < 2e-16 ***
ShelveLocGood      4.8486762  0.1528378  31.724  < 2e-16 ***
ShelveLocMedium    1.9532620  0.1257682  15.531  < 2e-16 ***
Age               -0.0579466  0.0159506  -3.633 0.000318 ***
Education         -0.0208525  0.0196131  -1.063 0.288361    
UrbanYes           0.1401597  0.1124019   1.247 0.213171    
USYes             -0.1575571  0.1489234  -1.058 0.290729    
Income:Advertising 0.0007510  0.0002784   2.698 0.007290 ** 
Price:Age          0.0001068  0.0001333   0.801 0.423812    
```

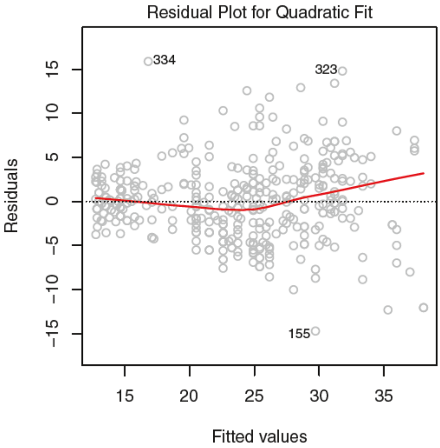
	Good	Medium
Bad	0	0
Good	1	0
Medium	0	1

Residuals vs Fitted Values

mpg on HP

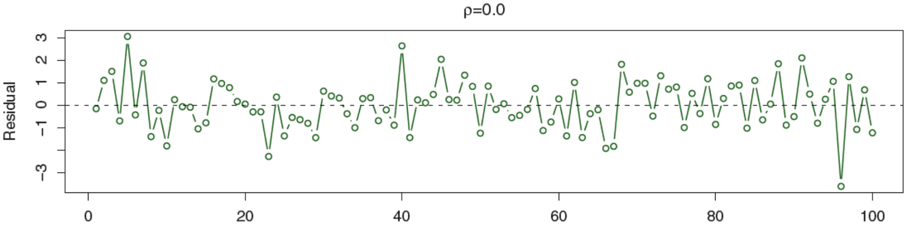


mpg on HP HP^2

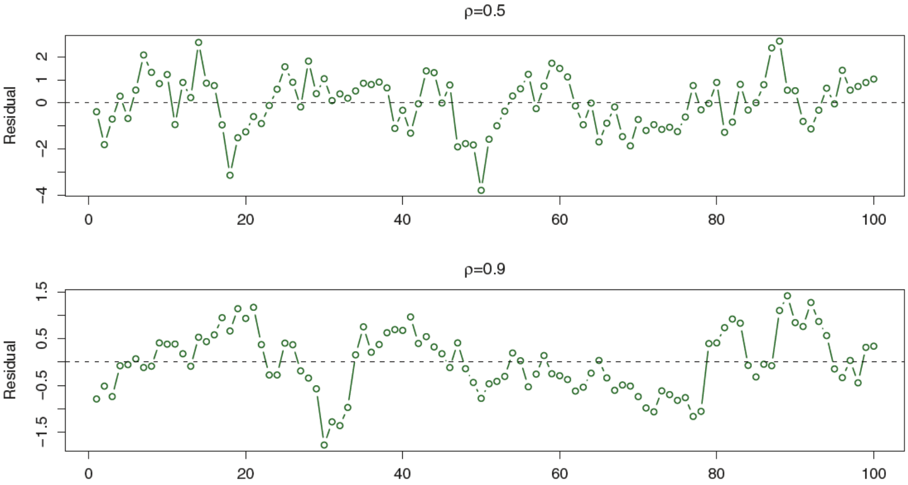


Uncorrelated Error Terms

$$Cov(\epsilon_t, \epsilon_s | X) = 0, \text{ for all } t \neq s$$

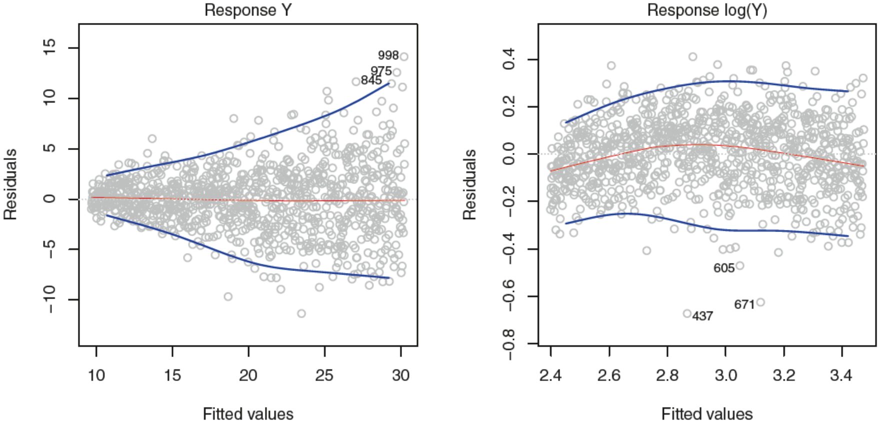


Correlated Error Terms



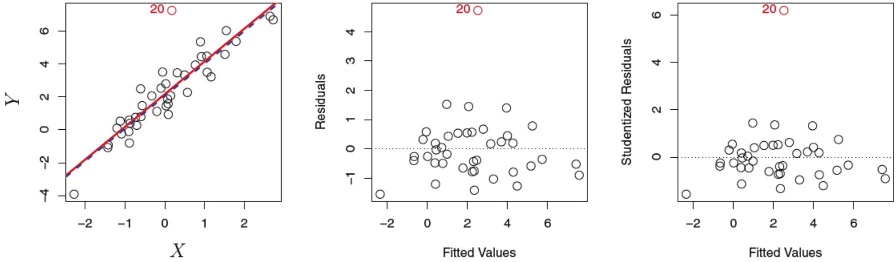
Homoscedasticity vs Heteroscedasticity

$$Var(\epsilon_i) = \sigma^2$$



Studentized Residuals (SR)

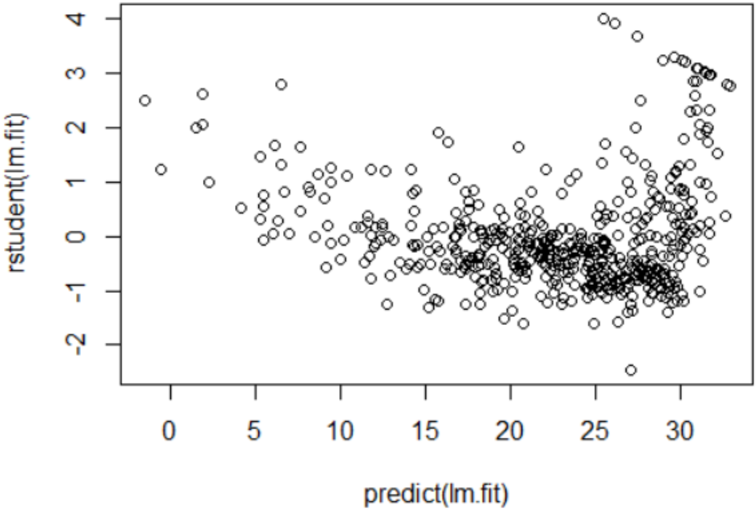
SR is computed by dividing each e_i by its estimated standard error



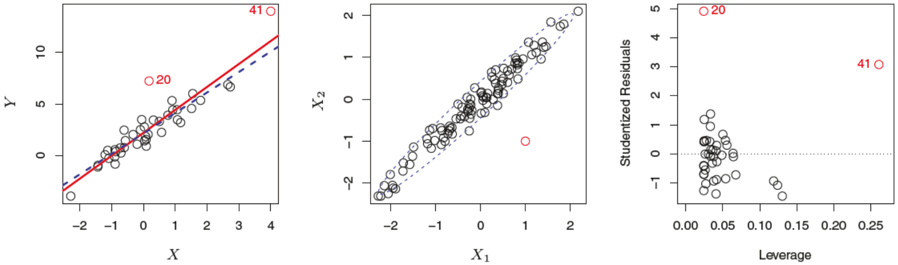
$>|3|$ are possible outliers

lm.fit=lm(medv~lstat,data=Boston)

plot(predict(lm.fit), rstudent(lm.fit))



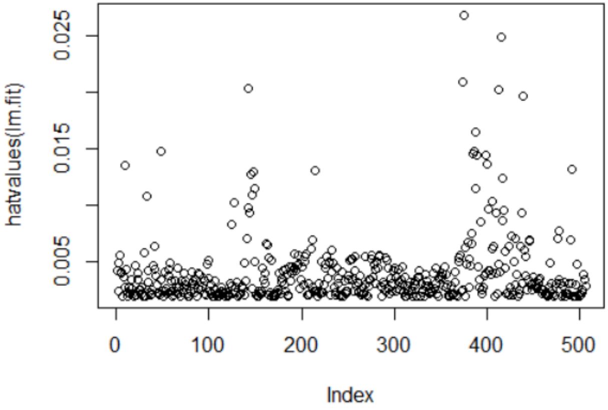
High Leverage Points (Unusual Value for x_i)



$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

lm.fit=lm(medv~lstat,data=Boston)

plot(hatvalues(lm.fit))



which.max(hatvalues(lm.fit)) **375**