introduction to statistical learning

chapter iii summary - linear regression

simple linear regression

$$Y \approx \beta_0 + \beta_1 X$$

estimating coefficients

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

estimating coefficients

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
$$e_i = y_i - \hat{y}_i$$

estimating coefficients

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$e_i = y_i - \hat{y}_i$$

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

estimating coefficients

coefficients value

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

estimating coefficients

coefficients error

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\sigma} = \sqrt{RSS/n-2}$$

assessing the coefficients

confidence interval beta

$$\hat{\beta}_0 \pm \mathbf{2} \cdot SE(\hat{\beta}_0)$$

$$\hat{\beta}_1 \pm \mathbf{2} \cdot SE(\hat{\beta}_1)$$

assessing the coefficients

hypothesis testing

$$H_0 : \hat{\beta}_1 = 0$$
$$H_1 : \hat{\beta}_1 \neq 0$$

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

assessing the coefficients

statsmodel

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                Lottery   R-squared:                       0.133
Model:                            OLS   Adj. R-squared:                  0.123
Method:                 Least Squares   F-statistic:                     12.89
Date:                Thu, 22 Aug 2019   Prob (F-statistic):           0.000555
Time:                        21:40:01   Log-Likelihood:                -392.11
No. Observations:                  86   AIC:                             788.2
Df Residuals:                      84   BIC:                             793.1
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     64.0896      6.265     10.230      0.000      51.631      76.548
Literacy      -0.5245      0.146     -3.590      0.001      -0.815      -0.234
==============================================================================
Omnibus:                        8.096   Durbin-Watson:                   1.946
Prob(Omnibus):                  0.017   Jarque-Bera (JB):                3.090
Skew:                           0.072   Prob(JB):                        0.213
Kurtosis:                       2.083   Cond. No.                         107.
==============================================================================
```

assessing the model

residual standard error

$$RSE = \sqrt{\frac{1}{n-p-1} RSS}$$

assessing the model

$$R^2$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

estimating the coefficients

coefficients value

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{i1} & x_{i2} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = (X'X)^{-1}X'Y$$

estimating the coefficients

coefficients error

$$SE(\hat{\beta})^2 = \sigma^2 (X'X)^{-1}$$

$$\hat{\sigma}^2 = \frac{e'e}{n-p}$$

assessing the coefficients and model

statsmodel

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Lottery   R-squared:                       0.223
Model:                            OLS   Adj. R-squared:                  0.195
Method:                 Least Squares   F-statistic:                     7.861
Date:                Thu, 22 Aug 2019   Prob (F-statistic):           0.000113
Time:                        21:41:13   Log-Likelihood:                -387.38
No. Observations:                  86   AIC:                             782.8
Df Residuals:                      82   BIC:                             792.6
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     37.6315     10.484      3.589      0.001      16.775      58.488
Literacy      -0.3304      0.153     -2.153      0.034      -0.636      -0.025
Donations      0.0003      0.000      0.695      0.489      -0.001       0.001
Infants        0.0009      0.000      2.892      0.005       0.000       0.001
==============================================================================
Omnibus:                       15.000   Durbin-Watson:                   1.891
Prob(Omnibus):                  0.001   Jarque-Bera (JB):                4.516
Skew:                           0.190   Prob(JB):                        0.105
Kurtosis:                       1.944   Cond. No.                     9.61e+04
==============================================================================
```

questions

F-Test

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$$

questions

Model selection

forward selection

backward selection

mixed selection

questions

model fit

R^2

RSE

questions

predictions

$$SE(\hat{y}_0) = \sqrt{\sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]}$$

$$\hat{y}_0 \pm t_{0.975, n-2} \cdot SE(\hat{y}_0)$$

other considerations

qualitative predictors

binary

$$x_{i2} = \begin{cases} 1 \text{ has some characteristic} \\ 0 \text{ doesn't have that characteristic} \end{cases}$$

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

$$y_i = \begin{cases} \beta_0 + \beta_1 x_{i1} + \beta_2 + \epsilon & \text{if the ith sample is positive} \\ \beta_0 + \beta_1 x_{i1} + \epsilon_i & \text{if the ith sample is not positive} \end{cases}$$

qualitative predictors

multiple

$$x_{i2} = \begin{cases} 1 \text{ if A} \\ 0 \text{ if not A} \end{cases}$$

$$x_{i3} = \begin{cases} 1 \text{ if B} \\ 0 \text{ if not B} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

$$y_i = \begin{cases} \beta_0 + \beta_1 x_{i1} + \beta_2 + \epsilon & \text{if A} \\ \beta_0 + \beta_1 x_{i1} + \beta_3 + \epsilon_i & \text{if B} \\ \beta_0 + \beta_1 x_{i1} + \epsilon_i & \text{if C} \end{cases}$$

extensions

removing additive assumption

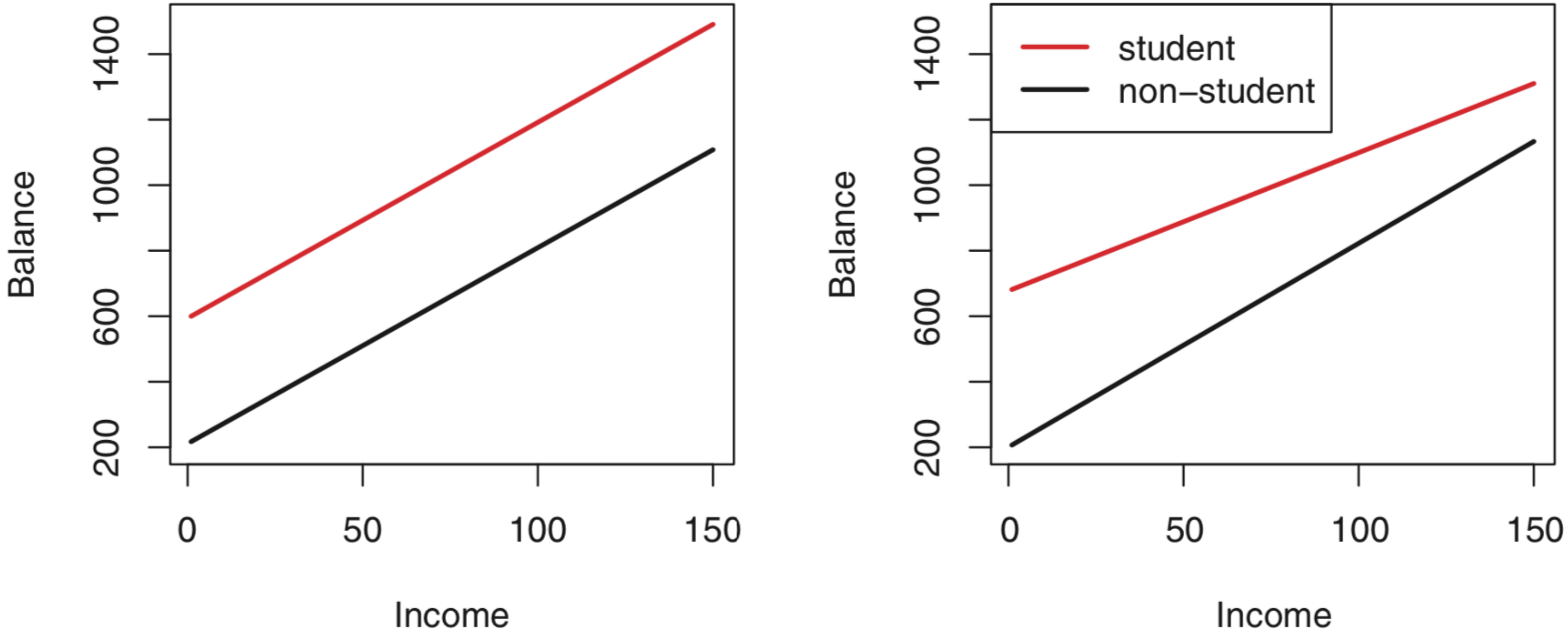
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

$$Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$$

$$Y = \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2 + \epsilon$$

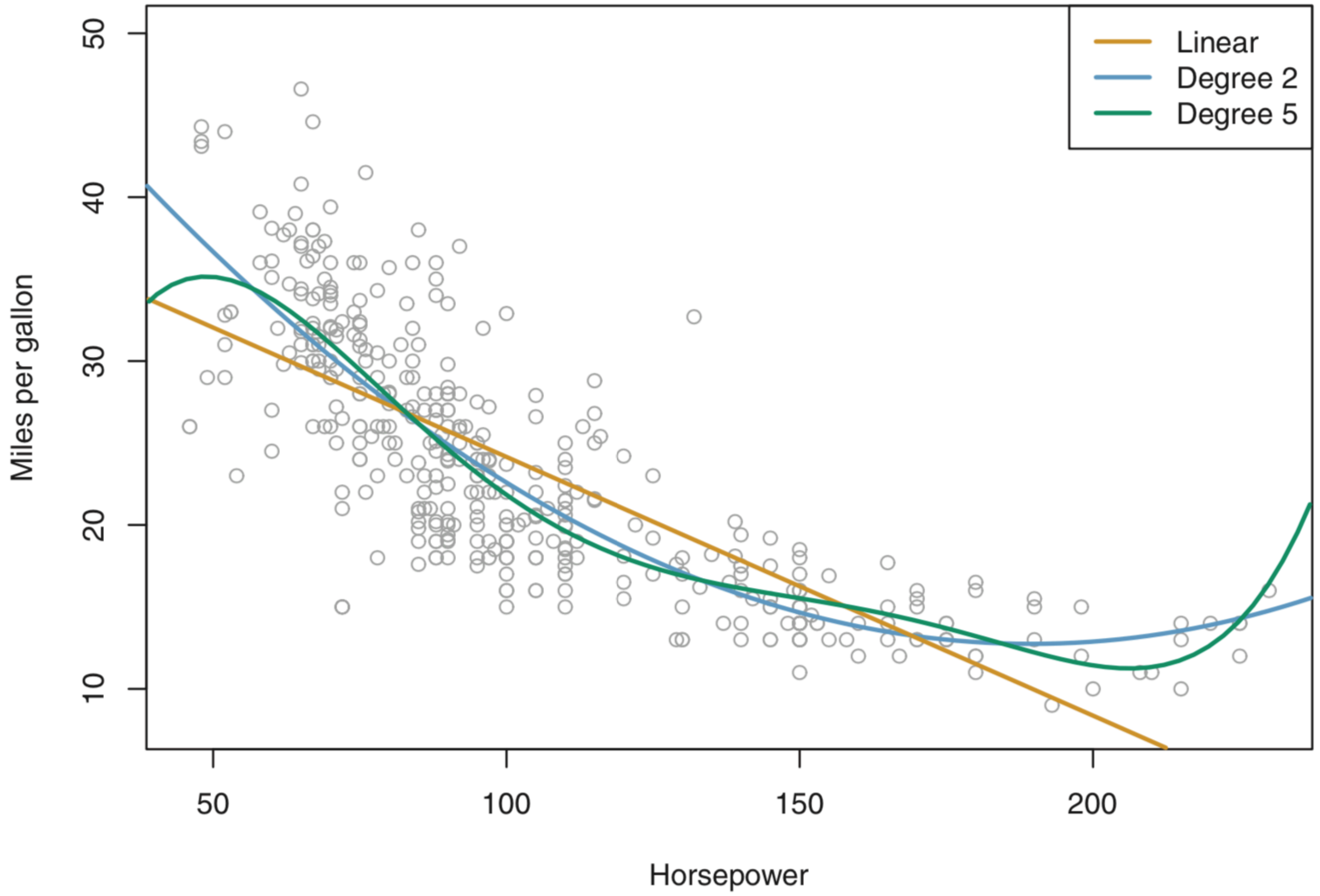
extensions

removing additive assumption



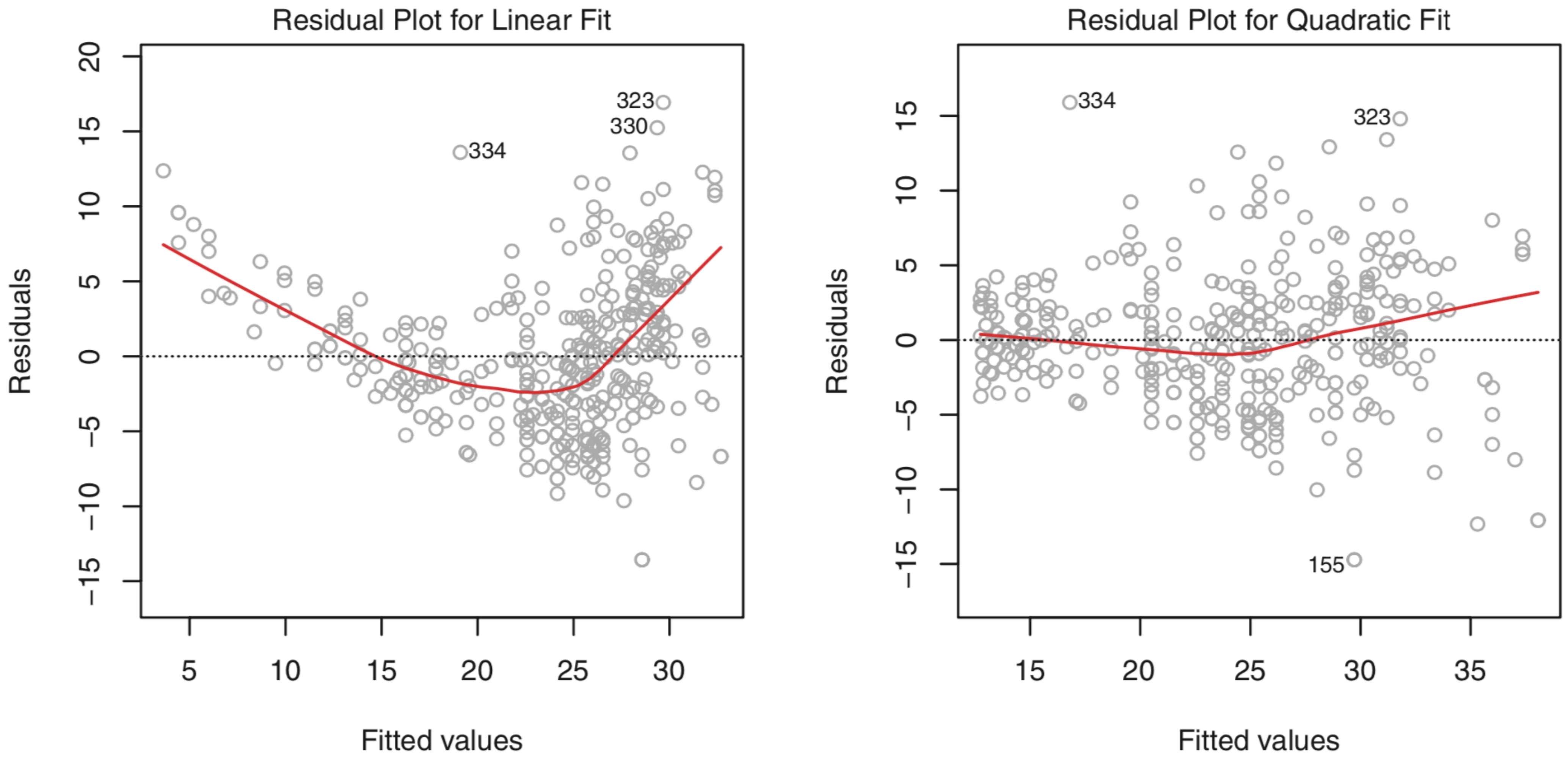
extensions

non-linear relationships



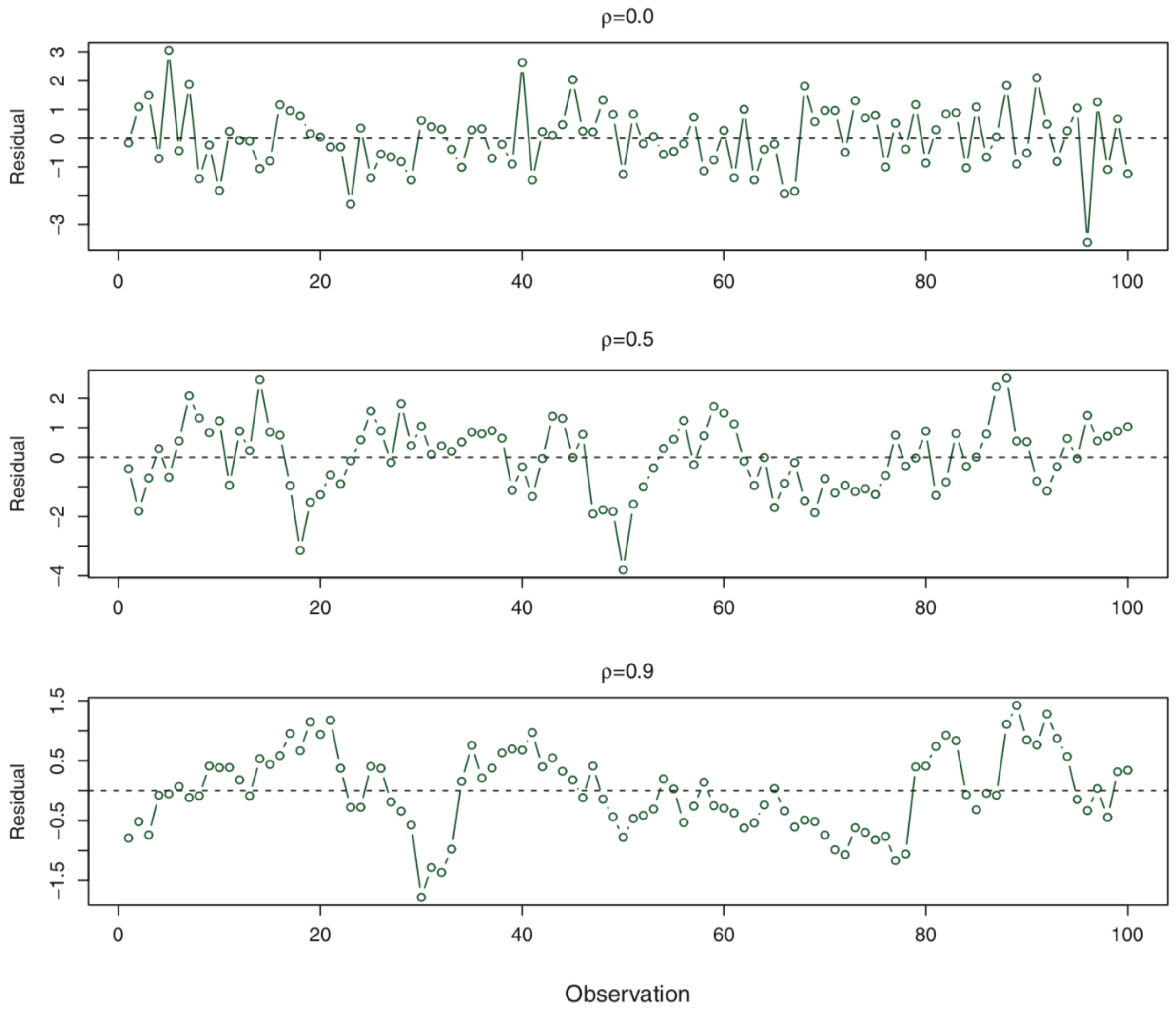
potential problems

non-linearity of the response-predictor relationships



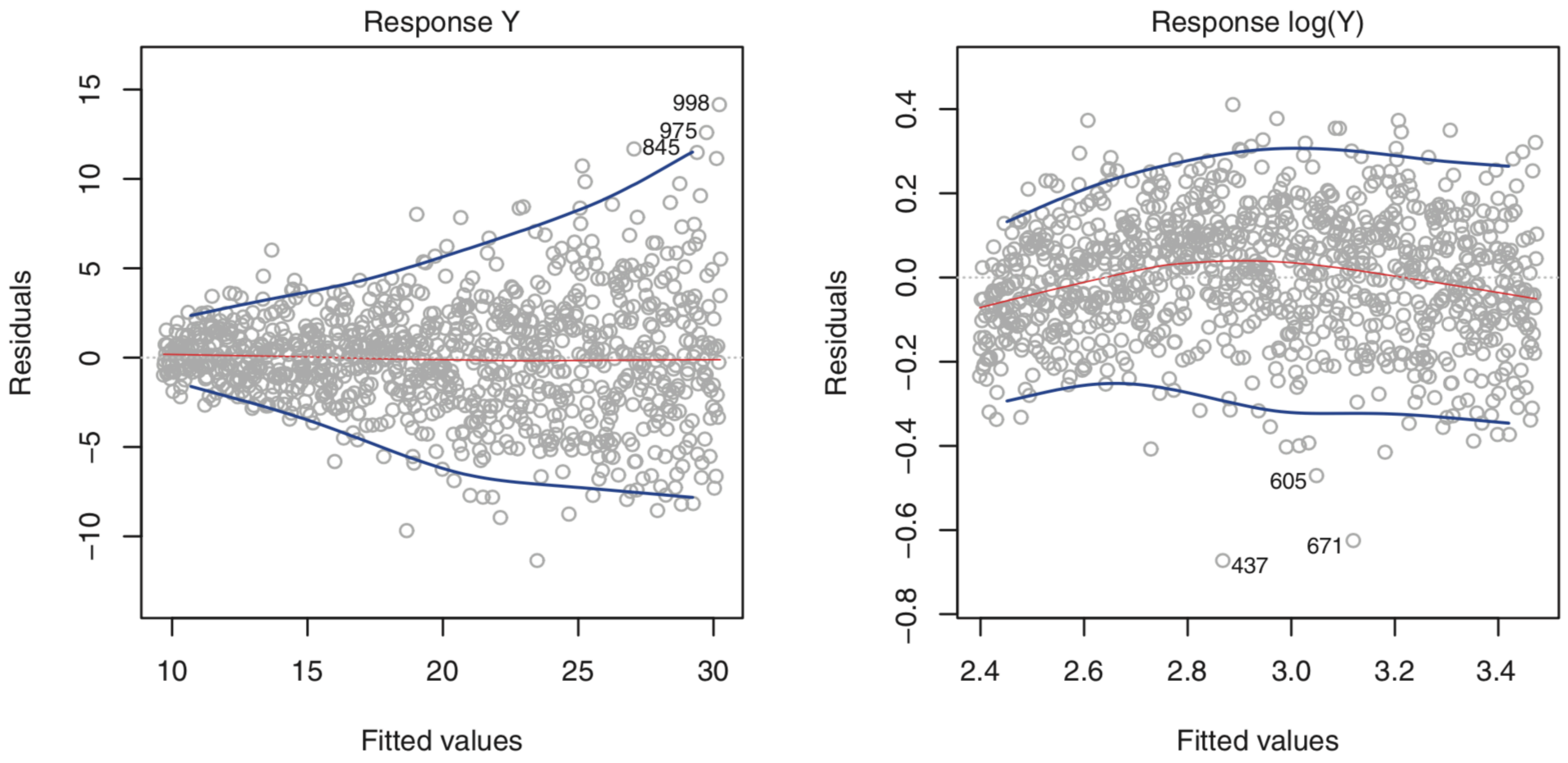
potential problems

correlation of the error terms



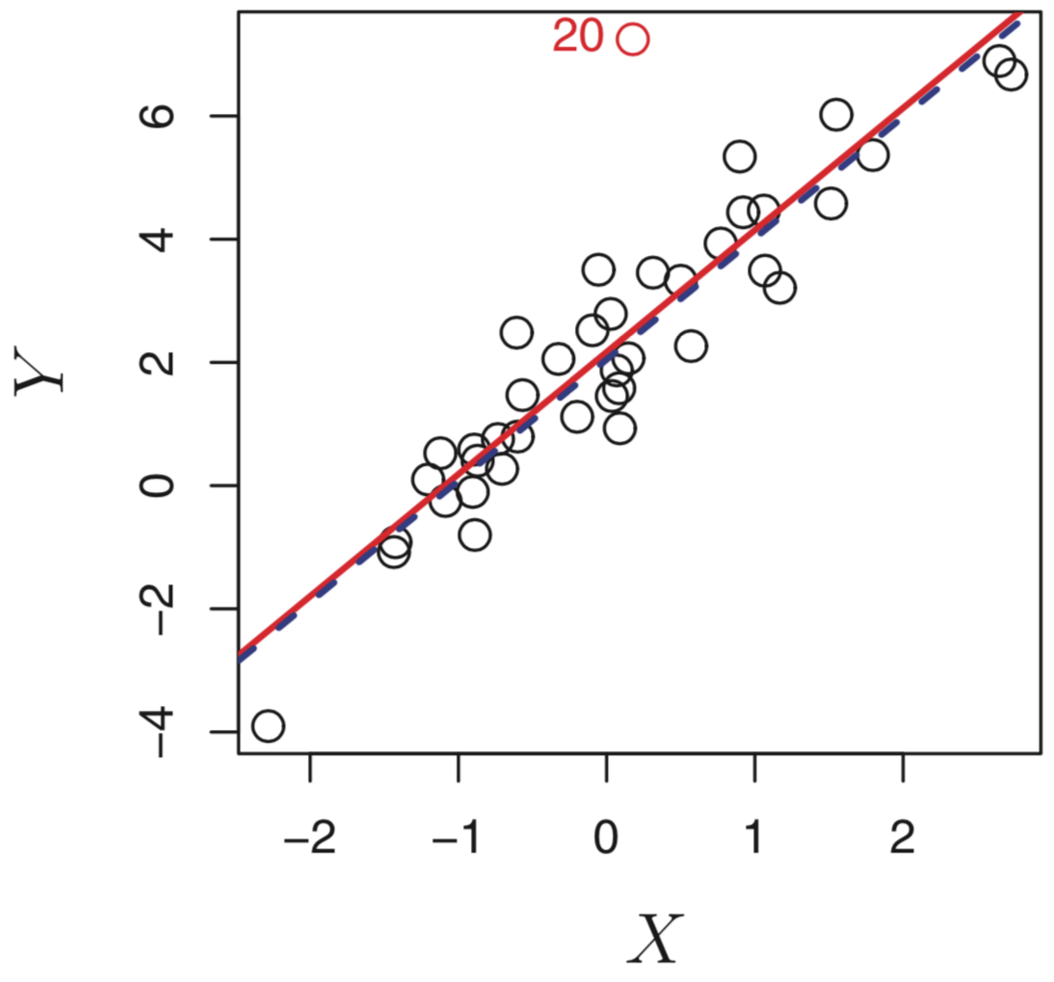
potential problems

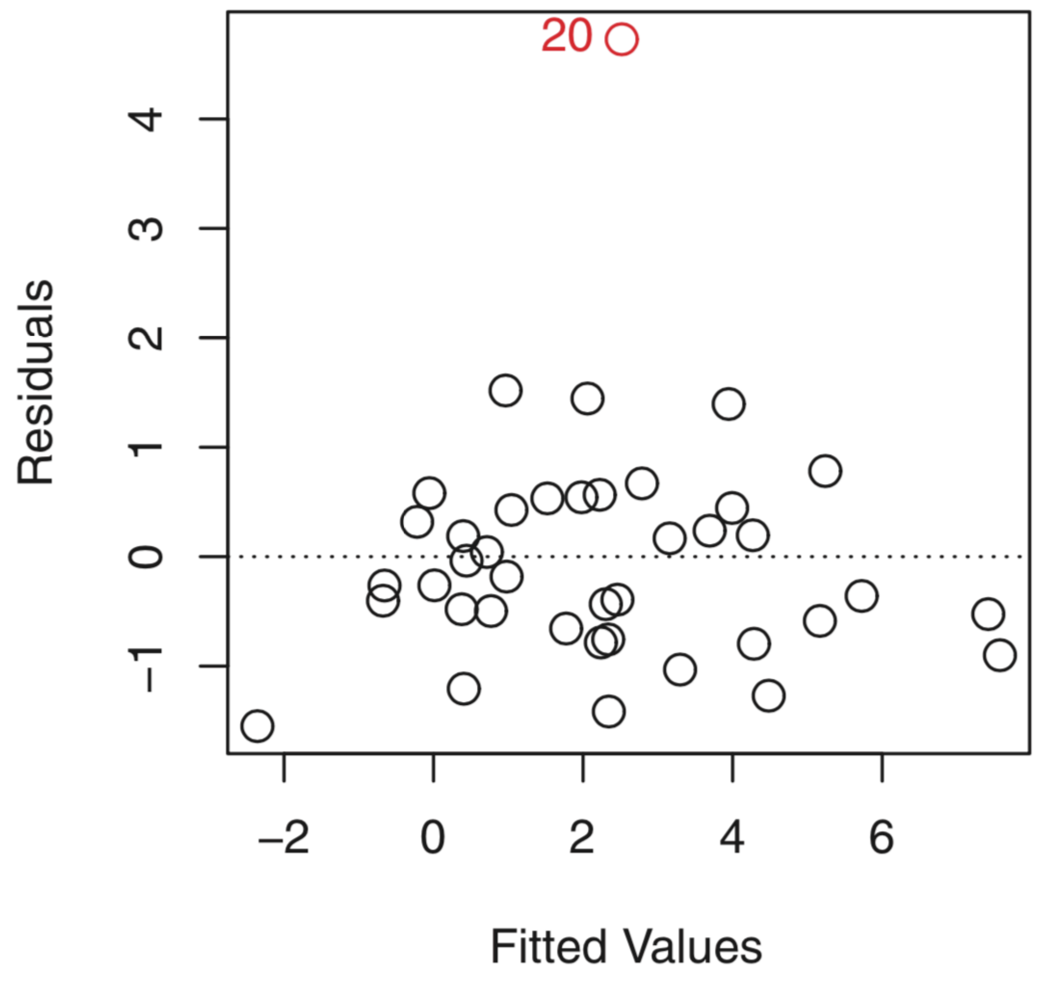
non-constant variance of error terms



potential problems

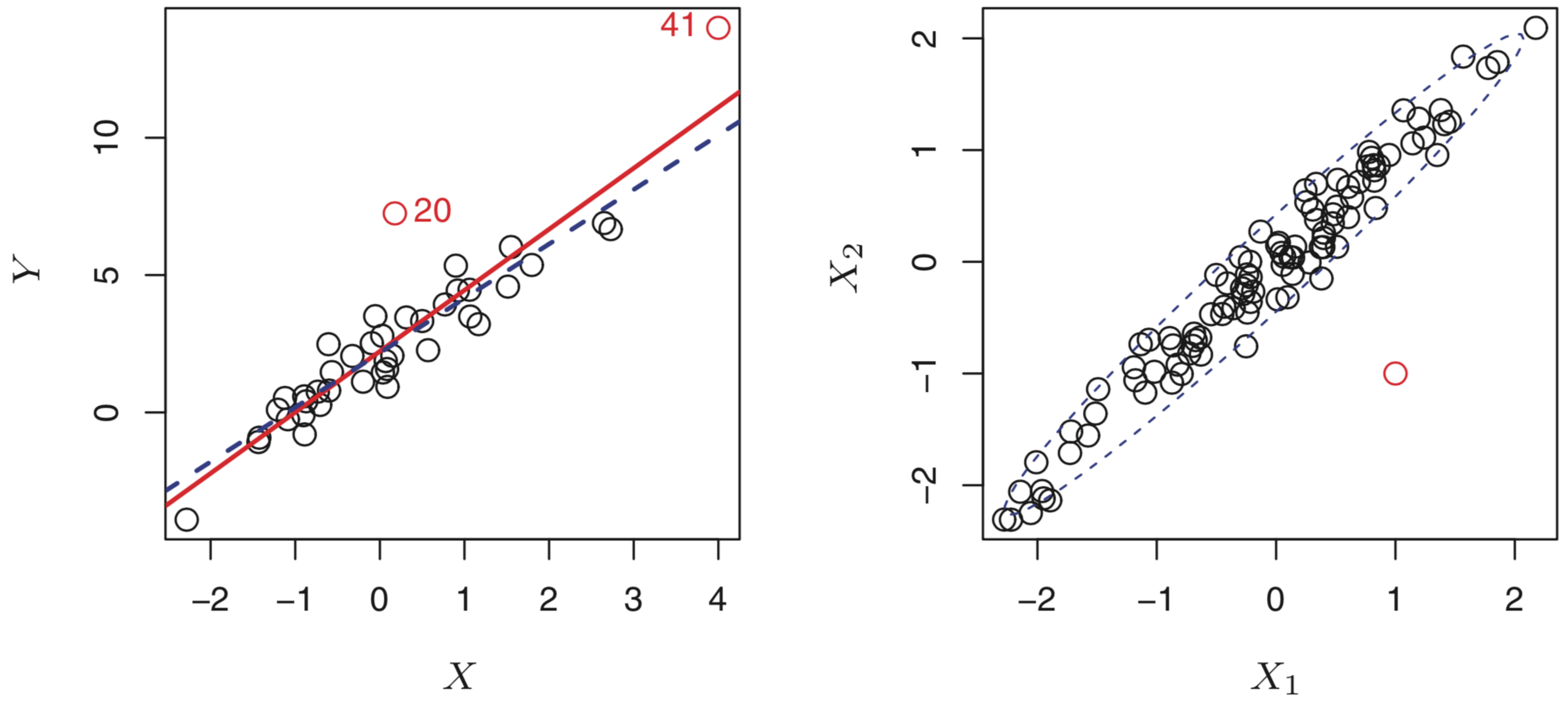
outliers





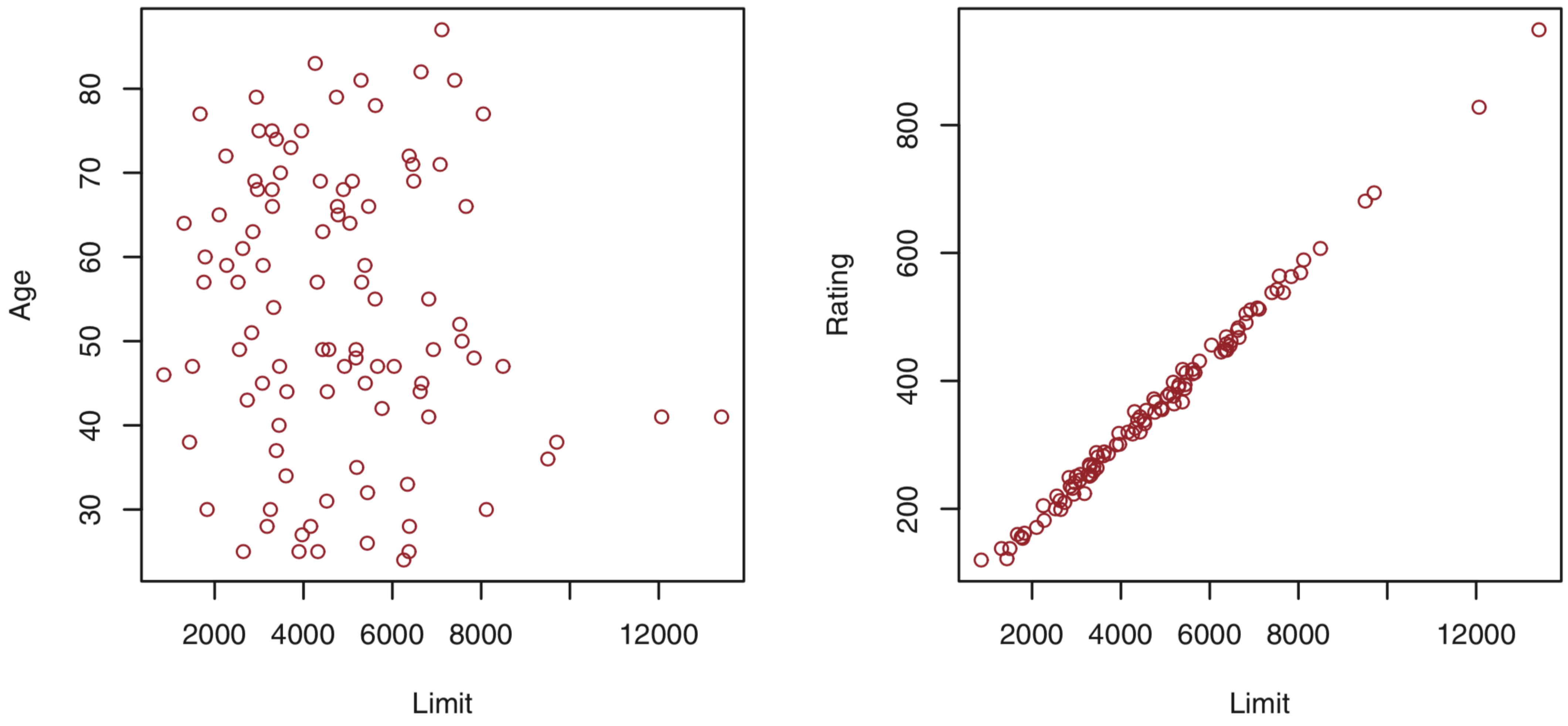
potential problems

high-leverage points



potential problems

collinearity



that's it