Regression

Statistical learning reading group

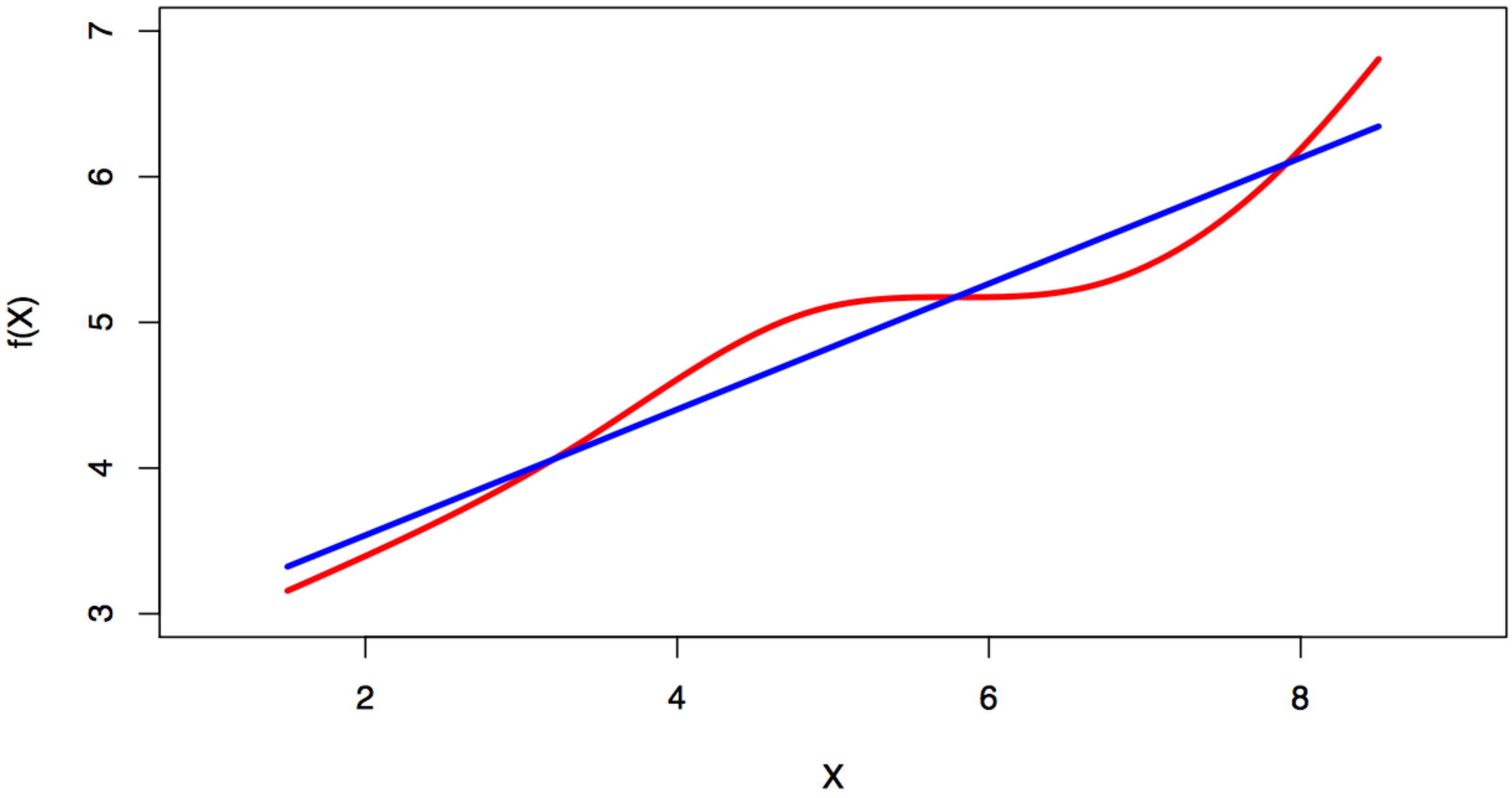
Tahira Jamil

17 November 2015

Linear Regression

- Linear approach to supervise learning
- We seek to identify (or estimate) a continuous variable y associated with a given input vector x.
- They are simple, sometimes outperform fancier nonlinear model (in prediction)
- In modern data analysis, data are high dimensional and we need better regression techniques to handle

True regression function are never linear



REVIEW OF LINEAR Regression analysis

- Simple linear Regression formula
  - In regression we assume that y is a function of x . The exact nature of function is governed by unknown parameters
  - The simple regression model can be represented as follows

slope

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

Error term

Intercept

dependent variable

independent variable / input / feature

Regression Analysis

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

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

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

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

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

Hypothesis testing

$$H_0 : \beta_1 = 0 \longrightarrow Y = \beta_0 + \epsilon,$$

$$H_A : \beta_1 \neq 0, \longrightarrow Y = \beta_0 + \beta_1 X + \epsilon,$$

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} \sim t_{(n-2)}$$

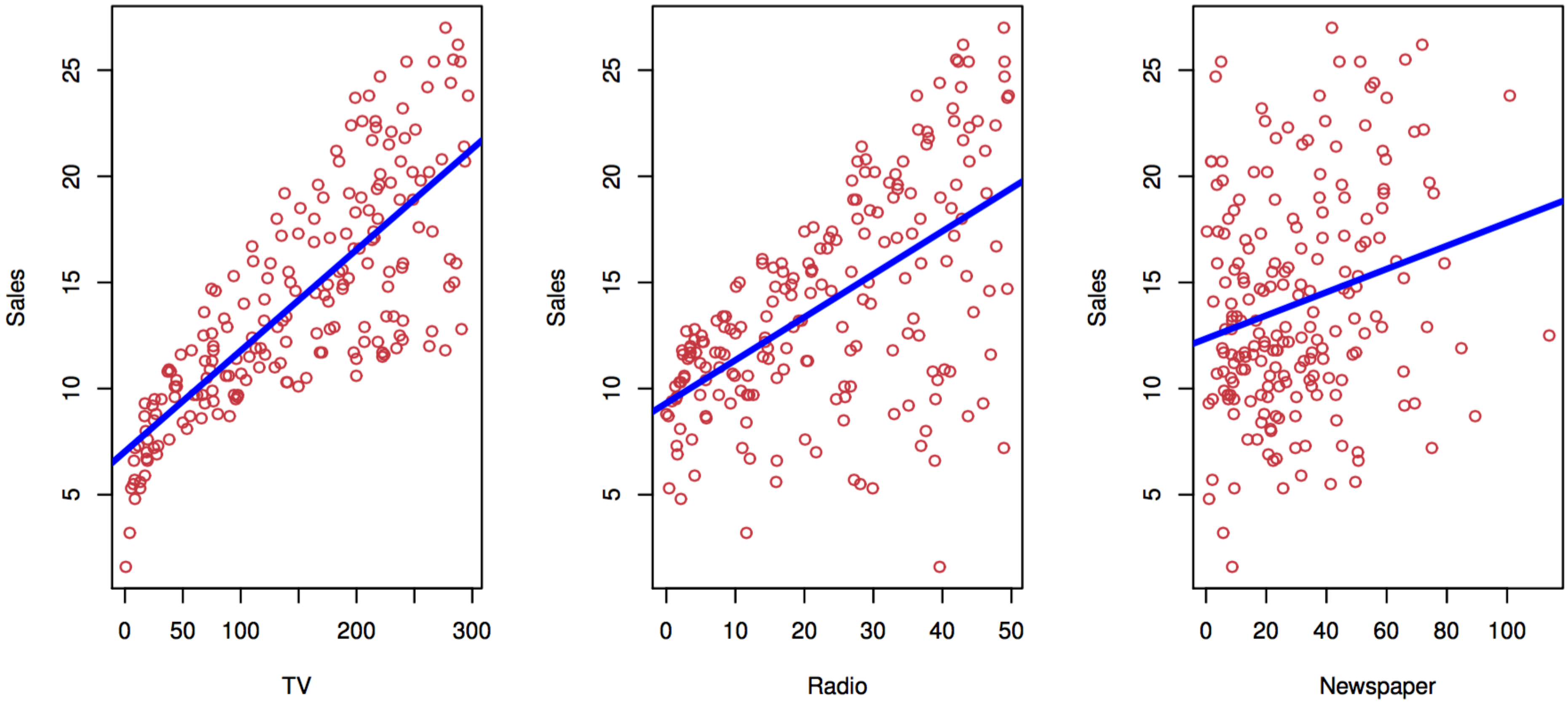
$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Regression Analysis

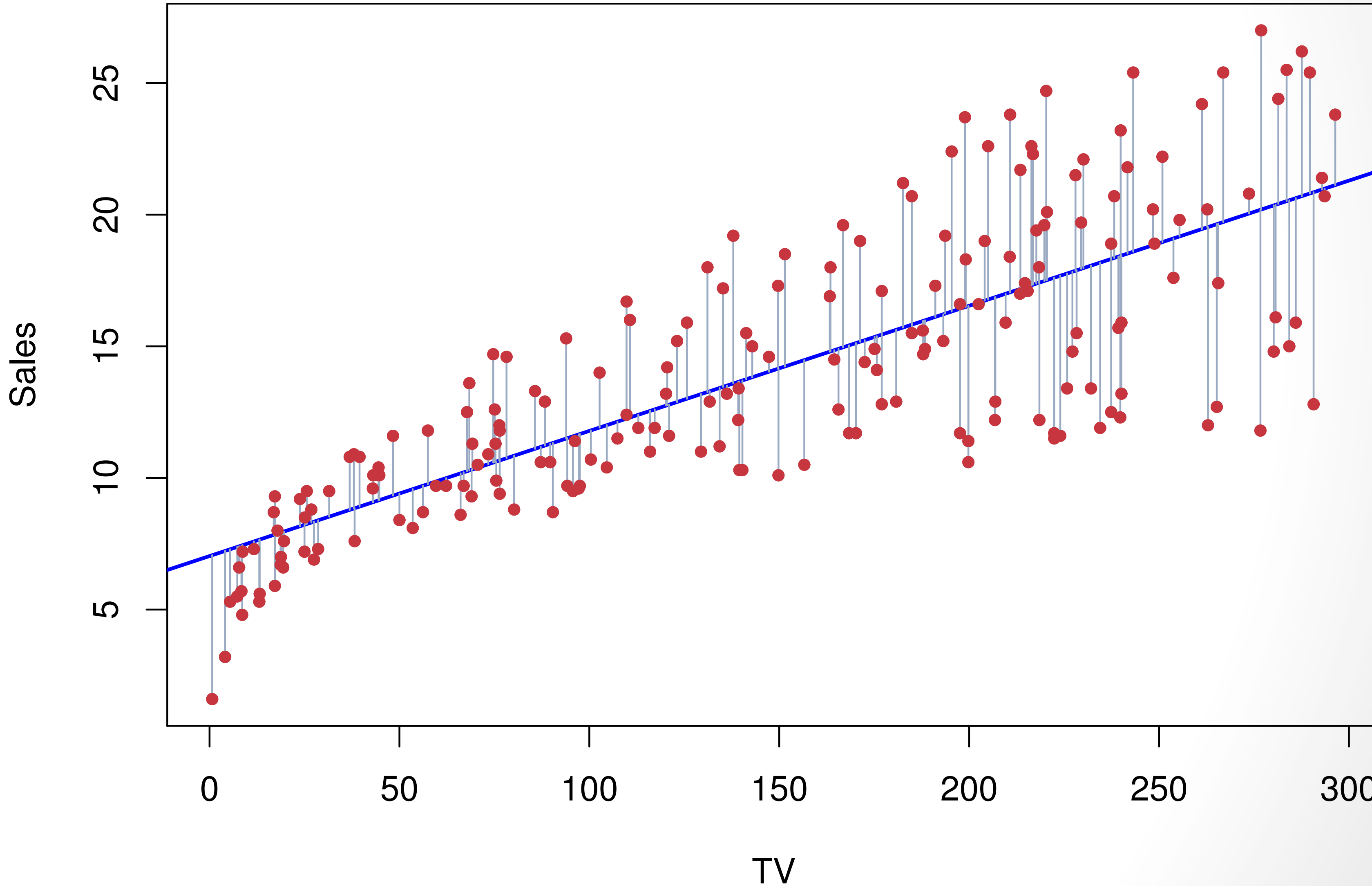
$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

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

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



Sales
25
20
15
10
5
0
50
100
200
300
TV
Sales
25
20
15
10
5
0
10
20
30
40
50
Radio
Sales
25
20
15
10
5
0
20
40
60
80
100
Newspaper


Sales
TV
0
50
100
150
200
250
300
5
10
15
20
25

Simple Linear Regression Analysis

The output of regression analysis will produce a coefficient table similar to the one below

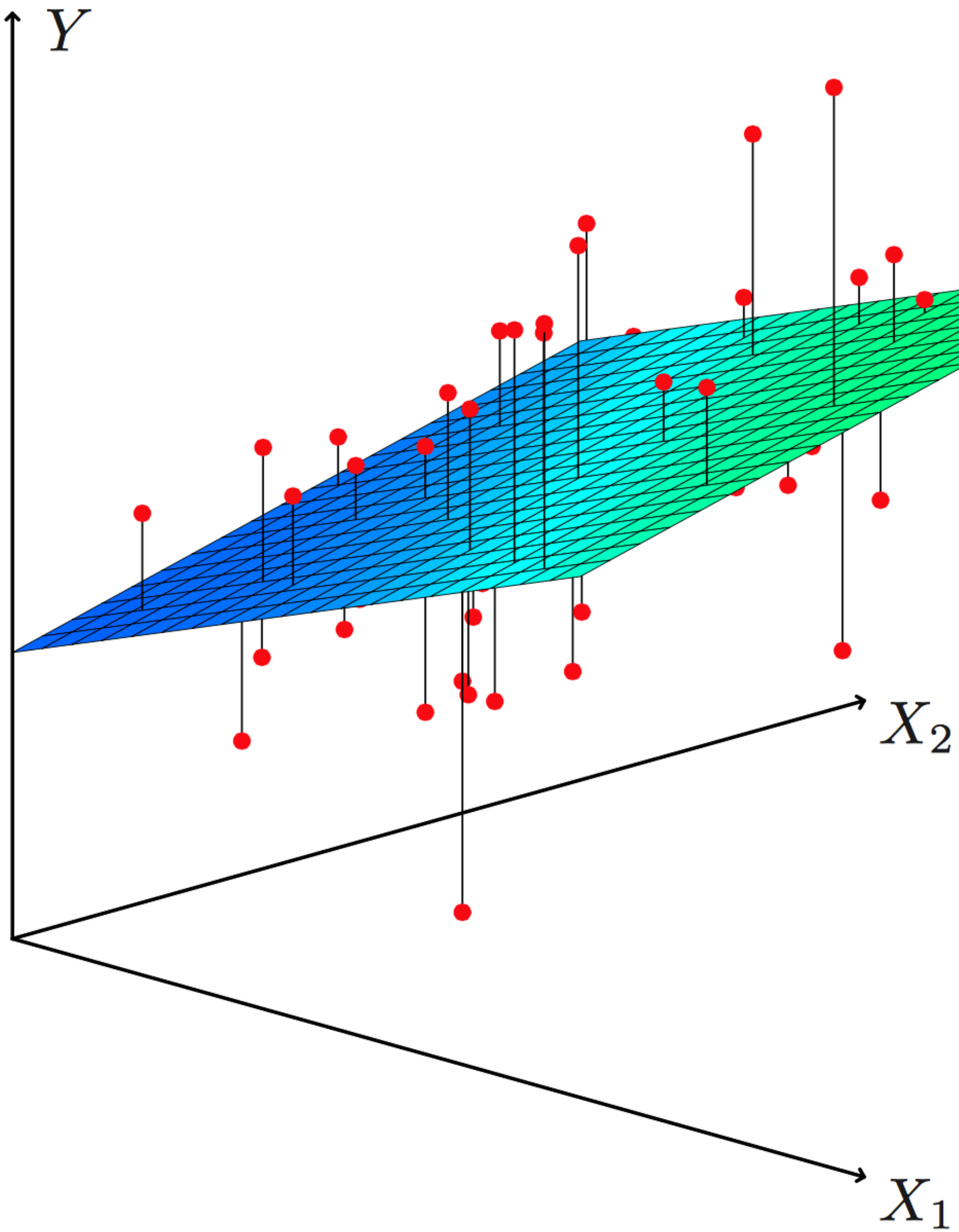
	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1

Multiple linear Regression

- A multiple linear regression is essentially the same the simple linear regression except there are multiple coefficients and independent variables
- Once we fit the function, we can use it to predict the y for new x

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


Y
X2
X1

Multiple linear Regression

- Once we fit the function, we can use it to predict the y for new x

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} \text{TV} \\ \text{Radio} \\ \text{Newspaper} \end{pmatrix}$$

$$Y = f(X) + \epsilon.$$

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

Results for advertising data

	Coefficient	Std. Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Correlations:

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Best subset selection

- For each k ∈ {0, 1, 2, . . . , p} the subset of size k that gives smallest residual sum of squares
- The question of how to choose k involves the tradeoff between bias and variance, along with the more subjective desire for parsimony.
- There are a number of criteria that one may use; typically we choose the smallest model that minimizes an estimate of the expected prediction error.
- These include Mallow's Cp, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R2 and Cross-validation (CV).

Forward- and Backward-Stepwise Selection

- Forward- stepwise selection starts with the intercept, and then sequentially adds into the model the predictor that most improves the fit.
- Computational; for large p we cannot compute the best subset sequence, but we can always compute the forward stepwise sequence (even when $p \gg N$).
- Statistical; a price is paid in variance for selecting the best subset of each size; forward stepwise is a more constrained search, and will have lower variance, but perhaps more bias.

Forward- and Backward-Stepwise Selection

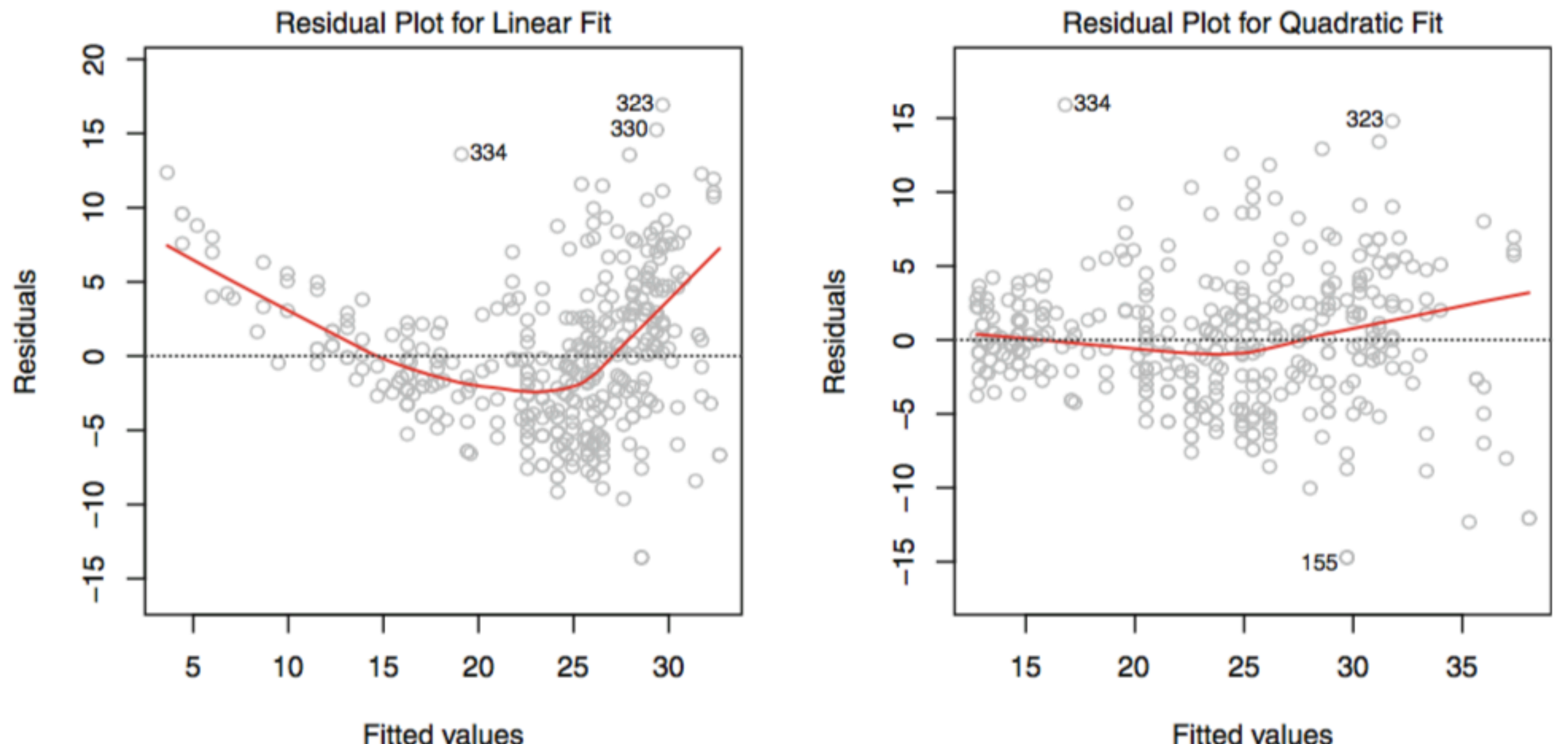
- Backward-stepwise selection starts with the full model, and sequentially deletes the predictor that has the least impact on the fit.
- Backward selection can only be used when $N > p$, while forward stepwise can always be used.
- Some software packages implement hybrid stepwise-selection strategies that consider both forward and backward moves at each step, and select the “best” of the two. For example in the R package the step function uses the AIC criterion for weighing the choices, which takes proper account of the number of parameters fit; at each step an add or drop will be performed that minimizes the AIC score.

Potential Problems

1. Non-linearity of the response-predictor relationships
2. Correlation of error terms
3. Non-constant variance of error terms.
   Outliers
4. High-leverage points
5. Collinearity

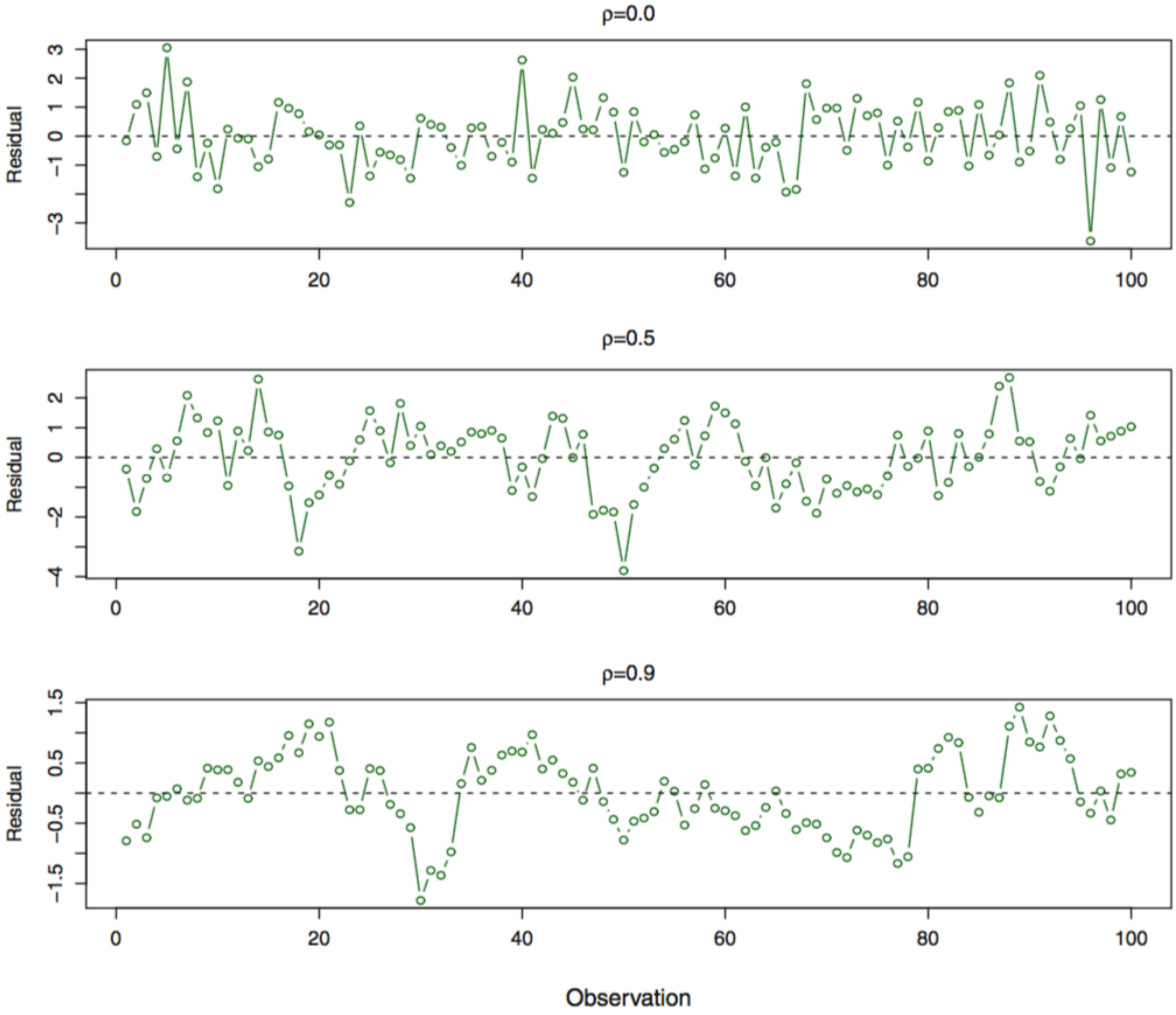
Non-linearity of the Data

- The residual plot will show no discernible pattern
- If indication of non-linear associations, then a simple approach is to use non-linear transformations of the predictors



Correlation of error terms

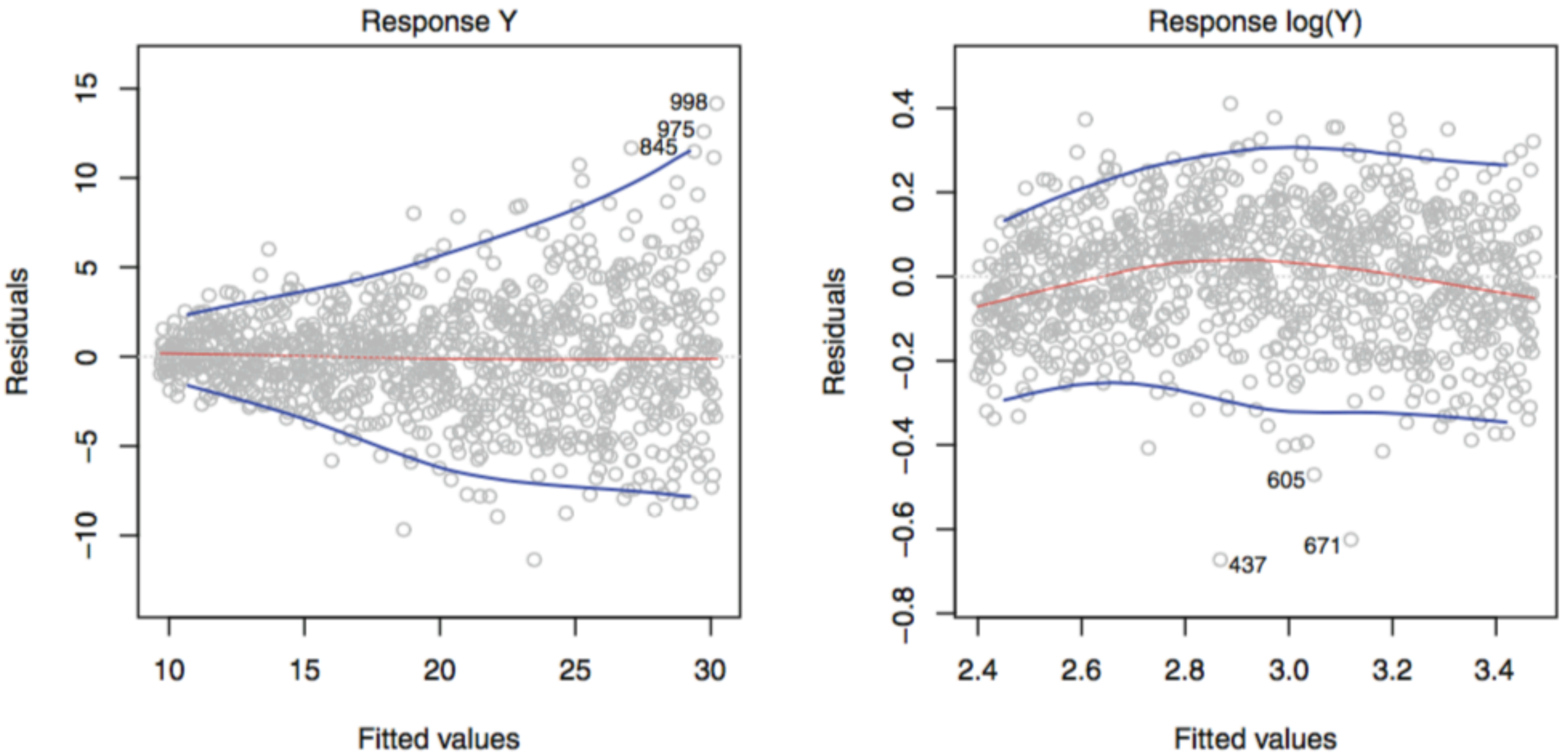
- The error terms, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, are uncorrelated
- If correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors. As a result, confidence and prediction intervals will be narrower than they should be.
- Common in time series data


ρ=0.0
ρ=0.5
ρ=0.9
Residual
Observation

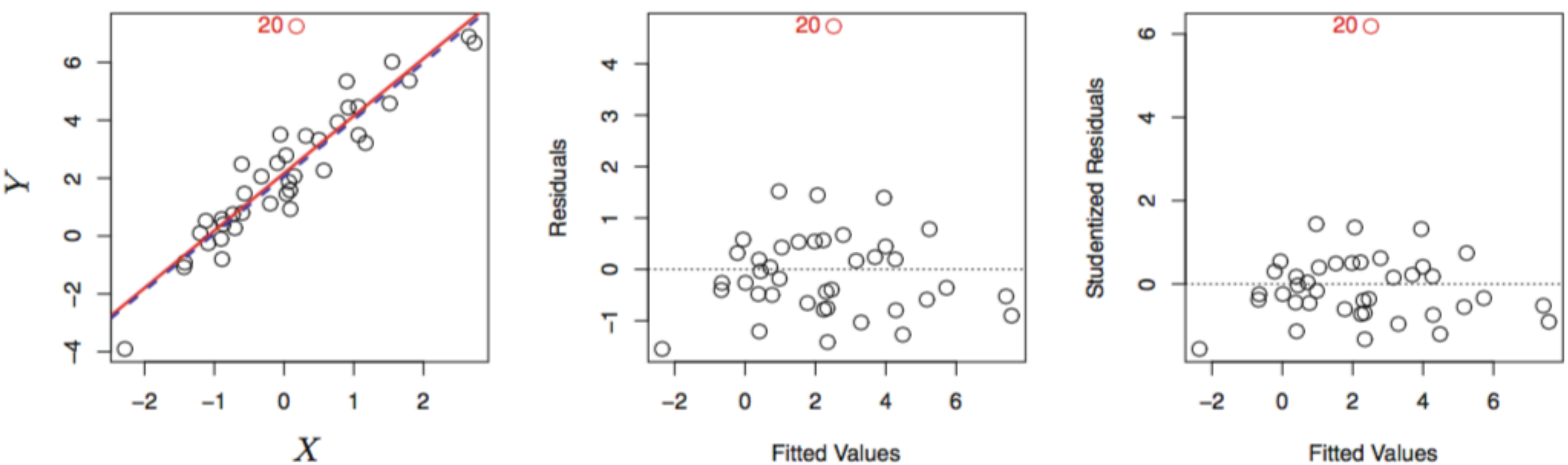
Non-constant Variance of Error Terms or heteroscedasticity

Remedy

- Transformation of response
- weighted least squares,

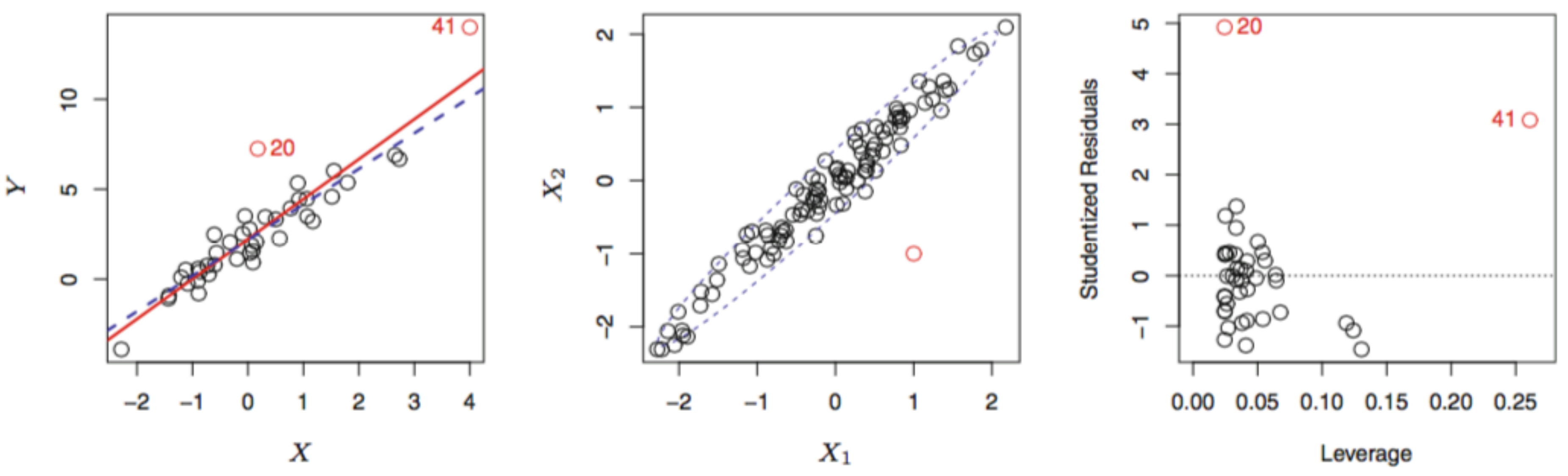


Outliers

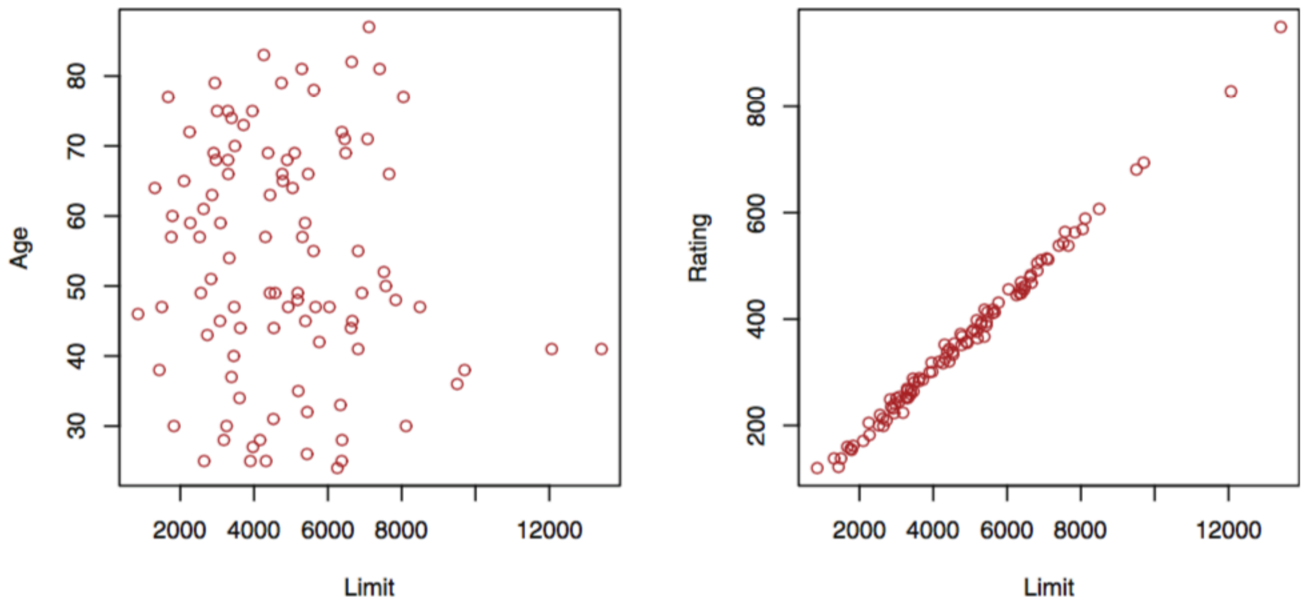

20
Y
X
Residuals
Fitted Values
Studentized Residuals
Fitted Values

High Leverage Points

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



Collinearity



$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

Collinearity

		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012