10) Multiple Linear Regression

Vitor Kamada

January 2020

Reference

Tables, Graphics, and Figures from

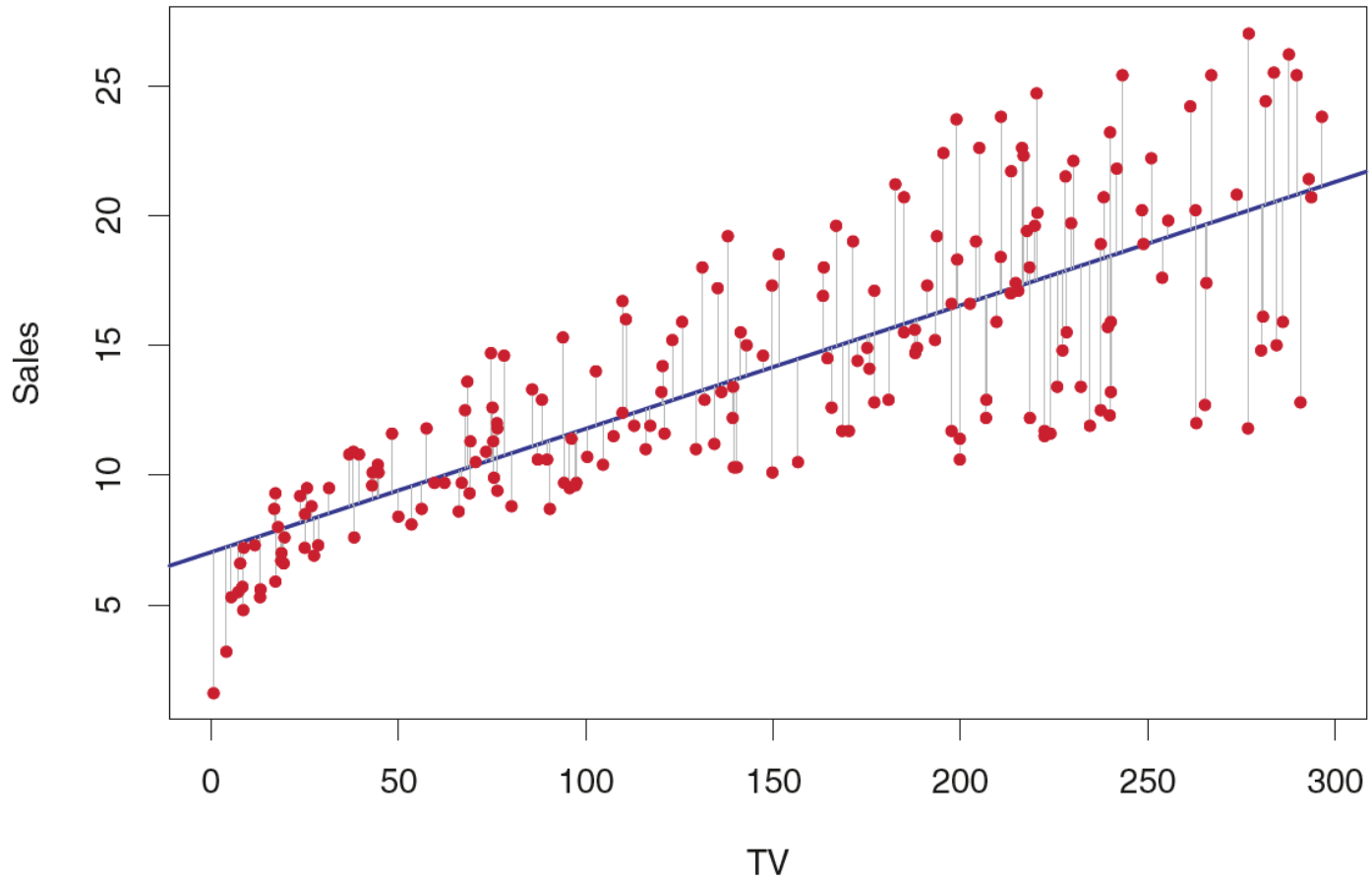
An Introduction to Statistical Learning

James et al. (2017): Chapters: 3.1, 3.2, 3.3

Simple Linear Regression

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 7.03 + 0.0475x$$



Residual Sum of Squares (RSS) $= e_1^2 + e_2^2 + ... + e_n^2$

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 = \sum_{i=1}^{n} e_i$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} [x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)] = 0$$

$$\sum_{i=1}^{n} x_i e_i = 0$$

Estimating the Coefficients

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

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

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

95% Confidence Interval

$$[\hat{\beta}_1 - 2SE(\hat{\beta}_1), \hat{\beta}_1 + 2SE(\hat{\beta}_1)]$$

CI for β_1 is $[0.042, 0.053]$

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

CI for β_0 is $[6.130, 7.935]$

Standard Errors

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

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

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

$$\sigma = RSE = \sqrt{\frac{RSS}{n-2}}$$

Hypothesis Tests

$$H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0$$

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

	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

Residual Standard Error

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

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

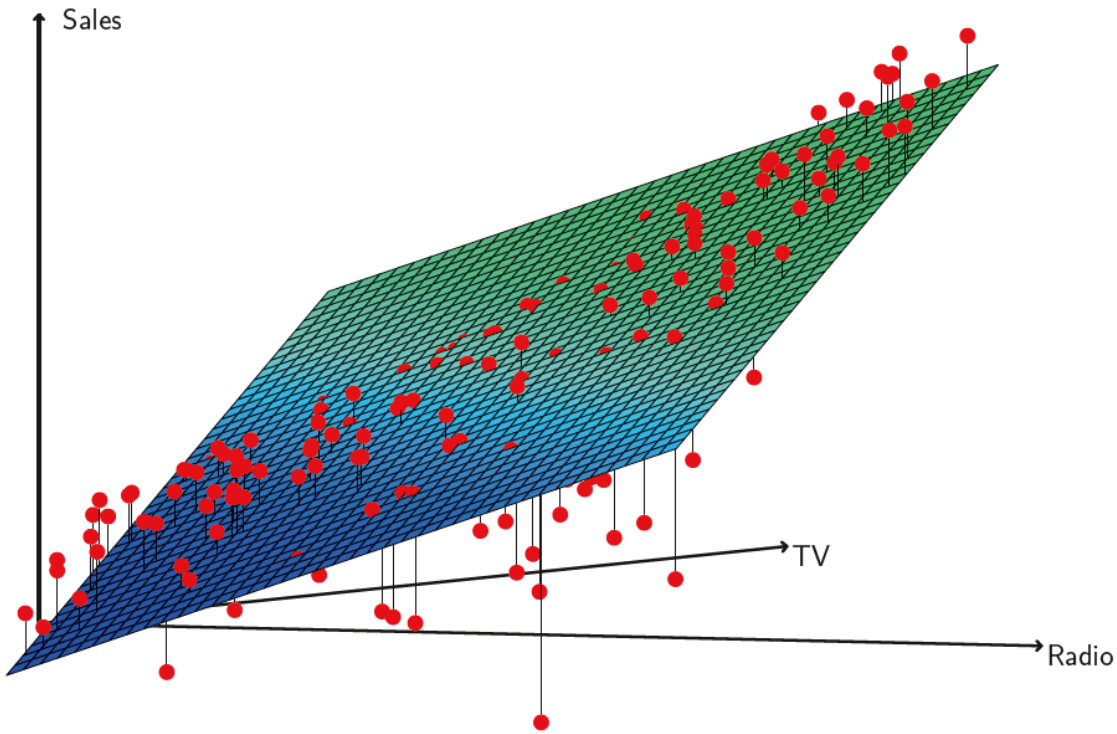
$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

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

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$



Simple vs Multiple Regression

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

	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

Correlation Matrix

$$r = \frac{cov(x,y)}{s_x s_y}$$

	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

F-statistic

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

$$H_a : \text{at least one } \beta_j \neq 0$$

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

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

Test if q Coefficients are Zero

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$$

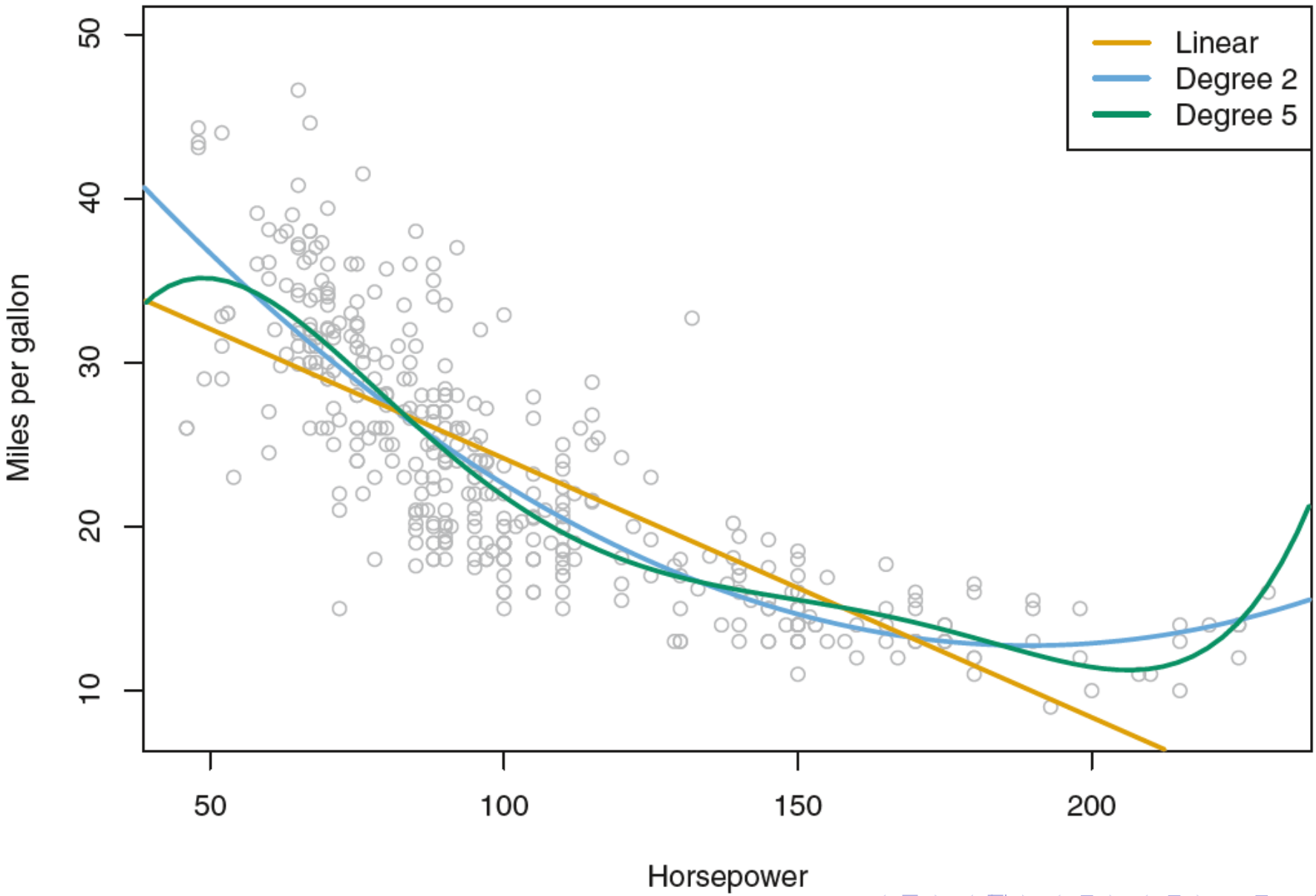
$$F = \frac{\frac{RSS_r - RSS_{ur}}{q}}{\frac{RSS_{ur}}{n-p-1}}$$

Interaction Effect

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

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Auto Data Set



Non-linear Relationships

$$mpg = \beta_0 + \beta_1 hp + \beta_2 hp^2 + \epsilon$$

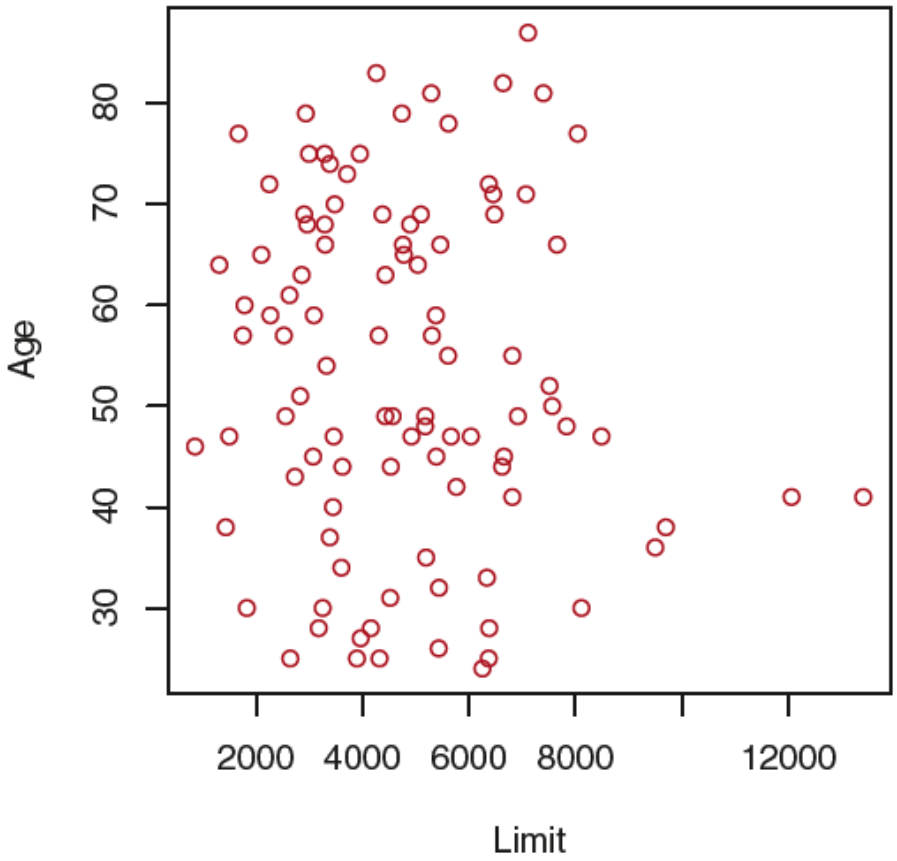
	Coefficient	Std. error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

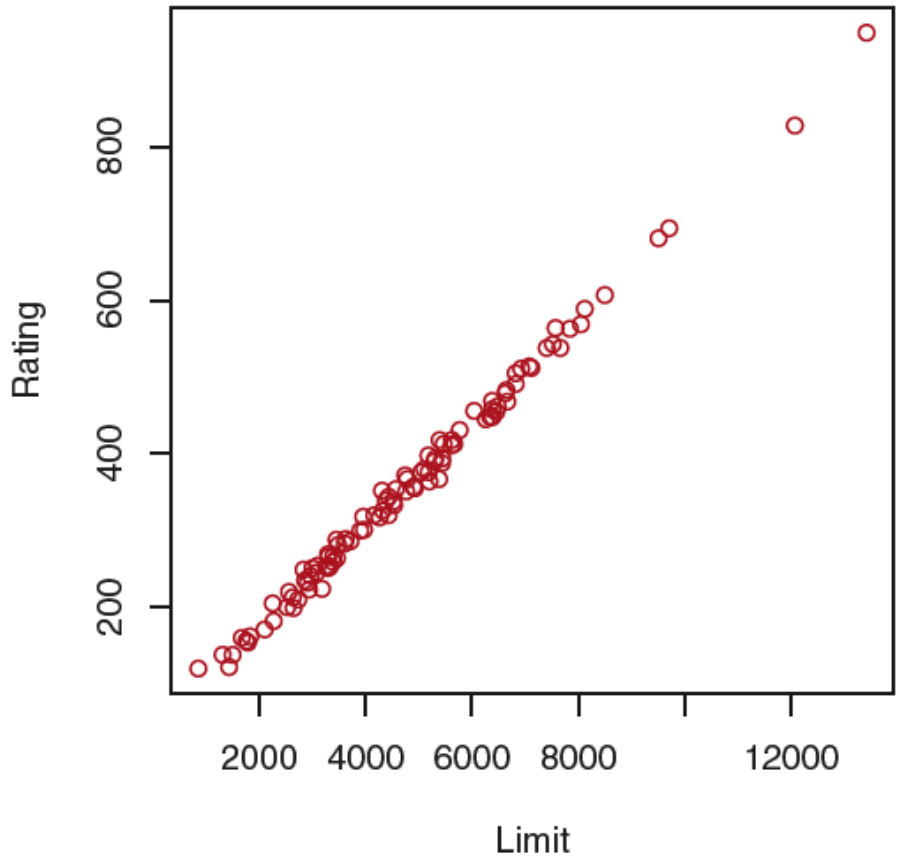
Multicollinearity

Y = balance

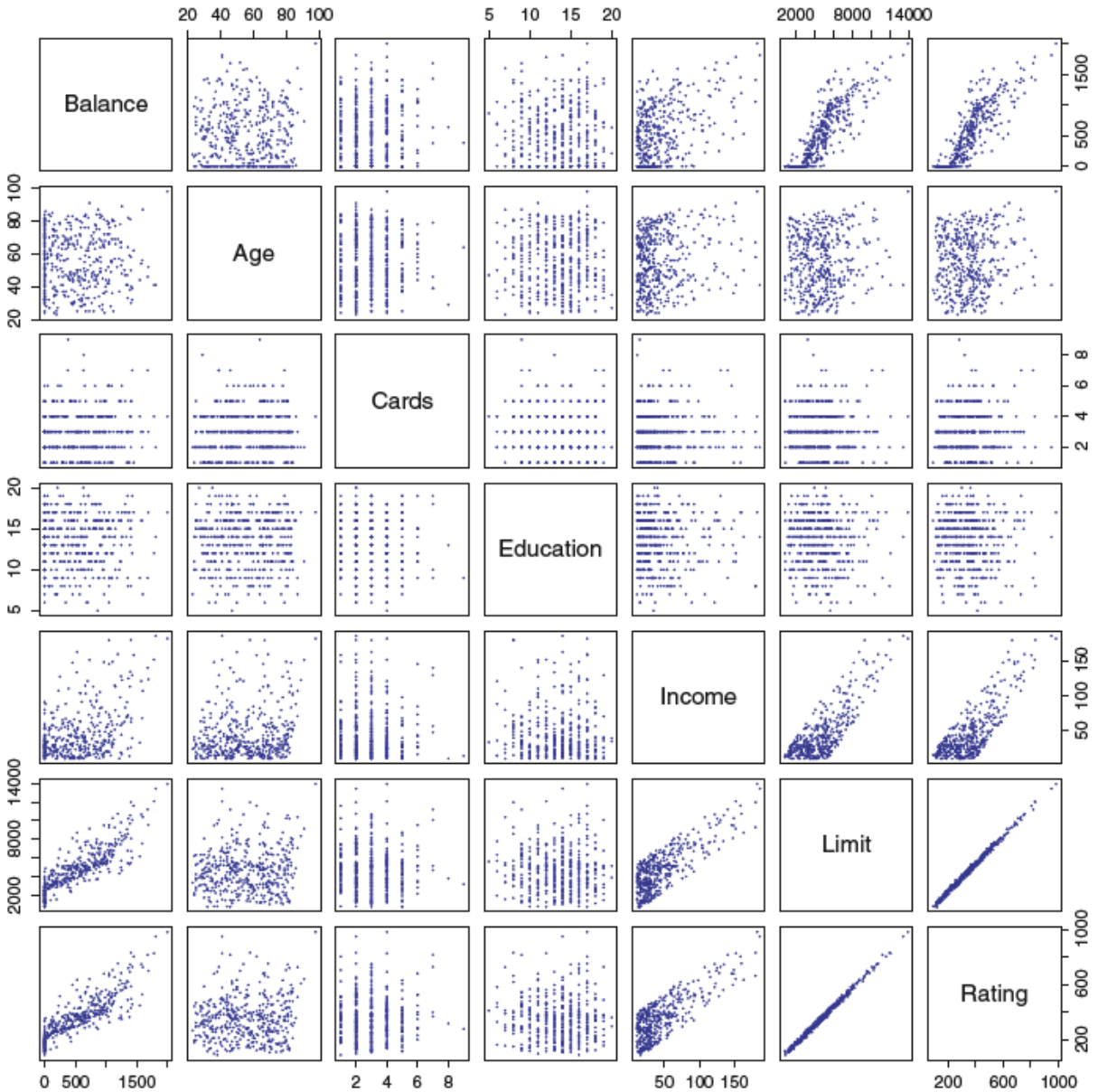
		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012

Scatterplots: Credit Data Set





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

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

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

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