SDSC 3006: Fundamentals of Machine Learning I

Topic 2. Linear Regression

Outline

- Simple linear regression
- Multiple linear regression
- Other considerations in regression
- Model diagnostics

Simple Linear Regression

- Predicting a quantitative response Y based on a single predictor variable X

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

Model parameters (coefficients)

β_0 ---- intercept (i.e., the average value of Y if $X = 0$)

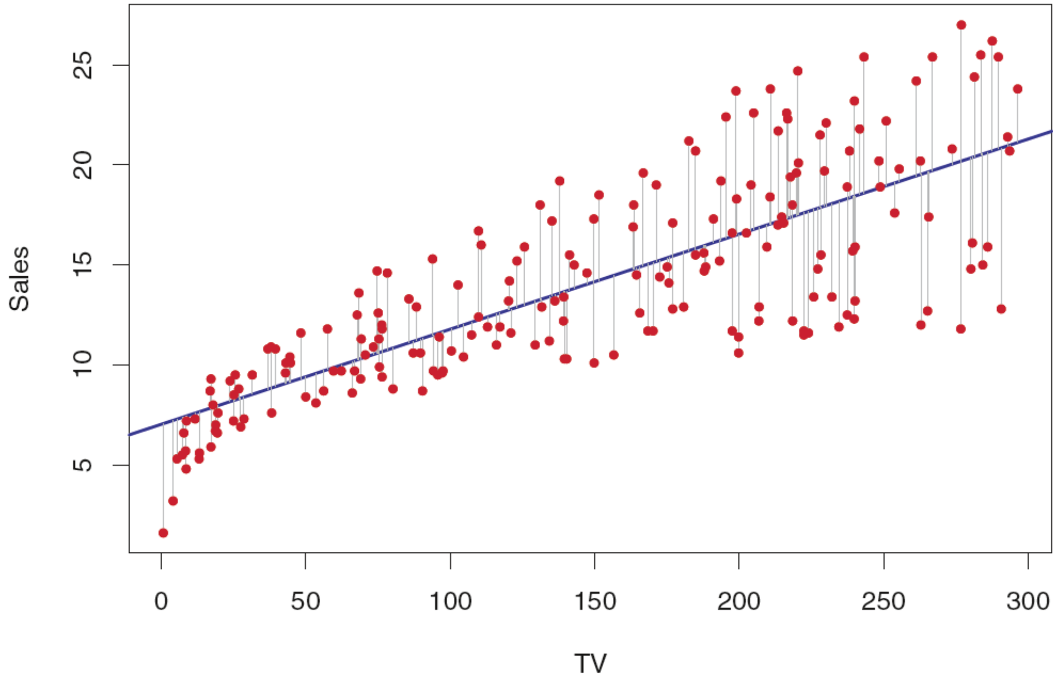
β_1 ---- slope (the average increase in Y when X is increased by 1)

1. Estimating Coefficients

- In model training, we want to find coefficient estimates $\hat{\beta}_0, \hat{\beta}_1$ based on the training data set

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

- Goal: $\hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \dots, n$, fit the actual data well.
- Example: Advertising data set, $X = TV, Y = sales, n = 200$



Least Squares (LS) Method

- Use the least squares criterion to find the coefficient estimates

  Prediction: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
  Residual: $e_i = y_i - \hat{y}_i$
  Residual sum of squares (RSS)

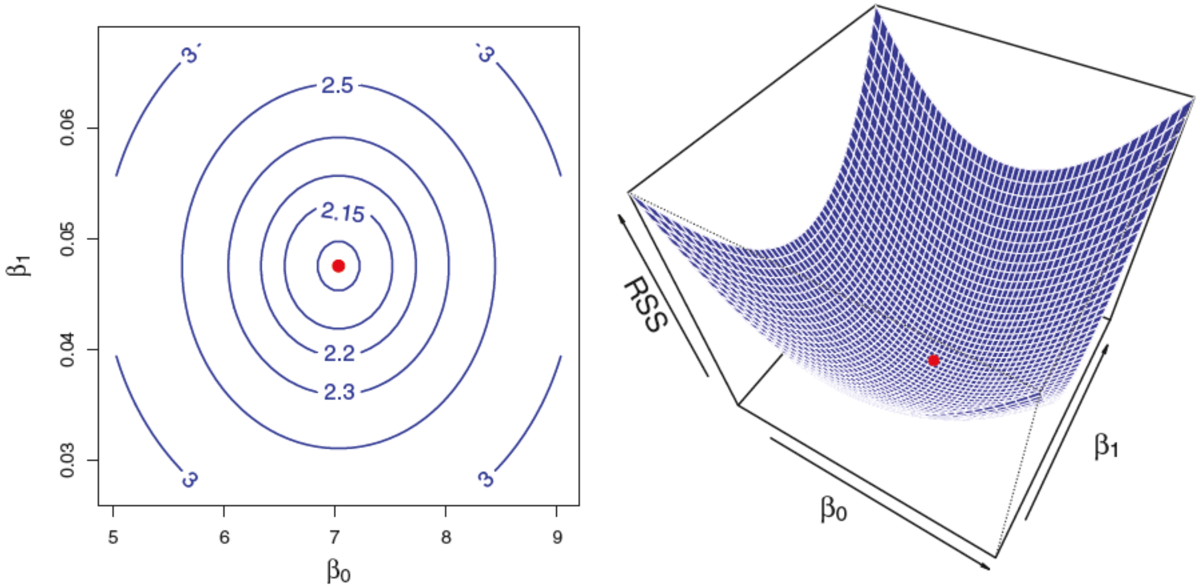
$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Find $\hat{\beta}_0, \hat{\beta}_1$ to minimize RSS

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

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

Results of the Advertising Example



- **Coefficient estimates:** $\hat{\beta}_0 = 7.03, \hat{\beta}_1 = 0.0475$
- **Fitted model:** $\hat{y}_i = 7.03 + 0.0475x_i$

$$sales \approx 7.03 + 0.0475 \times TV$$

- **Interpretation:** an additional $1000 spent on TV advertising is associated with selling approximately 47.5 additional units of the product.

2. Assessing Accuracy of Coefficient Estimates

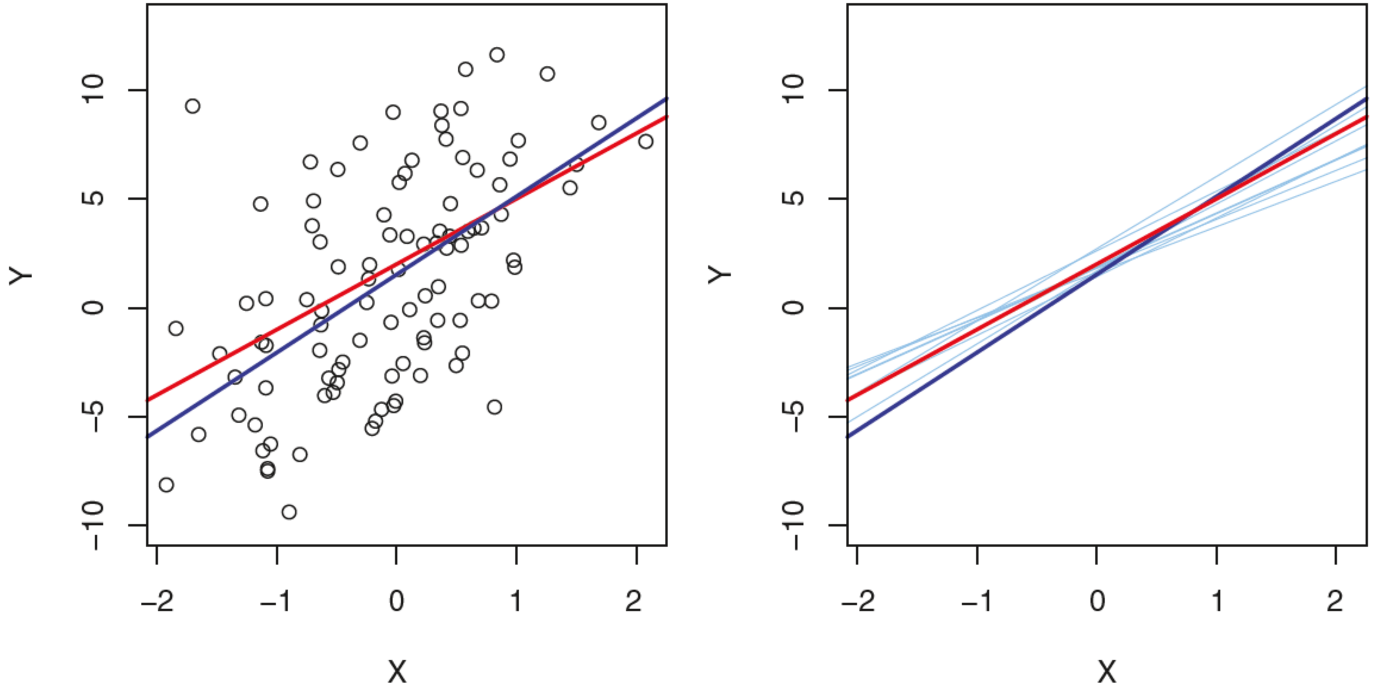
- Population regression line vs. least squares line

Population regression line: Best linear approximation to the true relationship

Model: $Y = 2 + 3X + \varepsilon$

Red: population regression line

Blue: least squares line



Standard Error of Coefficient Estimate

- Standard error (SE) of $\hat{\beta}_0$ and $\hat{\beta}_1$ tells us the average amount that the estimate of coefficient differs from true value.

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

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

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

Note: The estimates from data are $\widehat{\mathrm{SE}}(\hat{\beta}_0)$, $\widehat{\mathrm{SE}}(\hat{\beta}_1)$.

$$\hat{\beta}_0 \sim N(\beta_0, \mathrm{SE}(\hat{\beta}_0)^2), \; \hat{\beta}_1 \sim N(\beta_1, \mathrm{SE}(\hat{\beta}_1)^2)$$

Hypothesis Testing on Coefficients

- Hypothesis tests on the (true) coefficients

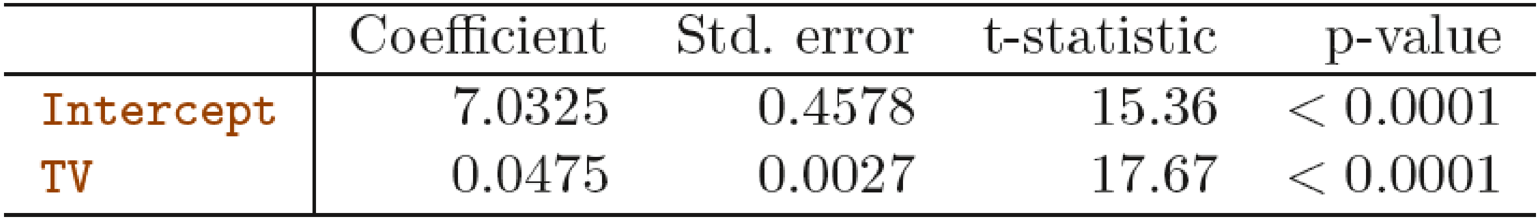
$$H_0: \beta_1 = 0 \text{ (There is no relationship between } X \text{ and } Y\text{)}$$

$$vs.$$

$$H_1: \beta_1 \neq 0 \text{ (There is some relationship between } X \text{ and } Y\text{)}$$

$$t\ test\colon\ t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \sim t(n-2)\ under\ H_0$$

Results of the Advertising Example



	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

- The estimate of β_1 is greater than 0, meaning that there is a positive relationship between TV advertising and sales.
- The standard errors are small relative to their associated coefficient estimates, meaning that the estimates are accurate.
- The p-values are very small, so we conclude that $\beta_0 \neq 0, \beta_1 \neq 0$. This indicates that there is a relationship between TV advertising and sales.

3. Assessing Model Fitting (Training)

- The quality of a linear regression fit (in training) is assessed by two quantities:

Residual standard error (RSE)

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

- An estimate of the standard deviation of ε
- A measure of *lack of fit* of the model to the data
- Not convenient to use since it is measured in the units of Y

Assessing Model Fitting

R^2 statistic

Measures the proportion of variance explained by the model

$$R^2 = 1 - \frac{RSS(residual\ sum\ of\ squares)}{TSS\ (total\ sum\ of\ squares)}$$

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

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

Results of the Advertising Example

- $RSE = 3.26$: this means that even if the model were correct and the true values of the unknown coefficients β_0 and β_1 were known exactly, any prediction of sales based on TV advertising would still be off by about 3260 units on average.

- $R^2 = 0.612$: about 61.2% variability in the response sales is explained by a linear regression on TV.

4. Intervals

- General formula of confidence interval

$$\hat{\theta} \pm C \times \mathrm{SE}(\hat{\theta})$$

- Confidence interval for coefficients β_0, β_1
- Confidence interval for the average response

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

- Prediction interval for the response

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Prediction interval is wider than confidence interval!

Multiple Linear Regression

- In practice there are often more than one predictors.
- Predicting the response based on multiple predictors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Model parameters (coefficients)

β_0 ---- intercept (i.e., the average value of Y if all inputs are zero)

β_j ---- slope for the *j*th predictor (the average increase in Y when X_j is increased by 1 and **all other predictors are held constant**)

1. Estimating Coefficients

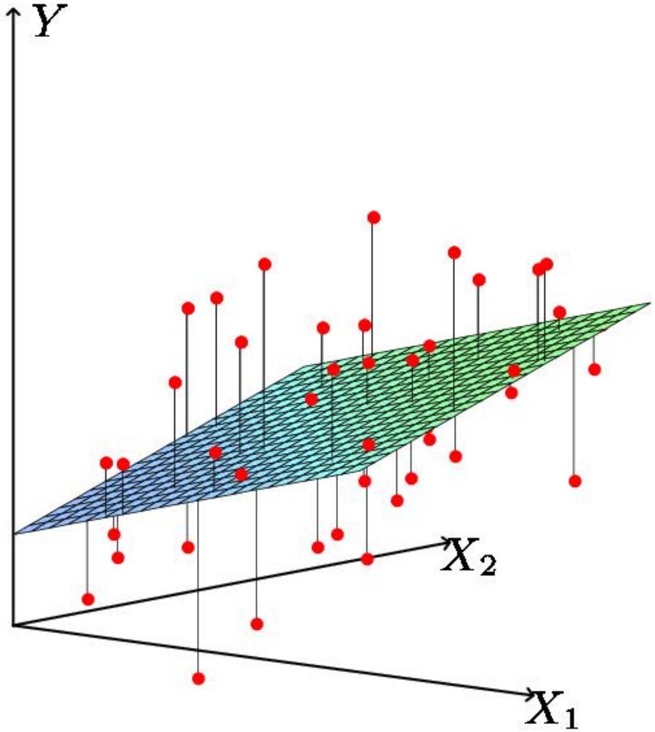
- Given a training data set

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$$
$$\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$$

- Use Least Squares (LS) method to find coefficient estimates

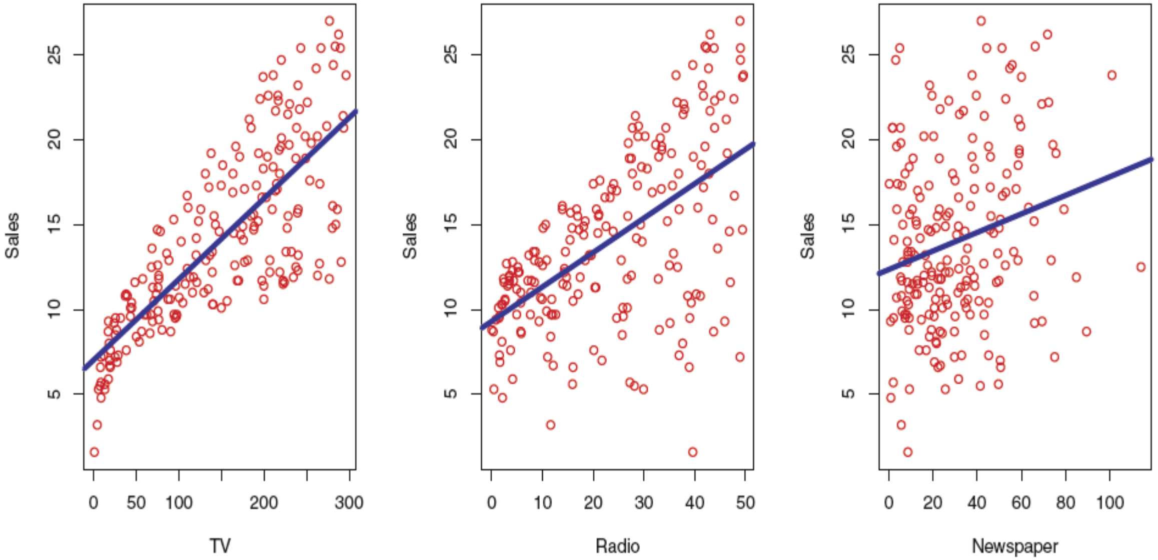
$$\text{minimize } RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_p x_p\right)^2$$



Advertising Example

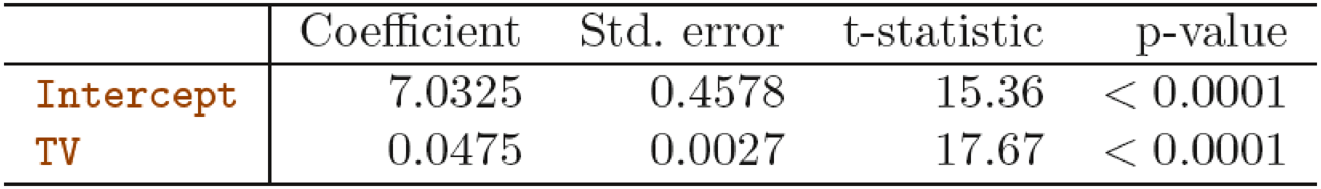
- Example: Advertising data set, $X_1 = TV, X_2 = radio, X_3 = newspaper, Y = sales, n = 200$



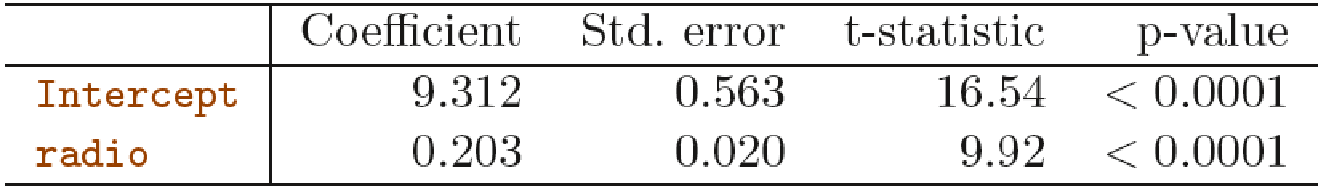
	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

2. Simple vs. Multiple Linear Regression

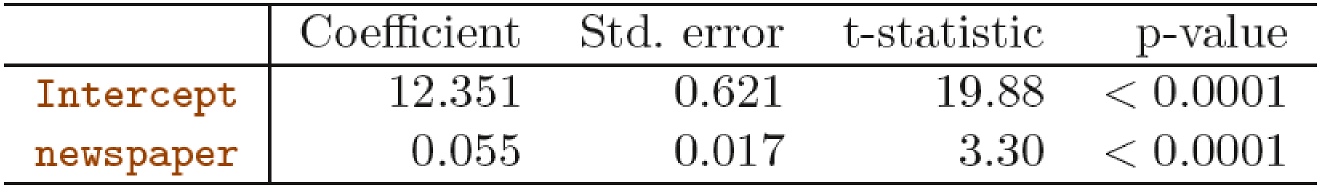
- Three simple regression models vs. multiple regression model



	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



	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	< 0.0001

- Based on the simple regression, newspaper has an effect on sales; however, based on the multiple regression, it does not.
- Correlation(newspaper, radio) = 0.35. newspaper gets “credit” from the effect of radio on sales.

3. Does Relationship Exist?

- Is there a relationship between the response and predictors?

  In the multiple regression with p predictors, we need to decide whether all of the coefficients are zero.

- F test on all coefficients

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

$$vs.$$

$$H_1: at\ least\ one\ \beta_j \neq 0$$

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

- F-statistic indicates the overall effect of predictors. If it is significant, it means at least one of the predictors is associated with the response.

4. Deciding on Important Variables

- Checking individual P values?
- Variable selection
  - Compare all possible models
  - Model search
  - Shrinkage

5. Assessing Model Fitting

- How to assess model fit for multiple regression?
- R^2: always increase when more variables are added to the model, even if those variables are only weakly associated with the response.

Advertising Example

Predictors in Model	R^2
TV	0.61
$TV, radio$	0.89719
$TV, radio, newspaper$	0.8972

Assessing Model Fitting

- Adjusted R^2: can decrease when a variable is added if it is weakly or not associated with the response at all.

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

$$R_a^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

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

R Output

```
lm.fit=lm(medv~lstat+age,data=Boston)
summary(lm.fit)
Call:
lm(formula = medv ~ lstat + age, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.981  -3.978  -1.283   1.968  23.158

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
age          0.03454    0.01223   2.826  0.00491 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.173 on 503 degrees of freedom
Multiple R-squared:  0.5513,    Adjusted R-squared:  0.5495
F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```

Other Considerations in Regression

- Qualitative predictors
- Interaction terms
- Nonlinear relationships

Qualitative Predictors

- Non-measurable predictor (often called a factor)
- Example: Credit data set

Response

balance: average credit card debt

Predictors

age

cards: number of credit cards

education: years of education

income: in thousands of dollars

limit: credit limit

rating: credit rating

gender: male/female

student: student status

status: marital status

Ethnicity: Caucasian/African American/Asian

Predictors With Two Levels

- Code the predictor with two levels as an indicator or dummy variable that takes two possible values
- For example, to predict balance (y) on gender (x)

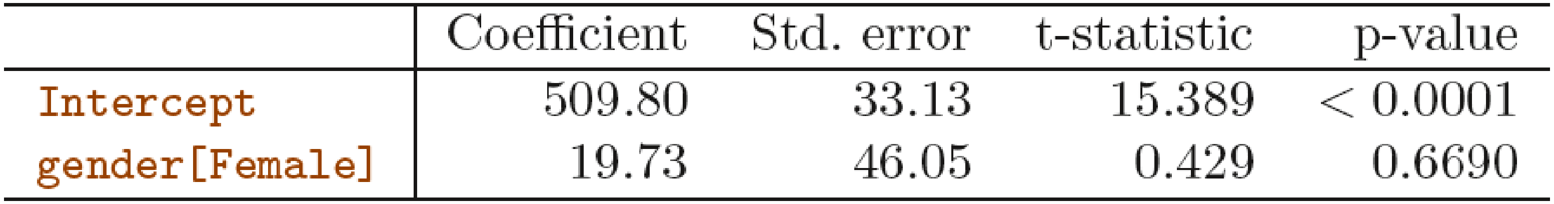
$$x_i = \begin{cases} 0 & if\ ith\ person\ is\ male \\ 1 & if\ ith\ person\ is\ female \end{cases}$$

- Regression model

$$y_i \approx \beta_0 + \beta_1 x_i = \begin{cases} \beta_0 & male \\ \beta_0 + \beta_1 & female \end{cases}$$

- β_0 represents the average credit card balance among males, $\beta_0 + \beta_1$ is the average balance among females, and β_1 is the average difference in balance between females and males.
- Males is coded as the “baseline”.

Interpretation



	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

- The average credit card debt for males is estimated to be $509.80.
- The average debt for females is $509.80+$19.73=$529.53.
- Females carry $19.73 in additional debt.

Other Coding Schemes

- There are other ways to code qualitative variables.
- An alternative way to code gender

$$x_i = \begin{cases} -1 & if\ ith\ person\ is\ male \\ 1 & if\ ith\ person\ is\ female \end{cases}$$

- Regression model

$$y_i \approx \beta_0 + \beta_1 x_i = \begin{cases} \beta_0 - \beta_1 & male \\ \beta_0 + \beta_1 & female \end{cases}$$

- β_0 represents the overall average credit card balance, and β_1 is the amount that females are above the average and males are below the average.

Predictors With More Than Two Levels

- When the qualitative predictor has more than two levels, it can be coded using multiple dummy variables.
- For example, to code ethnicity (Caucasian/African American/Asian)

$$x_{i1} = \begin{cases} 1 & if\ ith\ person\ is\ Asian \\ 0 & if\ ith\ person\ is\ not\ Asian \end{cases}$$

$$x_{i2} = \begin{cases} 1 & if\ ith\ person\ is\ Caucasian \\ 0 & if\ ith\ person\ is\ not\ Caucasian \end{cases}$$

- Regression model

$$y_i \approx \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} = \begin{cases} \beta_0 + \beta_1 & Asian \\ \beta_0 + \beta_2 & Caucasian \\ \beta_0 & African\ American \end{cases}$$

Interaction

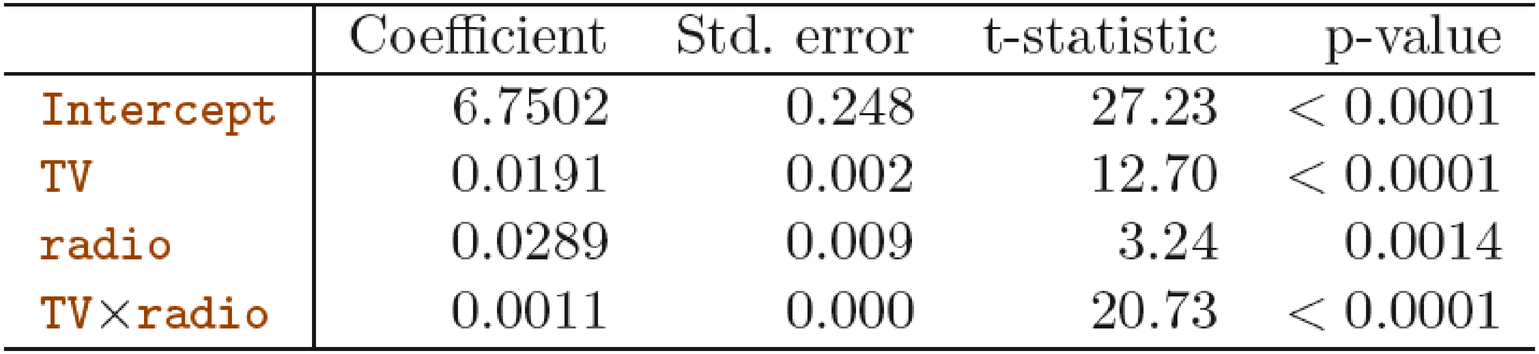
- When the effect on Y of increasing X_1 depends on another predictor (X_2)

$$Y = 2 + 3X_1 + 4X_2 + \varepsilon$$

$$Y = 2 + 3X_1 + 4X_2 + 2X_1X_2 + \varepsilon$$

- Advertising example
  - TV and radio advertising both increase sales.
  - Perhaps spending money on both of them may increase sales more than spending the entire amount on one alone? (*synergy* effect in marketing)

Interpretation in Advertising Example



	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

$$Sales \approx \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times TV \times Radio$$

- Spending $1000 extra on TV advertising will increase average sales by $19 + 1.1 \times radio$

$$Sales \approx \beta_0 + (\beta_1 + \beta_3 \times Radio) \times TV + \beta_2 \times Radio$$

- Spending $1000 extra on radio advertising will increase average sales by $28.9 + 1.1 \times TV$

$$Sales \approx \beta_0 + \beta_1 \times TV + (\beta_2 + \beta_3 \times TV) \times Radio$$

Interaction with Qualitative Predictors

- Credit data set: predicting balance using the income (quantitative) and student (qualitative) variables

$$student_i = \begin{cases} 1 & if\ ith\ person\ is\ a\ student \\ 0 & if\ ith\ person\ is\ not\ a\ student \end{cases}$$

- No interaction term

$$balance_i \approx \beta_0 + \beta_1 \times income_i + \beta_2 \times student_i$$

$$= \beta_1 \times income_i + \begin{cases} \beta_0 + \beta_2 & student \\ \beta_0 & nonstudent \end{cases}$$

***Interpretation:** the effect of income on balance does not depend on whether or not the person is a student.*

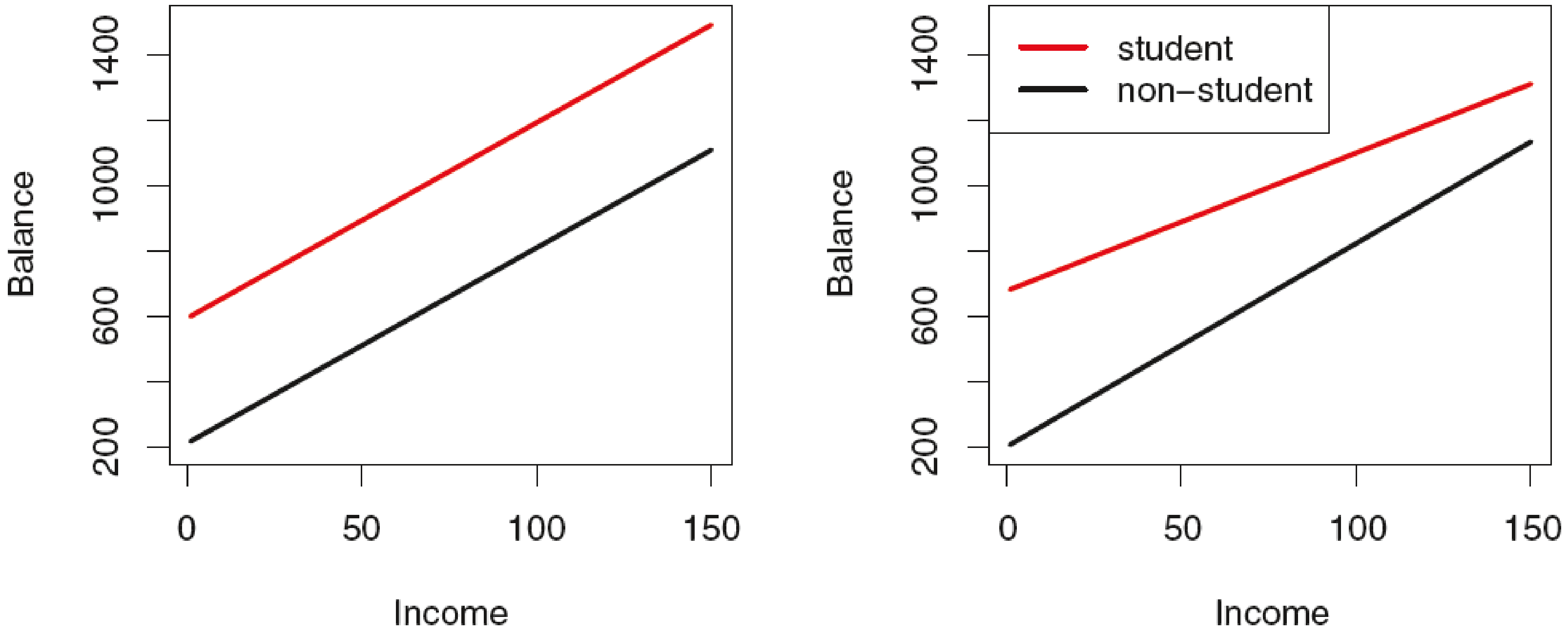
- Add interaction

$$balance_i \approx \beta_0 + \beta_1 \times income_i + \beta_2 \times student_i + \beta_3 \times student_i \times income_i$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times income_i & student \\ \beta_0 + \beta_1 \times income_i & nonstudent \end{cases}$$

***Interpretation:** the effect of income on balance is different for students and non-students.*

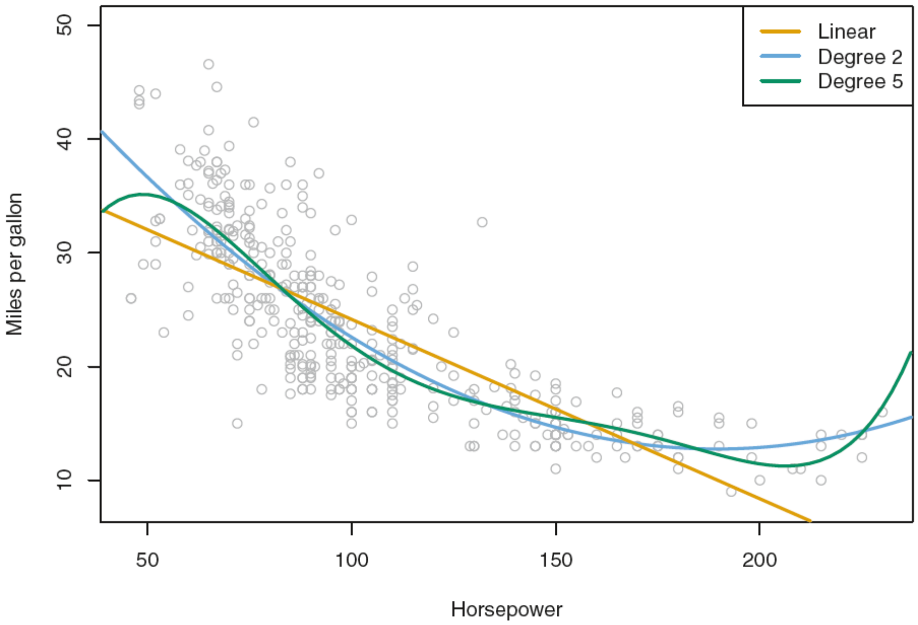
Regression Lines



- Without interaction term: regression lines are parallel (differ only in intercept)
- With interaction term: regression lines are not parallel (differ in both slope and intercept)

Nonlinear Relationships

- Extend linear model to accommodate nonlinear relationships
- Example: Auto data set, Y: mpg (gas mileage) X: horsepower



Linear: $mpg \approx \beta_0 + \beta_1 \times horsepower$

Degree 2: $mpg \approx \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2$

Degree 5: $mpg \approx \beta_0 + \beta_1 \times horsepower + \cdots + \beta_5 \times horsepower^5$

Assumptions of Linear Regression Model

- Linear regression model

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

Training data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

Assumptions on the random error term ε

Zero mean: $E(\varepsilon_i) = 0$

Constant variance: $Var(\varepsilon_i) = \sigma^2$

Normality: $\varepsilon_i \sim N(0, \sigma^2)$

Model Diagnostics

- After fitting a regression model to a given dataset, we need to check the model adequacy (potential problems that may invalid the model fit. For example, assumptions are satisfied? Any problems with the data such as outliers?). This is also called **model diagnostics**.

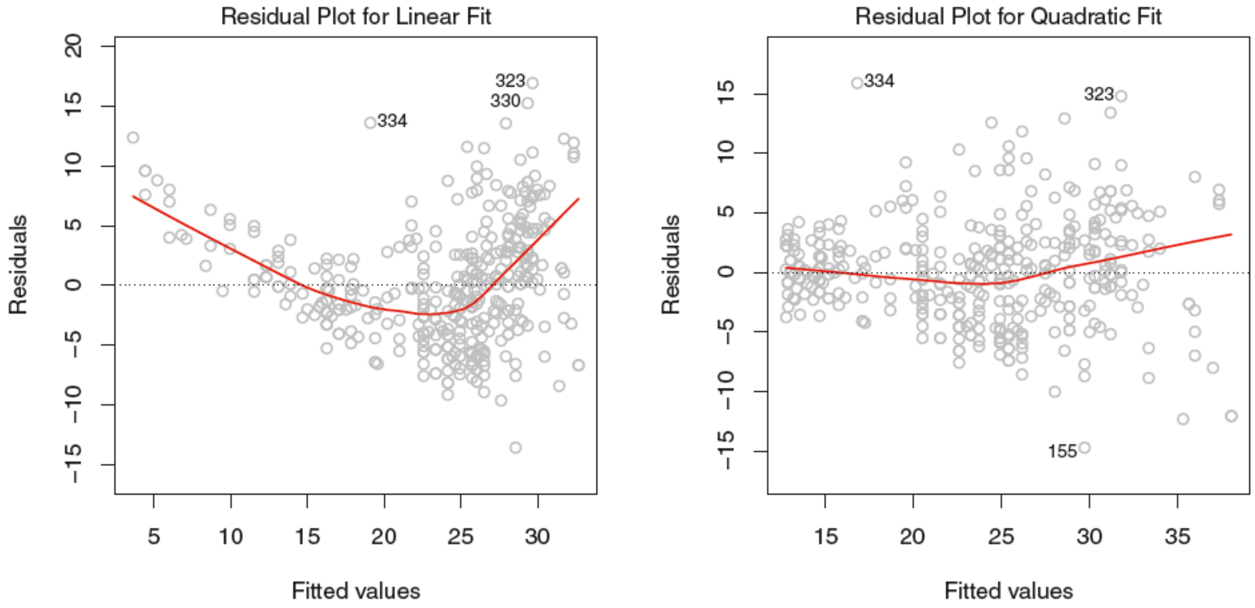
- **Five potential problems**

1. Non-linearity
2. Non-constant variance of error terms
3. Outliers
4. High-leverage points
5. Collinearity

1. Nonlinearity

- **Assumption**: linear regression model assumes a straight-line relationship between the response and predictors.
- **Detection**: check for *pattern* in residual plots ($e_i \; vs. \; \hat{y}_i$)

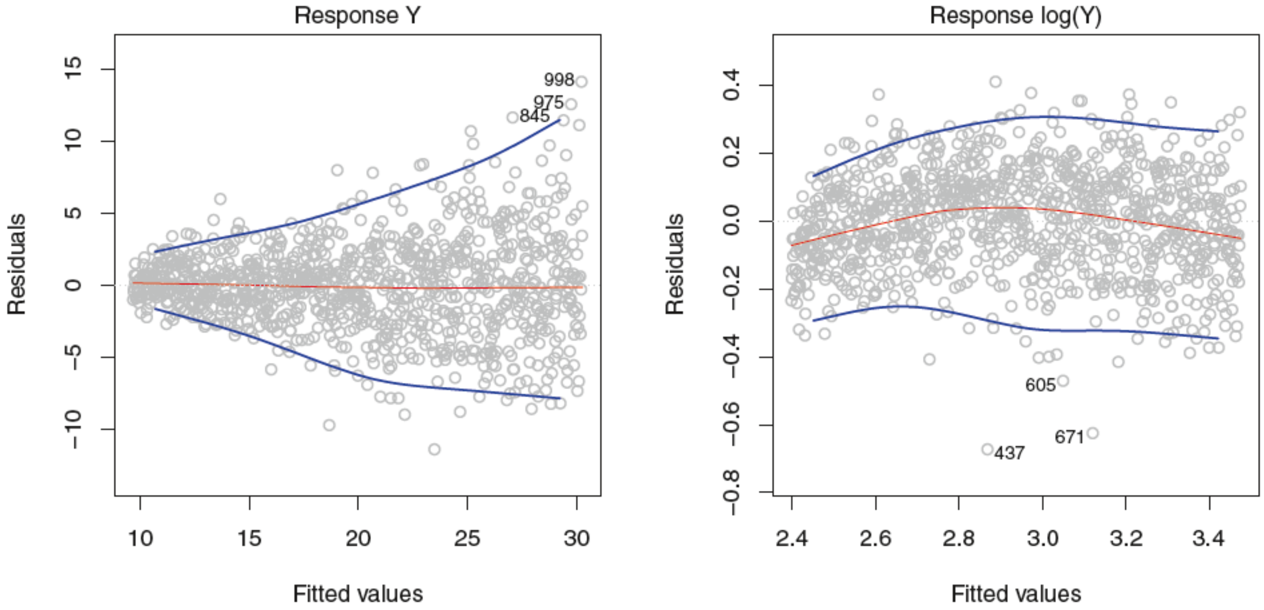
$mpg \approx \beta_0 + \beta_1 \times horsepower$ $\quad$ $mpg \approx \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2$



- **Solution**: use non-linear transformations of predictors (e.g., $x^2, \log(x), \sqrt{x}$)

2. Non-constant Variance of Error Terms

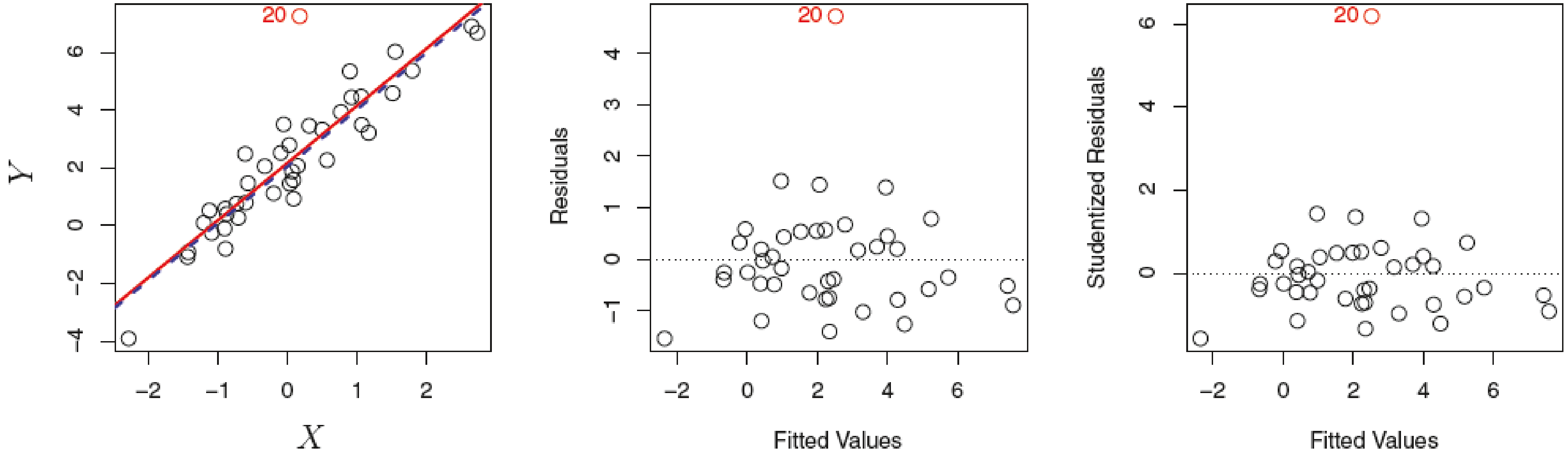
- **Assumption**: linear regression model assumes that the error terms have a constant variance, $Var(\varepsilon_i) = \sigma^2$.
- Non-constant variance of errors is called *heteroscedasticity*.
- **Detection**: check for *funnel shape* in the residual plot (e_i $vs.$ $\hat{y}_i$)



- **Solution**: transform the response (e.g., $\log(Y), \sqrt{Y}$) or use weighted least squares if information on variance of individual responses is available

3. Outliers

- **Outlier**: unusual y_i for given x_i
- **Detection**: check the plot of *studentized residuals* (dividing each residual by its estimated standard error) for potential outliers (absolute value>3)



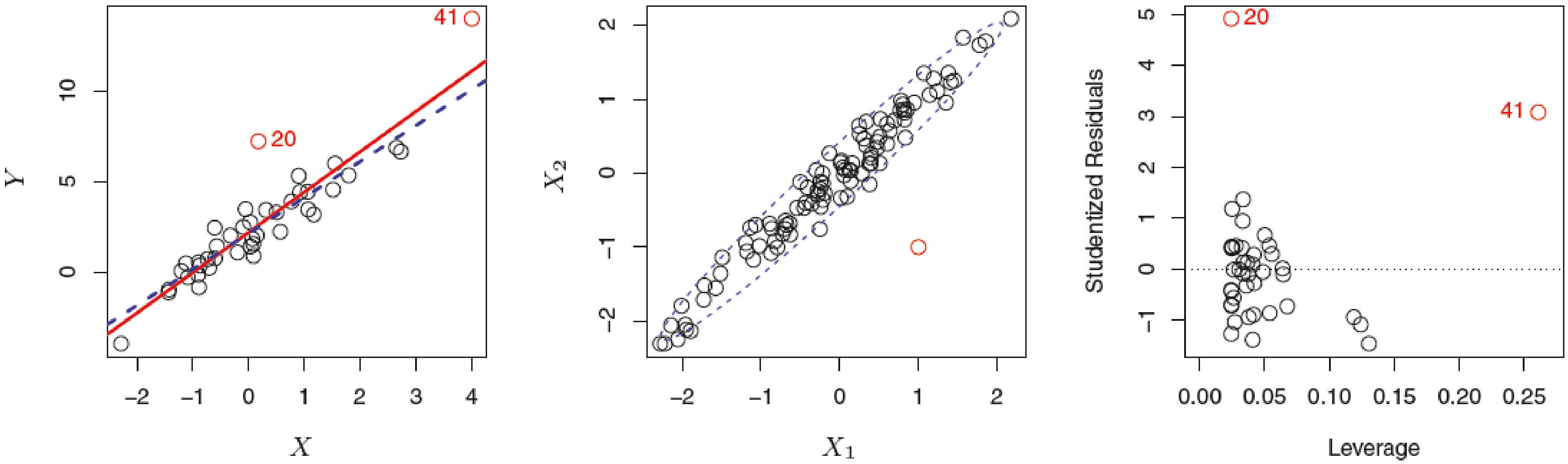
- **Solution**: if due to error in data collection or recording, correct/remove it. If valid, it cannot be removed.

4. High Leverage Points

- **High leverage points**: unusual value for x_i
- **Detection**: check the plot of leverages ($h_i \gg (p+1)/n$)

average leverage of all observations

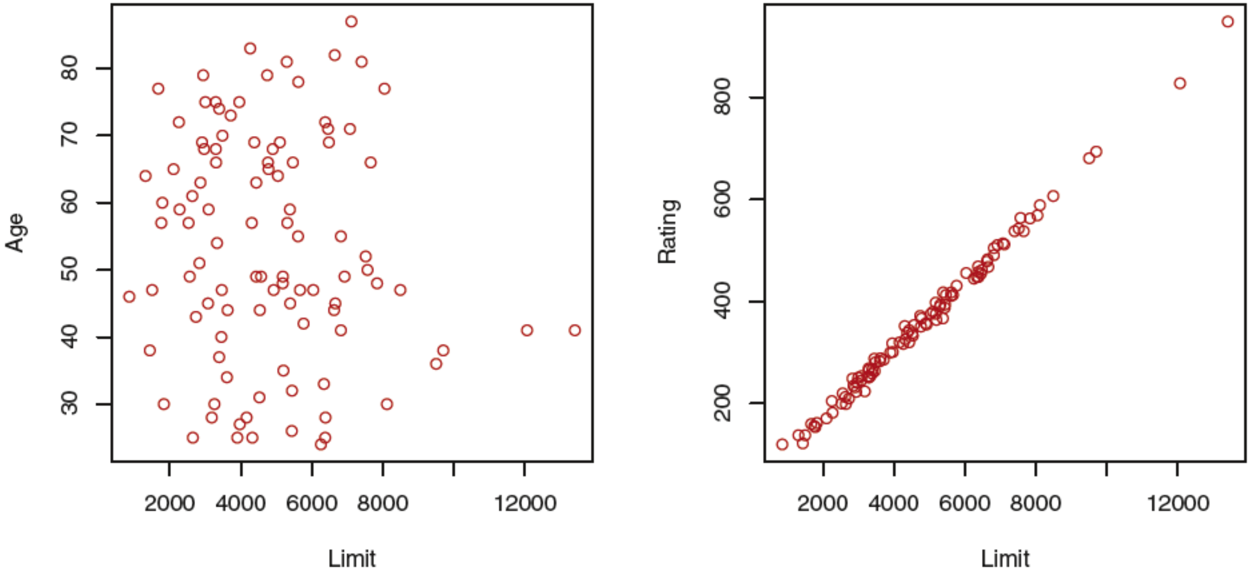
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$



- **Solution**: limit the values of x

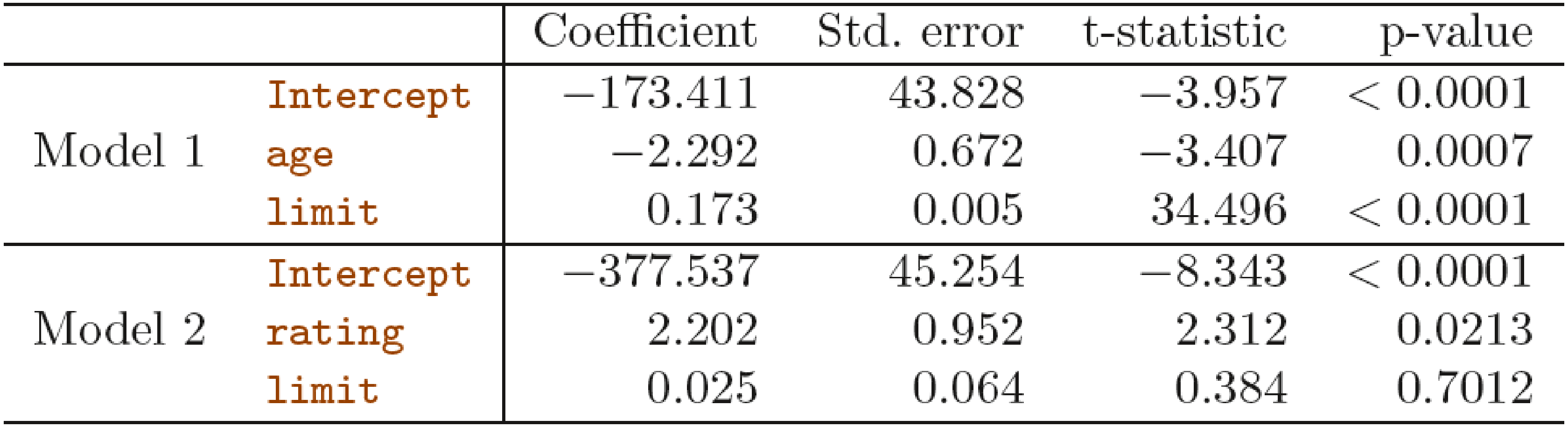
5. Collinearity

- **Collinearity**: two or more **predictors** are closely related to one another.
- It is difficult to determine how each one of the collinear variables separately affect the response.
- Example: Credit data set
  Response: balance
  Predictors: income, limit, age, rating, gender, etc.



Effect of Collinearity

- Increased standard errors in coefficient estimation
- Since the t test for each predictor is calculated by $\hat{\beta}_j / SE(\hat{\beta}_j)$, the importance of some predictors may be masked.



		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

Detect Collinearity and Solution

- **Variance inflation factor** (VIF): indicate how serious the collinearity is

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

R^2 from a regression of the *j*th predictor (X_j) onto all of the other predictors.

VIF=1, no collinearity

VIF>5, serious collinearity

- **Solution**: drop one of the problematic predictors from the regression model