Statistical Learning. Linear methods for regression

Jose Ameijeiras Alonso
Departamento de Estatística e Investigación Operativa (USC)

Máster Interuniversitario en Tecnologías de Análisis de Datos Masivos: Big Data

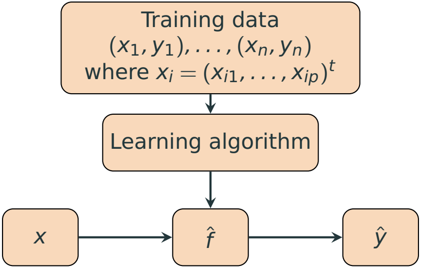
Introduction

- Example: Suppose that we observe a quantitative response Y and preditor variables $X = (X_1, \dots, X_p)$ and we assume that there is some relationship between Y and X. The relation can be written in general:

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

  - f is some fixed but unknown function
  - ϵ is a random error term and has mean zero
- How do we estimate f?
- We want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y)

Introduction



Training data
$(x_1, y_1), \ldots, (x_n, y_n)$
where $x_i = (x_{i1}, \ldots, x_{ip})^t$

Learning algorithm

x

$\hat{f}$

$\hat{y}$

- To evaluate the performance of a statistical learning method, we need some way to measure how well its predictions actually match the observed data.
- Consider the mean squared error (MSE) criterion for an arbitrary function f:

$$MSE(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

- We can also consider the residual sum of squares (RSS) criterion:

$$RSS(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Introduction

$$\boxed{MSE(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2}$$

- Note that $MSE(f)$ is computed using the training data (*training MSE*)
- Minimizing $MSE(f)$ leads to infinitely many solutions: any function $\hat{f}$ passing through the training points (x_i, y_i) is a solution
- However, we are interested in the accuracy of the predictions when we apply the method to previously unseen test data
- We must restrict the eligible solutions to a smaller set of functions
- Most statistical learning methods can be characterized as either parametric or non-parametric

Introduction

We observe a quantitative response Y and preditor variables $X = (X_1, \dots, X_p)$ and assume

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

Parametric Methods: we make an assumption about the functional form of f

- For example, one very simple assumption is that f is linear in X:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

- The problem of estimating f reduces to the problem of estimating a set of parameters (easier than to fit an entirely arbitrary function f)
- Disadvantage: it may happen that the model we choose does not match the true unknown form of f
- We can try to address this problem by choosing more flexible models
    - More flexible model requires estimating a greater number of parameters
    - More flexible models can lead to *overfitting*

Introduction

We observe a quantitative response Y and preditor variables $X = (X_1, \ldots, X_p)$ and assume

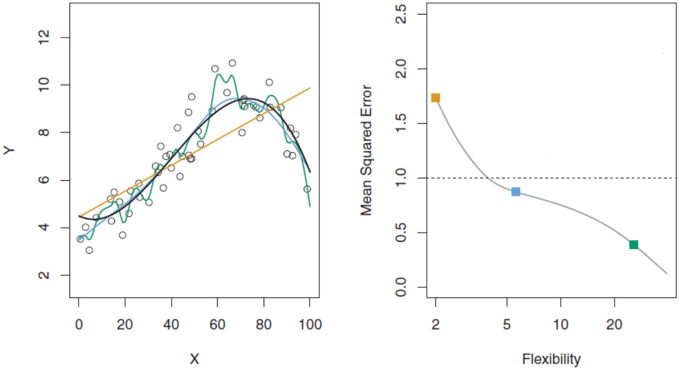
$$Y = f(X) + \epsilon$$

Non-parametric Methods: we do not make explicit assumptions about the functional form of f

- We seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly
- These methods allow great flexibility in the possible shape of f
- Disadvantage: a very large number of observations is required in order to obtain an accurate estimate for f
- Again, we must select a level of flexibility to avoid the *overfitting*

Introduction

$$\boxed{MSE(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}$$



In black,true line for f. In the right, in gray, training MSE for different fits with increasing flexibility.

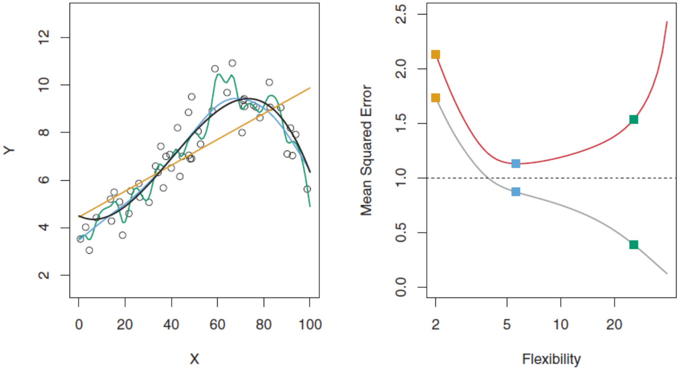
Image from "An Introduction to Statistical Learning with application in R"

Introduction

- Recall that we are interested in the accuracy of the predictions when we apply the method to previously unseen test data.
- As model flexibility increases, training MSE will decrease, but test MSE may not.

Introduction

$$MSE(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$



In black,true line for f. In the right, in gray, training MSE for different fits with increasing flexibility. In red, MSE in a test dataset.

Image from "An Introduction to Statistical Learning with application in R"

Introduction

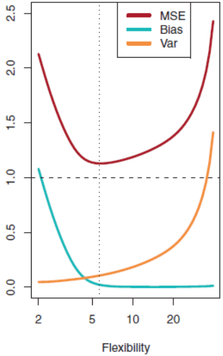
- Mathematically, we want to know whether $\hat{f}(x_0)$ is approximately equal to y_0, where (x_0, y_0) is a previously unseen test observation not used to train the statistical learning method.
- The expected test MSE, denoted by $\mathbb{E}((y_0 - \hat{f}(x_0))^2)$, is the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each at x_0
- It can be proved that it can be decomposed into the sum of three terms:

$$\mathbb{E}((y_0 - \hat{f}(x_0))^2) = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

Introduction

$$\mathbb{E}((y_0 - \hat{f}(x_0))^2) = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon)$$

- The first term $\text{Var}(\hat{f}(x_0))$ refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set
- The bias term $[\text{Bias}(\hat{f}(x_0))]^2$ refers to the error that is introduced by approximating a real-life problem by a much simpler model
- The term $\text{Var}(\epsilon)$ corresponds to the irreducible error



Linear methods for regression

- Suppose that we observe a quantitative response Y and preditor variables $X = (X_1, \ldots, X_p)$
- We write our model in general as

$$\boxed{Y = f(X) + \epsilon}$$

  where ϵ is a zero-mean error term that captures measurement errors and other discrepancies.
- The linear regression model assumes that

$$\boxed{f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}$$

Linear methods for regression

Example 1: Suppose that we are hired by a client to provide advice on how to improve sales of a particular product.

The Advertising data set consists of the sales of that product in different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper

Data set available at http://www-bcf.usc.edu/ gareth/ISL/data.html

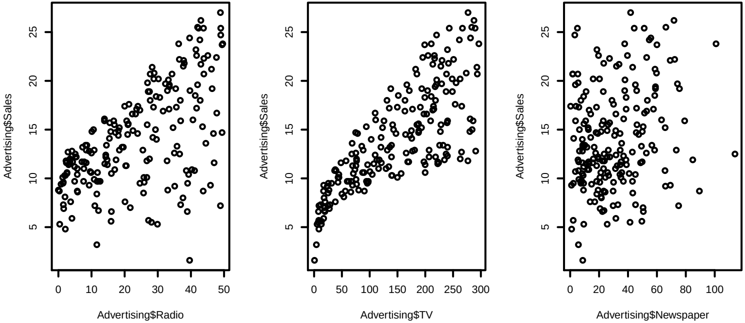
Images from http://www.vecteezy.com/

Linear methods for regression

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- Is the relationship linear?

Linear methods for regression

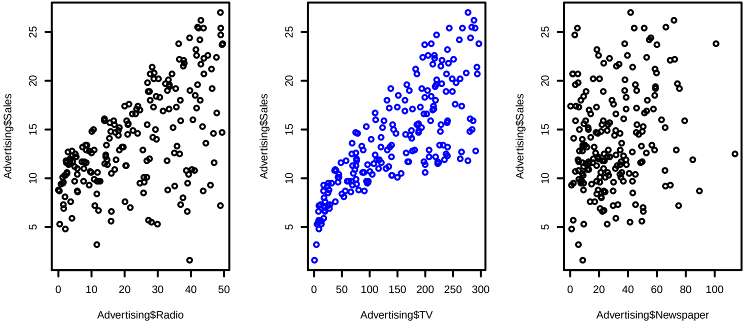
```
> Advertising <- read.csv("datasets/Advertising.csv")
> plot(Advertising$Radio, Advertising$Sales)
> plot(Advertising$TV, Advertising$Sales)
> plot(Advertising$Newspaper, Advertising$Sales)
```



Simple linear regression

- First, let us consider a very simple model. Let us assume that the response Y depends linearly on a single preditor variable X, that is

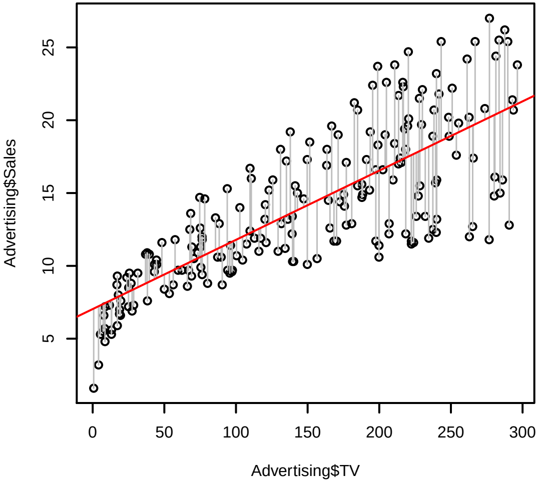
$$Y = \beta_0 + \beta_1 X + \epsilon$$



$$\boxed{\text{sales} \approx \beta_0 + \beta_1 \text{TV}}$$

Simple linear regression: estimation of the parameters

- We choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that $RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ is minimum
- The errors $\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ are called the residuals of the regression



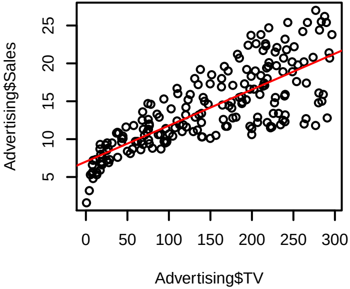
- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Simple linear regression: estimation of the parameters

```
> mod <- lm(Advertising$Sales ~ Advertising$TV)
> mod$coefficients
```

```
##    (Intercept) Advertising$TV
##     7.03259355     0.04753664
```

```
> plot(Advertising$TV, Advertising$Sales)
> abline(mod, col = 2)
```



- $\hat{\text{sales}} = 7.03 + 0.0475\text{TV}.$

Simple linear regression: properties of the estimators

- The model $Y = \beta_0 + \beta_1 X + \epsilon$ defines the population regression line (unobserved)
- The least squares line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ is the least squares estimate based on the observed data
    - The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased, that is, the mean of the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ are, respectively, β_0 and β_1
    - Regarding the standard error of $\hat{\beta}_0$ and $\hat{\beta}_1$, we have

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \quad \text{and} \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where $\sigma^2 = \mathrm{Var}(\epsilon)$

- In general σ is unknown and can be estimated by the residual standard error

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

Simple linear regression: confidence intervals for the parameters

- For linear regression, the 95% confidence interval for β_1 approximately takes the form
$$\left(\hat{\beta}_1 - 2\cdot \mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2\cdot \mathrm{SE}(\hat{\beta}_1)\right)$$
- For linear regression, the 95% confidence interval for β_0 approximately takes the form
$$\left(\hat{\beta}_0 - 2\cdot \mathrm{SE}(\hat{\beta}_0), \hat{\beta}_0 + 2\cdot \mathrm{SE}(\hat{\beta}_0)\right)$$

*To be precise, rather than the number 2, they should contain the 97.5% quantile of a t-distribution with $n-2$ degrees of freedom.

Simple linear regression: hypothesis tests

- In linear regression, the most common hypothesis test involves testing the null hypothesis of:

$$\boxed{H_0 : \text{There is no relationship between } X \text{ and } Y}$$

- Note that, since the model is $Y = \beta_0 + \beta_1 X + \epsilon$, this corresponds to testing

$$\boxed{\begin{array}{l} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{array}}$$

- So, we need to determine $\hat{\beta}_1$ is sufficiently far from zero that we can be confident that β_1 is non-zero
- In practice, we compute a t-statistic

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

  and determine the p-value (the probability of obtaining a more extreme statistic than we did if the null hypothesis were true). A small p-value indicates that it is unlikely to observe such value of the statistic if the null hypothesis is true

Simple linear regression

```
> mod <- lm(Advertising$Sales ~ Advertising$TV)
> summary(mod)

##
## Call:
## lm(formula = Advertising$Sales ~ Advertising$TV)
##
## Residuals:
##      Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.032594   0.457843   15.36   <2e-16 ***
## Advertising$TV 0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

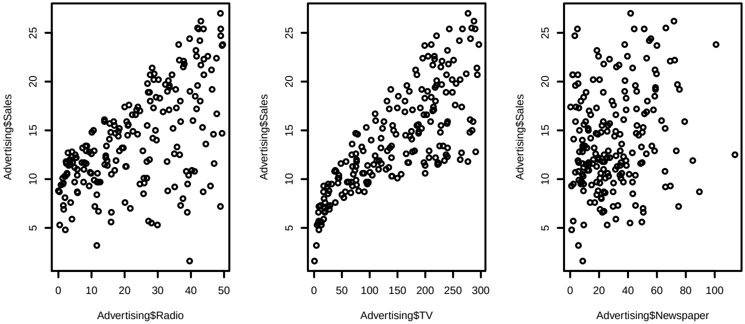
Simple linear regression: the R^2 statistic

- The RSE provides an absolute measure of lack of fit of the model, but it depends on the units of Y.
- As an alternative, the R^2 statistic measures the proportion of variability in Y that can be explained using X
  - It always takes on a value between 0 and 1
  - When it is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression
  - A number near 0 indicates that the regression did not explain much of the variability in the response

Multiple linear regression

- Now, let us consider a more general model. Let us assume that the response Y depends linearly on several preditor variables $X = (X_1, \ldots, X_p)$, that is

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



$$\boxed{\text{sales} \approx \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper}}$$

Multiple linear regression: estimation of the parameters

- Note that now, our training data is:

$$\begin{array}{cccc} y_1 & x_{11} & \cdots & x_{1p} \\ y_2 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{np} \end{array}$$

- We choose $\hat{\beta}_0$, $\hat{\beta}_1, \ldots, \hat{\beta}_p$ such that:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2 \text{ is minimum}$$

- The errors $\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip}$ are called the residuals of the regression

Multiple linear regression: estimation of the parameters

- Note that

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^t (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$

- It can be proved that, the values $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^t$ that minimize RSS are

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t \mathbf{X}\right)^{-1} \mathbf{X}^t \mathbf{y}$$

Multiple linear regression: estimation of the parameters

```
> mod2 <- lm(Advertising$Sales ~ Advertising$TV + Advertising$Radio
+           + Advertising$Newspaper)
> mod2$coefficients
```

```
##           (Intercept)        Advertising$TV     Advertising$Radio
##           2.938889369           0.045764645           0.188530017
## Advertising$Newspaper
##          -0.001037493
```

Multiple linear regression: hypothesis tests

- In linear regression, the most common hypothesis test involves testing the null hypothesis of:

$$\boxed{\text{H_0 : There is no relationship between X and Y}}$$

- Note that, since the model is $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$, this corresponds now to testing

$$\boxed{\begin{array}{l} H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \\ H_1 : \text{At least one } \beta_i \neq 0, \ i = 1, \ldots, p. \end{array}}$$

- That is, we are trying to determine if at least one of the predictors $X_1, X_2, \ldots, X_p$ is useful in predicting the response
- This hypothesis test is performed by computing the F-statistic. A small p-value indicates that it is unlikely to observe such value of the statistic if the null hypothesis is true

Multiple linear regression

```
> summary(mod2)

##
## Call:
## lm(formula = Advertising$Sales ~ Advertising$TV + Advertising$Radio
##     Advertising$Newspaper)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.938889   0.311908   9.422   <2e-16 ***
## Advertising$TV         0.045765   0.001395  32.809   <2e-16 ***
## Advertising$Radio      0.188530   0.008611  21.893   <2e-16 ***
## Advertising$Newspaper -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Multiple linear regression: hypothesis tests

- The first step in a multiple regression analysis is to compute the F-statistic
- If we conclude on the basis of the p-value that at least one of the predictors is related to the response, then it is natural to wonder which are the variabes that are related to the response.
- We could look at the individual p-values corresponding to the individual tests for $i = 1, \ldots, p$:

$$\boxed{\begin{array}{l} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{array}}$$

- But. . . be careful especially if the number of predictors p is large

Multiple linear regression: the R^2 statistic

- The residual standard error RSE provides an absolute measure of lack of fit of the model, but it depends on the units of Y.
- In the multiple lineal model, the RSE is estimated as

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

- As an alternative, the R^2 statistic measures the proportion of variability in Y that can be explained using X
- The problem is that the R^2 statistic will always increase when more variables are added to the model, even if those variables are only weakly associated with the response
- The adjusted R^2 statistic is a modified version of the R^2 statistic that has been adjusted for the number of predictors in the model

Multiple linear regression: predictions

- We predict the response Y on the basis of a set of values for the predictors $(X_1, \ldots, X_p)$ as follows:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p$$

- We can also compute confidence and prediction intervals:
    - Confidence interval for the mean value of Y for a given value of X
    - Prediction interval for a single value of Y for a given value of X

Multiple linear regression: problems may occur. . .

- Non-linearity of the response-predictor relationships.
- Correlation of error terms.
- Non-constant variance of error terms.
- Outliers.
- High-leverage points.
- Collinearity.