Statistical Learning

Tommaso Proietti

DEF Tor Vergata

Linear Regression

Specification

Let Y be a univariate quantitative response variable. We model Y as follows:

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

where $f(X)$ is the systematic part (that can be predicted using the inputs X) and ε is a disturbance (error) term, accounting for the all the variation sources different from X.
Assumptions:

- Linearity. The regression function is linear in the inputs:

$$f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

- Weak exogeneity: $\mathrm{E}(\varepsilon|X) = 0$, i.e. X and ε are independent in the mean.
- Homoscedasticity: $\mathrm{E}(\varepsilon^2|X) = \mathrm{Var}(\varepsilon|X) = \sigma^2$.

The inputs set X includes quantititative variables, nonlinear transformations and basis expansions (to be defined later), as well as dummy variables coding the levels of qualitative inputs.
Under these assumptions the regression function is interpreted as the conditional mean of Y given X

$$f(X) = \mathrm{E}(Y|X).$$

This is the optimal predictor of Y under square loss (minimum mean square estimator of Y).

Estimation

Set of training data $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, N\}$, with $\boldsymbol{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$.
We assume that the training sample is a sample of size N, drawn independently (with replacement) from the underlying population, so that for the i-th individual we can write

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \ldots, N$$

Writing the N equations in matrix form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with $\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}$. This means that ϵ_i's have the same variance and are uncorrelated: the error for the i-th unit is not correlated to that of any other unit in the sample.

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & & & \vdots \\ 1 & x_{i1} & x_{i2} & \dots & x_{ik} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & & & & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nk} & \dots & x_{Np} \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \\ \vdots \\ \beta_p \end{bmatrix}$$

We aim at estimating the unknown parameters $\beta_0, \dots, \beta_p$, and σ^2.

Learning by OLS

An important estimation method is Ordinary Least Squares (OLS).
OLS obtains an estimate of β denoted $\hat{\beta}$, as the minimizers of the residual sum of squares

$$\text{RSS}(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta).$$

There is a closed form solution to the above problem.
Computing the partial derivatives of $\text{RSS}(\beta)$ w.r.t. each β_j and equating to zero yields a $p+1$ linear system of equations in $p+1$ unknowns,

$$\boldsymbol{X}'\boldsymbol{X}\beta = \boldsymbol{X}'\boldsymbol{y}$$

which is solved w.r.t. β. The solution is unique provided that the X's are not collinear.
In matrix notation the solution is

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Example: $p = 1$ (simple linear regression)

$$\begin{aligned} N\beta_0 + \beta_1 \sum_i x_i &= \sum_i y_i \\ \beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2 &= \sum_i x_i y_i \end{aligned}$$

We solve this system by substitution: solving wrt β_0 the 1st eqn. yields

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

Replacing into the 2nd and solving wrt $\hat{\beta}_1$ gives:

$$\hat{\beta}_1 = \frac{\frac{1}{N}\sum_i x_i y_i - \bar{x}\bar{y}}{\frac{1}{N}\sum x_i^2 - \bar{x}^2} = \frac{\text{Cov}(x,y)}{\text{Var}(x)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

Predicted values and residuals

Predicted (fitted) values

$$\hat{y}_i = \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_1 x_{ip}, \quad i = 1, \ldots, N.$$

In matrix notation,

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the "hat" matrix, projecting $\boldsymbol{y}$ orthogonally onto the vector space generated by $\boldsymbol{X}$.

Residuals

The OLS residuals are $e_i = y_i - \hat{y}_i$, $i = 1, \ldots, N$. In matrix notation,

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y},$$

where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{H}$.

Properties of $\hat{y}$ and e

The fitted values and the LS residuals have properties that are simple to show and are good to know.

- Orthogonality. $\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}$. The residuals are uncorrelated with each of the variable in $\boldsymbol{X}$. In particular, $\sum_i e_i = 0$ (if the model includes the intercept, the residuals have zero mean) and $\sum_i x_{ik} e_i = 0, k = 1, \ldots, p$.
- $\hat{\boldsymbol{y}}'\boldsymbol{e} = \sum_i \hat{y}_i e_i = 0$
- $\text{RSS}(\hat{\boldsymbol{\beta}}) = \boldsymbol{e}'\boldsymbol{e} = \sum_i e_i^2$ (henceforth RSS)
- From $\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e}$ and the previous properties, $||\boldsymbol{y}||^2 = ||\hat{\boldsymbol{y}}||^2 + ||\boldsymbol{e}||^2$
- $\bar{y} = N^{-1}\boldsymbol{i}'\boldsymbol{y} = \bar{\hat{y}}$ (average of predicted values)

Estimation of σ^2

The OLS residual for the i-th observation is $e_i = y_i - \hat{y}_i$.
For estimating σ^2 (the variance of the error term) we could use the variance of the residuals corrected for the number of degrees of freedom:

$$\hat{\sigma}^2 = \frac{\sum_i e_i^2}{N - p - 1}$$

$N - p - 1$ is known as the number of degrees of freedom.
$\hat{\sigma}$ is the *standard error of regression* (SER).

Goodness of fit

Define
$TSS = \sum_i (y_i - \bar{y})^2$ (Total sum of squares - Deviance of y_i)
$ESS = \sum_i (\hat{y}_i - \bar{y})^2$ (Explained sum of squares - Deviance of $\hat{y}_i$)
$RSS = \sum_i e_i^2$ (Residual sum of squares - Deviance of e_i)
It is easy to prove that

$$TSS = ESS + RSS$$

A relative measure of g.o.f. is

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

This measures suffers from a serious drawback when used for model selection: the inclusion of (possibly irrelevant) additional regressors always produces an increase in R^2.

Adjusted R^2

$$\bar{R}^2 = 1 - \frac{RSS/(N-p-1)}{TSS/(N-1)} = 1 - \frac{N-1}{N-p-1}(1-R^2)$$

Properties of the OLS estimators

We are going to look at the statistical properties of the OLS estimators (and other related quantities, such as the fitted values and the residuals), hypothesizing that we can draw repeated training samples from the same population, keeping the X's fixed and drawing different Y's. A distribution of outcomes will arise.
Under the stated assumptions,

$$\mathrm{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{\beta}, \qquad \mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Also, $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$.
Moreover, the OLS estimators are Best Linear Unbiased (*Gauss Markov theorem*), also unconditionally: $\mathrm{E}(\hat{\beta}_j) = \beta_j, \forall j$ and $\mathrm{Var}(\hat{\beta}_j) = \mathrm{MSE}(\hat{\beta}_j)$ is the smallest among the class of all linear unbiased estimators.
Example: $p = 1$ (simple regression model)

$$\mathrm{Var}(\hat{\beta}_0|\boldsymbol{X}) = \sigma^2\left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right], \qquad \mathrm{Var}(\hat{\beta}_1|\boldsymbol{X}) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2},$$

If we further assume

$$\varepsilon_i|\boldsymbol{x}_i \sim \text{NID}(0, \sigma^2),$$

- $y_i|\boldsymbol{x}_i \sim \text{N}(\beta_0 + \beta_1 x_i, \sigma^2)$
- The OLS estimators have a normal distribution: $\hat{\boldsymbol{\beta}}|\boldsymbol{X} \sim \text{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$.
- $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator of the coefficients (we will say more about it).
- Letting $s.e.(\hat{\beta}_j)$ denote the estimated standard error of the OLS estimator, i.e. the j-th element of $\hat{\text{Var}}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$,

$$t_j = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim t_{N-p-1},$$

a Student's t r.v. with $N - p - 1$ d.o.f.
This result is used to test hypotheses on a single coefficient and to produce interval estimates.

- Suppose we wish to test for the significance of subset of the coefficients. The relevant statistic for the null that J coefficients are all equal to zero is

$$F = \frac{(RSS_0 - RSS)/J}{RSS/(N-p-1)} \sim F_{J,N-p-1}$$

where RSS_0 is the residual sum of squares of the restricted model (i.e. it has J less explanatory variables), and $F_{J,N-K}$ is Fisher's distribution with J and $N-p-1$ d.o.f.
As a special case, the test for $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (all the regression coefficients are zero, except for the intercept) is:

$$F = \frac{ESS/p}{RSS/(N-p-1)} = \frac{R^2/p}{(1-R^2)/(N-p-1)}$$

Example: clothing data. Regression of total sales (logs) on hours worked and store size (logs).

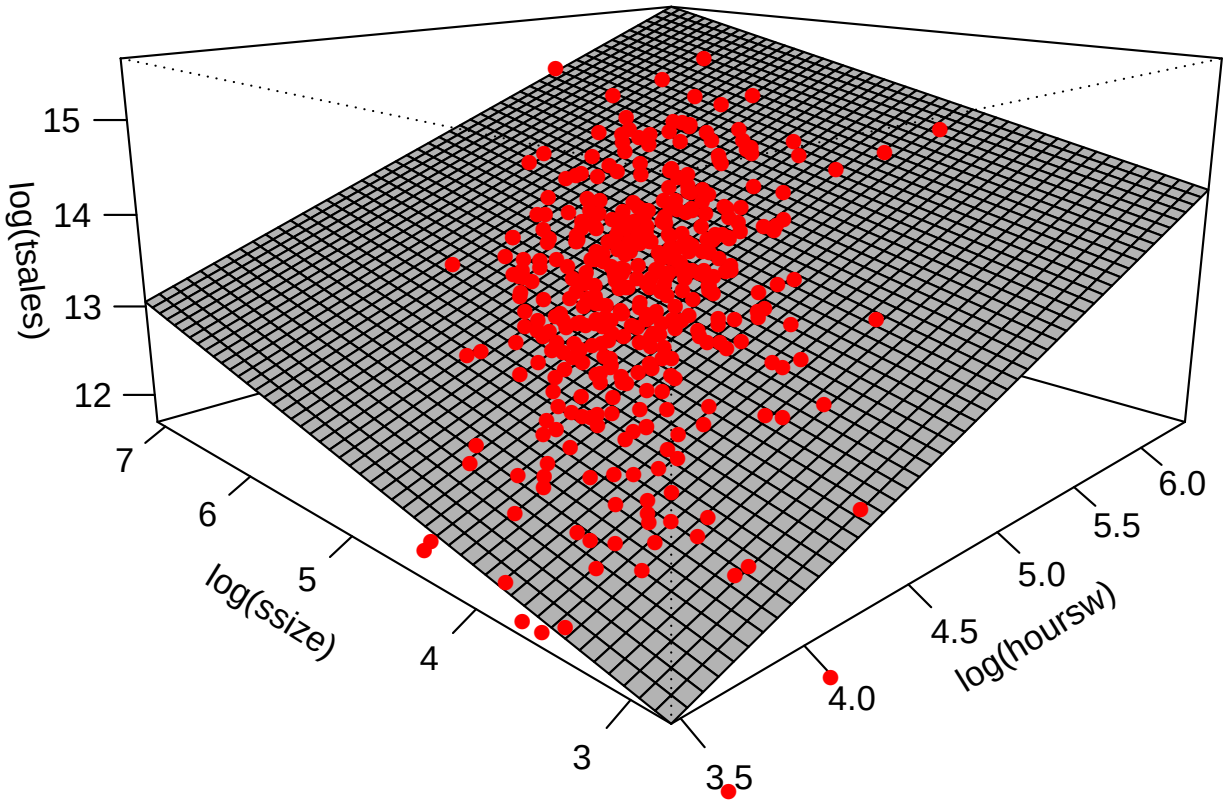
```
Call:
lm(formula = y ~ x1 + x2, data = clothing)

Residuals:
     Min       1Q   Median       3Q      Max
-1.89464 -0.21835  0.01462  0.29581  1.63896

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.73152    0.21850  35.385  < 2e-16 ***
x1           0.88949    0.05791  15.359  < 2e-16 ***
x2           0.31488    0.04308   7.309 1.49e-12 ***
---


Residual standard error: 0.4311 on 397 degrees of freedom
Multiple R-squared: 0.6342,Adjusted R-squared: 0.6324
F-statistic: 344.2 on 2 and 397 DF,  p-value: < 2.2e-16
```

Figure: Fitted values

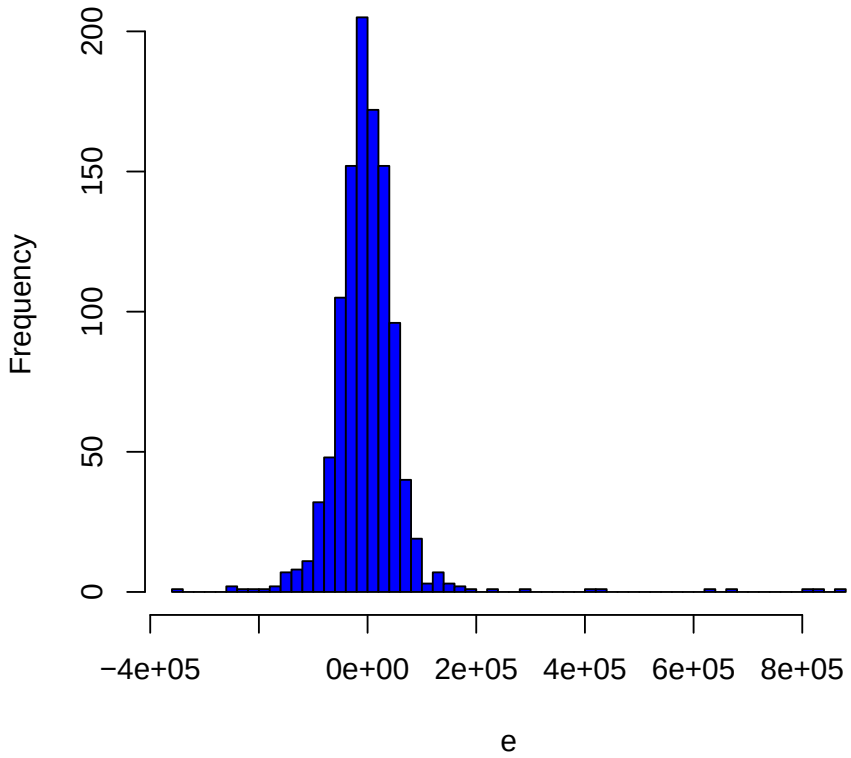


Example: house price regression in R

```
lm(formula = price ~ sqft + Age + Pool + Bedrooms + Pool + Fireplace
    Waterfront + DOM)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4336.851  12084.856   0.359    0.720
sqft            97.862      3.485  28.082  < 2e-16 ***
Age           -694.671    141.292  -4.917 1.02e-06 ***
Pool           815.028   8926.466   0.091    0.927
Bedrooms    -20923.878   4567.168  -4.581 5.16e-06 ***
Fireplace      -97.130   5152.184  -0.019    0.985
Waterfront   63376.186   9389.874   6.749 2.43e-11 ***
DOM            -20.988     24.841  -0.845    0.398
---

Residual standard error: 76390 on 1072 degrees of freedom
Multiple R-squared: 0.6162,Adjusted R-squared: 0.6137
F-statistic: 245.9 on 7 and 1072 DF,  p-value: < 2.2e-16
```

Figure: Histogram of residuals



Properties of the OLS residuals

- Does e_i reflect the properties of ϵ_i, and how?
- The i-th residual e_i has zero expectation (under the model's assumptions) and

$$\text{Var}(e_i|\boldsymbol{X}) = \sigma^2(1 - h_i),$$

where

$$h_i = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$$

is the **leverage** of observation i (the i-th diagonal element of $\boldsymbol{H}$).
- The leverage, h_i, is an important diagnostic quantity. It is a measure of remoteness of the inputs for the i-th individual in the space of the inputs.
- The barplot of h_i versus i illustrates the influence of the i-th observation on the fit.
- It is possible to show that $\frac{1}{N} \leq h_i \leq 1$, $\sum_i h_i = p + 1$ and that the average of the h_is is $(p + 1)/N$.
- The standardized residual is defined as

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}},$$

and is used in regression diagnostics (normal probability plot, check for homoscedasticity).

Figure: Anscombe Quartet: 4 datasets giving the same set fitted values

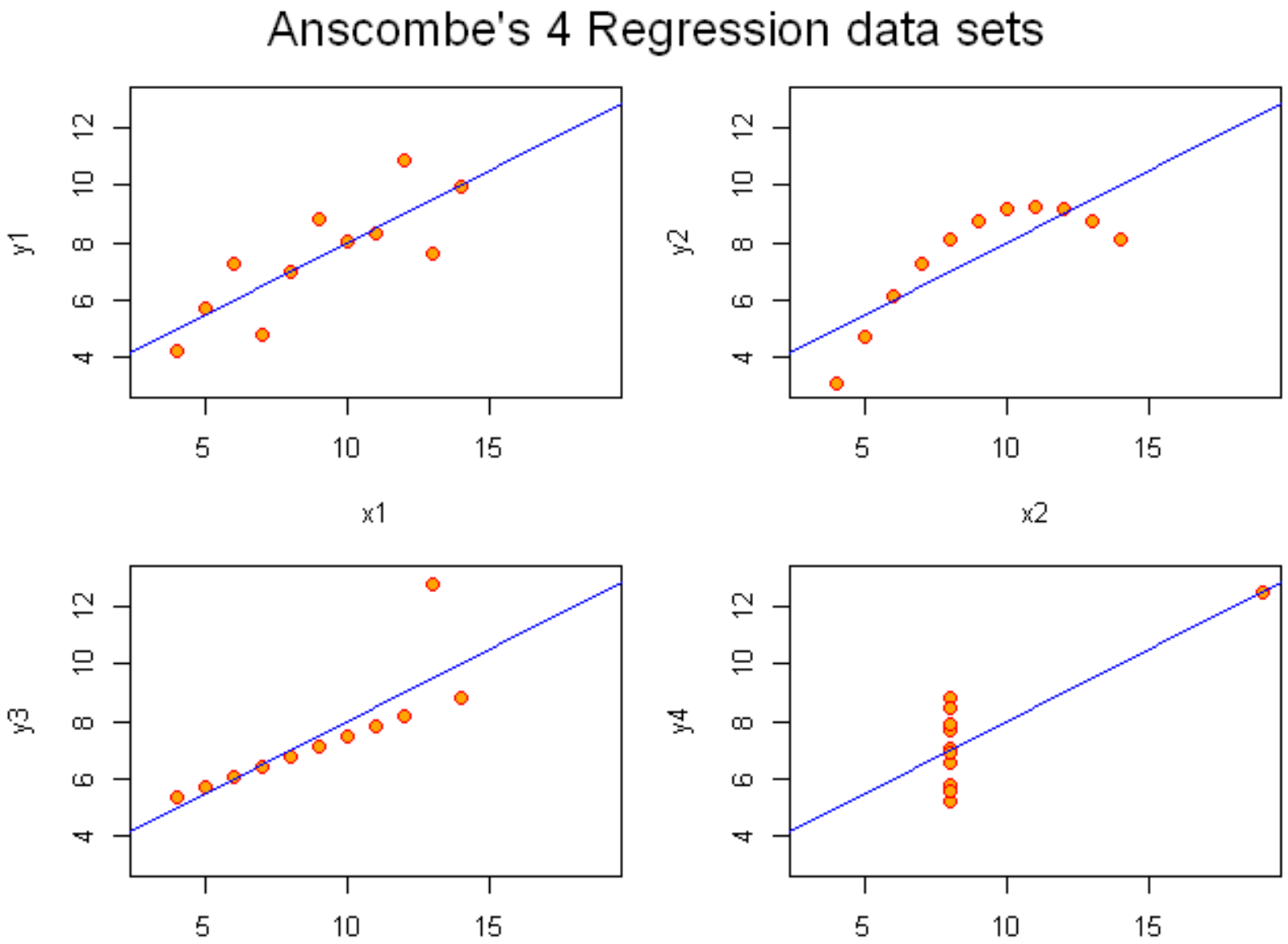
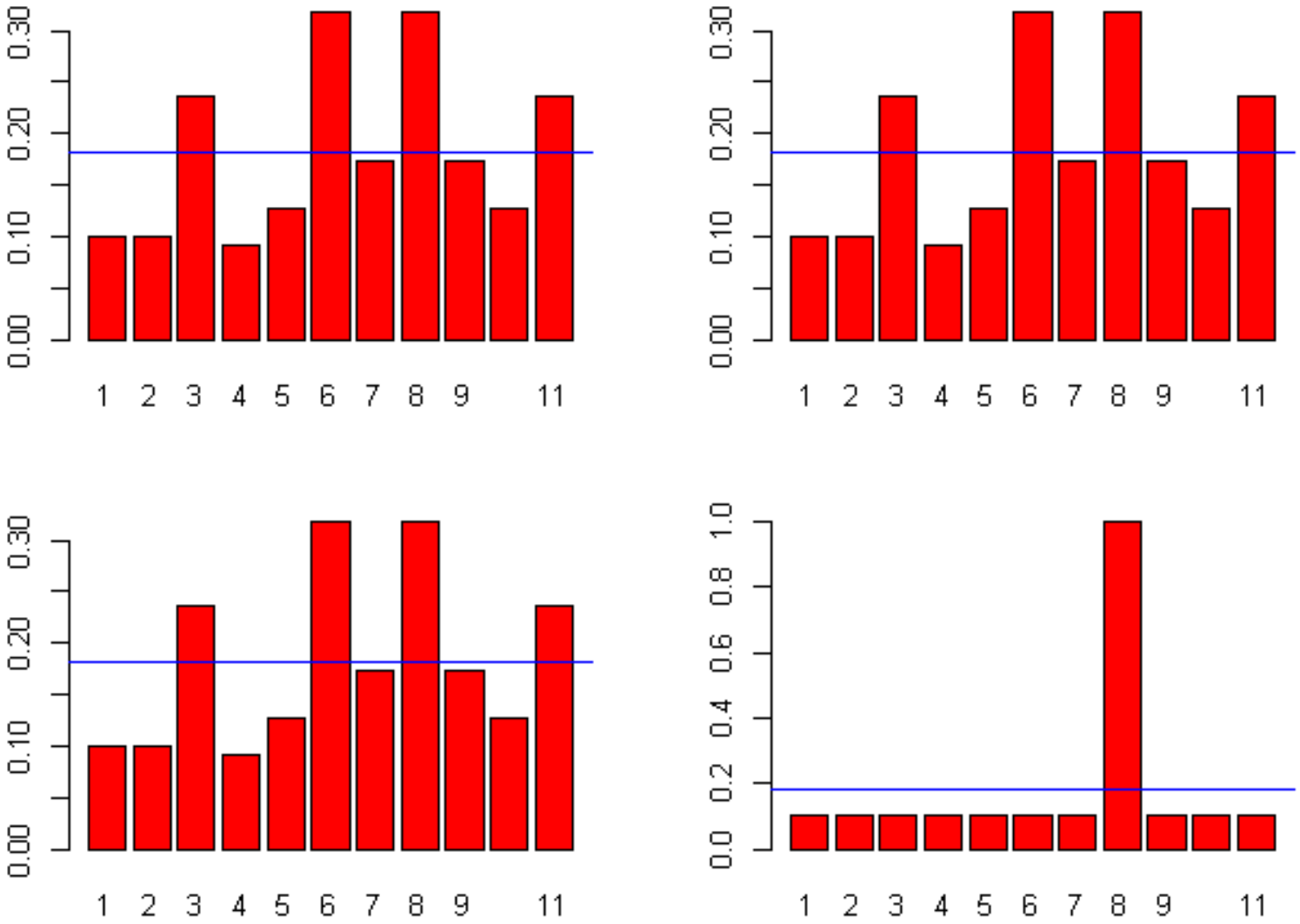


Figure: Anscombe Quartet: Leverage plot (h_i vs i)



Specification issues: omitted variable bias, collinearity, etc.

Consider the regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \gamma \boldsymbol{w} + \boldsymbol{\varepsilon}$, where $\boldsymbol{w}$ is a vector of N measurements on an added variable W. We can show that the OLS estimator of γ is

$$\hat{\gamma} = \frac{\boldsymbol{w}'\boldsymbol{M}\boldsymbol{y}}{\boldsymbol{w}'\boldsymbol{M}\boldsymbol{w}}, \quad \boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}',$$

and

$$\text{Var}(\hat{\gamma}) = \frac{\sigma^2}{\boldsymbol{w}'\boldsymbol{M}\boldsymbol{w}}.$$

These expressions have a nice interpretation: $\hat{\gamma}$ results from

- Regress Y on X and obtain the residuals $\boldsymbol{e}_{Y|X} = \boldsymbol{M}\boldsymbol{y}$;
- Regress W on X and obtain the residuals $\boldsymbol{e}_{W|X} = \boldsymbol{M}\boldsymbol{w}$;
- Regress $\boldsymbol{e}_{Y|X}$ on $\boldsymbol{e}_{W|X}$ to obtain the LS estimate $\hat{\gamma}$ (partial regr. coeff.).

Multicollinearity

- $\boldsymbol{Mw}$ are the residuals of the regression of W on X.
- $\boldsymbol{w}'\boldsymbol{Mw}$ is the residual sum of squares from that regression.
- If X explains most of the variation in W, this is very small, and $\text{Var}(\hat{\gamma})$ will be very high.
- We can conclude that the effect that W exerts on Y is poorly (i.e. very imprecisely) estimated.
- The variance inflation factor provides a measure of the increase in the variance of the OLS estimator due to multicollinearity: it is defined as $1/(1 - R_w^2)$, where R_w^2 is the R-squared of the regression of W on X (write $\boldsymbol{w}'\boldsymbol{Mw} = \boldsymbol{w}'\boldsymbol{w}(1 - R_w^2)$).

Omission of relevant variables

If W is omitted from the regression, the OLS estimator $\hat{\boldsymbol{\beta}}$ is biased,

$$\mathrm{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{w}\gamma,$$

but also more precise:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \leq \sigma^2(\boldsymbol{X}'\boldsymbol{M}_W\boldsymbol{X})^{-1};\ \boldsymbol{M}_W = \boldsymbol{I} - \boldsymbol{w}(\boldsymbol{w}'\boldsymbol{w})^{-1}\boldsymbol{w}',$$

where $\sigma^2(\boldsymbol{X}'\boldsymbol{M}_W\boldsymbol{X})^{-1}$ is the variance of the estimator of $\boldsymbol{\beta}$ when W is included in the model.
If W is irrelevant (true $\gamma = 0$) there is only a cost associated to including it in the model.
In any case, if W is orthogonal to X the distribution of $\hat{\boldsymbol{\beta}}$ is not affected by W.

Prediction of a new sample observation

We are interested in predicting Y for a new unit with values of X in the vector $\boldsymbol{x}_o = (1, x_{o1}, \ldots, x_{op})'$. The optimal predictor under square loss is

$$\hat{y}_o = \hat{f}(\boldsymbol{x}_o) = \hat{\beta}_0 + \hat{\beta}_1 x_{o1} + \cdots + \hat{\beta}_p x_{op} = \boldsymbol{x}_o' \hat{\boldsymbol{\beta}}$$

Under the stated assumptions, the predictor is unbiased, i.e. the prediction error has zero mean, $\mathrm{E}(Y_o - \hat{y}_o) = 0$, and

$$\mathrm{Var}(Y_o - \hat{y}_o) = \sigma^2(1 + \boldsymbol{x}_o'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_o)$$

Note: the stated assumptions concern the selection of the set of predictors in X and the properties of ε.

Predictive accuracy depends on

1. Ability to select all the relevant variable. Failure to do so will result in an inflated σ^2 and $\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) \neq 0$, so that $\hat{\boldsymbol{\beta}}$ is no longer optimal.
2. How removed is the unit profile $\boldsymbol{x}_o$ from the rest: $\boldsymbol{x}_o'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_o$ measures the (Mahalanobis) distance of the values of the $X_1, \ldots, X_p$ for the o-th unit in the space of the X's.

If the true model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \gamma \boldsymbol{w} + \boldsymbol{\varepsilon}, \mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}, \boldsymbol{w}) = \boldsymbol{0}, \mathrm{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X}, \boldsymbol{w}) = \sigma^2 \boldsymbol{I},$$

the above predictor is biased and its mean square error is

$$\mathrm{E}[(Y_0 - \hat{f}(\boldsymbol{x}_o))^2|\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{x}_o, w_o] = \sigma^2(1 + \boldsymbol{x}_o'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_o) + \mathrm{Bias}^2,$$

with

$$\mathrm{Bias} = (\boldsymbol{\beta} - \mathrm{E}(\hat{\boldsymbol{\beta}}))'\boldsymbol{x}_o + \gamma w_o = (w_o - \boldsymbol{w}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_o)\gamma$$

The term in parenthesis is the error of prediction of w_o using the information in $\boldsymbol{x}_o$.