PSI:
Linear Regression

Technological University Dublin City Campus

School of Computer Science

Statistical regression models

Statistical regression models

- are polynomials that express the dependent variable in terms of independent variables
- typically predict numeric dependent variables based on numeric independent variables
- can use categorical variables as predictors (independents) but in 'dummy' (one-hot encoded) form

	number of variables		**polynomial**
linear regression type	**independent**	**dependent**	**degree**
simple	1	1	1
multiple (multivariable) linear	>1	1	1
polynomial	1 or >1	1	>1
non-linear	1 or >1	1	NA e.g. exponential
multivariate	>1	>1	1

Why ’regression’?

Francis Galton’s heights diagram

Linear regression model

- A useful linear regression model can be generated if there is a linear relationship between the independent and dependent variable.

- The model is a **function** of the following form:

  $$f(x) = w_0 + w_1x_1 + w_2x_2 ... + w_nx_n$$

  where x_1, x_2...x_n are the independent variables and w_0, w_1, w_2... w_n are the coefficients of the linear function.

- The coefficients w_0 ... w_n are chosen (i.e. the model fitted) on the principle of minimizing the mean distance of the observations from the model line (or plane or hyperplane).

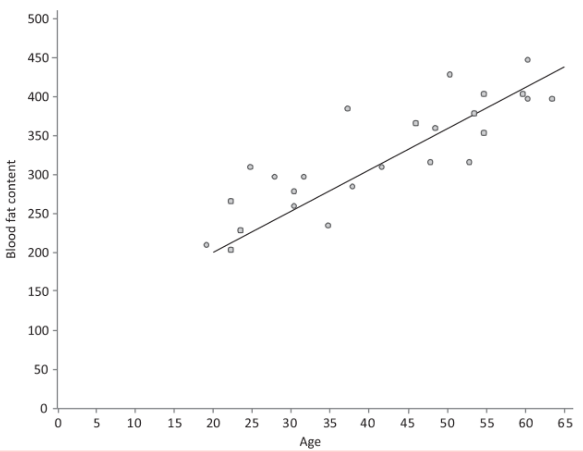
- The model *represents the relationship* between the independent variables on one side and the dependent variable on the other: given a new observation, the dependent variable value can be calculated from the values of the independent variables.

Fitting a simple regression model

- A linear regression model with one independent variable can be fitted using a couple of formulae.

- The regression line formula is: $y = a + bx$

  where y is the dependent variable, x is the independent variable and a and b are the coefficients of the linear function.



The picture shows an example of some data with one independent variable together with the fitted regression line.

Source: [MSD]

- The coefficient b, which is in face the slope of the line, is calculated as follows:

  $$b = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}$$

  where x_i and y_i are the independent and dependent variable values for the i^{th} observation, n is the number of observations and $\bar{x}$ and $\bar{y}$ are the means of the independent and dependent variable, respectively.

- The coefficient a can be calculated by substituting the variables in the equasion by the means:

  $$a = \bar{y} - b \times \bar{x}$$

- Fitting a model with more than one independent variable is more complicated and is generally done using statistics packages or programming language methods.

Other regression models

- **Multivariable regression** is analogous to simple regression but instead of a line, the shape being fitted is
  - a two-dimensional plane or level curve (with 2 independent variables)
  - a multi-dimensional (hyper)plane or curve (with 3 or more variables)

  The coefficients are fitted through matrix calculations.

- **Polynomial regression** includes terms of power > 1 and interaction terms
  - these models are also linear regression models with respect to the coefficients and are fitted in the same way
  - the independent variables must be standardised for reduction of multicollinearity
- **Non-linear regression** must be fitted using iterative numerical algorithms

Fitting a multiple linear regression model

This is a vector and matrix analogue to the simple regression fitting formula, based on least squares estimates:

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

where

$\hat{\boldsymbol{\beta}}$ is the estimated coefficient vector, $\mathbf{X}$ is the matrix of predictor variable values, $\mathbf{y}$ is the vector of dependent variable values, n is the number of observations and p is the number of independent (predictor) variables;

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

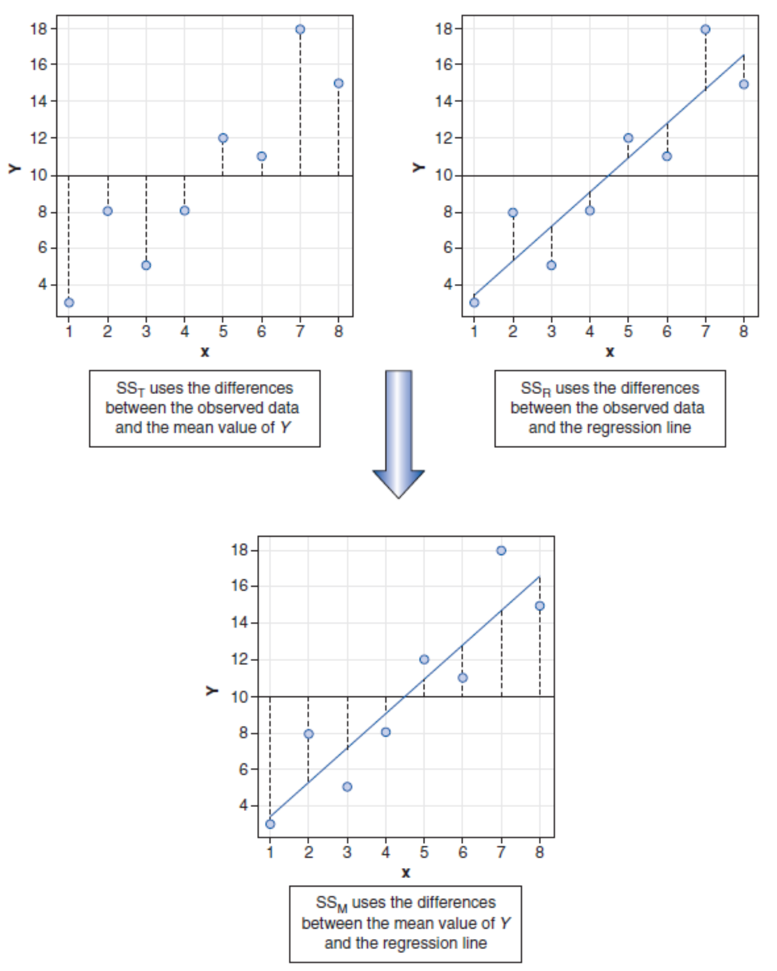
[Fitting a multiple linear regression model]

Values of the response (dependent) variable are estimated using the fitted model:

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

where $\hat{\mathbf{y}}$ is the vector of estimated dependent variable values and $\mathbf{H}$ is the 'hat matrix'.

Model evaluation



SS_T uses the differences between the observed data and the mean value of Y

SS_R uses the differences between the observed data and the regression line

SS_M uses the differences between the mean value of Y and the regression line

Source: [DSR]

Sums of squares

- SS_T total sum of squares: total variability

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

- SS_R residual sum of squares: variability outside of the model

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

- $SS_M = SS_T - SS_R$ model sum of squares

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

Measures of model performance

R^2 **and adjusted** R^2 (coefficient of determination): measures how much of the variation in the dependent variable can be predicted (explained) by the independent variables.

$$R^2 = \frac{SS_M}{SS_T}$$

The adjusted value accounts for the number of predictor variables.

F statistic: measures the correlation between the independent variables (as a group) and the dependent variable.

$$F = \frac{MS_M}{MS_R} = \frac{\frac{SS_M}{p-1}}{\frac{SS_R}{n-p}}$$

where p is the number of independent variables and n is the number of observations.

Regression coefficient p-values: measure individual significance of independent variables as predictors of the dependent variable.

General applicability of the model

- **Overfitting** typically occurs with linear regression when the number of independent variables is greater than the number of observations. As the model grows with independent variables being added into it, the move in the direction of overfitting manifests itself in a growing coefficient of determination, R^2.

- This can be avoided using **cross-validation**. This is a testing strategy whereby the model is fitted using part of the data (e.g. 70%) and the remainder of the data are held out of the fitting process, only to be used as yet unseen inputs for testing the model. During the test, the independent variables of the held-out section of the data are fed into the model and the output (generated values for the dependent variable) compared to the actual data using some measure, e.g. R^2.

- The following picture shows a fitting curve, including the mean squared errors for a training and testing subset of the data:

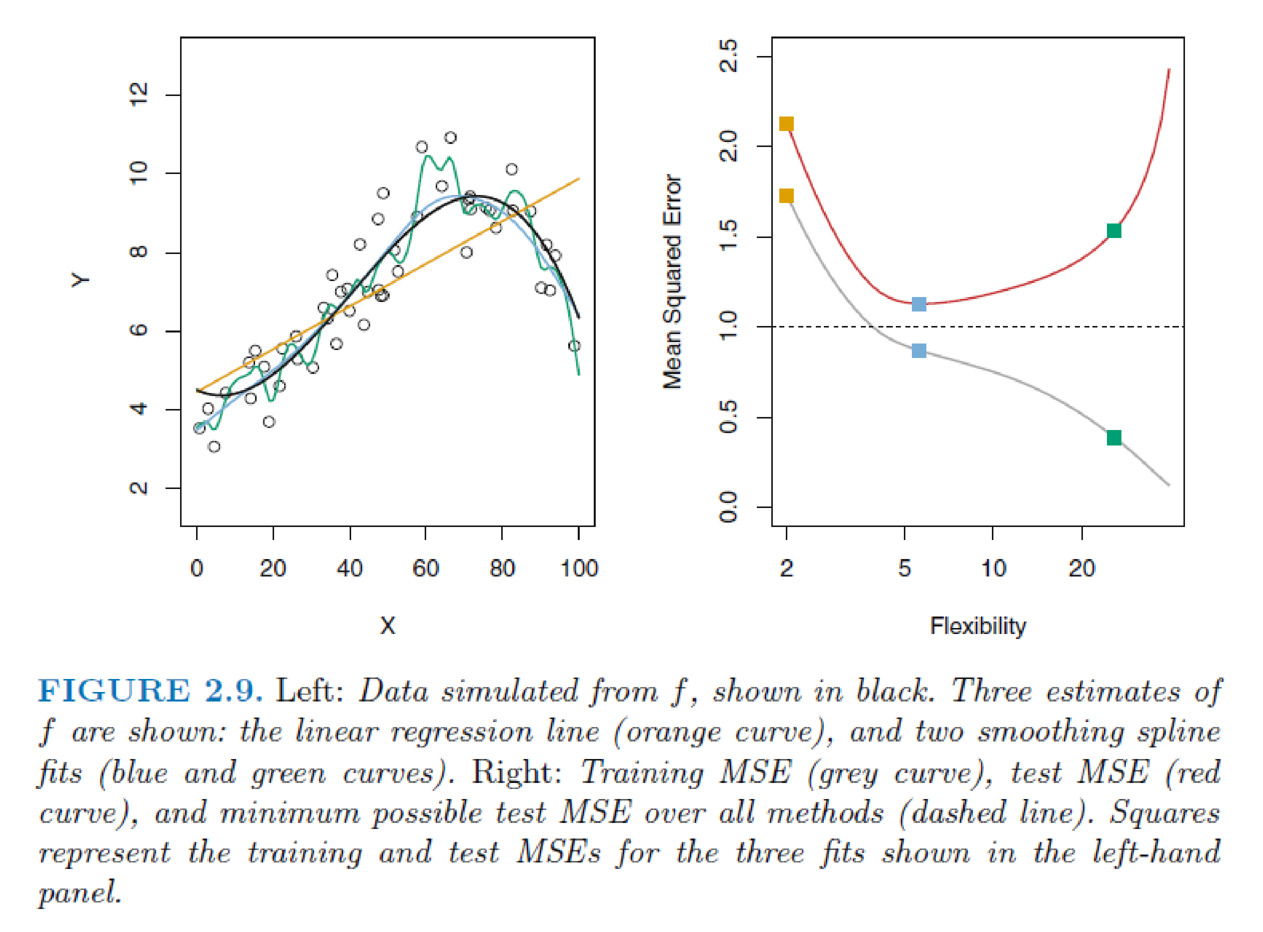
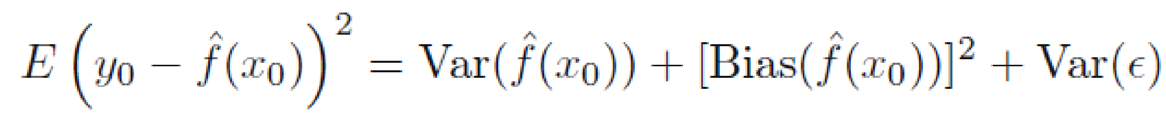


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

Source: [ISL]

- For the model to be the best possible representation of the population but not overfitted to the sample, a **bias-variance** tradeoff needs to take place, with both minimised (blue square) in the **test mean squared error**:

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

  Source: [ISL]

- There is a rough rule of thumb that there should be about 15 observations per independent variable in the data set for the model not to be overfitted simply by virtue of variable number.

- Another way of dealing with overfitting is to use a measure of model performance that accounts for the number of independent variables, e.g.:
  - **adjusted $\mathbf{R^2}$**, for which there are various formlae, with the following the oldest:

    $$adjustedR^2 = 1 - \frac{(n-1)}{(n-p-1)}(1-R^2)$$

    where n is the number of observations, p is the number of predictor variables and R^2 is the unajusted value.
  - **Akaike information criterion (AIC)**, which is a relative measure that has a meaning only when models are being compared in a specific context:

    $$AIC = n\,ln\left(\frac{SSE}{n}\right) + 2k$$

    where n is the number of observations and k is the number of predictor variables.
  - others such as **Mallow's $\mathbf{C_p}$** and **Bayesian information criterion (BIC)**

Predictor variable selection

- **Selection based on domain knowledge**

  A model may be fitted for the specific purpose of investigating particular properties of the subjects or instances that are being studied. Often, previous work would have established knowledge on the factors correlated with the dependent variable and these would be used, with the addition of some new variables of interest.

- **Best subset selection**

  This requires the investigation of all possible subsets of independent variables (numbering 2^p if p is the number of variables) and their comparison so that the one with the best performance could be selected. In the past, this method would have been computationally prohibitive.

- **Forward or backard stepwise selection**

  Forward selection starts with a no-variable model. Variables are added one-by one and model performance maximised at every step. This is the only method suitable for data sets with numbers of variables larger than observation counts. The model is built through the addition of variables one at a time and the process can stop before overfitting occurs.

  Backward selection works the other way around. The first investigated model contains all variables and they are removed one by one.

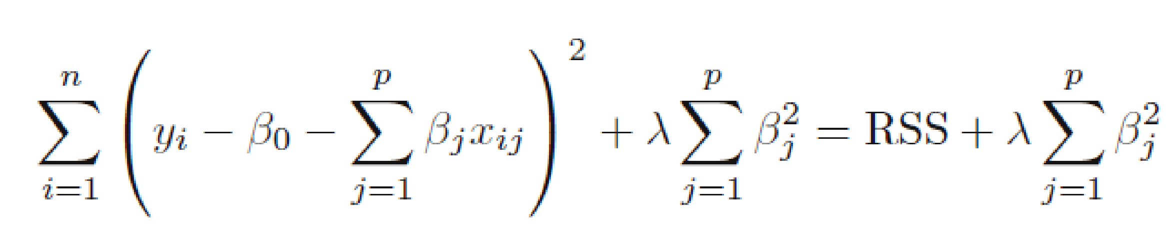
- With both best subset and stepwise selection, **comparison** between models of the same size (with respect to included number of variables) is performed using R^2 or MSE (mean squared error). Models of different sizes are compared using the adjusted R^2 or other measures that account for the variable number (e.g. AIC or BIC).

- **Problems**
  - *Stepwise selection* does not cover the entire variable combination space. Because of this and commonly present interaction between variables, the best combination of variables can be missed, which is a disadvantage. For example the best single-variable model may be one with variable **a** but the best two-variable model might be one with variables **b** and **c**. Because in forward stepwise selection variable **a** will have been added first, the combination of **b** and **c** would be 'missed'.
  - Because predictor selection is a statistical process itself, the cumulative probability of type I error can be large if not approached correctly. A Bonferroni correction or similar must be applied, otherwise fitting a model in this way could turn into a 'fishing' exercise (the 'squeezing' of data until it yields a significant result). For this reason, analysis with pre-stated hypotheses based on domain knowledge is often preferable.

Regularisation

An alternative to predictor variable selection, regularisation is an approach that includes all variables to start with but is based on minimisation of an objective function that includes a regularisation term in addition to the error squares term. The purpose of this additional term is to reduce overfitting by mathematically limitting the size of the regression parameters. Examples of regularisation are **ridge regression** and **lasso regression**. The lasso function is such that it can result in some variables being removed from the model i.e. in variable selection.

Ridge regression objective function:



$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

Source: [ISL]

Ridge regression effect:

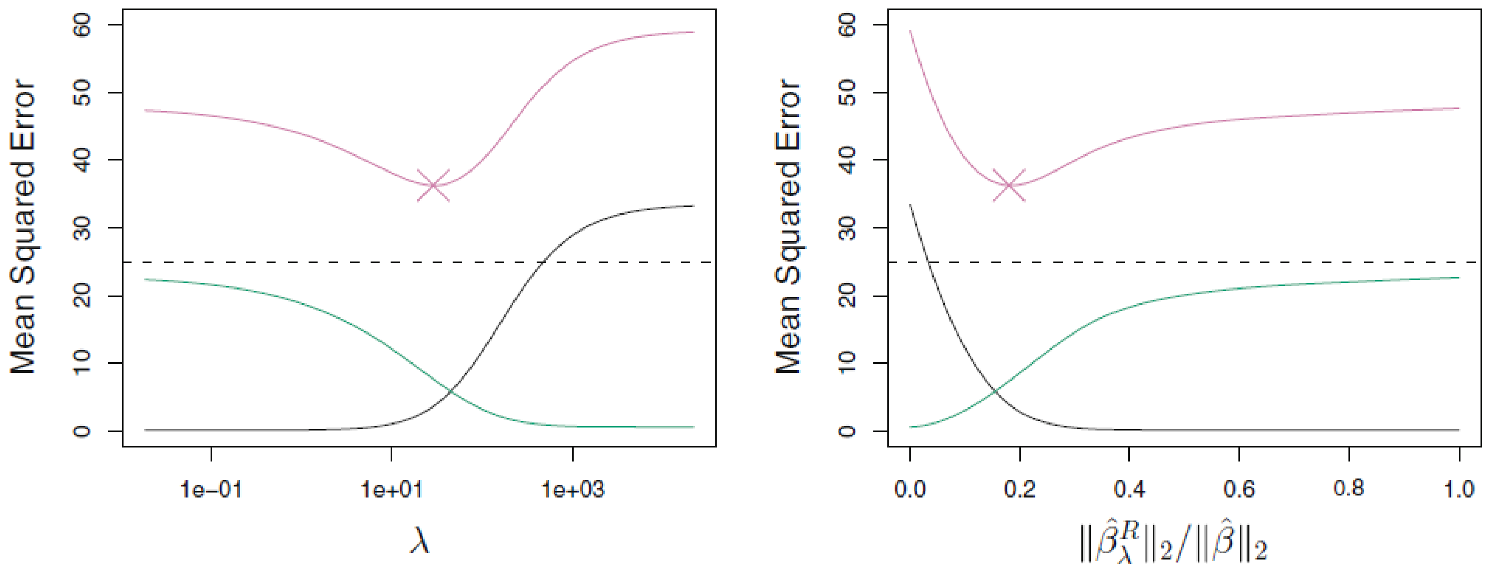


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Source: [ISL]

Linear regression diagnostics

With one variable, we had to investigate outliers as extreme cases. In the case of a predictor and dependent variable there are three types of extremes to worry about:

- **outliers** are observations whose residuals have extreme values with respect to the residual distribution (which is a t-distribution)

- **leverage** tells us how much **potential** a predictor value has to influence the regression model fit (in conjunction with the response value); for the predictor value x_i it is measured using h_{ii} from the H matrix, which contains factor that scale actual response values, y_i, to predicted values, $\hat{y}_i$ ($\boldsymbol{\hat{y}} = \boldsymbol{Hy}$); $h_{ii} > 3(\frac{p}{n})$ is considered to represent high leverage

- **influential cases** are cases that considerably modify the regression model in comaprison with what it would be if those observations were not present in the model
  - **difference of fits (DFFITS)** is a normalised difference between the predicted value for a dependent (response) observation in the full model and the that in a model built without that observation:

    $$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

    A DFFITS value is considered influential if $DFFITS_i > 2\sqrt{\frac{p+1}{n-p-1}}$

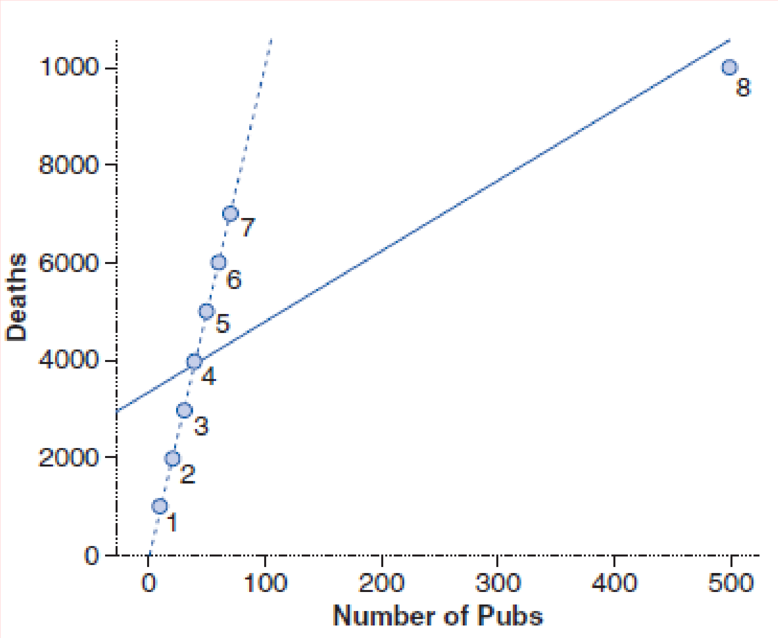
    Analogous to this are **DFBeta** values, for model parameters.

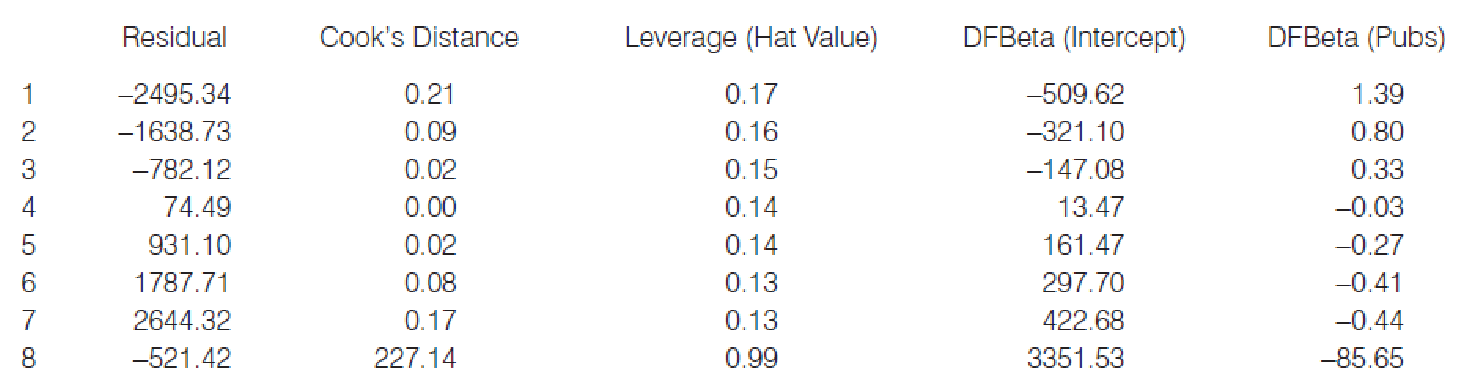
- **Cook's distance** is a value that combines the residual for an observation with its leverage:

  $$D_i = \frac{(y_i - \hat{y}_i)^2}{p \times MSE} \frac{h_{ii}}{(1 - h_{ii})^2}$$

  Values above 1 or values that stand out among other observations are considered to be influential.

An example:





	Residual	Cook's Distance	Leverage (Hat Value)	DFBeta (Intercept)	DFBeta (Pubs)
1	–2495.34	0.21	0.17	–509.62	1.39
2	–1638.73	0.09	0.16	–321.10	0.80
3	–782.12	0.02	0.15	–147.08	0.33
4	74.49	0.00	0.14	13.47	–0.03
5	931.10	0.02	0.14	161.47	–0.27
6	1787.71	0.08	0.13	297.70	–0.41
7	2644.32	0.17	0.13	422.68	–0.44
8	–521.42	227.14	0.99	3351.53	–85.65

Source: [ISL]

Model assessment: assumptions

- Data unbounded within its domain
- Non-zero variance in predictor variables
- No perfect multicollinearity
- Predictors are not correlated with any external variables
- Homoskedasticity
- Independent errors (no autocorrelation - can be checked with Durbin-Watson test)
- Normally distributed errors
- Independence of values in the outcome variable
- Linearity

Model assessment: multicollinearity

- **Multicollinearity** is high correlation between two or more independent variables in a regression model.
- It can be detected by calculating the correlation value for all pairs of independent variables in the data.
- Problems:
  - perfect correlation between some independent variables prevents model fitting (but R lm function will just drop one of the variables in order to proceed)
  - less confidence in the regression parameters
  - harder to tell how good the individual predictors are
- Measure: **variance inflation factor (VIF)** - values over 10 problematic
- What to do?
  - delete all but one of the collinear variables
  - combine them in some domain-appropriate way
  - combine them using principal component analysis

Using the model for prediction

- Once fitted, a model that meets all the assumptions can be used for prediction
- The values for the predictor variables on a new observation, with unknown response value, are substituted into the regression function and the response value calculated
- If the assumptions are not met, a prediction model can be built through random resampling of the data, with the use of bootstrapping