Understanding Data and Statistical Design (60117)

Chapter 7

Multiple Linear Regression I

Subject Coordinator: Stephen Woodcock
Lecture notes: Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

Chapter outline

Topics:

- fitting planes to data
    - model setup
    - method of least squares
    - vector/matrix form
- regression model
    - assumptions
- coefficient T-test
    - running the test
- model fit
    - ANOVA and F-test
    - ANOVA, R^2 and R^2_{adj}
    - leverage and influence
    - multicollinearity
- R example
    - model 1
    - model 2
    - model 2 (reduced dataset)

Fitting planes to data – model setup

So far we have described **simple linear regression**, where the least squares method is used to fit a **line** to data in $\mathbb{R}^2$, with the data the ordered pairs of dependent and independent variables.

We now develop **multiple linear regression**, where the least squares method is applied to fit a **plane** to data in $\mathbb{R}^{m+1}$, with the data the $(m+1)$–tuples of dependent and m independent variables.

We can also use multiple regression to fit curves instead of lines in $\mathbb{R}^2$, and indeed curved surfaces instead of planes in higher dimensions.

We do this by transforming the data to obtain linear relationships between the response and predictor variables and then fitting to the transformed data.

This will allow us to consider problems that don't appear linear in nature.

Fitting planes to data – model setup

We suppose the sample data to be modelled can be described as

$$y = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \epsilon.$$

This is the equation of a hyperplane disturbed by an observation of some RV ϵ which we call the **noise** or **error** term.

Our sample is one of many possible samples and we suppose it has been drawn from a population described by

$$Y = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \epsilon.$$

Notation

- $\beta_0, \beta_1, \ldots, \beta_m$ are unknown constants
- $x_1, \ldots, x_m$ are non-random
- y is an observation of the RV Y
- ϵ is an observation if used in the context of y and a RV if used in the context of Y

Fitting planes to data – model setup

The **estimates** $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ define the **fitted regression model**

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_j.$$

The fitted model is an observation of the **population regression model**

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_j. \tag{1}$$

Assuming $\mathbb{E}[\epsilon] = 0$, the fitted model is used to **estimate** the **mean** or **expected value**

$$\mathbb{E}[Y] = \beta_0 + \sum_{j=1}^{m} \beta_j x_j.$$

Notation

- $\hat{y}$ is an observation of the RV $\hat{Y}$
- $\hat{\beta}_0, \hat{\beta}_1 \ldots, \hat{\beta}_m$ are observations if used in the context of $\hat{y}$ and RVs if used in the context of $\hat{Y}$

Fitting planes to data – model setup

The **statistical model** for multiple linear regression is defined via the random sample

$$Y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \epsilon_i$$

for $i \in \{1, \ldots, n\}$ where

- Y_i is the i-th observation of the response
- β_0 is the intercept coefficient
- $\beta_1, \ldots, \beta_m$ are the slope coefficients
- x_{ij} is the i-th observation of the j-the predictor
- ϵ_i is the i-th noise or error term.

The **sample data** is an observation of the random sample

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \epsilon_i$$

for $i \in \{1, \ldots, n\}$.

Fitting planes to data – method of least squares

To fit the model we use a generalisation of the least squares method employed for simple linear regression.

The method of least squares is based on the idea of finding estimators $\hat{\beta}_0, \hat{\beta}_1 \ldots, \hat{\beta}_m$ such the **residual**

$$\begin{aligned}\hat{\epsilon} &= Y - \hat{Y} \\ &= Y - \hat{\beta}_0 - \sum_{j=1}^{m} \hat{\beta}_j x_j\end{aligned}$$

is minimised in some way.

The residual RV $\hat{\epsilon}$ is an **estimator** of the noise RV

$$\begin{aligned}\epsilon &= Y - \mathbb{E}[Y] \\ &= Y - \beta_0 - \sum_{j=1}^{m} \beta_j x_j\end{aligned}$$

assumed in the population model from which the sample data is drawn.

Fitting planes to data – method of least squares

The estimates $\hat{\beta}_0, \hat{\beta}_1 \ldots, \hat{\beta}_m$ are the values that $\beta_0, \beta_1, \ldots, \beta_m$ would take to minimise the **sum square error**

$$sse(\beta_0, \beta_1, \ldots, \beta_m) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2 = \sum_{i=1}^{n} \epsilon_i^2 .$$

We can write this problem mathematically as

$$(\hat{\beta}_0, \hat{\beta}_1 \ldots, \hat{\beta}_m) = \underset{(\beta_0, \beta_1, \ldots, \beta_m)}{\operatorname{argmin}} \; sse(\beta_0, \beta_1, \ldots, \beta_m)$$

and solve using techniques of calculus.

Fitting planes to data – method of least squares

In practice we don't need to do this ourselves, as R will perform all calculations for us.

However, we will outline the solution for those interested (ignore if not).

Through differentiation we define the **normal equations**

$$\frac{\partial}{\partial \beta_j} sse(\beta_0, \beta_1, \ldots, \beta_m)|_{\beta_j = \hat{\beta}_j} = 0$$

for all $j \in \{0, 1, \ldots, m\}$.

Fitting planes to data – method of least squares

After performing the differentiation, the normal equations become

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{m} \hat{\beta}_j x_{ij} \right) = 0$$

and

$$\sum_{i=1}^{n} x_{ij} \left(y_i - \hat{\beta}_0 - \sum_{k=1}^{m} \hat{\beta}_k x_{ik} \right) = 0, \quad j \in \{1, \ldots, m\},$$

which are solved for $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$.

Important. In multiple linear regression, a solution to the normal equations cannot always be found. It is easier to illustrate why this is the case using vector/matrix notation.

Fitting planes to data – method of least squares

There remains one other population parameter to find an estimator for, the variance σ^2 of the noise RV ϵ.

It turns out that an **unbiased** estimate for σ^2 is given by

$$s^2 = \frac{sse(\hat{\beta}_0, \hat{\beta}_1 \ldots, \hat{\beta}_m)}{n-m-1} = \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_{ij})^2}{n-m-1}$$
$$= \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-m-1}. \tag{2}$$

The least squares problem is solved.

Fitting planes to data – vector/matrix form

Alternative notation using vectors and matrices provides more compact depiction and simpler algebraic manipulation.

Arrange response RV and the sample data as

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \ \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \boldsymbol{x} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix}$$

where $\boldsymbol{x}$ has been augmented with a column of ones. Also define the fitted value, parameter, estimated parameter and residual vectors

$$\hat{\boldsymbol{y}} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_m \end{pmatrix}, \ \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_m \end{pmatrix}, \ \hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{pmatrix} \hat{\epsilon}_1 \\ \vdots \\ \hat{\epsilon}_n \end{pmatrix}.$$

Fitting planes to data – vector/matrix form

In matrix form, the model we are trying to fit is

$$\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{x}\boldsymbol{\beta}.$$

The estimate $\hat{\beta}$ minimises sse, which can be written in vector form as

$$sse(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}},$$

and is the solution of the **normal equation**

$$(\boldsymbol{x}^T \boldsymbol{x})\hat{\beta} = \boldsymbol{x}^T \boldsymbol{y}$$

which, provided the inverse exists, is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{x}^T \boldsymbol{x})^{-1} \boldsymbol{x}^T \boldsymbol{y}. \tag{3}$$

The inverse exists if the columns of $(\boldsymbol{x}^T \boldsymbol{x})$ are linearly independent. If not, remove superfluous independent variable(s) and continue.

Fitting planes to data – vector/matrix form

Then the model fitted to the sample data can be written as

$$\hat{\boldsymbol{y}} = \boldsymbol{x}\hat{\beta}. \tag{4}$$

The fitted vector $\hat{\boldsymbol{y}}$ is the **orthogonal projection** of the sample response vector $\boldsymbol{y}$ onto the column space of $\boldsymbol{x}$ or

$$\hat{\boldsymbol{y}} = \mathrm{proj}_{\mathrm{col}\,\boldsymbol{x}}\boldsymbol{y}.$$

This provides a different perspective of the least squares procedure and provides another justification for selecting as parameter estimates the values that minimise the sum of squared errors.

Regression model – assumptions

Although fitting the regression model using least squares requires no assumptions about the nature of the data, to go further and develop tools to analyse the fitted model does.

We make the assumptions:

- $\epsilon_i \sim N(0, \sigma)$, i.e. normally distributed with $\mathbb{E}[\epsilon_i] = 0$ and $\text{var}(\epsilon_i) = \sigma^2$
- ϵ_i are all independent from each other.

These assumptions can be re-stated in terms of the response variable as $Y_i \sim N(\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}, \sigma)$ and independent.

In summary, the assumptions are:

1. normality
2. constant variance
3. independence.

Coefficient T-test

The assumptions just listed enable us to determine the distributions of the parameter estimates $\hat{\beta}_0, \ldots, \hat{\beta}_m$ and the regression model $\hat{Y}$ given by (1).

This allows us to develop hypothesis tests and CIs for the parameter estimates and the model predictions.

T-tests can be performed on the parameters β_j, $j \in \{0, 1, \ldots, m\}$, using the RV

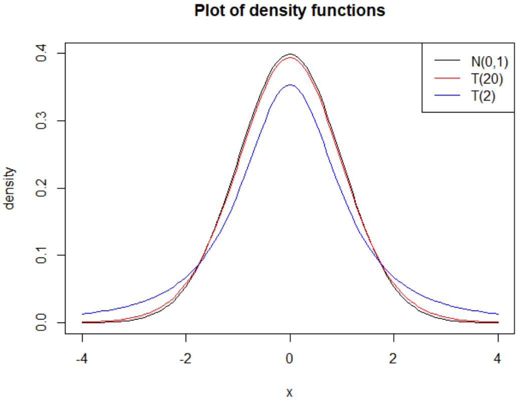
$$T_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}}, \tag{5}$$

which, under the assumptions, follows a Students' T-distribution with $n - m - 1$ degrees of freedom.

The standard error $S_{\hat{\beta}_j}$ is most conveniently expressed in matrix notation (details omitted).

Coefficient T-test

PDFs for Student's T-distributed RV for a range of degrees of freedom are displayed below.



The Student's T-distribution converges to the standard normal distribution as $n \to \infty$.

Coefficient T-test – running the test

Hypotheses

The **null hypothesis** for this test is

$$H_0\colon\ \beta_j = \beta_j^*, \quad j \in \{0, 1, \ldots, m\},$$

while the **alternative hypothesis** may be any of

$$H_A\colon\ \beta_j < \beta_j^* \text{ (lower tail test)}$$
$$H_A\colon\ \beta_j \neq \beta_j^* \text{ (two tail test)}$$
$$H_A\colon\ \beta_j > \beta_j^* \text{ (upper tail test)}$$

where β_j^* is the hypothesised value of β_j.

Test statistic

The test statistic is calculated from the sample data as

$$t^*_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j^*}{s_{\hat{\beta}_j}}.$$

Under H_0, $t^*_{\hat{\beta}_j}$ is an observation of the $T(n-m-1)$ RV in (5).

Coefficient T-test – running the test

Test decision – lower tail test
H_0 is rejected at significance level $0 < \alpha < 1$ if

$$t^*_{\hat{\beta}_j} < t_\alpha,$$

where the quantile t_α is from $T(n - m - 1)$ distribution.

Equivalently, H_0 is rejected if β^*_j falls outside the $100(1 - \alpha)\%$ **lower tail confidence interval (CI)** for μ given by

$$-\infty < \beta_j \leq \hat{\beta}_j + s_{\hat{\beta}_j} t_{1-\alpha}$$

or if the p-value

$$p = \text{Prob}(T < t^*_{\hat{\beta}_j}) < \alpha$$

where $T \sim T(n - m - 1)$.

The null hypothesis H_0 is retained in any other case.

Coefficient T-test – running the test

Test decision – two tail test
H_0 is rejected at significance level $0 < \alpha < 1$ if

$$|t^*_{\hat{\beta}_j}| > t_{1-\alpha/2},$$

where the quantile $t_{1-\alpha/2}$ is from $T(n-m-1)$ distribution.

Equivalently, H_0 is rejected if β^*_j falls outside the $100(1-\alpha)\%$ **two tail CI** for μ given by

$$\hat{\beta}_j - s_{\hat{\beta}_j} t_{1-\alpha/2} \leq \beta_j \leq \hat{\beta}_j + s_{\hat{\beta}_j} t_{1-\alpha/2}$$

or if the p-value

$$p = 2 \times \text{Prob}(T > |t^*_{\hat{\beta}_j}|) < \alpha$$

where $T \sim T(n-m-1)$.

The null hypothesis H_0 is retained in any other case.

Coefficient T-test – running the test

Test decision – upper tail test
H_0 is rejected at significance level $0 < \alpha < 1$ if

$$t^*_{\hat{\beta}_j} > t_{1-\alpha},$$

where the quantile $t_{1-\alpha}$ is from $T(n-m-1)$ distribution.

Equivalently, H_0 is rejected if β^*_j falls outside the $100(1-\alpha)\%$ **upper tail CI** for μ given by

$$\hat{\beta}_j - s_{\hat{\beta}_j} t_{1-\alpha} \leq \beta_j < \infty$$

or if the p-value

$$p = \text{Prob}(T > t^*_{\hat{\beta}_j}) < \alpha$$

where $T \sim T(n-m-1)$.

The null hypothesis H_0 is retained in any other case.

Model fit – ANOVA and F-test

Although it is possible to perform hypothesis tests on the individual parameter estimates $\hat{\beta}_0, \dots, \hat{\beta}_m$ using T-tests, it can be convenient to test them simultaneously using an F-test.

The F-test via ANOVA uses the idenity

$$SST = SSR + SSE$$

where (letting $\mathbf{1}$ be square matrix of ones) the **sum square total**

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \boldsymbol{Y}^T \boldsymbol{Y} - \frac{1}{n} \boldsymbol{Y}^T \mathbf{1} \boldsymbol{Y},$$

the **sum square regression**

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}^T \boldsymbol{Y} - \frac{1}{n} \boldsymbol{Y}^T \mathbf{1} \boldsymbol{Y}$$

and the **sum square error**

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}} = \boldsymbol{Y}^T \boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}^T \boldsymbol{Y}.$$

Model fit – ANOVA and F-test

Define the **mean square regression**

$$MSR = \frac{SSR}{m}$$

and the **mean square error**

$$MSE = \frac{SSE}{n - m - 1}$$

where the denominators are the **degrees of freedom**.

Under the assumptions that the noise terms $\epsilon_i \sim N(0, \sigma)$ and independent and that

$$\beta_1 = \cdots = \beta_m = 0,$$

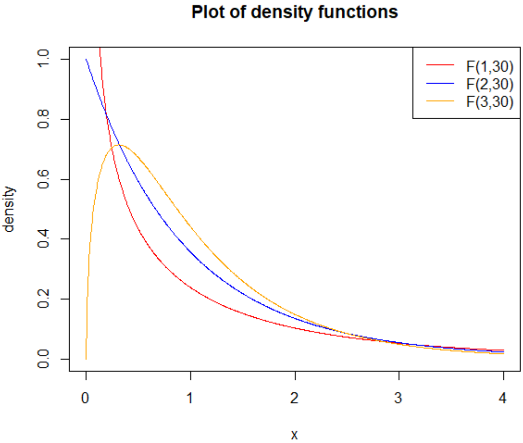
the RV

$$F^* = \frac{MSR}{MSE} \sim F(m, n - m - 1), \tag{6}$$

i.e. F-distributed with m numerator degree of freedom and $n - m - 1$ denominator degrees of freedom.

Model fit – ANOVA and *F*-test

The PDFs for *F*-distributed RVs with a variety of parameter values are displayed below.



The most extreme events occur in the upper tail, so the null hypothesis rejection area is in the upper tail.

Model fit – ANOVA and F-test

Hypotheses

In the context of multiple linear regression, the hypotheses are

$$H_0\colon\ \beta_1 = \cdots = \beta_m = 0$$
$$H_A\colon\ \text{at least one } \beta_j \neq 0.$$

Test statistic

The test statistic

$$f^* = \frac{msr}{mse}$$

is calculated from the data and is an observation of the RV F^* in (6).

Test decision

H_0 is rejected at significance level α if

$$f^* > f_{1-\alpha},$$

where the quantile $f_{1-\alpha}$ is from $F(m, n-m-1)$ distribution.

Model fit – ANOVA and F-test

Equivalently, H_0 is rejected if the p-value

$$p = \text{Prob}(F > f^*) < \alpha,$$

where $F \sim F(m, n - m - 1)$.

The null hypothesis H_0 is retained if this is not the case.

Model fit – ANOVA, R^2 and R^2_{adj}

The decomposition of total squared variation

$$SST = SSR + SSE$$

can also be used to provide a quantitative measure of model fit.

The larger the proportion of total squared variation explained by the model, the better the fit of the model to the data.

We define this proportion as the **coefficient of determination**

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

which satisfies $0 \leq R^2 \leq 1$.

Model fit – ANOVA, R^2 and R^2_{adj}

In the context of multiple regression, this statistic is deficient in that an increase in its value can be due to an increase in the number of predictors and not necessarily to an improvement in model fit.

In fact, $R^2 = 1$ for a model with number of β_j coefficients equal to the sample size (i.e. $m + 1 = n$).

Therefore, for multiple regression a modified version of R^2 is employed that accounts for this phenomenon.

This measure is called **adjusted $\boldsymbol{R^2}$** and is calculated as

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-m-1}$$

and, unlike R^2, be used to compare models fitted to the same sample data but with different number of predictors.

Note that R^2_{adj} may be negative.

Model fit – leverage and influence

Next we revisit the concepts of leverage and influence using a linear algebra perspective.

Using (3) in (4) allows us to write

$$\hat{\boldsymbol{y}} = \boldsymbol{x}\hat{\beta} = \boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{h}\boldsymbol{y}$$

where $\boldsymbol{h}$ is the **hat matrix** or **projection matrix**

$$\boldsymbol{h} = \boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T.$$

The *i*-th diagonal element of $\boldsymbol{h}$ is called the **leverage** of the *i*-th data point.

Model fit – leverage and influence

The hat matrix can be used to describe the covariance structure of the residuals as

$$\text{covar}(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\epsilon}}) = (\boldsymbol{I} - \boldsymbol{h})\sigma^2.$$

This gives us the variance of the *i*-th residual

$$\text{var}(\hat{\epsilon}_i) = (1 - h_{i,i})\sigma^2$$

which can be estimated as

$$\text{var}(\hat{\epsilon}_i) = (1 - h_{i,i})s^2,$$

with s^2 given in (2).

Model fit – leverage and influence

The fact that the residuals $\hat{\epsilon}_i$ do not have the same variance/covariance structure as the noise terms ϵ_i is behind the construction of the so-called "Studentised residuals".

The **internally Studentised residual** of the *i*-th point

$$\hat{t}_i = \frac{\hat{\epsilon}_i}{s\sqrt{1 - h_{ii}}}.$$

The **externally (deleted) Studentised residual** of the *i*-th point

$$\hat{d}_i = \frac{\hat{\epsilon}_i}{s^{(i)}\sqrt{1 - h_{ii}}},$$

where $s^{(i)}$ is the estimate s recalculated having excluded data point i.

Model fit – leverage and influence

To measure the influence of data points we use the internally deleted and externally deleted versions of the Studentised residuals to define **Cook's *D*** and **DFITS** respectively.

The description of these measures, and threshold values indicating potential points of influence, are contained in the previous chapter.

Essentially, if points of influence are identified they should be removed, the model re-run and analysed in comparison to the original.

Model fit – multicollinearity

When adding an extra predictor to a regression model we run the risk that this variable is closely related to predictor variables already in the model.

This can impact adversely on everything from the regression parameter estimates to the statistical tools developed to analyse model fit and to check model assumptions.

We even lose the intuitive interpretation of the associated beta parameter as the increase in the response variable for a unit increase in the predictor variable, all other predictors remaining unchanged.

The most extreme form of this phenomenon is when one predictor variable is **linearly dependent** on the others.

Model fit – multicollinearity

The j-th predictor x_j is linearly dependent if it can be expressed as

$$x_j = \sum_{k=1, k\neq j}^{m} c_k x_k$$

where c_k are constants, i.e. as a linear combination of the other predictor variables.

In this case, the inverse of the matrix $\boldsymbol{x}^T\boldsymbol{x}$, used in (3) to calculate $\hat{\beta}$, does not exist.

At least for this extreme case we will be aware of the problem – the parameter estimates will not compute.

However, sometimes the dependence will be more subtle and we may not be aware of the potential problem.

Model fit – multicollinearity

VARIANCE INFLATION FACTOR

A measure of collinearity of the *j*-th predictor can be obtained by regressing this predictor against the other $m-1$ predictors, i.e. by fitting

$$\hat{x}_j = \sum_{k=1, k\neq j}^{m} \hat{c}_k x_k,$$

and looking for the coefficient of determination R_j^2 to be very close to one (if $R_j^2 = 1$ then we have linear dependence).

Using R_j^2 we can define the **variance inflation factor (VIF)** of the *j*-th predictor variable as

$$VIF_j = \frac{1}{1 - R_j^2}$$

which is the inverse of a statistic called **tolerance**.

As a **rule of thumb**, the *j*-th predictor variable should be considered collinear if $VIF_j > 5$ (some statisticians use the threshold 10 instead).

R example

Now an example using R.

We are going to build a model that lets us predict average fuel consumption for a variety of vehicle characteristics.

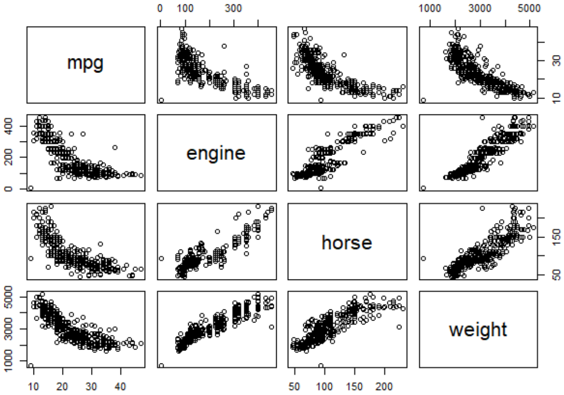
Our sample data contains observations for $n = 392$ vehicles, with the variables summarised in the table below (data in car.data.csv on Canvas).

Name	**Type**	**Description**
mpg	response	fuel consumption (mpg)
engine	predictor	engine displacement (cu in)
horse	predictor	engine horsepower (hp)
weight	predictor	vehicle weight (lb)

We will fit a multiple linear regression model, so the first thing we do is see if linear relationships can be found.

R example

Below is a **matrix scatter plot** produced by R.



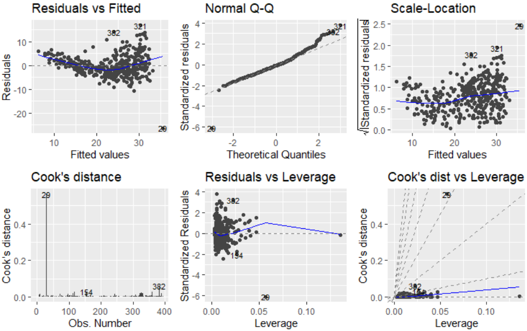
We see nonlinear relationships between the response *mpg* and the predictors.

R example

Fitting a regression to the population model

$$MPG = \beta_0 + \beta_e engine + \beta_h horse + \beta_w weight + \epsilon$$

results in the diagnostic plots below.

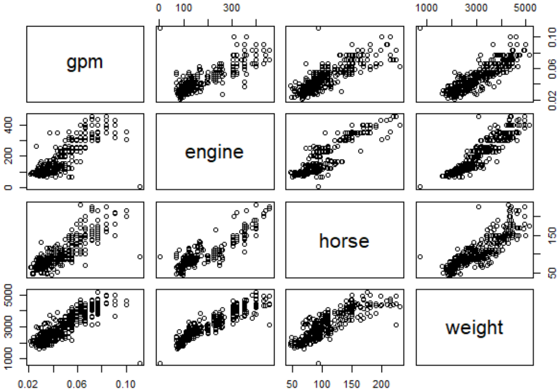


The "Residuals vs Fitted" plot shows curvature which violates the assumption of independent noise terms (we see increasing variance as well).

R example

Now consider the transformed response

$$gpm := \frac{1}{mpg}.$$

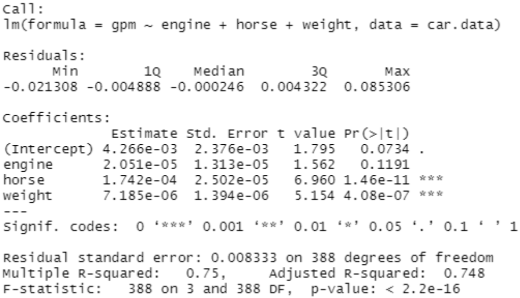


We now have linear relationships between *gpm* and the predictors.

R example – model 1

Now we fit a regression to the population model

$$GPM = \beta_0 + \beta_e engine + \beta_h horse + \beta_w weight + \epsilon.$$



```
Call:
lm(formula = gpm ~ engine + horse + weight, data = car.data)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.021308 -0.004888 -0.000246  0.004322  0.085306 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.266e-03  2.376e-03   1.795   0.0734 .  
engine      2.051e-05  1.313e-05   1.562   0.1191    
horse       1.742e-04  2.502e-05   6.960 1.46e-11 ***
weight      7.185e-06  1.394e-06   5.154 4.08e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.008333 on 388 degrees of freedom
Multiple R-squared:   0.75,     Adjusted R-squared:  0.748 
F-statistic:   388 on 3 and 388 DF,  p-value: < 2.2e-16
```

The fitted model is

$$\begin{aligned}\widehat{gpm} = {} & 4.266 \times 10^{-3} + 2.051 \times 10^{-5} engine \\ & + 1.742 \times 10^{-4} horse + 7.185 \times 10^{-6} weight.\end{aligned}$$

R example – model 1

To obtain the fitted model in terms of *mpg* we use

$$\widehat{mpg} = \frac{1}{\widehat{gpm}}.$$

From the output we see that the regression is significant at $\alpha = 0.05$ significance level (reject F-test hypothesis $\beta_e = \beta_h = \beta_w = 0$).

The model explains 75% of the variation in *gpm* about its mean ($R^2 = 0.75$).

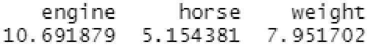
At $\alpha = 0.05$ significance level we see that the predictor

- *engine* is insignificant (retain T-test hypothesis $\beta_e = 0$)
- *horse* is significant (reject T-test hypothesis $\beta_h = 0$)
- *weight* is significant (reject T-test hypothesis $\beta_w = 0$).

Note this does not mean *engine* is not a useful predictor – it means that *engine* is insignificant in a model that also includes *horse* and *weight* as predictors.

R example – model 1

Below are the VIF statistics.



```
   engine     horse    weight 
10.691879  5.154381  7.951702 
```

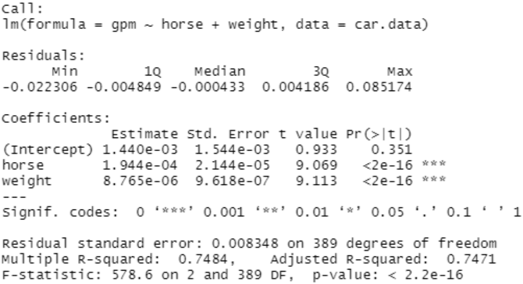
We see that collinearity problems exists as all VIFs are greater than 5.

Given *engine* has the highest variance inflation factor of $VIF_e = 10.69$ (and is also insignificant), this predictor should be removed and the model re-fitted.

R example – model 2

The next iteration is to fit a regression to the population model

$$GPM = \beta_0 + \beta_h horse + \beta_w weight + \epsilon.$$



```
Call:
lm(formula = gpm ~ horse + weight, data = car.data)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.022306 -0.004849 -0.000433  0.004186  0.085174 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.440e-03  1.544e-03   0.933    0.351    
horse       1.944e-04  2.144e-05   9.069   <2e-16 ***
weight      8.765e-06  9.618e-07   9.113   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.008348 on 389 degrees of freedom
Multiple R-squared:  0.7484,    Adjusted R-squared:  0.7471 
F-statistic: 578.6 on 2 and 389 DF,  p-value: < 2.2e-16
```

The fitted model is

$$\widehat{gpm} = 1.440 \times 10^{-3} + 1.944 \times 10^{-4} horse + 8.765 \times 10^{-6} weight.$$

R example – model 2

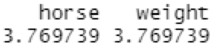
From the output we see that the regression is significant at $\alpha = 0.05$ significance level (reject F-test hypothesis $\beta_h = \beta_w = 0$).

The model explains 74.84% of the variation in *gpm* about its mean ($R^2 = 0.7484$).

At $\alpha = 0.05$ significance level we see that the predictor

- *horse* is significant (reject T-test hypothesis $\beta_h = 0$)
- *weight* is significant (reject T-test hypothesis $\beta_w = 0$).

Below are the VIF statistics.

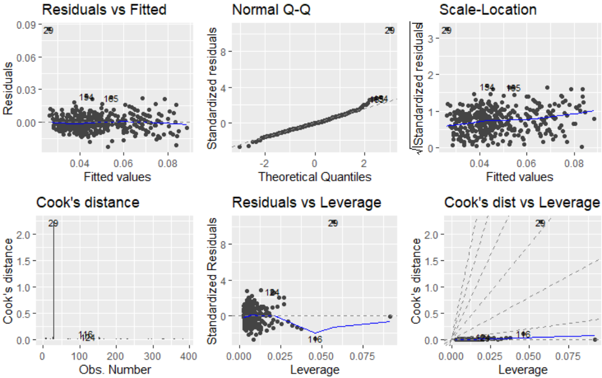


```
   horse   weight
3.769739 3.769739
```

We see there is no problem with multicollinearity ($VIF_h, VIF_w < 5.$)

R example – model 2

Now we check the diagnostic plots.

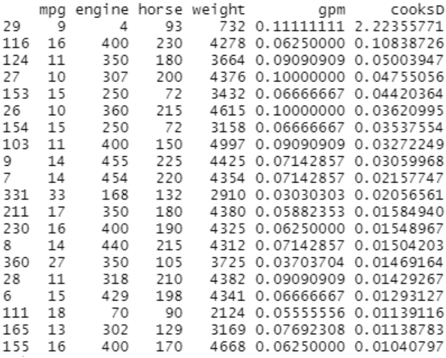


There is an obvious outlier which suggests we look for influential points.

R example – model 2

Data points that are potentially influential can be identified by looking for those with Cook's D in excess of

$$D_{\text{critical}} = \frac{4}{392 - 2 - 1} = 0.0103.$$

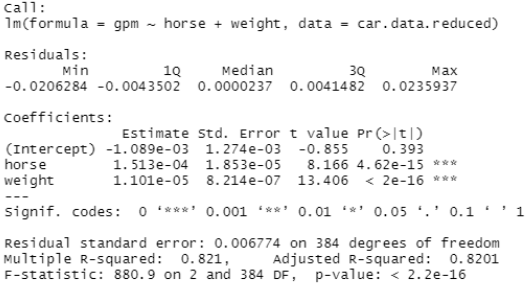


```
    mpg engine horse weight        gpm     cooksD
29    9      4    93    732 0.11111111 2.22355771
116  16    400   230   4278 0.06250000 0.10838726
124  11    350   180   3664 0.09090909 0.05003947
27   10    307   200   4376 0.10000000 0.04755056
153  15    250    72   3432 0.06666667 0.04420364
26   10    360   215   4615 0.10000000 0.03620995
154  15    250    72   3158 0.06666667 0.03537554
103  11    400   150   4997 0.09090909 0.03272249
9    14    455   225   4425 0.07142857 0.03059968
7    14    454   220   4354 0.07142857 0.02157747
331  33    168   132   2910 0.03030303 0.02056561
211  17    350   180   4380 0.05882353 0.01584940
230  16    400   190   4325 0.06250000 0.01548967
8    14    440   215   4312 0.07142857 0.01504203
360  27    350   105   3725 0.03703704 0.01469164
28   11    318   210   4382 0.09090909 0.01429267
6    15    429   198   4341 0.06666667 0.01293127
111  18     70    90   2124 0.05555556 0.01139116
165  13    302   129   3169 0.07692308 0.01138783
155  16    400   170   4668 0.06250000 0.01040797
```

We see quite a few data points that are potentially influential.

R example – model 2 (reduced dataset)

Finally we remove the five data points with Cook's D in excess of 0.04.



```
Call:
lm(formula = gpm ~ horse + weight, data = car.data.reduced)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0206284 -0.0043502  0.0000237  0.0041482  0.0235937 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.089e-03  1.274e-03  -0.855    0.393    
horse        1.513e-04  1.853e-05   8.166 4.62e-15 ***
weight       1.101e-05  8.214e-07  13.406  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.006774 on 384 degrees of freedom
Multiple R-squared:  0.821,	Adjusted R-squared:  0.8201 
F-statistic: 880.9 on 2 and 384 DF,  p-value: < 2.2e-16
```

The fitted model is

$$\widehat{gpm} = -1.089 \times 10^{-3} + 1.513 \times 10^{-4} horse + 1.101 \times 10^{-5} weight.$$

The large changes in the parameters estimates confirms the removed points as influential.

R example – model 2 (reduced dataset)

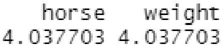
From the output we see that the regression is significant at $\alpha = 0.05$ significance level (reject F-test hypothesis $\beta_h = \beta_w = 0$).

The model explains 82.1% of the variation in *gpm* about its mean ($R^2 = 0.821$).

At $\alpha = 0.05$ significance level we see that the predictor

- *horse* is significant (reject T-test hypothesis $\beta_h = 0$)
- *weight* is significant (reject T-test hypothesis $\beta_w = 0$).

Below are the VIF statistics.

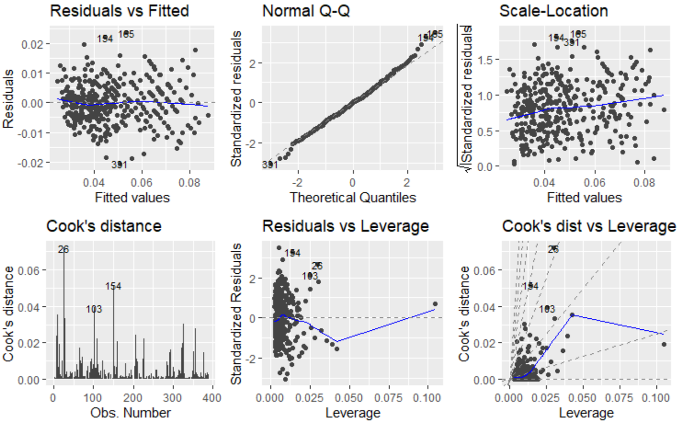


```
   horse   weight 
4.037703 4.037703 
```

We see there is no problem with multicollinearity ($VIF_h, VIF_w < 5$.)

R example – model 2 (reduced dataset)

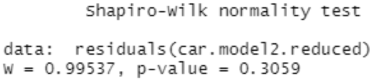
Finally we check the assumptions.



The are no obvious problems with the modelling assumptions (the diagonal strips in the "Residuals vs Fitted" plot are due to repeated values in the sample data).

R example – model 2 (reduced dataset)

We can perform a normality test on the residuals

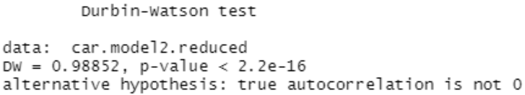


```
        Shapiro-Wilk normality test

data:  residuals(car.model2.reduced)
W = 0.99537, p-value = 0.3059
```

There are no problems with the normality assumption (retain hypothesis that the residuals are normally distributed).

To finish we check for autocorrelation.



```
        Durbin-Watson test

data:  car.model2.reduced
DW = 0.98852, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is not 0
```

Although not apparent from inspection of residuals plots, there is some statistical evidence of positive autocorrelation ($p < 0.05$, $dw < 1$).

References I