# Regression and Linear Models (37252)

# Lecture 4 - Multiple Linear Regression I

Lecturer: Joanna Wang
Notes adopted from Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

# Acknowledgement

These notes are based, in part, on earlier versions prepared by Dr Ed Lidums and Prof. James Brown.

# 4. Lecture outline

Topics:

- multidimensional least squares
  - scalar form
  - matrix form (*)
- model assumptions and statistical properties
- model fit
  - T-test
  - ANOVA and F-test
  - ANOVA, $R^2$ and $R^2_{\text{adj}}$
  - leverage and influence (*)
  - collinearity
  - collinearity and variance inflation factors
- R example

# 4. Multidimensional least squares – scalar form

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

This will allow us to consider problems that don't appear linear in nature and where transformations of the data do not assist (more on this next week).

*Sections marked (*) require knowledge of linear algebra. These sections are not assessable but are included for those interested.*

# 4. Multidimensional least squares – scalar form

The multiple regression problem is to approximate

$$Y|x_1, \ldots, x_m = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \epsilon. \tag{1}$$

This is the equation of plane in $\mathbb{R}^{m+1}$ disturbed by some RV $\epsilon$, making $Y$ a RV also.

*We could be more general and describe $\epsilon$ as representing the components of $Y$ not captured by the plane.*

We fit our model to the expectation

$$\begin{aligned} \mathbb{E}[Y|x_1, \ldots, x_m] &= \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \mathbb{E}[\epsilon] \\ &= \beta_0 + \sum_{j=1}^{m} \beta_j x_j \end{aligned} \tag{2}$$

assuming that $\mathbb{E}[\epsilon] = 0$ and $\text{var}(\epsilon) = \sigma^2$ for all $x_1, \ldots, x_m$.

# 4. Multidimensional least squares – scalar form

Our task is to find the estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ of $\beta_0, \beta_1, \ldots, \beta_m$ defining the model

$$\begin{aligned} \hat{Y}|x_1, \ldots, x_m &:= \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_j \\ &\approx \mathbb{E}[Y|x_1, \ldots, x_m]. \end{aligned} \tag{3}$$

A model that is optimum should be defined by parameter estimates such that the error

$$\hat{\epsilon} := Y - \hat{Y} = Y - \hat{\beta}_0 - \sum_{j=1}^{m} \hat{\beta}_j x_j$$

is minimised in some way, where the error $\hat{\epsilon}$ is our model's estimate of the RV

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

that we have assumed is built into underlying population model.

# 4. Multidimensional least squares – scalar form

As for simple linear regression, we choose as parameter estimates the values minimising the **sum squared error** over the sample data $(X_{i,1}, \ldots, X_{i,m}, Y_i)$; i.e.

$$\min_{(\beta_0,\ldots,\beta_m)} SSE(\beta_0, \ldots, \beta_m) = \min_{(\beta_0,\ldots,\beta_m)} \sum_{i=1}^{n} \left( Y_i - \beta_0 - \sum_{j=1}^{m} \beta_j X_{i,j} \right)^2$$

$$= \sum_{i=1}^{n} \left( Y_i - \hat{\beta}_0 - \sum_{j=1}^{m} \hat{\beta}_j X_{i,j} \right)^2. \tag{4}$$

Assuming that the minimum of $SSE$ exists, we can find $\hat{\beta}_0, \ldots, \hat{\beta}_m$ by differentiating $SSE$ with respect to each of these parameters, setting the derivatives to zero and solving the resulting system of $m + 1$ equations.

That is, the parameters values $\hat{\beta}_0, \ldots, \hat{\beta}_m$ will satisfy

$$\frac{\partial}{\partial \hat{\beta}_j} SSE(\beta_0, \ldots, \beta_j, \ldots, \beta_m) = 0 \tag{5}$$

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

# 4. Multidimensional least squares – scalar form

Differentiating and setting to zero gives the **least squares equations**

$$\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \sum_{k=1}^{m}\hat{\beta}_k X_{i,k}\right) = 0 \tag{6}$$

and

$$\sum_{i=1}^{n} X_{i,j}\left(Y_i - \hat{\beta}_0 - \sum_{k=1}^{m}\hat{\beta}_k X_{i,k}\right) = 0, \quad j \in \{1,\ldots,m\}. \tag{7}$$

The resulting regression plane

$$\hat{Y}|x_1,\ldots,x_m = \hat{\beta}_0 + \sum_{j=1}^{m}\hat{\beta}_j x_j$$

is that which minimises the *SSE* associated with the fitted points

$$\hat{Y}_i|X_{i,1},\ldots,X_{i,m} = \hat{\beta}_0 + \sum_{j=1}^{m}\hat{\beta}_j X_{i,j}. \tag{8}$$

# 4. Multidimensional least squares – matrix form (*)

Working with scalar notation in a multidimensional setting quickly becomes tedious, as anyone who has solved (6)-(7) in their current form would attest.

There is an alternative using vectors and matrices that provides more compact notation and simpler algebraic manipulation.

Arrange the sample data as

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \text{ and } \boldsymbol{X} = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,m} \end{pmatrix}$$

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

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

# 4. Multidimensional least squares – matrix form (*)

In matrix form the model we are trying to fit is

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}] = \boldsymbol{X}\beta. \tag{9}$$

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

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

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

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{10}$$

*Inverse exists if the columns of* $(\boldsymbol{X}^T\boldsymbol{X})$ *are linearly independent. If not, remove superfluous independent variable(s) and re-express.*

# 4. Multidimensional least squares – matrix form (*)

Then the fitted data can be written as

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta} \tag{11}$$

which is the matrix form of (8).

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

$$\hat{\boldsymbol{Y}} = \text{proj}_{\text{col}\,\boldsymbol{X}}\boldsymbol{Y}. \tag{12}$$

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

# 4. Model assumptions and statistical properties

In fitting the least squares model we have not made any assumptions, but to proceed and develop tools to analyse the model we need the following:

1. the underlying population model is given

$$Y|x_1, \ldots, x_m = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \epsilon,$$

   which we have already assumed in setting up our model;
2. the RV $\epsilon$ in the population model has $\mathbb{E}[\epsilon] = 0$ and $\text{var}(\epsilon) = \sigma^2$ for all $x_1, \ldots, x_m$;
3. the sample data errors $\epsilon_i \sim \text{N}(0, \sigma^2)$ and independent so that $Y_i|X_{i,1}, \ldots X_{i,m} \sim \text{N}(\beta_0 + \sum_{j=1}^{m} \beta_j X_{i,j}, \sigma^2)$ and independent.

An unbiased estimate of $\sigma^2$ is given by

$$S^2 = \frac{SSE(\hat{\beta}_0, \ldots, \hat{\beta}_m)}{n-m-1} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-m-1} = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-m-1}. \qquad (13)$$

# 4. Model fit – 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 fitted regression model $\hat{Y}$ given by (3).

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 $\beta_j$, $j \in \{0, 1, \ldots, m\}$, using the RV

$$T_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}}, \tag{14}$$

which 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, so we omit.

# 4. Model fit – T-test

**T-test hypotheses**

The null hypothesis is

$$H_0\colon\ \beta_j = \beta_j^*$$

with $\beta_j^*$ some hypothesised level of $\beta_j$ we hope to exclude.

The alternative hypothesis can be any of

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

**Test statistic**

The test statistic $t^*_{\hat{\beta}_j}$ is the value that the RV $T_{\hat{\beta}_j}$, given in (14), takes for the particular model.

# 4. Model fit – T-test

**Rejection of null hypothesis**

$H_0$ is rejected in favour of $H_A$ at **significance level** $0 < \alpha < 1/2$ if

$$t^*_{\hat{\beta}_j} > t_{1-\alpha} \text{ (**upper-tail test**)}$$

$$|t^*_{\hat{\beta}_j}| > t_{1-\alpha/2} \text{ (**two-tail test**)}$$

$$t^*_{\hat{\beta}_j} < t_{\alpha} \text{ (**lower-tail test**)}$$

where $t_\theta$ is the **quantile** satisfying

$$\text{Prob}(T_{\hat{\beta}_j} > t_\theta) = \theta.$$

# 4. Model fit – T-test

**Rejection of null hypothesis continued**

Equivalently, $H_0$ is rejected if

$$p < \alpha$$

where the **p-value**

$$p = \text{Prob}(T_{\hat{\beta}_j} > t^*_{\hat{\beta}_j}) \text{ (**upper-tail test**)}$$
$$p = 2 \times \text{Prob}(T_{\hat{\beta}_j} > |t^*_{\hat{\beta}_j}|) \text{ (**two-tail test**)}$$
$$p = \text{Prob}(T_{\hat{\beta}_j} < t^*_{\hat{\beta}_j}) \text{ (**lower-tail test**).}$$

The null hypothesis $H_0$ is **retained** in any other case.

**Interpretation of special case**

When the null hypothesis $\beta_j = 0$ is rejected, we can say the **predictor $x_j$ is statistically-significant** (at significance level $\alpha$).

# 4. Model fit – ANOVA and F-test

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

The F-test via ANOVA requires the **decomposition of total variation**

$$SST = SSR + SSE \tag{15}$$

where, letting $\mathbf{1}$ be square matrix of ones, the total sum of squares

$$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \boldsymbol{Y}^T\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^T\mathbf{1}\boldsymbol{Y}, \tag{16}$$

the total sum of squares due to the 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 of squared errors

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\hat{\epsilon}_i^2 = \hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}. \tag{17}$$

# 4. Model fit – ANOVA and F-test

Define the **mean square regression**

$$MSR = \frac{SSR}{m} \tag{18}$$

and the **mean square error**

$$MSE = \frac{SSE}{n - m - 1} \tag{19}$$

where the denominators are the **degrees of freedom**, the number of components necessary to calculate the sums $SSR$ and $SSE$.

Under the assumption that the residuals $\hat{\epsilon}_i \sim \mathrm{N}(0, \sigma^2)$ and independent, if

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

then the RV

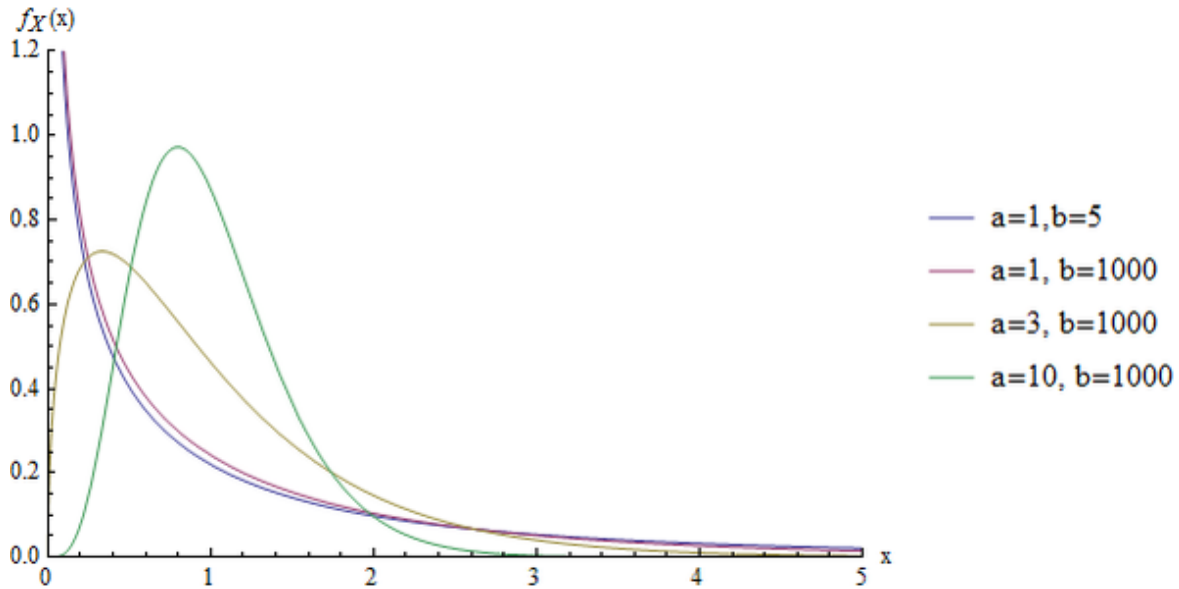
$$F = \frac{MSR}{MSE} \tag{20}$$

follows an $\mathrm{F}(\mathrm{m}, n - m - 1)$ distribution.

*$F$ is the quotient of two chi-squared RVs with $m$ and $n - m - 1$ degrees of freedom respectively.*

# 4. Model fit – ANOVA and F-test

The PDF for an F-distributed RV for a variety of parameter value is displayed below.




Example: PDF of $X \sim \mathrm{F}(a, b)$ RV

The most extreme events occur in the upper tail, so we use upper tail hypothesis tests.

# 4. Model fit – ANOVA and F-test

**F-test hypotheses**

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

**Test statistic**

The test statistic $f^*$ is the value that the RV $F$, given in (20), takes for the particular model.

**Rejection of null hypothesis**

$H_0$ is rejected in favour of $H_A$ at **significance level** $\alpha$ if

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

where $f_{1-\alpha}$ is the **quantile** satisfying

$$\text{Prob}(F > f_{1-\alpha}) = \alpha.$$

# 4. Model fit – ANOVA and F-test

**Rejection of null hypothesis continued**

Equivalently, $H_0$ is rejected if the **p-value**

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

The null hypothesis $H_0$ is **retained** in any other case.

The argument used here is that if it is so unlikely for an $F(m, n-m-1)$-distributed RV $F$ to take values at least as large as $f^*$, then perhaps $F$ cannot be so distributed.

But we know it must be if $\beta_1 = \cdots \beta_m = 0$ and so reject this hypothesis in favour of the alternative that at least one $\beta_j \neq 0$ for $j \in \{1, \ldots, m\}$.

**Interpretation**

When the null hypothesis is rejected, we know at least one parameter estimate is non-zero and can say the **regression model $\hat{\boldsymbol{Y}}$ is statistically-significant** (significance level $\alpha$).

# 4. Model fit – ANOVA, $R^2$ and $R^2_{\text{adj}}$

The decomposition of total variation in (15) can be used, as in simple regression, to calculate $R^2$ as

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

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

In the context of multiple regression, i.e. with $m \geq 2$, this statistic is deficient in that an increase in its value can be due to an increase in the number or parameters employed and not necessarily to an improvement in the fit of the model.

For multiple regression, a modified version of $R^2$ is employed that accounts for this phenomenon. This measure is called **adjusted** $R^2$ and is calculated as

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-m-1}. \tag{21}$$

Unlike $R^2$, $R^2_{\text{adj}}$ can be used to compare models fitted to different data sets. Note that $R^2_{\text{adj}}$ may be negative.

# 4. Model fit – leverage and influence (*)

Now that we have the necessary machinery we can take a closer look at leverage, residuals and influence.

To really see what is going on here requires the linear algebra perspective.

Using (10) in (11) 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**

$$\boldsymbol{h} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T. \tag{22}$$

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

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

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

# 4. Model fit – leverage and influence (*)

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

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

where the sample variance $S^2$ is given by (13).

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

$$\hat{t}_i = \frac{\hat{\epsilon}_i}{\sqrt{\text{var}(\hat{\epsilon}_i)}} = \frac{\hat{\epsilon}_i}{S\sqrt{1 - h_{ii}}} \sim \text{T}(n - m - 1)$$

Based on our model assumptions, $\hat{t}_i$ follows a Student's T distribution with $n - m - 1$ degrees of freedom. When we perform residual analysis we look for behaviour indicating departure from this assumption.

# 4. Model fit – leverage and influence (*)

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

$$\hat{d}_i = \frac{\hat{\epsilon}_i}{S^{(i)}\sqrt{1-h_{ii}}} \sim \mathrm{T}(n-m-2)$$

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

Based on our model assumptions, $\hat{d}_i$ follows a Student's T distribution with $n-m-2$ degrees of freedom. We look for departures from this assumption when performing residual analysis.

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 Lecture 3 Notes. Essentially, if points of influence are identified they should be removed, the model re-run and analysed in comparison to the original.

# 4. Model fit – collinearity

When adding an extra independent variable to a regression model we run the risk that this variable is related to independent 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 parameter as the *predicted increase in the dependent variable for a unit increase in the independent variable, all other independent variables remaining unchanged*.

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

# 4. Model fit – collinearity

Consider the case where there are $m$ independent variables. The $j$-th independent variable $X_j$ 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 independent variables.

In this case, the inverse of the matrix $\boldsymbol{X}^T \boldsymbol{X}$, used in (10) to calculate $\hat{\boldsymbol{\beta}}$, 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.

# 4. Model fit - collinearity and variance inflation factors

One way to approach the issue is to use the suspect independent variable as the dependent variable in a regression against the remaining independent variables, looking for $R^2$ to be very close to one.

If we label the value of $R^2$ returned by this regression model as $R_j^2$, we can define the **variance inflation factor (VIF)** of the $j$-independent variable as

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

As a rule of thumb, the $j$-th independent variable should be considered as potentially collinear if $VIF_j > 5$ and collinear if $VIF_j > 10$.

# 4. R example

Consider the data set cars.csv (available on Canvas) and the task of developing a regression model with fuel consumption as the dependent variable, measured for scaling reasons as *gallons* per one hundred miles (scaling the data also scales the model parameter estimates).

Potential independent variables include

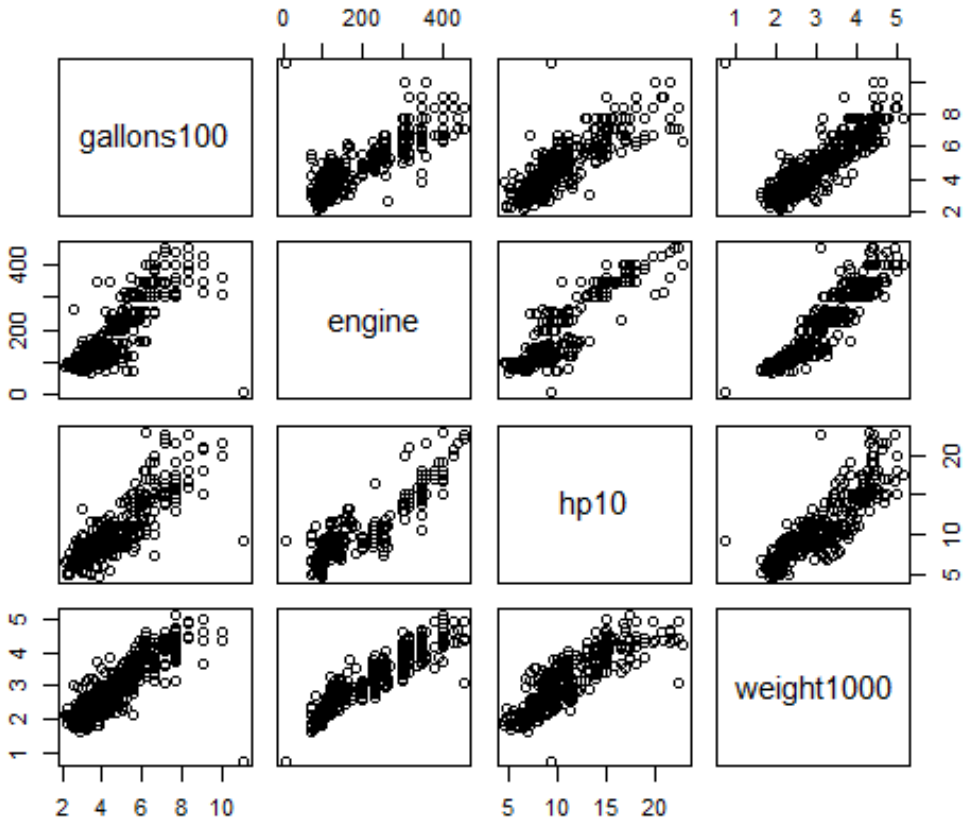
- car *weight* measured in thousands of lbs
- engine *displacement*
- engine *horsepower* measured in tens of hp.

The first step is to visually inspect the relationship between the dependent variable and each of the independent variables, checking for linear relationships that are suitable for such modelling.

If none are apparent, consider possible data transformations, examples of which were given in Lecture 2 Notes.

# 4. R example

The following scatter plots are computed.




Each dependent/independent variable scatter plot shows a linear relationship.

# 4. R example

We decide to use *horsepower* and *weight* as the dependent variables.

R returns the following Coefficients table.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12850    0.15356   0.837    0.403
hp10         0.19392    0.02133   9.093   <2e-16 ***
weight1000   0.88727    0.09567   9.275   <2e-16 ***
```

The estimated model is

$$\widehat{gallons} = 0.128 + 0.194 \times \frac{horsepower}{10} + 0.887 \times \frac{weight}{1000}$$

with T-tests showing the statistical significance of the model parameters.

# 4. R example

Summary statistics of model fit are also provided.

```
Residual standard error: 0.8304 on 389 degrees of freedom
  (15 observations deleted due to missingness)
Multiple R-squared: 0.7522, Adjusted R-squared: 0.751
F-statistic: 590.5 on 2 and 389 DF,  p-value: < 2.2e-16
```

We see that the regression model captures 75.2% of the total variation and that the F-test allows the hypothesis that the non-intercept parameters are zero to be rejected (as we also saw from the T-tests).

```
> summary(aov(mod1))
             Df Sum Sq Mean Sq F value Pr(>F)
hp10          1  755.0   755.0 1094.97 <2e-16 ***
weight1000    1   59.3    59.3   86.02 <2e-16 ***
Residuals   389  268.2     0.7
```

# 4. R example

We then consider adding *displacement* to the model but wonder about possible collinearity between this and the other independent variables – indeed, perhaps between them all. We investigate this with a correlation analysis.

```
> library('Hmisc')
> cars_cor <- as.matrix(cbind(cars$engine, cars$hp10, cars$weight1000))
> colnames(cars_cor)<-c("engine displacement", "horsepower", "vehicle
weight")
> cars_cor_res<-rcorr(cars_cor, type="pearson")
> round(cars_cor_res$r,3)
                    engine displacement horsepower vehicle weight
engine displacement               1.000      0.897          0.933
horsepower                        0.897      1.000          0.859
vehicle weight                    0.933      0.859          1.000
```

# 4. R example

```
> round(cars_cor_res$P,3)
                    engine displacement horsepower vehicle weight
engine displacement                  NA          0              0
horsepower                            0         NA              0
vehicle weight                        0          0             NA
```

We see this to be the case, with high Pearson's Correlation coefficients and very significant rejection of the zero correlation hypotheses.

# 4. R example

Although correlated, we still need to know whether this correlation is strong enough to call into question the regression model returned. We request a variance inflation analysis.

```
> mod2 <- lm(gallons100 ~ hp10 + weight1000 + engine, data =
cars)
> summary(mod2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.413388   0.236368   1.749   0.0811 .
hp10        0.173493   0.024888   6.971 1.36e-11 ***
weight1000  0.728051   0.138674   5.250 2.51e-07 ***
engine      0.002068   0.001306   1.583   0.1142
```

# 4. R example

We can also request a variance inflation analysis.

```
> vif(mod2)
      hp10 weight1000     engine
  5.154281   7.951901  10.691926
```

We see that $VIF_{displacement} > 10$, which is above the threshold indicating collinearity strong enough to disturb the model.

We also note weak collinearity for the other independent variables with $VIF_{horsepower}, VIF_{weight} > 5$.

Therefore we leave the model as is and do not enter *displacement* as an additional independent variable.

# References I

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd edition.