2.3 Hypothesis testing in Linear Models

1. Coefficient of determination

The strength of a linear relationship is usually measured through the sample correlation coefficient R .

- An R close to 1 indicates a **positive** linear relationship
- An R close to -1 indicates a **negative** linear relationship.

If we square R then this boils down to R^2 close to one giving a numerical indication of the strength of the regression.

The **residual sum of squares**

$$\mathrm{RSS} = \sum_{i=1}^{N} \varepsilon_i^2 = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

is minimised by the least squares estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, therefore

$$\widehat{\mathrm{RSS}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{y}$$

For the **minimal model**

$$Y_i = \beta_0 + \varepsilon_i,$$

$\mathbf{X}^\top \mathbf{X} = N$ and $\mathbf{X}^\top \mathbf{y} = \sum_{i=1}^{N} y_i$, then RSS is minimised for $\beta_0 = \bar{y}$, and then $\mathrm{RSS}_0 = \sum_{i=1}^{N} (y_i - \bar{y})^2$. This is the **worst possible** value for RSS. It is also sometimes called the **total sum of squares** (TSS).

If parameters are added to the model, then RSS *must decrease*. The relative amount of decrease

$$R^2 := \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} \tag{2.3.1}$$

is called the **coefficient of determination**. It is the proportion of the total variation in the data which is explained by the model.

For the **maximal model**, we will have $\mathrm{RSS} = 0$ and thus $R^2 = 1$: this means that the R^2 always

increases when more variables are added to the model, even if those variables are weakly associated with the response. The fact that adding a variable leads to *a small increase of the R^2 statistic provides evidence that the contribution of that variable is small*.

The coefficient of determination can be interpreted as the *proportion of variance explained by the model*.

In case of **one** covariate, the *coefficient of determination is the squared correlation between Y and x*. In multiple regression, it turns out that $R^2 = \mathbb{C}or(Y, \hat{Y})^2$ (the property of the least squares estimates is that they maximises the correlation among the responses and the fitted linear model among all the possible linear models).

Activity in R: Coefficient of determination

Question *Submitted Mar 16th 2023 at 9:32:20 pm*

Consider the regression output of the `basketball` data analysis one more time. What is the coefficient of determination in your analysis? What does it tell you about the model fit and how can it be interpreted?

```
basketball <- read.table("/course/data/baskball.txt", header=TRUE)
t(head(basketball))
attach(basketball)

# Perform any quick analysis here
```

asdf

Additional Activity

Question 1 *Submitted Mar 16th 2023 at 9:33:55 pm*

Residual analysis is an important part of regression. Which of the following statements are true about residuals in Linear Gaussian Models:

- [ ] the residuals are not assumed to be independent;
- [x] the variance of residuals is assumed to be constant;
- [x] standardising the residuals is recommended for better understanding of their magnitude;
- [x] residuals are assumed to be normally distributed

Question 2 *Submitted Mar 16th 2023 at 9:34:03 pm*

What is the leverage of an observation?

- ( ) Sum of squared differences between fitted values when the ith datum is removed.
- (•) The diagonal entry of the hat matrix.
- ( ) The proportion of variation in the data which is explained by the model.

2. The F-statistic in Linear Models

For the Linear Gaussian Model

$$\boldsymbol{E}[Y_i] = \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad Y_i \sim N(\mu_i, \sigma^2)$$

for independent random variables $Y_1, \ldots, Y_N$, we define the deviance to be:

$$D = \frac{1}{\sigma^2}(\mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{y}) \tag{2.3.2}$$

Now assume that we want to select between two competing models M_0 and M_1, i.e., let's consider a **null hypothesis** H_0 and an **alternative hypothesis** H_1.

$$H_0 = \boldsymbol{\beta} = \boldsymbol{\beta}_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}, \qquad H_1 = \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

The scaled deviance can be used for model comparison.

$$\begin{aligned} \Delta D = D_0 - D_1 &= \frac{1}{\sigma^2}\left[(\mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}_0^\top \mathbf{X}_0^\top \mathbf{y}) - (\mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}_1^\top \mathbf{X}_1^\top \mathbf{y})\right] \\ &= \frac{1}{\sigma^2}\left[\hat{\boldsymbol{\beta}}_1^\top \mathbf{X}_1^\top \mathbf{y} - \hat{\boldsymbol{\beta}}_0^\top \mathbf{X}_0^\top \mathbf{y}\right] \end{aligned}$$

Then we have $D_0 \sim \chi^2(N-q)$ and $D_1 \sim \chi^2(N-p)$, and thus

$$\Delta D \sim \chi^2(p-q)$$

when N is large.

If the values of ΔD *is in the critical region*, then we would *reject* H_0 *in favour of* H_1 on the grounds that model M_1 provides a significantly better description of the data.

The standard deviation σ^2, however, is unknown; we may replace it by its estimate $\hat{\sigma}^2$ results in $(2.3.2)$ being inaccurate.

But σ^2 may be eliminated by using the ratio

$$F = \frac{\frac{D_0 - D_1}{p-q}}{\frac{D_1}{N-p}} = \frac{\frac{\hat{\boldsymbol{\beta}}_1^\top \mathbf{X}_1^\top \mathbf{y} - \hat{\boldsymbol{\beta}}_0^\top \mathbf{X}_0^\top \mathbf{y}}{p-q}}{\frac{\mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}_1^\top \mathbf{X}_1^\top \mathbf{y}}{N-p}} \tag{2.3.3}$$

This may be explicitly calculated from the fitted values.

Under the null hypothesis H_0 ("the true model is represented by M_0"), against the alternative hypothesis H_1 ("the true model is represented by M_1"), the statistic F follows the $F(p-q, N-p)$ distribution. where q, p and N denote the number of parameters in M_0, M_1 and the number of observations, respectively.

Hence we reject H_0 if $F > F_\alpha(p-q, N-p)$, where α is the size of the test, typically chosen to be around 0.05, and . $F_\alpha(p-q, N-p)$ is the $1-\alpha$th quantile of the F(p-q, N-p) distribution. Alternatively we can also compute the corresponding P-value: $P(F_{(p-q,N-p)} > F)$.

The F-statistic is usually used to test the hypothesis

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_p = 0$$
$$H_1 : \text{at least one } \beta_j \text{ is non-zero}$$

and

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/(p-1)}{\mathrm{RSS}/(N-p)} \tag{2.3.4}$$

where $\mathrm{TSS} = \sum_{i=1}^{N} (\mathrm{Y}_i - \bar{\mathrm{Y}})^2$ and $\mathrm{RSS} = \sum_{i=1}^{N} (\mathrm{Y}_i - \hat{\mathrm{Y}}_i)^2$.

It is possible to show that $\mathbb{E}(\mathrm{RSS}/(N-p)) = \sigma^2$ and, under H_0, $\mathbb{E}[(\mathrm{TSS} - \mathrm{RSS})/p] = \sigma^2$. Therefore, under H_0, the F-statistic is expected to be close to 1, while under H_1, $\mathbb{E}[(\mathrm{TSS} - \mathrm{RSS})/p] > \sigma^2$ and the F-statistic is larger than 1.

There is a relationship between the F-statistics and the R^2 coefficient. We know that we can write:

$$R^2 \times \mathrm{TSS} = (\mathrm{TSS} - \mathrm{RSS}) \quad \text{and} \quad \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - R^2.$$

and from $(2.3.4)$ we have

$$
\begin{aligned}
F &= \frac{\frac{\mathrm{TSS}-\mathrm{RSS}}{p-1}}{\frac{\mathrm{RSS}}{N-p}} \\
&= \frac{R^2 \times \mathrm{TSS}}{\mathrm{RSS}} \frac{N-p}{p-1} \\
&= \frac{R^2}{1-R^2} \frac{N-p}{p-1} \sim F_{p-1,N-p}.
\end{aligned}
$$

Remark 1: The F-statistic is important because if you test for the effect of any predictor *separately* (without any correction) about 5% of the p-values will be under α (e.g. 0.05) by chance. The F-statistic does not suffer from this problem because it adjusts for the number of predictors.

Remark 2: The approach using the F-statistic works when $p < N$; however, sometimes we have a very large number of variables and $p > N$. In this case, multiple regression cannot be fitted and the F-statistic cannot be used.

Activity in R: The F statistic

Question *Submitted Mar 16th 2023 at 9:42:06 pm*

What is the F statistic value in our `basketball` data analysis? What does it tell you about the model adequacy?

```
basketball <- read.table("/course/data/baskball.txt", header=TRUE)
t(head(basketball))
attach(basketball)

# Perform any quick analysis here
```

asdf

Activity: Calculation of the F statistic using R-squared

As cheese ages, various chemical processes take place, which determines the taste of the final product. In a study of Cheddar cheese from the La Trobe Valley of Victoria, Australia, samples of cheese were analysed for their chemical composition and were subjected to taste tests. Overall taste scores were obtained by combining the scores from several tasters.

Data are measured on concentrations of various chemicals in 30 samples of mature Cheddar cheese, and a subjective measure of taste for each sample. The variables **`acetic`** and **`H2S`** are the natural logarithm of the concentration of acetic acid and hydrogen sulphide respectively. The variable **`lactic`** has not been transformed.

The multiple linear regression can be performed in R as follows:

```
library(faraway)

data("cheddar")
attach(cheddar)

cheese.lm<-lm(taste~Acetic+H2S+Lactic)
summary(cheese.lm)
```

Show how the F-statistic is calculated in R (you can use the reported value of R-squared) and use it to test whether any of the three possible predictors are significant, that is to test the following hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_1 : \text{ at least one } \beta_j \neq 0$$

Activity: Calculation of the F statistic using Deviance

The above model can be fitted using the $\mathbf{glm()}$ function, which produces the null Deviance D_0 (corresponding to the most simplified model including only the intercept) and residual Deviance D_1 for the model including the three predictor variables. Use these values to calculate the F statistic for the problem defined in the previous activity.

```
library(faraway)

data("cheddar")
attach(cheddar)

cheese.lm<-glm(taste~Acetic+H2S+Lactic, family=gaussian)
summary(cheese.lm)
```

Additional Activity

Question 1 *Submitted Mar 16th 2023 at 9:44:40 pm*

The F-statistic for Linear Gaussian Models can be calculated using:

- [x] the Deviance statistics;
- [x] the coefficient of determination;
- [ ] Cook's distances.

Question 2 *Submitted Mar 16th 2023 at 9:44:44 pm*

In the cheddar cheese example with three predictors, is there evidence that `H2S` is significant?

- (•) Yes
- ( ) No