STAT 513 fa 2020 Lec 08

Multiple linear regression

Karl B. Gregory

Multiple regression

- **Multiple regression model:** Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be fixed vectors in $\mathbb{R}^p$ and let $Y_1, \dots, Y_n$ be independent random variables such that

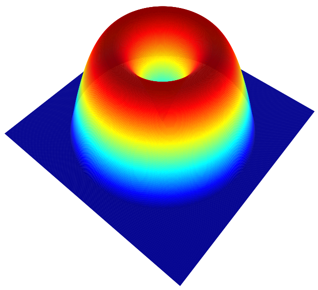
$$Y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \dots, n, \tag{1}$$

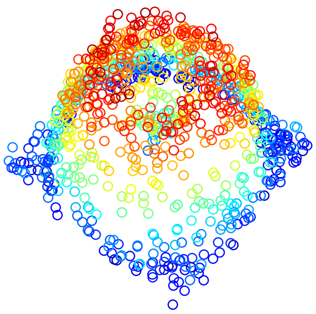
  for some function $f : \mathbb{R}^p \to \mathbb{R}$, where $\varepsilon_1, \dots, \varepsilon_n$ are independent identically distributed random variables with mean zero and variance σ^2.

- We observe a function plus noise, where the function is a "surface" in p-dimensional space. This is hard to visualize for $p > 2$, but for $p = 2$, we can depict the data we observe as points in 3-dimensional space floating above or below the surface given by the function f. For example, the plot below on the left shows the bunt-cake-like function

$$f(\mathbf{x}) = \mathbf{1}((x_1^2 + x_2^2)^{1/2} \leq 1) \cos(\pi \cdot [(x_1^2 + x_2^2)^{1/2} - 1/2])$$

  and the plot on the right plots some points (x_{i1}, x_{i2}, Y_i), for $i = 1, \dots, n$, where $Y_i = f(x_{i1}, x_{i2}) + \varepsilon_i$, $i = 1, \dots, 1000$, where $\varepsilon_1, \dots, \varepsilon_{1000}$ were generated from the $\text{Normal}(0, .1)$ distribution and the covariate values where generated as independent realizations from the $\text{Uniform}(-1, 1)$ distribution.





- We focus on the special case in which the function f is a linear combination of the covariates.
- **Multiple linear regression model:** Let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$ for $i = 1, \dots, n$ be fixed vectors in $\mathbb{R}^p$ and let $Y_1, \dots, Y_n$ be random variables such that

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, n, \tag{2}$$

  for some real numbers $\beta_0, \beta_1, \dots, \beta_p$, where $\varepsilon_1, \dots, \varepsilon_n$ are independent identically distributed random variables with mean zero and variance σ^2.

- When $p = 1$ the model in (2) is the simple linear regression model.
- Sometimes we include in the linear regression model nonlinear transformations of the covariates. For example, we might be interested in fitting the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i1}^2 + \varepsilon_i, \quad i = 1, \dots, n.$$

  This is still a linear regression model because, even though it involves some nonlinear transformations of the covariates, the parameters enter the model in a linear way. We can let

$$\begin{aligned} u_{i1} &= x_{i1} \\ u_{i2} &= x_{i2} \\ u_{i3} &= x_{i1} x_{i2} \\ u_{i4} &= x_{i1}^2 \end{aligned}$$

  and then consider the model

$$Y_i = \beta_0 + \beta_1 u_{i1} + \beta_2 u_{i2} + \beta_3 u_{i3} + \beta_4 u_{i4} + \varepsilon_i,$$

  which is equivalent.

Least-squares estimation in multiple linear regression

- We define the least-squares estimators of $\beta_0, \beta_1, \dots, \beta_p$ as

$$(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) = \underset{(\beta_0, \beta_1, \dots, \beta_p) \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]^2. \tag{3}$$

  It is very complicated to get expressions for $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ when $p > 1$; that is, unless we use matrices!

- **Matrix representation of the multiple linear regression model:** Let $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ be defined as

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

so that $\mathbf{Y}$ is an $n \times 1$ vector, $\mathbf{X}$ is an $n \times (p+1)$ matrix, $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector. Then we may express the multiple linear regression model in (2) as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{4}$$

We very often refer to the matrix $\mathbf{X}$ as the *design matrix* and to the observed values of the covariates $\mathbf{x}_1, \dots, \mathbf{x}_n$ as the *design points*. This language comes from experimental design; if a researcher designed an experiment, the covariate values $\mathbf{x}_1, \dots, \mathbf{x}_n$ might be determined by the design.

- For any vector $\mathbf{x} \in \mathbb{R}^d$, the quantity $\|\mathbf{x}\|_2 = (\mathbf{x}^T\mathbf{x})^{1/2} = (\sum_{j=1}^d x_j^2)^{1/2}$ is called the (Euclidean) norm of the vector $\mathbf{x}$, which is its length in d-dimensional Euclidean space. Its square $\|\mathbf{x}\|_2^2 = \sum_{j=1}^d x_j^2$ is the sum of the squared elements in $\mathbf{x}$.

- Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1 \dots, \hat{\beta}_p)^T$, where $\hat{\beta}_0, \hat{\beta}_1 \dots, \hat{\beta}_p$ are the least-squares estimators of $\beta_0, \beta_1, \dots, \beta_p$ defined in (3). Our matrix representation in (4) of the multiple linear regression model allows us to express the least-squares regression coefficients as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

- We can use calculus methods to find an expression for $\hat{\boldsymbol{\beta}}$. Let

$$Q_n(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

and let

$$\frac{\partial}{\partial \boldsymbol{\beta}} Q_n(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial}{\partial \beta_0} Q_n(\boldsymbol{\beta}) \\ \frac{\partial}{\partial \beta_1} Q_n(\boldsymbol{\beta}) \\ \vdots \\ \frac{\partial}{\partial \beta_p} Q_n(\boldsymbol{\beta}) \end{bmatrix}$$

be the vector of partial derivatives of $Q_n(\boldsymbol{\beta})$ with respect to $\beta_0, \beta_1, \dots, \beta_p$. The vector $\hat{\boldsymbol{\beta}}$ minimizes $Q_n(\boldsymbol{\beta})$ if and only if

$$\frac{\partial}{\partial \boldsymbol{\beta}} Q_n(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} = \mathbf{0},$$

where $\mathbf{0}$ is a $(p+1) \times 1$ vector of zeroes. Using matrix/vector calculus methods, we can get an expression in matrices for the vector of partial derivatives. We first write

$$Q_n(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}.$$

We now use the fact (see pg. 13–14 of [1]) that for $d \times 1$ vectors $\mathbf{a}$ and $\mathbf{u}$ and a $d \times d$ matrix $\mathbf{A}$ we have

$$\frac{\partial \mathbf{a}^T\mathbf{u}}{\partial \mathbf{u}} = \mathbf{a} \quad \text{and} \quad \frac{\partial \mathbf{u}^T\mathbf{A}\mathbf{u}}{\partial \mathbf{u}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{u}.$$

Applying these formulas gives

$$\frac{\partial}{\partial \boldsymbol{\beta}} Q_n(\boldsymbol{\beta}) = -2\mathbf{X}^T\mathbf{Y} + [\mathbf{X}^T\mathbf{X} + (\mathbf{X}^T\mathbf{X})^T]\boldsymbol{\beta} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta},$$

so the least-squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ must satisfy

$$-2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0} \iff \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}.$$

Provided $\mathbf{X}^T\mathbf{X}$ is invertible (non-singular), we can pre-multiply both sides of the above by $(\mathbf{X}^T\mathbf{X})^{-1}$, giving

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

- **Remark:** If $\mathbf{X}^T\mathbf{X}$ is not invertible, we cannot compute the least-squares estimator of $\boldsymbol{\beta}$. This occurs when rank of the matrix $\mathbf{X}$ is less than its number of columns, in which case we say that $\mathbf{X}$ is *rank-deficient* or that it does not have *full-column rank* (the *rank* of a matrix is the dimension of the space spanned by its columns, that is the dimension of the space containing all the points which can be reached with linear combinations of the columns). If $\mathbf{X}$ is rank-deficient, then it is possible to construct at least one of the columns of $\mathbf{X}$ with some linear combination of the other columns. This implies a kind of redundancy in the covariates. It could occur if one column of $\mathbf{X}$ contained measurements in inches, while another column contained the same measurements in centimeters; the latter column is equal to 2.54 times the former (there are 2.54 centimeters per inch), and these columns really contain the same information. In simple linear regression, with a single covariate, this occurs if the covariate takes only a single value, in which case the column of $\mathbf{X}$ containing $x_{11}, \dots, x_{n1}$ is just a multiple of the first column, which is a vector of ones. The problem of rank-deficiency always occurs if the number of columns in $\mathbf{X}$, which is $p + 1$, exceeds the sample size n. In this course, we will always assume that $p + 1$ is less than n.

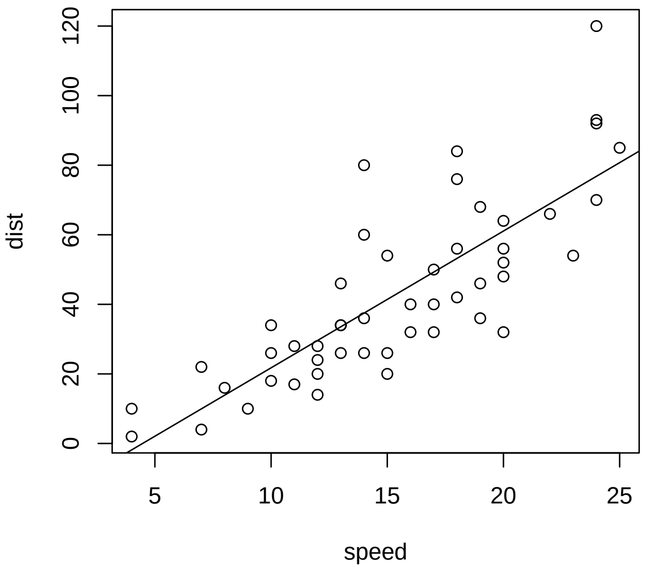
- We denote by $\hat{\boldsymbol{\varepsilon}}$ the vector $(\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)^T$ of residuals, which may be computed as

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}.$$

  We also use the notation $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ so that $\hat{\mathbf{Y}}$ is the vector of fitted values $(\hat{Y}_1, \dots, \hat{Y}_n)^T$.

- **Example:** The following code shows how to do these matrix calculations in R on a built-in R dataset called `cars`. The dataset is brought into the workspace by the command `data(cars)`. After bringing it into the workspace, type `?cars` to read more about the data set. The `cbind()` function concatenates matrices or column vectors together. The `%*%` operator performs matrix multiplication, the `solve()` function computes an inverse, and the `t()` function takes the transpose of a matrix. We will verify on these data that the previous calculations produce the same values of least-squares regression coefficients as the matrix calculations.

```
data(cars)

n <- nrow(cars)
Y <- cars$dist
x <- cars$speed

# using previous calculations:
beta1.hat <- cor(x,Y)*sd(Y)/sd(x)
beta0.hat <- mean(Y) - beta1.hat * mean(x)

# Matrices in R:
#
#           cbind() binds matrices together side-by-side
#           %*% does matrix multiplication
#           t() takes transpose
#           solve() takes inverse
#

# using matrices:
X <- cbind(rep(1,n),x) # rep(1,n) makes a vector of ones
beta.hat <- solve(t(X)%*%X)%*%t(X)%*%Y

plot(cars)
abline(beta.hat)
```



Mean and covariance matrix of least-squares estimators

- We begin this section with some general definitions and results about random vectors: A *random vector* is a vector in which each entry is a random variable.

- **Definition:** Let $\mathbf{U} = (U_1, \dots, U_d)^T$ be a $d \times 1$ random vector and let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)^T$ be the $d \times 1$ vector with entries given by $\mu_j = \mathbb{E}U_j$, for $j = 1, \dots, d$. Then $\boldsymbol{\mu}$ is called the *mean vector* (or just the *mean*) of $\mathbf{U}$, and we will use the notation $\mathbb{E}\mathbf{U} = \boldsymbol{\mu}$.

- **Definition:** Let $\mathbf{U} = (U_1, \dots, U_d)^T$ be a $d \times 1$ random vector and let $\boldsymbol{\Sigma}$ be the $d \times d$ matrix with entries given by $\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}(U_i, U_j)$, for $1 \le i, j \le d$. Then $\boldsymbol{\Sigma}$ is called the *covariance matrix* of the random vector $\mathbf{U}$, and we will use the notation $\mathrm{Cov}(\mathbf{U}) = \boldsymbol{\Sigma}$. Note that we have
$$\mathrm{Cov}(\mathbf{U}) = \mathbb{E}[(\mathbf{U} - \mathbb{E}\mathbf{U})(\mathbf{U} - \mathbb{E}\mathbf{U})^T].$$

- **Result:** Let $\mathbf{U} = (U_1, \dots, U_d)^T$ be a $d \times 1$ random vector and let $\mathbf{a} = (a_1, \dots, a_d)^T$ be a $d \times 1$ vector of real numbers. Then
$$\mathrm{Var}(\mathbf{a}^T\mathbf{U}) = \mathbf{a}^T \,\mathrm{Cov}(\mathbf{U})\mathbf{a}. \tag{5}$$
Derivation: We have
$$\begin{aligned}\mathrm{Var}(\mathbf{a}^T\mathbf{U}) &= \mathrm{Var}(\textstyle\sum_{j=1}^d a_j U_j)\\ &= \mathbb{E}(\textstyle\sum_{j=1}^d a_j U_j - \mathbb{E}\sum_{j=1}^d a_j U_j)^2\\ &= \mathbb{E}(\textstyle\sum_{j=1}^d a_j (U_j - \mathbb{E}U_j))^2\\ &= \mathbb{E}\textstyle\sum_{j=1}^d \sum_{k=1}^d a_j a_k (U_j - \mathbb{E}U_j)(U_k - \mathbb{E}U_k)\\ &= \textstyle\sum_{j=1}^d \sum_{k=1}^d a_j a_k \,\mathrm{Cov}(U_j, U_k)\\ &= \mathbf{a}^T \,\mathrm{Cov}(\mathbf{U})\mathbf{a}.\end{aligned}$$

- **Result:** Let $\mathbf{U} = (U_1, \dots, U_d)^T$ be a $d \times 1$ random vector and let $\mathbf{a} = (a_1, \dots, a_d)^T$ be a $d \times 1$ vector of real numbers and let $\mathbf{A}$ be a $d \times d$ matrix of real numbers with entries A_{ij}, $1 \le j, i \le d$. Then
$$\mathbb{E}(\mathbf{a} + \mathbf{A}\mathbf{U}) = \mathbf{a} + \mathbf{A}\mathbb{E}\mathbf{U} \tag{6}$$
$$\mathrm{Cov}(\mathbf{a} + \mathbf{A}\mathbf{U}) = \mathbf{A}\,\mathrm{Cov}(\mathbf{U})\mathbf{A}^T. \tag{7}$$
Derivation: We prove the second part:
$$\begin{aligned}\mathrm{Cov}(\mathbf{a} + \mathbf{A}\mathbf{U}) &= \mathbb{E}[(\mathbf{a} + \mathbf{A}\mathbf{U} - \mathbb{E}(\mathbf{a} + \mathbf{A}\mathbf{U}))(\mathbf{a} + \mathbf{A}\mathbf{U} - \mathbb{E}(\mathbf{a} + \mathbf{A}\mathbf{U}))^T]\\ &= \mathbb{E}[(\mathbf{A}\mathbf{U} - \mathbb{E}(\mathbf{A}\mathbf{U}))(\mathbf{A}\mathbf{U} - \mathbb{E}(\mathbf{A}\mathbf{U}))^T]\\ &= \mathbb{E}[\mathbf{A}(\mathbf{U} - \mathbb{E}\mathbf{U})(\mathbf{A}(\mathbf{U} - \mathbb{E}\mathbf{U}))^T]\\ &= \mathbb{E}[\mathbf{A}(\mathbf{U} - \mathbb{E}\mathbf{U})(\mathbf{U} - \mathbb{E}\mathbf{U})^T\mathbf{A}^T]\\ &= \mathbf{A}\mathbb{E}[(\mathbf{U} - \mathbb{E}\mathbf{U})(\mathbf{U} - \mathbb{E}\mathbf{U})^T]\mathbf{A}^T\\ &= \mathbf{A}\,\mathrm{Cov}(\mathbf{U})\mathbf{A}^T.\end{aligned}$$

- We now use the above results to show that the least-squares estimator $\hat{\boldsymbol{\beta}}$ of the vector of linear regression coefficients $\boldsymbol{\beta}$ is unbiased; that is, we will show $\mathbb{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. We also derive the covariance matrix $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ of $\hat{\boldsymbol{\beta}}$.

- **Unbiasedness of least-squares estimator:** The matrix representation of $\hat{\boldsymbol{\beta}}$ makes it very easy to show its unbiasedness. We have

  $$\mathbb{E}\hat{\boldsymbol{\beta}} = \mathbb{E}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbb{E}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}\boldsymbol{\varepsilon} = \boldsymbol{\beta},$$

  since $\mathbb{E}\boldsymbol{\varepsilon} = \mathbf{0}$.

- In order to find $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$, we first must find $\mathrm{Cov}(\boldsymbol{\varepsilon})$ and $\mathrm{Cov}(\mathbf{Y})$.

- The covariance matrix $\mathrm{Cov}(\boldsymbol{\varepsilon})$ of the vector of error terms $\boldsymbol{\varepsilon}$ is given by

  $$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n,$$

  where $\mathbf{I}_n$ is the $n \times n$ identity matrix, since for $1 \le i, j \le n$ we have

  $$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 & \text{for } i = j \\ 0 & \text{for } i \neq j, \end{cases}$$

  from the fact that $\varepsilon_1, \ldots, \varepsilon_n$ are independent.

- The covariance matrix $\mathrm{Cov}(\mathbf{Y})$ of the response vector $\mathbf{Y}$ is given by

  $$\mathrm{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}_n, \tag{8}$$

  since

  $$\mathrm{Cov}(\mathbf{Y}) = \mathrm{Cov}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n.$$

- **Covariance matrix of least-squares estimator:** We have

  $$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2. \tag{9}$$

  Derivation: From (8) and (7) we have

  $$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \mathrm{Cov}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathrm{Cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \sigma^2\mathbf{I}_n \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2. \end{aligned}$$

- **Mean and variance of estimated function at a point:** Consider a "new" point $\mathbf{x}_{\text{new}} \in \mathbb{R}^p$ at which we would like to estimate the regression function. Denote by $\tilde{\mathbf{x}}_{\text{new}}$ the vector $\tilde{\mathbf{x}}_{\text{new}} = (1, \mathbf{x}_{\text{new}}^T)^T$, so that
$$\beta_0 + \beta_1 x_{\text{new},1} + \cdots + \beta_p x_{\text{new},p} = \tilde{\mathbf{x}}_{\text{new}}^T \boldsymbol{\beta}.$$
We will estimate the value of the regression function f at the point $\mathbf{x}_{\text{new}}$ with $\tilde{\mathbf{x}}_{\text{new}}^T \hat{\boldsymbol{\beta}}$. We have
$$\mathbb{E}\tilde{\mathbf{x}}_{\text{new}}^T \hat{\boldsymbol{\beta}} = \tilde{\mathbf{x}}_{\text{new}}^T \boldsymbol{\beta} \qquad (10)$$
$$\text{Var}(\tilde{\mathbf{x}}_{\text{new}}^T \hat{\boldsymbol{\beta}}) = \tilde{\mathbf{x}}_{\text{new}}^T (\mathbf{X}^T\mathbf{X})^{-1} \tilde{\mathbf{x}}_{\text{new}} \sigma^2. \qquad (11)$$
Derivations: For the expectation we have
$$\mathbb{E}\tilde{\mathbf{x}}_{\text{new}}^T \hat{\boldsymbol{\beta}} = \tilde{\mathbf{x}}_{\text{new}}^T \mathbb{E}\hat{\boldsymbol{\beta}} = \tilde{\mathbf{x}}_{\text{new}}^T \boldsymbol{\beta},$$
since $\hat{\boldsymbol{\beta}}$ is unbiased. For the variance, we use (7) and (9).

- **Result:** For
$$\hat{\sigma}^2 = \frac{1}{n-p-1}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 = \frac{1}{n-p-1}\|\hat{\boldsymbol{\varepsilon}}\|_2^2$$
we have $\mathbb{E}\hat{\sigma}^2 = \sigma^2$. The proof of this result is beyond the scope of this course.

Inference in multiple linear regression

- In this section we will assume that the response values $Y_1, \ldots, Y_n$ are Normally distributed around the regression function. More precisely, we will assume that the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$ random variables. The Normality of the error terms leads to Normality of the least-squares regression coefficients, which enables inferential methods like hypothesis testing and the construction of confidence intervals based on Normal quantiles.

- In this section we will make use of the Multivariate Normal distribution, which we define next. Note that when we talk about the distribution of a random vector, we mean the joint distribution of the random variables comprising the random vector.

- **Multivariate Normal distribution:** The pdf of a random vector $\mathbf{U}$ having the Multivariate Normal distribution with mean vector $\boldsymbol{\mu}$ and (invertible) covariance matrix $\boldsymbol{\Sigma}$ is given by
$$f(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{u} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{u} - \boldsymbol{\mu})\right]$$
for all $\mathbf{u} \in \mathbb{R}^d$, where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. We will use $\text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the Multivariate Normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

- **Remark:** If a Multivariate Normal random vector $\mathbf{U}$ has dimension $d = 1$, then its pdf reduces to that of the "univariate" Normal distribution. If $d = 1$, then $\boldsymbol{\Sigma}$ is a 1×1 matrix, that is a scalar, which we may denote by σ^2, and $\boldsymbol{\mu}$ is a scalar, which we may denote by μ. In this case $\mathbf{U}$ has the pdf
$$f(u; \mu, \sigma^2) = (2\pi)^{-1/2}|\sigma^2|^{-1/2} \exp\left[-\frac{1}{2}\frac{(u-\mu)^2}{\sigma^2}\right]$$
for all $u \in \mathbb{R}$.

- If we assume that $\varepsilon_1, \dots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$ random variables, then the distribution of the random vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$, that is the joint distribution of $\varepsilon_1, \dots, \varepsilon_n$, is given by

$$\begin{aligned}\prod_{i=1}^{n} (2\pi)^{-1/2} \sigma^{-1} \exp\left[-\frac{1}{2}\frac{\varepsilon_i^2}{\sigma^2}\right] &= (2\pi)^{-n/2} |\sigma^2|^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{\varepsilon_i^2}{\sigma^2}\right] \\ &= (2\pi)^{-n/2} |\sigma^2 \mathbf{I}_n|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{\varepsilon}^T (\sigma^2 \mathbf{I}_n)^{-1} \boldsymbol{\varepsilon}\right],\end{aligned}$$

which we recognize this as the pdf of the $\text{Normal}(\mathbf{0}, \mathbf{I}_n\sigma^2)$ distribution. The second equality comes from the fact that for a diagonal matrix

$$\mathbf{D} = \begin{bmatrix} d_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & d_{nn} \end{bmatrix}$$

we have

$$|\mathbf{D}| = \prod_{i=1}^{n} d_{ii} \quad \text{and} \quad \mathbf{D}^{-1} = \begin{bmatrix} d_{11}^{-1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & d_{nn}^{-1} \end{bmatrix}.$$

- We will make much use of the following result concerning linear transformations of Multivariate Normal random vectors.

- **Result:** Let $\mathbf{U}$ be a $d \times 1$ random vector with the $\text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution and for some $r \times 1$ vector $\mathbf{a}$ and $r \times d$ matrix $\mathbf{A}$ let $\mathbf{V} = \mathbf{a} + \mathbf{AU}$. Then $\mathbf{V}$ is an $r \times 1$ random vector such that

$$\mathbf{V} \sim \text{Normal}(\mathbf{a} + \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T). \tag{12}$$

Proof: We will make use of multivariate moment generating functions. The mgf of the random vector $\mathbf{U}$ is given by

$$M_{\mathbf{U}}(\mathbf{t}) = \exp[\mathbf{t}^T\boldsymbol{\mu} + (1/2)\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}],$$

for all $\mathbf{t}$ in a rectangle in $\mathbb{R}^d$ that contains the origin. So the mgf of the random vector $\mathbf{V}$ is given by

$$\begin{aligned} M_{\mathbf{V}}(\mathbf{t}) &= M_{\mathbf{a}+\mathbf{AU}}(\mathbf{t}) \\ &= \mathbb{E}\exp[\mathbf{t}^T(\mathbf{a} + \mathbf{AU})] \\ &= \exp[\mathbf{t}^T\mathbf{a}]\mathbb{E}\exp[(\mathbf{A}^T\mathbf{t})^T\mathbf{U}] \\ &= \exp[\mathbf{t}^T\mathbf{a}]\exp[(\mathbf{A}^T\mathbf{t})^T\boldsymbol{\mu} + (1/2)(\mathbf{A}^T\mathbf{t})^T\boldsymbol{\Sigma}(\mathbf{A}^T\mathbf{t})] \\ &= \exp[\mathbf{t}^T(\mathbf{a} + \mathbf{A}\boldsymbol{\mu}) + (1/2)\mathbf{t}^T\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\mathbf{t}] \end{aligned}$$

which we recognize as the mgf of the $\text{Normal}(\mathbf{a} + \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$ distribution.

- From this result we see that if $\varepsilon_1, \dots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$ random variables we may write

$$\mathbf{Y} \sim \text{Normal}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n\sigma^2).$$

- We now present results about the sampling distribution of the least-squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ as well as of the unbiased estimator $\hat{\sigma}^2$ of σ^2.

- **Sampling distribution results:** If $\varepsilon_1, \dots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$ random variables, then

$$\hat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2) \tag{13}$$

$$\mathbf{a}^T\hat{\boldsymbol{\beta}} \sim \text{Normal}(\mathbf{a}^T\boldsymbol{\beta}, \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} \cdot \sigma^2) \tag{14}$$

  for any vector $\mathbf{a} \in \mathbb{R}^{p+1}$, and

$$(n-p-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p-1}. \tag{15}$$

  Results (13) and (14) follow from the result in (12). The proof of (15) is beyond the scope of this course. In addition, we have

$$\frac{\mathbf{a}^T\hat{\boldsymbol{\beta}} - \mathbf{a}^T\boldsymbol{\beta}}{\hat{\sigma}\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}} \sim t_{n-p-1}, \tag{16}$$

  which follows from (14) and (15) and from the fact that $\hat{\sigma}^2$ is independent from $\hat{\boldsymbol{\beta}}$ when $\varepsilon_1, \dots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$ random variables.

- These results can be put to use in the following ways:

- **Confidence intervals:** From (16), for any $\mathbf{a} \in \mathbb{R}^{p+1}$, a $(1-\alpha)100\%$ confidence interval for $\mathbf{a}^T\hat{\boldsymbol{\beta}}$ is given by

$$\mathbf{a}^T\hat{\boldsymbol{\beta}} \pm t_{n-p-1,\alpha/2}\hat{\sigma}\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}. \tag{17}$$

  Different choices of the vector $\mathbf{a}$ lead to confidence intervals for different quantities:

  - Choosing $\mathbf{a} = \tilde{\mathbf{x}}_{\text{new}} = (1, \mathbf{x}_{\text{new}}^T)^T$ gives a $(1-\alpha)100\%$ confidence interval for $f(x_{\text{new}}) = \tilde{\mathbf{x}}_{\text{new}}^T\boldsymbol{\beta}$:

$$\tilde{\mathbf{x}}_{\text{new}}^T\hat{\boldsymbol{\beta}} \pm t_{n-p-1,\alpha/2}\hat{\sigma}\sqrt{\tilde{\mathbf{x}}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\tilde{\mathbf{x}}_{\text{new}}}.$$

  - Define the *basis vectors* $\mathbf{e}_1, \dots, \mathbf{e}_{p+1}$ of $\mathbb{R}^{p+1}$ as

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \quad \dots \quad \mathbf{e}_{p+1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

    Then choosing $\mathbf{a} = \mathbf{e}_j$ gives a $(1-\alpha)100\%$ confidence interval for entry j of $\boldsymbol{\beta}$, which is

$$\mathbf{e}_j^T\hat{\boldsymbol{\beta}} \pm t_{n-p-1,\alpha/2}\hat{\sigma}\sqrt{\mathbf{e}_j^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{e}_j}.$$

    We may prefer to express this as

$$\hat{\boldsymbol{\beta}}_j \pm t_{n-p-1,\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\hat{\Omega}_{jj}^{1/2},$$

    where $\hat{\boldsymbol{\beta}}_j$ is entry j of the vector $\hat{\boldsymbol{\beta}}$ and $\hat{\Omega}_{jj}$ is entry (j,j) of $(n^{-1}\mathbf{X}^T\mathbf{X})^{-1}$.

- **Tests of hypotheses:** For some vector $\mathbf{a} \in \mathbb{R}^{p+1}$, consider testing hypotheses about the quantity $\mathbf{a}^T\boldsymbol{\beta}$ with respect to a null value a^*, and define

  $$T_n = \frac{\mathbf{a}^T\hat{\boldsymbol{\beta}} - a^*}{\hat{\sigma}\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}}.$$

  We have the following:

H_0	H_1	Reject H_0 at α iff	p-value
$\mathbf{a}^T\boldsymbol{\beta} \leq a^*$	$\mathbf{a}^T\boldsymbol{\beta} > a^*$	$T_n > t_{n-p-1,\alpha}$	$1 - F_{t_{n-p-1}}(T_n)$
$\mathbf{a}^T\boldsymbol{\beta} \geq a^*$	$\mathbf{a}^T\boldsymbol{\beta} < a^*$	$T_n < -t_{n-p-1,\alpha}$	$1 - F_{t_{n-p-1}}(T_n)$
$\mathbf{a}^T\boldsymbol{\beta} = a^*$	$\mathbf{a}^T\boldsymbol{\beta} \neq a^*$	$\lvert T_n\rvert > t_{n-p-1,\alpha/2}$	$2[1 - F_{t_{n-p-1}}(\lvert T_n\rvert)]$

  In the above, $F_{t_{n-p-1}}$ represents the cdf of the t_{n-p-1} distribution.

- **Example:** Consider the model

  $$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1, \dots, n,$$

  where $\varepsilon_1, \dots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$ random variables. Suppose you wish to test the hypotheses H_0: $\beta_2 = 0$ versus H_1: $\beta_2 \neq 0$. Then we perform the two-sided test from the table above with $\mathbf{a} = (0, 0, 1)^T$ and $a^* = 0$, since $(0, 0, 1)(\beta_0, \beta_1, \beta_2)^T = \beta_2$.

- **Prediction interval for new observation:** Suppose we are to observe the pair $(\mathbf{x}_{\text{new}}, Y_{\text{new}})$, for a known vector $\mathbf{x}_{\text{new}}$. We would like to construct an interval within which the as-yet-unobserved response Y_{new} will fall with probability $1 - \alpha$ for any $\alpha \in (0, 1)$.

  In order to construct such an interval, we consider the distribution of the residual $\hat{\varepsilon}_{\text{new}}$ which will results from our observing the value of Y_{new}. We have

  $$\hat{\varepsilon}_{\text{new}} = Y_{\text{new}} - \tilde{\mathbf{x}}_{\text{new}}^T\hat{\boldsymbol{\beta}},$$

  where $\tilde{\mathbf{x}}_{\text{new}} = (1, \mathbf{x}_{\text{new}}^T)^T$ as before. We may rewrite this as

  $$\begin{aligned}\hat{\varepsilon}_{\text{new}} &= Y_{\text{new}} - \tilde{\mathbf{x}}_{\text{new}}^T\boldsymbol{\beta} - \tilde{\mathbf{x}}_{\text{new}}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= \varepsilon_{\text{new}} - \tilde{\mathbf{x}}_{\text{new}}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),\end{aligned}$$

  where $\varepsilon_{\text{new}} = Y_{\text{new}} - \tilde{\mathbf{x}}_{\text{new}}^T\boldsymbol{\beta}$ is the error term corresponding to the pair $(\mathbf{x}_{\text{new}}, Y_{\text{new}})$, that is the difference between Y_{new} and the height $\tilde{\mathbf{x}}_{\text{new}}^T\boldsymbol{\beta}$ of the true regression function at $\mathbf{x}_{\text{new}}$. Since ε_{new} behaves just like the other error terms $\varepsilon_1, \dots, \varepsilon_n$, it is a $\text{Normal}(0, \sigma^2)$ random variable, and since it is independent of $\varepsilon_1, \dots, \varepsilon_n$, it is independent of $\hat{\boldsymbol{\beta}}$. From here we have

  $$\begin{aligned}\mathbb{E}\hat{\varepsilon}_{\text{new}} &= \mathbb{E}\varepsilon_{\text{new}} + \tilde{\mathbf{x}}_{\text{new}}^T\mathbb{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = 0 \\ \operatorname{Var}\hat{\varepsilon}_{\text{new}} &= \operatorname{Var}\varepsilon_{\text{new}} + \operatorname{Var}\tilde{\mathbf{x}}_{\text{new}}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sigma^2 + \tilde{\mathbf{x}}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\tilde{\mathbf{x}}_{\text{new}}\sigma^2.\end{aligned}$$

Since $\hat{\varepsilon}_{\text{new}}$ is a linear combination of Normal random variables, it also has a Normal distribution, so that

$$\hat{\varepsilon}_{\text{new}} \sim \text{Normal}(0, [1 + \tilde{\mathbf{x}}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\tilde{\mathbf{x}}_{\text{new}}]\sigma^2).$$

Combining this result with (15) gives

$$\frac{\hat{\varepsilon}_{\text{new}}}{\hat{\sigma}\sqrt{1 + \tilde{\mathbf{x}}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\tilde{\mathbf{x}}_{\text{new}}}} \sim t_{n-p-1}.$$

This allows us to write

$$P\left(-t_{n-p-1,\alpha/2} < \frac{Y_{\text{new}} - \tilde{\mathbf{x}}_{\text{new}}^T\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{1 + \tilde{\mathbf{x}}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\tilde{\mathbf{x}}_{\text{new}}}} < t_{n-p-1,\alpha/2}\right) = 1 - \alpha$$

for any $\alpha \in (0, 1)$, which is equivalent to

$$\begin{aligned} P\Big(\tilde{\mathbf{x}}_{\text{new}}^T\hat{\boldsymbol{\beta}} - t_{n-p-1,\alpha/2}\hat{\sigma}&\sqrt{1 + \tilde{\mathbf{x}}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\tilde{\mathbf{x}}_{\text{new}}} \\ &< Y_{\text{new}} < \tilde{\mathbf{x}}_{\text{new}}^T\hat{\boldsymbol{\beta}} + t_{n-p-1,\alpha/2}\hat{\sigma}\sqrt{1 + \tilde{\mathbf{x}}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\tilde{\mathbf{x}}_{\text{new}}}\Big) = 1 - \alpha. \end{aligned}$$

Therefore, if we wish to construct an interval around the fitted regression line at $\mathbf{x}_{\text{new}}$ which will contain the as-yet-unobserved Y_{new} with probability $1 - \alpha$, for any $\alpha \in (0, 1)$, the above suggests the interval given by

$$\tilde{\mathbf{x}}_{\text{new}}^T\hat{\boldsymbol{\beta}} \pm t_{n-p-1,\alpha/2}\hat{\sigma}\sqrt{1 + \tilde{\mathbf{x}}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\tilde{\mathbf{x}}_{\text{new}}}.$$

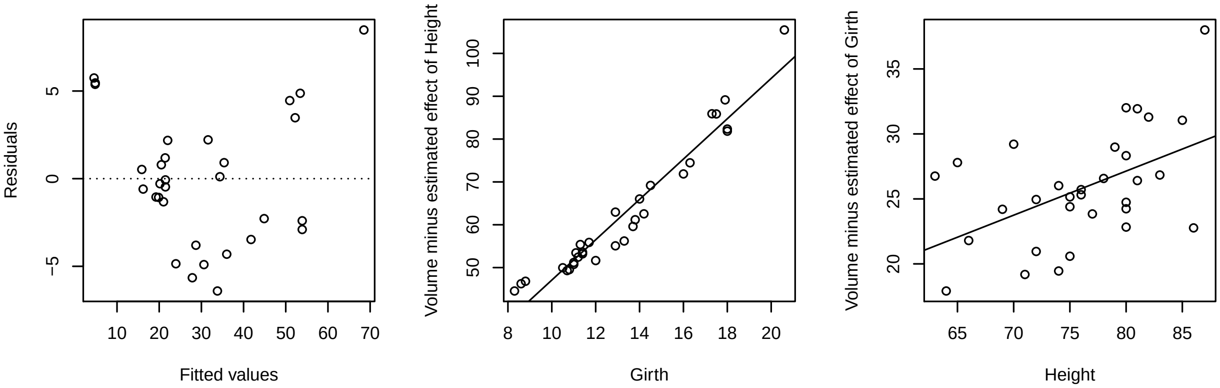
- **Exercise:** The following R code pulls into the workspace a built-in R data set called `trees`, which contains for each of $n = 31$ trees the girth, height, and volume of timber (type `?trees` into the console for more information about the data). We consider the multiple linear regression model in which the volume of timber is the response variable and girth and height are covariates; denote by $Y_1, \dots, Y_n$ the timber volumes, by $x_{11}, \dots, x_{n1}$ the girths, and by $x_{12}, \dots, x_{n2}$ the heights of the n trees.

  The R code computes the least-squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ and produces three plots. The left-most plot is a scatterplot of the residuals $\hat{\varepsilon}_{,} \dots, \hat{\varepsilon}_n$ against the fitted values $\hat{Y}_1, \dots, \hat{Y}_n$. This plot is used for diagnostic purposes—to see if the linear regression model is appropriate for the data. The second and third plots are scatterplots of the points

$$\begin{aligned} &(x_{1i}, Y_i - (\hat{\beta}_0 + \hat{\beta}_2 x_{2i})), \text{ for } i = 1, \dots, n \text{ and} \\ &(x_{2i}, Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i})), \text{ for } i = 1, \dots, n \end{aligned}$$

  with the lines given by $y = \hat{\beta}_1 x$ and $y = \hat{\beta}_2 x$, respectively, overlaid. The first of these two plots depicts the relationship between volume and girth after removing from the volumes the estimated effect of the heights. The second depicts the relationship between volume and height after removing from the volumes the estimated effect of the girths.

```
data(trees)

n <- nrow(trees) # count number of rows in data set
Y <- trees$Volume

X <- cbind( rep(1,n), trees$Girth, trees$Height )
beta.hat <- solve(t(X)%*%X) %*% t(X)%*% Y

par(mfrow=c(1,3)) # puts next three plots in a row

Y.hat <- X %*% beta.hat
e.hat <- Y - Y.hat
plot(e.hat~Y.hat,xlab="Fitted values",ylab="Residuals")
abline(h=0,lty=3)

plot( Y - X[,-2] %*% beta.hat[-2] ~ X[,2], xlab="Girth",
                        ylab="Volume minus estimated effect of Height")
abline(0,beta.hat[2])

plot(Y - X[,-3] %*% beta.hat[-3] ~ X[,3], xlab="Height",
                        ylab="Volume minus estimated effect of Girth")
abline(0,beta.hat[3])
```



Do the following:

i) Build a 99% confidence interval for β_1, the coefficient for girth.

ii) Build a 99% confidence interval for β_2, the coefficient for height.

iii) Get the p-value for testing H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ and interpret it.

iv) Get the p-value for testing H_0: $\beta_2 = 0$ versus H_1: $\beta_2 \neq 0$ and interpret it.

v) Build a 95% confidence interval for the average volume of trees which have girth equal to 15 and height equal to 70.

vi) Build a 95% prediction interval for the volume of a tree which has girth equal to 15 and height equal to 70.

Answers:

i) A 99% confidence interval for β_1 can be computed in R as follows:

```
sigma.hat <- sqrt(sum(e.hat^2)/(n - 2 - 1))
Omega.hat <- solve(t(X)%*%X / n) # Omega.hat[j,j] gives (j,j) entry

# 99% CI for Girth coefficient
tval <- qt(.995,n-2-1)
loci.Girth <- beta.hat[2] - tval * sigma.hat / sqrt(n) * sqrt(Omega.hat[2,2])
upci.Girth <- beta.hat[2] + tval * sigma.hat / sqrt(n) * sqrt(Omega.hat[2,2])
```

This gives the interval $(3.977928, 5.438393)$.

ii) Build a 99% confidence interval for β_1, the coefficient for height.

```
# 99% CI for Height coefficient
loci.Height <- beta.hat[3] - tval * sigma.hat / sqrt(n) * sqrt(Omega.hat[3,3])
upci.Height <- beta.hat[3] + tval * sigma.hat / sqrt(n) * sqrt(Omega.hat[3,3])
```

This gives the interval $(-0.02039064, 0.6988931)$.

iii) Get the p-value for testing H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ and interpret it.

```
# p-value for testing whether Girth coefficient is equal to zero:

a <- c(0,1,0) # pull beta.hat[2]
Tn <- abs(t(a) %*% beta.hat - 0)/(sigma.hat*sqrt(t(a) %*% solve(t(X)%*%X) %*% a))
Tn <- as.numeric(Tn)
pval <- 2*(1 - pt(Tn,n-2-1))
```

This gives the p-value ≈ 0.

iv) Get the p-value for testing H_0: $\beta_2 = 0$ versus H_1: $\beta_2 \neq 0$ and interpret it.

```
# p-value for testing whether Height coefficient is equal to zero:

a <- c(0,0,1) # pull beta.hat[3]
Tn <- abs(t(a) %*% beta.hat - 0)/(sigma.hat*sqrt(t(a) %*% solve(t(X)%*%X) %*% a))
Tn <- as.numeric(Tn)
pval <- 2*(1 - pt(Tn,n-2-1))
```

This gives the p-value 0.01449097.

v) Build a 95% confidence interval for the average volume of trees which have girth equal to 15 and height equal to 70.

```
# 95% conf. interval for avg Volume of trees with Girth = 15 and Height = 70

tval <- qt(.975,n-2-1)
a <- c(1,15,70)
loci.xnew <- t(a)%*%beta.hat - tval*sigma.hat/sqrt(n)*sqrt(t(a%*%Omega.hat%*%a)
upci.xnew <- t(a)%*%beta.hat + tval*sigma.hat/sqrt(n)*sqrt(t(a%*%Omega.hat%*%a)
```

This gives the confidence interval $(33.72289, 39.04178)$.

vi) Build a 95% prediction interval for the volume of a tree which has girth equal to 15 and height equal to 70.

```
# 95% pred. interval for Volume of a tree with Girth = 15 and Height = 70

a <- c(1,15,70)
lopi.xnew <- t(a)%*%beta.hat -tval*sigma.hat/sqrt(n)*sqrt(1+t(a)%*%Omega.hat%*%a)
uppi.xnew <- t(a)%*%beta.hat +tval*sigma.hat/sqrt(n)*sqrt(1+t(a)%*%Omega.hat%*%a)
```

This gives the prediction interval $(27.99782, 44.76685)$.

Power curves for tests about the slope parameter (optional)

- We will assume for this section that the response vector $\mathbf{Y}$ and each of the columns of the design matrix $\mathbf{X}$ have been centered by subtracting the mean from each entry. The effect of centering the covariates and the response is that it makes the least-squares estimate of the intercept term β_0 equal to zero, so that the intercept can be removed from the model, leaving $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ instead of $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$. Removing the intercept will allow us to retrieve β_j from the vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ as $\mathbf{e}_j^T\boldsymbol{\beta}$, whereas if we keep the intercept in the model, so that $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$, we need to retrieve β_j from $\boldsymbol{\beta}$ as $\mathbf{e}_{j+1}^T\boldsymbol{\beta}$. This would make the notation of this section cumbersome, so we assume a centered response and design, giving us the intercept-free model
$$Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, n.$$

- Consider testing H_0: $\beta_j = 0$ versus H_1: $\beta_j \neq 0$ for some $j = 1, \dots, p$ in the above model. The previous section gives the size-α test
$$\text{Reject } H_0 \text{ iff } \sqrt{n}\hat{\Omega}_{jj}^{-1/2}|\hat{\beta}_j|/\hat{\sigma} > t_{n-p,\alpha/2},$$
where $\hat{\Omega}_{jj}$ is the (j,j) element of $(n^{-1}\mathbf{X}^T\mathbf{X})^{-1}$. Note that without the intercept in the model, the degrees of freedom changes from $n-p-1$ to $n-p$.

- In order to find an expression for the power function of this test, we must use the noncentral t-distribution; we find that

  $$\sqrt{n}\hat{\Omega}_{jj}^{-1/2}\hat{\beta}_j/\hat{\sigma} \sim t_{n-p,\phi}, \qquad \text{with } \phi = \sqrt{n}\hat{\Omega}_{jj}^{-1/2}\beta_j/\sigma,$$

  where β_j is the true value of the slope coefficient for covariate j. Note that the noncentrality parameter ϕ is equal to zero when $\beta_j = 0$, that is when H_0 is true, so that the test statistic has a central t-distribution under the null. The power curve is given by

  $$\begin{aligned}\gamma(\beta_j) &= P_\beta(\sqrt{n}\hat{\Omega}_{jj}^{-1/2}|\hat{\beta}_j|/\hat{\sigma} > t_{n-p,\alpha/2})\\ &= P_\beta(\sqrt{n}\hat{\Omega}_{jj}^{-1/2}\hat{\beta}_j/\hat{\sigma} < -t_{n-p,\alpha/2}) + P_\beta(\sqrt{n}\hat{\Omega}_{jj}^{-1/2}\hat{\beta}_j/\hat{\sigma} > t_{n-p,\alpha/2})\\ &= 1 - F_{t_{n-p,\phi}}(t_{n-p,\alpha/2}) + F_{t_{n-p,\phi}}(-t_{n-p,\alpha/2}),\end{aligned}$$

  where the true value β_j of the regression coefficient is hidden in the noncentrality parameter ϕ. The power increases as ϕ moves away from 0 in either direction.

- To study what affects the power, we study the noncentrality parameter, which is a function of the true value of the regression coefficient β_j, the standard deviation σ of the error terms, and the somewhat enigmatic quantity $\hat{\Omega}_{jj}$. We find that understanding the quantity $\hat{\Omega}_{jj}$ can richly inform our discussions about power.

  It turns out that we can construct of the value of $\hat{\Omega}_{jj}$ as follows: Let $\mathbf{X}_j$ be column j of the matrix $\mathbf{X}$ and let $\mathbf{X}_{-j}$ be the matrix $\mathbf{X}$ after removing column j. Then define

  $$\hat{\boldsymbol{\gamma}}_{-j} = \underset{\boldsymbol{\gamma}_{-j}\in\mathbb{R}^p}{\text{argmin}}\|\mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\gamma}_{-j}\|_2^2,$$

  which is the least-squares estimator of the coefficients in the regression of $\mathbf{X}_j$ onto all the other columns of the design matrix $\mathbf{X}$. Then the value of $\hat{\Omega}_{jj}$ is given by

  $$\hat{\Omega}_{jj} = \left(\frac{1}{n}\|\mathbf{X}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\gamma}}_{-j}\|_2^2\right)^{-1}, \tag{18}$$

  which is one divided by the mean of the squared residuals of the least-squares regression in which the columns of $\mathbf{X}_{-1}$ are used to predict the values in $\mathbf{X}_j$. Plugging this into the expression for the non-centrality parameter, we have

  $$\phi = \sqrt{n}\hat{\Omega}_{jj}^{-1/2}\beta_j/\sigma = \|\mathbf{X}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\gamma}}_{-j}\|_2\beta_j/\sigma. \tag{19}$$

  This may not seem very informative so far, but we find that we can learn a great deal from the above expression. If the values of covariate j are highly correlated with the values of the other covariates, the residuals from regressing it onto the others will be small, so that the quantity $\|\mathbf{X}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\gamma}}_{-j}\|_2$ with be small, leading to low power (a small non-centrality parameter). On the other hand, if covariate j has very small correlations with the other covariates, $\|\mathbf{X}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\gamma}}_{-j}\|_2$ will be large, leading to high power (a large non-centrality parameter).

  On a more intuitive level, we may regard the size of $\|\mathbf{X}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\gamma}}_{-j}\|_2$ as representing the amount of *new* information contributed to the model by covariate j beyond the information contributed by the other covariates. If covariate j is highly correlated with the other covariates, then much of the information it carries is redundant information—information also carried by the other covariates. However, if covariate j is very weakly correlated with the other covariates, then most of the information it carries is unique information not possessed by the other covariates.

These observations point us toward a fundamental principal in multiple regression: When a covariate is closely related to the other covariates in the model, it is hard to distinguish its effect on the response from the effects of the other covariates; if it is less related to the other covariates in the model, its effects are easier to distinguish.

Effect of dimension on the power: Another very important observation we can make about the quantity $\|\mathbf{X}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\gamma}}_{-j}\|_2$ is that is it a strictly decreasing function of the number of covariates p in the model; every time we add a covariate to the model, this quantity *must* decrease (provided we are adding covariates which are not perfectly correlated with the ones already in the model)! This means that the power to reject H_0: $\beta_j = 0$ when it is false decreases as the total number of covariates in the model grows, regardless of the true value of β_j.

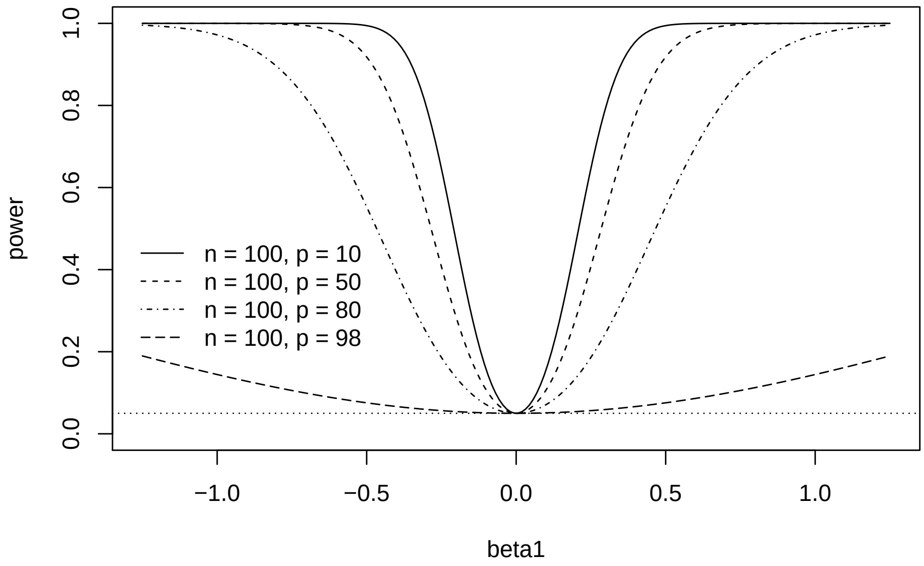
- The plot below shows, for testing H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$, the power curves of the test

$$\text{Reject } H_0 \text{ iff } \sqrt{n}\hat{\Omega}_{11}^{-1/2}|\hat{\beta}_1|/\hat{\sigma} > t_{n-p,0.025}$$

for $n = 100$ over the different total numbers of covariates $p = 10$, $p = 50$, $p = 80$, and $p = 98$ when the rows of the design matrix $\mathbf{X}$ are drawn from the $\text{Normal}(\mathbf{0}, \mathbf{I}_p)$ distribution. The power curve is given by

$$\gamma(\beta_1) = 1 - F_{t_{n-p,\phi}}(t_{n-p,0.025}) + F_{t_{n-p,\phi}}(-t_{n-p,0.025}),$$

where $\phi = \sqrt{n}\hat{\Omega}_{11}^{-1/2}\beta_1/\sigma$. Note that the value of $\hat{\Omega}_{11}$ depends on the design matrix $\mathbf{X}$, the rows of which we are generating as realizations from a multivariate Normal distribution. Therefore, under each setting, 200 datasets were generated, which resulted in 200 values of $\hat{\Omega}_{11}$, and thus 200 power curves. Each of the power curves plotted in the figure is the average of the 200 power curves based on the 200 values of $\hat{\Omega}_{11}$ under the corresponding setting. The value of σ was set equal to 1.



Note that as p increases, the power of the test over $\beta_1 \neq 0$ decreases; the decrease in power becomes very dramatic as p approaches n. This is because as more and more covariates are added to the model, the column $\mathbf{X}_1$ can be more and more accurately reconstructed using the columns of $\mathbf{X}_{-1}$, making the quantity $\|\mathbf{X}_1 - \mathbf{X}_{-1}\hat{\boldsymbol{\gamma}}_{-1}\|_2$, and thus the noncentrality parameter, smaller and smaller. In fact, as p approaches n, this quantity will approach zero, so that the test will tend towards having only trivial power (power no greater than the size) over $\beta_1 \neq 0$.

- Deriving the expression for $\hat{\Omega}_{jj}$ given in (18) takes a bit of work as well as some more advanced knowledge of linear algebra; specifically, one needs to know about projection matrices, which we do not cover in this class. We nevertheless present the details here.

  First partition the matrix $\mathbf{X}$ such that $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_{-1}]$, where $\mathbf{X}_1$ is the first column of $\mathbf{X}$ and $\mathbf{X}_{-1}$ is the matrix with the remaining columns of $\mathbf{X}$. Then we may write $\mathbf{X}^T\mathbf{X}$ as the block matrix

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_{-1} \\ \mathbf{X}_{-1}^T\mathbf{X}_1 & \mathbf{X}_{-1}^T\mathbf{X}_{-1} \end{bmatrix}.$$

  We will show

$$\hat{\Omega}_{11} = \frac{n}{\|(\mathbf{I} - \mathbf{P}_{-1})\mathbf{X}_1\|_2^2}, \tag{20}$$

  where $\mathbf{P}_{-1}$ is the projection matrix $\mathbf{P}_{-1} = \mathbf{X}_{-1}(\mathbf{X}_{-1}^T\mathbf{X}_{-1})^{-1}\mathbf{X}_{-1}^T$. The vector $(\mathbf{I} - \mathbf{P}_{-1})\mathbf{X}_1$ is the vector of residuals from regression the column $\mathbf{X}_1$ on the columns of $\mathbf{X}_{-1}$, so that our expression for $\hat{\Omega}_{11}$ in (20) matches the expression in (18). Since we can permute the columns of $\mathbf{X}$ to put any one of the columns as the first column, it is sufficient to find expression for $\hat{\Omega}_{11}$.

  We can obtain an expression for the $(1,1)$ element of the inverse of $\mathbf{X}^T\mathbf{X}$ using the following block-matrix inversion formula: We have

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{F}^{-1} & -\mathbf{F}^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{F}^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{F}^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix},$$

where $\mathbf{F} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$, provided the inverse exists. This formula gives

$$\begin{aligned}
[(\mathbf{X}^T\mathbf{X})^{-1}]_{(1,1)} &= \left(\mathbf{X}_1^T\mathbf{X}_1 - \mathbf{X}_1^T\mathbf{X}_{-1}(\mathbf{X}_{-1}^T\mathbf{X}_{-1})^{-1}\mathbf{X}_{-1}^T\mathbf{X}_1\right)^{-1} \\
&= \left(\mathbf{X}_1^T(\mathbf{I} - \mathbf{P}_{-1})\mathbf{X}_1\right)^{-1} \\
&= \left(\mathbf{X}_1^T(\mathbf{I} - \mathbf{P}_{-1})^T(\mathbf{I} - \mathbf{P}_{-1})\mathbf{X}_1\right)^{-1} \quad \text{(idempotence and symmetry of } \mathbf{I} - \mathbf{P}_{-1}) \\
&= \left([(\mathbf{I} - \mathbf{P}_{-1})\mathbf{X}_1]^T(\mathbf{I} - \mathbf{P}_{-1})\mathbf{X}_1\right)^{-1} \\
&= 1/\|(\mathbf{I} - \mathbf{P}_{-1})\mathbf{X}_1\|_2^2).
\end{aligned}$$

Now we have

$$\hat{\Omega}_{11} = [(n^{-1}\mathbf{X}^T\mathbf{X})^{-1}]_{(1,1)} = n[(\mathbf{X}^T\mathbf{X})^{-1}]_{(1,1)} = \frac{n}{\|(\mathbf{I} - \mathbf{P}_{-1})\mathbf{X}_1\|_2^2}.$$

This completes the derivation of the expression in (18) for $\hat{\Omega}_{jj}$. Letting $\mathbf{P}_{-j} = \mathbf{X}_{-j}(\mathbf{X}_{-j}^T\mathbf{X}_{-j})^{-1}\mathbf{X}_{-j}^T$, we have that $(\mathbf{I} - \mathbf{P}_{-j})\mathbf{X}_j$ is the vector of residuals from regressing the column $\mathbf{X}_j$ onto the remaining columns $\mathbf{X}_{-j}$. Then we can rewrite expression in (19) for the noncentrality parameter as

$$\phi = \sqrt{n}\hat{\Omega}_{jj}^{-1/2}\beta_j/\sigma = \frac{\beta_j}{\sigma}\|(\mathbf{I} - \mathbf{P}_{-j})\mathbf{X}_j\|_2.$$

Likelihood ratio test for “significance” of a subset of covariates

- We introduce in this section a test which is known as the full-reduced model F-test. To set things up, we first re-introduce the multiple linear regression model with Normal error terms:

- **Multivariate linear regression model with Normal errors:** Let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$ for $i = 1, \dots, n$ be fixed vectors in $\mathbb{R}^p$ and let $Y_1, \dots, Y_n$ be random variables such that

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, n, \tag{21}$$

  for some real numbers $\beta_0, \beta_1, \dots, \beta_p$, with $p+1 < n$, where $\varepsilon_1, \dots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$ random variables.

- In multivariate linear regression, it is often of interest to discover which of the covariates have an effect on the response. If a covariate has no effect on the response, then the corresponding regression coefficient will be equal to zero. In the case that only some of the covariates affect the response, several of the coefficients among $\beta_1, \dots, \beta_p$ will be equal to zero. This leads to an interest in sets of hypotheses of a certain form which we describe next.

- **Hypotheses for significance of a subset of covariates:** For some $r \in \{1, \dots, p-1\}$ we wish to test

$$H_0\text{: } \beta_{r+1} = \cdots = \beta_p = 0 \text{ versus } H_1\text{: } \beta_j \neq 0 \text{ for some } j \in \{r+1, \dots, p\}. \tag{22}$$

  If the null hypotheses is true, then all the relevant covariates are found among the covariates $1, \dots, r$ and all the covariates $r+1, \dots, p$ are irrelevant.

Note that we can always re-order the covariates, so that sets of hypotheses of this form can be used to test whether *any* subset of coefficients is equal to zero.

If we knew which covariates were irrelevant, we could ignore them. This would be advantageous, because the more covariates we include in the model, the more poorly we will estimate each one!

- **The likelihood ratio test for significance of a subset of covariates:** For any $r \in \{1, \dots, p-1\}$, we will derive the likelihood ratio test for the hypotheses in (22).

  Defining $\tilde{\mathbf{x}}_i = (1, \mathbf{x}_i^T)^T$ for $i = 1, \dots, n$ and the matrix $\mathbf{X} = [\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n]^T$ as well as the vectors $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ and $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, we may express the likelihood function as

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2; Y_1, \dots, Y_n, \mathbf{x}_1, \dots, \mathbf{x}_n) &= \prod_{i=1}^n (2\pi)^{-1/2} (\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(Y_i - \tilde{\mathbf{x}}_i^T \boldsymbol{\beta})^2\right] \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\right] \end{aligned}$$

  and the log-likelihood as

$$\ell(\boldsymbol{\beta}, \sigma^2; Y_1, \dots, Y_n, \mathbf{x}_1, \dots, \mathbf{x}_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

  Obtaining an expression for the likelihood ratio involves maximizing the likelihood over i) the null space, given by

$$\{(\boldsymbol{\beta}, \sigma^2) : \beta_j \in \mathbb{R} \text{ for } j \in \{0, 1, \dots, r\} \text{ and } \beta_j = 0 \text{ for } j \in \{r+1, \dots, p\}, \sigma^2 \geq 0\},$$

  and ii) the entire parameter space

$$\{(\boldsymbol{\beta}, \sigma^2) : \boldsymbol{\beta} \in \mathbb{R}^{p+1}, \sigma^2 \geq 0\}.$$

  Let $(\hat{\boldsymbol{\beta}}^*, \hat{\sigma}^{*2})$ be the $(\boldsymbol{\beta}, \sigma^2)$ pair which maximizes the likelihood over the null space. Then we have $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, \dots, \hat{\beta}_p^*)^T$, where

$$(\hat{\beta}_0^*, \hat{\beta}_1^*, \dots, \hat{\beta}_p^*) = \underset{\{(\beta_0, \beta_1, \dots, \beta_p) \in \mathbb{R}^{p+1} : \beta_j = 0 \text{ for } j \in \{r+1, \dots, p\}\}}{\operatorname{argmin}} \sum_{i=1}^n [Y_i - (\beta_0 + x_{i1}\beta_1 + \dots + x_{ip}\beta_p)]^2.$$

  We can easily compute $\hat{\boldsymbol{\beta}}^*$ as follows: Let $\mathbf{X}_{\mathcal{R}}$ be the matrix $\mathbf{X}$ after removing the columns associated with the covariates $r+1, \dots, p$, which are specified to be irrelevant in the null hypothesis, and define the $(r+1) \times 1$ vector $\hat{\boldsymbol{\beta}}_{\mathcal{R}}$ as

$$\hat{\boldsymbol{\beta}}_{\mathcal{R}} = (\mathbf{X}_{\mathcal{R}}^T \mathbf{X}_{\mathcal{R}})^{-1} \mathbf{X}_{\mathcal{R}}^T \mathbf{Y}.$$

  Then $\hat{\boldsymbol{\beta}}^*$ is given by

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\boldsymbol{\beta}}_{\mathcal{R}} \\ \mathbf{0} \end{bmatrix},$$

  where $\mathbf{0}$ is an $(p - r - 1) \times 1$ vector of zeroes. Then $\hat{\sigma}^{*2}$ is given by

$$\hat{\sigma}^{*2} = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*\|_2^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}\|_2^2,$$

noting that $\mathbf{X}\hat{\boldsymbol{\beta}}^* = \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}$.

The $(\boldsymbol{\beta}, \sigma^2)$ pair which maximizes the likelihood over the entire parameter space is $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_{\text{mle}})$, where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^T$, with

$$(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) = \underset{(\beta_0,\beta_1,\dots,\beta_p)\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^n [Y_i - (\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p)]^2,$$

which we recognize as the least-squares estimator of $\boldsymbol{\beta}$ and

$$\hat{\sigma}^2_{\text{mle}} = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2.$$

The likelihood ratio is thus given by

$$\begin{aligned}
\text{LR}(Y_1,\dots,Y_n,\mathbf{x}_1,\dots,\mathbf{x}_n) &= \frac{L(\hat{\boldsymbol{\beta}}^*, \hat{\sigma}^{*2}; Y_1,\dots,Y_n,\mathbf{x}_1,\dots,\mathbf{x}_n)}{L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_{\text{mle}}; Y_1,\dots,Y_n,\mathbf{x}_1,\dots,\mathbf{x}_n)} \\
&= \frac{(2\pi)^{-n/2}(\hat{\sigma}^{*2})^{-n/2}\exp\left[-\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*\|_2^2/(2\hat{\sigma}^{*2})\right]}{(2\pi)^{-n/2}(\hat{\sigma}^2_{\text{mle}})^{-n/2}\exp\left[-\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2/(2\hat{\sigma}^2_{\text{mle}})\right]} \\
&= \left[\frac{\hat{\sigma}^{*2}}{\hat{\sigma}^2_{\text{mle}}}\right]^{-n/2} \\
&= \left[\frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*\|_2^2}{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}\right]^{-n/2}.
\end{aligned}$$

With the substitution $\mathbf{X}\hat{\boldsymbol{\beta}}^* = \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}$, the likelihood ratio test is of the form

$$\text{Reject } H_0 \text{ iff } \left[\frac{\|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}\|_2^2}{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}\right]^{-n/2} < c$$

for some $c \in [0, 1]$.

- **The full-reduced model F-test:** The full-reduced model F-test is equivalent to the likelihood ratio test of the hypotheses in (22). Let

  $\text{SSE}_{\text{Red}} = \|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}\|_2^2$ be the "sum of squared errors for the reduced model", and

  $\text{SSE}_{\text{Full}} = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2$ be the "sum of squared errors for the full model".

  To give an interpretation to these quantities, SSE_{Full} is the sum of the squared residuals resulting from fitting the linear regression model with all the covariates included, i.e. the "full model", while SSE_{Red} is the sum of the squared residuals from fitting the linear regression model with only the covariates $1, \dots, r$ included, i.e., the model "reduced" to having only the first r covariates.

  The likelihood ratio test is based on the difference between SSE_{Red} and SSE_{Full}. It asks how much we lose in terms of being able to predict the response when we omit the covariates $r+1, \dots, p$. It is important to note that we always (when $\mathbf{X}$ is full-rank, which we are assuming) have

$$\text{SSE}_{\text{Red}} > \text{SSE}_{\text{Full}}\,.$$

That is, removing variables from the model will *always* increase the sum of the squared residuals, even if by a small amount. If SSE_{Red} is much larger than SSE_{Full}, then we will suspect that some important covariates have been omitted in the reduced model. If SSE_{Red} is only a little bit larger than SSE_{Full}, we will suspect that the covariates omitted in the reduced model were not important.

The full-reduced model F-test of the hypotheses

$$H_0\colon\ \beta_{r+1} = \cdots = \beta_p = 0 \text{ versus } H_1\colon\ \beta_j \neq 0 \text{ for some } j \in \{r+1, \dots, p\}$$

for any $r \in \{1, \dots, p-1\}$ is

$$\text{Reject } H_0 \text{ iff } \frac{(\text{SSE}_{\text{Red}} - \text{SSE}_{\text{Full}})/(p-r)}{\text{SSE}_{\text{Full}}\,/(n-p-1)} > F_{p-r,n-p-1,\alpha}. \tag{23}$$

- **The full-reduced model F-test as an LRT:** We now show that the full-reduced model F-test is equivalent to the size-α likelihood ratio test of the same hypotheses.

  Beginning from the rejection criterion of the likelihood ratio test, we make some manipulations to show that it is equivalent to that of the full-reduced model F-test. We have

$$\begin{aligned}
\left[\frac{\|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}\|_2^2}{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}\right]^{-n/2} &< c \\
\iff \frac{\|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}\|_2^2}{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2} &> c^{-2/n} \\
\iff \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 + \|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}\|_2^2 - \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2} &> c^{-2/n} \\
\iff \frac{\|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}\|_2^2 - \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2} &> c^{-2/n} - 1 \\
\iff \frac{(\|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}\|_2^2 - \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2)/(p-r)}{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2/(n-p-1)} &> (c^{-2/n} - 1)(n-p-1)/(p-r).
\end{aligned}$$

  It turns out (we will not prove these results in this course) that if H_0 is true, we have

$$\begin{aligned}
\|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}\|_2^2/\sigma^2 &\sim \chi^2_{n-r-1} && (24)\\
\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2/\sigma^2 &\sim \chi^2_{n-p-1} && (25)\\
(\|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}\|_2^2 - \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2)/\sigma^2 &\sim \chi^2_{p-r}, && (26)
\end{aligned}$$

  and that the quantities in (25) and (26) are independent. It follows that

$$\frac{(\|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}\|_2^2 - \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2)/(p-r)}{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2/(n-p-1)} \sim F_{p-r,n-p-1}. \tag{27}$$

  We can use this result to calibrate the rejection region of the likelihood ratio test so that it has a desired size; for any $\alpha \in (0,1)$, the size-α likelihood ratio test of the hypotheses in (22) is

$$\text{Reject } H_0 \text{ iff } \frac{(\|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\boldsymbol{\beta}}_{\mathcal{R}}\|_2^2 - \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2)/(p-r)}{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2/(n-p-1)} > F_{p-r,n-p-1,\alpha},$$

  which is the same rejection criterion as that of the full-reduced model F-test.

- **Exercise:** This exercise will use the built-in data set `swiss`, which contains for 47 French-speaking areas in Switzerland the following variables from ca. 1888: Fertility, Agriculture, Examination, Education, Catholic, and Infant Mortality. To bring the data set into the workspace, type `data(swiss)`, and type `?swiss` into the console for more information about the data. We will treat Fertility as the response variable and the other variables as covariates.

  Do the following:

  i) Give the least-squares regression coefficients for the full model.

  ii) Compute the value of SSE_{Full}.

  iii) Suppose we wish to test whether the covariates Examination and Education belong in the model, that is, whether their coefficients are equal to zero. Give the least-squares regression coefficients for the reduced model after omitting the covariates Examination and Education.

  iv) Compute the value of SSE_{Red} after omitting the covariates Examination and Education.

  v) Compute test statistic for the full-reduced model F-test for testing

  $$H_0\colon \beta_{\text{Examination}} = \beta_{\text{Education}} = 0$$

  against its alternative.

  vi) Decide whether to reject the null hypothesis at the $\alpha = 0.05$ significance level.

  vii) Compute the p-value for the full-reduced model F-test.

  Answers:

  i) The least squares regression coefficients are found with the following R code:

```
n <- nrow(swiss)
Y <- swiss$Fertility
X <- cbind( rep(1,n),
            swiss$Agriculture,
            swiss$Examination,
            swiss$Education,
            swiss$Catholic,
            swiss$Infant.Mortality)

beta.hat <- solve( t(X) %*% X ) %*% t(X) %*% Y
```

  We get

$$\begin{aligned}
\hat{\beta}_0 &= 66.9151817\\
\hat{\beta}_{\text{Agriculture}} &= -0.1721140\\
\hat{\beta}_{\text{Examination}} &= -0.2580082\\
\hat{\beta}_{\text{Education}} &= -0.8709401\\
\hat{\beta}_{\text{Catholic}} &= 0.1041153\\
\hat{\beta}_{\text{InfantMortality}} &= 1.0770481.
\end{aligned}$$

ii) The R code

```
e.hat <- Y - X %*% beta.hat
SSE.full <- sum(e.hat^2)
```

gives $\text{SSE}_{\text{Full}} = 2105.043$.

iii) The R code

```
X.red <- cbind(        rep(1,n),
                        swiss$Agriculture,
                        swiss$Catholic,
                        swiss$Infant.Mortality)

beta.hat.red <- solve( t(X.red) %*% X.red ) %*% t(X.red) %*% Y
```

gives

$$\begin{aligned}\hat{\beta}_0 &= 26.74754972\\ \hat{\beta}_{\text{Agriculture}} &= 0.14229420\\ \hat{\beta}_{\text{Catholic}} &= 0.08778473\\ \hat{\beta}_{\text{InfantMortality}} &= 1.63342374.\end{aligned}$$

As an aside: Note that some of these coefficient estimates are totally different from their values in the full model. For example, the effect of the Agriculture covariate appears to have been reversed by the removal of the Examination and Education variables from the model! This is a strong signal that we have removed something important from the model. This effect can occur when the covariates are correlated amongst themselves and when one of the removed covariates was a relevant one.

iv) The R code

```
e.hat.red <- Y - X.red %*% beta.hat.red
SSE.red <- sum(e.hat.red^2)
```

gives $\text{SSE}_{\text{Red}} = 4408.04$.

v) The R code

```
r <- 3 # Three variables left in reduced model
p <- 5 # Five variables in full model

F.test.stat <- ( (SSE.red - SSE.full) / ( p - r) ) / ( SSE.full / (n - p - 1))
```

gives the test statistic value 22.42778.

vi) The upper 0.05 quantile of the $F_{2,41}$ distribution is 3.225684, obtained from the following R code:

```
critical.val.05 <- qf(.95,p-r,n-p-1)
```

Since $22.42778 > 3.225684$ we reject the null hypothesis at the $\alpha = 0.05$ significant level, concluding that we should not remove both the variables Examination and Education from the model; one or the other or both of them should remain in the model.

vii) The R code

```
pval <- 1 - pf(F.test.stat,p-r,n-p-1)
```

gives the p-value 2.629152×10^{-07}. Therefore, we could reject the null hypothesis that Examination and Education are both irrelevant at very small significance levels. There is very strong evidence that one or the other or both should remain in the model.

Analysis of variance for linear regression:

- Analysis of variance (ANOVA) refers to decomposing the variation in a response variable into its different parts—the part accounted for by the model (in our case the linear regression model) and the part attributable to random error.

- The quantity representing the total amount of variation in a set of responses $Y_1, \dots, Y_n$ is $S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$, for which we define the new notation SST, so that

$$\text{SST} = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2.$$

  We will call this the *total sum of squares.*

- The quantity representing the amount of variation in $Y_1, \dots, Y_n$ accounted for by the model will be denoted by SSM. If a model produces the fitted values $\hat{Y}_1, \dots, \hat{Y}_n$ for the responses $Y_1, \dots, Y_n$, we define SSM as

$$\text{SSM} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2.$$

  We will call this the *model sum of squares* or the *sum of squares for the model.*

- The quantity representing the amount variation in $Y_1, \dots, Y_n$ attributable to random error is

$$\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2,$$

  We will call this the *error sum of squares* or the *sum of squared errors.*

- **Result:** In the linear regression model of (2), when the fitted values are defined by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots \hat{\beta}_p x_{ip}, \quad i = 1, \ldots, n,$$

where $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are the least-squares estimators of $\beta_0, \beta_1, \ldots, \beta_p$, we have

$$\mathrm{SST} = \mathrm{SSM} + \mathrm{SSE}\,.$$

- The decomposition of SST into the quantities SSM and SSE comes into play in a special case of the full-reduced model F-test in which we wish to test whether *any* of the covariates are relevant to the response. We sometimes call this test the *overall F-test* or the *overall test of significance* for the linear regression model.

- **Overall test of significance for the linear regression model:** Suppose we wish to test

$$H_0\colon\ \beta_1 = \cdots = \beta_p = 0 \text{ versus } H_1\colon\ \beta_j \neq 0 \text{ for at least one } j \in \{1, \ldots, p\}.$$

The null hypothesis specifies that *all* of the covariates are irrelevant. In this case the full-reduced model F-test with size α, using the notation introduced in this section, has the form

$$\text{Reject } H_0 \text{ iff } \frac{\mathrm{SSM}\,/p}{\mathrm{SSE}\,/(n-p-1)} > F_{p,n-p-1,\alpha}. \tag{28}$$

We can see this by noting firstly that SSE is the same as $\mathrm{SSE}_{\mathrm{Full}}$ and secondly, that the reduced model, which includes none of the covariates (we have $r = 0$), consists only of the intercept term β_0, for which the least-squares estimator is simply

$$\bar{Y}_n = \operatorname*{argmin}_{\beta_0} \sum_{i=1}^n (Y_i - \beta_0)^2.$$

This means that the fitted values of the reduced model are all equal to $\bar{Y}_n$, which gives

$$\mathrm{SSE}_{\mathrm{Red}} = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2 = \mathrm{SST}\,.$$

Thus we may write $\mathrm{SSE}_{\mathrm{Red}} - \mathrm{SSE}_{\mathrm{Full}} = \mathrm{SST} - \mathrm{SSE} = \mathrm{SSM}$.

- **Result:** Under the null hypothesis

$$H_0\colon\ \beta_1 = \cdots = \beta_p = 0,$$

the results stated in (24), (25), and (26) give, respectively,

$$\begin{aligned} \mathrm{SST}\,/\sigma^2 &\sim \chi^2_{n-1} \\ \mathrm{SSE}\,/\sigma^2 &\sim \chi^2_{n-p-1} \\ \mathrm{SSM}\,/\sigma^2 &\sim \chi^2_p. \end{aligned}$$

- We define two more quantities, which are sums of squares divided by the degrees of freedom of their corresponding chi-squared distributions: Let

  $$\text{MSM} = \text{SSM}\,/p \quad \text{and} \quad \text{MSE} = \text{SSE}\,/(n-p-1).$$

  This leads to an even simpler formulation of the overall F-test. The rejection criterion in (28) can now be written as

  $$\text{Reject } H_0 \text{ iff } \frac{\text{MSM}}{\text{MSE}} > F_{p,n-p-1,\alpha}.$$

  The ratio $\text{MSM}\,/\,\text{MSE}$ is sometimes called the F-statistic, being the test statistic for the overall F-test.

- With these various quantities defined, it is now time to introduce the so-called ANOVA table, which is a standard part of output produced by statistical software when linear models are fitted. The ANOVA table has the following form (though it may vary some depending on software), where $F_n = \text{MSM}\,/\,\text{MSE}$:

	df	SS	MS	F_n	p-value
Model	p	SSM	MSM	F_n	$1 - F_{F_{p,n-p-1}}(F_n)$
Error	$n-p-1$	SSE	MSE		
Total	$n-1$	SST			

  The function $F_{F_{p,n-p-1}}$ is the cdf of the $F_{p,n-p-1}$ distribution, so the p-value given in the ANOVA table is the p-value for the overall test of significance of the linear model.

- **Exercise:** For the dataset `swiss` in R that was discussed previously, compute all the values in the ANOVA table.

Answer: The following R code computes the values:

```
n <- nrow(swiss)
Y <- swiss$Fertility
X <- cbind( rep(1,n),
            swiss$Agriculture,
            swiss$Examination,
            swiss$Education,
            swiss$Catholic,
            swiss$Infant.Mortality)

beta.hat <- solve(t(X) %*% X) %*%t(X) %*% Y

Y.hat <- X %*% beta.hat

SSE <- sum((Y - Y.hat)^2)
SSM <- sum((Y.hat - mean(Y))^2)
SST <- sum((Y - mean(Y))^2)

p <- 5

MSM <- SSM / p
MSE <- SSE / ( n - p - 1 )

Fn <- MSM / MSE

pval <- 1 - pf( Fn, p, n - p - 1 )
```

We get the following values:

	df	SS	MS	F_n	p-value
Model	5	5072.912	1014.582	19.76106	5.593799×10^{-10}
Error	41	2105.043	51.34251		
Total	46	7177.955			

- **The coefficient of determination:** Another value associated with linear regression that is included as standard output by most statistical software is called the *coefficient of determination*. It is denoted by R^2 and is defined as
$$R^2 = \frac{\text{SSM}}{\text{SST}}.$$
Recall that $\text{SST} = \text{SSM} + \text{SSE}$, so R^2 is the proportion of the total variation in the responses $Y_1, \dots, Y_n$ accounted for by the model. We see that R^2 takes values in the interval $[0, 1]$. If R^2 is close to 1, the model explains a lot of the variation in the responses; if R^2 is close to zero, the model does not explain much of the variation in the response.

We do not use R^2 directly for inference; a large value of R^2 does not necessarily mean that we should reject the null hypothesis that all the regression coefficients are zero. However, we do find that we can express the test statistic F_n for the overall F-test in terms of R^2. We have

$$\begin{aligned} F_n &= \frac{\mathrm{SSM}\,/p}{\mathrm{SSE}\,/(n-p-1)} \\ &= \left(\frac{\mathrm{SSM}}{\mathrm{SST}-\mathrm{SSM}}\right)\frac{n-p-1}{p} \\ &= \left(\frac{\mathrm{SSM}\,/\,\mathrm{SST}}{1-\mathrm{SSM}\,/\,\mathrm{SST}}\right)\frac{n-p-1}{p} \\ &= \left(\frac{R^2}{1-R^2}\right)\frac{n-p-1}{p}, \end{aligned}$$

which tells us that F_n is an increasing function of R^2.

We also have that R^2 is given by the square of Pearson's correlation coefficient between the values $Y_1, \ldots, Y_n$ and the fitted values $\hat{Y}_1, \ldots, \hat{Y}_n$.

References

[1] John F Monahan. *A primer on linear models.* CRC Press, 2008.