Ordinary Least Squares (OLS) estimators: introduction

We consider the general linear regression model of the form

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \qquad \text{with } \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),\ i = 1, ..., n$$

The matrix representation of the model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \text{with } \boldsymbol{\epsilon} \overset{iid}{\sim} N(0, \sigma^2 I_n)$$

The Ordinary Least Squares (OLS) estimator for $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. Indeed, $\hat{\boldsymbol{\beta}}$ is the minimizer of the objective function

$$\min_{\boldsymbol{\beta}} \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Let us now derive the distribution of $\hat{\boldsymbol{\beta}}$.

Expectation of the OLS estimators

Given that $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, we have

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E[\mathbf{y}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

So we conclude that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$ since its expectation is equal to the parameter vector. The next step is to derive the covariance matrix of the OLS estimator.

Covariance matrix of the OLS estimators

Again, provided that $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, and noting that $var(\mathbf{y}) = \sigma^2 I_n$ where I_n is the identity matrix of order n and σ^2 is a scalar, we have

$$\begin{aligned} var(\hat{\boldsymbol{\beta}}) &= var\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T var(\mathbf{y})((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2 I_n)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned}$$

Eventually, let us note that the standard error associated to $\hat{\beta}_j$ is given by $\sigma\sqrt{\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{jj}}$, where $\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{jj}$ is a scalar on the diagonal of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$.

Sampling distribution of the OLS estimators

Recall that one of the assumptions of the linear model is that $\boldsymbol{\epsilon} \overset{iid}{\sim} N(0, \sigma^2 I_n)$. This implies that $\mathbf{y}$ is also multivariate Normally distirbuted. Indeed, we have $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$. And in turn it implies that $\hat{\boldsymbol{\beta}}$ is multivariate Normally distributed. Its sampling distribution is given by

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}\right)$$

Using this result, it is now possible to do inference in the linear model, that is for example doing hypothesis testing and deriving confidence intervals for the model parameters.

Simulation study in R (1/2)

The R code below generates synthetic data and computes the OLS estimators.

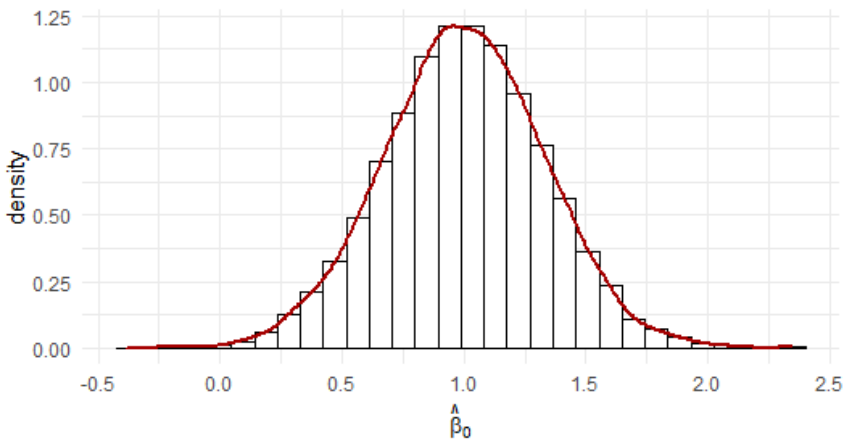
```r
# create synthetic data
set.seed(2022)
n <- 150
X <- matrix(c(rep(1,n), runif(n, min = 0, max = n)), ncol = 2, byrow = FALSE)
beta0 <- 1 ; beta1 <- 2.5 ; sigma <- 2
Beta <- matrix(c(beta0, beta1))
epsilon <- rnorm(n, sd = sigma)

# OLS estimators
y <- X %*% Beta  + epsilon
Beta.hat <- solve(t(X) %*% X) %*% t(X) %*% y
Beta.hat
#         [,1]
# [1,] 0.9785
# [2,] 2.4987

matrix(c(coef(lm(y ~ X[,-1]))[[1]], coef(lm(y ~ X[,-1]))[[2]]))
#         [,1]
# [1,] 0.9785
# [2,] 2.4987
```

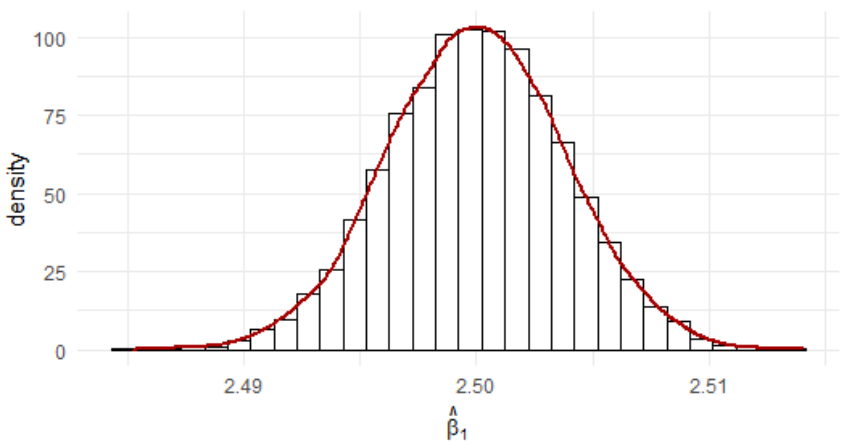
Simulation study in R (2/2)

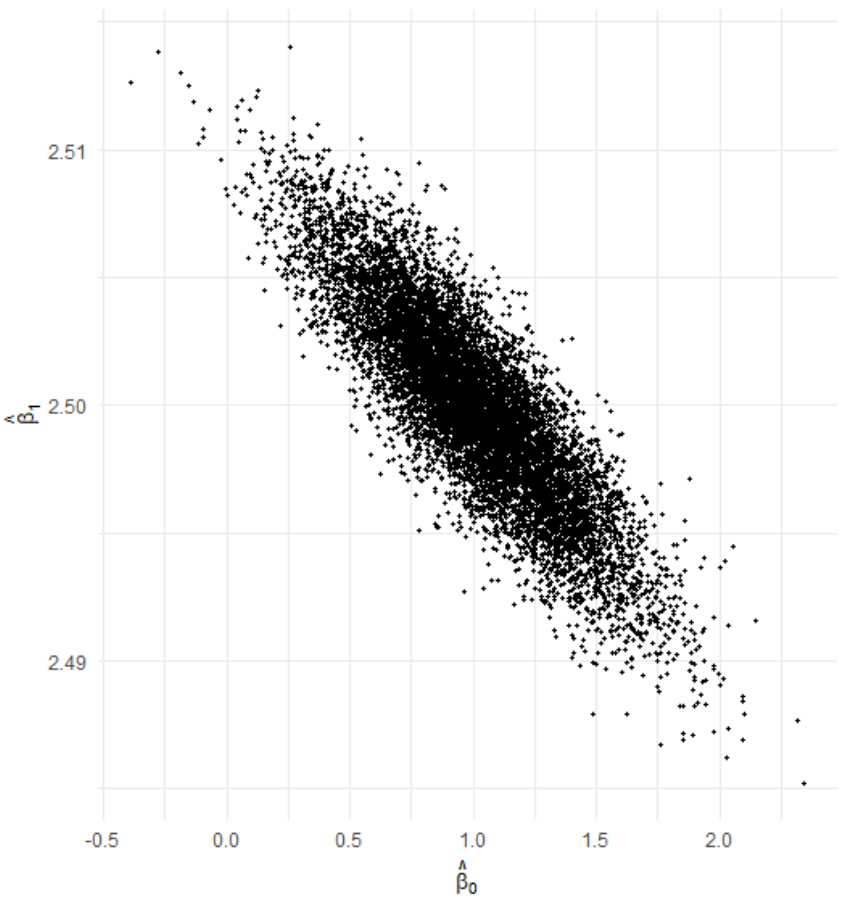
This R code simulates $10,000$ estimates for the true model parameters.

```
# Simulation study
set.seed(1986)
nsim = 10000
OLS <- matrix(rep(0, 2*nsim), ncol = 2, nrow = nsim)

for (i in 1:nsim){

  X <- matrix(c(rep(1,n), runif(n, min = 0, max = n)), ncol = 2, byrow = FALSE)
  epsilon <- rnorm(n, sd = sigma)
  y <- X %*% Beta  + epsilon
  OLS[i, ] <- t(solve(t(X) %*% X) %*% t(X) %*% y)

}

head(OLS)
#         [,1]   [,2]
# [1,] 0.5326 2.505
# [2,] 0.4589 2.506
# [3,] 0.8918 2.499
# [4,] 1.0718 2.496
# [5,] 0.9344 2.501
# [6,] 1.0746 2.502
```

Visualizing the sampling distribution of the OLS estimators







Distribution of individual coefficients

Let us now consider any individual regression coefficient $\hat{\beta}_j$. We have demonstrated in the last video that

$$\hat{\beta}_j \sim N\left(\beta_j, \sigma^2 \left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{jj}\right)$$

where $\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{jj}$ is a scalar on the diagonal of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$. It follows that

$$\hat{\beta}_j - \beta_j \sim N\left(0, \sigma^2 \left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{jj}\right)$$

and eventually we have that

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{jj}}} \sim N(0,1)$$

We note that for the quantity on the left, β_j and σ are unknown.

Unbiased estimator for the regression variance

It can be shown that $\hat{\sigma}^2 = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-(p+1)}$ is an unbiased estimator for σ^2.
It is therefore also the Mean Squared Error (MSE) of this estimator.
It corresponds to the Residual Sum of Squares (RSS) divided by the number of degrees of freedom for the model.

Now, since we estimate σ^2 by $\hat{\sigma}^2$, we can prove that

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{jj}}} \sim t_{n-(p+1)}$$

where $t_{n-(p+1)}$ is the Student distribution with $n-(p+1)$ degrees of freedom.

Confidence intervals for individual coefficients

With that in mind, it is easy to derive a $(1-\alpha)\times 100\%$ confidence interval for an individual regression coefficient $\hat{\beta}_j$. We have

$$\hat{\beta}_j \pm t_{1-\alpha/2,n-(p+1)} \underbrace{\hat{\sigma}\sqrt{\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{jj}}}_{\hat{se}(\hat{\beta}_j)}$$

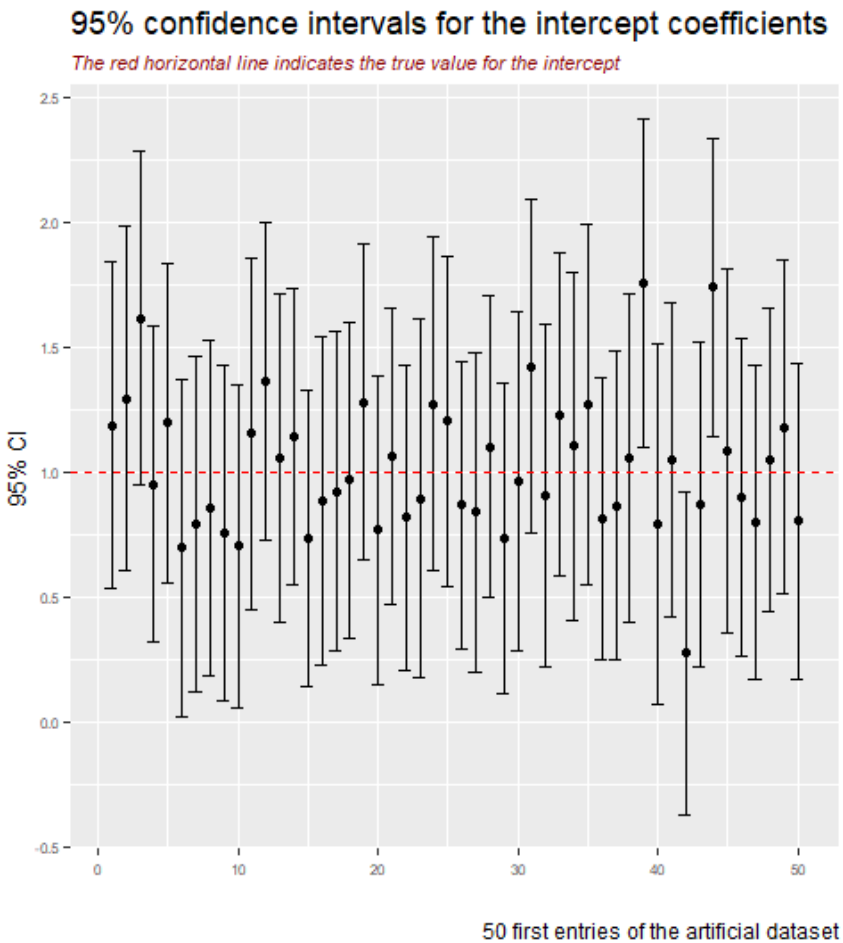
where the critical value $t_{1-\alpha/2,n-(p+1)}$ is the quantile of order $1-\alpha/2$ of a Student distribution with $n-(p+1)$ degrees of freedom and $\hat{\sigma} = \sqrt{\frac{RSS}{n-(p+1)}} = \sqrt{\frac{\boldsymbol{e}^T\boldsymbol{e}}{n-(p+1)}} = \sqrt{\frac{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{n-(p+1)}}$ is the estimated standard deviation of the model.

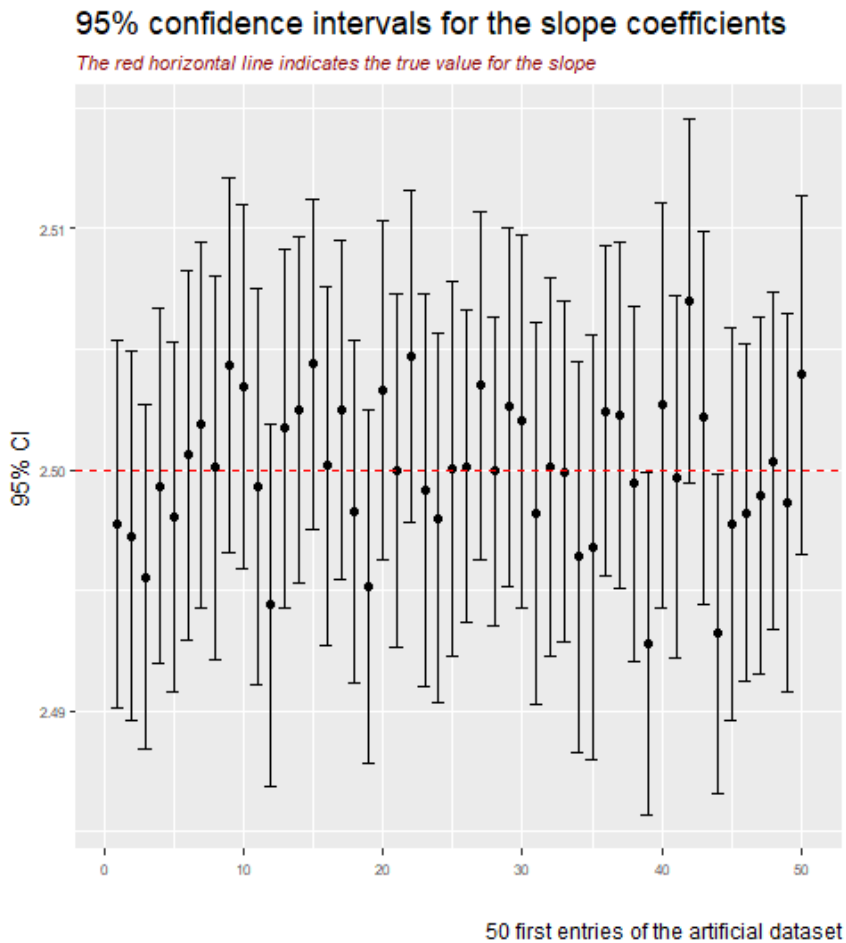
Simulation study in R

The R code below computes 95% confidence intervals for the regression coefficients.

```
# create synthetic data
set.seed(1986)
n <- 150
X <- matrix(c(rep(1,n), runif(n, min = 0, max = n)), ncol = 2, byrow = FALSE) ;
beta0 <- 1 ; beta1 <- 2.5 ; sigma <- 2
Beta <- matrix(c(beta0, beta1)) ; epsilon <- rnorm(n, sd = sigma)

# simulation study
nsim = 10000
OLS <- matrix(rep(0, 2*nsim), ncol = 2, nrow = nsim)
sehat.beta0hat = sehat.beta1hat = numeric(nsim)
lowbeta0 = lowbeta1 = highbeta0 = highbeta1 = numeric(nsim)

for (i in 1:nsim){
  X <- matrix(c(rep(1,n), runif(n, min = 0, max = n)), ncol = 2, byrow = FALSE)
  epsilon <- rnorm(n, sd = sigma)
  y <- X %*% Beta  + epsilon
  OLS[i, ] <- t(solve(t(X) %*% X) %*% t(X) %*% y)
  sehat.beta0hat[i] <- sqrt(((t(y-X %*% OLS[i, ]) %*% (y-X %*% OLS[i, ]))) /
                       (n-(2+1))) * sqrt(diag(solve(t(X) %*% X))[1])
  sehat.beta1hat[i] <- sqrt(((t(y-X %*% OLS[i, ]) %*% (y-X %*% OLS[i, ]))) /
                       (n-(2+1))) * sqrt(diag(solve(t(X) %*% X))[2])
  lowbeta0[i] <- OLS[i,1] - (qt(1-0.05/2, n-(2+1)) * sehat.beta0hat[i])
  lowbeta1[i] <- OLS[i,2] - (qt(1-0.05/2, n-(2+1)) * sehat.beta1hat[i])
  highbeta0[i] <- OLS[i,1] + (qt(1-0.05/2, n-(2+1)) * sehat.beta0hat[i])
  highbeta1[i] <- OLS[i,2] + (qt(1-0.05/2, n-(2+1)) * sehat.beta1hat[i])
}
```

Visualizing the confidence intervals for the regression coefficients





Testing for one individual coefficient

We want to test

$$\begin{cases} H_0 : \beta_j = \beta_{j0} \\ H_1 : \beta_j \neq \beta_{j0} \end{cases}$$

where β_{j0} is some arbitrary value. Under H_0, the test statistic

$$t = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{\sigma}\sqrt{\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{jj}}}$$

obeys a Student distribution with $n - (p + 1)$ degrees of freedom with p being the number of covariates or regressors in the model and 1 corresponds to the intercept. We indeed have $p + 1$ parameters to estimate. For this two-sided test setting, the null hypothesis H_0 is rejected at level α if

$$| \, t \, | > t_{1-\alpha/2, n-(p+1)}$$

The Fisher Test of overall significance

We want to test

$$\begin{cases} H_0 : \beta_0 = \beta_1 = ... = \beta_p = 0 \\ H_1 : \text{at least one } \beta_j \neq 0 \end{cases}$$

Under H_0, the test statistic

$$F = \frac{\frac{(\boldsymbol{y}-\bar{\boldsymbol{y}})^T(\boldsymbol{y}-\bar{\boldsymbol{y}})-\boldsymbol{e}^T\boldsymbol{e}}{p}}{\frac{\boldsymbol{e}^T\boldsymbol{e}}{n-(p+1)}}$$

obeys a Fisher distribution with appropriate degrees of freedom and where $\boldsymbol{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is a vector of length n of residuals and $\bar{\boldsymbol{y}}$ is a vector of length n whose elements are the sample mean of the dependent variable. R returns as output p-values associated with the test(s) for individual coefficient(s) and the Fisher Test for overall significance.

Example in R

Minimal example of linear model on the 'swiss' dataset in R programming.

```
head(swiss)
#              Fertility Agriculture Examination Education Catholic Infant.Mortality
# Courtelary        80.2        17.0          15        12     9.96             22.2
# Delemont          83.1        45.1           6         9    84.84             22.2
# Franches-Mnt      92.5        39.7           5         5    93.40             20.2

lm1 <- lm(Fertility ~ Agriculture + Education + Catholic, data = swiss)
summary(lm1)

# Residuals:
#     Min      1Q  Median      3Q     Max
# -15.178  -6.548   1.379   5.822  14.840
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 86.22502    4.73472  18.211  < 2e-16 ***
#   Agriculture -0.20304    0.07115  -2.854  0.00662 **
#   Education   -1.07215    0.15580  -6.881 1.91e-08 ***
#   Catholic     0.14520    0.03015   4.817 1.84e-05 ***
#   ---
#   Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1
#
# Residual standard error: 7.728 on 43 degrees of freedom
# Multiple R-squared:  0.6423,  Adjusted R-squared:  0.6173
# F-statistic: 25.73 on 3 and 43 DF,  p-value: 1.089e-09
```

References

Bijma, F., Jonker M., Van der Vaart, A. (2016), An Introduction to Mathematical Statistics. Amsterdam University Press. ISBN 978 94 6298 5100

The R Project for Statistical Computing:
https://www.r-project.org/

Course notes