1. Multivariate normal

Let $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ be a bivariate normal random vector with mean $\boldsymbol{\mu} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$ and covariance $\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$.

(a) Find the distribution of $X_1 - 2X_2$.

(b) Let x_2 be a real number. Find the conditional distribution of X_1 given $X_2 = x_2$.

(c) Find a constant c such that X_1 and $X_1 + cX_2$ are independent

2. Multiple linear regression

We consider the multiple linear regression model (model A)

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$, for $i = 1, \ldots, n$ and all ε-s independent.

The model summary (R output for model A) for $n = 65$ observations is given on page 2.

We also consider a reduced model, model B,

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i$$

where the covariate x_4 from model A is not included.

The model summary for the 65 observations is given on page 3.

(a) Fill in the missing values (3 question marks) in the R output for model A.

Use the t_{n-p} - distributed statistic

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\widehat{\mathrm{SE}}(\hat{\beta}_j)}$$

to derive the expression for a $(1 - \alpha) \cdot 100\%$ confidence interval for a coefficient β_j.

Calculate a 95% confidence interval for β_1 using the R output for model A.

Which model do you prefer, model A or model B? Justify your answer.

We continue with model B and consider a new point $\mathbf{x}_0$, where the first element of the vector corresponds to the intercept, then x_1, x_2, x_3. Let Y_0 denote a new observation at $\mathbf{x}_0$, independent of previous observations $Y_1, \ldots, Y_{65}$.

(b) What is the distribution of the prediction $\hat{Y}_0 = \mathbf{x}_0^T \hat{\boldsymbol{\beta}}$? Use $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1}$, where X is the 65×4 design matrix.

And what is the distribution of the prediction error $\hat{\varepsilon}_0 = Y_0 - \hat{Y}_0$?

Calculate the prediction $\hat{y}_0$ at $\mathbf{x}_0 = (1, 0, 3, 0)^T$ using the R output for model B, and also calculate a 95% prediction interval using

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 1.8 & 0.0 & -0.4 & -0.3 \\ 0.1 & 0.2 & 0.0 & 0.0 \\ -0.4 & 0.0 & 0.2 & 0.0 \\ -0.3 & 0.0 & 0.0 & 0.1 \end{bmatrix}.$$

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.2067 -2.2215 -0.1877  3.0400  9.3978 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.9807     1.8846   2.112  0.03884 *  
x1            2.4094     0.4262   5.653 4.64e-07 ***
x2            1.2718     0.4808   2.645  0.01040 *  
x3           -1.0141       ?     -3.377  0.00129 ** 
x4           -0.1014     0.3236  -0.313  0.75519    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.885 on 60 degrees of freedom
Multiple R-squared:   ? , Adjusted R-squared:  0.4597 
F-statistic: 14.61 on ? and 60 DF,  p-value: 2.172e-08
```

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3124 -2.2275 -0.2984  3.0837  9.4508 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.5642     1.3256   2.689 0.009235 ** 
x1            2.4355     0.4149   5.870 1.93e-07 ***
x2            1.2523     0.4732   2.647 0.010331 *  
x3           -1.0280     0.2948  -3.487 0.000913 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.857 on 61 degrees of freedom
Multiple R-squared:  0.4926, Adjusted R-squared:  0.4677 
F-statistic: 19.74 on 3 and 61 DF,  p-value: 4.625e-09
```

3. Partial F-test

Consider a linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \sigma^2 \text{I}_n)$ and $\boldsymbol{\beta}$ is a vector of length p. Let $\mathbf{X}_0$ denote the design matrix corresponding to only the first r columns of $\mathbf{X}$. Divide $\boldsymbol{\beta}$ into $\boldsymbol{\beta}_0$ of length r (including the intercept) and $\boldsymbol{\beta}_1$ of length $p - r$ so that the full model can be written as

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}.$$

We will test $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ against $H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}$ using a partial F-test.

The restricted model (under H_0) is

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}.$$

Define projection matrices $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$ for the full model and $\mathbf{H}_0 = \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0$ for the restricted model. The residual vector from fitting the full model is then $(\mathbf{I} - \mathbf{H})\mathbf{Y}$ and for the restricted model $(\mathbf{I} - \mathbf{H}_0)\mathbf{Y}$.

(a) A statistic for the partial F-test can be expressed in terms of the differences in error sums of squares between the full (SSE) and restricted (SSE_0) model;

$$F_1 = \frac{(\text{SSE}_0 - \text{SSE})/(p - r)}{\text{SSE}/(n - p)}.$$

Show that, when H_0 is true, $(\text{SSE}_0 - \text{SSE})/\sigma^2$ is χ^2-distributed with $p - r$ degrees of freedom.

Show that, when H_0 is true, the test statistic F_1 has an F-distribution with $(p-r, n-p)$ degrees of freedom.

(b) Another statistic for the partial F-test is constructed from the estimator $\hat{\boldsymbol{\beta}}_1$ (the last $p - r$ elements of the estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$) and its estimated covariance $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_1) = \frac{\hat{\sigma}^2}{\sigma^2}\text{Cov}(\hat{\boldsymbol{\beta}}_1)$;

$$F_2 = \frac{1}{p - r}\hat{\boldsymbol{\beta}}_1^T \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_1)^{-1}\hat{\boldsymbol{\beta}}_1.$$

Show that $F_1 = F_2$. You may use that for a null hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$, the residuals for the restricted model can be expressed as

$$(\mathbf{I} - \mathbf{H}_0)\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y} + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T(\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T)^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}).$$

4. 2-level fractional factorial designs

Consider a 2^{5-1} fractional factorial design with generator ABCD = E.

(a) In this design, main effects are aliased with 4-factor interaction effects. Which effects are the 2-factor interaction effects aliased with? What is the resolution of the design?

We will assume that all 3-factor interactions and higher are negligible. Furthermore, we will only estimate two-factor interaction effects including the factor A.

(b) Write down the regression model one would use for estimation of main effects and the relevant two-factor interactions.

Explain briefly all the components of the model and the model assumptions.

The R output from fitting the regression model with $n = 16$ observations collected from one experimental run of the 2^{5-1} fractional factorial design is given on page 6.

(c) Using a Bonferroni correction to control the FWER at 5%, which effects are significant? Do not perform a test of the intercept.

Sketch an interaction effects plot for the 2-factor interaction between factors A and E. Use the levels of A on the x-axis. Show calculations. Use your sketch to give a brief interpretation of the interaction effect.

The total sums of squares (SST) is 205.09. Use the R output to calculate the proportion of the total sums of squares that is accounted for by the main effect A in the model.

```
Call:
lm(formula = y ~ A + B + C + D + E + A * B + A * C + A * D + A * E)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.5366 -0.1178 -0.1011  0.4016  1.1023 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 29.78024    0.29019 102.623 5.77e-11 ***
A            2.03032    0.29019   6.997 0.000425 ***
B           -0.72472    0.29019  -2.497 0.046693 *  
C            0.98461    0.29019   3.393 0.014622 *  
D            1.75879    0.29019   6.061 0.000915 ***
E           -0.07614    0.29019  -0.262 0.801810    
A:B          1.02634    0.29019   3.537 0.012267 *  
A:C         -0.20709    0.29019  -0.714 0.502249    
A:D          0.03434    0.29019   0.118 0.909669    
A:E         -1.58100    0.29019  -5.448 0.001590 ** 
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.161 on 6 degrees of freedom
Multiple R-squared:  0.9606, Adjusted R-squared:  0.9015 
F-statistic: 16.25 on 9 and 6 DF,  p-value: 0.001479
```