TMA4267 Linear Statistical Models – 4 June 2016

Problem 1 Independence

a) It is known that if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $\boldsymbol{Y} = \boldsymbol{A}_{q\times p}\boldsymbol{X} \sim N_q(\boldsymbol{A}\boldsymbol{X}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}})$. We have $\boldsymbol{A} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix} 5 \\ 3 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{A}\,\mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 5 \\ 3 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$$

$$\mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}$$

Answer: $\boldsymbol{Y} \sim N_2(\begin{pmatrix} 2 \\ -2 \end{pmatrix}, \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix})$.

It is known that if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $\boldsymbol{Y} = \boldsymbol{A}_{q\times p}\boldsymbol{X}$ and $\boldsymbol{Z} = \boldsymbol{B}_{r\times p}\boldsymbol{X}$ are independent iff $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^{\mathrm{T}} = \boldsymbol{0}$.

$$\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^{\mathrm{T}} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 2 & b \\ a & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 2 & b \\ a & 1 \end{pmatrix}$$

$$\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^{\mathrm{T}} = \boldsymbol{0} \Leftrightarrow \begin{pmatrix} 2-a & b-1 \\ -2+a & -b+1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Element (1,1) and (2,1) gives the same equation for a: $2-a=0$, so $a=2$, and element (1,2) and (2,2) gives the same equation for b: $b-1=0$, so $b=1$. Thus, $(a,b)=(2,1)$ to give $\boldsymbol{Y} = \begin{pmatrix} X_1 - X_2 \\ -X_1 + X_2 \end{pmatrix}$ independent of $\boldsymbol{Z} = \begin{pmatrix} 2X_1 + 2X_2 \\ X_1 + X_2 \end{pmatrix}$.

Problem 2 Plant stress

a) **T-statistic in Intercept row:** $t_0 = \frac{\hat\beta_0 - 0}{\sqrt{\widehat{\mathrm{Var}}(\hat\beta_0)}} = \frac{16.15942}{0.04140} = 390.3$. Meaning: this is the test statistic for testing the null hypothesis $H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$.

Std.Error in row named $D:T$: in general, $t_j = \frac{\hat\beta_j - 0}{\sqrt{\widehat{\mathrm{Var}}(\hat\beta_j)}}$ so that $\sqrt{\widehat{\mathrm{Var}}(\hat\beta_j)} = \frac{\hat\beta_j}{t_j} = \frac{-0.00242}{-0.058} = 0.04$. Alternatively, we may conclude that the Std.Error for $\hat\beta_{D:T}$ is 0.04140 since we have orthogonal columns in our design matrix and the std.error is the same for all estimated regression parameters in the model. Meaning: the estimated standard deviation for the regression coefficient estimate. Mathematically we find this by looking at the diagonal element corresponding to $D:T$ of the square root of $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}s^2$, where s^2 is the estimate for the regression variance σ^2. For our orthogonal design $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is a diagonal matrix with 32 on the diagonal. We read off s from the print-out "Residual standard error=0.2342". Thus, Std.Error=$0.2342 \cdot \frac{1}{\sqrt{32}} = 0.04$.

***p*-value in row named $D:F:T$**: two tails of t-distribution with 24 degrees of freedom, observed t-statistics is 2.198. Can't find precise value, but from the table on page 4 of "Tabeller og formler i statistikk" we see that the critical value in the t-distribution with 24 degrees of freedom is 2.064 for $\alpha = 0.025$ and 2.492 for $\alpha = 0.01$. This means that the p-value must be between 0.02 and 0.05.
Meaning: Test the null hypothesis that $\beta_{D:F:T} = 0$ vs. $\beta_{D:F:T} \neq 0$, (with the other seven covariates and intercept present in the model), and produce a p-value of the test. Reject the null hypothesis if the p-value is smaller than the chosen significance level.

Multiple R-squared (also just called R^2): R^2=1-SSE/SST, so we need SSE and SST. We find SSE from s since SSE=$s^2 \cdot (n-p) = 0.2342^2 \cdot 24 = 1.32$, but SST is more difficult (not impossible, may be found from the F-statistic). But, it is easiest to find R^2 from R^2-adjusted (Adjusted R-squared), since Adjusted R-squared: $1-(1-R^2)(n-1)/(n-p) = 0.9594$ is given, and we know that $n = 32$ and $n-p = 24$. Thus, $R^2 = 1 - \frac{n-p}{n-1}(1-R^2_{\text{adj}}) = 1 - \frac{24}{31}(1-0.9594) = 0.9686$. Differences in answers is due to rounding.

For completeness: SST will be SSE in a model where only intercept is included. The F-test for the null hypothesis that all regression coefficients (except the intercept) equals zero gives test statistic $F = \frac{\frac{\text{SST}-\text{SSE}}{31-24}}{\frac{\text{SSE}}{24}} = 105.6$, and SSE in the full model we found above to be 1.32. Solving for SST yields 39.2. Finally, $R^2 = \frac{\text{SST-SSE}}{\text{SST}} = \frac{39.3-1.32}{39.2} = 0.966$.

b) I would judge the model fit to be good. The model explains 96% of the variability in the data and the model is significant (from the F-test). The plot of standardized residuals vs fitted values shows no clear structure, and the qq-plot to follow a straight line. The Anderson-Darling normality test doesn't reject the null hypothesis of normal data.

The main effect of damage: when we compare the estimated effect of damage $D = 1$ with the estimated effect of no damage $D = -1$ (keeping the F and T constant at some level), our estimate is $2 \cdot \hat{\beta}_D = 2 \cdot 0.93739 = 1.87$. So, keeping F and T fixed, the effect of damage raises the gene activity with 1.87.

The interaction plot for D and F is found both in cell (1,2) and (2,1). In cell (1,2) the two lines are for $D = -1$ (red) and $D = 1$ (black). The red line goes from $(15.2 + 14.5)/2 = 14.85$ ($F = -1$ and $D = -1$) to $(16.3 + 14.9)/2 = 15.6$ ($F = 1$ and $D = -1$, and shows the effect of F when D is kept at $D = -1$ (no damage) (15.6-14.85=0.75). The numbers taken from the cube plot. The black line goes from $(17.4 + 16.4)/2 = 16.9$ ($F = -1$ and $D = 1$) to $(17.9 + 16.7)/2 = 17.3$ ($F = 1$ and $D = -1$), and shows the effect of F (17.3-16.9=0.4) when D is kept at $D = 1$ (damage). The two lines are not exactly parallel, since the black line is less steep than the red line (however not much). The estimated interaction effect for $D : F$ is $2 \cdot \hat{\beta}_{D:F} = 2 \cdot (-0.08878) = -0.1775$ - or equivalently $0.4/2 - 0.75/2 = 0.2 - 0.375 = -0.175$ (change due to rounding in cube plot numbers).

A natural estimator for γ is

$$\hat{\gamma} = 2^{\hat{\beta}_F - \hat{\beta}_D}$$

where $\hat{\beta}_F$ and $\hat{\beta}_D$ are the appropriate elements of the vector of parameter estimates $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$, were the $\boldsymbol{X}$ is the design matrix and $\boldsymbol{Y}$ is the vector of responses.

We turn to first order Taylor approximations, but first observe that since $2^x = \exp(x \ln 2)$ then $\frac{d(2^x)}{dx} = 2^x \cdot \ln 2$.

$$\begin{aligned} h(\hat{\beta}_F, \hat{\beta}_D) &= 2^{\hat{\beta}_F - \hat{\beta}_D} \\ \frac{\partial h(\hat{\beta}_F, \hat{\beta}_D)}{\partial \hat{\beta}_F} &= \ln 2 \cdot 2^{\hat{\beta}_F - \hat{\beta}_D} \\ \frac{\partial h(\hat{\beta}_F, \hat{\beta}_D)}{\partial \hat{\beta}_D} &= -\ln 2 \cdot 2^{\hat{\beta}_F - \hat{\beta}_D} \end{aligned}$$

where the random variable $\hat{\beta}_F$ has $\mathrm{E}(\hat{\beta}_F) = \beta_F$ and $\mathrm{Var}(\hat{\beta}_F) = \frac{1}{n}\sigma^2$, and $\hat{\beta}_D$ has $\mathrm{E}(\hat{\beta}_D) = \beta_D$ and $\mathrm{Var}(\hat{\beta}_D) = \frac{1}{n}\sigma^2$. Further, $\mathrm{Cov}(\hat{\beta}_F, \hat{\beta}_D) = 0$ since we have an orthogonal design matrix.

Define

$$\begin{aligned} h'_{\beta_F}(\beta_F, \beta_D) &= \frac{\partial h(\hat{\beta}_F, \hat{\beta}_D)}{\partial \hat{\beta}_F} \mid_{\hat{\beta}_F = \beta_F, \hat{\beta}_D = \beta_D} = \ln 2 \cdot 2^{\beta_F - \beta_D} \\ h'_{\beta_D}(\beta_F, \beta_D) &= \frac{\partial h(\hat{\beta}_F, \hat{\beta}_D)}{\partial \hat{\beta}_D} \mid_{\hat{\beta}_F = \beta_F, \hat{\beta}_D = \beta_D} = -\ln 2 \cdot 2^{\beta_F - \beta_D} \end{aligned}$$

The first order Taylor approximation for two independent RVs gives:

$$\begin{aligned} \mathrm{E}(h(\hat{\beta}_F, \hat{\beta}_D)) &\approx h(\beta_F, \beta_D) = 2^{\beta_F - \beta_D} \\ \mathrm{Var}(h(\hat{\beta}_F, \hat{\beta}_D)) &\approx (h'_{\beta_F}(\beta_F, \beta_D))^2 \, \mathrm{Var}(\hat{\beta}_F) + (h'_{\beta_D}(\beta_F, \beta_D))^2 \, \mathrm{Var}(\hat{\beta}_D) \\ &= (\ln 2 \cdot 2^{\beta_F - \beta_D})^2 \frac{1}{n}\sigma^2 + (-\ln 2 \cdot 2^{\beta_F - \beta_D})^2 \frac{1}{n}\sigma^2 = \frac{2(\ln 2)^2 \sigma^2}{n} \cdot 2^{2(\beta_F - \beta_D)} \end{aligned}$$

Estimates using numerical values $\hat{\beta}_F = 0.28546$, $\hat{\beta}_D = 0.93739$, $s^2 = 0.2342^2$ (estimat for σ^2), $n = 32$.

$$\hat{\mathrm{E}}(h(\hat{\beta}_F, \hat{\beta}_D)) \approx 2^{0.28546-0.93739} = 2^{-0.65} = 0.64$$

$$\widehat{\mathrm{Var}}(h(\hat{\beta}_F, \hat{\beta}_D)) \approx \frac{2(\ln 2)^2 0.2342^2}{32} \cdot 2^{2(0.28546-0.93739)} = 0.001647 \cdot 2^{-1.3} = 6.67 \cdot 10^{-4}$$

c) The hypothesis test can be performed as a general linear hypothesis:

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{d} \text{ vs. } H_1 : \boldsymbol{C\beta} \neq \boldsymbol{d}$$

with

$$\boldsymbol{C} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \ \boldsymbol{d} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

and $\boldsymbol{\beta} = (\beta_0, \beta_D, \beta_F, \beta_T, \beta_{D:F}, \beta_{D:T}, \beta_{F:T}, \beta_{D:F:T})$. To test the hypothesis we have worked with the test statistics F_{obs}:

$$F_{obs} = \frac{1}{r}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})^{\mathrm{T}}[\hat{\sigma}^2 \boldsymbol{C}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{C}^{\mathrm{T}}]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})$$

where r is the number of hypotheses being tested (here $r = 3$), $\hat{\sigma}^2$ is the unbiased estimator for σ^2 (previously we have used s^2 for $\hat{\sigma}^2$) and $\hat{\boldsymbol{\beta}}$ is the least squares estimator for $\boldsymbol{\beta}$ (in the full model, where we have p=8 regression parameters). When the null hypothesis is true F_{obs} follows a Fisher distribution with r and $n - p$ degrees of freedom. We have that othogonal columns of the design matrix, and thus $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is a 8×8 diagonal matrix with $n = 32$ on the diagonal, and $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ is a 8×8 diagonal matrix with $\frac{1}{32}$ on the diagonal. Further, $\boldsymbol{C}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{C}^{\mathrm{T}}$ is a 3×3 matrix with $\frac{1}{32}$ on the diagonal, and finally $[\boldsymbol{C}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{C}^{\mathrm{T}}]^{-1}$ is a 3×3 matrix with 32 on the diagonal. This means that F_{obs} will be a sum with three terms – one for each regression parameter to be tested.

$$\begin{aligned} F_{obs} &= \frac{32}{3\hat{\sigma}^2}(\hat{\beta}^2_{D:T} + \hat{\beta}^2_{F:T} + \hat{\beta}^2_{D:F:T}) \\ &= \frac{32}{3 \cdot 0.2342^2}[(-0.00242)^2 + (-0.12614)^2 + (0.09099)^2] = 4.705 \end{aligned}$$

The F-distribution with 3 and 24 degrees of freedom has critical value 3.01 for $\alpha = 0.05$ and 3.72 for $\alpha = 0.025$, so we reject the null hypothesis at level 0.025.

d) Let us assume that an intercept term is present in our regression model. In all-subsets model selection we consider all possible $2^7 = 128$ regression models. Let the model complexity be the number of regression parameters fitted, that is, our model complexity is 1 (only intercept) - 8 (full model). First the best model (minimum SSE, maximum R^2 and minimum s) for each model complexity is found, and is presented in the print-out in Figure 5. E.g. the best model with 2 regression parameters is the one with intercept and β_D. Then, we use R^2_{adj} to choose between each of these 7 best models.

The reason we don't use R^2 to choose between models of different complexity is that R^2 will increase when a regressors is added to the model, even if the new regressors are independent of the response. Why? The least squares estimator will minimize SSE and if the regression coefficient for the new regressor is estimated to be a value different from zero, this means that the SSE of this larger model will be smaller than the SSE of the smaller model.

The R^2_{adj} is constructed to also include information about the number of parameters estimated and the number of observations in the data set. In our example the best model is according to this strategy the model with 6 covariates in addition to the intercept (only the $\beta_{D:T}$ is not included). This model has an R^2_{adj} of 0.961. The fitted regression for this model is found by selecting the estimated regression parameter in Figure 1 (due to orthogonal columns) for the non-zero coefficients.

$$\hat{y} = 16.2 + 0.94D + 0.29F - 0.52T - 0.09D \cdot F - 0.13F \cdot T + 0.09D \cdot F \cdot T$$

However, there are very minor differences between this best model and smaller models. The model with 4 covariates (in addition to the intercept) has R^2_{adj} equal to 0.95, so other choices for the "best model" are possible - if we want model parsimony (which we often want).

Lasso regression adds a penalty term to the least squares criterion to make the model more sparse. This may give a robust fit and avoid overfitting. The penalty term for lasso is the sum of the absolute value of the regression coefficients, and the optimization procedure is to minimize with respect to $\boldsymbol{\beta}$ the following quantity

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p-1} \mid \beta_j \mid$$

(The intercept is not part of the penalty term, and the sum is thus over the other $p - 1 = 7$ parameters in our case.) Cross-validation is used for model choice, and we see from Figure 5 that a value of -4.7 for $\log(\lambda)$ is chosen. This gives a regression model with $\hat{\beta}_{D:T} = 0$.

The fitted lasso regression model is:

$$\hat{y} = 16.2 + 0.93D + 0.28F - 0.51T - 0.08D \cdot F - 0.12F \cdot T + 0.08D \cdot F \cdot T$$

Both all-subset model selection and lasso regression came up with the same non-zero covariates, but in the result from model selection the least-squares estimates for the non-zero coefficients were kept, while for the lasso the least-squares estimates were shrunken (due to the lasso penalty). However, looking at the fitted models we see that the differences between the least squares solution and the lasso solution are minor.

e) The design of our experiment is a full factorial 2^3 design done in four replications. This means that the design matrix (both of the full model and the reduced model) will be an orthogonal matrix. This means that $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ will be a diagonal matrix with n on the diagonal and thus $\hat{\boldsymbol{\beta}}_k = [(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}]_k = \frac{1}{n}\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{Y}$ where $\boldsymbol{x}_k$ is the kth column of the design matrix, i.e. the $\hat{\beta}_k$ will only be a function of $\boldsymbol{x}_k$ and $\boldsymbol{Y}$. Further, $\mathrm{Var}(\hat{\beta})_k = \frac{1}{n}\sigma^2$ and $\mathrm{Cov}(\hat{\beta}_k, \hat{\beta}_j) = 0$ for $j \neq k$. This is the reason why the estimated regression parameters are the same in the full and reduced model.

The full and reduced model will give different predictions and also different residuals, and thus different estimates for the error variance, and thus different estimated standard deviations for the estimated regression parameters between the full and reduced model.

Finally, prediction and prediction interval. In the reduced model the vector of regression parameters is $(\beta_0, \beta_D, \beta_F, \beta_T, \beta_{D:F})$.The prediction is to be made at $D = 1, F = 1, T = -1$, which gives $\boldsymbol{x}_0 = (1, 1, 1, -1, 1)$ as coding for covariates in the reduced model. The prediction is given as $\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} = (1, 1, 1, -1, 1)^{\mathrm{T}}(16.16, 0.94, 0.29, -0.52, -0.09) = 16.16 + 0.94 + 0.29 + 0.52 - 0.09 = 17.82$.

For the interval we need to observe that $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is a 5×5 diagonal matrix with 32 on the diagonal, and thus $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ is a 5×5 diagonal matrix with $\frac{1}{32}$ on the diagonal. Further, $\boldsymbol{x}_0^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0 = 5/32$, since a quadratic form with a diagonal matrix $\boldsymbol{A}$ and a vector $\boldsymbol{x}$ is just $\sum_{i=1}^5 x_i^2 A_{ii}$. The t critical number is found from Figure 7 to be 2.05, and we have $s = 0.2782$ from Figure 7.

$$\begin{aligned}
&\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{\frac{\alpha}{2}, n-p} \cdot s \cdot \sqrt{1 + \boldsymbol{x}_0^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0} \\
=&17.82 \pm 2.05 \cdot 0.2782 \cdot \sqrt{1 + \frac{5}{32}} = 17.82 \pm 0.61 = [17.2, 18.4]
\end{aligned}$$

Problem 3

a) $\mathbf{1}^{\mathrm{T}}\mathbf{1} = n$ and $\mathbf{1}\mathbf{1}^{\mathrm{T}}$ is a $n \times n$ matrix were each entry is equal 1, further, the matrix $\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ has entries

$$\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}} = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & -\frac{1}{n} & \ddots & -\frac{1}{n} \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{pmatrix} = \begin{pmatrix} \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & -\frac{1}{n} & \ddots & -\frac{1}{n} \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & \frac{n-1}{n} \end{pmatrix}$$

So, for $n = 4$ $\mathbf{1}^{\mathrm{T}}\mathbf{1} = 4$ and $\mathbf{1}\mathbf{1}^{\mathrm{T}}$ is a 4×4 matrix were each entry is equal 1, and

$$\boldsymbol{I} - \frac{1}{4}\mathbf{1}\mathbf{1}^{\mathrm{T}} == \begin{pmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

Key characteristics: 1) Symmetric: We see that this matrix is symmetric, since $(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})^{\mathrm{T}} = \boldsymbol{I}^{\mathrm{T}} - \frac{1}{n}(\mathbf{1}\mathbf{1}^{\mathrm{T}})^{\mathrm{T}} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$. 2) Idempotent: We start by showing that $\frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ is idempotent, and then that $\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ is idempotent.

$$\begin{aligned}
(\frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})^2 = \frac{1}{n^2}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{1}\mathbf{1}^{\mathrm{T}} &= \frac{1}{n^2}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \\
&= \frac{1}{n^2}\begin{pmatrix} n & n & \cdots & n \\ \vdots & \vdots & \vdots & \vdots \\ n & n & \cdots & n \end{pmatrix} = \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} = \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}} \\
(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})^2 &= \boldsymbol{I} - 2\frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}} + \frac{1}{n^2}(\mathbf{1}\mathbf{1}^{\mathrm{T}})(\mathbf{1}\mathbf{1}^{\mathrm{T}}) = \boldsymbol{I} - 2\frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}} + \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}
\end{aligned}$$

For a symmetric idempotent matrix the rank of the matrix equals the trace (sum of diagonal elements) of the matrix. $\mathrm{tr}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}) = \mathrm{tr}(\boldsymbol{I}) - \mathrm{tr}(\frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}) = n - 1$, so the rank of $\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ is $n - 1$.

Finally, $\mathrm{E}(S^2)$.

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y} \\
\mathrm{E}(S^2) &= \frac{1}{n-1}\,\mathrm{E}(\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}) \\
\mathrm{E}(\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}) &= \mathrm{tr}((\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\sigma^2\boldsymbol{I}) + \mu\mathbf{1}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\mu\mathbf{1} \\
&= \sigma^2\,\mathrm{tr}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}) + \mu^2(\mathbf{1}^{\mathrm{T}}\mathbf{1} - \mathbf{1}^{\mathrm{T}}\frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{1}) = \sigma^2(n-1) + \mu^2(n - \frac{1}{n}\cdot n\cdot n) \\
&= \sigma^2(n-1) \\
\mathrm{E}(S^2) &= \frac{1}{n-1}\sigma^2(n-1) = \sigma^2
\end{aligned}$$

b) It is known from the course curriculum and lectures that if $\boldsymbol{Y} \sim N_n(\mathbf{0}, \boldsymbol{I})$, and $\boldsymbol{R}$ is a symmetric and idempotent matrix with rank r, then

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Y} \sim \chi^2_r$$

We have $\boldsymbol{Y} \sim N_n(\mu\mathbf{1}, \sigma^2\boldsymbol{I})$, and $\boldsymbol{R} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ is symmetric and idempotent with rank $n-1$. We need to consider a transformation of $\boldsymbol{Y}$ with mean $\mathbf{0}$ and covariance $\boldsymbol{I}$. The transformed $\boldsymbol{Y}^* = \frac{1}{\sigma}(\boldsymbol{Y} - \mu\mathbf{1}) \sim N_n(\mathbf{0}, \boldsymbol{I})$, since $\mathrm{E}(\frac{1}{\sigma}(\boldsymbol{Y} - \mu\mathbf{1})) = \frac{1}{\sigma}(\mu\mathbf{1} - \mu\mathbf{1}) = \mathbf{0}$ and $\mathrm{Cov}(\frac{1}{\sigma}(\boldsymbol{Y} - \mu\mathbf{1})) = \frac{1}{\sigma^2}\mathrm{Cov}(\boldsymbol{Y}) = \frac{1}{\sigma^2}\sigma^2\boldsymbol{I} = \boldsymbol{I}$.

We then have

$$(\boldsymbol{Y}^*)^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}^* \sim \chi^2_{n-1}$$

and need to relate this to $\frac{1}{\sigma^2}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}$.

$$\begin{aligned}
&(\boldsymbol{Y}^*)^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}^* = \frac{1}{\sigma}(\boldsymbol{Y} - \mu\mathbf{1})^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\frac{1}{\sigma}(\boldsymbol{Y} - \mu\mathbf{1}) \\
&= \frac{1}{\sigma^2}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y} - 2\frac{1}{\sigma^2}\mu\mathbf{1}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y} + \frac{1}{\sigma^2}\mu\mathbf{1}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\mu\mathbf{1} \\
&= \frac{1}{\sigma^2}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y} - 2\frac{1}{\sigma^2}\boldsymbol{Y}^{\mathrm{T}}\mu(\mathbf{1} - \frac{n}{n}\mathbf{1}) + \frac{1}{\sigma^2}\mu\mathbf{1}^{\mathrm{T}}\mu(\mathbf{1} - \frac{n}{n}\mathbf{1}) = \frac{1}{\sigma^2}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}
\end{aligned}$$

Thus, $\frac{1}{\sigma^2}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y} \sim \chi^2_{n-1}$. It is known that the variance of a χ^2_ν-distributed variable equals 2ν.

$\mathrm{Var}(S^2)$:

$$\begin{aligned}
Var(S^2) &= \mathrm{Var}(\frac{\sigma^2}{\sigma^2}\frac{1}{n-1}\boldsymbol{Y}\mathrm{T}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}) \\
&= \frac{\sigma^4}{(n-1)^2}\mathrm{Var}(\frac{1}{\sigma^2}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}) = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1}
\end{aligned}$$

Distribution of $\boldsymbol{AY}$: $\frac{1}{n}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y} = \bar{Y} \sim N_1(\mu, \frac{\sigma^2}{n})$ since $\frac{1}{n}\mathbf{1}^{\mathrm{T}}\mu\mathbf{1} = \mu$ and $\frac{1}{n}\mathbf{1}^{\mathrm{T}}\sigma^2\boldsymbol{I}\mathbf{1}\frac{1}{n} = \frac{\sigma^2}{n}$.
Distribution of $\boldsymbol{BY}$: $(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}$: It is known that if $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $\boldsymbol{Z} = \boldsymbol{A}_{q\times n}\boldsymbol{Y} \sim N_n(\boldsymbol{AY}, \boldsymbol{A\Sigma A}^{\mathrm{T}})$. Here we have $\boldsymbol{\mu} = \mu\mathbf{1}$, $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$ and $\boldsymbol{A} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$.

$$\begin{aligned}
\mathrm{E}((\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}) &= (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\mu\mathbf{1} = \mathbf{0} \\
\mathrm{Cov}((\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{Y})\boldsymbol{Y}) &= (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\sigma^2\boldsymbol{I}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})^{\mathrm{T}} = \sigma^2(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})
\end{aligned}$$

This gives: $(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y} \sim N_n(\mathbf{0}, \sigma^2(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})$.

Since $\boldsymbol{Y}$ is multivariate normal with covariance matrix $\sigma^2\boldsymbol{I}$, then $\boldsymbol{AY}$ and $\boldsymbol{BY}$ are independent iff $\boldsymbol{A}\sigma^2\boldsymbol{IB}^{\mathrm{T}} = \mathbf{0}$.

$$\begin{aligned}
\boldsymbol{A}\sigma^2\boldsymbol{IB}^{\mathrm{T}} &= \frac{1}{n}\mathbf{1}^{\mathrm{T}}\sigma^2\boldsymbol{I}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{1})^{\mathrm{T}} = \frac{\sigma^2}{n}(\mathbf{1}^{\mathrm{T}}\boldsymbol{I} - \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{1}\mathbf{1}^{\mathrm{T}}) \\
&= \frac{\sigma^2}{n}(\mathbf{1}^{\mathrm{T}} - \mathbf{1}^{\mathrm{T}}) = \mathbf{0}
\end{aligned}$$

This means that $\frac{1}{n}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y}$ and $(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}$ are independent random vectors.

Finally, to find the distribution of

$$\frac{n(\frac{1}{n}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y} - \mu)^2}{\frac{1}{n-1}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}},$$

we know that the numerator and denominator of the expression are independent since the numerator is a function of $\frac{1}{n}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y}$ and the denominator is a function of $(\boldsymbol{I} - \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{1})^{\mathrm{T}}\boldsymbol{Y}$, and we have shown that these two are independent. Further, observe that

$$\frac{(\frac{1}{n}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y} - \mu)^2}{\frac{\sigma^2}{n}} \sim \chi^2_1$$
$$\frac{1}{\sigma^2}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y} \sim \chi^2_{n-1}$$

We will rewrite the expression under study to include these two independent χ^2-distributed expressions.

$$\frac{n(\frac{1}{n}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y} - \mu)^2}{\frac{1}{n-1}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}} = \frac{\frac{n}{\sigma^2}(\frac{1}{n}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y} - \mu)^2}{\frac{1}{n-1}\frac{1}{\sigma^2}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}} = \frac{\chi^2_1/1}{\chi^2_{n-1}/(n-1)} \sim F_{1,n-1}$$

Thus, the expression follows a Fisher distribution with 1 and $n-1$ degrees of freedom. Observe that this is the squared t-statistic for one sample inference, and that $t^2_{n-1} = F_{1,n-1}$.