TMA4267 Linear Statistical Models
Part 3: Hypothesis testing and ANOVA
Solutions to recommended exercise 5 - V2017

February 28, 2017

Problem 1: Plant stress

a) T-statistic in Intercept row: $t_0 = \frac{\hat{\beta}_0 - 0}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_0)}} = \frac{16.15942}{0.04140} = 390.3$. 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, $t_j = \frac{\hat{\beta}_j - 0}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}}$ so that $\sqrt{\widehat{\text{Var}}(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{t_j} = \frac{-0.00242}{-0.058} = 0.04$. 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 $(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}s^2$, where s^2 is the estimate for the regression variance σ^2. For our orthogonal design $\boldsymbol{X}^{\text{T}}\boldsymbol{X}$ 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, $R^2 = 1 - \frac{n-p}{n-1}(1 - R^2_{\text{adj}}) = 1 - \frac{31}{24}(1 - 0.9594) = 0.9686$. 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 response 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), $n = 32$.

$$\begin{aligned}
\hat{\mathrm{E}}(h(\hat{\beta}_F, \hat{\beta}_D)) &\approx 2^{0.28546 - 0.93739} = 2^{-0.65} = 0.64 \\
\widehat{\mathrm{Var}}(h(\hat{\beta}_F, \hat{\beta}_D)) &\approx \frac{2(\ln 2)^2 0.2342^2}{32} \cdot 2^{2(0.28546 - 0.93739)} = 0.001647 \cdot 2^{-1.3} = 6.67 \cdot 10^{-4}
\end{aligned}$$

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

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

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).

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 2: Multiple testing with plant stress

a) We are testing m hypotheses and we select a cut-off on p-values that leads to rejecting R hypotheses. Out of these R rejected hypotheses V is the number of Type I errors (the true hypotheses among the R rejected).

FWER is the probability of one or more false positive finding (at least one fake news), and is mathematically $P(V > 0)$.

FDR is the expected proportion of false positive findings among the rejections (expected proportion of news that are fake): $\mathrm{E}(V/R)$ (and when $R = 0$ then 0).

b) The Bonferroni rule choose $\alpha_{\text{loc}} = \frac{\alpha}{m}$, and for us we choose $\alpha = 0.05$ and we have $m = 10000$ so $\alpha_{\text{loc}} = \frac{0.05}{10000} = 5 \cdot 10^{-6}$.

Bonferronis method can always be used, for any dependency structure between the p-values.

People often state that the method of Bonferroni is conservative, but what they most often mean is that FWER is a very strict criterion. Controlling that the probability of one or more Type I error when m is large is very strict. On the other hand FDR is less strict and can be used.

Other reasons for saying that the Bonferroni rule is conservative is related to that the rule is valid for all types of dependency structures, also when the p-values from the m hypotheses are independent. Often, at least in genetical applications the tests performed are dependent on each other because the genes or genetic markers tested are correlated. Then more elaborate methods that take into account this dependency structure will give a much higher value for the p-value cut-off.

c) It is now known that $m_0 = 9000$ of the hypotheses are true and and $m_1 = 1000$ are false. This gives the following numbers in our table (see R-code below). In particular we do not have any type I errors, $V = 0$.

In the table below - when the true nature of which hypotheses are true and false are unknown to us - then we only can observe m and R.

	Not reject H_0	Reject H_0	Total
H_0 true	9000	0	9000
H_0 false	981	19	1000
Total	9981	19	10000

d) Now $\alpha_{\text{loc}} = 0.05$ and we do as on c), and get the following table. Observe that we have 428 false positives (fake news).

	Not reject H_0	Reject H_0	Total
H_0 true	8572	428	9000
H_0 false	178	822	1000
Total	8750	1250	10000

Problem 3: F test and partial F test in multiple linear regression

a) We already know from the theory that $\boldsymbol{I} - \boldsymbol{H}$ and $\boldsymbol{H}_0$ are idempotent. $(\boldsymbol{H} - \boldsymbol{H}_0)^2 = \boldsymbol{H}^2 - \boldsymbol{H}\boldsymbol{H}_0 - \boldsymbol{H}_0\boldsymbol{H} + \boldsymbol{H}_0^2 = \boldsymbol{H} - \boldsymbol{H}\boldsymbol{H}_0 - \boldsymbol{H}_0\boldsymbol{H} + \boldsymbol{H}_0$. Since $\boldsymbol{H}_0$ is a projection onto the column space of $\boldsymbol{X}_0$, a further projection by $\boldsymbol{H}$ to the column space of $\boldsymbol{X}$ will do nothing further, and $\boldsymbol{H}\boldsymbol{H}_0 = \boldsymbol{H}_0$. Also, $\boldsymbol{I} - \boldsymbol{H}$ projects onto a subspace orthogonal to the column space of $\boldsymbol{X}$, so a further projection by $\boldsymbol{H}_0$ gives $\boldsymbol{H}_0(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{0}$, and $\boldsymbol{H}_0\boldsymbol{H} = \boldsymbol{H}_0$. So $(\boldsymbol{H} - \boldsymbol{H}_0)^2 = \boldsymbol{H} - \boldsymbol{H}_0$.

b) The first identity is true in general since $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = 0$. For the second identity, under the null hypothesis $\boldsymbol{X}\boldsymbol{\beta}$ is in the column space of $\boldsymbol{X}_0$, so that $\boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{H}_0\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta}$, and $(\boldsymbol{H} - \boldsymbol{H}_0)\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$.

c) We know from the theory that $\operatorname{rank}(\boldsymbol{I} - \boldsymbol{H}) = n - p$, $\operatorname{rank}\boldsymbol{H} = p$ and that $\operatorname{rank}\boldsymbol{H}_0 = r$. Then $\operatorname{rank}(\boldsymbol{H} - \boldsymbol{H}_0) = \operatorname{tr}(\boldsymbol{H} - \boldsymbol{H}_0) = \operatorname{tr}\boldsymbol{H} - \operatorname{tr}\boldsymbol{H}_0 = p - r$. The statement about the distribution of the F-statistic follows from

1. the independence of the two quadratic forms $\boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H}_0)\boldsymbol{Y}$ and $\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ (which is itself a consequence of the orthogonality of the projections $\boldsymbol{H} - \boldsymbol{H}_0$ and $\boldsymbol{I} - \boldsymbol{H}$; see point 4 of Theorem B.8 on p. 651 of Fahrmeir et al.) and
2. the definition of the F-distribution as the distribution of a ratio of two independent chi-square distributed random variables, each divided by its number of degrees of freedom (Definition B.14 on p. 645 in Fahrmeir et al.).

d)
```
# How does pH in Nowegian lakes depend on sulfate, nitrate, calcium, aluminium
# and organic content (x1, ... x5), area of lake (x6) and location (x7 = 0,
# Telemark, or x7 = 1, Tr{\o}ndelag)? Data from the Norwegian pollution control
# authority (Statens forurensningstilsyn, 1986). Here, 26 lakes located in
# Telemark and Tr{\o}ndelag have been randomly selected from among 1005 lakes.

acidrain <-
  read.table("http://www.math.ntnu.no/~mettela/TMA4267/Data/acidrain.txt",
  header=TRUE)

attach(acidrain)

n <- length(y)

x <- cbind(rep(1,n),acidrain[,2:8]) # we want intercept in model
names(x)[1] <- 1
x <- as.matrix(x)
p <- dim(x)[2]

i <- diag(n)
h <- x%*%solve(t(x)%*%x)%*%t(x)

# Test of submodel where coefficients of x2, x4, x5, x6, x7 are zero:
r <- 3
x0 <- x[,c(1,2,4)]
h0 <- x0%*%solve(t(x0)%*%x0)%*%t(x0)
f <- t(y)%*%(h-h0)%*%y/(p-r)/(t(y)%*%(i-h)%*%y/(n-p))
```

```
f
pf(f,p-r,n-p,lower.tail=FALSE) # p-value - cannot reject null hypothesis
# or by R functions (1 is added by R as a covariate):
fit <- lm(y~x1+x2+x3+x4+x5+x6+x7)
fit0 <- lm(y~x1+x3)
anova(fit0,fit)

# Test of whether all coefficients except the intercept are zero
r <- 1
x0 <- x[,1]
h0 <- x0%*%solve(t(x0)%*%x0)%*%t(x0)
f <- t(y)%*%(h-h0)%*%y/(p-r)/(t(y)%*%(i-h)%*%y/(n-p))
f
pf(f,p-r,n-p,lower.tail=FALSE) # p-value - cannot reject null hypothesis
# or by R functions:
fit0 <- lm(y~1)
anova(fit0,fit)
# or:
fit2 <- lm(y~x)
anova(fit2)
```

e) In this case $\boldsymbol{X}_0 = \mathbf{1}$, so that $\boldsymbol{H}_0 = \boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^T = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T = \mathbf{1}n^{-1}\mathbf{1}^T = \frac{1}{n}\mathbf{1}\mathbf{1}^T$. R code is given above.

Problem 4: One- and two-way ANOVA – and the linear model

a)
```
income <- c(300, 350, 370, 360, 400, 370, 420, 390,
              400, 430, 420, 410, 300, 320, 310, 305,
              350, 370, 340, 355, 370, 380, 360, 365)
gender <- c(rep("Male", 12),rep("Female",12))
place <- rep(c(rep("A",4),rep("B",4),rep("C",4)),2)
data <- data.frame(income,gender=factor(gender,levels=c("Female","Male")),
            place=factor(place,levels=c("A","B","C")))
data

pairs(data)
plot(income~place,data=data)
plot(income~gender,data=data)
interaction.plot(data$gender, data$place, data$income)
plot.design(income~place+gender, data = data)
```

b)
```
X = cbind(rep(1,length(data$income)),data$place=="A",
   data$place=="B",data$place=="C")
X
XtX <- t(X)%*%X
qr(XtX)$rank
```

c)
```
model = lm(income~place-1,data=data,x=TRUE)
model$x # design matrix
summary(model)
```

```
anova(model)
```

This is a parametrization without intercept, and with three estimated effects for place.

d)
```
options(contrasts=c("contr.treatment","contr.poly"))
model1 = lm(income~place,data=data,x=TRUE)
model1$x
summary(model1)
anova(model1)
```

Treatment contrast parametrization codes the factor at the lowest level (which is `A` here) as 0, so that the value of the intercept will be the estimate for the level `A`. Compare this with the model above.

```
model$coeff
model1$coeff

options(contrasts=c("contr.sum","contr.poly"))
model2 = lm(income ~ place,data=data,x=TRUE)
model2$x
summary(model2)
model2$coeff
```

Sum-to-zero contrast parametrization puts `C` as `-A-B` so that `A+B+C=0`.

e) Using linear hypothesis - Starting with model 1:

```
r=2
C=cbind(rep(0,r),diag(r))
d=matrix(rep(0,r),ncol=1)

betahat=matrix(model1$coefficients,ncol=1)
sigma2hat=summary(model1)$sigma^2
X=model.matrix(model1)

Fobs1=(t(C%*%betahat-d)%*%solve(C%*%solve(t(X)%*%X)%*%t(C))%*%
            (C%*%betahat-d))/(r*sigma2hat)
Fobs1
1-pf(Fobs1,r,n-length(betahat))

betahat=matrix(model2$coefficients,ncol=1)
sigma2hat=summary(model2)$sigma^2
X=model.matrix(model2)

Fobs2=(t(C%*%betahat-d)%*%solve(C%*%solve(t(X)%*%X)%*%t(C))%*%
            (C%*%betahat-d))/(r*sigma2hat)
Fobs2
1-pf(Fobs2,r,n-length(betahat))
```

Same result of hypothesis test. What about the no intercept that was in b) (not asked for)?

```
r=2
C=matrix(c(1,-1,0,0,1,-1),ncol=3,byrow=TRUE)
C
d=matrix(rep(0,r),ncol=1)

betahat=matrix(model$coefficients,ncol=1)
sigma2hat=summary(model)$sigma^2
X=model.matrix(model)

Fobs=(t(C%*%betahat-d)%*%solve(C%*%solve(t(X)%*%X)%*%t(C))%*%
          (C%*%betahat-d))/(r*sigma2hat)
Fobs
1-pf(Fobs,r,n-length(betahat))
```

This also gives the same result.

f)
```
options(contrasts=c("contr.treatment","contr.poly"))
model3 = lm(income~place+gender,data=data,x=TRUE)
model3$x
anova(model3)
summary(model3)

options(contrasts=c("contr.sum","contr.poly"))
model4 = lm(income~place+gender,data=data,x=TRUE)
model4$x
summary(model4)
anova(model4)
```

Testing the place effect in model 4, and then the gender effect:

```
betahat=matrix(model4$coefficients,ncol=1)
sigma2hat=summary(model4)$sigma^2
X=model.matrix(model4)

r=2
Cplace=cbind(rep(0,r),diag(r),rep(0,r)) #add gener coeff last column
d=matrix(rep(0,r),ncol=1)

Fobsplace=(t(Cplace%*%betahat-d)%*%
           solve(Cplace%*%solve(t(X)%*%X)%*%t(Cplace))%*%
           (Cplace%*%betahat-d))/(r*sigma2hat)
Fobsplace
1-pf(Fobsplace,r,n-length(betahat))
```

There's no need to test the significance of `gender`, since only one parameter can be read off of the summary. This gives the same result as using `anova(model4)`.

```
options(contrasts=c("contr.sum","contr.poly"))
model5 = lm(income~place*gender,data=data,x=TRUE)
summary(model5)
X=model5$x
anova(model5)
```

The interaction is not significant. Now perform the same test (significance of `place:gender`-interaction, given that all main effects are in the model) using the $\mathbf{C}\boldsymbol{\beta}$-approach.

```
r=2
Csamspill=cbind(rep(0,r),rep(0,r),rep(0,r),rep(0,r),diag(r))
# add gender coef. to last column
d=matrix(rep(0,r),ncol=1)

betahat=model5$coefficients
betahat
Csamspill%*%betahat
sigma2hat=summary(model5)$sigma^2
Fobssamspill=(t(Csamspill%*%betahat-d)%*%solve(Csamspill%*%
               solve(t(X)%*%X)%*%t(Csamspill))%*%
               (Csamspill%*%betahat-d))/(r*sigma2hat)
Fobssamspill
1-pf(Fobssamspill,r,n-length(betahat))
```

This gives the same result as above. Finally, repeat the same test using dummy variable coding (`contr.treatment`).

```
options(contrasts=c("contr.treatment","contr.poly"))
model5 = lm(income~place*gender,data=data,x=TRUE)
summary(model5)
X=model5$x
anova(model5)
r=2
Csamspill=cbind(rep(0,r),rep(0,r),rep(0,r),rep(0,r),diag(r))
d=matrix(rep(0,r),ncol=1)

betahat=model5$coefficients
betahat
Csamspill%*%betahat
sigma2hat=summary(model5)$sigma^2
Fobssamspill=(t(Csamspill%*%betahat-d)%*%
               solve(Csamspill%*%solve(t(X)%*%X)%*%t(Csamspill))%*%
               (Csamspill%*%betahat-d))/(r*sigma2hat)
Fobssamspill
1-pf(Fobssamspill,r,n-length(betahat))
```

This also gives the same result.

Problem 5: Teaching reading

a) We would like to investigate if the expected reading score varies between the teaching methods.
Write down the null- and alternative hypothesis and perform one hypothesis test based on the summary statistics in the table above.

What are the assumptions you need to make to use this test?
What is the conclusion from the test? Hypotheses:
Let μ_A, μ_B and μ_C be the expected reading scores for each of the three methods.

$$H_0 : \mu_A = \mu_B = \mu_C \text{ vs. } H_1 : \text{at least one pair differs}$$

This hypothesis can be tested using one-way analysis of variance. We need to fill in the ANOVA table (SS, MS, df, F), which can be calculated from the summary statistics.

Let $\bar{x}_A$ denote the average and s_A the standard deviation of method A. Ditto for methods B and C. Let $\bar{x}$ denote the grand mean.

$$\begin{aligned} SSA &= n_A(\bar{x}_A - \bar{x})^2 + n_B(\bar{x}_B - \bar{x})^2 + n_C(\bar{x}_C - \bar{x})^2 \\ &= 22 \cdot (41.05 - 44.02)^2 + 22 \cdot (46.73 - 44.02)^2 + 22 \cdot (44.27 - 44.02)^2 \\ &= 357.005 \\ SSE &= (n_A - 1)s_A^2 + (n_B - 1)s_B^2 + (n_C - 1)s_C^2 \\ &= 25511.712 \end{aligned}$$

Source	SS	df	MS	F
Method	357.005	2	178.5	4.47
Error	2511.712	63	39.9	
Total	2868.717	65		

The F statistic, here observed to be 4.47, should be compared with the critical value $f_{0.05,2,63}$. We find $f_{0.05,2,60} = 3.15$ in Table A.6, and we thus reject the null hypothesis. (We know that $f_{0.05,2,63} < f_{0.05,2,60}$.)

Assumptions:
The one-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where the error terms are independent and normally distributed with the same variance across treatment groups.

Conclusion:
There is reason to believe that the expected reading score is not the same for all the methods.

b) $\boldsymbol{C}$ has dimension 2×3, since we are testing 2 linear hypotheses and we have 3 parameters in our regression.

But, since we only have summary data - and not pairs of response and method, we can't fit a regression model directly. It is however possible to calculate the F-test statistics with some effort, see R-code at course webpage, but that involves at bit of fiddling with the terms and is not within the core of the course. Only those especially interested should check out the R-code.

c) Let $\bar{X}_B$ be the mean of a random sample from using method A and $\bar{X}_C$ the mean of a random sample from using method C. A natural estimator for γ is

$$\hat{\gamma} = \frac{\bar{X}_B}{\bar{X}_C}$$

We turn to first order Taylor approximations with

$$\begin{aligned} h(\bar{X}_B, \bar{X}_C) &= \frac{\bar{X}_B}{\bar{X}_C} \\ \frac{\partial h(\bar{X}_B, \bar{X}_C)}{\partial \bar{X}_B} &= \frac{1}{\bar{X}_C} \\ \frac{\partial h(\bar{X}_B, \bar{X}_C)}{\partial \bar{X}_C} &= -\frac{\bar{X}_B}{\bar{X}_C^2} \end{aligned}$$

where the random variable $\bar{X}_B$ has $\mathrm{E}(\bar{X}_B) = \mu_B$ and $\mathrm{Var}(\bar{X}_B) = \sigma_B^2/n_B$, and $\bar{X}_C$ has $\mathrm{E}(\bar{X}_C) = \mu_C$ and $\mathrm{Var}(\bar{X}_C) = \sigma_C^2/n_C$.

Define

$$\begin{aligned} h'_B(\mu_B, \mu_C) &= \frac{\partial h(\bar{X}_B, \bar{X}_C)}{\partial \bar{X}_B} \mid_{\bar{X}_B=\mu_B, \bar{X}_C=\mu_C} = \frac{1}{\mu_C} \\ h'_C(\mu_B, \mu_C) &= \frac{\partial h(\bar{X}_B, \bar{X}_C)}{\partial \bar{X}_C} \mid_{\bar{X}_B=\mu_B, \bar{X}_C=\mu_C} = -\frac{\mu_B}{\mu_C^2} \end{aligned}$$

We assume that the two samples are independent. The first order Taylor approximation for two independent samples:

$$\begin{aligned} \mathrm{E}(h(\bar{X}_B, \bar{X}_C)) &\approx h(\mu_B, \mu_C) = \frac{\mu_B}{\mu_C} \\ \mathrm{Var}(h(\bar{X}_B, \bar{X}_C)) &\approx (h'_B(\mu_B, \mu_C))^2 \,\mathrm{Var}(\bar{X}_B) + (h'_C(\mu_B, \mu_C))^2 \,\mathrm{Var}(\bar{X}_C) \\ &= \left(\frac{1}{\mu_C}\right)^2 \cdot \frac{\sigma_B^2}{n_B} + \left(-\frac{\mu_B}{\mu_C^2}\right)^2 \cdot \frac{\sigma_C^2}{n_C} \end{aligned}$$

Estimates using numerical values $n_B = 22$, $n_C = 22$, $\hat{\mu}_B = \bar{x}_B = 46.73$, $\hat{\mu}_C = \bar{x}_C = 44.27$, $\hat{\sigma}_B^2 = s_B^2 = 7.388^2$, $\hat{\sigma}_C^2 = s_C^2 = 2^2$ are as follows.

$$\begin{aligned} \hat{\gamma} &= \frac{46.73}{44.27} = 1.06 \\ \mathrm{E}(h(\bar{X}_B, \bar{X}_C)) &\approx \frac{46.73}{44.27} = 1.06 \\ \mathrm{Var}(h(\bar{X}_B, \bar{X}_C) &\approx \left(\frac{1}{44.27}\right)^2 \cdot \frac{7.388^2}{22} + \left(\frac{46.73}{44.27^2}\right)^2 \cdot \frac{5.767^2}{22} \\ &= 0.00127 + 0.00086 = 0.00212 \\ \mathrm{SD}(h(\bar{X}_B, \bar{X}_C)) &\approx \sqrt{0.00212} = 0.046 \end{aligned}$$