Department of Mathematical Sciences

TMA4267 Linear statistical models
Recommended exercises 7 − solutions

Problem 1   **Inference about a new observation in multiple linear regression**

a) $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X^{\mathrm{T}}X)^{-1})$, so that the linear transformation $\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ of $\hat{\boldsymbol{\beta}}$ has the distribution $N(\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2\boldsymbol{x}_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x}_0)$, which is univariate. So its expected value is the EY_0 specified by the model, and is thus unbiased.

b) $U = (n-p)\hat{\sigma}^2/\sigma^2 = \mathrm{SSE}/\sigma^2 \sim \chi^2_{n-p}$, where n is the length of $\boldsymbol{Y}$ and p is the length of $\boldsymbol{\beta}$, and it is independent of $\hat{\boldsymbol{\beta}}$ and thus of $\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}$. Standardize $\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ and divide it by $\sqrt{U/(n-p)}$ to get

$$T = \frac{\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} - EY_0}{\hat{\sigma}\sqrt{\boldsymbol{x}_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x}_0}} \sim t_{n-p}.$$

Solve the double inequality $-t_{n-p} \leq T \leq t_{n-p}$, where $-t_{\alpha/2}$ is the $\alpha/2$-quantile of a t_{n-p} variable, to get a confidence interval having bounds

$$\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{x}_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x}_0}.$$

c) Since Y_0 is independent of $\boldsymbol{\epsilon}$ and thus of $\boldsymbol{x}_0^{T}\hat{\boldsymbol{\beta}}$, $Y_0 - \boldsymbol{x}_0^{T}\hat{\boldsymbol{\beta}} \sim N(0, \sigma^2(1 + \boldsymbol{x}_0^{T}(X^{T}X)^{-1}\boldsymbol{x}_0))$. Standardize $Y_0 - \boldsymbol{x}_0^{T}\hat{\boldsymbol{\beta}}$ and proceed as in (b) to get a prediction interval having bounds

$$\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x}_0}.$$

d)
```
acidrain <-
   read.table("https://www.math.ntnu.no/emner/TMA4267/2018v/acidrain.txt",header=TRUE)
fit <- lm(y~.,data=acidrain) # lm: linear model
x <- model.matrix(fit)
n <- dim(x)[1]; p <- dim(x)[2]

# New observation:
x0 <- c(1,3,50,1,50,2,1,0) # remember intercept (first 1)
prediction <- sum(x0*fit$coefficients)
prediction

# Confidence interval:
halflength <-
   qt(.025,n-p,lower.tail=FALSE)*summary(fit)$sigma*sqrt(x0%*%solve(t(x)%*%x)%*%x0)
```

```
prediction-halflength
prediction+halflength

# Prediction interval:
halflength <-
  qt(.025,n-p,lower.tail=FALSE)*summary(fit)$sigma*sqrt(1+x0%*%solve(t(x)%*%x)%*%x0)
prediction-halflength
prediction+halflength

# By R functions instead:
newdata <- data.frame(x1=3,x2=50,x3=1,x4=50,x5=2,x6=1,x7=0)
# or the following two lines
newdata <- data.frame(3,50,1,50,2,1,0)
names(newdata) <- names(coefficients(fit)[-1])
# In either case:
predict(fit, newdata, level=.95, interval="confidence")
predict(fit, newdata, level=.95, interval="prediction")
```

e) The design matrix in this case is

$$X = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}^{\mathrm{T}},$$

and, with $\boldsymbol{x}_0 = (1 \ \ x_0)^{\mathrm{T}}$, you can verify that

$$\boldsymbol{x}_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x}_0 = \frac{\sum_i (x_i - x_0)^2}{n \sum_i (x_i - \bar{x})^2}.$$

Next, write $x_i - x_0 = (x_i - \bar{x}) + (\bar{x} - x_0)$ to get $\sum_i (x_i - x_0)^2 = \sum_i (x_i - \bar{x})^2 + n(\bar{x} - x_0)^2$ (crossterms vanish). The bounds given follow.

Problem 2   Plant stress

a) **t value in (Intercept) row:** $\hat{\beta}_0/\sqrt{\widehat{\mathrm{Var}\,\hat{\beta}_0}} = 16.15942/0.04140 = 390.3$. This is the test statistic for testing the null hypothesis $H_0\colon \beta_0 = 0$ vs. $H_1\colon \beta_0 \neq 0$, where β_0 is the intercept.

Std. Error in row named D:T: in general, $t_j = \hat{\beta}_j/\sqrt{\widehat{\mathrm{Var}\,\hat{\beta}_j}}$, so that $\sqrt{\widehat{\mathrm{Var}\,\hat{\beta}_j}} = \hat{\beta}_j/t_j = -0.00242/(-0.058) = 0.04$. This is the estimated standard deviation for the coefficient estimator.

Alternatively, we may conclude that **Std. Error** for $\hat{\beta}_{D:T}$ is 0.04140 since we have orthogonal columns in our design matrix and the standard error estimates are the same for

all estimated regression parameters in the model. This is the square root of the diagonal entry corresponding to $D : T$ of $(X^{\mathrm{T}}X)^{-1}\hat{\sigma}^2$, where X is the design matrix and $\hat{\sigma}^2$ is the estimate for the regression variance σ^2. For our orthogonal design, $X^{\mathrm{T}}X$ is a diagonal matrix with all diagonal entries equal to 32. We read off $\hat{\sigma}$ from the print-out, `Residual standard error: 0.2342`. Thus, `Std. Error` $= 0.2342/\sqrt{32} = 0.04$.

`Pr(>|t|)` **in row named** `D:F:T:` The probability that the absolute value of a t-distributed variable with 24 degrees of freedom is greater than the observed t-statistic, 2.198. Using R, we find

```
> 2 * pt(2.198, 24, lower.tail = FALSE)
[1] 0.03783603
```

This is the p-value when testing $H_0\colon \beta_{D:F:T} = 0$ vs. $H1\colon \beta_{D:F:T} \neq 0$. Reject the null hypothesis if the p-value is smaller than the chosen significance level. Using only the statistical tables, *Tabeller og formler i statistikk*, we cannot find a precise value, but from the table on page 4, 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 is between 0.02 and 0.05.

`Multiple R-squared:` $R^2 = 1 - \mathrm{SSE}/\mathrm{SST}$, so we need SSE and SST. We find SSE from $\hat{\sigma}$, $\mathrm{SSE} = (n - p)\hat{\sigma}^2 = 24 \cdot 0.2342^2 = 1.32$. The F-test for the null hypothesis that all regression coefficients (except the intercept) equals zero gives test statistic $F = \frac{(\mathrm{SST}-\mathrm{SSE})/(31-24)}{\mathrm{SSE}/24} = 105.6$, and SSE in the full model we found above to be 1.32. Solving for SST yields 39.2. Finally, $R^2 = 1 - \mathrm{SSE}/\mathrm{SST} = 1 - 1.32/39.2 = 0.97$.

We can also find R^2 from the adjusted R^2 (`Adjusted R-squared`). $R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SSE}/(n-p)}{\mathrm{SST}/(n-1)}$, which gives $R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{n-p}{n-1}(1 - R^2_{\mathrm{adj}}) = 1 - \frac{31}{24}(1 - 0.9594) = 0.97$.

b) 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 entries of the vector of parameter estimates $\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\boldsymbol{Y}$, were the X is the design matrix and $\boldsymbol{Y}$ is the vector of responses.

The first-order Taylor approximation of $h(\hat{\beta}_F, \hat{\beta}_D) = 2^{\hat{\beta}_F - \hat{\beta}_D}$ is, noting that $d\, 2^x/dx = d\exp(x \ln 2)/dx = 2^x \ln 2$,

$$2^{\hat{\beta}_F - \hat{\beta}_D} = h(\hat{\beta}_F, \hat{\beta}_D) \approx h(\beta_F, \beta_D) + h_{\hat{\beta}_F}(\beta_F, \beta_D)(\hat{\beta}_F - \beta_F) + h_{\hat{\beta}_D}(\beta_F, \beta_D)(\hat{\beta}_D - \beta_D)$$
$$= 2^{\beta_F - \beta_L} + 2^{\beta_F - \beta_L}(\hat{\beta}_F - \beta_F)\ln 2 - 2^{\beta_F - \beta_L}(\hat{\beta}_D - \beta_D)\ln 2,$$

where $h_{\hat{\beta}_F}$ and $h_{\hat{\beta}_D}$ denote partial derivatives with respect to the first and second variable, respectively, so that

$$E\, 2^{\hat{\beta}_F - \hat{\beta}_D} \approx 2^{\beta_F - \beta_L},$$

since $E\hat{\beta}_F = \beta_F$ and $E\hat{\beta}_D = \beta_D$, and

$$\operatorname{Var} 2^{\hat{\beta}_F - \hat{\beta}_D} \approx (2^{\beta_F - \beta_L} \ln 2)^2 \operatorname{Var} \hat{\beta}_F + (2^{\beta_F - \beta_L} \ln 2)^2 \operatorname{Var} \hat{\beta}_L$$
$$= (\ln 2)^2 \, 2^{2(\beta_F - \beta_L)} (\operatorname{Var} \hat{\beta}_F + \operatorname{Var} \hat{\beta}_D).$$

Then we have estimates

$$\widehat{E2^{\hat{\beta}_F - \hat{\beta}_D}} = 2^{\hat{\beta}_F - \hat{\beta}_D} = 2^{0.28546 - 0.93739} = 0.636$$

and

$$\widehat{\operatorname{Var} 2^{\hat{\beta}_F - \hat{\beta}_D}} = (\ln 2)^2 \, 2^{2(\hat{\beta}_F - \hat{\beta}_L)} (\operatorname{Var} \hat{\beta}_F + \operatorname{Var} \hat{\beta}_D)$$
$$= (\ln 2)^2 \, 2^{2(0.28546 - 0.93739)} (0.04140^2 + 0.04140^2) = 6.67 \cdot 10^{-4}.$$

c) The hypothesis test can be performed as

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

with

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

and $\boldsymbol{\beta} = (\beta_0 \ \ \beta_D \ \ \beta_F \ \ \beta_T \ \ \beta_{D:F} \ \ \beta_{D:T} \ \ \beta_{F:T} \ \ \beta_{D:F:T})^{\mathrm{T}}$. We can use the test statistic

$$F = (C\hat{\boldsymbol{\beta}} - \boldsymbol{d})^{\mathrm{T}} C(X^{\mathrm{T}}X)^{-1} C^{\mathrm{T}})^{-1} (C\hat{\boldsymbol{\beta}} - \boldsymbol{d}) / (r\hat{\sigma}^2),$$

where r is the number of hypotheses being tested (here $r = 3$), $\hat{\sigma}^2$ is the unbiased estimator for σ^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, $F \sim F_{r,n-p}$. The design matrix has orthogonal columns, and thus $X^{\mathrm{T}}X$ is a diagonal 8×8 matrix with all diagonal entries equal to $n = 32$, and $(X^{\mathrm{T}}X)^{-1}$ is a diagonal 8×8 matrix with all diagonal entries equal to $1/32$. Further, $C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}}$ is a diagonal 3×3 matrix with all diagonal entries $1/32$, and finally $(C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}})^{-1}$ a diagonal 3×3 matrix with all diagonal entries 32. This means that F will involve a sum of three terms – one for each regression parameter to be tested,

$$F = \frac{32}{3\hat{\sigma}^2} (\hat{\beta}_{D:T}^2 + \hat{\beta}_{F:T}^2 + \hat{\beta}_{D:F:T}^2)$$
$$= \frac{32}{3 \cdot 0.2342^2} \left((-0.00242)^2 + (-0.12614)^2 + (0.09099)^2 \right) = 4.705$$

The $F_{3,24}$ distribution 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) The design matrix of both the full and the reduced model has orthogonal columns (the experiment is what we in the last part of the course will call a full factorial 2^3 design done in four replications). This means that $X^\mathrm{T}X$ is a diagonal matrix with all diagonal entries equal to $n = 32$, so that the jth entry of $\hat{\boldsymbol{\beta}} = (X^\mathrm{T}X)^{-1}X^\mathrm{T}\boldsymbol{Y}$ is $\hat{\beta}_j = \frac{1}{n}\boldsymbol{x}_j^\mathrm{T}\boldsymbol{Y}$, where $\boldsymbol{x}_j$ is the jth column of X, i.e., $\hat{\beta}_j$ is a function only of $\boldsymbol{x}_j$ and $\boldsymbol{Y}$. This is the reason why the coefficient estimates are the same in the full and reduced model.

Further, $\operatorname{Var}\hat{\beta}_j = \sigma^2/n$ and $\operatorname{Cov}(\hat{\beta}_j, \hat{\beta}_k) = 0$ for $j \neq k$. You may be surprised that the standard error estimates are greater for the reduced model than for the full model. What we know, is that the true standard errors for a reduced model is at most those of the full. In the current models with orthogonal design, they are in fact equal, σ^2/n, and the estimate of σ was somewhat higher for the reduced model than for the full: The two models give different predictions and also different residuals, and thus different estimates for the error variance.

In the reduced model the coefficient vector is $(\beta_0 \ \beta_D \ \beta_F \ \beta_T \ \beta_{D:F})^\mathrm{T}$.The prediction is to be made at $D = 1, \ F = 1, \ T = -1$, which gives covariate vector $\boldsymbol{x}_0 = (1 \ 1 \ 1 \ -1 \ 1)^\mathrm{T}$. The prediction is $\boldsymbol{x}_0^\mathrm{T}\hat{\boldsymbol{\beta}} = (1 \ 1 \ 1 \ -1 \ 1)(16.16 \ 0.94 \ 0.29 \ -0.52 \ -0.09)^\mathrm{T} = 16.16 + 0.94 + 0.29 + 0.52 - 0.09 = 17.82$.

For the prediction interval, we need $(X^\mathrm{T}X)^{-1}$, which is a 5×5 diagonal matrix with all diagonal entries $1/32$, and $\boldsymbol{x}_0^\mathrm{T}(X^\mathrm{T}X)^{-1}\boldsymbol{x}_0 = 5/32$. The t critical value is found from Figure 2 to be $t_{\alpha/2} = 2.05 \ (n - p = 32 - 5 = 27)$, and we have $\hat{\sigma} = 0.2782$, also from Figure 2. The 95% prediciton interval has bounds

$$\boldsymbol{x}_0^\mathrm{T}\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^\mathrm{T}(X^\mathrm{T}X)^{-1}\boldsymbol{x}_0} = 17.82 \pm 2.05 \cdot 0.2782 \cdot \sqrt{1 + \frac{5}{32}} = 17.82 \pm 0.61,$$

and we have the prediction interval $[17.2, 18.4]$.

Problem 3   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, let V be the number of Type I errors (true hypotheses rejected).

FWER is the probability of one or more false positive finding, $P(V > 0)$.

FDR is the expected proportion of false positive findings among the rejections, $E(V/R)$ (or 0, if $R = 0$).

b) Using Bonferroni's method, we choose $\alpha_{\mathrm{loc}} = \alpha/m = 0.05/10000 = 5 \cdot 10^{-6}$. We reject 19 null hypotheses.

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

Bonferroni's method is often called conservative, but what is most often meant is that FWER is a very strict criterion. Controlling the probability of one or more Type I error when m is large may be very strict.

Another reason for saying that Bonferroni's rule, is that the rule is valid for all types of dependency structures, also when the p-values from the m hypotheses are independent. Often, the tests performed are dependent on each other, meaning that the "effective" number of tests is less than m. In genetical applications, tests can be dependent because the genes or genetic markers tested are correlated. Then more elaborate methods that take into account this dependency structure (such as the multivariate distribution of the test statistics) can give a much higher value for the p-value cut-off, while still controlling the FWER.

c) The filled-in table:

	H_0 not rejected	H_0 rejected	Total
H_0 true	9000	0	9000
H_0 false	981	19	1000
Total	9981	19	10000

There are none false positives. When we do not know which hypotheses are true and false, as in a normal multiple hypothesis testing situation, only $m = 10000$ and $R = 10$ (in the last row) are known.

d) Without correcting for multiple testing:

	H_0 not rejected	H_0 rejected	Total
H_0 true	8572	428	9000
H_0 false	178	822	1000
Total	8750	1250	10000

We have 428 false positives.