Norwegian University of Science and Technology
Department of Mathematical Sciences

TMA4267 Linear statistical models
Recommended exercises 7

Problem 1 Inference about a new observation in multiple linear regression

Let X be an $n \times p$ matrix of rank p. Consider a multiple linear regression model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 I)$. Assume that $Y_0 = \boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_0$ is a new observation, with $\epsilon_0 \sim N(0, \sigma^2)$, independent of $\boldsymbol{\epsilon}$.

a) Show that $\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ is an unbiased estimator of EY_0, where $\hat{\boldsymbol{\beta}}$ is the least-square estimator of $\boldsymbol{\beta}$. Find the distribution of the estimator.

Let $\hat{\sigma}^2 = \mathrm{SSE}/(n-p)$ be the usual unbiased estimator of σ^2 and $-t_{\alpha/2}$ the $\alpha/2$-quantile of a t_{n-p} variable.

b) Show that the interval having bounds $\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{x}_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x}_0}$ is a $100(1-\alpha)\%$ confidence interval for EY_0.

c) Show that the interval having bounds $\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x}_0}$ is a $100(1-\alpha)\%$ prediction interval for Y_0, that is, an interval that will contain Y_0 with probability $1-\alpha$.

d) Use the acid rain data to find a confidence interval for the expected value of a new observation having covariates (`x1` $\cdots$ `x7`) $=$ (3 50 1 50 2 1 0). Also find a prediction interval for such a new observation.

e) From the theory of simple linear regression, you know that the bounds of the confidence interval are

$$\hat{y}_0 \pm t_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}},$$

where $\hat{y}_0$ is the estimator of EY_0, $-t_{\alpha/2}$ the $\alpha/2$-quantile of a t_{n-2} variable, n the number of observations, x_i the covariates and x_0 the new covariate. Show that this is the same confidence interval as found above.

Problem 2 Plant stress

(Adapted from Exam TMA4267 2016 spring, Problem 2.)

At the Department of Biology at NTNU, researchers use the model plant *Arabidopsis thaliana* to study the response of a plant to different sources of stress. In an experiment, *Arabidopsis thaliana* seedlings were subject to a stress situation. The following factors (categorical covariates are often called *factors*, and their values *levels*) were fitted:

D (damage): $D = 1$ means that the plant was damaged mechanically by cutting into the leaves of the plant by a pair of scissors. $D = -1$ means damage was not inflicted (no cutting).

F (flagellin): $F = 1$ means that the pathogen-derived peptide flagellin was sprayed on the leaves of the plant. $F = -1$ means water (not flagellin) was sprayed.

T (time): Plants were harvested at two different time points after the stress situation. $T = 1$ means that the plant was harvested 60 minutes after the stress situation and $T = -1$ means that the plant was harvested 30 minutes after the stress situation.

Thus, we have three factors, D, F and T, each at two levels. In the study, experiments for all possible combinations of the three factors were performed four times yielding 32 experiments in total.

The response measured in the experiment, was the observed gene activity level (a continuous measurement) of each of around 40 000 genes. We will only focus on the gene activity level of one of these genes, the AT1G32920 gene, and we denote the gene activity level of this gene by Y. It is known that this gene is active in response to wounding.

For experiment number i (where $i = 1, \dots, 32$): Y_i is the observed response, D_i is chosen value of D, F_i is chosen value of F, and T_i is chosen value of T. A multiple regression model with all main effects (the factors), and two- and three-way interactions (products of two or three factors), was considered,

$$Y_i = \beta_0 + \beta_D D_i + \beta_F F_i + \beta_T T_i + \beta_{D:F} D_i F_i + \beta_{D:T} D_i T_i + \beta_{F:T} F_i T_i + \beta_{D:F:T} D_i F_i T_i + \epsilon_i,$$

where $i = 1 ,\dots, 32$, and we assume ϵ_i independent and identically normally distributed with mean 0 and variance σ^2. We refer to this as the *full model.* The vector of regression parameters is $\boldsymbol{\beta} = (\beta_0 \ \ \beta_D \ \ \beta_F \ \ \beta_T \ \ \beta_{D:F} \ \ \beta_{D:T} \ \ \beta_{F:T} \ \ \beta_{D:F:T})^{\mathrm{T}}$, and the ith row of the design matrix X is $(1 \ \ D_i \ \ F_i \ \ T_i \ \ D_iF_i \ \ D_iT_i \ \ F_iT_i \ \ D_iF_iT_i)$.

In Figure 1, you find R commands and print-out from fitting the full model.

a) In the print-out from `summary(fit)` in Figure 1, four numerical values are replaced by question marks. Calculate numerical values for each of these, and explain what each of the values means.

```
# data is in "standard order" in data frame with name "ds"
> ds #showing only rows 1-6 and 27-32 for space considerations
          Y  D  F  T
1  15.45169 -1 -1 -1
2  15.15908 -1 -1 -1
3  14.93064 -1 -1 -1
4  15.06569 -1 -1 -1
5  14.51032 -1 -1  1
6  14.76922 -1 -1  1
...
27 18.23645  1  1 -1
28 17.70327  1  1 -1
29 16.66523  1  1  1
30 16.96046  1  1  1
31 16.73133  1  1  1
32 16.57248  1  1  1
> fit=lm(Y~D*F*T,data=ds) # model formula to generate all interactions easier
> model.matrix(fit) #only showing rows 1-6 and 27-32
   (Intercept)  D  F  T D:F D:T F:T D:F:T
1            1 -1 -1 -1   1   1   1    -1
2            1 -1 -1 -1   1   1   1    -1
3            1 -1 -1 -1   1   1   1    -1
4            1 -1 -1 -1   1   1   1    -1
5            1 -1 -1  1   1  -1  -1     1
6            1 -1 -1  1   1  -1  -1     1
...
27           1  1  1 -1   1  -1  -1    -1
28           1  1  1 -1   1  -1  -1    -1
29           1  1  1  1   1   1   1     1
30           1  1  1  1   1   1   1     1
31           1  1  1  1   1   1   1     1
32           1  1  1  1   1   1   1     1
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.15942    0.04140  ?       < 2e-16
D            0.93739    0.04140  22.644  < 2e-16
F            0.28546    0.04140   6.896 3.93e-07
T           -0.52354    0.04140 -12.647 4.18e-12
D:F         -0.08878    0.04140  -2.145  0.04231
D:T         -0.00242    ?        -0.058  0.95386
F:T         -0.12614    0.04140  -3.047  0.00555
D:F:T        0.09099    0.04140   2.198  ?
Residual standard error: 0.2342 on 24 degrees of freedom
Multiple R-squared:      ?, Adjusted R-squared:  0.9594
F-statistic: 105.6 on 7 and 24 DF,  p-value: < 2.2e-16
```

Figure 1: Printout from R commands and statistical analyses for the plant stress data set. Four numbers are replaced by question marks.

Let $\gamma = 2^{\beta_F - \beta_D}$ be a new parameter of interest.

b) Suggest an estimator, $\hat{\gamma}$, for γ. Use approximate methods to find the expected value and variance of this estimator, that is, $E\hat{\gamma}$ and $\operatorname{Var}\hat{\gamma}$. Use results in Figure 1 to calculate numerical value for $\hat{\gamma}$, and estimated numerical values for $E\hat{\gamma}$ and $\operatorname{Var}\hat{\gamma}$.

Hint: Consider the first-order Taylor approximation of a function of two variables: $h(x,y) \approx h(a,b) + h_x(a,b)(x-a) + h_y(a,b)(y-b)$, where h_x and h_y denote partial derivitives with respect to the first and second variable, respectively. Further, you may use that $2^x = \exp(x \ln 2)$, where ln is the natural logarithm.

The researchers want to test the hypothesis

$$\begin{aligned} &H_0\colon \beta_{D:T} = \beta_{F:T} = \beta_{D:F:T} = 0 \qquad \text{vs.} \\ &H_1\colon \text{at least one of } \beta_{D:T},\ \ \beta_{F:T},\ \ \beta_{D:F:T} \text{ is different from 0.} \end{aligned}$$

c) Perform the hypothesis test at a significance level of your own choice. All the numerical values you need for the calculations are found in Figure 1.

Hint: $X^{\mathrm{T}}X$ and $(X^{\mathrm{T}}X)^{-1}$ are easy to calculate.

The researchers choose to use the following *reduced model* for prediction:

$$Y_i = \beta_0 + \beta_D D_i + \beta_F F_i + \beta_T T_i + \beta_{D:F} D_i F_i + \epsilon_i,$$

where $i = 1, \ldots, 32$, and we assume ϵ_i independent and identically normally distributed with mean 0 and variance σ^2. Output from fitting the reduced model is given in Figure 2.

d) Compare the estimated regression coefficients and the estimated standard deviations of the estimated regression coefficients for the full model (Figure 1) and the reduced model (Figure 2), and explain what you observe.

Based on the reduced model (Figure 2), provide a prediction, and a 95% prediction interval, for the gene activity level for the factor combination $D = 1,\ F = 1,\ T = -1$.

Hint: See problem 1(c), and Figure 2 for some values of $t_{\alpha/2}$.

```
> fitRED=lm(Y~D+F+T+D:F,data=ds)
> summary(fitRED)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.15942    0.04919 328.528  < 2e-16
D            0.93739    0.04919  19.057  < 2e-16
F            0.28546    0.04919   5.804 3.56e-06
T           -0.52354    0.04919 -10.644 3.66e-11
D:F         -0.08878    0.04919  -1.805   0.0822
Residual standard error: 0.2782 on 27 degrees of freedom
Multiple R-squared:   0.95,Adjusted R-squared:  0.9426
F-statistic: 128.4 on 4 and 27 DF,  p-value: < 2.2e-16
> qt(0.025,32,lower.tail=FALSE)
[1] 2.036933
> qt(0.025,27,lower.tail=FALSE)
[1] 2.051831
> qt(0.025,24,lower.tail=FALSE)
[1] 2.063899
```

Figure 2: Print-out from R performing linear regression on the reduced model for the plant stress data set.

Problem 3 Multiple testing with plant stress

In this problem we work with the same biological problem as in the previous problem, but we will here focus on multiple hypothesis testing.

We want to study the gene activity level of each of $m = 10\,000$ genes (in Problem 2 we only looked at one of the genes), and test for relationship with damage. Let μ_j denote the regression coefficient of damage for gene j, $j = 1, \ldots, m$ (previously denoted β_D for one gene). Our aim is now to identify genes that respond to damage, and we do this by performing the two-sided hypothesis tests

$$H_0\colon \mu_j = 0 \qquad \text{vs.} \qquad H_1\colon \mu_j \neq 0$$

for $j = 1, \ldots, m$ (all genes). Assume that we have performed the hypothesis tests and calculated p-values $p_1, \ldots, p_m$.

Note: For each gene we have performed one regression, with the gene activity level of that gene as the response, so the μ_js are from m different regression models with different responses (not to be confused with different regression parameters).

You can read the p-values into R:

```
pvalues <- scan("https://www.math.ntnu.no/emner/TMA4267/2018v/damagePvalues.txt")
m <- length(pvalues)
```

a) Explain *family-wise error rate* (FWER) and *false discovery rate* (FDR).

We choose to control the FWER at level $\alpha = 0.05$ by the Bonferroni method.

b) Calculate the Bonferroni local significance level α_{loc} for our data, to be used as a cut-off for the p-values. How many null hypotheses will you reject? Are there any requirements for when we can use the Bonferroni method? Why do people say that the Bonferroni method is conservative?

The p-values given above have been artificially generated, and the truth is that the 9000 first p-values are from true null hypotheses and the last 1000 are from the false null hypotheses.

c) Based on your rule in (b), fill in numerical values for the quantities denoted S, T, U, V, R, m and m_0 in the following table. What is the number of false positives?

	H_0 not rejected	H_0 rejected	Total
H_0 true	U	V	m_0
H_0 false	T	S	$m - m_0$
Total	$m - R$	R	m

In a multiple hypothesis setting where the true nature of the hypotheses are not known, which of the entries in the table above are actually known?

d) What if we instead consider all p-values below 0.05 as a significant result? What would then S, T, U, V, R, m and m_0 become? What is the number of false positives now?

In scientific research it is important to avoid false positive results (fake news). That is why a cut-off of 0.05 never should be used when more than one hypothesis test is conducted.