Exercise/lecture note – Linear models

A)

The data set 'plants' contains data on the height of tomato plants grown under different temperatures, light conditions and soil types in a greenhouse experiment. Start by reading the data into R:

```
plants = read.csv("http://www.finse.uio.no/events/international-
workshops/introduction-to-estimation/data/plants.csv")
```

This data set is quite small, so you can look at the whole `data.frame`:

```
plants
```

and we can get a summary of the data:

```
summary(plants)
```

'`height`' is in centimeters, '`temp`' is temperature in Celsius, '`light`' is a factor describing the light intensity with three levels (low, medium and high) and there are two types of '`soil`' (A and B). Look at the numbers you get from `summary(plants)` and make sure you understand what they mean. These data are from a balanced design with one plant measured for each combination of '`temp`', '`light`' and '`soil`'.

Get the mean '`height`' from each of the 6 groups defined by unique values of '`light`' and '`soil`':

```
tapply(plants$height, list(plants$light, plants$soil), mean)
```

Compare these numbers to the parameter estimates you get when you fit a linear model where the expected heights depend on '`light`', '`soil`' and the combination of '`light`' and '`soil`' (the '`light:soil`' interaction effect), and where the height measurements are normal distributed:

```
lm(height ~ light + soil + light:soil, data=plants)
```

Q1: What is the relationship between these parameter estimates and the means you got R to compute earlier?

The parameter estimates from the model fit can be extracted from a model object with the function 'coef':

```
fit = lm(height ~ light + soil + light:soil, data=plants)

coef(fit)
```

Q2: If you call these 6 parameter estimates $\widehat{\beta_1}$, $\widehat{\beta_2}$, $\widehat{\beta_3}$, $\widehat{\beta_4}$, $\widehat{\beta_5}$, and $\widehat{\beta_6}$ respectively, can you write down with math the expected mean for each of the 6 groups of plants?

Now try adding the effect of 'temp' as a continuous predictor variable to the model.

```
fit2 = lm(height ~ light + soil + light:soil + temp, data=plants)
beta = coef(fit2)
beta
```

Q3: What does the parameter estimate 'temp' mean?

Q4: Why does the estimate of the 'intercept' change?

We can also extract the data you have used from the model object in a special format called the "model matrix" or the "design matrix":

```
model.matrix(fit2)
```

This matrix can also be computed from the data without fitting the model to the data:

```
X = model.matrix(~ light + soil + light:soil + temp, data=plants)
X
```

Q3: Can you make any sense out of this matrix if you compare it to the original data?

I'll explain what this matrix is during the lecture, but for now you can think of it as a "reformatting of the data". I will here show you a few useful things this matrix can be used for:

a) If you multiply this matrix with the parameter vector, you will get the "fitted predictions" (often called $\hat{y}$):

```
y.hat = X %*% beta
y.hat
```

These are the expected height for plants that have the same set of predictor variables as the plants in each row of the data.

b) If you subtract the observed values to the fitted predictions, you get the residuals (below I make a histogram of these):

```
residuals = plants$height - y.hat
hist(residuals)
```

c) Just as two parameters are not independent (c.f., Andy's presentation), the fitted predictions are not independent either. The whole variance-covariance matrix for the predictions can be computed as $\mathbf{XSX}'$, where $\mathbf{X}$ is the model matrix (or "design matrix") we have computed above, $\mathbf{S}$ is the variance-covariance matrix of the parameter estimates, and $\mathbf{X}'$ is the transpose of the model matrix (columns tuned into rows and vice versa). In R we compute this as:

```
S = vcov(fit2)

S

X %*% S %*% t(X)
```

This is a very large matrix, which we are not so often interested in. The diagonal of this matrix contains the variances of the fitted predictions, and if we take the square root of these, we get the standard errors of the predictions

```
var.fitted.preds = diag(X %*% S %*% t(X))

se.fitted.preds = sqrt(var.fitted.preds)

se.fitted.preds
```

d) We can get the predictions for any plant in the population with a given '`temp`', '`light`' and '`soil`' by coding this into a vector and performing the same calculations as above. For example, the predicted height of a plant grown under medium light conditions in soil-type A in 20 ˚C is $\widehat{\beta_1} + \widehat{\beta_3} + \widehat{\beta_5} \cdot 20$ (note that this is an extrapolation since 20 is outside the range of the temperature values in the data). This can be computed as

```
x = c(1,0,1,0,20,0,0)

pred = x %*% beta

pred
```

We can compute the standard error of this prediction in the same way as above:

```
se.pred = sqrt(x %*% S %*% x)

se.pred
```

(Note that we don't transpose the last x here because a R-vector is treated as either a column vector or a row vector – which ever fits with the expression (`t(x)` becomes a row vector and you get an error message)).

e) We can compute an approximate 95% confidence interval for this prediction as the estimate ±2SE:

```
pred + c(-2,2)*se.pred
```

Note that this is the confidence interval for the expectation (or the large sample mean) of all plants in the population that have these values for '`temp`', '`light`' and '`soil`'. If we increase the sample size this confidence interval will become narrower, and if we increase the sample size a lot the standard error will move towards zero. However, if we want to construct a confidence interval for the expected measurement of one individual plant, we have to include the variance of the residuals in the calculations. The standard deviation of the residual term can be obtained from the model object as

```
summary(fit2)$sigma
```

The variance for the prediction of a single individual plant is the variance for the prediction of the mean plus the variance of the residual term:

```
se.pred^2 + summary(fit2)$sigma^2
```

And the square root of this is the standard error of the predictions for a single individual plant:

```
se.pred.ind = sqrt(se.pred^2 + summary(fit2)$sigma^2)
```

The confidence interval based on this is

```
pred + c(-2,2)*se.pred.ind
```

f) We can find the standard errors, co-variances and confidence intervals of any linear combination of the parameters. For example, the difference between the expected height of plants grown under medium and low light conditions (but same temperature), when the plants are grown in soil type B, is $(\beta_1 + \beta_3 + \beta_4 + \beta_5\{\text{temp}\} + \beta_7) - (\beta_1 + \beta_2 + \beta_4 + \beta_5\{\text{temp}\} + \beta_6) = \beta_3 + \beta_7 - \beta_2 - \beta_6$. Hence, by the same procedure as before we get

```
x = c(0,-1,1,0,0,-1,1)

est = x %*% beta # Estimated difference

est

se = sqrt(x %*% S %*% x) # SE of this difference

se

est + c(-2,2)*se # approximately 95% c.i.
```

For the same difference when the plants are grown in soil type A, we get

```
x = c(0,-1,1,0,0,0,0)

est = x %*% beta # Estimated difference

est

se = sqrt(x %*% S %*% x) # SE of this difference

se

est + c(-2,2)*se # approximately 95% c.i.
```