Module 2: MULTIPLE LINEAR REGRESSION

TMA4315 Generalized linear models H2018 Week 1

Mette Langaas, Department of Mathematical Sciences, NTNU
– with contributions from Øyvind Bakke and Ingeborg Hem

30.08 and 06.09 [PL], 31.08 and 07.09 [IL]

SECOND WEEK

What to remember?

NEED TO CHANGE: GLM WAY!!!

Model:

$$\mathbf{Y} = \mathbf{X} \ +$$

with full rank design matrix. And classical *normal* linear regression model when

$$\varepsilon \sim N_n(\mathbf{0}, \ ^{\mathbf{2}}\mathbf{I}).$$

Parameter of interest is β and σ^2 is a nuisance. Maximum likelihood estimator

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

has distribution: $\widehat{\beta} \sim N_p(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.

Restricted maximum likelihood estimator for $\ ^2$:

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{Y} - \mathbf{X}\widehat{\beta})^T(\mathbf{Y} - \mathbf{X}\widehat{\beta}) = \frac{\mathsf{SSE}}{n-p}$$

Inference

We will consider confidence intervals and prediction intervals, and then test single and linear hypotheses.

Confidence intervals (CI)

In addition to providing a parameter estimate for each element of our parameter vector β we should also report a $(1-\alpha)100\%$ confidence interval (CI) for each element. (We will not consider simultanious confidence regions in this course.)

We focus on element j of β, called β_j. It is known that $T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}}$ follows a t-distribution with $n-p$ degrees of freedom. Let $t_{\alpha/2,n-p}$ be such that $P(T_j > t_{\alpha/2,n-p}) = \alpha/2$. REMARK: our textbook would here look at area to the left instead of to the right - but we stick with this notation. Since the t-distribution is symmetric around 0, then $P(T_j < -t_{\alpha/2,n-p}) = \alpha/2$. We may then write

$$P(-t_{\alpha/2,n-p} \leq T_j \leq t_{\alpha/2,n-p}) = 1-\alpha$$

```r
library(ggplot2)
```

Inserting $T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}}$ and solving so β_j is in the middle gives:

$$P(\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{c_{jj}}\hat{\sigma} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{c_{jj}}\hat{\sigma}) = 1 - \alpha$$

A $(1-\alpha)$% CI for β_j is when we insert numerical values for the upper and lower limits: $[\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{c_{jj}}\hat{\sigma}, \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{c_{jj}}\hat{\sigma}]$.

CIs can be found in R using `confint` on an `lm` object. (Here dummy variable coding is used for `location`, with average as reference location.)

```
library(gamlss.data)
fit = lm(rent ~ area + location + bath + kitchen + cheating
confint(fit)
```

```
##                   2.5 %      97.5 %
## (Intercept) -44.825534   0.8788739
## area          4.354674   4.8029443
## location2    28.579849  49.9405909
## location3    92.970636 159.1443278
## bath1        52.076412  96.0311030
```

Prediction intervals

Remember, one aim for regression was to “construct a model to predict the reponse from a set of (one or several) explanatory variables- more or less black box”.
Assume we want to make a prediction (of the response - often called Y_0) given specific values for the covariates - often called $\mathbf{x}_0$.
An intuitive point estimate is $\hat{Y}_0 = \mathbf{x}_0^T \hat{\beta}$ - but to give a hint of the uncertainty in this prediction we also want to present a prediction interval for the Y_0.

To arrive at such an estimate we start with the difference between the unobserved response Y_0 (for a given covariate vector $\mathbf{x}_0$) and the point prediction $\hat{Y}_0$, $Y_0 - \hat{Y}_0$. First, we assume that the unobserved response at covariate $\mathbf{x}_0$ is independent of our previous observations and follows the same distibution, that is $Y_0 \sim N(\mathbf{x}_0^T \beta, \sigma^2)$. Further,

$$\hat{Y}_0 = \mathbf{x}_0^T \hat{\beta} \sim N(\mathbf{x}_0^T \beta, \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0).$$

Then, for $Y_0 - \mathbf{x}_0^T \hat{\beta}$ we have

$\mathsf{E}(Y_0 - \mathbf{x}_0^T \hat{\beta}) = 0$ and $\mathsf{Var}(Y_0 - \mathbf{x}_0^T \hat{\beta}) = \mathsf{Var}(Y_0) + \mathsf{Var}(\mathbf{x}_0^T \hat{\beta}) = \sigma^2 + \sigma^2 \mathbf{x}_0^T ($

so that

$$Y_0 - \mathbf{x}_0^T \hat{\beta} \sim N(0, \sigma^2 (1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0))$$

Inserting our REML-estimate for σ^2 gives

$$T = \frac{Y_0 - \mathbf{x}_0^T \hat{\beta}}{\hat{\sigma} \sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}} \sim t_{n-p}.$$

PIs can be found in R using `predict` on an `lm` object, but make sure that `newdata` is a `data.frame` with the same names as the original data. We want to predict the rent - with PI - for an appartment with area 50, location 2 (“good”), nice bath and kitchen and with central heating.

```
library(gamlss.data)
fit = lm(rent ~ area + location + bath + kitchen + cheating
newobs = rent99[1, ]
newobs[1, ] = c(NA, NA, 50, NA, 2, 1, 1, 1, NA)
predict(fit, newdata = newobs, interval = "prediction", typ
```

```
##          fit      lwr      upr
## 1 602.1298 315.5353 888.7243
```

Q (and A):

1. When is a prediction interval of interest?
2. Explain the result from `predict` above.
3. What is the interpretation of a 95% prediction interval?

Single hypothesis testing set-up

In single hypothesis testing we are interesting in testing one null hypothesis against an alternative hypothesis. In linear regression the hypothesis is often about a regression parameter β_j:

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0$$

Remark: we implicitly say that our test is done given that the other variables are present in the model, that is, the other β_is $(j \neq i)$ are not zero.
###Two types of errors:

- "Reject H_0 when H_0 is true"="false positives" = "type I error" ="miscarriage of justice". These are our *fake news*, which are very important for us to avoid.
- "Fail to reject H_0 when H_1 is true (and H_0 is false)"="false negatives" = "type II error"= "guilty criminal go free".

We choose to reject H_0 at some significance level α if the p-value of the test (see below) is smaller than the chosen significance level. We say that : Type I error is “controlled” at significance level α, which means that the probability of miscarriage of justice (Type I error) does not exceed α.

Q: Draw a 2 by 2 table showing the connection between

- “truth” (H_0 true or H_0 false) - rows in the table, and
- “action” (reject H_0 and accept H_0) - columns in the table,

and place the two types of errors in the correct position within the table.

What else should be written in the last two cells?

Hypothesis test on β_j (t-test)

In linear regression models our test statistic for testing $H_0 : \beta_j = 0$ is

$$T_0 = \frac{\hat{\beta}_j - 0}{\sqrt{c_{jj}}\hat{\sigma}_\varepsilon} \sim t_{n-2}$$

where $c_{jj}\hat{\sigma}^2_\varepsilon = \widehat{\text{Var}}(\hat{\beta}_j)$.
Inserted observed values (and estimates) we have t_0.
We would in a two-sided setting reject H_0 for large values of abs(t_0). We may rely on calculating a p-value.

The p-value

A p-value is a test statistic satisfying $0 \leq p(\mathbf{Y}) \leq 1$ for every vector of observations $\mathbf{Y}$.

- Small values give evidence that H_1 is true.
- In single hypothesis testing, if the p-value is less than the chosen significance level (chosen upper limit for the probability of committing a type I error), then we reject the null hypothesis, H_0. The chosen significance level is often referred to as α.
- A p-value is *valid* if

$$P(p(\mathbf{Y}) \leq \alpha) \leq \alpha$$

  for all α, $0 \leq \alpha \leq 1$, whenever H_0 is true, that is, if the p-value is valid, rejection on the basis of the p-value ensures that the probability of type I error does not exceed α.
- If $P(p(\mathbf{Y}) \leq \alpha) = \alpha$ for all α, $0 \leq \alpha \leq 1$, the p-value is called an *exact* p-value.

In our linear regression we use the t-distibution to calculate p-values for our two-sided test situation $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$. Assume we have observed that our test statistic T_0 takes the numerical value t_0. Since the t-distribution is symmetric around 0 we have

$$p\text{-value} = P(T_0 > \mathsf{abs}(t_0)) + P(T_0 < -\mathsf{abs}(t_0)) = 2 \cdot P(T_0 > \mathsf{abs}(t_0)).$$

We reject H_0 if our calculated p-value is below our chosen signficance level. We often choose as significance level $\alpha = 0.05$.

Munich rent index hypothesis test

We look at print-out using summary from fitting `lm`.

```
library(gamlss.data)
colnames(rent99)
fit = lm(rent ~ area + location + bath + kitchen + cheatin
summary(fit)
```

```
## [1] "rent"      "rentsqm"   "area"      "yearc"     "locatio
## [8] "cheating" "district"
##
## Call:
## lm(formula = rent ~ area + location + bath + kitchen + c
##     data = rent99)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -633.41  -89.17   -6.26   82.96 1000.76
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

Testing linear hypotheses in regression

We study a normal linear regression model with $p = k + 1$ covariates, and refer to this as model A (the larger model). We then want to investigate the null and alternative hypotheses of the following type(s):

$$\begin{aligned} H_0 : \beta_j &= 0 \text{ vs. } H_1 : \beta_j \neq 0 \\ H_0 : \beta_1 &= \beta_2 = \beta_3 = 0 \text{ vs. } H_1 : \text{ at least one of these } \neq 0 \\ H_0 : \beta_1 &= \beta_2 = \cdots = \beta_k = 0 \text{ vs. } H_1 : \text{ at least one of these } \neq 0 \end{aligned}$$

We call the restricted model (when the null hypotesis is true) model B, or the smaller model.
These null hypotheses and alternative hypotheses can all be rewritten as a linear hypothesis

$$H_0 : \mathbf{C} \ = \mathbf{d} \text{ vs. } \mathbf{C} \ \neq \mathbf{d}$$

by specifying $\mathbf{C}$ to be a $r \times p$ matrix and $\mathbf{d}$ to be a column vector of length p.

Testing a set of parameters - what is $\mathbf{C}$ and $\mathbf{d}$?

We consider a regression model with intercept and five covariates, $x_1, \dots, x_5$. Assume that we want to know if the covariates x_3, x_4, and x_5 can be dropped (due to the fact that none of the corresponding β_js are different from zero). This means that we want to test:

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0 \text{ vs. } H_1 : \text{ at least one of these } \neq 0$$

This means that our $\mathbf{C}$ is a 6×3 matrix and $\mathbf{d}$ a 3×1 column vector

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } \mathbf{d} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Testing one regression parameter

If we set $\mathbf{C} = (0, 1, 0, \cdots, 0)^T$, a row vector with 1 in position 2 and 0 elsewhere, and $\mathbf{d} = (0, 0, \ldots, 0)$, a column vector with 0s, then we test

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0.$$

Now $\mathbf{C}\hat{\beta} = \beta_1$ and $\mathbf{C}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{C}^\mathsf{T} = c_{11}$, so that F_{obs} then is equal to the square of the t-statistics for testing a single regression parameter.

$$F_{obs} = (\hat{\beta}_1 - 0)^T[\hat{\sigma}^2 c_{jj}]^{-1}(\hat{\beta}_1 - 0) = T_1^2$$

Repeat the argument with β_j instead of β_1.
Remark: Remember that $T_\nu^2 = F_{1,\nu}$.

Testing “significance of the regression”

If we set $\mathbf{C} = (0, 1, 1, \cdots, 1)^T$, a row vector with 0 in position 1 and 0 elsewhere, and $\mathbf{d} = (0, 0, \ldots, 0)$, a column vector with 0s, then we test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ vs. } H_1 : \text{at least one different from zero.}$$

This means we test if at least one of the regression parameters (in addition to the intercept) is different from 0. The small model is then the model with only the intercept, and for this model the SSE_{H_0} is equal to SST (sums of squares total, see below). Let SSE be the sums-of-squares of errors for the full model. If we have k regression parameters (in addition to the intercept) then the F-statistic becomes

$$F_{obs} = \frac{\frac{1}{k}(\text{SST} - \text{SSE})}{\frac{\text{SSE}}{n-p}}$$

with k and $n - p$ degrees of freedom under H_0.

```r
library(gamlss.data)
```

Relation to Wald test

Since Cov$(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, then Cov$(\mathbf{C}\hat{\beta}) = \mathbf{C}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T$, so that $\mathbf{C}\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T$ can be seen as an estimate of Cov$(\mathbf{C}\hat{\beta})$. Therefore, F_{obs} can be written

$$F_{obs} = \frac{1}{r}(\mathbf{C}\hat{} - \mathbf{d})^{\mathsf{T}}[\widehat{\mathsf{Cov}}(\mathbf{C}\hat{\beta})]^{-1}(\mathbf{C}\hat{} - \mathbf{d}) = \frac{1}{r}W$$

where W is a socalled Wald test. It is known that $W \sim \chi^2_r$ asymptotically as n becomes large. We will study the Wald test in more detail later in this course.

Asympotic result

It can in general be shown that

$$rF_{r,n-p} \overset{n\to\infty}{\longrightarrow} \chi^2_r.$$

That is, if we have a random variable F that is distributed as Fisher with r (numerator) and $n-p$ (denominator) degrees of freedom, then when n goes to infinity (p kept fixed), then rF is approximately χ^2-distributed with r degrees of freedom.
Also, if our error terms are not normally distributed then we can assume that when the number of observation becomes very large then $rF_{r,n-p}$ is approximately χ^2_r.

Focus on likelihood: Likelihood ratio test and deviance

The likelihood ratio test

An alternative to the Wald test is the likelihood ratio test (LRT), which compares the likelihood of *two models*.

We use the following notation. A: the larger model and B: the smaller model (under H_0), and the smaller model is nested within the larger model (that is, B is a submodel of A).

- First we maximize the likelihood for model A (the larger model) and find the parameter estimate $\hat{\beta}_A$. The maximum likelihood is achieved at this parameter estimate and is denoted $L(\hat{\beta}_A)$.
- Then we maximize the likelihood for model B (the smaller model) and find the parameter estimate $\hat{\beta}_B$. The maximum likelihood is achieved at this parameter estimate and is denoted $L(\hat{\beta}_B)$.

The likelihood of the larger model (A) will always be larger or equal to the likelihood of the smaller model (B). Why?

The likelihood ratio statistic is defined as

$$-2\ln\lambda = -2(\ln L(\hat{\beta}_B) - \ln L(\hat{\beta}_A))$$

Analysis of variance decomposition and coefficient of determination, R^2

It is possible to decompose the total variability in the data, called SST (sums-of-squares total), into a part that is explained by the regression SSR (sums-of-squares regression), and a part that is not explained by the regression SSE (sums-of-squares error, or really residual).

Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, and $\hat{Y}_i = \mathbf{x}_i^T \hat{\beta}$. Then,

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \mathbf{Y}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{Y}$$

$$\text{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \mathbf{Y}^T(\mathbf{H} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{Y}$$

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}.$$

Analysis of variance tables - with emphasis on sequential Type I ANOVA

It is possible to call the function anova on an lm-object. What does that function do?

```
library(gamlss.data)
fit = lm(rent ~ area + location + bath + kitchen + cheating
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: rent
##             Df   Sum Sq  Mean Sq  F value    Pr(>F)
## area         1 40299098 40299098 1911.765 < 2.2e-16 ***
## location     2  1635047   817524   38.783 < 2.2e-16 ***
## bath         1  1676825  1676825   79.547 < 2.2e-16 ***
## kitchen      1  2196952  2196952  104.222 < 2.2e-16 ***
## cheating     1  7317894  7317894  347.156 < 2.2e-16 ***
## Residuals 3075 64819547    21080
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

What is produced is a *sequential* table of *the reductions in residual sum of squares (SSE) as each term in the regression formula is added in turn*. This type of ANOVA is often referred to as "Type 1" (not to be confused with type I errors).

We can produce the same table by fitting larger and larger regression models.

```r
library(gamlss.data)
fit = lm(rent ~ area + location + bath + kitchen + cheating
fit0 <- lm(rent ~ 1, data = rent99)
fit1 <- update(fit0, . ~ . + area)
fit2 <- update(fit1, . ~ . + location)
fit3 <- update(fit2, . ~ . + bath)
fit4 <- update(fit3, . ~ . + kitchen)
fit5 <- update(fit4, . ~ . + cheating)
anova(fit0, fit1, fit2, fit3, fit4, fit5, test = "F")
# anova(fit0,fit1) # compare model 0 and 1 - NOT sequentia
# anova(fit0,fit5) # compare model 0 and 5 - NOT sequentia
```

```
## Analysis of Variance Table
```

###Details on the test `anova(fit)`

When running `anova` on one fitted regression the F-test in `anova` is calculated as for “testing linear hypotheses” - but with a slight twist. Our large model is still the full regression model (from the fitted object), but the smaller model is replaced by the *the change from one model to the next*.

Let SSE be the sums-of-squares-error (residual sums of squares) from the full (large, called A) model - this will be our denominator (as always). For our rent example the denominator will be SSE/(n-p)=64819547/3075 (see print-out above).

However, for the numerator we are not comparing one small model with the full (large) one, we are instead looking at the change in SSE between two (smaller) models (calles model B1 and B2). So, now we have in the numerator the difference in SSE between models B1 and B2, scaled with the difference in number of parameters estimated in model B1 and B2 =“number in B2 minus in B1” (which is the same as the difference in degrees of freedom for the two models).

A competing way of thinking is called *type 3 ANOVA* and instead of looking sequentially at adding terms, we (like in `summary`) calculated the contribution to a covariate (or factor) given that all other covariates are present in the regression model. Type 3 ANOVA is available from library `car` as function `Anova` (possible to give type of anova as input).

Check : Take a look at the print-out from `summary` and `anova` and observe that for our rent data the p-values for each covariate are different due to the different nature of the H_0s tested (sequential vs. "all other present").

If we had orthogonal columns for our different covariates the type 1 and type 3 ANOVA tables would have been equal.

Optional (beyond the scope of this course)

There is also something called a type 2 ANOVA table, but that is mainly important if we have interactions in our model, so we do not consider that here. If you want to read more this blogplot http://goanna.cs.rmit.edu.au/~fscholer/anova.php is a good read. And, in combination with different variants of dummy and effct

Model choice

(read for yourself)

Quality measures

To assess the quality of the regression we can report the R^2 coefficient of determination. However, since adding covariates to the linear regression can not make the SSE larger, this means that adding covariates can not make the R^2 smaller. This means that SSE and R^2 are only useful measures for comparing models with the same number of regression parameters estimated.
If we consider two models with the same model complexity then SSE can be used to choose between (or compare) these models.
But, if we want to compare models with different model complexity we need to look at other measures of quality for the regression.

R^2 adjusted (corrected)

$$R^2_{\text{adj}} = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \frac{n-1}{n-p}(1 - R^2)$$

Choose the model with the *largest* R^2

AIC Akaike information criterion

AIC is one of the most widely used criteria, and is designed for likelihood-based inference. Let $l(\hat{\beta}_M, \tilde{\sigma}^2)$ be the maximum of the log-likelihood of the data inserted the maximum likelihood estimates for the regression and nuisance parameter. Further, let $|M|$ be the number of estimated regression parameters in our model.

$$\text{AIC} = -2 \cdot l(\hat{\beta}_M, \tilde{\sigma}^2) + 2(|M| + 1)$$

For a normal regression model:

$$\text{AIC} = n \ln(\tilde{\sigma}^2) + 2(|M| + 1) + C$$

where C is a function of n (will be the same for two models for the same data set). Remark that $\tilde{\sigma}^2 = SSE/n$ - our ML estimator (not our unbiased REML), so that the first term in the AIC is just a function of the SSE. For MLR the AIC and the Mallows Cp gives the same result when comparing models.
Choose the model with the minimum AIC.

BIC Bayesian information criterion.

The BIC is also based on the likelihood (see notation above).

R packages

```r
install.packages(c("gamlss.data", "tidyverse", "GGally", "N
```

References and further reading

- Slightly different presentation (more focus on multivariate normal theory): Slides and written material from TMA4267 Linear Statistical Models in 2017, Part 2: Regression (by Mette Langaas).
- And, same source, but now [Slides and written material from TMA4267 Linear Statistical Models in 2017, Part 3: Hypothesis testing and ANOVA] (http://www.math.ntnu.no /emner/TMA4267/2017v/TMA4267V2017Part3.pdf)