Module 2: MULTIPLE LINEAR REGRESSION
Week 2

TMA4315 Generalized linear models H2018

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

30.08.24

What to remember?

Model:

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

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

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

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

Restricted maximum likelihood estimator for σ^2:

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

with $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.

Statistic for inference

Statistic for inference about β_j, c_{jj} is diagonal element j of $(\mathbf{X}^T\mathbf{X})^{-1}$.

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t_{n-p}$$

This requires that $\hat{\beta}_j$ and $\hat{\sigma}$ are independent.

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.)

Confidence intervals (CI)

We focus on element j of β, called β_j. We know

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}}$$

follows a t-distribution with $n - p$ degrees of freedom.

Let $t_{\alpha/2,n-p}$ be such that $P(T_j > t_{\alpha/2,n-p}) = \alpha/2$.

Since the t-distribution is symmetric around 0, then $P(T_j < -t_{\alpha/2,n-p}) = \alpha/2$. We may then write

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

Pretty Pictures

(Blue lines at $\pm t_{\alpha/2, n-p}$.)

Solving for a CI

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 in R

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
## kitchen1     94.907671 145.9621578
## cheating1   144.427555 178.4000215
```

Prediction intervals

One aim for regression was to construct a model to predict the response from a set of (one or several) explanatory variables

Assume we want to make a prediction (Y_0) given specific values for the covariates, $\mathbf{x}_0$. 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.

Prediction intervals

First, assume the unobserved response at covariate $\mathbf{x}_0$ follows the same distribution, i.e. $Y_0 \sim N(\mathbf{x}_0^T\beta, \sigma^2)$ and is independent of our previous observations. 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

$$\mathsf{Var}(Y_0 - \mathbf{x}_0^T\hat{\beta}) = \mathsf{Var}(Y_0) + \mathsf{Var}(\mathbf{x}_0^T\hat{\beta}) = \sigma^2 + \sigma^2\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0$$

so that

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

Prediction intervals

Inserting our REML estimate for σ^2 gives

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

Prediction intervals

Then, we start with

$$P(-t_{\alpha/2,n-p} \leq \frac{Y_0 - \mathbf{x}_0^T\hat{\beta}}{\hat{\sigma}\sqrt{1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}} \leq t_{\alpha/2,n-p}) = 1-\alpha$$

and solve so that Y_0 is in the middle, which gives

$$P(\mathbf{x}_0^T\hat{\beta} - t_{\alpha/2,n-p}\hat{\sigma}\sqrt{1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0} \leq Y_0 \leq \mathbf{x}_0^T\hat{\beta} + t_{\alpha/2,n-p}\hat{\sigma}\sqrt{1+$$

A $(1-\alpha)$% PI for Y_0 is when we insert numerical values for the upper and lower limits:
$[\mathbf{x}_0^T\hat{\beta} - t_{\alpha/2,n-p}\hat{\sigma}\sqrt{1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}, \mathbf{x}_0^T\hat{\beta} + t_{\alpha/2,n-p}\hat{\sigma}\sqrt{1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}]$.

Prediction intervals

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 apartment with area 50 m^2, 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")
```

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

Questions:

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

Hypothesis Testing

General Strategy

- Invent a hypothesis
  - Then derive H_0 and H_1
- find a test statistic to distinguish between H_0 and H_1
- work out the distribution of test statistic under H_0
- Calculate the p-value, i.e. the probability of the data (or something more extreme) under H_0
- Report the result of the test

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.

Invent a hypothesis

In linear regression the hypothesis is often about a regression parameter β_j, i.e.

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

find a test statistic to distinguish between H_0 and H_1

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

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

where $c_{jj}\hat{\sigma}_\varepsilon^2 = \widehat{\mathsf{Var}}(\hat{\beta}_j)$.

Inserted observed values (and estimates) we have t_0.

(more on test statistics shortly)

Work out the distribution of test statistic under H_0

We know that

$$T_0 \sim t_{n-2}$$

so this is easy.

Calculate the p-value, i.e. the probability of the data (or something more extreme) under H_0

```r
pt(...)
```

The p-value

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

- In single hypothesis testing, if the p-value is less than the chosen significance level, 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.
- 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.

Reporting the Result: Strict interpretation

We choose to reject H_0 at some significance level α if the p-value of the test 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.

Strict Interpretations of significance tests

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?

Two types of errors:

- “Reject H_0 when H_0 is true”=“false positives” = “type I error”
- “Fail to reject H_0 when H_1 is true (and H_0 is false)”=“false negatives” = “type II error”

STOPPED HERE

Test Statistics

We can use any statistic that will distinguish between H_0 and H_1, but some are better than others. A good test statistic will

- be good at distinguishing between H_0 and H_1
  - discussed as the power of the test
- have a nice distribution under the null hypothesis
  - if not we can simulate our way to a solution

In this course we will use three approaches to developing test statistics:

- Wald tests
- Likelihood Ratio Tests
- Score Tests

Wald Tests

Wald statistics measure the standardised distance between H_0 and H_1 in the parameter space, i.e. between $\beta_j = 0$ and $\hat{\beta}_j$.

For our single regression parameter we can use

$$W = \frac{(\hat{\beta}_j - \beta_j)^2}{\widehat{\mathsf{Var}}(\hat{\beta}_j)},$$

Note that this is calculated at H_1

Likelihood Ratio Tests

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).

LRT: model fitting

- 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?

Defining the LRT

The likelihood ratio statistic is defined as

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

(so, -2 times small minus large).

Under weak regularity conditions the test statistic is approximately χ^2-distributed with degrees of freedom equal the difference in the number of parameters in the large and the small model. This is general - and not related to the GLM! More in TMA4295 Statistical Inference!

P-values are calculated in the upper tail of the χ^2-distribution.

The LRT can be performed using `anova()`.

Deviance (something new!)

The *deviance* is used to assess model fit and also for model choice, and is based on the likelihood ratio test statistic. It is used for all GLMs in general - and replaces using SSE in multiple linear regression.

Saturated model: If we were to provide a perfect fit to our data t

This "imaginary model" is called the *saturated* model. This would be a model where each observation was given its own parameter.

Candidate model: The model that we are investigated can be thought of as a *candidate* model. Then we maximize the likelihood and get $\hat{\beta}$.

Deviance

The *deviance* is then defined as the likelihood ratio statistic, where we put the saturated model in place of the larger model A and our candidate model in place of the smaller model B:

$$D = -2(\ln L(\text{candidate model}) - \ln L(\text{saturated model}))$$

For the maximal model, we have parameters $\theta_1, \dots, \theta_n$, where $\theta_i = E[Y_i]$ The log-likelihood for this model is

$$l(\hat{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \theta)^T(\mathbf{y} - \theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2$$

(because $\hat{\theta}_i = y_i$, something you may work out for yourselves)

Deviance

The deviance is then

$$\begin{aligned} D &= -2\left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta}) - (-\frac{n}{2}\ln(2\pi\sigma^2)\right) \\ &= \frac{1}{\sigma^2}(\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta}) \end{aligned}$$

Note the connection with the RSS! Under the null hypothesis that the model fits the data well, $D \sim \chi^2_{n-p}$ exactly (in this case).

Score Tests

Our third test is a *score test*. This uses (as you might have guessed) the score, i.e. the partial derivative of the log likelihood. the intuition is that at the MLE the slope is 0, so if the score at the null hypothesis is too large, we can reject the null.

More formally, the test statistic for a score test is

$$u(\beta_0) = s^T(\beta_0)F^{-1}(\beta_0)s(\beta_0)$$

where $s(\beta_0)$ is the score, and $F(\beta_0)$ the expected Fisher information. Note that these are calulated at the value of β_0 from the null hypothesis, we do not need to calculate it at $\hat{}$.

Normal Score tests

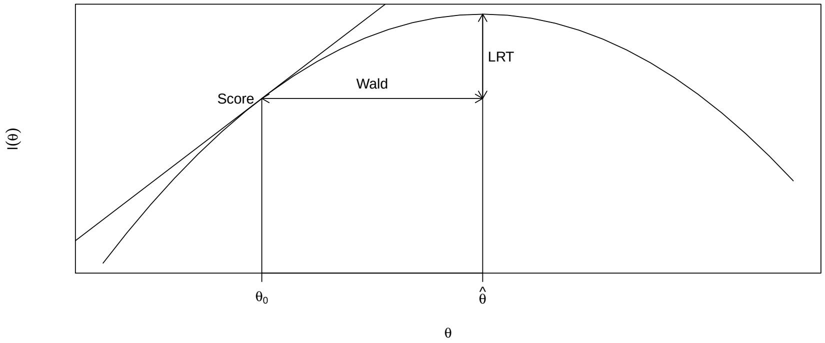
For the normal distribution the score is $s(\beta) = \Sigma^{-1}(X^T X \ - X^T \mathbf{y})$ and the Fisher information is $\Sigma^{-1}(X^T X)$

So

$$
\begin{aligned}
u(\beta_0) &= \sigma^{-2}(X^T X \ - X^T \mathbf{y})^T \sigma^{-2}((X^T X))^{-1}(\sigma^{-2} X^T X \ - X^T \mathbf{y}) \\
&= \sigma^{-2}(X^T \mu - X^T \mathbf{y})^T (X^T X)^{-1}(X^T \mu - X^T \mathbf{y}) \\
&= \sigma^{-2}(\mu - \mathbf{y})^T X (X^T X)^{-1} X^T (\mu - \mathbf{y})
\end{aligned}
$$

(because σ^{-2} is a scalar, and $X \ $ is the predicted value of y)

Summary of the Tests



Munich rent index hypothesis test

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

```
library(gamlss.data)
fit = lm(rent ~ area + location + bath + kitchen + cheating
knitr::kable(summary(fit)$coefficients, digits = 3)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-21.973	11.655	-1.885	0.059
area	4.579	0.114	40.055	0.000
location2	39.260	5.447	7.208	0.000
location3	126.057	16.875	7.470	0.000
bath1	74.054	11.209	6.607	0.000
kitchen1	120.435	13.019	9.251	0.000
cheating1	161.414	8.663	18.632	0.000

Question

Q (and A):

1. Where is hypothesis testing performed here, and which are the hypotheses rejected at level 0.01?
2. Will the test statistics and p-values change if we change the regression model?
3. What is the relationship between performing an hypothesis test and constructing a CI interval?

More Complex Testing of 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 hypothesis is true) model B, or the smaller model.

Hypotheses

These null hypotheses and alternative hypotheses can all be rewritten as a linear hypothesis

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

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

The Wald test and LRT statistics are the same: the test is called F_{obs} and so can be formulated in two ways:

$$F_{obs} = \frac{\frac{1}{r}(SSE_{H_0} - SSE)}{\frac{SSE}{n-p}} \tag{1}$$

$$F_{obs} = \frac{1}{r}(\mathbf{C}\hat{\beta} - \mathbf{d})^\mathsf{T}[\hat{\sigma}^2\mathbf{C}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{C}^\mathsf{T}]^{-1}(\mathbf{C}\hat{\beta} - \mathbf{d}) \tag{2}$$

where SSE is from the larger model A, SSE_{H_0} from the smaller

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 $\mathbf{d} = (0, 0, \dots, 0)$, a column vector with 0s, then we test

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs. H_1 : 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

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

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

Significant??

Is the regression significant?

```r
summary(fit)$fstatistic
```

```
##     value     numdf     dendf
##  420.0427    6.0000 3075.0000
```

Relation to Wald test

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

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

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.

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).

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

$$\begin{aligned}
\mathsf{SST} &= \mathsf{SSR} + \mathsf{SSE} \\
\mathsf{SST} &= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \mathbf{Y}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{Y} \\
\mathsf{SSR} &= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \mathbf{Y}^T(\mathbf{H} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{Y} \\
\mathsf{SSE} &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}.
\end{aligned}$$

R^2

Based on this decomposition we may define the *coefficient of determination* (R^2) as the ratio between SSR and SST, that is

$$R^2 = \text{SSR}/\text{SST} = 1 - \text{SSE}/\text{SST}$$

1. The interpretation of this coefficient is that the closer it is to 1 the better the fit to the data. If $R^2 = 1$ then all residuals are zero - that is, perfect fit to the data.

2. In a simple linear regression the R^2 equals the squared correlation coefficient between the reponse and the predictor. In multiple linear regression R^2 is the squared correlation coefficient between the observed and predicted response.

3. If we have two models M1 and M2, where model M2 is a submodel of model M1, then
$$R^2_{M_1} \geq R^2_{M_2}.$$
This can be explained from the fact that $\text{SSE}_{M_1} \leq \text{SSE}_{M_2}$. (More in the Theoretical questions.)

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)
fit1 = lm(rent ~ area + location + bath, data = rent99)
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: rent
##             Df   Sum Sq  Mean Sq  F value    Pr(>F)
## area         1 40299098 40299098 1668.142 < 2.2e-16 ***
## location     2  1635047   817524   33.841 2.901e-15 ***
## bath         1  1676825  1676825   69.410 < 2.2e-16 ***
## Residuals 3077 74334393    24158
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Sequential

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.

```
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)
```

ANOVAs

```
anova(fit0, fit1, fit2, fit3, 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
##
## Model 1: rent ~ 1
## Model 2: rent ~ area
## Model 3: rent ~ area + location
## Model 4: rent ~ area + location + bath
##   Res.Df       RSS Df Sum of Sq        F    Pr(>F)
## 1   3081 117945363
## 2   3080  77646265  1  40299098 1668.142 < 2.2e-16 ***
## 3   3078  76011217  2   1635047   33.841 2.901e-15 ***
## 4   3077  74334393  1   1676825   69.410 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Change the Order

If we had changed the order of adding the covariates to the model, then our anova table might also change. You might check that if you want.

Notes

See the last page of the classnotes 04.09.2017 for mathematical notation on the sequential test in `anova`, and details on the print-out comes next - NEW: now with formulas!

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 above).

The logic is that the full model provides an estimate of σ^2: the others may not.

T to Normality

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 (called 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).

Test Stats

This means that the test statistics we use are:

$$F_0 = \frac{\frac{\mathsf{SSE}_{B1} - \mathsf{SSE}_{B2}}{\mathsf{df}_{B1} - \mathsf{df}_{B2}}}{\frac{\mathsf{SSE}_A}{\mathsf{df}_A}}$$

Remark: notice that the denominator is just the $\hat{\sigma^2}$ from the larger model A.

This makes our F-test statistic: $f_0 = \frac{40299098/1}{64819547/3075} = 1911.765$ (remember that we swap from capital to small letters when we insert numerical values).

To produce a p-value to the test that

H_0 : "Model B1 and B2 are equally good" vs H_1 : "Model B2 is better tl

and then the $F \sim \mathsf{df}_{B1} - \mathsf{df}_{B2}, \mathsf{df}_A$.

Example

In our example we compare to an F-distribution with 1 and 3075 degrees of freedom. The p-value is the "probability of observing a test statistic at least as extreme as we have" so we calculate the p-value as $P(F > f_0)$. This gives a p-value that is practically 0.

If you then want to use the asymptotic version (relating to a chi-square instead of the F), then multiply your F-statistic with $\text{df}_{B1} - \text{df}_{B2}$ and relate to a χ^2 distribution with $\text{df}_{B1} - \text{df}_{B2}$ degrees of freedom, where $\text{df}_{B1} - \text{df}_{B2}$ is the difference in number of parameters in models B1 and B2. In our example $\text{df}_{B1} - \text{df}_{B2} = 1$.

Anova

For the anova table we do this sequentially for all models from starting with only intercept to the full model A. This means you need to calculate SSE and df for models of all sizes to calculate lots of these F_0s. Assume that we have 4 covariates that are added to the model, and call the 5 possible models (given the order of adding the covariates)

- model 1: model with only intercept
- model 2: model with intercept and covariate 1
- model 3: model with intercept and covariate 1 and covariate 2
- model 4: model with intercept and covariate 1 and covariate 2 and covariate 3
- model 5: model with intercept and covariate 1 and covariate 2 and covariate 3 and covariate 4

Output

Fit a linear model (lm) for each model 1-5, and store SSE and degrees of freedom=df (number of observations minus number of covariates estimated) for each of the models. Call these SSE_1 to SSE_5 and df_1 to df_5$.

The anova output has columns: Df Sum Sq Mean Sq F value Pr(>F) and one row for each covariate added to the model.

For example

model 2 vs model 1: Df=df_1-df_2, Sum Sq=SSE_1-SSE_2, Mean Sq=Sum Sq/Df, F value=(Mean Sq)/(SSE_5/df_5)=f_0, Pr(>F)=pvalue=$P(F > f_0)$.

model 3 vs model 2: Df=df_2-df_3, Sum Sq=SSE_2-SSE_3, Mean Sq=Sum Sq/Df, F value=(Mean Sq)/(SSE_5/df_5)=f_0, Pr(>F)=pvalue=$P(F > f_0)$.

In R the p-value is calculated as `1-pf(f0,Df)` or as `1-pchisq(Df*f0,Df)` if the asymptotic chisquare distribution is used

Question

**Q*: What if you change the order of the covariates into the model?

Type III ANOVA

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.

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)

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

Choose the model with the *largest* R^2_{adj}.

But, if we are choosing model, we should use hypothesis tests, or techniques to comapre models!

Information Criteria

Rather than find the best model by hypothesis testing, we can instead use a measure of model adequacy (defined somehow!) to compare models.

There are different ways of defining model adequacy, which lead to different statistics

AIC Akaike information criterion

AIC is one of the most widely used criteria, and is designed for likelihood-based prediction. 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.

Choose the model with the minimum AIC.

BIC Bayesian information criterion.

The BIC is also based on the likelihood, nad is designed for inference (finding the "true" model).

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

For a normal regression model:

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

Choose the model with the minimum BIC.

BIC has a larger penalty than AIC ($\log(n)$ vs. 2), and will often give a smaller model (=more parsimonious models) than AIC.

In general we would not like a model that is too complex.

Model selection strategies

- All subset selection: use smart “leaps and bounds” algorithm, works fine for number of covariates in the order of 40.
- Forward selection: choose starting model (only intercept), then add one new variable at each step - selected to make the best improvement in the model selection criteria. End when no improvement is made.
- Backward elimination: choose starting model (full model), then remove one new variable at each step - selected to make the best improvement in the model selection criteria. End when no improvement is made.
- Stepwise selection: combine forward and backward.

R packages

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

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)