Practical 1

Mikis Stasinopoulos Bob Rigby Gillian Heller
Fernanda De Bastiani Niki Umlauf

The 1980s Munich rent data

The `rent` data come from a survey conducted in April 1993 by Infratest Sozialforschung, where a random sample of accommodation with new tenancy agreements or increases of rents within the last four years in Munich was selected, including single rooms, small apartments, flats and two-family houses. The data were analysed by Stasinopoulos, Rigby, and Fahrmeir (2000) and they are in the package `gamlss.data` (which is automatically loaded when `gamlss` is loaded). There are 1,969 observations on nine variables in the data set but, for the purpose of demonstrating GAMLSS, we will use only the following five variables:

```
library(gamlss.ggplots)
library(broom)
library(knitr)
library(gamlss.ggplots)
# remove two variables
da <- rent[, -c(4,5, 6, 8)]
da  |> head() |> kable(digits = c(2, 0, 0, 0, 0,0,0), format="pipe")
```

Table 1: Variables in Munich rent data

R	Fl	A	H	loc
693.3	50	1972	0	2
422.0	54	1972	0	2
736.6	70	1972	0	2
732.2	50	1972	0	2
1295.1	55	1893	0	2
1195.9	59	1893	0	2

```
library(gamlss.prepdata)
data_xyplot(da, response=R)
```

```
100 % of data are ploted,
that is, 1969 observations.
```

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

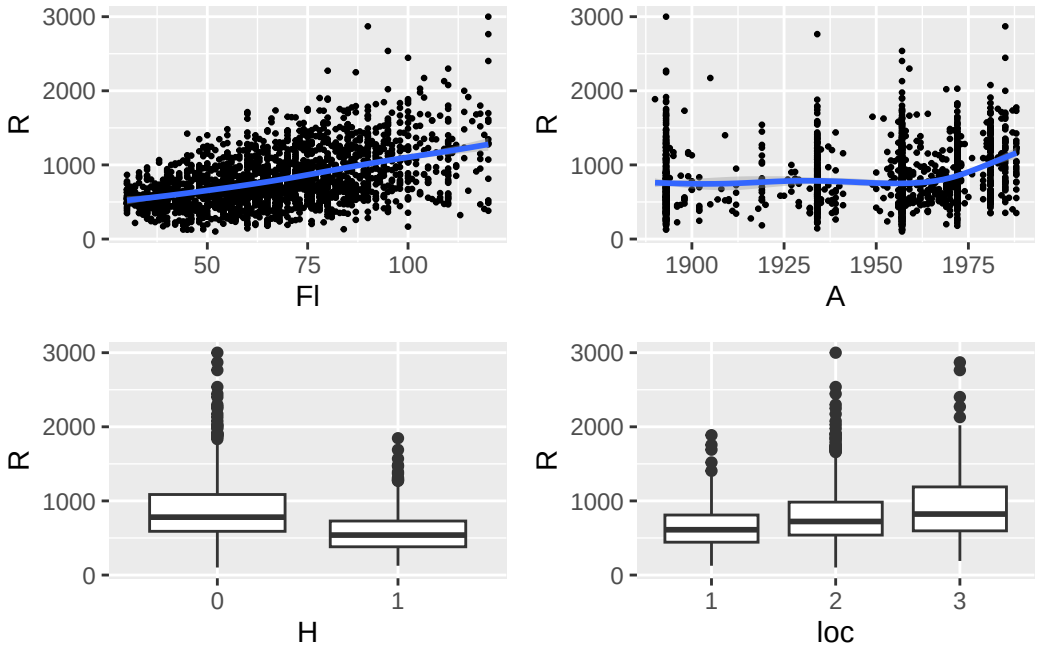


Figure 1: Plot of the rent `R` against explanatory variables `Fl`, `A`, `H` and `loc`.

Figure **??** shows plots of the rent, `R`, against each of the explanatory variables. Although these are bivariate exploratory plots and take no account of the interplay between the explanatory variables, they give an indication of the complexity of these data. The first two explanatory variables, `Fl` and `A`, are continuous. %the plots also show exploratory univariate

The plot of rent, `R`, against floor space, `Fl`, suggests a positive relationship, with increased variation for larger floor spaces, with the result that an assumption of homogeneity of variance would be violated here. There is also some indication of positive skewness in the distribution of rent, `R`. The peculiarity of the plot of `R` against year of construction, `A`, is due to the method of data collection. Many of the observations of `A` were collected on an interval scale and assigned the value of the interval midpoint, while for the rest the actual year of construction

was recorded. The plot suggests that for flats up to 1960 the median rent is roughly constant but, for those constructed after that year, there is an increasing trend in the median rent. The two boxplots display how the rent varies according to the explanatory factors. The median rent increases if the flat has central heating, and increases as the location changes from below average to average and then to above average. There are no surprises in the plots here but again the problem of skewness is prominent, with asymmetrical boxes about the median and longer upper than lower whiskers.

In summary, any statistical model used for the analysis of the rent data should be able to deal with the following statistical problems:

- **Complexity of the relationship between rent and the explanatory variables**. The dependence of the median of the response variable rent on floor space and age of construction is nonlinear, and nonparametric smoothing functions may be needed. Median rent may also depend on linear or nonlinear interactions between the explanatory variables.
- **Non-homogeneity of variance of rent.** There is clear indication of non-homogeneity of the variance of rent. The variance of rent may depend on its mean and/or explanatory variables. A statistical model in which this dependence can be modelled explicitly, is needed.
- **Skewness in the distribution of rent.** There is clear indication of positive skewness in the distribution of rent which may depend on explanatory variables and this has to be accounted for within the statistical model.

The linear regression model

Linear regression is a simple but effective model, which served the statistical community well for most of the last century. With response variable Y, r covariates $x_1, ..., x_r$ and sample size n, it is defined as

$$\begin{aligned} & Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_r x_{ir} + \epsilon_i \\ \text{where} \qquad & \epsilon_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2) \ , \qquad \text{for } i = 1, 2, ..., n \end{aligned}$$

i.e. ϵ_i for $i = 1, 2, ..., n$ are independently distributed each with a normal distribution with mean zero and variance σ^2. This specification is equivalent to

$$\begin{aligned} & Y_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma^2) \\ \text{where} \qquad & \mu_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_r x_{ir} \ , \qquad \text{for } i = 1, 2, ..., n \ . \end{aligned}$$

We rewrite model (Equation **??**) in matrix form as:

$$\begin{aligned} \mathbf{Y} &\overset{\text{ind}}{\sim} \mathcal{N}(\mu, \sigma^2 \mathbf{I}) \\ \mu &= \mathbf{X}\beta \end{aligned}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is the response vector, $\mathbf{X}$ is the $n \times p$ design matrix ($p = r + 1$) containing the r covariate columns, plus a column of ones (if the constant is required), $\beta = (\beta_0, \ldots, \beta_r)^\top$ is the coefficient vector, and $\mu = (\mu_1, \ldots, \mu_n)^\top$ is the mean vector. Note that in order for the model to be fitted, both β and σ^2 have to be estimated from the data. The usual practice is to estimate β using the least squares estimator, obtained by minimizing the sum of squared differences between the observations y_i and the fitted means $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_r x_{ir}$, with respect to the $\hat{\beta}$'s . In matrix form this is written as In matrix form this is written as

$$\hat{\beta} = \text{argmin}_{\hat{\beta}} (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta})$$

which has solution

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

It can be shown that $\hat{\beta}$ is also the maximum likelihood estimator (MLE) of β. Let

$$\hat{\mu} = \mathbf{X}\hat{\beta}$$

be the fitted values of the model and $\hat{\epsilon} = \mathbf{Y} - \hat{\mu}$ the standard residuals (i.e. fitted errors). Then the MLE for σ^2 is

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}^\top \hat{\epsilon}}{n} \ ,$$

which is a biased estimator, i.e. $E(\hat{\sigma}^2) \neq \sigma^2$. An unbiased estimator of σ^2 is given by

$$s^2 = \frac{\hat{\epsilon}^\top \hat{\epsilon}}{n - p} \ .$$

Sometimes s^2 is referred as the REML (Restricted Maximum Likelihood) estimator of σ^2.

A linear regression model can be fitted in `R` using the function `lm()`. Here we compare the results from `lm()` to the ones obtained by `gamlss2()`. The notation

```
R ~ Fl+A+H+loc
```

refers to a formula in `R` for more information type `?formula`.

```
library(gamlss2)
r1 <-  gamlss2(R ~ Fl+A+H+loc, family=NO,  data=rent, trace=FALSE)
l1 <- lm(R ~ Fl+A+H+loc,data=rent)
coef(r1)
```

```
   mu.p.(Intercept)             mu.p.Fl              mu.p.A             mu.p.H1
       -2775.038803            8.839445            1.480755         -204.759562
          mu.p.loc2           mu.p.loc3 sigma.p.(Intercept)
          134.052349          209.581472            5.731647
```

```
coef(l1)
```

```
 (Intercept)           Fl            A           H1         loc2         loc3
-2775.038803     8.839445     1.480755  -204.759562   134.052349   209.581472
```

The coefficient estimates for the μ parameter of the two fits are identical. Note the `gamlss2` produce an extra coefficient from the variance model which is a constant. Note that the two factors of the `rent` data, `H` and `loc`, are fitted as dummy variables as explained in more detail in later section Section.

The fitted objects `r1` and `l1` use the methods `fitted()` and `resid()` to obtain fitted values and residuals respectively. Note that the `gamlss2` object residuals are the normalized (randomized) quantile residuals as explained in the lecture and not the usual residuals $\hat{\epsilon}$ that might be expected.

The MLE of σ can be obtained from a `gamlss2` fitted object using the command `fitted(r1, type="parameter", what="sigma")[1]`. (Here `[1]` shows the first element of the fitted vector for σ) since it is constant for all observations. `summary()` will show the standard errors and t-tests of the estimated coefficients. The method used to calculate standard errors in the `summary()` function of a `gamlss2` model are the standard methods based on the second derivative of the likelihood function.

```
head(fitted(r1, type="parameter"),5)
```

```
        mu    sigma
1 721.0349 308.4768
2 756.3927 308.4768
3 897.8238 308.4768
4 721.0349 308.4768
5 648.2525 308.4768
```

```
summary(r1)
```

```
Call:
gamlss2(formula = R ~ Fl + A + H + loc, data = rent, family = NO,
    ... = pairlist(trace = FALSE))
```

```
---
Family: NO
Link function: mu = identity, sigma = log
*--------
Parameter: mu
---
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2775.0388   526.8431  -5.267 1.54e-07 ***
Fl              8.8394     0.3386  26.108  < 2e-16 ***
A               1.4808     0.2673   5.540 3.43e-08 ***
H1           -204.7596    19.3784 -10.566  < 2e-16 ***
loc2          134.0523    25.1343   5.333 1.07e-07 ***
loc3          209.5815    27.1218   7.727 1.74e-14 ***
*--------
Parameter: sigma
---
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.73165    0.01594   359.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
*--------
n = 1969 df =  7 res.df =  1962
Deviance = 28159.0039 Null Dev. Red. = 2.8%
AIC = 28173.0039 elapsed =  0.01sec
```

The fitted model is given by

$$y \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$$

where

$$\begin{aligned} \hat{\mu} &= -2775.03 + 8.83\,Fl + 1.48\,A - 204.75\texttt{if H=1)} + \\ &\quad + 134.0\texttt{if loc=2)} \quad + 209.5\texttt{(if loc=3)} \\ \log(\hat{\sigma}) &= 5.73 \end{aligned}$$

Note that σ is fitted on the log scale (indicated by the log link function, so its fitted value is computed from its intercept as

$$\hat{\sigma} = \exp(5.73).$$

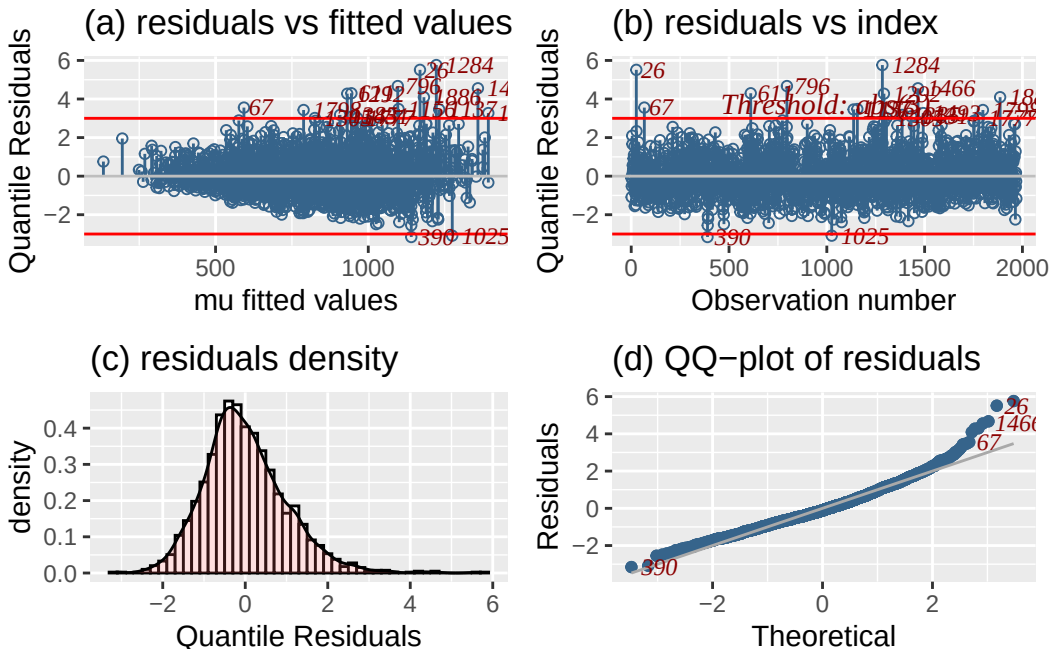
R^2 is obtained from the `gamlss` fitted object as

```
Rsq(r1)
```

```
[1] 0.3372029
```

One way of checking the adequacy of a model is to examine the residuals.

```
resid_plots(r1)
```



The important issue here is that the distributional assumption of normality is easily rejected by looking at the normal Q-Q plot (bottom right panel, Figure **??**). There is a systematic departure from a linear relationship between the observed (normalized quantile) residuals and their approximate expected values, indicating that the residuals are positively skewed. Note also that the plot of residuals against fitted values (top left panel, Figure **??**) is not randomly scattered about a horizontal line at 0, but fans out, indicating variance heterogeneity, in particular that the variance increases with the mean.

Given that the normal (or Gaussian) assumption is violated because of the positive skewness, we consider the generalized linear model next.

The generalized linear model (GLM)

The generalized linear model (GLM) was introduced by Nelder and Wedderburn (1972) and further developed in McCullagh and Nelder (1989). There are three major innovations in their