KU LEUVEN

MASTER OF STATISTICS

[G0A18A]

Project

2018-19 GLM Course KULeuven

Vincent Buekers
(vincent.buekers@student.kuleuven.be)
John Valen
(johntheodore.valen@student.kuleuven.be)
Sriram Sivaraman
(srirammeyappan.sivaraman@student.kuleuven.be)
Yao Kou
(yao.kou@student.kuleuven.be)

June 2019

Bike Sharing

In this first part of the project, we examine the effect of several covariates on the number of bike rentals per day. The variables *temperature, wind speed,* and *humidity* are considered as possible continuous predictors, while the *weather situation* and the *seasons* are considered as categorical predictors. Lastly, a binary variable indicating whether or not it is a working day is also included in the covariate space. Various techniques are implemented in order to obtain an appropriate model that allows for reliable predictions regarding the amount of bikes rented per day. The following R packages are used: `MASS, mgcv, lmtest, sandwich.`

1. ***Set up a Poisson regression model that predicts the number of total rental bikes (cnt) in a frequentist manner with the R software used in the course.***

As instructed, a Poisson regression model including the aforementioned predictors is considered in order to model the number of bike rentals per day. Note that a categorical approach has been taken with respect to the *weather situation* and the *season* to adequately represent the differences between the respective levels, rather than to obtain a general estimate for those differences. Moreover, the continuous covariates - *temperature* (°C), *humidity* and *wind speed* - are already on normalized scales (i.e. divided by the highest observed value). For now, we will keep these normalizations as they are. The resulting model output is displayed in Figure 1.1.

```
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept)  7.377159   0.006979 1057.038   <2e-16 ***
weathersit2 -0.095911   0.002440  -39.308   <2e-16 ***
weathersit3 -0.693882   0.007047  -98.468   <2e-16 ***
workingday   0.017148   0.001954    8.778   <2e-16 ***
temp         1.209646   0.008699  139.053   <2e-16 ***
hum         -0.227331   0.008817  -25.784   <2e-16 ***
windspeed   -0.610726   0.013434  -45.462   <2e-16 ***
season2      0.503511   0.003917  128.555   <2e-16 ***
season3      0.451313   0.004753   94.957   <2e-16 ***
season4      0.629664   0.003496  180.124   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 228892  on 364  degrees of freedom
Residual deviance:  48542  on 355  degrees of freedom
AIC: 52160
```

Figure 1.1 : Linear Poisson Model Output

While one might be satisfied with the high degree of significance associated with each covariate, the residual deviance reveals the issue of overdispersion for the assumed Poisson distribution. Under the correct model, the residual deviance should approximate the expected value of a Chi-squared distribution, to which it converges in distribution. Indeed $D = 48{,}542$ is not even remotely close to $n - p$ = 355. Consequently, the model fit is inappropriate as the model inadequately captures the underlying variance in the data.

2. ***Select the most predictive regressors in a classical GLM manner and interpret your results.***

To determine whether variables from the previously obtained Poisson model should be dropped, one can use the likelihood ratio test as an appropriate test for nested models. The p-values associated with the single deletions of predictors indicate that they are all significant components of the model; however, the LRT also points out that some covariates tend to be more predictive than others.

```
Single term deletions

Model:
cnt ~ weathersit + workingday + temp + hum + windspeed + season
            Df Deviance   AIC   LRT  Pr(>Chi)
<none>           48542 52160
weathersit  2    59870 63484 11328 < 2.2e-16 ***
workingday  1    48620 52235    77 < 2.2e-16 ***
temp        1    67969 71585 19427 < 2.2e-16 ***
hum         1    49204 52820   661 < 2.2e-16 ***
windspeed   1    50624 54240  2082 < 2.2e-16 ***
season      3    85294 88905 36751 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1.2: Single Term Deletions for Full Poisson Model from figure 1.1

As the objective is to select the most predictive regressors, we consecutively remove the variables *working day*, then *hum*, and lastly *wind speed*. These single deletions are based on the LRT results at each deletion step (for the other deletion steps we refer to the R code). The remaining covariates - which are supposed to be the most predictive - are *weather, temperature* and *season*. While the model including *weather*, *temperature* and *season* can be improved upon, we may draw a few preliminary conclusions. Below are the parameter estimates for the linear Poisson model containing only the aforementioned most predictive regressors (rounded to 3 decimal points).

Variable	Estimate
Intercept	7.142
Weather situation 2	-0.136
Weather situation 3	-0.787
Temperature	1.204
Season = Spring	0.499
Season = Summer	0.462
Season = Fall	0.640

Table 1.1: Linear Poisson Model with Covariates `weather`, `temperature`, `season`

The estimated mean number of rental bikes rented during winter (*season 1*), when it is clear or partly cloudy (*weather 1*), and *temperature* (normalized) is equal to zero is exp(7.142), or about 1,263 bikes. Keeping all other covariates equal, this estimated mean in bikes rented decreases multiplicatively as the weather worsens (i.e. *weather 2, 3)*. Conversely, as seasons progress from winter (*reference category)* to fall (*season 4)*, the expected number of bikes rented increases as one would expect. Lastly, the temperature plays an important role as well. That is, for a one-unit increase in normalized temperature, the expected amount of rental bikes increases by a factor of exp(1.20) = 3.32, assuming *season* and *weather* are unchanged. However, an increase by 1 unit doesn't make sense on the normalized temperature scale as this would mean that the temperature increases from 0 to 41 °C.

Note that these interpretations all surface from an inadequate linear Poisson model that can be improved further. Before addressing problems such as the distribution and the covariate scales, we first examine an extended additive model in the following section.

3. *Also select the most predictive regressors in an extended GAM manner.*

Next, a generalized additive model (GAM) is considered. The continuous covariates are modelled via a non-linear relationship by means of smoothing techniques. As such, the objective is to find out if a smooth relationship between the continuous covariates (*windspeed*, *temp* and *hum*) and the expected amount of bike rentals is appropriate.

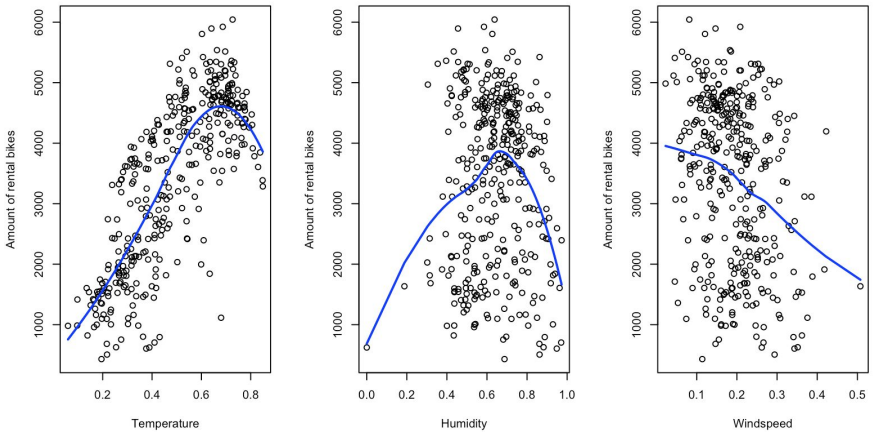

Figure 1.3: Plots of Continuous Covariates vs. cnt

As seen in Figure 1.3, smoothing the relationship between *temp* and log(*cnt*) seems appropriate, whereas it less clear with *hum* and *wind speed*. The non-linearity in *hum* (center plot) arises due to the two lowest humidity values which seem to be outliers in the X-space. The covariate *wind speed* (right plot) does not look like an appropriate candidate for smoothing.

To prevent overfitting we use penalized splines based on B-splines, which penalizes excessive curvature. Although we do not need to specify the number of knots and their position with this approach, we must set the penalty parameter λ. However, one can also rely on the generalized cross-validation criterion (GCV) to obtain the optimal lambda rather

than to set it manually. Note that the additive Poisson model is still implemented with the canonical log link. As such, the results shown in Figure 1.4 are obtained.

```
Parametric coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept)  7.776667   0.003592 2164.827  < 2e-16 ***
weathersit2 -0.067790   0.002551  -26.575  < 2e-16 ***
weathersit3 -0.482194   0.009019  -53.466  < 2e-16 ***
workingday   0.013040   0.002053    6.351 2.14e-10 ***
season2      0.339816   0.004197   80.960  < 2e-16 ***
season3      0.366697   0.004927   74.420  < 2e-16 ***
season4      0.471300   0.003905  120.702  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
               edf Ref.df Chi.sq p-value
s(temp)      8.989  9.000  35206  <2e-16 ***
s(hum)       7.966  7.999   6045  <2e-16 ***
s(windspeed) 8.997  9.000   5856  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.876   Deviance explained =   88%
UBRE =  74.65  Scale est. = 1         n = 365
```

Figure 1.4: Model Output for Poisson GAM with Smooth Continuous Covariates

The GAM with smoothing of the continuous covariates seems to explain a large portion of variability and produces multiple significant covariates; however, we must investigate the nature of the smoothed relationships before interpretation. By examining the partial residuals, it is immediately apparent that smoothing overfits the outliers in *hum* (center plot) and *wind speed* (rightmost plot).

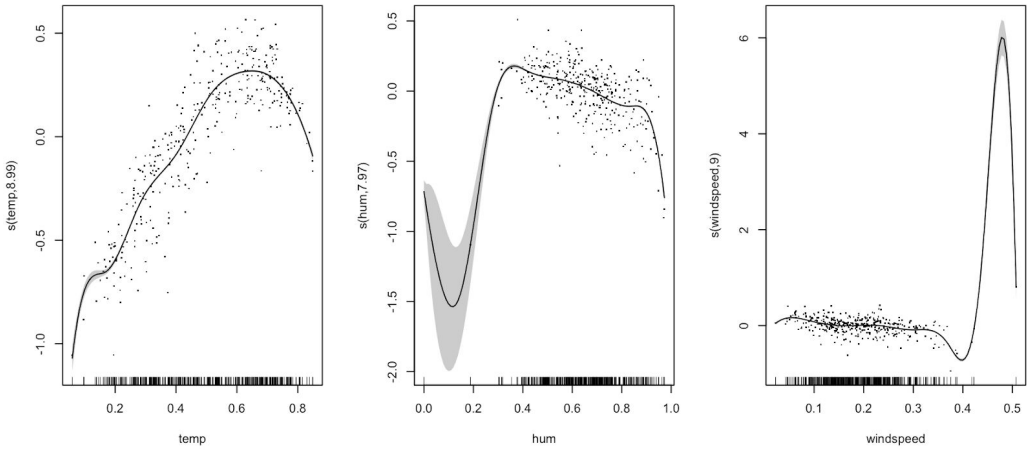

Figure 1.5: Partial Residual Plots for Smoothed Terms

It could be justified to remove the 2 humidity outliers and the single wind speed outlier. Note that these outliers are, in fact, only from 2 observations as one of them has extreme observations for both *hum* and *windspeed*.

Additionally, one could decrease the amount of knots to reduce the wiggliness of the smoothed relationships. The knot amounts were set to 10 for both *temp* and *windspeed*, and to 9 for *hum* via the automated GCV. In a next step, these are lowered to 5 for *temp* and

windspeed, and 4 for *hum* to reduce the wiggliness of smoothing. By reducing the knots and removing the influential observations, the improvements are quite apparent as depicted in Figure 1.6. The covariate *temp* is now fitted using a nicely smoothed function, whereas the need to smooth *windspeed* or *humidity* seems to have partially disappeared.

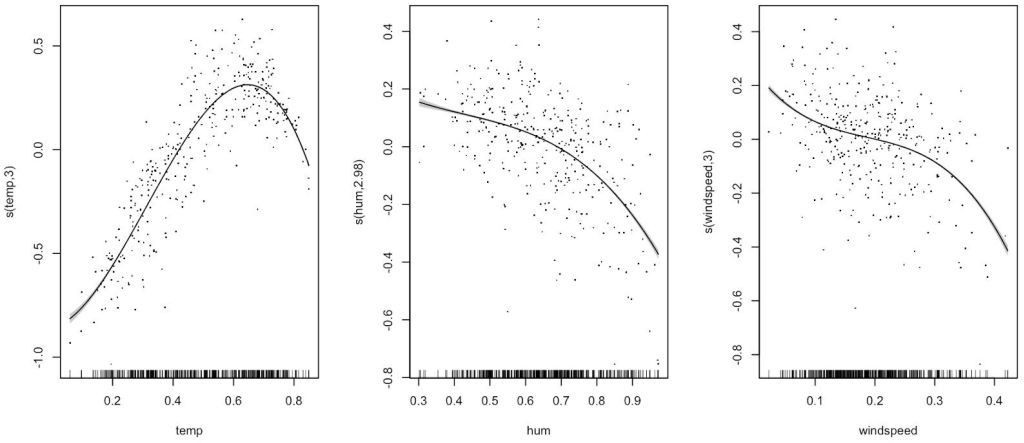

Figure 1.6: Improved Partial Residual Plots using Reduced Knots

As the objective is to select the most predictive regressors in an extended GAM manner, several non-nested models are compared based on AIC in table 1.2 below. First we modelled the GAM without excluding the outlying observations, which produced overly smooth functions for the continuous covariates.

Next, we removed the outliers and drastically reduced the number of knots for each of the smoothed terms (fit.gam2). Given the results of Figure 1.6, *hum* and *windspeed* were no longer fitted using a smooth relationship (fit.gam3). The remaining models (fit.gam4, fit.gam5 and fit.gam6) are obtained by consecutively dropping *workingday*, *windspeed* and *humidity.* Removing *windspeed* and *hum* noticeably increases the AIC, whereas dropping *workingday* seems to have a rather minimal impact on the model. As such, we choose fit.gam4 as our model of preference.

Model	AIC	Model Modification
fit.gam2	34476.05	Outliers removed & reduced knots
fit.gam3	35940.00	No smoothing for *hum, windspeed*
fit.gam4	35957.43	Drop *workingday*
fit.gam5	39972.70	Drop *windspeed*
fit.gam6	42239.04	Drop *hum*

Table 1.2: Model Selection based on AIC

As opposed to the linear Poisson model, the remaining covariates which are deemed most predictive in the GAM model are *season*, *weather situation, temperature (smooth), humidity*

and windspeed. Before interpreting the results, the adequateness of the additive model and the Poisson distribution will be examined in more detail.

4. ***Do the necessary checks to verify that the chosen model(s) fit the data well, e.g. check the link function, the scale of the covariates, etc. In other words, choose the most appropriate procedure. Illustrate your findings with appropriate graphics and illustrate how good prediction is.***

Link function

To assess the appropriateness of the canonical log link function for the Poisson model we compare it with the square root link. The log and square root link functions link the set of linear predictors to the log and the square root of the expected amount of rental bikes per day, respectively. This is shown graphically for the continuous covariate *temperature* in Figure 1.7.

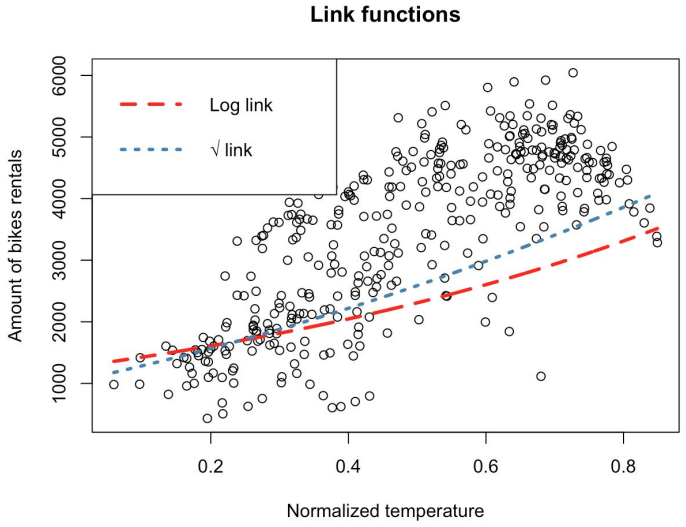

Figure 1.7: Log & Square Root Link Functions for `temperature`

The square root link function does seem to link the average values slightly better than the canonical link. However, it is unlikely that the model would improve as our main problem, overdispersion, has not yet been adequately accounted for.

Scale of Covariates

The normalized scales for the continuous covariates obscure an intuitive interpretation. Temperatures (°C) were all divided by the maximum value of 41, resulting in values between 0 and 1. The same procedure was used for *hum* and *windspeed,* with divisions by 100 and 67, respectively.

To allow for clearer interpretation, we back-transform these covariates by multiplying them by their respective maxima, after which we standardise them using a Z-score approach (subtracting the mean from each value and dividing by the standard deviation of the covariate). Since overdispersion is still unaddressed, the previously discussed models will not be re-fit with these Z-scores. However, these appropriately standardised measures will be used in subsequent models in which overdispersion will be addressed.

Residual Diagnostics

Figure 1.8 below shows unmistakable improvements of the additive smooth model (fit.gam4) over the linear Poisson model. This is unsurprising as a smooth relationship aims at reducing the MSE, thereby reducing the residuals considerably.

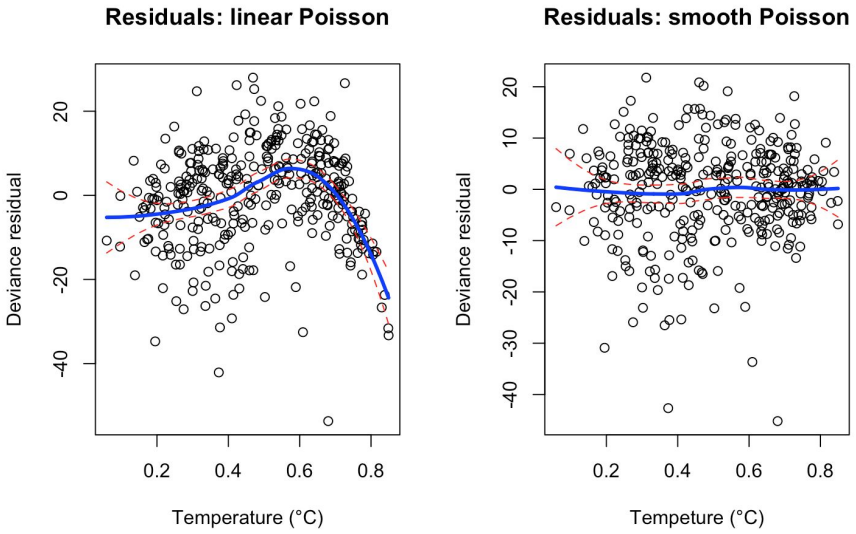

Figure 1.8: Deviance Residuals for the Linear and the Smooth Poisson Model

Model Adequacy

As outlined in the section on the additive models, our model of choice was fit.gam4. The corresponding model diagnostics are displayed in figure 1.9 below. The model seems to be relatively satisfying apart from the distribution of the deviance residuals.

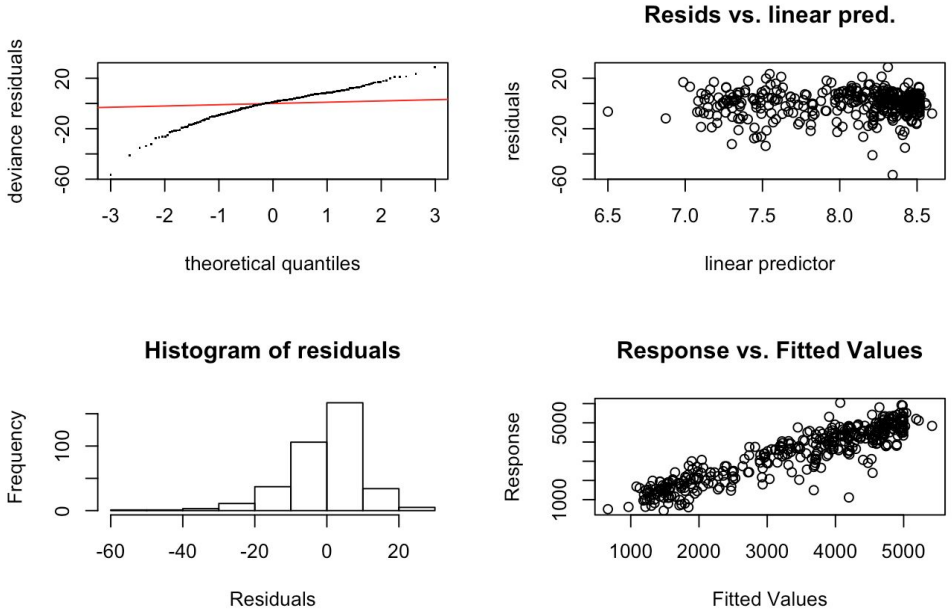

Figure 1.9: GAM Model Diagnostics

However the residuals are approximately centered around zero. Moreover, the recorded responses (i.e. the number of rental bikes per day) do tend to correspond quite well to the values fitted under the chosen GAM model (fit.gam4). Apart from some minor deviations among the residuals, there seems to be no problem with heteroscedasticity.

Prediction

Lastly, we need to examine the performance of our chosen model (fit.gam4) by comparing the observed counts for the bike rentals with the model-based predictions for those counts. Graphically, this correspondence is displayed in figure 1.10. Generally, the deviations from the observed counts seem to be rather contained, yet the model still heavily overestimates the extreme counts (i.e. low or high number of rental bikes)

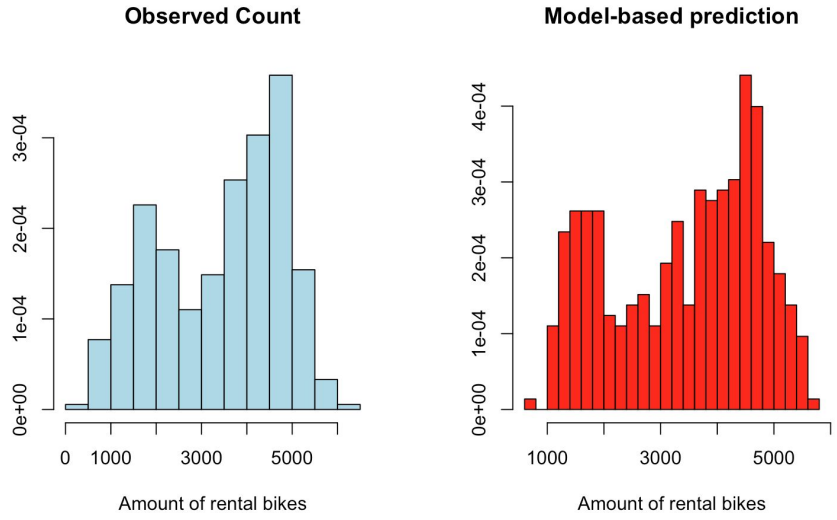

Figure 1.10: Observed and predicted counts

5. ***Fit a negative binomial model and a quasi-Poisson model in a frequentist manner if there is overdispersion.***

Overdispersion

Overdispersion of the Poisson distribution is an important, unaddressed problem with the previously evaluated models. This is immediately clear when we evaluate the distribution of the number of daily bike rentals (Figure 1.11). Although imperfect, the negative binomial distribution is far more appropriate than the one assumed under a Poisson model.

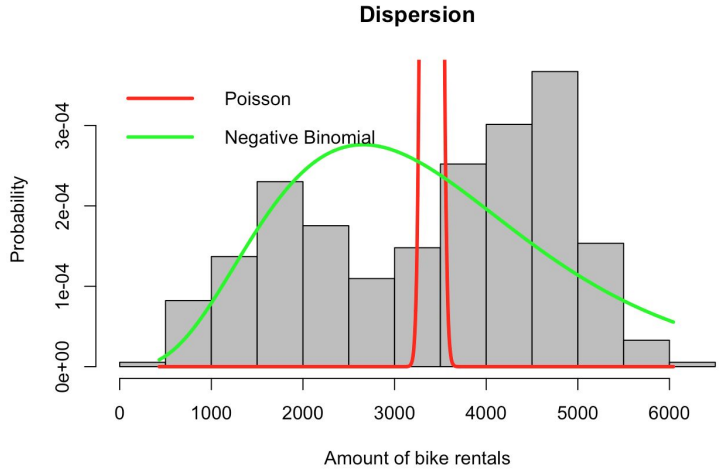

Figure 1.11: Dispersion of the Poisson and Negative Binomial Distribution

Negative Binomial Fit

The results corresponding to a negative binomial model with the canonical log link are given in figure 1.12 below. Fortunately the negative binomial provides considerable improvements in terms of model fit. The residual deviance (367.13) is indeed quite close to the degrees of freedom of the $\chi 2$-distribution (354). The AIC (5808) has also drastically decreased compared to any of the Poisson models. Under the negative binomial model, the most predictive regressors retained are *weather situation*, *temperature*, humidity, *windspeed*, and *season*. So in fact only workinday was considered insignificant according to this model.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.74847    0.03634 213.195  < 2e-16 ***
weathersit2 -0.07208    0.03258  -2.212 0.026943 *
weathersit3 -0.62265    0.07738  -8.046 8.54e-16 ***
temp         0.28883    0.02192  13.178  < 2e-16 ***
hum         -0.05767    0.01741  -3.312 0.000928 ***
windspeed   -0.05454    0.01306  -4.177 2.95e-05 ***
season2      0.43653    0.04519   9.660  < 2e-16 ***
season3      0.36057    0.05944   6.066 1.31e-09 ***
season4      0.60153    0.03982  15.105  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(19.5002) family taken to be 1)

    Null deviance: 1570.37  on 362  degrees of freedom
Residual deviance:  367.13  on 354  degrees of freedom
AIC: 5808
```

Figure 1.12: Linear Negative Binomial Model Output

The parameter estimates for these covariates are quite different from the Poisson models, although they generally point to the same conclusions for the categorical covariates (see Figure 1.1). The expected amount of rental bikes during winter time (*season 1*) and under favorable weather conditions (i.e. few or no clouds) is exp(7.74847) or about 2318 rental bikes, assuming *temperature* and *windspeed* are at their average (Z-score of 0).

Worsening weather conditions decrease this amount by factors of exp(-0.14) = 0.83 and exp(-0.76) = 0.47 for weather situation 2 and 3, respectively. Compared to winter, and all else equal, the expected amount of rental bikes increases by a factor of exp(0.43) = 1.54, exp(0.37) = 1.45 and exp(0.59) = 1.80 for the respective seasons Spring, Summer and Fall.

Unit increases in *temperature*, *wind speed and humidity* (i.e. increases by 1 standard deviation since expressed as Z-scores) correspond to multiplicative changes of exp(0.27) = 1.31, exp(-0.04) = 0.96 and exp(-0.05767) = respectively.

Quasi-Poisson

As an alternative to the negative binomial distribution, a quasi-Poisson model is considered to allow for more flexibility w.r.t. the mean-variance structure. As such, a working variance function is used rather than the one obtained from the corresponding exponential distribution. As the name implies, the estimation is done via a quasi-likelihood estimation.

For quasi-Poisson models, the simple overdispersion will have that the working variance is proportional to the model variance, with the proportionality factor being the dispersion parameter. With this option, the estimates are consistent with the MLE obtained from a regular Poisson, yet the standard errors have to be obtained by means of the robust covariance estimator. Results for a quasi-Poisson model with the simple overdispersion and the canonical log link are displayed in figure 1.13 below. Note that the standard errors of these estimates are based on the sandwich estimator.

```
z test of coefficients:

             Estimate Std. Error  z value  Pr(>|z|)
(Intercept)  7.706619   0.045562 169.1448 < 2.2e-16 ***
weathersit2 -0.089952   0.025726  -3.4966 0.0004712 ***
weathersit3 -0.660184   0.095811  -6.8905 5.561e-12 ***
workingday   0.016343   0.023970   0.6818 0.4953676
season2      0.497340   0.051118   9.7292 < 2.2e-16 ***
season3      0.442835   0.061767   7.1695 7.528e-13 ***
season4      0.625342   0.040188  15.5605 < 2.2e-16 ***
temp         0.232404   0.021816  10.6527 < 2.2e-16 ***
hum         -0.038086   0.015491  -2.4585 0.0139501 *
windspeed   -0.046409   0.012631  -3.6743 0.0002385 ***
```

Figure 1.13: Quasi Poisson model output (robust standard errors)

The dispersion parameter has been set to about 129, which essentially means that the working variance is assumed to be larger by that factor in order to properly correspond to the empirical variance. This would confirm our graphical illustration of 1.12, since it shows that the assumed variance under the regular Poisson model was considerably lower than the variance in the data.

In terms of significance of predictors, the quasi-Poisson model is in line with the indications of the negative binomial model. However, when looking at the residual deviance (48237 on 353 degrees of freedom), the problem seems to be only partially solved by the quasi approach. Sadly, we can't compare this model with the negative binomial model obtained above since the AIC is not present for quasi-likelihood models. However, since the negative binomial model was more adequate in terms of dealing with the problem of overdispersion, we keep our conclusions as they were inferred from the negative binomial setting.

IMDb

In this analysis, we examine how the profit of a movie (in dollars) is affected by the covariates *content rating* (categorical variable consisting of four levels indicating the suitable audience), *budget* (in dollars), and *director Facebook likes* (the number of likes of the director on his Facebook page). The following packages are used: `mgcv, lattice, ggplot2, grid, splines, fmsb`.

1. ***Make a descriptive analysis to look at the relationship between the covariates and the profit. Covariates used in this analysis are content_rating, budget, director_facebook_likes.***

Before moving on to the modelling of movie profit, we try to get some preliminary insights into the IMDb data by means of exploratory analysis. To that end, we examine the relationship between the movie profit and it's content rating, it's budget and also the amount of facebook likes the director of that particular movie has. As the content rating is a categorical covariate, it is natural to consider the pairwise relationships between profit and the numerical covariates in a grouped manner. As such, figure 2.1 gives the relationships between *director Facebook likes* and *profit,* and *budget* and *profit* respectively.

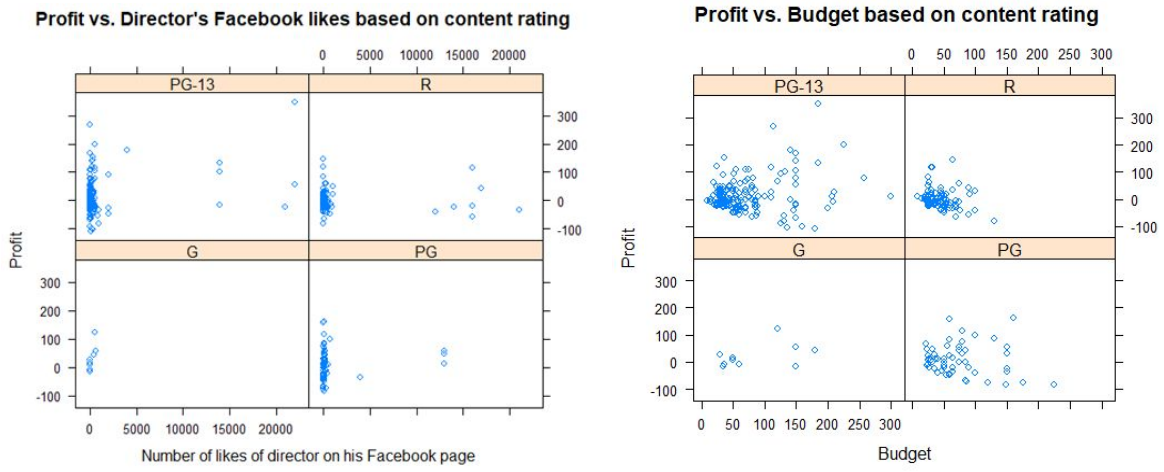

Figure 2.1: Grouped scatter plots

Notice that the "classes" are not balanced: in particular, there are far more observations in the PG-13 factor level as opposed to the movies that are G-rated for instance. Based on the scatter plot of profit versus the facebook likes, there does not seem to be an apparent relationship at first glance since profit seems to vary regardless of the amount of facebook likes. Moreover, the director's who have a substantial amount of facebook likes are the exception rather than the rule. It remains to be seen whether or not these observations ought to be classified as outliers.

When looking at the plot between the profit and the budget per group, the positive relationship seems to be most present in the PG-13 group. Moreover, the variability of profit

in all groups seems to increase as the budget increases. One could already expect that this will likely lead to heteroscedasticity issues when implementing a regular linear regression model. To further investigate the difference in variability across the different content rating groups, a grouped box-plot is examined in Figure 2.2.

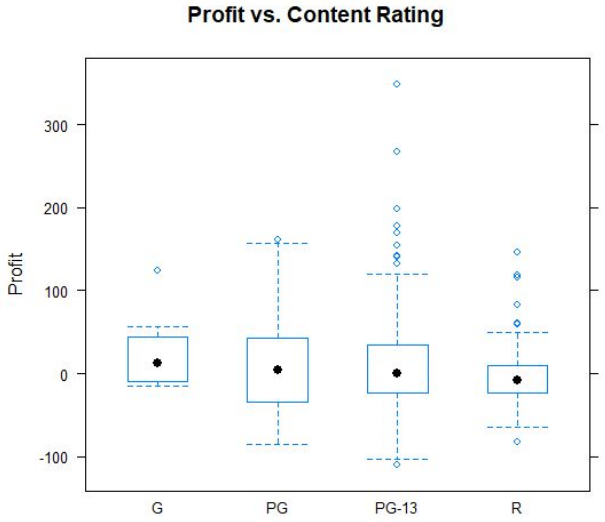

Figure 2.2: Grouped Box-plot

As previously mentioned, the variability is highest in the PG-13 level. It is clearly also the content rating that has the most outliers, followed by the R rating. Note that this outlier classification is simply based on the computation classically used in box-plots, namely observations outside of the interval [Q1 - 1.5*IQR, Q3 + 1.5*IQR]. In the end, it is difficult to tell whether or not these differences in variability and the occurrence of outliers are due to an imbalance in terms of observations across the factor levels.

Finally, Table 2.1 gives the correlations (p-values) between each covariate and the response based on the group, where the p-values based on the Pearson correlation test identify the significance of the relationship (these are linear correlations: note that no claim for or against linearity is being made here). Nonlinear modelling of these relationships follows in the subsequent sections.

Group (based on content rating)	Profit, Budget	Profit, Director's Facebook likes	Budget, Director's Facebook likes
G	0.448 (0.195)	0.786 (0.007)	0.751 (0.012)
PG	-0.108 (0.429)	0.121 (0.374)	-0.011 (0.936)
PG-13	0.214 (0.008)	0.295 (0.000)	0.326 (0.000)
R	-0.210 (0.044)	0.002 (0.982)	0.274 (0.008)

Table 2.1: Correlation matrix between numerical attributes (including the respective p-values in parentheses)

2. ***Model the profit as a function of budget only, use the following techniques: Polynomial regression model, Truncated polynomial splines of degree 2 (consider k=2, 3 and 5 knots), B-splines of degree 2 (consider m=3, 5 and 8 knots), Cubic P-splines (consider k=5, 8 and 20 knots)***

Polynomial Regression model

A first attempt at modelling the movie profit as a function of the movie budget consists of a linear polynomial regression model. Hence, the polynomial regression model can be defined as follows:

$$Profit_i = \beta_0 + \beta_1 budget_i + \beta_2 budget_i^2 + ... + \beta_l budget_i^l + \varepsilon_i$$

To find out what degree would be appropriate for relationship between movie profit and budget, we estimate several models with varying degrees (degrees 2, 3 and 7). The regression curves corresponding to these models are given in Figure 2.3, together with the 95% confidence band.

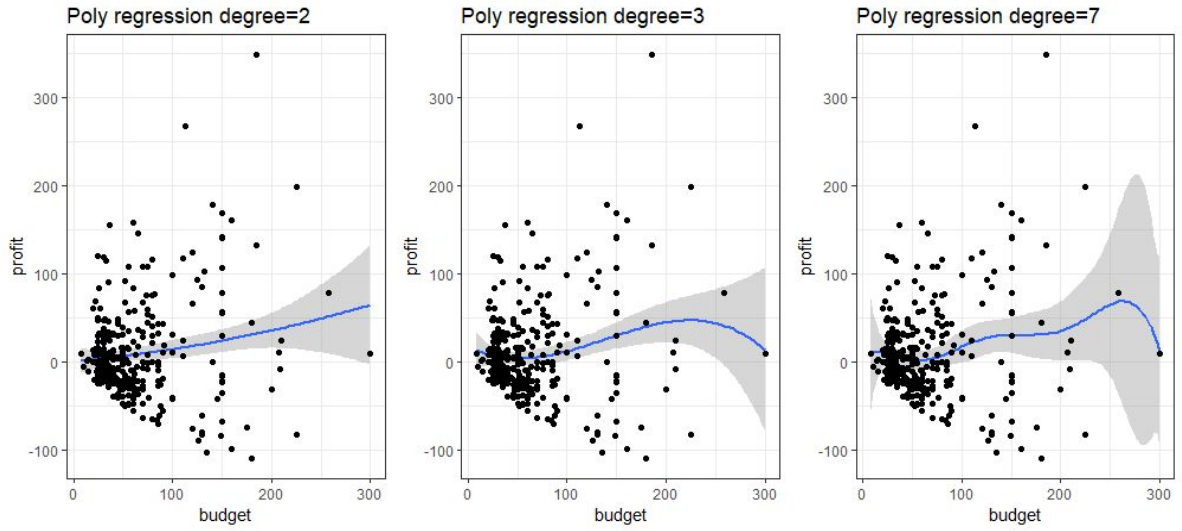

Figure 2.3 Polynomial regression with l = 2, 3, 7

Clearly the fitted curve for the model with degree 7 is heavily influenced by the local data points, especially across the higher budgets. This holds for the regression with degree 3 as well, although to a lesser extent. Regardless of the degree of the polynomial, the regression curve fails to properly fit all of the observations due to the very large variability of the response variable. Moreover, the variability of the movie profit also increases drastically for larger budgets, indicating a problem with heteroscedasticity.

Since polynomial regression models make use of global smoothing, this method will not be appropriate for data of this kind. With this in mind, local smoothing techniques could be more

appropriate. Local smoothing should allow to better approximate the nonconstant variability of the data by fitting local polynomials over different ranges of the movie budget.

Truncated polynomial splines

The first local smoothing technique that we consider are the truncated polynomial splines. The local polynomials in each interval depend on the amount of knots and the splines degree. As instructed, the degree of the polynomial is set to 2 and the amount of knots are set to 2, 3 and 5. The results corresponding to these respective knot amounts are is shown in figure 2.4 below.

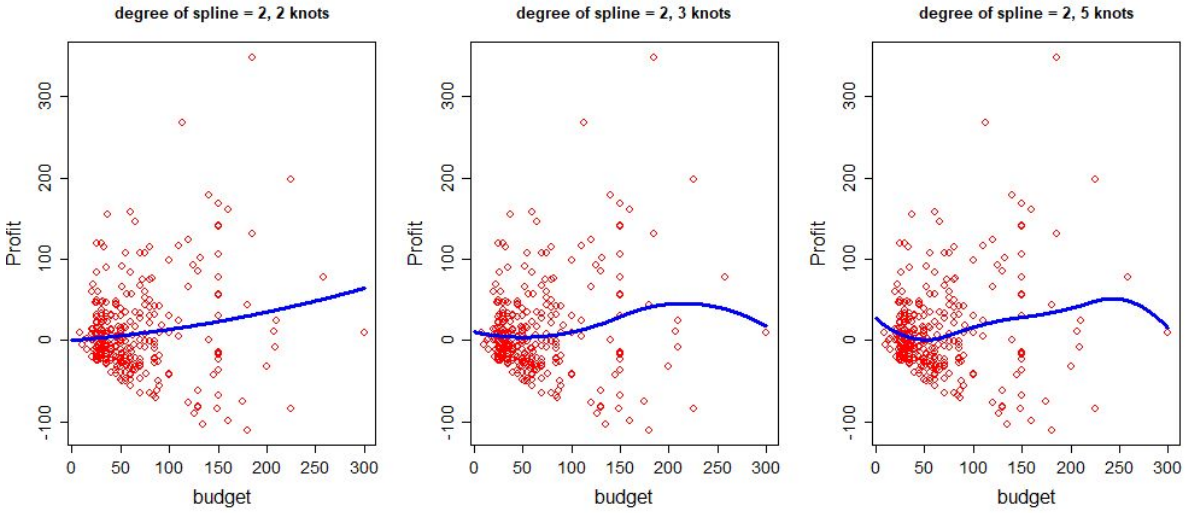

Figure 2.4 Truncated polynomial splines with k = 2, 3, 5

One can see that when budget is partitioned over 5 knots, the curve captures the data in a way similar to the higher-degree polynomial regressions of the previous section. It is quite obvious that 2 knots fail to fit the data well, whereas using 3 knots already shows a significant improvement. Increasing the knots to 5 seems to only yield marginal improvements.

B-splines

The second local smoothing method we use are the B-splines. With this method, the regression function *f(x)* is considered as a linear composition of $d = m + l - 1$ basis functions.

$$f(x) = \sum_{j=1}^{d} \gamma_j \, B_j(x)$$

where the b-spline basis contains (l+1) polynomial pieces of degree l, and m is the number of the inner knots, Each B-spline basis is continuous over the range of the knots, so that the combination of the B-basis as the regression function is also continuous. Note that the number of knots has a significant influence on the basis and, as we will show, on the resulting fit.

As instructed, we took into consideration k=3, 5, 8 with basis functions of degree 2. The fitted curves corresponding to these model specifications are given below in figure 2.5.

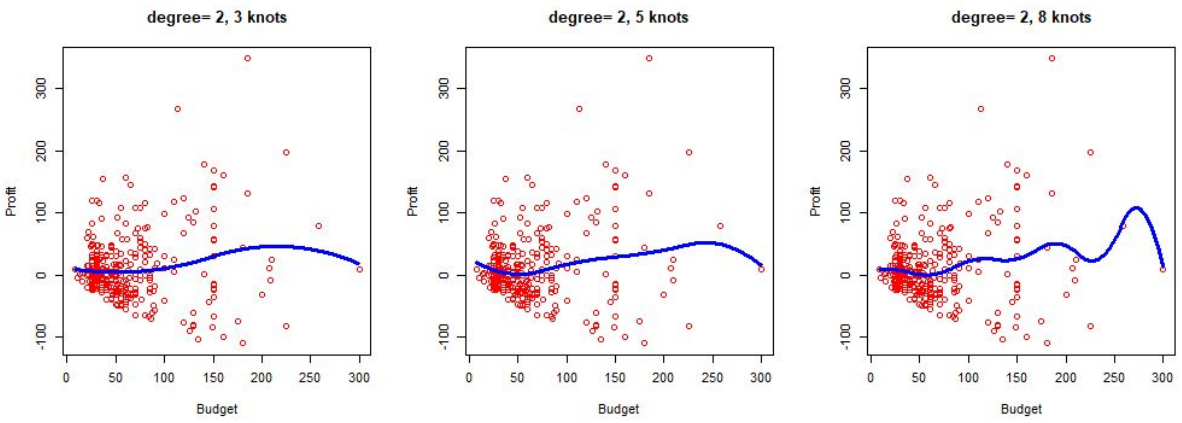

Figure 2.5: B-splines of degree 2 with k = 3, 5 and 8.

Clearly, the fitted curves are more influenced by the local data points as the knot amount increases. As such, 8 knots seems to be overfitting on the extreme observations in the X-space (i.e. the highest movie budgets). Whereas 3 knots doesn't answer to local data characteristics, a knot amount of 5 seems to be reasonable for a second degree B-spline regression.

Cubic P-splines

Lastly, a penalized local smoothing technique is examined, which introduces bias as seen in e.g. ridge regression. More specifically we use a penalized cubic polynomial (i.e. degree = 3), which is more commonly referred to as cubic P-splines. We consider three different knot amounts, namely 5, 8 and 20. The fitted curves corresponding to those models are summarized below in figure 2.6.

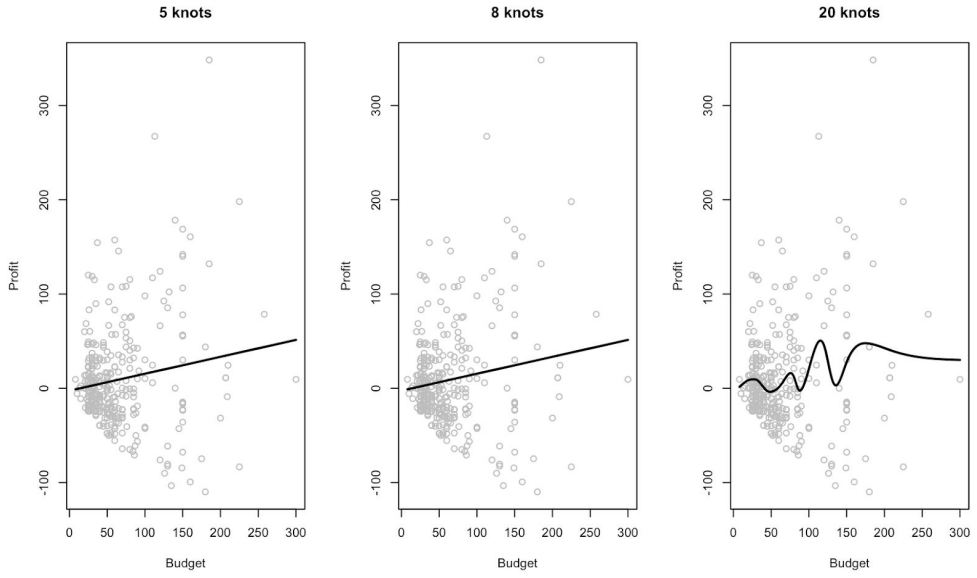

Figure 2.6: Cubic P-splines with k = 5, 8, 20

As it turns out, setting the knot amount to 20 seems to produce a good fit to the data, without overfitting on outliers. In contrast, choosing k=5,8 are completely reduced to a linear approximation. Further inspection of the smoothing parameter should provide more insight.

Since the P-splines introduce a penalty for the roughness of the function approximation, this parameter also has to be estimated. Fortunately, this means the estimated fit no longer depends on the amount of knots, which has been problematic for the other techniques. The optimal value for the smoothing parameter is obtained via the Generalized Cross-Validation criterion (GCV). For this application, these values are given below in table 2.2.

Knot amount	*Optimal Lambda*
5 knots	*7.38341e+11*
8 knots	*8.722682e+12*
20 knots	*114.9466*

Table 2.2: Optimal lambda for the respective models

As was expected based on the visualization in figure 2.6, the penalty for the knot amounts 5 and 8 are very large, to the extent that the fitted function is equivalent to the linear approximation of the function. Although still large, the penalty is about 115 when 20 knots are specified.

Upon closer examination of the significance of these smooth terms, some interesting results are obtained. The approximate F-value for the smoothing term of the cubic spline with 20 knots is 1.407 (p=0.147), with 11.57 effective and 13 reference degrees of freedom respectively. Although only an approximate significance indicator, it was also found that the smoothing terms for the models with 5 and 8 knots (i.e. the models that are essentially linear) were, in fact, considered significant. Consequently, we will compare the smooth cubic model using 20 knots with a linear model as shown in table 2.3.

Model	df	AIC
Linear model	3	3388.471
Cubic P-spline with 20 knots	13.565	3394.103

Table 2.3: Model comparison

Based on the AIC, there seems to be a slight preference for the linear model, even though the difference is quite marginal. Note that the AIC values for the cubic models with knots 5 and 8 indeed coincided with the AIC values of the linear model.

In conclusion, the linear model ultimately seems most appropriate since our smoothing

techniques either overfitted on local data or resulted in penalties that reduced the smoothing approximation to a linear fit.

3. ***Include the other covariates in the model and determine what variables show an impact on the profit.***

In this part of the analysis, the covariates *content rating, duration and director facebook likes* are taken into consideration as additional regressors to the previously obtained linear model. As it turns out, the only significant addition to the model are the facebook likes of the director. This was found out by initially adding in all the considered covariates, after which we sequentially removed the covariates *duration and content rating* as indicated by the F-test. Indeed, when comparing these nested models based on their F-values, only director_facebook_likes seems to have a significant impact on the movie profit next to the movie budget. This is shown in the analysis of variance table below (figure 2.7).

```
Analysis of Variance Table

Model 1: profit ~ budget + director_facebook_likes
Model 2: profit ~ budget + director_facebook_likes + content_rating
Model 3: profit ~ budget + director_facebook_likes + content_rating +
    duration
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    309 912730
2    306 899324  3     13406 1.5155 0.2105
3    305 899323  1         1 0.0003 0.9854
```

Figure 2.7: Analysis of Variance table

Consequently, we can now take a look at the obtained results for the model consisting of both the budget variable and the facebook likes of the directors. This corresponds to the model output shown below in figure 2.8. For interpretative purposes, the covariate budget has been standardised by means of a Z-score. After all, interpretation for a movie with zero budget is not quite appealing.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.9825116  3.2000622   2.182  0.02986 *
budget.Z                6.9167833  3.1472271   2.198  0.02871 *
director_facebook_likes 0.0022863  0.0008684   2.633  0.00889 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.35 on 309 degrees of freedom
Multiple R-squared:  0.04545,   Adjusted R-squared:  0.03927
F-statistic: 7.356 on 2 and 309 DF,  p-value: 0.0007566
```

Figure 2.8: Model output

Although it doesn’t explain a lot of the variability in the movie profit, the model is considered globally significant according to the F-test ($p<0.001$). A movie with an average budget ($Z = 0$) directed by someone with 0 likes on facebook, has an estimated profit of about 7 US dollars. For a budget increase of 1 standard deviation, the profit increases by approximately 6.9 dollars, given that the facebook likes are kept constant. Conversely, the likes affect the

movie profit by about 2.3 dollars for a thousand additional likes on the directors facebook page, keeping budget constant.

4. ***A movie is defined as successful when the profit is positive. Fit a model that relates the probability of success and the covariates considered above.***

Lastly, it is of interest to examine the probability for a movie to be profitable, as a function of the covariates selected in our previous analysis. To that end, the movie profit is binarized into two classes, namely a success in case of profit and a fail in case of recorded losses. With this labeling, we can now implement a logistic regression analysis, where we use a logit-link function and include *budget* and the *facebook likes of the director* as regressors. The resulting model output is displayed below in figure 2.9.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.760e-01  1.937e-01  -0.909    0.363
budget                   2.128e-03  2.437e-03   0.873    0.382
director_facebook_likes  1.003e-05  3.222e-05   0.311    0.755

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 432.47  on 311  degrees of freedom
Residual deviance: 431.45  on 309  degrees of freedom
AIC: 437.45
```

Figure 2.9: Logistic model output

Interpretation of the estimated model's coefficients depends on the choice of the link function, which is the *logistic* link function in this case. As such, interpretations of coefficients in this model are based on the fact that they represent expected changes in the log odds of the mean response (i.e. the probability to be profitable), for unit increases in the covariates. These are marginal effects however, since the other covariates in the model are held at some fixed values. For the *budget* covariate, we can interpret the marginal effect as follows; while the facebook likes are fixed, the change in the odds ratio for a dollar increase in the *budget* is by a factor of *exp*(0.0021) = 1.0021. Conversely, this is *exp*(0.00001) = 1.0001 under the same conditions. Clearly, the coefficients show no impact at all in a logistic setting, as indicated by the significance results. Also the Analysis of deviance table below (figure 2.10) shows these variables are not valuable additions to the model.

```
Model: binomial, link: logit

Response: indicator

Terms added sequentially (first to last)


                         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                       311     432.47
budget                    1  0.92063       310     431.55   0.3373
director_facebook_likes   1  0.09743       309     431.45   0.7549
```

Figure 2.10: Anova-table for deviance and degrees of freedom

Model fit

An idea about the goodness of fit for this model can be formed by looking at the Hosmer-Lemeshow statistic. Since only continuous covariates are involved, the appropriate convergence of the Pearson and Deviance statistic can not be guaranteed since those only apply when categorical variables are present. The Hosmer-Lemeshow statistic for the model depicted above has a Chi-squared value of 6.4842 (p = 0.5932). As such, the null hypothesis is not rejected and the model fit is deemed appropriate.

Note however that the model exhibits overdispersion to some extent since D >> n - p. Fitting a quasi-binomial logistic model did not yield any improvements; the dispersion parameter was taken to be 1.01, which is essentially the regular model.

There is also not really a problem with separation or multicollinearity of any kind, as can be derived from figure 2.11 below. The covariates behave quite independently, while there also isn't a clear hyperplane that separates the profit yielding movies from the loss incurring ones.

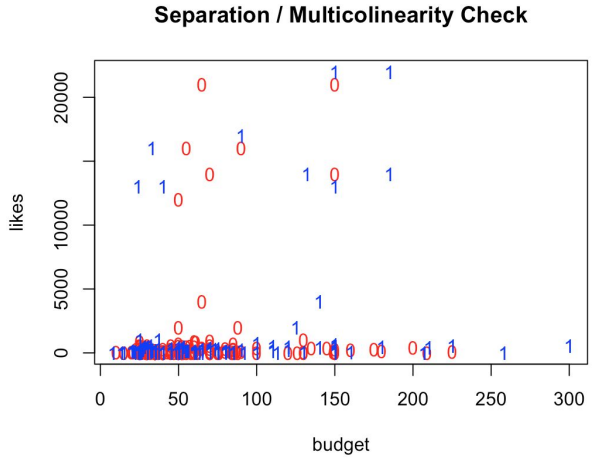

Figure 2.11: Separation / Multicollinearity check

Quality of Prediction

Next, we can also evaluate the quality of prediction by looking at an appropriately adapted coefficient of determination for logistic regression, such as the nagelkerke R^2. The model outlined above measures in at about 0.0043 whereas when we exclude the *likes of the director,* we obtain a value of 0.0039. The model with only the intercept has a value of 0 exactly. The inclusion of the covariates only very slightly improve predictions as is clear from these values.

An alternative is to look at the Concordance Measure, C, which essentially evaluates how much of the predictions are in the same direction as the observed outcomes. For this model C = 49.90%. This shows, once again, that the predictive performance of the model is not quite promising.

Residual Diagnostics

Perhaps the residuals can show if there are any outliers or bad fitting observations. Indeed the high variability in the movie budget seems to be problematic, as depicted in figure 2.11.

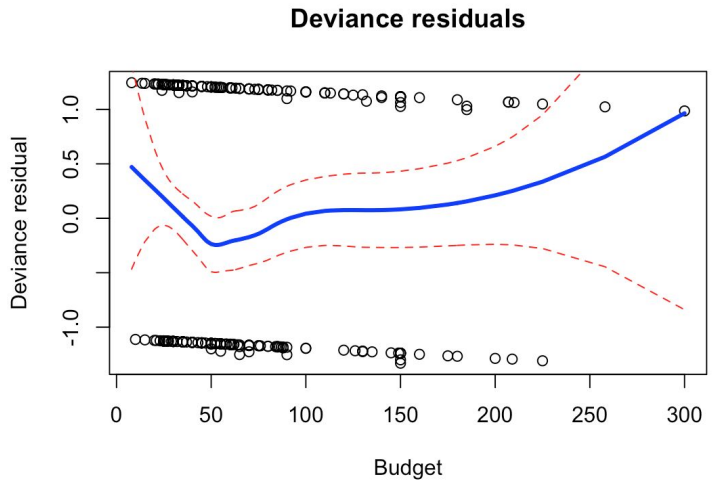

Figure 2.11: Deviance Residuals

Consequently, it might be worthwhile to identify these observations. To that end, the Cook's distance values are examined in Figure 2.12. One can now see which of the movies are considered as relative outliers in the data.

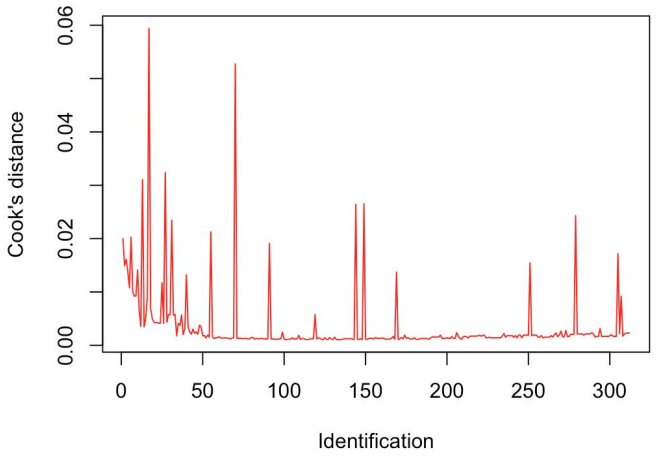

Figure 2.12: Influence plot based on Cook's distance

Link function

One more thing worth mentioning is that the model is robust to the choice of link function, as coefficient estimates are very similar when using the probit and the complementary log-log links. With the probit link, we have D = 431.35 on 309 df, while the complementary log-log gives 431.38 on 305 df.