Estimating Sales Prices

Philipp Burckhardt

August 12, 2014

1 Introduction

It is easy to see that the price a certain residential property is sold at is determined by different factors: The size of the building, the number of bedrooms, but also external factors like the quality of the neighborhood, whether it is located in a popular town or on the countryside. All these play a major role in determining the final house price. As the recent housing bubble in the U.S. has shown, cyclical behaviour is also an influencing factor. For an analysis of the macroeconomic forces driving the housing market, see Tsatsaronis and Zhu (2004). To this day, the exact interplay between all these factors remains obscure, though.

In the present analysis, we use information about the sales prices of residential property from a single town. The data set was obtained from a States Assessor's Office and comprises information about 444 individual residential properties sold in the time period from 1879 to 2010. Besides our variable of interest, the *price* of the house in U.S. dollars, the set of collected features contains: *garage*, the number of car spaces in the garage (0-4), the number of working *fireplaces* in the property, the *area* of the property (measured in square feet) and the *year* in which the building was completed. All variables are treated as numeric. Although the values of *garages* and *fireplaces* are discrete-valued and their influence on the response is not entirely linear, we don't treat them as categorical. Our reasons are two-fold: Firstly, we do not want an bloated set of predictors, which would be the consequence of the inclusion of dummy variables for factors. Secondly, if they were treated as categorical, there would be very few observations for some levels of the factors, giving unreliable results of effect sizes.

Using standard linear modelling techniques, our goal is to devise a model which enables prediction of sales prices given the set of explanatory variables. Predicting housing prices is a standard example in the machine learning and statistical literature, and many different methods have been used in this context (see Hotel and June (2004) for a study using neural networks).

Our reasoning for using linear modelling is that it may allow the researcher to gain a deeper understanding of the relationships between the variables, which is in stark contrast to many black-box methods like neural networks (Hastie et al., 2009).

The rest of the paper is organized as follows: In Section 2, we conduct an exploratory analysis and perform outlier detection. The used methodology for model fitting and model selection is presented in Section 3. Section 4 discloses the obtained results. The adequacy of the final model is investigated in Section 5 by residual diagnostics and other model checking practices. An alternative modelling approach using resistant regression is presented in Section 6. Section 7 concludes.

2 Data Exploration and Cleaning

Common sense tells us that the area size of a property will be positively correlated with the realized price of sale. But is this connection valid for all observations, or are there some properties which are e.g. so large that they must be even sold at a lower price in order to balance the enormous upkeep costs?

In the U.S., housing prices have risen roughly 2% per year over the last decades (Glaeser et al., 2005). Having this in mind, one would expect that the year a house had been built was positively correlated with its price. Is this really the case, or are there so many enthusiasts for old buildings as to inflate prices?

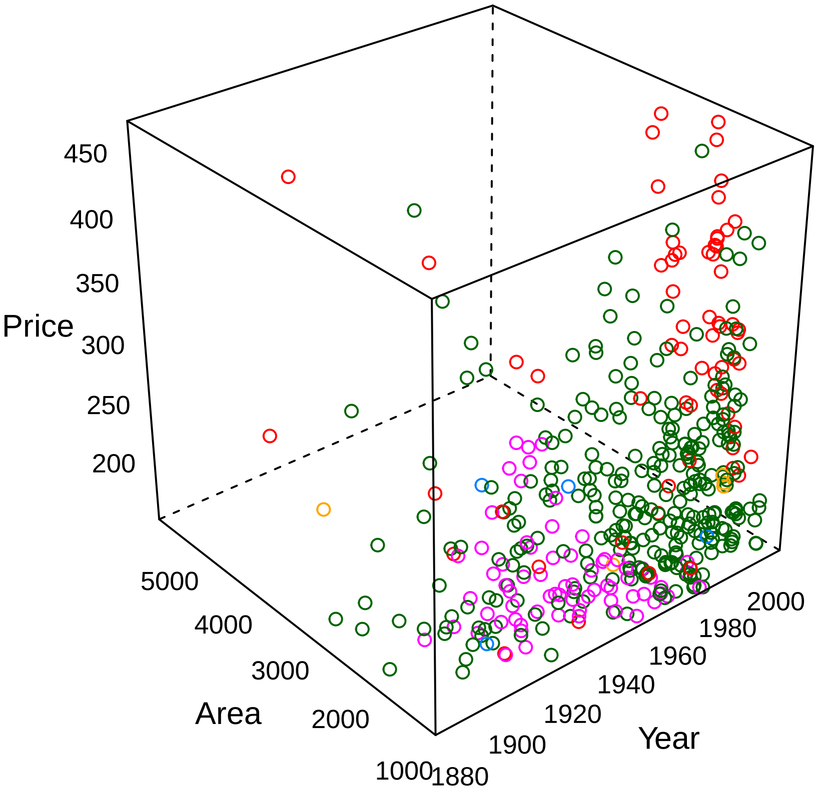

Figure 1: 3d scatterplot with dependent variable *price* on the z-axis and numeric predictors *year* and *area* on the x and y-axis. The count of *garages* is mapped to the color of the scatter points, with red standing for three garages, green for two and magenta indicating one garage. The rare cases of no garage or four garages are colored in blue and orange, respectively.

To shed some light on these questions, Figure 1 presents a 3d scatterplot of the three variables. Furthermore, *garage* was mapped to the color attribute of the points. First, it is noticeable that a later year is indeed associated with a higher sales price. If one looks at the marginal distribution of *price* against *year*, this pattern will become even more obvious. In linear modelling, the question of model choice is of utmost importance. One decision which needs to be made is whether it is necessary to include higher-order terms of the predictors. A simple linear regression of the sales prices on a cubic polynomial of *year* shows that the terms for all three degrees are clearly significant (p-values < 0.001). By including all of them in the set of potential covariates, we achieve two goals: We do not prohibit variables which have a significant influence to be part of the final regression equation and we also permit *year* to have not only a monotonic increasing, but also a potentially decreasing influence. This might become a desirable feature as more data comes in and the model is updated, since the astronomic prices which had been realized for new property during the housing bubble have likely fallen.

From the colored cluster, we can further deduce that in recent years there had been a trend towards larger garages: While building built in the 1920s to 1950s generally possess garages with only one parking spot, contemporary buildings are equipped with garages in which two or even three cars fit in.

Although not clearly noticeable from the 3d scatterplot, *area* is indeed positively affecting the sales price. This can be best seen by looking at the marginal distribution which is depicted in Figure 2. The point cloud clearly exhibits an upward trend, which is reflected by the blue regression line which features a positive slope.

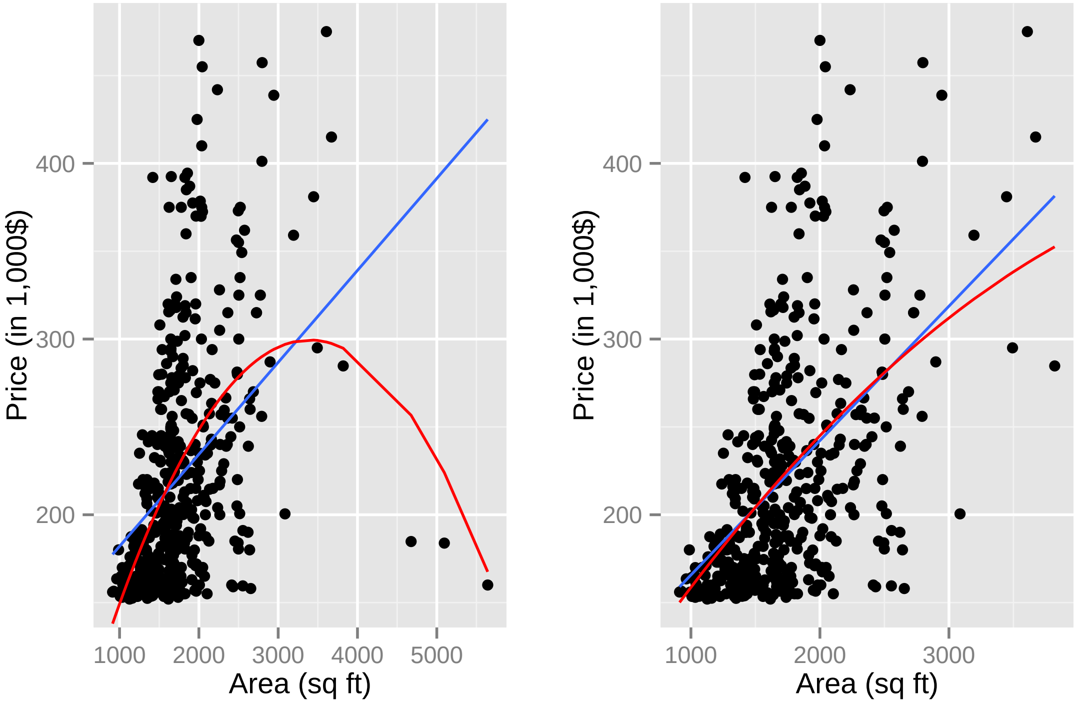

Figure 2: On the left-hand side, a scatterplot of *area* against *price* is displayed. Superimposed are regression lines from a simple regression of *price* on *area* and a second-order polynomial regression. The right-hand side shows the same plot after the three outliers in the lower-right corner were removed.

The red regression curve on the plot of the left-hand side in Figure 2 could make us believe that it is necessary to include a higher-order term for *area*: The parabola is nowhere near the ordinary regression line, indicating that the second-order term exerts a major influence. And while this is indeed true, this behaviour is almost exclusively determined by the three extreme points in the lower-right corner (they have indices 343, 438 and 443). If they were removed, the regression lines look all of a sudden very similar. Naturally, the question arises what to do with the outliers. If we leave them untouched, the results of standard linear modelling using OLS will be severely influenced by them.

To decide how to proceed, we must check whether no measurement error is present and assess how likely it is that they stem from the same population as the other observations. Concerning the first question, we have no reason to believe that measurement error is present: Sales prices around $200,000$$ are not unreasonable, and the existence of properties with an area of more than $5,000$ square feet raises no eyebrows either.

However, they do possess unusually large values of predictor *area* when compared to the other properties. Combined with the fact that their sales prices are so low, this does seem strange. One should raise the question: why were they sold so cheaply? We don't know, but most likely they have either very high upkeep costs or differ from the other locations in some other respect. In any case, it is likely that they are structurally different from the rest of the observations and therefore should not play a major influence. Therefore, we omit them from the following analysis, restricting our focus of investigation to residential properties whose area is between $1000 - 4000$ square feet. A modelling strategy which does not involve deleting the outliers is pursued in Section 6.

3 Methodology

3.1 Multiple Linear Regression

In the linear modelling context, the simplest model is undoubtedly linear regression. After having dealt with the outliers, there is no other problem which would necessitate the use of a more complicated model: The response variable is numerical and there is no multi-level structure present in the data set. Written in matrix notation, the linear regression model is given by

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \quad \text{with} \quad \varepsilon \sim N(0, \sigma^2\mathbf{I}), \tag{1}$$

where $\mathbf{Y}$ is a $n \times 1$ vector of response variables, $\mathbf{X}$ is the $n \times p$ design matrix of explanatory variables and ε is an error vector coming from a multivariate normal distribution where the error terms are supposed to be uncorrelated with each other and to have constant variance.

Recall that linear regression can only be carried out in the absence of perfect multicollinearity. Perfect multicollinearity occurs if any of the predictor variables can be represented as a linear combination of the other ones. If this is the case, the matrix $\mathbf{X}$ is singular and cannot be inverted. This type of multicollinearity is rarely encountered in real life. However, strongly correlated predictors also pose a problem insofar as they render the estimates of the regression coefficients β unstable. Luckily, the variables in $\mathbf{X}$ are not multicollinear, with all variables having correlations of less than 0.5 among each other.

3.2 Model Selection

When fitting a linear model, one first needs to decide upon the set of regressors which should belong to the model. In Section 2, we already noted that there are interaction effects present, with the amount of garages varying over time. Thus, we allow all second-order interaction terms to be present in the model, but prohibit higher-order interactions in order to to prevent overfitting and to ensure that the final model can be interpreted. We also exclude interactions with the higher-order terms of *year*.

Having decided upon the full model, we install an automated backward selection procedure to assess which terms are not necessary parts of the model and might be dropped. The used procedure is based on functionality by Ripley (1996), who in their R package *MASS* provide an automated function for stepwise-selection based on AIC as a model comparison criterion.

Since their advent, methods for automatic variable selection have come under scrutiny. Critics argue that while they can be beneficial, they are often mis-used to the point that the drawn statistical conclusions become invalid, for example by not accounting for the multiple comparisons carried out in stepwise procedures.

To prevent such problems from arising, different proposals have been put forward. A thorough discussion of the associated issues is given by Chateld (1995), who sees potential in resampling techniques as means to deal with aforementioned problems. Harrell (2001) argues strongly in favour of bootstrapping methods, emphasizing their advantages compared to standard cross-validation procedures. A recent technique which combines an automated stepwise-selection procedure with bootstrap resampling has been proposed by Austin and Tu (2004). They argue that one is likely to arrive at a final model which will include spurious variables if one used automatic selection procedures in isolation. This in turn would result in inferior prediction performance when the model was evaluated on independent training data. Using bootstrap resampling then may serve as a means to separate variables that are independent predictors of the outcome from noise variables since the former will be selected in a a majority of the bootstrap samples, "whereas noise variables would be identified as predictors in only a minority of samples" (Austin and Tu, 2004).

Drawing $B = 100$ times with replacement from our original data set $\mathcal{D}$, 100 bootstrap replicates $\mathcal{D}^*_1, \ldots, \mathcal{D}^*_{100}$ are created. Backwards selection using AIC as a model comparison criterion is then carried out for each sample $\mathcal{D}^*_k$ and the resulting models are recorded. To carry out the calculations, we utilize an implementation in statistical programming environment R written by Rizopoulos (2009).

The returned models then serve as our candidate set from which we need to pick our final model. Training each candidate model on $B = 100$ bootstrapped samples $\mathcal{D}^*_1, \ldots, \mathcal{D}^*_{100}$ of original data set $\mathcal{D}$, we calculate the RMSE of prediction when the fitted model is evaluated on the original data set, i.e. we calculate

$$\text{RMSE}^{(b)} = \sqrt{\tfrac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i^{(b)}\right)^2}$$

where $\hat{Y}_i^{(b)}$ denotes the i-th fitted value from the model trained on the bth bootstrap sample and $b = 1 \ldots B$. The final RMSE estimate for a model is then given by

$$\text{RMSE} = \frac{1}{B}\sum_{b=1}^{B}\text{RMSE}^{(b)}.$$

From the list of candidate models we then choose the one associated with the lowest RMSE as our final model. Although this estimate of RMSE will likely be over-optimistic since we have based our decision-rule on it and optimized it, this does not invalidate the model selection process as all candidate models will likely benefit similarly from this bias. To get an accurate estimate of RMSE, it would nonetheless be necessary to evaluate the final model on a "true" validation set.

4 Results

Implementing the procedure described in the previous section, we have obtained $B = 100$ candidate models which deserve further attention. Tables 1 and 2 disclose information about the selected models. Table 1 shows for each predictor how often it ended up in the final model chosen by the stepwise-selection procedure, whereas Table 2 displays the proportion of times the regression coefficients had a negative or positive regression coefficient.

Although these results are encouraging insofar as many predictors have been chosen in all bootstrap samples as parts of the chosen model, one might wonder why some of them do not have consistent regression coefficients. This fact can be explained by multicollinearity among the predictor variables. For example, from the main effects *area* has the least consistent pattern concerning the sign of its regression coefficient. Looking at the left table, the source of this behaviour becomes apparent: The interaction terms of *area* with the other predictors are not very often the same in the selected models, with some ending up in less than half of the final candidate set. Since these interactions are highly correlated with *area*, the estimated coefficients will depend on whether they are included in the model or not.

	(%)
Area	100.00
Fireplaces	100.00
Garage	100.00
I(Year^3)	100.00
Year	100.00
Bath	98.00
I(Year^2)	98.00
Garage:Area	100.00
Garage:Year	96.00
Fireplaces:Year	73.00
Bath:Year	71.00
Garage:Bath	70.00
Fireplaces:Area	63.00
Area:Year	44.00
Bath:Area	43.00
Fireplaces:Bath	19.00
Garage:Fireplaces	16.00

Table 1: For every predictor, this table lists the percentage of all $B = 100$ bootstrap iterations in which the respective predictor was chosen by the backward selection procedure.

	+ (%)	- (%)
I(Year^3)	100.00	0.00
Year	98.00	2.00
Bath	96.94	3.06
Area	25.00	75.00
Fireplaces	20.00	80.00
Garage	0.00	100.00
I(Year^2)	0.00	100.00
Fireplaces:Year	100.00	0.00
Garage:Area	100.00	0.00
Garage:Year	100.00	0.00
Fireplaces:Area	98.41	1.59
Area:Year	79.55	20.45
Fireplaces:Bath	63.16	36.84
Garage:Fireplaces	31.25	68.75
Bath:Year	2.82	97.18
Bath:Area	2.33	97.67
Garage:Bath	1.43	98.57

Table 2: This Table displays the percentage of times each predictor had a positive and negative regression coefficient. Coefficients flip sign because of multicollinearity among predictors.

Running our bootstrap resampling scheme to obtain an estimate of RMSE for each model in the candidate set, we end up with the model depicted in Table 3 as our final model. Notice that this model is mostly comprised of variables deemed important in a large number of the bootstrap replicates (all predictors appear in at least 60% of the candidate models), signaling that we did not pick spurious variables to be part of our model. Most regression coefficients are significant at the 5% level, and an $R^2 = 0.662$ shows that the model suffices to explain a large amount of the original variance of the response variable. The enormous value of the intercept might look troublesome at first glance, however this can be easily reconciled by observing that this negative value is balanced by predictor *year*, which has at least a value of 1879, the earliest date in the data set. In follow-up studies, we would advise to drop the *year* variable and instead include the age of the building as a predictor. This conveys the same information, but leads to more naturally interpretable coefficients.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1633240.424	425342.632	-3.840	0.000
Garage	-961.256	270.944	-3.548	0.000
Fireplaces	-498.754	235.332	-2.119	0.035
Bath	399.682	233.269	1.713	0.087
Area	-0.029	0.017	-1.669	0.096
Year	2532.510	653.901	3.873	0.000
I(Year^2)	-1.308	0.335	-3.905	0.000
I(Year^3)	0.000	0.000	3.935	0.000
Garage:Bath	-17.882	5.536	-3.230	0.001
Garage:Area	0.048	0.007	6.549	0.000
Garage:Year	0.471	0.138	3.417	0.001
Fireplaces:Area	0.010	0.007	1.495	0.136
Fireplaces:Year	0.258	0.117	2.197	0.029
Bath:Year	-0.194	0.119	-1.626	0.105

Table 3: Regression output of the model chosen by the selection procedure outlined in Section 3.

A look at Figure 3 reveals that our final model did not stand out too much from the competition, though. All models chosen by the stepwise selection using AIC provide a reasonably good RMSE, with their confidence intervals overlapping.

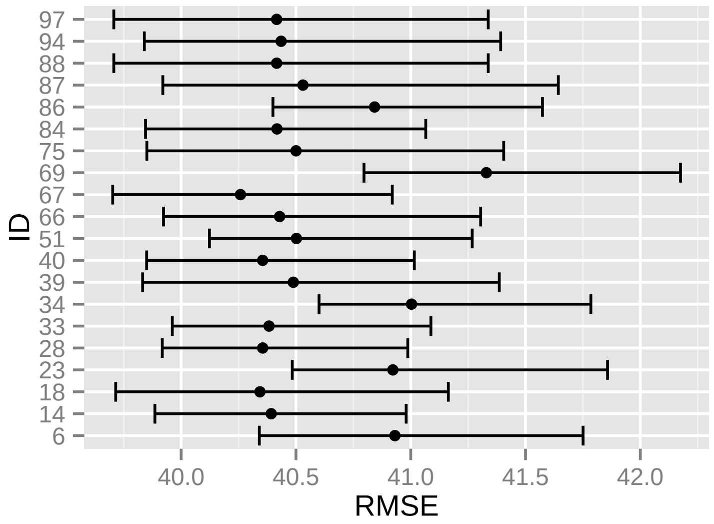

Figure 3: For a randomly chosen subset of size 20 of the models selected by backward elimination using AIC, this plot depicts RMSE along with its 95% confidence interval. For each model, the point estimates for RMSE were estimated by averaging across the 100 bootstrap replications. The endpoints of the associated confidence intervals were chosen as the 2.5- and 97.5-percentile of the respective empirical distribution of RMSE.

5 Residual Diagnostics

The diagnostic plots displayed in Figure 4 can be used to consider whether the fit of the selected model is satisfactory or if there were blatant violations of the model assumptions.

The scatterplot of the residuals does not indicate any form of heteroskedasticity, a sign in favor of the current model specification. The fact that the residuals look constrained to the left-side can be best explained by the fact that the domain of house prices is restricted by natural circumstances (e.g. resource prices). Likewise, the plot in the bottom left does not point out any observations with an unusually large value of Cook's Distance, suggesting that we have appropriately dealt with the outliers and do not have major problems with highly influential observations anymore.

Yet, a look at the residual histogram and the associated quantile-quantile plot reveals that the residuals do not follow a normal distribution. Instead, they have a longer-tailed distribution with larger mass in the right tail than expected. This brings about the question whether this non-normality poses a threat for drawing valid inference. Ramsey and Schafer (2002) note that non-normality is usually a minor concern. They point out that the only situation in which it causes problems is a long-tailed error distribution combined with a small sample size. Also, prediction intervals based on the normal approximation are no longer viable. The reason that regression coefficients can still be assumed to follow a normal distribution is given by the central limit theorem (CLT), which states that the regression coefficients will tend to a normal distribution as $n \to \infty$ regardless of the underlying distribution of the error terms.

One can verify this by using the bootstrap. We apply the residual bootstrap method to our fitted model. Here, the fitted values from the original regression are retained, and new bootstrap samples are created by drawing with replacement from the original residuals and adding the bootstrapped

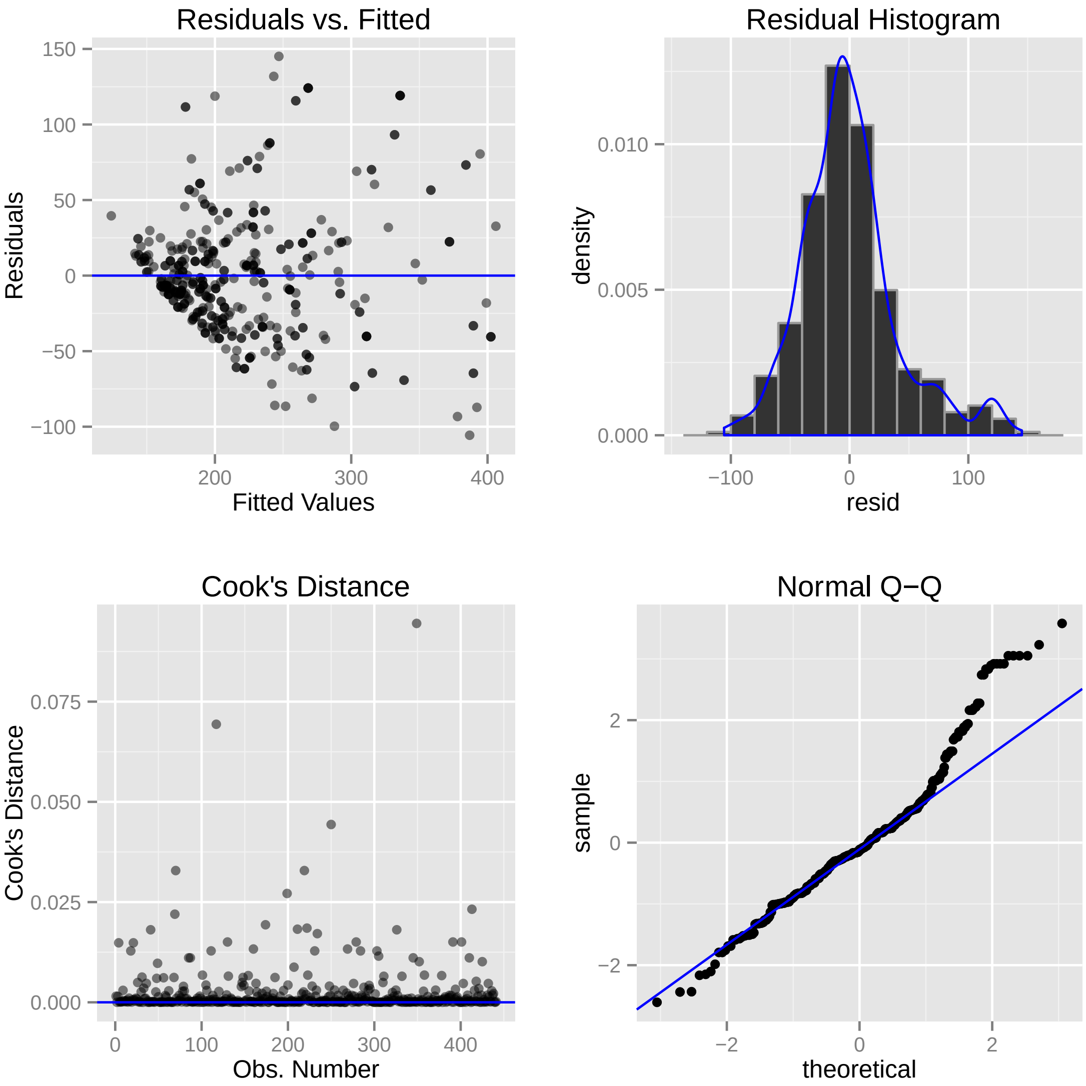

Figure 4: Diagnostic plots for the model in Table 3. In the plot on the top-left, the fitted values $\hat{Y}$ are plotted against the residuals $e = \hat{Y} - Y$. On the top-right, a histogram of the residuals along with a kernel density line is provided, whereas the plot on the bottom left shows Cook's Distance, which is a measure of the importance of each data point in determining the fitted regression line. Finally, the plot on the bottom right displays a quantile-quantile plot of the studentized residuals.

residuals to the fitted values in order to form a bootstrap response. Performing $B = 2000$ iterations, we then get an empirical distribution for all regression coefficients. We then apply the Shapiro-Wilk test for normality to assess whether they come from a normal distribution. The results indicate that for all regression coefficients besides *garage*, normality cannot be rejected at the 5% significance level.

6 Alternative: Resistant Regression

Instead of removing the outliers and fitting a normal linear regression model, one could also use a resistant regression method since these are not affected by outliers. In contrast to OLS, the least absolute deviations (LAD) model does not attempt to minimize the sum of squared residuals, but instead the sum of absolute residuals. So one attempts to solve

$$\min_{\beta} \sum_{i=1}^{n} |y_i - x_i^{\mathsf{T}} \beta|,$$

which can be formulated as a linear program and be solved with the simplex algorithm. LAD regression is a special case of a larger set of models which are known as quantile regression models. LAD constitutes the special case when the median is estimated. An introduction to the subject is given by Koenker and Hallock (2001).

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-1575786.298	274658.135	-5.737	0.000
Garage	-668.802	151.455	-4.416	0.000
Fireplaces	-312.782	120.598	-2.594	0.010
Bath	438.809	134.789	3.256	0.001
Area	-0.011	0.010	-1.165	0.245
Year	2440.986	421.885	5.786	0.000
I(Year^2)	-1.260	0.216	-5.833	0.000
I(Year^3)	0.000	0.000	5.881	0.000
Garage:Bath	-10.261	3.548	-2.892	0.004
Garage:Area	0.031	0.004	8.027	0.000
Garage:Year	0.329	0.077	4.252	0.000
Fireplaces:Area	0.014	0.003	5.538	0.000
Fireplaces:Year	0.157	0.061	2.571	0.010
Bath:Year	-0.217	0.069	-3.145	0.002

Table 4: This Table displays the regression output for the model equation from Table 3 when fitted by the least absolute deviations (LAD) regression model. In this model, the error terms are not bound to follow a normal distribution with mean zero. It is instead assumed that they have a median of zero. The regression coefficients are found not by minimizing the residual sum of squares, but the sum of absolute residual deviations.

The results of fitting this model are displayed in Table 4. Compared to the earlier obtained results in Table 3, there are some noticeable differences. Which model fares better? Using the same resampling technique as described in Section 3, we get 41.325 as an estimate for RMSE. This is worse than most models found by the stepwise procedure, as a look at Figure 3 confirms. However, it must be noted that this comparison is necessarily unfair: Instead of re-running the model selection procedure, we used the same set of covariates as before in fitting the LAD model. This puts the model necessarily at a disadvantage since its final set of regressors was not properly chosen.

7 Conclusion

In this paper, we used linear modelling to come up with a predictive model for the sales price of residential property. In performing exploratory data analysis, several outliers were detected. They were removed from the subsequent calculations. Building upon earlier work by Ripley (1996) and Austin and Tu (2004), we devised a two-step procedure for model selection. In a first step, a set of candidate models was assembled by iteratively applying a stepwise selection procedure on a resampled version of the data set. These candidate models were then ranked according to their performance according to RMSE when fitted on 100 bootstrap samples of the original data set and evaluated on the original training set. The model with the best average performance was chosen as the final model.

As an alternative way to deal with the outliers, a resistant regression model was fitted and the results of the two models were compared. The resistant regression model performed worse, but this is owing to the circumstance that we did not properly optimized it using our procedure but instead fitted the same model formula as before, the reason being that the R functions we relied on are not able to deal with LAD objects.

This poses an interesting field for new research projects, as automated selection procedures become increasingly important in the wake of "big data". With the ever increasing size of data sets, data exploration and model assessment become more and more inapplicable. Using resistant and robust regression techniques whose results do not depend on single observations might help in these scenarios.

References

AUSTIN, P. AND J. TU (2004): "Bootstrap methods for developing predictive models," *Am. Stat.*, 58, 131–137.

CHATELD, C. (1995): "Model uncertainty, data mining and statistical inference," *Soc. A*, 158, 419–466.

GLAESER, E., J. GYOURKO, AND R. SAKS (2005): "Why have housing prices gone up?" .

HARRELL, F. E. (2001): *Regression Modeling Strategies*, Springer.

HASTIE, T., R. TIBSHIRANI, AND J. FRIEDMAN (2009): "The Elements of Statistical Learning," *Elements*, 1, 337–387.

HOTEL, B. C. AND N. Z. JUNE (2004): "House Price Prediction : Hedonic Price Model vs. Artificial Neural Network," .

KOENKER, R. AND K. F. HALLOCK (2001): "Quantile Regression," *J. Econ. Perspect.*, 15, 143–156.

RAMSEY, F. L. AND D. W. SCHAFER (2002): "The Statistical Sleuth," .

RIPLEY, B. D. (1996): *Pattern Recognition and Neural Networks*, vol. 40, Cambridge University Press.

RIZOPOULOS, D. (2009): *bootStepAIC: Bootstrap stepAIC.*

TSATSARONIS, K. AND H. ZHU (2004): "What drives housing price dynamics: cross-country evidence," *Bis Q. Rev.*, 65–78.

Appendix

R Code

```r
set.seed(1215)  # for reproducibility

# use of knitr (Sweave successor) to create this report
opts_chunk$set(fig.align = "center", cache = TRUE, message = FALSE,
    echo = FALSE, eval = TRUE)
options(replace.assign = TRUE, width = 85, digits = 3)

Housing <- read.csv(file = "HousingData.csv")
Housing$SalePrice <- Housing$SalePrice/1000  # recode such that it is measured in 1,000 U.S. Dollars

# remove outliers
Housing.clean <- Housing[-c(343, 438, 443), ]

# no signs of multicollinearity: variables in X are not
# highly correlated (cor<0.5 for all variables) cor(Housing)
# create 3d scatterplot (Figure 1)
require(lattice)
par.set <- list(axis.line = list(col = "transparent"), clip = list(panel = "off"))

cloud(SalePrice ~ Year + Area, data = Housing, groups = Garage,
    scales = list(arrows = FALSE), par.settings = par.set, zlab = "Price")
lm.fit <- lm(formula = SalePrice ~ Garage + Fireplaces + Bath +
    Area + Year, data = Housing.clean)
lm.fit.full <- lm(formula = SalePrice ~ (Garage + Fireplaces +
    Bath + Area + Year)^2 + I(Year^2) + I(Year^3), data = Housing.clean)
fit.quad <- fitted(lm(SalePrice ~ Area + I(Area^2), data = Housing))
fit.quad.clean <- fitted(lm(SalePrice ~ Area + I(Area^2), data = Housing.clean))

p <- ggplot(aes(x = Area, y = SalePrice), data = Housing) + geom_point() +
    scale_color_brewer(palette = "Set1") + stat_smooth(method = "lm",
    se = FALSE) + geom_line(y = fit.quad, col = "red") + labs(x = "Area (sq ft)",
    y = "Price (in 1,000$)")

q <- ggplot(aes(x = Area, y = SalePrice), data = Housing.clean) +
    geom_point() + scale_color_brewer(palette = "Set1") + stat_smooth(method = "lm",
    se = FALSE) + geom_line(y = fit.quad.clean, col = "red") +
    labs(x = "Area (sq ft)", y = "Price (in 1,000$)")

grid.arrange(p, q, ncol = 2)
set.seed(666)
library(bootStepAIC)
stepAIC.res <- boot.stepAIC(lm.fit.full, data = Housing.clean,
    direction = "backward")
CandidateModels <- stepAIC.res$BootStepAIC
save(list = c("CandidateModels", "stepAIC.res"), file = "bootResults.RData")
load(file = "bootResults.RData")
print(xtable(stepAIC.res$Covariates), floating = FALSE, hline.after = NULL,
    add.to.row = list(pos = list(-1, 0, nrow(stepAIC.res$Covariates)),
        command = c("\\toprule\n", "\\midrule\n", "\\bottomrule\n")))
print(xtable(stepAIC.res$Sign), floating = FALSE, hline.after = NULL,
    add.to.row = list(pos = list(-1, 0, nrow(stepAIC.res$Sign)),
        command = c("\\toprule\n", "\\midrule\n", "\\bottomrule\n")))
# select best via cross-validation
set.seed(13)
my.resamples <- createResample(y = Housing.clean$SalePrice, times = 100)

evaluate.model <- function(fitObj) {
    RMSE.list <- lapply(X = my.resamples, FUN = function(x) {
        Resampled.Data <- Housing.clean[x, ]
        temp.fit <- update(fitObj, data = Resampled.Data)
        preds <- predict(object = temp.fit, newdata = Housing.clean)
        RMSE <- sqrt(mean((preds - Housing.clean$SalePrice)^2))
        RMSE
    })
    return(RMSE.list)
}

models.RMSE.vec <- lapply(X = CandidateModels, evaluate.model)
models.RMSE.avg <- unlist(lapply(models.RMSE.vec, function(x) mean(unlist(x))))

final.fit <- CandidateModels[[which.min(models.RMSE.avg)]]
RMSE.sd <- sd(unlist(models.RMSE.vec[[which.min(models.RMSE.avg)]]))
xtable(summary(final.fit), digits = 3, caption = paste0("Regression output of the model chosen by ",
    "the selection procedure outlined in Section \\ref{sec:methods}."),
    label = "final")
perf <- lapply(models.RMSE.vec, function(x) unlist(x))
perf.df <- data.frame(matrix(unlist(perf), nrow = 100, byrow = F))
perf.df.mean <- colMeans(perf.df)
perf.df.upper <- apply(perf.df, 2, quantile, 0.975)
perf.df.lower <- apply(perf.df, 2, quantile, 0.025)

df3 <- data.frame(mean = perf.df.mean, upper = perf.df.upper,
    lower = perf.df.lower, id = 1:100)
# plot a sample of 20 models
df3 <- df3[sample(100, size = 20), ]
df3$id <- as.factor(df3$id)
ggplot(df3, aes(x = id, y = mean)) + geom_errorbar(aes(ymin = lower,
    ymax = upper)) + geom_point(aes(x = id, y = mean)) + geom_smooth(aes(x = id,
    y = mean)) + labs(x = "ID", y = "RMSE") + coord_flip()
ResidPlot <- function(lm.obj) {
    p <- qplot(x = fitted(lm.obj), y = resid(lm.obj), geom = "blank") +
        geom_point(alpha = 0.5) + geom_hline(yintercept = 0,
        col = "blue") + labs(x = "Fitted Values", y = "Residuals",
        title = "Residuals vs. Fitted")

    q <- gg_QQplot(lm.obj)

    df1 <- data.frame(resid = resid(lm.obj))
    r <- ggplot(aes(x = resid, y = ..density..), data = df1) +
        geom_histogram(binwidth = 20, col = "grey60") + geom_density(col = "blue") +
        labs(title = "Residual Histogram")

    cooks <- fortify(lm.obj)$.cooksd
    s <- qplot(x = seq_along(fitted(lm.obj)), y = cooks, geom = "blank") +
        geom_point(alpha = 0.5) + geom_hline(yintercept = 0,
        col = "blue") + labs(x = "Obs. Number", y = "Cook's Distance",
        title = "Cook's Distance")

    grid.arrange(p, r, s, q)
}
ResidPlot(final.fit)
# bootstrap coefficient distributions
lm.boot <- Boot(final.fit, method = "residual", f = function(obj) coef(obj),
    R = 2000)
# normality assumption cannot be rejected for bootstrapped
# coefficients at alpha=0.05 except garage
apply(lm.boot$t, 2, shapiro.test)
# fit the resistant LAD model
final.formula <- formula(final.fit)
lad.fit <- rq(formula = final.formula, tau = 0.5, data = Housing)

# compute RMSE estimate
mean(unlist(evaluate.model(lad.fit)))
xtable(summary(lad.fit, se = "iid")$coef, digits = 3, caption = paste0("This Table displays the regression
    " when fitted by the least absolute deviations (LAD) regression model. ",
    "In this model, the error terms are not bound to follow a normal distribution with mean zero. ",
    "It is instead assumed that they have a median of zero. ",
    "The regression coefficients are found not by minimizing the residual sum of squares, ",
    "but the sum of absolute residual deviations."), label = "final_lad")
```