End of Chapter - Data Exercises - 200692

Dec 24, 2020

Chapter 7.3

Analyze the hotel price-distance to the center pattern for another city. Select a city from the large hotels-Europe data set. Keep hotels only, those with 3 to 4 stars, for a November 2017 weekday. Examine the distribution of the distance variable and drop observations if you think you should. First estimate a bin scatter with four bins and then a lowess regression. Visualize the results and summarize the most important findings from these non-parametric regressions. Then move on to estimate a simple linear regression and interpret its coefficients. Compare your results to what we found for Vienna.

I choose *Milan* as the target city to study. First, let's do some housekeeping work to find the data required(city of Milan, hotels with 3 to 4 stars, a November 2017 weekday). The input is a csv file of hotelbookingdata, and the output is hotel_Milan.csv.

Examine the Distribution of the Distance

As a result, we get a sample of 403 observations, where we filter hotels as required and excluded hotels with prices exceeding 1000 Euros. From Table 1 and Figure 1 we notice that:

1. All hotels are within the range of [0, 30] miles, with most of the hotels within 5 miles to the city center.
2. The mean value of hotel distance is far larger than median value, so its distribution plot is skewed right, with a long right tail.

There is a gap between hotels within 20 miles and farther than 20 miles to the city center, and apparently those fall outside the range of (0,20) are outside Milan. So, I decided to drop these observations.

To learn about the average price of hotels with differing distances to the city center, we create bin scatters(Figure 2) by splitting the data into four bins([0,5],[5,10],[10,15] and [15,20]), with equal bandwidth of 5 miles. Also, we create a lowess non-parametric model(Figure 3) to approximate the regression.

Table 1: Summary statistics of Hotel Distance

Obs	n	Mean	Median	Min	Max	Std.	Skew
distance	403	6.07	2.7	0.2	30	6.92	1.63

Bin Scatter and Lowess Regression

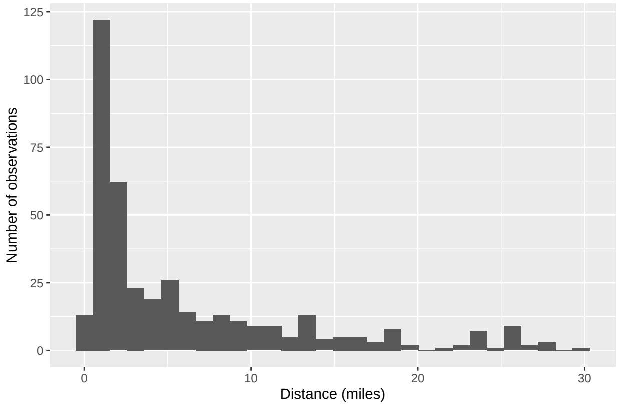

Figure 1. Histogram of Distance

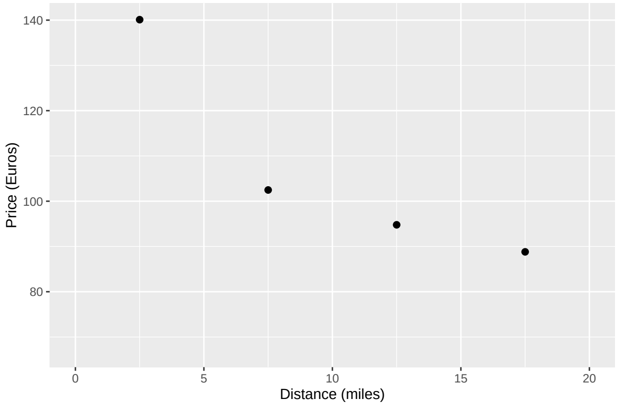

Figure 2. Hotel price and distance to the city center: bin scatters

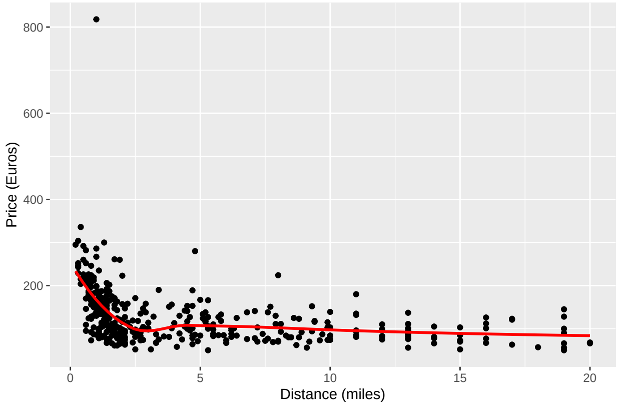

Figure 3. Hotel price and distance to the city center: lowess regression and scatterplot

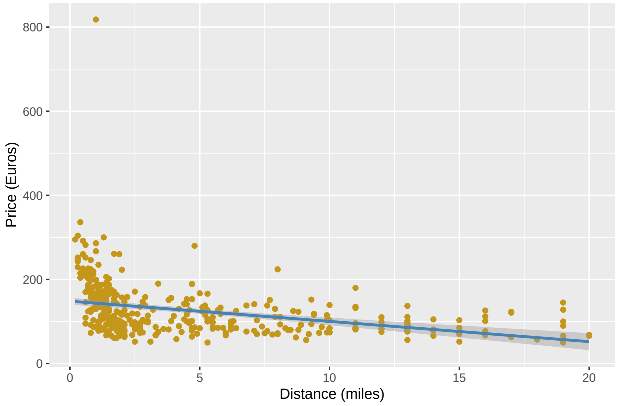

Figure 4. Hotel price to the city center: linear regression and scatterplot

From Figure 2 and Figure 3, we notice that:

- With four distance categories, we can see that the relationship appears to be monotonic but nonlinear: the difference in average y between the adjacent bins are not always the same. There is a larger negative difference between the [0,5] miles and [5,10] miles bins than between adjacent bins at higher distances.
- The smooth non-parametric regression is steeper at small distances and flatter at longer distances, but in general the line keeps a negative slope, except that in the subset of bin (2.5, 5) there is a local positive trend.
- Both two figures suggest a negative pattern of association between hotel price and distance to the center, hotels further away from the city center are, on average, less expensive.

We've known that hotels further away from the city center are less expensive on average, but we wonder how much less expensive they can be. This is a quantitative question which requires linear regression.

Simple Linear Regression

Figure 4 shows the regression line together with the scatter plot, and Table 2 produces an intercept of 148.3 and a slope of -4.8:

- The intercept is 148.3, suggesting that the average price of hotels right in the city center is 148.3 euros. But we are not sure if there is such a hotel right in the city center.
- The slope of the linear regression is -4.8. Hotels that are 1 mile further away from the city center are, on average, 4.8 euros cheaper in our data.

Table 2: Modelling Hotel Price and Distance to City center

	Estimate	Std. Error	t value	Pr(>|t|)	CI Lower	CI Upper	DF
(Intercept)	148.2847	5.0882	29.1428	0	138.2797	158.2897	375
distance	-4.8091	0.5414	-8.8820	0	-5.8737	-3.7444	375

Table 3: Summary Statistics of Lexus Car Age and Price

Variable	n	Min	1st IQR	Median	3rd IQR	Max	Mean	Std.	Skew
Age	1106	0	3.00	4.0	6	21	5.17	3.40	1.68
Price	1106	3588	21498.25	28959.5	34998	88989	28765.25	11369.55	0.75

Comparison with Hotels in Vienna

Compared with our findings with hotels in Vienna, we conclude that:

- In both cities, hotels further away from the city center are, on average, less expensive.
- However, with 1 mile further away from the city center, hotels in Milan are much cheaper than in Vienna, with difference of 9.2 euros per mile, on average.
- Beyond 5 miles, hotel price in Vienna have a rather positive pattern(i.e., with 1 mile further away, the hotel becomes more expensive, on average); but the negative pattern remains the same for hotels in Milan.

Chapter 7.4

Collect data on used cars of a specific brand and type, and analyze price and age of the car. First estimate a bin scatter with two bins, then one with four bins, and then estimate a lowess regression. Visualize the results and summarize the most important findings from these non-parametric regressions. Then move on to estimate a simple linear regression and interpret its coefficients. Finally, use the results from the simple linear regression to list candidate cars for the best deal using the residuals of the linear regression.

I am interested in prevalent used **Lexus** cars in the market, so, with the help of Chrome extention Selector-Gadget, I scrape all basic information of Lexus cars on autotrader, for required explanatory and regression analysis.

Bin Scatter and Lowess Regression

As a result, we get a sample of more than 1,000 observations, with car price and age specifics in table 3.

Also, we create binned scatters(with 2 and 4 bins, separately) and lowess regression(Figure 5, 6 and 7) to capture the pattern of association between Lexus used car age and prices:

Table 4: Modelling Second-handed Lexus Price and Age

	Estimate	Std. Error	t value	Pr(>|t|)	CI Lower	CI Upper	DF
(Intercept)	41550.29	489.14	84.95	0	40590.55	42510.03	1104
Age	-2474.24	74.79	-33.08	0	-2620.98	-2327.49	1104

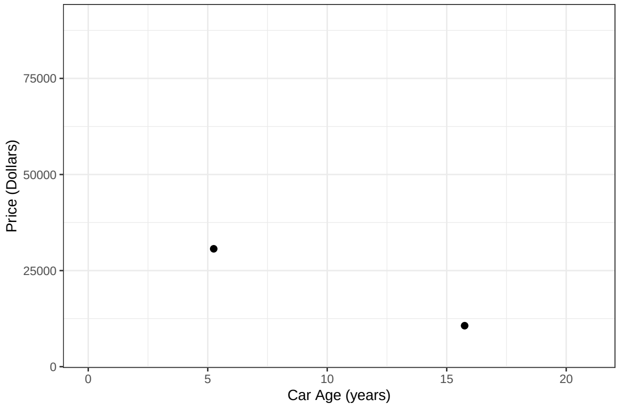

Figure 5. Used Lexus Car Price to Car Age: 2−bin scatters

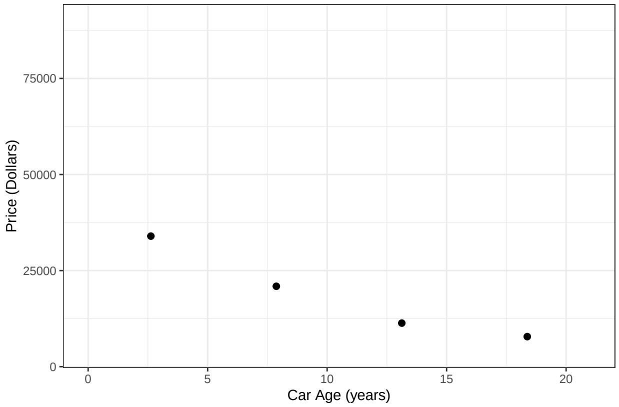

Figure 6. Used Lexus Car Price to Car Age: 4−bin scatters

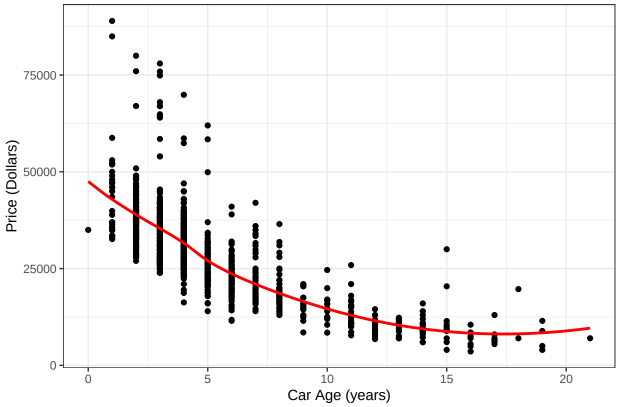

Figure 7. Used Lexus Car Price to Car Age: lowess regression and scatterplot

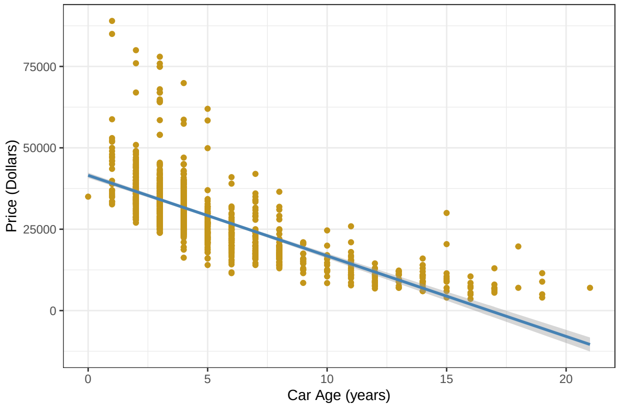

Figure 8. Used Lexus Car Price to Car Age: linear regression and scatterplot

- Both bin scatters(Figure 5, and Figure 6) suggest a negative pattern of association between used Lexus car price and age.
- With 4 bins, we can see that the relationship between car age and price is monotonic but nonlinear. And there is a larger difference between the [0,5] years and [5,10] years bins than between adjacent bins at older ages.
- Figure 7 shows the lowess non-parametric regression, together with the scatter plot.The smooth line appears to be deeper at earlier ages and flatter at older ages.
- In general, we uncovered across these regressions a negative slope in general, which is, used Lexus cars with older age are, on average, cheaper.

Linear Regression

Still, we wonder, how much cheaper an used Lexus car can be, if it's one year older? Thus, we move on to simple linear regression to answer this quantitative question. Table 4 and Figure 8 demonstrate the result of linear regression, with an intercept of 42,306.77 and a slope of -2,605.67:

- The intercept is 42,306.77, suggesting that the average price of brand new Lexus cars is 42,306.77 dollars, on average.

Table 5: Most Under-priced Lexus Second-hand Cars, Top 5

Name	Price	y_pred2	res2	Year	Age
Used 2015 Lexus CT 200h	11749	26704.88	-14955.88	2015	6
Used 2017 Lexus CT 200h	16250	31653.35	-15403.35	2017	4
Used 2015 Lexus CT 200h	11500	26704.88	-15204.88	2015	6
Used 2016 Lexus CT 200h	15990	29179.11	-13189.11	2016	5
Used 2016 Lexus CT 200h	13995	29179.11	-15184.11	2016	5

- The slope of the linear regression is -2,605.67. Lexus cars that are 1 year older are, on average, 2605.67 dollars cheaper in our data.

Residual Analysis: Find the Best Deal

We've had a basic grasp of used Lexus cars, but how should we find a good deal among them? The regression line in Figure 8 shows the predicted values, while the actual value is on each scatter. The difference between them, the residual, will help us find an answer to the question. Candidates for a good deal are cars that re under-priced relative to their ages; in other words, those with most negative residuals, as shown in table 5.

Chapter 8.5

Download data on used cars of a specific brand and type, and analyze price vs age of cars in order to find a good deal. Estimate a regression that fits the pattern of association well enough. Use the results from the simple linear regression to list candidate cars for the best deal using the residuals of the linear regression. (During your analysis you may want to consider taking logs, inspecting and dealing with influential observations, experimenting with functional forms, and so on.

So, we will continue with used cars cases. Except for simple linear regression in chapter 7.4, I will also transform X or Y by taking logs.

Transformation of Variables

Taking natural logs of variables can better approximate some non-linear patterns to linear regressions, especially when the data is right-skewed. I list four models for further review:

- level-level: Price = alpha + beta * Age
- log-level: ln_Price = alpha + beta * Age
- level-log: Price = alpha + beta * ln_Age
- log-log: ln_Price = alpha + beta * ln_Age

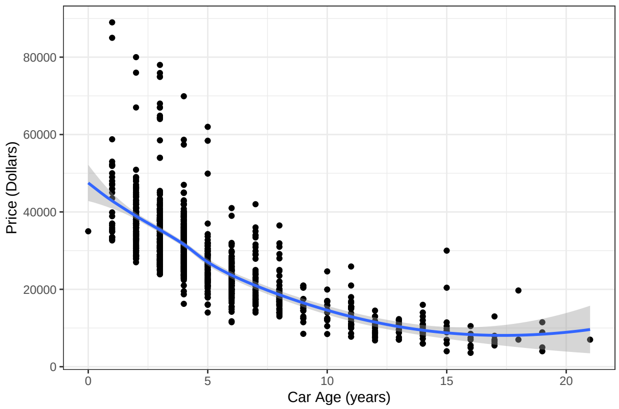

Figure 9. Regression of Used Lexus Car Price to Car Age: level Price, level Age

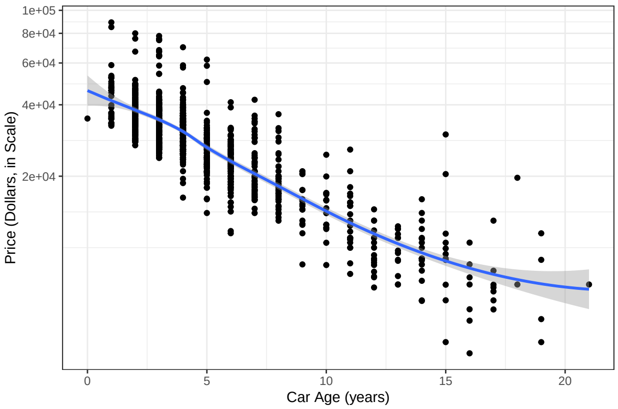

Figure 10. Regression of Used Lexus Car Price to Car Age: log Price, level Age

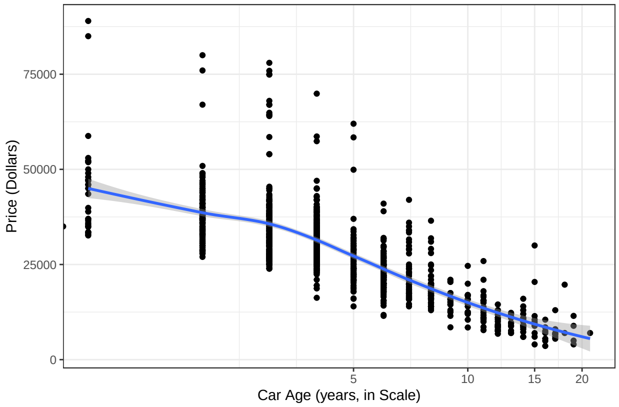

Figure 11. Regression of Used Lexus Car Price to Car Age: level Price, log Age

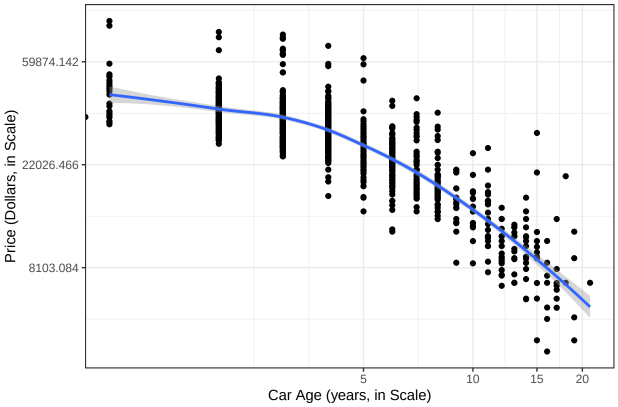

Figure 12. Regression of Used Lexus Car Price to Car Age: log Price, log Age

Based on comparison, we select **Log-Level Model**

- Substantive Reasoning: Price fluctuation measured in percentage changes is easier to interpret with age changes in absolute terms. Also, choosing relative terms means being free from arbitrary units of measurement.
- Statistical Reasoning: the graph gives better approximation: the scatter plot suggests a good linear pattern.

Presentation of Model Choice

We chose various regression models to capture the non-linearity, including linear regression, quadratic regression, and piecewise linear regression. To prepare for aforementioned regression, we take logs of Price and square terms of Age; also we remove negative log values to maintain the validity of regression.

Considering summary statistics and regression figure, we choose Log-Level Model as the final one. As it is obvious that all three figures have similar association pattern, but the Log-Level Model is much better to interpret than other two.

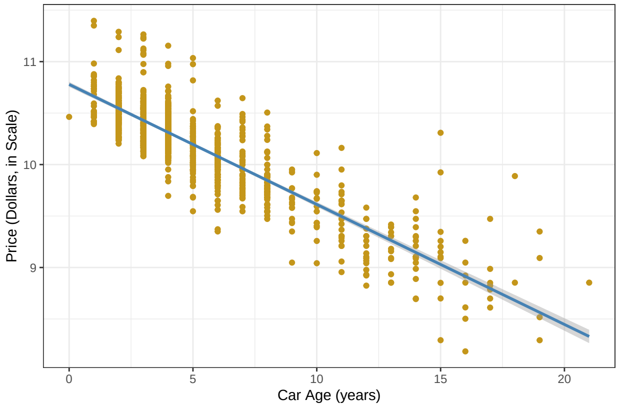

Figure 13. Regression of Used Lexus Car Price to Car Age: log Price, level Age

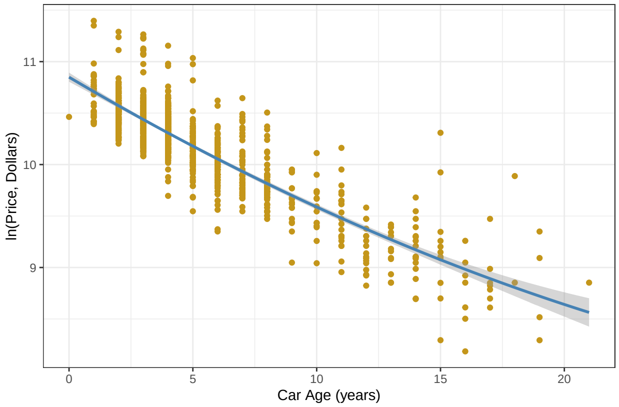

Figure 14. Used Lexus Car Price and Car Age: Quadratic function

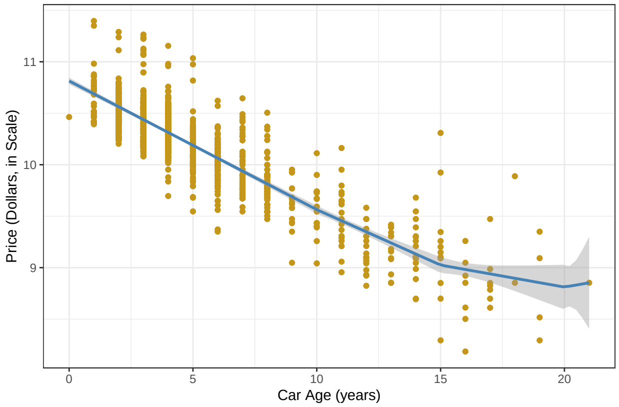

Figure 15. Used Lexus Car Price and Car Age: Piecewise linear spline

Interpretation

With model comparison table(See last page) and plot visualization, we can conclude:

- For second-hand Lexus cars with one year older of use, prices are 12% lower, on average.
- r-square = 0.74 means 74% of the variation in ln(Price) is captured by the regression, and 26% is left for residual variation.

Residual Analysis

We can use residual result to find the best deals. The most negative residuals will be the best car deals for us.

Name	Age	Price	l_reg_y_pred	l_reg_res	Year
Used 2015 Lexus CT 200h	6	11749	10.079166	-0.7076431	2015
Used 2006 Lexus IS 250 AWD	15	4000	9.029810	-0.7357604	2006
Used 2015 Lexus CT 200h	6	11500	10.079166	-0.7290642	2015
Used 2012 Lexus CT 200h	9	8499	9.729381	-0.6816772	2012
Used 2005 Lexus ES 330	16	3588	8.913215	-0.7278646	2005

	Ln_Price	Ln_Price	Ln_Price
(Intercept)	10.78 ***	10.81 ***	10.85 ***
	(0.02)	(0.02)	(0.03)
Age	-0.12 ***		-0.14 ***
	(0.00)		(0.01)
lspline(Age, c(10, 15, 20))1		-0.12 ***	
		(0.00)	
lspline(Age, c(10, 15, 20))2		-0.11 ***	
		(0.02)	
lspline(Age, c(10, 15, 20))3		-0.04	
		(0.05)	
lspline(Age, c(10, 15, 20))4		0.04	
		(0.23)	
Age_sq			0.00 *
			(0.00)
nobs	1106	1106	1106
r.squared	0.75	0.75	0.75
adj.r.squared	0.75	0.75	0.75
statistic	1473.34	11776.26	1119.83
p.value	0.00	0.00	0.00
df.residual	1104.00	1101.00	1103.00
nobs.1	1106.00	1106.00	1106.00
se_type	HC2.00	HC2.00	HC2.00

*** p < 0.001; ** p < 0.01; * p < 0.05.