STAT 503 Case Study 1: Restaurant Tipping

1 Description

Food server's tips in restaurants may be influenced by many factors including the nature of the restaurant, size of the party, table locations in the restaurant, ... To make appropriate assignments (which tables the food server waits on) for the food servers, restaurant managers need to know what these factors are. For the sake of staff morale they must avoid either the substance or appearance of unfair treatment of the food servers, for whom tips are a major component of pay. In one restaurant, a food server recorded the following data on all customers they had served during a interval of two and a half months in early 1990. The restaurant, located in a suburban shopping mall, was one of a national chain and served a varied menu. In observance of local law the restaurant offered seating in a non-smoking sections to patrons who requested it. The data was assigned to those days and during those times when the food server was routinely assigned to work.

The data available are:

TOTBILL	Total bill, including tax, in dollars
TIP	Tip in dollars
SEX	Sex of person paying bill (0=male, 1=female)
SMOKER	Smoker in party? (0=No, 1=Yes)
DAY	3=Thur, 4=Fri, 5=Sat, 6=Sun
TIME	0=day, 1=night
SIZE	Size of the party

This is a great data set in many ways because it is clearly a pilot study. There should be no temptation to make inference from the data, and emphasis should be on poking around the data to formulate hypotheses, and design a careful study.

The main question is: "Which factors most affect tip rate?"

2 Suggested Approaches

Approach	Reason	Type of questions addressed
Data Restructuring		
Make new variable Tip Rate from Tip/Total	Tip is usually referred to by percentage points, or as a rate. This also 'calibrates' the variable according to the bill total and allows us to compare values across the other variables such as size of the party.	
Make dummy variable for day of the week	This is a categorical variable so it is not appropriate to treat it as an ordinal value. This is especially important for a regression analysis.	
Summary statistics (marginal and conditional)	extract location/scale information	"What is the average tip rate at the restaurant?", "Are tips higher on Saturdays than on Thursdays?", "What is the average party size?"
Histograms (with different bin widths)	explore univariate distributions	"Are there unusually large or small tips?", "Is there a pattern to the way people tip?"
Pairwise Scatterplots (marginal and conditional)	explore bivariate distribution and correlation structure	"Are there unusually large or small tips relative to the total bill?", "Is there a pattern to the way people tip relative to the total bill?"
Mosaic Plots	explore multivariate (categorical) distributions	"Does the proportion of females paying the bill change with the day of the week?"
Regression	Determining the most important factors to tip rate	"Which factors contribute to higher tips?"

We don't have to do anything really sophisticated with this data. It is almost entirely categorical except for tip and total bill. So we will make extensive use of conditional plots.

3 Actual Approaches

3.1 Summary Statistics

Number of Observations = 244.

Table 1: Averages and Standard Deviations of Variables. The average bill is $20, average tip is $3 and average tip rate is $16. The average number of diners per party are 2.5.

	TOTBILL	TIP	TIPRATE	SIZE
Mean	$19.78	$3.00	16.1%	2.57
SD	$8.90	$1.38	6.1%	0.95

Table 2: Counts (Proportions) for Gender of the Bill Payer and Smoking Parties. More males paid the bills than females, and more parties were seated in the non-smoking section.

Gender/Smoke	No	Yes	Total
Male	97 (0.40)	60 (0.25)	157 (0.64)
Female	54 (0.22)	33 (0.14)	87 (0.36)
Total	151 (0.62)	93 (0.38)	244 (1.00)

Table 3: Counts (Proportions) for Time of the Day and the Day of the Week. There were no dining parties for this waiter at lunch time on Saturday and Sunday, and only one dining party at dinner on Thursday. Most od the dining parties served by this waiter were Thursday at lunch and Saturday and Sunday at dinner.

Time/Day	Thur	Fri	Sat	Sun	Total
Lunch	61 (0.25)	7 (0.03)	0 (0.00)	0 (0.00)	68 (0.28)
Dinner	1 (0.00)	12 (0.05)	87 (0.36)	76 (0.31)	176 (0.72)
Total	62 (0.25)	19 (0.08)	87 (0.36)	76 (0.31)	244 (1.00)

Table 4: Tip Rate (total number of diners) broken down by Day of the week and Time of Day. Over the days of the week Friday has very few diners, and Saturday has the most diners. There are no dining parties for lunch on the weekends. During the week there are more dining parties for lunch than dinner, but on weekends there are more dining parties for dinner. The lowest tip rate occurs on Saturday night. The highest is Friday lunch but there were only 7 dining parties then.

	Thurs	Fri	Sat	Sun
Day	16.1 (61)	18.9 (7)	0	0
Night	16.0 (1)	15.9 (12)	15.3 (87)	16.7 (76)
Total	16.1 (62)	17.0 (19)	15.3 (87)	16.7 (76)

Table 5: Tip Rate (total number of diners) broken down by Sex and Smoker. The average tip rate paid by females is higher than that paid by males, non-smokers is lower than for smokers. But note that female smokers gave the highest average tip, and that male non-smokers paid a higher average tip rate than did female non-smokers.

Tip Rate (Total Number)	Male	Female	Total
Non-Smoker	16.1 (97)	15.7 (54)	15.9 (151)
Smoker	15.3 (60)	18.2 (33)	16.3 (93)
Total	15.8 (157)	16.6 (87)	16.1 (244)

Table 6: Tip Rate (total number of diners) broken down by Size of the Party. The most dining parties were of size 2, and there were very few dining parties of size 1, 5 or 6. Parties of size 2,3 and 4 show a decreasing trend in tip rate as size increases.

Size of Party	1	2	3	4	5	6
Tip Rate (Total Number)	21.7 (4)	16.6 (156)	15.2 (38)	14.6 (37)	14.1 (5)	15.6 (4)

3.2 Histograms of tips: varying bar width

3.2.1 $1 barwidth

The shape of the data is skewed to the right, which says that there are more small tips, and fewer large tips.

It is unimodal with a peak at $1-2 and gradually falling off to the right. This says that there are more groups of diners that pay tips in the range $1-4. Assuming there is a relationship between tip and total bill this suggests that the total bills may be mostly $15-60 roughly. It suggests the restaurant is not an expensive one.

There are very few tips less than $1.

3.2.2 $0.50 barwidth

It is no longer unimodal but bi- or even tri-modal. There is more to the tipping habits than implied by the large barwidth of $1.

Also several very high tips are now apparent in the data. These are outliers, but from a waiting perspective these are the customers that you like!

3.2.3 $0.25 barwidth

It is clearly multimodal in shape now. There are large peaks at $2,3,4,5 which suggests that patrons have a habit of rounding their tips to the nearest dollar amount.

(Personal comment: I wonder if this behavior is different depending on whether the bill paid by credit card or cash.)

3.2.4 $0.10 barwidth

This is a very fine resolution plot. There are now secondary peaks visible at $1.50,2.50,3.50 which suggests rounding not just to the nearest dollar amount but many people round to the nearest 50c.

Also, it is clear that there are no tips recorded which are less than $1.

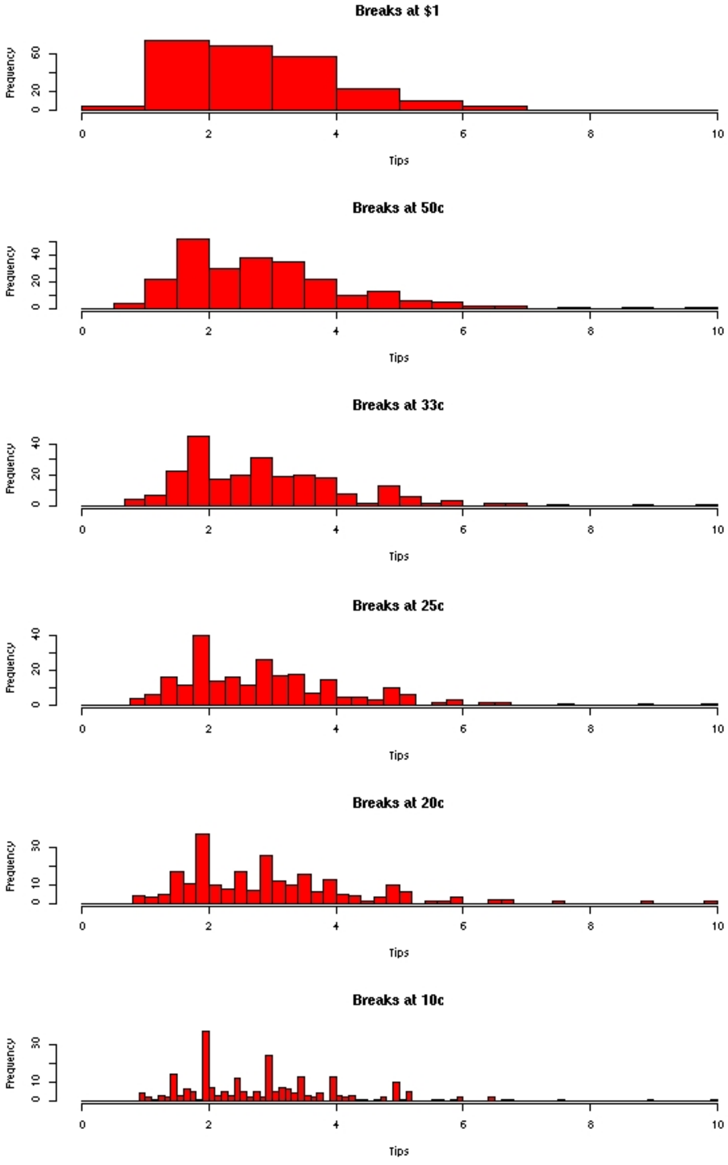

Figure 1: Histograms of Actual Tips with differing barwidth: $1, 50c, 33c, 25c, 20c, 10c.

3.3 Pairwise Scatterplots

3.3.1 Tip versus Total Bill

There are more patrons that pay lower tips and lower total bills, and fewer and fewer spend at the high end, that is, it is skewed as we observed when looking at tips alone.

There are lines of points visible at the tips level of $2,3,4 and to the keen eye also at the $1.50,2.50,3.50 amounts. This says that same as what we learned from the histograms that patrons tend to round their tips to the nearest dollar or half dollar amounts.

There is also a lower bound of tip at $1.

There is a linear relationship between tip and total bill, as you might expect. If you eyeball the the trend it looks to be about 15%. But there are a lot more points below the diagonal than above, that is, more patrons tip below the recommended level of 15% than above.

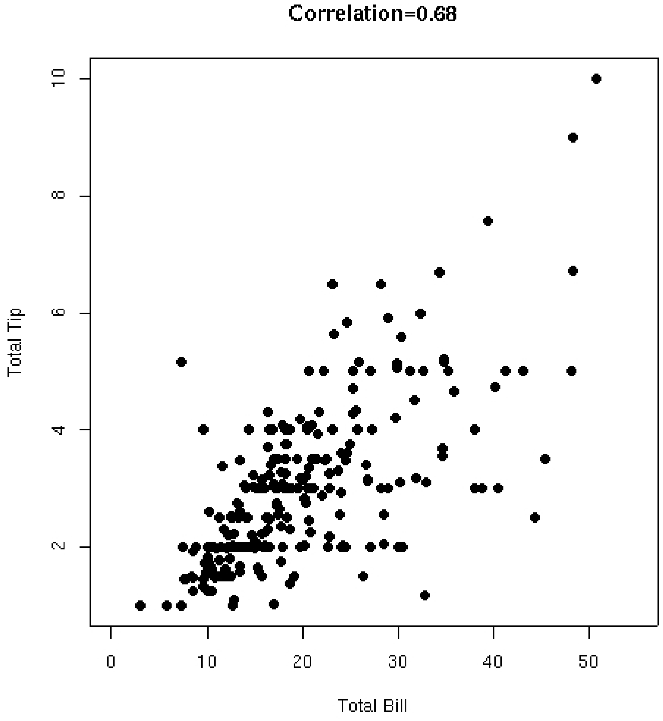

Figure 2: Scatterplot of Total Tip vs Total Bill. More points in the bottom right indicate more cheap tippers than generous tippers.

3.3.2 Tip vs Total Bill by Sex of the Bill Payer

(Note: The scale on both plots is the same, to allow for direct comparison.)

There are more larger bills paid by males, but not too many. The larger tips/larger bills are paid by males.

The eyeballed slope looks higher for the males than for the females.

(Note that we don't have any more information on the remaining members of the dining parties or even the size at this point, so we cannot tell whether the groups are dining together for business, as friends or for a romantic rendezvous. So we cannot infer whether a guy is trying to impress his girlfriend on their first date or a business colleague.)

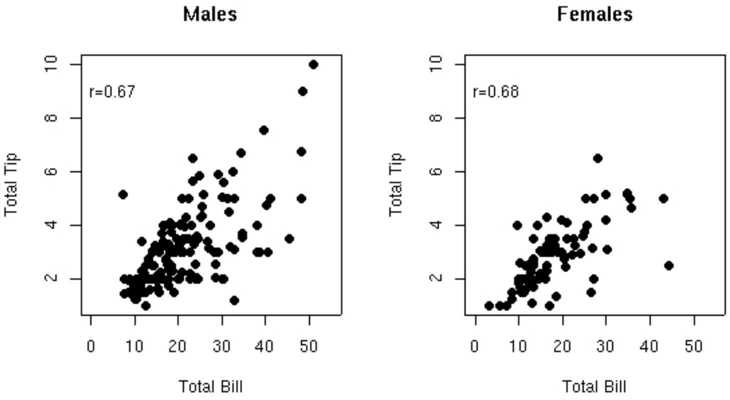

Figure 3: Scatterplot of Total Tip vs Total Bill by Sex.

3.3.3 Tip vs Total Bill by Sex of the Bill Payer and whether there was a Smoker in the Party or Not

Smokers plots show a lot more variability than the non-smokers. There is a little relationships between the total bill and the tip for smoking parties! (What are they smoking?)

Female non-smokers, with the exception of 3 low tippers are very consistent tippers as observed by the lack of variability.

3.4 Histograms of Tip Rate

There are two unusually high tips. Observation 173 has a tip rate of 71%. This tip was paid by a male, tip of $5.15 for a $7.25 bill consumed by smoking party with 2 people, at dinner time on Sunday. Observation 179 has a tip rate of 42%, which was paid by a female for a tip of $4.00 on a $9.60 bill for items consumed by a smoking party of 2 at dinner time on Sunday.

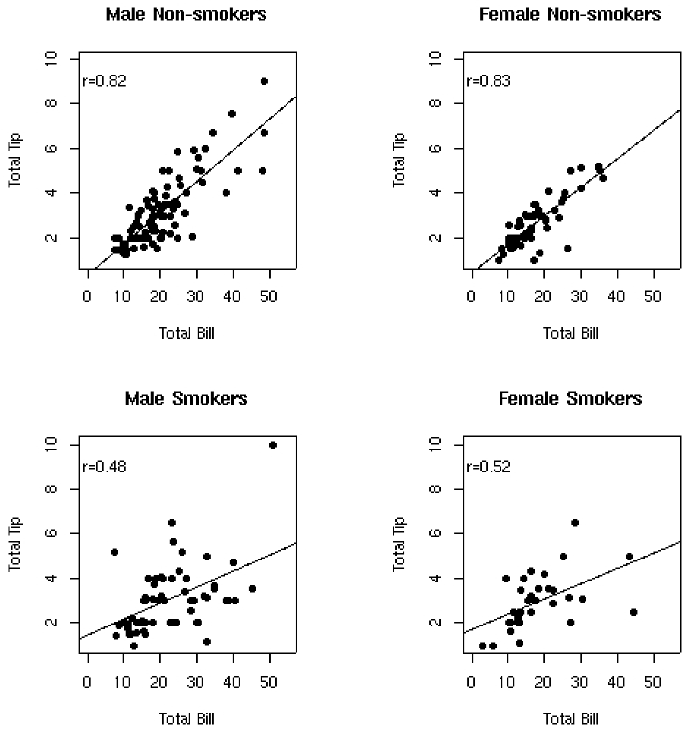

Figure 4: Scatterplot of Total Tip vs Total Bill by Sex and Smoker.

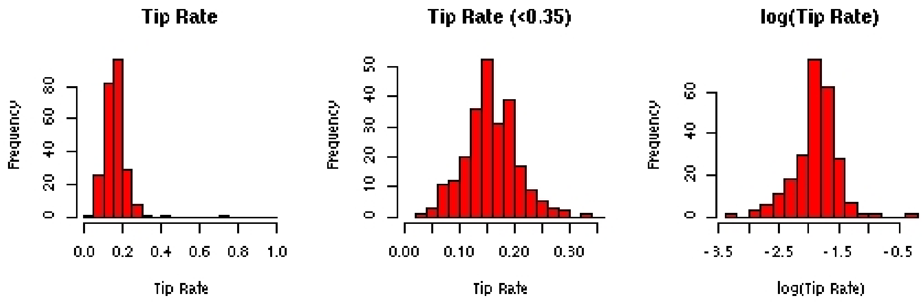

Figure 5: Histograms of Tip Rate: (left) all data, (middle) less than 35%, (right) log of tiprate. Tip rate looks skewed but it is mostly due to the two large tips.

3.5 Mosaic Plots

A mosaic plot is a multivariate extension of the histogram, or to be more precise a spine plot. It is specifically designed to explore conditional relationships between multiple categorical variables. In a histogram the height of the bar corresponds to the count for that category. In a spine plot it is the width of the bar that corresponds to the category count. The mosaic plot begins by splitting the horizontal axis into bars with width corresponding to the first variable. The vertical axis is then used for the second variable, with the height of the bar corresponding to the relative count of the second variable conditional on the category of the first variable. This creates a mosaic of rectangles corresponding to the conditional count for each category of the second variable given the category of the first variable. Further variables can be introduced by interleaving them in the horizontal and vertical directions. The order of the entry into the mosaic plot changes the information that can be perceived.

Figure 6 shows a mosaic plot of the variables Day of the Week and Sex of the Bill Payer. We can see the conditional distributions of Sex given the day of the week. On Thursday and Friday there are roughly equal numbers of males and females paying the bills but Saturday and Sunday see more males paying the bills.

Figure 7 shows a mosaic plot of Smoking Party and Sex of the Bill Payer. There is no relationship between Smoking and Sex, as seen by the roughly equal rectangles across categories. It also displays a mosaic plot of Day of the Week vs Time of Day which shows that Saturday and Sunday there are no measurements taken during the day. Thursday there are very few evening measurements, and Friday there are similar numbers of day to evening measurements.

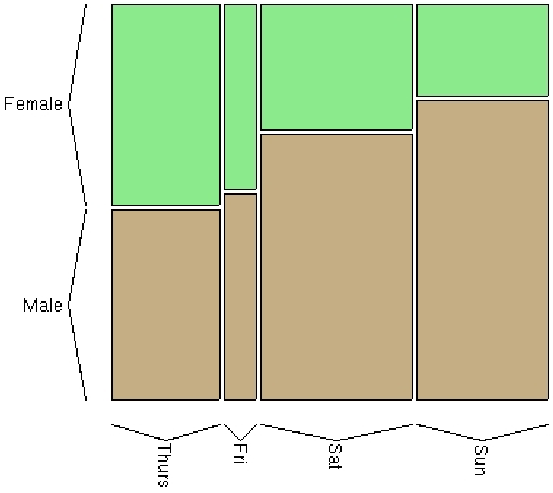

Figure 6: Mosaic Plot of Day of the Week vs Sex of the Bill Payer.

Figure 8 shows a mosaic plot of Size of the Party and Day of the Week and Sex of the Bill Payer. It can be seen that there are mostly parties of size 2 (something noticed in the summary statistics). The distribution of male to female bill payers replicates the trend noticed earlier only in parties of size 2. Otherwise in parties of size 1, the sex of the bill payer depends on the day of the week: mostly males on Friday, but mostly females on Thursday and Saturday, and no single diners on Sunday. In parties of size 3 Thursday saw roughly equal male and female bill payers, Friday mostly females, Saturday more males, but Sunday roughly equal. In parties of size 4 roughly equal males and females paid the bills on Thursday but more males paid the bills on Friday to Sunday.

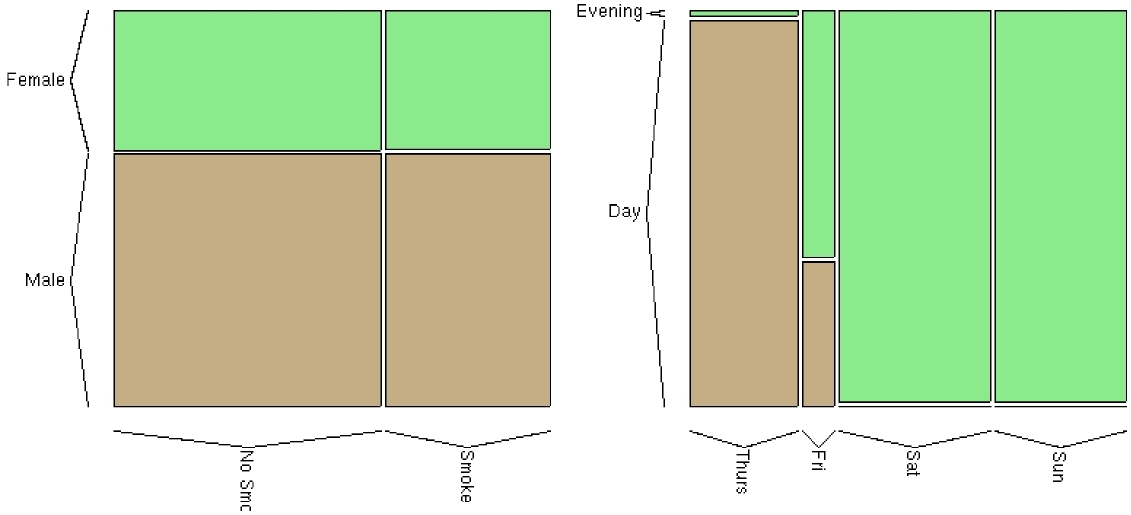

Figure 7: Mosaic Plot of (Left) Smoking Party vs Sex of the Bill Payer, (Right) Day of the Week vs Time of Day.

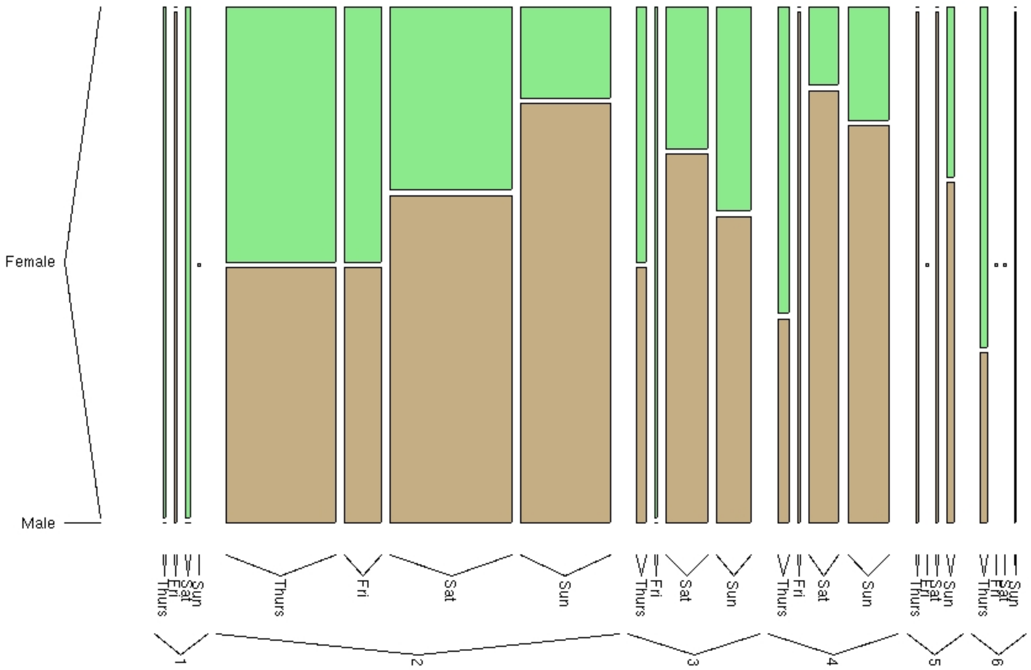

Figure 8: Mosaic Plot of Size of the Party vs Day of the Week vs Sex of the Bill Payer.

3.6 Regression

A general linear model is fit with Tip Rate as the response variable against all other variables. The details are below:

$$\begin{aligned}\hat{TipRate} &= 0.2147 + 0.0085 \times Sex + 0.0036 \times Smoker - 0.0234 \times Time - 0.0096 \times Size - \\ &\quad 0.0348 \times Thurs - 0.0167 \times Fri - 0.0184 \times Sat\end{aligned}$$

Term	Estimate	Std. Error	t-value	p-value
Intercept	0.2147	0.0300	7.183	0.00^{***}
Sex	0.0085	0.0083	1.023	0.31
Smoker	0.0036	0.0085	0.428	0.67
Time	-0.0234	0.0261	-0.895	0.37
Size	-0.0096	0.0042	-2.282	0.02^{*}
Thurs	-0.0348	0.0278	-1.253	0.21
Fri	-0.0167	0.019	-0.876	0.38
Sat	-0.0184	0.0098	-1.878	$0.06^{\cdot}$

For a model to explain the data perfectly the deviance for the model will be 0. The deviance for the above model is 0.868 which is just 0.038 less than deviance for the null model (intercept term only) of 0.906. This difference can be compared using a χ^2_7 distribution which yields a p-value equal to 1, that is the model doesn't fit the data well at all.

In the model, size of the party is the most important predictor of tip rate, and Saturday is the next most important.

We fit the simpler model of tip rate with just the size of the party. (Its not clear how day of the week should be used, since only Sat is important.)

$$\hat{TipRate} = 0.1844 - 0.0092 \times Size$$

Term	Estimate	Std. Error	t-value	p-value
Intercept	0.1844	0.0112	16.475	0.00^{***}
Size	-0.0092	0.0041	-2.245	0.0256^{*}

The explanation for the model is that tip rate drops by roughly 1% from 18% for each additional dining party member. The model deviance is 0.888 which is only 0.018 less than the null deviance of 0.906. This is also a non-significant model, when the difference is compared to a chi^2_1 critical value. Using $R^2 = 0.02$ the variable size of the party explains only 2% of the variation in tip rate.

In the data there are two extremely large tips, 71% and 42% in parties of size 2. It is interesting to see the effect of these two data values on the model. When they are excluded size of the party is the only important variable and the model changes to become:

$$\hat{TipRate} = 0.1757 - 0.0071 \times Size$$

The deviance for this model is 0.525 which is 0.011 less than the null model 0.536. This means that the model doesn't explain significant amount of the variation in tip rate, despite the fact that the coefficient for size of the party is significant;y different from 0. Using R^2 the model explains just 2% of the variation in tip rate. The estimated coefficients for the model with all variables is:

Term	Estimate	Std. Error	t-value	p-value
Intercept	0.2011	0.0231	8.715	0.00***
Sex	0.0074	0.0065	1.146	0.25
Smoker	-0.0077	0.0066	-1.169	0.24
Time	-0.0251	0.0201	-1.248	0.21
Size	-0.0069	0.0033	-2.108	0.04*
Thurs	-0.0243	0.0215	-1.130	0.26
Fri	0.0018	0.0148	0.121	0.90
Sat	-0.0041	0.0077	-0.540	0.59

Using log(tip rate) the results are effectively the same, although its trickier to interpret the coefficients. Thus we prefer to use the untransformed tip rate in the model.

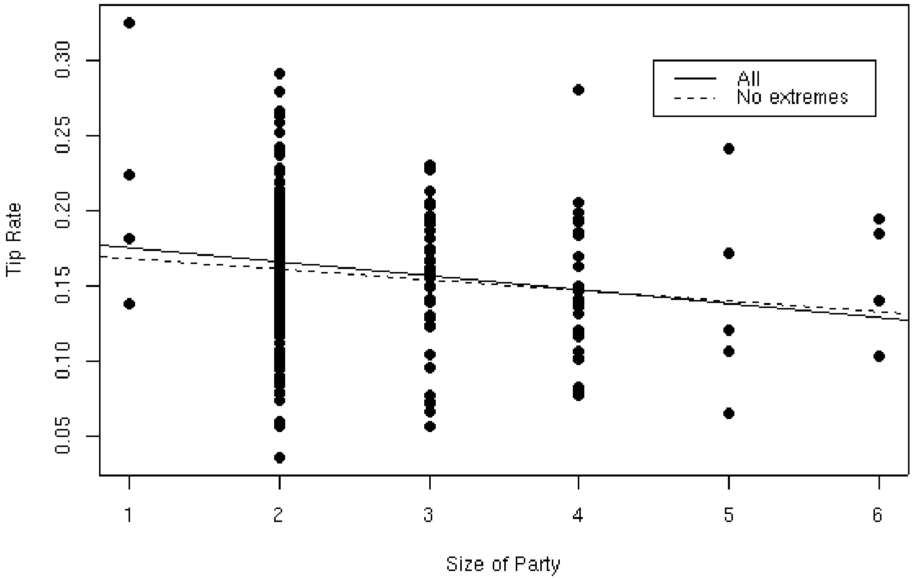

Figure 9: Scatterplot of Tip Rate vs Size of Party, with least squares fit overlaid. The scale on the vertical axis excludes the two extreme tip values, 0.71 and 0.42. The model without these two cases is fairly similar to the model that includes them.

Earlier we noticed a relationship between sex of the bill payer, smoking or not in relation to the tip and total bill. We can use regression to explore this relationship. Here we fit a model of tip rate against these two variables including an interaction term (with the two extremely large tips excluded).

Whats interesting is that when the interaction between sex and smoke is included in the model both variables become important predictors of the tip rate, whereas they are not important when the interaction term is excluded. Size of the party remains important as well. Then the model becomes:

$$Tip\hat{R}ate = 0.1793 - 0.0069 \times Size - 0.0046 \times Sex - 0.0187 \times Smoke + 0.0343 \times Sex * Smoke$$

The deviance for this model is 0.504 which is only 0.033 less than the null deviance, and not a significant model. A summary of the estimated coeffiecients is:

Term	Estimate	Std. Error	t-value	p-value
Intercept	0.1793	0.0097	18.398	0.00***
Size	-0.0069	0.0128	2.678	0.008**
Sex	-0.0046	0.0078	-0.583	0.561
Smoker	-0.0187	0.0076	-2.454	0.015*
Sex*Smoke	0.0343	0.0032	-2.182	0.030*

This model says that from a base tip rate of 18% each additional party member reduces the rate 0.7%, if a female pays the bill then the rate is reduced a further 0.5%, if it is a smoking party then the rate is reduced a further 1.9% and if it is a female who pays the bill of a smoking party then the rate is increased by 3%. This relationship roughly matches the summary statistics computed earlier. But it should be noted that although the coefficients are significantly different from 0, the model explains an insignificant amount of the variation in tip rate.

4 Summary of Findings

- Size of the party is the most important factor in predicting tip rate. The best regression model is $Tip\hat{R}ate = 0.1757 - 0.0071 \times Size$ which says that for each additional person in the dining party tip rate decreases a little less than a percentage point from 17.5%. A more complex model is $Tip\hat{R}ate = 0.1793 - 0.0069 \times Size - 0.0046 \times Sex - 0.0187 \times Smoke + 0.0343 \times Sex * Smoke$, which says that from a base tip rate of 18% each additional party member reduces the rate 0.7%, if a female pays the bill then the rate is reduced a further 0.5%, if it is a smoking party then the rate is reduced a further 1.9% and if it is a female who pays the bill of a smoking party then the rate is increased by 3%. Our data is really only adequate to measure this for dining parties of size 2, 3 and 4, because the other party size have very few data points.

- There is a tendency to give "cheap" tips rather than "generous" tips relative to the total bill.

- There is a tendency to round the tip to the nearest dollar, and to a lesser extent to the nearest 50c.

- There is much more variation in the tips given by smoking parties than non-smoking parties. And the female non-smokers are very consistent tippers.

- There are a couple very large tips. The largest, 71%, was given by a male smoker on Saturday night in a party of size 2. The next largest, 42%, was given by a female smoker on Saturday night in a party of size 2.

- Depending on the day of the week there appears to be increasingly more males paying the bill towards the weekend days but this is true only for the parties of size 2.

References

[Bryant and Smith, 1995] Bryant, P. G. and Smith, M. A. (1995). *Practical Data Analysis: Case Studies in Business Statistics*. Richard D. Irwin Publishing, Homewood, IL.