Unit 1 Regression Models-SRM (Solution)

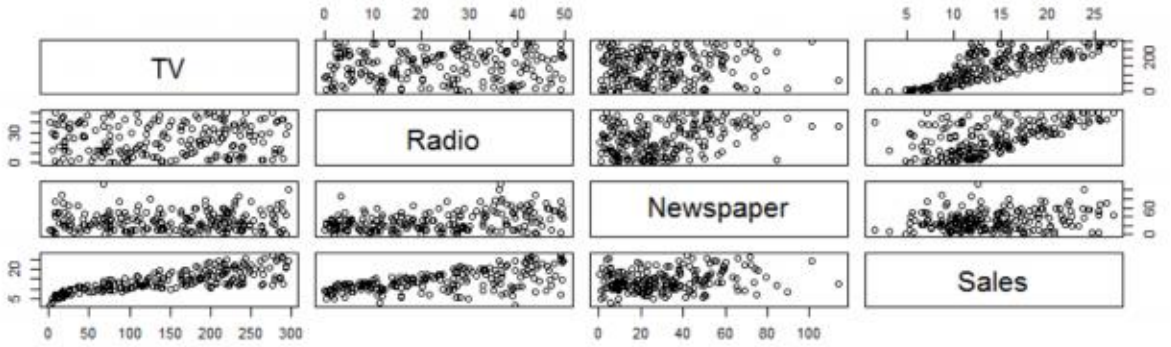
Name: Solution

Advertising Data

This data set includes "Sales (in thousands of units)" for a particular product as a function of "Advertising budgets (in thousands of dollars)" for TV, Radio, and Newspaper media.

Address a few important questions below;

1. Is there a relationship between advertising budget and sales?



Yes. Based on the pair wise correlation plot above, there are very clear positive linear relationship between TV and Sales. Also, the positive linear relationship is existed between Radio and Sales. However, the relationship between Newspapr and Sales is relatively

2. How strong is the relationship between advertising budget and sales?

```
> cor (Advertising [,-1])
                 TV      Radio  Newspaper     Sales
TV         1.00000000 0.05480866 0.05664787 0.7822244
Radio      0.05480866 1.00000000 0.35410375 0.5762226
Newspaper  0.05664787 0.35410375 1.00000000 0.2282990
Sales      0.78222442 0.57622257 0.22829903 1.0000000
```

According to the rule of thumb, there are a strong linear relationship between TV and Sales (r=.78), a moderate linear relationship between Radio and Sales (r=.58), and a weak linear relationship between Newspaper and Sales (r=.228).

3. Is the relationship linear?

Yes. All three relationships are linear beased on the scatter plot in part 1.

4. Which media contribute to sales?

All three media positively contribute to sales independently. The most effective media to increase sales is Radio by comparing the magnitude of coeffiients. However, the association between TV and Sales explains the greater variability of the data.

Media Type	Coefficients	R-squares
TV	.048***	.612
Radio	.203***	.332
Newspaper	.055**	.052

Note: *** p<.001; **p<.01

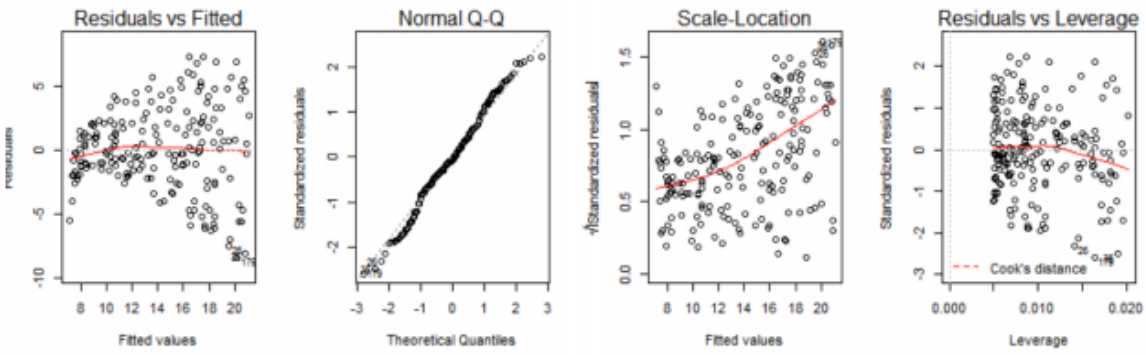
In real world, company allocates their budget to the all three media together, not all in the budget for one media. To reflect this reality, we need to refit the model by including all the variables together so that we can examine the relationships between media and sales simultaneously. We will talk how to extend the simple regression model next week.

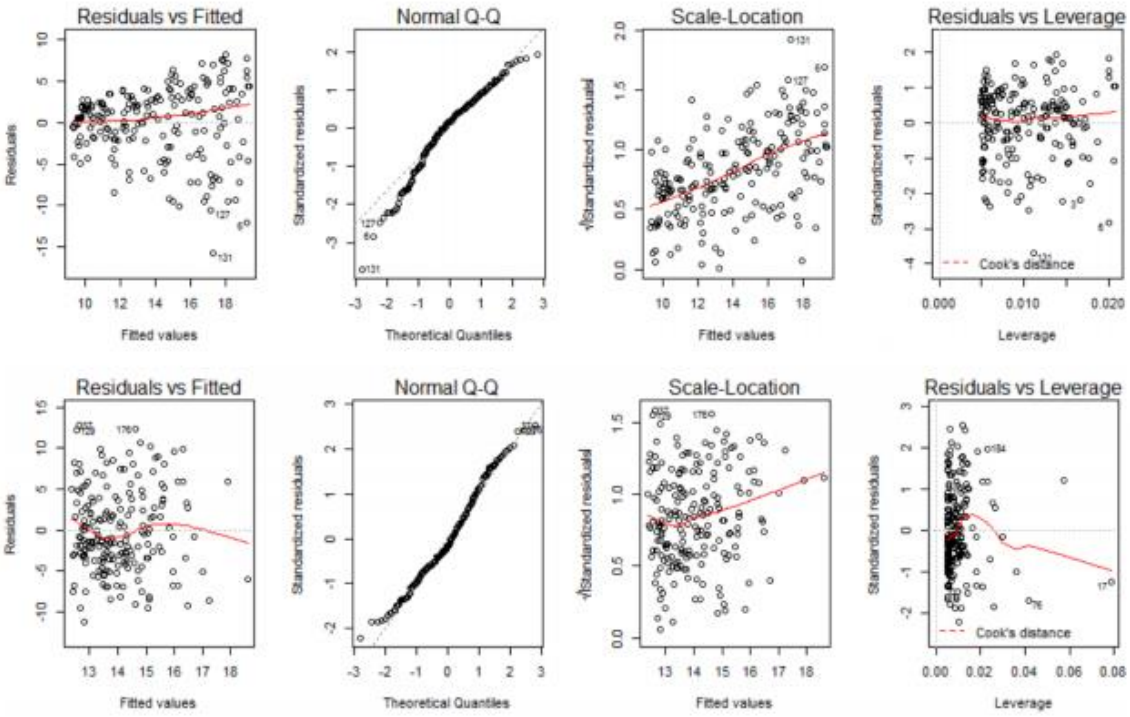
5. How accurately can we estimate the effect of each medium on sales?

Models	Predictor	Residual standard error	F-test
Model1	TV	3.259	F=312.1***
Model2	Radio	4.275	F=98.42***
Model3	Newspaper	5.092	F=10.89**

Note: *** p<.001; **p<.01

Models are reasonable which means that the estimations for the effects of each medium on sales are accurate since (1) residual standard errors are small and (2) models provide better fir to the data than a model that contains only intercept (significant F values). However, it is necessary to examine the potential outliers or add quadratic term (or interaction terms) to the model because the assumptions of the model are not satisfied baased on the residual plots and Normal Q-Q plot.





6. How accuracy can we predict future sales?

Generate New values and compute predicted values using the model 1 which is the best model;

```
> new
   TV
1 200
2 250
3 300
4 350
> pred.w.plim1
       fit      lwr      upr
1 16.53992 10.09162 22.98822
2 18.91675 12.45146 25.38205
3 21.29359 14.80049 27.78668
4 23.67042 17.13886 30.20198
> pred.w.plim2
       fit      lwr      upr
1 16.53992 16.00567 17.07418
2 18.91675 18.20619 19.62732
3 21.29359 20.36346 22.22371
4 23.67042 22.50160 24.83924
```

There are two ways to calculate errors in order to check the prediction accuracy.

First, we can split the whole data set into two parts, test set and training set. Then, we can fit a model using the training set and find the best model with the smallest error and the largest R-

square. Once we the best model is chosen, we need to refit the best model using the test set. Then, compare the errors.
Second, if we have actual future data, we can calculate predictive errors ($y_{actual} - \hat{y}_{predict}$). We will talk about this end of the semester when we learn time series models.

7. Is there synergy among the advertising media?

The synergy effects among the advertising media can be captured by interaction terms such as $TV \times Radio, TV \times Newspaper$, and $Radio \times Newspaper$.

We will talk about the details about the interaction terms later.

Amusement park data

For this in-class activity, we will use simulated data for a hypothetical survey of visitors to an amusement park. This dataset comprises a few objective measures: whether the respondent visited on a weekend (which will be the variable "weekend" in the data frame), the number of children brought (num.child), and distance traveled to the part (distance). There are also subjective measures of satisfaction: expressed satisfaction overall (overall) and satisfaction with the rides, games, waiting time, and cleanliness (rides, games, wait, and clean, respectively).
Using this data, address a few important questions below;

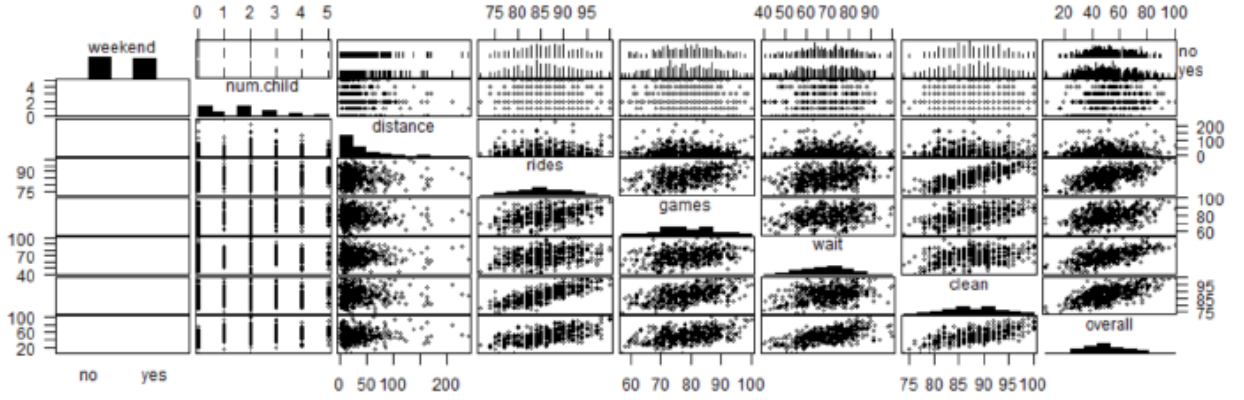
1. Create new data "sat.df" by adding subjective measures of satisfaction.

2. Justify the variable types of the "sat.df" data set

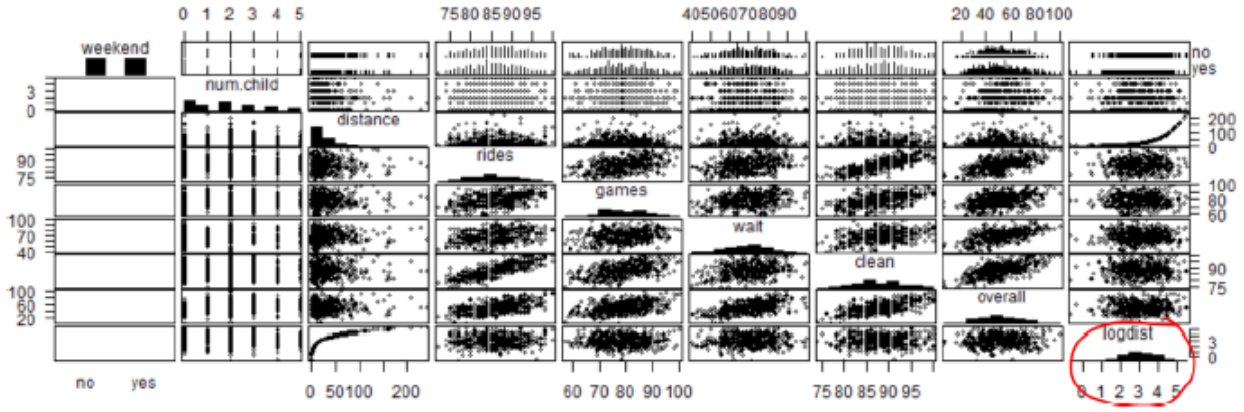
Amusement park dataset contains 8 variables with 500 observations. Among the 8 variables, only 'weekend' is a qualitative variable and the other 7 variables are all quanlitative variables. Using the "str()" function, we can easily check the variable types.

```
> str(sat.df)
'data.frame':   500 obs. of  8 variables:
 $ weekend  : Factor w/ 2 levels "no","yes": 2 2 1 2 1 1 2 1 1 2 .
 $ num.child: int  0 2 1 0 4 5 1 0 0 3 ...
 $ distance : num  114.6 27 63.3 25.9 54.7 ...
 $ rides    : int  87 87 85 88 84 81 77 82 90 88 ...
 $ games    : int  73 78 80 72 87 79 73 70 88 86 ...
 $ wait     : int  60 76 70 66 74 48 58 70 79 55 ...
 $ clean    : int  89 87 88 89 87 79 85 83 95 88 ...
 $ overall  : int  47 65 61 37 68 27 40 30 58 36 ...
```

3. Are there any significant relationships between subjective measures and overall satisfaction?
   Tip: Create a pairwise plot and transformed the variable if needed.



Distance is highly skewed to the right and there is curve-linear shape on the pairwise correlation plot. In order to improve this variable, we can consider the log-transformation.



After log-transformation, the distribution of the distance variable is more symmetric.

4. How strong are the relationships between subjective measures and overall satisfaction?

```
> cor(sat.df[ , c(2, 4:9)])
             num.child       rides       games        wait       clean     overall      logdist
num.child  1.000000000 -0.04026024 0.004658171 -0.02097292 -0.01345167 0.31948036 -0.004592229
rides     -0.040260243  1.00000000 0.455185111  0.31419951  0.78956505 0.58598628 -0.011027676
games      0.004658171  0.45518511 1.000000000  0.29910498  0.51697987 0.43746787  0.001868728
wait      -0.020972921  0.31419951 0.299104980  1.00000000  0.36788467 0.57262166  0.017460929
clean     -0.013451671  0.78956505 0.516979874  0.36788467  1.00000000 0.63939818  0.022123740
overall    0.319480357  0.58598628 0.437467872  0.57262166  0.63939818 1.00000000  0.076327893
logdist   -0.004592229 -0.01102768 0.001868728  0.01746093  0.02212374 0.07632789  1.000000000
```

Moderate relationships: # of child-satisfaction (marginally), rides-satisfaction, games-satisfaction, wait-satisfaction, clean-satisfaction
Weak relationship: log(distance)-satisfaction

5. Which subject measures to overall satisfaction? (For this group activity, students fit the model, $\widehat{Overall\ Satisfaction} = b_0 + b_1 Rides$)
   1) Fit simple linear regression models using lm().
   2) Interpret the slope of each model.

$\beta_{Num.Child} = 3.391$;For every additional number of children brought, the satisfaction level increases by 3.391 unit, on average.
$\beta_{Distance} = .042$; For every additional distance traveled to the park, the satisfaction level increases by 0.042 unit, on average.
$\beta_{Rides} = 1.70$; For every additional number of rides, the satisfaction level increases by 1.70 unit, on average.
$\beta_{Games} = .855$; For every additional number of games, the satisfaction level increases by .855 unit, on average.
$\beta_{Wait} = .844$; For every additional waiting time, the satisfaction level increases by .844 unit, on average.
$\beta_{Clean} = 1.99$; For every additional cleaness (in unit), the satisfaction level increases by 1.99 unit, on average.

3) Test significance of the coefficients.

Models	Predictors	Coefficients	Residual standard errors	R-square	F-test	
Model 1	Weekend	-1.684	15.87	.003	1.41	Do not reject
Model 2	# of child	3.391***	15.06	.102	56.61***	Reject
Model 3	Distance	.042**	15.83	.008	3.81**	Reject
Model 4	Rides	1.70***	12.88	.343	260.4***	Reject
Model 5	Games	.855***	14.29	.191	117.9***	Reject
Model 6	Wait	.844***	13.03	.328	243***	Reject
Model 7	Cleaness	1.99***	12.22	.409	344.4***	Reject
Model 8	Log(Distance)	1.23	15.85	.006	2.92	Do not reject

Note: *** p<.001; **p<.05

6. How accurately can we estimate the effect of each subjective measures on the overall satisfaction? Look at the summary of the model and interpret the statistics of the model evaluation.

The estimated model is $Overall\ \widehat{Satisfaction} = -94.96 + 1.70 \times Rides$ with relatively low standard errors (=12.88). This simple regression model is reasonable (or significant) since the f-test is significant with small p-value. In addition, the linear relationship between Rides and Overall satisfaction is addressed variability of the data set about 34%.

Considering 8 separate simple regression models, we can make comments like as following;

According to the F-test, model 4-7 accurately measure the effect of subjective factors on the overall satisfaction. The relationships between overall satisfaction and each subjective measures explain the data 19%-41%. Still the models capture the data below 50%, but it is acceptable. Also, all the slopes of the subjective measures are statistically significant at $\alpha = 0.05$.