Relationship between Pace of Life and Heart Disease

Shivangi Sinha

2/24/2019

INTRODUCTION

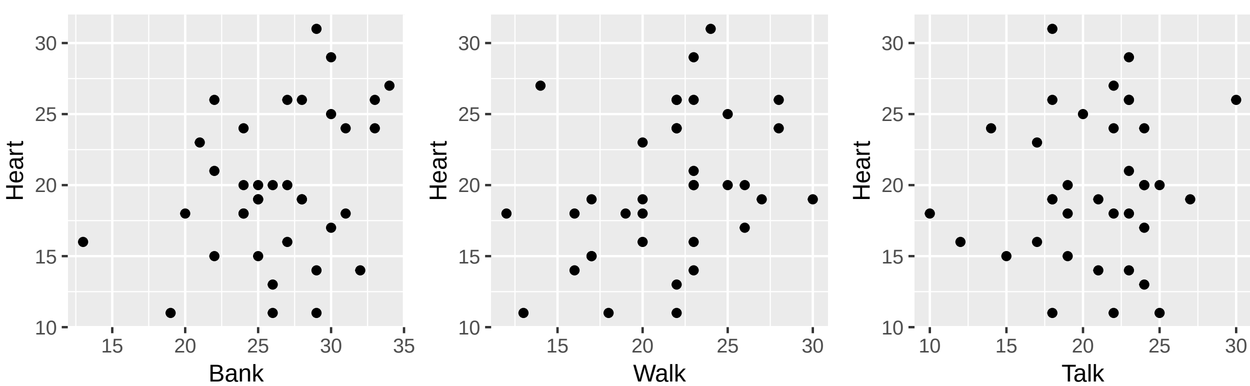
In four regions of the US (Northeast, Midwest, South and West), the researchers chose three different metropolitan cities, three medium sized cities, three smaller cities, and then measured the indicators of pace of life. The aim of the study was to understand the relationship between heart diseases and pace of life. They used following variables to measure pace of life:

- **Walk -** The variable Walk is the walking speed of pedestrians over a distance of 60 feet during business hours on a clear summer day alog a main downtown street.
- **Bank -** It is the average time a sample of bank clerks take to make change of two $20 bills.
- **Talk -** It is the recorded responses of postal clerks explaining the difference between regular,cerified, and insured mail and by diving the total number of syllables by the time of their responses.
- **Heart -** The researchers also obtained the age adjusted death rates from ischemic heart disease for each city.

Numerical Summary

	Bank	Walk	Talk	Heart
Mean	26.36	21.42	20.75	19.81
Minimum	13	12	10	11
Maximum	334	30	30	31

Graphical Summary



From graphical analysis, it seems like there is no relationship between heart disease and talking and a very mild linear relationships between heart disease and bank time and walking speed.

METHODS

Using "Heart" variable as the response and rest of the variables are predictors,we tried to fit all the possible linear regression equation and chose the one with the highest adjusted R^2.

The linear equation fitted on the data is

$$Y_i = \beta_0 + \beta_1 X_{(Walk)i} + \beta_2 X_{(Talk)i} + \beta_1 X_{(Bank)i} + \epsilon_i, i = 0, 1...36$$

- Y_i is the observed value of age adjusted death rates.
- X_i's are the explanatory variables corresponding to walking speed, postal * clerk talking speed and bank clerk speed vrespectively.
- β's are the parameters that are to be estimated.
- ϵ_i is the error terms

Using this model we will be a conducting F test to check if pace of life indicators are related to the response that is the heart disease death rate. To check if the all of the explanatory variables are related with the death rate we will be using following F test

$H_0 = \beta_1 = \beta_2 = \beta_3 = 0$ vs H_A=At least one of them is not zero

$$F = \frac{MSR}{MSE} \sim F(p-1, n-p)$$

n is the number of observations which is 36 in our case and p is the number of β's which is 4. * MSR is the mean regression sum of squares * MSE is the mean square error

In order to test for each individial explanatory variable we will be using F-test for reduced model. Null Hypothesis Let Ω denote a larger model of interest with p parameters and ω a smaller model that represents some simplification of Ω with q parameters. Null hypothesis: the simplification to Ω implied by the simpler model, ω

F statistic

$$F = \frac{(RSS_\omega - RSS_\Omega)/(p-q)}{RSS_\Omega/(n-p)} \sim F(p-q, n-p)$$

Assumptions for Inference

- We are assuming that heart disease rate is linearly related to Walking speed,Bank clerk speed and Postal clerk talking speed.
- All the observations are independent to each other.
- The error term follows normal distribution with mean 0 and constant variance σ^2

RESULTS

After fitting the above model, we derived the following estimates-

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.179	6.337	0.5016	0.6194
Bank	0.4052	0.1971	2.056	0.04803
Walk	0.4516	0.2009	2.248	0.03158
Talk	-0.1796	0.2222	-0.8083	0.4249

Table 3: Fitting linear model: Heart ~ .

Observations	Residual Std. Error	R^2	Adjusted R^2
36	4.805	0.2236	0.1509

F statistics and degree of freedom

value	numdf	dendf
3.073	3	32

Conducting a F test between these two models to check in β associated with Talk is zero.

$$Model1 Y_i = \beta_0 + \beta_1 X_{(Walk)i} + \beta_2 X_{(Talk)i} + \beta_1 X_{(Bank)i} + \epsilon_i, i = 0, 1...36$$

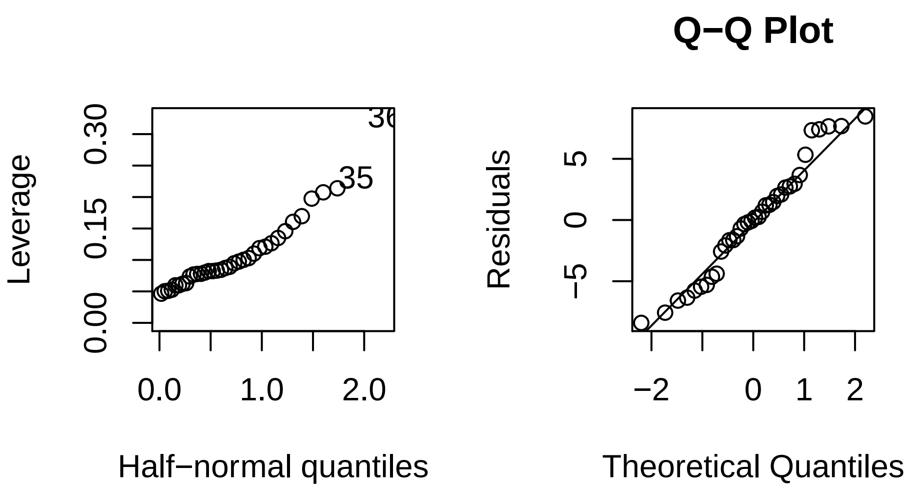
$$Model2 = Y_i = \beta_0 + \beta_1 X_{(Walk)i} + \beta_2 X_{(Bank)i} + \epsilon_i, i = 0, 1...36$$

Table 5: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
32	738.8	NA	NA	NA	NA
33	753.9	-1	-15.08	0.6533	0.4249

Since p value is > 0.05, we fail to reject the null hypothesis.

Odd observations and Normality



Obervation number 36 and 35 seem to have high leverage. While checking for assumption of normality we see that, except the end points, we see that residuals do moderately follow normal distribution.

CONCLUSIONS

- Estimates : With one unit change in the speed of bank clerk,speed of talking of postal clerk and walking speed of pedestrian the change in the death rate is equal to 0.4052,-0.17960 and 0.4516 respectively. The adjusted R^2 is 0.1509 which is low for a linear model. Thus, I think we should try to get a bigger sample to get more accurate results.
- F-Statistic and Estimated Sigma : Sum of square residuals is equal to 4.805 and F statistic for overall regression is equal to 3.073 whose p-value is less than 0.05. Thus, we find convincing evidence that pace of life indicators are related to the heart disease.
- Reduced Model for Regression : While conducting F-test for model with all the explanatory variables and the model with only speed of walking and bank clerk, we find that the p-value is less than 0.05. Thus,there is convincing result that talking speed of postal clerk may or may not be associated with death rate.
- Odd observation : By ploting halfnormal plot,we see that observation number 35 and 36 have high lervage. Thus, the regression line with these values will tend to tend to move towards those points while the regression line without them will move away.