ST552 Homework 7

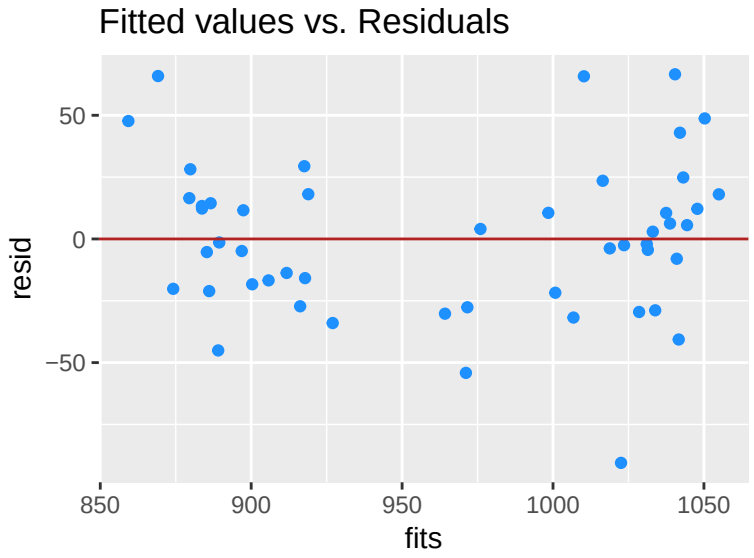
Nick Sun

February 24, 2019

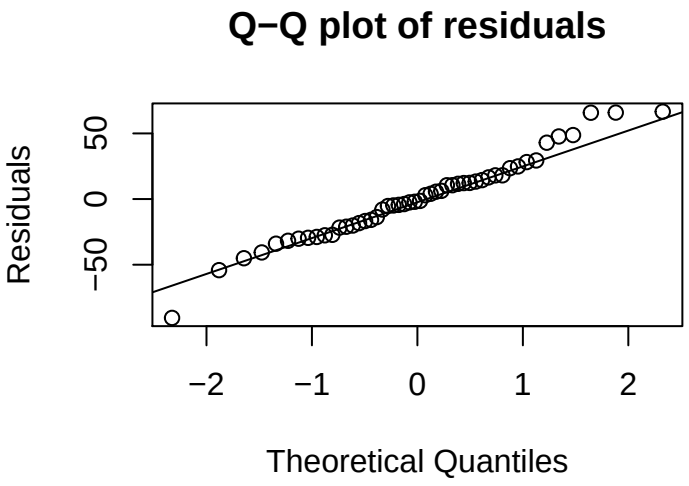
Faraway 6.1

Taking a look at SAT data! Sure does take me back.

We can begin by checking the constant variance assumption of the errors. To do this, we can examine a residual plot versus the fitted values.



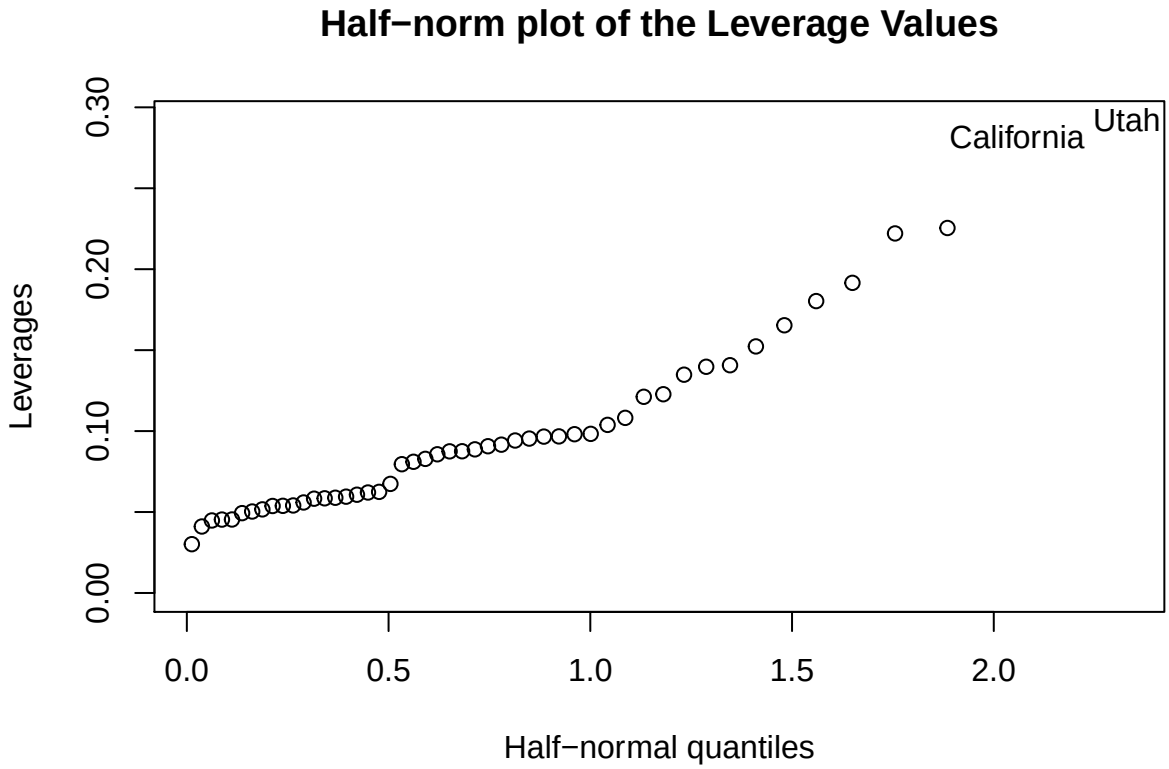
The residuals look relatively evenly distributed and centered around 0, so at least the residuals appear homoscedastic. We can further examine normality using a Q-Q plot.



This doesn't look awful, but the tails are worrying. Also there appears to be a rise in the middle of the quantile plot. We can formally test this using the Shapiro-Wilk Test.

From the results of the Shapiro-Wilk test, we don't have evidence to say that the residuals are not distributed normally.

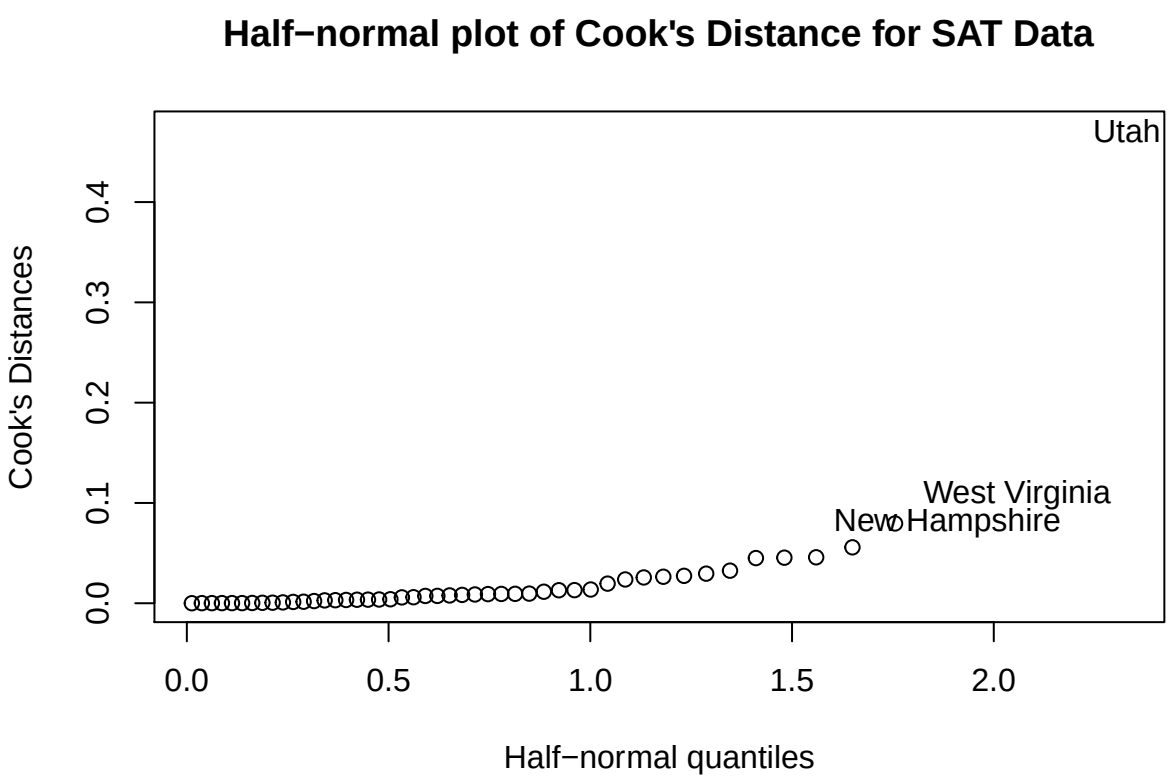
We can check other diagnostics such as leverage.



There are two observations with very high leverage, California and Utah. We can look at the values for these rows and see if we find anything strange when compared with the rest of the data.

We can see that Utah has the lowest amount of *takers* while California is near the top in ratio, salary, and total score. We might be interested in examining these high leverage points as outliers.

We can use Cook's Distance to find influential points in the data:



Utah appears here again and it is really not even close. This data point has a **ton** on influence and leverage.

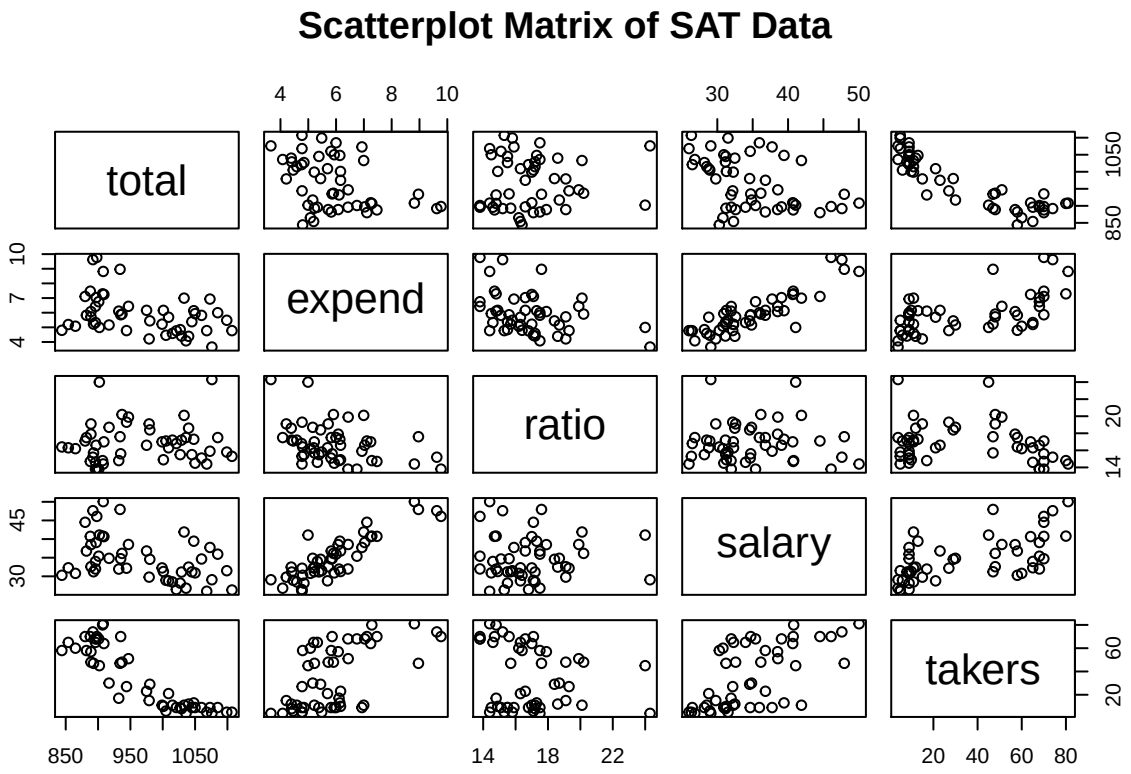
We can take a look at the studentized residuals and perform a t-test to get a firmer grasp on which observations are outliers.

```
##  West Virginia         Nevada South Carolina        Wyoming          Texas
##      -3.124428      -1.732004      -1.468832      -1.311890      -1.070881
```

The state with the highest studentized residual is actually West Virginia. Go Mountaineers! Surprisingly, Utah and California do not appear in the top 5 highest studentized residuals, indicating that while they are high leverage they are probably not outliers.

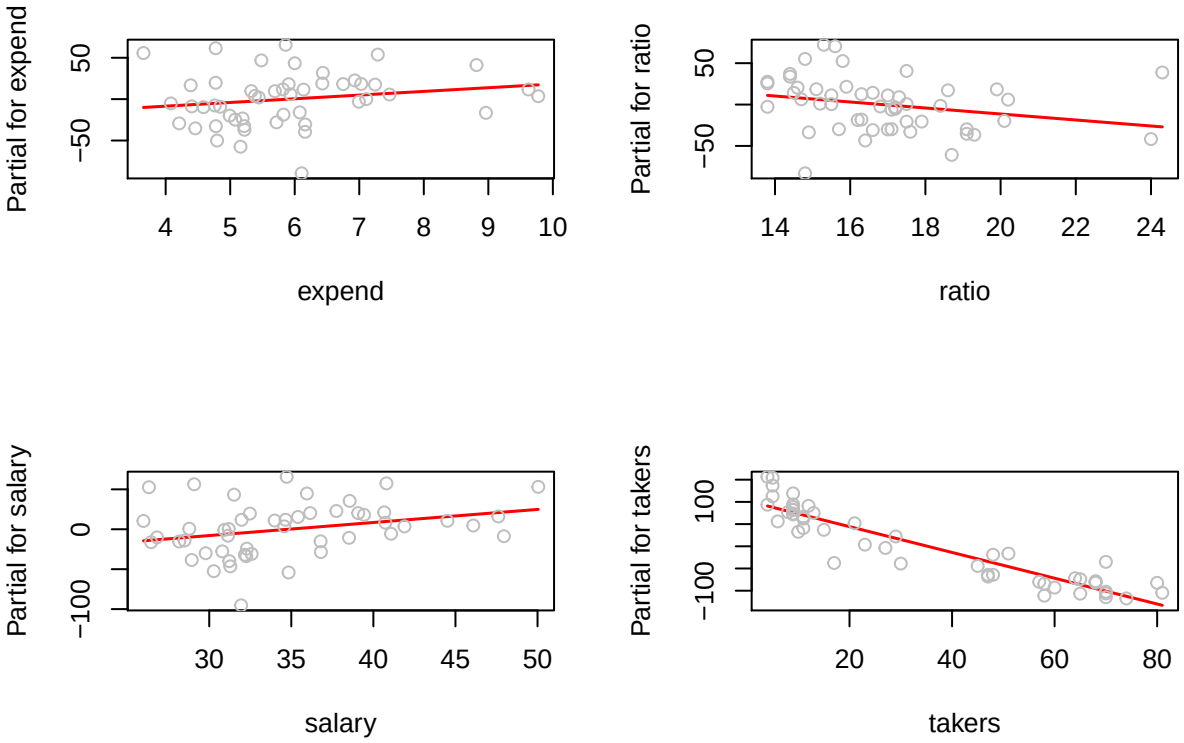
Using a t-test with a Bonferroni corrected p-value we have the following critical value: -3.52. Our largest studentized residual is -3.124 and is smaller than our critical value. Therefore, we have no statistically significant evidence to say that West Virgina is an outlier. This conclusion comes with caveats in that we are using a very conservative correction - if we were to use a less conservative procedure such as Holm or Benjamini-Hochberg, we might arrive at a different conclusion.

We can examine the relationship structure of the predictors and the covariates using a scatterplot matrix.



There are some interesting relationships here! There appears to be a distinct negative correlation between **total** and **takers** – as the number of **takers** increase, the total score on average appears to go down!

Salary and **expend** appears to have a positive correlation - as **salary** increases, expenditure on education increases.



These partial residual plots are made by plotting the predictor variables with the partial residuals of that predictor variables. Partial residuals are calculated in R by first centering the predictor values and then multiplying them by their respective coefficients. These values are referred to as *terms*[1]

These terms are then added to the residuals to make partial residuals. The main takeaway from these plots is the slope of the red line. This corresponds to the effect of that particular predictor when accounting for the other variables. If there is a strong linear relationship, then we have some indication that the particular variable has an effect on the response. Using partial residual plots helps us identify collinearity since some variables might appear to have a relationship with the predictor *only* because they have a relationship with another variable. Once that relationship is accounted for, it is clearer to see that there is really no association with the response.

In our case, there is one variable *takers* which really stands out. The clear negative relationship corresponds to the significant p-value for the t-test as well as the effect it has on the response. The other variables have less of an effect on our response. In particular, *expend* and *salary* only have a slight positive relationship with the response. These termplots are more or less centered around 0, so we might even have evidence to say that *expend* and *salary* really have no effect on the response when the other variables are accounted for. The t-tests back up this argument, since neither *expend* not *salary* are significant at $\alpha = .05$

Similarly, *ratio* only has a slight negative relationship with the response. The red line is centered close to 0, so we might also be suspicious of the variables effect (or lack thereof) on response.

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 1045.971536  52.869760  19.7839283 7.857530e-24
## expend         4.462594  10.546528   0.4231339 6.742130e-01
## salary         1.637917   2.387248   0.6861110 4.961632e-01
```

[1]For more detail see: http://www.clayford.net/statistics/tag/termplot/

```
## ratio         -3.624232   3.215418  -1.1271418 2.656570e-01
## takers        -2.904481   0.231260 -12.5593745 2.606559e-16
```

Faraway 7.4

GNP data! Cool stuff.

Let's first look at condition numbers

Condtional numbers are defined by:

$$\kappa = \sqrt{\frac{\lambda_1}{\lambda_p}}$$

or alternatively we can find a list of condtional numbers by looking at

$$\kappa = \sqrt{\frac{\lambda_1}{\lambda_i}} \text{ where } i = 1, \ldots, p$$

Table 1: Conditional Numbers for the longley Linear Model

Eigenvalues	Condtional Numbers
66652992.883	1.00000
209072.969	17.85504
105355.048	25.15256
18039.760	60.78472
24.557	1647.47771
2.015	5751.21560

These eigenvalues are pretty spread out and the most of the conditional numbers are large. This tells us that there are probably mulitple linear combinations are play here that we should be wary of when trying to reduce collinearity in our model.

Correlation between predictors

We can further analyze the correlation between predictors using the `cor()` function.

```
##              GNP.deflator  GNP Unemployed Armed.Forces Population Year
## GNP.deflator         1.00 0.99       0.62         0.46       0.98 0.99
## GNP                  0.99 1.00       0.60         0.45       0.99 1.00
## Unemployed           0.62 0.60       1.00        -0.18       0.69 0.67
## Armed.Forces         0.46 0.45      -0.18         1.00       0.36 0.42
## Population           0.98 0.99       0.69         0.36       1.00 0.99
## Year                 0.99 1.00       0.67         0.42       0.99 1.00
```

There's really large correlation between **GNP.deflator** and many of the other covariates in the model, especially **GNP**, **Population**, and **Year**. **Year** is correlated very highly with **GNP** as well.

Variance Inflation Factors

We can analyze the variance inflation factors as well. A high VIF for predictor j indicates that that particular predictor may have a linear relationship with the other predictors in the model. We might want to remove it to reduce multicollinearity.

```
## GNP.deflator          GNP   Unemployed Armed.Forces   Population
##    135.53244   1788.51348     33.61889      3.58893    399.15102
##         Year
##    758.98060
```

A rule of thumb from Faraway is that a VIF greater than 10 is a sign that there is multicollinearity present in the model matrix. Most of these predictors have a VIF far greater than 10! We can get rid of some of these redundant predictors and the model may perform better with reduced multicollinearity. Let's see what happens when we remove **GNP.deflator**, **Year**, and **Population**.

```
##          GNP   Unemployed Armed.Forces
##     3.140867     2.596610     2.058847
```

Once we remove these variables, the variance inflation factors of the remaining variables decrease precipitously.

Faraway 7.5

This problem is similar to the last one! Let's get into it

```
## [1] "lcavol"  "lweight" "age"     "lbph"    "svi"     "lcp"     "gleason"
## [8] "pgg45"   "lpsa"
```

Table 2: Conditional Numbers for the prostate Linear Model

Eigenvalues	Condtional Numbers
479082.631	1.00000
61907.037	2.78186
210.904	47.66094
175.633	52.22787
64.799	85.98499
44.524	103.73114
20.239	153.85414
8.093	243.30248

Most of these condition numbers are pretty big. We should investigate this dataset further for possible multicollinearity.

Correlation matrix

```
##         lcavol lweight  age  lbph   svi   lcp gleason pgg45
## lcavol    1.00    0.19 0.22  0.03  0.54  0.68    0.43  0.43
## lweight   0.19    1.00 0.31  0.43  0.11  0.10    0.00  0.05
```

```
## age       0.22    0.31 1.00  0.35  0.12  0.13    0.27  0.28
## lbph      0.03    0.43 0.35  1.00 -0.09 -0.01    0.08  0.08
## svi       0.54    0.11 0.12 -0.09  1.00  0.67    0.32  0.46
## lcp       0.68    0.10 0.13 -0.01  0.67  1.00    0.51  0.63
## gleason   0.43    0.00 0.27  0.08  0.32  0.51    1.00  0.75
## pgg45     0.43    0.05 0.28  0.08  0.46  0.63    0.75  1.00
```

We don't appear to have insane collinearity here, but there correlations that could potentially be problematic: *lcp* and *lcavol*, *svi*, *pgg45* have somewhat strong correlations over .6, and *gleason* and *pgg45* has an even stronger correlations over .7

More fun with VIFs

Let's further examine this using Variance Inflation Factors.

```
##   lcavol  lweight      age     lbph      svi      lcp  gleason    pgg45
## 2.054115 1.363704 1.323599 1.375534 1.956881 3.097954 2.473411 2.974361
```

These VIFs are all actually pretty low though, so there probably isn't as much collinearity here as the correlation matrix and condition numbers would suggest.

Pace of Life Analysis

Introduction

Our analysis will aim to answer the question: *"Does a faster pace of life affect the average heart disease death rate?"* The data was gathered in 1990 and consists of 36 observations from 36 different cities and four variables which proxy for pace of life:

- **Bank**, which corresponds to the average time for bank clerks to make change of two $20 bills
- **Walk**, which corresponds to the average walking speed of pedestrians along a main street in that city
- **Talk**, which corresponds to the average talking speed of postal clerks
- and the response variable **Heart**, which corresponds to age adjusted death rate due to heart disease

It's important to note is that the variables here have been **standardized** so that there are no units of measurement.

Methods

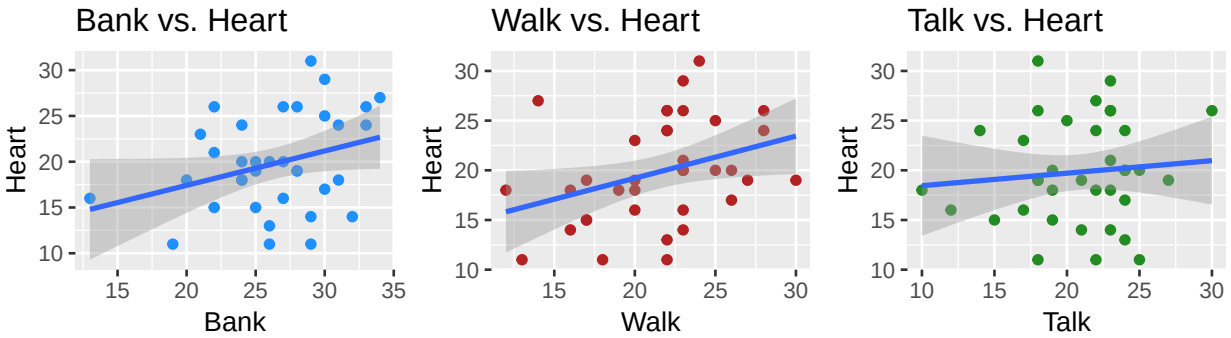
Our principle tool here for finding association with heart disease and pace of life will be inference using a multiple linear regression. In order to make sure that it is appropriate we should check the required assumptions for inference: linearity, constant and normal errors centered around zero. Additionally, we might want to check for outliers and influential points. Assuming assumptions are met, several models will be fit starting starting with a full model with all possible predictors. We will select a final model based on regression diagnostics such as R^2 and use this model to draw our conclusions.

Results

Let's begin with numerical and graphical summaries of the data:

Table 3: Summary Statistics

	min	Q1	median	Q3	max	mean	sd
Heart	11	16.00	19.0	24.00	31	19.80556	5.214373
Bank	13	24.00	26.5	29.25	34	26.36111	4.408775
Talk	10	18.00	22.0	23.25	30	20.75000	4.129165
Walk	12	18.75	22.0	23.25	30	21.41667	4.285357



The plots of the covariates against the response indicate that **Bank** and **Walk** may have a postive linear

relationship with the response. **Talk** against the response appears to have little to no linear relationship.

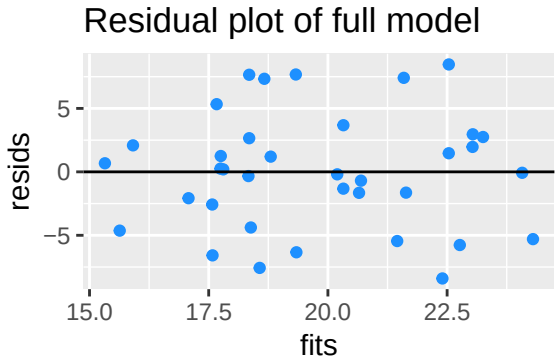
Since we have some evidence that there might be a linear relationship here, we can continue with fitting a linear regression model. To begin, let's fit a full model with all the possible covariates.

Table 4: Coefficients of the Full Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.1786957	6.3369459	0.5016132	0.6193734
Bank	0.4052170	0.1971021	2.0558738	0.0480291
Walk	0.4516011	0.2008735	2.2481862	0.0315839
Talk	-0.1796096	0.2222154	-0.8082681	0.4249046

Bank and **Walk** are significant predictors according to the individual t-tests while the **Intercept** and **Talk** are not. Our F-statistic from this model is which has a p-value of .041 indicating that this model at least performs better than just using the mean. However, the R^2 is quite low: 0.22.

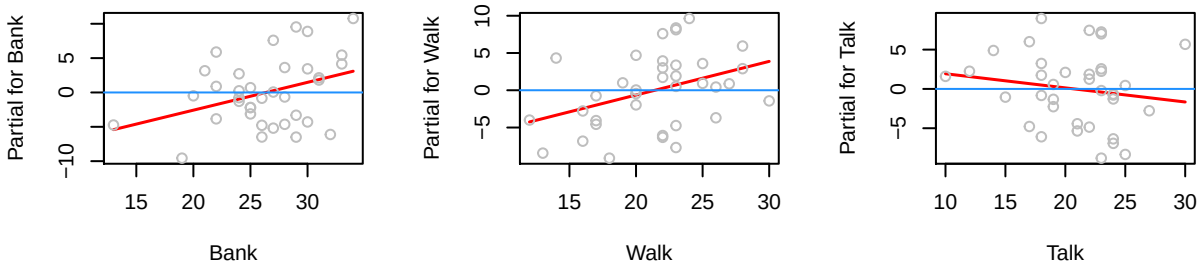
This may be an effect of the data being standardized, but we should check here that the regression assumptions are all being met. We have already examined the linear relationships with the predictors and the response, so now we should check the constant spread and normality of errors as well as try to find points of influence.



We find that these residuals look relatively well behaved: constant spread and centered around 0. Furthermore, the Shapiro-Wilk test reports a p-value of 0.313 so we have no evidence to reject the normality of errors.

Now we can check for potential outliers. Upon checking the studentized residuals for this model, we find that the largest one is -3.12 which is lower than our critical t-value for this model 3.5 so we don't have any significant evidence for outliers.

While the model's R^2 values are unlikely to be improved by removing variables, it might be of interest to us to see if we can drop **Talk** from the model. This will simplify the model, making it easier for us to gather data later on. Also, if **Talk** is collinear with another predictor, we still have a shot at improving the model. We can explore the structure of the relationships in the data by examining the partial residual plots to visualize the effect of a single predictor on the response.



The blue lines on these partial plots represent what the red line would look like if the variable in question had no effect on the response. Its red line for **Talk** is closer to the blue line than the other two variables, indicating it may not have any practical effect on the response. It may be better to drop it from the model.

We can formally examine this using an F-test between our full model and a reduced model. Doing so gives us p-value of .425 which indicates we have no significant evidence to say that the the full model is better than the reduced model. Furthermore, the R^2 values between the two models are quite similar, verifying our choice of going with the reduced model.

The coefficents for our final selected model are:

Table 5: Coefficients of the Reduced Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.0518854	6.1491366	0.3336867	0.7407271
Bank	0.3494715	0.1836691	1.9027241	0.0658322
Walk	0.3988115	0.1889587	2.1105744	0.0424718

with estimated σ 4.78, 3, 33, 3 degrees of freedom, and an F-statistic of 4.33 which has a pvalue of .021.

Conclusions

Since the data is standardized and have no units, we cannot make a comment on the original scales of the variables nor can we make a comment on if pace of life *causes* an increase in heart rate. However, our model does allow us to say that bank clerk rate and walking speed rate are associated with heart disease rate. Since the coefficients are positive, the effect of each is that as they increase, so does heart disease rate.

While our final model was statistically significant, we should be cautious in using it for any kind of prediction or interpolation since the R^2 is very low. From the analysis here, this low predictive power might just be due to the fact that while the predictors have some kind of statistically significant association with heart disease, they *are not* strongly associated and therefore not good predictors. Heart disease is a very complex response that has many contributing factors, which we fail to account for here when just examining pace of life.

In future studies, we would probably want to have more data available. We only had three possible predictors here, but pace of life can probably be measured in many different ways. Addtionally, if the goal of future studies is prediction, then additionally we want to include data probably of a clinical nature - average weight, cholesterol, exercise level, etc.