Ugur Eren Canakci
Linear Statistical Models
Worksheet 1
Answers

Dataset “redlining.csv” was imported to RStudio, onto the variable ‘redlining’. The minimum significance level that is considered as valid is 0.01 since the topic in hand is very unforgiving for false conclusions.

Question 1:

The code that is executed for this fitting is given as:

```
> md_redlining_rate_race = lm (rate ~ race, redlining)
> summary(md_redlining_rate_race)
```

The output of this code is given as:

```
Call:
lm(formula = rate ~ race, data = redlining)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7496 -0.2479 -0.1487  0.3129  1.1724 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.129218   0.096611   1.338    0.188    
race        0.013882   0.002031   6.836 1.78e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4488 on 45 degrees of freedom
Multiple R-squared:  0.5094,	Adjusted R-squared:  0.4985 
F-statistic: 46.73 on 1 and 45 DF,  p-value: 1.784e-08
```

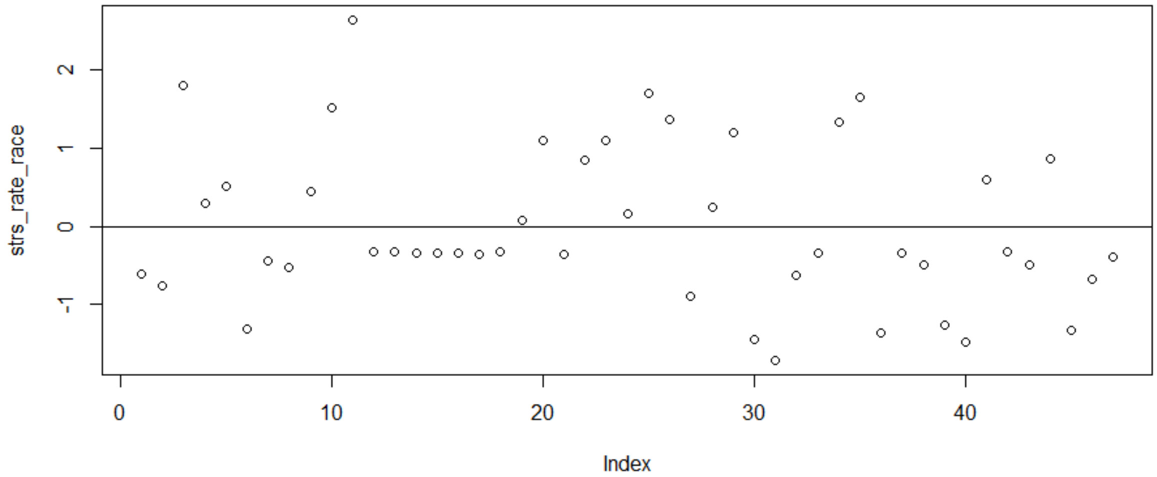
Significance level: 0.01
Hypotheses:
H0: Coefficient of race as a regressor is equal to 0.
Ha: Coefficient of race as a regressor is greater than 0.

The p-value of race having an estimate coefficient of 0.0139 is 1.78*e^{-8}, which is way smaller than the significance level 0.001 for right-tail test, signified by three *** asterisks at the right of the p-value of race. Hence, we have enough evidence to reject H0, i.e rates of minority have significant effect on rate of decline of insurance.

Now we look at the normal probability plot of the standardized residuals:

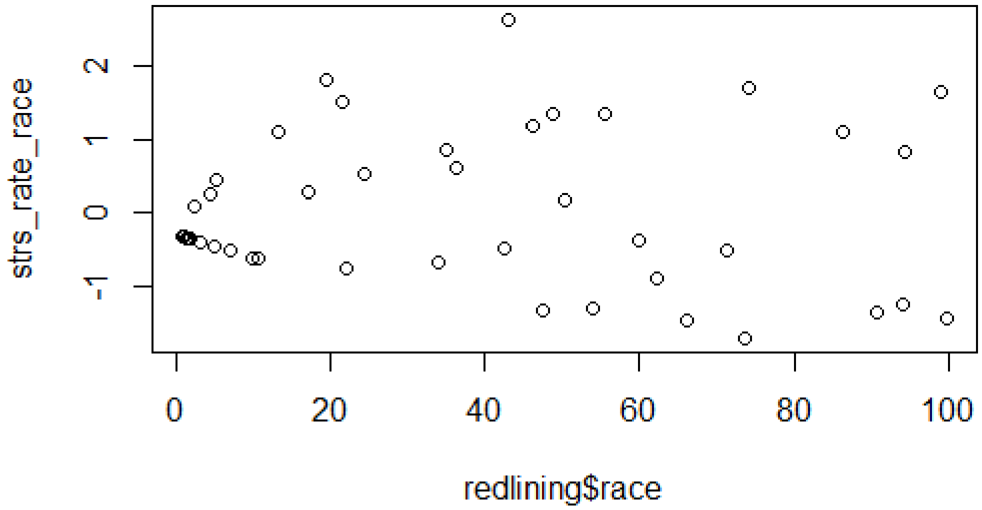
```
> strs_rate_race = rstandard(md_redlining_rate_race)
> plot(strs_rate_race)
> abline(h=0)
```

Plot result:



The test statistics of most of the residual errors fall between [-2,2] as seen in the plot above.

```
> plot(redlining$race, strs_rate_race)
```



The predictive level is pretty low, given as:

```
Multiple R-squared:  0.5094,	Adjusted R-squared:  0.4985
```

```
> fitted_y_rate_race = predict(md_redlining_rate_race)
> print(redlining$rate - fitted_y_rate_race)
          1           2           3           4           5           6 
7           8           9          10 
-0.26804156 -0.33740626  0.79868786  0.13061727  0.23066433 -0.57886508 
-0.19724156 -0.22778273  0.19720550  0.67231139 
         11          12          13          14          15          16 
17          18          19          20 
 1.17245257 -0.14448862 -0.14310038 -0.15281803 -0.15142979 -0.15004156 
-0.15420627 -0.14310038  0.03607609  0.48475844 
         21          22          23          24          25          26 
27          28          29          30 
-0.15938272  0.36028787  0.47412317  0.07388786  0.74071140  0.60031139 
-0.39408860  0.10969962  0.52941727 -0.61328860 
         31          32          33          34          35          36 
37          38          39          40 
-0.74957096 -0.27775920 -0.15004156  0.59332316  0.69781729 -0.58695919 
-0.14865332 -0.21764154 -0.53554742 -0.64684154 
         41          42          43          44          45          46 
47 
 0.26546433 -0.14310038 -0.21921802  0.38351139 -0.58724155 -0.30121802 
-0.17225332 
```

Above is the difference between data points and fitted values for the SLR we did. The difference is not negligible overall, which is shown by multiple R^2 and adjusted R^2 statistics given before. Hence, one needs to do more investigation on this matter, by maybe doing MLR.

Question 2:

The question asserts the assumption that there is a linear relation between race and fire and between race and theft.

```
> md_theft_race = lm(theft ~ race, redlining)
> summary(md_theft_race)

Call:
lm(formula = theft ~ race, data = redlining)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.528  -9.511  -4.025   3.563 111.984 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  26.2578     4.6907   5.598 1.23e-06 ***
race          0.1745     0.0986   1.770   0.0836 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 21.79 on 45 degrees of freedom
Multiple R-squared:  0.06506,	Adjusted R-squared:  0.04428 
F-statistic: 3.131 on 1 and 45 DF,  p-value: 0.08358
```

There isn’t conclusive information to reject that “race and theft statistics have no correlation” because the p-value of race is larger than 0.01.

```
> md_fire_race = lm(fire ~ race, redlining)
> summary(md_fire_race)

Call:
lm(formula = fire ~ race, data = redlining)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.7819  -4.3364  -1.6295   0.6215  24.8467 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.35869    1.63038   3.900 0.000318 ***
race         0.16922    0.03427   4.938 1.13e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.574 on 45 degrees of freedom
Multiple R-squared:  0.3514,	Adjusted R-squared:  0.337 
F-statistic: 24.38 on 1 and 45 DF,  p-value: 1.132e-05
```

In this case, p-value of race is smaller than 0.001, meaning that we can refute the hypothesis that race and theft have no correlation. Yet, since R^2 scores are not close to 1, there could be many other factors for the number of fires being very high, piled around the intercept estimate (which is concluded by the p-value of the intercept being smaller than 0.001).

One could argue not taking the fire statistic into the investigation of refusal rates since the proportion of minorities is already a good regressor for rate of insurance refusal. For this case, we will do two MLR regressions on our dataset ‘redlining’, one relating redlining$fire to redlining$rate and one not.

```
> md_redlining = lm(rate ~ race + fire + theft + income + age, redlining)
> summary(md_redlining)

Call:
lm(formula = rate ~ race + fire + theft + income + age, data = redlining)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.84428 -0.15804 -0.04093  0.18116  0.80828 
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.608979   0.495260  -1.230 0.225851
race         0.009133   0.002316   3.944 0.000307 ***
fire         0.038817   0.008436   4.602    4e-05 ***
theft       -0.010298   0.002853  -3.610 0.000827 ***
income       0.024500   0.031697   0.773 0.443982
age          0.008271   0.002782   2.973 0.004914 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3351 on 41 degrees of freedom
Multiple R-squared:  0.7508,	Adjusted R-squared:  0.7204
F-statistic: 24.71 on 5 and 41 DF,  p-value: 2.159e-11

> md_redlining_no_fire = lm(rate ~ race + theft + income + age, redlining)
> summary(md_redlining_no_fire)

Call:
lm(formula = rate ~ race + theft + income + age, data = redlining)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81890 -0.24795 -0.03815  0.29804  0.93772

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.053065   0.576592   0.092 0.927111
race         0.011530   0.002745   4.200 0.000136 ***
theft       -0.003353   0.002946  -1.138 0.261431
income      -0.024043   0.036366  -0.661 0.512130
age          0.008688   0.003382   2.569 0.013855 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4077 on 42 degrees of freedom
Multiple R-squared:  0.6221,	Adjusted R-squared:  0.5861
F-statistic: 17.29 on 4 and 42 DF,  p-value: 1.871e-08
```

Just in case, we look at “no theft” case as well:

```
> md_redlining_no_theft = lm(rate ~ race + fire + income + age, redlining)
> summary(md_redlining_no_theft)

Call:
lm(formula = rate ~ race + fire + income + age, data = redlining)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95452 -0.18195  0.01301  0.20936  0.85078

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -0.170457   0.544560  -0.313  0.75582   
race         0.008330   0.002614   3.186  0.00272 **
fire         0.022709   0.008119   2.797  0.00775 **
income      -0.012084   0.034063  -0.355  0.72455   
age          0.005707   0.003050   1.871  0.06830 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3801 on 42 degrees of freedom
Multiple R-squared:  0.6716,	Adjusted R-squared:  0.6404 
F-statistic: 21.48 on 4 and 42 DF,  p-value: 1.053e-09
```

Finally, we do an anova comparison between the models we fitted:

```
> anova(md_redlining_no_fire, md_redlining_no_theft, md_redlining)
Analysis of Variance Table

Model 1: rate ~ race + theft + income + age
Model 2: rate ~ race + fire + income + age
Model 3: rate ~ race + fire + theft + income + age
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     42 6.9828                                  
2     42 6.0680  0   0.91479                     
3     41 4.6047  1   1.46328 13.029 0.0008269 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As seen above, all data considered at the same time gives the best result. Hence, the whole set of variables is taken as the variables affecting the residual rates.

Question 3:

```
> anova(md_redlining_rate_race, md_redlining)
Analysis of Variance Table

Model 1: rate ~ race
Model 2: rate ~ race + fire + theft + income + age
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     45 9.0653                                  
2     41 4.6047  4    4.4606 9.9291 1.033e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Anova test clearly denounces the MLR “md_redlining” over the SLR “md_redlining_rate_race” as the better model fitting. p-value for this test is 1.033*10^{-5}.

Question 4:

Means are prepared:

```
> race_mean = mean(redlining$race)
> fire_mean = mean(redlining$fire)
> theft_mean = mean(redlining$theft)
> income_mean = mean(redlining$income)
> age_mean = mean(redlining$age)
```

Difference between means are prepared:

```
> err_race = redlining$race - race_mean
> err_fire = redlining$fire - fire_mean
> err_theft = redlining$theft - theft_mean
> err_income = redlining$income - income_mean
> err_age = redlining$age - age_mean
```

Polynomial regression models are prepared:

```
> md_sq_race =lm(rate ~ race + I(err_race^2) + fire + theft + income + age,
redlining)
> md_sq_fire =lm(rate ~ race + I(err_fire^2) + fire + theft + income + age,
redlining)
> md_sq_theft =lm(rate ~ race + I(err_theft^2) + fire + theft + income + age,
redlining)
> md_sq_income =lm(rate ~ race + I(err_income^2) + fire + theft + income +
age, redlining)
> md_sq_age =lm(rate ~ race + I(err_age^2) + fire + theft + income + age,
redlining)
```

Anova test between the original model and these new models:

```
> anova(md_redlining, md_sq_race, md_sq_fire, md_sq_theft, md_sq_income,
md_sq_age)
Analysis of Variance Table

Model 1: rate ~ race + fire + theft + income + age
Model 2: rate ~ race + I(err_race^2) + fire + theft + income + age
Model 3: rate ~ race + I(err_fire^2) + fire + theft + income + age
Model 4: rate ~ race + I(err_theft^2) + fire + theft + income + age
Model 5: rate ~ race + I(err_income^2) + fire + theft + income + age
Model 6: rate ~ race + I(err_age^2) + fire + theft + income + age
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     41 4.6047
2     40 4.6047  1  0.000014 1e-04 0.9913
3     40 4.3217  0  0.282955
4     40 4.5898  0 -0.268077
5     40 4.6038  0 -0.013992
6     40 4.4082  0  0.195625
```

RSS statistics are quite similar to each other, implying no significant change between the first model and the others. No need to include quadratic terms into our model.

Question 5:

```
> summary(md_redlining)

Call:
lm(formula = rate ~ race + fire + theft + income + age, data = redlining)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.84428 -0.15804 -0.04093  0.18116  0.80828 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.608979   0.495260  -1.230 0.225851    
race         0.009133   0.002316   3.944 0.000307 ***
fire         0.038817   0.008436   4.602    4e-05 ***
theft       -0.010298   0.002853  -3.610 0.000827 ***
income       0.024500   0.031697   0.773 0.443982    
age          0.008271   0.002782   2.973 0.004914 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3351 on 41 degrees of freedom
Multiple R-squared:  0.7508,	Adjusted R-squared:  0.7204 
F-statistic: 24.71 on 5 and 41 DF,  p-value: 2.159e-11
```

Our model tells us that we can represent rate of insurance rejection as

rate = -0.609 + 0.0091*race + 0.0388*fire - 0.0102*theft + 0.0245*income + 0.0083*age

with residual standard error 0.3351.

- We cannot reject the claims that the intercept and the coefficient of income are 0 since their p-values are greater than the significance level 0.0005; there is no significant evidence for intercept and the coefficient of income being nonzero.
- We can reject the claim that the coefficient of race is 0 since its p-value is less than the significance level 0.001.
- We can reject the claims that the coefficients of fire and theft are 0 since their p-value are less than the significance level 0.0005.
- We can reject the claim that the coefficient of age is 0 since its p-value is less than the significance level 0.0005.
- We can also reject the claim that all coefficients including the intercept are 0 because our MLR model has a p-value of 0.2159*10^{-11} < 0.0005

Question 6:

The evidence suggests the existence of racial profiling, based on the rate of minority residence in areas; the coefficient is around 0.0091 with mean standard error of 0.0023 and p-value 0.0003, significantly smaller than 0.001 which is chosen in sight of right-tail test. Yet, the suggested coefficient is very ineffective: assuming all other factors stay the same, a 100 percent increase on the race statistic implies a 0.91 percent increase in insurance rejection rate. The case where the biggest nominal effect would happen in the rejection rate by the

0.91 percent increase is the case that there has been no rejection yet. So, the rejection rate goes from 0 to 0.91 percent. For the racial profiling to be recognized by people, we would expect a sample of houses in a neighborhood of size n where n*0.0091>=1 at the very least and probably many many more for people to catch a trend. We infer that n has to be at least 110. Even by assuming the worst case scenario, which is statistically impossible unless a racially motivated displacement action is at play, we cannot find or recreate enough trials for the racial profiling to be recognized by people in this situation. This is the reason we need statistical analysis in this case; it's impossible to extrapolate the idea that racial profiling exists on insurance declines by purely looking at the numbers but doing linear regression with many factors under inquisition gives us a clear picture on the case.