Homework 02

Sky Liu

Septemeber 25, 2018

Introduction

In homework 2 you will fit many regression models. You are welcome to explore beyond what the question is asking you.

Please come see us we are here to help.

Data analysis

Analysis of earnings and height data

The folder `earnings` has data from the Work, Family, and Well-Being Survey (Ross, 1990). You can find the codebook at http://www.stat.columbia.edu/~gelman/arm/examples/earnings/wfwcodebook.txt

```
gelman_dir <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
heights    <- read.dta (paste0(gelman_dir,"earnings/heights.dta"))
```

Pull out the data on earnings, sex, height, and yearbn

1. In R, check the dataset and clean any unusually coded data.

```
#pull out earning, sex, height, and yearbn
refine_heights <- heights[,c(1,4,8,9)]
#Clean the rows with NAs
clean_heights <- data.frame()
for (i in 1:2029){
  if ((is.na(heights[i,1]) == FALSE) & (heights[i,1] != 0)){
    clean_heights <- rbind(clean_heights,refine_heights[i,])
  }
}
age <- 118 - clean_heights$yearbn
male <- 2 - clean_heights$sex #male coded as 1, female coded as 0
clean_heights <- cbind(clean_heights,age,male)
```

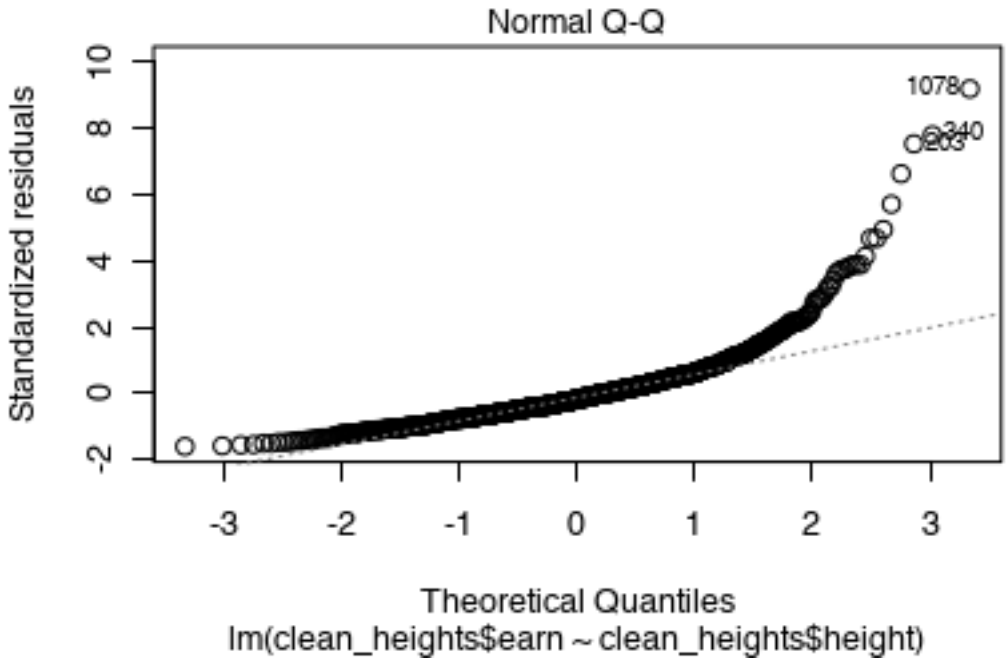
2. Fit a linear regression model predicting earnings from height. What transformation should you perform in order to interpret the intercept from this model

as average earnings for people with average height?

```
reg <- lm(clean_heights$earn ~ clean_heights$height)
summary(reg)
```

```
##
## Call:
## lm(formula = clean_heights$earn ~ clean_heights$height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30166 -11309  -3428   6527 172953
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -61316.3     9525.2  -6.437 1.76e-10 ***
## clean_heights$height   1262.3      142.1   8.883  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18870 on 1190 degrees of freedom
## Multiple R-squared:  0.06218,    Adjusted R-squared:  0.06139
## F-statistic:  78.9 on 1 and 1190 DF,  p-value: < 2.2e-16
```

```
plot(reg,which=2)
```



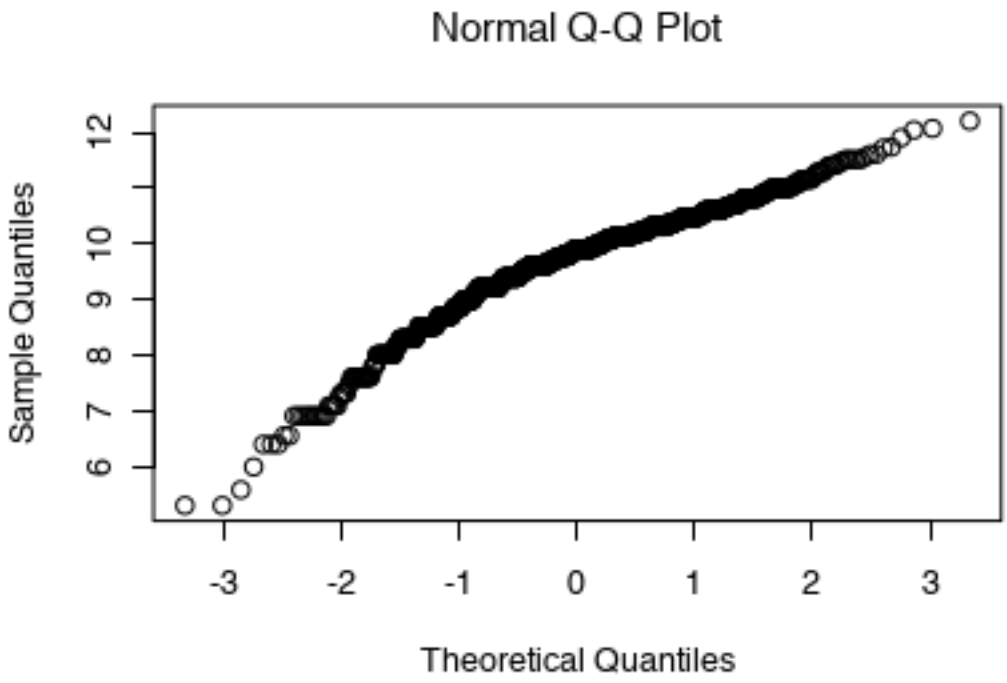
If we do a regular linear model on earning and heights we can see from the qq plot that the line is skewed. Thus, here we want to do a log transfermation and center the height by average, which is 66.92

```
centered_height <- clean_heights$height-mean(clean_heights$height)
centered_age <- clean_heights$age-mean(clean_heights$age)
reg_log <- lm(log(clean_heights$earn) ~ centered_height)
summary(reg_log)
```

```
##
## Call:
```

```
## lm(formula = log(clean_heights$earn) ~ centered_height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4209 -0.3975  0.1394  0.5833  2.3536
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.714349   0.025867 375.544   <2e-16 ***
## centered_height 0.058817   0.006728   8.743   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8931 on 1190 degrees of freedom
## Multiple R-squared:  0.06035,    Adjusted R-squared:  0.05957
## F-statistic: 76.44 on 1 and 1190 DF,  p-value: < 2.2e-16
```

```
qqnorm(log(clean_heights$earn))
```



Now, from the qqplot we can see that log(clean_heights$earn) is basically normally distributed.

Based on the model summary presented above, we obtain that the model is: $log(earning) = 9.71 + 0.06 * centeredheight$

That is:

$earning = e^{9.71} * e0.06 * height = 16481 * 1.06^{centeredheight}$

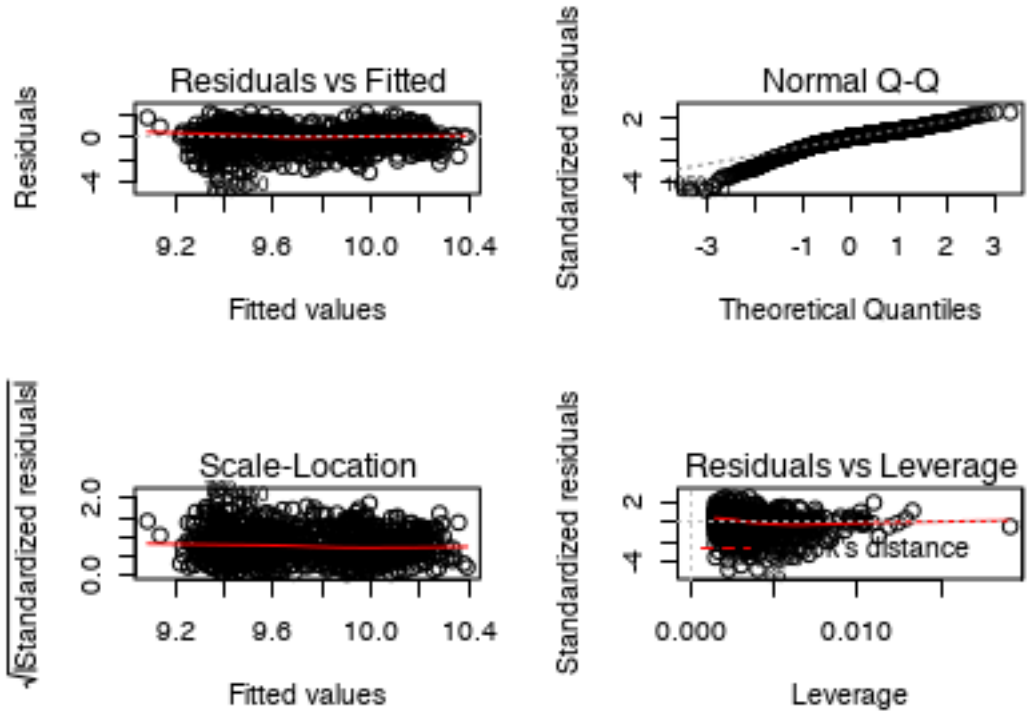
From this model, we could interpret that the average earning for a person with average height is 16481. The average earning will increase 6% if the height of this person increases by one unit.

3. Fit some regression models with the goal of predicting earnings from some combination of sex, height, and age. Be sure to try various transformations and interactions that might make sense. Choose your preferred model and justify.

```
reg_log3 <- lm(log(clean_heights$earn) ~ centered_height + clean_heights$male + centered_age)
summary(reg_log3)
```

```
##
## Call:
## lm(formula = log(clean_heights$earn) ~ centered_height + clean_heights$male +
##     centered_age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1876 -0.3812  0.1786  0.5648  2.2267
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         9.541217   0.039641 240.692  < 2e-16 ***
## centered_height     0.025896   0.009319   2.779  0.00554 **
## clean_heights$male 0.408660   0.071990   5.677 1.73e-08 ***
## centered_age        0.007069   0.001612   4.384 1.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8742 on 1188 degrees of freedom
## Multiple R-squared:  0.1011, Adjusted R-squared:  0.09883
## F-statistic: 44.54 on 3 and 1188 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(reg_log3)
```



$log(earning) = 9.54 + 0.025 * height + 0.409 * male + 0.007 * age$

$earning = e^{9.54} * e^{0.025 * height} + e^{0.409 * male} + e^{0.007 * age}$

$earning = 13904.948 * 1.025^{height} * 1.505^{male} * 1.007^{age}$

From the summary, we can see that this model explains 87% of data. From the residual plot we can see the variance is pretty constant. However, the normality is questionable.

4. Interpret all model coefficients.

The average earning for a female with average height and age is 13904.948.

The average earning for a male will be 50.5% higher than a female with the same age and height.

The average earning will be 2.5% higher if the height of the person is incremented by one unit, holding other variables constant.

The average earning will be 0.7% higher if the person is one year older, holding other variables constant.

5. Construct 95% confidence interval for all model coefficients and discuss what they mean.

```
confint(reg_log3, level=0.95)
```

```
##                          2.5 %     97.5 %
## (Intercept)        9.463443782 9.61899092
## centered_height    0.007612598 0.04417875
```

```
## clean_heights$male 0.267418209 0.54990232
## centered_age       0.003905583 0.01023275
```

The intercept falls in [9.46,9.61] with 95% of possibility.

The coefficient of height falls in [0.008, 0.044] with 95% of possibility.

The coefficient of male falls in [0.267, 0.550] with 95% of possibility.

The coefficient of age falls in [0.004, 0.010] with 95% of possibility.

Analysis of mortality rates and various environmental factors

The folder `pollution` contains mortality rates and various environmental factors from 60 U.S. metropolitan areas from McDonald, G.C. and Schwing, R.C. (1973) 'Instabilities of regression estimates relating air pollution to mortality', Technometrics, vol.15, 463-482.

Variables, in order:

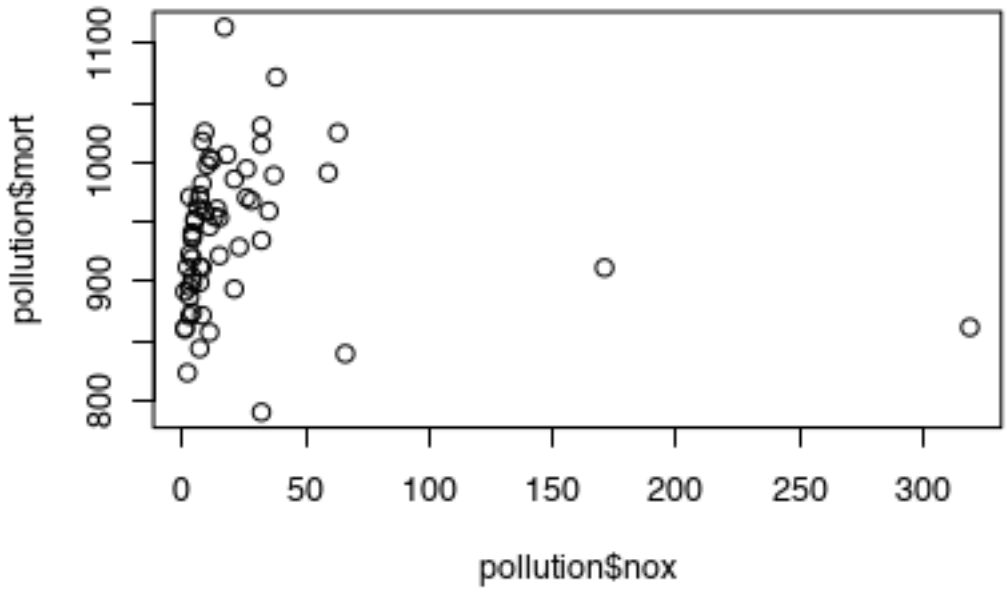
- PREC Average annual precipitation in inches
- JANT Average January temperature in degrees F
- JULT Same for July
- OVR65 % of 1960 SMSA population aged 65 or older
- POPN Average household size
- EDUC Median school years completed by those over 22
- HOUS % of housing units which are sound & with all facilities
- DENS Population per sq. mile in urbanized areas, 1960
- NONW % non-white population in urbanized areas, 1960
- WWDRK % employed in white collar occupations
- POOR % of families with income < $3000
- HC Relative hydrocarbon pollution potential
- NOX Same for nitric oxides
- SO@ Same for sulphur dioxide
- HUMID Annual average % relative humidity at 1pm
- MORT Total age-adjusted mortality rate per 100,000

For this exercise we shall model mortality rate given nitric oxides, sulfur dioxide, and hydrocarbons as inputs. This model is an extreme oversimplification as it combines all sources of mortality and does not adjust for crucial factors such as age and smoking. We use it to illustrate log transformations in regression.

```
gelman_dir   <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
pollution    <- read.dta (paste0(gelman_dir,"pollution/pollution.dta"))
```

1. Create a scatterplot of mortality rate versus level of nitric oxides. Do you think linear regression will fit these data well? Fit the regression and evaluate a residual plot from the regression.

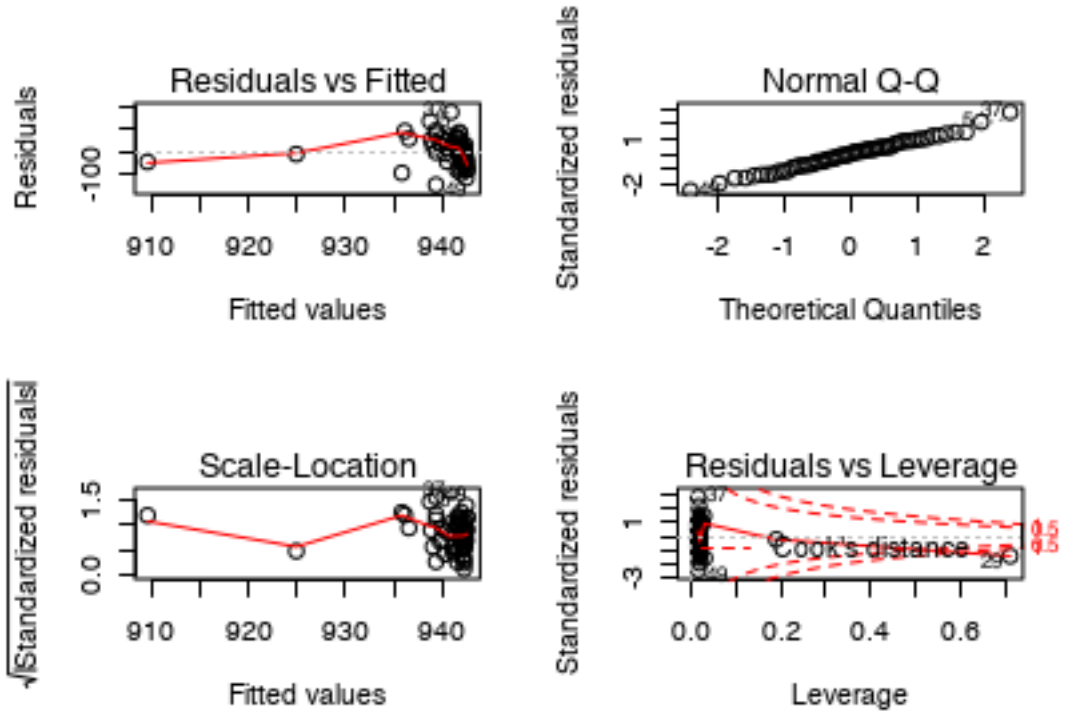
```
plot(pollution$nox,pollution$mort)
```




```
reg_pol1 <- lm(pollution$mort~pollution$nox)
summary(reg_pol1)
```

```
##
## Call:
## lm(formula = pollution$mort ~ pollution$nox)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -148.654  -43.710    1.751   41.663  172.211
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   942.7115     9.0034 104.706   <2e-16 ***
## pollution$nox  -0.1039     0.1758  -0.591    0.557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.55 on 58 degrees of freedom
## Multiple R-squared:  0.005987,   Adjusted R-squared:  -0.01115
## F-statistic: 0.3494 on 1 and 58 DF,  p-value: 0.5568
```

```
par(mfrow=c(2,2))
plot(reg_pol1)
```



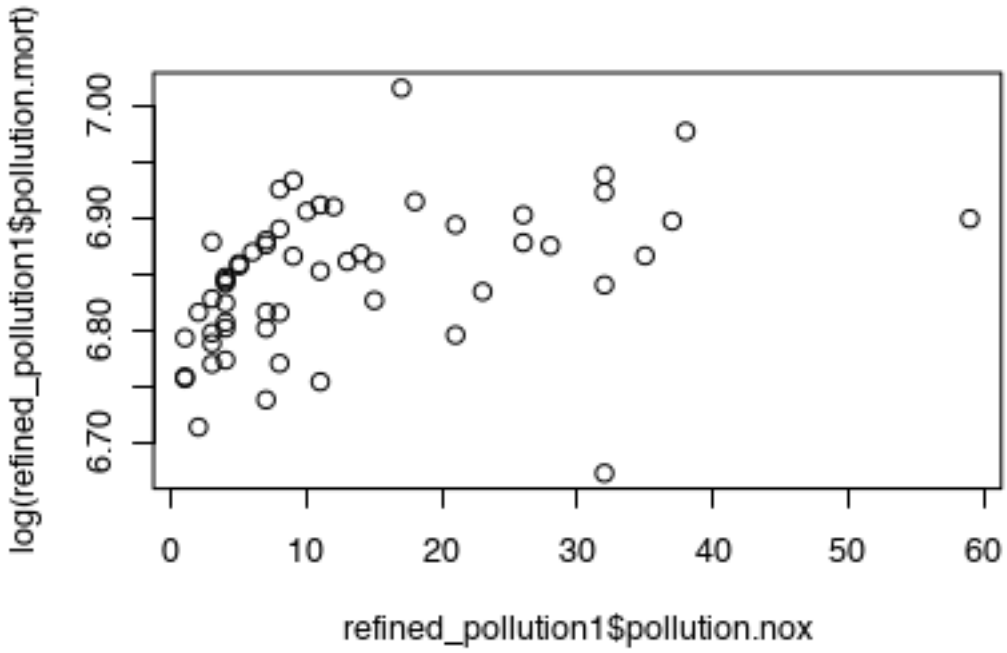
Based on the regression model summary ($R^2 = 0.6\%$) and residual plot, this fit is not ideal at all.

2. Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

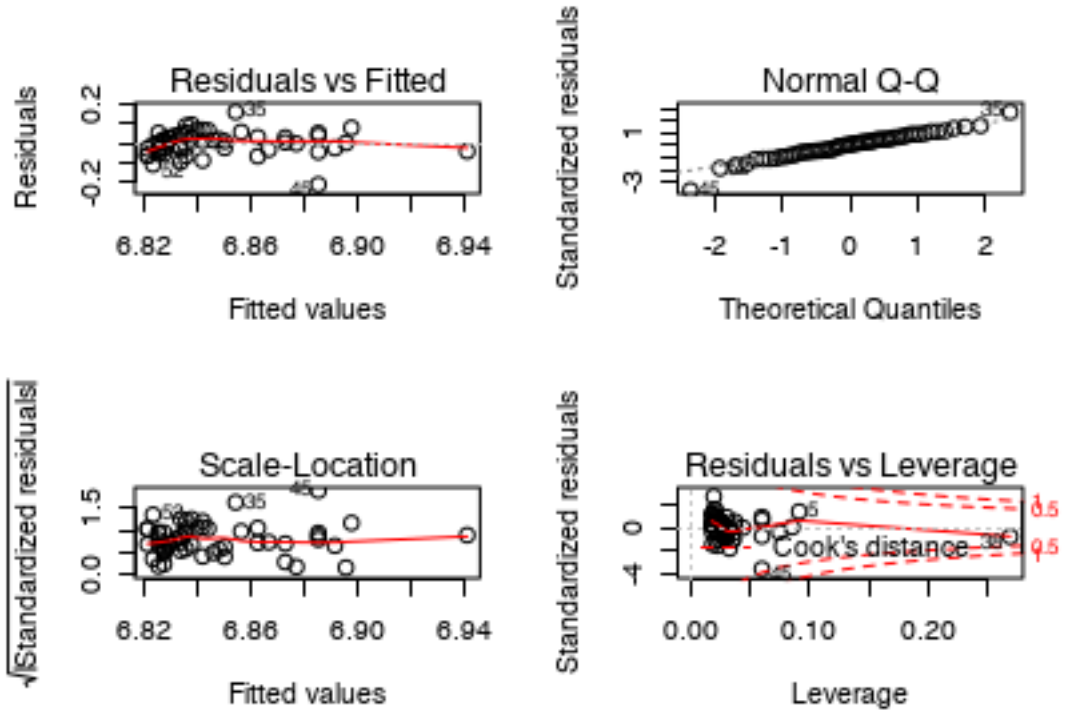
```
refined_pollution <- data.frame(pollution$nox,pollution$mort)


outlierlist <- which(pollution$nox > 60)
refined_pollution1 <- rbind(refined_pollution[1:11, ],refined_pollution[13:28, ],refined_pollution[30:4
```

```
reg_pol2 <- lm(log(refined_pollution1$pollution.mort)~refined_pollution1$pollution.nox)
plot(refined_pollution1$pollution.nox,log(refined_pollution1$pollution.mort))
```




```
par(mfrow=c(2,2))
plot(reg_pol2)
```



```r
summary(reg_pol2)
```

```
##
## Call:
## lm(formula = log(refined_pollution1$pollution.mort) ~ refined_pollution1$pollution.nox)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.212183 -0.032245  0.003643  0.037557  0.160631
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     6.8193938  0.0118574 575.116  < 2e-16 ***
## refined_pollution1$pollution.nox 0.0020547  0.0006593   3.116  0.00293 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06019 on 54 degrees of freedom
## Multiple R-squared:  0.1524, Adjusted R-squared:  0.1367
## F-statistic: 9.713 on 1 and 54 DF,  p-value: 0.00293
```

This model2 looks much better than the last one. The residuals are normally distributed with equavarience. The R^2 also increases from 0.6% to 15%.

3. Interpret the slope coefficient from the model you chose in 2.

$mort = e^{6.82} + e^{0.002*nox} = 916 * 1.002^{nox}$

The average total age-adjusted mortality rate per 100,000 is 916 if the relative nitric oxides pollution potential is 0.

The average total age-adjusted mortality rate per 100,000 will increase by 0.2% if the relative nitric oxides pollution potential is incremented by 1 unit.

4. Construct 99% confidence interval for slope coefficient from the model you chose in 2 and interpret them.

```r
confint(reg_pol2,level=0.99)
```

```
##                                       0.5 %      99.5 %
## (Intercept)                     6.7877346383 6.851052917
## refined_pollution1$pollution.nox 0.0002943819 0.003814976
```

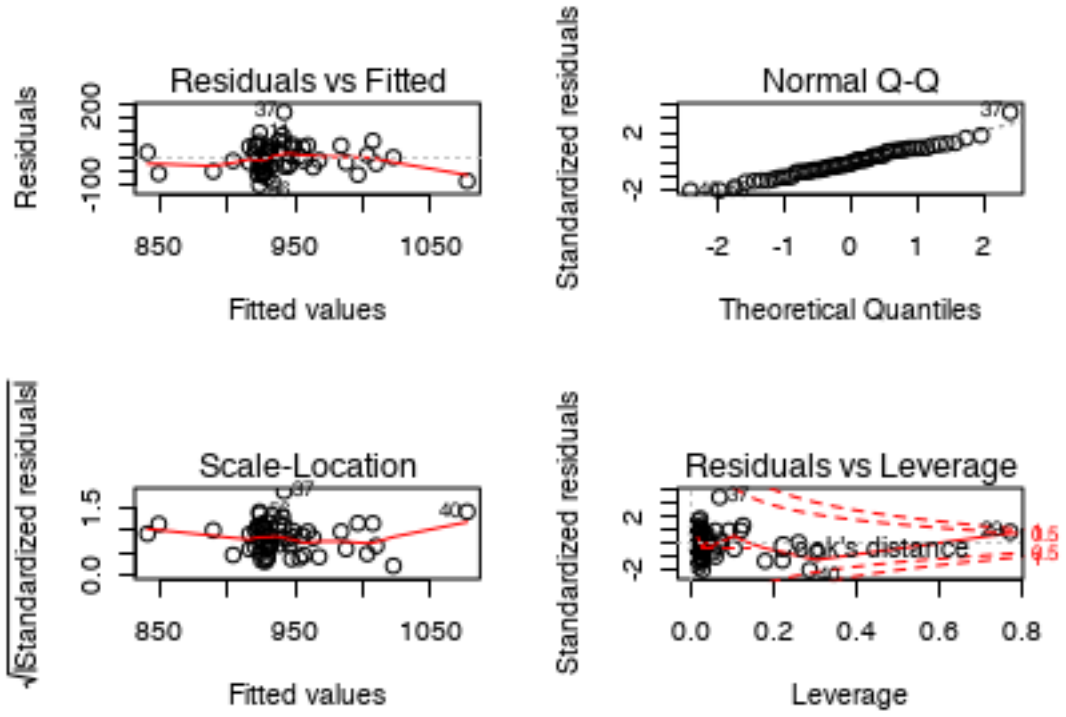
The true coefficient for nox falls in [0.0003, 0.0038] with 99% of possibility.

5. Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.

```
reg_pol3 <- lm(pollution$mort~pollution$nox+pollution$so2+pollution$hc)
summary(reg_pol3)
```

```
##
## Call:
## lm(formula = pollution$mort ~ pollution$nox + pollution$so2 +
##     pollution$hc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.020  -33.058   -5.287   38.398  171.163
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   924.1670     8.9731 102.993   <2e-16 ***
## pollution$nox   2.9350     1.2668   2.317   0.0242 *
## pollution$so2   0.2006     0.1728   1.161   0.2507
## pollution$hc   -1.6135     0.6069  -2.659   0.0102 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.84 on 56 degrees of freedom
## Multiple R-squared:  0.3407, Adjusted R-squared:  0.3054
## F-statistic: 9.647 on 3 and 56 DF,  p-value: 3.131e-05
```

```
par(mfrow=c(2,2))
plot(reg_pol3)
```



The model is $mort = 924 + 2.9 * nox + 0.2 * so2 - 1.6 * hc$

The average total age-adjusted mortality rate per 100,000 will be 924, if the relative pollution potential of nox, so2 and hc are all 0.

The average total age-adjusted mortality rate per 100,000 will be increased by 2.9, if the relative pollution potential of nox increases by 1 unit, holding other variables constant.

The average total age-adjusted mortality rate per 100,000 will be increased by 0.2, if the relative pollution potential of so2 increases by 1 unit, holding other variables constant.

The average total age-adjusted mortality rate per 100,000 will be decrease by 1.6, if the relative pollution potential of hc increases by 1 unit, holding other variables constant.

6. Cross-validate: fit the model you chose above to the first half of the data and then predict for the second half. (You used all the data to construct the model in 4, so this is not really cross-validation, but it gives a sense of how the steps of cross-validation can be implemented.)

```
pollution_half1 <- pollution[1:30,]
pollution_half2 <- pollution[31:60,]
reg_pol4 <- lm(pollution_half1$mort~pollution_half1$nox+pollution_half1$so2+pollution_half1$hc)
summary(reg_pol4)
```

```
##
## Call:
```

```
## lm(formula = pollution_half1$mort ~ pollution_half1$nox + pollution_half1$so2 +
##     pollution_half1$hc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.618 -28.975  -5.018  33.383  84.599
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         925.6763    11.1996  82.653   <2e-16 ***
## pollution_half1$nox   0.9431     2.1741   0.434   0.6680
## pollution_half1$so2   0.4206     0.2372   1.773   0.0879 .
## pollution_half1$hc   -0.6423     1.0320  -0.622   0.5391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.28 on 26 degrees of freedom
## Multiple R-squared:  0.3836, Adjusted R-squared:  0.3124
## F-statistic: 5.393 on 3 and 26 DF,  p-value: 0.005073
```

```
pollution_half2_pred <- as.data.frame(predict(reg_pol4,pollution_half2,interval = 'prediction', level =
which(pollution_half2$mort > pollution_half2_pred$upr)
```

```
## [1] 7
```

```
which(pollution_half2$mort < pollution_half2_pred$lwr)
```

```
## [1] 19 26
```

3 out of 30 data point from the second half of mort data are not in the 95% confident interval conducted from the first half data model. I think the first half data model is not good enough.

Study of teenage gambling in Britain

```
data(teengamb)
?teengamb
```

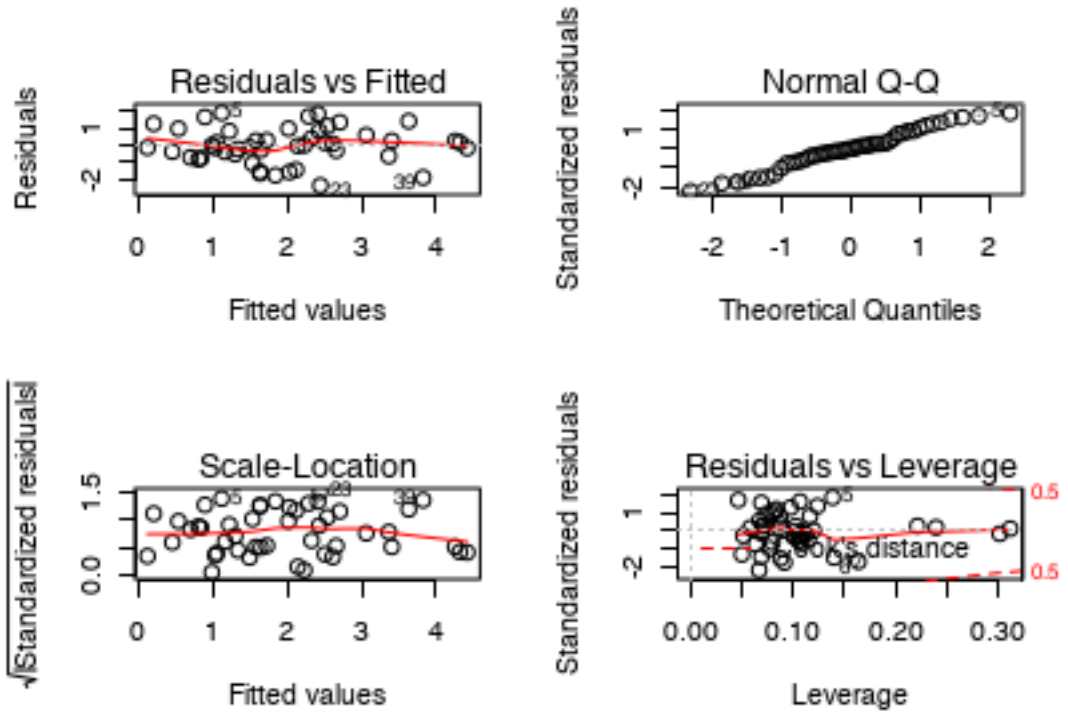
1. Fit a linear regression model with gamble as the response and the other variables as predictors and interpret the coefficients. Make sure you rename and transform the variables to improve the interpretability of your regression model.

```
gambdata <- data.frame(teengamb)
centered_income<-teengamb$income-mean(teengamb$income)
centered_status<-teengamb$status-mean(teengamb$status)
centered_verbal<-teengamb$verbal-mean(teengamb$verbal)
reg_gamb1 <- lm(log(gambdata$gamble+1) ~ factor(gambdata$sex) + centered_status + centered_income + cen
summary(reg_gamb1)
```

```
##
## Call:
```

```
## lm(formula = log(gambdata$gamble + 1) ~ factor(gambdata$sex) +
##     centered_status + centered_income + centered_verbal, data = gambdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.35012 -0.56865  0.00413  0.71512  1.90319
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.32410    0.22417  10.368 3.78e-13 ***
## factor(gambdata$sex)1 -0.87120    0.39268  -2.219   0.0320 *
## centered_status          0.02983    0.01344   2.219   0.0320 *
## centered_income          0.21565    0.04904   4.398 7.33e-05 ***
## centered_verbal         -0.26165    0.10388  -2.519   0.0157 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.085 on 42 degrees of freedom
## Multiple R-squared:  0.5206, Adjusted R-squared:  0.475
## F-statistic:  11.4 on 4 and 42 DF,  p-value: 2.347e-06
```

```
par(mfrow=c(2,2))
plot(reg_gamb1)
```



$$gamble = e^{2.32} * e^{-0.87*sex} * e^{0.03*status} * e^{0.22*income} * e^{-0.26*verbal}$$

$$gamble = 10.18 * 0.42^{*sex} * 1.03^{status} * 1.25^{income} * 0.77^{verbal}$$

The average expenditure on gambling in pounds per year is 10.18 for a male with average socioeconomic status score based on parents' occupation, average weekly income, average verbal score in words out of 12 correctly defined.

The average expenditure on gambling in pounds per year will decrease by 58% for a female than a male with the same socioeconomic status score based on parents' occupation, weekly income, verbal score in words out of 12 correctly defined.

The average expenditure on gambling in pounds per year will increase by 3% if the person's socioeconomic status score based on parents' occupation increases by 1 unit, holding other variables constant.

The average expenditure on gambling in pounds per year will increase by 25% if the person's weekly income increases by 1 unit, holding other variables constant.

The average expenditure on gambling in pounds per year will decrease by 23% if the person's verbal score in words out of 12 correctly defined increases by 1 unit, holding other variables constant.

2. Create a 95% confidence interval for each of the estimated coefficients and discuss how you would interpret this uncertainty.

```
confint(reg_gamb1,level = 0.95)
```

```
##                              2.5 %      97.5 %
## (Intercept)             1.87171038  2.77649881
## factor(gambdata$sex)1 -1.66365707 -0.07873377
## centered_status         0.00269987  0.05696050
## centered_income         0.11668468  0.31460764
## centered_verbal        -0.47128110 -0.05200895
```

```
exp(-.08)
```

```
## [1] 0.9231163
```

The intercept falls in [1.87,2.78] with 95% of possibility.

The coefficient of gender falls in [-1.66, -0.079] with 95% of possibility.

The coefficient of average status falls in [0.003, 0.057] with 95% of possibility.

The coefficient of average income falls in [0.12, 0.31] with 95% of possibility.

The coefficient of average verbal score falls in [-0.47, -0.05] with 95% of possibility.

3. Predict the amount that a male with average status, income and verbal score would gamble along with an appropriate 95% CI. Repeat the prediction for a male with maximal values of status, income and verbal score. Which CI is wider and why is this result expected?

```
male_mean <- as.data.frame(
  cbind(mean(gambdata$status), mean(gambdata$income), mean(gambdata$verbal) ))
male_mean_prdc <- predict(reg_gamb1, male_mean,level = 0.95,interval  = "prediction")
```

```
## Warning: 'newdata' had 1 row but variables found have 47 rows
```

```
male_max <- as.data.frame(
  cbind(max(gambdata$status), max(gambdata$income), max(gambdata$verbal) ))
compare_male <- rbind(male_mean,male_max)
male_max_prdc <- predict(reg_gamb1,compare_male,level = 0.95,interval  = "prediction")
```

```
## Warning: 'newdata' had 2 rows but variables found have 47 rows
```

A male with maximal values of status, income and verbal score will have larger CI, because it has larger standard deviation

School expenditure and test scores from USA in 1994-95

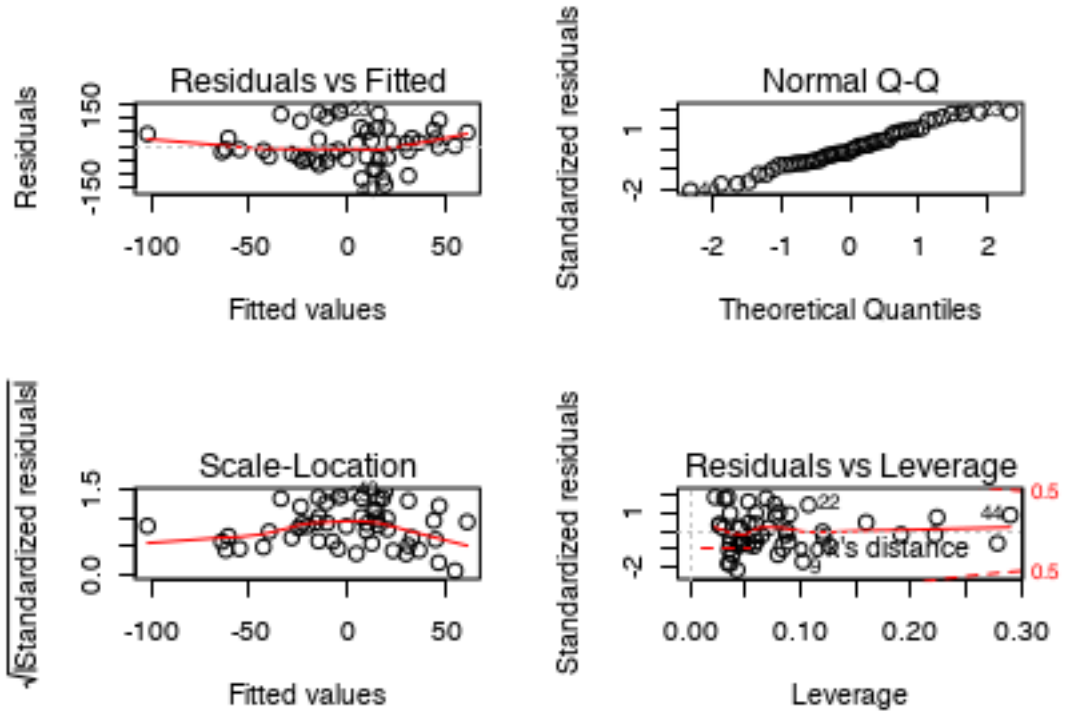
```
data(sat)
?sat
```

1. Fit a model with total sat score as the outcome and expend, ratio and salary as predictors. Make necessary transformation in order to improve the interpretability of the model. Interpret each of the coefficient.

```
centered_total <- sat$total - mean(sat$total)
reg_sat <- lm(centered_total ~ sat$expend + sat$ratio + sat$salary)
summary(reg_sat)
```

```
##
## Call:
## lm(formula = centered_total ~ sat$expend + sat$ratio + sat$salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  103.314    110.925   0.931   0.3565
## sat$expend    16.469     22.050   0.747   0.4589
## sat$ratio      6.330      6.542   0.968   0.3383
## sat$salary    -8.823      4.697  -1.878   0.0667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

```
par(mfrow = c(2,2))
plot(reg_sat)
```



The model is $theaveragetotalsatscore = 103.31 + 16.47 * expend + 6.33 * ratio - 8.82 * salary$

The average total sat score is 103.31 if the current expenditure per pupil in average daily attendance in public elementary and secondary schools, the average pupil/teacher ratio in public elementary and secondary schools, and the estimated average annual salary of teachers in public elementary and secondary schools is 0.

The average total sat score wil increases by 16.47 if the current expenditure per pupil in average daily attendance in public elementary and secondary schools increases by one unit, holding other variable constant.

The average total sat score wil increases by 6.33 if the average pupil/teacher ratio in public elementary and secondary schools increases by one unit, holding other variable constant.

The average total sat score wil decreases by 8.82 if the estimated average annual salary of teachers in public elementary and secondary schools increases by one unit, holding other variable constant.

2. Construct 98% CI for each coefficient and discuss what you see.

```
confint(reg_sat, level = 0.98)
```

```
##                     1 %       99 %
## (Intercept) -164.035801 370.664137
## sat$expend   -36.675540  69.613271
## sat$ratio     -9.437308  22.097842
## sat$salary   -20.142788   2.497524
```

The intercept falls in [-164,370] with 95% of possibility. 0 included, not significant.

The coefficient of expenditure falls in [-37,70] with 95% of possibility.0 included, not significant.

The coefficient of ratio falls in [-9,22] with 95% of possibility.0 included, not significant.

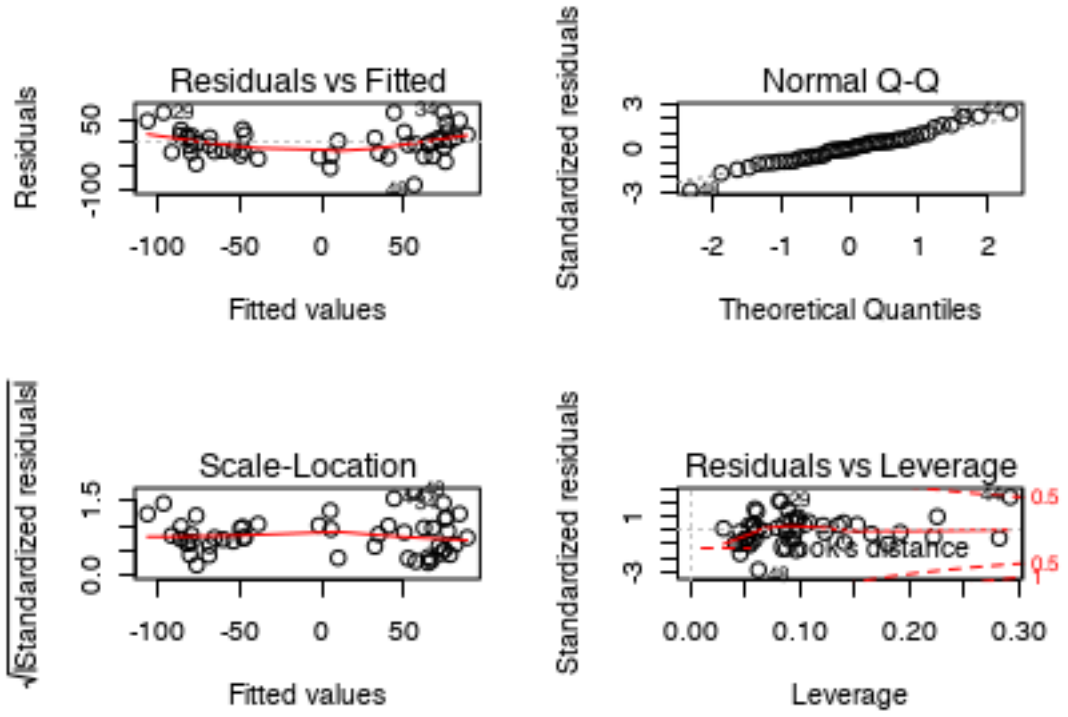
The coefficient of salary falls in [-20,2] with 95% of possibility.0 included, not significant.

3. Now add takers to the model. Compare the fitted model to the previous model and discuss which of the model seem to explain the outcome better?

```
reg_sat1 <- lm(centered_total ~ sat$expend + sat$ratio + sat$salary + sat$takers)
summary(reg_sat1)
```

```
##
## Call:
## lm(formula = centered_total ~ sat$expend + sat$ratio + sat$salary +
##     sat$takers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  80.0515    52.8698   1.514    0.137
## sat$expend    4.4626    10.5465   0.423    0.674
## sat$ratio    -3.6242     3.2154  -1.127    0.266
## sat$salary    1.6379     2.3872   0.686    0.496
## sat$takers   -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(reg_sat1)
```



The model with takers is much better. The R^2 increases from 20% to 82%, and the takers' coefficient appears to be significant. Also, the residual plot of takers model is not evenly spreaded.

Conceptual exercises.

Special-purpose transformations:

For a study of congressional elections, you would like a measure of the relative amount of money raised by each of the two major-party candidates in each district. Suppose that you know the amount of money raised by each candidate; label these dollar values D_i and R_i. You would like to combine these into a single variable that can be included as an input variable into a model predicting vote share for the Democrats.

Discuss the advantages and disadvantages of the following measures:

- The simple difference, $D_i - R_i$

Adv: easy to make comparison and visualization

Dis: It only shows the total value, not the propotion value. Also, the measurement unit has to be the same to make the subtraction make sense.

- The ratio, D_i/R_i

Adv: easy to show relative propotion ratio.

Dis: 50/10 and 500/100 will be the same. But in reality, it makes a huge difference.

- The difference on the logarithmic scale, $logD_i - logR_i$

Adv: Great to show the relative propotion ratio of skewed data.

Dis: log may possiblly lead to calculation error.

- The relative proportion, $D_i/(D_i + R_i)$.

Adv:easy to show relative participation ratio.

Dis: Same the dis. for the ratio, D_i/R_i

Transformation

For observed pair of x and y, we fit a simple regression model

$$\mathrm{y} = \alpha + \beta \mathrm{x} + \epsilon$$

which results in estimates $\hat{\alpha} = 1$, $\hat{\beta} = 0.9$, $SE(\hat{\beta}) = 0.03$, $\hat{\sigma} = 2$ and $r = 0.3$.

1. Suppose that the explanatory variable values in a regression are transformed according to the $\mathrm{x}^\star = \mathrm{x} - 10$ and that y is regressed on $\mathrm{x}^\star$. Without redoing the regression calculation in detail, find $\hat{\alpha}^\star$, $\hat{\beta}^\star$, $\hat{\sigma}^\star$, and $r^\star$. What happens to these quantities when $\mathrm{x}^\star = 10\mathrm{x}$? When $\mathrm{x}^\star = 10(\mathrm{x} - 1)$?

Given $\mathrm{x}^\star = \mathrm{x} - 10$,

We obtain $\hat{\alpha}^\star = \hat{\alpha} + 100.9 = 10$,

$\hat{\beta}^\star$, $\hat{\sigma}^\star$ and $r^\star$ stay the same.

Given $\mathrm{x}^\star = 10\mathrm{x}$,

We obtain $\hat{\beta}^\star = \hat{\beta}/10 = 0.09$

$\hat{\sigma}^\star = \hat{\sigma}/10 = 0.2$

$\hat{\alpha}^\star$ and $r^\star$ stays the same.

Given $\mathrm{x}^\star = 10(\mathrm{x} - 1)$,

We obtain $\hat{\alpha}^\star = \hat{\alpha} + \hat{\beta} = 1.9$

$\hat{\beta}^\star = \hat{\beta}/10 = 0.09$

$\hat{\sigma}^\star = \hat{\sigma}/10 = 0.2$

$r^\star$ stays the same.

2. Now suppose that the response variable scores are transformed according to the formula

$\mathrm{y}^{\star\star} = \mathrm{y} + 10$ and that $\mathrm{y}^{\star\star}$ is regressed on x. Without redoing the regression calculation in detail, find $\hat{\alpha}^{\star\star}$, $\hat{\beta}^{\star\star}$, $\hat{\sigma}^{\star\star}$, and $r^{\star\star}$. What happens to these quantities when $\mathrm{y}^{\star\star} = 5\mathrm{y}$? When $\mathrm{y}^{\star\star} = 5(\mathrm{y} + 2)$?

Given $\mathrm{y}^{\star\star} = \mathrm{y} + 10$,

We obtain $\hat{\alpha}^\star = \hat{\alpha} + 10 = 11$,

$\hat{\beta}^\star$, $\hat{\sigma}^\star$ and $r^\star$ stay the same.

Given $\mathrm{y}^{\star\star} = 5\mathrm{y}$,

We obtain $\hat{\alpha}^\star = 5\hat{\alpha} = 5$

$\hat{\beta}^\star = 5\hat{\beta} = 4.5$

$\hat{\sigma}^\star = 5\hat{\sigma} = 10$

$r^\star$ stays the same.

Given $\mathrm{y}^{\star\star} = 5(\mathrm{y} + 2)$,

We obtain $\hat{\alpha}^\star = 5(\hat{\alpha} + 2) = 15$

$\hat{\beta}^\star = 5\hat{\beta} = 4.5$

$\hat{\sigma}^\star = 5\hat{\sigma} = 10$

$r^\star$ stays the same.

3. In general, how are the results of a simple regression analysis affected by linear transformations of y and x?

The slope and $\hat{\sigma}$ will not be affected by adding a constant to x or y, however, they will be affected by x or y multiplying to some number.

The intercept will be affected if y is multiplied by some number, not if x is multiplied by some number.

r will not be affected by linear transformation

4. Suppose that the explanatory variable values in a regression are transformed according to the $\mathrm{x}^\star = 10(\mathrm{x} - 1)$ and that y is regressed on $\mathrm{x}^\star$. Without redoing the regression calculation in detail, find $SE(\hat{\beta}^\star)$ and $t_0^\star = \hat{\beta}^\star / SE(\hat{\beta}^\star)$.

Becasue $\hat{\beta}^\star = \frac{\hat{\beta}}{10} = 0.09$ and $SE(\hat{\beta}^\star) = SE(\hat{\beta})/10 = 0.003$,

We obtain that $t_0^\star = \hat{\beta}^\star / SE(\hat{\beta}^\star) = 30$.

5. Now suppose that the response variable scores are transformed according to the formula $\mathrm{y}^{\star\star} = 5(\mathrm{y} + 2)$ and that $\mathrm{y}^{\star\star}$ is regressed on x. Without redoing the regression calculation in detail, find $SE(\hat{\beta}^{\star\star})$ and $t_0^{\star\star} = \hat{\beta}^{\star\star} / SE(\hat{\beta}^{\star\star})$.

Becasue $\hat{\beta}^{\star\star} = 5 * \hat{\beta} = 4.5$ and $SE(\hat{\beta}^{\star\star}) = 5 * SE(\hat{\beta}^\star) = 0.15$,

We obtain that $t_0^{\star\star} = \hat{\beta}^{\star\star} / SE(\hat{\beta}^{\star\star}) = 30$.

6. In general, how are the hypothesis tests and confidence intervals for β affected by linear transformations of y and x?

The confidence interval is $[\hat{\beta} \pm t_{\frac{\alpha}{2}} * SE(\beta)]$

From the formula we can see that CI will not be affected if adding a number to x or y, however, if x is multiplied by a constant, CI will be the original CI divided by that contant. If y is multiplied by a constant, CI will be the original CI multiplied by that constant.

Feedback comments etc.

If you have any comments about the homework, or the class, please write your feedback here. We love to hear your opinions.