# Homework 02

*Weiling Li*

*Septemeber 21, 2019*

# 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"))
#wfw90    <- read.table (paste0(gelman_dir,"earnings/wfw90.dat"))
```

Pull out the data on earnings, sex, height, and weight.

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

The dataset has 9 variables, from the code book, they mean the following:

- earn: personal income during year 1989, in dollars.
- height1: height in inches
- height2: height in inches
- sex:
  - male: 1
  - female: 2
- race:
  - white: 1
  - black: 2
  - asian: 3
  - native amerian: 4
  - others: 9
- hisp:
  - hispanic origin: 1
  - otherwise: 2
- ed: highest grade or years in school(highest grade converted to years in school) from 0 to 18, integers.
- yearbn: year of born in 19xx.

- height: interviewee's height in inches, rounded to nearest integer.

```r
## Earnings has NA value and 0 value, which needs to be cleaned.
## Also, to better managing data. original data had been put in tibble form.
## For the purpose of this HW, only earn,sex,race,ed & height were kept
h1 <- filter(filter(as_tibble(heights),!is.na(earn)),earn >0)%>%select(earn,sex,race,yearbn,ed,height)
```

```
## Warning: `lang()` is deprecated as of rlang 0.2.0.
## Please use `call2()` instead.
## This warning is displayed once per session.

## Warning: `new_overscope()` is deprecated as of rlang 0.2.0.
## Please use `new_data_mask()` instead.
## This warning is displayed once per session.

## Warning: `overscope_eval_next()` is deprecated as of rlang 0.2.0.
## Please use `eval_tidy()` with a data mask instead.
## This warning is displayed once per session.
```

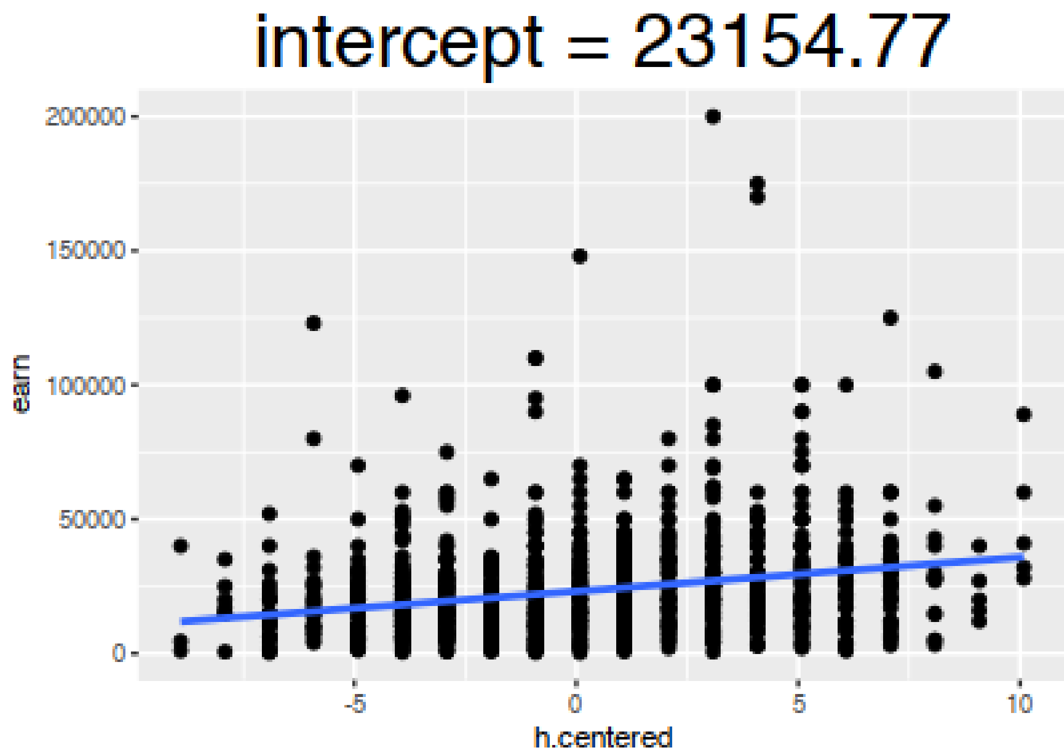
```r
#hist(log(h1$earn),probability = T)
#hist(h1$ed,probability = T)
#hist(h1$height,probability = T)
```

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?

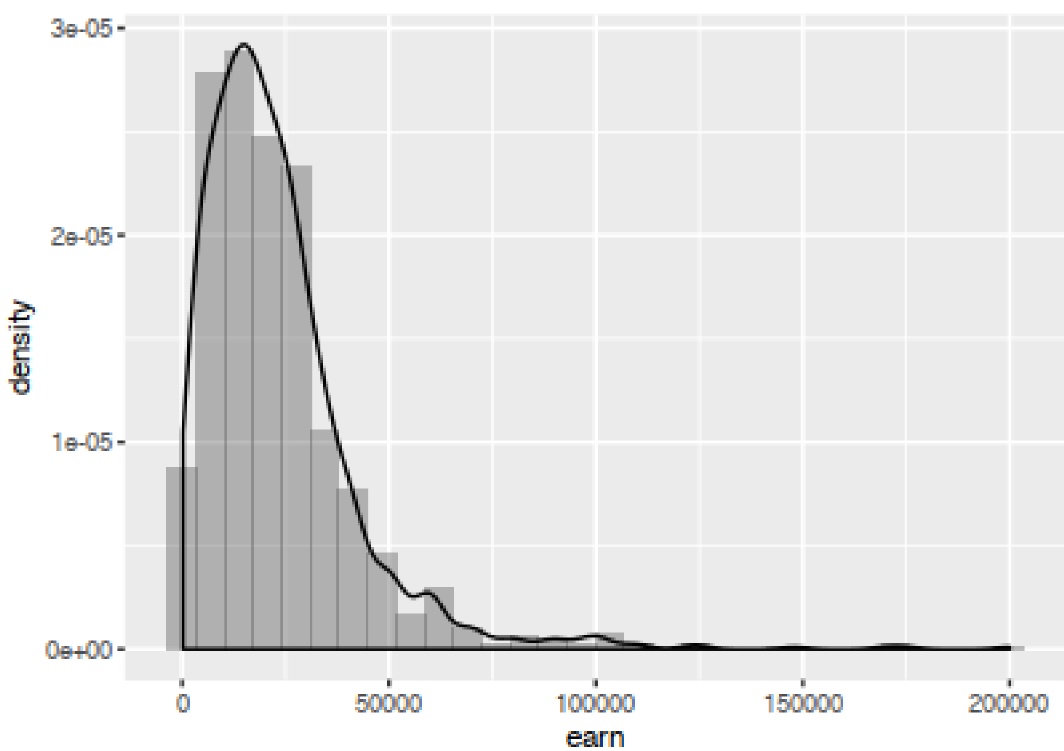
```r
#to achieve what exactly asked. we only need to subtract height with mean.
h1 <- mutate(h1,h.centered = height - mean(h1$height))
```

```
## Warning: The `printer` argument is deprecated as of rlang 0.3.0.
## This warning is displayed once per session.
```

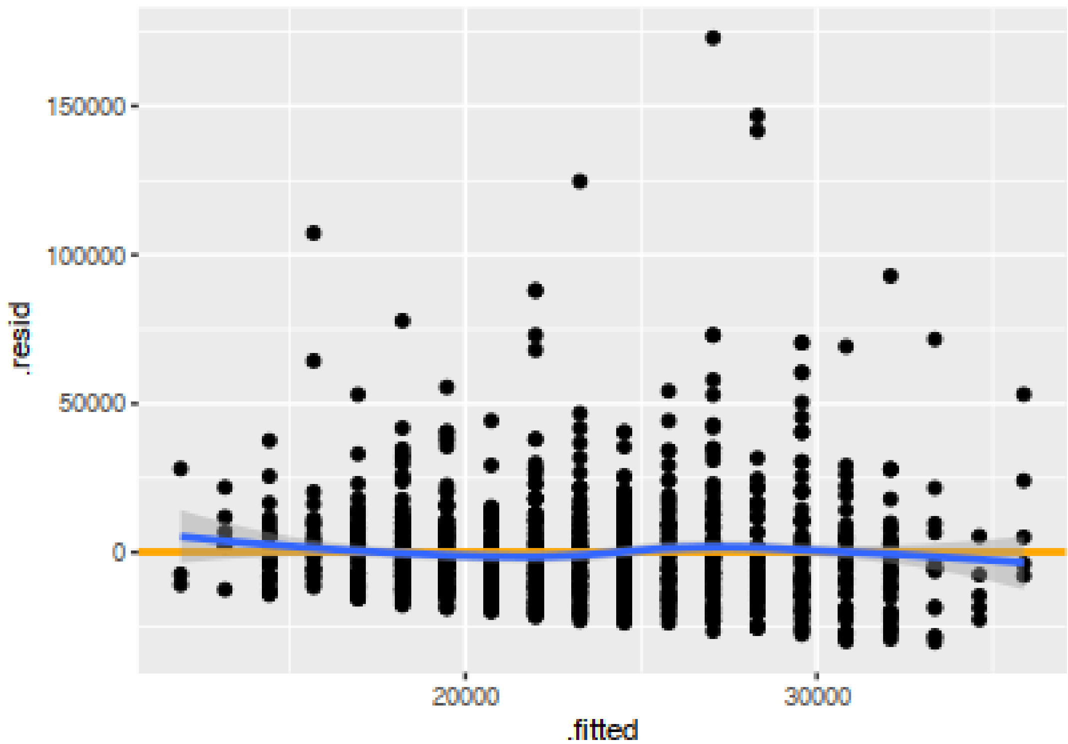
```r
#hist(height.centered)
#hist(h1$height)
ggplot(h1)+
  aes(x = h.centered, y = earn)+geom_point()+geom_smooth(method = 'lm',se = F)+ggtitle(paste0('intercep
```




```r
#however, as can be seen from the residual plot and the histogram of earn, the skewness of the earn cau
ggplot(h1)+aes(x = earn)+geom_histogram(bins = 30,alpha = .4,aes(y=..density..))+geom_density()
```




```r
ggplot(lm(data = h1, earn ~ h.centered)) + aes(x=.fitted, y=.resid)+
  geom_point()+geom_abline(intercept = 0,slope = 0,color='orange',size=1)+geom_smooth(method = 'loess',a
```




```
#this violates the assumption of linear regression, to better achieve the results. One should at 1st re
```

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.
   census data: us 1990 census
   codebook: wfwcodebook.txt

```
## for better interpretation and simplicity, the following transformation of the data set is done as fo
## take log transformation of the earning
## substract sex with 1 to make the variable a binary with 0 indicate male and 1 indicate female.
## 90 - yearbn to get the approximate age for the interviewee as they report their earning. to ensure n
h.transformed <- mutate(h1,log.earn = log(earn)) %>% mutate(sex = sex - 1) %>% mutate(age = ifelse(year
## A sample of the transformed data is shown below
kable(sample_n(h.transformed,10,replace = T)%>%select(log.earn,sex,h.centered,age),format = 'latex',dig
```

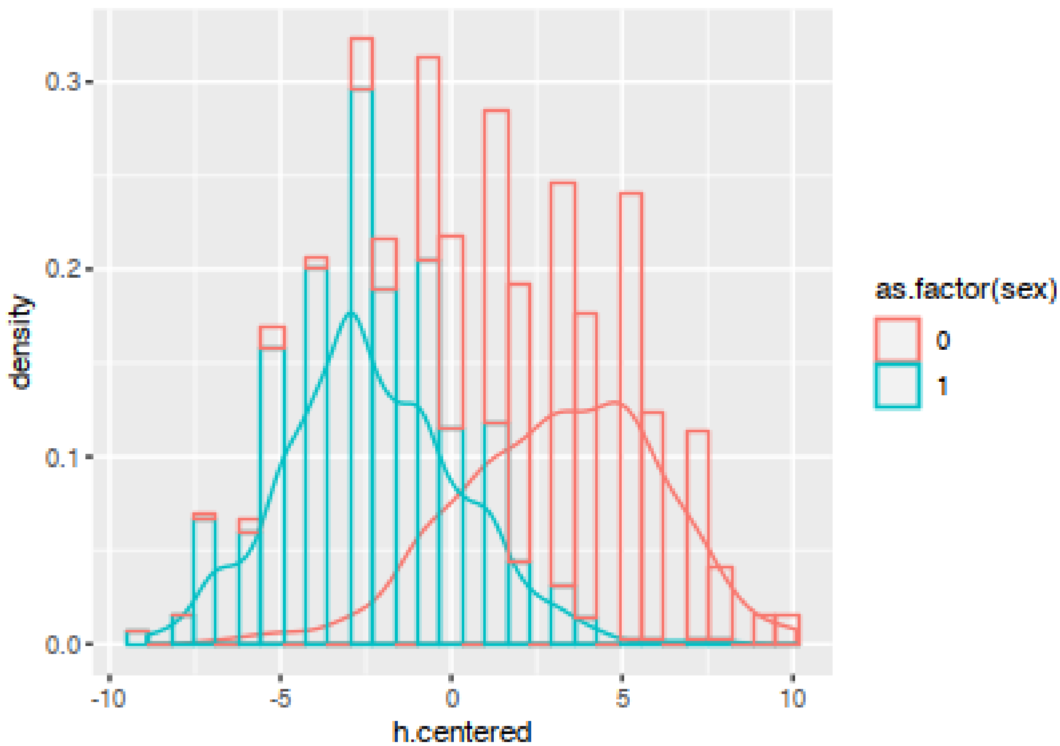
| log.earn | sex | h.centered | age |
|---|---|---|---|
| 10.37 | 1 | -1.92 | 28 |
| 9.78 | 1 | -2.92 | 30 |
| 9.62 | 0 | 1.08 | 28 |
| 11.00 | 0 | 7.08 | 54 |
| 10.00 | 1 | -4.92 | 47 |
| 10.09 | 1 | -2.92 | 30 |
| 8.29 | 0 | 5.08 | 21 |
| 9.84 | 1 | -3.92 | 36 |
| 8.52 | 1 | 2.08 | 24 |
| 10.13 | 0 | 0.08 | 26 |

```r
##check race factor:
kable(group_by(h.transformed,race)%>%summarise(count.race = n()),format = 'latex')
```

| race | count.race |
|---|---|
| 1 | 1051 |
| 2 | 112 |
| 3 | 15 |
| 4 | 11 |
| 9 | 3 |

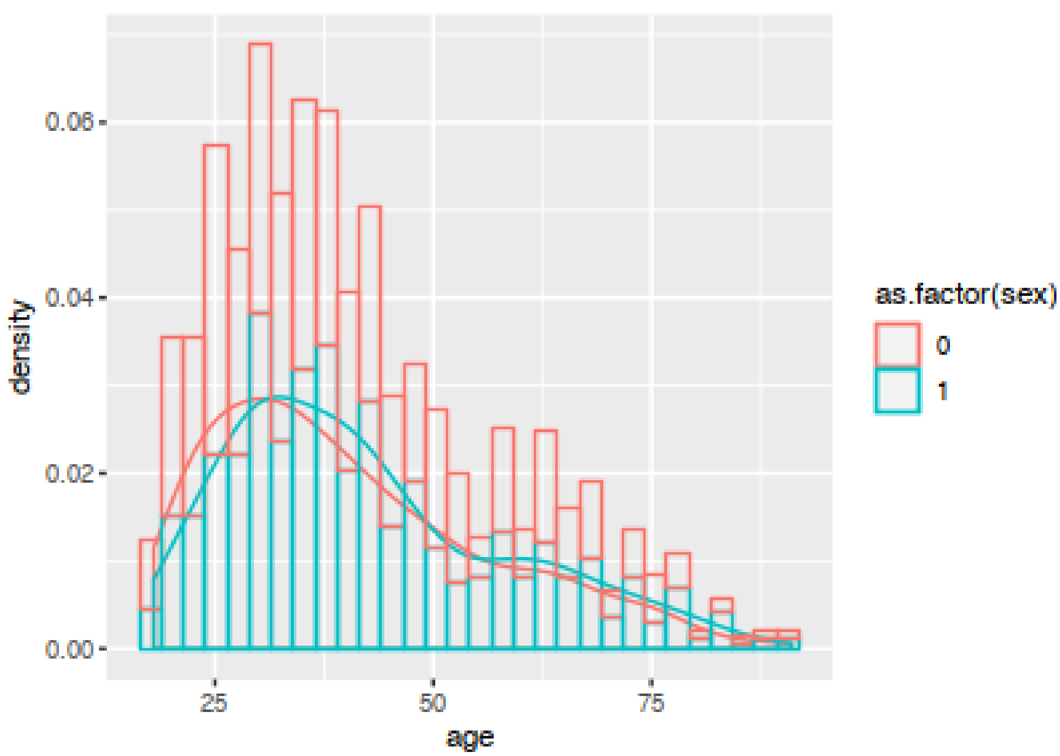
```r
## as can be seen the majority interviewees are white, less than 10% are black and less than 2% are oth
## The 1990 census data shows that about 80% american population is white, 12% is black and the remaini
#ggplot(h.transformed)+aes(x = h.centered,y = log.earn,color = as.factor(sex))+geom_point()+geom_smooth

## Lets see how our variables natrually distributed under gender factors
## height
ggplot(h.transformed)+aes(x = h.centered,color = as.factor(sex))+geom_density( alpha = .4)+geom_histogra
```



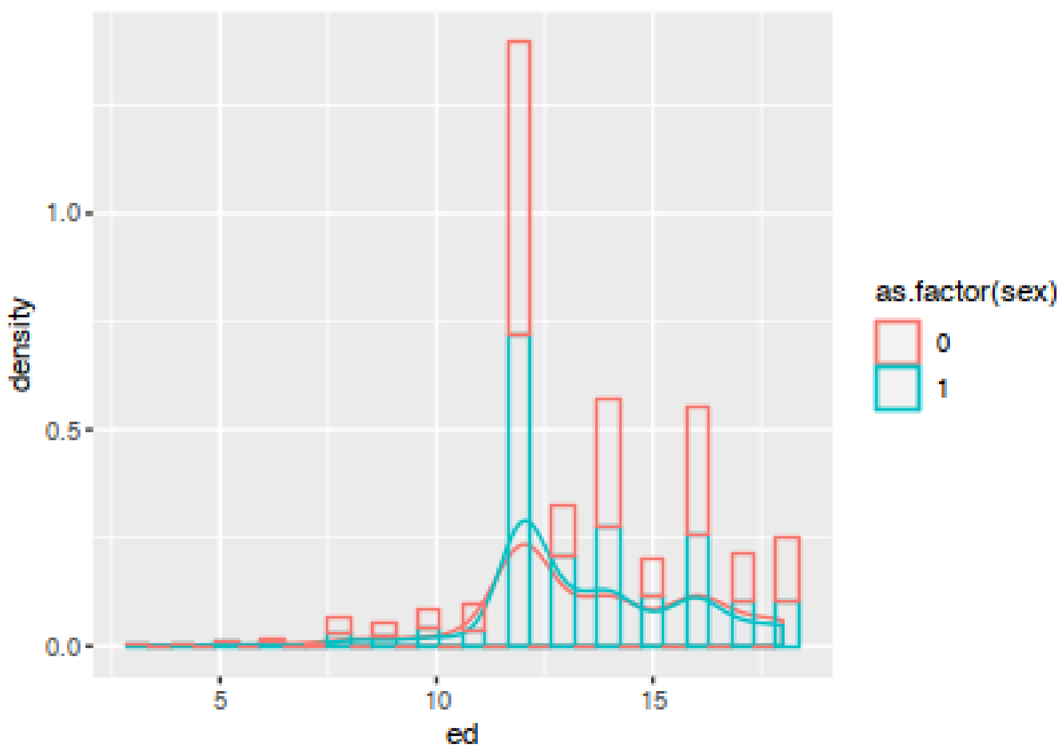

```r
ggplot(h.transformed)+aes(x = h.centered,y = log.earn,color = as.factor(sex))+geom_jitter()+geom_smooth
```


```
## age
ggplot(h.transformed)+aes(x = age,color = as.factor(sex))+geom_density( alpha = .4)+geom_histogram(bins
```

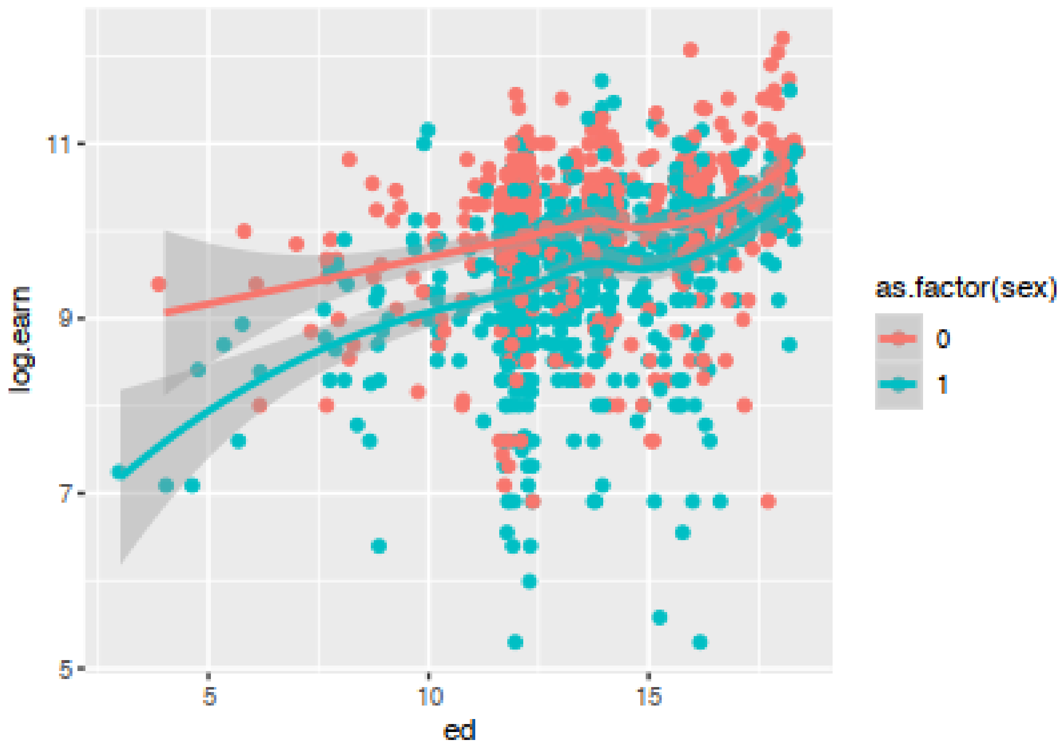


```
ggplot(h.transformed)+aes(x = age,y = log.earn,color = as.factor(sex))+geom_jitter()+geom_smooth(method=
```


```
## education
ggplot(h.transformed)+aes(x = ed,color = as.factor(sex))+geom_density( alpha = .4)+geom_histogram(bins =
```



```r
ggplot(h.transformed)+aes(x = ed,y = log.earn,color = as.factor(sex))+geom_jitter()+geom_smooth(method=
```




One can observed from the histogram that height variable is unevenly distributed between genders. female has a mean roughly 3 inches below average while male roughly 4 inches above average with both genders have an approximated sample standard error at 2.5 inches. This fact approximately puts the mean of one gender out of 2 standard errors of the opposite gender. in this case, when we are predicting using height variable, we inevitably have to use gender to separate them. Also, because gender had seperated heights into two clusters, it is better to introduce interaction terms $height \times gender$ to ensure linear regression models the two genders differently.
The model's coefficient and it's residual plots was given below:

```r
MD1 = lm(data = h.transformed,log.earn ~ sex + h.centered + sex*h.centered)
summary(MD1)$call
```

```
## lm(formula = log.earn ~ sex + h.centered + sex * h.centered,
##     data = h.transformed)
```

```r
summary(summary(MD1))
```

```
##               Length Class  Mode
## call             3   -none- call
## terms            3   terms  call
## residuals     1192   -none- numeric
## coefficients    16   -none- numeric
## aliased          4   -none- logical
## sigma            1   -none- numeric
## df               3   -none- numeric
## r.squared        1   -none- numeric
## adj.r.squared    1   -none- numeric
```

```
## fstatistic       3   -none- numeric
## cov.unscaled    16   -none- numeric
```

```r
summary(MD1)$coef
```

```
##                     Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)      9.946321548 0.05736331 173.391704 0.000000e+00
## sex             -0.419713073 0.07301657  -5.748190 1.145709e-08
## h.centered       0.024454484 0.01330615   1.837834 6.633649e-02
## sex:h.centered -0.007446534 0.01863502  -0.399599 6.895237e-01
```

```r
summary(MD1)$r.squared
```

```
## [1] 0.08668346
```

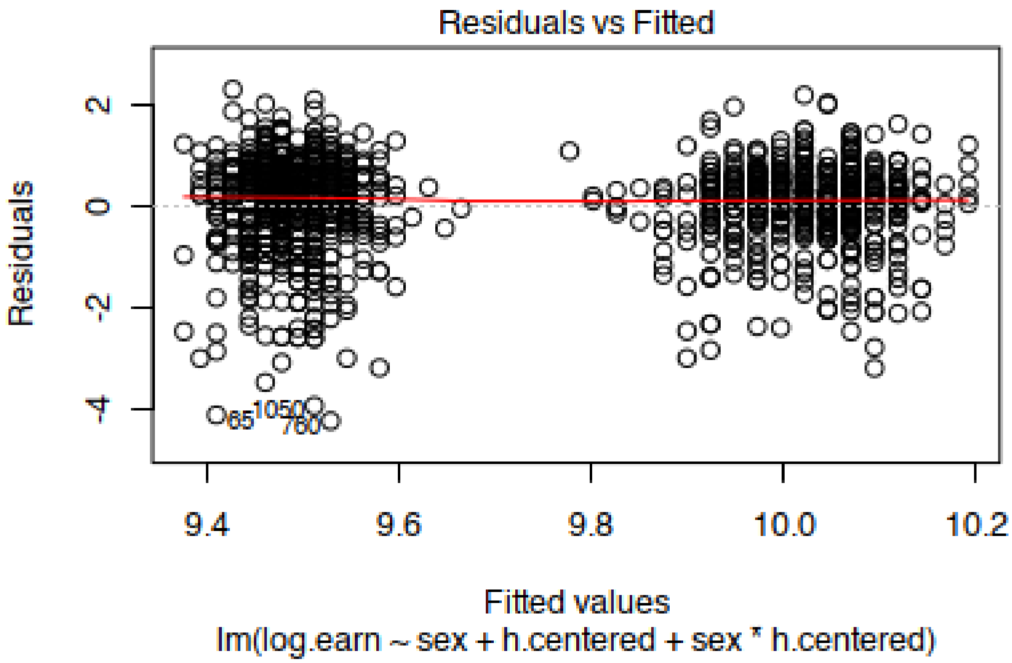
```r
summary(MD1)$call
```

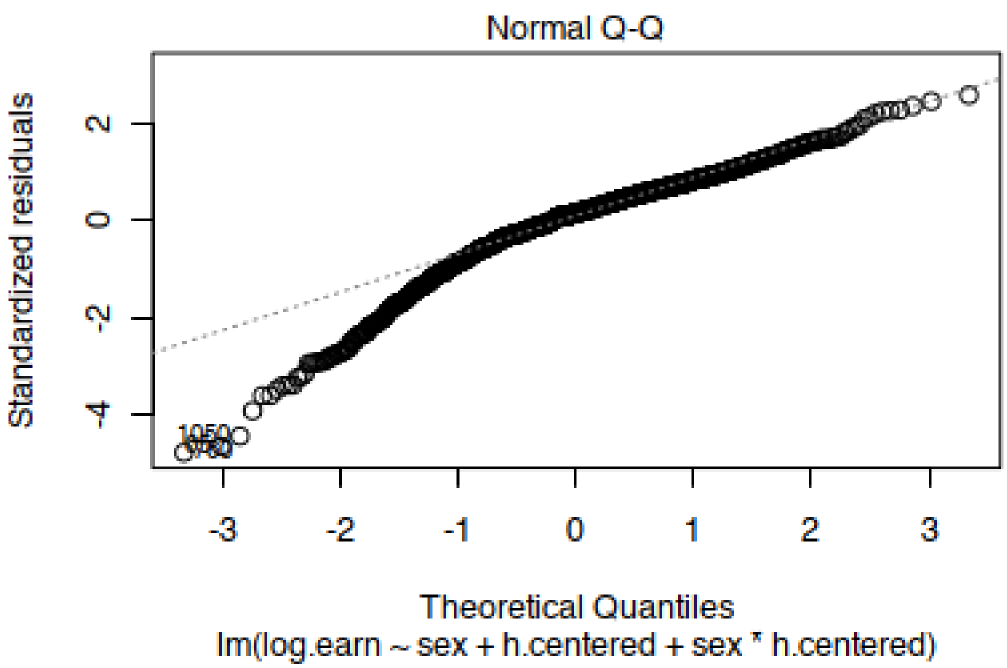
```
## lm(formula = log.earn ~ sex + h.centered + sex * h.centered,
##     data = h.transformed)
```

```r
summary(MD1)
```

```
##
## Call:
## lm(formula = log.earn ~ sex + h.centered + sex * h.centered,
##     data = h.transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2297 -0.3720  0.1388  0.5646  2.2940
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.946322   0.057363 173.392  < 2e-16 ***
## sex            -0.419713   0.073017  -5.748 1.15e-08 ***
## h.centered      0.024454   0.013306   1.838   0.0663 .
## sex:h.centered -0.007447   0.018635  -0.400   0.6895
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8812 on 1188 degrees of freedom
## Multiple R-squared:  0.08668,    Adjusted R-squared:  0.08438
## F-statistic: 37.58 on 3 and 1188 DF,  p-value: < 2.2e-16
```

```r
plot(MD1,which=1:2)
```




```
ggplot(h.transformed)+aes(x = h.centered,y = log.earn,color = as.factor(sex))+geom_jitter()+geom_smooth
```


From the plot shown, the residual show a tiny bias but stays really close to 0, but the variance is quite high and not equal across all the heights. From the QQ-plot we can see that the model works reasonably well for higher height values but have problems with lower height values.
From the age vs log.earn plot we can clearly see that there can be a nonlinear relationship between the two, for comparison we construct the model 2 as $y = gender + age + age^2 + gender * age + gender * age^2$.

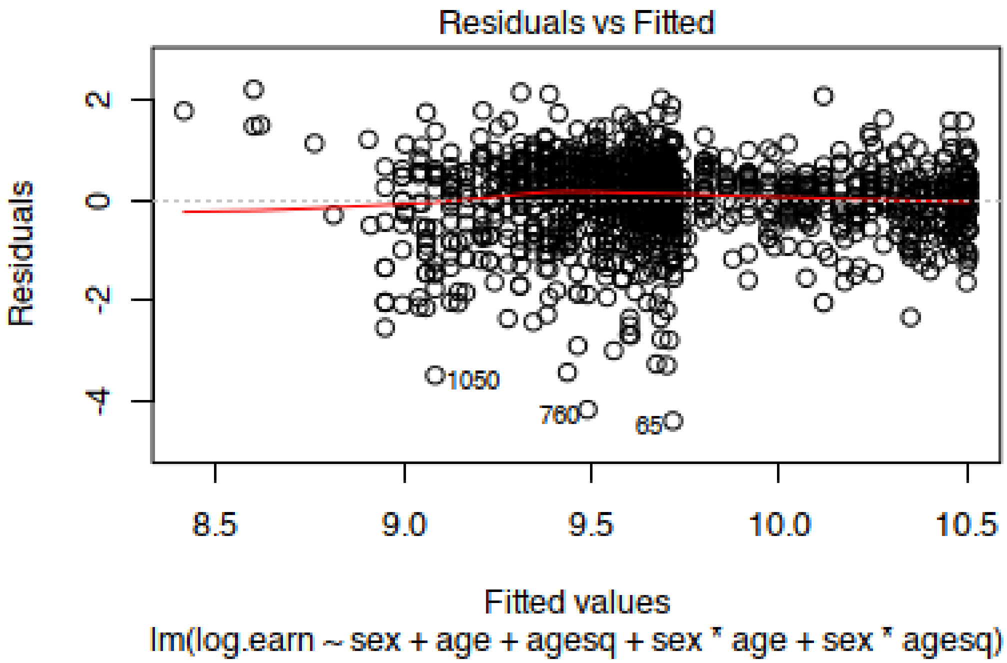
```
## substract age with min(age) & add one colomn age^2
h.transformed <- mutate(h.transformed,age = age-min(age))%>%mutate(agesq = age^2)
MD2 = lm(data = h.transformed, log.earn ~ sex + age + agesq + sex*age + sex *agesq)
summary(MD2)$coef[,1]
```

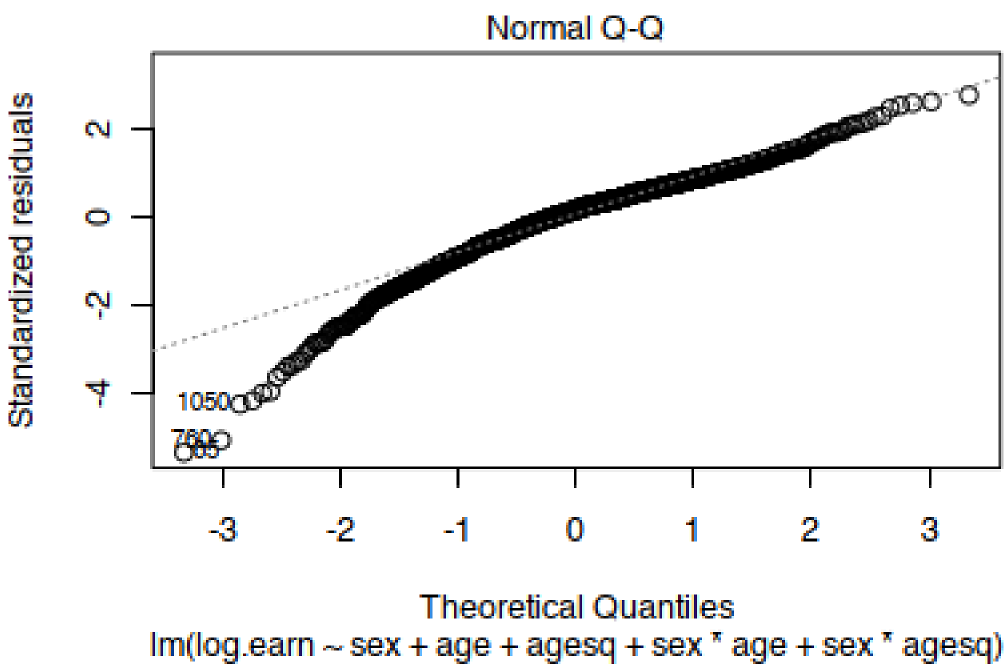
```
##   (Intercept)           sex           age         agesq       sex:age
##  9.0601894067 -0.1091444448  0.0872425976 -0.0013161904 -0.0407576635
##     sex:agesq
##  0.0006137022
```

```
md2coef <- summary(MD2)$coef[,1]
summary(MD2)$r.squared
```

```
## [1] 0.19855
```

```
plot(MD2,which = 1:2)
```




4. Interpret all model coefficients. Both models discussed above shows although using different variables, but shows a similar results in residual analysis. Thus, for the purpose of this HW, model 2 using age & gender is selected to interpret here.

```r
kable(t(md2coef),format = 'latex',digits = ,align = 'c')
```

| (Intercept) | sex | age | agesq | sex:age | sex:agesq |
|---|---|---|---|---|---|
| 9.060189 | -0.1091444 | 0.0872426 | -0.0013162 | -0.0407577 | 0.0006137 |

```r
kable(exp(t(md2coef)),format = 'latex',digits = ,align = 'c')
```

| (Intercept) | sex | age | agesq | sex:age | sex:agesq |
|---|---|---|---|---|---|
| 8605.78 | 0.8966009 | 1.091161 | 0.9986847 | 0.9600618 | 1.000614 |

The model can be written down as:

$$log.earn = \beta_0 + \beta_1 \cdot gender + \beta_2 \cdot age + \beta_3 \cdot age^2 + \beta_4 \cdot gender \cdot age + \beta_5 \cdot gender \cdot age^2$$

$$earn = exp(\beta 0) \cdot exp(\beta_1)^{gender} \cdot exp(\beta_2)^{age} \cdot exp(\beta_3)^{age^2} \cdot exp(\beta_4)^{gender \cdot age} \cdot exp(\beta_5)^{gender \cdot age^2}$$

The effect of the binary term *gender* causes the regression function to be separated for male and female: male will have the function as:

$$earn = exp(\beta 0) \cdot exp(\beta_2)^{age} \cdot exp(\beta_3)^{age^2}$$

female will have:

$$earn = exp(\beta 0 \cdot \beta_1) \cdot exp(\beta_2 \cdot \beta_4)^{age} \cdot exp(\beta_3 \beta_5)^{age^2}$$

The interpretation follows as:

- $exp(\beta_0) = 8605.78$ is the mean earning for minimun age 18 as male which 8605.78.
- $exp(\beta_1) = 0.90$ means at same age, female with minimun age 18 earns 90
- $exp(\beta_2) = 1.09$ means for male, 1 years older averagely result in 9
- $exp(\beta_3) = 0.998$ means for male, 1 unit larger in $age^2$ result in 0.2
- $exp(\beta_4) = 0.96$ means for female, compare to male, can expect 4
- $exp(\beta_5) = 1.00$ means for female, compare to male, can expect the same 0.2

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

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.06 | 0.09 | 96.19 | 0.0 |
| sex | -0.11 | 0.13 | -0.84 | 0.4 |
| age | 0.09 | 0.01 | 11.10 | 0.0 |
| agesq | 0.00 | 0.00 | -10.03 | 0.0 |
| sex:age | -0.04 | 0.01 | -3.88 | 0.0 |
| sex:agesq | 0.00 | 0.00 | 3.58 | 0.0 |

```
## Warning: Setting row names on a tibble is deprecated.
```

| | Est | StandardError | lower95CI | upper95CI |
|---|---|---|---|---|
| (Intercept) | 9.06 | 0.09 | 8.87 | 9.25 |
| sex | -0.11 | 0.13 | -0.37 | 0.15 |
| age | 0.09 | 0.01 | 0.07 | 0.10 |
| agesq | 0.00 | 0.00 | 0.00 | 0.00 |
| sex:age | -0.04 | 0.01 | -0.06 | -0.02 |
| sex:agesq | 0.00 | 0.00 | 0.00 | 0.00 |

**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:

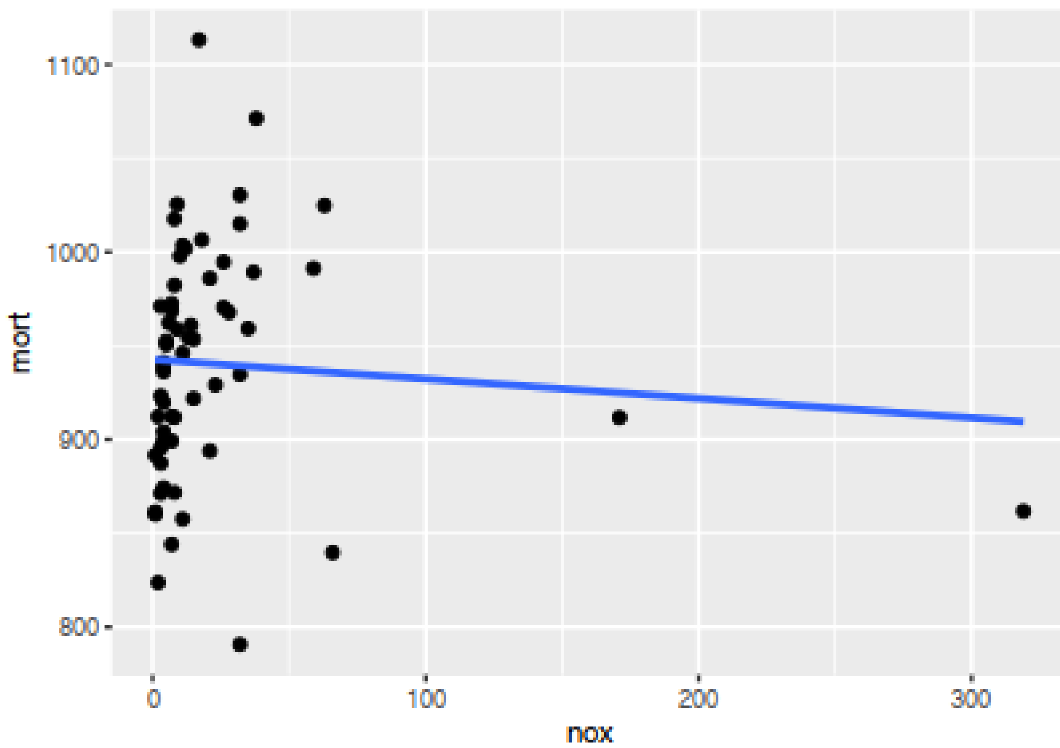
- PREC Average annual precipitation in inches
- JANT Average January temperature in degrees F
- JULT Same for July
- OVR65 % of 1960 SMSA population aged 65 or older
- POPN Average household size
- EDUC Median school years completed by those over 22
- HOUS % of housing units which are sound & with all 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.

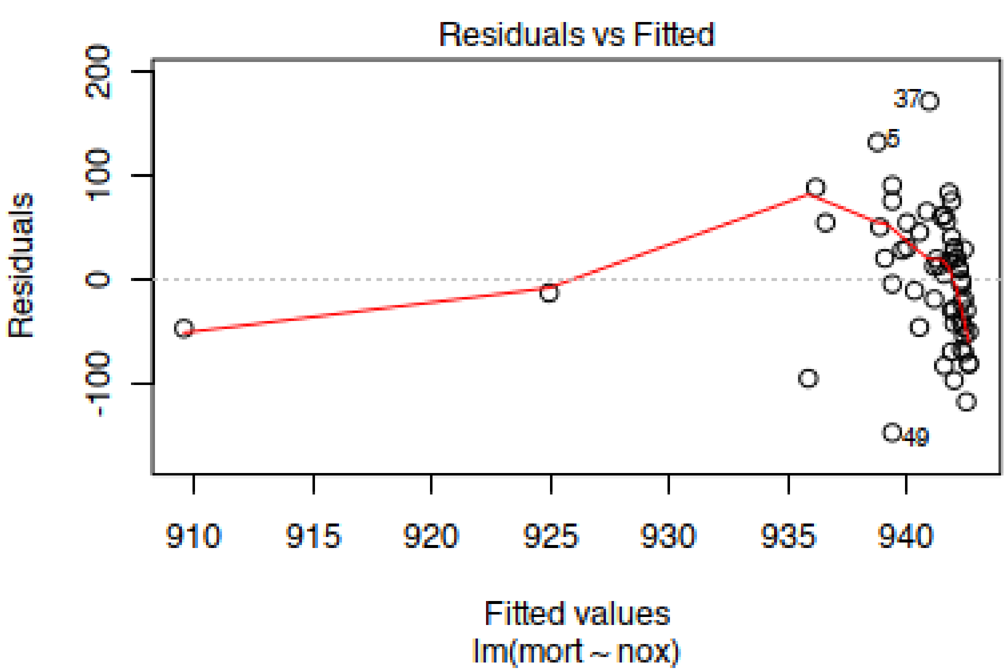
```r
gelman_dir   <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
pollution    <- read.dta (paste0(gelman_dir,"pollution/pollution.dta"))
```

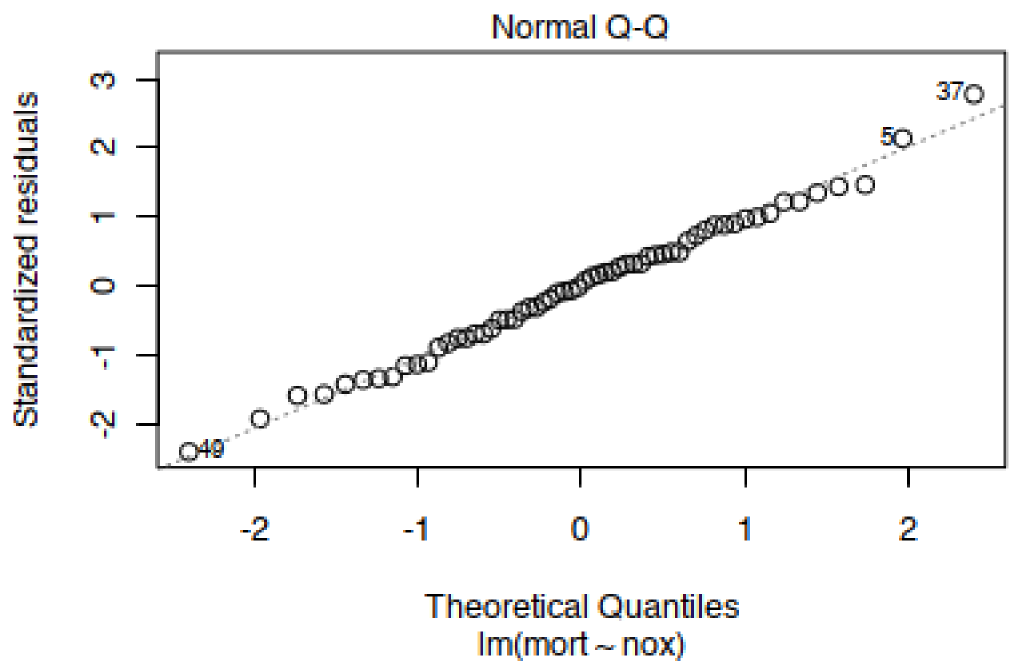
1. Create a scatterplot of mortality rate versus level of nitric oxides. Do you think linear regression will fit these data well? Fit the regression and evaluate a residual plot from the regression.

```r
ggplot(pollution)+aes(x = nox,y = mort)+geom_point()+geom_smooth(method = 'lm',se=F)
```




```
polMD1 = lm(data = pollution, mort ~ nox)
plot(polMD1,which=1:2)
```





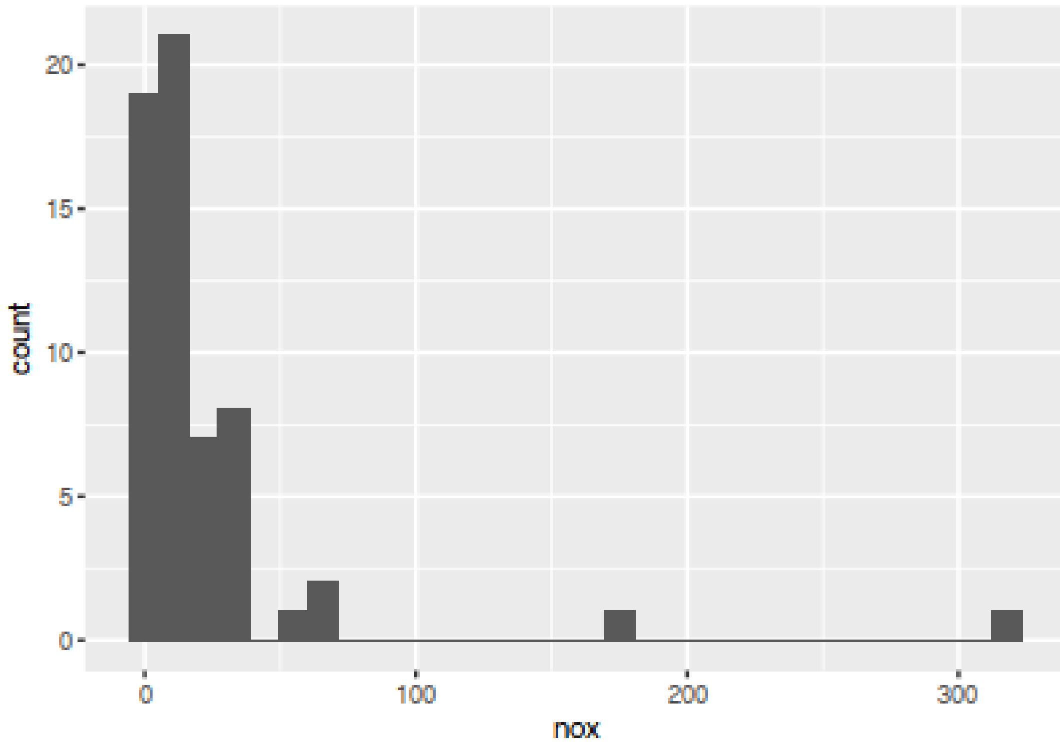

```r
summary(polMD1)
```

```
##
## Call:
## lm(formula = mort ~ nox, data = pollution)
##
## 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 ***
## 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
```

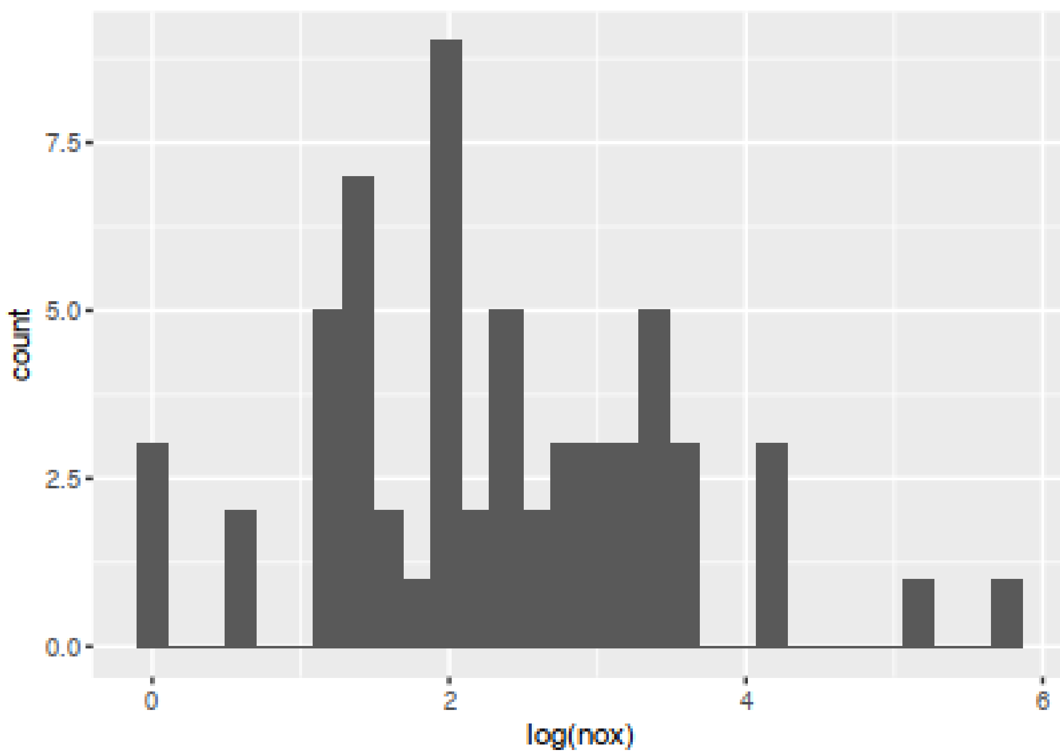
Judging from the regression plot and the residual plot, the fit is really bad.

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.

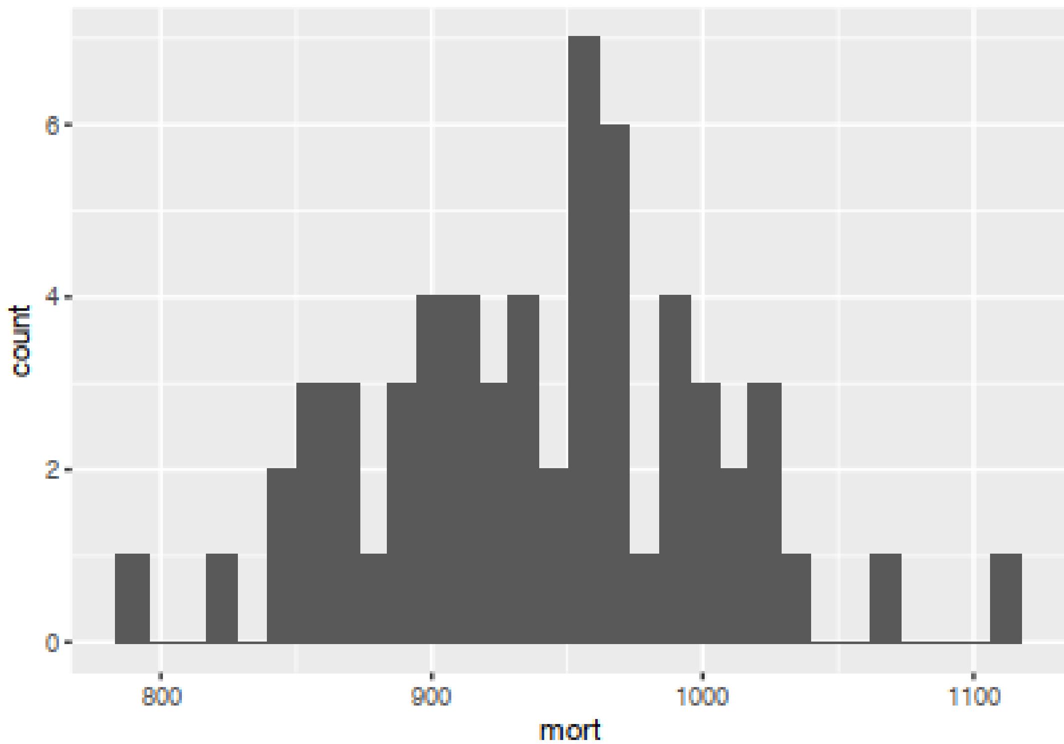
```
ggplot(pollution)+aes(x = nox)+geom_histogram(bins=30)
```



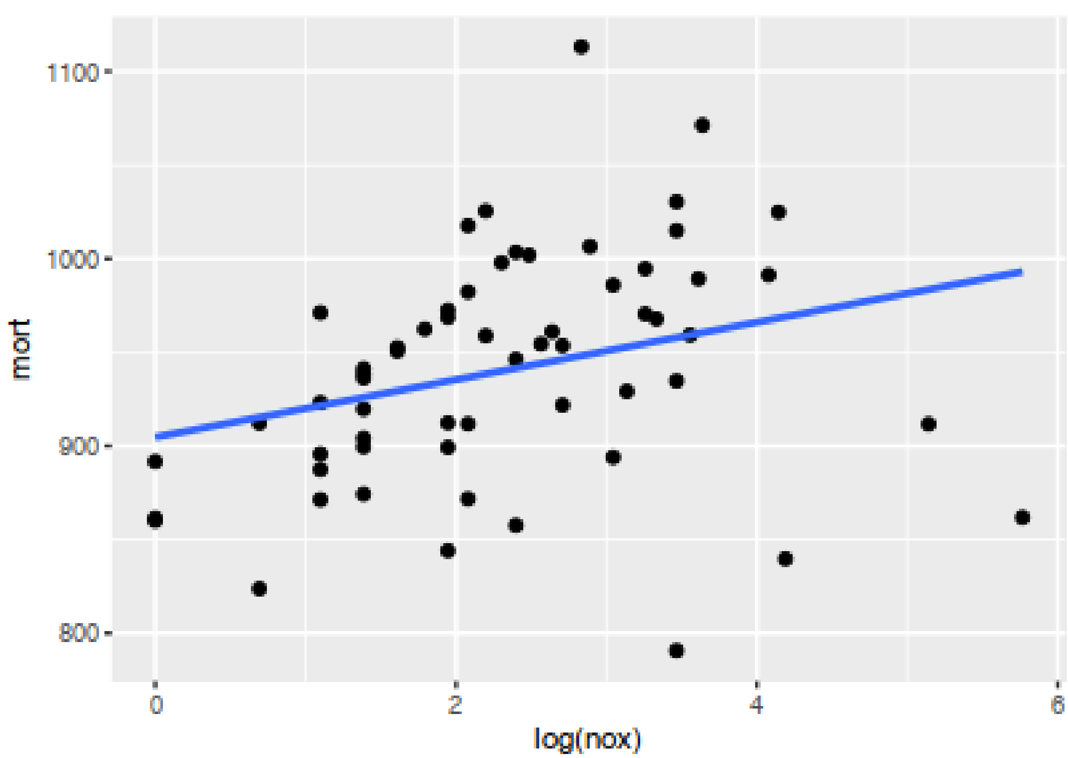

```
ggplot(pollution)+aes(x = log(nox))+geom_histogram(bins=30)
```



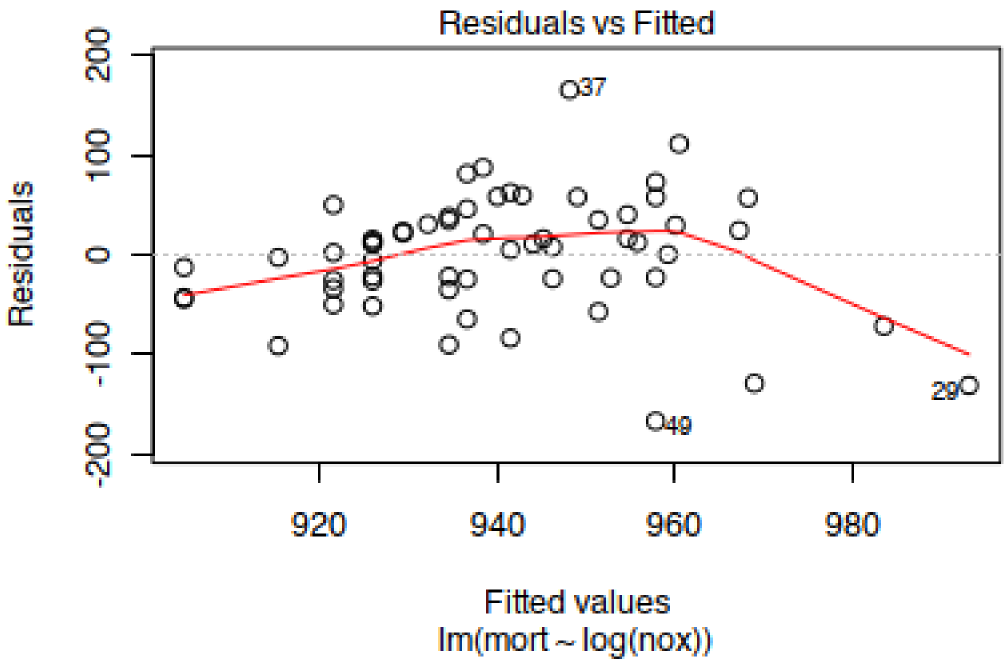
```
ggplot(pollution)+aes(x = mort)+geom_histogram(bins=30)
```

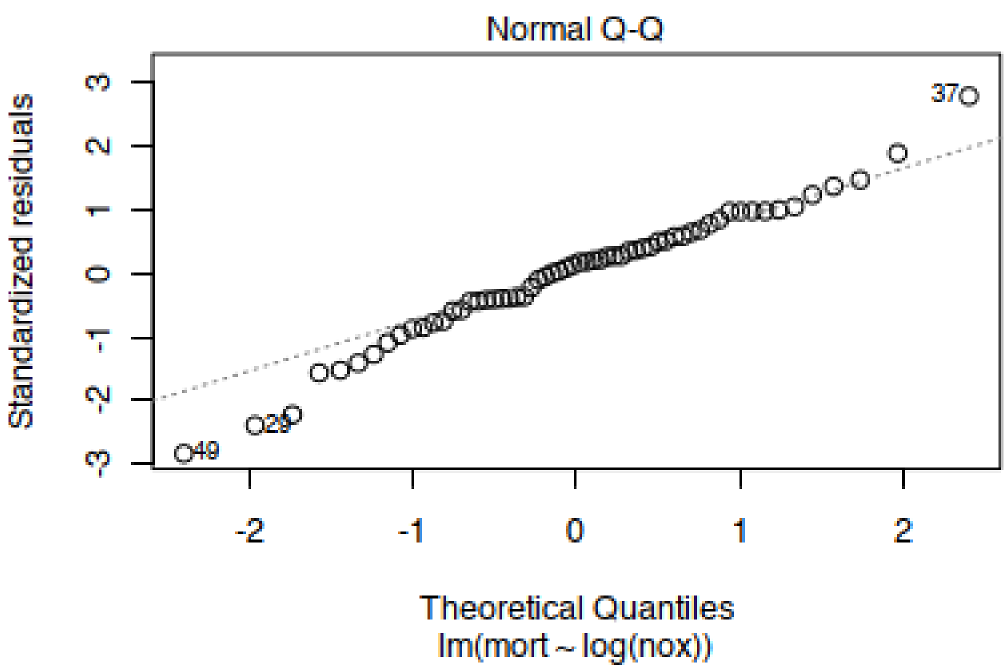



```
ggplot(pollution)+aes(x=log(nox),y=mort)+geom_point()+geom_smooth(method='lm',se=F)
```



```r
polMD2 = lm(data = pollution, mort ~ log(nox))
plot(polMD2,which=1:2)
```



```r
summary(polMD2)
```

```
##
## Call:
## lm(formula = mort ~ log(nox), data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -167.140  -28.368    8.778   35.377  164.983
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  904.724     17.173  52.684   <2e-16 ***
## log(nox)      15.335      6.596   2.325   0.0236 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.01 on 58 degrees of freedom
## Multiple R-squared:  0.08526,    Adjusted R-squared:  0.06949
## F-statistic: 5.406 on 1 and 58 DF,  p-value: 0.02359
```

The new model fits better to the dataset than the previous one. The residual looks better, and the R-square had increased in a factor of 10.

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

```r
summary(polMD2)
```


```
##
## Call:
## lm(formula = mort ~ log(nox), data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -167.140  -28.368    8.778   35.377  164.983
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  904.724     17.173  52.684   <2e-16 ***
## log(nox)      15.335      6.596   2.325   0.0236 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.01 on 58 degrees of freedom
## Multiple R-squared:  0.08526,    Adjusted R-squared:  0.06949
## F-statistic: 5.406 on 1 and 58 DF,  p-value: 0.02359
```


The intercept 904.724 is the mean mortality rate per 100,000 for nox concentration of 1.
4. Construct 99% confidence interval for slope coefficient from the model you chose in 2 and interpret them.

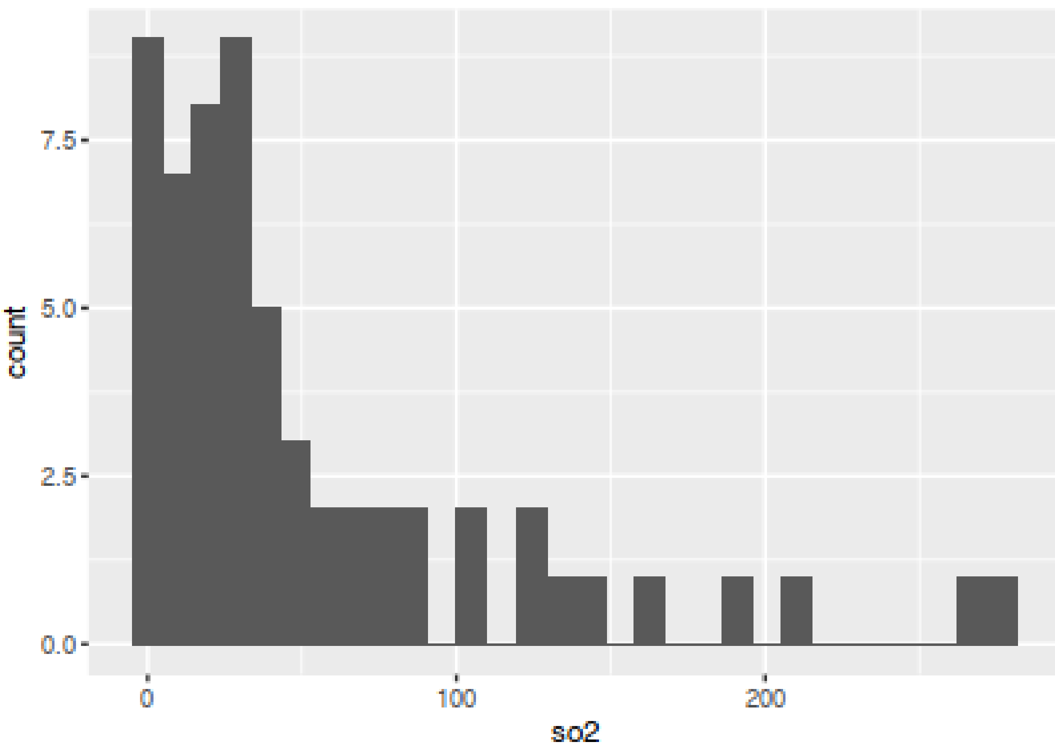
```r
kable(t(c(15.335-3*6.596,15.335+3*6.596)),col.names = c('99%CIlowerbound','99%CIupperbound'),align = 'c
```

| 99%CIlowerbound | 99%CIupperbound |
|---|---|
| -4.453 | 35.123 |

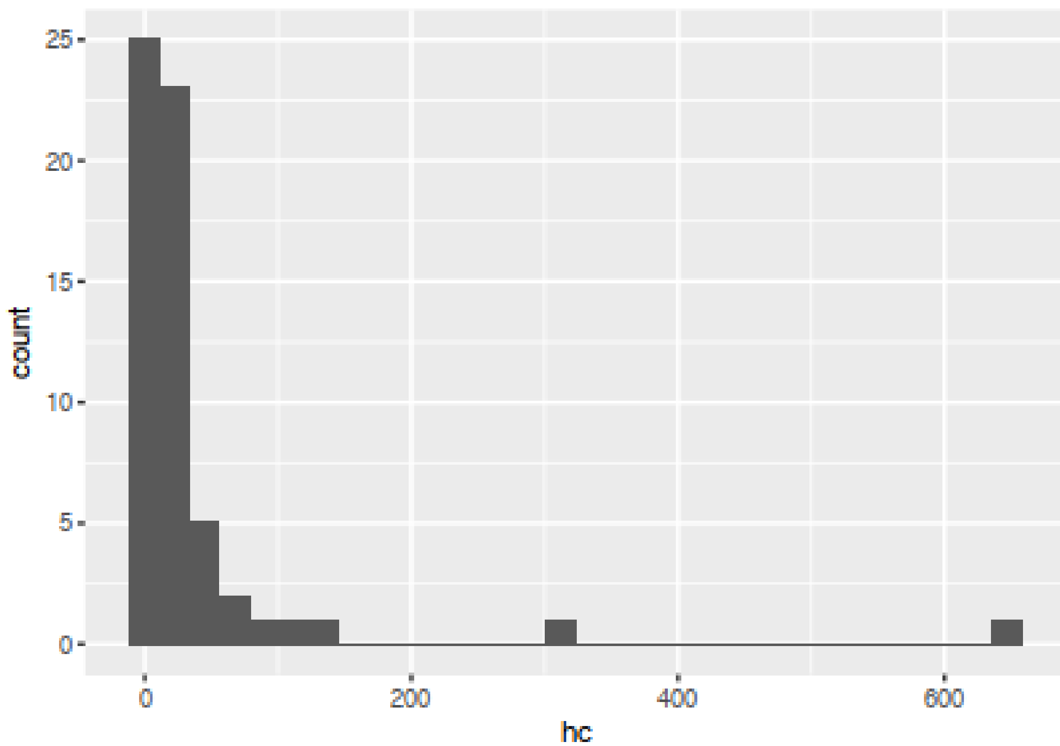
if the same experiment can be done 100 times, one can expect 99 of the times the true slope value lies within

the rage of −4.453 to 35.123.
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.

```r
ggplot(pollution)+aes(x = so2)+geom_histogram(bins=30)
```

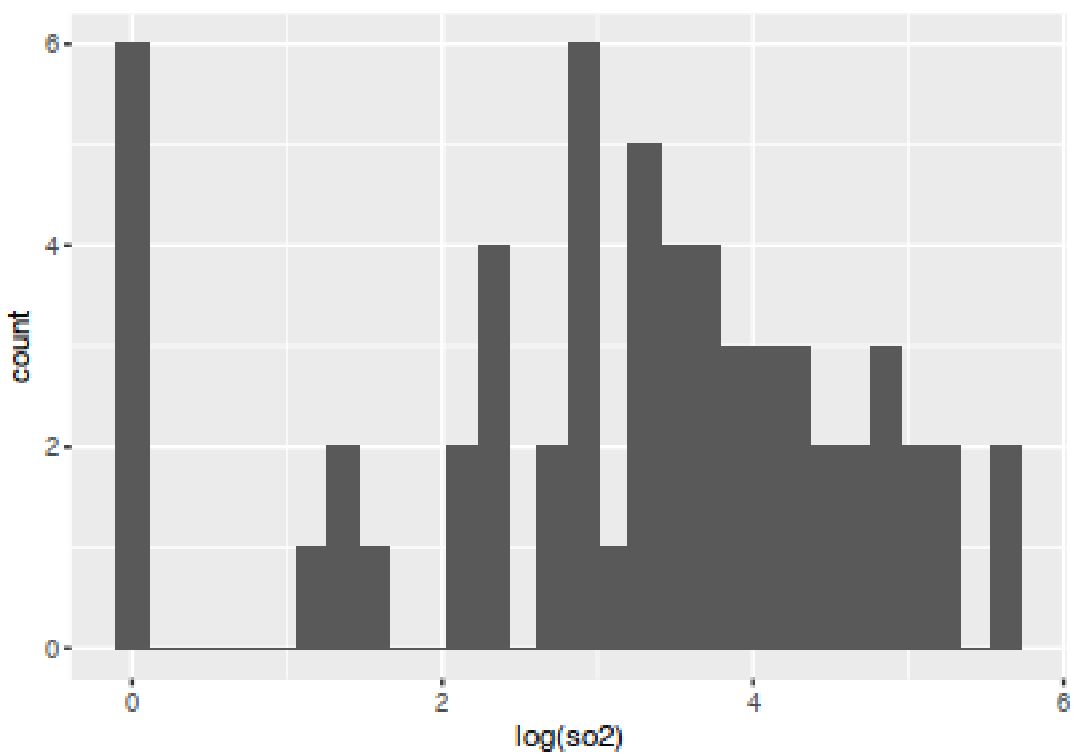



```r
ggplot(pollution)+aes(x = hc)+geom_histogram(bins=30)
```

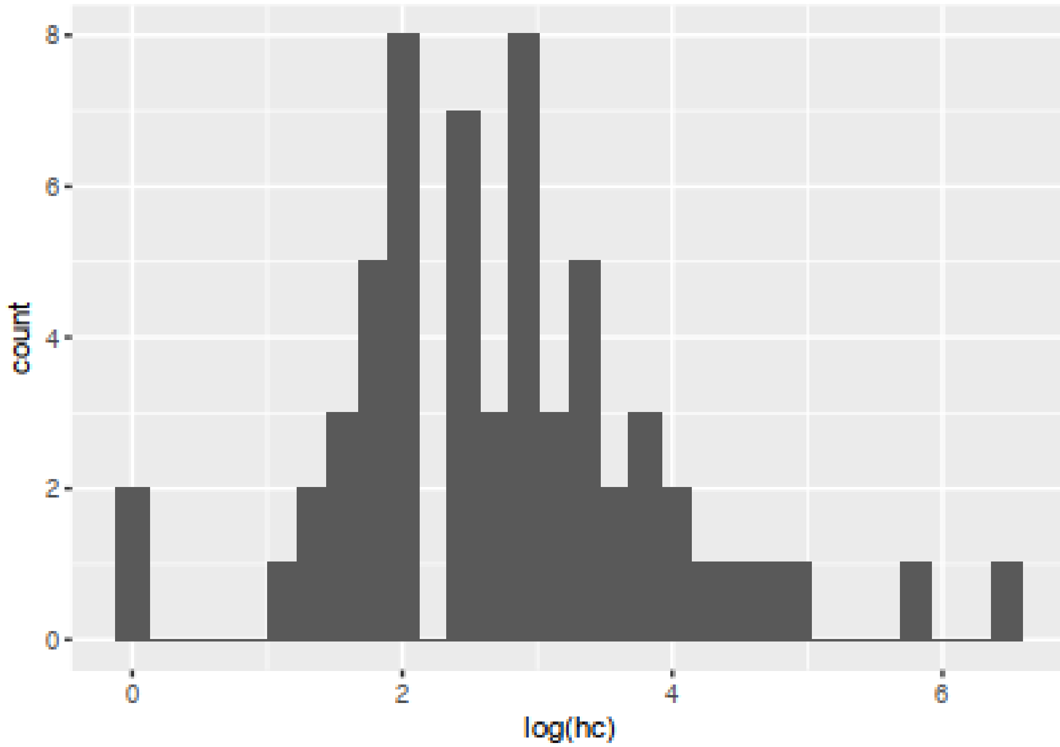



both $SO^2$ and $HC$ is heavily right skewed, so we can try log transformation:

```
ggplot(pollution)+aes(x = log(so2))+geom_histogram(bins=30)
```



```r
ggplot(pollution)+aes(x = log(hc))+geom_histogram(bins=30)
```




log transformation looks fine, thus we can use this to perform our prediction.

```r
polMD3 = lm(data = pollution, mort ~ log(nox) + log(so2) + log(hc))
summary(polMD3)
```

```
##
## Call:
## lm(formula = mort ~ log(nox) + log(so2) + log(hc), data = pollution)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -97.793 -34.728  -3.118  34.148 194.567
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  924.965     21.449  43.125  < 2e-16 ***
## log(nox)      58.336     21.751   2.682  0.00960 **
## log(so2)      11.762      7.165   1.642  0.10629
## log(hc)      -57.300     19.419  -2.951  0.00462 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.36 on 56 degrees of freedom
## Multiple R-squared:  0.2752, Adjusted R-squared:  0.2363
## F-statistic: 7.086 on 3 and 56 DF,  p-value: 0.0004044
```

```r
pollution1 = mutate(pollution,predt = predict(polMD3))
ggplot(pollution1)+geom_point(mapping = aes(x = log(nox),y = mort,color = 'Observed Mortality'))+geom_po
```


Under this model, intercept means the average mortality rate is 924.965 per 100,000 people with nox, so2 and hc all equals to 1.
coefficient for $\ln nox$ means that for each 1 unit increase in $\ln nox$ one can averagly expect 58.336 increase in mortality rate per 100,000 people
coefficient for $\ln so_2$ means that for each 1 unit increase in $\ln so_2$ one can averagly expect 11.762 increase in mortality rate per 100,000 people
coefficient for $\ln hc$ means that for each 1 unit increase in $\ln hc$ one can averagly expect 57.300 decrese in mortality rate per 100,000 people
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.)

```
pol1 = sample_frac(pollution,0.5,replace = F)
pol2 = setdiff(pollution,pol1)%>%select(nox,so2,hc,mort)

polMD4 = lm(data = pol1, mort ~ log(nox) + log(so2) + log(hc))
summary(polMD4)$r.squared
```

```
## [1] 0.5507843
```

```
pol2 = mutate(pol2, predt = predict(object = polMD4,newdata = pol2))%>%mutate(y = mort - mean(mort))%>%r

##R-squared on test set
(1-((pol2$err%*%pol2$err)/(pol2$y%*%pol2$y)))[1]
```

```
## [1] -0.5548842
```

### Study of teenage gambling in Britain

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.

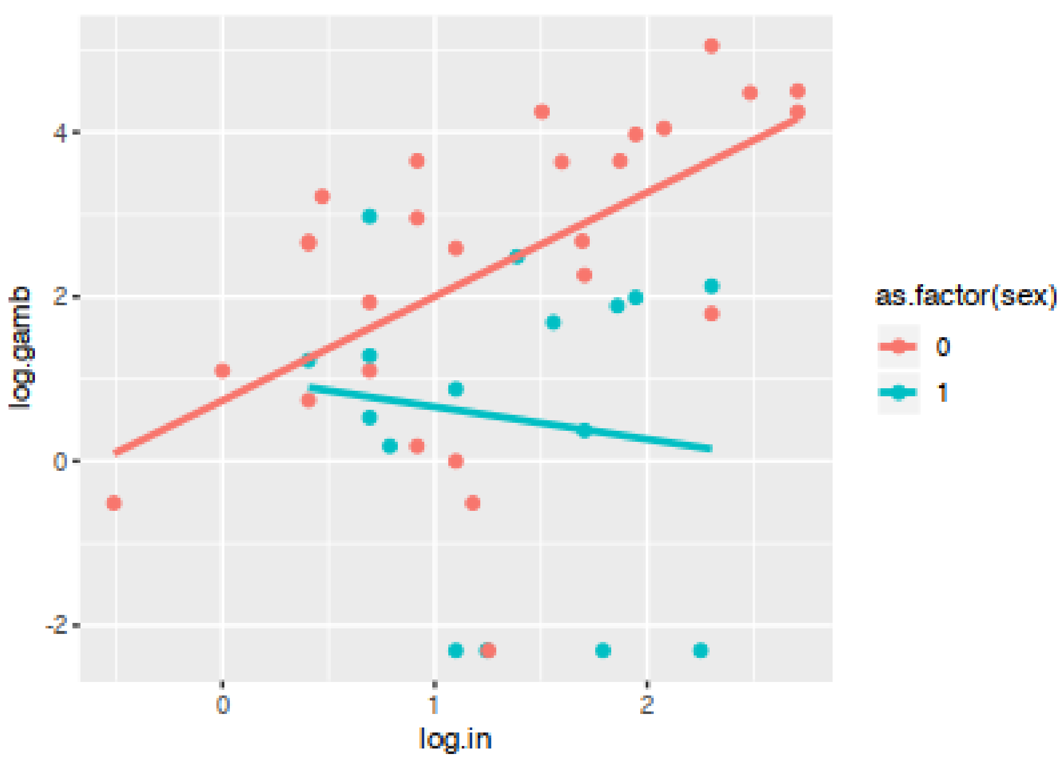
```
## construct log income and log gambling
teengamb1 = mutate(teengamb1,log.in = log(income))%>%mutate(log.gamb = log(gamble))
## choosing log.income and sex to build model
teenMD = lm(data = teengamb1,log.gamb ~ sex + log.in + sex * log.in)
teenMDcoef = exp(as.tibble(summary(teenMD)$coef)[,1])
kable(t(teenMDcoef),format = 'latex',digits = 3, align = 'c',col.names = c('Intercept','Gender','Log.in
```

| | Intercept | Gender | Log.in | Gen : Log.in |
|---|---|---|---|---|
| Estimate | 2.091 | 1.373 | 3.545 | 0.19 |

```
kable(summary(teenMD)$coef,format = 'latex', digits = 3, align = 'c')
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 0.738 | 0.600 | 1.229 | 0.226 |
| sex | 0.317 | 1.224 | 0.259 | 0.797 |
| log.in | 1.265 | 0.392 | 3.232 | 0.003 |
| sex:log.in | -1.659 | 0.828 | -2.005 | 0.052 |

```
ggplot(teengamb1)+aes(x = log.in, y=log.gamb,color = as.factor(sex))+geom_point()+geom_smooth(method =
```




The intercept means the mean for male with 1 income per week, will gamble 2.091 per year in pounds
The coefficient for gender means the for female with 1 income per week is predicted to put 37.3% more on gambling per year
The coefficient for log.income means that for male, unit increase in log income per week will increase gambing per year by 254.5%
The coefficient for Gen:log.in means that for female, the increase in gambling per year will decrease by 81% compare to male
2. Create a 95% confidence interval for each of the estimated coefficients and discuss how you would interpret this uncertainty.

```r
gambsum = as_tibble(summary(teenMD)$coef)%>%select(1:2)%>%`colnames<-`(c('Est','StandardError'))%>%muta
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```r
kable(gambsum,format = 'latex',digits = 2,align = 'c')
```

| | Est | StandardError | lower95CI | upper95CI |
|---|---|---|---|---|
| (Intercept) | 0.74 | 0.60 | -0.46 | 1.94 |
| sex | 0.32 | 1.22 | -2.13 | 2.77 |
| log.in | 1.27 | 0.39 | 0.48 | 2.05 |
| sex:log.in | -1.66 | 0.83 | -3.31 | 0.00 |

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?

```r
teengamb1 = mutate(teengamb1,mean.status = status-mean(status))%>%mutate(mean.logincome = log.in - mean
teenMD2 = lm(data = teengamb1,log.gamb ~ mean.status + mean.logincome + mean.verbal)
teenMD3 = lm(data = teengamb1,log.gamb ~ max.status + max.logincome + max.verbal)
sum1 = t(as_tibble(summary(teenMD2)$coef[1,1:2]))
##lower95CI for mean
sum1[1]-2*sum1[2]
```

```
## [1] 1.129144
```

```r
##upper95CI for mean
sum1[1]+2*sum1[2]
```

```
## [1] 2.238994
```

```r
sum2 = t(as_tibble(summary(teenMD3)$coef[1,1:2]))
##lower95CI for mean
sum2[1]-2*sum2[2]
```

```
## [1] 1.822091
```

```r
##upper95CI for mean
sum2[1]+2*sum2[2]
```

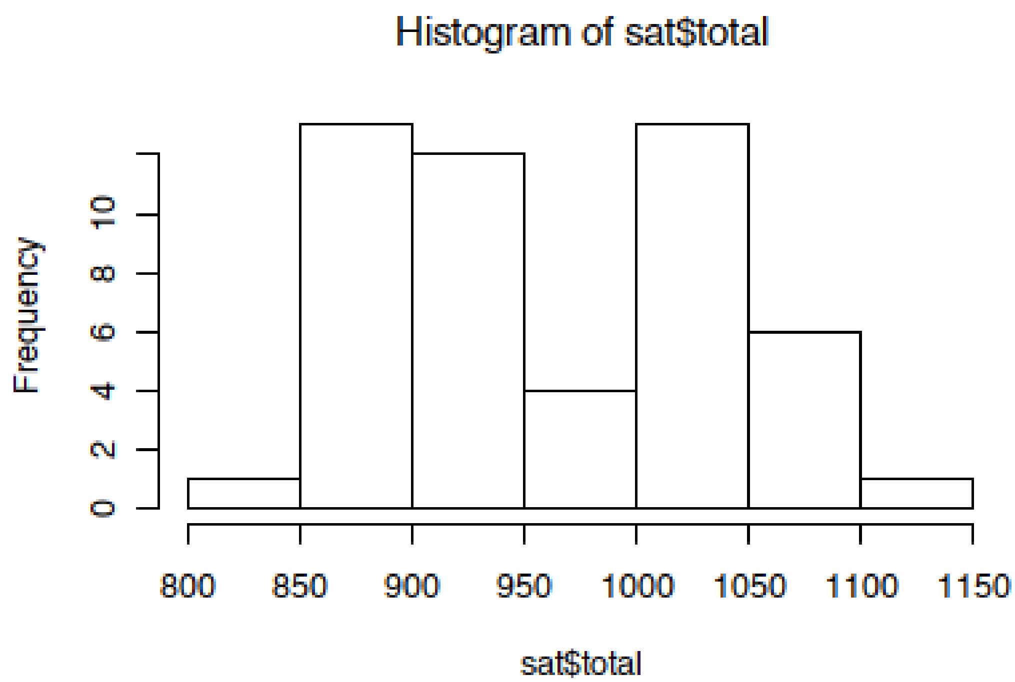
```
## [1] 5.49428
```

The maximal CI is wider, because locally there is fewer points than the mean ones, which results in bigger variance. ### School expenditure and test scores from USA in 1994-95

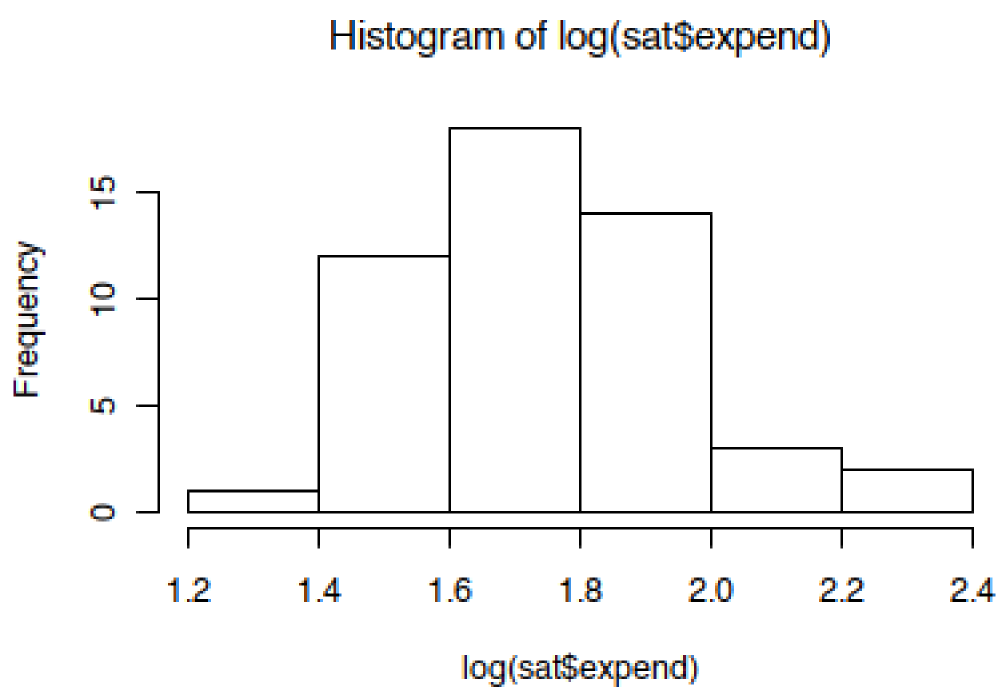
```r
sat = sat
```

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.

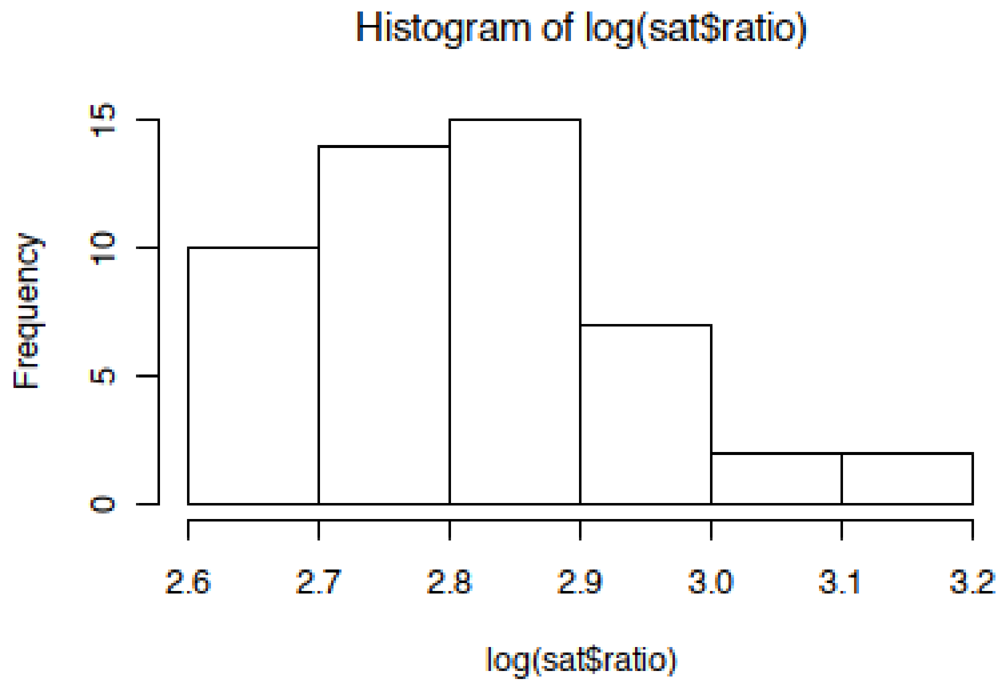
```r
hist(sat$total)
```




```r
hist(log(sat$expend))
```




```r
hist(log(sat$ratio))
```

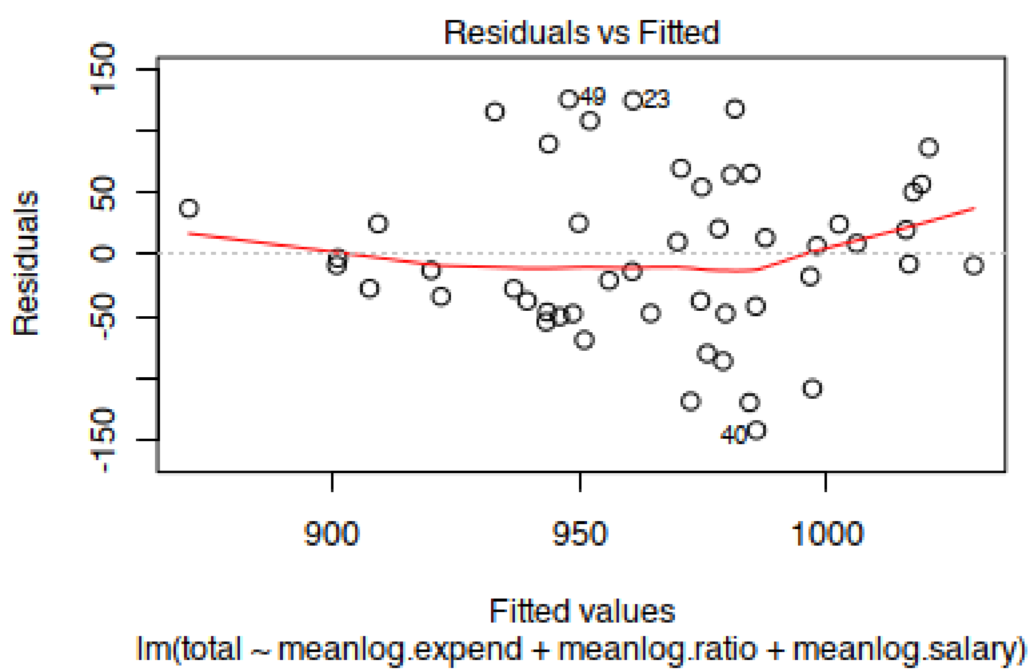


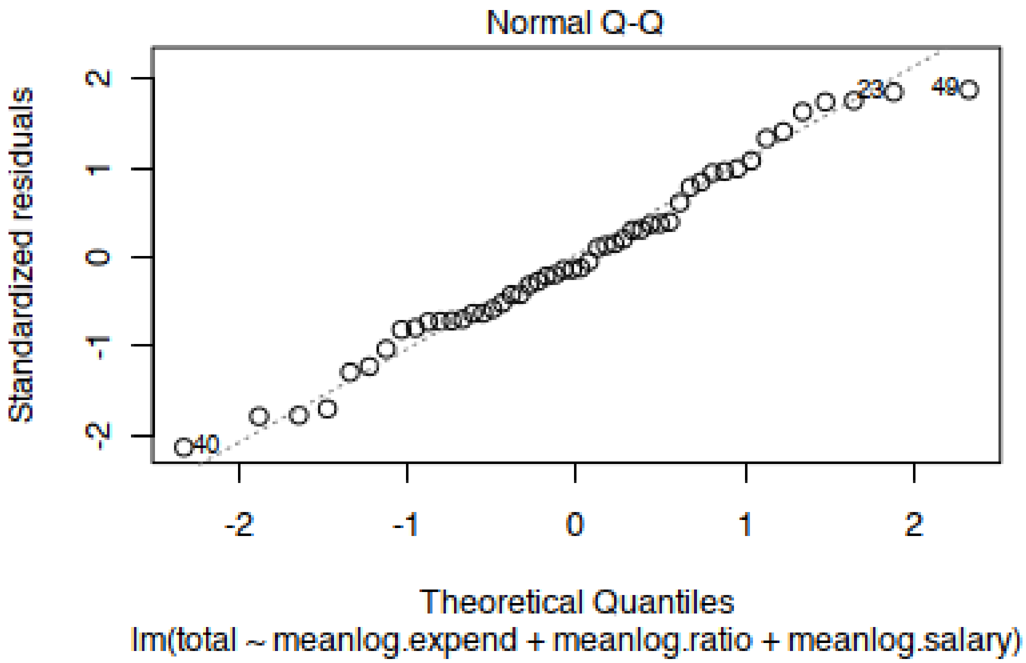

```
hist(log(sat$salary))
```


```
sat1 = mutate(sat,log.expend = log(expend))%>%mutate(log.ratio = log(ratio))%>%mutate(log.salary = log(s
satMD1 = lm(data = sat1, total ~ meanlog.expend + meanlog.ratio + meanlog.salary)
summary(satMD1)
```

```
##
## Call:
## lm(formula = total ~ meanlog.expend + meanlog.ratio + meanlog.salary,
##     data = sat1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -141.883  -45.280   -8.312   47.040  125.150
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     965.920      9.627 100.332   <2e-16 ***
## meanlog.expend   92.895    133.651   0.695   0.4905
## meanlog.ratio   117.352    121.224   0.968   0.3381
## meanlog.salary -311.093    161.183  -1.930   0.0598 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.08 on 46 degrees of freedom
## Multiple R-squared:  0.2229, Adjusted R-squared:  0.1722
## F-statistic: 4.397 on 3 and 46 DF,  p-value: 0.008403
```

```
plot(satMD1,which = 1:2)
```






intercept means average SAT score for average log.expend, average log.ratio and average log.salary
The coef of meanlog.expend for every unit increase in log expend, SAT scores would expect to increase 92.895.
The coef of meanlog.ratio means for every unit increase in log ratio, SAT scores would expect to increase 117.352
the coef of meanlog.salary means for every unit increse in log salary, SAT scores would expect to decrese 311.093

2. Construct 99% CI for each coefficient and discuss what you see.

```
satsum = as_tibble(summary(satMD1)$coef)%>%select(1:2)%>%`colnames<-`(c('Est','StandardError'))%>%mutat
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
kable(satsum,format = 'latex',digits = 2,align = 'c')
```

| | Est | StandardError | lower95CI | upper95CI |
|---|---|---|---|---|
| (Intercept) | 965.92 | 9.63 | 946.67 | 985.17 |
| meanlog.expend | 92.90 | 133.65 | -174.41 | 360.20 |
| meanlog.ratio | 117.35 | 121.22 | -125.10 | 359.80 |
| meanlog.salary | -311.09 | 161.18 | -633.46 | 11.27 |

the takeaway is that, there might not be an actual correlation between SAT scores and the predictors
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?

```
satMD2 = lm(data = sat1, total ~ meanlog.expend + meanlog.ratio + meanlog.salary + meantakers)
anova(satMD1,satMD2)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ meanlog.expend + meanlog.ratio + meanlog.salary
## Model 2: total ~ meanlog.expend + meanlog.ratio + meanlog.salary + meantakers
```

```
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
## 1     46 213174                                  
## 2     45  48030  1    165144 154.73 3.657e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(satMD1)
```

```
## 
## Call:
## lm(formula = total ~ meanlog.expend + meanlog.ratio + meanlog.salary, 
##     data = sat1)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -141.883  -45.280   -8.312   47.040  125.150 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     965.920      9.627 100.332   <2e-16 ***
## meanlog.expend   92.895    133.651   0.695   0.4905    
## meanlog.ratio   117.352    121.224   0.968   0.3381    
## meanlog.salary -311.093    161.183  -1.930   0.0598 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 68.08 on 46 degrees of freedom
## Multiple R-squared:  0.2229, Adjusted R-squared:  0.1722 
## F-statistic: 4.397 on 3 and 46 DF,  p-value: 0.008403
```

```r
summary(satMD2)
```

```
## 
## Call:
## lm(formula = total ~ meanlog.expend + meanlog.ratio + meanlog.salary + 
##     meantakers, data = sat1)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -92.613 -20.727   0.343  13.809  67.984 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    965.9200     4.6202 209.063  < 2e-16 ***
## meanlog.expend  15.9424    64.4382   0.247    0.806    
## meanlog.ratio  -79.9294    60.2999  -1.326    0.192    
## meanlog.salary  71.8433    83.2543   0.863    0.393    
## meantakers      -2.9199     0.2347 -12.439 3.66e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 32.67 on 45 degrees of freedom
## Multiple R-squared:  0.8249, Adjusted R-squared:  0.8093 
## F-statistic:    53 on 4 and 45 DF,  p-value: < 2.2e-16
```

clearly from the anova test, adding takers significantly improve the fit # 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, $D_i - R_i$
  - advantages includes
    - substraction provides quick classification from plus and minus sign.
    - know the exact difference
  - disadvantages includes
    - zero and negative value can be a problem for future transformation.
    - lose the knowledge of how close or how different the two party is(eg: 5500$ difference with each raised over 1mil vs. 500$ difference with each raised 5000ish)
- The ratio, $D_i/R_i$
  - advantages includes
    - quick classification comparing to '1'
    - always greater than zero, easy to transform
  - disadvantages includes
    - lose the exact amount of difference in raised money
    - when $R_i$ isa small, increases rapidly, not evenly distributed.
- The difference on the logarithmic scale, $logD_i - logR_i$
  - is esentailly taking log on $D_i/R_i$
  - resolve some distribution issue, but still lose the exact amount difference info.
- The relative proportion, $D_i/(D_i + R_i)$.
  - well organized, scale from 0 to 1.
  - but lose info on actual amount

**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 $\mathrm{x}^\star = 10(\mathrm{x} - 1)$?
   When $\mathrm{x}^\star = \mathrm{x} - 10$
   $$\alpha^\star = \alpha - 10 \cdot \beta$$
   $$\beta^\star = \beta$$
   $$\sigma^\star = \sigma$$
   When $\mathrm{x}^\star = 10\mathrm{x}$
   $$\alpha^\star = \alpha$$
   $$\beta^\star = \beta/10$$
   $$\sigma^\star = \sigma$$
   When $\mathrm{x}^\star = 10(\mathrm{x} - 1)$
   $$\alpha^\star = \alpha - \beta$$
   $$\beta^\star = \beta/10$$
   $$\sigma^\star = \sigma$$

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 $\mathrm{y}^{\star\star} = 5(\mathrm{y} + 2)$?
   When $\mathrm{y}^{\star\star} = a(\mathrm{y} + b), a \neq 0$

$$\alpha^{\star\star} = a(\alpha + b)$$

$$\beta^{\star\star} = a \cdot \beta$$

$$\sigma^{\star\star} = \sqrt{a} \cdot \sigma$$

3. In general, how are the results of a simple regression analysis affected by linear transformations of y and x? rules:
   multiplication on $x$ equals to divide the $\beta$ term, substraction of $x$ equals to add substraction $\times\beta$ to the intercept
   multiplication on $y$ equals to multiplication on all coeficients, substraction on $y$ equals to substract on intercept

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 $t_0^{\star} = \hat{\beta}^{\star}/SE(\hat{\beta}^{\star})$.

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 $t_0^{\star\star} = \hat{\beta}^{\star\star}/SE(\hat{\beta}^{\star\star})$.

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

# 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.

modeling took too much time to finish, and they are not really different from one another, conceptual questions are much more fun