Homework 02

Jinfei Xue

Septemeber 21, 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

round(confint(r),2)

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 weight.

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

```
require(foreign)
require(arm)
require(ggplot2)

#exclude NA's
heights <- na.omit(heights)

#scale earnings(divided by 1000)
heights$earn <- heights$earn / 1000
```

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?

```
# centralize "height"
heights$height <- heights$height - mean(heights$height)
```

```r
# Model 1
r_1 <- lm(earn ~ height, data=heights)
summary(r_1)
```

```
##
## Call:
## lm(formula = earn ~ height, data = heights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.031 -12.497  -3.215   7.474 174.659
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.0149     0.5077   39.42   <2e-16 ***
## height        1.5631     0.1334   11.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.85 on 1377 degrees of freedom
## Multiple R-squared:  0.09061,    Adjusted R-squared:  0.08995
## F-statistic: 137.2 on 1 and 1377 DF,  p-value: < 2.2e-16
```

In order to interpret the intercept as average earnings for people with average height, I perform the centralization on the independent variable "height", which makes heights equal to original heights data minus its mean.

According to the regression result, the average earnings for people with average height is 20.0149 thousand dollars.

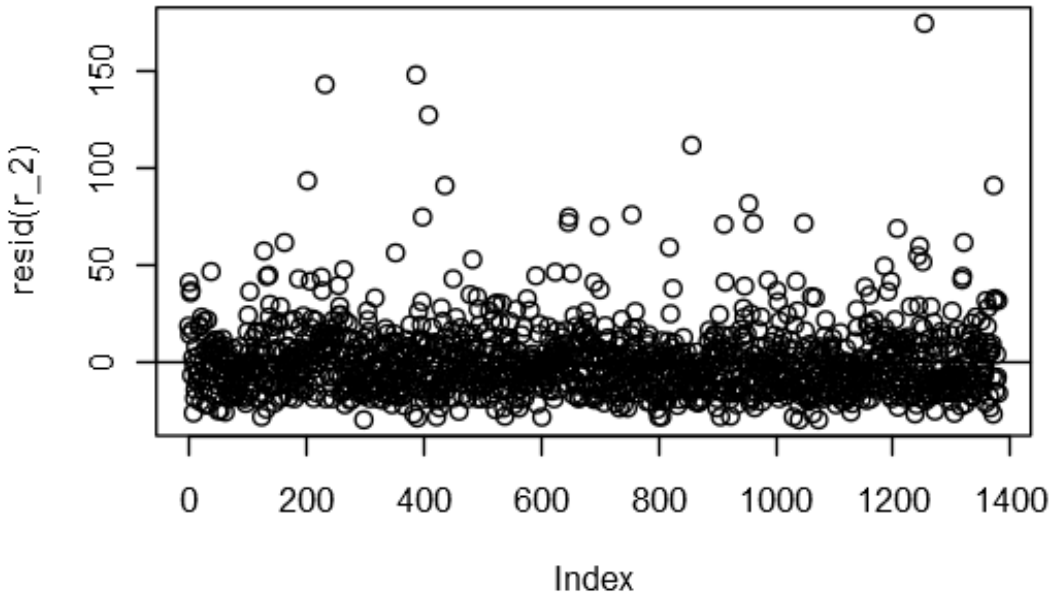
3. Fit some regression models with the goal of predicting earnings from some combination of sex, height, and weight. Be sure to try various transformations and interactions that might make sense. Choose your preferred model and justify.

```r
# Model 2: transformation by standardizing height
heights$height<-heights$height/sd(heights$height)
r_2 <- lm(earn ~ height, data=heights)
summary(r_2)
```

```
##
## Call:
## lm(formula = earn ~ height, data = heights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.031 -12.497  -3.215   7.474 174.659
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

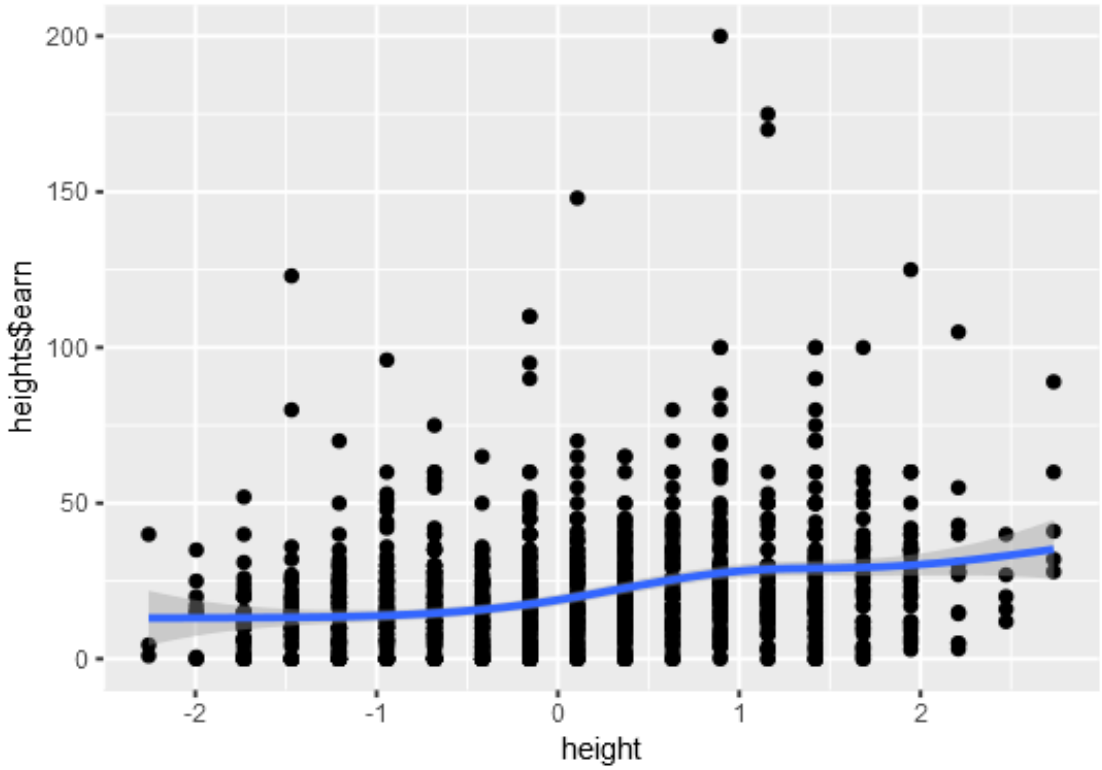
```
## (Intercept)  20.0149     0.5077   39.42   <2e-16 ***
## height        5.9493     0.5079   11.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.85 on 1377 degrees of freedom
## Multiple R-squared:  0.09061,    Adjusted R-squared:  0.08995
## F-statistic: 137.2 on 1 and 1377 DF,  p-value: < 2.2e-16
```

```
plot(resid(r_2))
abline(h=0)
```




```
ggplot(r_2)+aes(y=heights$earn,x=height)+geom_point(color="black")+geom
_smooth()
```

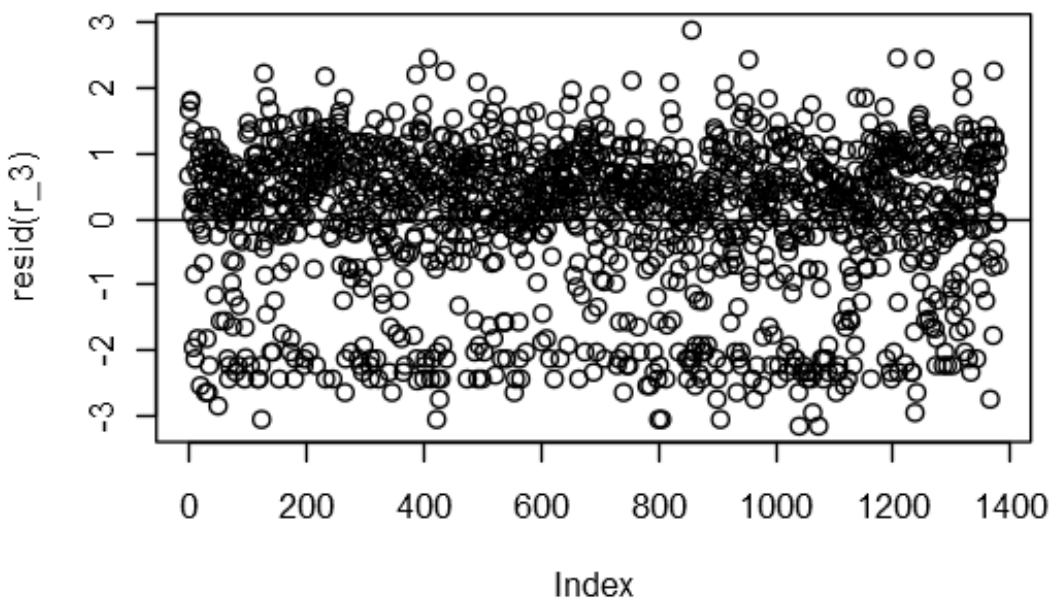
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")
'
```




```r
#Model 3:log transformation on log(earn+1)
r_3<-lm(log(earn+1)~height,data = heights)
summary(r_3)
```

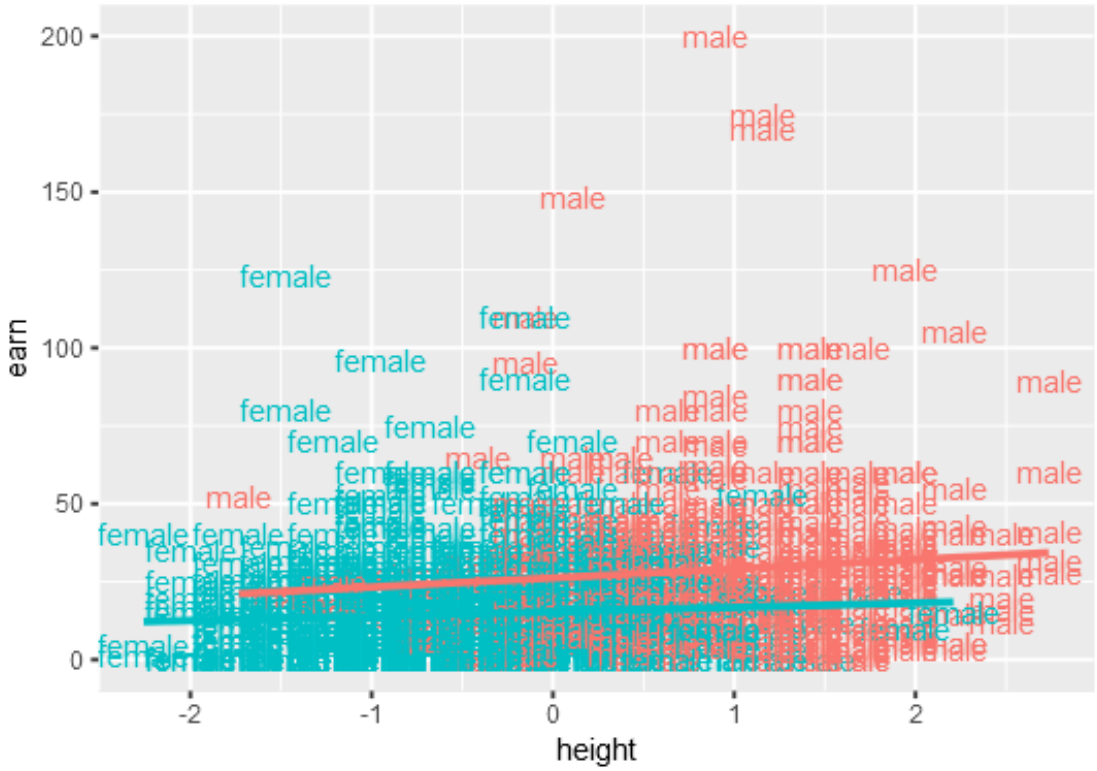
```
##
## Call:
## lm(formula = log(earn + 1) ~ height, data = heights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1661 -0.5302  0.3144  0.8429  2.8906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.50589    0.03181   78.77   <2e-16 ***
## height       0.39214    0.03182   12.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.181 on 1377 degrees of freedom
## Multiple R-squared:  0.09931,    Adjusted R-squared:  0.09866
## F-statistic: 151.8 on 1 and 1377 DF,  p-value: < 2.2e-16
```

```r
plot(resid(r_3))
abline(h=0)
```




```
#Model 4: combination of sex, height, and weight
heights$sex <- factor(heights$sex, labels=c("male", "female"))
r_4<-lm(earn~height+sex+height*sex, data = heights)
summary(r_4)

##
## Call:
## lm(formula = earn ~ height + sex + height * sex, data = heights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.209 -12.591  -3.172   7.223 171.109
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        26.259      1.248  21.041  < 2e-16 ***
## height              2.940      1.047   2.809  0.00505 **
## sexfemale         -10.868      1.492  -7.286 5.36e-13 ***
## height:sexfemale   -1.536      1.412  -1.088  0.27670
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.46 on 1375 degrees of freedom
## Multiple R-squared:  0.1296, Adjusted R-squared:  0.1277
## F-statistic: 68.22 on 3 and 1375 DF,  p-value: < 2.2e-16
```

```
ggplot(heights)+aes(x=height,y=earn,label=sex,color=sex)+theme(legend.p
osition="none")+
   geom_text()+geom_smooth(method="lm",se=FALSE)
```



From the regression result, we can see the coefficient of $height \cdot sex$ is not significant. Besides, the height versus earn plot shows the interaction does not fit very well.

According to the regression results of models shown above, I think Model 3 fit the data best because all of its coefficients are significant and the adjusted R-square is biggest.

4. Interpret all model coefficients.

For Model 3, the regression linear model is
$log(earn + 1) = 2.50589 + 0.39214 height.$

$beta_0 = 2.50589$ *means if z-score height value equals to 0, then the expected value of earn would be* $e^{(}2.50589) - 1$.

if we change z-score height value by 1 unit, we'd expect our earn variable to increase by 39.214 percent plus 0.39214 unit

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

```
confint(r_3,level = 0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 2.4434867 2.5683017
## height      0.3297131 0.4545734
```

we have 95% of confidence that the range [2.4434867,2.5683017] will include the intercept coefficient's true value.

we have 95% of confidence that that the range [0.3297131,0.4545734] will include the height coefficient's true value.

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:

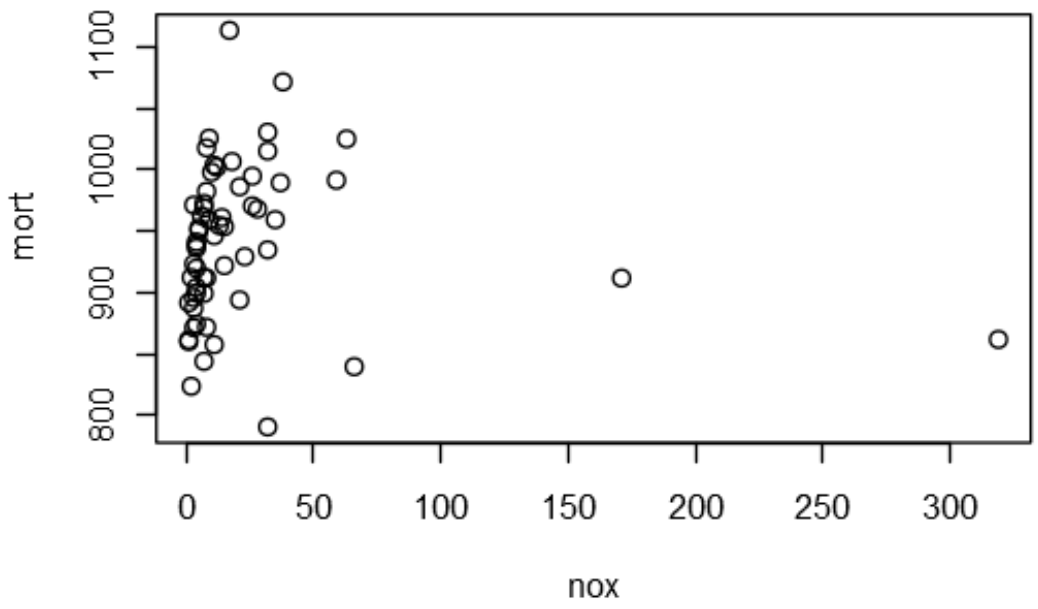
- PREC Average annual precipitation in inches
- JANT Average January temperature in degrees F
- JULT Same for July
- OVR65 % of 1960 SMSA population aged 65 or older
- POPN Average household size
- EDUC Median school years completed by those over 22
- HOUS % of housing units which are sound & with all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

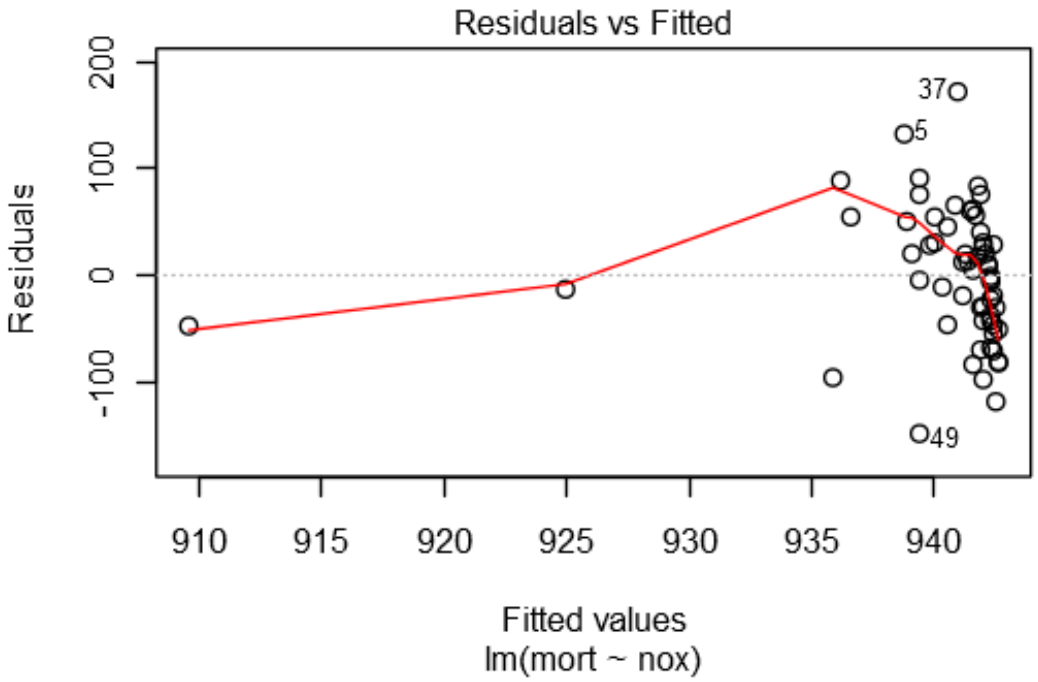
```r
#Model 1
mort<-pollution$mort
nox<-pollution$nox
#Create a scatterplot of mortality rate versus level of nitric oxides
plot(x=nox,y=mort)
```




```r
#Regression Model
r_1<-lm(mort~nox)
summary(r_1)
```

```
##
## Call:
## lm(formula = mort ~ 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 ***
## 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
```

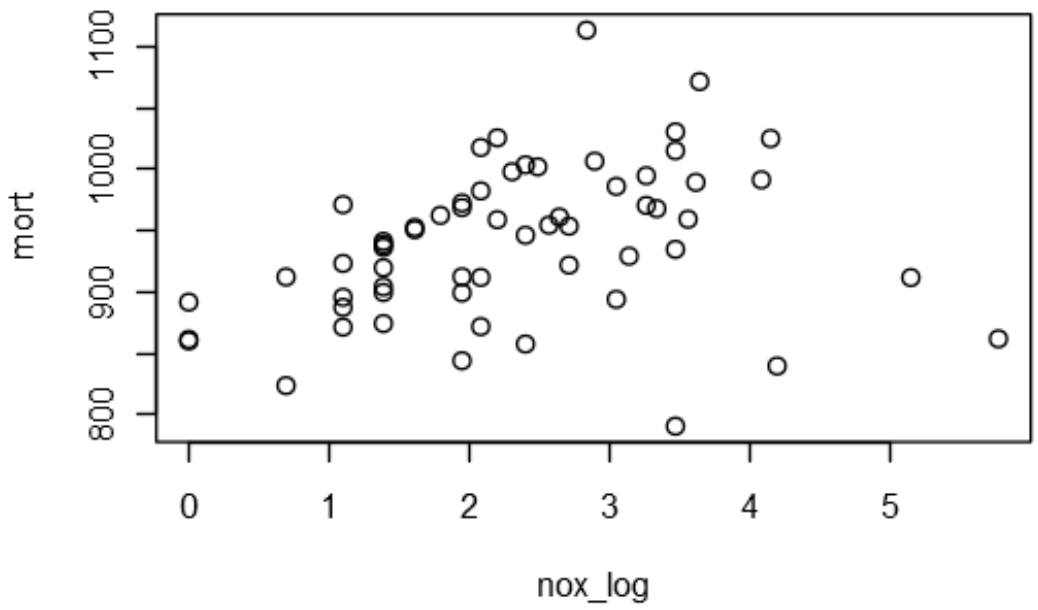
```
## Multiple R-squared:  0.005987,   Adjusted R-squared:  -0.01115
## F-statistic: 0.3494 on 1 and 58 DF,  p-value: 0.5568

#residual plot from the regression
plot(r_1,which=1)
```



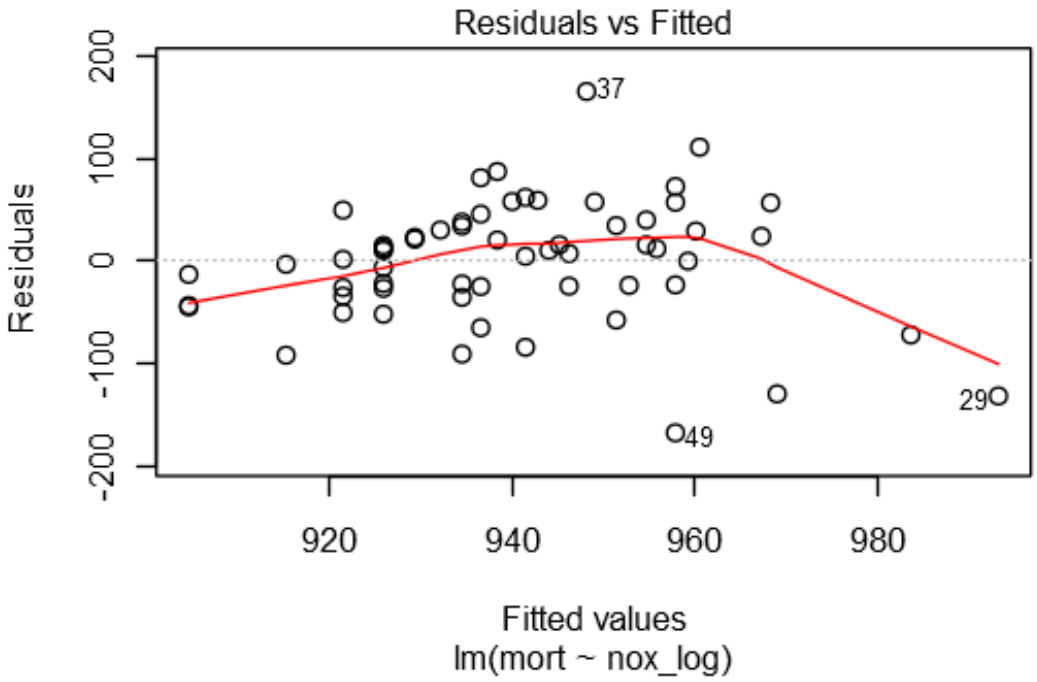
The regression does not fit the data well, because the first plot shows right-skewness and they do not have linear relationship. Besides, the second plot shows residuals are not evenly distributed around the dotted line.

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.

```
#Model 2
nox_log<-log(nox)
plot(nox_log,mort)
```




```
r_2<-lm(mort~nox_log)
summary(r_2)

##
## Call:
## lm(formula = mort ~ nox_log)
##
## 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 ***
## nox_log       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

plot(r_2,which=1)
```



In this case, the residuals evenly distributed on both sides of dotted line in the scatter plot, so Model 2 is more appropriate than Model 1.

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

According to the regression result, the linear regression model is $mort = 904.724 + 15.335log(nox)$.

The slope coefficient means if we increase nox by one percent, we expect mort to increase by 0.15335 units of mort.

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

```
confint(r_2,level = 0.99)
```

```
##                 0.5 %    99.5 %
## (Intercept) 858.988556 950.46037
## nox_log      -2.230963  32.90196
```

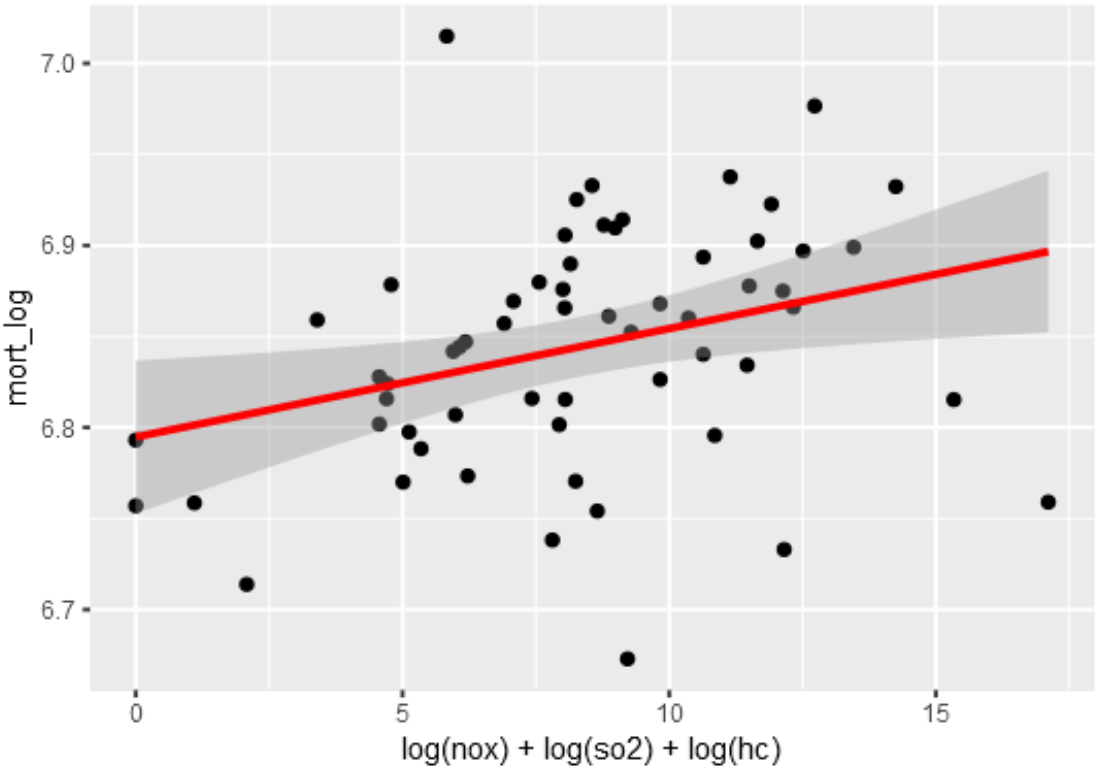
**we have 99% of confidence that that the range [-2.230963,32.90196] will include the slope coefficient's true value.*

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
so2<-pollution$so2
hc<-pollution$hc
mort_log<-log(mort)
##Model 3
r_3<-lm(mort_log~log(nox)+log(so2)+log(hc))
summary(r_3)
```

```
##
## Call:
## lm(formula = mort_log ~ log(nox) + log(so2) + log(hc))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10874 -0.03574 -0.00218  0.03709  0.20085
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.826749   0.022701 300.726  < 2e-16 ***
## log(nox)     0.059837   0.023021   2.599  0.01192 *
## log(so2)     0.014309   0.007584   1.887  0.06436 .
## log(hc)     -0.060812   0.020553  -2.959  0.00452 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05753 on 56 degrees of freedom
## Multiple R-squared:  0.2852, Adjusted R-squared:  0.2469
## F-statistic: 7.449 on 3 and 56 DF,  p-value: 0.0002777
```

```r
#Plot the fitted regression model
library(ggplot2)
ggplot(r_3) + aes(x=log(nox)+log(so2)+log(hc),y=mort_log) +
  geom_point() + stat_smooth(method = "lm",col = "red")
```



According to the regression result, the linear regression model is $log(mort) = 6.826749 + 0.059837log(nox) + 0.014309log(so2) - 0.060812log(hc)$.

1> The intercept coefficient means when nox=so2=hc=1, the expected value of mort would be e^6.826749.

2> With the same so2 and hc levels, if we increase nox by one percent, we'd expect mort to increase by 0.059837 percent.

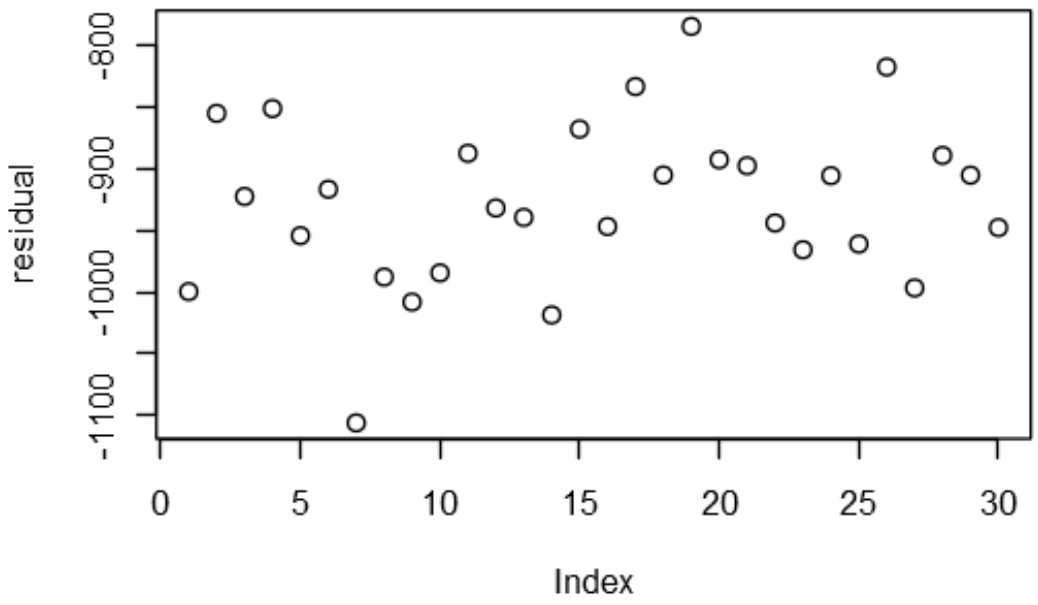
3> With the same nox and hc levels, if we increase so2 by one percent, we'd expect mort to increase by 0.014309 percent.

4> With the same nox and so2 levels, if we increase hc by one percent, we'd expect mort to decrease by 0.060812 percent.

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.)

```
dim(pollution)

## [1] 60 16

data1<-pollution[1:30,]
data2<-pollution[31:60,]
r_data1<-lm(log(mort)~log(nox)+log(so2)+log(hc),data=data1)
prediction<-predict(object = r_data1,newdata =
```

```
                        data.frame(nox=data2$nox,so2=data2$so2,hc=data2$h
c),
                      interval= "prediction")
#the difference between the actual values and the predicted values
residual<-prediction[,1]-pollution[31:60,]$mort
plot(residual)
```



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.

```
gamble_log<-log(teengamb$gamble+1)
sex<-teengamb$sex
status_zscore<-(teengamb$status-mean(teengamb$status))/sd(teengamb$stat
us)
income<-teengamb$income
verbal<-teengamb$verbal
r_1<-lm(gamble_log~sex+status_zscore+income+verbal)
summary(r_1)

##
## Call:
```

```
## lm(formula = gamble_log ~ sex + status_zscore + income + verbal)
##
## 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)    3.06554    0.74198   4.132 0.000168 ***
## sex           -0.87120    0.39268  -2.219 0.031975 *
## status_zscore  0.51496    0.23208   2.219 0.031951 *
## income         0.21565    0.04904   4.398 7.33e-05 ***
## verbal        -0.26165    0.10388  -2.519 0.015673 *
## ---
## 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
```

According to the regression result, the linear regression model is $log(gamble + 1) = 3.06554 - 0.87120sex + 0.51496status_zscore + 0.21565income - 0.26165verbal$.

1> The intercept coefficient means when the teenager is male, his socioeconomic status score equals to the mean of status score, his income and verbal score are zero, the expected value of his expenditure on gambling in pounds per year would be e^3.06554-1.

2> With the same status score, income and verbal score levels, the expected value of expenditure on gambling in pounds per year plus 1 of a female would be 87.120 percent less than that of a male.

3> With the same sex, income and verbal score levels, if we increase status score by one unit, we'd expect expenditure on gambling in pounds per year plus 1 to increase by 51.496 percent.

4> With the same sex, status score and verbal score levels, if we increase income by one unit, we'd expect expenditure on gambling in pounds per year plus 1 to increase by 21.565 percent.

5> With the same sex, status score and income levels, if we increase verbal score by one percent, we'd expect expenditure on gambling in pounds per year plus 1 to decrease by 26.165 percent.

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

```
confint(r_1,level = 0.95)
```

```
##                    2.5 %      97.5 %
## (Intercept)    1.56816814  4.56290788
```

```
## sex           -1.66365707 -0.07873377
## status_zscore  0.04660771  0.98330592
## income         0.11668468  0.31460764
## verbal        -0.47128110 -0.05200895
```

1> We have 95% of confidence that the range [1.56816814,4.56290788] will include the intercept coefficient's true value.

2> We have 95% of confidence that the range [-1.66365707,-0.07873377] will include the sex coefficient's true value.

3> We have 95% of confidence that the range [0.04660771,0.98330592] will include the z-score status coefficient's true value.

4> We have 95% of confidence that the range [0.11668468,0.31460764] will include the income coefficient' true value.

5> We have 95% of confidence that the range [-0.47128110,-0.05200895] will include the verbal coefficient's true value.

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?

```
#calculate CI of average value
prediction_average<-predict(object = r_1,newdata=
  data.frame(sex=0,status_zscore=0,income=mean(teengamb$income),
             verbal=mean(teengamb$verbal)),level = 0.95,interval = "pre
diction")
CI_1<-prediction_average[3]-prediction_average[2]

#calculate CI of maximal value
prediction_max<-predict(object = r_1,newdata=
  data.frame(sex=0,status_zscore=max(status_zscore),
             income=max(teengamb$income),verbal=max(teengamb$verbal)),
             level = 0.95,interval = "prediction")
CI_2<-prediction_max[3]-prediction_max[2]

#Compare CI_1 with CI_2
CI_1<CI_2

## [1] TRUE
```

The result of logical code is "TRUE". That is to say, CI of the amount that a male with maximal values of status, income and verbal score would gamble is wider than that of a male with average values of status, income and verbal score would gamble.

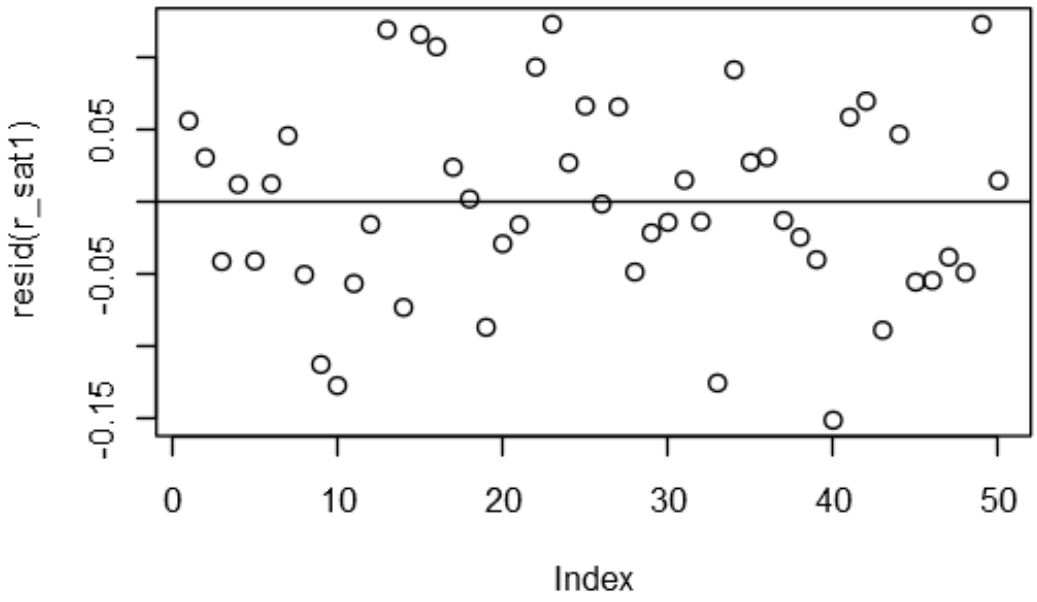
The reason is that the length of CI is $2 * \frac{s}{\sqrt{n}} * t_{\frac{\alpha}{2}}$*. For a male with maximal values of status, income and verbal score, the standard error s is larger than that of a male with average status, income and verbal score.*

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.

```
total<-sat$total
expend_zscore<-(sat$expend-mean(sat$expend))/sd(sat$expend)
ratio_zscore<-(sat$ratio-mean(sat$ratio))/sd(sat$ratio)
salary_zscore<-(sat$salary-mean(sat$salary))/sd(sat$salary)
r_sat1<-lm(log(total)~expend_zscore+ratio_zscore+salary_zscore,data = s
at)
summary(r_sat1)

##
## Call:
## lm(formula = log(total) ~ expend_zscore + ratio_zscore + salary_zsco
re,
##     data = sat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.151140 -0.046616 -0.006997  0.046837  0.123402
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.87016    0.01001 686.348   <2e-16 ***
## expend_zscore  0.02391    0.03098   0.772   0.4442
## ratio_zscore   0.01540    0.01529   1.008   0.3189
## salary_zscore -0.05443    0.02877  -1.892   0.0648 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07078 on 46 degrees of freedom
## Multiple R-squared:  0.2082, Adjusted R-squared:  0.1566
## F-statistic: 4.032 on 3 and 46 DF,  p-value: 0.01256

plot(resid(r_sat1))
abline(h=0)
```




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

(Intercept) 6.87016 0.01001 686.348 <2e-16 *** expend_zscore 0.02391 0.03098 0.772 0.4442
ratio_zscore 0.01540 0.01529 1.008 0.3189
salary_zscore -0.05443 0.02877 -1.892 0.0648 .

According to the regression result, the linear regression model is $log(total) = 6.87016 + 0.02391 expend_z score + 0.01540 ratio_z score - 0.05443 salary_z score.$

1> The intercept coefficient means when the expenditure, ratio and salary of a student equal to their corresponding means, the expected value of the student's total score in SAT would be e^6.87016.

2> With the same ratio and salary levels, if we increase expend_zscore by one unit, we'd expect the student's total score in SAT to increase by 2.391 percent.

3> With the same expenditure and salary levels, if we increase ratio_zscore by one unit, we'd expect the student's total score in SAT to increase by 1.54 percent.

4> With the same expenditure and ratio levels, if we increase salary_zscore by one unit, we'd expect the student's total score in SAT to decrease by 5.443 percent.

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

```
confint(object = r_sat1,level = 0.98)
```

```
##                       1 %       99 %
## (Intercept)    6.84603569 6.89428636
## expend_zscore -0.05075940 0.09857730
## ratio_zscore  -0.02143981 0.05224324
## salary_zscore -0.12377316 0.01490439
```

1> We have 98% of confidence that the range [6.84603569,6.89428636] will include the intercept coefficient's true value.

2> We have 98% of confidence that the range [-0.05075940,0.09857730] will include the expend_zscore coefficient's true value.

3> We have 98% of confidence that the range [-0.02143981,0.05224324] will include the ratio_zscore status coefficient's true value.

4> We have 98% of confidence that the range [-0.12377316,0.01490439] will include the salary_zscore coefficient' true value.

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?

```
takers_zscore<-(sat$takers-mean(sat$takers))/sd(sat$takers)
r_sat2<-lm(log(total)~expend_zscore+ratio_zscore+salary_zscore+takers_z
score,data = sat)
summary(r_sat2)
```

```
##
## Call:
## lm(formula = log(total) ~ expend_zscore + ratio_zscore + salary_zsco
re +
##     takers_zscore, data = sat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.091157 -0.023196 -0.000844  0.015822  0.070993
##
## Coefficients:
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)    6.870161   0.004682 1467.352   <2e-16 ***
## expend_zscore  0.006954   0.014551    0.478    0.635
## ratio_zscore  -0.007976   0.007377   -1.081    0.285
## salary_zscore  0.009965   0.014359    0.694    0.491
## takers_zscore -0.080545   0.006266  -12.855   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03311 on 45 degrees of freedom
## Multiple R-squared:  0.8305, Adjusted R-squared:  0.8155
## F-statistic: 55.13 on 4 and 45 DF,  p-value: < 2.2e-16
```

This model is better compared to the previous one. The reasons are as follows:

1> The R squared of the second model is much more larger.

2> The coefficient of takers_zscore we added is a statistically significant.

3> The p-value of F statistics in the second value is much more smaller, which means the total independent variables in the second model can explain the dependent variable better.

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: This measurement is easy to interpret, we can just say one unit of the numerical difference between two parties' amount will lead to how many unit of difference.

Disadvantages: This doesn't take difference proportion into account.

- The ratio, D_i/R_i

Advantages: It is easy to interpret how the ratio of money raised by two candidates can effect the result.

Disadvantages: It can only indicate the effect of ratio. For example, when $D_i = 200, R_i = 100$*, the result is same with the situation in which* $D_i = 400, R_i = 200$*.*

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

Advantages: This measurement gives the inforamtion about proportion.

Disadvantages: This measurement makes the model less interpretable.

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

Advantages: The total amount is taken into account. It gives us an idea of the influence of the percentage of D_i *in total amount.*

Disadvantages: This measurement makes the model difficult to explain.

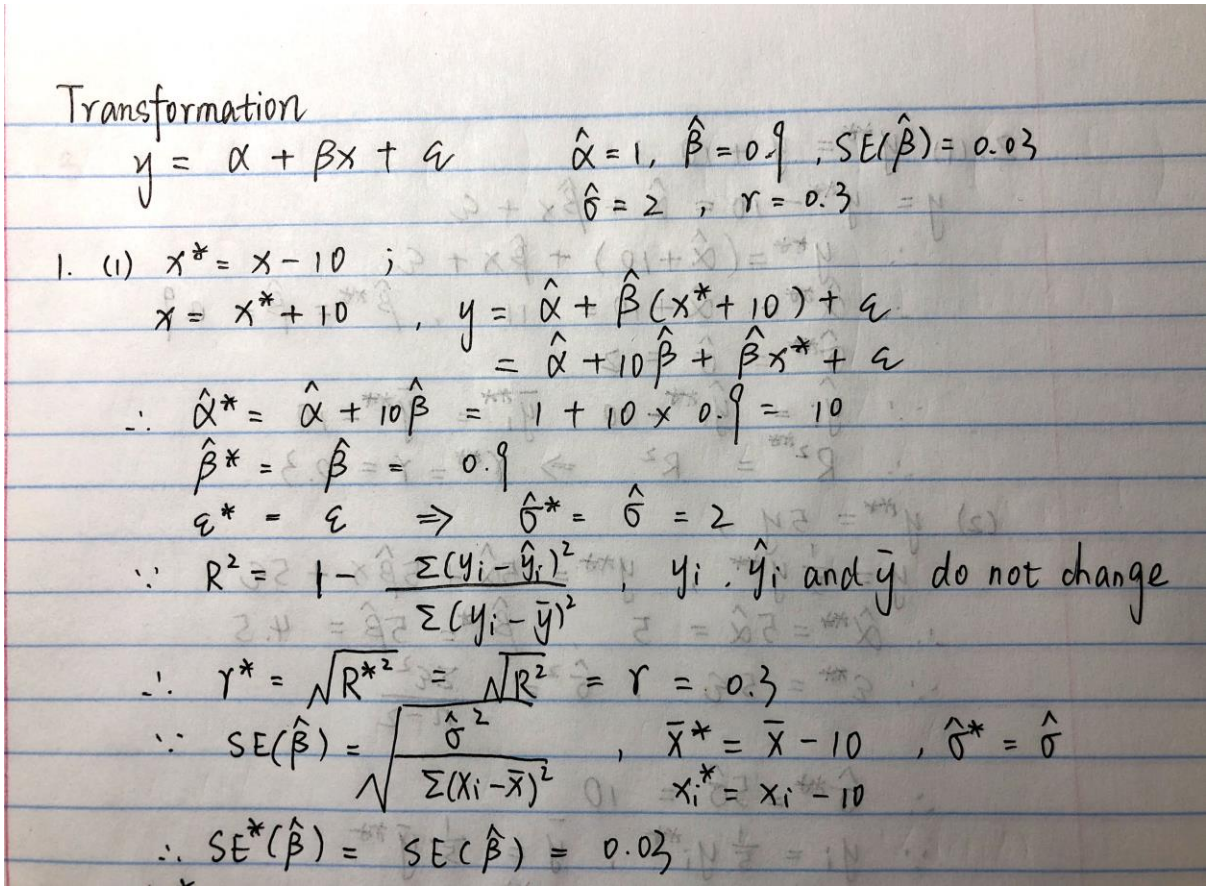
Transformation

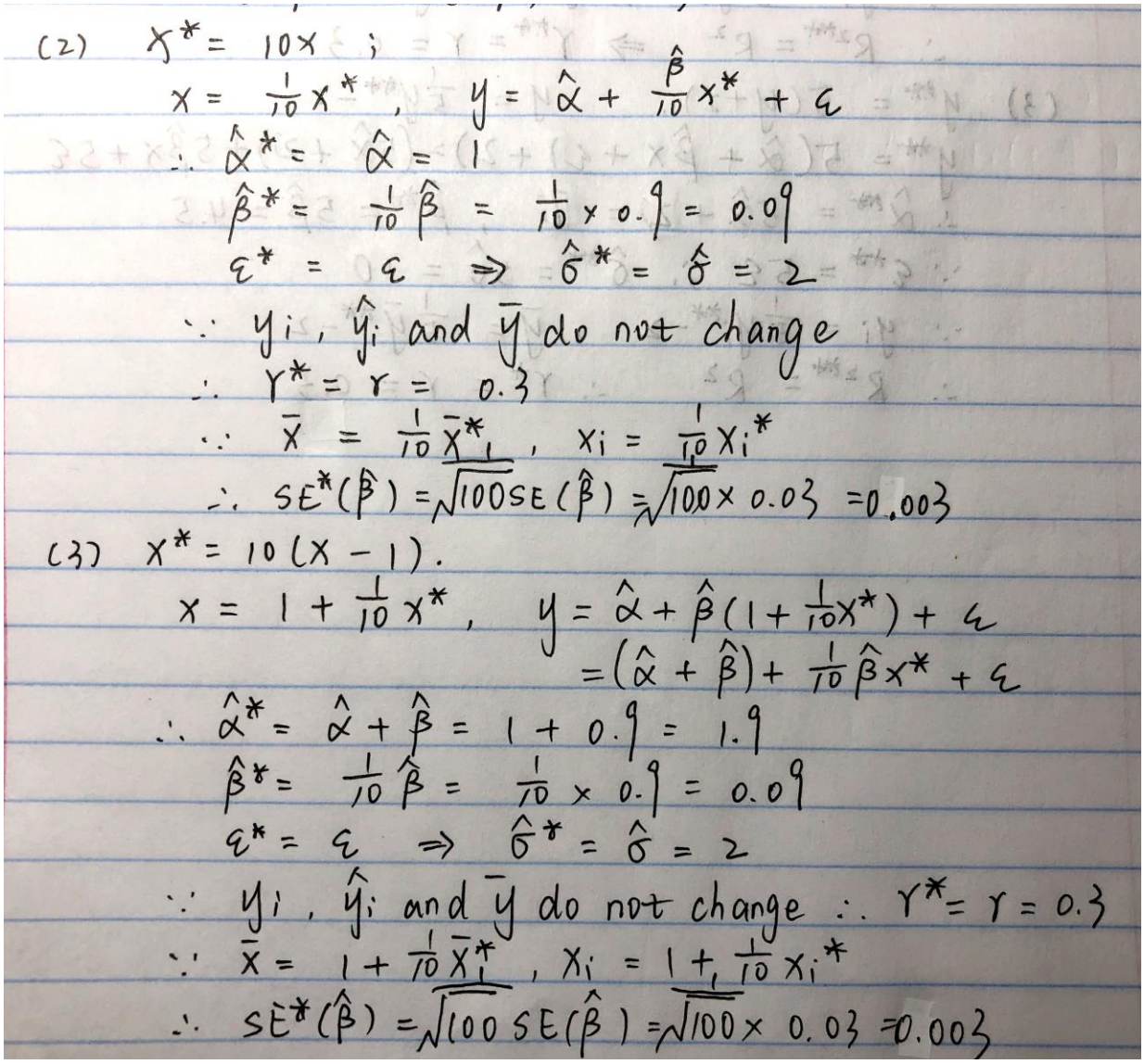
For observed pair of x and y, we fit a simple regression model

$$y = \alpha + \beta x + \epsilon$$

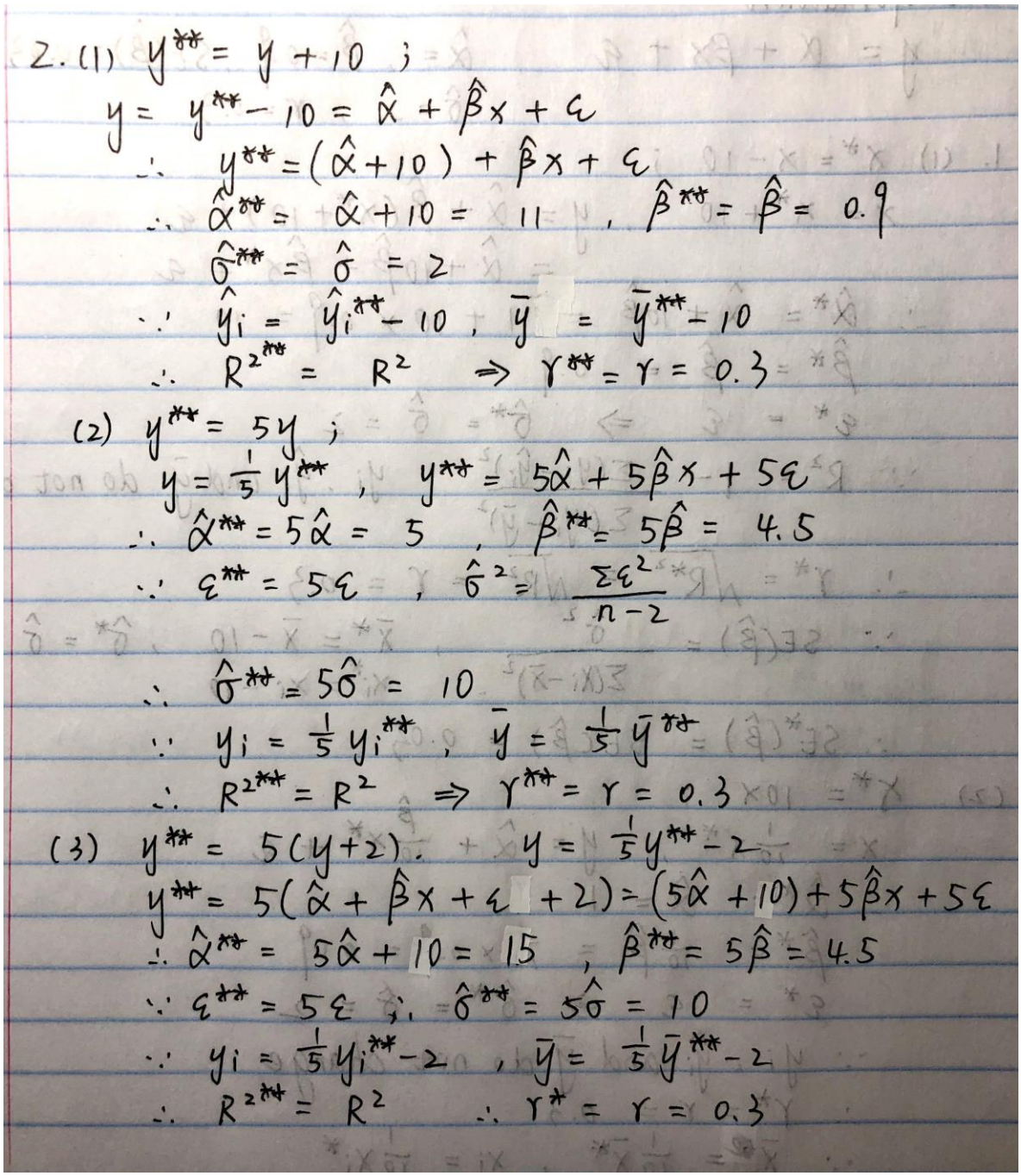
which results in estimates $\hat{\alpha} = 1, \hat{\beta} = 0.9, SE(\hat{\beta}) = 0.03, \hat{\sigma} = 2$ and $r = 0.3$.

1. Suppose that the explanatory variable values in a regression are transformed according to the $x^\star = x - 10$ and that y is regressed on $x^\star$. 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 $x^\star = 10x$? When $x^\star = 10(x - 1)$?





2. Now suppose that the response variable scores are transformed according to the formula $y^{**} = y + 10$ and that y^{**} is regressed on x. Without redoing the regression calculation in detail, find $\hat{\alpha}^{**}$, $\hat{\beta}^{**}$, $\hat{\sigma}^{**}$, and r^{**}. What happens to these quantities when $y^{**} = 5y$? When $y^{**} = 5(y + 2)$?



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

1> linear transformations of x *do not affect* ϵ *and* R^2

x + *c will result in the intercept change to* $\hat{\alpha} - c\,\hat{\beta}$, *but* $\hat{\beta}$ *do not change.*

x · *d will result in the* $\hat{\beta}$ *change to* $\hat{\beta}\,/d$, *but* $\hat{\alpha}$ *do not change.*

2> linear transformations of y *do not affect* R^2, *but affect* ϵ.

$y + c$ *will result in the intercept change to* $\hat{\alpha} + c$*, but* $\hat{\beta}$ *do not change.*

$y \cdot d$ *will result in the* $\hat{\alpha}$ *change to* $\hat{\alpha} \cdot d$*, and* $\hat{\beta}$ *will change to* $\hat{\beta} \cdot d$*,and* $\hat{\sigma}$ *will change to* $\hat{\sigma} \cdot d$

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

$$SE(\hat{\beta}^{*}) = SE(\hat{\beta})/10 = 0.003$$

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

$$t_0^{*} = t_0 = 30$$

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

$$\hat{\beta}^{**} = 5\,\hat{\beta} = 4.5$$

$$SE(\hat{\beta}^{*}) = 5SE(\hat{\beta}) = 0.15$$

$$t_0^{**} = t_0 = 30$$

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

1> In hypothesis test, H_0:$\mu = 0$, H_1:$\mu \neq 0$

$t = \frac{\overline{\beta}}{SE(\beta)} \sim t(n-1)$

According to the conclusion in last question, We can say that linear transformations on x or y will not change t value so that they will not change the result of test.

2> confidence intervals

$\frac{\overline{\beta} - \mu_0}{SE(\beta)} \sim t(n-1)$

Confidence Interval is $[\overline{\beta} - t_{\alpha/2} * SE(\beta), \overline{\beta} + t_{\alpha/2} * SE(\beta)]$

If $x^{=} cx$, then $\overline{\beta^{*}} = \overline{\beta}/c$, CI is $[\overline{\beta}/c - t_{\alpha/2} * SE(\beta)/c, \overline{\beta}/c + t_{\alpha/2} * SE(\beta)/c]$

If $y^{=} dy$, then $\overline{\beta^{*}} = \overline{\beta} * d$, CI is $[\overline{\beta} * d - t_{\alpha/2} * SE(\beta) * d, \overline{\beta} * d + t_{\alpha/2} * SE(\beta) * d]$

Therefore, the linear transformations will on x or y will change confidence intervals.

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.