

Homework 02

yourname

Septemeber 16, 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

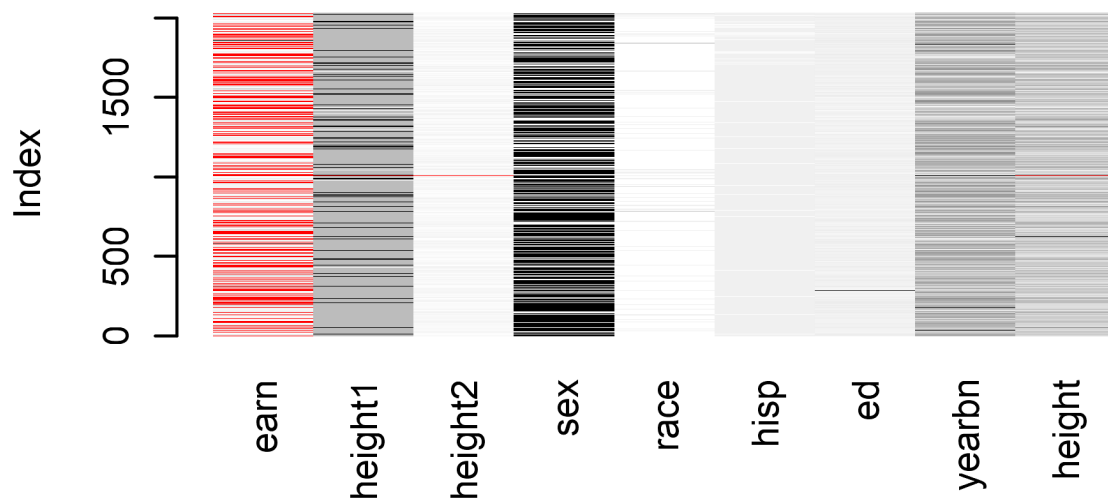
The folder **earnings** has data from the Work, Family, and Well-Being Survey (Ross, 1990). You can find the codebook at <http://www.stat.columbia.edu/~gelman/arm/examples/earnings/wfwcodebook.txt>

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

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

Table 1: Quality of Data

	earn	height1	height2	sex	race	hisp	ed	yearbn	height
Number of Unique Value	135.00	4	14.0	2	5	3	19	74	24
Number of Missing Value	650.00	8	6.0	0	0	0	0	0	8
Number of 0	187.00	0	194.0	0	0	0	0	3	0
% of Missing Value	0.41	0	0.1	0	0	0	0	0	0

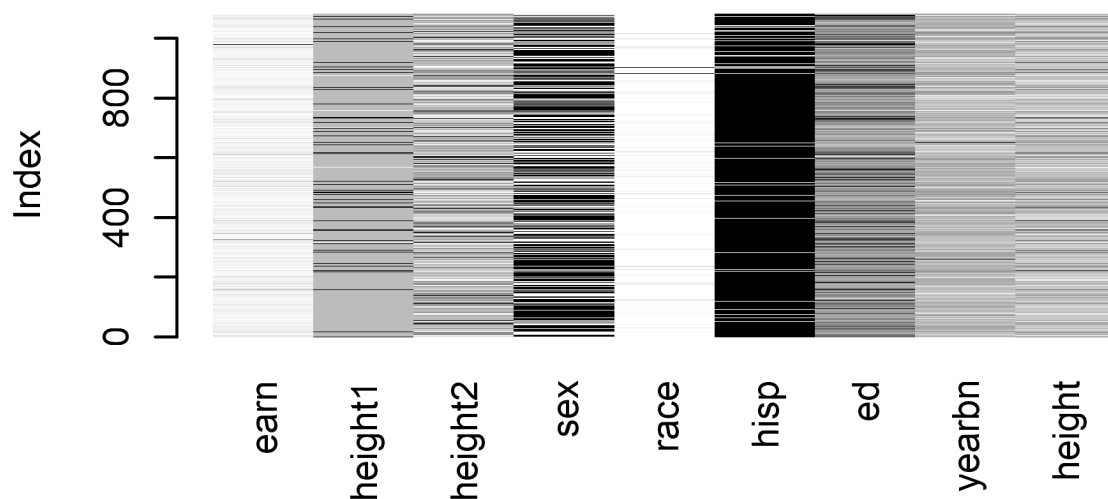


Above chart and figure show the distribution of missing value and 0 among variables, in the figure, the red means there is a missing value, and for the rest, the darker the color, the higher the value.

So according to the analysis above, the most of the missing value ,which is represented by ‘NA’ and ‘0’, contained in variable “earning”, about 41% of data doesn’t have variable ‘earning’, which is reasonable, beacause earning is kond of a private quetion, due to the extent of missing, we can hardly apply any imputaion, so simply delete it.

Table 2: Quality of Data

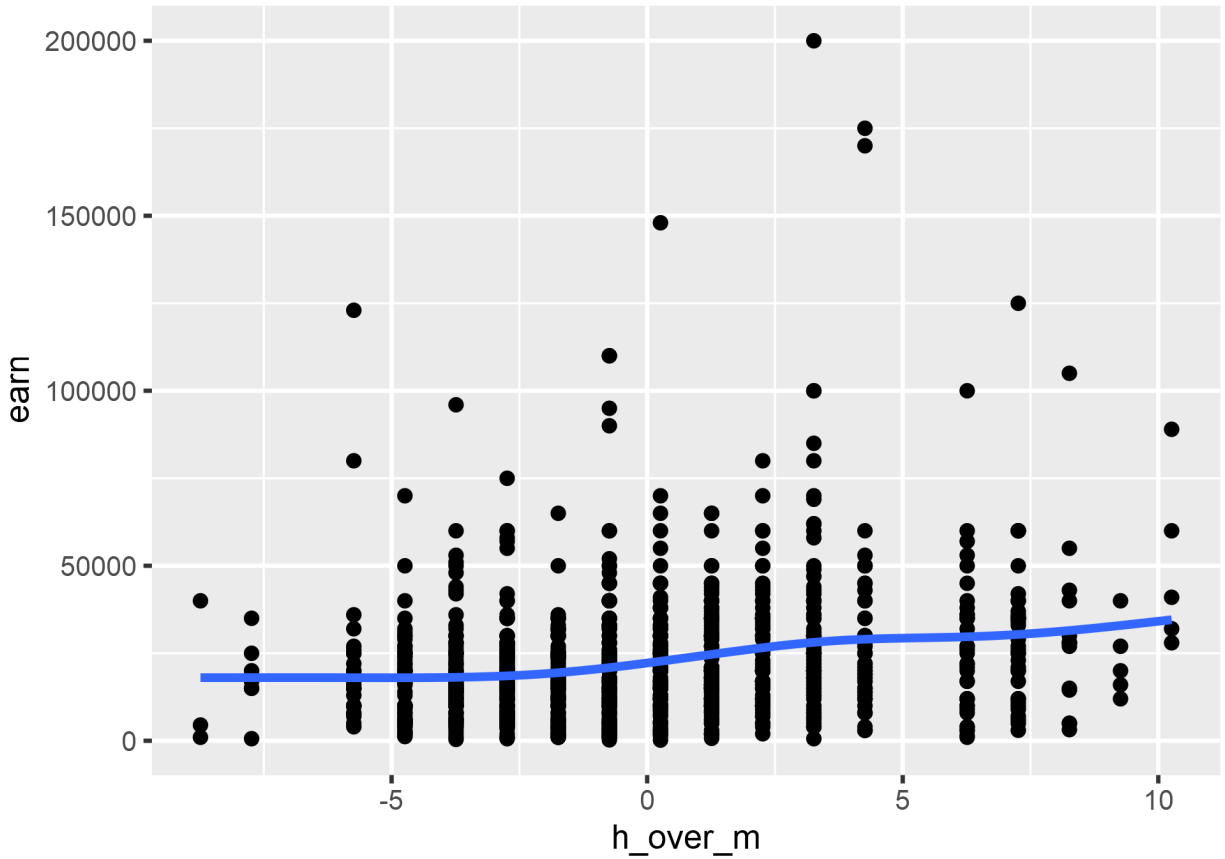
	earn	height1	height2	sex	race	hisp	ed	yearbn	height
Number of Unique Value	130	3	11	2	5	2	16	71	18
Number of Missing Value	0	0	0	0	0	0	0	0	0
Number of 0	0	0	0	0	0	0	0	0	0
% of Missing Value	0	0	0	0	0	0	0	0	0



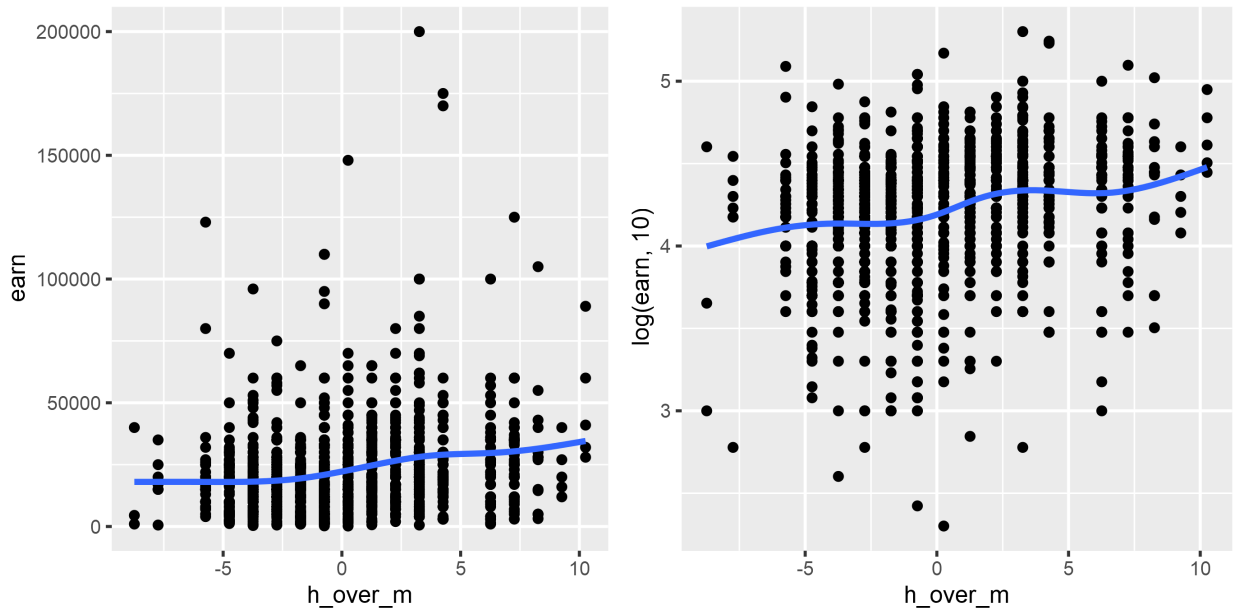
So now the data is clean and ready to be analysed.

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?

Look at the data, we can find three variables that related to height, “height 1”, “height 2” and “height”, after reading the original article, we know that real height should be “hight 1” feet and “height 2” inches, so $RealHeight = 12 * height_1 + height_2$, which is exactly the variable “height”.



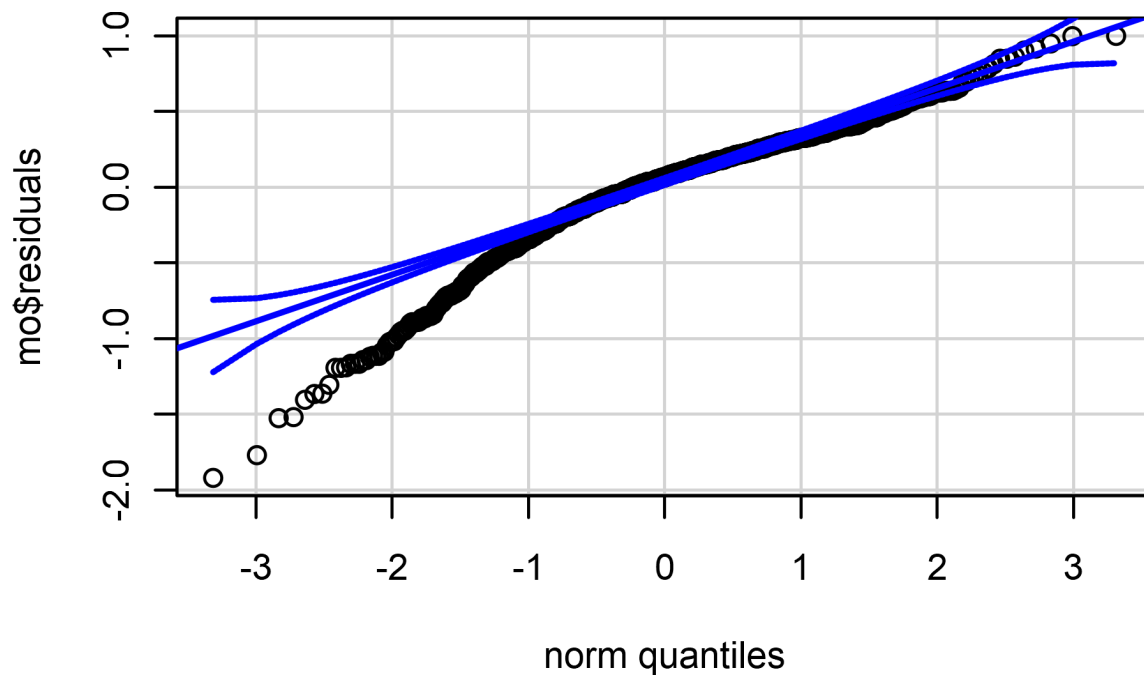
Look at the plot of height over average against earning, we can easily find some outliers, and a linear relationship may not apply to this situation, so we apply some non-linear transformation, first we consider use $\log(\text{earn})$ replace earn to eliminate the effect of outliers. As for the predictor, we try to include the polynomial in our model.



From above plot, we can tell that replacing earn by $\log_{10} \text{Earn}$ does condense the data and alleviate

the effect of outliers a bit. Then apply regression analysis.

```
##
## Call:
## lm(formula = log(earn, 10) ~ h_over_m + I(h_over_m^2) + I(h_over_m^3),
##     data = hi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.9194 -0.1696  0.0658  0.2459  0.9998
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.2130970   0.0153305  274.818 < 2e-16 ***
## h_over_m       0.0281404   0.0057486   4.895 1.13e-06 ***
## I(h_over_m^2)  0.0003175   0.0008552   0.371  0.710
## I(h_over_m^3) -0.0001484   0.0001499  -0.990  0.322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.386 on 1077 degrees of freedom
## Multiple R-squared:  0.04665,    Adjusted R-squared:  0.044
## F-statistic: 17.57 on 3 and 1077 DF,  p-value: 3.872e-11
```

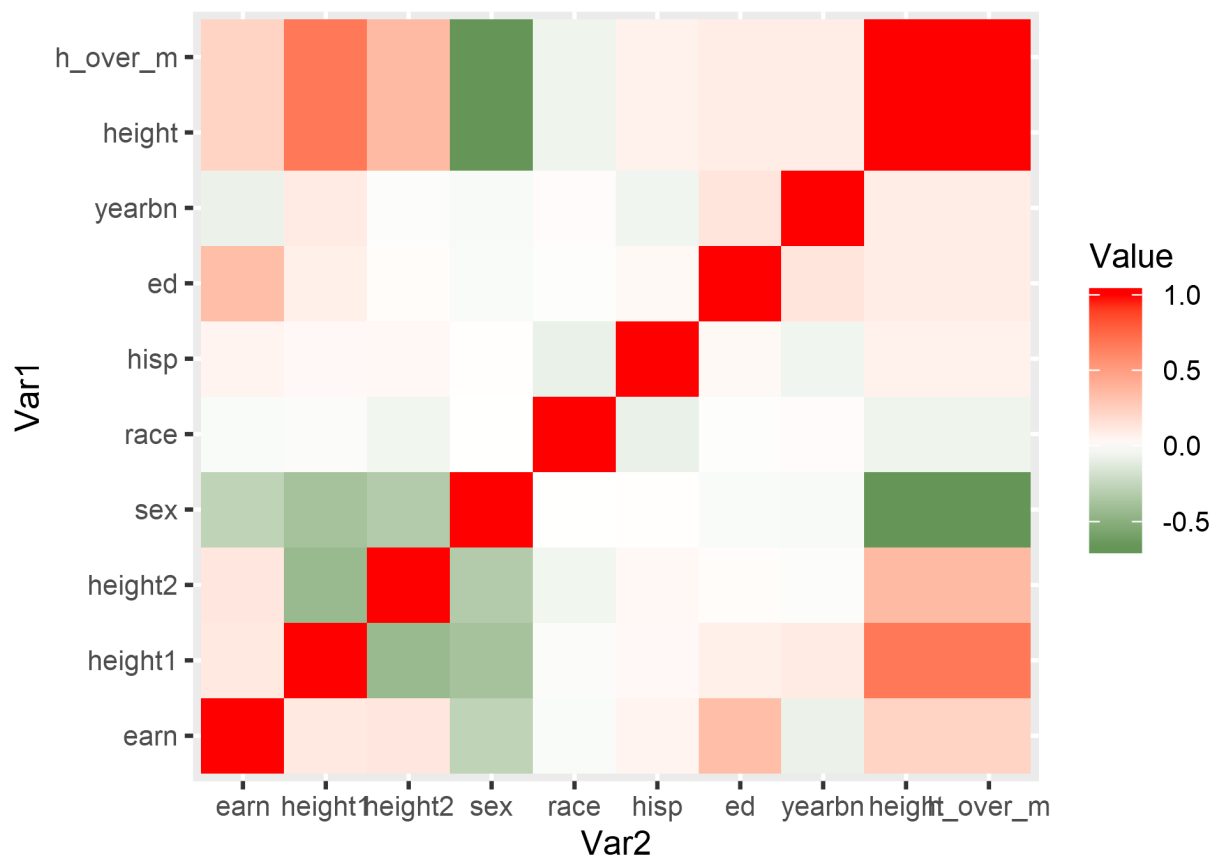


From result above, we can see that the performance of the model is poor, whose residuals are not normally distributed and even not close. I tried different combination of predictors, including heighs and the

power of height. The coefficient of model shows that only the coefficient of height is significant, Which means one unit increase(1 inch) will increase the earn by around 2.8%.

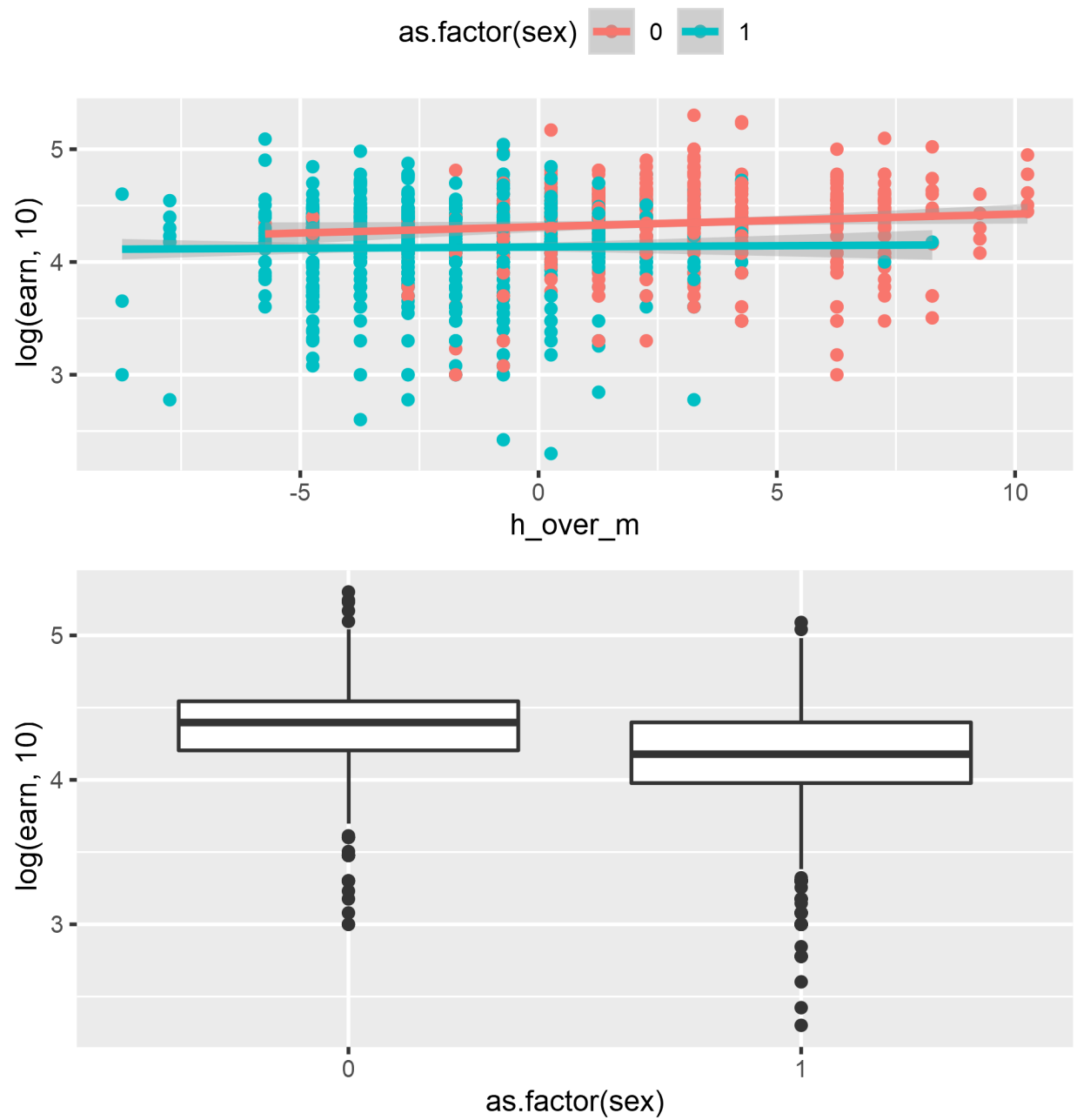
3. Fit some regression models with the goal of predicting earnings from some combination of sex, height, and weight. Be sure to try various transformations and interactions that might make sense. Choose your preferred model and justify.

To select proper predictors, we first do some analysis.

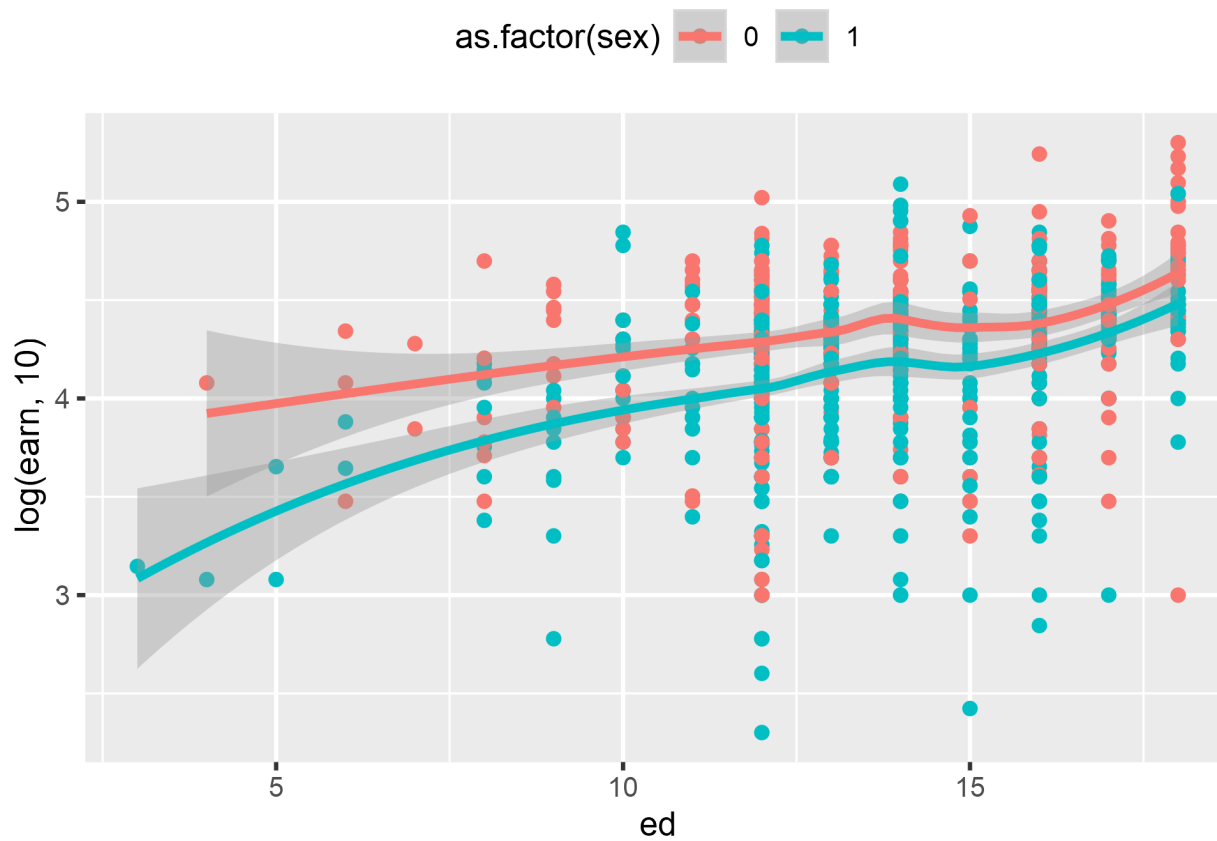


We can tell that sex is highly related to height, earn is related to earn, race is not obviously related to any other variables. So we may consider put sex, ed into our model.

Before we analyze sex, we have to transform sex into a binary variable, where 0 represent male, 1 represent female.

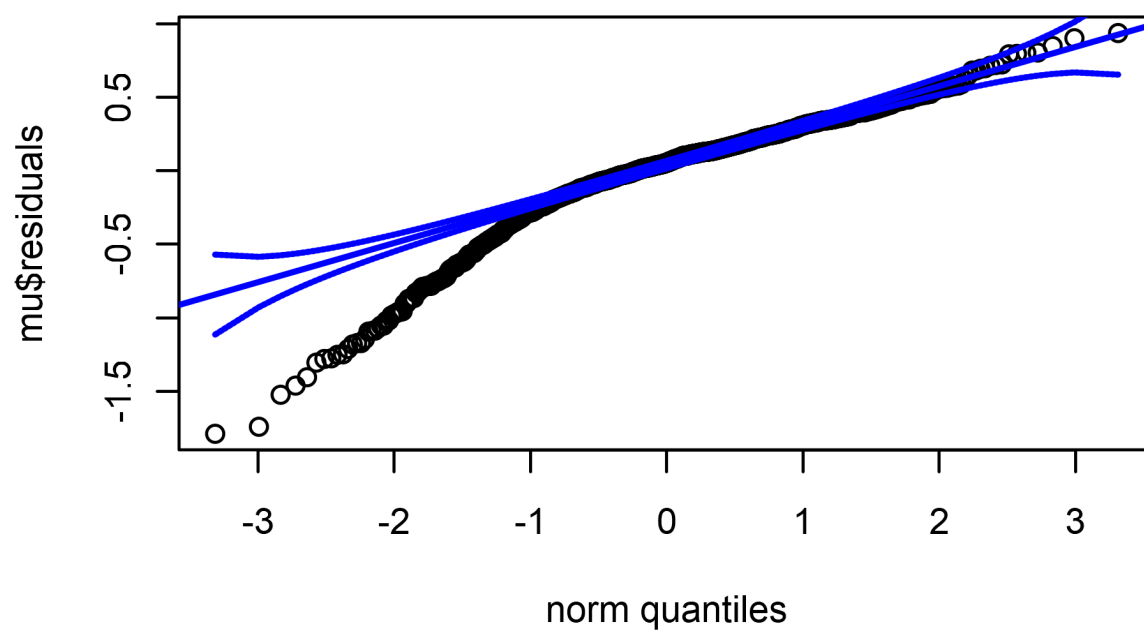
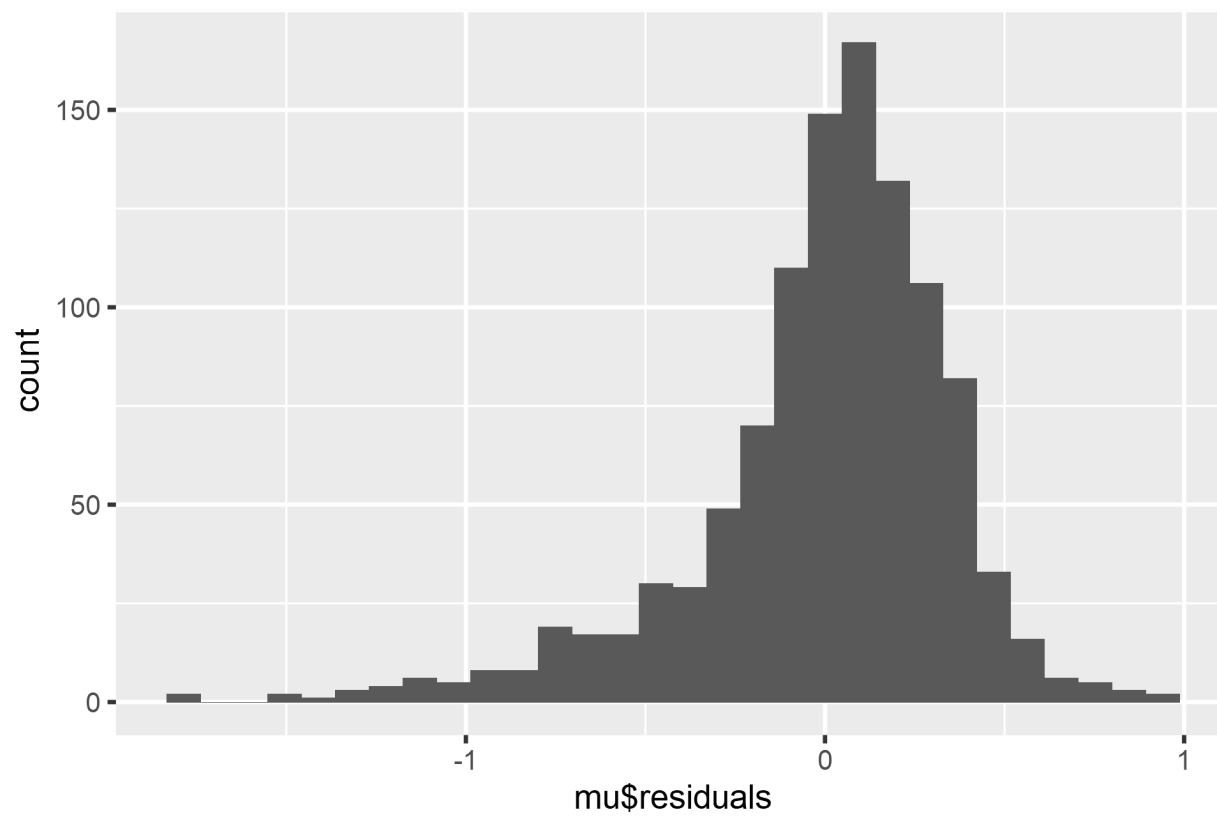


For sex, we can reasonably guess that sex does affect the average of earning and the way height affecting earn.



We can tell the education level is related to earn, rough judge from above figure, we can tell that higher education level, higher earning, and sex has no obvious effect on the slope of education level.

```
##
## Call:
## lm(formula = log(earn, 10) ~ h_over_m + sex + I(log(ed)) + sex:h_over_m +
##     I(log(ed) * sex), data = hi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78936 -0.13797  0.05559  0.22213  0.93727
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.023693   0.220592  13.707 < 2e-16 ***
## h_over_m       0.008637   0.005733   1.507  0.13218
## sex           -0.815275   0.293800  -2.775  0.00562 **
## I(log(ed))     0.500305   0.085104   5.879 5.51e-09 ***
## I(log(ed) * sex) 0.238336   0.112889   2.111  0.03498 *
## h_over_m:sex   -0.013923   0.008179  -1.702  0.08900 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3588 on 1075 degrees of freedom
## Multiple R-squared:  0.1779, Adjusted R-squared:  0.174
## F-statistic: 46.51 on 5 and 1075 DF,  p-value: < 2.2e-16
```



After adding new variables to our model, R^2 increased a little, but the residuals of model is still not normally distributed.

4. Interpret all model coefficients.

Clearly, this model is bad, low R^2 means this model could explain little of the variance of output, and half of the coefficients are not significant, but due to some of them are highly correlated, we have to keep them to control the influence of endogeneity. Here is the interpretation of coefficients:

h_over_m and h_over_sex: when keeping others unchanged, one unit higher than average height, earning will increase around 0.863% for male, on the contrary, will decrease around 0.529%

sex: on average, keeping others unchanged, earning of female will be around 48% lower than male;

I(log(ed)) and I(log(ed)): on average, keeping others unchanged, 1% increase in education level, for male, earning will increase around 50%, for female, earning will increase 73%;

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

Table 3: Confident Interval of Coefficients

	2.5 %	97.5 %
(Intercept)	2.5908531	3.4565330
h_over_m	-0.0026110	0.0198853
sex	-1.3917609	-0.2387883
I(log(ed))	0.3333158	0.6672949
I(log(ed) * sex)	0.0168270	0.4598443
h_over_m:sex	-0.0299719	0.0021262

95% Confident Interval means that each interval has 95% chance that contain real coefficient.

Analysis of mortality rates and various environmental factors

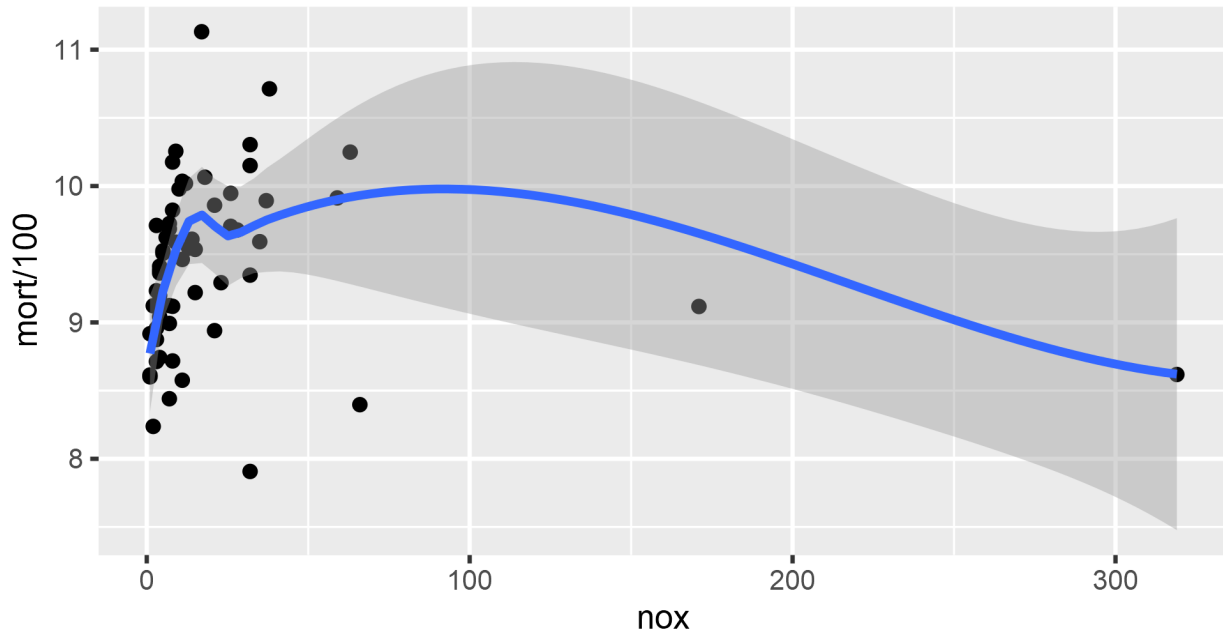
The folder `pollution` contains mortality rates and various environmental factors from 60 U.S. metropolitan areas from McDonald, G.C. and Schwing, R.C. (1973) 'Instabilities of regression estimates relating air pollution to mortality', *Technometrics*, vol.15, 463-482.

Variables, in order:

- PREC Average annual precipitation in inches
- JANT Average January temperature in degrees F
- JULY 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

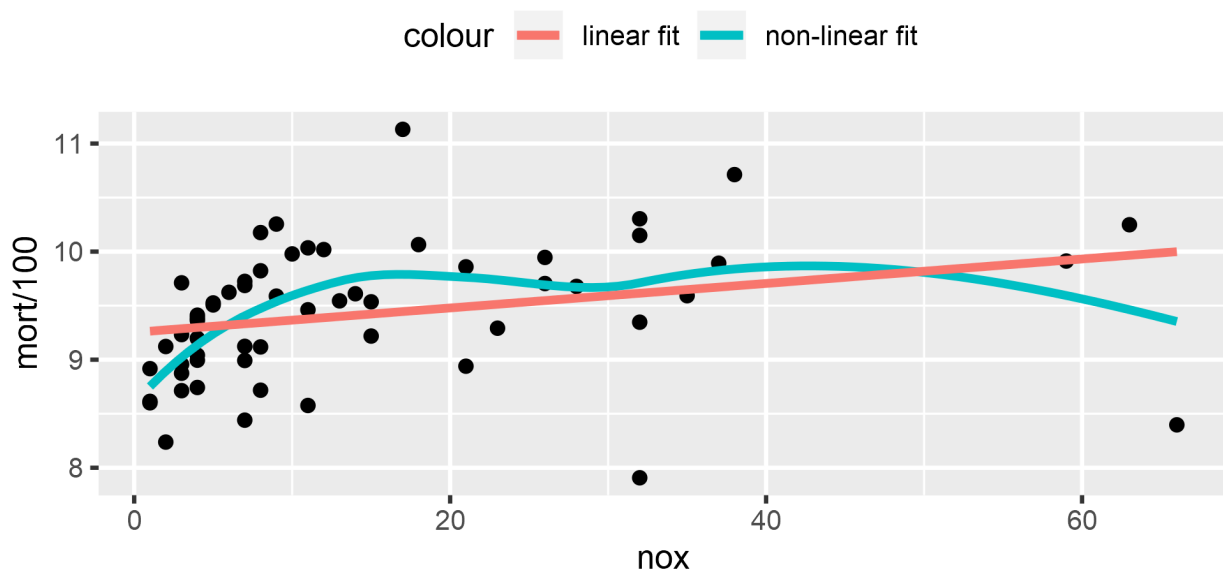
For this exercise we shall model mortality rate given nitric oxides, sulfur dioxide, and hydrocarbons as inputs. This model is an extreme oversimplification as it combines all sources of mortality and does not adjust for crucial factors such as age and smoking. We use it to illustrate log transformations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

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.



From above plot we can tell that linear relationship between mortality rate and nitric oxides won't fit well, and we can also tell from the smooth line that two outliers does influent the result.

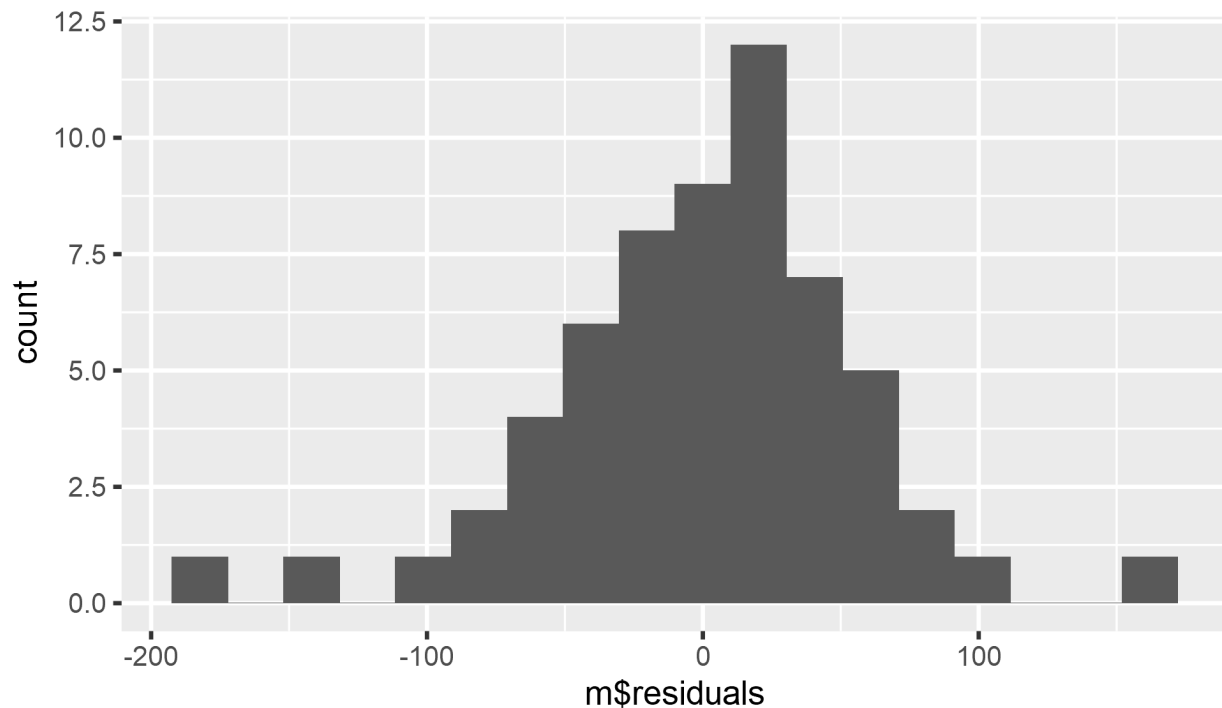
2. Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

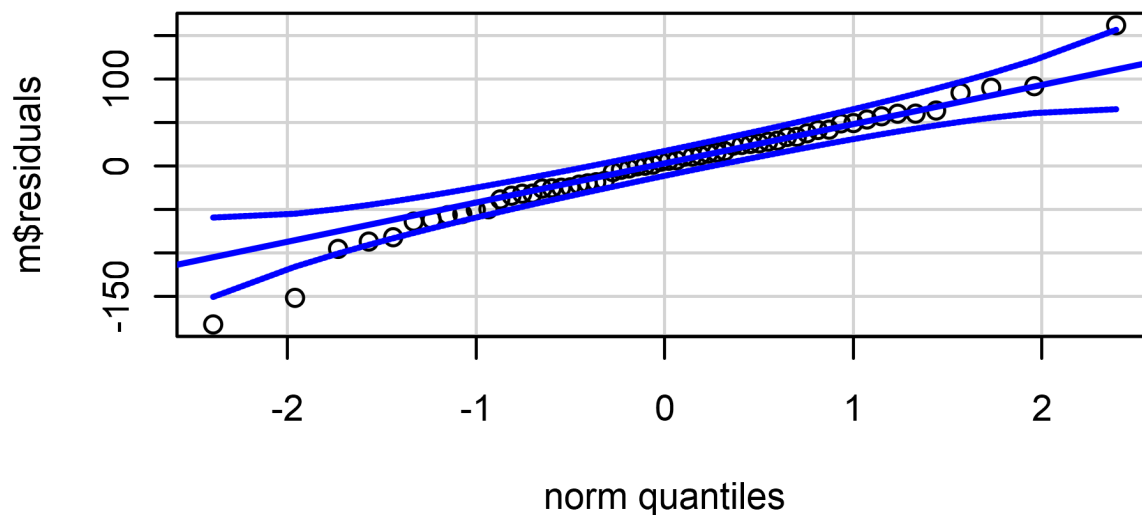


Exclude those two outliers, we can tell that we may try to add nox^3 to the predictors.

```
##
## Call:
## lm(formula = mort ~ nox + I(nox^2) + I(nox^3), data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -182.430  -27.441    5.511   33.449  161.985
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.142e+02  1.235e+01  74.016 < 2e-16 ***
## nox          2.582e+00  8.902e-01   2.900  0.00532 **
## I(nox^2)     -2.468e-02  9.784e-03  -2.523  0.01451 *
## I(nox^3)      5.048e-05  2.352e-05   2.146  0.03622 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.62 on 56 degrees of freedom
## Multiple R-squared:  0.1572, Adjusted R-squared:  0.1121
## F-statistic: 3.483 on 3 and 56 DF,  p-value: 0.02165
```

As we can see, the coefficients of the power of nox are all significant, but due to the small number of predictors, the R^2 of model is quite small.

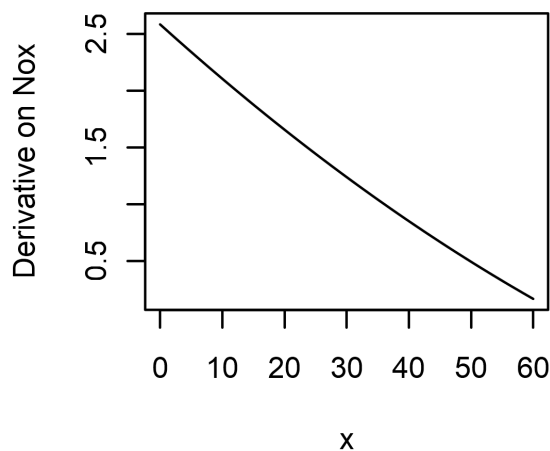
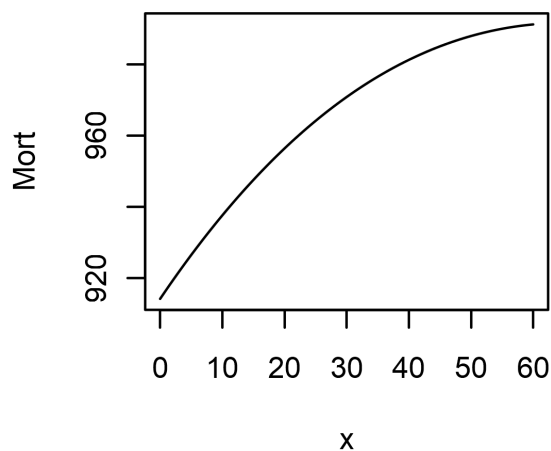




Even the R^2 is small, but the residuals comply with normal distribution.

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

In question 2, the regression model is $mort = 914.2 + 2.582 \cdot nox - 0.02468 \cdot nox^2 + 0.00005048 \cdot nox^3$, the derivative on nox is $2.582 - 0.04936 \cdot nox + 0.00015144 \cdot nox^2$, which means, assume $nox = 1$, every unit increase of nox , mortality rate increase by 2.253791, and from the figure below we can tell that the marginal increase of $mort$ is decrease.

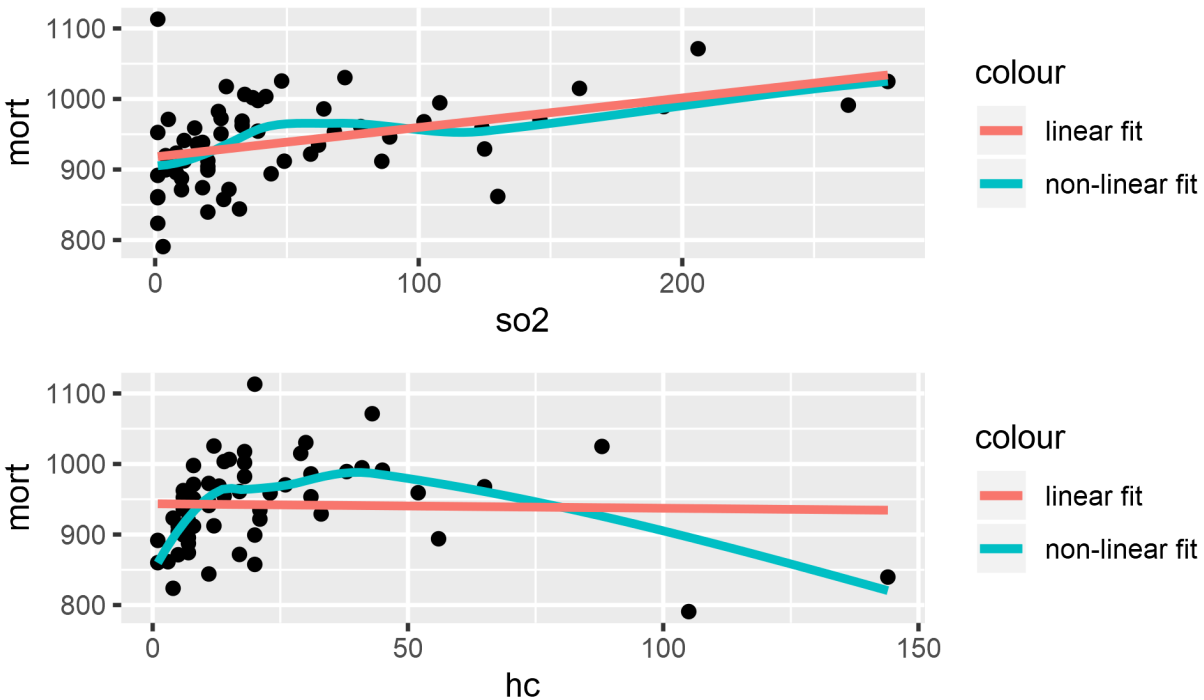


4. Construct 99% confidence interval for slope coefficient from the model you chose in 2 and interpret them.

Table 4: Confident Interval of Coefficients

	0.5 %	99.5 %
(Intercept)	881.2313287	947.0992026
nox	0.2081142	4.9555736
I(nox ²)	-0.0507714	0.0014063
I(nox ³)	-0.0000122	0.0001132

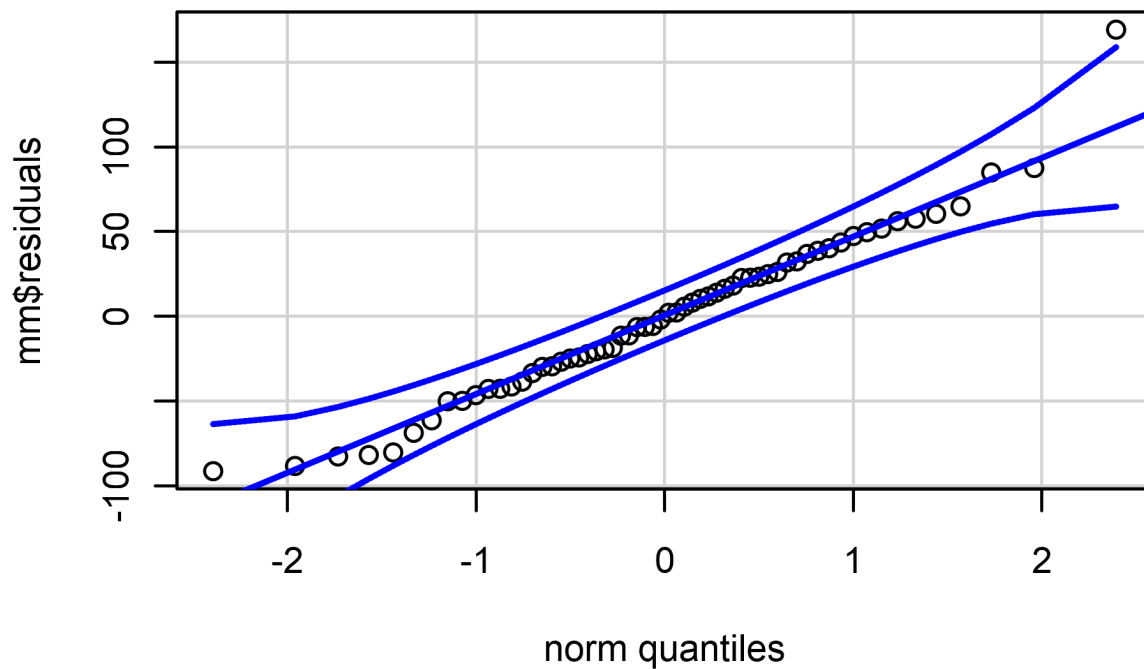
5. Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.



We can tell that for sulfur dioxide, linear relationship with mortality rate is proper, but for hydrocarbons, a transformation to the power of hydrocarbons is necessary.

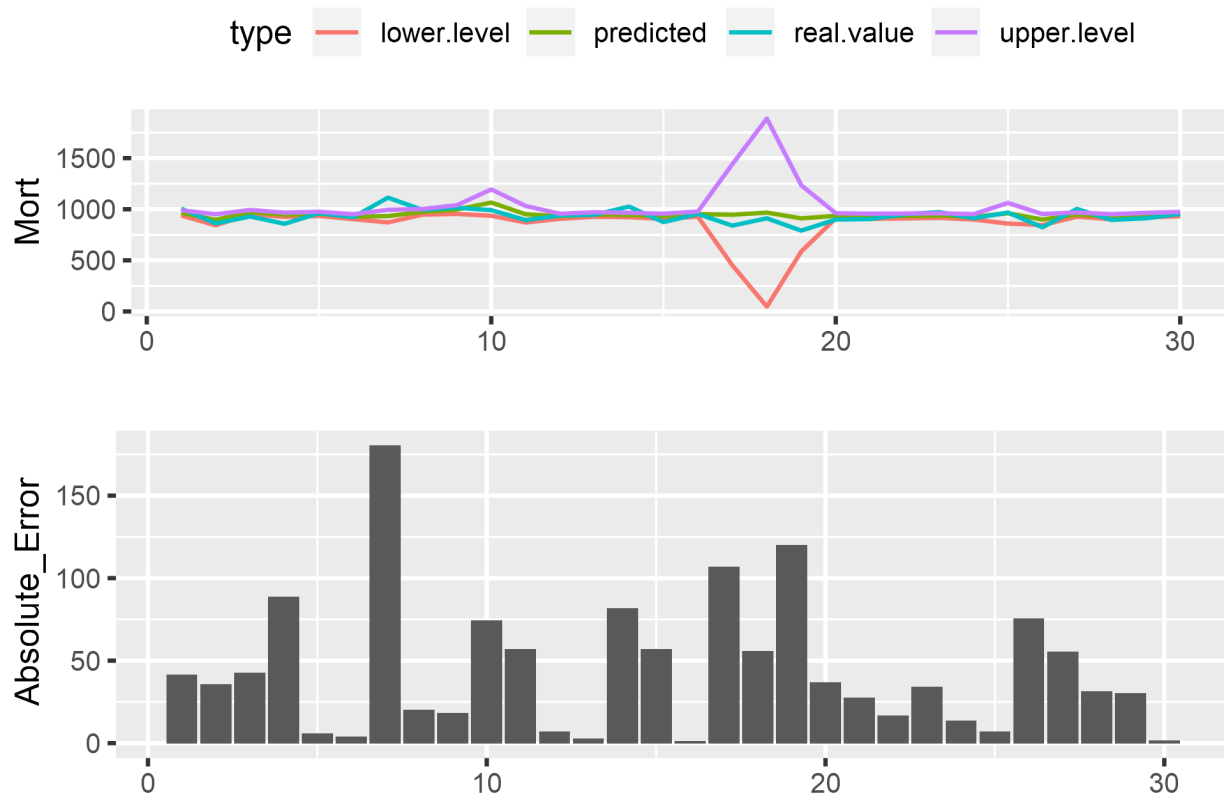
```
##
## Call:
## lm(formula = mort ~ nox + I(nox^3) + I(so2^(-1)) + I(hc^2), data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -91.426 -30.785   0.061  31.891 169.173
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.145e+02  1.124e+01  81.374 < 2e-16 ***
## nox          2.824e+00  5.955e-01   4.742 1.54e-05 ***
## I(nox^3)     1.271e-04  2.735e-05   4.646 2.15e-05 ***
## I(so2^(-1)) -1.434e+01  2.395e+01  -0.599  0.552
```

```
## I(hc^2)      -1.208e-02  2.446e-03  -4.937 7.75e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.03 on 55 degrees of freedom
## Multiple R-squared:  0.3727, Adjusted R-squared:  0.3271
## F-statistic: 8.171 on 4 and 55 DF,  p-value: 3.026e-05
```



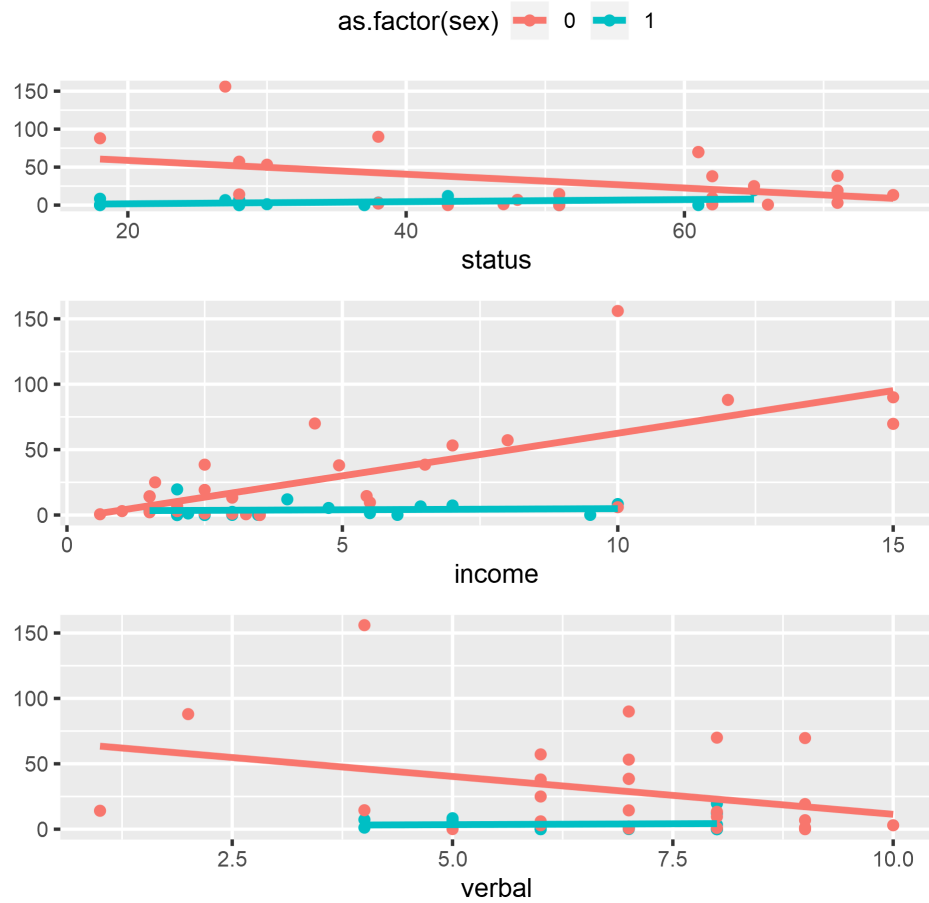
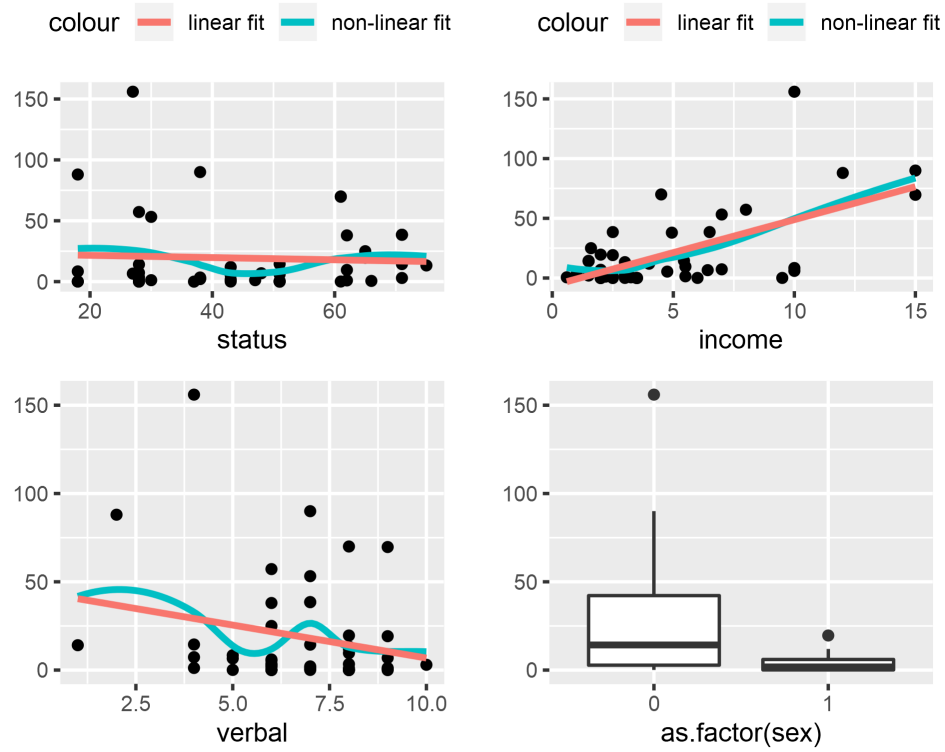
According to the result of the model, the residuals comply with normal distribution, the coefficients of the power of nox and hc are significant, but the coefficient of so2 are not even I have tried several different transformation, but due to the correlation between nox, so2 and hc is high (nox~so2:0.4093936), it's better to delete this variables so that the coefficients of nox and hc will not be affected.

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.)



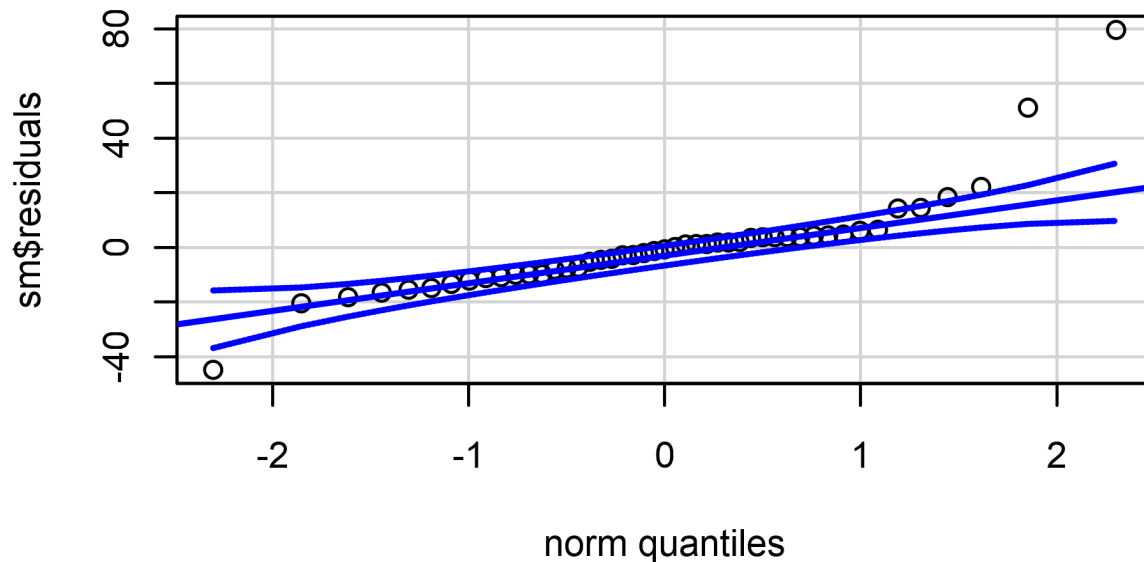
Study of teenage gambling in Britain

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.



According to figure above, we consider adding the power of status and verbal as predictors, and also we may consider the interaction between sex and other predictors.

```
##
## Call:
## lm(formula = gamble ~ status + I(status^2) + income + verbal +
##     sex + sex:status + sex:income, data = teengamb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -44.776  -9.817  -0.680   3.813  79.603
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  95.33067   32.79573   2.907  0.00599 **
## status       -3.56884    1.40587  -2.539  0.01524 *
## I(status^2)   0.03307    0.01355   2.442  0.01926 *
## income        5.70833    0.99385   5.744 1.18e-06 ***
## verbal       -0.94242    1.99482  -0.472  0.63925
## sex          -43.88391   27.86806  -1.575  0.12340
## status:sex     1.08107    0.53555   2.019  0.05044 .
## income:sex    -5.84836    2.27461  -2.571  0.01406 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.63 on 39 degrees of freedom
## Multiple R-squared:  0.671, Adjusted R-squared:  0.6119
## F-statistic: 11.36 on 7 and 39 DF,  p-value: 9.442e-08
```



From above result, most of the coefficients are significant, and residuals comply with normal distribution.

Coefficients interpretation:

status and status:sex: because the coefficient of the power of status is too small compared to the coefficient of status, so we can ignore it, so the coefficient of status means that one unit change in social status, keeping others unchanged, the average expenditure on gambling per year will decrease around 3.5 pound for male, for female, it will only decrease 2.42 pounds;

income and income:sex: one pound increase of weekage, keeping others unchanged, expenditure on gambling will increase by 5.7 pounds for male, for female, it will decrease by 0.1 pounds;

verbal: one unit increase in verbal score, keeping others unchanged, expenditure on gambling will decrease by 0.9 pounds;

sex: the average expenditure on gambling for female is 43 pounds less than male.

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

Table 5: Confident Interval of Coefficients

	0.5 %	99.5 %
(Intercept)	6.5226872	184.1386521
status	-7.3758201	0.2381323
I(status^2)	-0.0036061	0.0697557
income	3.0170676	8.3995929
verbal	-6.3442253	4.4593909
sex	-119.3482034	31.5803854
status:sex	-0.3691609	2.5312994
income:sex	-12.0078001	0.3110861

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?

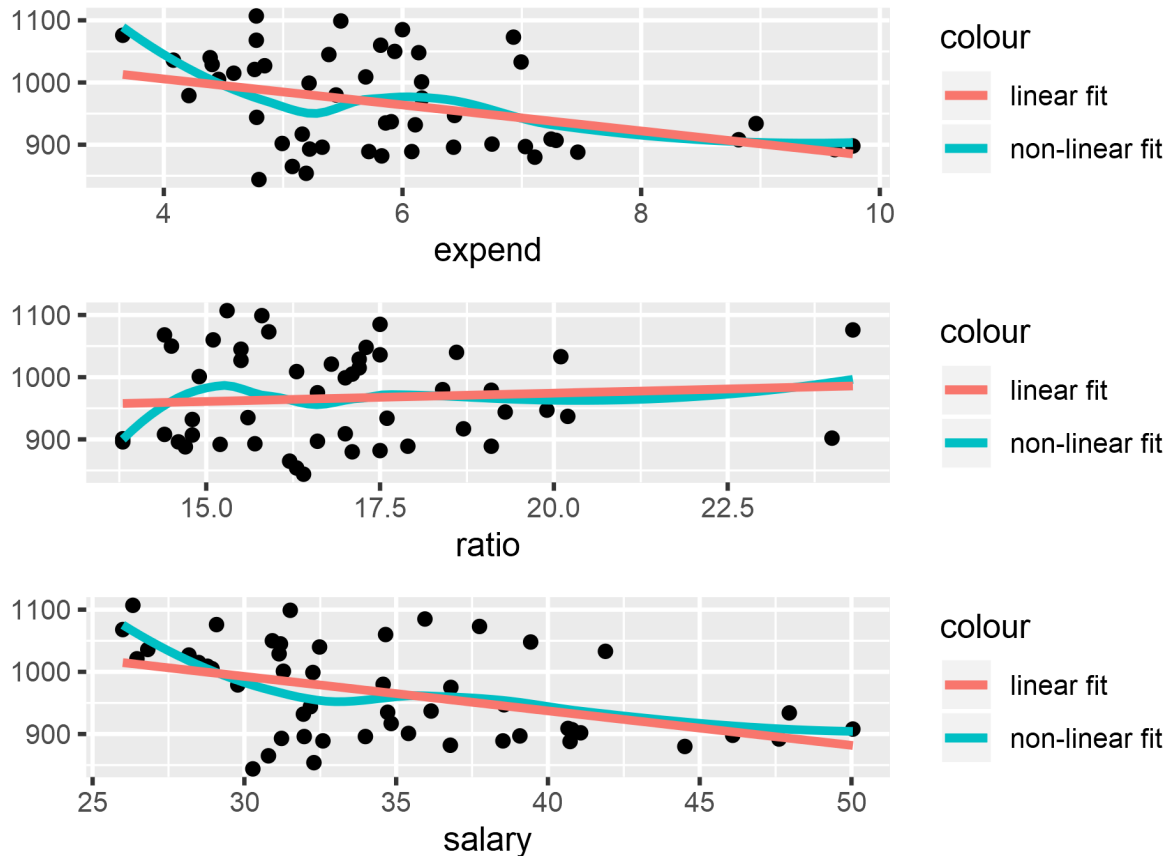
Table 6: prediction of gamble

	fit	lwr	upr	variate
Mean level	21.79382	11.14272	32.44492	21.30220
max level	89.91373	60.70990	119.11755	58.40766

From above result, the CI of prediction using max level is bigger.

School expenditure and test scores from USA in 1994-95

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.



we can tell from the above figure, that a linear relationship may be enough.

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary, data = sat)
##
## 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)  1069.234    110.925   9.639 1.29e-12 ***
## expend       16.469     22.050   0.747  0.4589
## ratio        6.330      6.542   0.968  0.3383
## 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
```

Clearly that the SAT score has nothing to do with *expend* and *ratio*, but *salary* has some influence on sat score. Coefficients interpretation:

expend: One unit increase in *expend* may increase the sat score by 11.4, keeping others unchanged;

ratio: One unit increase in ratio may decrease the Sat score by 0.436, keeping other unchanged;

salary: Due to the coefficient of $salary^2$ is too small compared to the coefficient of salary, we just ignore the coefficient of $salary^2$, so the coefficient of salary means that one unit increase insalary will decrease the SAT score by 19.06.

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

Table 7: Confident Interval of Coefficients

	1 %	99 %
(Intercept)	801.88	1336.58
expend	-36.68	69.61
ratio	-9.44	22.10
salary	-20.14	2.50

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?

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary + takers, data = sat)
##
## 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) 1045.9715    52.8698  19.784 < 2e-16 ***
## expend         4.4626    10.5465   0.423  0.674
## ratio        -3.6242     3.2154  -1.127  0.266
## salary         1.6379     2.3872   0.686  0.496
## 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
```

Adding takers into the model improves the performance of this model a lot, mainly represented by the bigger R^2 , which increases 61%.

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, $D_i - R_i$

- The ratio, D_i/R_i
- The difference on the logarithmic scale, $\log D_i - \log R_i$
- The relative proportion, $D_i/(D_i + R_i)$.

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 $x^* = x - 10$ 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 $x^* = 10x$? When $x^* = 10(x - 1)$?

If we transform x into $x - 10$, $\hat{\alpha}^* = \hat{\alpha} + 10 \times \hat{\beta} = 1 + 10 \times 0.9 = 9.1$, $\hat{\beta}^*$, $\hat{\sigma}^*$, r remains.

If we change x into $10x$, $\hat{\beta}^* = \frac{\hat{\beta}}{10} = 9 \times 10^{-2}$, $\hat{\sigma}^* = \frac{\hat{\sigma}}{10^2} = 3 \times 10^{-4}$, $\hat{\alpha}^*$, r remains.

If we transform x into $10(x - 1)$, $\hat{\alpha}^* = 9.1$, $\hat{\beta}^* = 9 \times 10^{-2}$, $\hat{\sigma}^* = 3 \times 10^{-4}$, r remains.

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 ?
4. Suppose that the explanatory variable values in a regression are transformed according to the $x^* = 10(x - 1)$ and that y is regressed on x^* . Without redoing the regression calculation in detail, find $SE(\hat{\beta}^*)$ and $t_0^* = \hat{\beta}^*/SE(\hat{\beta}^*)$.
5. Now suppose that the response variable scores are transformed according to the formula $y^{**} = 5(y + 2)$ and that y^{**} is regressed on x . Without redoing the regression calculation in detail, find $SE(\hat{\beta}^{**})$ and $t_0^{**} = \hat{\beta}^{**}/SE(\hat{\beta}^{**})$.
6. In general, how are the hypothesis tests and confidence intervals for β 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.