Analysis of mortality rates and various environmental factors

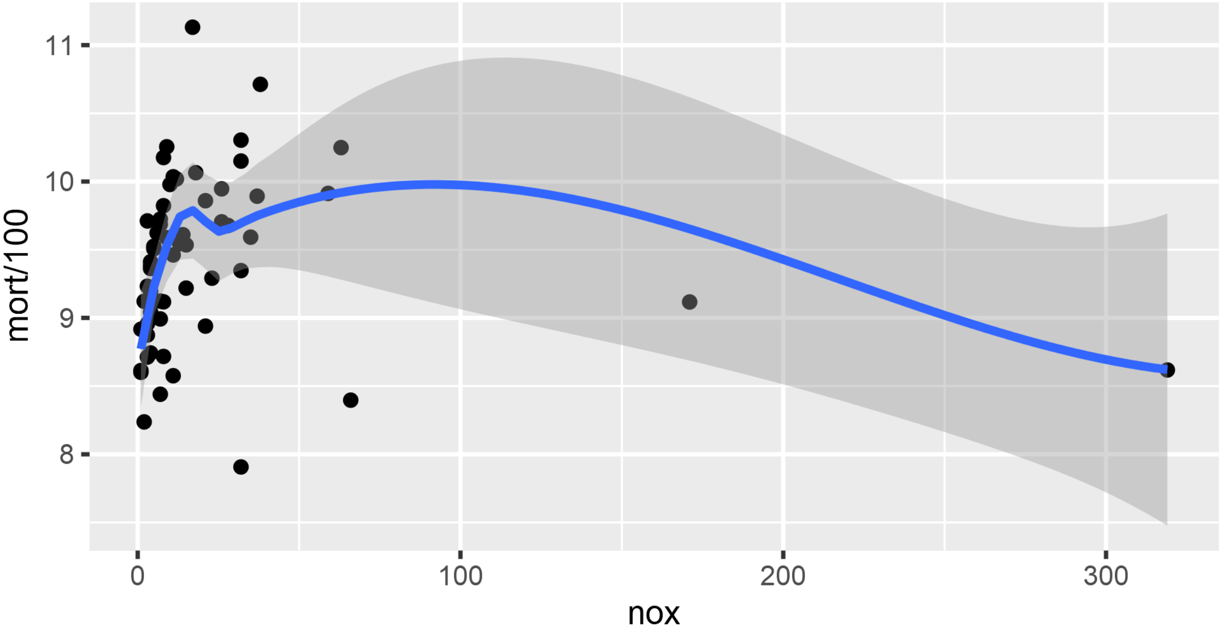
The folder `pollution` contains mortality rates and various environmental factors from 60 U.S. metropolitan areas from McDonald, G.C. and Schwing, R.C. (1973) 'Instabilities of regression estimates relating air pollution to mortality', Technometrics, vol.15, 463-482.

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

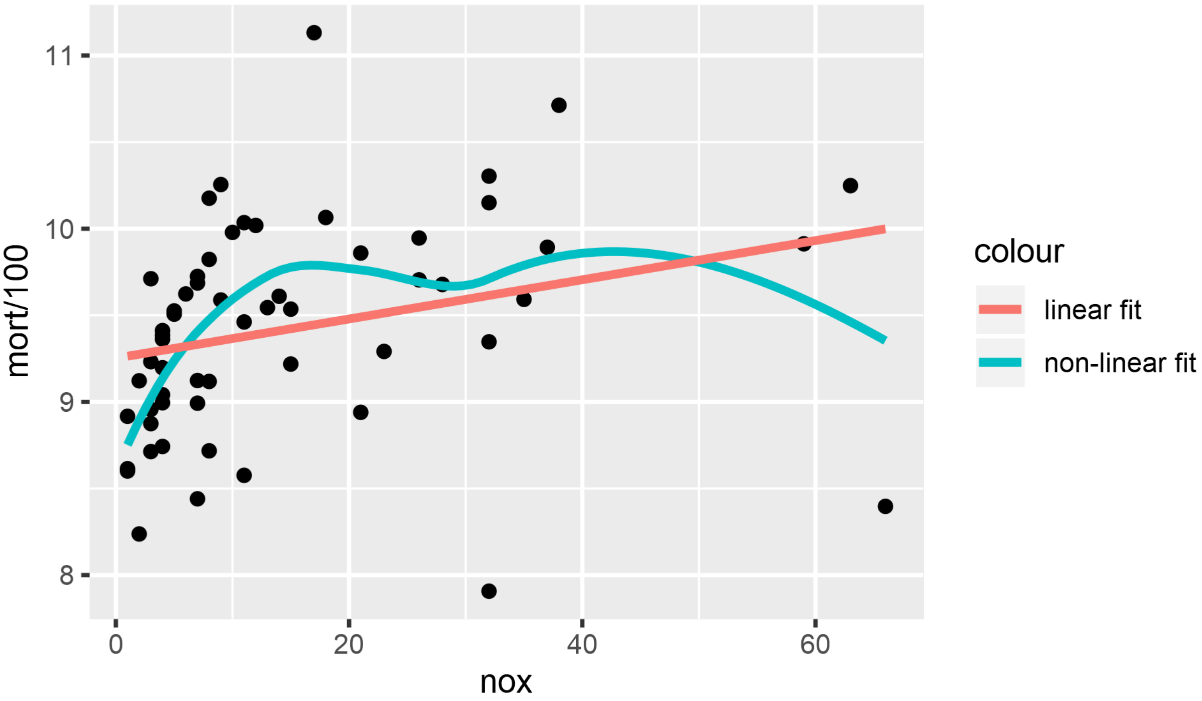
For this exercise we shall model mortality rate given nitric oxides, sulfur dioxide, and hydrocarbons as inputs. This model is an extreme oversimplification as it combines all sources of mortality and does not adjust for crucial factors such as age and smoking. We use it to illustrate log transformations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

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.

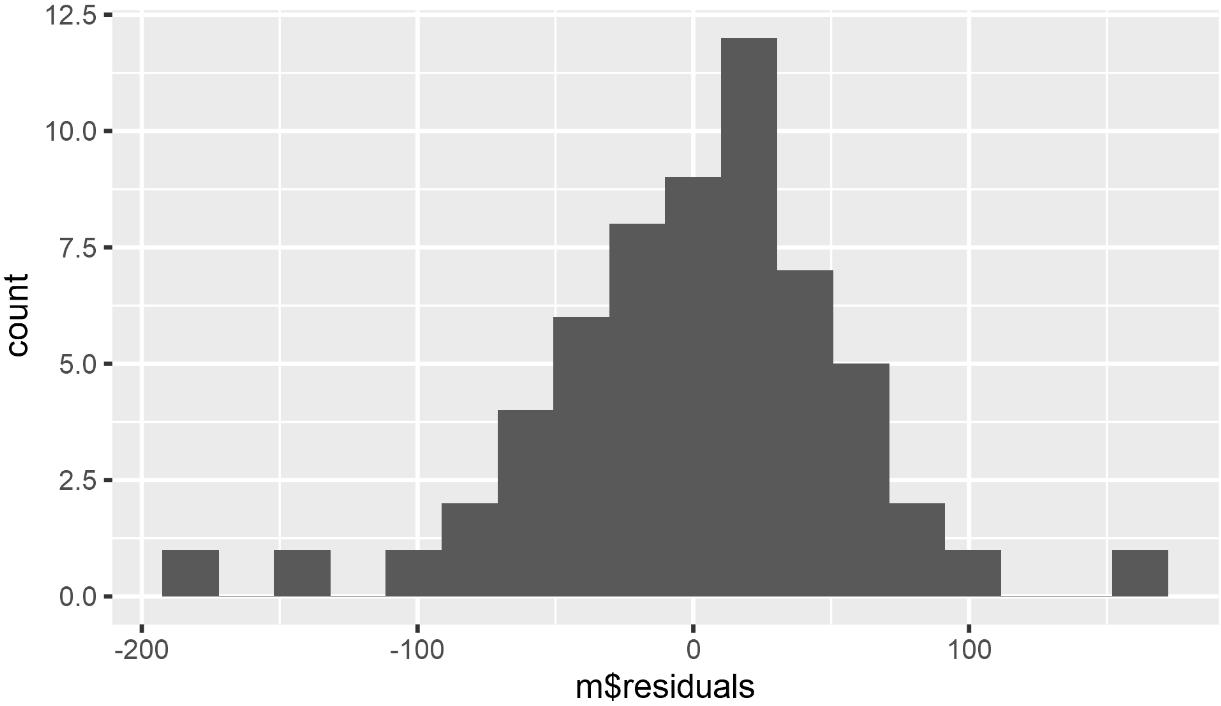
2. Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

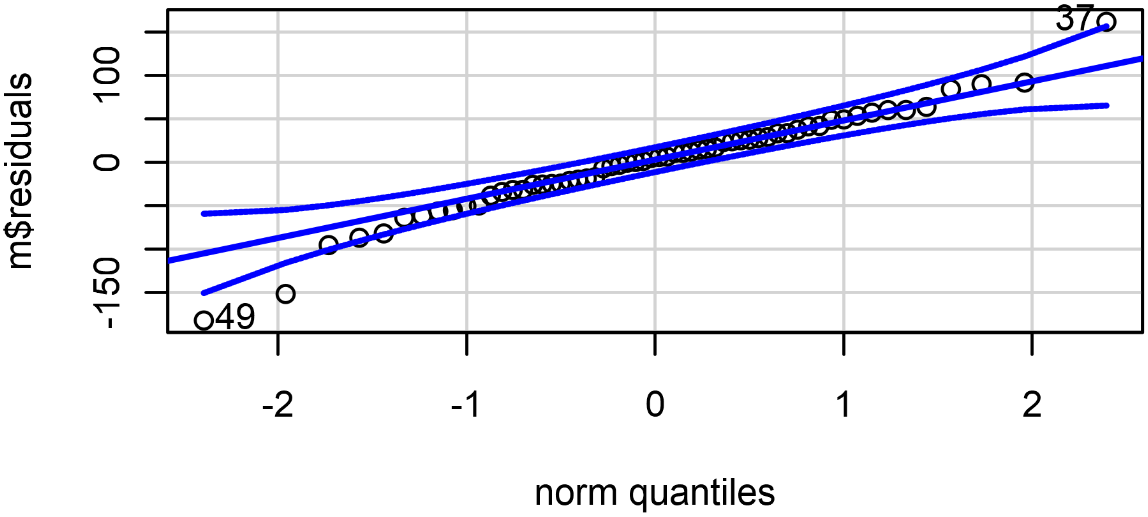


Exclude those two outliers, we can tell that we may try to add nox^3 to the predictors.

```
##
## 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.



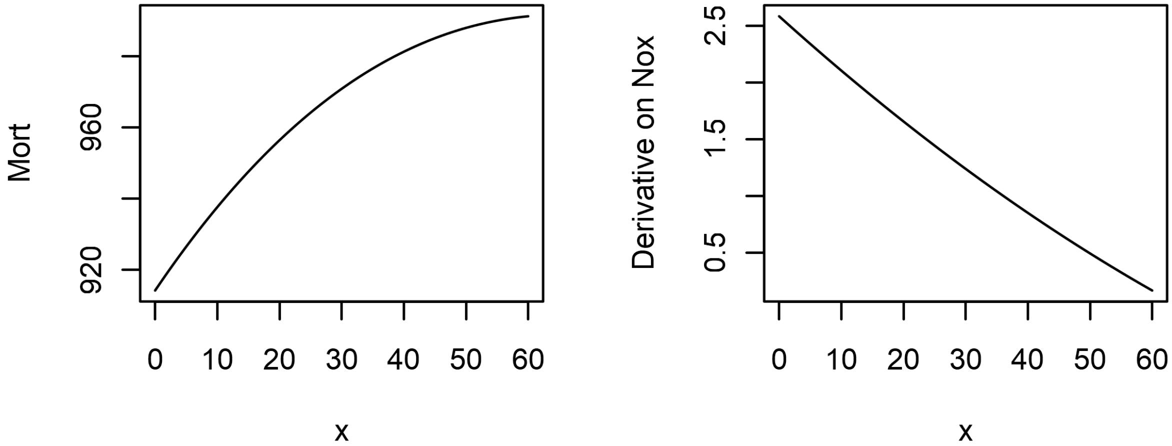


```
## [1] 49 37
```

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, motality rate increase by 2.253791, and from the figure below we can tell that the marginal increase of mort is decrease.

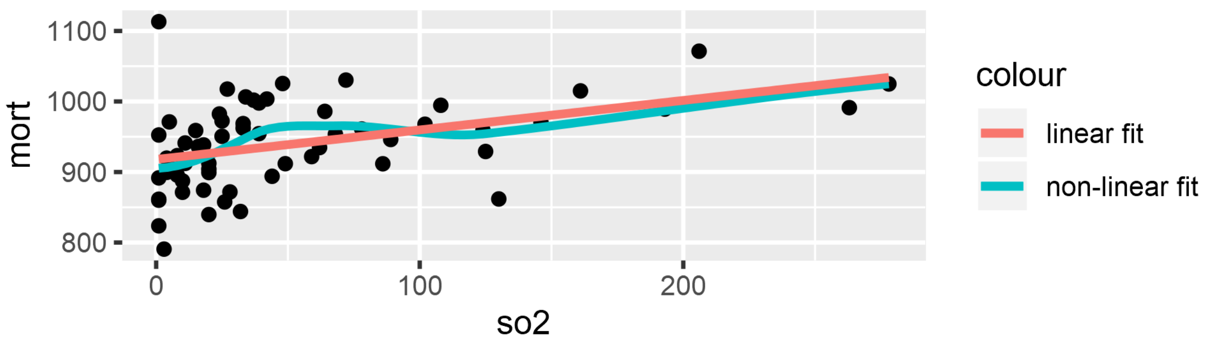


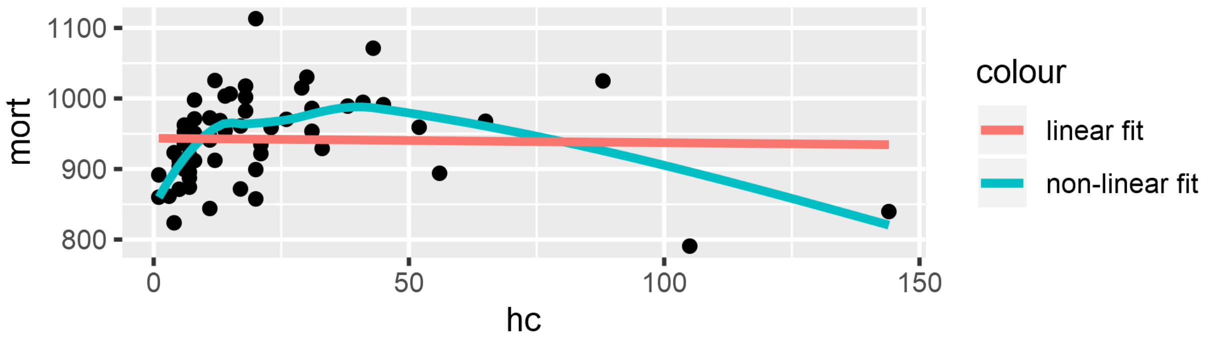
4. Construct 99% confidence interval for slope coefficient from the model you chose in 2 and interpret them.

Table 1: 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

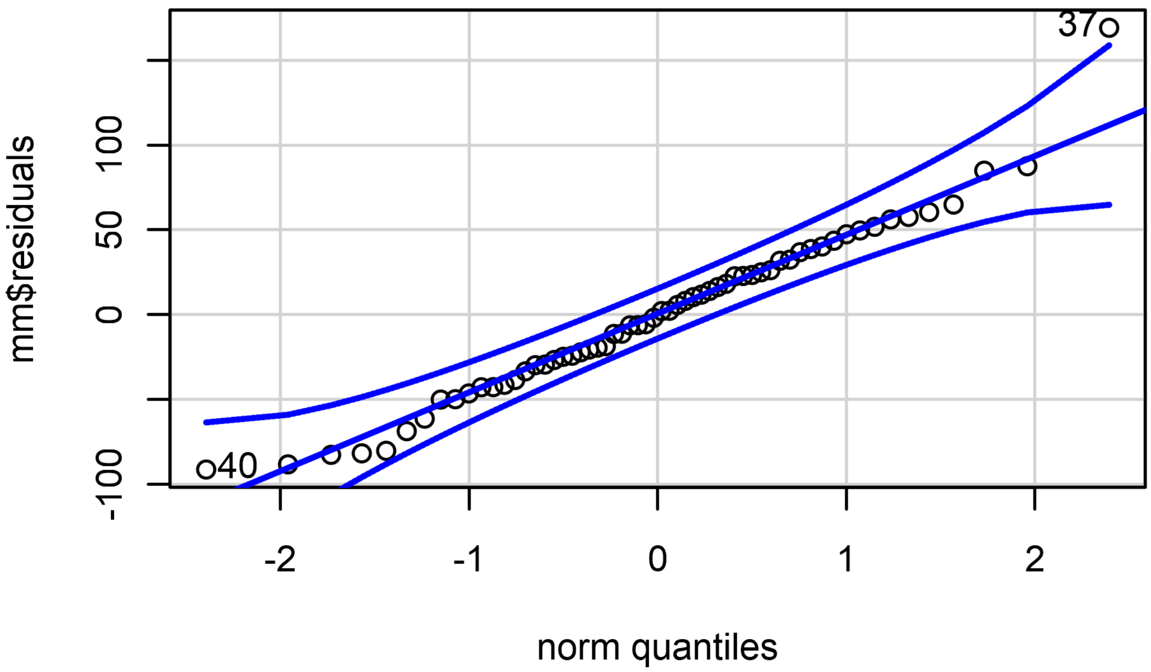
5. Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.





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

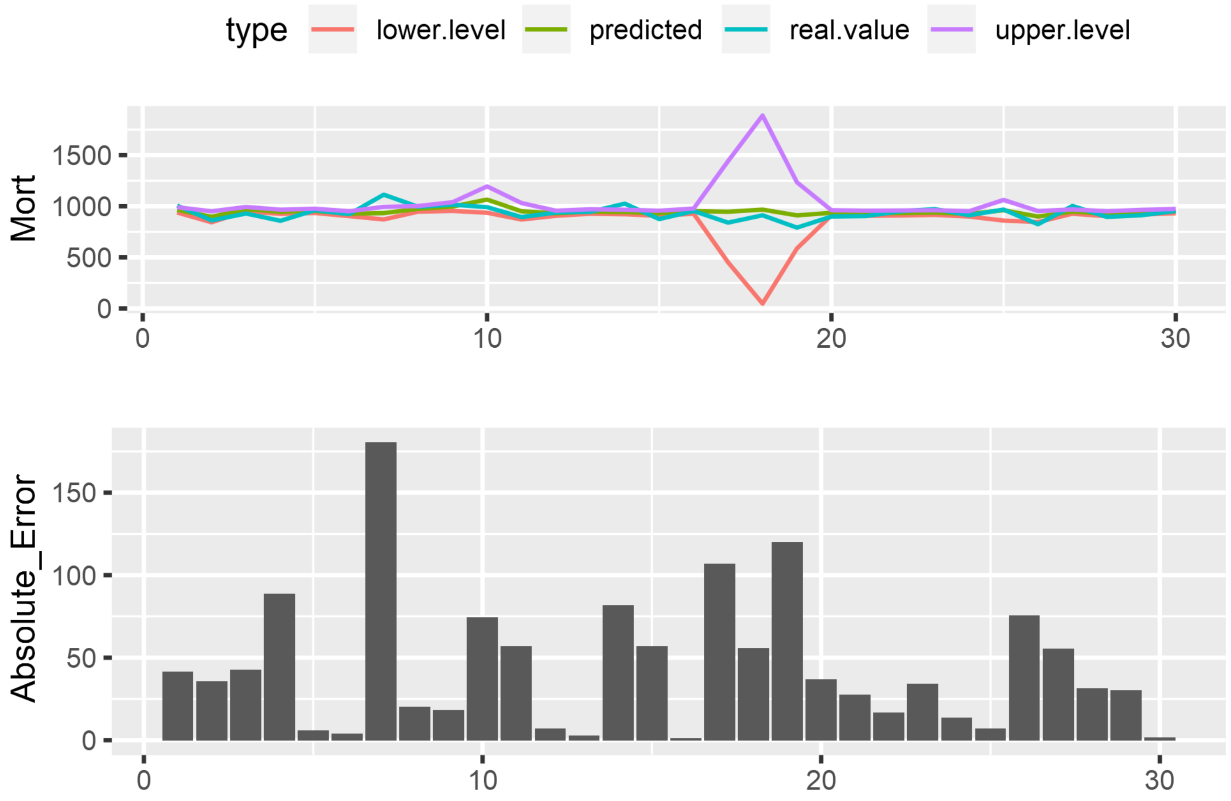
```
##
## Call:
## lm(formula = mort ~ nox + I(nox^3) + I(so2^(-1)) + I(hc^2), data = pollution)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.426 -30.785   0.061  31.891 169.173
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.145e+02  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
```



```
## [1] 37 40
```

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

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

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

Transformation

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

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

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

1. Suppose that the explanatory variable values in a regression are transformed according to the $\mathrm{x}^\star = \mathrm{x} - 10$ and that y is regressed on $\mathrm{x}^\star$. Without redoing the regression calculation in detail, find $\hat{\alpha}^\star$, $\hat{\beta}^\star$, $\hat{\sigma}^\star$, and $r^\star$. What happens to these quantities when $\mathrm{x}^\star = 10\mathrm{x}$? When $\mathrm{x}^\star = 10(\mathrm{x} - 1)$?
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 $\mathrm{y}^{\star\star} = 5(\mathrm{y} + 2)$?
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 $\mathrm{x}^\star = 10(\mathrm{x} - 1)$ 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 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.