Homework 02

Sample Solution

Septemeber 19, 2017

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_df    <- read.dta (paste0(gelman_dir,"earnings/heights.dta"))
```

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

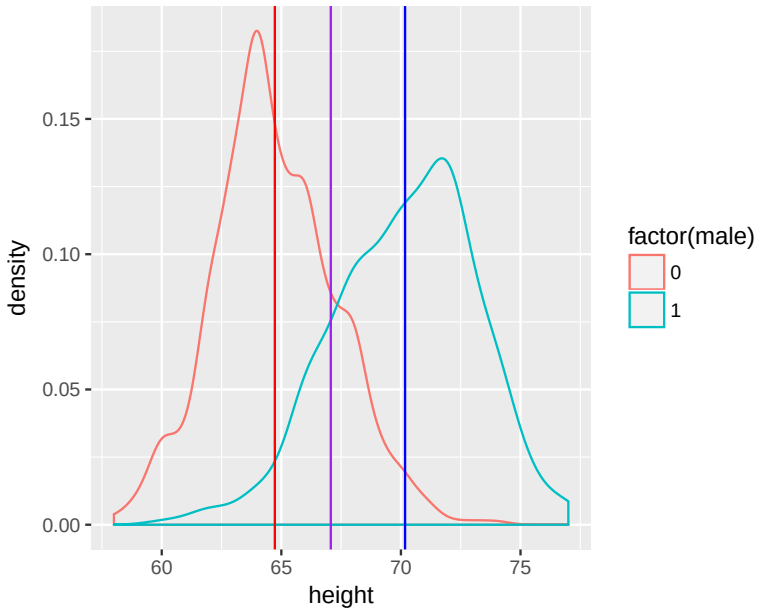
1. In R, check the dataset and clean any unusually coded data.

```
heights <- heights_df

# create variables for age and ethnicity categories

heights$age <- 90 - heights$yearbn  # survey was conducted in 1990
heights$age[heights$age<18] <- NA
heights$age.category <- ifelse (heights$age<35, 1, ifelse (heights$age<50, 2, 3))
heights$eth <- ifelse (heights$race==2, 1, ifelse (heights$hisp==1, 2, ifelse (heights$race==1, 3, 4)))
heights$male <- 2 - heights$sex

# education 99 is NA
heights$ed[heights$ed==99] <- NA

# (for simplicity) remove cases with missing data
# and restrict to people with positive earnings born after 1925

heights$ok <-!is.na (heights$earn+heights$height+heights$sex+heights$age) &
  heights$earn>0 & heights$yearbn>25
heights.dt <- heights[heights$ok==TRUE,]
n <- nrow (heights.dt)
```

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?

Average height is a tricky concept when you have both male and female in your sample. If you look at the distribution for height for each gender you will see that they are very different.

```
ggplot(heights.dt)+geom_density()+aes(x=height,color=factor(male))+
  geom_vline(aes(xintercept = mean(height) ),color="purple") +
  geom_vline(aes(xintercept = mean(height[male==1])),color="blue") +
  geom_vline(aes(xintercept = mean(height[male==0])),color="red")
```



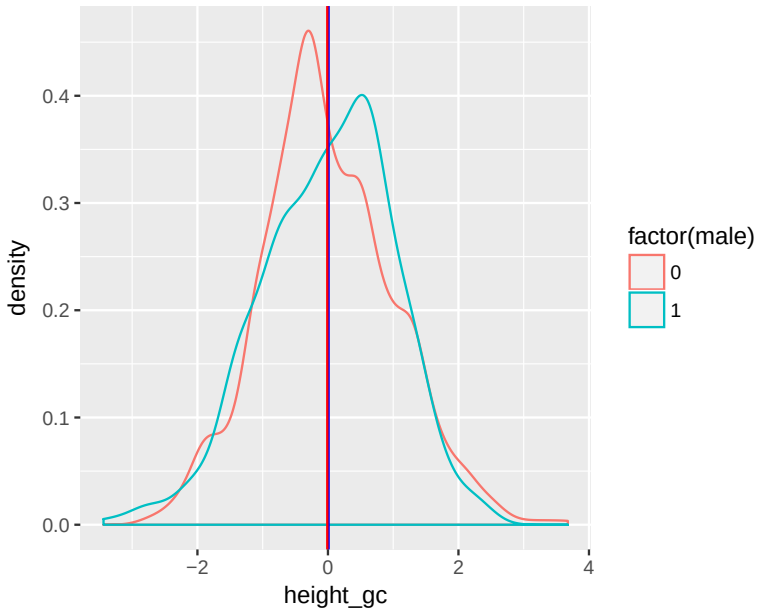
When we subtract the sample means what we are doing is centering the height at a height that is on a larger end of female and lower end of male. So we should ask whether we are interested in absolute height or the relative height. For simplicity, we will go with absolute height.

```
heights.dt$height_c<- heights.dt$height-mean(heights.dt$height)
fit_height_01 <- lm(earn~height_c,data=heights.dt)
display(fit_height_01)
```

```
## lm(formula = earn ~ height_c, data = heights.dt)
##             coef.est coef.se
## (Intercept) 23720.70   593.30
## height_c     1255.98   155.06
## ---
## n = 1059, k = 2
## residual sd = 19307.17, R-Squared = 0.06
```

However, if you are interested in gender centered height we should do something like

```
heights.dt$height_gc<- (heights.dt$height-
  by(heights.dt$height,heights.dt$male,mean)[heights.dt$male+1])/
  by(heights.dt$height,heights.dt$male,sd)[heights.dt$male+1]

ggplot(heights.dt)+geom_density()+aes(x=height_gc,color=factor(male))+
  geom_vline(aes(xintercept = mean(height_gc) ),color="purple") +
  geom_vline(aes(xintercept = mean(height_gc[male==1])+0.01),color="blue") +
  geom_vline(aes(xintercept = mean(height_gc[male==0])-0.01),color="red")
```



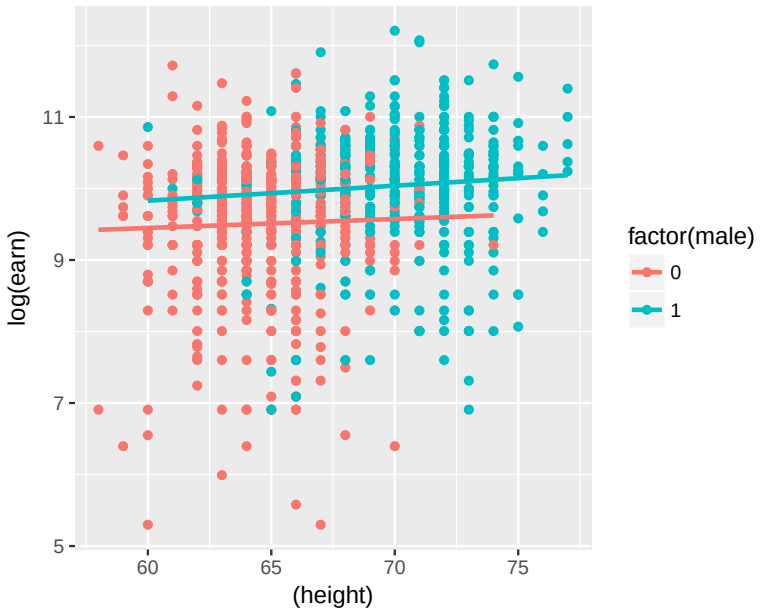

```
fit_height_012 <- lm(earn~height_gc,data=heights.dt)
display(fit_height_012)
```

```
## lm(formula = earn ~ height_gc, data = heights.dt)
##             coef.est coef.se
## (Intercept) 23720.70   610.77
## height_gc     925.79   611.35
## ---
## n = 1059, k = 2
## residual sd = 19875.79, R-Squared = 0.00
```

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.

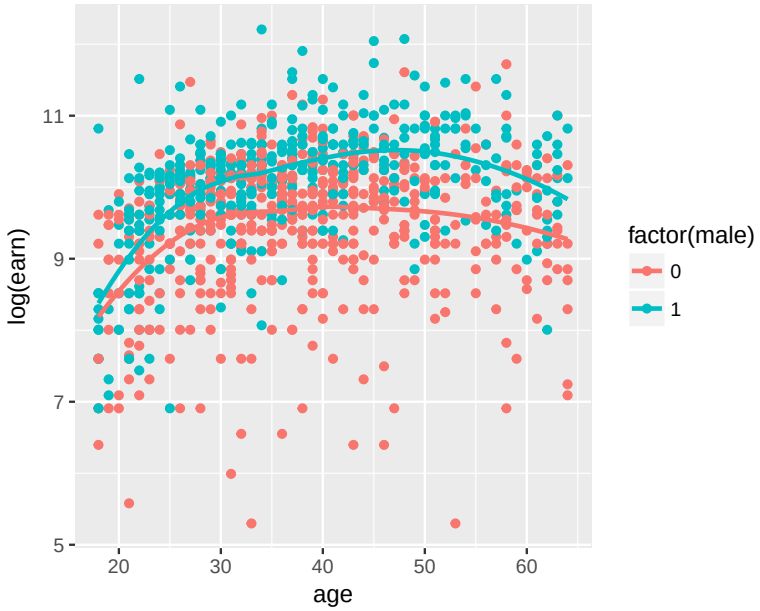
After couple of EDA figures

```
ggplot(heights.dt)+geom_point()+aes(x=(height),y=log(earn),color=factor(male))+geom_smooth(method="lm",s
```



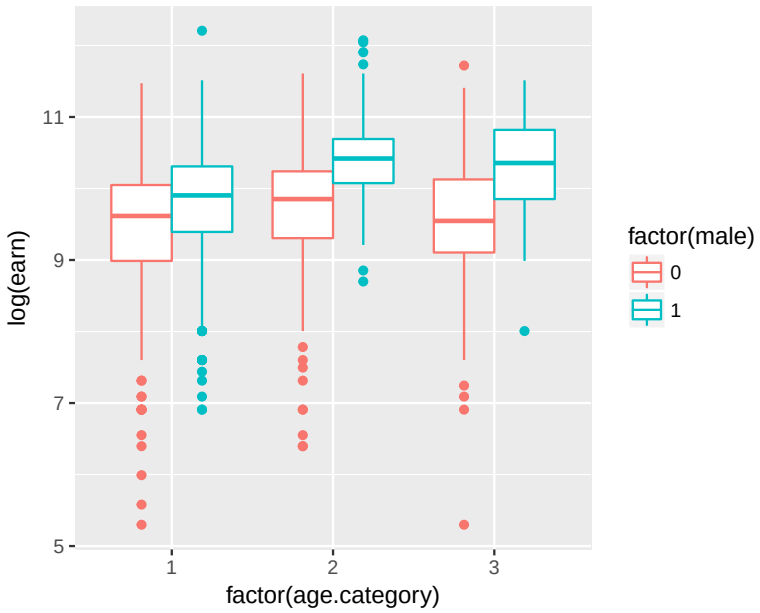

```
ggplot(heights.dt)+geom_point()+aes(x=age,y=log(earn),color=factor(male))+geom_smooth(se=FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```



You might come to think that age has nonlinear effect, which from substantive perspective makes sense because not everyone's income will increase as they get older. One way to address this is to add nonlinear effect to the model. Another is to discretize the age so that we have 3 categories 35<, 35> & 65<, 65>.

```
ggplot(heights.dt)+geom_boxplot()+aes(x=factor(age.category),y=log(earn),color=factor(male))
```

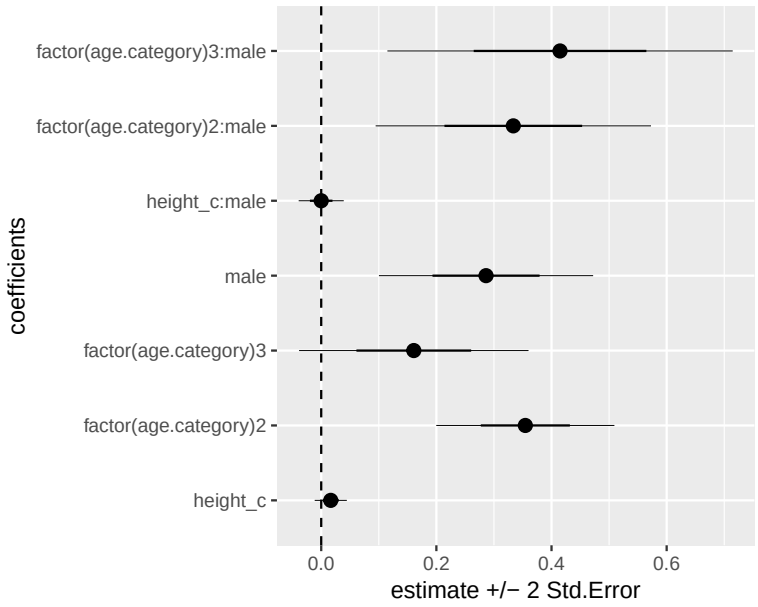


We will go with simpler approach.

```
fit_height_02 <- lm(log(earn)~(height_c+factor(age.category))*male,data=heights.dt)
display(fit_height_02)
```

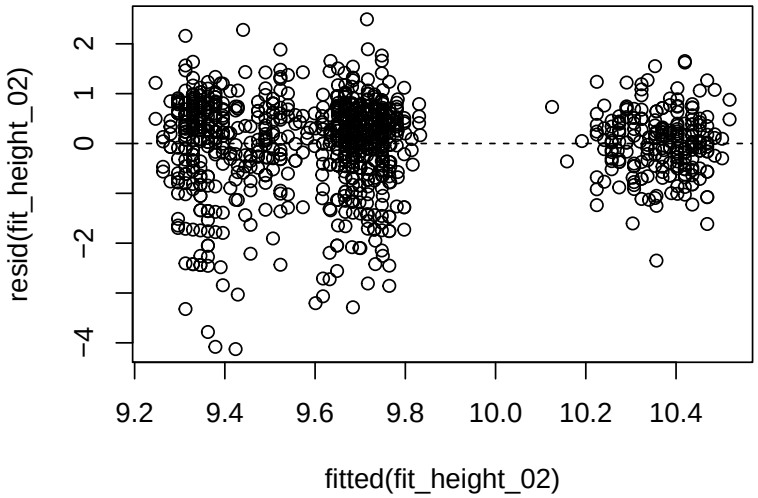
```
## lm(formula = log(earn) ~ (height_c + factor(age.category)) *
##     male, data = heights.dt)
##                             coef.est coef.se
## (Intercept)                 9.38     0.06
## height_c                    0.02     0.01
## factor(age.category)2       0.35     0.08
## factor(age.category)3       0.16     0.10
## male                        0.29     0.09
## height_c:male               0.00     0.02
## factor(age.category)2:male 0.33     0.12
## factor(age.category)3:male 0.41     0.15
## ---
## n = 1059, k = 8
## residual sd = 0.86, R-Squared = 0.15
```

```
coefplot_my(fit_height_02)
```



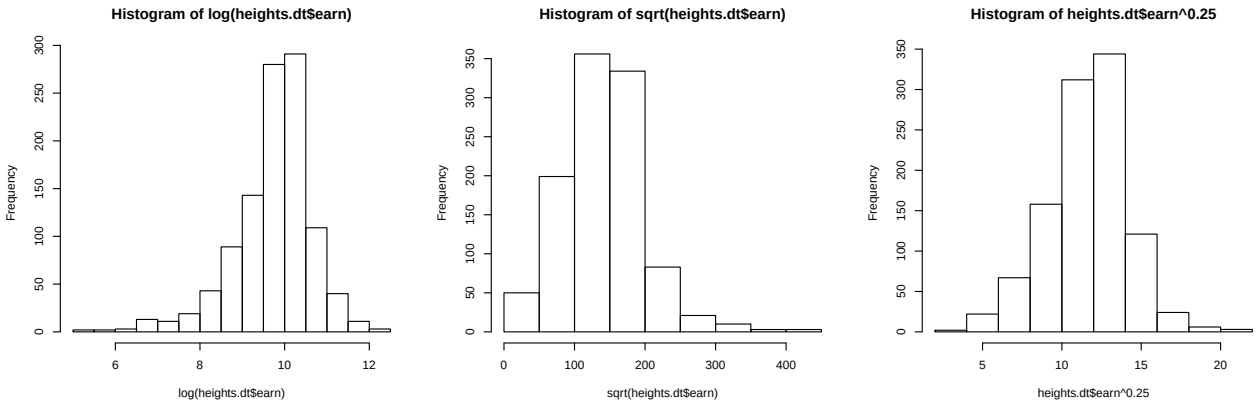
residual plot shows us three clusters for different age categories. The equal variance assumption seems to be violated in this case due to left skewness of the log earning.

```
plot(fitted(fit_height_02),resid(fit_height_02)); abline(h=0,lty=2)
```



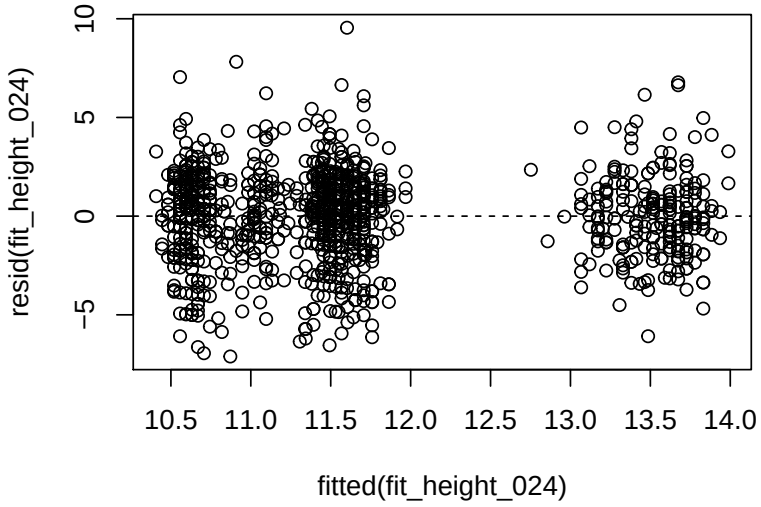
We can choose different transformations to remove the left skewness.

```
par(mfrow=c(1,3))
hist(log(heights.dt$earn))
hist(sqrt(heights.dt$earn))
hist(heights.dt$earn^0.25)
```



the quartic root seems to do a fairly good job in getting a symmetric distribution.

```
fit_height_024 <- lm((earn)^0.25~(height_c+factor(age.category))*male,data=heights.dt)
plot(fitted(fit_height_024),resid(fit_height_024)); abline(h=0,lty=2)
```



However, since this will affect our interpretability in the next question we will work with the logs instead. Again this decision depends on the purpose of the study. If the predictive accuracy is of interest, you may go with the quartic root.

4. Interpret all model coefficients.

- Intercept is log earning of female with average height and age category 1, which is less than 35 years old.
- With every inch increase in height female earning would be about 2% more.
- For female in age between 35 to 50 you'd expect the average log earning to be exp(0.35) = 1.42 times more than those less than 35.
- For female in age over 50 you'd expect the average log earning to be exp(0.16)=1.17 times more than those less than 35.
- Males make exp(0.29)=1.34 times more than female on average.
- There is no difference in effect of height for male compared to female.
- For a male in age between 35 to 50 you'd expect the average log earning to be exp(0.33)=1.39 times more than those less than 35.
- For a male in age over 50 you'd expect the average log earning to be exp(0.41)=`round(exp(0.41),2)` times more than those less than 35.

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

```
round(confint(fit_height_02),2)
```

```
##                              2.5 % 97.5 %
## (Intercept)                   9.26   9.50
## height_c                     -0.01   0.04
## factor(age.category)2         0.20   0.51
## factor(age.category)3        -0.03   0.36
## male                          0.10   0.47
## height_c:male                -0.04   0.04
## factor(age.category)2:male    0.10   0.57
## factor(age.category)3:male    0.12   0.71
```

What you see is that effect of height is inconclusive once you control for age and gender.

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.

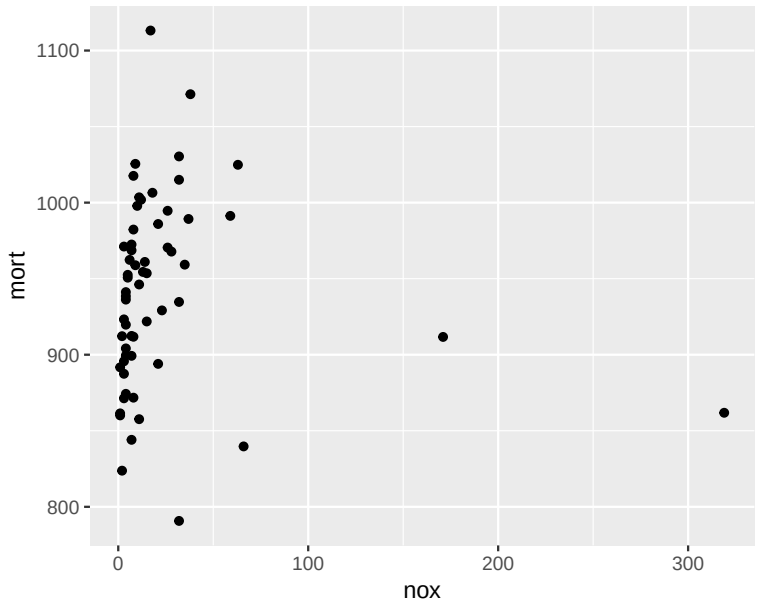
Variables, in order:

- 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

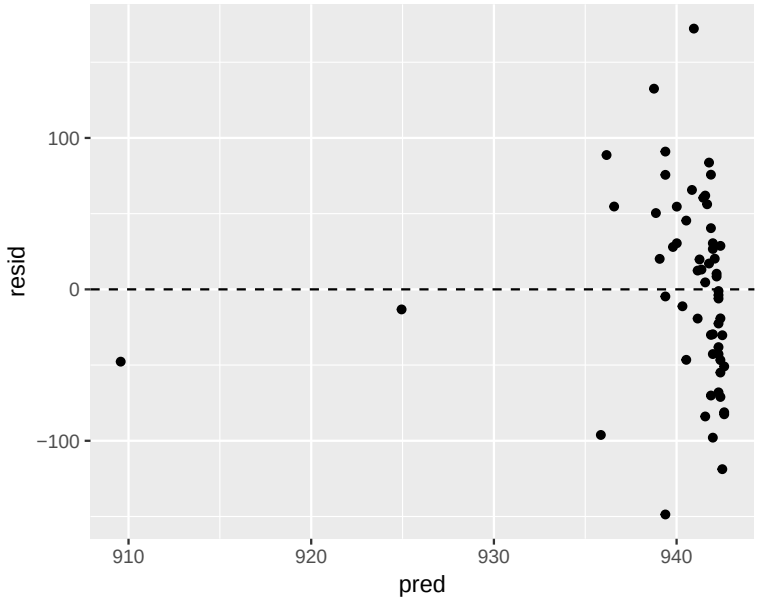
For this exercise we shall model mortality rate given nitric oxides, sulfur dioxide, and hydrocarbons as inputs. This model is an extreme oversimplification as it combines all sources of mortality and does not adjust for crucial factors such as age and smoking. We use it to illustrate log transformations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

1. Create a scatter plot 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.

```
gelman_dir <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
pollution <- data.table(read.dta (paste0(gelman_dir,"pollution/pollution.dta")))
ggplot(pollution)+geom_point()+aes(x=nox,y=mort)
```

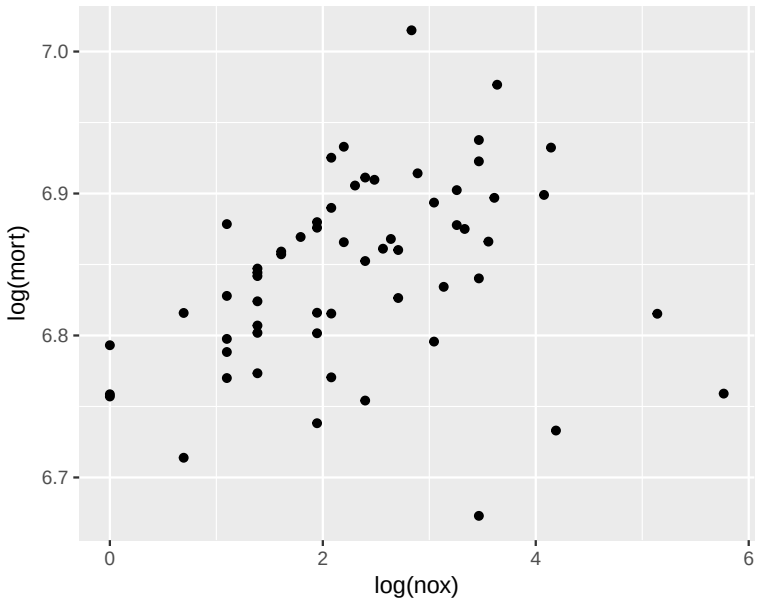



```
fit_mort_01<-pollution[,lm(mort~nox)]
pollution<-pollution[,resid:=resid(fit_mort_01)]
pollution<-pollution[,pred:=predict(fit_mort_01)]
ggplot(pollution)+geom_point()+aes(x=pred,y=resid)+geom_hline(yintercept=0,lty="dashed")
```

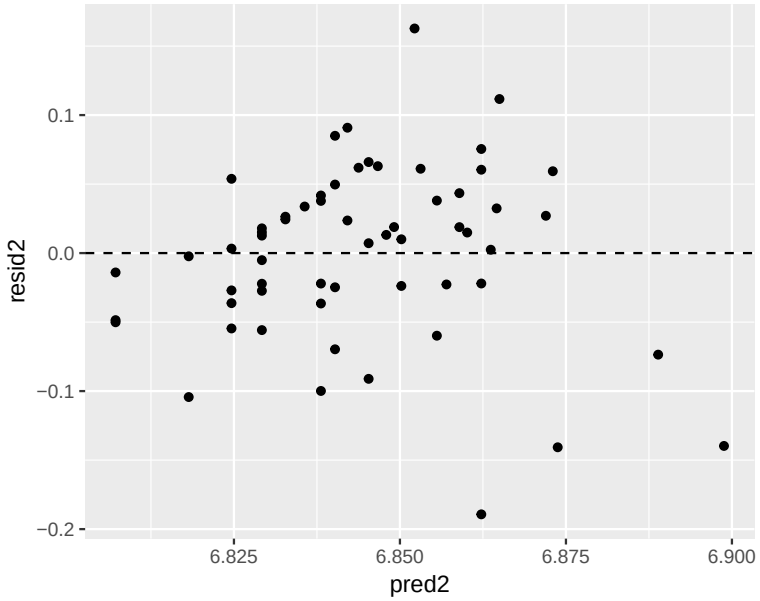


2. Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

```
ggplot(pollution)+geom_point()+aes(x=log(nox),y=log(mort))
```




```
pollution<-pollution[,log_nox_c:=scale(log(nox),center=TRUE,scale=FALSE)]
fit_mort_02<-pollution[,lm(log(mort)~log_nox_c)]
pollution<-pollution[,resid2:=resid(fit_mort_02)]
pollution<-pollution[,pred2:=predict(fit_mort_02)]
ggplot(pollution)+geom_point()+aes(x=pred2,y=resid2)+geom_hline(yintercept=0,lty="dashed")
```



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

```
display(fit_mort_02)
```

```
## lm(formula = log(mort) ~ log_nox_c)
##             coef.est coef.se
## (Intercept) 6.84     0.01
## log_nox_c   0.02     0.01
## ---
## n = 60, k = 2
```

```
## residual sd = 0.06, R-Squared = 0.08
```

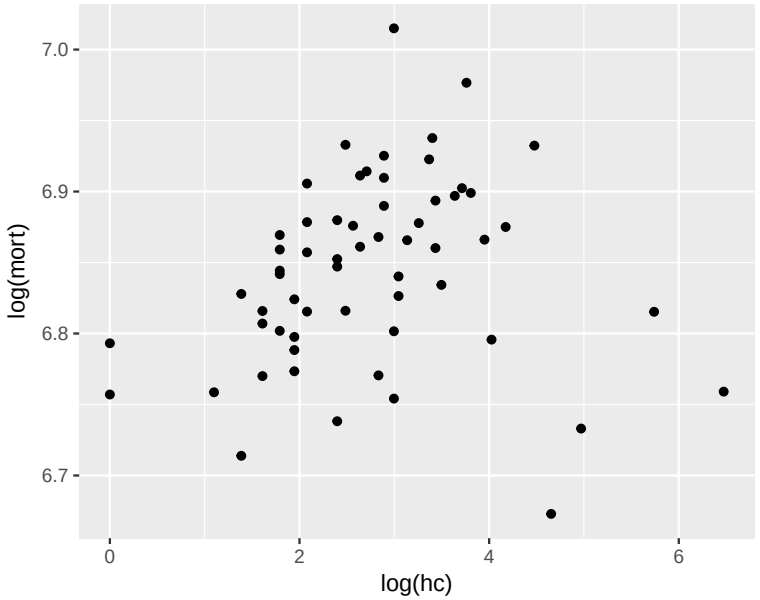
4. Construct 99% confidence interval for slope coefficient from the model you chose in 2 and interpret them.

```
round(confint(fit_mort_02,level=0.99),2)
```

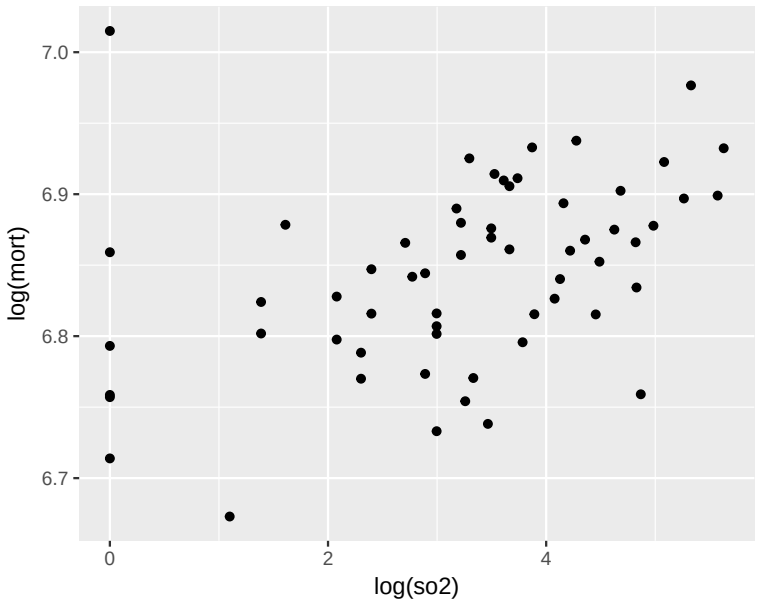
```
##             0.5 % 99.5 %
## (Intercept)  6.82   6.87
## log_nox_c    0.00   0.03
```

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.

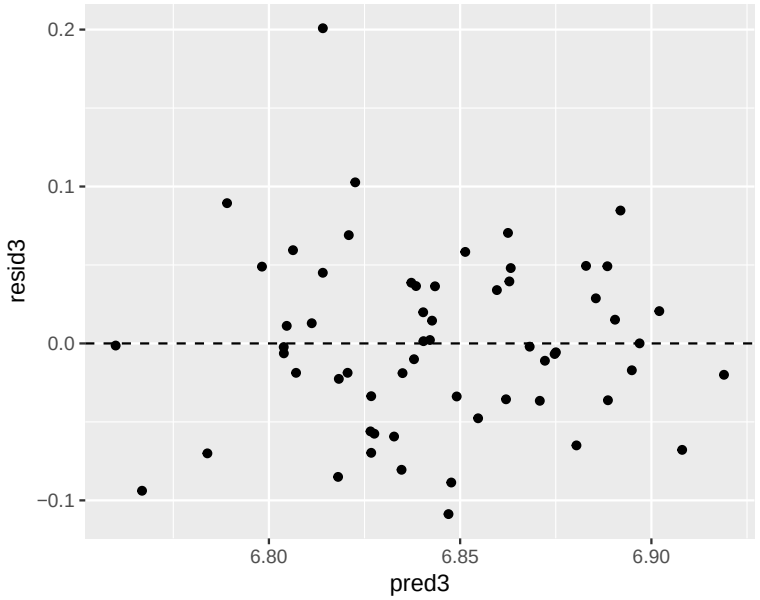
```
ggplot(pollution)+geom_point()+aes(x=log(hc),y=log(mort))
```




```
ggplot(pollution)+geom_point()+aes(x=log(so2),y=log(mort))
```

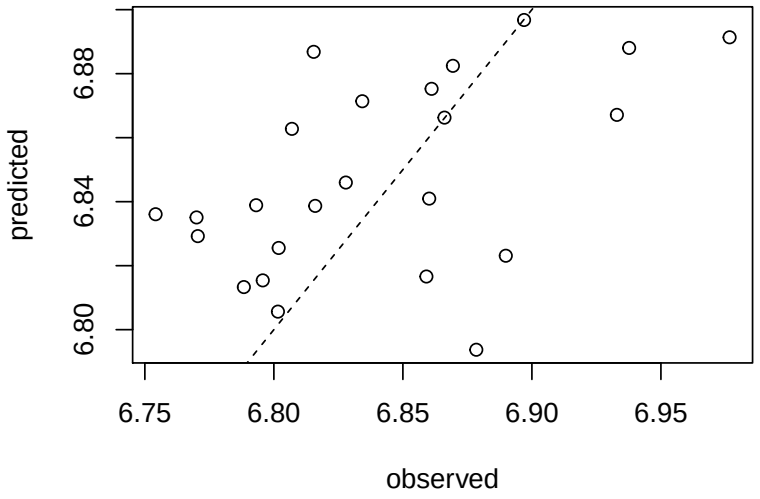


```
pollution<-pollution[,log_hc_c:=scale(log(hc),center=TRUE,scale=FALSE)]
pollution<-pollution[,log_so2_c:=scale(log(so2),center=TRUE,scale=FALSE)]
fit_mort_03<-pollution[,lm(log(mort)~log_nox_c+log_hc_c+log_so2_c)]
pollution<-pollution[,resid3:=resid(fit_mort_03)]
pollution<-pollution[,pred3:=predict(fit_mort_03)]
ggplot(pollution)+geom_point()+aes(x=pred3,y=resid3)+geom_hline(yintercept=0,lty="dashed")
```



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.)

```
n<-nrow(pollution)
set.seed(123456)
train_flag<- rbinom(n,1,0.5)
pollution <-pollution[,train:=train_flag]
fit_mort_04<-pollution[train_flag==1,lm(log(mort)~log_nox_c+log_hc_c+log_so2_c)]
plot(pollution[train_flag==0,log(mort)],predict(fit_mort_04,newdata=pollution[train_flag==0,]),xlab="obs
abline(0,1,lty=2)
```



Study of teenage gambling in Britain

```
library(faraway)
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.

```
status_dm <- teengamb$status-mean(teengamb$status) # difference from the mean status
income_dm <- teengamb$income-mean(teengamb$income) # difference from the mean income
verbal_dm <- teengamb$verbal-mean(teengamb$verbal) # difference from the mean verbal score
rg_g1 <- lm(gamble ~ sex + status_dm + income_dm + verbal_dm + sex:income_dm,
            data = teengamb)
summary(rg_g1)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status_dm + income_dm + verbal_dm +
##     sex:income_dm, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.109  -6.162  -0.938   2.267  86.503
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    28.45576    4.29594   6.624 5.61e-08 ***
## sex           -25.81559    7.62425  -3.386  0.00157 **
## status_dm      -0.04876    0.25978  -0.188  0.85203
## income_dm       6.19885    1.02591   6.042 3.77e-07 ***
## verbal_dm      -2.60864    1.99386  -1.308  0.19805
## sex:income_dm  -6.43683    2.14337  -3.003  0.00454 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.79 on 41 degrees of freedom
## Multiple R-squared:  0.6121, Adjusted R-squared:  0.5647
## F-statistic: 12.94 on 5 and 41 DF,  p-value: 1.417e-07
```

- The intercept predicts that the male teenage with average status, income and verbal score would spend 28.46 pounds on gambling per year.
- The coefficient of sex predicts the difference of gambling expenditure between female and male that both have the average income and the same level of status and verbal score. The female teenagers who have the average income spend 25.81559 pounds less than the male teenagers that have the average income and the same level of status and the same verbal score.
- The coefficient of status_dm(difference from the mean status) predicts that the 1 unit increase of the difference from average status of those who has the same sex, average income and verbal score would cause 0.04876 unit decrease of their expenditure on gambling in pounds per year.
- The coefficient of income_dm(difference from the mean income) predicts that the 1 unit increase of the difference from average income of those who has the same sex, average status and verbal score would cause 6.19885 unit increase of their expenditure on gambling in pounds per year.
- The coefficient of verbal_dm(difference from the mean verbal) predicts that the 1 unit increase of the difference from average verbal scores of those who has the same sex, average status and income would

cause 2.60864 unit decrease of their expenditure on gambling in pounds per year.
- The coefficient of the interaction term predicts the difference of the slop of income_dm for male and female when they have the same level of status and same verbal score. In such a case, the slop for female's income is 6.43683 smaller than the slop for male's income.

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

```
round(confint(rg_g1),2)
```

```
##                2.5 % 97.5 %
## (Intercept)    19.78  37.13
## sex           -41.21 -10.42
## status_dm      -0.57   0.48
## income_dm       4.13   8.27
## verbal_dm      -6.64   1.42
## sex:income_dm -10.77  -2.11
```

- There are 95% confidence that the interval(19.7799317, 37.1315889) contains the true value of the intercept.
- There are 95% confidence that the interval(-41.2130799, -10.4180961) contains the true value of the coefficient of sex.
- There are 95% confidence that the interval(-0.5734002, 0.4758751) contains the true value of the coefficient of status_dm and 0.
- There are 95% confidence that the interval(4.1269732, 8.2707189) contains the true value of the coefficient of income_dm.
- There are 95% confidence that the interval(-6.6353263, 1.4180392) contains the true value of the coefficient of verbal_dm and 0.
- There are 95% confidence that the interval(-10.7654601, -2.1081980) contains the true value of the coefficient of the interaction term.

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?

```
predict(rg_g1, data.frame(sex=0, status_dm=0, income_dm=0, verbal_dm=0),
        interval = 'confidence', level = 0.95)
```

```
##        fit      lwr      upr
## 1 28.45576 19.77993 37.13159
```

```
ms <- max(teengamb$status)-mean(teengamb$status)
mi <- max(teengamb$income)-mean(teengamb$income)
mv <- max(teengamb$verbal)-mean(teengamb$verbal)
predict(rg_g1, data.frame(sex=0, status_dm=ms, income_dm=mi, verbal_dm=mv),
        interval = 'confidence', level = 0.95)
```

```
##        fit      lwr      upr
## 1 82.49849 54.79436 110.2026
```

- The 95% CI for the amount a male with maximal values of status, income and verbal score would gamble along is wider than the 95% CI for male with average status, income and verbal score. The standard error for male with maximal values of status, income and verbal score is bigger than the standard error for male with average status, income and verbal score. Thus the 95% CIs are different.

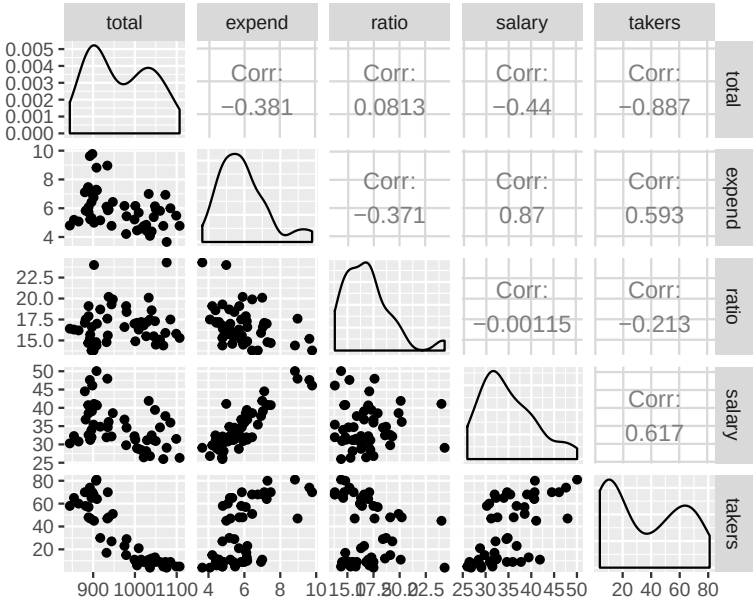
School expenditure and test scores from USA in 1994-95

```
data(sat)
?sat
sat_dt <- data.table(sat)
library(GGally)
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:faraway':
##
##     happy
```

```
ggpairs(sat_dt[,list(total,expend,ratio,salary,takers)])
```



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.

```
fit_sat_01<-sat_dt[,lm(total~expend+ratio+salary)]
display(fit_sat_01)
```

```
## lm(formula = total ~ expend + ratio + salary)
##             coef.est coef.se
## (Intercept) 1069.23   110.92
## expend        16.47    22.05
## ratio          6.33     6.54
## salary        -8.82     4.70
## ---
## n = 50, k = 4
## residual sd = 68.65, R-Squared = 0.21
```

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

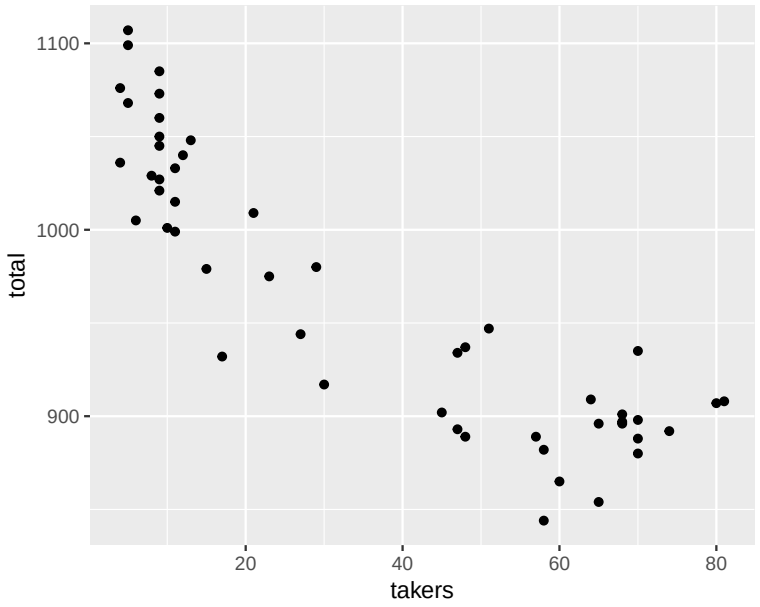
```
round(confint(fit_sat_01,level=0.98),2)
```

```
##                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?

```
ggplot(sat)+geom_point()+aes(x=takers,y=total)
```




```
fit_sat_02<-sat_dt[,lm(total~expend+ratio+salary+takers)]
round(confint(fit_sat_02,level=0.98),2)
```

```
##                1 %    99 %
## (Intercept) 918.44 1173.50
## expend      -20.98   29.90
## ratio       -11.38    4.13
## salary       -4.12    7.40
## takers       -3.46   -2.35
```

Conceptual exercises.

Special-purposetransformations:

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, $D_i/(D_i + R_i)$.

Answer:

- The simple difference, $D_i - R_i$
  - Advantage: symmetric and centered at zero, which is when $D_i = R_i$. The unit is in dollars so easy to interpret.
  - Disadvantage: You loose the scale since 7M and 5M will be the same as 2M and 0M. If absolute difference is what matters this may be a good transformation.
- The ratio, D_i/R_i
  - Advantage:
  - Disadvantage: Center is at 1 when $D_i = R_i$. Asymmetric and unstable when $R_i \approx 0$. Unit is defined relative to R_i so 20M and 1M is the same as 0.2M and 0.01M overly emphasizing the smaller district.
- The difference on the logarithmic scale, $logD_i - logR_i$
  - Advantage: this is same as $log(D_i/R_i)$ but if the outcome is on log scale the interpretation may be easier.
  - Disadvantage:
- The relative proportion, $D_i/(D_i + R_i)$.
  - Advantage: Symmetric and stable unless $D_i = R_i = 0$.
  - Disadvantage: Center is at 0.5 and unit become hard to interpret.
-
-
-

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 $\mathrm{x}' = \mathrm{x} - 10$ 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 $\mathrm{x}' = 10\mathrm{x}$? When $\mathrm{x}' = 10(\mathrm{x} - 1)$?

For $\mathrm{x}' = \mathrm{x} - 10$, $\hat{\alpha}' = 10$, $\hat{\beta}' = 0.9$, $\hat{\sigma} = 0.03$, and $r' = 0.3$ For $\mathrm{x}' = 10\mathrm{x}$, $\hat{\alpha}' = 1$, $\hat{\beta}' = 0.09$, $\hat{\sigma} = 0.03$, and $r' = 0.3$ For $\mathrm{x}' = 10(\mathrm{x} - 1)$, $\hat{\alpha}' = 1.9$, $\hat{\beta}' = 0.09$, $\hat{\sigma} = 0.03$, and $r' = 0.3$

2. Now suppose that the response variable scores are transformed according to the formula $\mathrm{y}'' = \mathrm{y} + 10$ 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 $\mathrm{y}'' = 5\mathrm{y}$? When $\mathrm{y}'' = 5(\mathrm{y} + 2)$?

For $\mathrm{y}'' = \mathrm{y} + 10$, $\hat{\alpha}'' = 11$, $\hat{\beta}'' = 0.9$, $\hat{\sigma}'' = 0.03$, and $r'' = 0.3$ For $\mathrm{y}'' = 5\mathrm{y}$, $\hat{\alpha}'' = 5$, $\hat{\beta}'' = 4.5$, $\hat{\sigma}'' = 0.15$, and $r'' = 0.3$ For $\mathrm{y}'' = 5(\mathrm{y} + 2)$, $\hat{\alpha}'' = 15$, $\hat{\beta}'' = 4.5$, $\hat{\sigma} = 0.15$, and $r'' = 0.3$

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

For $X' = a(X - b)$, $\hat{\alpha}' = \alpha + b\beta$, $\hat{\beta}' = \beta/a$, $\hat{\sigma} = \sigma$, and $r' = r$ For $Y' = a(Y + b)$, $\hat{\alpha}' = a(\alpha + b)$, $\hat{\beta}' = a\beta$, $\hat{\sigma} = a\sigma$, and $r' = r$

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

For $X^{'} = a(X - b)$, $B^{'} = \frac{B}{a}$, $S_E^{'} = S_E$, $\therefore SE(B^{'}) = \sqrt{\frac{S_e^{'2}}{\sum (X_i^{'} - \bar{X}')^2}} = \sqrt{\frac{S_e^2}{a^2 \sum (X_i - \bar{X})^2}} = \frac{SE}{a}$ $\therefore t_0^{'} = \frac{B^{'}}{SE(B')} = \frac{\frac{B}{a}}{\frac{SE}{a}} = t_0$

Thus: When $X^{'} = 10(X - 1)$, $a = 10$ We can have $SE(B^{'}) = \frac{SE(B)}{10}$, $t_0^{'} = t_0 \backslash$

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

For $Y^{"} = aY + b$, $B^{'} = aB$, $S_E^{'} = aS_E$, $\therefore SE(B^{'}) = \sqrt{\frac{S_e^{'2}}{\sum (X_i - \bar{X})^2}} = \sqrt{\frac{a^2 S_e^2}{\sum (X_i - \bar{X})^2}} = aSE$ $\therefore t_0^{'} = \frac{B^{'}}{SE(B')} = \frac{aB}{aSE} = t_0$

Thus: When $Y^{"} = 5(Y + 2)$, $a = 5$ We can have $SE(B^{'}) = 5SE$, $t_0^{'} = t_0$

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

The general results for how are hypothesis tests and confidence intervals for β affected by linear transformations of X and Y have been given at the first parts of (a) and (b).

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.