Homework 02

Weiling Li

Septemeber 21, 2019

Introduction

In homework 2 you will fit many regression models. You are welcome to explore beyond what the question is asking you.

Please come see us we are here to help.

Data analysis

Analysis of earnings and height data

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

```r
gelman_dir <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
heights    <- read.dta (paste0(gelman_dir,"earnings/heights.dta"))
#wfw90    <- read.table (paste0(gelman_dir,"earnings/wfw90.dat"))
```

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

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

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

- earn: personal income during year 1989, in dollars.
- height1: height in inches
- height2: height in inches
- sex:
  - male: 1
  - female: 2
- race:
  - white: 1
  - black: 2
  - asian: 3
  - native amerian: 4
  - others: 5
- hisp:
  - hispanic origin: 1
  - otherwise: 2
- ed: highest grade or years in school(highest grade converted to years in school) from 0 to 18, integers.
- yearbn: year of born in 19xx.

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

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

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

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

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

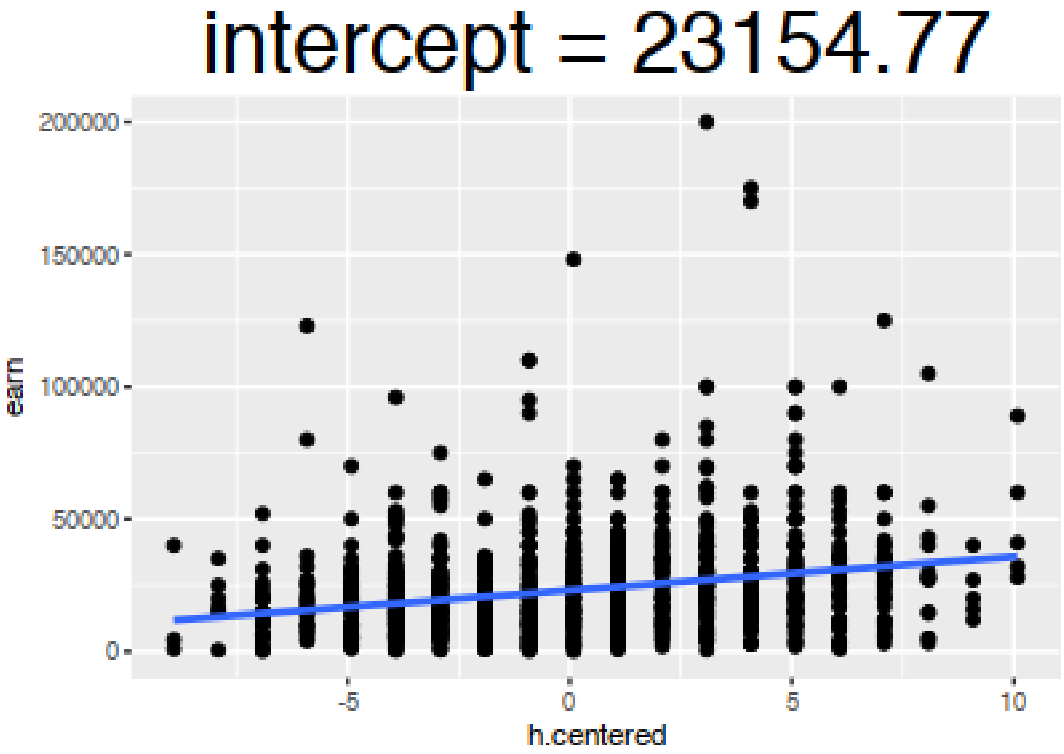
```
#hist(log(h1$earn),probability = T)
#hist(h1$ed,probability = T)
#hist(h1$height,probability = T)
```

2. Fit a linear regression model predicting earnings from height. What transformation should you perform in order to interpret the intercept from this model as average earnings for people with average height?

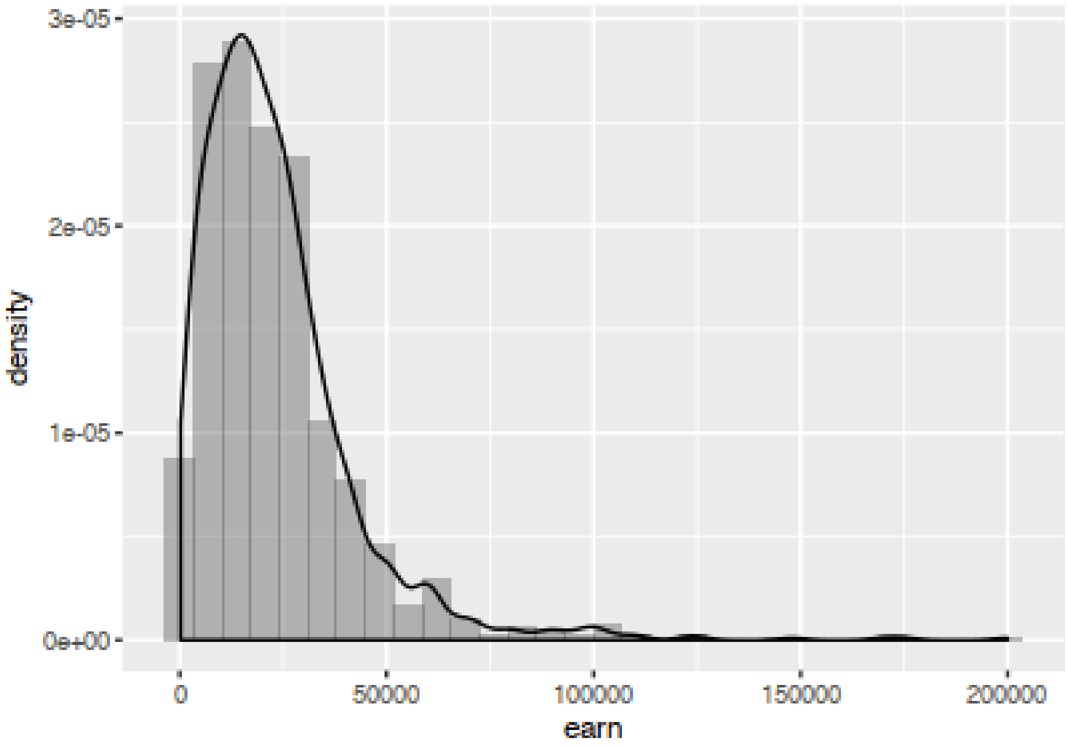
```
#to achieve what exactly asked. we only need to subtract height with mean.
h1 <- mutate(h1,h.centered = height - mean(h1$height))
```

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

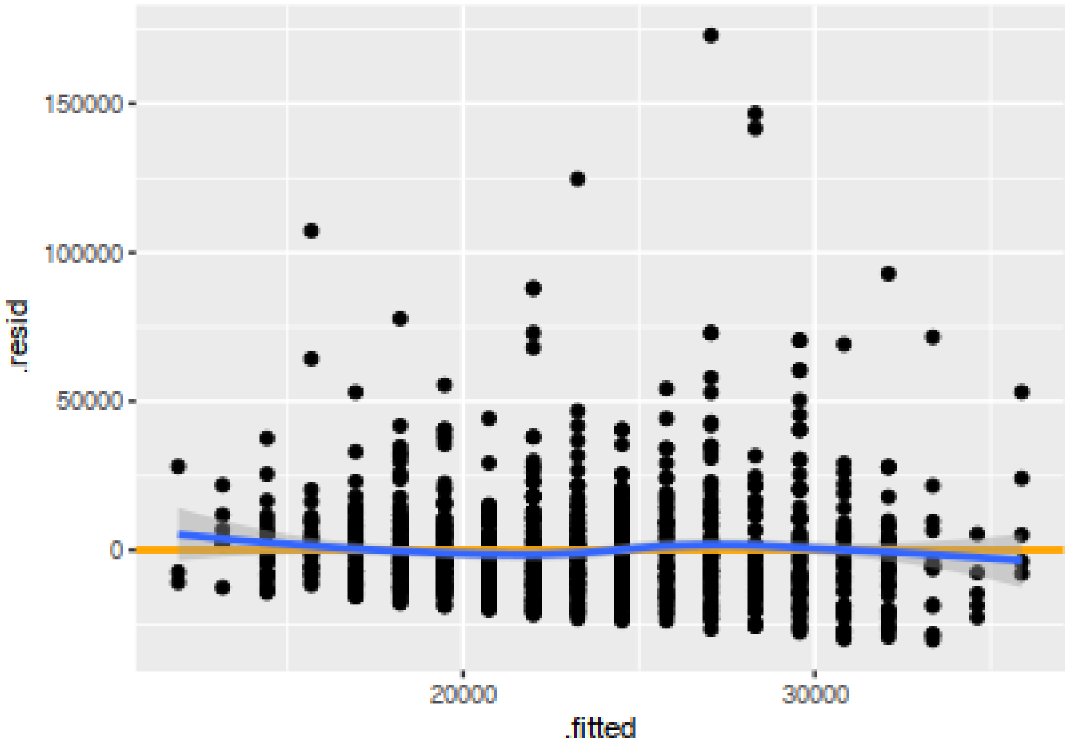
```
#hist(height.centered)
#hist(h1$height)
ggplot(h1)+
  aes(x = h.centered, y = earn)+geom_point()+geom_smooth(method = 'lm',se = F)+ggtitle(paste0('intercep
```



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



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



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

3. Fit some regression models with the goal of predicting earnings from some combination of sex, height, and age. Be sure to try various transformations and interactions that might make sense. Choose your preferred model and justify.
   census data: us 1990 census

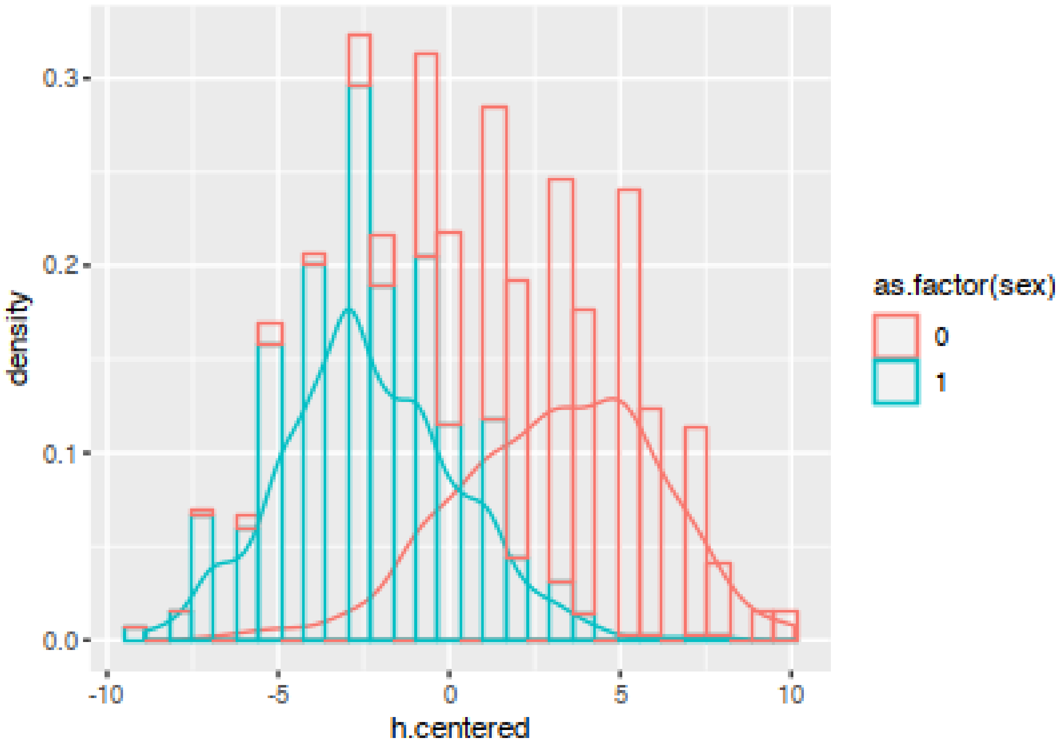
```
## for better interpretation and simplicity, the following transformation of the data set is done as fo
## take log transformation of the earning
## substract sex with 1 to make the variable a binary with 0 indicate male and 1 indicate female.
## 89 - yearbn to get the approximate age for the interviewee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

log.earn	sex	h.centered	age
8.99	1	-2.92	18
11.00	1	-2.92	42
5.30	1	0.08	32
8.52	1	0.08	18
11.16	1	-4.92	37
10.37	0	1.08	29
9.90	0	2.08	57
10.28	1	-4.92	48
10.13	0	0.08	25
10.09	1	-1.92	26

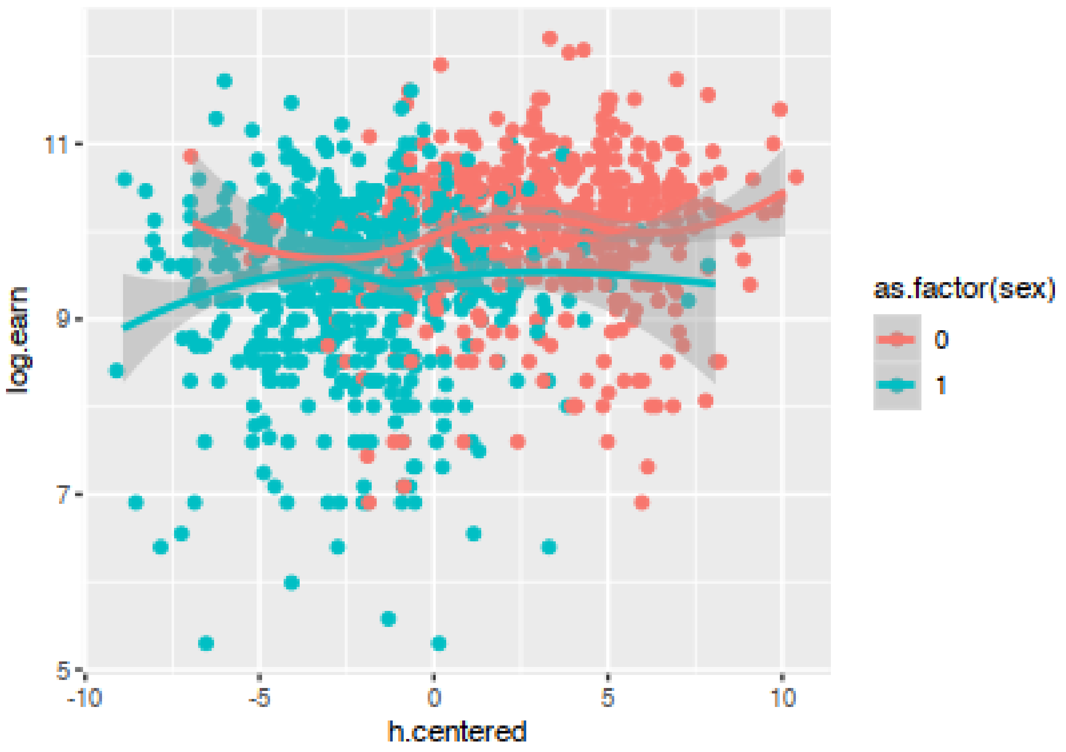
```
##check race factor:
kable(group_by(h.transformed,race)%>%summarise(count.race = n()),format = 'latex')
```

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

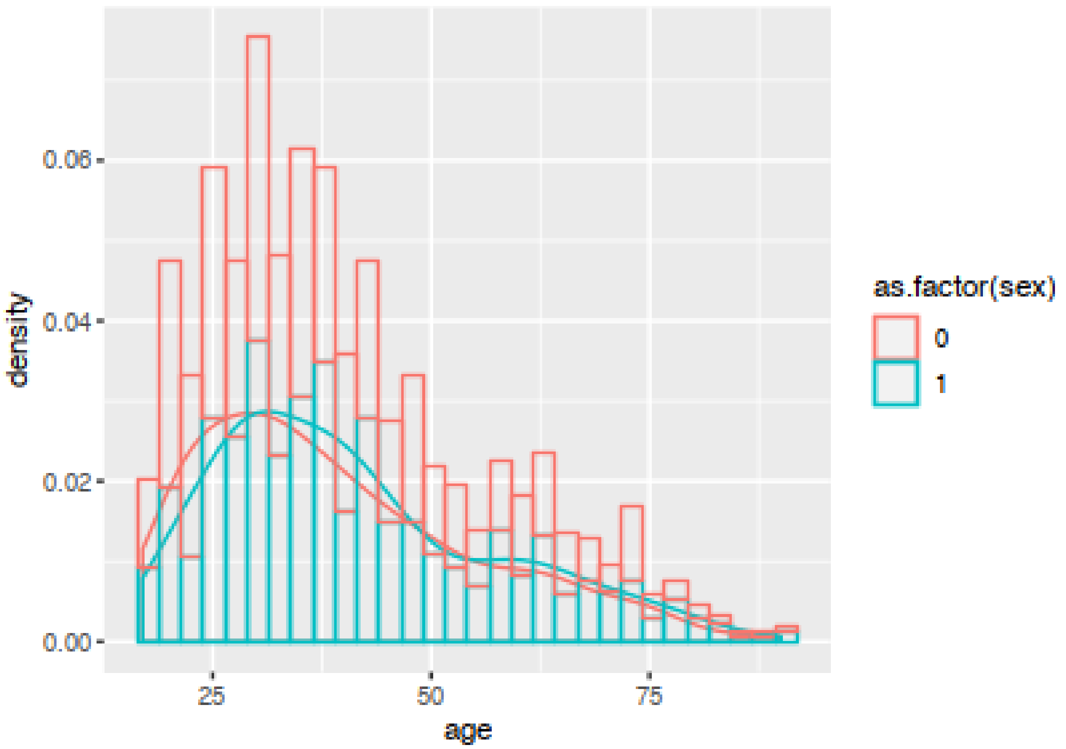
```
## as can be seen the majority interviewees are white, less than 10% are black and less than 2% are oth
## The 1990 census data shows that about 80% american population is white, 12% is black and the remaini
#ggplot(h.transformed)+aes(x = h.centered,y = log.earn,color = as.factor(sex))+geom_point()+geom_smooth

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



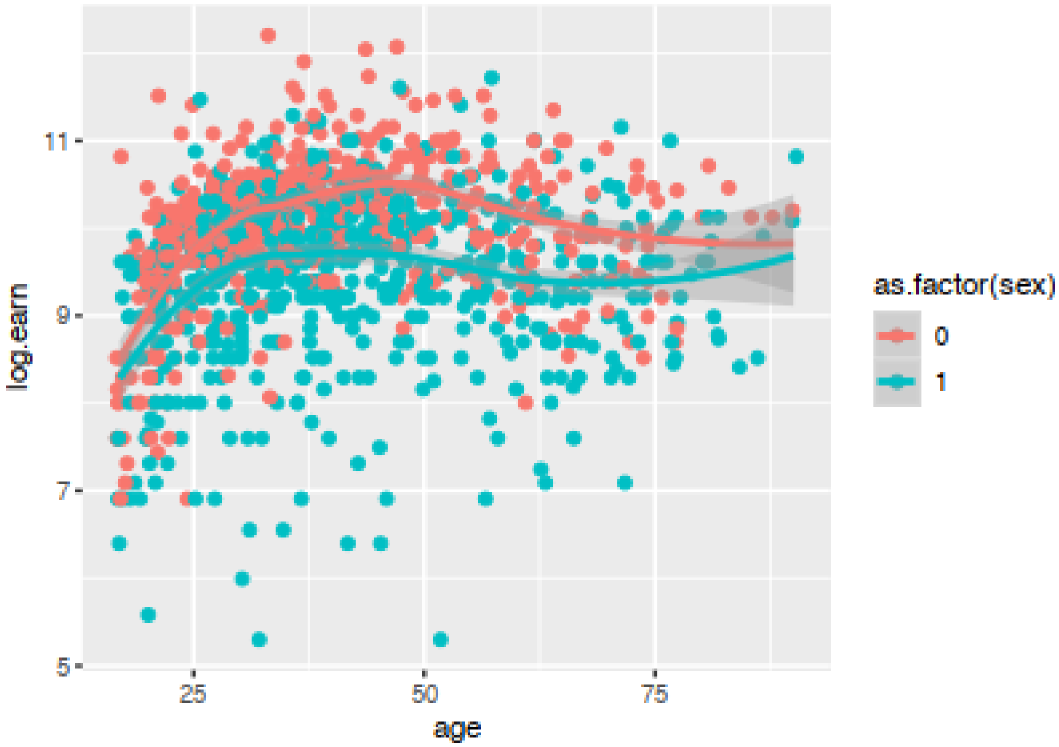
```
ggplot(h.transformed)+aes(x = h.centered,y = log.earn,color = as.factor(sex))+geom_jitter()+geom_smooth
```



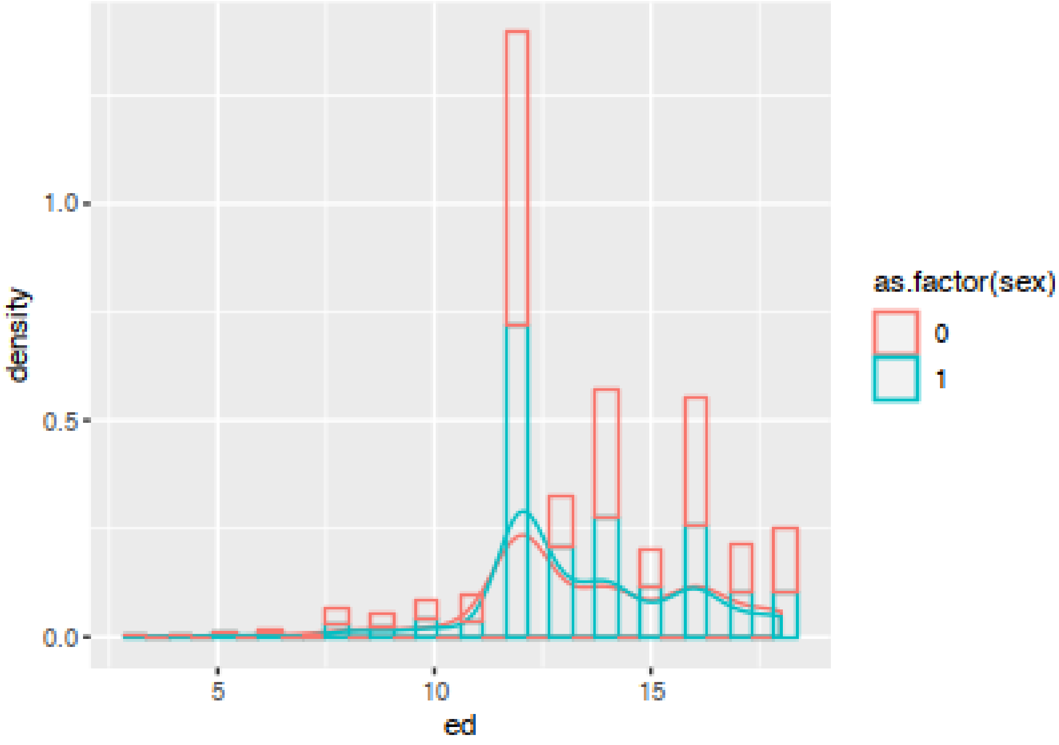
```
## age
ggplot(h.transformed)+aes(x = age,color = as.factor(sex))+geom_density( alpha = .4)+geom_histogram(bins
```



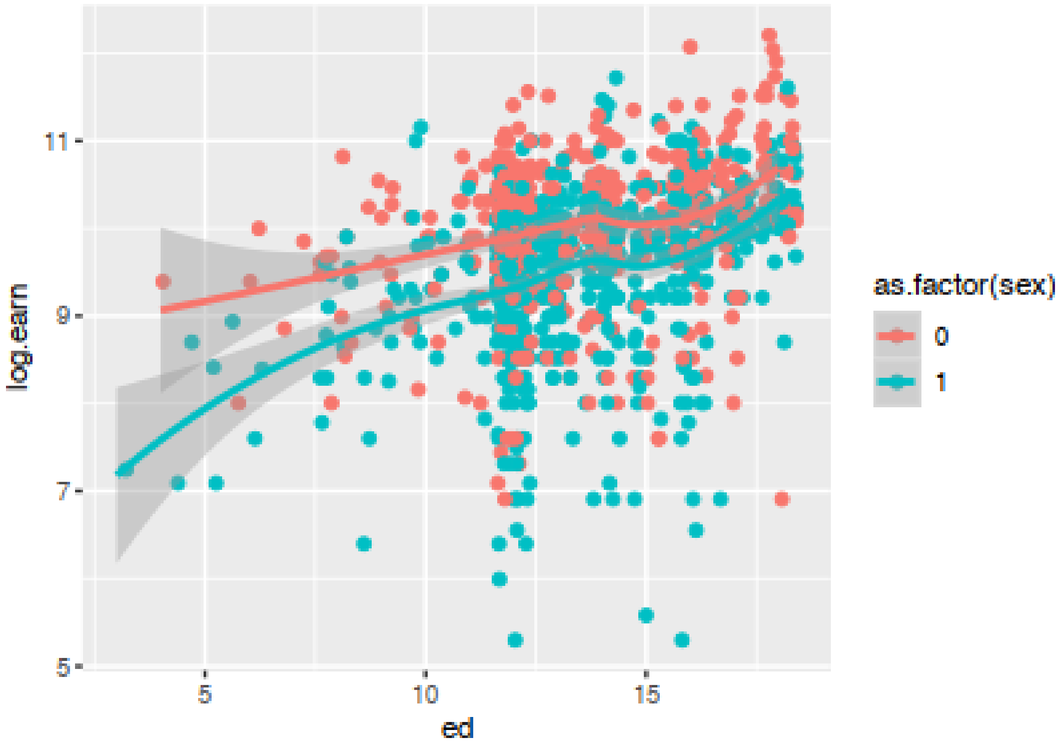
```
ggplot(h.transformed)+aes(x = age,y = log.earn,color = as.factor(sex))+geom_jitter()+geom_smooth(method
```



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



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



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

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

```
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.
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

```
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.
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.