MA678 Homework 2

Megha Pandit

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

```
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.

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

#Removing the data points for which earnings are NA
d <- na.omit(dt)

#Removing data points for which the year born is after 1990
d <- d[d$yearbn < 90,]

#Changing the yearbn column to age for easier interpretation
d$yearbn <- 90 - d$yearbn
names(d)[8] <- paste("age")

#Removing the height1 and height2 columns since they are redundant
d <- d[-c(2,3)]

#Factorizing the education variable into categories
d$ed <- d$ed[which(d$ed != 98)]
d$ed <- d$ed[which(d$ed != 99)]

#Removing data points with zero earnings
d <- d[which(d$earn != 0),]
rownames(d) <- 1:nrow(d)
```

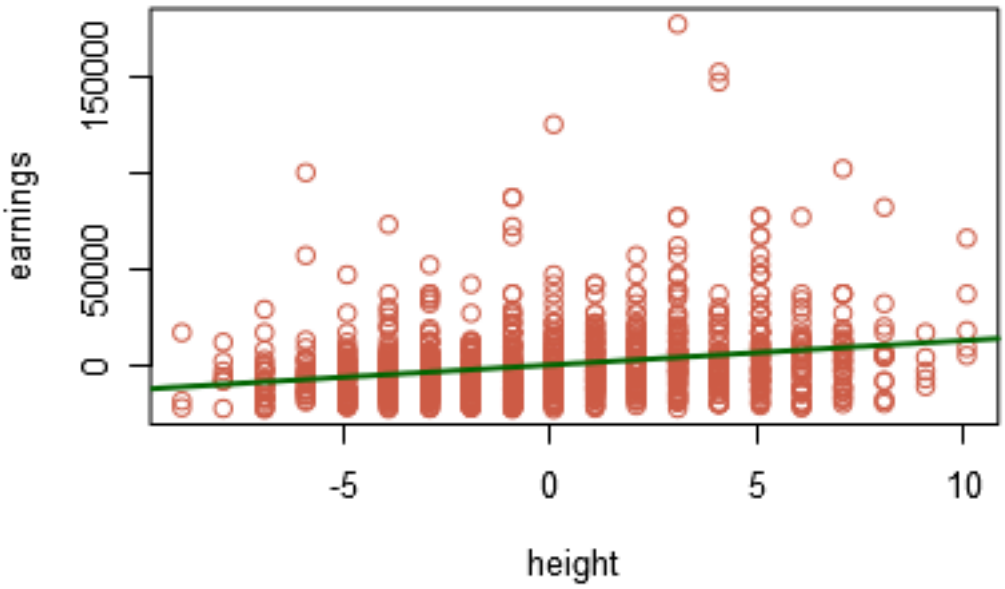
2. Fit a linear regression model predicting earnings from height. What transformation should you perform in order to interpret the intercept from this model as average earnings for people with average height?

```r
#Fitting a regression model for predicting earnings from height
fit <- lm(d$earn ~ d$height)
summary(fit)
```

```
##
## Call:
## lm(formula = d$earn ~ d$height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30211 -11318  -3403   6579 172953
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -61929.5     9540.7  -6.491 1.25e-10 ***
## d$height      1271.1      142.3   8.930  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18870 on 1187 degrees of freedom
## Multiple R-squared:  0.06295,    Adjusted R-squared:  0.06216
## F-statistic: 79.74 on 1 and 1187 DF,  p-value: < 2.2e-16
```

To interpret the intercept as average earnings for people with height, we can center the earnings and the height by subtracting their respective means from their data points. Therefore, centering the earnings and height by subtracting their means from the data points, we get

```r
earn_c <- d$earn - mean(d$earn)
height_c <- d$height -mean(d$height)
lm_c <- lm(earn_c ~ height_c)
plot(height_c, earn_c, col = "coral3", xlab = "height", ylab = "earnings")
abline(lm_c, col = "darkgreen", lwd = 2)
```



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.

```
#Fitting a regression model for predicting earnings from a combination of sex, height and age
lm1 <- lm(earn ~ height + age + sex, data = d)
summary(lm1)
```

```
##
## Call:
## lm(formula = earn ~ height + age + sex, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31859 -11174  -2626   6527 171047
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5004.79   15223.24  -0.329 0.742395
## height        547.12     197.23   2.774 0.005623 **
## age           132.75      34.55   3.843 0.000128 ***
## sex         -8853.87    1522.87  -5.814 7.84e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18490 on 1185 degrees of freedom
## Multiple R-squared:  0.1014, Adjusted R-squared:  0.09909
## F-statistic: 44.55 on 3 and 1185 DF,  p-value: < 2.2e-16
```

```r
#Transforming the earnings into log earnings
lmg <- lm(log(earn) ~ height + age + sex, data = d)
```

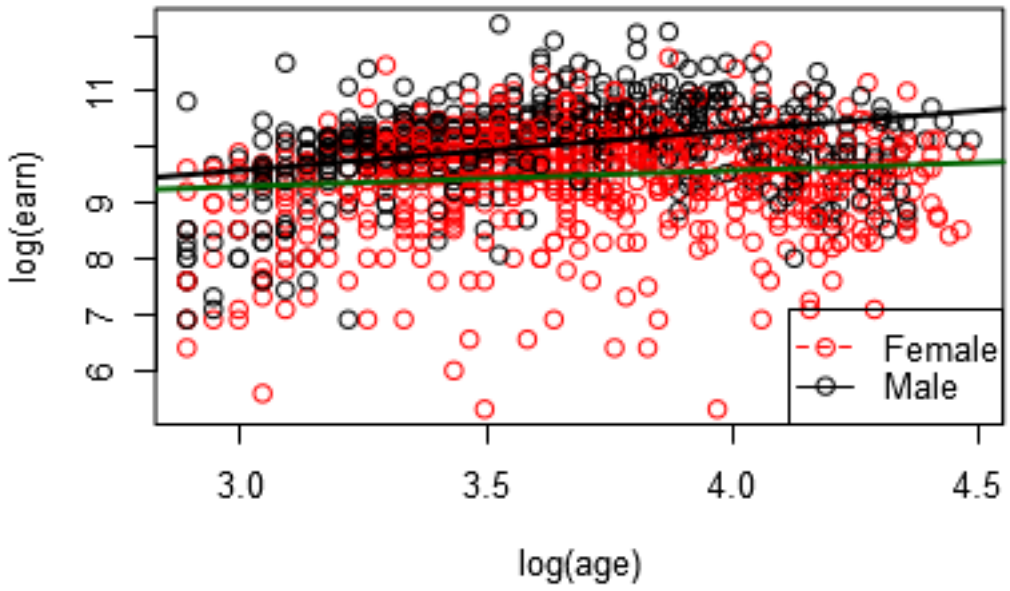
```r
#Plotting the earnings against height while differentiating between 'Male' and 'Female'
#plot(d$height, log(d$earn), col = factor(d$sex))
#abline(lm(log(d$earn[which(d$sex == 1)]) ~ d$height[which(d$sex == 1)]))
#abline(lm(log(d$earn[which(d$sex == 2)]) ~ d$height[which(d$sex == 2)]))
#hist(resid(lm(d$earn ~ d$height)))
#summary(lm(d$earn ~ d$height))

#Considering the interaction of the height and sex variables
#lm_2 <- lm(log(d$earn) ~ d$height + d$age + d$sex + d$height:d$sex)
#summary(lm_2)
#hist(resid(lm_2))
#plot(lm_2)
```

Considering the interaction between height and sex variables does not explain the variation in earnings very well, since the coefficient estimates of height and sex and even the interaction coefficient do not seem to be statistically significant.

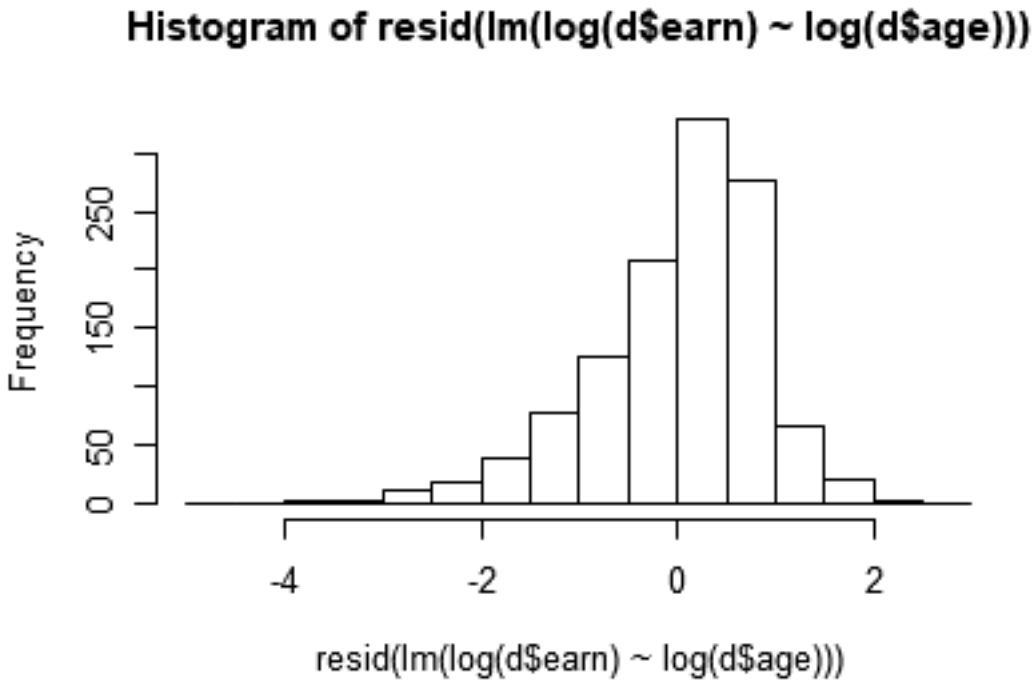
Therefore, Plotting the earnings against age while categorizing the sex variable into 'Male' and 'Female', we see that the difference in slopes for males and females is very distinct. Hence, we can consider an interaction between the age and sex variables.

Owing to the large variation in earnings, i.e., from 200 - 200,000 dollars, a log transformation on the earnings would improve the fit and also make the interpretability easier. Also, since the ages have quite a range of variability, performing a log transformation on the ages may make interpreting the coefficients easier.

```r
plot(log(d$age), log(d$earn), col = factor(d$sex), xlab = "log(age)", ylab = "log(earn)")
abline(lm(log(d$earn[which(d$sex == 1)]) ~ log(d$age[which(d$sex == 1)])), lwd = 2)
abline(lm(log(d$earn[which(d$sex == 2)]) ~ log(d$age[which(d$sex == 2)])), col = "darkgreen", lwd = 2)
legend("bottomright", c("Female", "Male"), lty = c(2,1), pch = c(1,1), col = c("red", "black") )
```




```
hist(resid(lm(log(d$earn) ~ log(d$age))), breaks = 20)
```

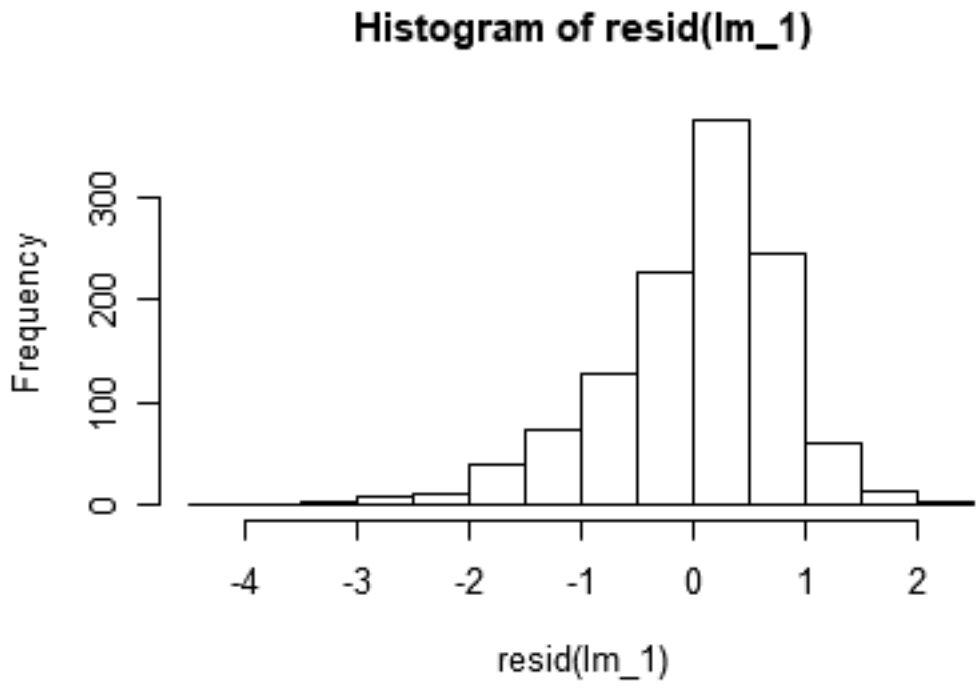



```
#Regressing log of earnings on height, age and sex, considering the interaction between age and sex,
height_c <- (d$height - mean(d$height))/sd(d$height)
age_c <- d$age - 17
```

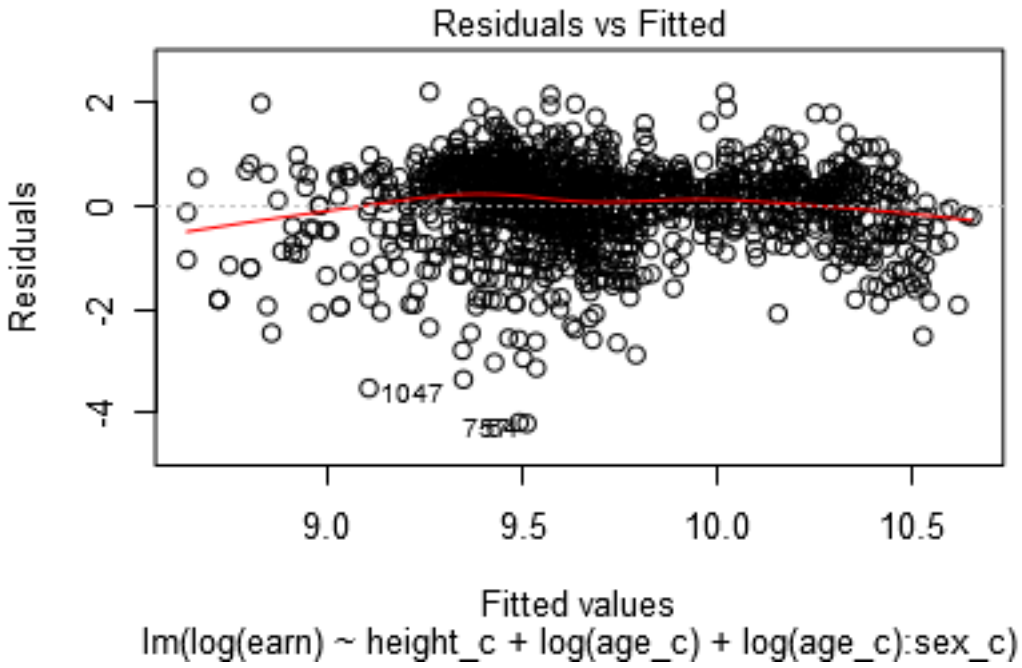
```r
sex_c <- d$sex - 1
lm_1 <- lm(log(earn) ~ height_c + log(age_c) + log(age_c):sex_c, data = d)
summary(lm_1)
```

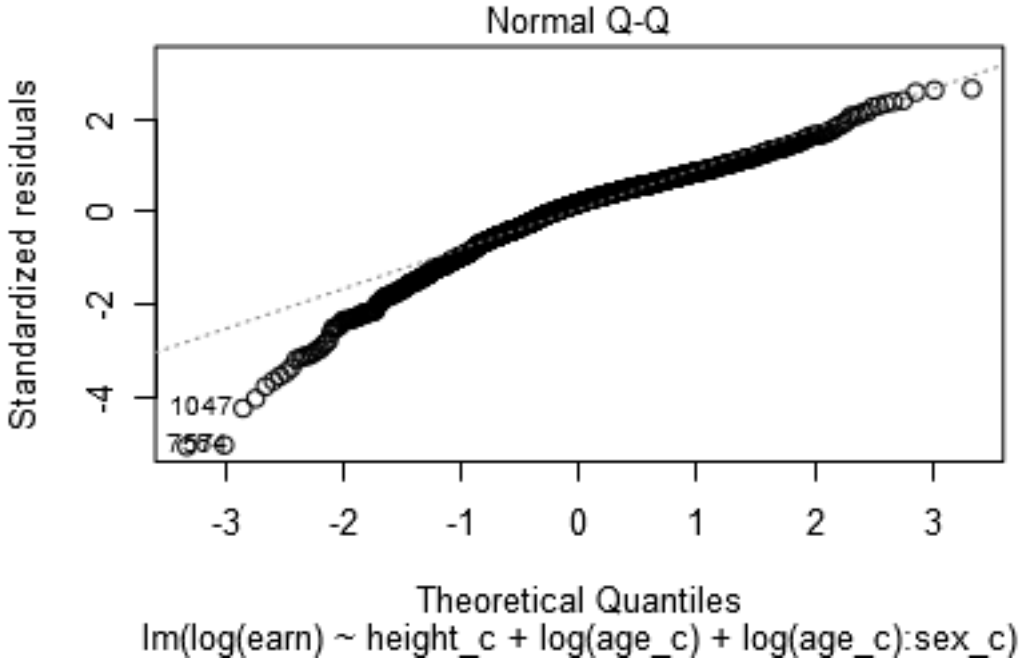
```
##
## Call:
## lm(formula = log(earn) ~ height_c + log(age_c) + log(age_c):sex_c,
##     data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2133 -0.4249  0.1573  0.5379  2.2096
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.77485    0.09080  96.635  < 2e-16 ***
## height_c           0.10410    0.03331   3.125  0.00182 **
## log(age_c)         0.40956    0.03206  12.774  < 2e-16 ***
## log(age_c):sex_c -0.15171    0.02218  -6.840 1.26e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8364 on 1185 degrees of freedom
## Multiple R-squared:  0.178,  Adjusted R-squared:  0.1759
## F-statistic: 85.51 on 3 and 1185 DF,  p-value: < 2.2e-16
```

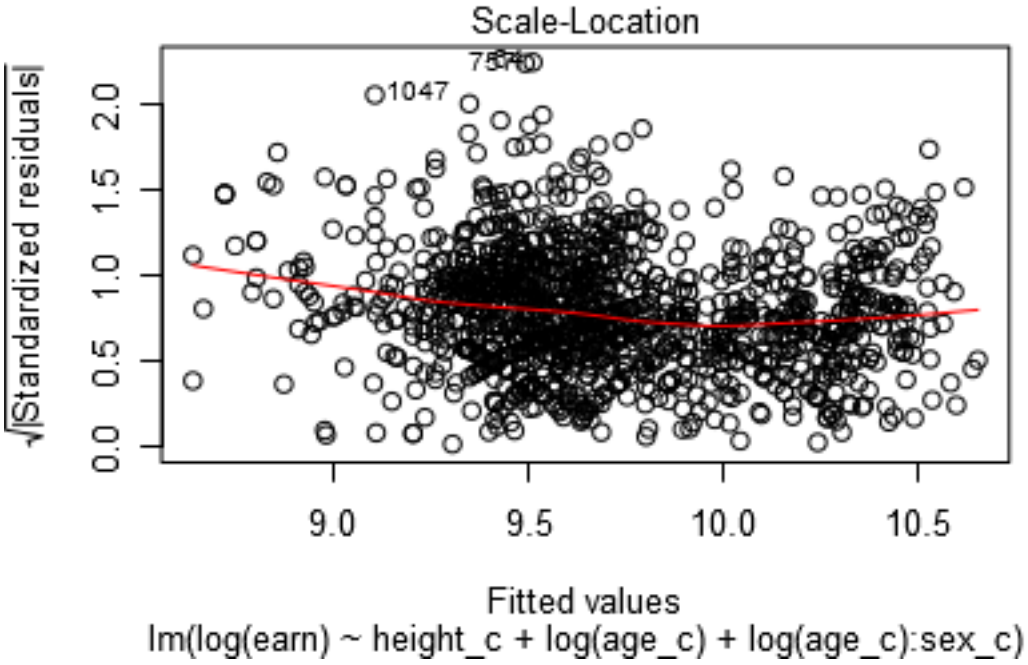
```r
hist(resid(lm_1))
```

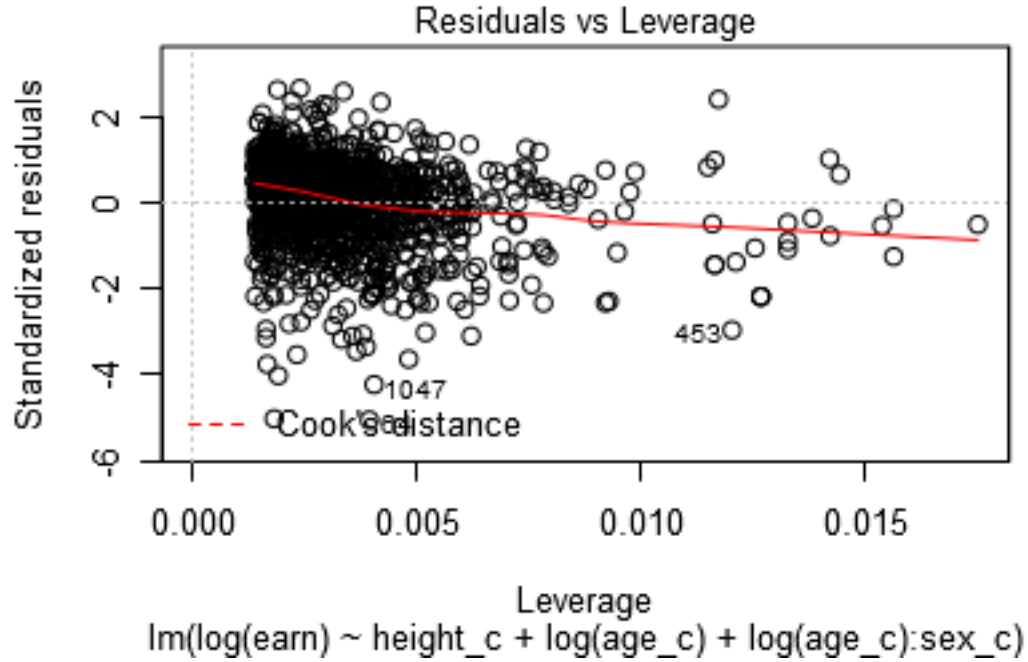



```r
plot(lm_1)
```


Residuals vs Fitted
Residuals
2
0
-2
-4
1047
9.0
9.5
10.0
10.5
Fitted values
lm(log(earn) ~ height_c + log(age_c) + log(age_c):sex_c)


Normal Q-Q
Standardized residuals
2
0
-2
-4
1047
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(log(earn) ~ height_c + log(age_c) + log(age_c):sex_c)





Having an R-Squared value of around 0.178,this model explains the 17.8% of the variation in earnings. The intercept and slope coefficients have small standard errors relative to their estimates and are statistically significant. Though the coefficient of sex is not statistically significant, we can keep it because the coefficient of interaction between log age and sex explains

a decent amount of the variation in log earnings.

4. Interpret all model coefficients.

Intercept The intercept is the predicted log earnings if height and sex are zero, and age is 1. It does not make sense to consider the height being zero. Therefore, the heights can be scaled to have a mean of 0 and standard deviation of 1. The ages can be centered around 17. Therefore, for a male of 18 years of age whose height is 66.91 inches, the predicted log earnings is 8.77. Predicted earnings = exp of 8.70 = 6438.17

height Coefficient The coefficient of height is the difference in the predicted log earnings for a difference of one standard deviation in height. For a male of age 18 and a difference of 3.84 inches in height, the estimated predictive difference in earnings is 10.4%

log age Coefficient The coefficient of log age is the predicted difference in earnings for a 1% difference in age. For a 10% difference in age, the predicted earnings differ by 4%.

Interaction Coefficient The log age sex interaction coefficient is the difference in slopes predicting the log earnings on age, comparing males to females. An increase in the age corresponds to a decrease in the earnings while going from males to females.

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

```
confint(lm_1, level = 0.95)
```

```
##                         2.5 %     97.5 %
## (Intercept)       8.59669949  8.9530089
## height_c          0.03874267  0.1694637
## log(age_c)        0.34665504  0.4724695
## log(age_c):sex_c -0.19523002 -0.1081991
```

Since none of the confidence intervals cross 0, there is evidence that the predictor variables and the response variable are related. We can be confident that if we perform the same regression 100 times, in 95 out of the 100 times, the above intervals will contain the true values of the coefficients we estimated.

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

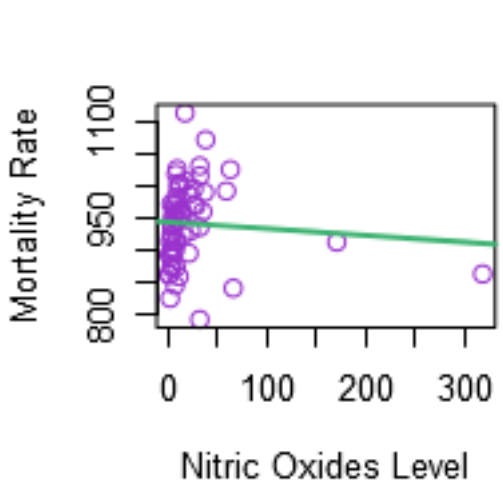
- HUMID Annual average % relative humidity at 1pm
- MORT Total age-adjusted mortality rate per 100,000

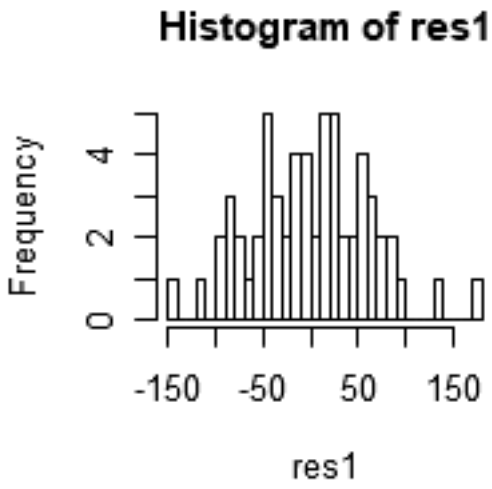
For this exercise we shall model mortality rate given nitric oxides, sulfur dioxide, and hydrocarbons as inputs. This model is an extreme oversimplification as it combines all sources of mortality and does not adjust for crucial factors such as age and smoking. We use it to illustrate log transformations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

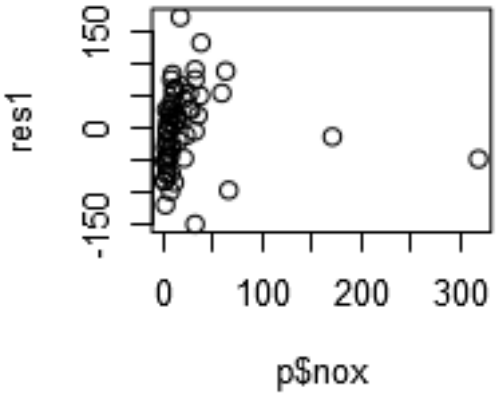
```
gelman_dir   <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
pollution    <- read.dta (paste0(gelman_dir,"pollution/pollution.dta"))
```

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.

```
gelman_dir   <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
p <- read.dta (paste0(gelman_dir,"pollution/pollution.dta"))
lm <- lm(p$mort ~ p$nox)
plot(p$nox, p$mort, col = "darkorchid3", xlab = "Nitric Oxides Level", ylab = "Mortality Rate")
abline(lm, col = "mediumseagreen", lwd = 2)
res1 <- resid(lm)
hist(res1, breaks = 30)
plot(p$nox, res1)
```





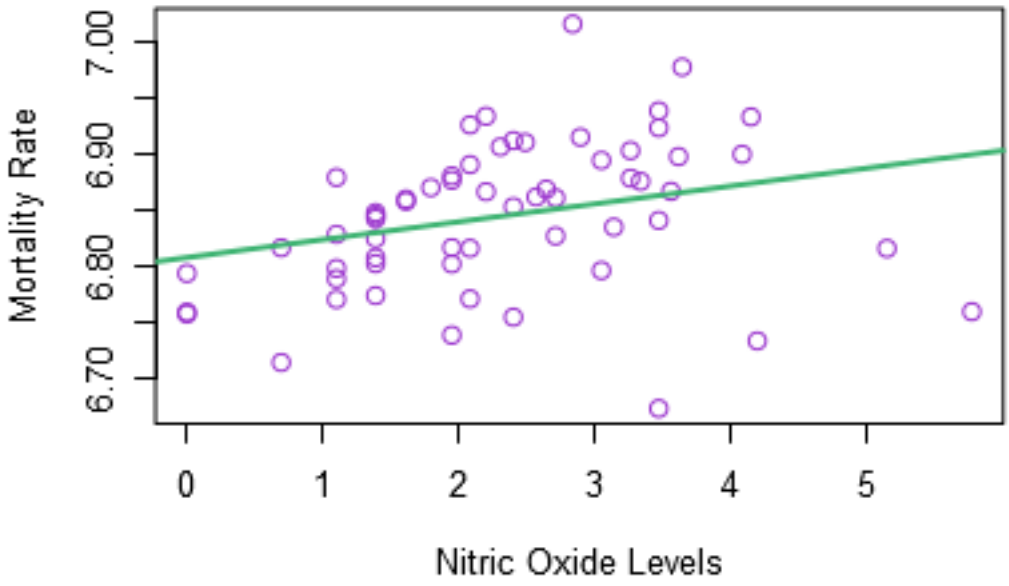


Linear regression does not fit the data well as seen from the regression plot of mortality rate on nitric oxide levels.

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.

Taking log of the nitric oxide levels and mortality rates makes the data more appropriate for linear regression.

```
lm_n <- lm(log(mort) ~ log(nox), data = p)
plot(log(p$nox), log(p$mort), col = "darkorchid3", xlab = "Nitric Oxide Levels", ylab = "Mortality Rate
abline(lm_n, col = "mediumseagreen", lwd = 2)
```

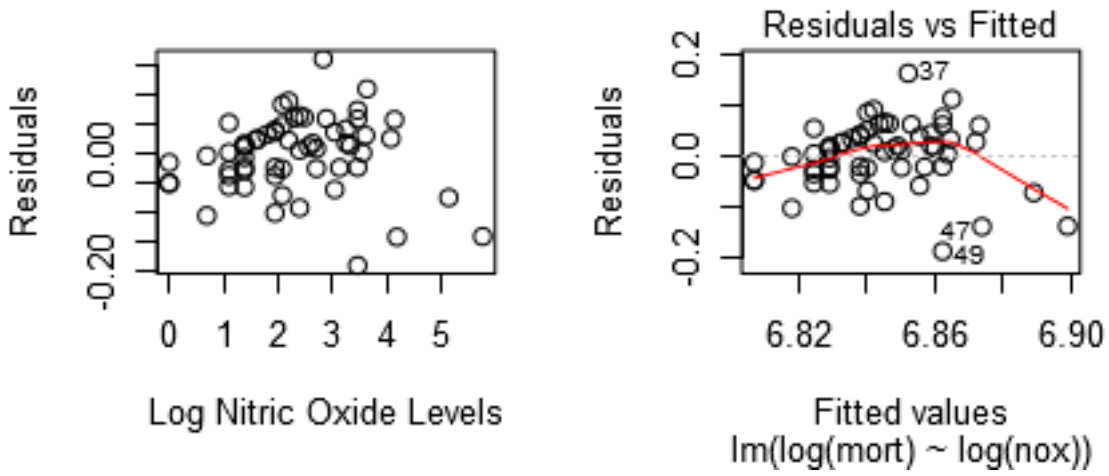


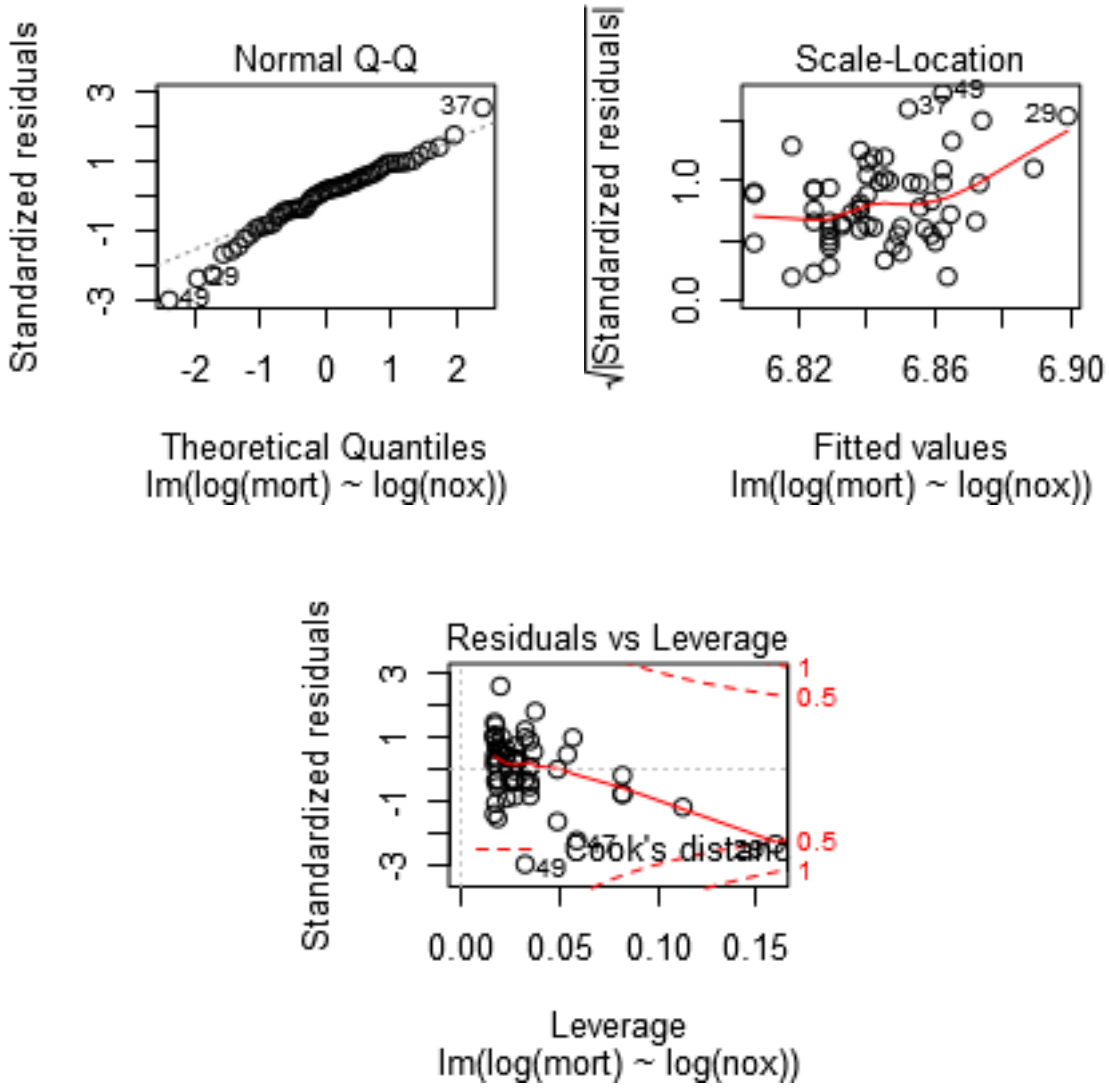
```r
summary(lm_n)
```

```
##
## Call:
## lm(formula = log(mort) ~ log(nox), data = p)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18930 -0.02957  0.01132  0.03897  0.16275
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.807175   0.018349 370.975   <2e-16 ***
## log(nox)    0.015893   0.007048   2.255   0.0279 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06412 on 58 degrees of freedom
## Multiple R-squared:  0.08061,    Adjusted R-squared:  0.06476
## F-statistic: 5.085 on 1 and 58 DF,  p-value: 0.02792
```

```r
res_n <- resid(lm_n)
```

```r
plot(log(p$nox), res_n, xlab = "Log Nitric Oxide Levels", ylab = "Residuals")
plot(lm_n)
```





The residuals Vs. nitric oxide levels shows that the residuals are randomly distributed. But the residuals vs fitted plot shows a parabolic trend of the residuals which may be due to a few outliers that exist, or may be due to the presence of the predictor variable in squared form.

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

The slope coefficient is the percentage of predicted difference in mortality rate for a 1% difference in the nitric oxide level. For a 10% increase in the nitric oxide level, the mortality rate increases by 0.15%

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

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

```
##                    0.5 %     99.5 %
## (Intercept)  6.758304991 6.85604444
## log(nox)    -0.002876882 0.03466334
```

The confidence intervalfor the slope coefficient shows that if we perform the regression a 100

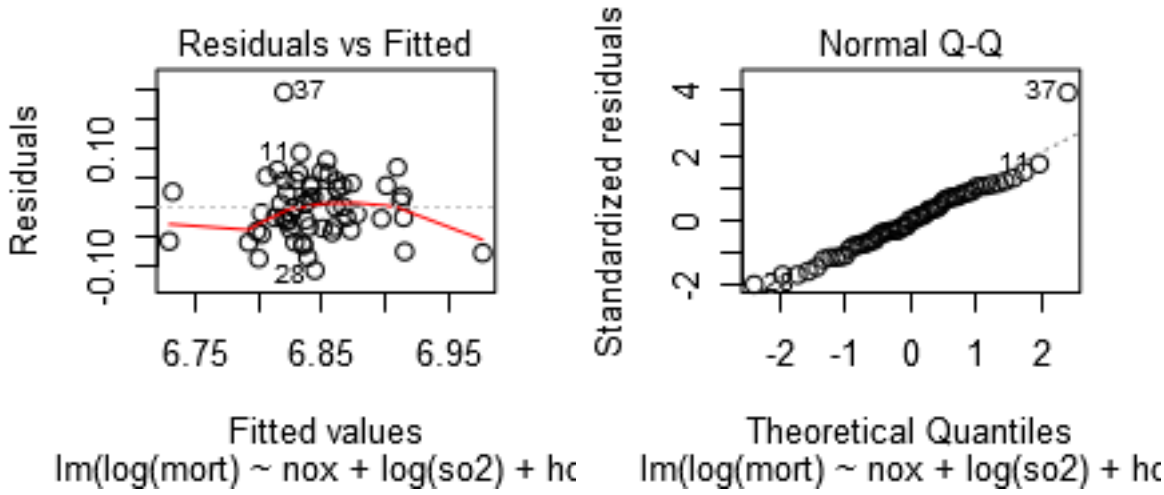
times, in 99 times out of 100, the interval will contain the true value of the slope coefficient that we estimated. But, the confidence interval for the slope coefficient crosses 0, which implies that the predictor and response variables may not be well related.

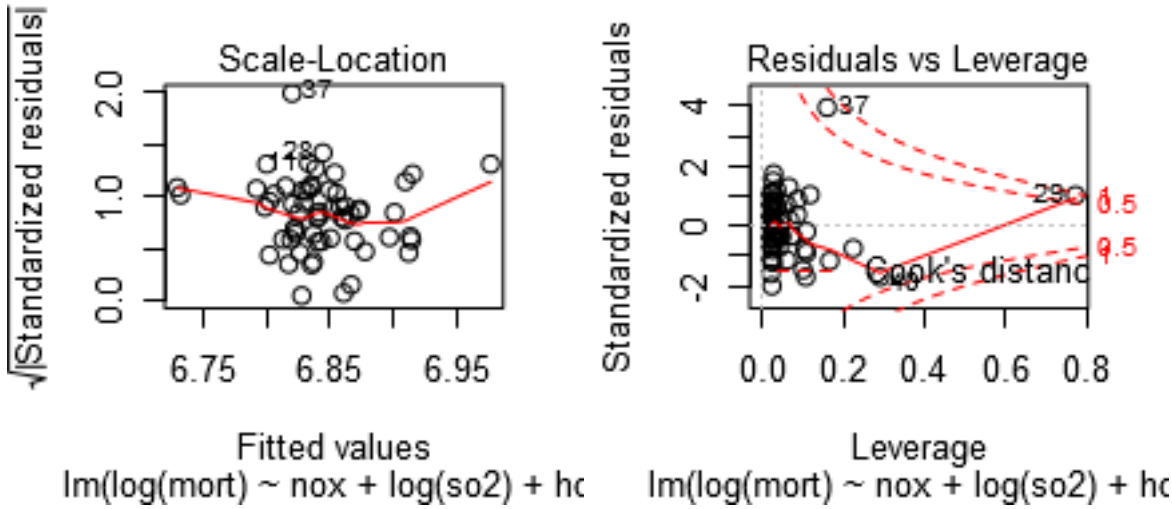
5. Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.

```
lm_4 <- lm(log(mort) ~ nox + log(so2) + hc, data = p)
summary(lm_4)
```

```
##
## Call:
## lm(formula = log(mort) ~ nox + log(so2) + hc, data = p)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.106355 -0.035855  0.000073  0.036057  0.194545
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.8011963  0.0175864 386.730  < 2e-16 ***
## nox          0.0032249  0.0010211   3.158 0.002558 **
## log(so2)     0.0116525  0.0058664   1.986 0.051901 .
## hc          -0.0017805  0.0004981  -3.575 0.000731 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05375 on 56 degrees of freedom
## Multiple R-squared:  0.3761, Adjusted R-squared:  0.3427
## F-statistic: 11.25 on 3 and 56 DF,  p-value: 6.983e-06
```

```
plot(lm_4)
```





The intercept is the predicted log mortality rate for zero nitric oxide and hydrocarbon levels, and a sulphur dioxide level of 1. Mortality rate = exp of 6.8 = 897.84 The nox coefficient is the predicted difference in the log mortality rate when so2 is 1 and hc is zero. exp of 0.003 = 1.003 which implies that for a unit difference in nox level, the mortality rate changes by .3%. The log so2 coefficient is the predicted difference in mortality rate for a 1% difference in the so2 level, when the nox and hc levels are 0. For an increase of 10% in so2, the mortality rate increases by 0.1% The hc coefficient is the predicted difference in the estimated mortality rate when so2 is i and nox is 0. exp of -0.0017 = 0.99 implies that for a unit increase in the hc level, the estimated mortality rate decreases by around 1%

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.)

```
train <- p[c(1:30),]
test <- p[c(31:60),]
fit <- lm(log(mort) ~ nox + log(so2) + hc, data = train)
predict(fit)
```

```
##        1        2        3        4        5        6        7        8
## 6.869668 6.867727 6.859349 6.844248 6.910728 6.896130 6.903884 6.820011
##        9       10       11       12       13       14       15       16
## 6.859367 6.846067 6.846225 6.913535 6.900772 6.871411 6.832623 6.796558
##       17       18       19       20       21       22       23       24
## 6.843274 6.847975 6.886055 6.831284 6.796558 6.834563 6.819511 6.832282
##       25       26       27       28       29       30
## 6.798711 6.853330 6.818870 6.855095 6.764891 6.913372
```

```
predict(fit, newdata = test, type = "response")
```

```
##       31       32       33       34       35       36       37       38
## 6.873152 6.794277 6.884292 6.849202 6.876950 6.831959 6.806311 6.878607
##       39       40       41       42       43       44       45       46
## 6.910774 6.953786 6.836614 6.845250 6.880115 6.864686 6.844110 6.860646
##       47       48       49       50       51       52       53       54
## 6.811399 6.851628 6.757271 6.836944 6.848159 6.850446 6.850952 6.834199
##       55       56       57       58       59       60
```

```
## 6.854206 6.795101 6.864092 6.828538 6.867630 6.866776
```

Study of teenage gambling in Britain

```
teen <- 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.

Since income has a skewed distribution, we do a log transformation on income.

```
data(teengamb)
log.income <- log(teengamb$income)
lm_5 <- lm(gamble ~ verbal + log.income +  status + sex, data = teengamb)
summary(lm_5)
```

```
##
## Call:
## lm(formula = gamble ~ verbal + log.income + status + sex, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.254 -14.986  -1.249   9.283  95.050
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.61359   18.21959   1.461 0.151535
## verbal       -2.80723    2.28742  -1.227 0.226563
## log.income   20.80599    5.05179   4.119 0.000175 ***
## status       -0.08675    0.28939  -0.300 0.765845
## sex         -27.74490    8.39750  -3.304 0.001955 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.9 on 42 degrees of freedom
## Multiple R-squared:  0.4749, Adjusted R-squared:  0.4249
## F-statistic: 9.497 on 4 and 42 DF,  p-value: 1.463e-05
```

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

```
confint(lm_5, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) -10.1550302  63.382212
## verbal       -7.4234227   1.808960
## log.income   10.6110654  31.000914
## status       -0.6707582   0.497268
## sex         -44.6917459 -10.798048
```

According to the intervals shown, the interval for the intercept is wide and the intervals for the intercept, verbal, status and sex coefficients cross zero, rendering the coefficients statistically insignificant. In contrast, the r-squared value of around 0.45 implies that the model explains around 45% of the variation in the response variable.

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?

```
male_data <- teengamb[teengamb$sex == 0,]
male_data_new <- rbind (c(mean(male_data$status), mean(male_data$income), mean(male_data$verbal)), c(ma
colnames(male_data_new) <- c("status", "income", "verbal")

lm_p <- lm(gamble ~ status + log(income) + verbal, data = teengamb)
predict(lm_p, newdata = as.data.frame(male_data_new), interval = "prediction", level = 0.95)
```

```
##        fit        lwr      upr
## 1 29.22793 -25.090643  83.5465
## 2 51.21241  -8.482823 110.9076
```

The confidence interval for the average values of of predictor variables is narrower than that for the maximal values. For the average values, the standard deviation, x - avg of x is zero and mathematically, the confidence interval is narrower. When values deviate largely from the mean, they have a larger standard deviation which gives a wider confidence interval.

School expenditure and test scores from USA in 1994-95

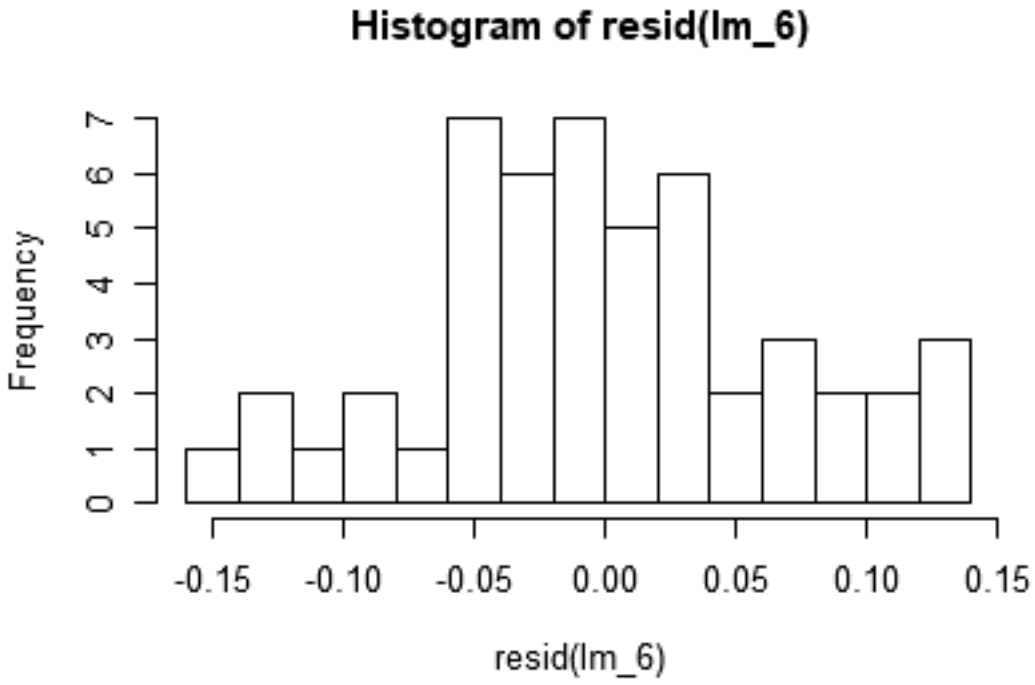
```
data(sat)
?sat
```

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.

```
data(sat)
sats <- as.data.frame(sat)
lm_6 <- lm(log(total) ~ expend + ratio + log(salary), data = sats)
summary(lm_6)
```

```
##
## Call:
## lm(formula = log(total) ~ expend + ratio + log(salary), data = sats)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.151335 -0.043010 -0.009961  0.038443  0.128399
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.857333   0.389248  20.186   <2e-16 ***
## expend       0.018998   0.020689   0.918   0.3633
## ratio        0.007533   0.006503   1.158   0.2527
## log(salary) -0.346733   0.158920  -2.182   0.0343 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06995 on 46 degrees of freedom
## Multiple R-squared:  0.2266, Adjusted R-squared:  0.1762
## F-statistic: 4.493 on 3 and 46 DF,  p-value: 0.007574
```

```
hist(resid(lm_6), breaks = 10)
```



Intercept- The intercept is the predicted log total score when the current expenditure is zero, the average pupil to teacher ratio is zero and the salary is 1. Expenditure Coefficient- The expend coefficient is the predicted difference in log total score for a unit change in the expenditure per pupil, when average pupil to teacher is zero and salary is 1. exp of 0.018 = 1.019 implies that the predicted difference in total score for a unit difference in expenditure is 1.9%. Average Pupil to Teacher Ratio Coefficient- When expenditure is 0 and salary is 1, the ratio coefficient is the predicted difference in log total score for a unit change in the ratio. exp of 0.0075 = 1.007 implies that for a unit change in the average ratio of pupil to teacher, the totl score changes by 0.7%. log salary Coefficient- The log salary coefficient is the predicted difference in log total score when expenditure and average pupil to teacher ratio are 0. For a 10% increase in salary, the predicted total score decreases by 3.4%.

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

```
confint(lm_6, level = 0.98)
```

```
##                      1 %       99 %
## (Intercept)  6.919171742 8.79549491
## expend      -0.030866844 0.06886250
## ratio       -0.008139979 0.02320581
## log(salary) -0.729760059 0.03629338
```

The confidence intervals for the expenditure, ratio and the salary coefficients cross zero rendering the coefficients statistically insignificant. Only the intercept is statistically significant. In 98 out of 100 times the regression is performed, the interval will contain the true value of the intercept we are estimating through the regression model.

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?

```r
lm_7 <- lm(log(total) ~ expend + ratio + log(salary) + takers, data = sats)
summary(lm_7)
```

```
##
## Call:
## lm(formula = log(total) ~ expend + ratio + log(salary) + takers,
##     data = sats)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.092258 -0.023434 -0.000306  0.015038  0.070711
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.8396839  0.2015356  33.938   <2e-16 ***
## expend       0.0074762  0.0098596   0.758    0.452
## ratio       -0.0030084  0.0031967  -0.941    0.352
## log(salary)  0.0403375  0.0814073   0.496    0.623
## takers      -0.0029973  0.0002375 -12.621   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03319 on 45 degrees of freedom
## Multiple R-squared:  0.8296, Adjusted R-squared:  0.8145
## F-statistic: 54.79 on 4 and 45 DF,  p-value: < 2.2e-16
```

The addition of takers variable boosted teh R-Squared value of the model from 0.22 to 0.82 implying that the new model explains 82% of the variation in total score as compared to 22% for the previous fit. The coefficient of takers has a small standard error comapred to its estimate and is statistically significant to explain the variation in total score. But, the other three predictor variables continue to remain statistically insignificant.

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 difference between D_i and R_i } ***could be a good measure because it is symmetric and centered. But, this measure is not proportional. For example, if the average money raised by the parties is 7million and 5 million, the difference Di - Ri is 2million. If the money raised were 3million and 1 million, the same difference would have corresponded to a much closer gap in the first case than in the second case.***

- The ratio, D_i/R_i

This measure is not very appropriate because it is centered at 1. If the Ri is much larger than Di, then the ratio tends to zero and it tends to infinity when Di is much larger than Ri.

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

This measure is similar to the first measure but it is proportional in terms of the magnitude of difference in the money raised by both the parties, i.e., a 2million difference will have a lesser value even when the counties raise 100million. This measure is less sensitive to outliers as well.

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

This measure is centered at 0.5 and is symmetric. When Ri is much larger than Di, the ratio tends to zero but when Di is much larger than Ri, the ratio tends to 1.

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 $\mathrm{x}^\star = 10(\mathrm{x} - 1)$?

When x is transformed to x - 10, the slope coefficient remains intact but the intercept changes. $\hat{\alpha}^\star = 10$ ***The intercept will now correspond to the value of the predicted y when x = 10.*** $\hat{\beta}^\star$ ***is equal to the value of*** $\hat{\beta} = 0.9$ $\hat{\sigma}^\star$ ***remains the same as*** $\hat{\sigma} = 2$

When $\mathrm{x}^\star = 10\mathrm{x}$, ***the intercept remains the same but the slope coefficient, i.e. the coefficient of x gets scaled by 10. The*** $\hat{\beta}$ ***is now equal to 9. r is not affected by scaling.***

For $\mathrm{x}^\star = 10(\mathrm{x} - 1)$, ***the intercept becomes 1.9, the slope becomes 0.09, the standard error of regression coefficient becomes 0.003 and the standard deviation becomes 0.2. The r remains intact***

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 $\mathrm{y}^{\star\star} = 5(\mathrm{y} + 2)$?

For y + 10 transformation, the intercept becomes 11 and all the other coefficients remain the same. When y becomes 5y, all the coefficients become 5 times their original values. Intercept becomes 5, regression coefficient becomes 4.5, the standard error of beta becomes 0.15 and the standard deviation becomes 10. For 5(y+2), the intercept becomes 15 and the slope coefficient becomes 4.5.

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

Centering x and y or adding a constant to x and y changes only the intercept and does not affect the slope/regression coefficients. Scaling x scales the slope coefficient of x but does not affect the intercept. Scaling y scales both the intercept and slope coefficients.

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

Standard error of beta hat becomes 0.003 whereas the t value remains unchanged.

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

Standard error becomes 0.15 and the t values remains unchanged.

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

When x=ax, beta= a(beta), the standard error of beta is also multiplied by a. Since beta is scaled by a, the confidence interval for beta become wider. When y is ay, the beta is scaled by a and hence the confidence interval for beta becomes wider.

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.