Homework 02

XUAN ZHU

Septemeber 21, 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.

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

```
ok <- !is.na (heights$earn+heights$height+heights$sex) & heights$earn> 0 & heights$yearbn>25
heights.clean <- cbind.data.frame(earn=heights$earn,sex=heights$sex,height=heights$height)[ok,]
```

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?

```
print("Centering the variable 'height' by subtracting the mean of the data")
```

```
## [1] "Centering the variable 'height' by subtracting the mean of the data"
```

```
c.mean.height <- heights.clean$height-mean(heights.clean$height)
model.height <- lm(earn~c.mean.height,data = heights.clean)
summary(model.height)
```

```
##
## Call:
## lm(formula = earn ~ c.mean.height, data = heights.clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30439 -11419  -3579   6336 172598
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    23748.8      592.4   40.09  < 2e-16 ***
## c.mean.height   1245.8      154.8    8.05 2.21e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 19300 on 1060 degrees of freedom
## Multiple R-squared:  0.05761,    Adjusted R-squared:  0.05672
## F-statistic:  64.8 on 1 and 1060 DF,  p-value: 2.213e-15
```

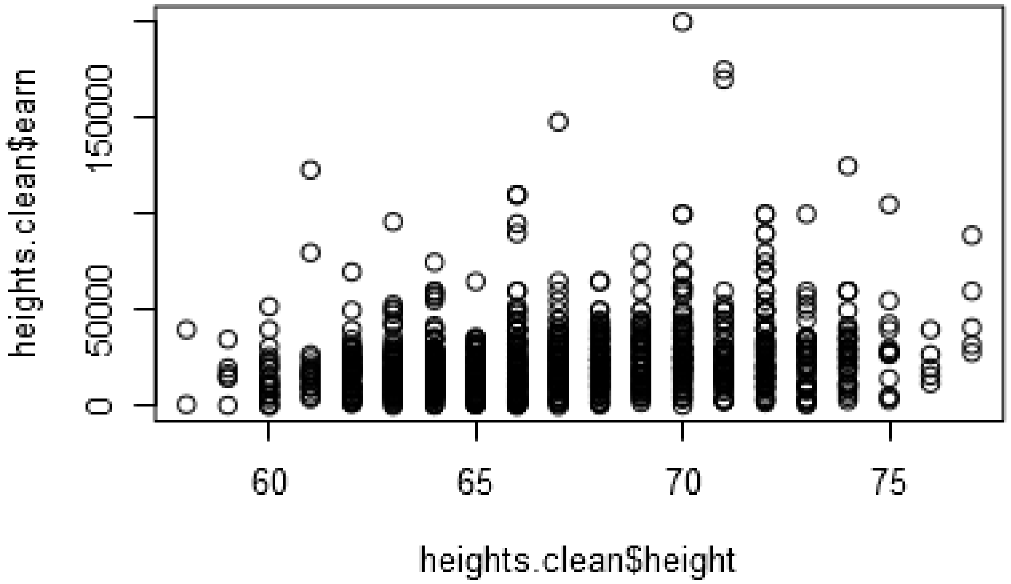
3. Fit some regression models with the goal of predicting earnings from some combination of sex, height, and weight. Be sure to try various transformations and interactions that might make sense. Choose your preferred model and justify.

Statement:

We can learn from the model.height that it does not fit the data very well 'cuz the residual standard error is extremely large. So the first thing I do in this problem is to draw the plot and to see what is the possibly reasonable relationship between earnings and height. It seems that some people have particularly high earnings than others with the same height. So there must be other variables afftecting the earnings. But since the problem only asks me to predict earnings from sex and height, what I can do is to take log of earnings to make the effect of large numbers smaller on the model than before.

Now we have better models with small RSE. When the interaction term is included, it looks like the coeff is not statistically significant. But we can still keep it in the model.

```r
plot(x = heights.clean$height,y = heights.clean$earn)
```



```r
log.earn <- log(heights.clean$earn)
heights.clean$sex[heights.clean$sex ==1] <- 0
heights.clean$sex[heights.clean$sex ==2] <- 1
model.try <- lm(log.earn~sex+height, data = heights.clean)
summary(model.try)
```

```
##
## Call:
## lm(formula = log.earn ~ sex + height, data = heights.clean)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2498 -0.3475  0.1427  0.5605  2.2714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.87802    0.71018  12.501  < 2e-16 ***
## sex         -0.44219    0.07812  -5.661 1.94e-08 ***
## height       0.01660    0.01010   1.643    0.101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8933 on 1059 degrees of freedom
## Multiple R-squared:  0.08251,    Adjusted R-squared:  0.08078
## F-statistic: 47.62 on 2 and 1059 DF,  p-value: < 2.2e-16
```

```r
model.adjust <- lm(log.earn~sex+height+height:sex,data =heights.clean)
summary(model.adjust)
```

```
##
## Call:
## lm(formula = log.earn ~ sex + height + height:sex, data = heights.clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2394 -0.3394  0.1503  0.5682  2.2543
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.567061   0.994832   8.612   <2e-16 ***
## sex          0.166222   1.364775   0.122    0.903
## height       0.021033   0.014163   1.485    0.138
## sex:height  -0.009026   0.020215  -0.447    0.655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8936 on 1058 degrees of freedom
## Multiple R-squared:  0.08269,    Adjusted R-squared:  0.08009
## F-statistic: 31.79 on 3 and 1058 DF,  p-value: < 2.2e-16
```

4. Interpret all model coefficients.

Because the 'sex' var is taken on 1 or 2, the explanation is a little bit complicated. So I convert the values to 0 and 1. It does not affect any inferences of models, just to make the interpretation of coeffs in the <model.adjust> easy.

<model.try>

Intercept: no meaning as no one is 0 meter high.

sex: A female earns 35.74% less than a male does with the same height

height: An increase of one unit in height leads to 1.67% iencrease in a person's income.

<model.adjust>

height: A male earns 2.13% more than he did in the past if his height increases by 1 unit.

sex: no meaning as no one is 0 meter high.

height:sex: the difference in the slope for height, compared female with male.

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

```
round(confint(model.height, level=.95) ,2 )
```

```
##                   2.5 %   97.5 %
## (Intercept)   22586.44 24911.16
## c.mean.height   942.13  1549.48
```

```
round(confint(model.try, level=.95) ,2 )
```

```
##             2.5 % 97.5 %
## (Intercept)  7.48  10.27
## sex         -0.60  -0.29
## height       0.00   0.04
```

```
round(confint(model.adjust, level=.95) ,2 )
```

```
##             2.5 % 97.5 %
## (Intercept)  6.61  10.52
## sex         -2.51   2.84
## height      -0.01   0.05
## sex:height  -0.05   0.03
```

We have 95% of confidence that the true coneffs lie within these CIs. If the CI contains 0, it means that the coeff is not statistically significant.

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:

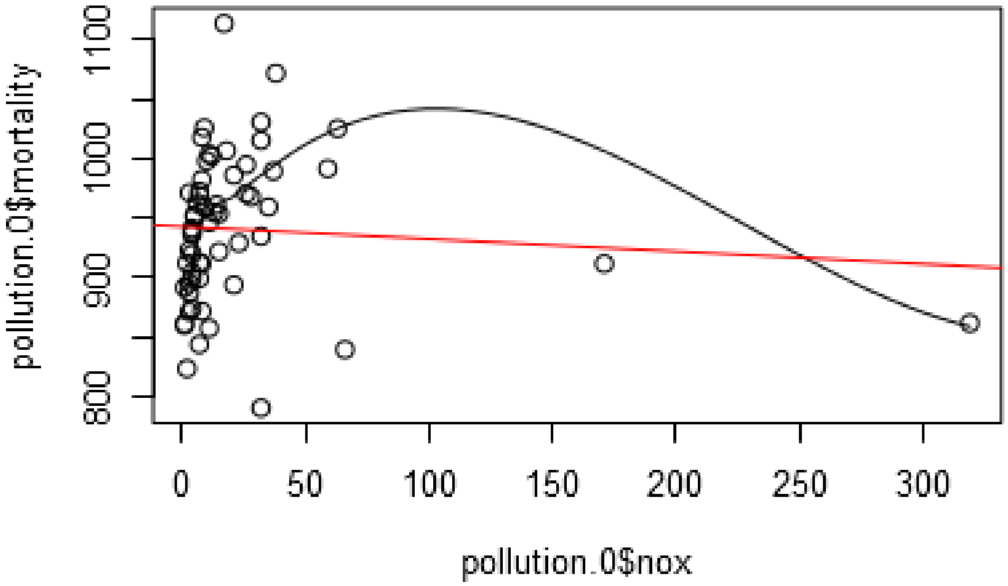
- PREC Average annual precipitation in inches
- JANT Average January temperature in degrees F
- JULT Same for July
- OVR65 % of 1960 SMSA population aged 65 or older
- POPN Average household size
- EDUC Median school years completed by those over 22
- HOUS % of housing units which are sound & with all 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.

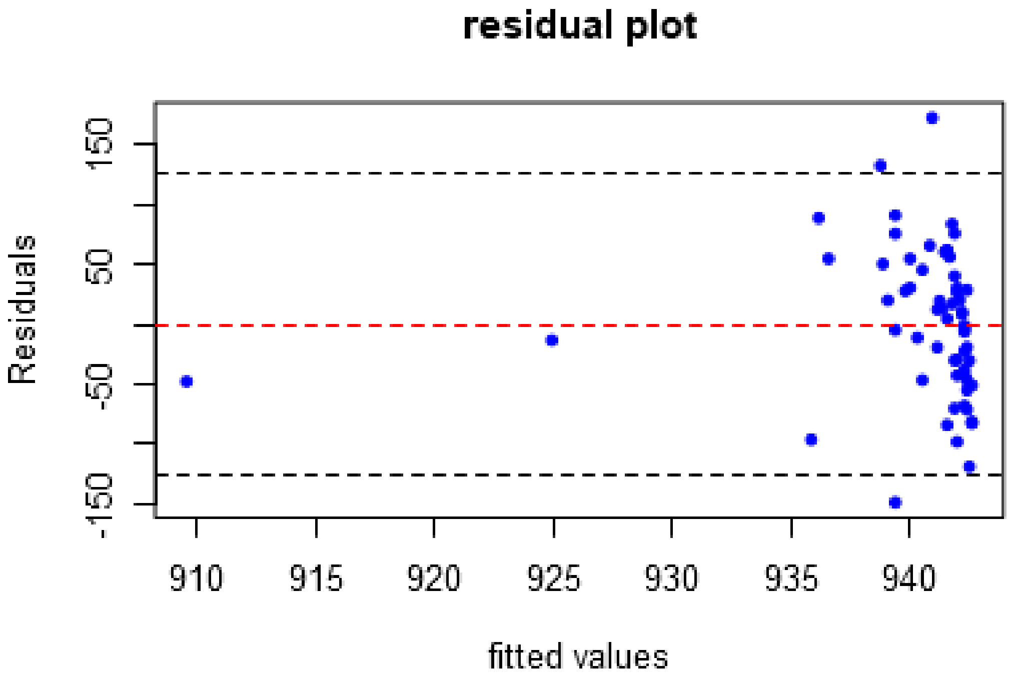
```
gelman_dir   <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
pollution    <- read.dta (paste0(gelman_dir,"pollution/pollution.dta"))
```

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.

```
pollution.0 <- cbind.data.frame(mortality=pollution$mort,nox=pollution$nox,so2=pollution$so2,hc=polluti
scatter.smooth(x = pollution.0$nox,y = pollution.0$mortality)
l.m <- lm(data = pollution.0, mortality~nox)
abline(l.m,col = 'red')
```



```
y.hat <- fitted(l.m)
u <- resid(l.m)
sigma <- sigma.hat(l.m)
residual.plot(y.hat, u, sigma,xlab = "fitted values",main = "residual plot")
```



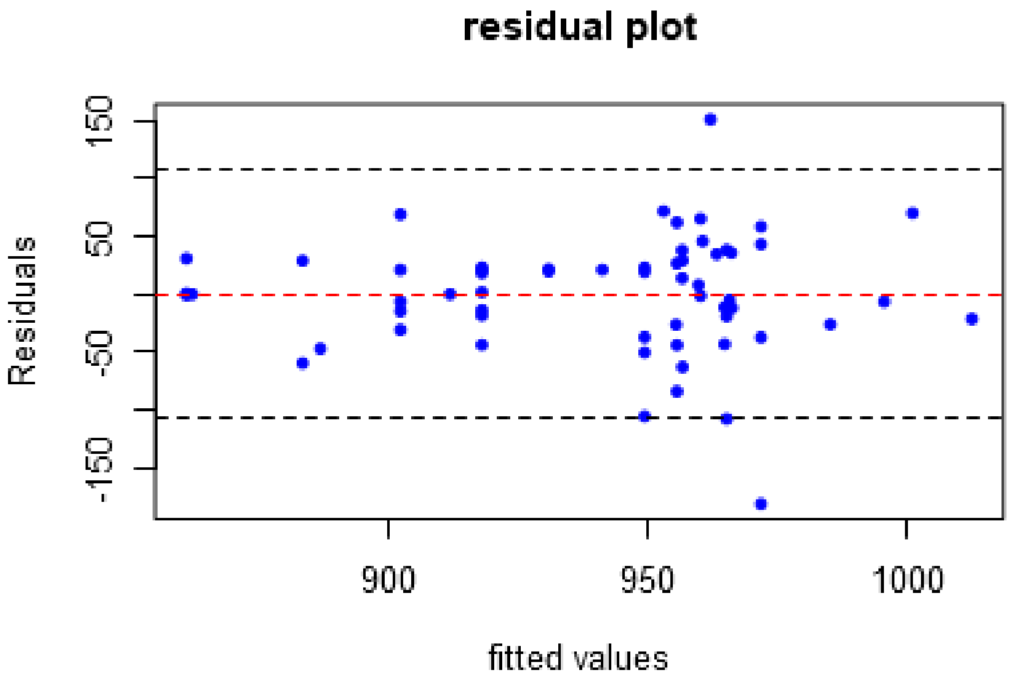
Obviously, the linear regression is a bad choice. From the residual plot we can tell that redisuals are not random.

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.

```
ploymodel <- lm(data=pollution.0, mortality~poly(nox,6))
summary(ploymodel)
```

```
##
## Call:
## lm(formula = mortality ~ poly(nox, 6), data = pollution.0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -181.086  -26.083    0.376   29.099  151.097
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      940.358      6.948 135.351   <2e-16 ***
## poly(nox, 6)1    -36.973     53.815  -0.687   0.4951
## poly(nox, 6)2   -136.784     53.815  -2.542   0.0140 *
## poly(nox, 6)3    125.791     53.815   2.337   0.0232 *
## poly(nox, 6)4   -129.383     53.815  -2.404   0.0197 *
## poly(nox, 6)5     54.001     53.815   1.003   0.3202
## poly(nox, 6)6   -138.773     53.815  -2.579   0.0127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.82 on 53 degrees of freedom
## Multiple R-squared:  0.3277, Adjusted R-squared:  0.2516
## F-statistic: 4.305 on 6 and 53 DF,  p-value: 0.001316
```

```
y.hat1 <- fitted(ploymodel)
u1 <- resid(ploymodel)
sigma1 <- sigma.hat(ploymodel)
residual.plot(y.hat1, u1, sigma1,xlab = "fitted values",main = "residual plot")
```



When we add more and higher polynomial terms, the residual pts spread out and the plots look better than before.

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

When the relative pollution potential of nox is 0, the mortality rate is 940 per 10000. The coeffs in polynomial terms are hard to interpret.

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

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

```
##                      0.5 %     99.5 %
## (Intercept)    921.79583 958.921040
## poly(nox, 6)1 -180.75798 106.812678
## poly(nox, 6)2 -280.56946   7.001196
## poly(nox, 6)3  -17.99399 269.576668
## poly(nox, 6)4 -273.16832  14.402341
## poly(nox, 6)5  -89.78436 197.786296
## poly(nox, 6)6 -282.55788   5.012783
```

We are 99% confident that the true beta0 falls within 921 ~ 959.

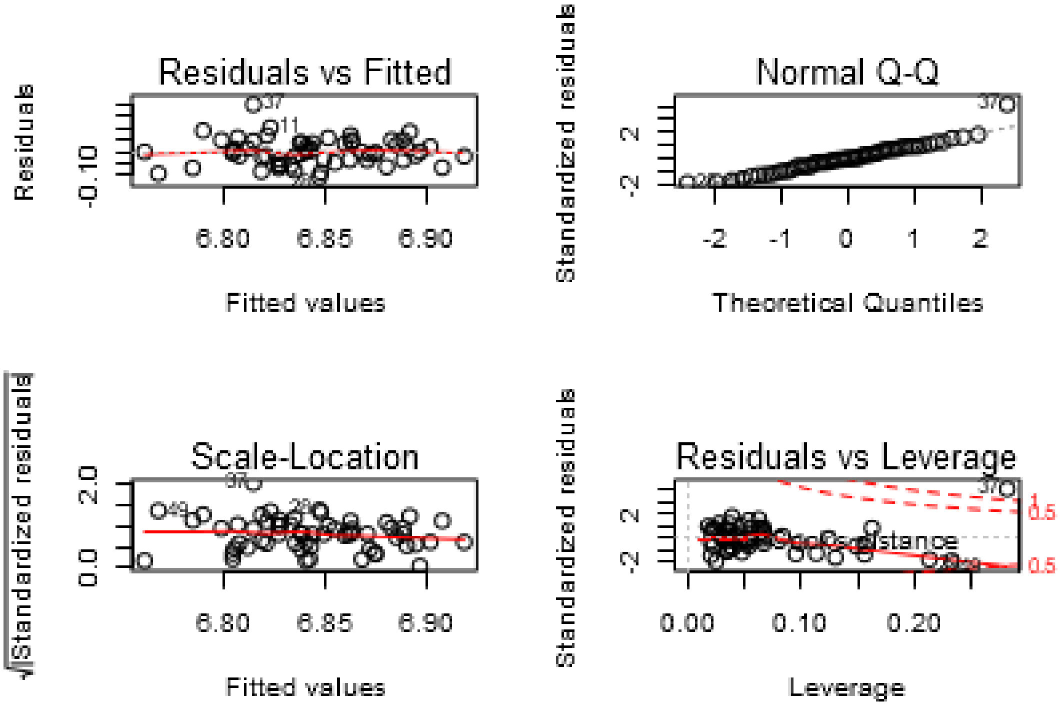
5. Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons

as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.

```
finalmodel <- lm(data=pollution.0,log(mortality)~log(nox)+log(so2)+log(hc))
summary(finalmodel)
```

```
##
## Call:
## lm(formula = log(mortality) ~ log(nox) + log(so2) + log(hc),
##     data = pollution.0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10874 -0.03574 -0.00218  0.03709  0.20085
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.826749   0.022701 300.726  < 2e-16 ***
## log(nox)     0.059837   0.023021   2.599  0.01192 *
## log(so2)     0.014309   0.007584   1.887  0.06436 .
## log(hc)     -0.060812   0.020553  -2.959  0.00452 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05753 on 56 degrees of freedom
## Multiple R-squared:  0.2852, Adjusted R-squared:  0.2469
## F-statistic: 7.449 on 3 and 56 DF,  p-value: 0.0002777
```

```
par(mfrow=c(2,2))
plot(finalmodel)
```



We use a log-log model here. For the interpretations, 1% increase in the each var will lead to (coeff*100)% change in mortality.

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.)

```
randomrows <- sample(x=c(1:nrow(pollution.0)),size=nrow(pollution.0)/2)
half1 <- pollution.0[randomrows,]
half2 <- pollution.0[-randomrows,]
reg.half1 <- lm(data=half1,log(mortality)~log(nox)+log(so2)+log(hc))
pre.half2 <- predict(reg.half1,half2,level=0.95,interval="prediction")
```

Study of teenage gambling in Britain

```
data(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.

```
colnames(teengamb)[colnames(teengamb)=="sex"] <- "female"
c.status <- teengamb$status-mean(teengamb$status)
c.income <- 48* (teengamb$income - mean(teengamb$income))
c.verbal <- teengamb$verbal-mean(teengamb$verbal)
log.gamble <- log(teengamb$gamble+0.001)
modelteen <- lm(log.gamble~c.verbal+female+c.income+c.status,data=teengamb)
summary(modelteen)
```

```
##
## Call:
## lm(formula = log.gamble ~ c.verbal + female + c.income + c.status,
##     data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1176 -0.9049  0.5461  1.4017  3.6411
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.615278   0.527280   3.063  0.00381 **
## c.verbal    -0.588189   0.244338  -2.407  0.02054 *
## female      -1.635434   0.923639  -1.771  0.08388 .
## c.income     0.006891   0.002403   2.868  0.00644 **
## c.status     0.060508   0.031621   1.914  0.06251 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.552 on 42 degrees of freedom
## Multiple R-squared:  0.3854, Adjusted R-squared:  0.3269
## F-statistic: 6.585 on 4 and 42 DF,  p-value: 0.0003304
```

Intercept: A male with average verbal score & average yearly income & average Socioeconomic status score spends 5.027 pounds on gambling per year.

female: A female will spend 80.5% pounds less than a male on gambling per year, holding all other vars constant.

c.income: One pound increase in yearly income leads to 0.69% pounds increase in gambling expenditure, holding all other vars constant.

c.status: One point increase in Socioeconomic status score leads to 6.24% pounds increase in gambling expenditure,holding all other vars constant.

c.verbal: One point increase in verbal score leads to 44.46% decrease in gambling expenditure,holding all other vars constant.

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

```
round(confint(modelteen,level = 0.95),5)
```

```
##                2.5 %   97.5 %
## (Intercept)  0.55118  2.67937
## c.verbal    -1.08128 -0.09510
## female      -3.49941  0.22855
## c.income     0.00204  0.01174
## c.status    -0.00331  0.12432
```

We have 95% of confidence that the true coneffs lie within these CIs. If the CI contains 0, it means that the coeff is not statistically significant.

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?

```
prediction.average <- predict(modelteen,newdata = data.frame(female=0,c.status=0,c.income=0,c.verbal=0)
exp(prediction.average)
```

```
##        fit        lwr      upr
## 1 5.029284 0.02613837 967.6846
```

```
prediction.max <- predict(modelteen,newdata = data.frame(female=0,c.status=max(c.status),c.income=max(c
exp(prediction.max)
```

```
##        fit       lwr     upr
## 1 131.3392 0.2941355 58646.4
```

Let's say p1 and p2 here. p2's PI is wider because p1 and p2 use different standard errors to calculate intervals. From p1 we are receiving the value of the intercept, so its PI is similar to its CI. And for the SE we use in calculating p2, it includes the SE from p1. So we expect that p2 is larger than p1.

School expenditure and test scores from USA in 1994-95

```
data(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.

```
c.expend <- sat$expend-mean(sat$expend)
c.ratio <- sat$ratio-mean(sat$ratio)
c.salary <- sat$salary-mean(sat$salary)
```

```r
model.sat <- lm(sat$total~c.expend+c.ratio+c.salary+c.expend:c.salary)
summary(model.sat)
```

```
##
## Call:
## lm(formula = sat$total ~ c.expend + c.ratio + c.salary + c.expend:c.salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -145.97  -40.36   -5.99   32.91  132.04
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        958.821     12.262  78.194   <2e-16 ***
## c.expend             8.780     23.512   0.373   0.7106
## c.ratio              5.630      6.590   0.854   0.3975
## c.salary            -8.411      4.722  -1.781   0.0816 .
## c.expend:c.salary    1.029      1.083   0.950   0.3474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.73 on 45 degrees of freedom
## Multiple R-squared:  0.2251, Adjusted R-squared:  0.1563
## F-statistic: 3.269 on 4 and 45 DF,  p-value: 0.01952
```

Intercept: Average expenditure per pupil + Average ratio + average salary lead to an average SAT score of 958.821.

c.expend: 1000 dollars increase in expenditure per pupil with average salary and average ratio leads to 8.78 pts increase in SAT score.

c.ratio: One point increase in pupil/teacher ratio leads to 5.63 pts increase in SAT score, with average expend & average salary.

c.salary: 1000 dollars increase in teachers' annual salary leads to 8.411 decrease in SAT score with average expend & average ratio.

c.expend:c.salary: the difference in the slope for c.expend is 1.029, compared the average salary with 1000 dollars more

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

```r
confint(model.sat,level =0.98)
```

```
##                            1 %       99 %
## (Intercept)       929.243858 988.398927
## c.expend          -47.933710  65.493534
## c.ratio           -10.266755  21.527043
## c.salary          -19.800335   2.978642
## c.expend:c.salary  -1.584281   3.641325
```

All CIs contain 0 except for the CI of intercept. This suggests that our model has problems.

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
model.takers <- lm(sat$total~c.expend+takers+c.ratio+c.salary+c.expend:c.salary,data=sat)
summary(model.takers)
```

```
##
## Call:
## lm(formula = sat$total ~ c.expend + takers + c.ratio + c.salary +
##     c.expend:c.salary, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.714 -21.418  -1.957  17.739  66.616
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1066.5724    10.5662 100.942  < 2e-16 ***
## c.expend             3.0884    11.3067   0.273    0.786
## takers              -2.8934     0.2355 -12.285 8.13e-16 ***
## c.ratio             -3.7155     3.2567  -1.141    0.260
## c.salary             1.6740     2.4127   0.694    0.491
## c.expend:c.salary    0.1899     0.5249   0.362    0.719
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.02 on 44 degrees of freedom
## Multiple R-squared:  0.8251, Adjusted R-squared:  0.8052
## F-statistic: 41.51 on 5 and 44 DF,  p-value: 1.409e-15
```

Model.takers is better with smaller RSE and larger adjusted Rˆ2. # 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, $D_i - R_i$

This measurement directly shows how the differnece in the money raising for two parties would affect the vote share. But this number can be quite large when the amount of money is large. In other words, it is hard for us to feel the difference between 25m and 26m.

- The ratio is easy for us to interpret it as the proportion, but it loses some info such as the amount of the money raised. For example, both 2m/1m and 10k/5k equal 2, but they may have a different impact on the vote share.
- The difference on the logarithmic scale, $logD_i - logR_i$

If the data is skewed, the log transformation helps the data fit a linear model. But log are restricted ed to positive values.

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

Give us about the proportion of Di in the total money raised. But it loses some info.

Transformation

For observed pair of x and y, we fit a simple regression model y $= \alpha + \beta$x$+ \epsilon$ which results in estimates $\hat{\alpha} = 1$, $\hat{\beta} = 0.9$, $SE(\hat{\beta}) = 0.03$, $\hat{\sigma} = 2$ and r=0.3.

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

Linear transformations do not change the variance and resiuals. So

$\hat{\alpha}^\star = 11$, $\hat{\beta}^\star = 0.9$, $\hat{\sigma}^\star = 2$, and $r^\star = 0.3$.

when x* = 10x,

$\hat{\alpha}^\star = 1$, $\hat{\beta}^\star = 0.09$, $\hat{\sigma}^\star = 0.2$, and $r^\star = 0.3$.

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

$\hat{\alpha}^\star = 1.9$, $\hat{\beta}^\star = 0.09$, $\hat{\sigma}^\star = 0.2$, and $r^\star = 0.3$.

2. Now suppose that the response variable scores are transformed according to the formula y** = y + 10 and that y** is regressed on x. Without redoing the regression calculation in detail, find $\hat{\alpha}^{\star\star}$, $\hat{\beta}^{\star\star}$, $\hat{\sigma}^{\star\star}$, and $r^{\star\star}$. What happens to these quantities when y** = 5y ? When y** = 5(y + 2)?

$\hat{\alpha}^{\star\star} = 11$, $\hat{\beta}^{\star\star} = 0.9$, $\hat{\sigma}^{\star\star} = 2$, and $r^{\star\star} = 0.3$.

when y** = 5y, $\hat{\alpha}^{\star\star} = 5$ $\hat{\beta}^{\star\star} = 4.5$, $\hat{\sigma}^{\star\star} = 10$, and $r^{\star\star} = 0.3$.

When y** = 5(y + 2) $\hat{\alpha}^{\star\star} = 15$ $\hat{\beta}^{\star\star} = 4.5$, $\hat{\sigma}^{\star\star} = 10$, and $r^{\star\star} = 0.3$.

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

We plug the new x into the linear model and formulas of se and residuals. Basically, linear transformation doesn't affect residuals. Changes to beta affect SE. Adding a number on y or x only changes the intercept. Multipication on y or x changes beta.

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

$SE(\hat{\beta}^\star) = 0.003$ and $t_0^\star = \hat{\beta}^\star / SE(\hat{\beta}^\star) = 30$.

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

$SE(\hat{\beta}^{\star\star}) = 0.15$ and $t_0^{\star\star} = \hat{\beta}^{\star\star} / SE(\hat{\beta}^{\star\star}) = 30$.

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

They are affected by multiplication.

If $x^\star = c * x$, $\hat{\beta}^\star = \hat{\beta}/c$, $SE(\hat{\beta}^\star) = SE(\hat{\beta})/c$,so CI is $[\frac{\hat{\beta}}{c} \pm t_{\frac{\alpha}{2}} \frac{SE(\beta)}{c}]$.

If $y^\star = c * y$, $\hat{\beta}^\star = c\hat{\beta}$, $SE(\hat{\beta}^\star) = cSE(\hat{\beta})$,thus CI is$[c\hat{\beta} \pm t_{\frac{\alpha}{2}} cSE(\beta)]$.