Assignment3

Linfeng Zhou

October 3, 2015

Question 1

1. Python users, use Python's pandas library to read in the Stata-formatted dataset used in class called "SI Sales.dta". R users, use R's foreign library to read in the Stata-formatted dataset called "SI Sales Old.dta". Replicate all of the regression results for this dataset that I presented in class.

```
library(foreign)
setwd("~luke/Dropbox/Applied_Data_Science/Assignment 3")
data <- read.dta("SI Sales Old.dta")
str(data)
```

```
## 'data.frame':    31680 obs. of  6 variables:
##  $ price     : num  327500 346314 349830 325000 285000 ...
##  $ unit_size : num  142.7 195.1 179.5 174.8 97.1 ...
##  $ land_size : num  232 239 201 228 272 ...
##  $ age       : num  32 10 1 106 104 23 29 13 5 51 ...
##  $ todt      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ sales_year: num  2011 2006 2006 2005 2003 ...
##  - attr(*, "datalabel")= chr ""
##  - attr(*, "time.stamp")= chr "27 Sep 2015 18:41"
##  - attr(*, "formats")= chr  "%8.0g" "%9.0g" "%9.0g" "%9.0g" ...
##  - attr(*, "types")= int  255 254 254 254 254 254
##  - attr(*, "val.labels")= chr  "" "" "" "" ...
##  - attr(*, "var.labels")= chr  "" "" "" "" ...
##  - attr(*, "version")= int 12
```

```
cor(data)
```

```
##                  price   unit_size  land_size          age         todt
## price       1.00000000  0.54970588 0.50535567 -0.064087848 0.352150749
## unit_size   0.54970588  1.00000000 0.44610503 -0.201536247 0.326306695
## land_size   0.50535567  0.44610503 1.00000000  0.274696779 0.346801031
## age        -0.06408785 -0.20153625 0.27469678  1.000000000 0.005157383
## todt        0.35215075  0.32630670 0.34680103  0.005157383 1.000000000
## sales_year  0.09609229  0.02341565 0.07740216  0.125040347 0.012260027
##             sales_year
## price       0.09609229
## unit_size   0.02341565
## land_size   0.07740216
## age         0.12504035
## todt        0.01226003
## sales_year  1.00000000
```

```
fit1<-lm(price~unit_size,data)
summary(fit1)
```

```
##
## Call:
## lm(formula = price ~ unit_size, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2167011   -81808   -10257    65075  7275309
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 135043.87    2542.33   53.12   <2e-16 ***
## unit_size     1658.37      14.16  117.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 200100 on 31678 degrees of freedom
## Multiple R-squared:  0.3022, Adjusted R-squared:  0.3022
## F-statistic: 1.372e+04 on 1 and 31678 DF,  p-value: < 2.2e-16
```

```r
fit2<-lm(price ~ unit_size + land_size,data)
summary(fit2)
```

```
##
## Call:
## lm(formula = price ~ unit_size + land_size, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1598276   -62449     3382    62065  7269034
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.179e+05  2.398e+03   49.18   <2e-16 ***
## unit_size   1.221e+03  1.483e+01   82.34   <2e-16 ***
## land_size   2.684e+02  4.064e+00   66.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 187600 on 31677 degrees of freedom
## Multiple R-squared:  0.3867, Adjusted R-squared:  0.3866
## F-statistic:  9985 on 2 and 31677 DF,  p-value: < 2.2e-16
```

```r
fit3<-lm(price ~ unit_size + land_size + age,data)
summary(fit3)
```

```
##
## Call:
## lm(formula = price ~ unit_size + land_size + age, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1761510   -64418     1124    60437  7293473
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 149555.067   2940.848   50.85   <2e-16 ***
## unit_size     1110.981     15.927   69.75   <2e-16 ***
## land_size      302.429      4.445   68.04   <2e-16 ***
## age           -696.560     37.871  -18.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 186600 on 31676 degrees of freedom
## Multiple R-squared:  0.3931, Adjusted R-squared:  0.3931
## F-statistic:  6840 on 3 and 31676 DF,  p-value: < 2.2e-16
```

```r
fit4<-lm(price ~ unit_size + land_size + age + todt,data)
summary(fit4)
```

```
##
## Call:
## lm(formula = price ~ unit_size + land_size + age + todt, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1746982   -65094      751    60976  7297417
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 167397.698   2985.210   56.08   <2e-16 ***
## unit_size     1033.986     16.017   64.56   <2e-16 ***
## land_size      275.181      4.514   60.96   <2e-16 ***
## age           -671.391     37.469  -17.92   <2e-16 ***
## todt        357431.153  13454.780   26.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 184600 on 31675 degrees of freedom
## Multiple R-squared:  0.4064, Adjusted R-squared:  0.4063
## F-statistic:  5421 on 4 and 31675 DF,  p-value: < 2.2e-16
```

```r
data$priceper1000= data$price/1000
head(data)
```

```
##    price unit_size land_size age todt sales_year priceper1000
## 1 327500 142.69901  232.2575  32    0       2011      327.500
## 2 346314 195.09630  239.3181  10    0       2006      346.314
## 3 349830 179.48860  201.4137   1    0       2006      349.830
## 4 325000 174.75055  227.6124 106    0       2005      325.000
## 5 285000  97.08363  271.7413 104    0       2003      285.000
## 6 445000 196.48984  278.7090  23    0       2007      445.000
```

```r
fit5<-lm(priceper1000 ~ unit_size + land_size + age + todt,data)
summary(fit5)
```

```
##
## Call:
## lm(formula = priceper1000 ~ unit_size + land_size + age + todt,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1747.0   -65.1     0.8    61.0  7297.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 167.397698   2.985210   56.08   <2e-16 ***
## unit_size     1.033986   0.016017   64.56   <2e-16 ***
## land_size     0.275181   0.004514   60.96   <2e-16 ***
## age          -0.671391   0.037469  -17.92   <2e-16 ***
## todt        357.431153  13.454780   26.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 184.6 on 31675 degrees of freedom
## Multiple R-squared:  0.4064, Adjusted R-squared:  0.4063
## F-statistic:  5421 on 4 and 31675 DF,  p-value: < 2.2e-16
```

```r
data1<-log(data)
fit6<-lm(log(priceper1000)~log(unit_size)+
           log(land_size)+log(age+1)+todt,data)
summary(fit6)
```

```
##
## Call:
## lm(formula = log(priceper1000) ~ log(unit_size) + log(land_size) +
##     log(age + 1) + todt, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2065 -0.1250  0.0680  0.2092  2.9673
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.869106   0.029426   97.50   <2e-16 ***
## log(unit_size)  0.340766   0.006215   54.83   <2e-16 ***
## log(land_size)  0.263042   0.003770   69.77   <2e-16 ***
## log(age + 1)   -0.047558   0.001957  -24.30   <2e-16 ***
## todt            0.442606   0.024897   17.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3592 on 31675 degrees of freedom
## Multiple R-squared:  0.3644, Adjusted R-squared:  0.3643
## F-statistic:  4540 on 4 and 31675 DF,  p-value: < 2.2e-16
```

Question 2

2. There is an additional feature in this dataset called "sales_year", which captures the year the sale of a house in Staten Island occurred. From this feature, generate a feature that is linear time trend. (A linear time trend is a feature that takes on value "1" in the initial year and increments by "1" each subsequent year. For example, if 2003 were "1", 2004 would be "2", 2005 would be "3", and so forth.) Run a linear regression model that relates the sales price to unit size, land size, age, the Todt Hill indicator, and the linear time trend. How would you interpret the estimated coefficient associated with the linear time trend? What is the 95% confidence interval of your interpretation? Based on your regression diagnostics, have you improved the fit of the house price sales data by including the linear time trend as an additional explanatory feature?

```r
data$year <- data$sales_year-2002
head(data)
```

```
##    price unit_size land_size age todt sales_year priceper1000 year
## 1 327500 142.69901  232.2575  32    0       2011      327.500    9
## 2 346314 195.09630  239.3181  10    0       2006      346.314    4
## 3 349830 179.48860  201.4137   1    0       2006      349.830    4
## 4 325000 174.75055  227.6124 106    0       2005      325.000    3
## 5 285000  97.08363  271.7413 104    0       2003      285.000    1
## 6 445000 196.48984  278.7090  23    0       2007      445.000    5
```

```r
fit7<-lm(price ~ unit_size + land_size + age + todt + year,data)
summary(fit7)
```

```
##
## Call:
## lm(formula = price ~ unit_size + land_size + age + todt + year,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1757395   -62255    -1432    58866  7303161
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 143506.455   3295.435   43.55   <2e-16 ***
## unit_size     1025.378     15.955   64.27   <2e-16 ***
## land_size      273.393      4.496   60.81   <2e-16 ***
## age           -741.640     37.538  -19.76   <2e-16 ***
## todt        359808.337  13396.318   26.86   <2e-16 ***
## year          6325.690    376.930   16.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 183800 on 31674 degrees of freedom
## Multiple R-squared:  0.4116, Adjusted R-squared:  0.4115
## F-statistic:  4431 on 5 and 31674 DF,  p-value: < 2.2e-16
```

```r
confint(fit7,level=0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) 137047.2751 149965.6348
## unit_size      994.1062   1056.6500
## land_size      264.5808    282.2048
## age           -815.2156   -668.0641
## todt        333551.0339 386065.6405
## year          5586.8922   7064.4880
```

Comment:

- According to the summary of year added model, house prices will increase 6325 by year per unit increase.
- The 95% confidence interval results are below argument `confint(fit7,level=0.95)` .
- The adjusted r-squared was increased after including the linear time trend as an additional explanatory feature.

Question 3

3. As noted in class, the unit size and land size features are measured in squared meters. Suppose I ask you to re-express these features using the Imperial system of square feet rather than square meters, but I express a concern that the interpretation of the estimated coefficients, such as age, would be changed. Without acutally doing any statistical learning, what would you say to me about my concern? Rerun the linear regression in 2. using the dwelling size and land size measured in square feet (rather than square meters). What, if anything, has changed in your estimated coefficients?

```
data$unit_sizesf <- data$unit_size * 10.7639
data$land_sizesf <- data$land_size * 10.7639

fit8<-lm(price~unit_sizesf + land_sizesf + age + todt + year,data)
summary(fit8)
```

```
##
## Call:
## lm(formula = price ~ unit_sizesf + land_sizesf + age + todt +
##     year, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1757395   -62255    -1432    58866  7303161
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.435e+05  3.295e+03   43.55   <2e-16 ***
## unit_sizesf  9.526e+01  1.482e+00   64.27   <2e-16 ***
## land_sizesf  2.540e+01  4.177e-01   60.81   <2e-16 ***
## age         -7.416e+02  3.754e+01  -19.76   <2e-16 ***
## todt         3.598e+05  1.340e+04   26.86   <2e-16 ***
## year         6.326e+03  3.769e+02   16.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 183800 on 31674 degrees of freedom
## Multiple R-squared:  0.4116, Adjusted R-squared:  0.4115
## F-statistic:  4431 on 5 and 31674 DF,  p-value: < 2.2e-16
```

Comment:

- You don't have to worry about features like age, todt and year. The changes of unit-size and land-size only cause themselves' changes.
- According to my results, changed parameters are land size and unit size, which prove my opinion.

Question 4

4. (Challenging question. Feel free to work together to the extent that it assists you.) Assume the following data generating process (DGP) governs a random sample of size 10,000: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ for $\epsilon_i \sim N(0, 1)$. Further assume for this DGP that $\beta_0 = \beta_1 = \beta_2 = 1$. (a) Suppose the following process governs your features: $x_{1i} \sim N(0, 1)$ and $x_{2i} \sim N(0, 1)$ are independent. Using R or Python, calculate the correlation between features x_{1i} and x_{2i}. Mistakenly, you decide to estimate a linear regression that includes only the feature x_{1i}. Using R or Python, simulate this DGP and run the mistaken linear regression that includes only feature x_{1i}. What value do you obtain for the coefficient associated with with feature x_{1i}? (b) Suppose instead that the follow process governs your features: $x_{1i} = z_i + \eta_i$ and $x_{2i} = -z_i + \omega_i$, where $z_i \sim N(0, 1)$, $\eta_i \sim N(0, 1)$, and $\omega_i \sim N(0, 1)$ are independent. Using R or Python, calculate the correlation between features x_{1i} and x_{2i}. Again, you mistakenly decide to estimate a linear regression that includes only the feature x_{1i}. Using R or Python, simulate this DGP and run the mistaken linear regression that includes only feature x_{1i}. What value do you obtain for the coefficient associated with with feature x_{1i}? (c) Are there any conclusions you can draw from your results in (a) and (b)?

```r
set.seed(1335)

x1 <- rnorm(10000, mean=0, sd=1)
x2 <- rnorm(10000, mean=0, sd=1)
e1 <- rnorm(10000, mean=0, sd=1)
y1 <- 1 + x1 + x2 + e1
cor(x1,x2)
```

```
## [1] -0.007774456
```

```r
fit9<-lm(y1~x1)
summary(fit9)
```

```
##
## Call:
## lm(formula = y1 ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1354 -0.9724  0.0088  0.9646  5.3097
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.00826    0.01418   71.11   <2e-16 ***
## x1           0.99205    0.01405   70.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.418 on 9998 degrees of freedom
```

```
## Multiple R-squared:  0.3328, Adjusted R-squared:  0.3327
## F-statistic:  4987 on 1 and 9998 DF,  p-value: < 2.2e-16
```

Comment:

The correlation between x1 and x2 is -0.007774456, and the coefficient of x1 is 0.99205.

```
z1<-rnorm(10000, mean=0, sd=1)
q1<-rnorm(10000, mean=0, sd=1)
w1<-rnorm(10000, mean=0, sd=1)

x2_1<- z1+q1
x2_2<- -z1+w1
y2 = 1 + x2_1 + x2_2 +e1
cor(x2_1,x2_2)
```

```
## [1] -0.494366
```

```
fit10<-lm(y2~x2_1)
summary(fit10)
```

```
##
## Call:
## lm(formula = y2 ~ x2_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7377 -1.0694  0.0047  1.0604  6.6174
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.98441    0.01601   61.49   <2e-16 ***
## x2_1         0.49216    0.01139   43.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.601 on 9998 degrees of freedom
## Multiple R-squared:  0.1574, Adjusted R-squared:  0.1574
## F-statistic:  1868 on 1 and 9998 DF,  p-value: < 2.2e-16
```

Comment:

Our regression model is a combanation of three standard normal distribution. The R-square is calculated by variance. In this situation, the first model only include one normal situation, as you can see, the r-square is only 0.33. The second model r-square can use the same rule to explain. For intercept, standard normal distributions' mean is zero, therefore, it can be correctly estimated whatever how many features you use. As we can see from coefficient results, the standard normal distribution's properties still work on this problem. It didn't change the first model estimation. For the second model, however, both x1 and x2 are combinations of two standard normal distribution. If you only fit the model with one parameter, it only has half of the original settings. If you estimated model by using both x1 and x2, it can give you correctly coefficient results.