DARIEN OLIVA
ECON 453 - PSET3

1a) Write your estimated equation. Carefully interpret the meaning of the estimated parameters. Although successful in business, Mr. Dimeworth never completed his college having flunked his statistics courses. He thinks that college education is esoteric, far removed from reality with no practical value, and is for party-minded brats. You feel an obligation to show him that useful and practical applications can be made using econometric techniques.

```
> model1 <-lm(data=data1, Price~Area+Age+D_Pool+D_Location)
> summary(model1,digits=5)

Call:
lm(formula = Price ~ Area + Age + D_Pool + D_Location, data = data1)

Residuals:
   Min     1Q Median     3Q    Max
-31632 -12096   -596  10584  29464

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 128258.45   22459.37   5.711 8.38e-07 ***
Area            66.72       9.59   6.956 1.18e-08 ***
Age          -1138.64     206.85  -5.505 1.69e-06 ***
D_Pool        5859.00    4903.19   1.195   0.2384
D_Location    9287.42    4727.85   1.964   0.0557 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15110 on 45 degrees of freedom
Multiple R-squared:  0.6952,    Adjusted R-squared:  0.6681
F-statistic: 25.66 on 4 and 45 DF,  p-value: 4.086e-11
```

The estimated equation/code to calculate this regression in R is represented by the code in blue on the left. A regression analysis allows us to estimate the relationships between variables through statistical processes. For this problem set, we are making estimations about how variables such as area, age, pool, and location affect the price of a house. The price of the house is the dependent variable and the area, age, pool, and location are the independent variables. This equation/code performs the regression analysis for us, and gives us information to make our estimates such as t and p values, f statistics, standard errors, degrees of freedom, and more.

1b) Does the location of the house matter, i.e., is there a premium for a house located in the prestigious “northern” part of the town?

The p-value that corresponds to the t-statistic is 0.0557 and slightly larger than the significance level of 0.05. Based on this significance level, this means that the location (North or not-North) of the house does not have a strong relationship with the prices of houses. Therefore, from this information we can see that there isn’t a premium or at least a significant one, for a house located in the northern part of town.

1c) Does the presence of a swimming pool influence the selling price of a house?

The p-value that corresponds to the t-statistic is 0.2384 and this is higher than the significance level of 0.05. Based on this significance level, this means that there is not a significant relationship between whether a house has a pool or not, and the price of the house.

1d) Does your estimated equation support your hypothesis that "size of house matters?" Interpret the parameter associated with the "house size" variable.

The results from the equation confirm the hypothesis that the size of the house matters because the relationship between square footage in the house and its price is the strongest relationship we can see in this model. The p value that corresponds to the t-statistic is 1.18e-08, and that is extremely smaller than our significance level.

1e) Mr. Dimeworth believes that newer houses fetch a greater price, other things being equal. Test his hypothesis. Each additional year in the age of the house contributes how much to the price of the house?

The relationship between how old the house is and its price is an interesting relationship. This is because houses that are new, as well as houses that are old have high value. The data/regression results shows that there is a high correlation between the house price and the age of the house, but it is not one-sided because old houses are of as high of value as the new houses.

1f) Do a test of the overall significance of the regression equation. State your null hypothesis and conclusion clearly.

The null hypothesis is that our model is a significant model, and the alternative is that our model is not good enough to predict the value of the dependent variable/price of a house. When testing the overall significance of the regression equation, we get an extremely small p-value. This means that our model is very different from the restricted model, and our model/the coefficients are significant, and how significant power to predict the price of a house. Therefore, we reject our null hypothesis.

1g) What is your best forecast of the price for a 2,000 square foot house built in 1990 with a swimming pool, a fireplace, and a garage?

Based on the data we have, the best estimate of the price for a 2,000 square foot house built in 1990 with a swimming pool, a fireplace, and a garage would be approximately 216,000 to 260,000. This was determined by filtering the data and matching the descriptions to find a range of prices for houses that met that description. Because our data does not have any information about whether the houses had garages and/or fireplaces or not, we can not tell how that would affect the price in question.

1h) Considering everything, explain why you think you have a (un)reasonable model?

There are several factors that go into the prices that people list homes for, as well as numerous different factors that make houses the perfect fit for different people. Due to all of these factors, I do not believe that the model is broad enough to accurately determine what aspects of a house significantly affect the price. However, we know that if the p-value associated with the f-statistic is less than our significance level, then the regression model better fits our data than if we had a model with no predictors.

1i) You present your results to Mr. Dimeworth. Trying not to let his admiration for your work out, he asks, "What does your model tell us about the value of a garage?" Can you use your estimated equation to impute the value of a garage? If not, how do you modify your model so you can measure the value of a garage?

Our current estimated equation for this problem set does not include any data or information regarding if the houses sampled had garages or not. If we wanted to impute the value a garage has on the price of a house we would need a set of data that also stated if the houses have garages. With this data we can make conclusions about the relation between garages and house price, and overall we would have a better/more accurate model because there would be more variables and data to incorporate in our results.

1j) Your estimated equation indicates that the effect of house size on price (i.e., slope parameter associated with area of the house) is the same for houses in all locations (North and not-North). How do you modify your equation so that the effect of house size on price depends on the location? For full credit, estimate a model that allows the effect of house size on house price to vary across the two locations. Write your estimated equation(s) and interpret key parameter estimates. Test the hypothesis that the effect of house size on price is the same for both locations (North and not-North).

The null hypothesis is that the effect of house size on price is the same for both locations, and the alternative is that the effect of house size on is different for both locations. If we wanted to estimate an equation to test this, we add an additional independent variable to our current equation, which is Area*D_Location. This variable is the size of the house multiplied by either 1, or 0, depending on if the house is in a North or Not-North location. When we rerun the regression, we can see that this is a more accurate model because our p-value associated with the f-statistic is extremely low. Therefore we can reject our null hypothesis because the effect of house size on price is different for the two locations.

1k) Suppose you measure house price in thousands of dollars instead of dollars and rerun the regression. What happens to the estimated coefficients? R-Square? Significance of the parameters?

If this regression was rerun using thousands of dollars instead of dollars, the estimated coefficients and R-Squared would not change because the relationship between the variables would still have the same variance and correlation.

ADDENDUM - R CODE

```
# DARIEN OLIVA - PSET 3 - ECON 453

> library(readxl)
> data1 <- read_excel("Downloads/pset3_data.xlsx")
> View(data1)
> summary(data1)

   Price              Area               Year          Pool             Location
 Length:58          Length:58          Min.   :63.00   Length:58          Length:58
 Class :character   Class :character   1st Qu.:67.00   Class :character   Class :character
 Mode  :character   Mode  :character   Median :73.00   Mode  :character   Mode  :character
                                       Mean   :76.12
                                       3rd Qu.:85.00
                                       Max.   :95.00
                                       NA's   :8
> str(data1)
tibble [58 × 5] (S3: tbl_df/tbl/data.frame)
 $ Price   : chr [1:58] "224700" "162200" "212300" "217300" ...
 $ Area    : chr [1:58] "1778" "1465" "1805" "2181" ...
 $ Year    : num [1:58] 95 76 65 80 95 72 70 63 85 88 ...
 $ Pool    : chr [1:58] "no" "no" "no" "no" ...
 $ Location: chr [1:58] "North" "Not-North" "North" "North" ...
> data1=data1[c(1:50), ]
> str(data1)
tibble [50 × 5] (S3: tbl_df/tbl/data.frame)
 $ Price   : chr [1:50] "224700" "162200" "212300" "217300" ...
 $ Area    : chr [1:50] "1778" "1465" "1805" "2181" ...
 $ Year    : num [1:50] 95 76 65 80 95 72 70 63 85 88 ...
 $ Pool    : chr [1:50] "no" "no" "no" "no" ...
 $ Location: chr [1:50] "North" "Not-North" "North" "North" ...
> data1$Price <- as.numeric(data1$Price)
> data1$Area <- as.numeric(data1$Area)
> str(data1)
tibble [50 × 5] (S3: tbl_df/tbl/data.frame)
 $ Price   : num [1:50] 224700 162200 212300 217300 236600 ...
 $ Area    : num [1:50] 1778 1465 1805 2181 1770 ...
 $ Year    : num [1:50] 95 76 65 80 95 72 70 63 85 88 ...
 $ Pool    : chr [1:50] "no" "no" "no" "no" ...
 $ Location: chr [1:50] "North" "Not-North" "North" "North" ...
> data1$D_Pool=ifelse((data1$Pool =="yes"), 1, 0)
> data1$D_Location=ifelse((data1$Location =="North"), 1, 0)
> str(data1)
tibble [50 × 7] (S3: tbl_df/tbl/data.frame)
 $ Price     : num [1:50] 224700 162200 212300 217300 236600 ...
```

```
 $ Area     : num [1:50] 1778 1465 1805 2181 1770 ...
 $ Year     : num [1:50] 95 76 65 80 95 72 70 63 85 88 ...
 $ Pool     : chr [1:50] "no" "no" "no" "no" ...
 $ Location : chr [1:50] "North" "Not-North" "North" "North" ...
 $ D_Pool   : num [1:50] 0 0 0 0 1 0 1 1 1 0 ...
 $ D_Location: num [1:50] 1 0 1 1 1 0 0 1 0 0 ...
> var(data1$Price)
[1] 688150127
> mean(data1$Price)
[1] 213774
> data1$Age=2022-(1900+data1$Year)
```

1A)

```
> model1 <-lm(data=data1, Price~Area+Age+D_Pool+D_Location)
> summary(model1,digits=5)

Call:
lm(formula = Price ~ Area + Age + D_Pool + D_Location, data = data1)

Residuals:
  Min    1Q Median    3Q   Max
-31632 -12096   -596 10584 29464

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 128258.45   22459.37   5.711 8.38e-07 ***
Area           66.72       9.59   6.956 1.18e-08 ***
Age         -1138.64     206.85  -5.505 1.69e-06 ***
D_Pool       5859.00    4903.19   1.195   0.2384
D_Location   9287.42    4727.85   1.964   0.0557 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15110 on 45 degrees of freedom
Multiple R-squared:  0.6952,  Adjusted R-squared:  0.6681
F-statistic: 25.66 on 4 and 45 DF,  p-value: 4.086e-11
```

1E)

```
> model1$coefficients["Age"]
     Age
-1138.645
```

1F)

```
> linearHypothesis(model1, c("Age = 0","Area=0","D_Pool=0","D_Location=0"))
Linear hypothesis test

Hypothesis:
Age = 0
Area = 0
D_Pool = 0
D_Location = 0

Model 1: restricted model
Model 2: Price ~ Area + Age + D_Pool + D_Location

  Res.Df       RSS Df  Sum of Sq     F    Pr(>F)
1     49 3.3719e+10
2     45 1.0277e+10  4 2.3442e+10 25.66 4.086e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

1J)

```
> model2=lm(data=data1, Price~Area+Age+D_Pool+D_Location+Area*D_Location)
> summary(model2)

Call:
lm(formula = Price ~ Area + Age + D_Pool + D_Location + Area *
    D_Location, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-31202.4 -12477.3    -23.4  10230.2  25127.7
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    139649.56   26435.72   5.283 3.77e-06 ***
Area             62.01      11.19   5.541 1.59e-06 ***
Age           -1184.54     214.92  -5.512 1.75e-06 ***
D_Pool         6579.52    4997.69   1.317    0.195
D_Location    -28688.40   46295.26  -0.620    0.539
Area:D_Location    19.33      23.44   0.825    0.414
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15170 on 44 degrees of freedom
Multiple R-squared:  0.6998,  Adjusted R-squared:  0.6657
F-statistic: 20.52 on 5 and 44 DF,  p-value: 1.611e-10

> data1$inter= data1$Area*data1$D_Location
> model3 <-lm(data=data1, Price~Age+Area+D_Pool+D_Location+inter)
> summary(model3)

Call:
lm(formula = Price ~ Age + Area + D_Pool + D_Location + inter,
   data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-31202.4 -12477.3    -23.4  10230.2  25127.7

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 139649.56   26435.72   5.283 3.77e-06 ***
Age         -1184.54     214.92  -5.512 1.75e-06 ***
Area           62.01      11.19   5.541 1.59e-06 ***
D_Pool       6579.52    4997.69   1.317    0.195
D_Location  -28688.40   46295.26  -0.620    0.539
inter          19.33      23.44   0.825    0.414
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15170 on 44 degrees of freedom
Multiple R-squared:  0.6998,  Adjusted R-squared:  0.6657
F-statistic: 20.52 on 5 and 44 DF,  p-value: 1.611e-10
```