Midterm1

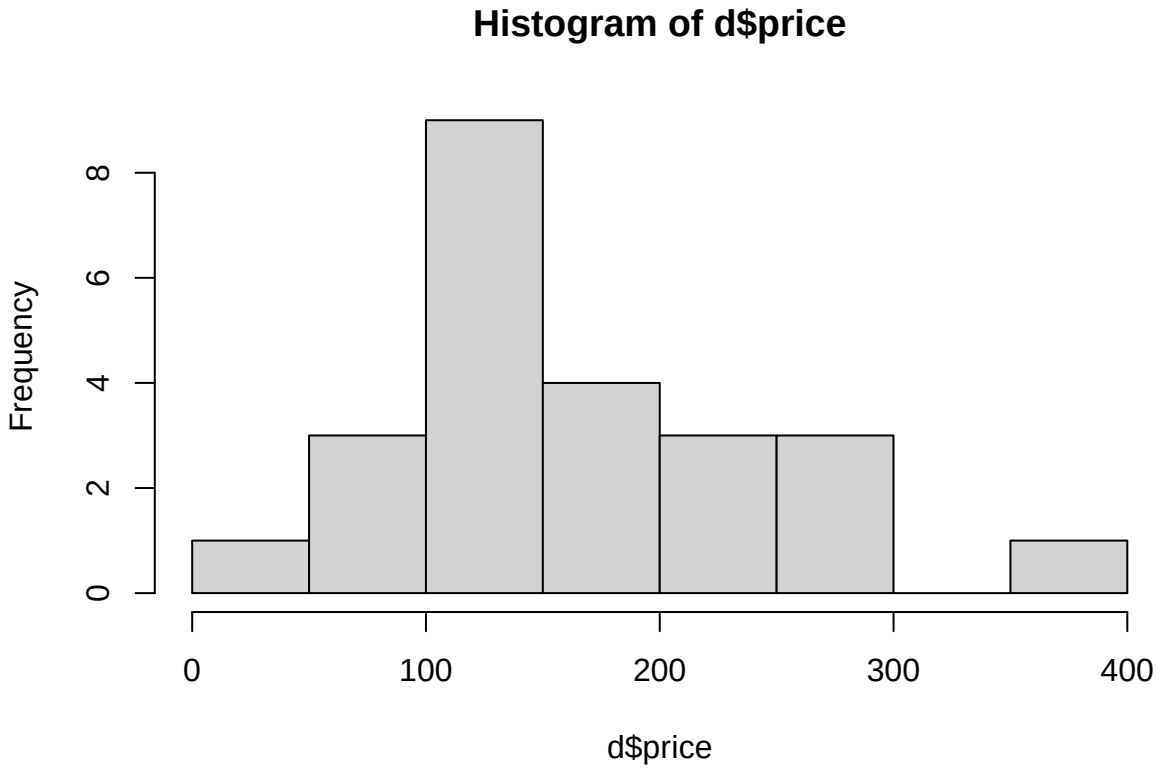
Aaron Graybill

3/14/2021

Problem 1.

a.

I plot the house prices as follows:



The data is definitely skewed right (because house prices have a lower bound at zero among other things), so the mean may not be the best measure of central tendency. I compare the mean to median with the following output.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    42.0   117.2   148.5   168.9   220.2   355.0
```

```
## [1] "sd=73.4935174110566"
```

As is expected with a skew right distribution, the mean is greater than the median. As such, I will report the typical house price as $148,500. The standard deviation is 73 thousand dollars which is quite high.

b.

Now I repeat a similar procedure for house size. The distribution is:

The distrubtions of sizes is much more uniform. It is still slightly skew right, but I predict the standard deviation will be lower relative to the mean when compared to price.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1040    1391    1845    1834    2180    2758
```

```
## [1] "sd=463.992689144678"
```

The median and the mean are relatively similar, so using the median we can safely say the typical price size is 1845 square feet. The standard deviation here is relatively lower than the price.

c.

The problem at hand can be setup in the following way:

$$\begin{cases} H_0: & \mu = 150,000 \\ H_a: & \mu \geq 150,000 \end{cases}$$

For ease of coding, I test against whether or not `d$price` is greater than 150 instead of converting everything to dollars. `R` reports everything that we need for this question from its `t.test` function:

```
t.test(d$price,alternative = "greater",mu=150)
```

```
##
##  One Sample t-test
##
## data:  d$price
## t = 1.261, df = 23, p-value = 0.11
## alternative hypothesis: true mean is greater than 150
## 95 percent confidence interval:
##  143.2055      Inf
## sample estimates:
## mean of x
```

```
##  168.9167
```

Reading from the output above, the test statistic is 1.261, the degrees of freedom are $n - 1 = 23$ and the p-value is .11. This is insufficient evidence to reject the null hypothesis at the $\alpha = .05$ level. We are not able to conclude that the mean house price is significantly above $150,000.

d.

Similar to before, we set up the null and alternative hypothesis as follows:

$$\begin{cases} H_0: & \mu = 2,000 \\ H_a: & \mu \leq 2,000 \end{cases}$$

We can compute the desiried numbers in the following way:

```
t.test(d$size,alternative = "less",mu=2000)
```

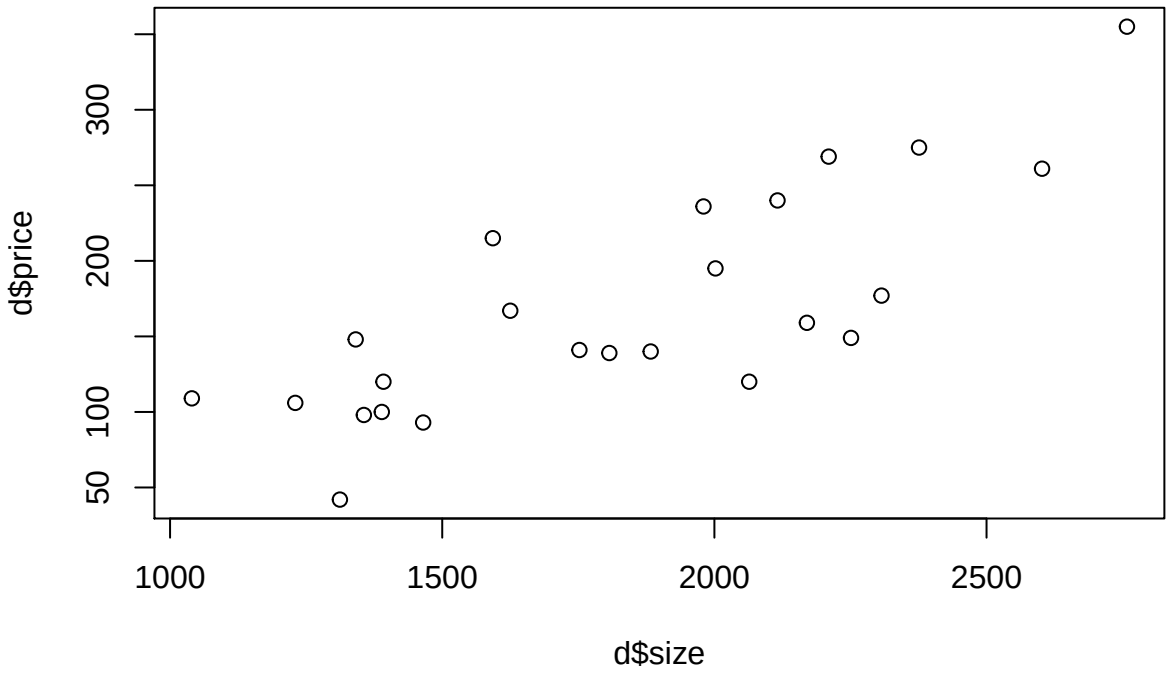
```
##
##  One Sample t-test
##
## data:  d$size
## t = -1.7505, df = 23, p-value = 0.04668
## alternative hypothesis: true mean is less than 2000
## 95 percent confidence interval:
##      -Inf 1996.533
## sample estimates:
## mean of x
##  1834.208
```

In this case the test statistic is -1.75, the degrees of freedom remains $n - 1 = 23$ and the p-value is 0.04668 which is below the $\alpha = .1$ level, so we reject the null hypothesis that the true mean house size is 2000 and we assert that the true mean house size is less than 2000.

Problem 2.

a.

Let's see if a linear model is appropriate for the data with the following scatter plot:



There seems to be quite a bit of variance from the regression line, but the relationship does appear to be rather linear, so a linear model should be appropriate. We can write the linear model as follows:

$$Price_i = \beta_0 + \beta_1 Size_i + e_i$$

where $e_i \sim N(0, \sigma)$.

b.

I believe this question is asking whether or not the estimated $\hat{\beta}_1$ is signiifically different from zero (indicating that the size term does add to the strength of the model), however, it may be asking to run an ANOVA to see if the model has explanatory power at all (so including the intercept term). I will proceed with the following null and alternative hypotheses:

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_a: & \beta_1 \neq 0 \end{cases}$$

Thankfully, `R` computes this test with its standard `summary(out)` values, so we can read of the table below:

```
##
## Call:
## lm(formula = price ~ size, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.701 -31.008   0.486  38.555  76.298
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -60.84201   39.05465  -1.558    0.134
## size          0.12526    0.02067   6.061 4.23e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 45.99 on 22 degrees of freedom
## Multiple R-squared:  0.6254, Adjusted R-squared:  0.6084
## F-statistic: 36.73 on 1 and 22 DF,  p-value: 4.231e-06
```

Above we see that the t-value, the test stat, on the size term is 6.061 which implies a p-value of 4.23×10^{-6} which is well less than the $\alpha = .01$ significance level. Therefore, we reject the null hypothesis that $\beta = 0$ and accept that there is some linear relationship between size and price.

c.

This question is a bit of a trick because it has a very similar setup to the previous in that we have the following hypotheses:

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_a: & \beta_1 \geq 0 \end{cases}$$

So we will have the exact same test statistic, the only thing that will change is that we can cut off one of the tails when computing the p-value which implies that the p-value is simply the previous value divided by two which gives that the p value is 4.12×10^{-6} which is well below the stated $\alpha = .1$ level.

d.

This question is also a little tricky to setup because we have to keep in mind the units. So, it's asking if a 100 square foot increase leads to a $10,000$ increase in price which is equivalent to a 1 square foot increase causing a \$100 increase. But we have constructed the units such that price is in thousands, so in out regression model we have 1 square foot increase causing a .1 thousands of dollar increase. The question is asking if the β_1 term is signifcantly at or above .1. The hypothesis, in our units is as follows:

$$\begin{cases} H_0: & \beta_1 = .1 \\ H_a: & \beta_1 > .1 \end{cases}$$

As far as I know we have to compute the test stat and p-value manaully using the formula given found class:

$$\frac{\hat{\beta}_1 - \beta_1^0}{S/\sqrt{SXX}}$$

When computing the p-value we get that $p = .1173$, which implies that we fail to reject the null hypothesis that the true slope is $= .1$.

e.

`R` quite compactly computes the desired confidence interval as follows:

```
##                        5 %      95 %
## (Intercept) -127.90447210 6.2204563
## size           0.08977309 0.1607532
```

At this stage we only need to focus on the second confidence interval which states that we are 90% confident that the true slope falls between .090 and .16.

f.

Bonferonni requires that if we are computing 2 confidence intervals at a joint confidence level $\alpha^* = ..05$, that we compute the individual confidence intervals as though they are at $\alpha^i = 1 - \frac{\alpha^*}{2} = .975$ Computing this gives:

```
##                     1.25 %    98.75 %
## (Intercept) -154.78689241 33.1028766
## size           0.07554667  0.1749796
```

Therefore we are 95% certain that the true estimates of β_0 and β_1 fall in the region $\{[-154.786, 33.102], [0.0755, 0.1749]\}$.

g.

This question is simply asking for the R^2 of the model, so reading off of the summary table above we have that 62.54% of the variation in price can be explained through the linear relationship with size and the intercept term.

h.

We can get a good portion of this informtion by creating a prediction interval at 2000 square feet as the size input we have:

```
## $fit
##        fit      lwr      upr
## 1 189.6842 108.8677 270.5008
##
## $se.fit
## [1] 9.993722
##
## $df
## [1] 22
##
## $residual.scale
## [1] 45.9912
```

The model estimates that the price of a house with 2000 square feet of floor space would be $\$189,684$. The standard error of this prediction requires some massaging to find. We can compute it in the following way:

```
## [1] "Standard prediction error is: 47.0644803580569"
```

i.

The output above gives the prediction interval, but to summarise the results, are 90% sure that the true value of price when size is 2000 is between $\$108,868$ and $\$270,501$.

j.

Now doing a very similar thing but with confidence intervals gives:

```
## $fit
##        fit      lwr      upr
## 1 189.6842 172.5236 206.8449
##
## $se.fit
## [1] 9.993722
##
## $df
## [1] 22
##
## $residual.scale
## [1] 45.9912
```

Here we are 90% sure that the mean price when size is 2000 lies between $\$172,524$ and $\$206,845$.

k.

This question is asking us to compute three confidence intervals simultaneously using the bonferroni method at a group confidence level of $\alpha^* = .1$, but since we are computing 3 intervals, the individual confidence levels need to be $1 - \alpha^*/3$. Computing this gives:

```
##        fit      lwr      upr
## 1 127052.7 115110.2 138995.2
## 2 189684.2 179444.2 199924.3
## 3 252315.8 235247.3 269384.3
```

The joint confidence intervals are the three intervals above given by `lwr` and `upr` respectively. The numbers have already been multiplied by 1000 to get into the regular $ units.

l.

Only two observations are truly above $4/n$, this are observations 3 and 8 which correspond to houses of size 2758 and 1040 respectively. These are the minimum and maximum of the sample. Also of note is observation 19 which misses the cutoff by a hair and has a size of 2602, also quite large.

m.

The smallest standardized residual is: -1.73549717 which occurs at observation 17, and the largest standardized residual is 1.72576558 which occurs at observation three. Neither of these values have an absolute value greater than two, so cannot be considered outliers.

n.

`R` identifies observations 3, 11, and 14 as having unusually large cooks distance values, but the real standout is observation 3 which has a cooks distance equal to .405, 11 and 14 have cooks distances of 0.1107206238 and 0.1060215040 respectively. All of these cooks distances are fairly small.

o.

We can test normality using the Shapiro-Wilk test which tests for normality. If we get a low p-value, it is unlikely that our errors are normal.

```
##
##  Shapiro-Wilk normality test
##
## data:  MASS::stdres(out)
## W = 0.96485, p-value = 0.5432
```

The p-value of .5 indicates that we do not have a ton of evidence that the errors aren't normal, so assumning normality is a relatively safe bet.

Problem 3.

a.

We showed in class that $\hat{\beta}_1 = \frac{SXX}{SXY}$ and that $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$. That means the regression line is given by:

$$\hat{y} = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x$$

Plugging in $x = \overline{x}$ immediately cancels down to:

$$\hat{y}_{\overline{x}} = \overline{y}$$

b.

We are given the following:

$$MSReg = \frac{\sum (\hat{y}_i - \overline{y})^2}{1} = \sum (\hat{y}_i - \overline{y})^2$$

Using the substitutions that we have and doing some algebra gives:

$$\begin{aligned}
MSReg &= \sum (\hat{y}_i - \overline{y})^2 \\
&= \sum \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - \overline{y}\right)^2 \\
&= \sum \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - \overline{y}\right)^2 \\
&= \sum \left(\overline{y} - \beta_1 \overline{x} + \hat{\beta}_1 x_i - \overline{y}\right)^2 \\
&= \sum \left(\beta_1 \overline{x} + \hat{\beta}_1 x_i\right)^2 \\
&= \hat{\beta}_1^2 \sum (\overline{x} - x_i)^2 \\
&= \hat{\beta}_1^2 \sum (x_i - \overline{x})^2
\end{aligned}$$

Everything until that last step is fairly straightforward. In the last step I use the fact that since the quantity is being squared I am free to multiply the inside by -1 without changing the value.

Now applying the expectation operator and using the properties given:

$$\begin{aligned}
E[MSReg] &= E\left[\hat{\beta}_1^2 \sum (x_i - \overline{x})^2\right] \\
&= \left(\sum (x_i - \overline{x})^2\right) E\left[\hat{\beta}_1^2\right] \\
&= \left(\frac{\sigma^2}{\sum (x_i - \overline{x})^2} + \beta_1^2\right) \sum (x_i - \overline{x})^2 \\
&= \sigma^2 + \beta_1^2 \sum (x_i - \overline{x})^2
\end{aligned}$$

The second step comes from the fact that the x_is are not random variables so the sum over them can be treated a constant and factored out of the expectation.