Data Science HW4

Matt McCullough

1) Least squares through $(\bar{x}, \bar{y})$

$\hat{Y} = \hat{B}_0 + \hat{B}_1 x$

$x = \bar{x}$

$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}$

$\hat{B}_1 = \frac{Sxy}{Sxx}$

$\hat{Y} = \bar{y} - \hat{B}_1 \bar{x} + \hat{B}_1 \cdot \bar{x}$

$SSx = \sum_{i=1}^{n} (x_i - \bar{x})^2$

$\hat{y} = \bar{y} - 0$

$SSy = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

$\hat{y} = \bar{y}$ when $x = \bar{x}$

Therefore, the least squares regression line passes through $(\bar{x}, \bar{y})$

2) Show that $\sum_{i=1}^{n} e_i = 0$ where $\hat{y}_i = \hat{B}_0 + \hat{B}_1 x_i$

$$\begin{aligned}
\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} (y_i - \hat{B}_0 - \hat{B}_1 x_i) \\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{B}_0 - \hat{B}_1 \sum_{i=1}^{n} x_i \\
&= \sum_{i=1}^{n} y_i - n\hat{B}_0 - \hat{B}_1 \sum_{i=1}^{n} x_i \\
&= n\bar{y} - n\hat{B}_0 - \hat{B}_1 n\bar{x} \\
&= n(\bar{y} - \hat{B}_0 - \hat{B}_1 \bar{x}) \\
&= 0
\end{aligned}$$

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

$\sum_{i=1}^{n} = n\bar{x}$

$\bar{y}$ from last proof

$\bar{y} - (\hat{B}_0 - \hat{B}_1 \bar{x}) = 0$

3) a) Find the least squares estimator of B_1

$$SSE = \sum_{i=1}^{n} [y_i - \hat{B}_1 x_i]^2$$

minimize SSE $\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - B_1 x_i)^2$

$$SSE(B_1) = \sum_{i=1}^{n} (y_i - B_1 x_i)^2$$

$$\frac{dSSE}{dB_1} = \frac{dSSE}{dB_1} \sum_{i=1}^{n} (y_i - B_1 x_i)^2 = \sum_{i=1}^{n} \frac{d}{dB_1} (y_i - B_1 x_i)^2$$

$$= \sum_{i=1}^{n} 2(y_i - B_1 x_i)(-x_i) = -2\sum_{i=1}^{n} 2(y_i - B_1 x_i^2) = 0$$

$$\rightarrow \sum_{i=1}^{n} (x_i y_i - B_1 x_i^2) = 0 \qquad \text{critical min point}$$

$$\sum_{i=1}^{n} x_i y_i - B_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$-B_1 \sum_{i=1}^{n} x^2 = -\sum_{i=1}^{n} x_i y_i$$

$$-B_1 = \frac{-\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$\hat{B}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

double check

$$\frac{d^2 SSE}{dB_1^2} = -2\sum_{i=1}^{n}(-x_i^2) = 2\sum_{i=1}^{n} x_i^2 \geq 0$$

✓

b) EC

Replace x_i with $f(x_i)$ $\Rightarrow$ $\hat{B}_1 = \dfrac{\sum_{i=1}^{n} f(x_i) y_i}{\sum_{i=1}^{n} (f(x_i))^2}$

pretty much the same proof as a

4) Show that the slope can also be derived from the formula

$\hat{B}_1 = r\frac{sy}{sx}$ where sx and sy are the standard deviations of the x and y values respectively

have to show $r\frac{sy}{sx} = \frac{SSxy}{SSxx}$

$$r \cdot \frac{sy}{sx} = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(\frac{x_i - \bar{x}}{sx}\right)\cdot\left(\frac{y_i - \bar{y}}{sy}\right)\right]\cdot\frac{sy}{sx} =$$

$$\frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{(x_i - \bar{x})(y_i - \bar{y})}{sx\,sy}\right]\cdot\frac{sy}{sx} =$$

$$\frac{SSxy}{(n-1)\,sx\,sy}\cdot\frac{sy}{sx} = \frac{SSxy}{(n-1)\,sx^2}$$

Show that $(n-1)\,sx^2 = SSxx$

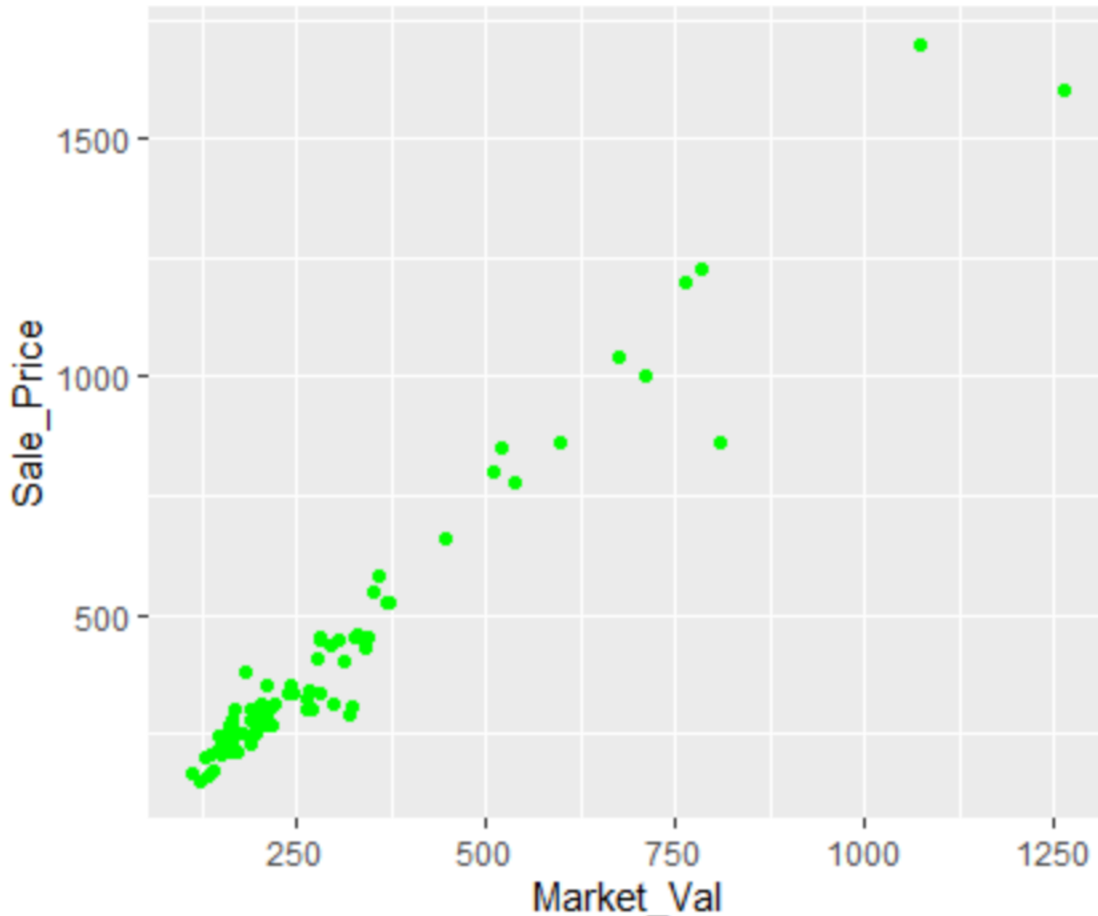
$$(n-1)(sx^2) = (n-1)\left(\sqrt{\frac{SSxy}{n-1}}\right)^2 = (n-1)\cdot\frac{SSxx}{n-1} = SSxx$$

Therefore $r\frac{sy}{sx} = \frac{SSxy}{SSxx}$ and can be used for slope

5a.

The explanatory variable is the market value, and the response variable is the sale price

this scatterplot does show an approximately linear relationship between the two variables



b.

```
> summary(reg)

Call:
lm(formula = Sale_Price ~ Market_Val, data = TAMPALMS)

Residuals:
     Min       1Q   Median       3Q      Max 
-282.171  -24.829    1.807   29.791  188.792 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.35868   13.76817   0.099    0.922    
Market_Val   1.40827    0.03693  38.132   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68.76 on 74 degrees of freedom
Multiple R-squared:  0.9516,    Adjusted R-squared:  0.9509 
F-statistic:  1454 on 1 and 74 DF,  p-value: < 2.2e-16
```

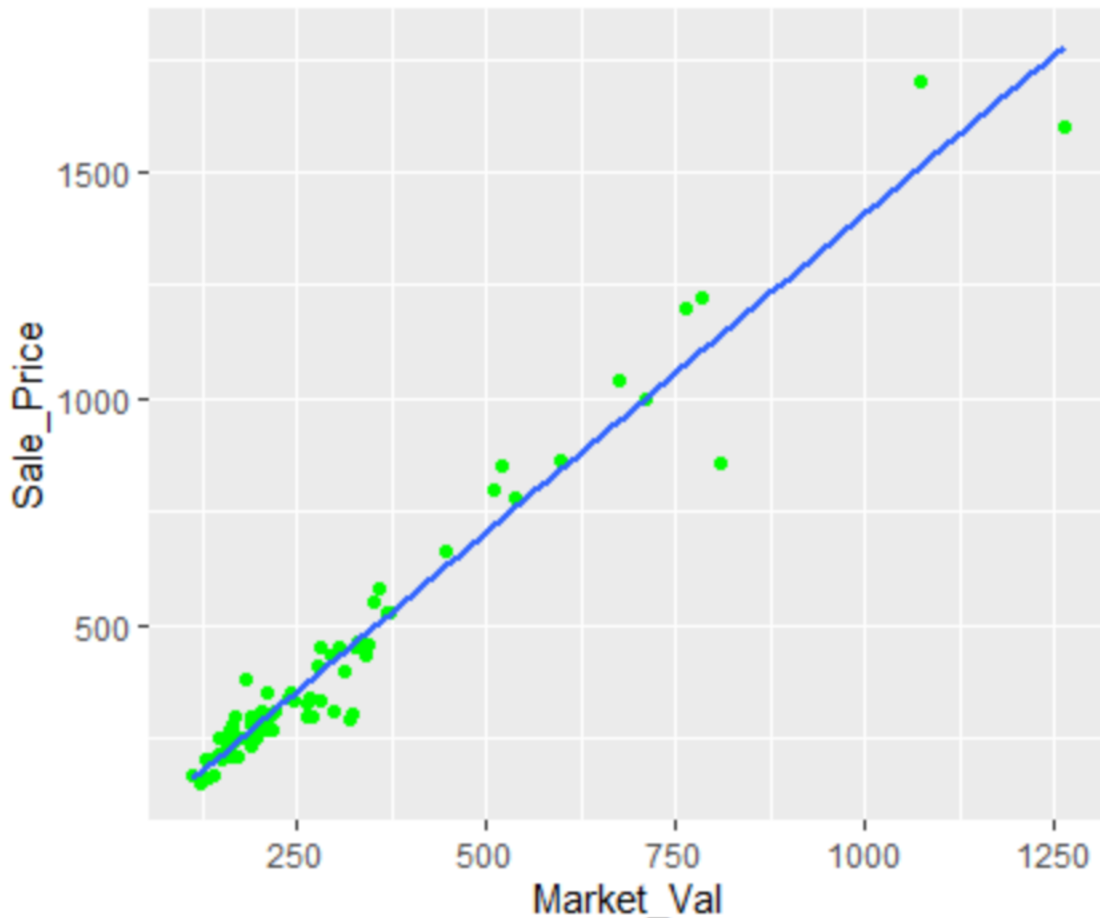
.

```
> anova(reg)
Analysis of Variance Table

Response: Sale_Price
           Df  Sum Sq Mean Sq F value    Pr(>F)    
Market_Val  1 6874034 6874034  1454.1 < 2.2e-16 ***
Residuals  74  349833    4727                      
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c.

The least squares regression line is y = 1.408271x + 1.358681



d.

It would be appropriate the slope of this linear model, as for every increase of market value (which is 1.408271 thousand dollars) to the market value of a property, there is an increase of 1.35881 thousand dollars of to the predicted sale price on average.

e.

null: B1 = 0

alternate B1 > 0

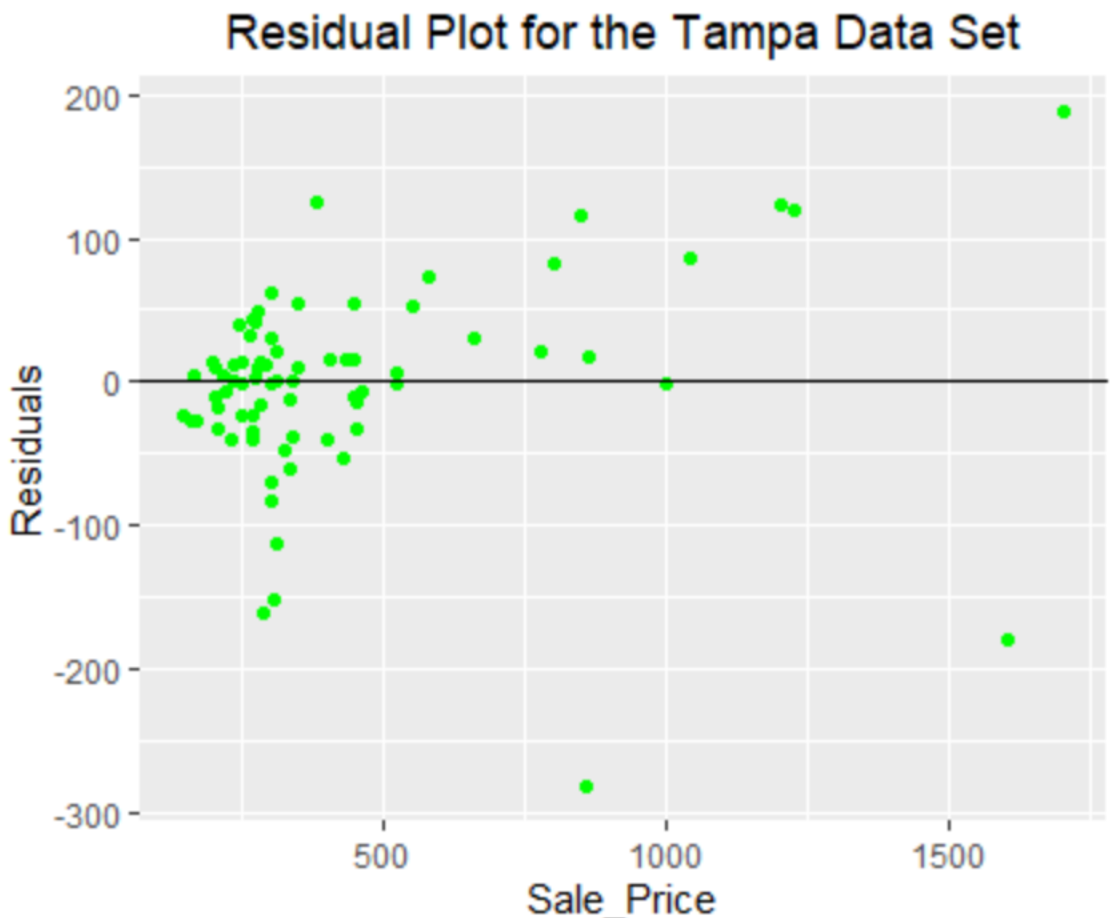
There is sufficient evidence that there is a positive linear relationship between appraised and property value and sale price for residential properties sold. β1 the slope of the straight-line model, is positive as t was 38.132 and the P value was very small as it's < .0001

The formula I used was P(T > 38.132) = <2e-16 (given by summary) < .05, so the null hypothesis can be rejected and the slope is positive f.

The 95% confidence interval for the slope is 1.334683 - 1.481858

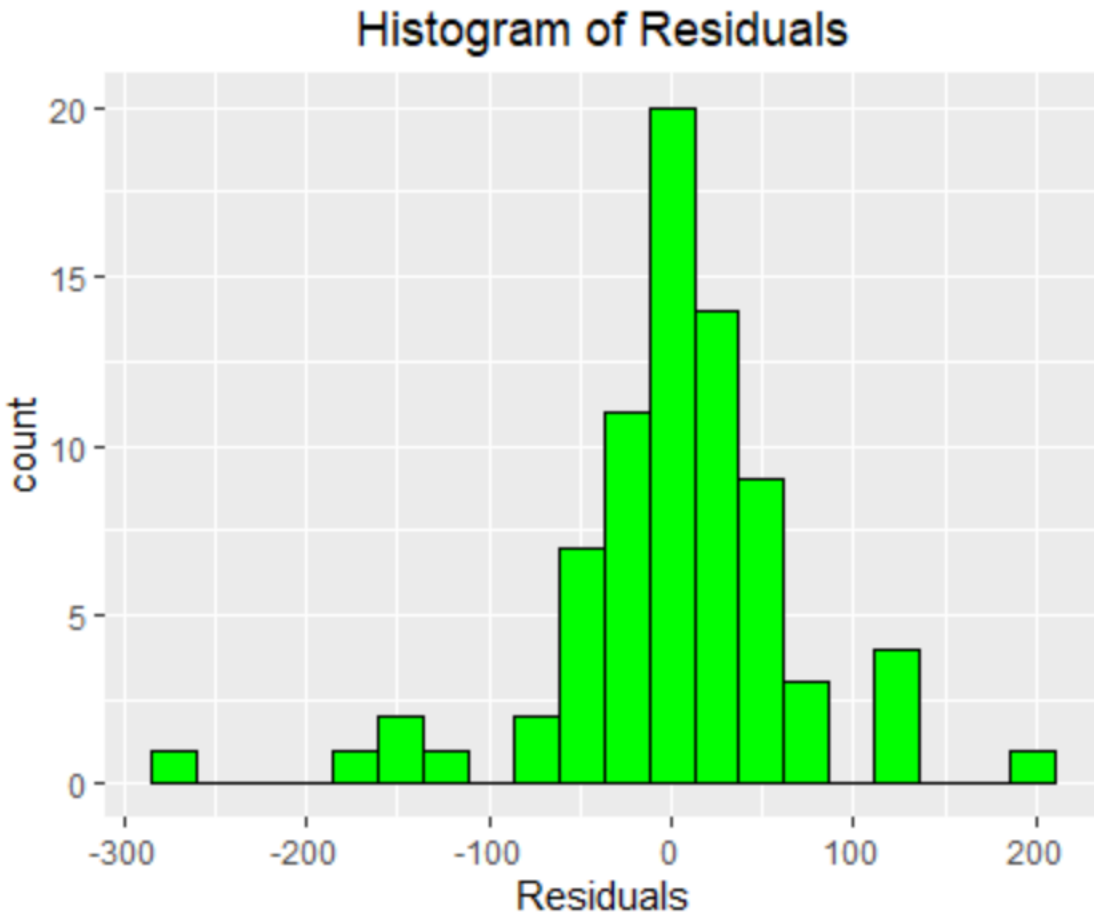
g.

This is a good residual plot as most of the plots are relatively near the line at 0. The more of the data near the line, the higher r^2 will be.



h.

Yes, this distribution of the residuals is close to a normal distribution with a mean of zero as its center is near zero and looks like a mound shape distribution



i.

Residual standard error: 68.76 on 74 degrees of freedom (s is 68.76)

Approximately 95% of the data points lie within 2s or 137.5132 thousand dollars

j.

95.16% of the variability can be explained by the regression. Yes, I would consider this regression to be a success as r^2 is .9516 which is very close to 1 and because of that can make mostly accurate predictions

k.

The 95% prediction interval of the regression line for the estimate of 300000 dollars is 285.9404 - 561.7394