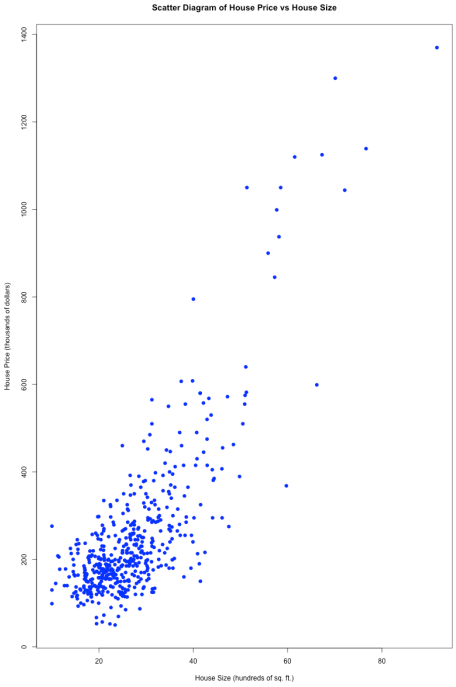
HW0303

Question 2.17

Part a:



I think from the initial scatter plot we can see that majority of houses fall into the category of being under 4000 square feet and under the price of 400 000 US dollars. There is a bit of spread in terms of some bigger houses can be on the cheaper while smaller ones can be more expensive but overall there seems to be a trend where the bigger the house the higher it's likely to cost.

Part b:

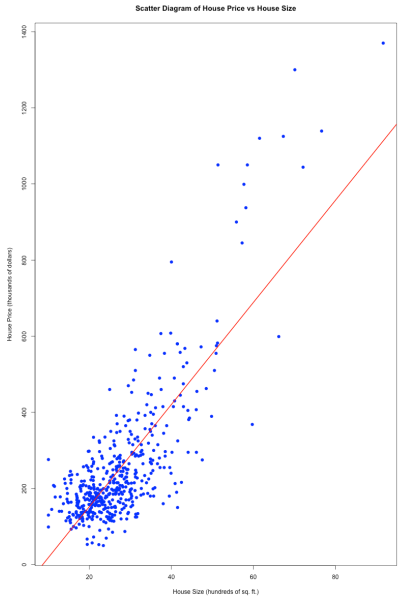
I estimated a linear regression and these were the results I received:

```
Call:
lm(formula = price ~ sqft, data = collegetown)

Residuals:
    Min      1Q  Median      3Q     Max
-316.93  -58.90   -3.81   47.94  477.05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -115.4236    13.0882  -8.819   <2e-16 ***
sqft          13.4029     0.4492  29.840   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 102.8 on 498 degrees of freedom
Multiple R-squared:  0.6413,    Adjusted R-squared:  0.6406
F-statistic: 890.4 on 1 and 498 DF,  p-value: < 2.2e-16
```

From these results I can estimate the intercept and coefficient for the linear regression. From the results we see that the linear regression is:

$$\widehat{\text{price}} = -115.4236 + 13.4029 \times \text{sqft}$$

Using the estimate for the linear regression, I plotted a fitted line over the scatter plot I created earlier.



Part c:

Using R I estimated the quadratic regression model with the following result:

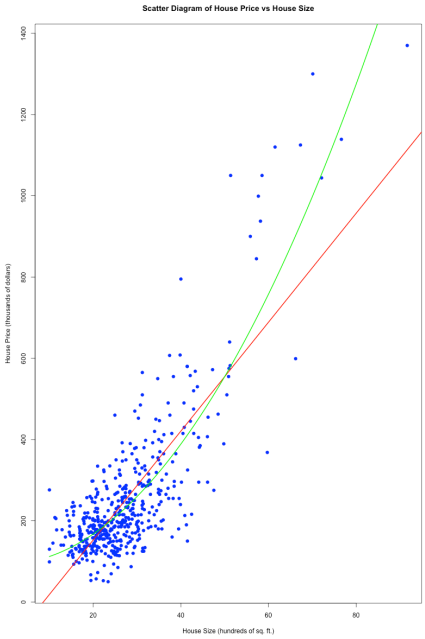
```
Call:
lm(formula = price ~ I(sqft^2), data = collegetown)

Residuals:
    Min      1Q  Median      3Q     Max
-383.67  -48.39   -7.50   38.75  469.70

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 93.565854   6.072226   15.41   <2e-16 ***
I(sqft^2)    0.184519   0.005256   35.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 92.08 on 498 degrees of freedom
Multiple R-squared:  0.7122,    Adjusted R-squared:  0.7117
F-statistic:  1233 on 1 and 498 DF,  p-value: < 2.2e-16
```

From here I estimate that the regression formula will look something like this:

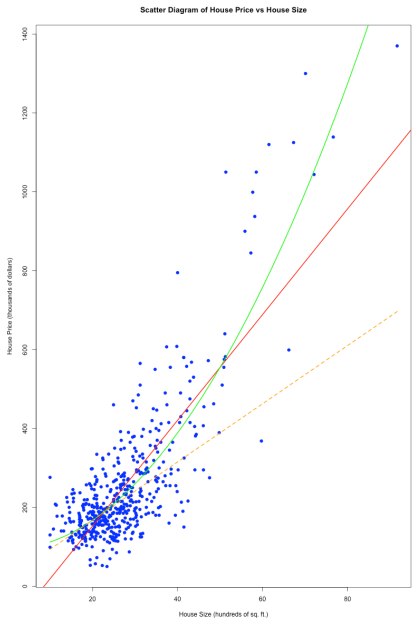
$$\widehat{\text{PRICE}} = 93.565854 + 0.184519\,(\text{SQFT})^2 + e\,.$$

I used R to estimate that the marginal effect of an additional 100 square feet of living area in a home with 2000 square feet of living space would be around $7.38076 (or it could also be 73807.6$ - which I think is more accurate).

Part d:



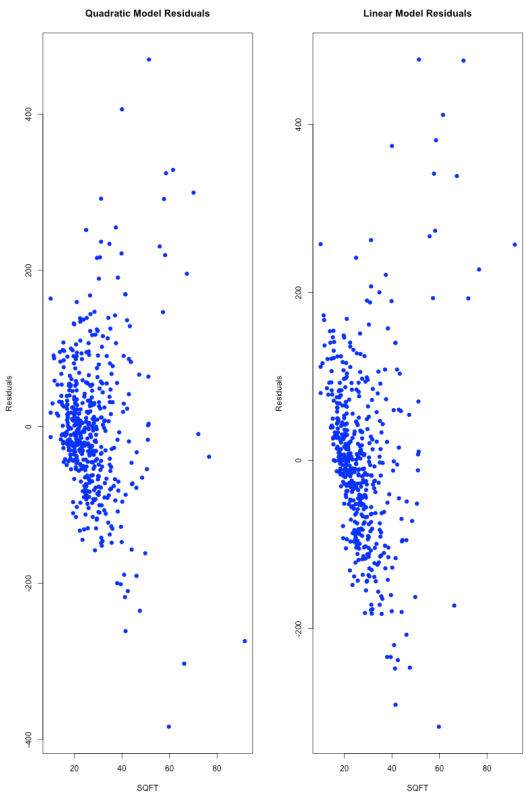
The graph to the left represents a fitted quadratic regression line for house price to square feet. While the graph on the right also depicts the tangent (orange dotted line).



Part e:

I did the calculations in r and I got an elasticity of: 0.8819511, meaning that the price is inelastic. To put in a more simple way - it's cheaper to buy bigger properties. For the step by step solution check the relevant r file.

Part f:



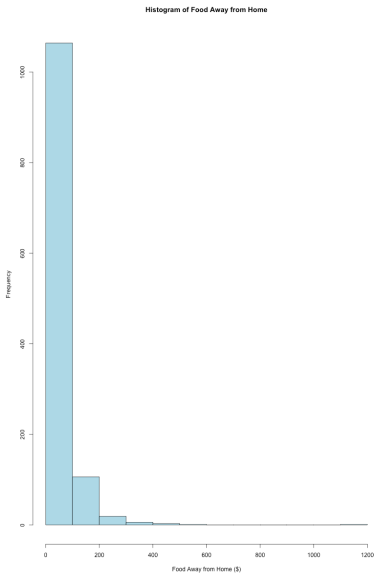
Based on what I can see in the residual plot the quadratic model seems to be the best fit. It does confirm my earlier assumption in terms which model would be a better for this specific data set. I came to this conclusion based on the clustering and spread of the residuals - the quadratic model seems to have less of a trend and seems to be more even than the linear model.

Part g:

The quadratic model has a lower SSR. From the analysis carried out in R, I got that the linear model has an SSR of 5262847, while the linear model has an SSR of 4222356. A lower SSR means that, on average, the difference between the observed values and the model's predicted values is smaller. So the quadratic model seems to be the best fit for this data set.

Question 2.25

Part a:



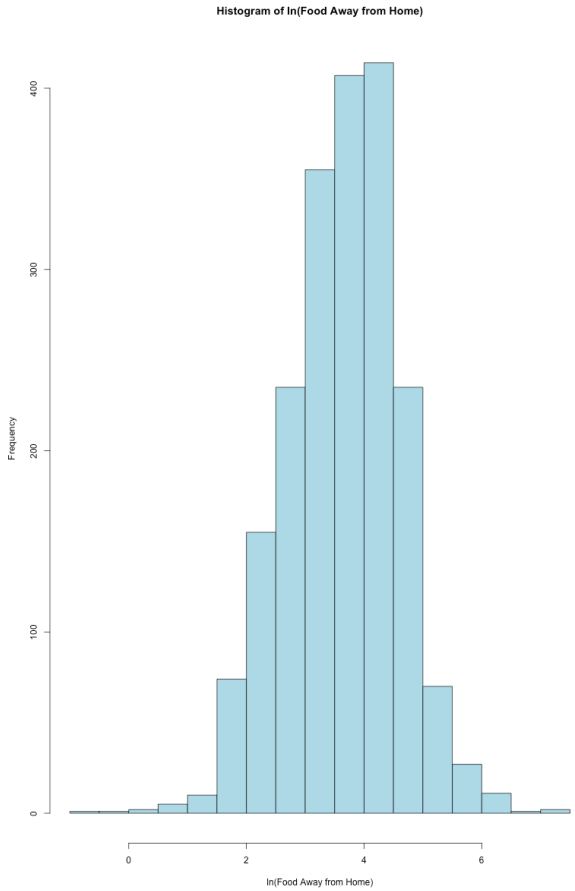

```
> mean(cex5$foodaway, na.rm = TRUE)
[1] 51.45131
> median(cex5$foodaway, na.rm = TRUE)
[1] 33.145
> quantile(cex5$foodaway, probs = c(0.25, 0.75), na.rm = TRUE)
    25%     75%
14.1175 69.4400
>
```

The histogram on the right represents the distribution of household spending on food consumed while away from home. It shows the frequency of observations for different spending amounts. As can be seen it's skewed to the left because the data is not normalized and majority of households tend to spend a relatively small amount, as confirmed by the mean spending, $51.45, being higher than the median spending, $33.15. Most families tend to spend between $14.12 and $69.44 as indicated by the 25 and 75 percentiles.

Part b:

Education Level	Mean Foodaway ($)	Median Foodaway ($)
Advanced Degree	75.37	48.34
College Degree (No Advanced)	54.90	36.11
No College or Advanced Degree	38.32	25.02

The results indicate a positive association between education level and expenditure on food consumed away from home. With families who have at least one family with an advanced degree tend to spend more on food while away from home. I think this could be an issue of preference - as the data set is based on US consumption families who can afford to send one family member tend to have a higher socio-economic standing, hence they may lean towards more expensive place compared to families where no family member has a degree.



Part c:

The variables FOODAWAY and ln(FOODAWAY) differ in the number of observations because the natural logarithm is undefined for non-positive values (i.e., zero or negative numbers). As a result, any household with zero expenditure on food away from home is excluded from the log-transformed variable. This reduction in valid observations leads to a smaller sample size for ln(FOODAWAY) compared to FOODAWAY.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.734   3.076   3.756   3.659   4.280   7.406
```

Part d:

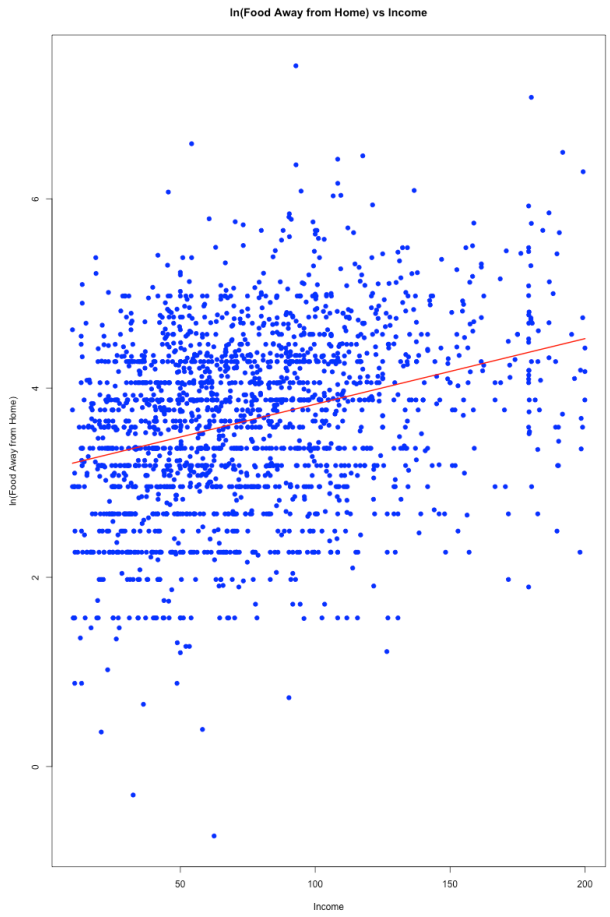
$$\ln(\text{FOODAWAY}_i) = 3.1365 + 0.0069 \cdot \text{INCOME}_i + \varepsilon_i$$

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.3034 -0.6003  0.0654  0.6119  3.6262

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.1364599  0.0420567   74.58   <2e-16 ***
income      0.0069283  0.0004898   14.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.898 on 2003 degrees of freedom
Multiple R-squared:  0.09083,   Adjusted R-squared:  0.09038
F-statistic: 200.1 on 1 and 2003 DF,  p-value: < 2.2e-16
```

The estimated model regresses the natural logarithm of household expenditure on food consumed away from home (\ln(\text{FOODAWAY})) on household income. The coefficient on INCOME is 0.0069, which is statistically significant at the 1% level. This suggests that, holding other factors constant, a $1 increase in income is associated with approximately a 0.69% increase in food-away-from-home

spending. The model indicates a positive and statistically significant relationship between income and food-away spending, consistent with the expectation that higher-income households are more likely to spend more when dining outside the home.

Part e:



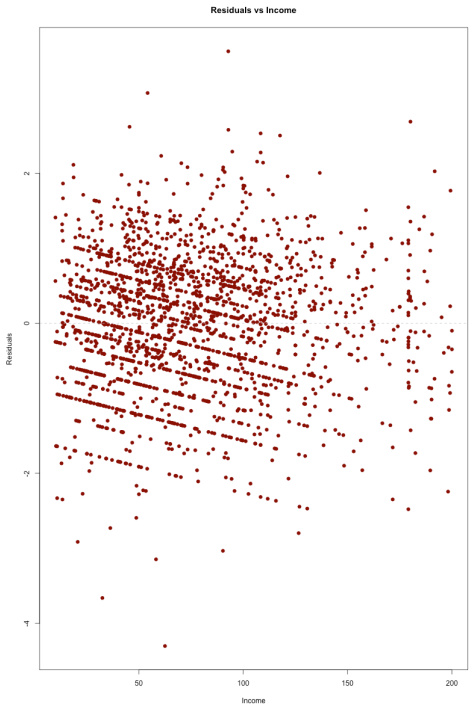
I plotted this in R, but I don't know why the line is straight even though it s logarithm and should be curved.

Question 2.17

Part a:

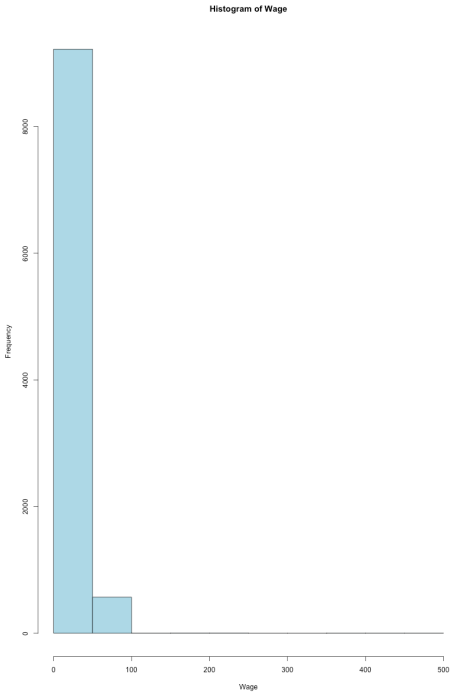
Part f:

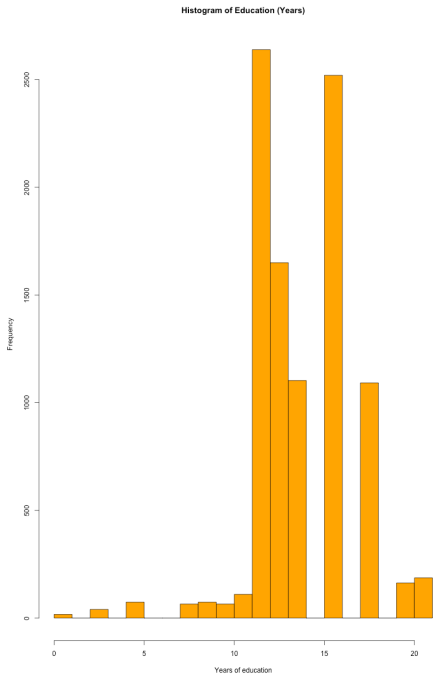
The residuals are generally centered around zero but show signs of heteroscedasticity, as their variance appears to increase with income. Additionally, the banding pattern suggests the outcome variable is discretely measured. These observations suggest that while the linear specification may be appropriate in form, the model could benefit from robust standard errors or alternative specifications that address heteroscedasticity.



Question 2.28

Part a:





Statistic	Wage ($/hr)	Education (Years)
Minimum	2.50	0.00
1st Quartile	13.00	12.00
Median	19.23	14.00
Mean	23.46	14.21
3rd Quartile	29.07	16.00
Maximum	466.00	21.00

The wage distribution is right-skewed, with most individuals earning between $10–30/hour and a few extreme outliers pushing the maximum to $466/hour, indicating notable wage inequality. In contrast, education levels are clustered around key milestones—12, 14, and 16 years—suggesting many individuals complete high school or college, with a fairly symmetric distribution overall. While wage data

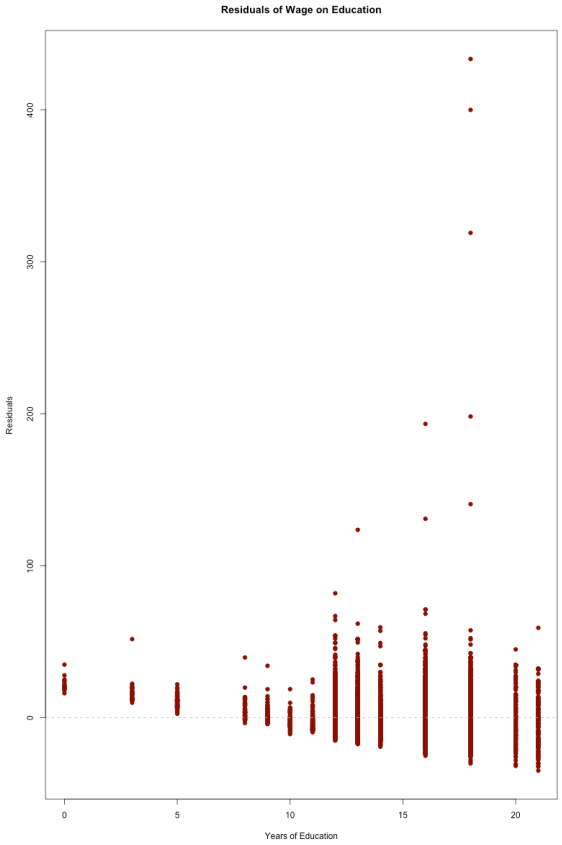
shows high variability and outliers, education is more structured and evenly distributed, making it a potential predictor of earnings.

Part b:

```
Call:
lm(formula = wage ~ educ, data = cps5)

Residuals:
   Min     1Q Median     3Q    Max
-34.82  -8.36  -3.11   5.67 433.38

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.90593    0.77426  -14.09   <2e-16 ***
educ          2.41817    0.05348   45.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.62 on 9797 degrees of freedom
Multiple R-squared:  0.1727,    Adjusted R-squared:  0.1726
F-statistic:  2044 on 1 and 9797 DF,  p-value: < 2.2e-16
```

The regression results show a strong, statistically significant positive relationship between education and wage. As can be seen in the above results each additional year of education increases hourly wage by about $2.42 (which in all honesty is depressing if you think about it). However, the model’s R-squared is 0.173, meaning education alone explains just 17.3% of the variation in wages (which makes sense, because while education does usually translate to higher earning potential the spread just within the same role can be massive) . The negative intercept isn’t meaningful in practice but is needed to fit the line. Overall, while education is an important factor, other variables likely influence wages and should be included for a more complete model.

Part c:



The residual plot shows a clear increase in the spread of residuals as education increases, indicating heteroskedasticity. This violates assumption SR3, which requires constant error variance. If the classical regression assumptions SR1–SR5 held, the residuals would appear randomly scattered with no visible pattern or trend.

Part d:

Group	Intercept	Educ Coefficient	Std. Error (Educ)	R-squared	Residual Std. Error	Obs (df + 1)
Male	-9.51	**2.45**	0.064	**0.215**	13.39	5423
Female	-14.83	**2.53**	0.091	0.150	15.73	4374
Black	-11.58	**2.26**	0.158	0.189	11.27	876
White	-10.64	**2.42**	0.059	0.165	14.97	8413

I ran the regression models for males, females, blacks and whites. I used ChatGPT to turn the results into an easily presentable table, just so that they can be easily compared side by side. Some core insights are that females seem to have the highest estimated return to education ($2.53/hour), but the model fits better for males (highest R^2 = 21.5%). Blacks show a slightly lower return ($2.26) compared to whites

($2.42), with both groups having R^2 below 20%. Which is something unfortunately confirmed by other studies, for example a paper which compared customer satisfaction across different ethnic groups found that minorities and women are more likely to receive negative or critical feedback from untrained evaluators (in other words everyday customers). As such all models are statistically significant ($p < 2e\text{-}16$), but R-squared values are modest, showing that education alone doesn't fully explain wage differences across groups (something which was confirmed earlier.

Part e:

Estimated quadratic regression model:

```
Call:
lm(formula = wage ~ educ2, data = cps5)

Residuals:
   Min     1Q Median     3Q    Max
-39.03  -7.96  -2.61   5.17 432.34

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.776106   0.421996   11.32   <2e-16 ***
educ2       0.089143   0.001888   47.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.51 on 9797 degrees of freedom
Multiple R-squared:  0.1854,    Adjusted R-squared:  0.1853
F-statistic:  2229 on 1 and 9797 DF,  p-value: < 2.2e-16
```