Does length increase prices in diamonds?

A sample of inference on plots, models, and hypothesis tests for Aengus White, using the "diamonds" data set in ggplot2. d1 is the set of diamonds with less than average length, and d2 is the set of diamonds with more than (or equal to) the average length for the whole set diamonds.

```
t.test(d1$price, d2$price, alternative = "two.sided", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  d1$price and d2$price
## t = 223.21, df = 27533, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5581.940 5680.843
## sample estimates:
## mean of x mean of y
##  6852.479  1221.087
```

Hypothesis Test Output

Using a Welch two sample t-test:

$h_o : \mu_{d1} = \mu_{d2}$
$h_a : \mu_{d1} \neq \mu_{d2}$

From this test, with test statistic 223.21 and p-value $< .001$, we reject the null hypothesis. Thus, we conclude that diamonds with less than average length ($x < 5.731mm$) and diamonds with greater than or equal to average length ($x \geq 5.731mm$) do not have the same average price.

Singular Linear Regression Output

Next, let's move on to what effect the size actually has, using a linear model. We'll be using the whole diamonds set and observe what the variable x - the length of the diamond - has on the price.

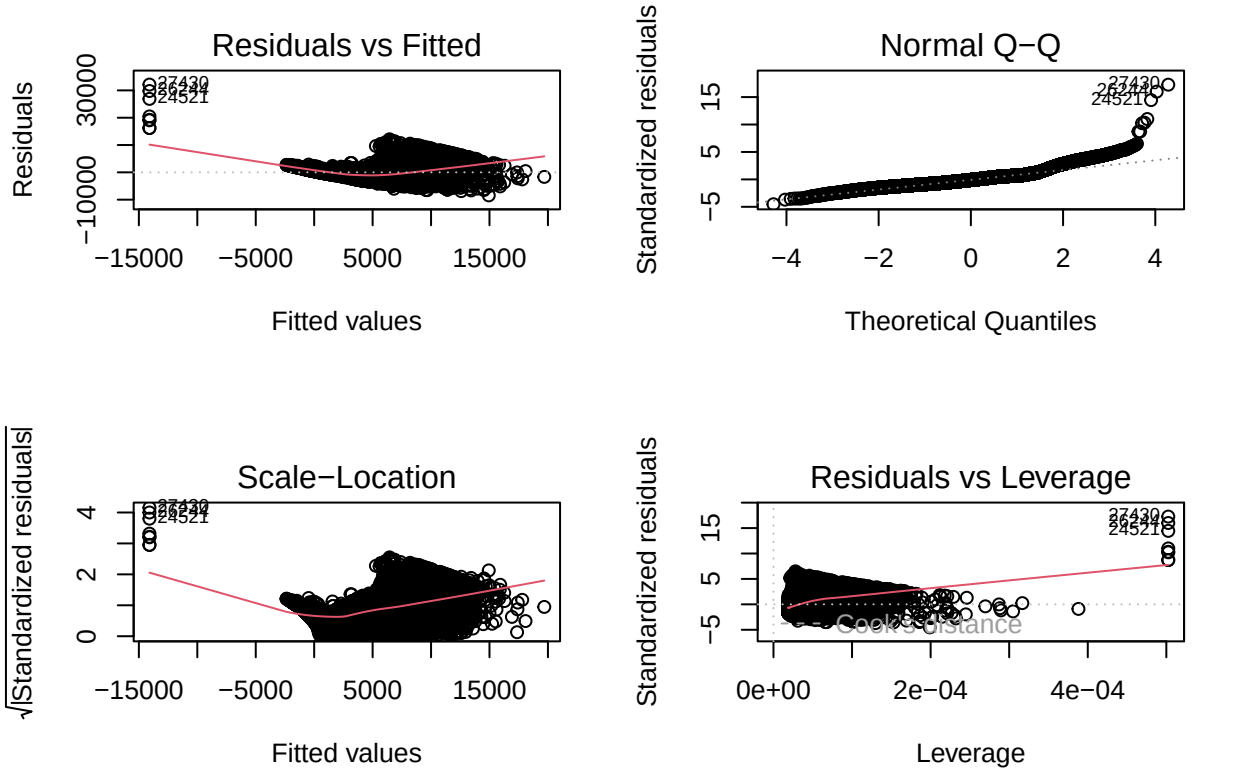
```
model1 <- lm(price ~ x , data = diamonds)
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ x, data = diamonds)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8426  -1264   -185    973  32128
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14094.056     41.732  -337.7   <2e-16 ***
## x             3145.413      7.146   440.2   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1862 on 53938 degrees of freedom
## Multiple R-squared:  0.7822, Adjusted R-squared:  0.7822
## F-statistic: 1.937e+05 on 1 and 53938 DF,  p-value: < 2.2e-16
```

We see that, based on our output, the coefficient on x, our length, is positive, which answers the question of "does length increase price?" Yes, it does. Since our coefficient is positive, we know that for each unit increase, our dependent variable will increase: contextually and in Layman's, for each mm increase of length on a diamond, the price of it goes up by about $3145. Jumping a little further into the output, we see that the p-value ($2e^{-16}$) on our coefficient and variable is very small, falling into every standard significance level. This is also the same p-value from our earlier hypothesis test. Again, simply, we can see here that the length of a diamond changes the price by about $3145 per additional mm. The keen will notice that for appropriately small diamonds, they apparently sell for a negative price - unfortunately, they do not. So, why does our regression tell us this? We need to center our diamond price variable around the mean, so that a diamond with average length has an average price. Lastly, this'll pop up later, but we have an R-Squared of .78, meaning that approximately 78% of the variability is accounted for, or rather than our model is not a perfect predictor, or even further, that we do cannot predict the price of a diamond accurately without more information of a different variable.

```
par(mfrow=c(2,2))
plot(model1)
```



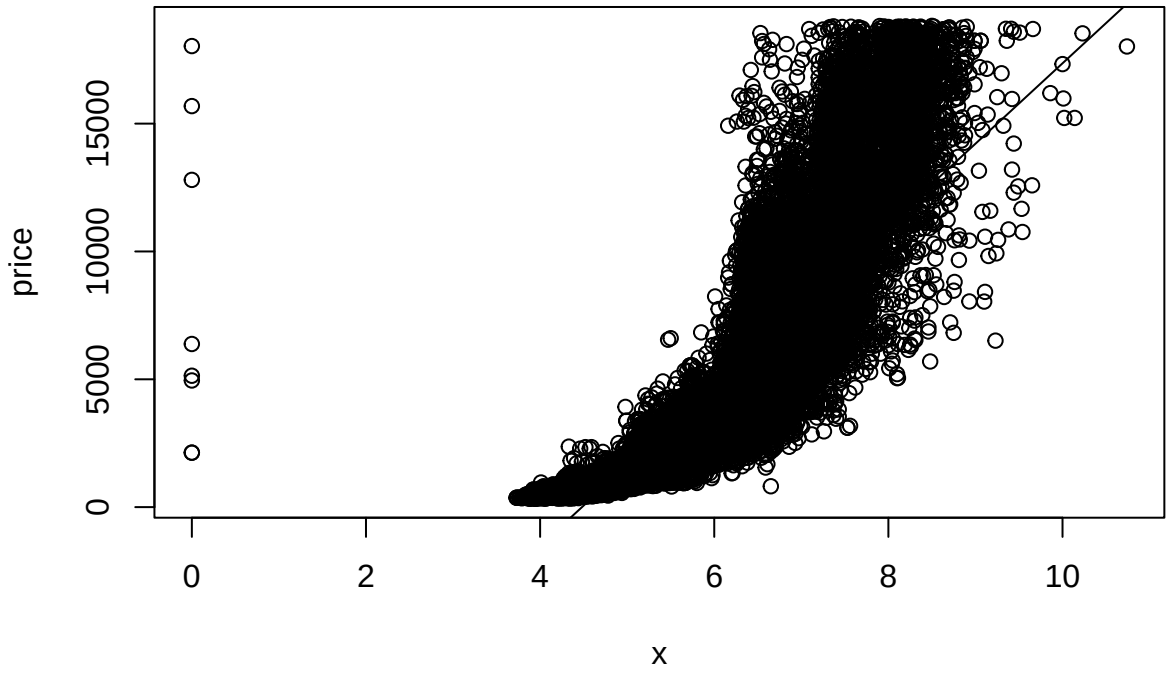
Technically, we are supposed to look look at these plots, which validate (or invalidate in our case) our assumptions. Since we already know that our model isn't well made, I won't harp on this too much. In the two leftmost graphs, the red line (and the data accompanying it) should have no correlation, or basically

have the red line be straight and the data should just be around it without any pattern. These are 'passable' results, in the way that it is good enough for me to interpret rather than fix. On the Normal Q-Q plot, we should see that all of our points follow that dotted line (that we can't see too much of), but normally we allow for a little drifting at the ends. In our Normal Q-Q, I think there's a little too much drifting, but again neither here nor there. Lastly, on the residuals vs leverage, we're actually doing pretty well. Usually, there's a dotted red line that shows if there are outliers in our set, and since we cannot see a line, we are not even close to having any outliers. Overall, if I had to rate the model off of just these plots, I'd say the model is a 6/10. Not great, but good enough to interpret.

Plot Inference

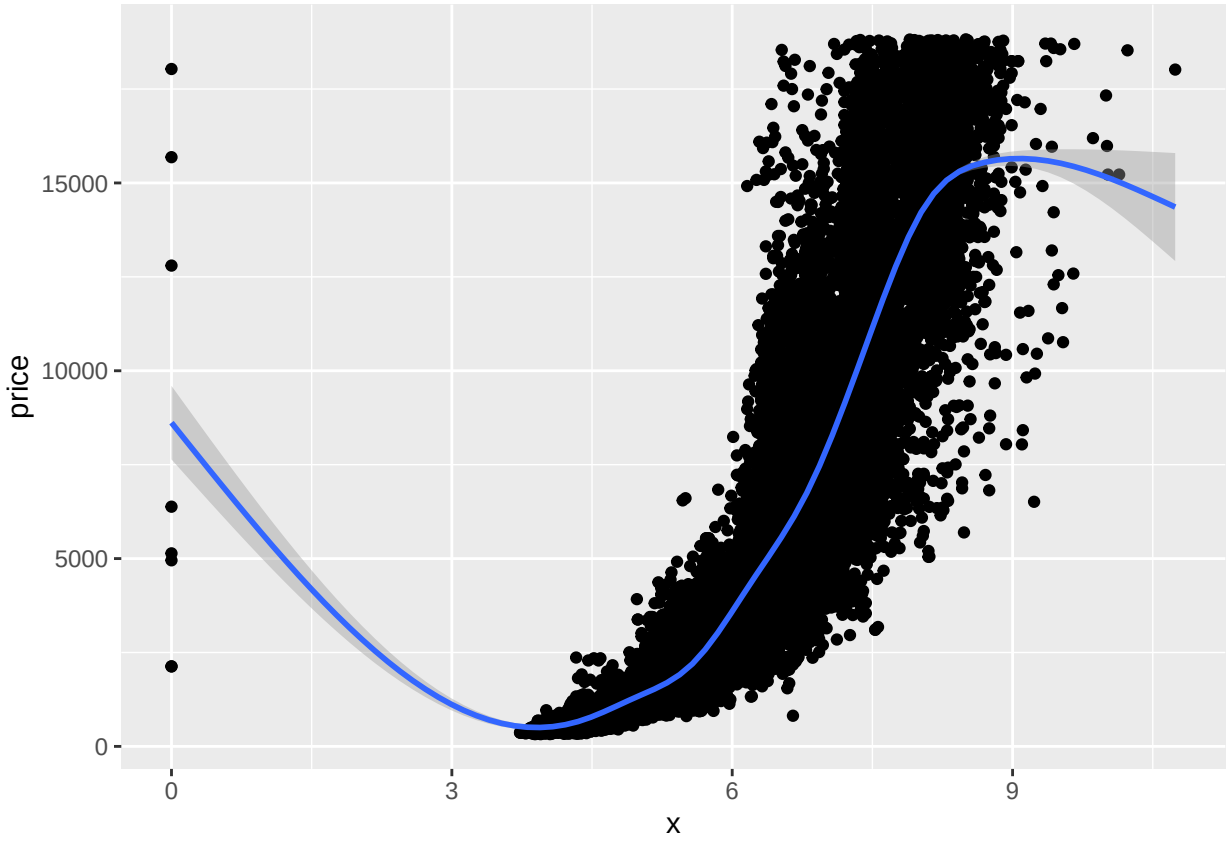
Lastly, let's look at a plot of price vs length, just to round out this work sample.

```
plot( price ~ x, data = diamonds)
abline(model1)  #can't see LOB
```




```
ggplot(diamonds, mapping = aes(x, price)) +
  geom_point()+
  stat_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



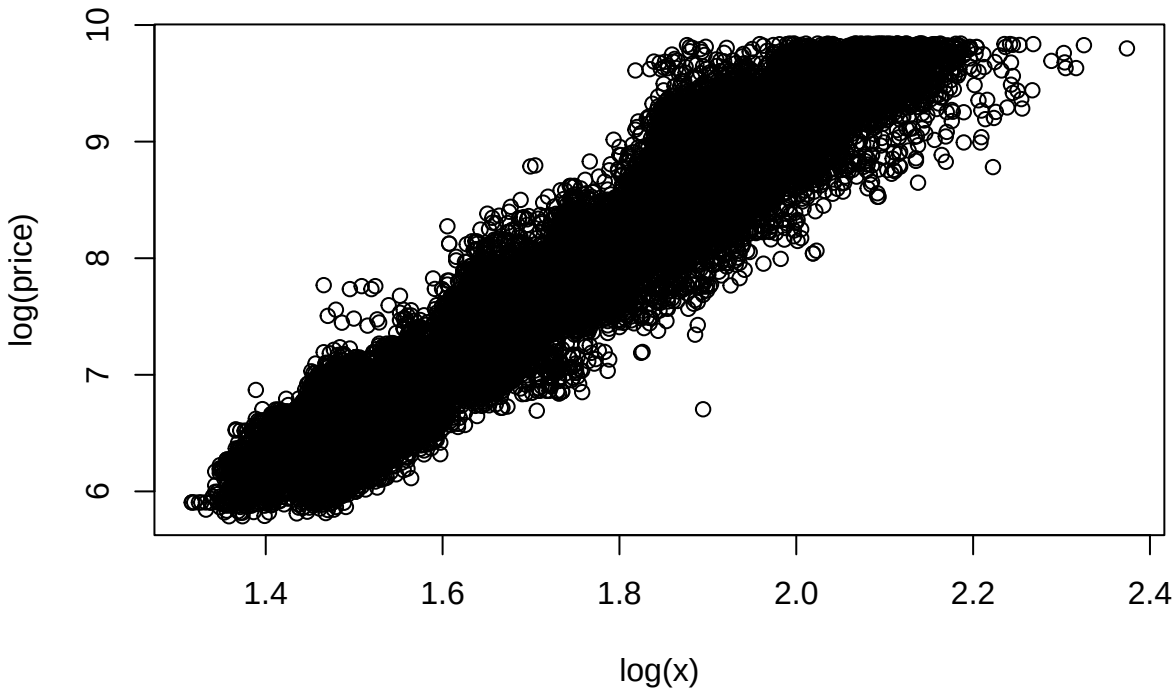
First, the first plot has a linear line of best fit, and I can't change the color of the line. I promise it's there, but again, this is inference and not an R test. We can kind of see it at the top and bottom. We'll come back to it later though, so let's turn our attention to the second plot.

Here, we can see a non-linear line of best fit in blue, and this plot basically shows us where our model would predict what the price to be at certain values for length. As we can see, the line is curvy in all the right places, so it fits kind of well. Regardless, we see that, for the most part, as length increases, so does the price. We've got 2 problem areas though, at the ends. If I had to guess, those diamonds at $x = 0$ are just a data error. If not, maybe they're super thick but have almost no length. We would explore this more if this was an R test, but alas. Next, at the $x < 9$ part of the plot, we see a little slumping. If we did a multiple linear regression, we'd see that there's a slew of different things that can affect a diamonds price, namely cut, size in other directions, clarity, depth, and whatever other metrics we can think up. However, for a concise answer, there's other variables at play here, meaning that our model doesn't account for every bit of variability. More simply, our model is not a perfect predictor, because we are lacking some information, in the form of the aforementioned metrics.

So, back to the first plot: what does it mean? Since our data has 50,000 points, it's hard to see the trend line, but the line is positively sloped, meaning that as length increases, so does price. In order to make this plot legible, we'd probably have to log the price, but then we cannot directly say that more length makes the price go up. We'd have to take a more roundabout way.

Interperting the log(price)

```
plot( log(price) ~ log(x), data = diamonds)
```




```
# model2 <- lm(log(price) ~ log(x), data = diamonds), NA values in x
```

Above, we see the scatter plot of both variables logged, which would be super easy to make a linear model for. However, we've got some bad values in our diamonds set for x, so we cannot make a model without removing these instances, which is enough to deter me. Regardless, let's interpret the plot. The equation for the LOB is $log(price) = k + \beta log(x)$, where k is some intercept and β is the coefficient on our variable, x. This does not make any sense to those without upper level math skills. So, let's get back to regular price and x: taking both sides as the exponent to e, we have some happy cancellations. $price = e^k + x^\beta$. If we do this, we technically do not have a linear equation, but we can conclude that each mm increase in length is exponential, meaning that the price increases exponentially for each mm longer a diamond is. This is actually what we had seen in our first and second plots, with that super curvy line of best fit and the mass of points going up quickly! So, to interpret it in no unclear terms, as length increases, the price of a diamond increases exponentially. We can't find exactly what kind of exponential though, due to a data issue.