Deeper Analysis on an in-class (LING 5011) replication of Allen, Miller & DeSteno (2002)

Jacob Johnson

Background

An important question in Lab Phonology regards talker-specific differences. We know from daily experience that different people have voices that sound different, but how is it that we are still able to understand each other in spite of these differences? An early theory distinguished between two types of properties: indexical properties (which stem from accent, physiological differences, etc., and carry only personal or social information about the speaker) and phonetically-relevant acoustic properties (which carry linguistic information, and make the difference between similar-sounding words). This theory supposed that indexical properties accounted for all variation between different voices, but that phonetically-relevant acoustic properties were constant across all proficient speakers of a particular language, in order to ensure that speakers understand each other correctly.

Voice-Onset-Time (VOT) is one such phonetically-relevant acoustic property. This refers to the time (on the order of milliseconds) between the opening of the vocal tract (whether at the lips, the tongue, or the back of the mouth) to allow airflow and the beginning of vocal cord vibration. This measurement distinguishes between English consonant sounds known as “voiced” – /b/ as in “bravo”, /d/ as in “delta”, /g/ as in “golf” – and the “voiceless” consonants – /p/ as in “papa”, /t/ as in “tango”, /k/ as in “kilo”. You can feel the difference between these two sets of consonants by resting a finger on your voice box and repeating the sounds without surrounding vowels. You will feel vibration for /b/, /d/, and /g/ but not for /p/, /t/, and /k/, even though your mouth is in the same position for /b/ and /p/, for /d/ and /t/, and for /g/ and /k/. Speak these consonants again at the beginning of a full word, with surrounding vowel context, and you will feel vibrations each time, both for the voiced and voiceless consonants. In the case of the voiceless consonants these vibrations are coming from the succeeding vowel, delayed by a tiny – but measurable – interval. The length of that interval is Voice-Onset-Time.

In 2002, J. Sean Allen, Joanne L. Miller, and David DeSteno performed an experiment to test whether VOT – a phonetically-relevant acoustic property – can vary from speaker to speaker. They recorded several participants’ speech while saying words beginning with the above voiceless consonants (/p/, /t/, /k/) and measured the VOTs for the utterances. In 2021, my Lab Phonology class partially replicated this same study with a simpler analysis. I retained the data we used, in order to do a more complete analysis on it below.

Considerations: It had been previously discovered that VOT varies with speaking rate. This makes sense: as you talk faster, all aspects of speech must happen faster, including breaks in vocal fold vibrations. I thus analyzed the relationship between VOT and Duration, to see if there is any statistically-significant and satisfactorily explanatory linear regression model. If there is, then the hypothesis that phonetically-relevant acoustic properties are constant cannot be refuted by our findings:

$$H_0 : \text{there exists a linear regression that explains observed VOTs as a function of Speaking Rate}$$

If, on the other hand, no satisfactorily explanatory relationship is found, it suggests that VOT may also vary independently from Speaking Rate across (and within) speakers!

$$H_a : \text{no such linear regression exists, and it is concluded that VOT is not constant for a given Speaking Rate}$$

(Hint: This is what the original 2002 study found!)

Note: For all individual hypothesis tests on the statistical-significance of regression coefficients, $alpha = .05$ will be used.

Overview of Data

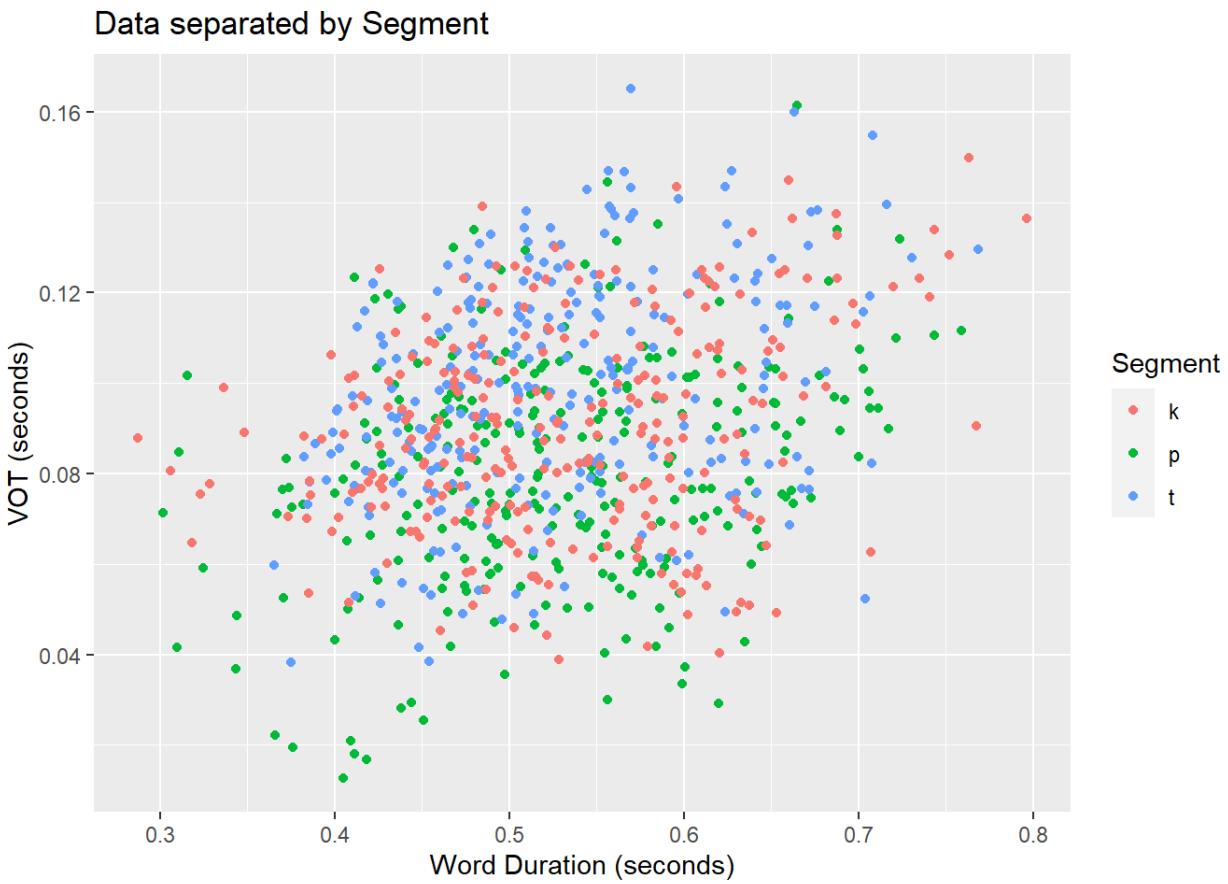
```
#read in data from file
consonants <- read.csv("consonants.csv")
#set all non-numeric values to be factors
consonants <- data.frame(VOT <- consonants$VOT, Duration <- consonants$Duration, Segment <- as.factor(c
onsonants$Segment), Word <- as.factor(consonants$Word), Speaker <- as.factor(consonants$Speaker))

#import ggplot2 for plots
library(ggplot2)
```

The Data was collected from nine students plus one teacher in my LING5011 class. The speakers were all proficient speakers of English, but not from any cohesive dialect group, and several participants were L2 (non-native) speakers of English. Each speaker recorded themselves saying 18 words five times each, and measured the Word Durations using the free PRAAT software (see https://www.fon.hum.uva.nl/praat/ (https://www.fon.hum.uva.nl/praat/)). Since Word Duration is inversely proportional to speaking rate, this measurement (or at times, its inverse) will be used as the independent variable over the course of this analysis, while VOT itself will be used as the dependent variable.

The 18 words began with either /p/, /t/, or /k/. It is possible that the individual word recorded (“pause” vs “pen” vs “pill”) may have slightly impacted VOT values, but there should be very little coarticulation (interactions between neighboring speech segments) for the chosen words, so this detail will be ignored in my analysis. Exactly one data point, likely a measurement error by one of my classmate-colleagues, was removed because it was a distant outlier with a Duration of over one second. Linguistically, this is not a realistic measurement for the Duration of a single syllable word.

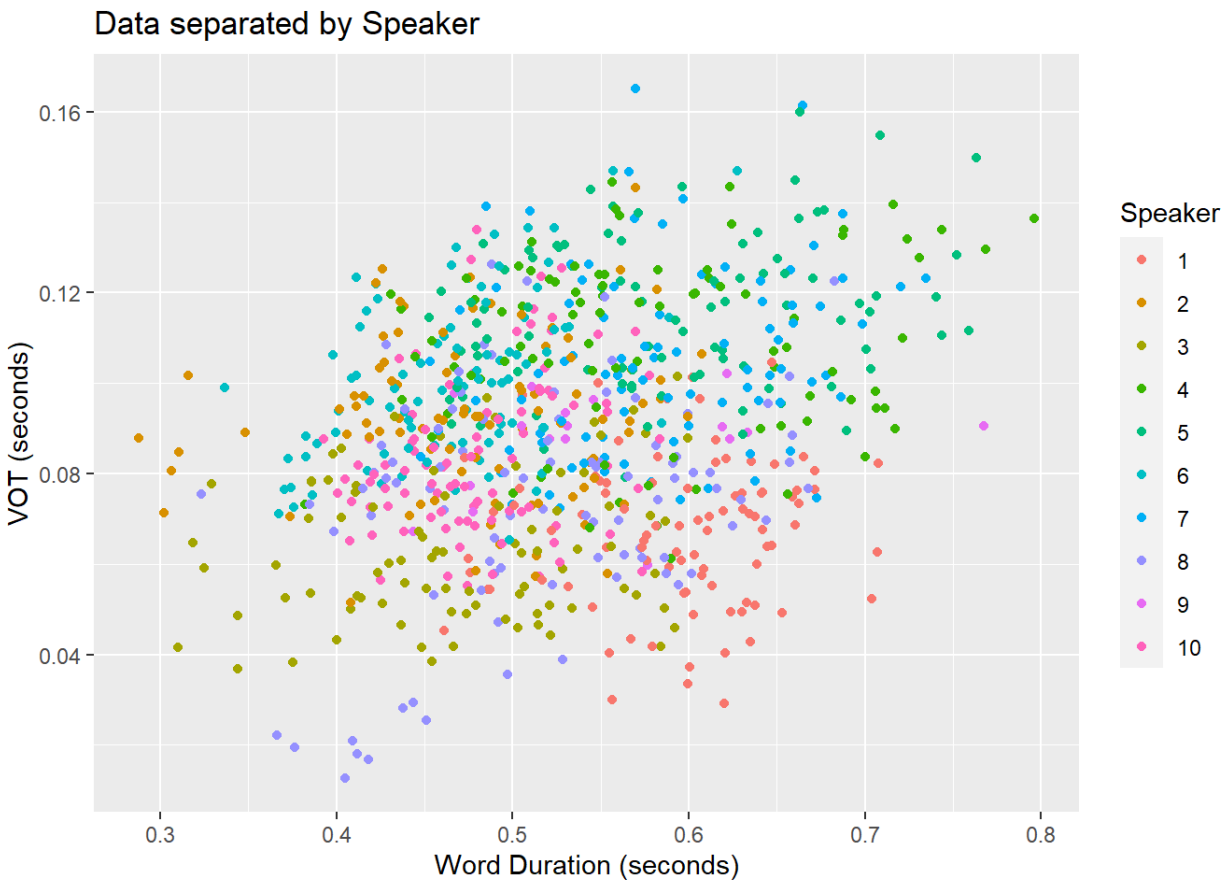
First, lets plot the data:

```
ggplot(consonants, aes(Duration, VOT, colour = Segment)) + geom_point() + ggtitle("Data separated by Se
gment") + xlab("Word Duration (seconds)") + ylab("VOT (seconds)")
```



There is a visually-obvious slight upward trend in the data as a whole. Separated by segment (/p/, /t/, or /k/), it is visible that there is a lot of overlap in these three phonemes (more on this later).

```
ggplot(consonants, aes(Duration, VOT, colour = Speaker)) + geom_point() + ggtitle("Data separated by Speaker") + xlab("Word Duration (seconds)") + ylab("VOT (seconds)")
```

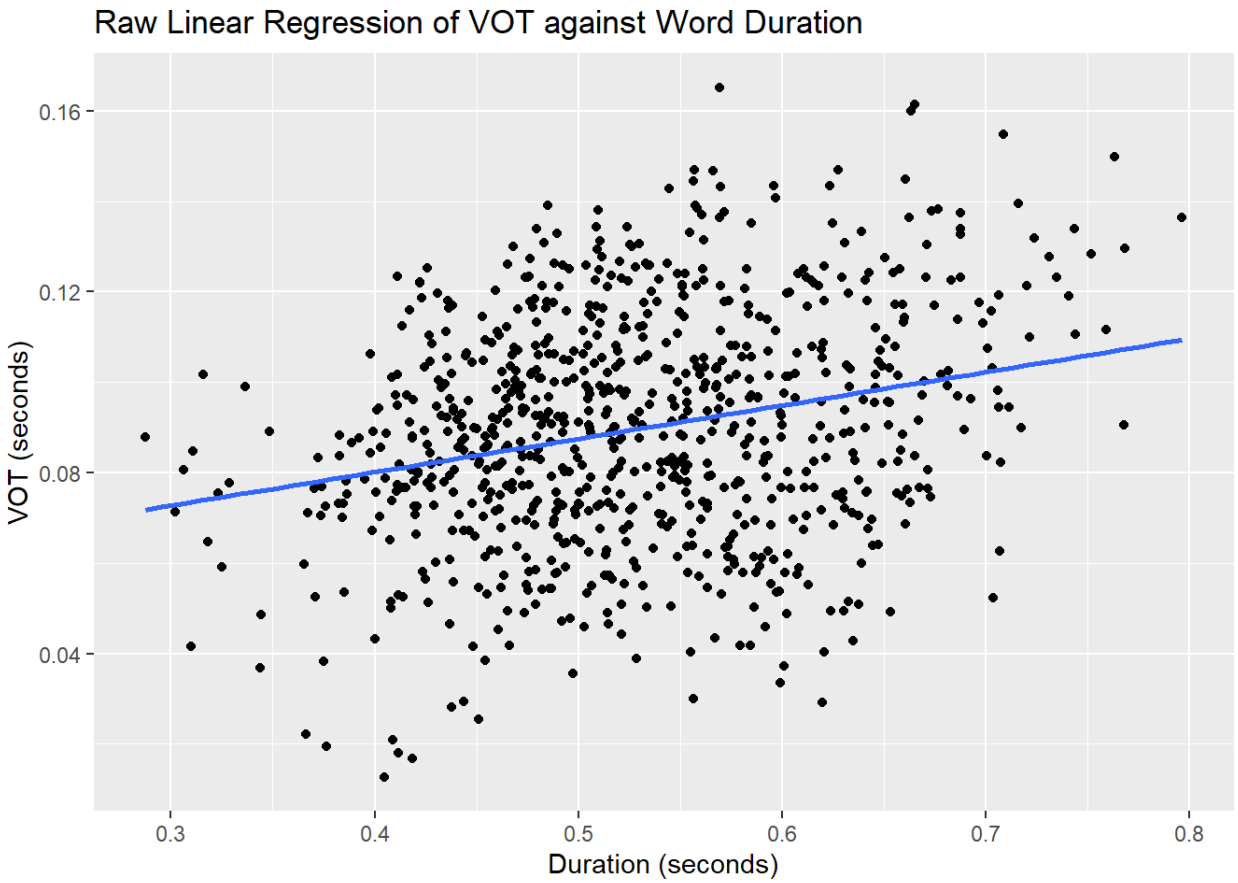


Separated by speaker, you can start to get the idea that individual speakers might actually differ in their VOTs. We expected to see across-Speaker differences in Word Duration – Speaking Rate is, after all, an indexical property – but notice how, for a particular Duration, some speakers cluster near a lower VOT, and some cluster near a higher VOT; compare, for example, speakers 5 and 10 at a duration of around .6 seconds.

Linear Model - No Transformation, No Regard for Categories

Our very first step is to simply create a linear regression model from the raw measurements themselves. No transformation of the measured values, no regard for different phoneme categories. In accordance with our null hypothesis, we expect to see that there is an increased VOT correlating to an increased Duration:

```
ggplot(consonants, aes(Duration, VOT)) + geom_point() + geom_smooth(method = "lm", se = FALSE) + ggtitle("Raw Linear Regression of VOT against Word Duration") + xlab("Duration (seconds)") + ylab("VOT (seconds)")
```




```
processData(data.frame(x <- consonants$Duration, y <- consonants$VOT))
```

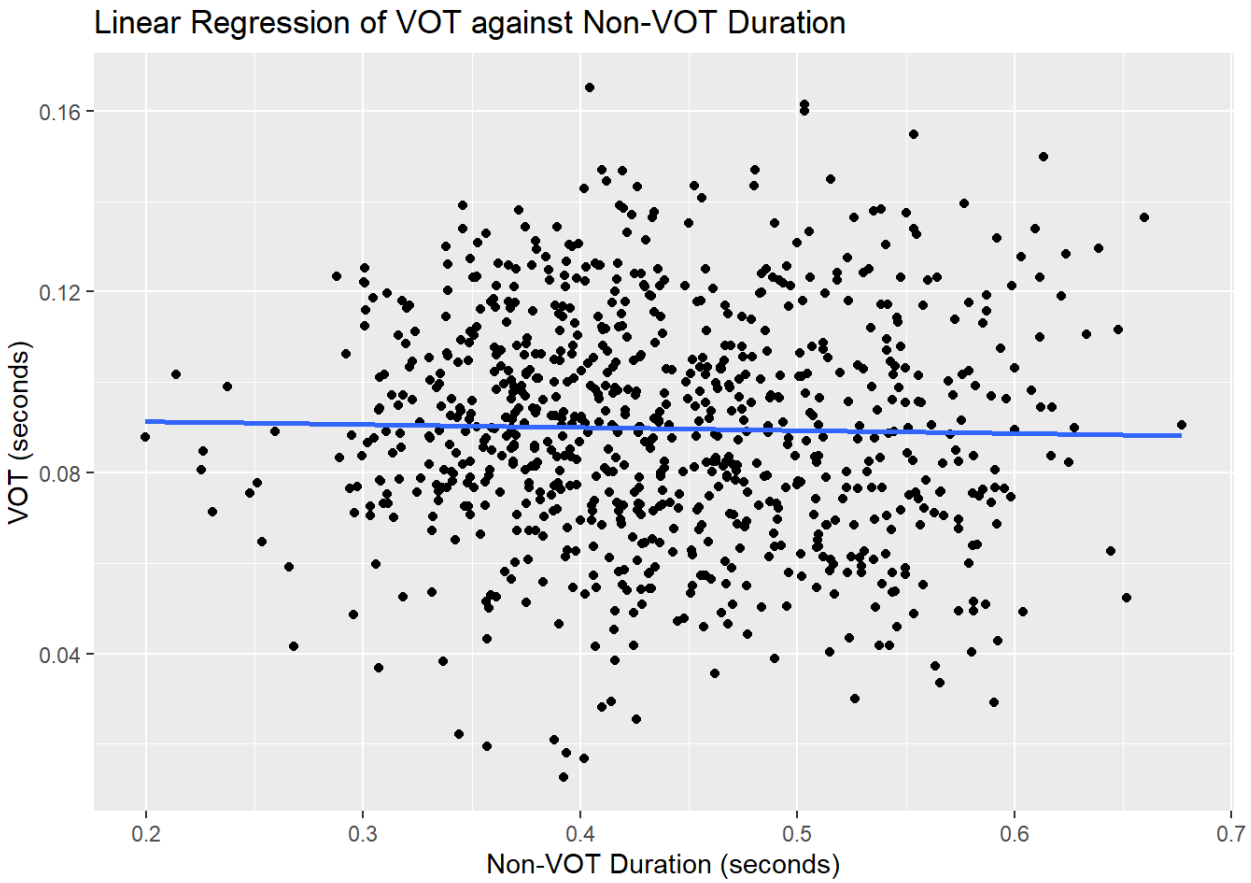
```
## [1] y = 0.0737858941346734x + 0.0506206240422417
## [1] r-squared: 0.0681276956995878
## [1] Does this linear regression model specify a statistically-significant relationship?  TRUE
## [1] (p-value: 9.07202608808816e-16)
```

We see from this linear model that there is indeed a statistically-significant correlation between increased word Duration and increased VOT! However, look at the R-squared value! That is very low (only about 6.8% of variation in VOT can be explained by variation in word duration), which means that this relationship is not enough to account for the observed data. We will look at more contributing factors soon (Speaker, in particular), but first lets try transforming the data to see if that improves the result any:

Linear Model - Duration Transformation, No Regard for Categories

The original study tried this transformation to help fit the data. The logic here is sound: if VOT were to increase, it would also increase the overall Duration. So in this transformed dataset, the y-axis still corresponds to the same VOT, but the x-axis refers to the non-VOT word duration, that is, the part of the word that was spoken after the vocal fold vibrations kicked in. Let's see:

```
ggplot(consonants, aes(Duration-VOT, VOT)) + geom_point() + geom_smooth(method = "lm", se = FALSE) + gg
title("Linear Regression of VOT against Non-VOT Duration") + xlab("Non-VOT Duration (seconds)") + ylab(
"VOT (seconds)")
```




```
processData(data.frame(x <- consonants$Duration - consonants$VOT, y <- consonants$VOT))
```

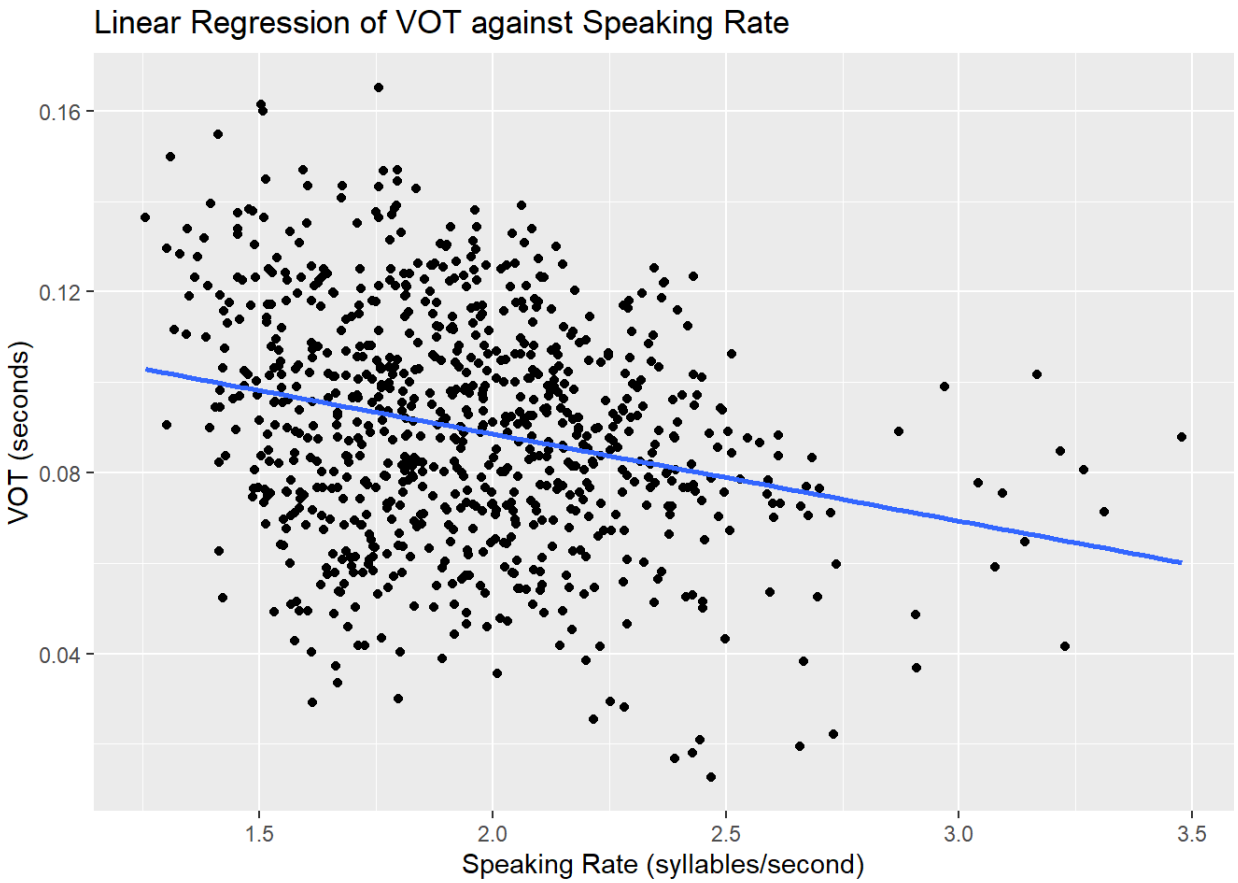
```
## [1] y = -0.00657283092849164x + 0.0926500488199514
## [1] r-squared: 0.000504031049697873
## [1] Does this linear regression model specify a statistically-significant relationship?  FALSE
## [1] (p-value: 0.250698603476465)
```

This transformation did not help our dataset. Not only did it reduce our R-squared (now only a laughable .05% of variation is explained by our model. Not 5%, .05%.), but it also made our model not even be statistically significant. Let's Try something else.

Linear Model - Reciprocal Transformation, No Regard for Categories

For this data transformation, we will take the reciprocal of the measured word Duration, to instead calculate the Speaking Rate, in syllables per second:

```
ggplot(consonants, aes(1/Duration, VOT)) + geom_point() + geom_smooth(method = "lm", se = FALSE) + ggti
tle("Linear Regression of VOT against Speaking Rate") + xlab("Speaking Rate (syllables/second)") + ylab
("VOT (seconds)")
```




```
processData(data.frame(x <- 1/(consonants$Duration), y <- consonants$VOT))
```

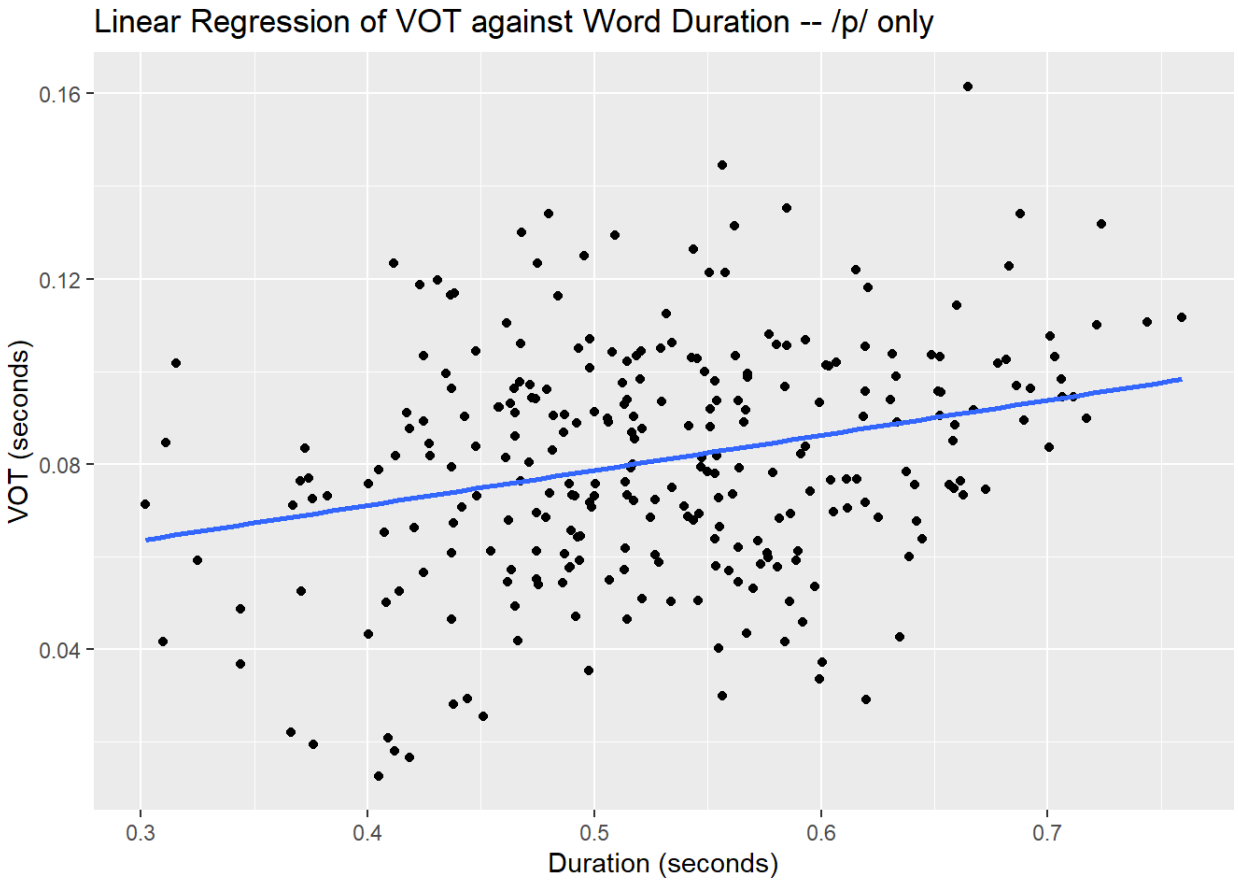
```
## [1] y = -0.0192424430793453x + 0.127055170417744
## [1] r-squared: 0.0669626879604952
## [1] Does this linear regression model specify a statistically-significant relationship?  TRUE
## [1] (p-value: 1.60211645980815e-15)
```

Here we see that this IS statistically-significant (which is better than our last data transformation attempt), but it slightly reduces our R-squared value (now only 6.7% of the variation can be explained by our modeled relationship). Looking at the plot, it is visually obvious that the data still does not very closely cluster around the best-fit line. Let's see how else we can try to improve this.

Finding a Linear Model - Categorized by Segment

We are going to separate the data by the segment that was spoken: did the word start with /p/, /t/, or /k/? Linguistically, this is known to be different across different consonants. In fact, this study chose to use words beginning with these unvoiced consonants specifically because their average VOTs were already known to be more different from each other than were those of the voiced consonants /b/, /d/, and /g/. We will also go back to using the untransformed data, because that gave us the best combination of statistical significance and high explanatory value (high R-squared). At this stage, we could alternatively construct a multiple regression with Segment as a factor, but just applying the same regression to subsets of the data will yield some nice plots of the data.

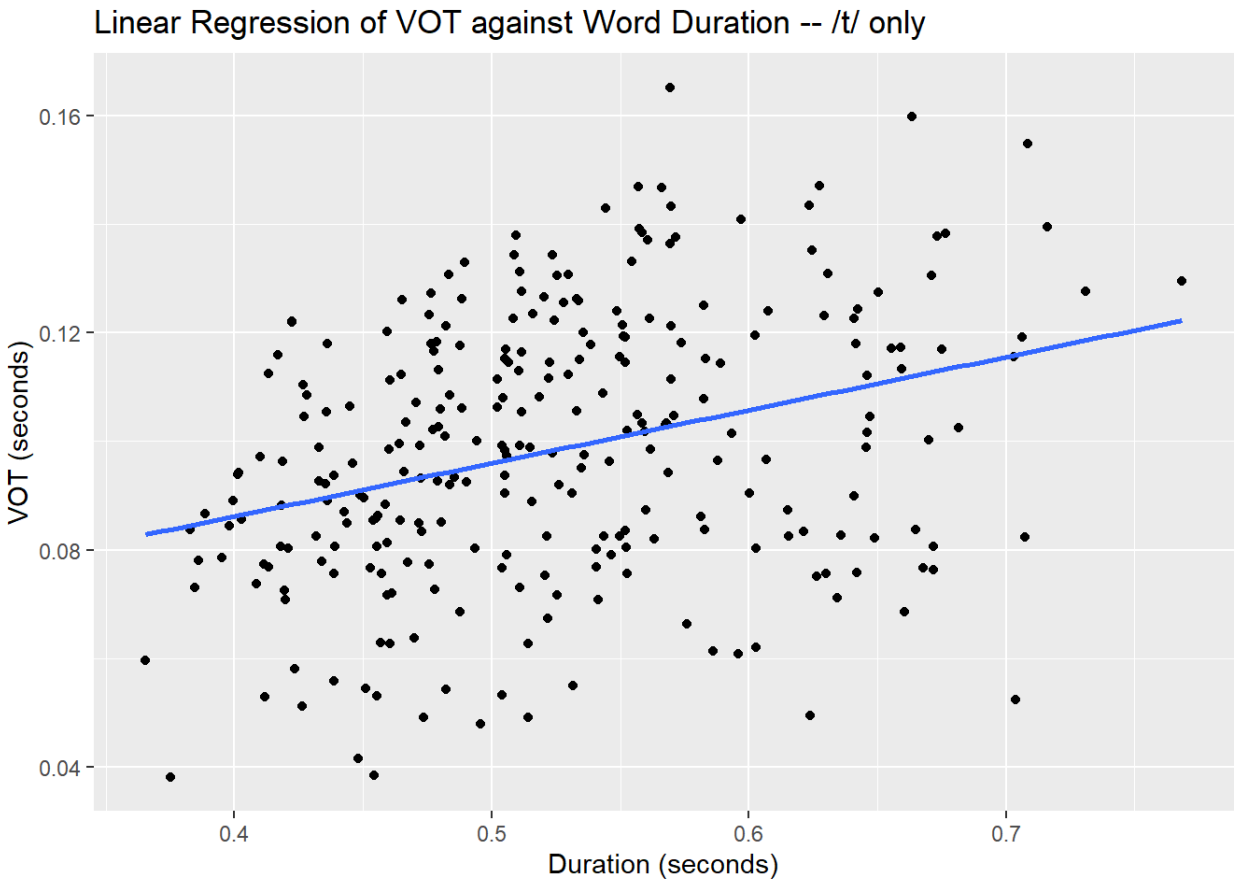
```
#subset the data and format as data.frame
pData <- subset(consonants, Segment == "p")
pData <- data.frame(VOT = pData$VOT, Duration = pData$Duration, Speaker = pData$Speaker, Word = pData$W
ord)
#plot and processData
ggplot(pData, aes(Duration, VOT)) + geom_point() + geom_smooth(method = "lm", se = FALSE) + ggtitle("Li
near Regression of VOT against Word Duration -- /p/ only") + xlab("Duration (seconds)") + ylab("VOT (se
conds)")
```




```
processData(data.frame(x <- pData$Duration, y <- pData$VOT))
```

```
## [1] y = 0.0756258385629936x + 0.0408611887871109
## [1] r-squared: 0.0748411717841837
## [1] Does this linear regression model specify a statistically-significant relationship?  TRUE
## [1] (p-value: 7.83089249319488e-07)
```

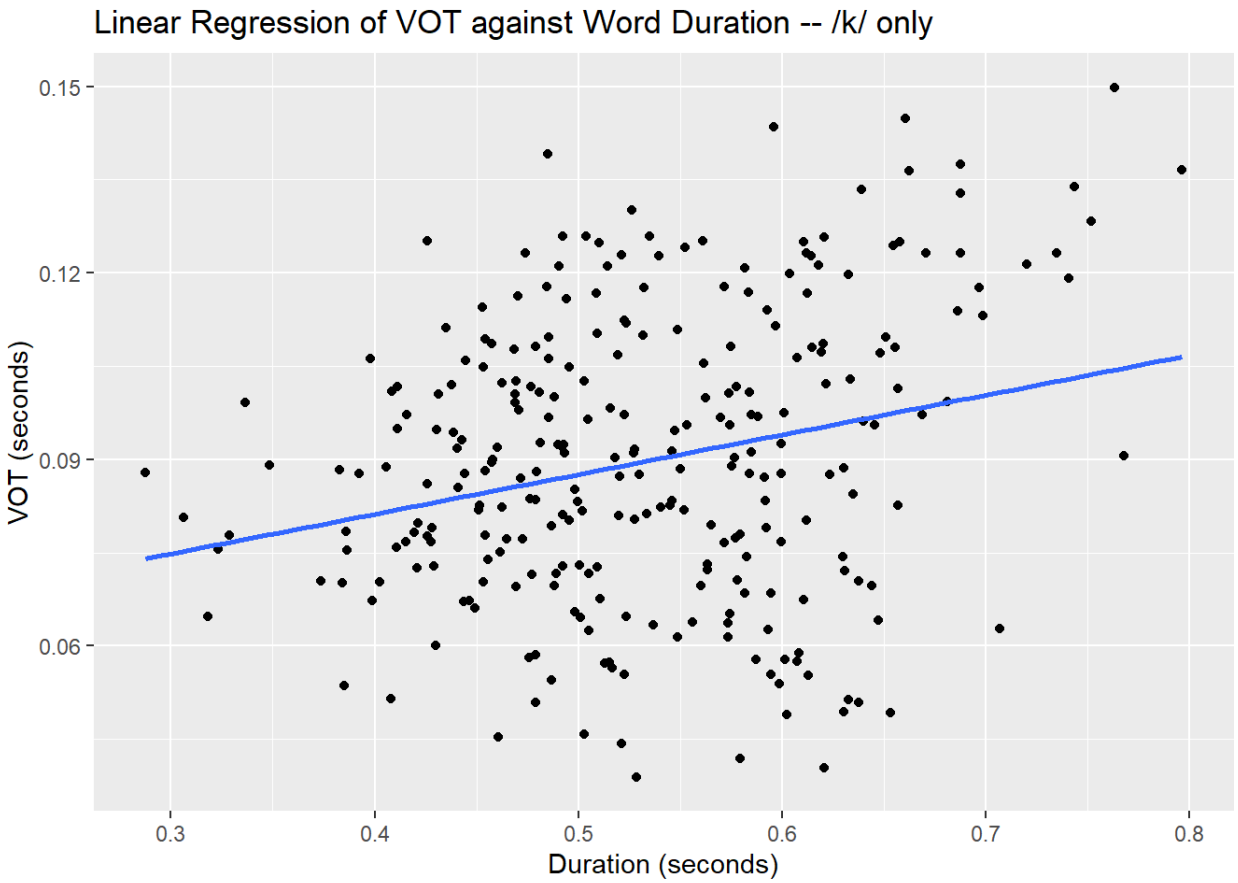
```
#subset the data and format as data.frame
tData <- subset(consonants, Segment == "t")
tData <- data.frame(VOT = tData$VOT, Duration = tData$Duration, Speaker = tData$Speaker, Word = tData$W
ord)
#plot and processData
ggplot(tData, aes(Duration, VOT)) + geom_point() + geom_smooth(method = "lm", se = FALSE) + ggtitle("Li
near Regression of VOT against Word Duration -- /t/ only") + xlab("Duration (seconds)") + ylab("VOT (se
conds)")
```




```
processData(data.frame(x <- tData$Duration, y <- tData$VOT))
```

```
## [1] y = 0.0979296015441924x + 0.0470264404020791
## [1] r-squared: 0.103239479669138
## [1] Does this linear regression model specify a statistically-significant relationship?  TRUE
## [1] (p-value: 6.2163029395069e-09)
```

```
#subset the data and format as data.frame
kData <- subset(consonants, Segment == "k")
kData <- data.frame(VOT = kData$VOT, Duration = kData$Duration, Speaker = kData$Speaker, Word = kData$W
ord)
#plot and processData
ggplot(kData, aes(Duration, VOT)) + geom_point() + geom_smooth(method = "lm", se = FALSE) + ggtitle("Li
near Regression of VOT against Word Duration -- /k/ only") + xlab("Duration (seconds)") + ylab("VOT (se
conds)")
```




```
processData(data.frame(x <- kData$Duration, y <- kData$VOT))
```

```
## [1] y = 0.0638361560804537x + 0.0556594149205832
## [1] r-squared: 0.0704955231965478
## [1] Does this linear regression model specify a statistically-significant relationship?  TRUE
## [1] (p-value: 1.55409038894052e-06)
```

Looking at these scatter plots with their individual best-fit lines, all have a better R-squared value than they did together. This corresponds to the discovery that different consonants have statistically different VOT times. However, even taking this into account is not nearly enough to fully explain the observed variation of VOT.

Conclusion

From all these analyses, it is clear that Speaking Rate or Duration is a predictor for VOT, albeit one that explains only a small proportion of the variation in VOT, even when only one individual phoneme category is considered. As such, it is NOT adequately explanatory.

Seeing as how our original overall null hypothesis is not supported by the data, it is concluded that individual speakers do vary, not only in the indexical properties of their speech, but also in phonetically-relevant acoustic properties such as VOT! This finding adds to our understanding of the complexity of language, and several competing theories have sprung up to answer the remaining question of how we manage to understand each other!

Finale!

To finish off this analysis, here is a base r ANOVA (ANalysis Of VAriance) for the data. This compares the observed VOTs to one another, based on which Speaker the measurement came from, and answers the question of whether it looks like they are from a single population with a particular distribution, OR if it looks like they come from populations whose

distributions differ in a statistically significant way.

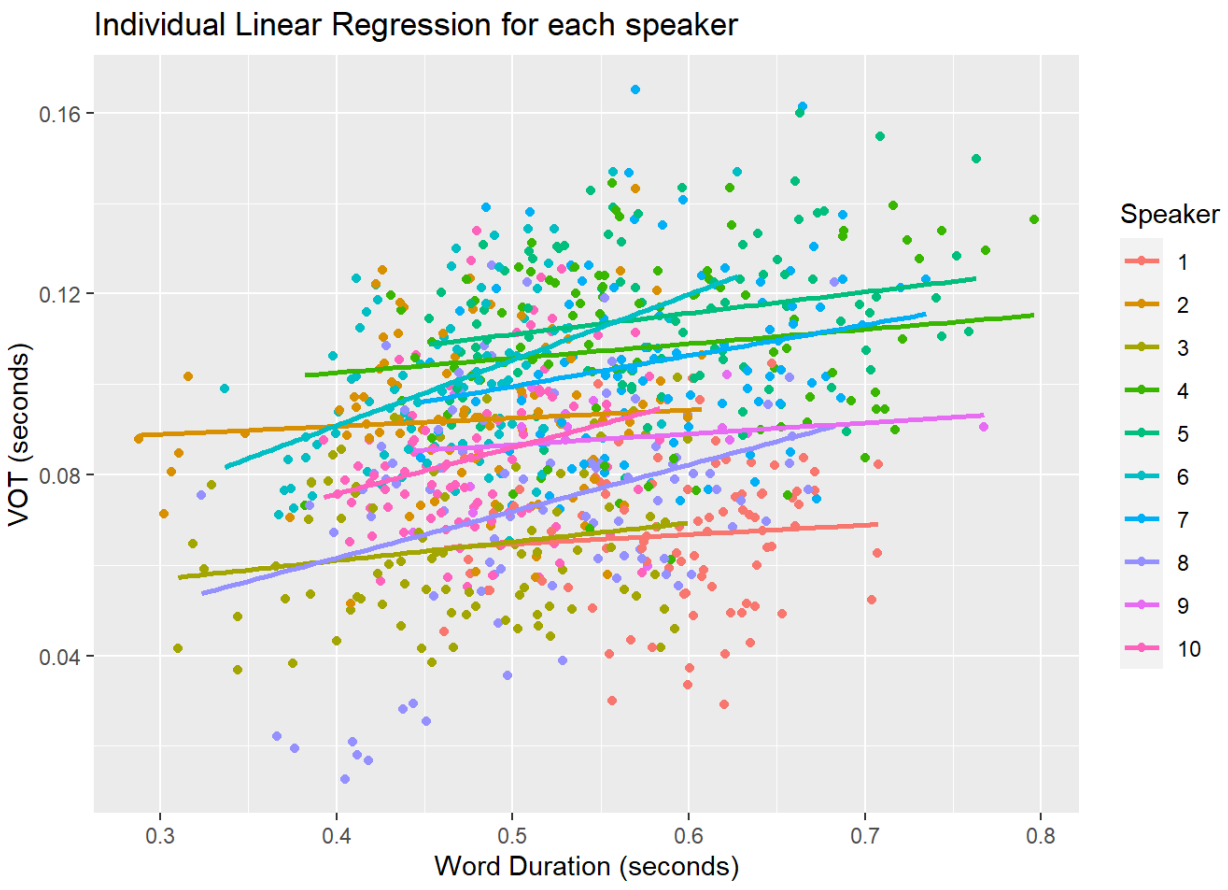
```
summary(aov(VOT~Speaker+Duration+Speaker*Duration, consonants))
```

```
##                   Df  Sum Sq  Mean Sq F value   Pr(>F)
## Speaker            9 0.25572 0.028414  91.650  < 2e-16 ***
## Duration           1 0.01265 0.012647  40.793 2.74e-10 ***
## Speaker:Duration   9 0.00572 0.000635   2.049   0.0316 *
## Residuals        879 0.27251 0.000310
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *** on the lines with Speaker or Duration mean that there is a highly statistically-significant result suggesting that different Speakers statistically have different VOTs and that a different word Duration leads to a different VOT; we already knew this from the scatter plot in the introduction to the data. The * on the line with Speaker:Duration means that there is also a statistically-significant result suggesting that change in Duration affects different speakers differently! This is the explanation we were looking for!

We can use base R to create a multiple-regression, where, depending on the speaker, different regression coefficients (which will look like Speaker-specific pairings of slopes and intercepts) result in a different best fit line for each Speaker. Let's look at this on a scatter plot:

```
ggplot(consonants, aes(x = Duration, y = VOT, colour = Speaker)) + geom_point() + geom_smooth(method =
"lm", se = FALSE) + ggtitle("Individual Linear Regression for each speaker") + xlab("Word Duration (sec
onds)") + ylab("VOT (seconds)")
```



Look at how closely the regression matches, especially Speakers 3 and 8!

Let's see what the adjusted R-squared value is for the above model:

```
summary(lm(VOT~Duration*Speaker, consonants))
```

```
##
## Call:
## lm(formula = VOT ~ Duration * Speaker, data = consonants)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.049424 -0.012516  0.000081  0.011941  0.060933
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.054229   0.019974   2.715  0.00676 **
## Duration            0.020955   0.033447   0.627  0.53115
## Speaker2            0.029299   0.023829   1.230  0.21919
## Speaker3           -0.009806   0.023392  -0.419  0.67517
## Speaker4            0.035580   0.023423   1.519  0.12911
## Speaker5            0.033499   0.024133   1.388  0.16546
## Speaker6           -0.021305   0.025875  -0.823  0.41052
## Speaker7            0.011391   0.026014   0.438  0.66159
## Speaker8           -0.033821   0.023457  -1.442  0.14970
## Speaker9            0.020341   0.023903   0.851  0.39502
## Speaker10          -0.019557   0.027984  -0.699  0.48481
## Duration:Speaker2  -0.002898   0.042992  -0.067  0.94628
## Duration:Speaker3   0.020576   0.042093   0.489  0.62509
## Duration:Speaker4   0.011005   0.039416   0.279  0.78016
## Duration:Speaker5   0.025760   0.040464   0.637  0.52453
## Duration:Speaker6   0.123975   0.048584   2.552  0.01089 *
## Duration:Speaker7   0.046960   0.043821   1.072  0.28418
## Duration:Speaker8   0.082092   0.040927   2.006  0.04518 *
## Duration:Speaker9   0.003230   0.041013   0.079  0.93725
## Duration:Speaker10  0.081778   0.052599   1.555  0.12037
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01761 on 879 degrees of freedom
## Multiple R-squared:  0.5014, Adjusted R-squared:  0.4907
## F-statistic: 46.53 on 19 and 879 DF,  p-value: < 2.2e-16
```

Impressive! Notice the R-squared and adjusted R-squared on the second to last line. This regression explains around HALF of the variation in the observed VOT, even after the R-squared is adjusted to account for the high number of model parameters, and even though this does not take into account the different Segments! This is a huge improvement over our previous models, and is solid evidence that there are differences across Speakers.

Let's add Segment back into our model, to see what our model predicts. Instead of splitting up the data into three charts, we can ask R to just consider the Segment factor and its interactions with Duration and with Speaker. We will look at just the adjusted R-squared for this final multiple regression.

```
summary(lm(VOT~Duration*Speaker*Segment, consonants))$adj.r.squared
```

```
## [1] 0.6209066
```

This model can account for 62% of variance in VOT! The remainder of the variation is due to trace phonetic coarticulation effects, as well as within-Speaker variation, both interesting linguistic phenomena in their own right.

Appendix - Code used to process datasets

```
## function(data){
##   sumX <- sum(data$x)
##   sumY <- sum(data$y)
##   sumX2 <- sum(data$x^2)
##   sumY2 <- sum(data$y^2)
##   sumXY <- sum(data$x * data$y)
##   n <- length(data$x)
##   xBar <- sumX / n
##   yBar <- sumY / n
##   #more descriptive statistics
##   sXY <- sumXY - xBar*sumY
##   sXX <- sumX2 - xBar*sumX
##   sYY <- sumY2 - yBar*sumY
##
##   #find linear regression parameters
##   beta1Hat <- sXY / sXX
##   beta0Hat <- yBar - beta1Hat*xBar
##
##   #find r2
##   SSE <- sumY2 - beta0Hat * sumY - beta1Hat * sumXY
##   SST <- sYY
##   r2 <- 1 - SSE/SST
##
##   #find standard deviation of epsilon and beta1Hat
##   sdEpsilon <- sqrt(SSE/(n-2))
##   sdBeta1Hat <- sdEpsilon/sqrt(sXX)
##
##   #test whether the linear regression model specifies a statistically-significant relationship
##   #H_0 : beta1 = 0; H_a : beta1 != 0
##   a <- .05 #alpha = .05
##   t <- (beta1Hat - 0) / sdBeta1Hat
##   pValue <- pt(abs(t), df = n-2, lower.tail = FALSE)
##   tCrit <- qt(a/2, df = n-2, lower.tail = FALSE)
##
##   #print outputs
##   print(paste("y = ", beta1Hat, "x + ", beta0Hat, sep = ""), quote = FALSE)
##   print(paste("r-squared:", r2), quote = FALSE)
##   print(paste("Does this linear regression model specify a statistically-significant relationship?
", abs(t) > tCrit), quote = FALSE)
##   print(paste("(p-value: ", pValue, ")", sep = ""), quote = FALSE)
## }
## <bytecode: 0x000000002159d500>
```