

# Assignment #1

*Jeremy Yeaton*

*November 25, 2017*

## Question 1:

- (a) *Give a numerical summary of FEV1 (mean, standard deviation and range) for each smoking category (recoded as a categorical variable with appropriate levels), and for all subjects (grand mean and overall standard deviation). Results should be printed in one or two Tables.*

The table below shows the descriptive statistics for each of the groups:

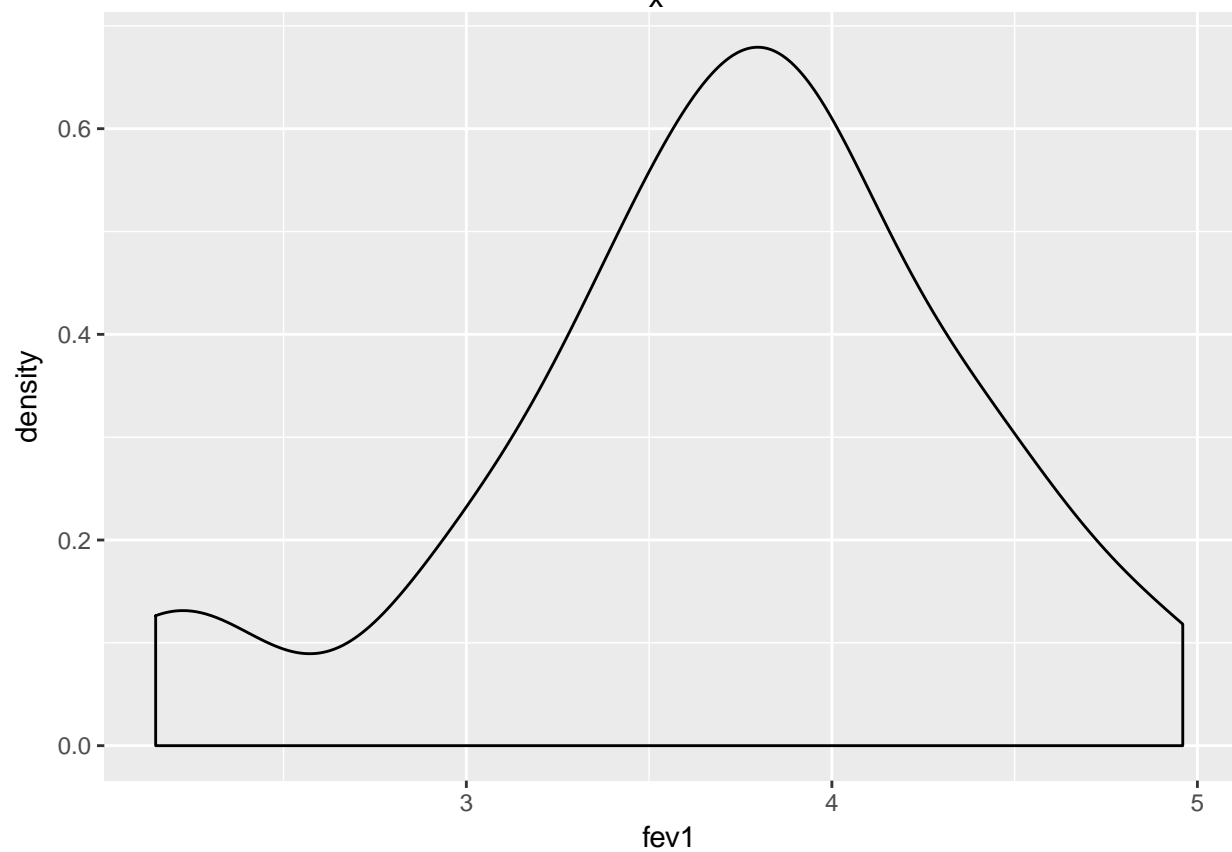
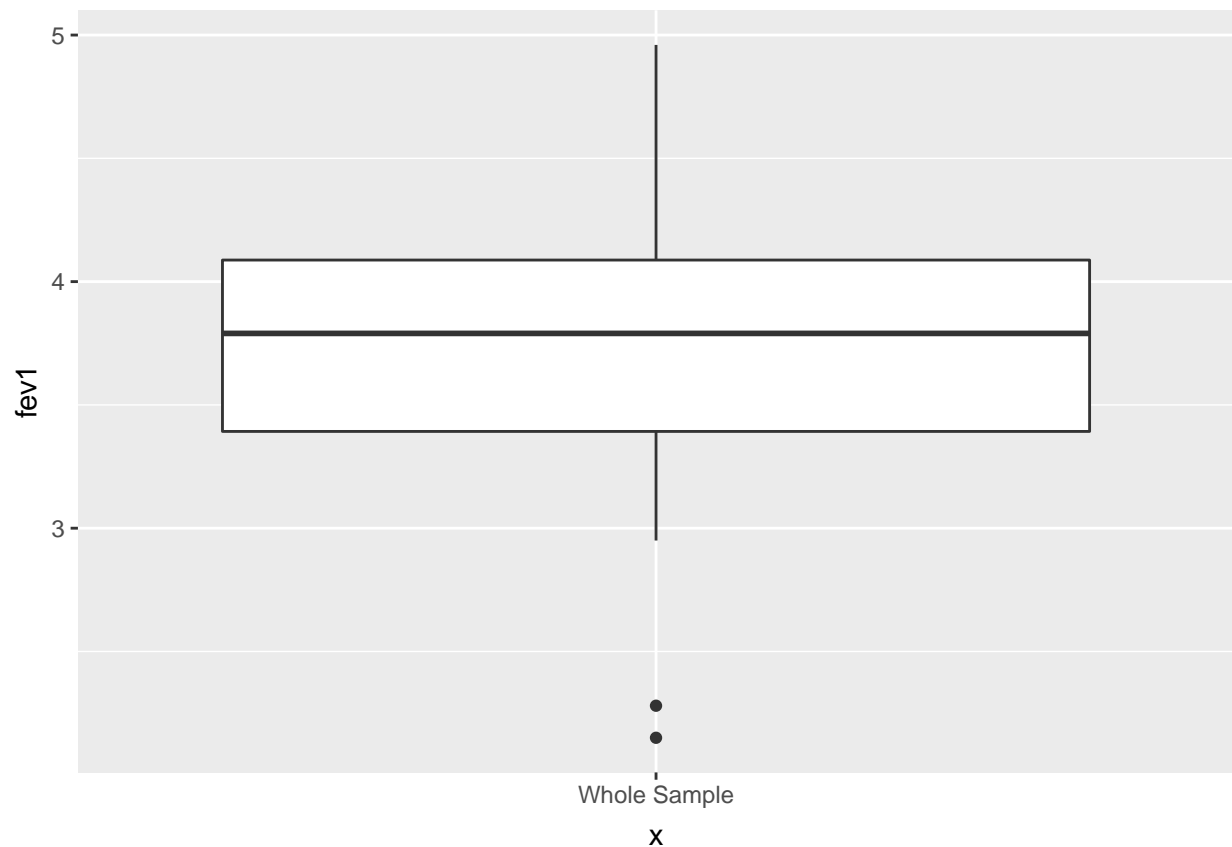
cat.f	mean	sd	range
current	3.220000	0.6758106	1.82
early	3.938333	0.2545912	0.69
non-smoker	4.220000	0.5726081	1.46
recent	3.460000	0.7128534	2.12

And descriptive statistics of all of the data together:

grand mean	overall sd	total range
3.709583	0.6749202	2.81

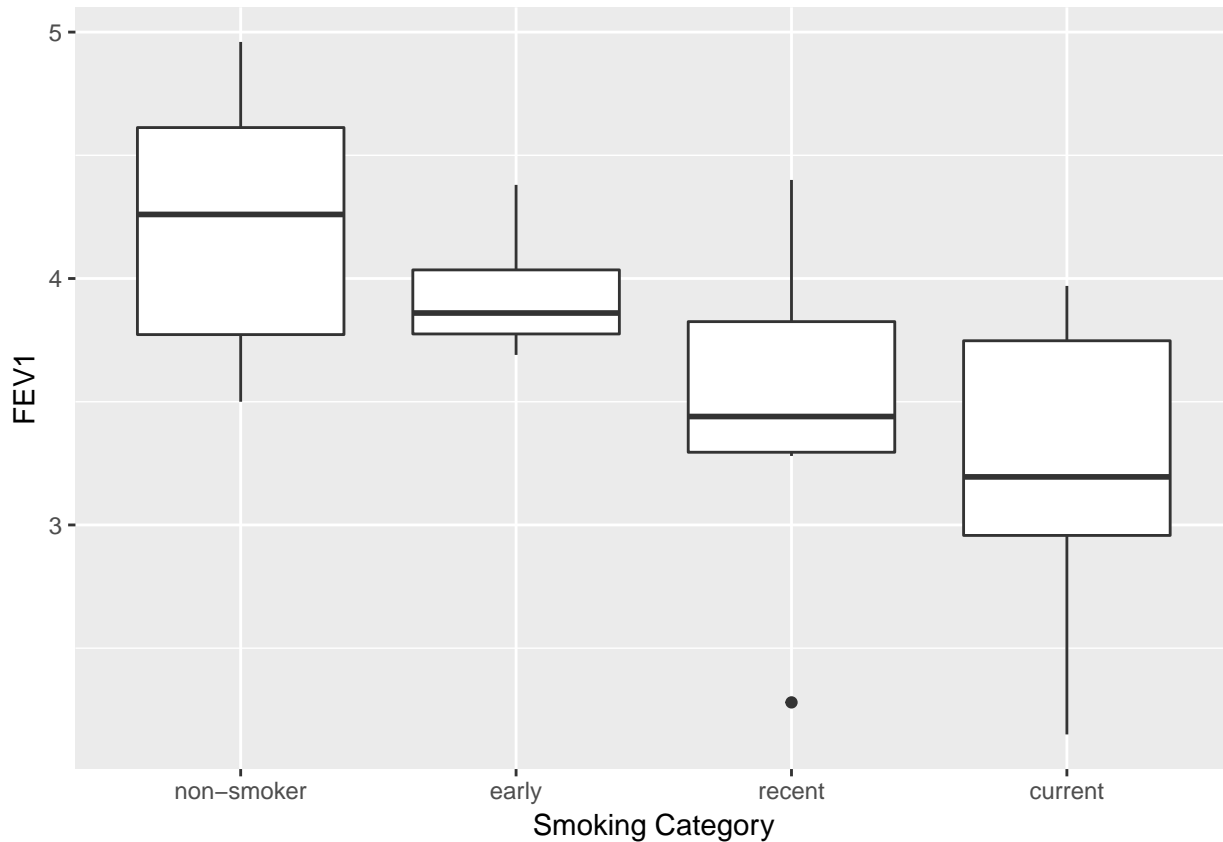
- (b) *Use box-and-whiskers charts or density plots to show the distribution of individual values.*

Below is a box-and-whisker plot of all of the points in the sample. A density plot of the same data is also included below it.



The following box-and-whisker chart displays the distribution of the individual values by group, and provides

a more useful insight into the data than the one above. A density plot of all the data together is also included below.



## Question 2:

Carry out a one-way ANOVA to test the null hypothesis that FEV1 does not depend on smoking category.

```
## $ANOVA
##   Effect DFn DFd      F      p p<.05      ges
## 1  cat.f   3  20 3.623135 0.03085325 * 0.3521093
##
## $`Levene's Test for Homogeneity of Variance`
##   DFn DFd      SSn      SSd      F      p p<.05
## 1   3   20 0.4522792 2.229833 1.352206 0.2859163
```

(a) *Formulate your conclusion in plain English, and*

Based on our ANOVA, it seems to be the case that FEV1 *does* depend on smoking category, with a very low p-value and a high F-value, and as such, we can reject the null-hypothesis.

(b) *report the percentage of explained variance.*

The percentage of variance explained by this test is about 35.2%.

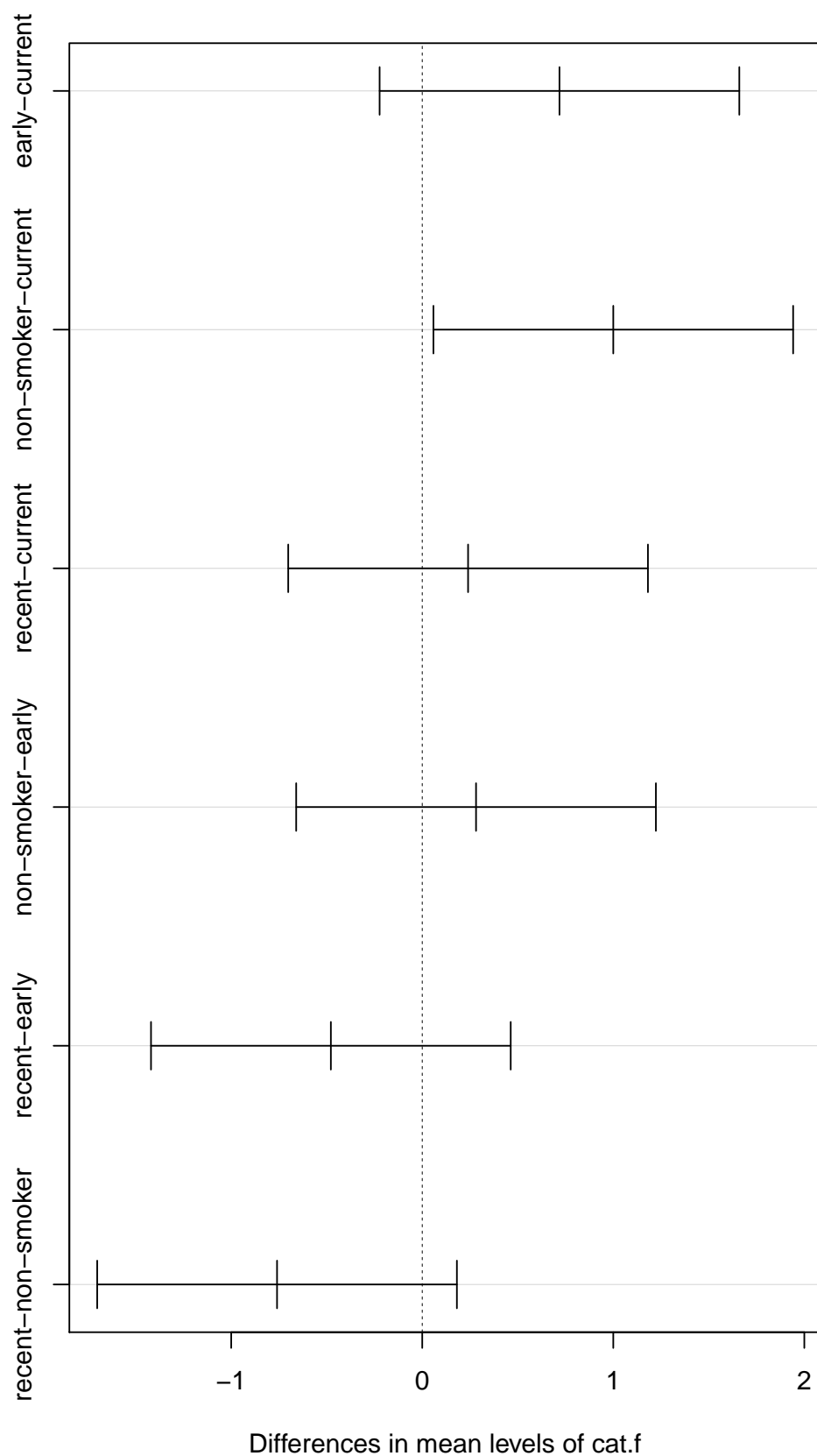
### Question 3:

- (a) Use post-hoc Tukey HSD tests (R command: `TukeyHSD`) to compare all pairs of means among the four groups of smokers. Summarize point estimates and 95% confidence intervals in a Table or graphical display, and indicate which pairs of means are found to be significantly different.

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## cat.f      3  3.689  1.2297    3.623 0.0309 *
## Residuals 20  6.788  0.3394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = fev1 ~ cat.f, data = f_clean)
##
## $cat.f
##           diff          lwr          upr      p adj
## early-current    0.7183333 -0.22308912  1.6597558 0.1760748
## non-smoker-current 1.0000000  0.05857755  1.9414224 0.0348503
## recent-current    0.2400000 -0.70142245  1.1814224 0.8905477
## non-smoker-early   0.2816667 -0.65975578  1.2230891 0.8360677
## recent-early     -0.4783333 -1.41975578  0.4630891 0.5008038
## recent-non-smoker -0.7600000 -1.70142245  0.1814224 0.1415657
```

**95% family-wise confidence level**



Based on the Tukey HSD test, it would appear that only the difference between the **current smoker** and **non-smoker** groups is significant, since that is the only pair where the adjusted p-value is less than .05.

- (b) Compare those results with results from all pairwise comparisons for mean FEV1 using the Bonferroni method (R command: *pairwise.t.test*).

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data:  f_clean$fev1 and f_clean$cat.f
##
##           current early non-smoker
## early      0.272    -      -
## non-smoker 0.045    1.000    -
## recent     1.000    1.000 0.211
##
## P value adjustment method: bonferroni
```

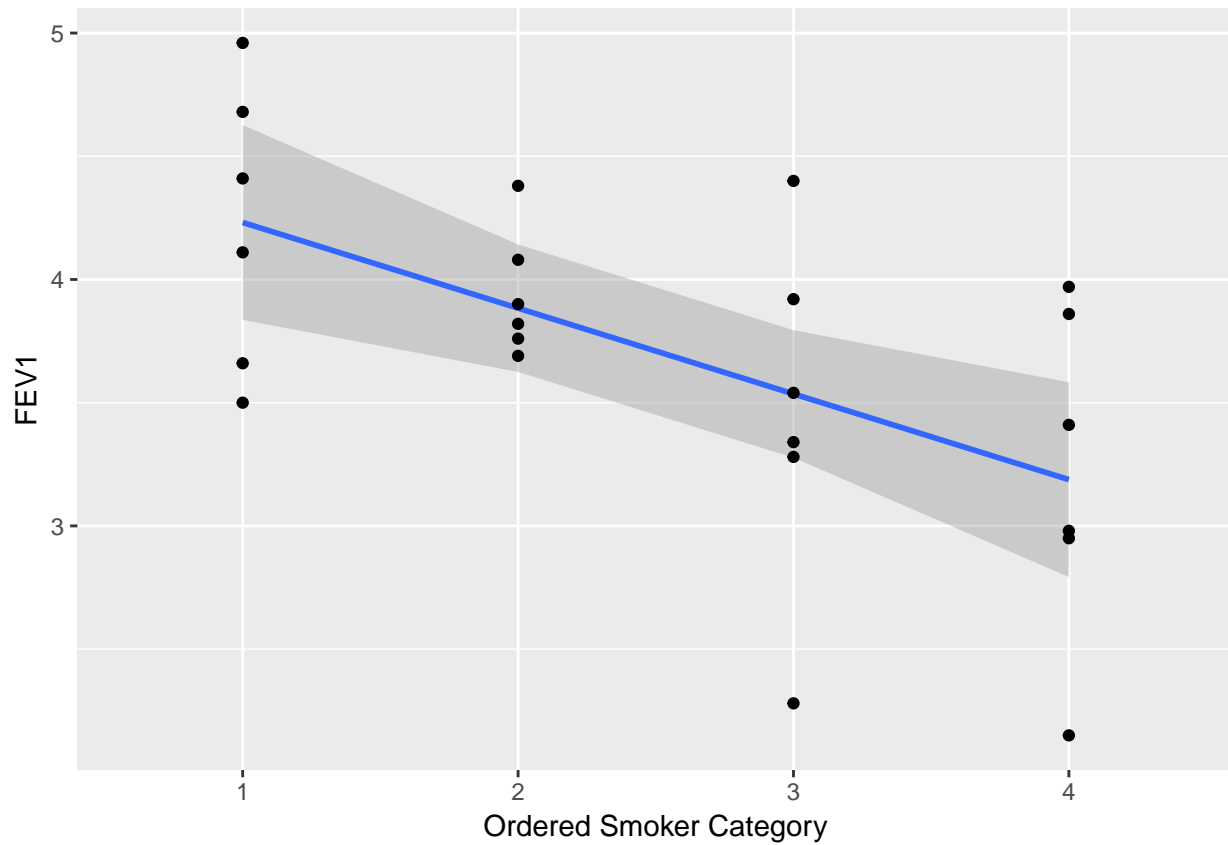
We see the same result here, with only the pair **current/non-smoker** returning a p-value less than .05, and as such, is the only pair where we can assume a significant difference.

## Question 4:

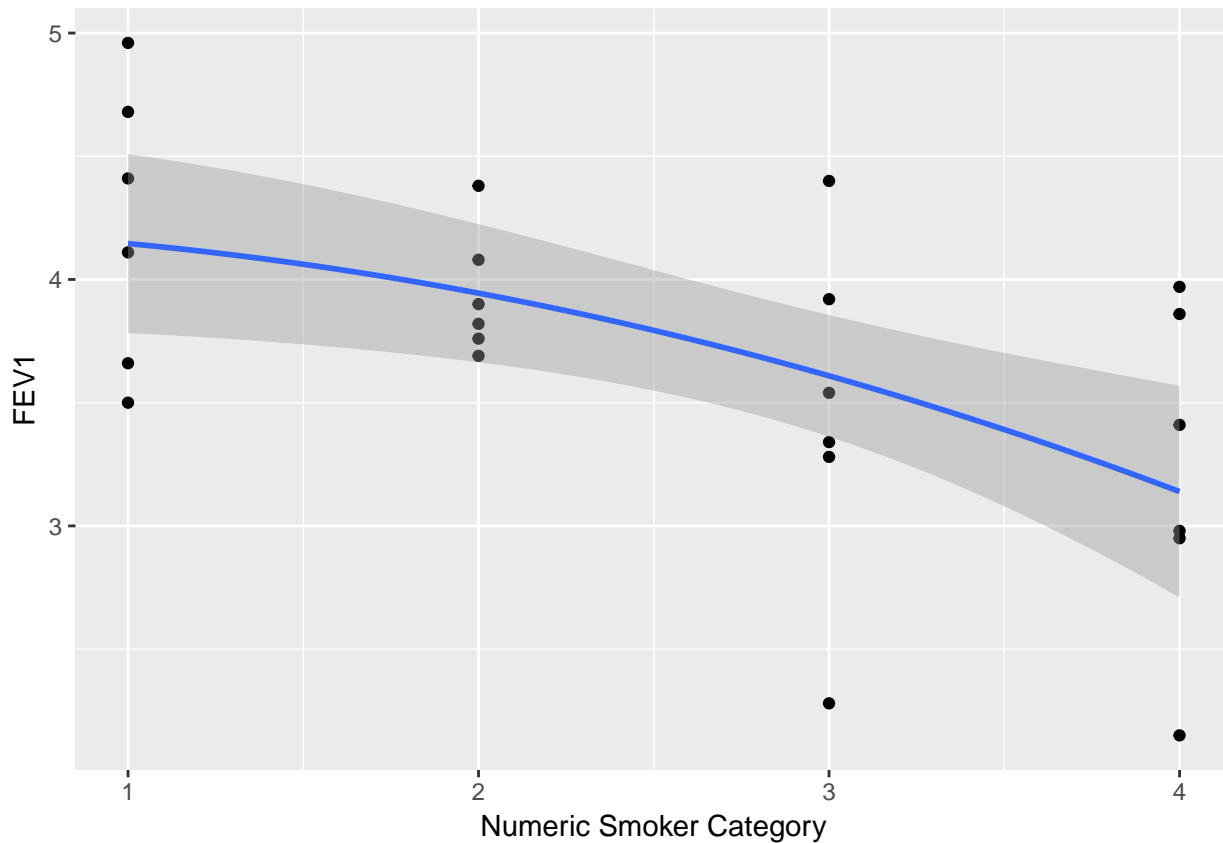
Is there any evidence for a linear or quadratic trend for mean FEV1 when considering smoking status as ordered factor levels:  $1 < 2 < 3 < 4$  (use the R command *factor* with the *ordered = TRUE* option)

```
##
## Call:
## lm(formula = fev1 ~ factor(cat.f, ordered = T), data = f_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18000 -0.24208 -0.07417  0.44625  0.94000
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.7096     0.1189  31.195 < 2e-16 ***
## factor(cat.f, ordered = T).L   0.2240     0.2378   0.942  0.35756
## factor(cat.f, ordered = T).Q  -0.7392     0.2378  -3.108  0.00554 **
## factor(cat.f, ordered = T).C  -0.1353     0.2378  -0.569  0.57582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5826 on 20 degrees of freedom
## Multiple R-squared:  0.3521, Adjusted R-squared:  0.2549
## F-statistic: 3.623 on 3 and 20 DF,  p-value: 0.03085
```

This model would account for about 25.5% of the variance. A linear model of this data would produce the following graph along its categorical x-axis:



Based on the low p-value of the quadratic model above, there seems to be a downward quadratic trend in the data. This quadratic model is presented below. It is important to note, however, that this graph is not strictly commparable to the one above, since it relies on numerical x-values instead of categorical ones.

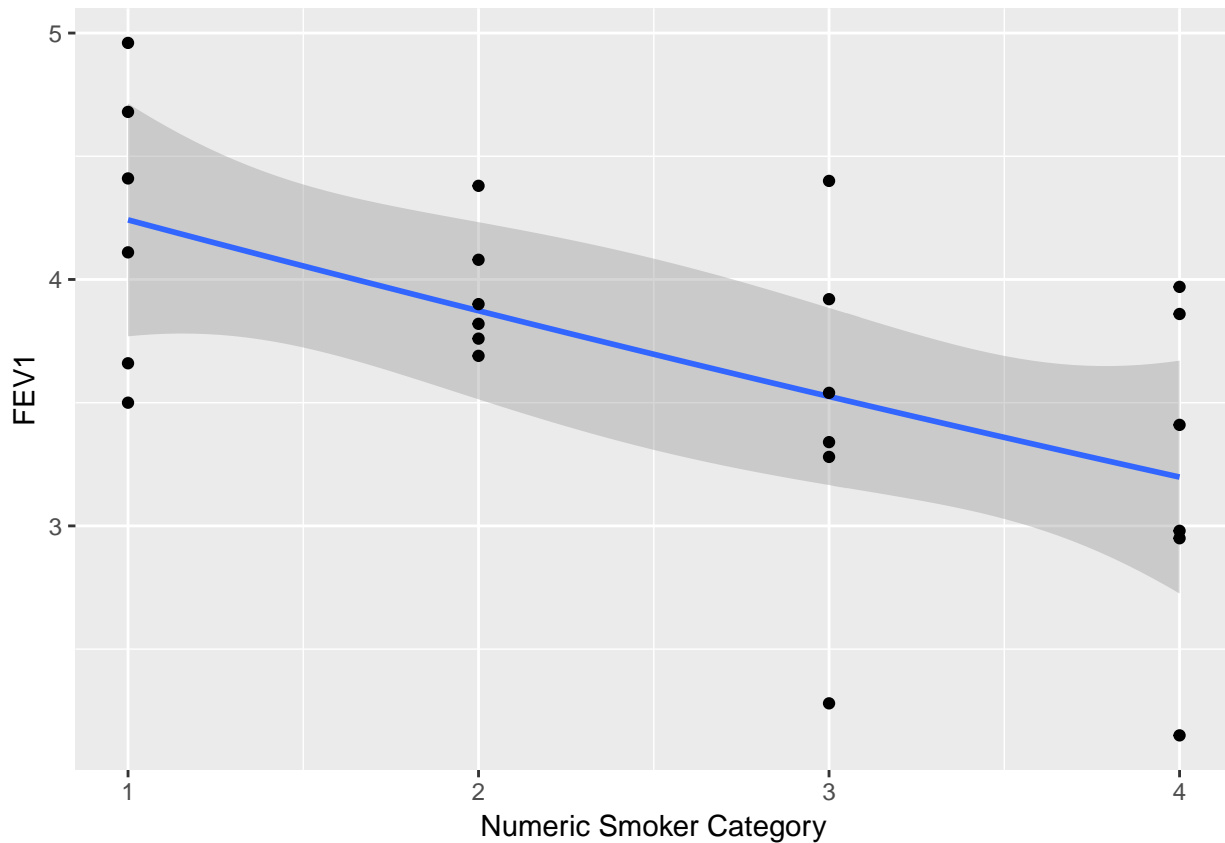


## Question 5:

- (a) Compare the preceding results with the conclusion that would be reached by using a regression approach where one considers smoking status as a numerical variable, as well as its square, i.e., using the R command `lm` with a formula like `FEV1 ~ smoking + I(smoking)^2`.

```
##
## Call:
## lm(formula = fev1 ~ cat + I(cat^2), data = lm_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24525 -0.22500 -0.01917  0.40562  0.87475
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.63125    0.64885   7.138 4.87e-07 ***
## cat          -0.39992    0.59193  -0.676   0.507
## I(cat^2)       0.01042    0.11654   0.089   0.930
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5709 on 21 degrees of freedom
## Multiple R-squared:  0.3467, Adjusted R-squared:  0.2845
## F-statistic: 5.572 on 2 and 21 DF,  p-value: 0.01145
```





(b) *What could explain the difference, if any?*

```
##
## Call:
## lm(formula = fev1 ~ I(cat^2), data = lm_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32895 -0.25803 -0.05667  0.34219  0.83065
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.21273     0.19065   22.10  < 2e-16 ***
## I(cat^2)     -0.06709     0.02027   -3.31  0.00318 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5638 on 22 degrees of freedom
## Multiple R-squared:  0.3325, Adjusted R-squared:  0.3021
## F-statistic: 10.96 on 1 and 22 DF,  p-value: 0.003183
```

## Appendix

f\_clean

```
##      cat fev1      cat.f ident
```

## 1	1	4.41	non-smoker	1
## 2	1	4.96	non-smoker	2
## 3	1	3.50	non-smoker	3
## 4	1	3.66	non-smoker	4
## 5	1	4.68	non-smoker	5
## 6	1	4.11	non-smoker	6
## 7	2	3.69	early	7
## 8	2	3.90	early	8
## 9	2	3.82	early	9
## 10	2	4.08	early	10
## 11	2	3.76	early	11
## 12	2	4.38	early	12
## 13	3	3.54	recent	13
## 14	3	4.40	recent	14
## 15	3	3.28	recent	15
## 16	3	2.28	recent	16
## 17	3	3.34	recent	17
## 18	3	3.92	recent	18
## 19	4	2.98	current	19
## 20	4	2.95	current	20
## 21	4	2.15	current	21
## 22	4	3.41	current	22
## 23	4	3.97	current	23
## 24	4	3.86	current	24