**STAT 614 - HW 7**

By Sihyuan Han

**Due**: Monday, November 23, 2020 in Blackboard by 11:59pm.

**Instructions**: Please type your solutions in a separate document and upload the document in

Blackboard as a pdf. I will not be collecting syntax for this assignment. You will need concepts from Chapters 7 & 8 on the simple linear regression model (in addition to past models!). HW 8 will address multiple regression.

Forced expiratory volume (FEV) is an index of pulmonary function that measures the volume of air expelled after 1 second of constant effort. The data set FEV.csv in Blackboard contains determinations of FEV for 654 children ages 3 through 19 who were seen in the Childhood Respiratory Disease (CRD) Study in East Boston, Massachusetts. These data are part of a longitudinal study to follow the change in pulmonary function over time in children. Variables in the data set are the participant ID number, Age (in years), FEV (in liters), Height (in inches), a binary Sex indicator (0 = female/1 = male), and Smoking status (0 = non-smoker/1 = current smoker).

1. Characterize the association between pulmonary function (FEV) and smoking status. To do this answer the following questions:
   1. Use the natural **log transformation** of FEV and examine **an independent two-sample procedure** to test for differences in the population mean LN(FEV) between the two smoking groups. Provide a brief summary of the model results. Is there evidence of an association between (transformed) pulmonary function and smoking status? If so, estimate the extent of the association (that is, give estimate and confidence interval for the parameter of interest).

**Ans**:

μ1- mean of non-smoker FEV, μ2- mean of current smoker FEV

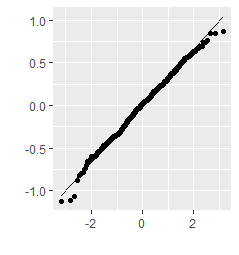
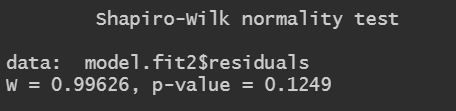
H0: μ1 = μ2

Ha: μ1 ≠ μ2

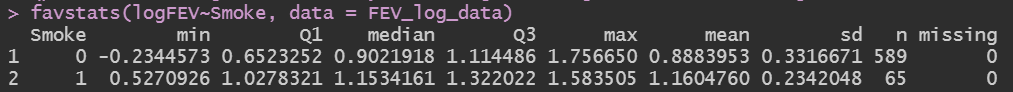
Assumptions:

1. Two samples of independent observations from two distinct population
2. Samples are independent
3. Both normally distributed with unknown mean and standard deviation
4. standard deviations are equal

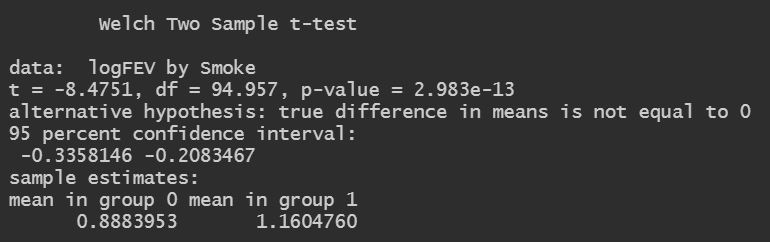
check for normality (after log transformation)

From Shapiro test the p-value is 0.1249 > 0.05, and the plot shows it is mostly linear!



standard deviations are close!



Based on the test, the p-value is 2.983e-13 which is < 0.05, so we have enough evidence to reject the null hypothesis and have 95% confidence to conclude that there is association between pulmonary function (FEV) and smoking status.

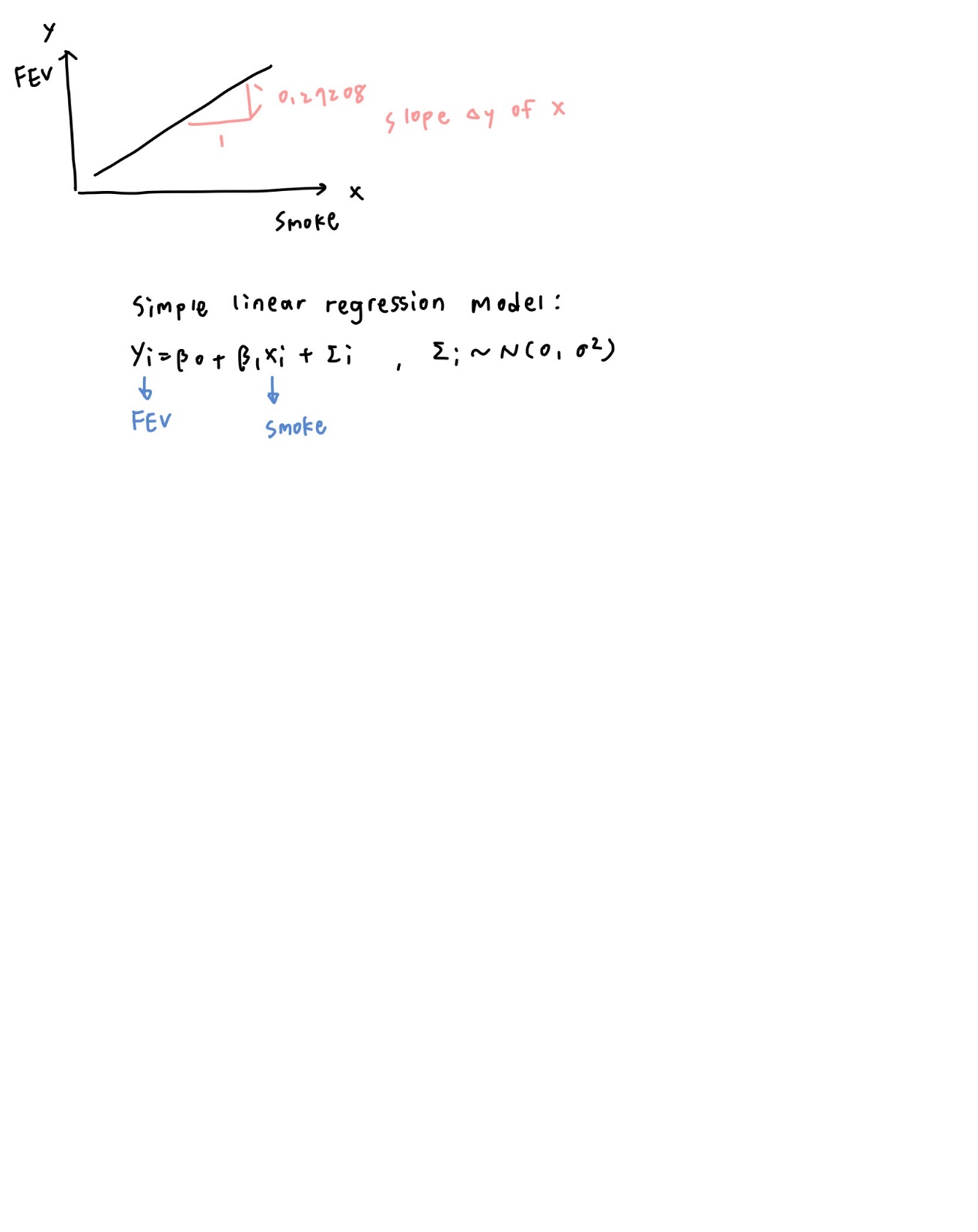
* 1. Are you surprised by the results in (a)? (Note that this is the *unadjusted* association.)

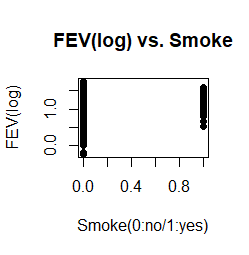
**Ans**:

No, this is a really large sample size, so the result could be significant to whether reject the H0 or not.

1. The smoking status variable is an *indicator* variable in that it takes the value 0 for non-smokers and 1 for current smokers. Non-smokers are considered the *reference* group. Even though this is a *categorical* (i.e. qualitative or grouping) variable, we can use indicator variables to designate groups in the regression procedure. (This is solely due to the 0/1 status of the variable!) Fit the simple linear regression model of FEV (use the natural log transformed FEV) with smoking status as the explanatory variable.
   1. Test the null hypothesis of no association between smoking status and FEV in the simple linear regression model.

**Ans**:

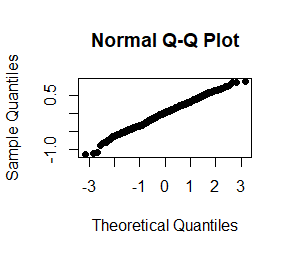
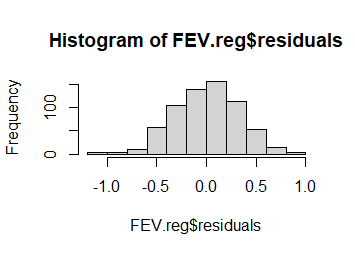
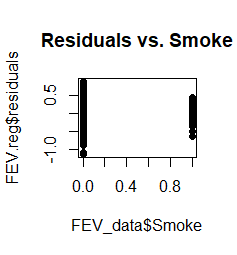




H0: β1 = 0 v.s. Ha: β1 ≠ 0

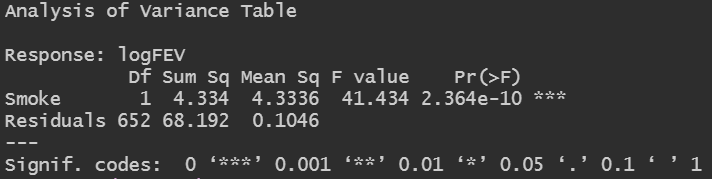
Assumptions:

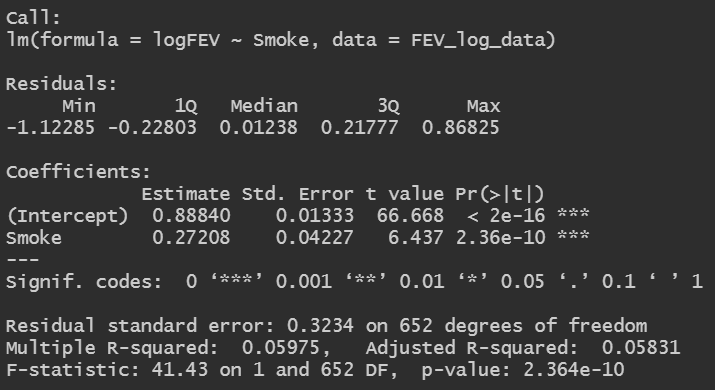
1. Normally distributed
2. Equal variances
3. Outliers
4. Independents
5. Linear association

Residuals are mostly linear from the QQ plot.

We can see some unequal variances.





Learn from the test that:

ˆβ0 = 0.88840 SE(ˆβ0) = 0.01333

ˆβ1 = 0.27208 SE(ˆβ1) = 0.04227

MSE = 0.1046

Test statistics T = 0.27208/0.04227 = 6.437 ~t(652)

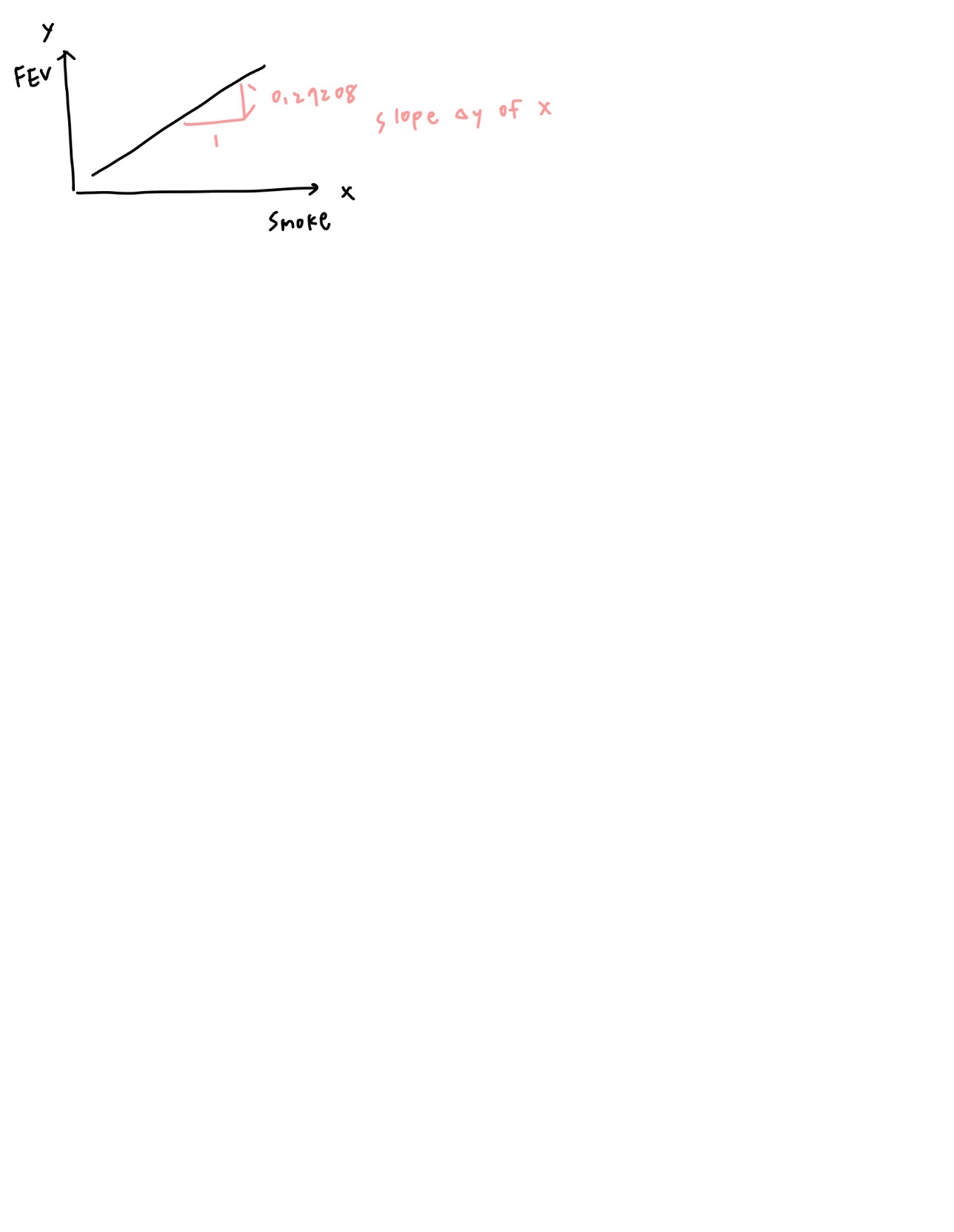
The 0.025 cut-off for t-distribution with 652 df is qt(0.025, df = 652): -1.963609. Thus a 95% confidence interval for β1 has endpoints 0.272-1.964(0.042) = 0.189 and 0.272+1.964(0.042) = 0.354, so we have 95% confident to say slope β1 is between 0.189 and 0.354.

P-value of the test of no association is 2.364e-10, which is very small, so there is a lot of evidence against the null hypothesis of no association between smoking status and FEV in the simple linear regression model.

* 1. Interpret the slope coefficient.

**Ans**:

slope coefficient = 0.27208



* 1. Compare your results to those in part 1a. You should draw *identical* conclusions. Do you? (Same estimated difference in mean LN(FEV) between non-smokers and smokers, same confidence interval, same p-value.)

**Ans**:

**two-sample t test**

* estimated difference in mean: 1.1604760 - 0.8883953 = 0.272
* CI: 95%, 0.208~0.335
* p-value: 2.983e-13

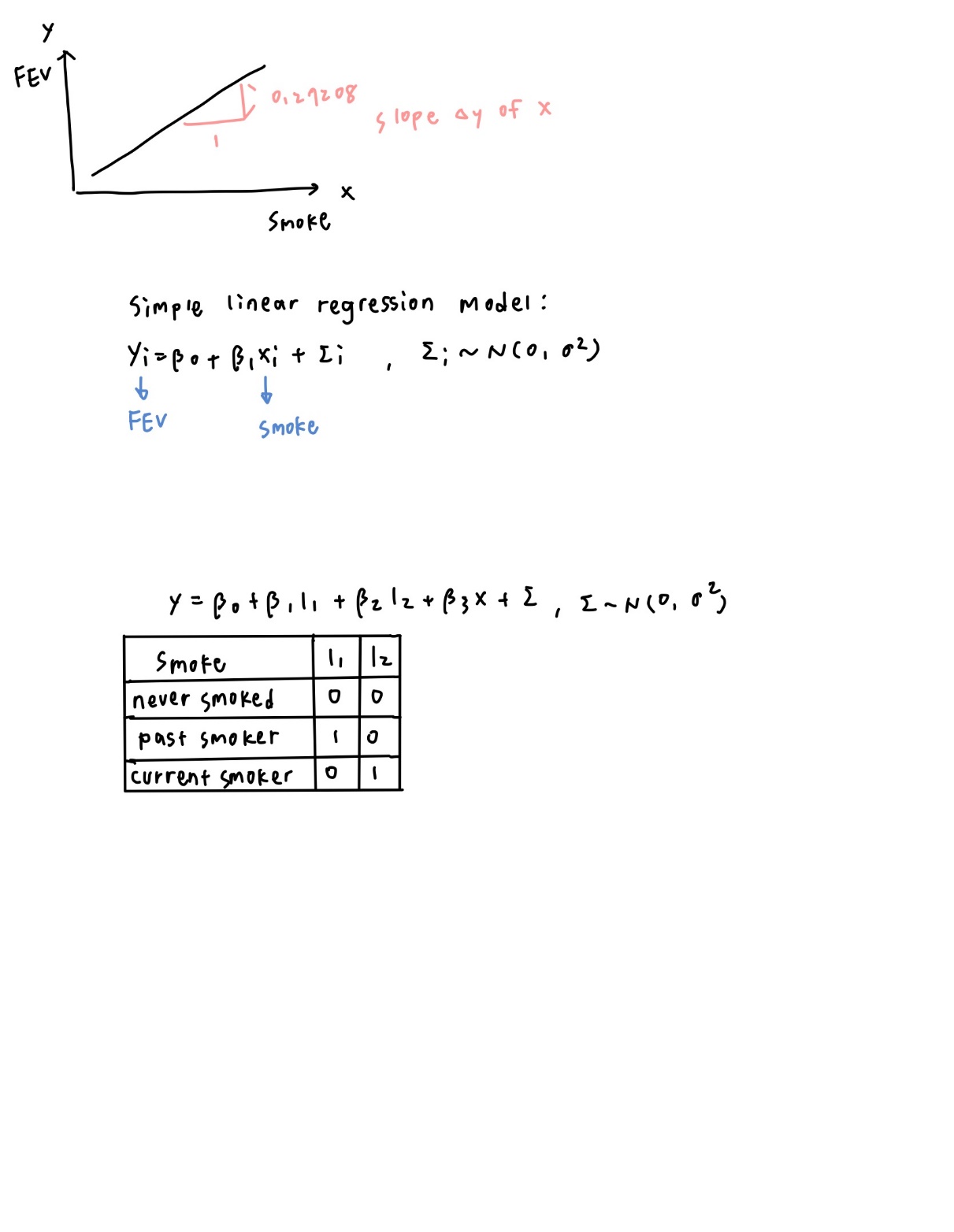
**regression**

* estimated difference in mean: 0.272(slope coefficient)
* CI: 95%, 0.189~0.354
* p-value: 2.364e-10

From the data listed above we can conclude that the results are nearly the same, it both indicates that there is a lot of evidence against the null hypothesis of no association between smoking status and FEV.

* 1. Explain how to use a regression model with indicator variables to include a three (or more) category explanatory variable. For example, *if* smoking status was 0 for never smoked, 1 for past smoker, and 2 for current smoker, how would you incorporate the three smoking levels into a regression model? (Note: This gives us a way to incorporate both quantitative and qualitative variables into a model!)

**Ans**:



So now we have 3 groups we need 3-1 = 2 indicator variables to further fit the multiple linear regression model. Set never smoked as “reference group”.