STAT 337 Tutorial 3 – Module 3 – Confounding and effect modification

Section 3.1: Causation and Confounding

Causation: A change in X CAUSES a direct change in Y

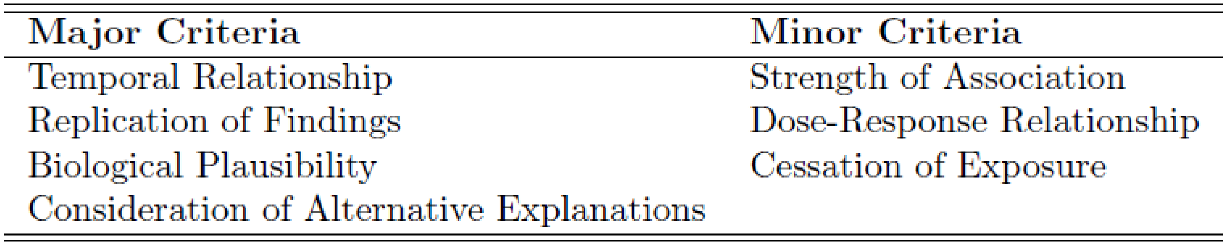
Association: As X changes, Y also appears to change, i.e. as X increases, Y tends to increase etc.

The goal in medical research is to determine if there's an association between an exposure and disease by conducting studies which examine the characteristics of the individuals with and without the disease. To investigate these questions the following order is typically used:

A. Descriptive studies – based on clinical observations or available data. Used to generate hypothesis and refine medical research questions.
B. Observational Studies – typically start with Case-Control and/or Cross sectional since relatively quick and then can move to cohort.
C. Experimental Studies – Randomised control trials.

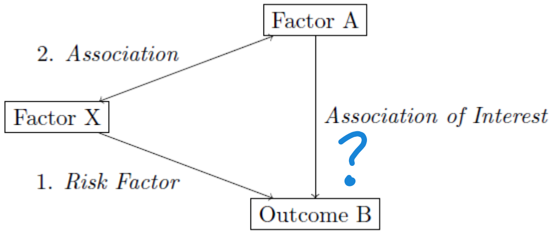
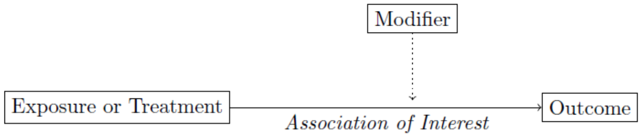
Guidelines for judging whether an association is Causal:

Set of Criteria created by Sir Austin Bradford Hill



Major Criteria	Minor Criteria
Temporal Relationship	Strength of Association
Replication of Findings	Dose-Response Relationship
Biological Plausibility	Cessation of Exposure
Consideration of Alternative Explanations	

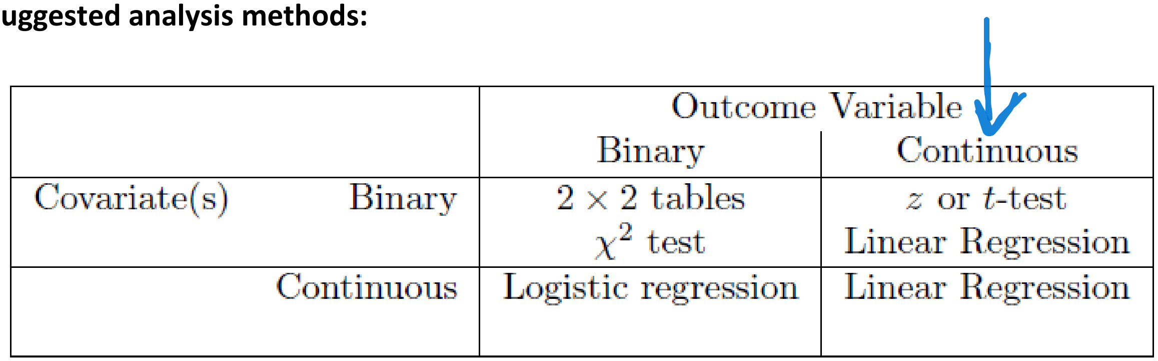
Bias, Confounding and Effect Modification:

Bias	**Confounding**	**Effect Modification**
A systematic error in the design or conduct of a study that results in an erroneous estimate of the true association between exposure and outcome, i.e creates an association that is not true: Measurement error/ bias Nonresponse bias, Recall bias Selection bias, Information bias, Misclassification bias	Occurs when the association between an exposure of interest and disease/outcome is distorted by the presence of another factor. Put a simpler way - Two variables are said to be confounded if it is impossible to separate their effects on the response. i.e. describes an association that exists but potentially is misleading. If the confounder is known and measured, we can adjust for it in our modelling. **Formal definition:** 1.Factor X is a risk factor for Outcome B 2.Factor X is associated with Factor A BUT not a direct result of Factor A. Then X is a confounder  Factor A 2. *Association* Factor X *Association of Interest* ? 1. *Risk Factor* Outcome B -Positive confounding – true association is enhanced (looks stronger). -Negative confounding – true association is dimmed (looks weaker)	Occurs when a variable differently modifies the true association between the exposure of interest and the disease/ outcome, i.e there is a different association in different groups.  Modifier Exposure or Treatment *Association of Interest* Outcome

Carefully planning for your study can help reduce bias	At Design stage can minimise effect by matching, stratification, randomisation if unknown. At Analysis stage, if have data can include in analysis to adjust/control for its effect. Linear regression, logistic regression, post-stratification or restriction	Typically investigated using stratification and/or interaction terms in linear or logistic regression.

Section 3.2: Multiple Linear Regression

Suggested analysis methods:



		Outcome Variable	
		Binary	Continuous ✓
Covariate(s)	Binary	2 × 2 tables χ^2 test	z or t-test Linear Regression
	Continuous	Logistic regression	Linear Regression

Simple Linear Regression:

Purpose of model:

1. Describe the associations between the outcome of interest and one or more explanatory variables.
2. Predict values of the outcome variable for new levels of the explanatory variable.
3. Adjust for other covariates and confounding variables.

In simple linear regression there is only one explanatory variable and the notation used is as follows:

- Response Variable, Y
- Explanatory Variable, X

We are interested in estimating the line of best fit, where a straight-line relating Y (response variable) to *x* (explanatory variable) has an equation of the form:

$$\underline{\boldsymbol{\mu_{Y|X}}} = \boxed{\boldsymbol{\beta_0 + \beta_1 x}}$$

Where

- $\boldsymbol{\mu_{Y|X} = E(Y|x)}$ and represents the true mean value of Y for a given value of x.
- $\boldsymbol{\mu_{Y|X}}$ is typically referred to as the **deterministic part** of the model and captures the **known variation.**

Visualising the line of best fit:

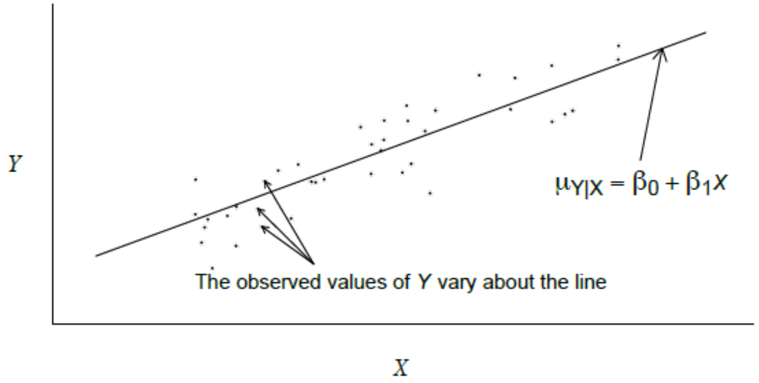


Figure 15.2: Illustration of the linear regression model for a simulated data set.

More conventionally we write the model as:

$$\boldsymbol{Y = \beta_0 + \beta_1 x + \epsilon}$$

Total Variation = known variation + unknown variation

Where

- Y is the response variable (observed)
- x is the explanatory variable (observed)
- β_0 is the Y-intercept (unknown parameter and needs to be estimated)
- β_1 is the slope (unknown parameter and needs to be estimated)
- ϵ is a random error term (residual), and represents the unknown variation

Once the unknown parameters are estimated using the data collected, we can use our **estimated regression line** to make predictions on the response variable:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x = \hat{\mu}_{Y|X}$$

Where

- $\hat{Y}$ represents the **predicted value of Y** for a given value of X, strictly speaking is captures $\hat{\mu}_{Y|X}$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ represent the statistics that estimate β_0 and β_1 respectively.

Interpreting the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$:

- $\boldsymbol{\hat{\beta}_0}$ is the estimated average response when **x=0** (may not be of interest depending on whether x=0 has meaning or not)
- $\boldsymbol{\hat{\beta}_1}$ is the estimated change in the average response for a **one unit increase** in x.

This notion can be extended into Multiple Linear Regression:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \qquad \text{with} \qquad \text{iid } \epsilon_i \sim N(0, \sigma^2)$$

The β_j's are the regression coefficients. Here β_0 is the intercept (i.e. the expected value of y at $x_1 = \cdots = x_p = 0$) and for $j = 1, \ldots, p$, β_j represents the expected change in y for a one unit increase in x_j holding all other x's constant.

Testing the significance of an explanatory variables, i.e is it really associated with the outcome, while adjusting for the other covariates.

Step 1: State your Hypothesis - $H_0: \beta_i = 0$ $vs.$ $H_a: \beta_i \neq 0$

Step 2: Solve for test statistic value

The test statistic is

$$t^* = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)}$$

which has a $\underline{t_{n-(p+1)}}$ distribution (note $p+1$ parameters in the model).

n = sample size

p+1 = total # of parameters.

Step 3: Solve for p-value or critical value

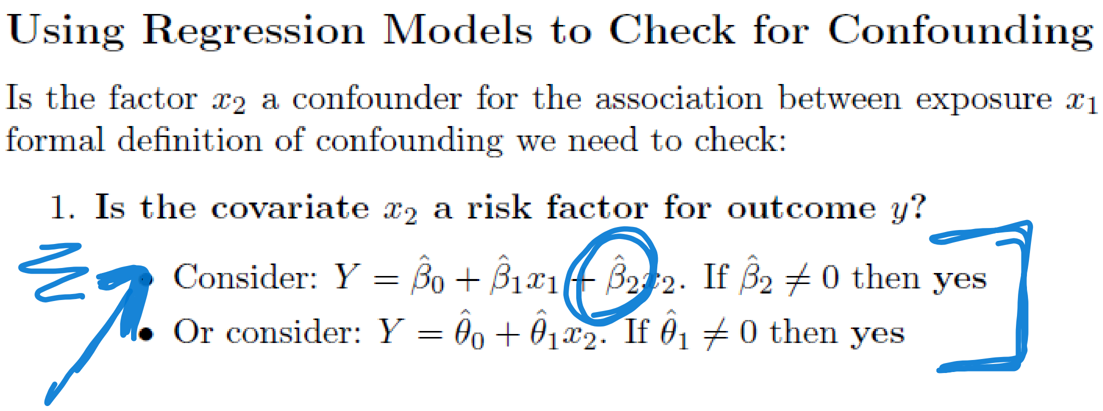
Step 4: Make conclusion – if p-value < alpha we Reject the null hypothesis.

Interpreting the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_i s$:

- $\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$ is the estimated average response when **x=0** (may not be of interest depending on whether x=0 has meaning or not)
- $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ is the estimated change in the average response for a **one unit increase** in the corresponding x_i, while holding all other covariates fixed.

Using Regression Models to check for Confounding:

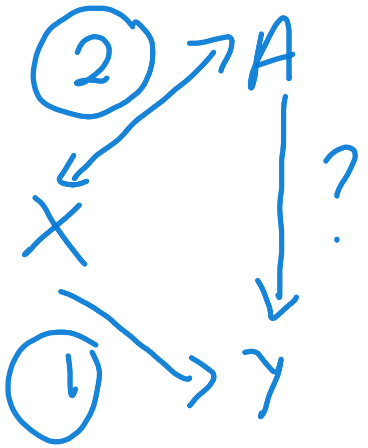
Recall confounding Occurs when the association between an exposure of interest and disease/outcome is distorted by the presence of another factor. Put a simpler way - Two variables are said to be confounded if it is impossible to separate their effects on the response. i.e. describes an association that exists but potentially is misleading. If the confounder is known and measured, we can adjust for it in our modelling.



Using Regression Models to Check for Confounding

Is the factor x_2 a confounder for the association between exposure x_1 and outcome Y? Using the formal definition of confounding we need to check:

1. **Is the covariate x_2 a risk factor for outcome y?**
   - Consider: $Y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. If $\hat{\beta}_2 \neq 0$ then yes
   - Or consider: $Y = \hat{\theta}_0 + \hat{\theta}_1 x_2$. If $\hat{\theta}_1 \neq 0$ then yes



continuous

2. **Is x_2 associated with exposure x_1 (but not a direct consequence of it)?**
   - Consider a model like $x_1 = \hat{\gamma}_0 + \hat{\gamma}_1 x_2$. If $\hat{\gamma}_1 \neq 0$ then **yes**
   - Use scientific context of the study to determine if x_2 is a direct consequence of x_1

If we answer yes to both of the above then we conclude that x_2 is a confounder. Note in above tests often use a larger significance level (i.e. $\alpha = 0.10$ or 0.20).

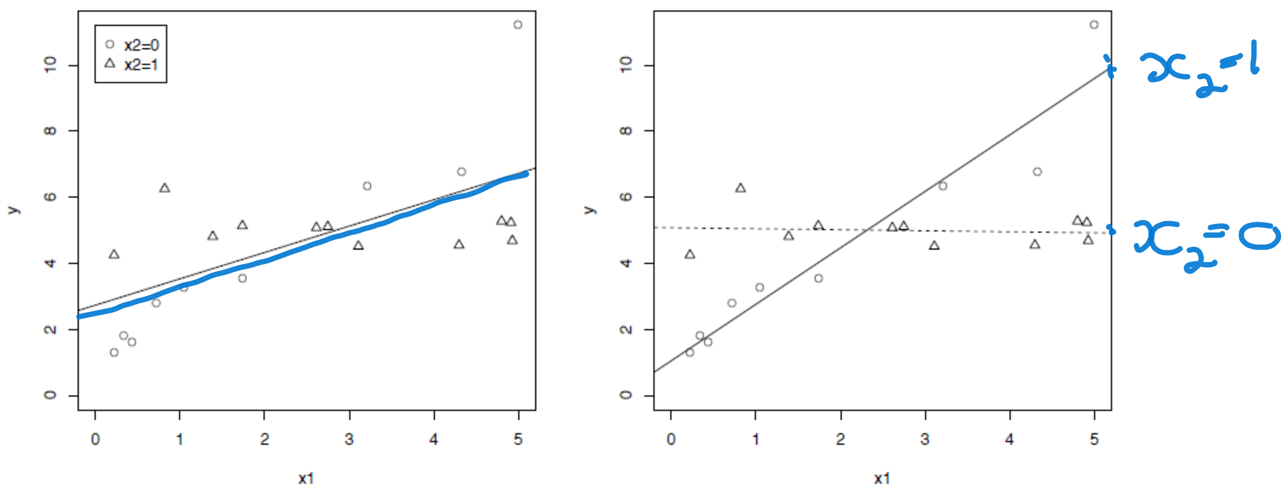
Using Regression Models to check for effect modification:

Recall effect modification occurs when a variable differently modifies the true association between the exposure of interest and the disease/ outcome, i.e there is a different association in different groups.

For this we need to introduce an interaction term leading to the following

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i \qquad \text{with} \qquad \text{iid } \epsilon_i \sim N(0, \sigma^2)$$

The fitted models are shown below.



$$Y = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 \qquad\qquad Y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$$

Let's examine the interaction model. When $x_2 = 0$ the model becomes

$$Y = \beta_0 + \beta_1 x_1$$

When $x_2 = 1$ the model is becomes

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1$$

Example: 18

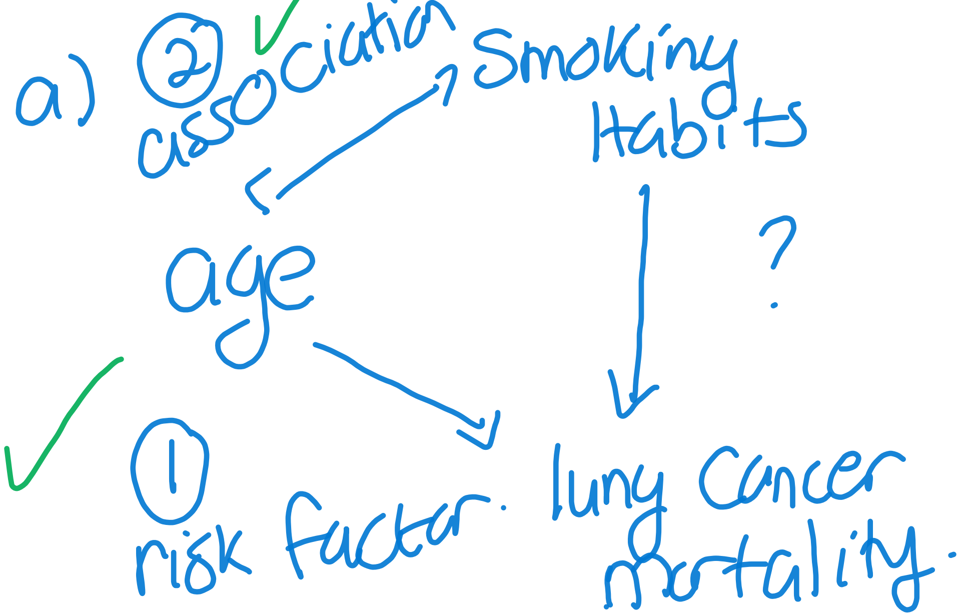
This question is based on the following paper:

> Doll, Richard and Hill, A. Bradford. *The Mortality of Doctors in Relation to Their Smoking Habits.* British Medical Journal, June 26, 1954, pp 145-1455.

This is the first major report in what would be a long running prospective cohort study.

18. (a) Referring back to the paper in question 16. (in particular Table I) would you consider age to be a confounder for the relationship between smoking and lung cancer mortality? Use the formal definition of confounding in your solution.

(b) If you were to run a similar cohort study today give at least two other potential confounders would you want to be sure to consider.



① Age is a risk factor for mortality since mortality rates are known to increase as age increases.

②

smoking. doesn't cause age.

an association are possibly exists

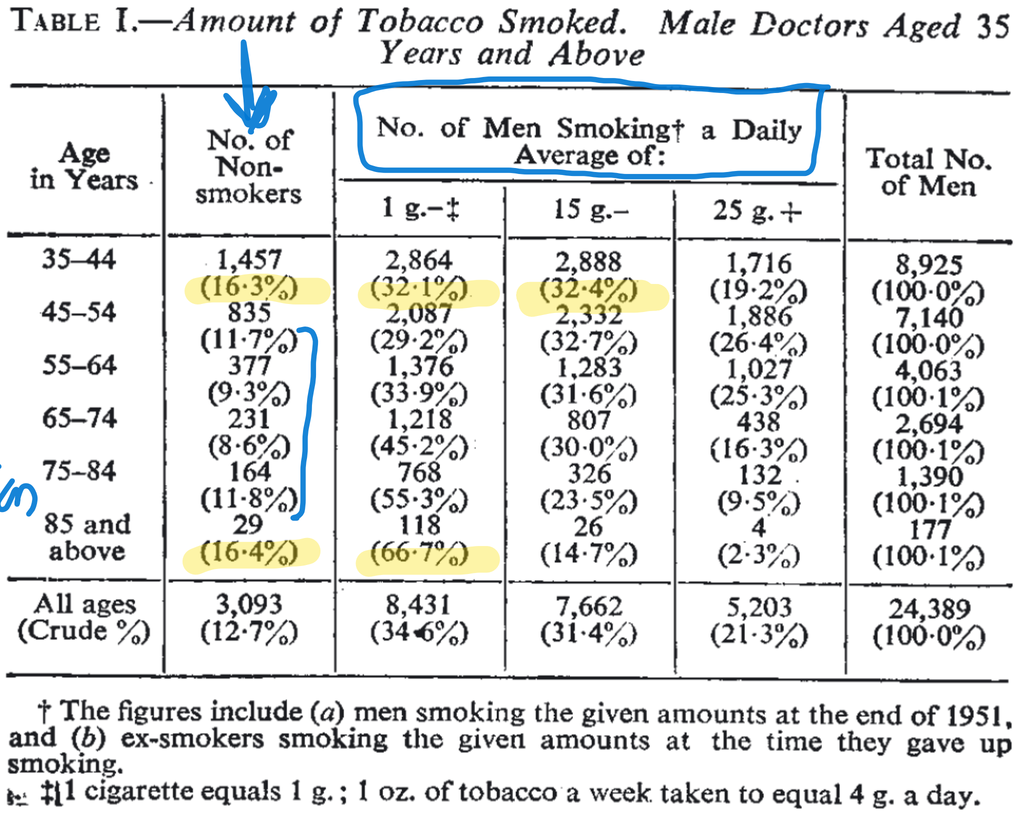


Table I.—*Amount of Tobacco Smoked. Male Doctors Aged 35 Years and Above*

Age in Years	No. of Non-smokers	No. of Men Smoking† a Daily Average of:			Total No. of Men
		1 g.–‡	15 g.–	25 g.+	
35–44	1,457 (16·3%)	2,864 (32·1%)	2,888 (32·4%)	1,716 (19·2%)	8,925 (100·0%)
45–54	835 (11·7%)	2,087 (29·2%)	2,332 (32·7%)	1,886 (26·4%)	7,140 (100·0%)
55–64	377 (9·3%)	1,376 (33·9%)	1,283 (31·6%)	1,027 (25·3%)	4,063 (100·1%)
65–74	231 (8·6%)	1,218 (45·2%)	807 (30·0%)	438 (16·3%)	2,694 (100·1%)
75–84	164 (11·8%)	768 (55·3%)	326 (23·5%)	132 (9·5%)	1,390 (100·1%)
85 and above	29 (16·4%)	118 (66·7%)	26 (14·7%)	4 (2·3%)	177 (100·1%)
All ages (Crude %)	3,093 (12·7%)	8,431 (34·6%)	7,662 (31·4%)	5,203 (21·3%)	24,389 (100·0%)

† The figures include (*a*) men smoking the given amounts at the end of 1951, and (*b*) ex-smokers smoking the given amounts at the time they gave up smoking.

‡1 cigarette equals 1 g.; 1 oz. of tobacco a week taken to equal 4 g. a day.

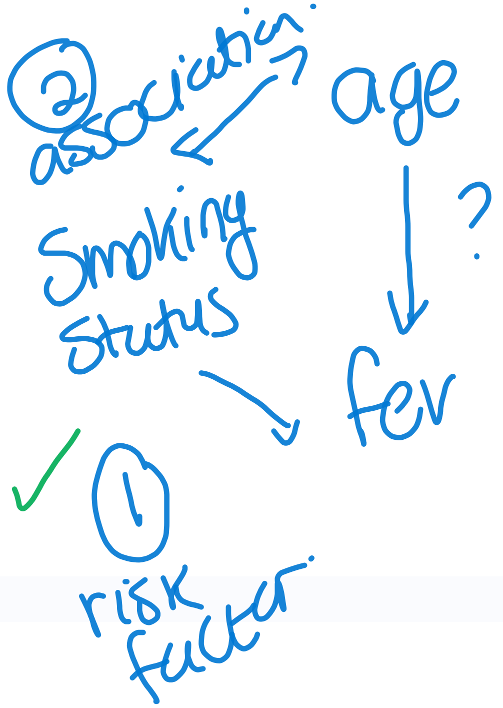
Since the table shows more non-smokers at a young and old age groups than in the middle.

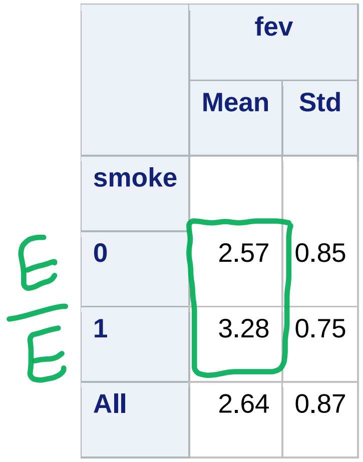
Example:

The dataset ‘FEV.xls’ contains information to investigate forced expiratory volume (FEV measured in liters) as a primary indicator of lung function. FEV corresponds to the volume of air that can forcibly be blown out in the first second after full inspiration. The variables included are FEV, age in years, height, sex (0: Female, 1: Male), and exposure to smoke (0: exposed, 1: not exposed).

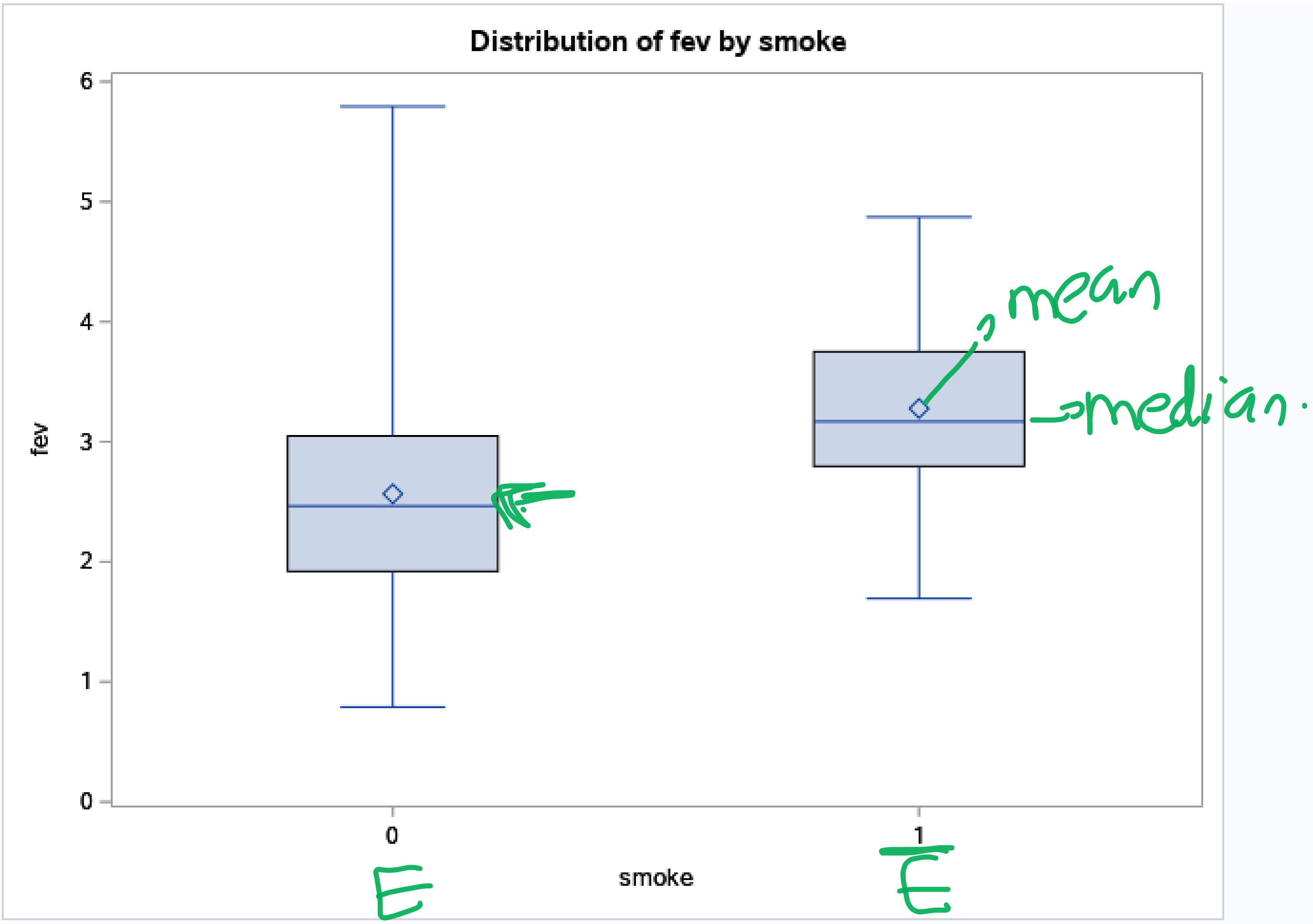
a) The investigators are interested in determining whether there is an association between FEV and Age. However, they suspect that smoking status is a confounding variable.
   I. Using the formal definition of confounding determine whether smoking status is a confounding variable.
   II. Fit appropriate regression models to determine whether smoking status is a confounding variable.

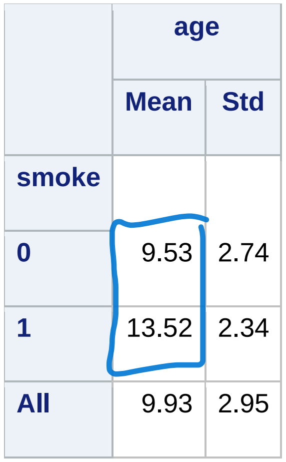
b) The investigators are again interested in determining whether there is an association between FEV and Age. However, they now want to determine whether sex is an effect modifier.
   I. Create an appropriate scatter plot and comment on whether sex appears to be an effect modifier.
   II. Fit an appropriate regression model to determine whether sex is an effect modifier.




E
$\bar{E}$

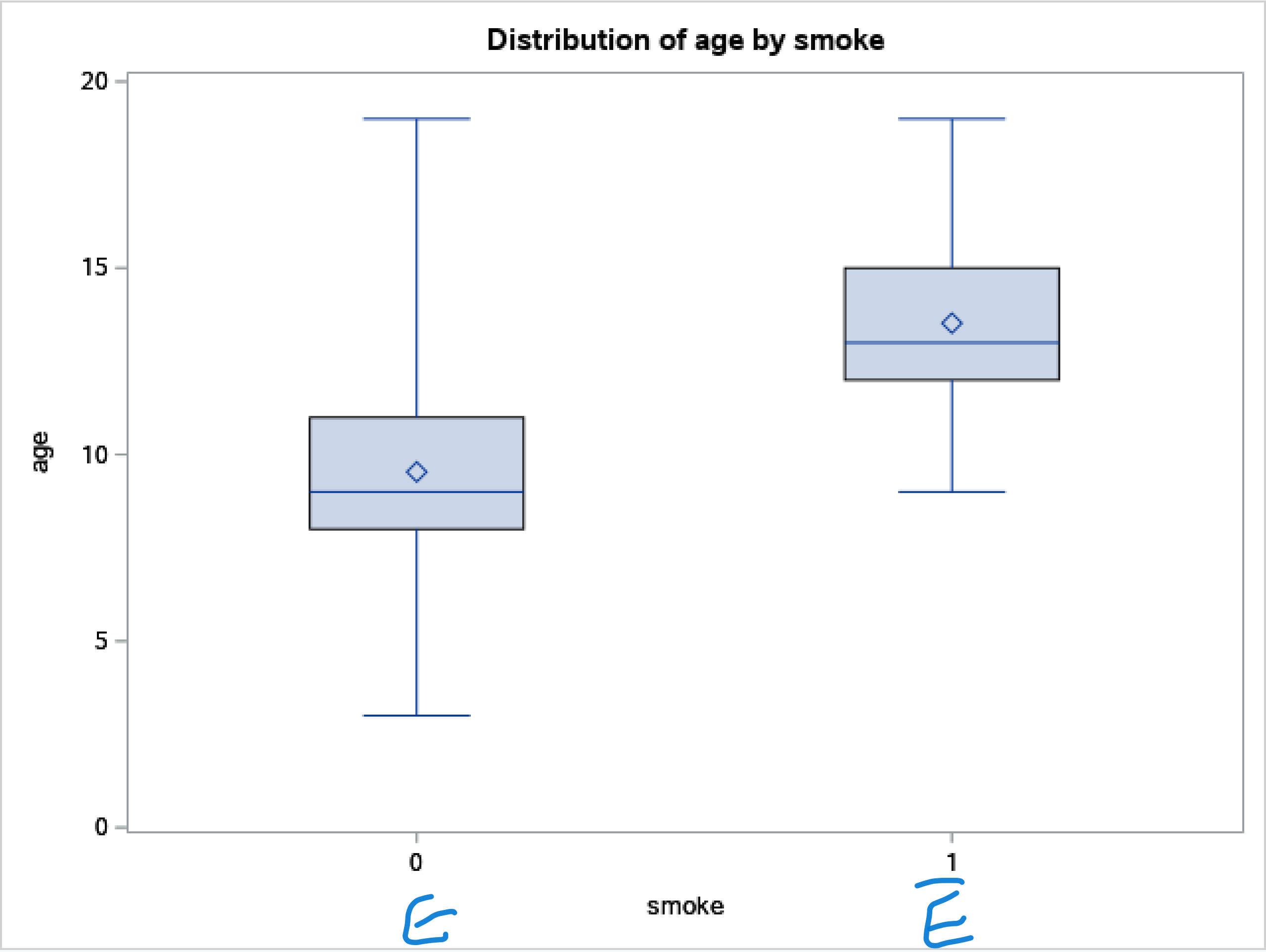
	fev	
	Mean	Std
smoke		
0	2.57	0.85
1	3.28	0.75
All	2.64	0.87


Distribution of fev by smoke
fev
6
5
4
3
2
1
0
mean
median.
0
1
smoke
E
Ē



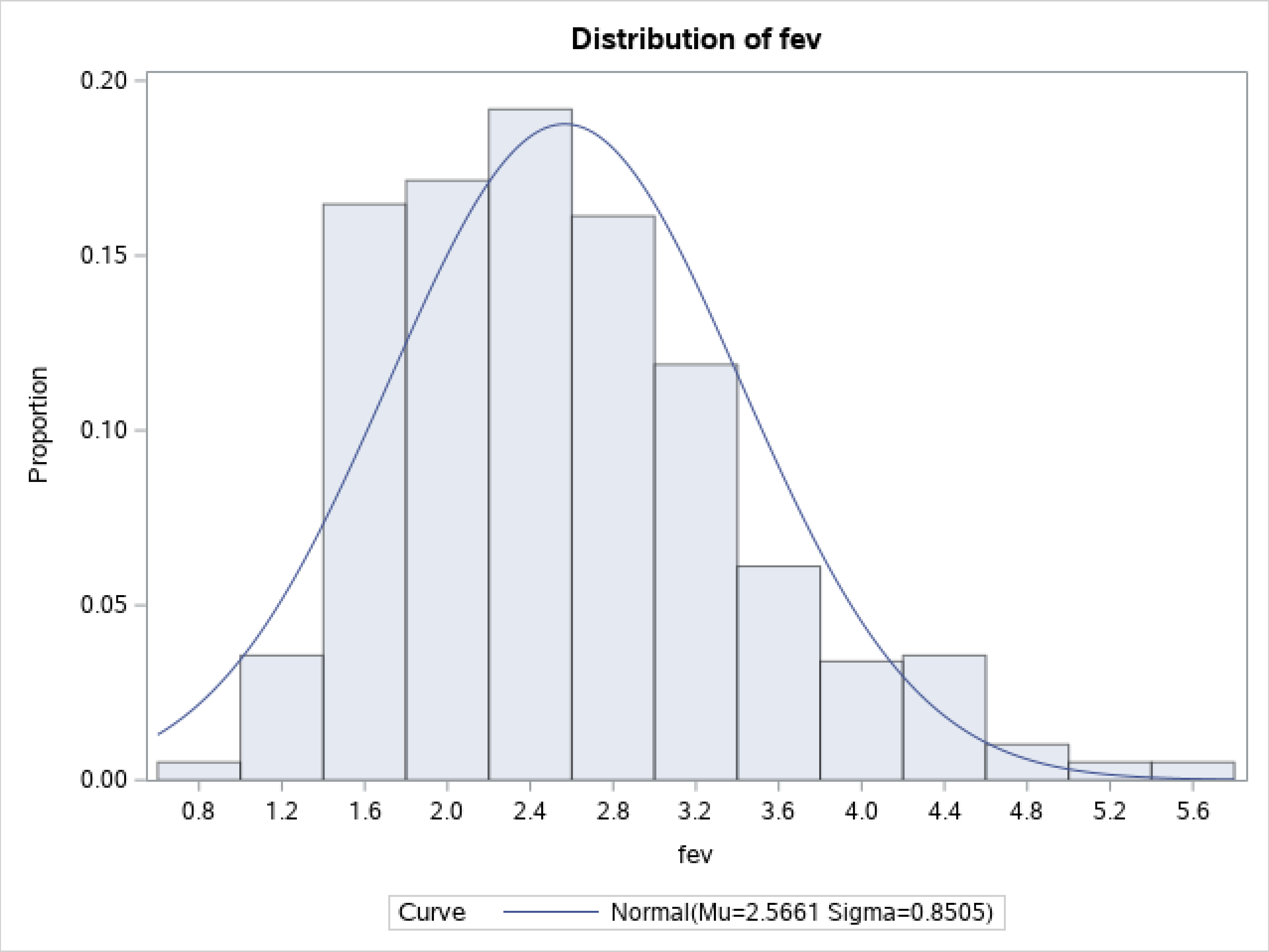
	age	
	Mean	Std
smoke		
0	9.53	2.74
1	13.52	2.34
All	9.93	2.95

$\frac{E}{E}$



The UNIVARIATE Procedure

smoke=0



The UNIVARIATE Procedure
Fitted Normal Distribution for fev (fev)

smoke=0

Parameters for Normal Distribution

Parameter	Symbol	Estimate
Mean	Mu	2.566143
Std Dev	Sigma	0.850522

Goodness-of-Fit Tests for Normal Distribution

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.05669864	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.69524680	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	4.87043523	Pr > A-Sq	<0.005

Quantiles for Normal Distribution

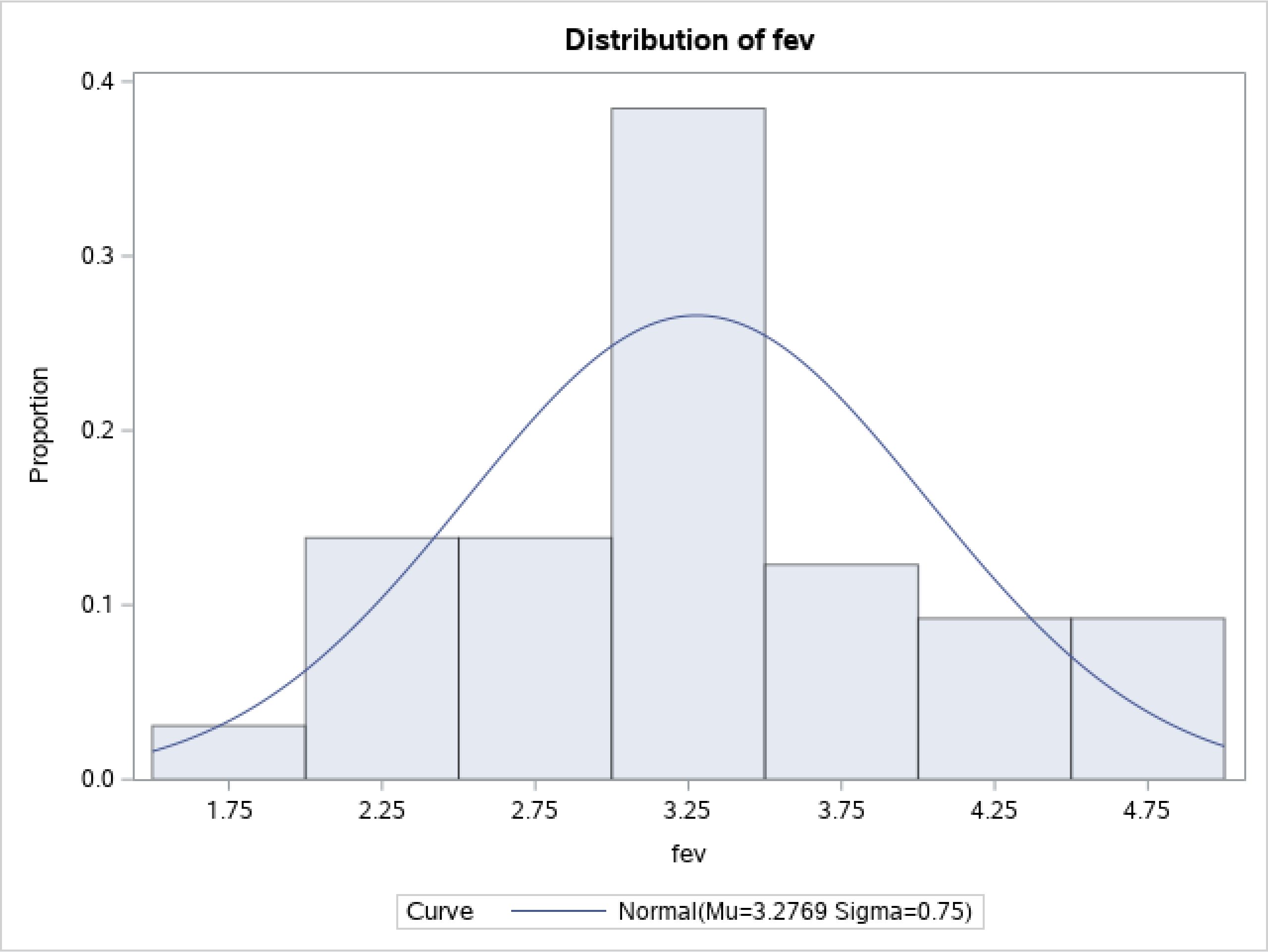
	Quantile	
Percent	Observed	Estimated
1.0	1.09200	0.58753
5.0	1.42300	1.16716
10.0	1.58900	1.47616
25.0	1.92000	1.99247

Quantiles for Normal Distribution

	Quantile	
Percent	Observed	Estimated
50.0	2.46500	2.56614
75.0	3.04800	3.13981
90.0	3.74100	3.65613
95.0	4.23200	3.96513
99.0	5.08300	4.54475

The UNIVARIATE Procedure

smoke=1



The UNIVARIATE Procedure
Fitted Normal Distribution for fev (fev)

smoke=1

Parameters for Normal Distribution

Parameter	Symbol	Estimate
Mean	Mu	3.276862
Std Dev	Sigma	0.749986

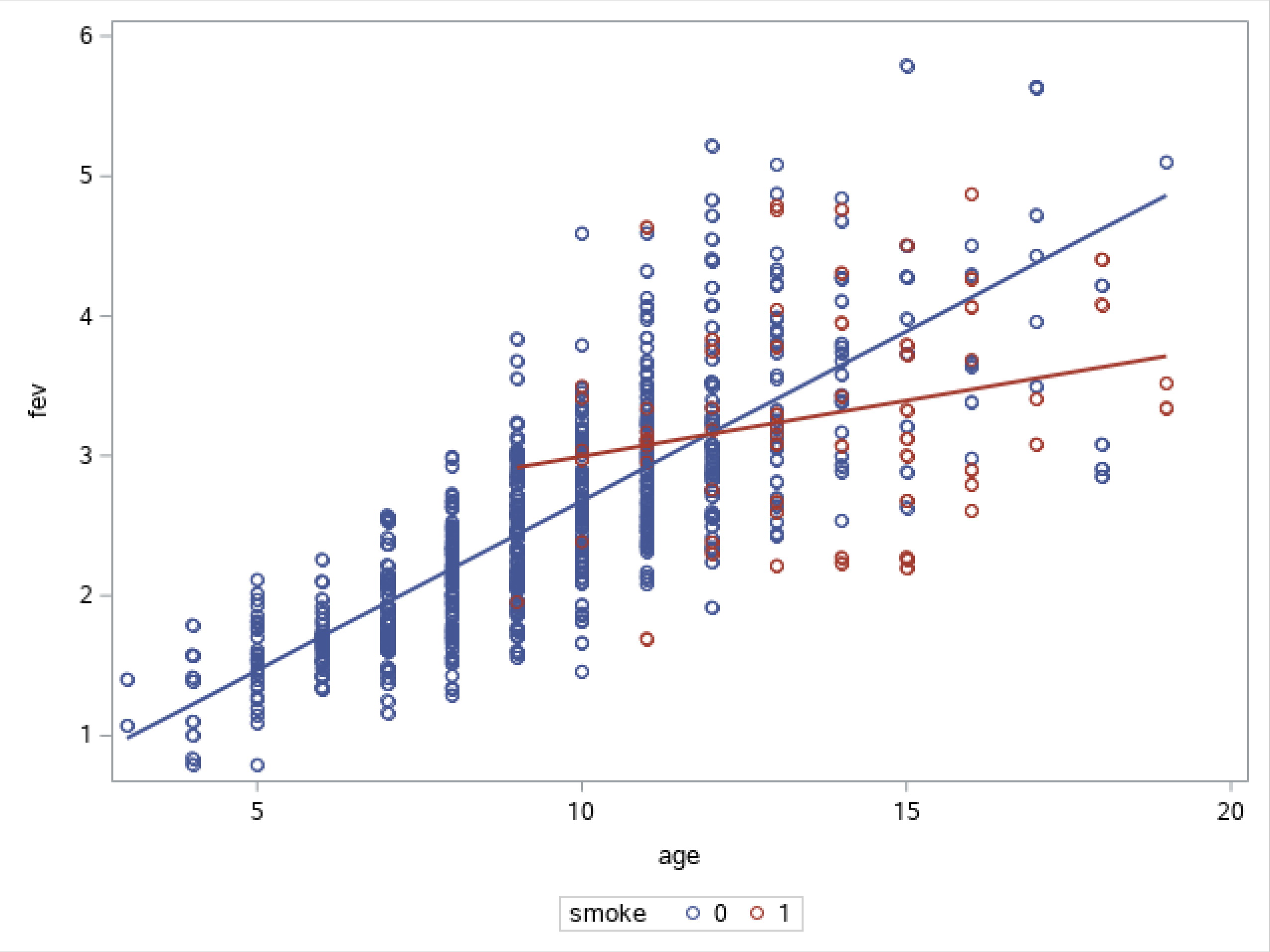
Goodness-of-Fit Tests for Normal Distribution

Test	Statistic		p Value	
Kolmogorov-Smirnov	**D**	0.09706844	**Pr > D**	0.131
Cramer-von Mises	**W-Sq**	0.10252212	**Pr > W-Sq**	0.103
Anderson-Darling	**A-Sq**	0.61174156	**Pr > A-Sq**	0.108

Quantiles for Normal Distribution

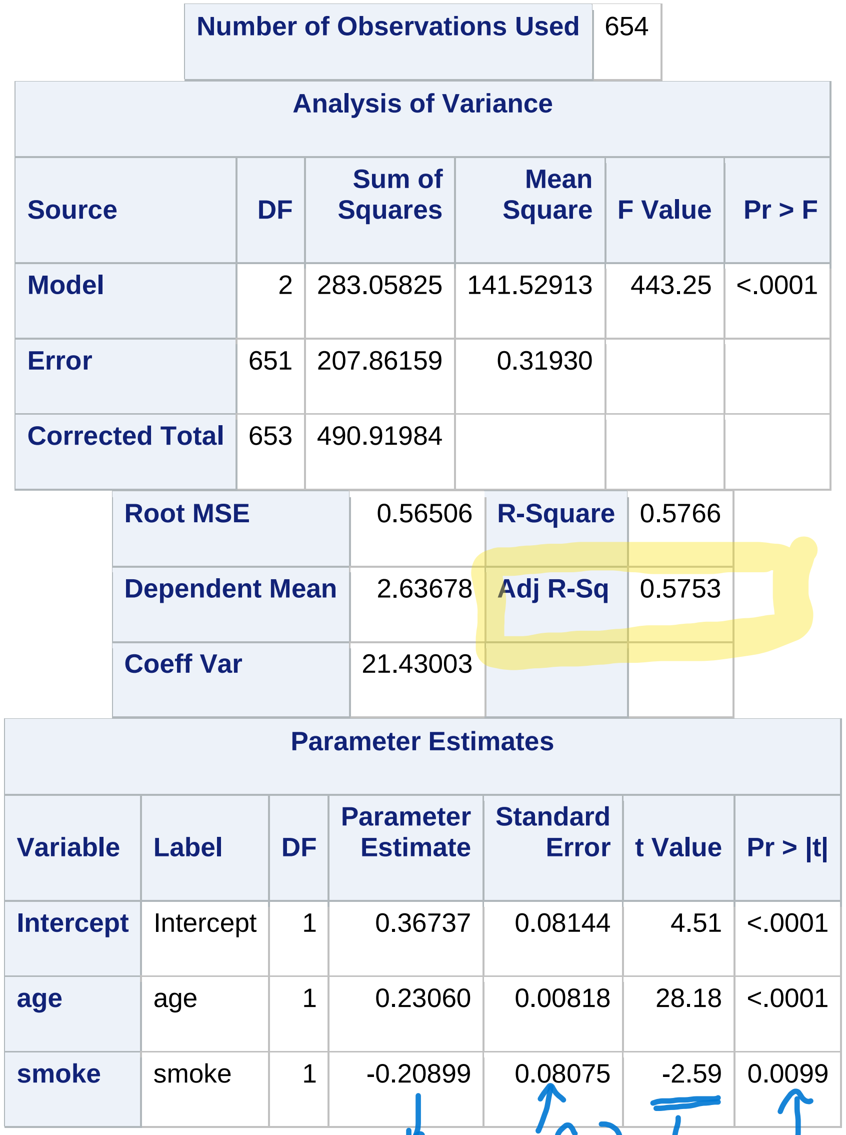
	Quantile	
Percent	Observed	Estimated
1.0	1.69400	1.53213
5.0	2.21600	2.04324
10.0	2.27600	2.31572
25.0	2.79500	2.77100

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
50.0	3.16900	3.27686
75.0	3.75100	3.78272
90.0	4.40400	4.23801
95.0	4.75600	4.51048
99.0	4.87200	5.02159



The REG Procedure
Model: MODEL1
Dependent Variable: fev fev

Number of Observations Read	654



Number of Observations Used	654

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	283.05825	141.52913	443.25	<.0001
Error	651	207.86159	0.31930		
Corrected Total	653	490.91984			

Root MSE	0.56506	R-Square	0.5766
Dependent Mean	2.63678	Adj R-Sq	0.5753
Coeff Var	21.43003		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.36737	0.08144	4.51	<.0001
age	age	1	0.23060	0.00818	28.18	<.0001
smoke	smoke	1	-0.20899	0.08075	-2.59	0.0099

$\hat{\beta}_2$

$se(\hat{\beta}_2)$

$$\frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)}$$

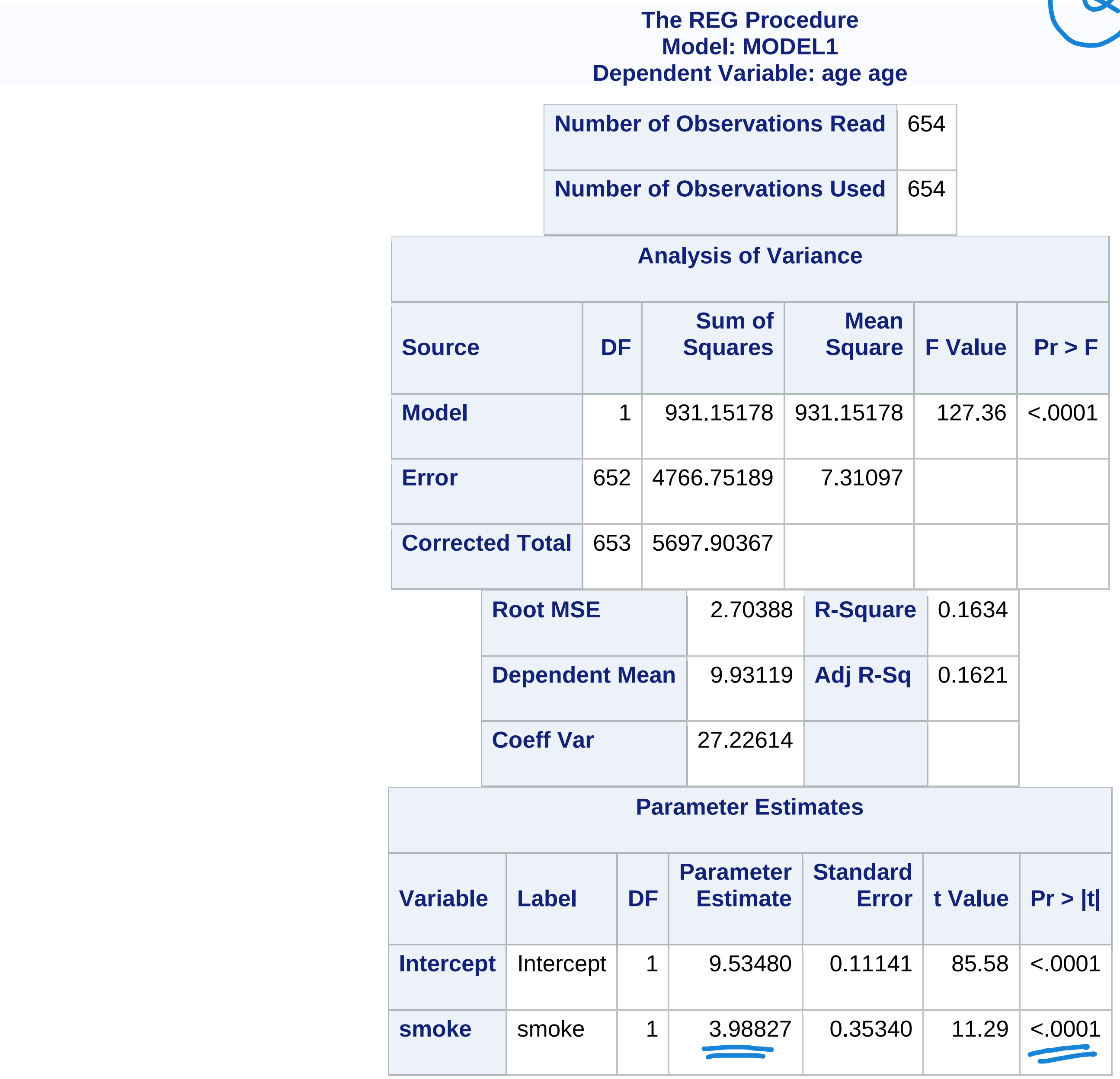
p-value.

① is smoking a risk factor for fev?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

p-value $= 0.0099 < \alpha = 0.1$

$\Rightarrow$ smoking is a risk factor on fev.



The REG Procedure
Model: MODEL1
Dependent Variable: age age

Number of Observations Read	654
Number of Observations Used	654

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	931.15178	931.15178	127.36	<.0001
Error	652	4766.75189	7.31097		
Corrected Total	653	5697.90367			

Root MSE	2.70388	R-Square	0.1634
Dependent Mean	9.93119	Adj R-Sq	0.1621
Coeff Var	27.22614		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9.53480	0.11141	85.58	<.0001
smoke	smoke	1	3.98827	0.35340	11.29	<.0001

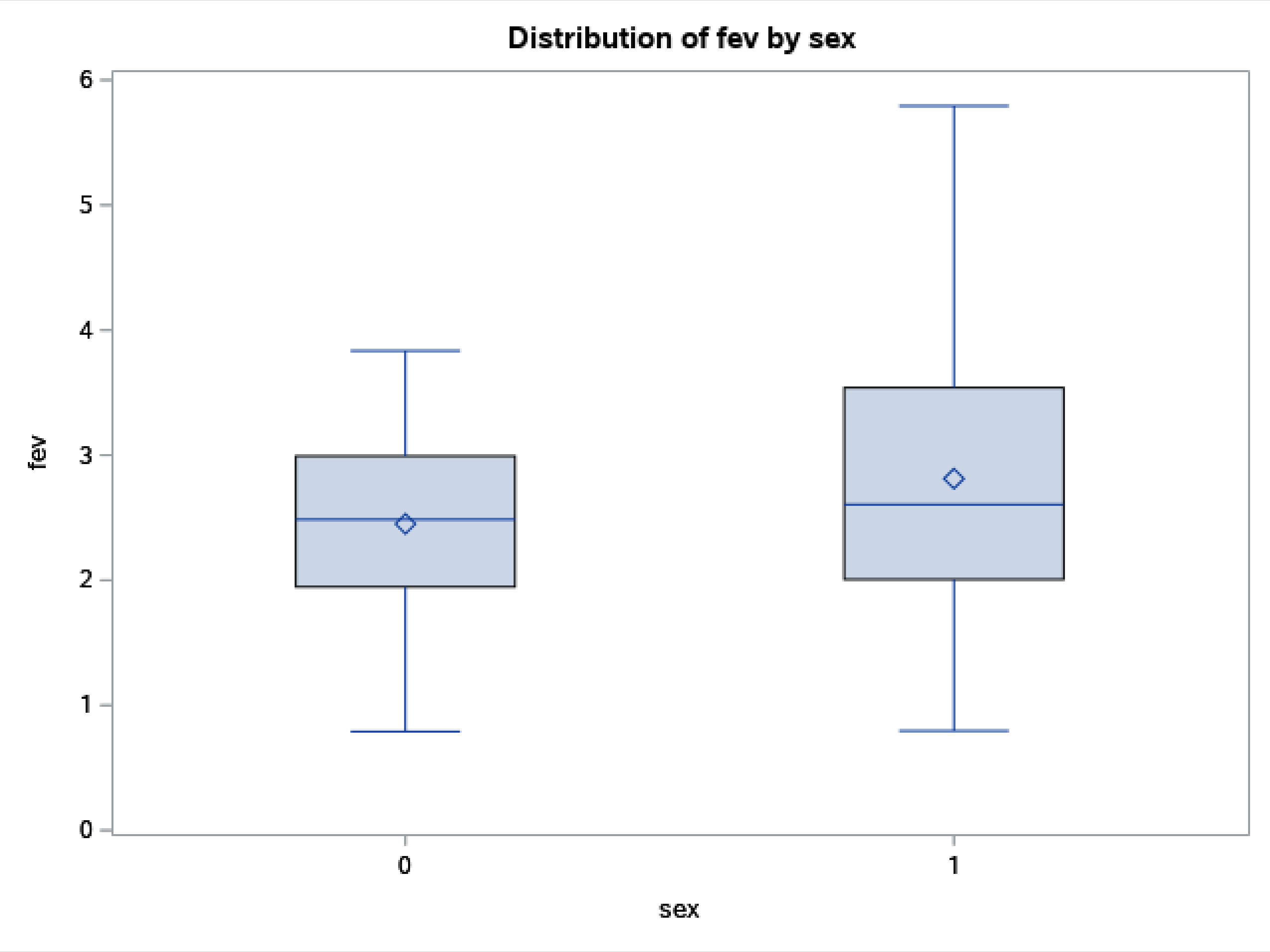
2 is smoking associated with age but not a direct cause of age?

$x_1 = \beta_0 + \beta_1 x_2 + \varepsilon.$

p-value < 0.0001
$< \alpha = 0.1$

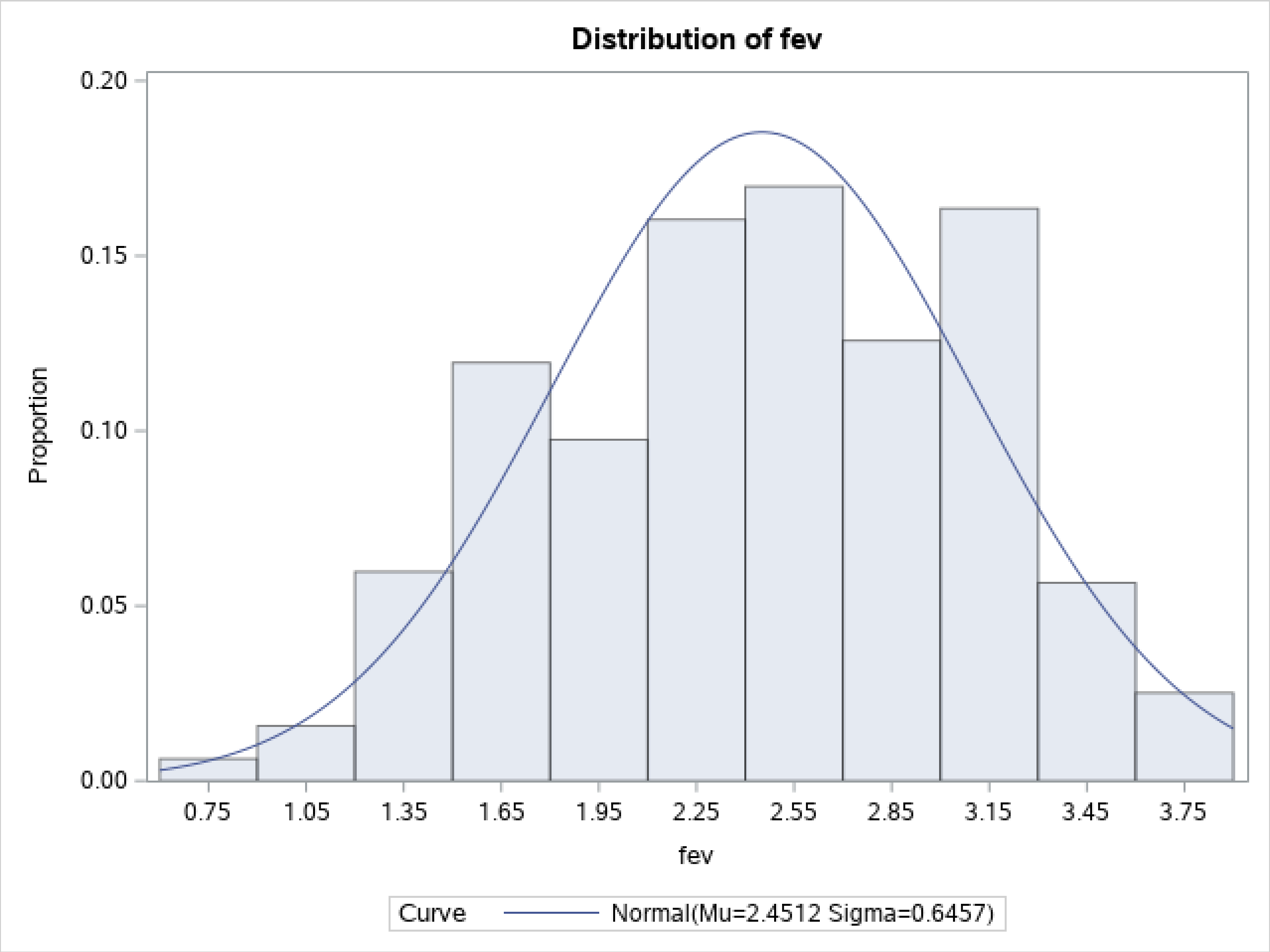
$\Rightarrow$ smoking is associated with age

	fev	
	Mean	Std
sex		
0	2.45	0.65
1	2.81	1.00
All	2.64	0.87



The UNIVARIATE Procedure

sex=0



The UNIVARIATE Procedure
Fitted Normal Distribution for fev (fev)

sex=0

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	2.45117
Std Dev	Sigma	0.645736

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.05822940	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.21055286	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	1.38204856	Pr > A-Sq	<0.005

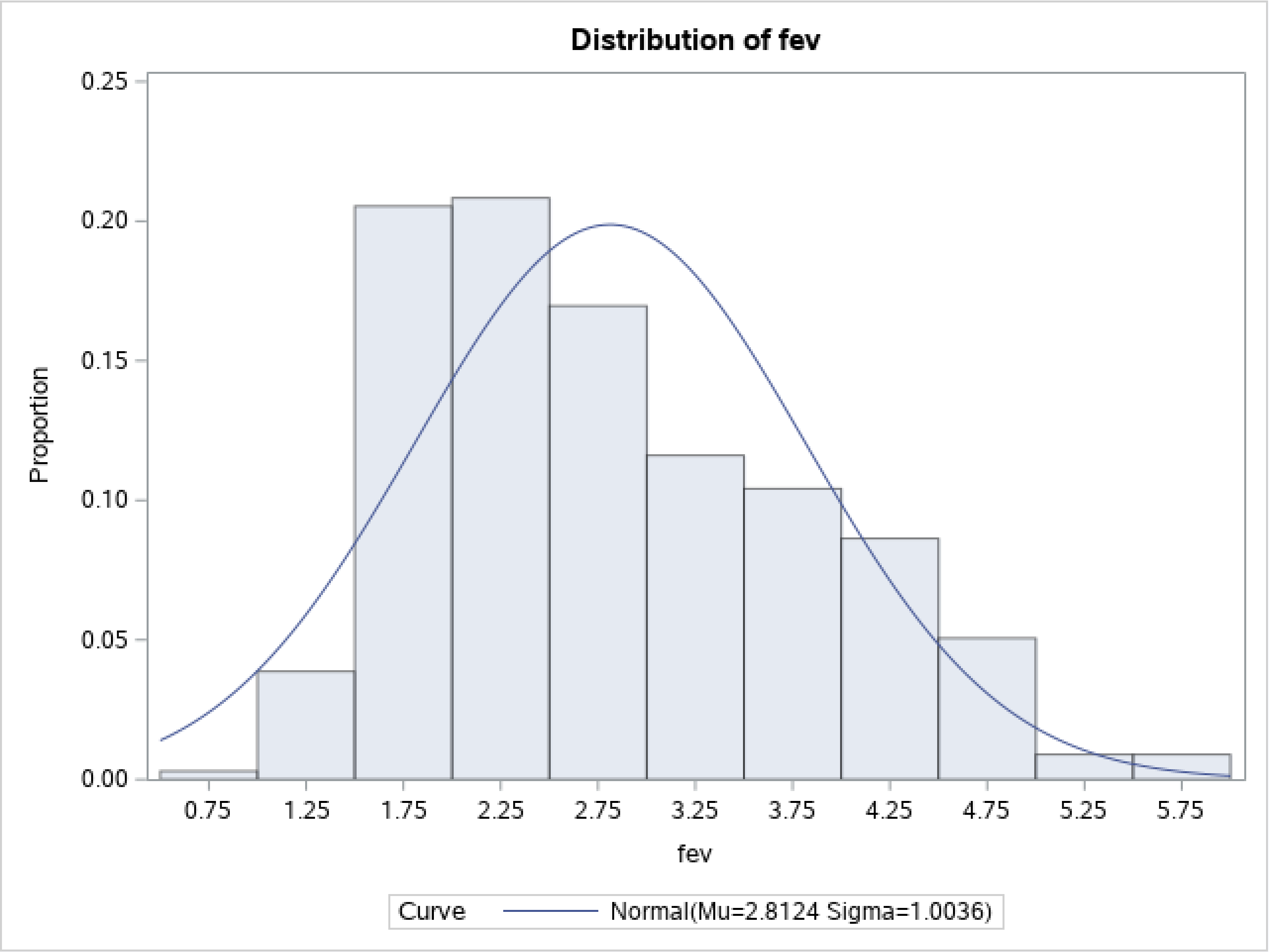
Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	1.09200	0.94896
5.0	1.37000	1.38903
10.0	1.55200	1.62363
25.0	1.94700	2.01563

Quantiles for Normal Distribution

	Quantile	
Percent	Observed	Estimated
50.0	2.48600	2.45117
75.0	2.99300	2.88671
90.0	3.23600	3.27871
95.0	3.42800	3.51331
99.0	3.77400	3.95338

The UNIVARIATE Procedure

sex=1



The UNIVARIATE Procedure
Fitted Normal Distribution for fev (fev)

sex=1

Parameters for Normal Distribution

Parameter	Symbol	Estimate
Mean	Mu	2.812446
Std Dev	Sigma	1.003598

Goodness-of-Fit Tests for Normal Distribution

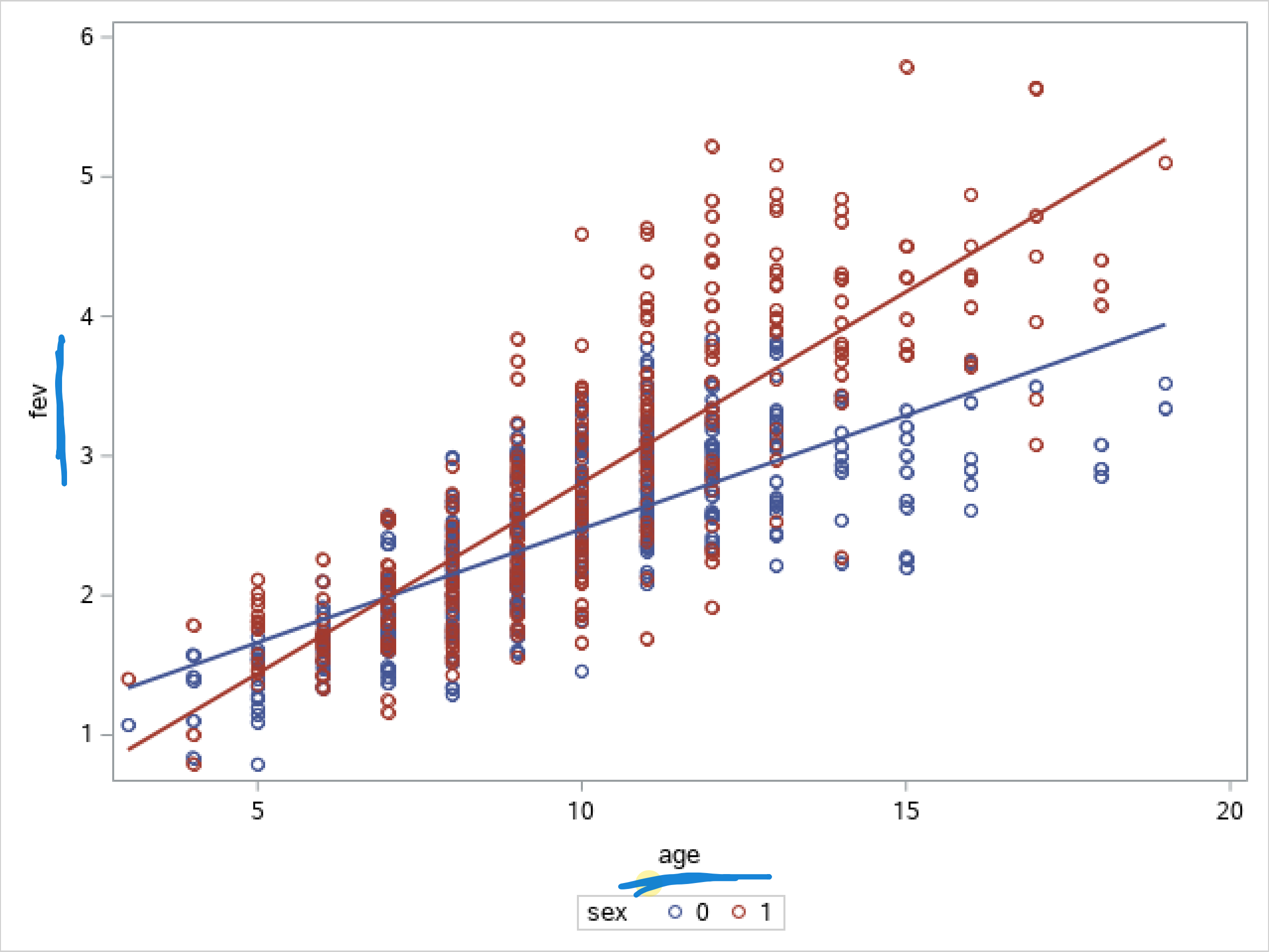
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.08680796	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.78154058	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	4.75442138	Pr > A-Sq	<0.005

Quantiles for Normal Distribution

	Quantile	
Percent	Observed	Estimated
1.0	1.25300	0.47773
5.0	1.52700	1.16168
10.0	1.66500	1.52628
25.0	2.00700	2.13553

Quantiles for Normal Distribution

	Quantile	
Percent	Observed	Estimated
50.0	2.60600	2.81245
75.0	3.53950	3.48936
90.0	4.28400	4.09861
95.0	4.63700	4.46322
99.0	5.22400	5.14716



The REG Procedure
Model: MODEL1
Dependent Variable: fev fev

Number of Observations Read	654

$\hat{\beta}_1 = 0.16273$

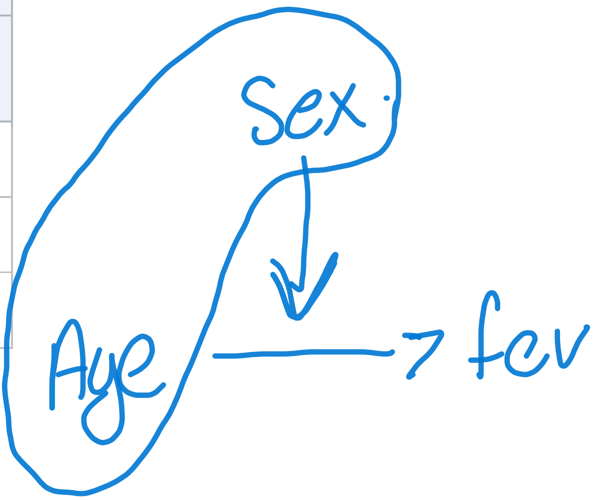
$\hat{\beta}_1 + \hat{\beta}_3 = 0.16273 + 0.11075 = 0.27348$

Number of Observations Used	654

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	315.41042	105.13681	389.37	<.0001
Error	650	175.50942	0.27001		
Corrected Total	653	490.91984			

Root MSE	0.51963	R-Square	0.6425
Dependent Mean	2.63678	Adj R-Sq	0.6408
Coeff Var	19.70696		



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.84947	0.10220	8.31	<.0001
age	age	1	0.16273	0.00995	16.35	<.0001
sex	sex	1	-0.77587	0.14275	-5.44	<.0001
AgebySex		1	0.11075	0.01379	8.03	<.0001

$< \alpha = 0.05$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_3} x_1 x_2 + \varepsilon.$$

APPENDIX:

SAS Code

```
proc import datafile='/home/ddawoud0/sasuser.v94/FEV.xls' DBMS=xls out=FEV;

run;


/*Investigate association between FEV and Smoke*/


proc sort data=FEV;

by smoke;

run;


proc tabulate data=FEV;

class smoke;

var fev;

table (smoke ALL), fev*(MEAN STD);

run;


proc boxplot data=FEV;

plot fev*smoke;

run;
```

when sex=0

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

when sex=1

$$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \varepsilon$$

$\hat{\beta}_1 = 0.16273 \Rightarrow$ for every 1 unit increase in age the average response increased by 0.16273 while holding sex fixed.

```
proc tabulate data=FEV;

class smoke;

var age;

table (smoke ALL), age*(MEAN STD);

run;


proc boxplot data=FEV;

plot age*smoke;

run;


proc univariate data=FEV noprint;

histogram fev/ normal vscale=proportion;

by smoke;

run;


proc sgplot data=FEV;

scatter y=fev x=age / group=smoke;

reg y=fev x=age / group=smoke;

run;
```

$\hat{\beta}_2 = -0.77587$

=0 as we go from sex=0 to sex=1 the average response decrease by 0.78 units while keeping age fixed.

$\hat{\beta}_3 = 0.11075$ added change on the average response as age increase by 1 unit when we go from sex=0 to sex=1.

```
/*Determine is smoking is a confounder*/
proc reg data=FEV plots=none;
model fev= age smoke; /*smoke is statistically significant i.e is a risk factor of FEV*/
run;

proc reg data=FEV plots=none;
model age = smoke; /*smoke is associated with age, however is smoking a direct consequence of age?*/
run;


/*Investigate associations between FEV and Sex*/
proc sort data=FEV;
by sex;
run;

proc tabulate data=FEV;
class sex;
var fev;
table (sex ALL), fev*(MEAN STD);
```

```
run;

proc boxplot data=FEV;
plot fev*sex;
run;

proc univariate data=FEV noprint;
histogram fev/ normal vscale=proportion;
by sex;
run;

proc sgplot data=FEV;
scatter y=fev x=age / group=sex;
reg y=fev x=age / group=sex;
run;

/*Determine if Sex is an effect modifier*/
data FEV1;
set FEV;
AgebySex = age*sex;
```

```
run;


proc reg data=FEV1 plots=none;

model fev= age sex AgebySex; /*add interaction term*/

run;
```