

# Smoking And Cancer

## 1 Abstract

Government statisticians in England conducted a study of the relationship between smoking and lung cancer. The data concern 25 occupational groups and are condensed from data on thousands of individual men. The explanatory variable is the number of cigarettes smoked per day by men in each occupation relative to the number smoked by all men of the same age. This smoking ratio is 100 if men in an occupation are exactly average in their smoking, it is below 100 if they smoke less than average, and above 100 if they smoke more than average. The response variable is the standardized mortality ratio for deaths from lung cancer. It is also measured relative to the entire population of men of the same ages as those studied, and is greater or less than 100 when there are more or fewer deaths from lung cancer than would be expected based on the experience of all English men.

## 2 Reference

*Occupational Mortality: The Registrar General's Decennial Supplement for England and Wales, 1970-1972*, Her Majesty's Stationary Office, London, 1978.

## 3 Description

Data summarizes a study of men in 25 occupational groups in England. Two indices are presented for each occupational group. The smoking index is the ratio of the average number of cigarettes smoked per day by men in the particular occupational group to the average number of cigarettes smoked per day by all men. The mortality index is the ratio of the rate of deaths from lung cancer among men in the particular occupational group to the rate of deaths from lung cancer among all men.

## 4 Number of cases

25

## 5 Variable Names

1. Occupational\_Group: Occupational Group
2. Smoking: Smoking index (100 = average)
3. Mortality: Lung cancer mortality index (100 = average)

## 6 The Data

Occupational_Group	Smoking	Mortality
Farmers, foresters, and fishermen	77	84
Miners and quarrymen	137	116
Gas, coke, and chemical makers	117	123
Glass and ceramics makers	94	128
Furnace, forge, foundry, and rolling mill workers	116	155
Electrical and electronics workers	102	101
Engineering and allied trades	111	118
Woodworkers	93	113
Leather workers	88	104
Textile workers	102	88
Clothing workers	91	104
Food, drink, and tobacco workers	104	129
Paper and printing workers	107	86
Makers of other products	112	96
Construction workers	113	144
Painters and decorators	110	139
Drivers of stationary engines, cranes, etc.	125	113
Laborers not included elsewhere	133	146
Transport and communications workers	115	128
Warehousemen, storekeepers, packers, and bottlers	105	115
Clerical workers	87	79
Sales workers	91	85
Service, sport, and recreation workers	100	120
Administration and managers	76	60
Professionals, technical workers, and artists	66	51

Analyze this data. Calculate summary statistics (including graphs) for both variables individually. Make both a least squares regression model and a median-median model. Check for outliers and influential points. Do a scatterplot and also a residual plot. Calculate both  $r$  and  $r^2$  to help you assess how good the lsrl model is. Suppose a particular occupation had a smoking index of 90. What would you predict for their lung cancer mortality index?