Elementary Statistics Lecture 3
Association: Contingency, Correlation and Regression

Chong Ma

Department of Statistics
University of South Carolina
chongm@email.sc.edu

Outline

Association

Association Between Two Variables

An **association** exists between two variables if particular values for one variable are more likely to occur with certain values of the other variable.

Response Variable and Explanatory Variable

The **response variable** is the outcome variable on which comparisons are made for different values of the explanatory variable.

Examples

- survival status is the response variable and smoking status is the explanatory variable.

Association

Association Between Two Variables

An **association** exists between two variables if particular values for one variable are more likely to occur with certain values of the other variable.

Response Variable and Explanatory Variable

The **response variable** is the outcome variable on which comparisons are made for different values of the explanatory variable.

Examples

- survival status is the response variable and smoking status is the explanatory variable.
- CO_2 is the response variable and country's amount of gasoline use for automobiles is the explanatory variable.

Association

Association Between Two Variables

An **association** exists between two variables if particular values for one variable are more likely to occur with certain values of the other variable.

Response Variable and Explanatory Variable

The **response variable** is the outcome variable on which comparisons are made for different values of the explanatory variable.

Examples

- survival status is the response variable and smoking status is the explanatory variable.
- CO_2 is the response variable and country's amount of gasoline use for automobiles is the explanatory variable.
- GPA is the response variable and the number of hours a week spent studying is the explanatory variable.

Outline

Association Between Two Categorical Variables

	Pesticide Status		
Food Type	**Present**	**Not Present**	**Total**
Organic	29	98	**127**
Conventional	19,485	7,086	**26,571**
Total	**19,514**	**7,184**	**26,698**

Table 1: Frequencies for food type and pesticide status.

Association Between Two Categorical Variables

	Pesticide Status		
Food Type	**Present**	**Not Present**	**Total**
Organic	29	98	**127**
Conventional	19,485	7,086	**26,571**
Total	**19,514**	**7,184**	**26,698**

Table 1: Frequencies for food type and pesticide status.

	Pesticide Status		
Food Type	**Present**	**Not Present**	**Total**
Organic	0.228	0.772	**1**
Conventional	0.733	0.267	**1**
Total	**0.734**	**0.266**	**1**

Table 2: Conditional proportions on pesticide status for two food types.

Proportions

Conditional Proportion $\hat{P}(\text{present}|\text{Organic}) = \frac{29}{127} = 0.228$

$\hat{P}(\text{present}|\text{Conventional}) = \frac{19{,}514}{26{,}698} = 0.733$

Marginal Proportion $\hat{P}(\text{present}) = \frac{19{,}485}{26{,}698} = 0.734$

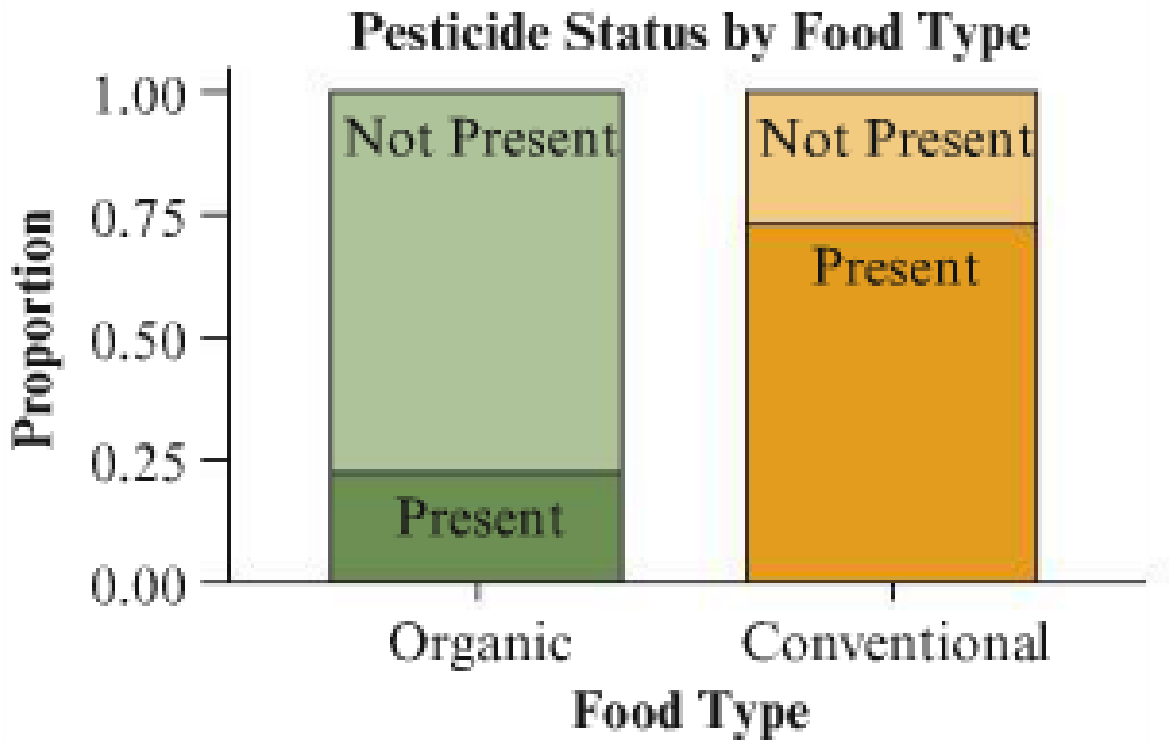

Figure 1: Stacked Bar Graph. It stacks the proportion of pesticides present and not present on top of each other.

Independence and Dependence(Association)

Naive approach

Association exists if the proportion with pesticide present had a "big" difference between the food types and vice verse. How big is big?

Independence and Dependence(Association)

Naive approach

Association exists if the proportion with pesticide present had a "big" difference between the food types and vice verse. How big is big?

Chi-Squared Test

$$\chi^2_{\text{df}} = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

Where

$$\text{Expected cell count} = \frac{\text{Row total} \times \text{Column Total}}{\text{Total sample size}}$$

and

$$\text{df} = (r-1) \times (c-1)$$

Independence and Dependence(Association)

Naive approach

Association exists if the proportion with pesticide present had a "big" difference between the food types and vice verse. How big is big?

Chi-Squared Test

$$\chi^2_{\text{df}} = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

Where

$$\text{Expected cell count} = \frac{\text{Row total} \times \text{Column Total}}{\text{Total sample size}}$$

and

$$\text{df} = (r-1) \times (c-1)$$

Rule of Thumb

The larger the χ^2(smaller the p-value), the more statistical evidence for existence of association between the categorical variables.

Chi-Square(χ^2) Distribution

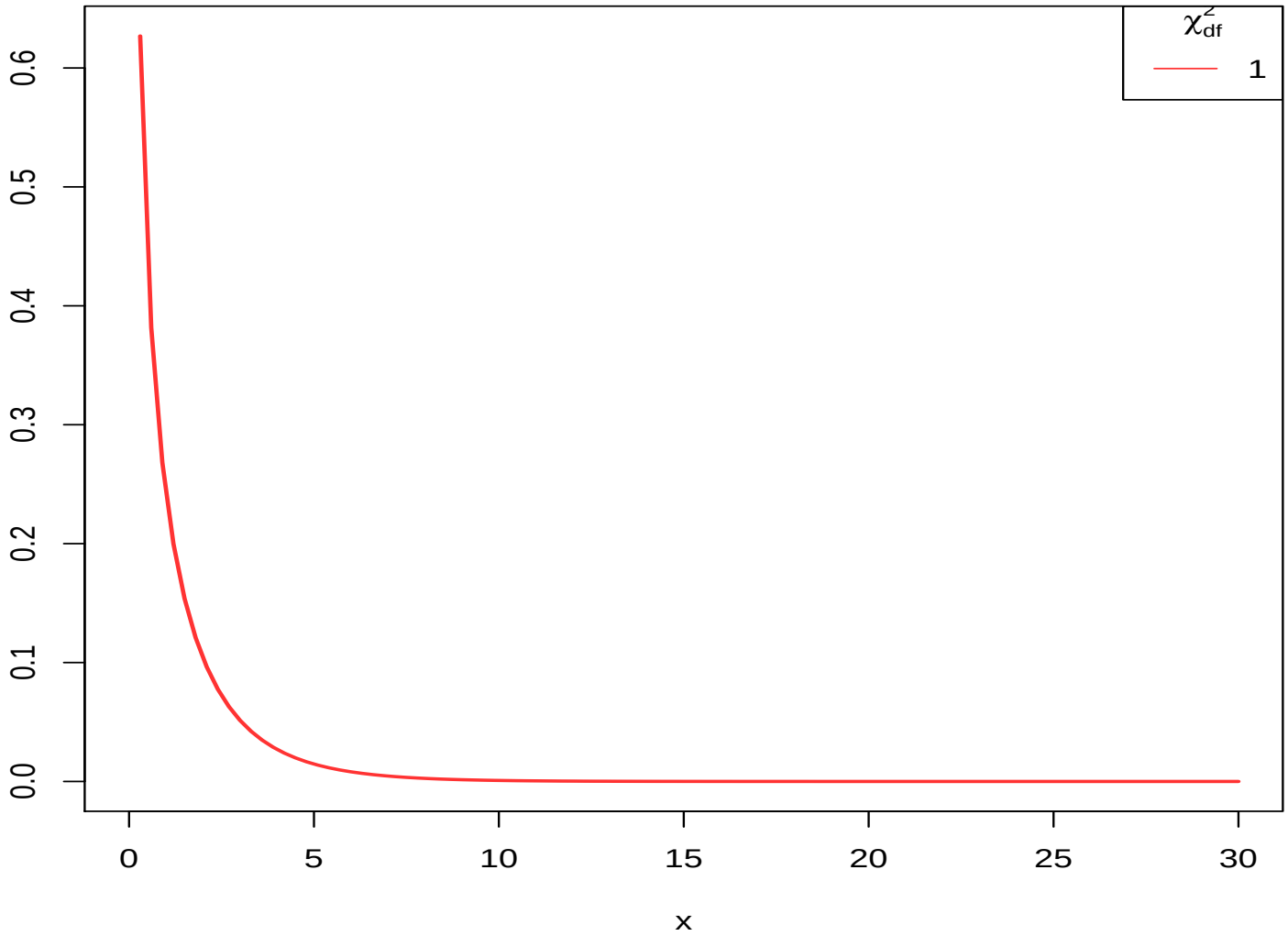

Figure 2: χ^2 distribution with degrees of freedom 1.

Chi-Square(χ^2) Distribution

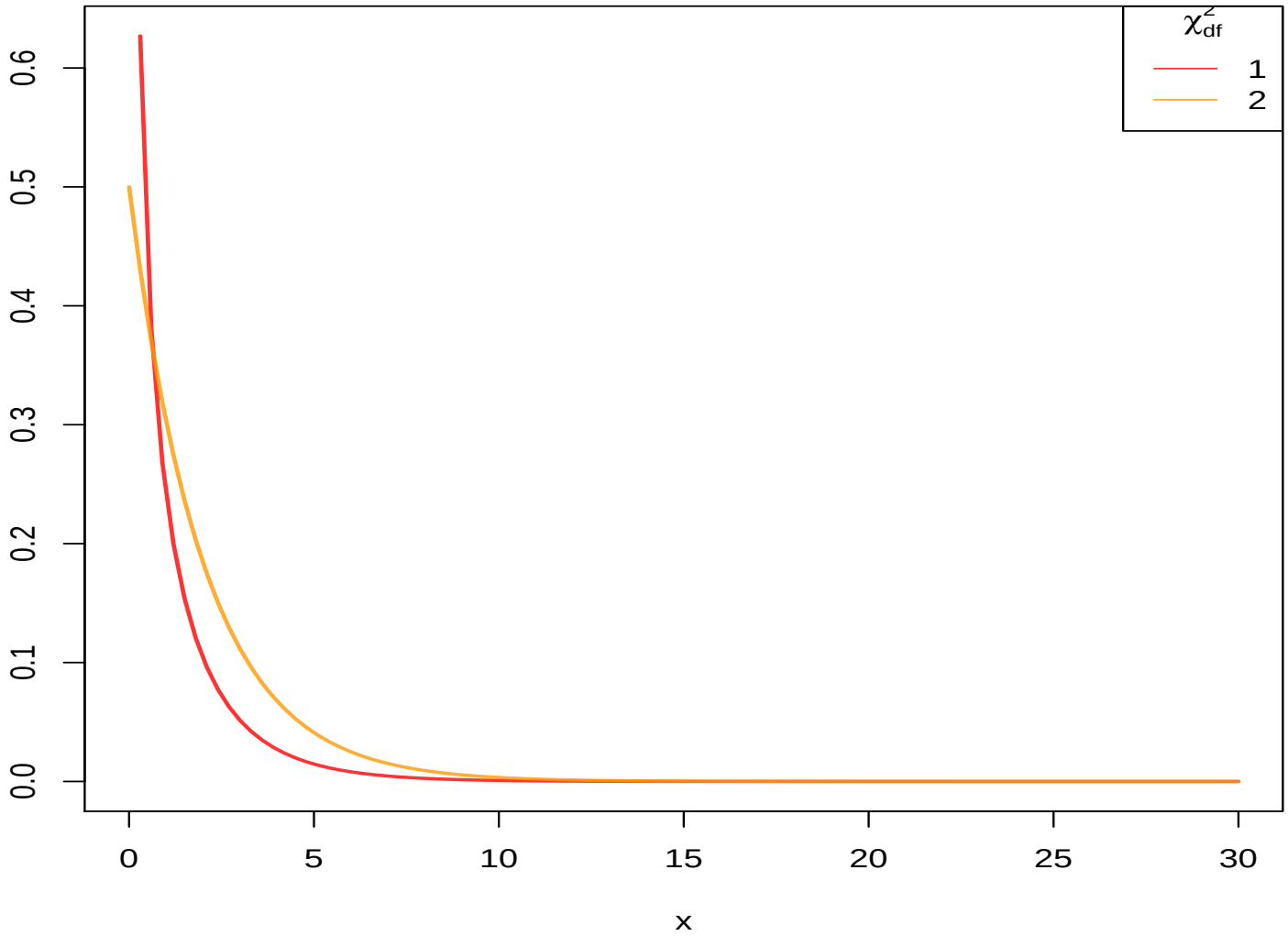

Figure 3: χ^2 distribution with degrees of freedom 1 and 2, respectively.

Chi-Square(χ^2) Distribution

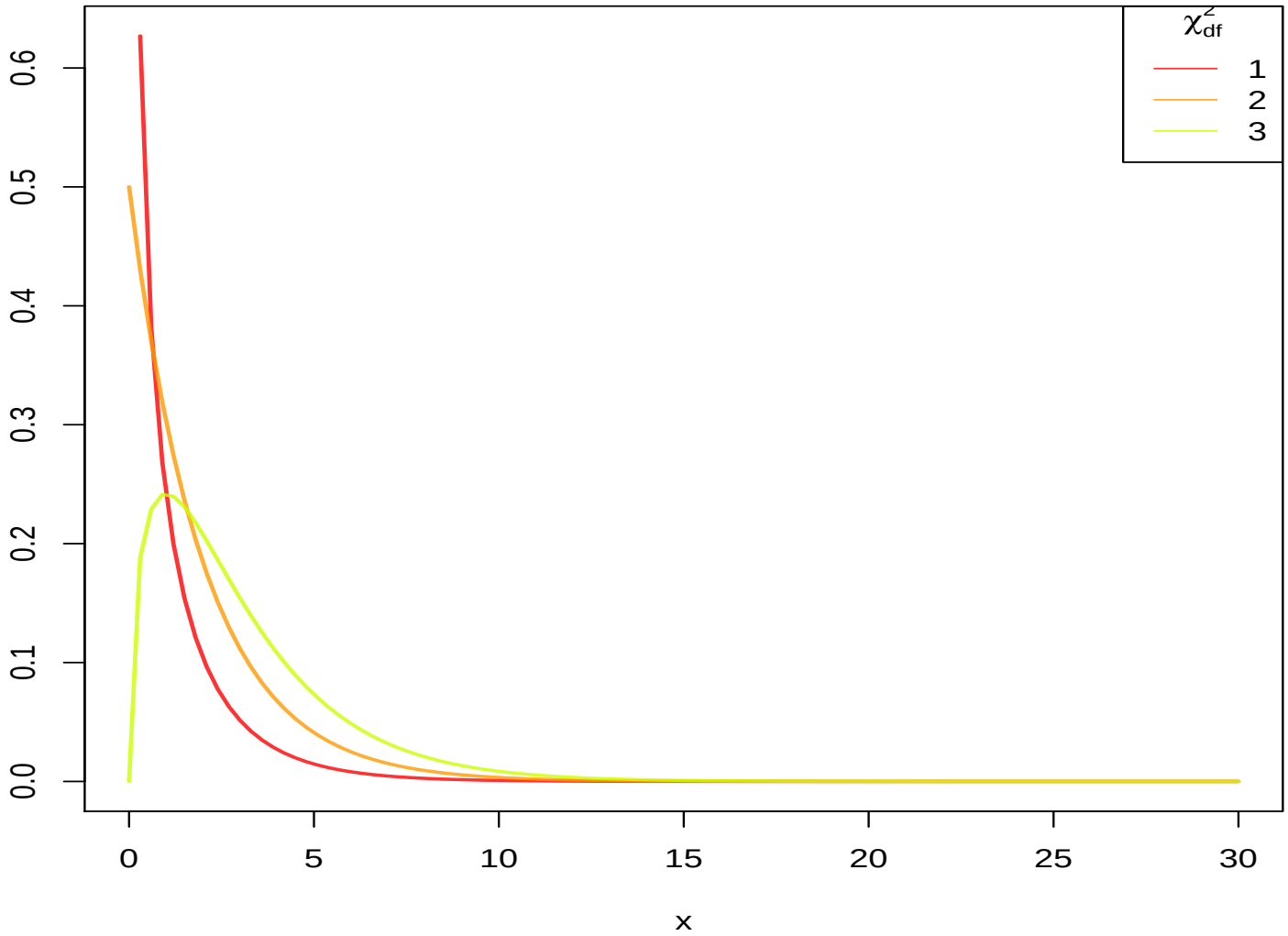

Figure 4: χ^2 distribution with degrees of freedom 1, 2 and 3, respectively.

Chi-Square(χ^2) Distribution

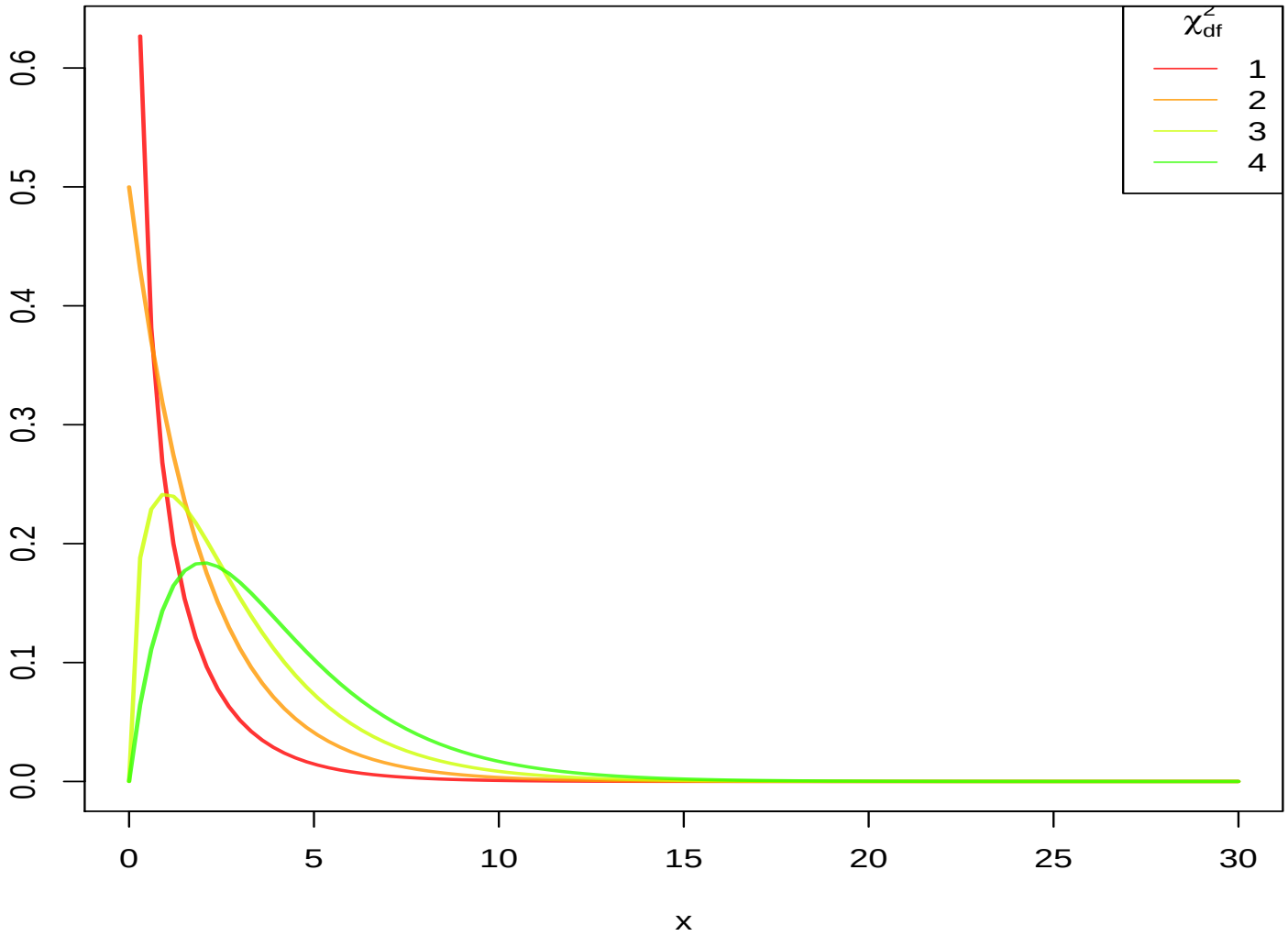

Figure 5: χ^2 distribution with degrees of freedom 1, 2, 3 and 4, respectively.

Chi-Square(χ^2) Distribution

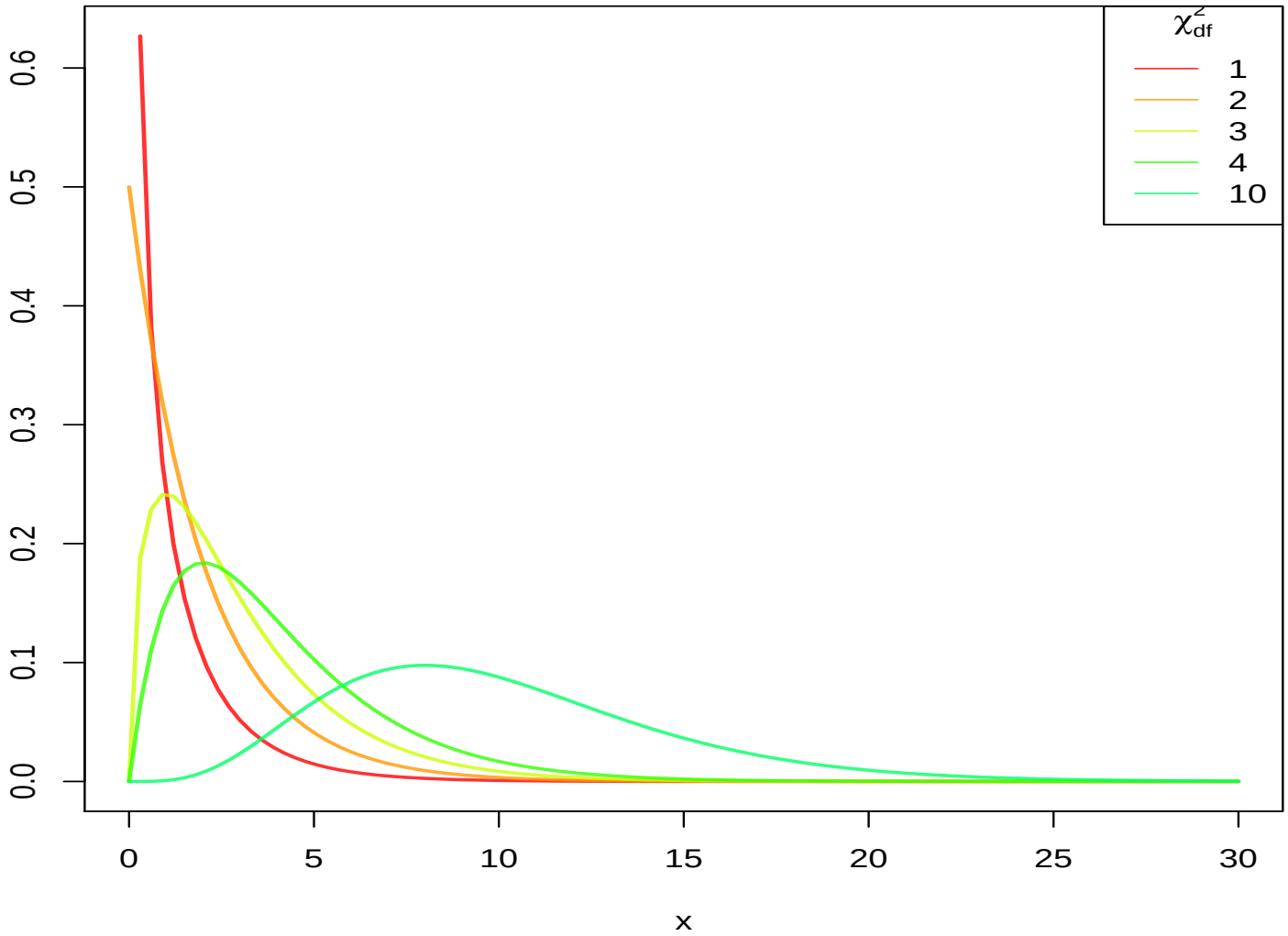

Figure 6: χ^2 distribution with degrees of freedom 1, 2, 3, 4 and 5, respectively.

Pesticide in Organic Foods

	Pesticide Status		
Food Type	**Present**	**Not Present**	**Total**
Organic	29	98	**127**
	93	34	Expected Count
	43.89	119.21	Contribution to χ^2
Conventional	19,485	7,086	**26,571**
	19,421	7,150	Expected Count
	0.21	0.57	Contribution to χ^2
Total	**19,514**	**7,184**	**26,698**

Table 3: Chi-squared Test Table of independence(not association) of food type and pesticide status.

```
Pearson's Chi-squared test with Yates' continuity correction
data:  pes
X-squared = 161.32, df = 1, p-value < 2.2e-16
```

Strength Of The Association

Difference of Proportions

	Stress			Depression		
Gender	**Yes**	**No**	**Total**	**Yes**	**No**	**Total**
Female	44%	56%	100%	11%	89%	100%
Male	20%	80%	100%	6%	94%	100%

Table 4: Survey of college freshmen on feeling frequently stressed and feeling frequently depressed at UCLA in 2013

Strength Of The Association

Difference of Proportions

	Stress			Depression		
Gender	**Yes**	**No**	**Total**	**Yes**	**No**	**Total**
Female	44%	56%	100%	11%	89%	100%
Male	20%	80%	100%	6%	94%	100%

Table 4: Survey of college freshmen on feeling frequently stressed and feeling frequently depressed at UCLA in 2013

$$\hat{P}(\text{Stress}|\text{Female}) - \hat{P}(\text{Stress}|\text{Male}) = 44\% - 20\% = 22\%$$

$$\hat{P}(\text{Depression}|\text{Female}) - \hat{P}(\text{Depression}|\text{Male}) = 11\% - 6\% = 5\%$$

Insights

There is evidence of a greater difference between females and males on their feelings about stress than on depression.

The Ratio of Proportions: Relative Risk

	Developed Flu		
Group	**Yes**	**No**	**Total**
Vaccine	26	3874	3900
Control	70	3830	3900

Table 5: Results from a Vaccine Efficacy Study. Subjects are either vaccinated with a cell-derived flu vaccine or a placebo(Control Group)

The Ratio of Proportions: Relative Risk

	Developed Flu		
Group	**Yes**	**No**	**Total**
Vaccine	26	3874	3900
Control	70	3830	3900

Table 5: Results from a Vaccine Efficacy Study. Subjects are either vaccinated with a cell-derived flu vaccine or a placebo(Control Group)

Difference of proportions

$$\begin{aligned}\hat{P}(\text{Yes}|\text{Control}) - \hat{P}(\text{Yes}|\text{Vaccine}) &= \frac{70}{3900} - \frac{26}{3900} \\ &= 0.0179 - 0.0067 \\ &= 0.0112\end{aligned}$$

Insight

The difference of proportions itself is in very small magnitude, which might mistakenly lead us to a conclusion of independence between the developed flu status and vaccine status.

The Ratio of Proportions: Relative Risk

	Developed Flu		
Group	**Yes**	**No**	**Total**
Vaccine	0.0067	0.9933	1
Control	0.0179	0.9821	1

Table 6: Conditional proportions of flu status for vaccinated and unvaccinated (control) groups.

Ratio of proportions(relative risk)

$$\text{Relative Risk} = \frac{\hat{P}(\text{Yes}|\text{Control})}{\hat{P}(\text{Yes}|\text{Vaccine})} = \frac{0.0179}{0.0067} = 2.67$$

Insight

Unvaccined subjects are 2.67 times more likely to develop the flu, where 2.67 times more likely refers to the relative risk.

Odds and Odds Ratio

In the table 6, denote $\hat{p}_1 = \hat{P}(\text{Yes}|\text{Vaccine})$ and $\hat{p}_2 = \hat{P}(\text{Yes}|\text{Control})$.

Odds the ratio of the two conditional proportions in each row.

$$\text{Odds}_1 = \hat{p}_1/(1-\hat{p}_1) = 0.0067/0.9933 = 0.0067 \approx 1/150$$
$$\text{Odds}_2 = \hat{p}_2/(1-\hat{p}_2) = 0.0179/0.9821 = 0.0182 \approx 1/50$$

Odds Ratio The ratio of the two odds

$$\frac{\hat{p}_2/(1-\hat{p}_2)}{\hat{p}_1/(1-\hat{p}_1)} = \frac{0.0182}{0.0067} = 2.72$$

Insights

- The odds compare the proportion of subjects developing flu to those who don't in the vaccinated group and unvaccinated group.
- The odds ratio tells us the difference in times compared one odds to the other.

Odds and Odds Ratio

Interpretation

- In the control group, for every one subject developing the flu, 50 are not developing it.
- In the Vaccinated group, for every one subject developing the flu, 150 are not developing it.
- The odds of developing fle in the control group are 2.7 (about 3) times the odds in the vaccinated group.

Properties of the Odds Ratios

- Could be any nonnegative value.
- odds ratio equals 1 $\Leftrightarrow$ independence
- values farther from 1 represents stronger association.

Smoking and Health

A survey of $1,314$ women in the United Kingdom that asked each woman whether she was a smoker. Twenty years later, a follow-up survey observed whether each woman was dead or still alive.

	Survival Status		
Smoker	**Dead**	**Alive**	**total**
Yes	139	443	582
No	230	502	732
Total	369	945	1314

Table 7: Smoking Status and 20-Year Survival in Women

Question

- What's the response variable? Explanatory variable?
- Construct a contingency table that shows the conditional proportion of survival status given smoker status.

Smoking and Health

	Age Group							
	18-34 Survival?		35-54 Survival?		55-64 Survival?		65+ Survival?	
Smoker	**Dead**	**Alive**	**Dead**	**Alive**	**Dead**	**Alive**	**Dead**	**Alive**
Yes	5	174	41	198	51	64	42	7
No	6	213	19	180	40	81	165	28

Table 8: Four contingency tables relating smoking status and survival status for these $1,134$ women sperated into four age groups.

Example

Construct the contingency table that shows conditional proportions of deaths for smokers and nonsmokers by age.

Smoking and Health

	Age Group			
Smoker	18-34	35-54	55-64	65+
Yes	2.8%	17.2%	44.3%	85.7%
No	2.7%	9.5%	33.1%	85.5%
Difference	0.1%	7.7%	11.2%	0.2%

Table 9: Conditional Proportion of deaths for smokers and nonsmokers by age.

Insights

When looking at the data separately by age group, we can see that smokers have lower survival rates than nonsmokers.

Outline

Association Between Two Quantitative Variables

- Interested in if the value of one variable X (response variable) can be reliably predicted by the value of the other variable Y (explanatory or predictor variable).
- Need examine the nature and strength of the relationship between two variables.
- Develop a statistical model to predict new values of the response variable using the predictor variable.

  simple linear straight line

  polynomial curve

Correlation

The **correaltion** summarizes the strength and direction of the linear (straight-line) association between two quantitative variables. Denoted by r, it takes values between -1 and $+1$.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$
$$= \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i}$$

Where n is the number of observations, $\bar{x}$ and $\bar{y}$ are means, and s_x and s_y are standard deviations for x and y.

Correlation: Direction and Strength

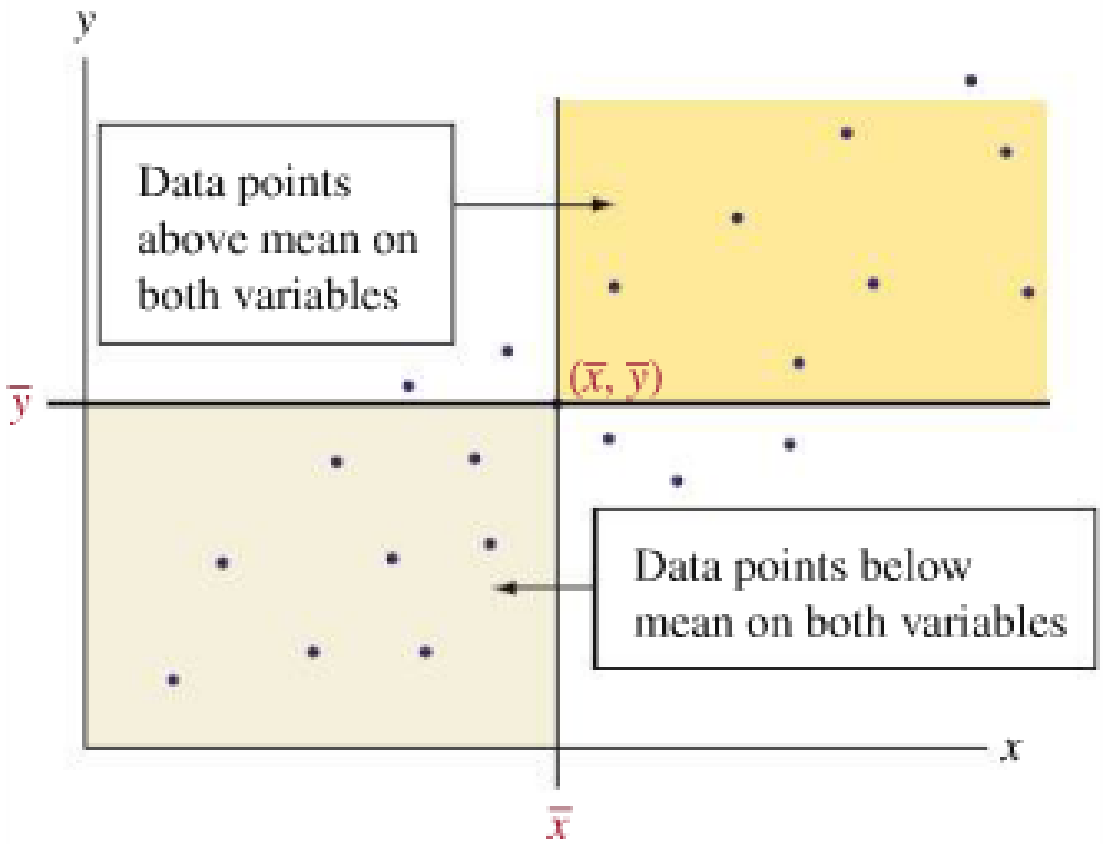

Figure 7: The data points in the upper-right and lower-left quadrants make a positive contribution to the correlation value of r, and it is opposite for the data points in the upper-left and the lower-right quadrants.

Correlation: Direction and Strength

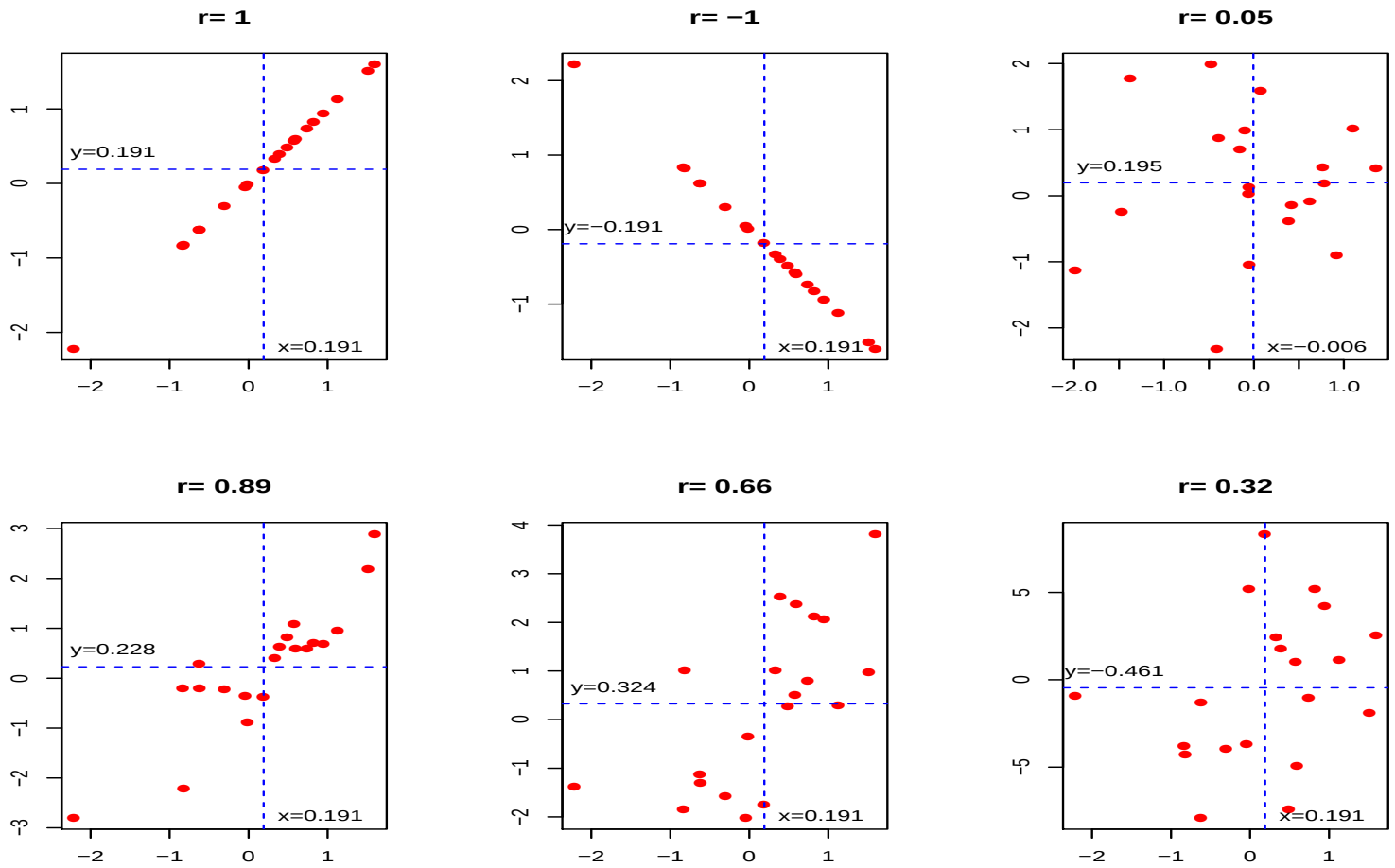

Figure 8: The closer the data points fall to a straight line, the closer the correlation r gets to ± 1, and the stronger the linear association between x and y.

Correlation: Direction and Strength

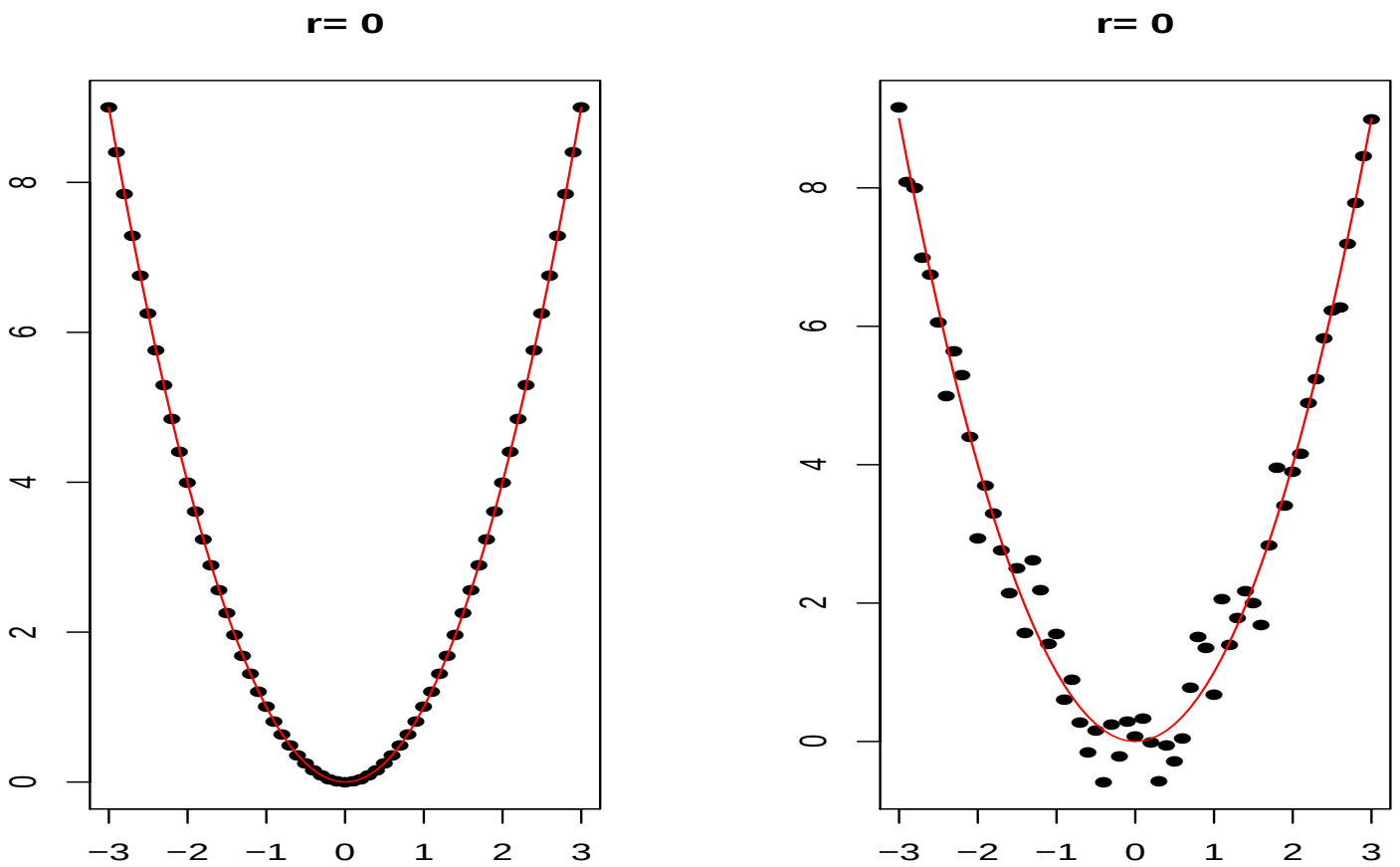

Figure 9: The correlation r poorly describes the association when the relationship is curved. Always construct a scatterplot to display a relationship between two quantitative variables.

Simple Linear Regression

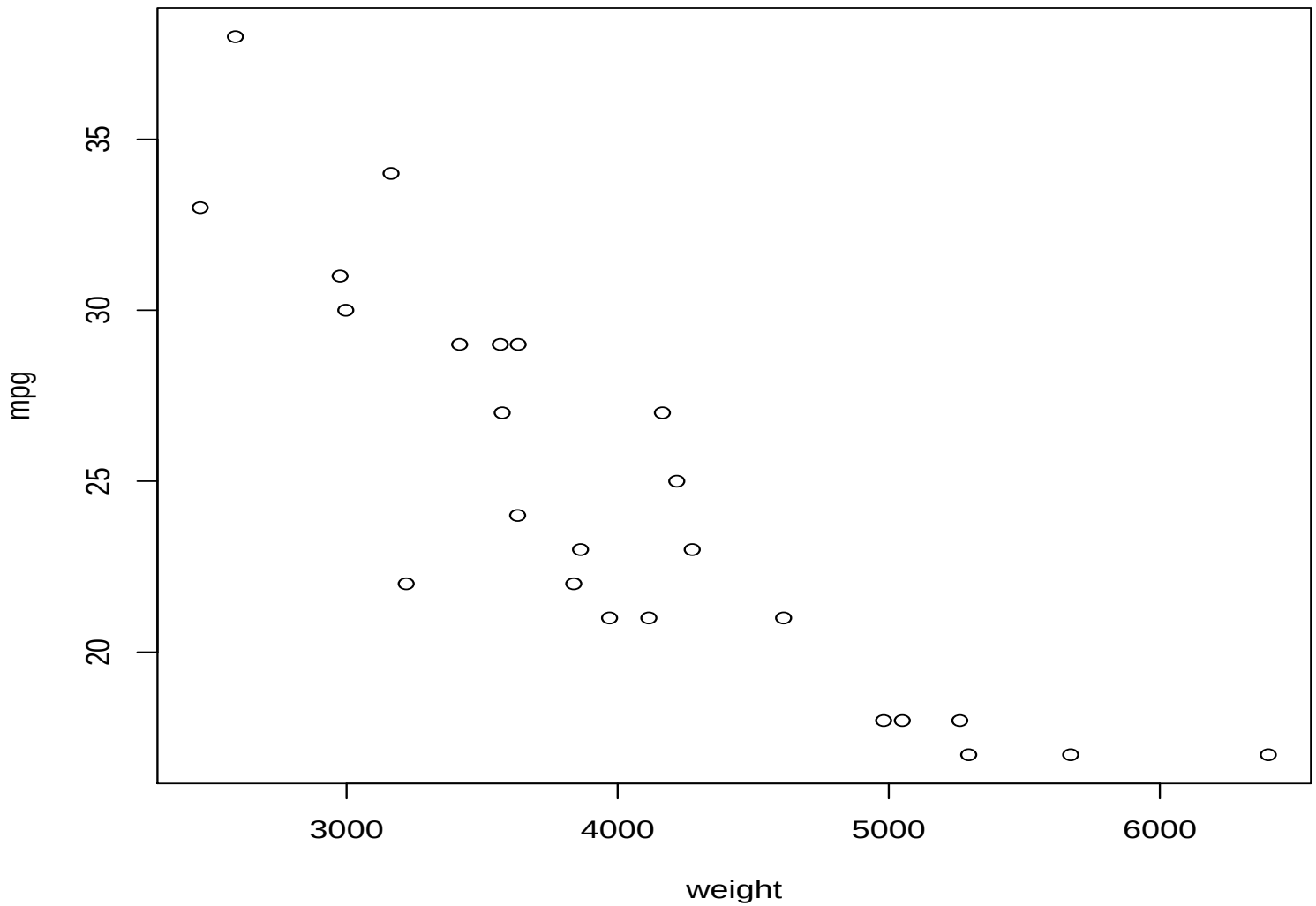

Figure 10: The scatterplot of x weights (in pounds) and y mileage (miles per gallon of gas) for 2004 model cars with automatic transmission.

Simple Linear Regression

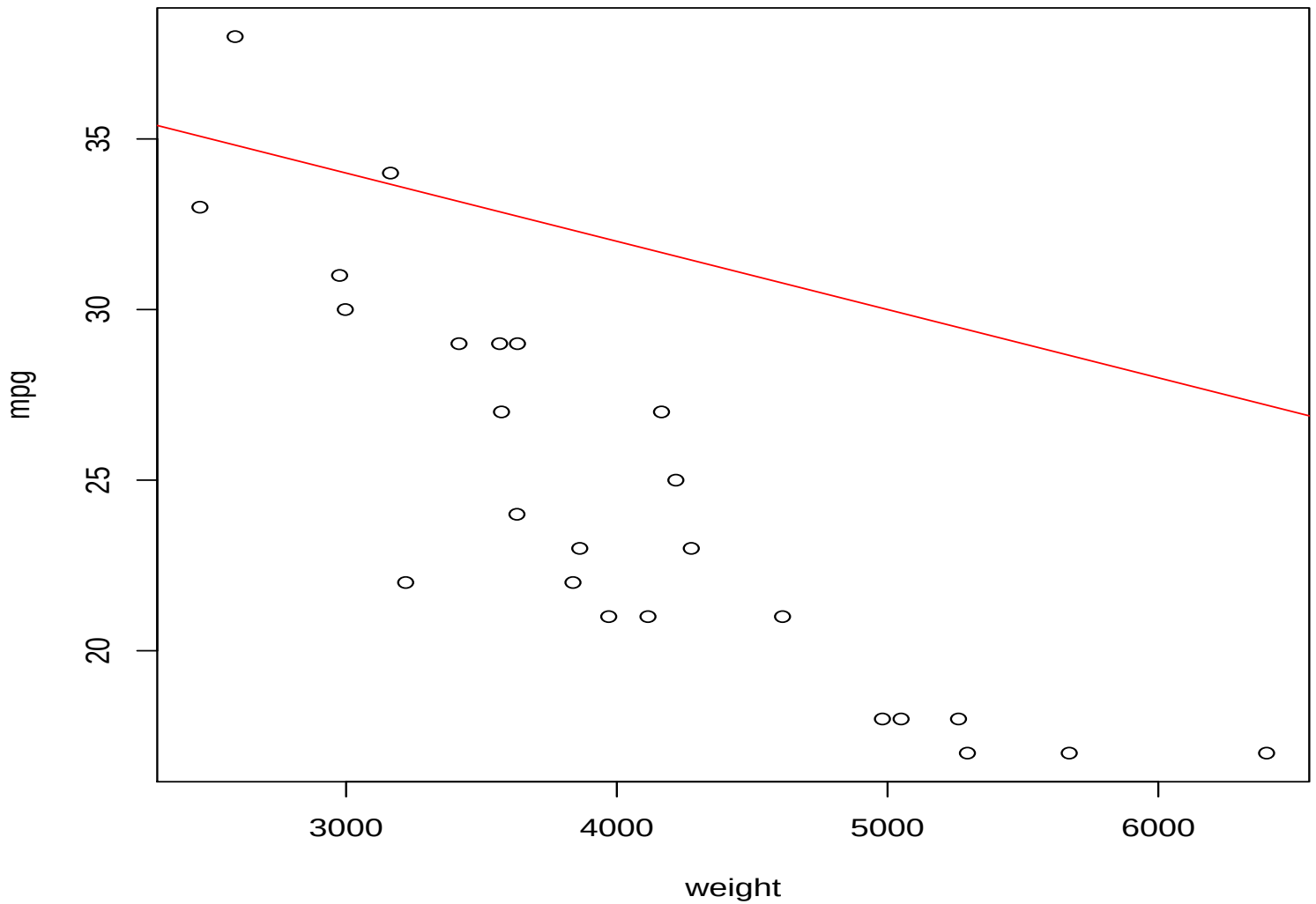

Figure 11: The scatterplot of x weights (in pounds) and y mileage superimposed with an arbitrary straight line.

Simple Linear Regression

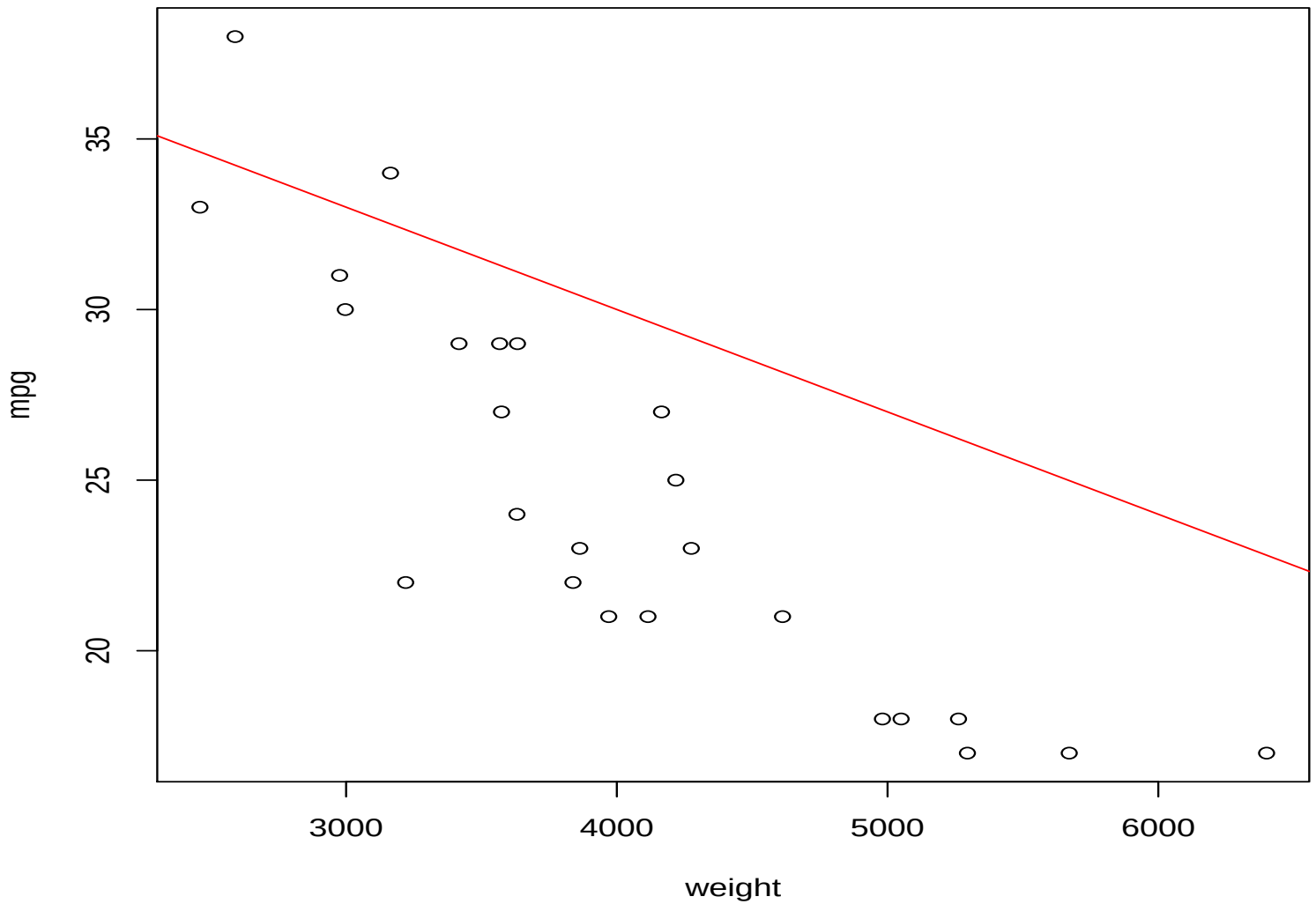

Figure 12: The scatterplot of x weights (in pounds) and y mileage superimposed with an arbitrary straight line.

Simple Linear Regression

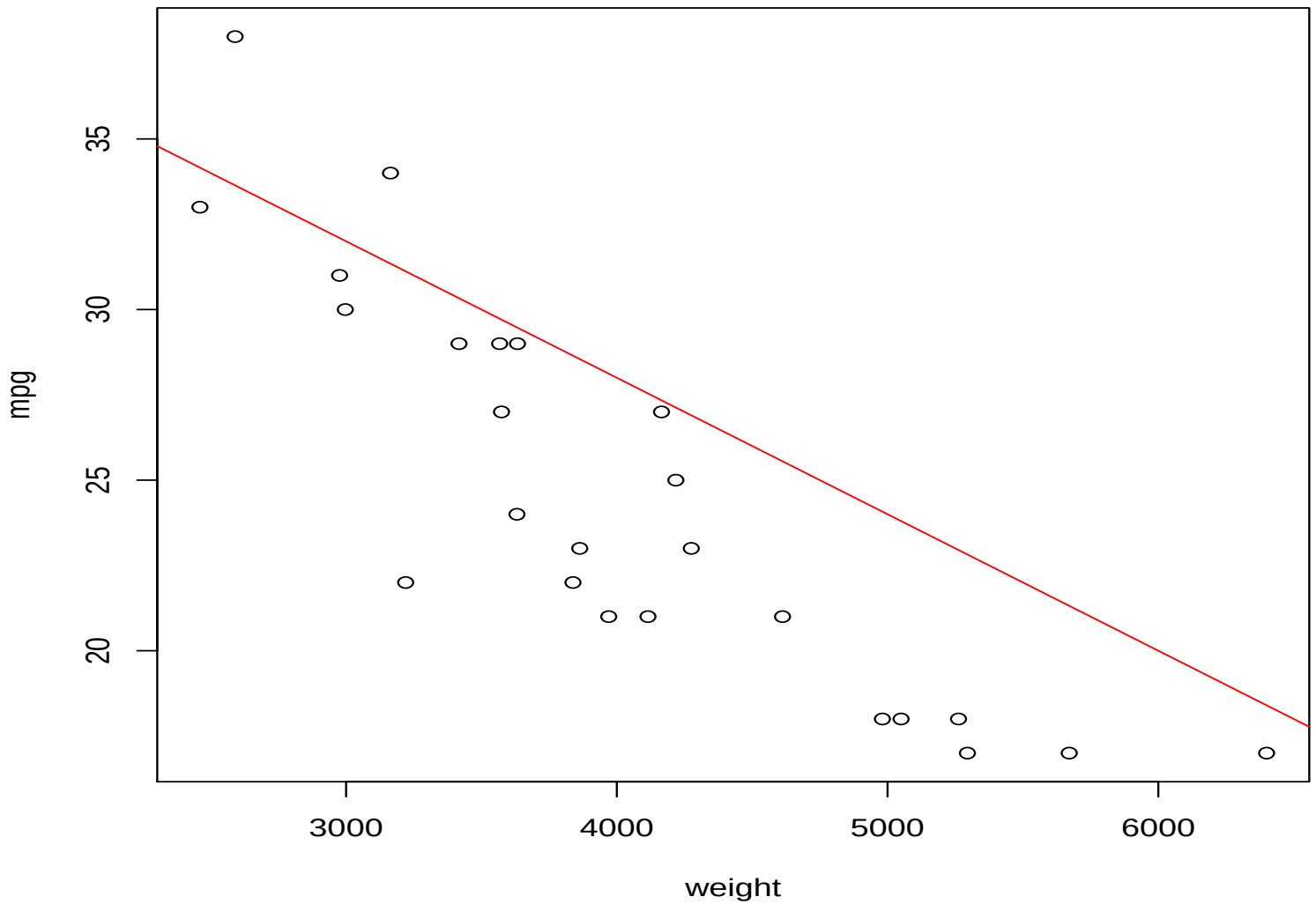

Figure 13: The scatterplot of x weights (in pounds) and y mileage superimposed with an arbitrary straight line.

Simple Linear Regression

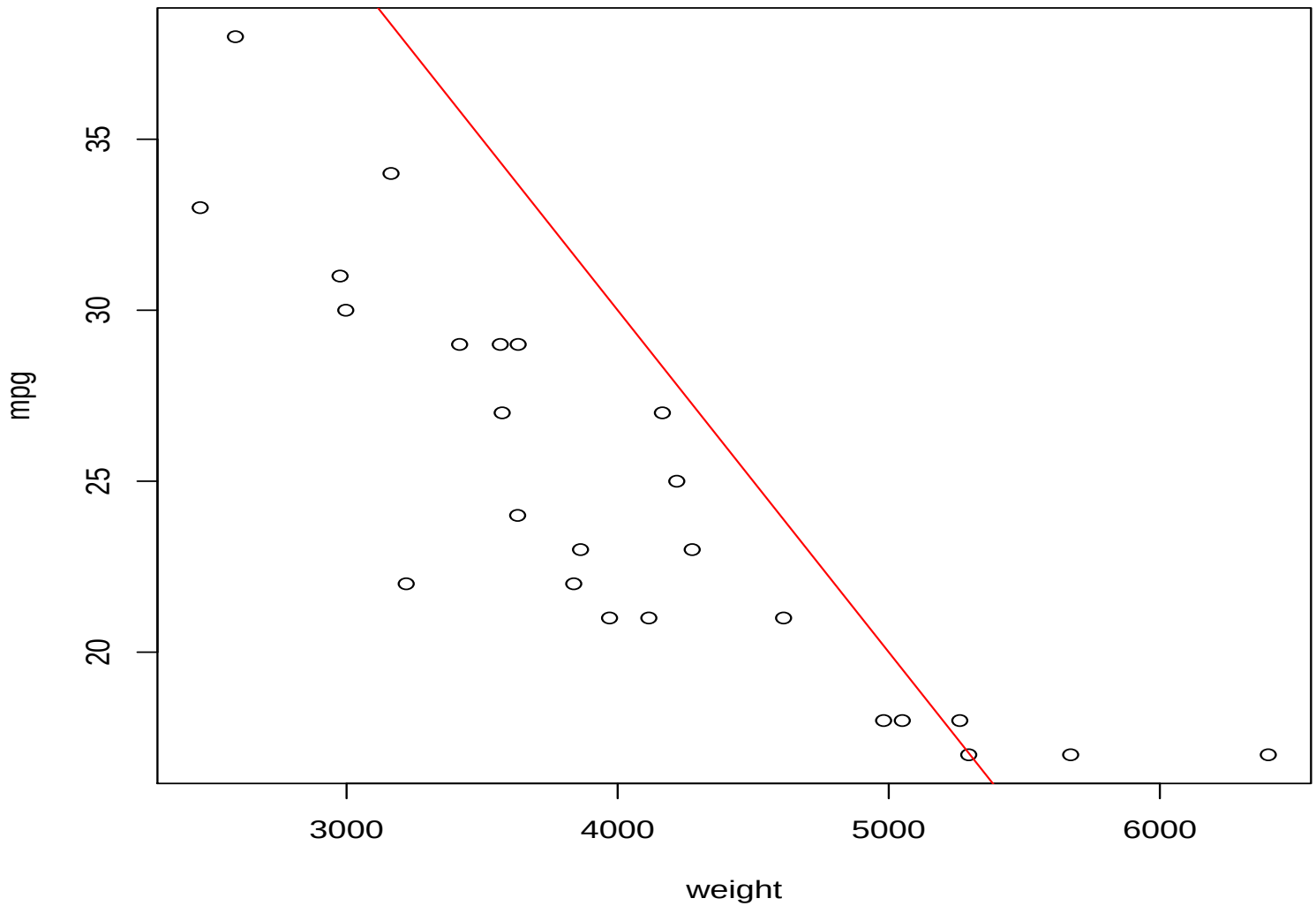

Figure 14: The scatterplot of x weights (in pounds) and y mileage superimposed with an arbitrary straight line.

Simple Linear Regression

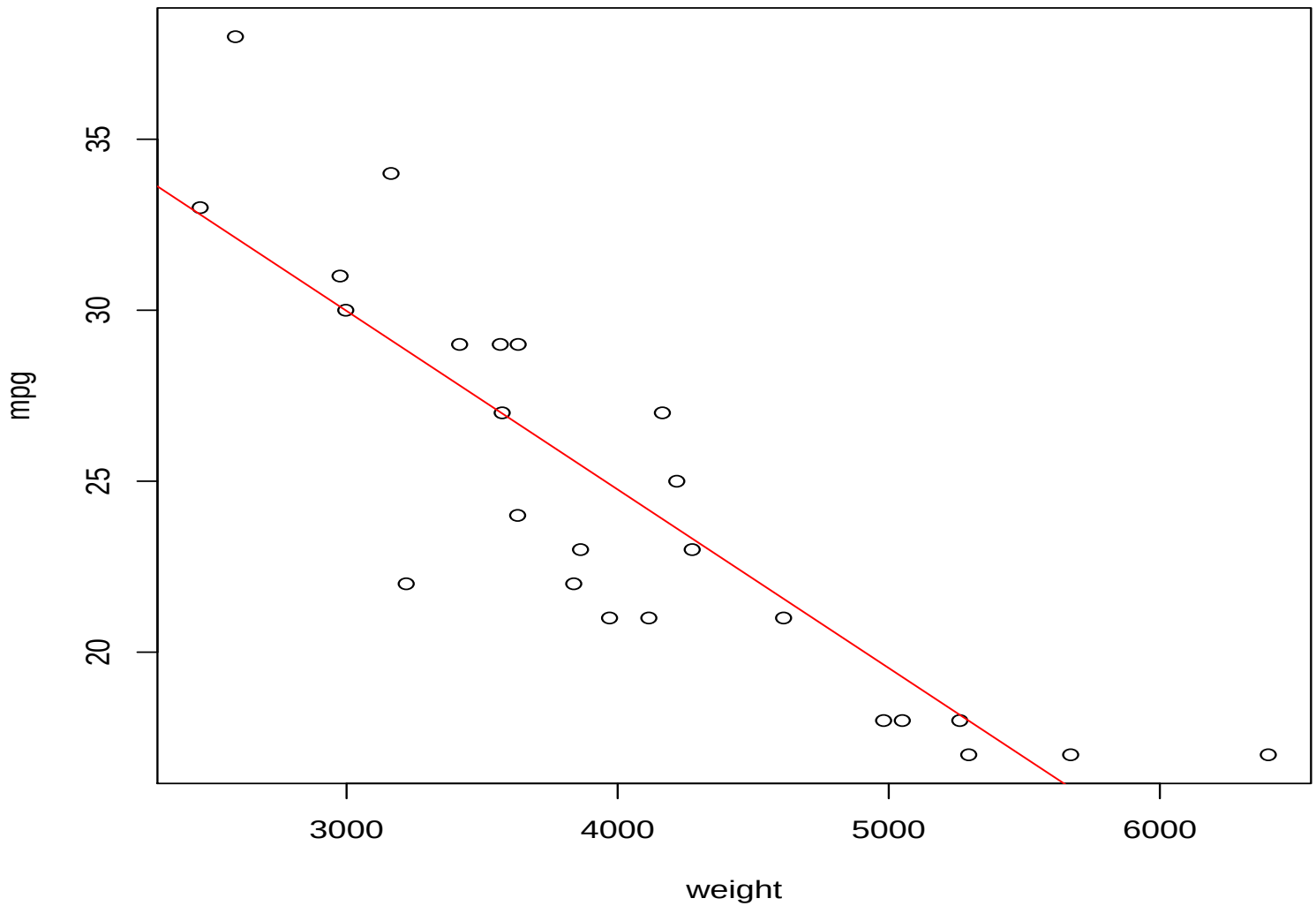

Figure 15: The scatterplot of x weights (in pounds) and y mileage superimposed with the straight line of best fit.

Least Squares Method

Assume that there is an underlying straight-line relationship between the predictor variable X and the response variable Y. This relationship can be written as a regression equation, denoted by

$$\mu_Y = \beta_0 + \beta_1 X$$

Where μ_Y denotes the population mean of Y for all the subjects at a particular value of X.

In practice, we estimate the regression equation using the sample data, and the estimated regression equation is denoted by

$$\hat{y} = b_0 + b_1 x$$

Least Squares Method

The least squared method guarantees the optimal line to fit through the data points by the sum of squared residuals as small as possible.

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Where the optimal line is called the regression line. $y_i - \hat{y}_i$ is the residual for the i^{th} subject with the observed response variable value y_i and the predictor variable value x_i and $\hat{y}_i = b_0 + b_1 x_i$.

Car Weights and MPGs

```
Call:
lm(formula = car$mileage ~ car$weight)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.6453593  2.6027584  17.537 8.30e-15 ***
car$weight  -0.0052220  0.0006271  -8.328 2.14e-08 ***

Residual standard error: 3.016 on 23 degrees of freedom
Multiple R-squared:  0.751,Adjusted R-squared:  0.7401
F-statistic: 69.35 on 1 and 23 DF,  p-value: 2.138e-08
```

The estimated regression equation is

$$\hat{y} = 45.65 - 0.005x$$

Car Weights and MPGs

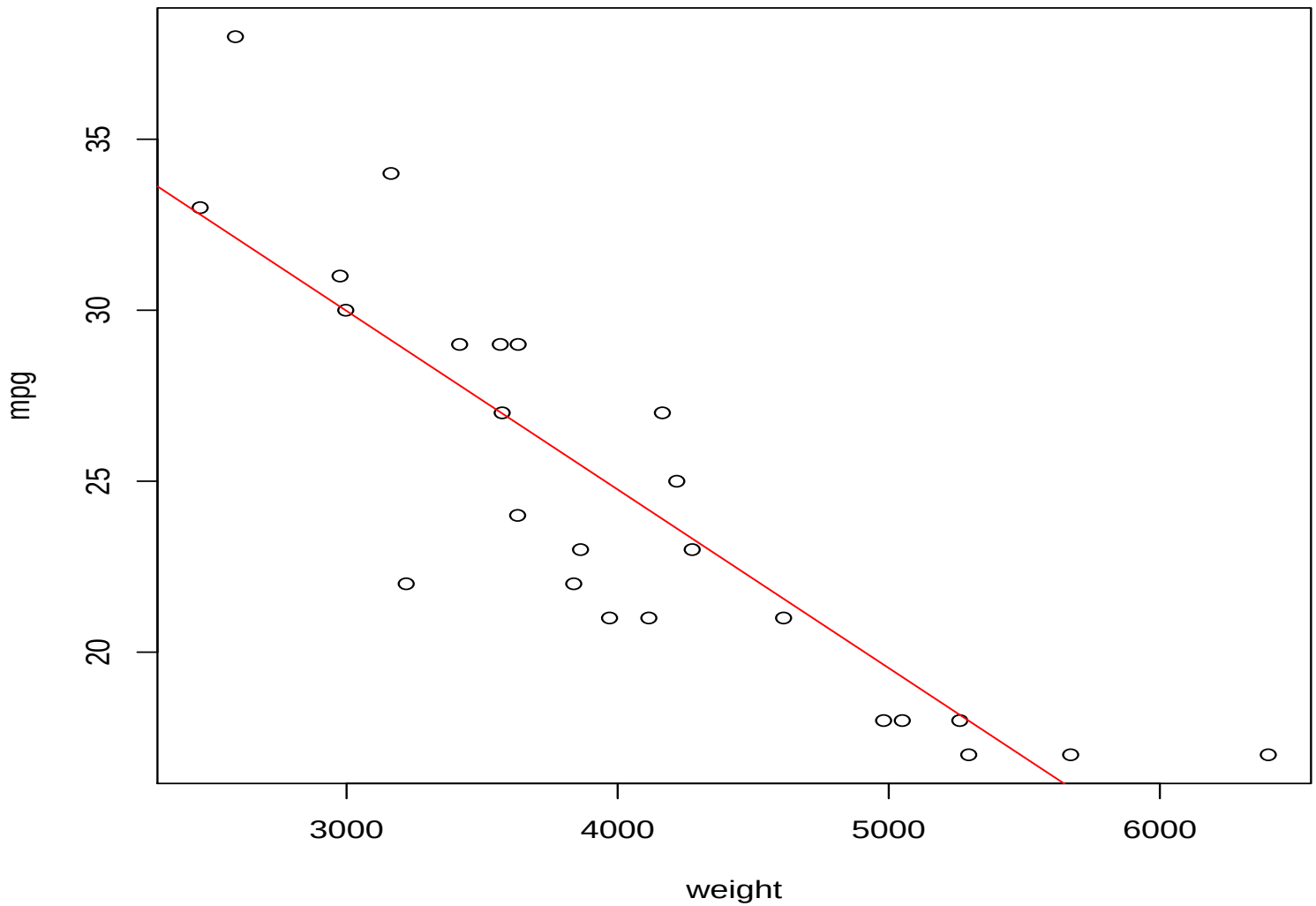

Figure 16: The scatterplot of x weights (in pounds) and y mileage superimposed with the estimated regression line $\hat{y} = 45.65 - 0.005x$.

Intercept and Coefficient

For the estimated regression equation $\hat{y} = 45.65 - 0.005x$, what do the y-intercept and x-coefficient mean?

$b_0 = 45.65$ The predicted value of MPG(y) when weight $x = 0$, which is meaningless. It is a kind of extrapolation and extrapolation is too dangerous to be used!

$b_1 = -0.005$ The average miles per gallon of gas decreases for an additional pound increase of a car.

Residuals

Given the estimated regression equation $\hat{y} = 45.65 - 0.005x$, complete the table by calculating the estimated MPGs and residuals.

model	weight(x)	mileage(y)	$\hat{y}$	$y - \hat{y}$
Toyota Corolla	2590	38		
Jeep Liberty	4115	21		
Hummer H2	6400	17		

Residuals

model	weight(x)	mileage(y)	$\hat{y}$	$y - \hat{y}$
Toyota Corolla	2590	38	32.12	5.88
Jeep Liberty	4115	21	24.16	-3.16
Hummer H2	6400	17	12.22	4.78

Table 10: Estimated MPGs and residuals for 3 different cars.

R-square

How to assess the quality of the regression line? Ignoring x and its relationship with y, we can predict y merely using the sample mean $\bar{y}$. Then, the total sum of squares equals

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Instead, using the estimated regression equation, the sum of squared residuals is

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The R-square gives the summary measure of association and describes the predictive power

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}}$$

R-square

The R-square(R^2) summarizes the reduction in the prediction error when using the regression equation rather than $\bar{y}$ to predict y. R^2 is also interpreted as how much of the variability in y can be explained by x using the simply linear regression. Especially, R^2 is related to the correlation r in formula

$$R^2 = r^2$$

Where

$$r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y})$$

For instance, in the car weight and MPG example, $R^2 = 0.75$ indicates that 75% of variability in the MPGs that is explained by the simple linear regression between weight and MPG.

Polynomial Regression

Assume that there is an underlying curve relationship between the predictor variable X and the response variable Y. This relationship can be written as a polynomial regression equation, denoted by

$$\mu_Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

Where μ_Y denotes the population mean of Y for all the subjects at a particular value of X.

In practice, we estimate the regression equation using the sample data, and the estimated regression equation is denoted by

$$\hat{y} = b_0 + b_1 x + b_2 x^2$$

Polynomial Regression

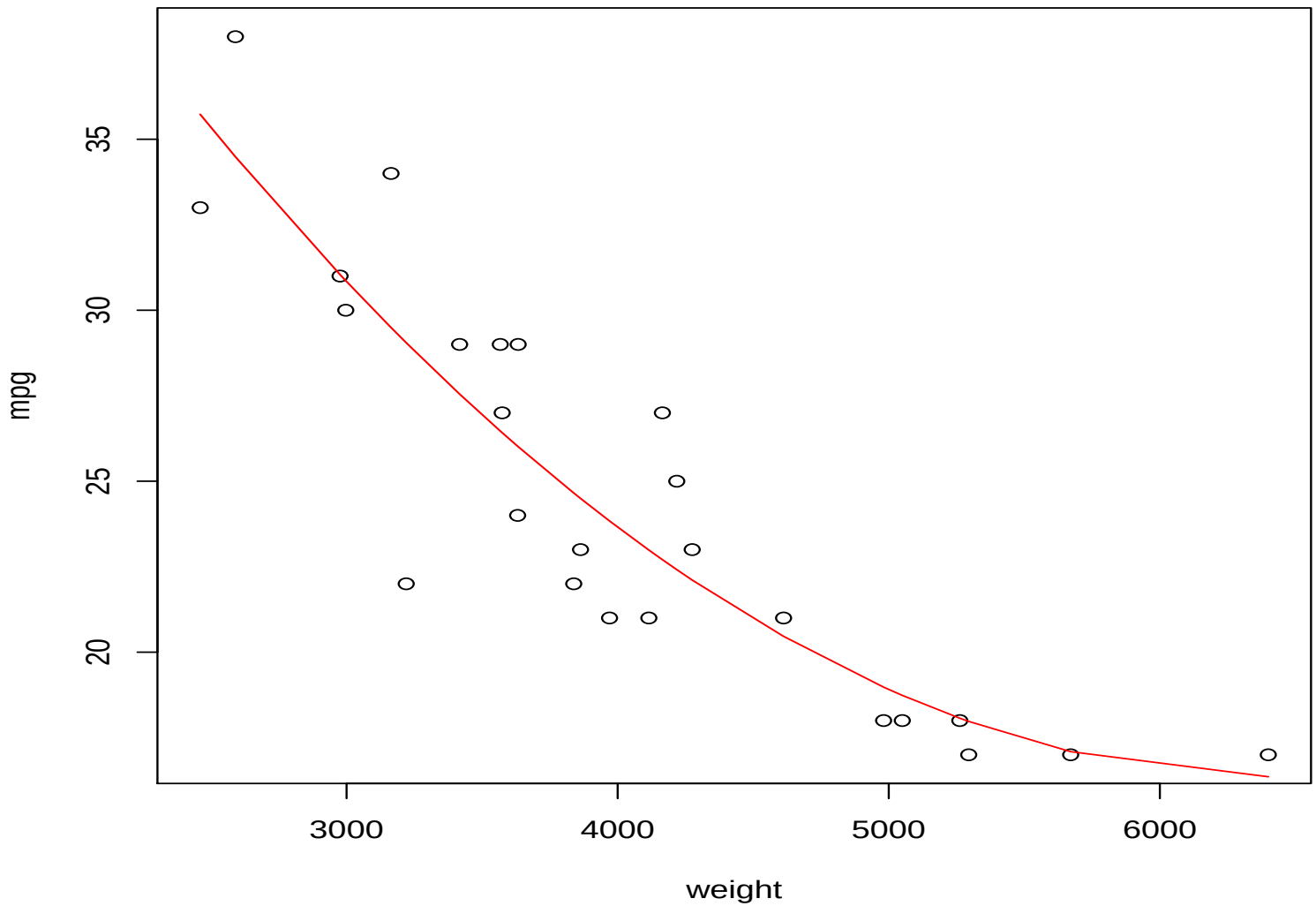

Figure 17: The scatter plot of weight(x) and MPG(y) with superimposed the estimated polynomial regression $\hat{y} = 67 - 0.016x + 1.2 \times 10^{-6}x^2$.

Polynomial Regression

```
Call:
lm(formula = car$mileage ~ car$weight + I(car$weight^2))

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.700e+01  8.843e+00   7.577 1.44e-07 ***
car$weight      -1.571e-02  4.224e-03  -3.718   0.0012 **
I(car$weight^2)  1.218e-06  4.862e-07   2.505   0.0202 *

Residual standard error: 2.72 on 22 degrees of freedom
Multiple R-squared:  0.8062,Adjusted R-squared:  0.7886
F-statistic: 45.77 on 2 and 22 DF,  p-value: 1.447e-08
```

The estimated regression equation is

$$\hat{y} = 67 - 0.016x + 1.2 \times 10^{-6} x^2$$

Polynomial Regression

Using the polynomial regression, $R^2 = 0.81$, indicates that 81% of variability in the response variable MPG that can be explained by the polynomial regression between the predicator variable weight and the response variable MPG.

Polynomial Regression

Using the polynomial regression, $R^2 = 0.81$, indicates that 81% of variability in the response variable MPG that can be explained by the polynomial regression between the predicator variable weight and the response variable MPG.

Calculate the predicted MPGs and residuals using the estimated polynomial equation

$$\hat{y} = 67 - 0.016x + 1.2 \times 10^{-6}x^2$$

model	weight(x)	weight^2(x^2)	mileage(6)	$\hat{y}$	$y - \hat{y}$
Toyota Corolla	2590	6.7×10^6	38		
Jeep Liberty	4115	16.93×10^6	21		
Hummer H2	6400	40.9×10^6	17		

Polynomial Regression

model	weight(x)	weight^2(x^2)	mileage(y)	$\hat{y}$	$y - \hat{y}$
Toyota Corolla	2590	6.7×10^6	38	34.49	3.51
Jeep Liberty	4115	16.93×10^6	21	22.99	-1.99
Hummer H2	6400	40.9×10^6	17	16.36	0.64

Table 11: Predicted MPGs and residuals using the estimated polynomial regression equation $\hat{y} = 67 - 0.016x + 1.2 \times 10^{-6}x^2$.

Model Selection

Given two fitted model, the simply linear regression model and the polynomial regression model, which model should you select to make predictions?

Rule of Thumb

- Visual fitting
- Larger R^2
- simpler model