Introduction to Regression

Hypothesis Testing

Statistic	**Dist**	**Dependent Variable**	**Independent Variable**
One-Sample Hypothesis	t/Z	Interval-ratio	Nominal or Ordinal (1 value, compared against population mean)
Two-Sample Hypothesis	t/Z	Interval-ratio	Nominal or Ordinal (2 values a.k.a. binary)
ANOVA	F	Interval-ratio	Nominal or Ordinal (3 or more values, but usually no more than 5)
Chi-Square	χ^2	Nominal or Ordinal	Nominal or Ordinal

Hypothesis Testing: Two Interval-Ratio Level Variables

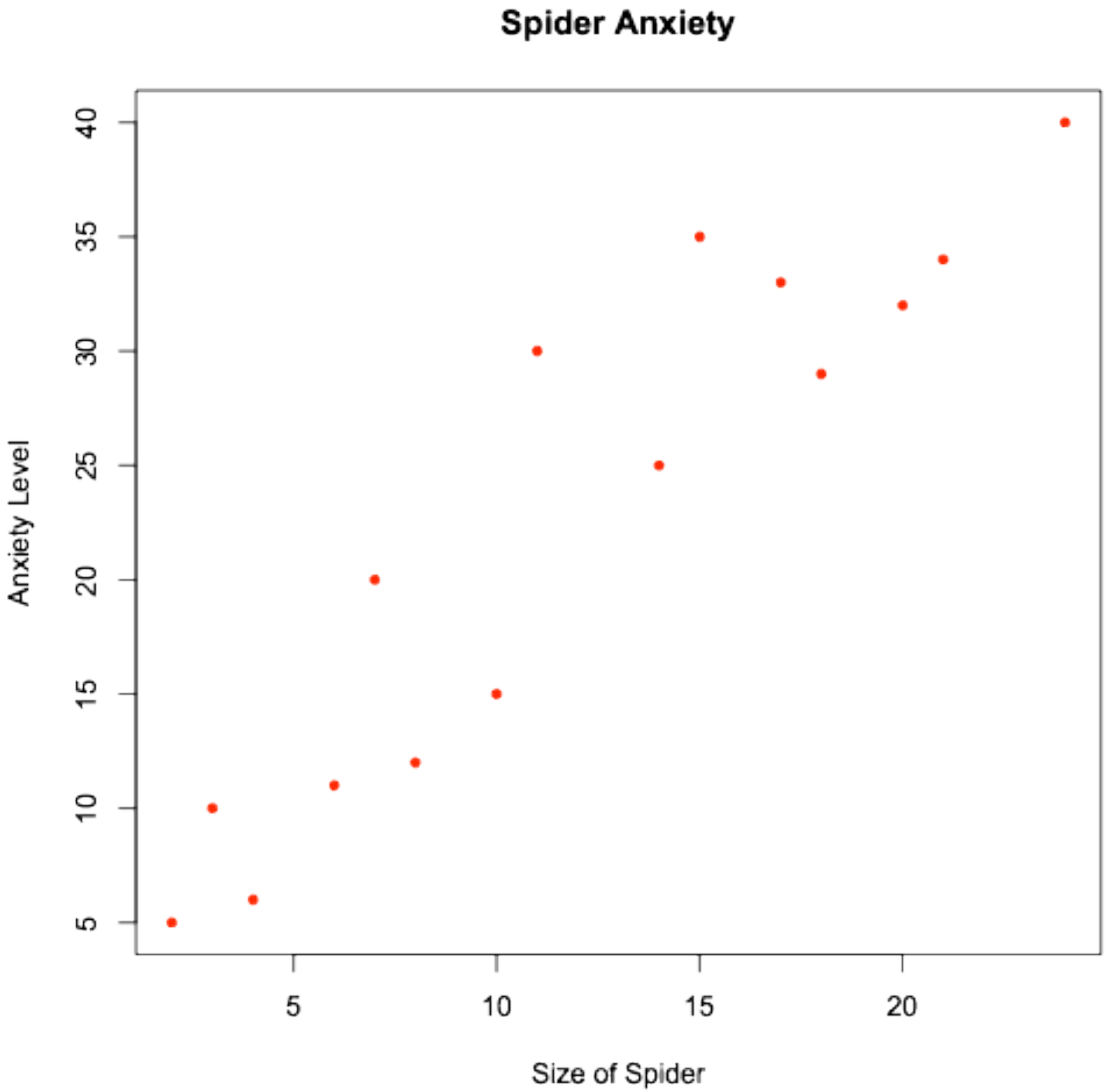
- Regression (Linear)
- Least Squares Method

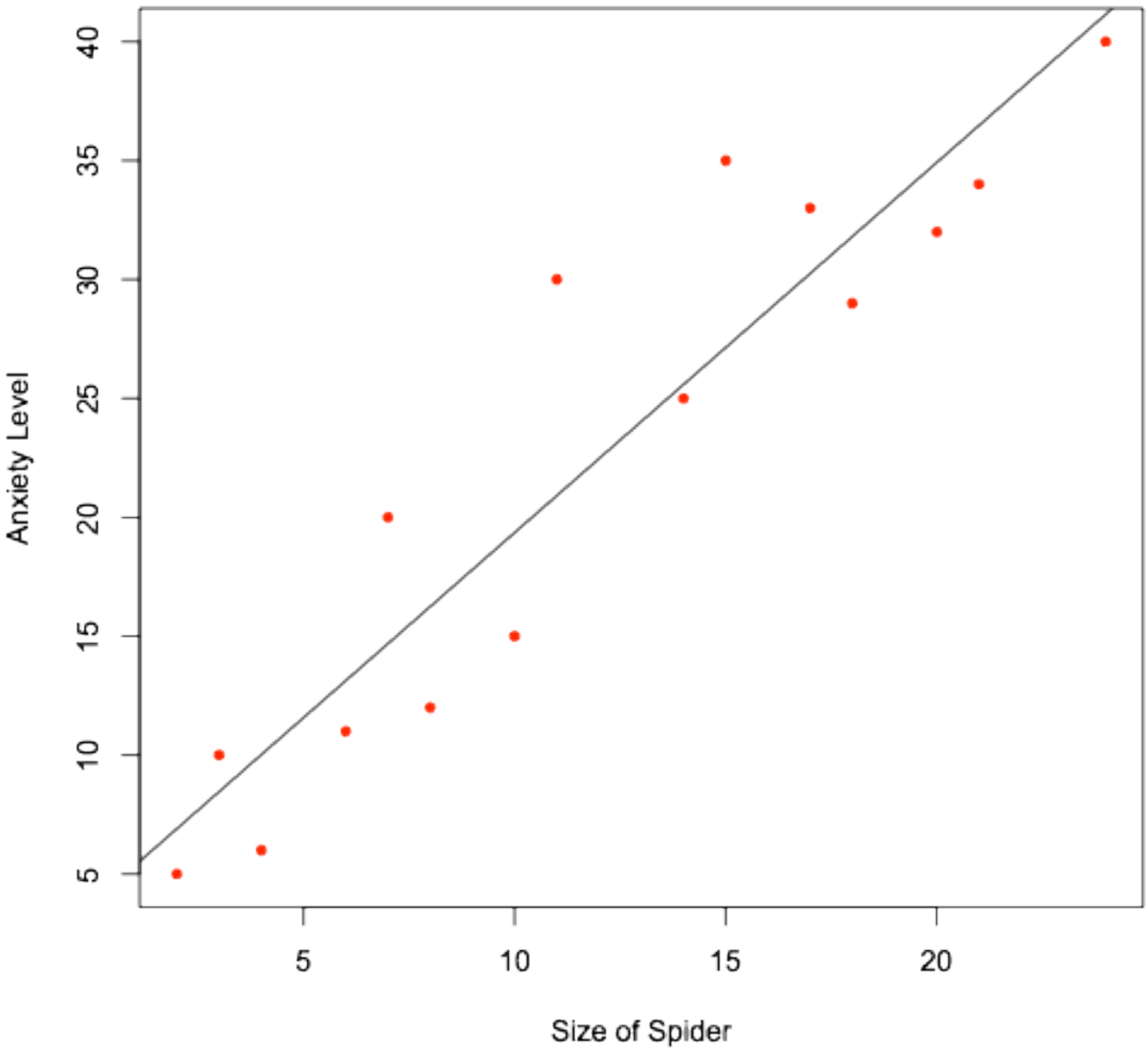
Spiders!

Data on Spider Anxiety

anx(y)	size(x)
5	2
10	3
6	4
11	6
12	8
20	7
15	10
30	11
25	14
35	15
33	17
29	18
32	20
34	21
40	24

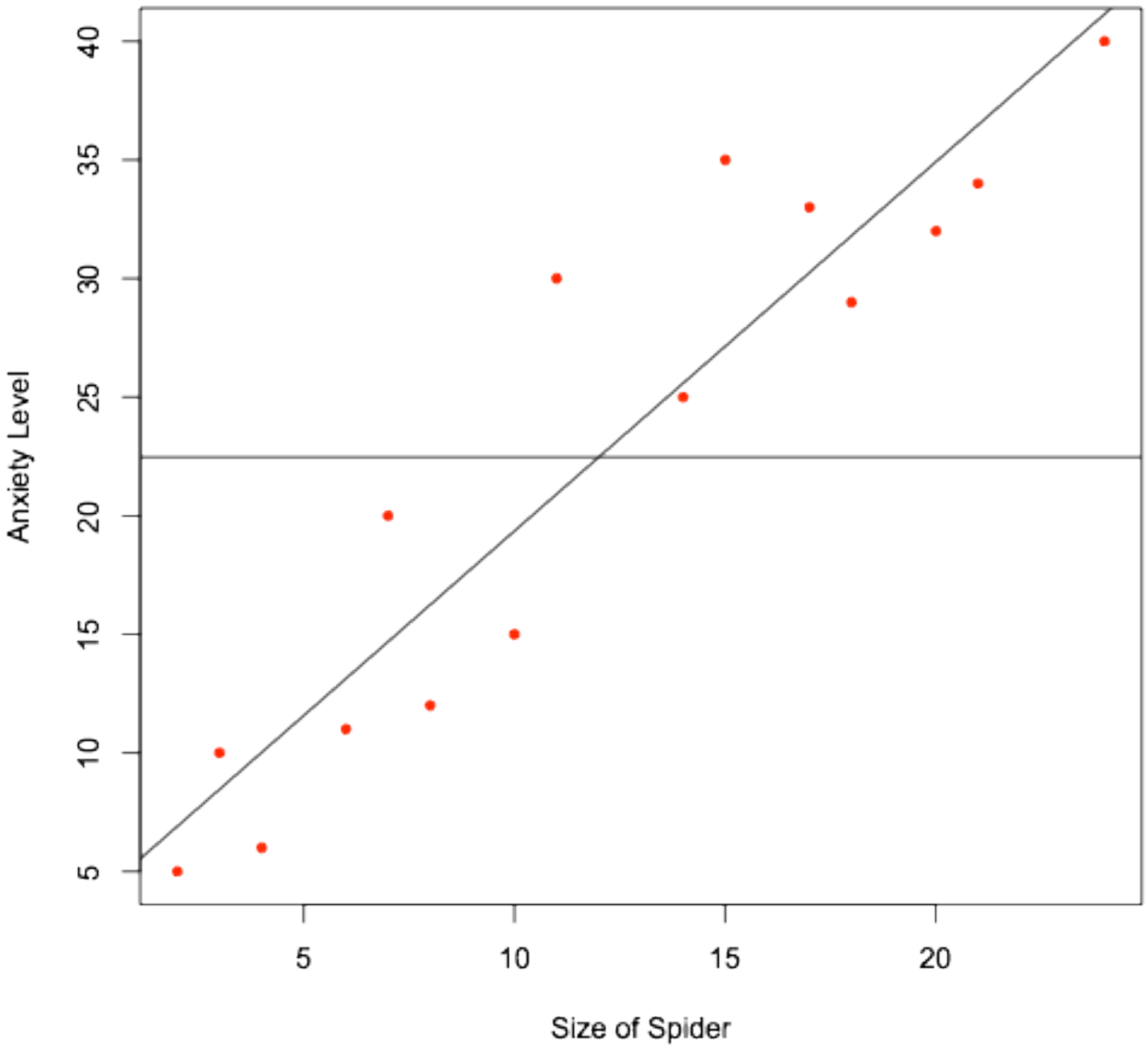
Scatter Plot




Anxiety Level
5
10
15
20
25
30
35
40
5
10
15
20
Size of Spider

Properties of this Line

- Best fit, thanks to Least-Squares (circa 1795)
- Perfectly straight


Anxiety Level
5
10
15
20
25
30
35
40
5
10
15
20
Size of Spider

Simple Regression

- Define Dependent Variable (Y) and Independent Variable (X); both should be interval-ratio!!!
- Find the best fitting line:
  - $Y = a + bX$
  - formulae for b and a are 14.2 and 14.3 in the textbook

The Hypothesis Test

- Uses the F distribution, and Analysis of Variance
- SST = $\sum(Y_i - Ybar)^2$
- SSR = $\sum(Y_i - (a + bX))^2$
- SSM = SST - SSR
- dfM = k
- dfR = N-k-1

The Hypothesis Test

- MSM = SSM / dfM
- MSR = SSR / dfR
- F = MSM/ MSR
- R^2 = SSM / SST
- R = SQRT(R^2)

The Hypothesis Test

- Uses the F distribution, and Analysis of Variance
- SST = Total Sum of Squares
- SSR = Residual Sum of Squares
- SSM = Model Sum of Squares
- dfM = Model degrees of freedom
- dfR = Residual degrees of freedom
- k = number of independent variables

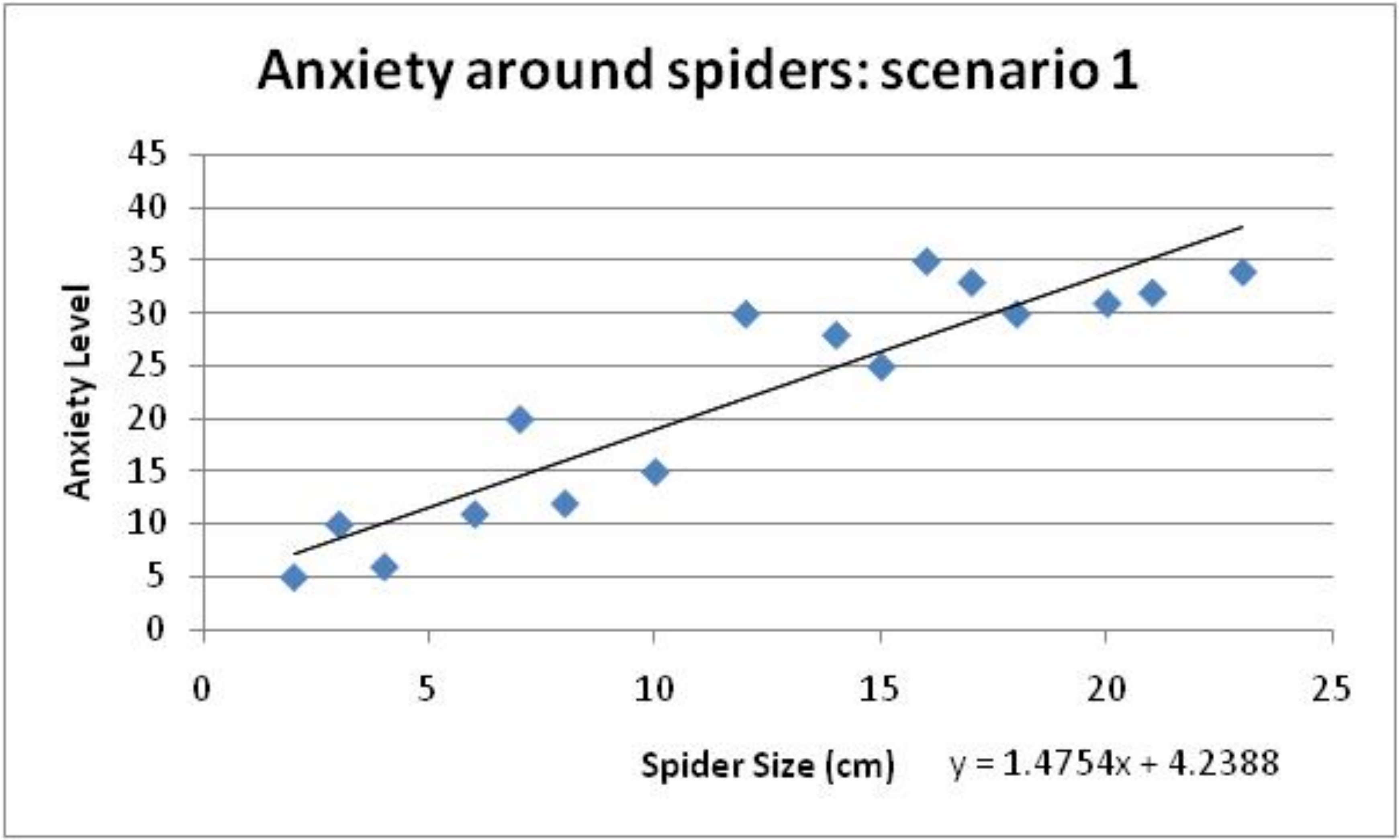
Results

```
Analysis of Variance TableResponse: anx
Df     Sum Sq       Mean Sq      F value      Pr(>F)    size
1      1671.70      1671.70      81.081       5.987e-07 ***
13     268.03       20.62
```

```
Coefficients:                Estimate    Std. Error   t value      Pr(>|t|)
(Intercept)                 3.7884      2.3827       1.590        0.136
size                        1.5565      0.1729       9.005        5.99e-07
```

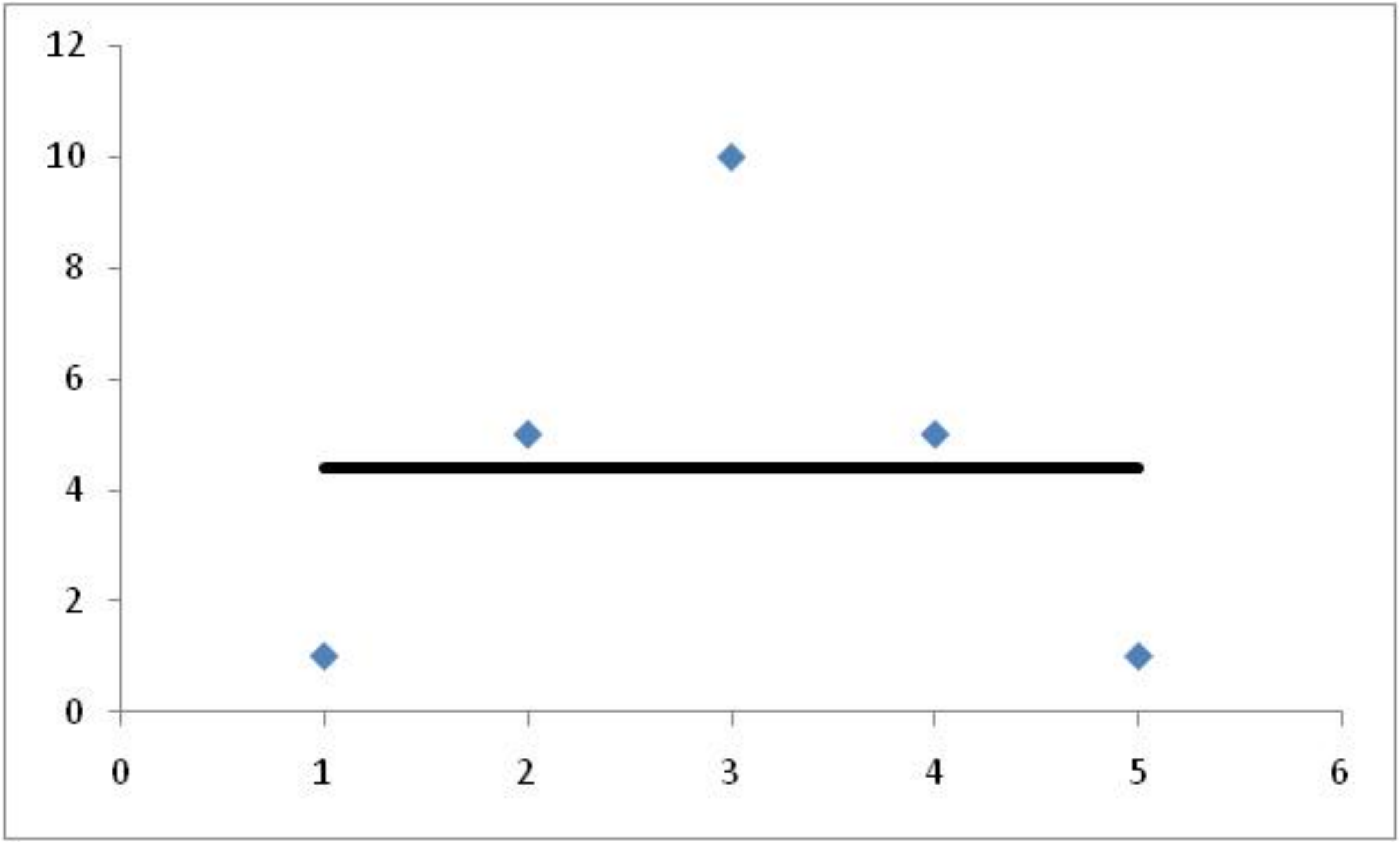
```
***---Multiple R-squared: 0.8618,  Adjusted R-squared: 0.8512
F-statistic: 81.08 on 1 and 13 DF,  p-value: 5.987e-07
```

PREDICTION with regression

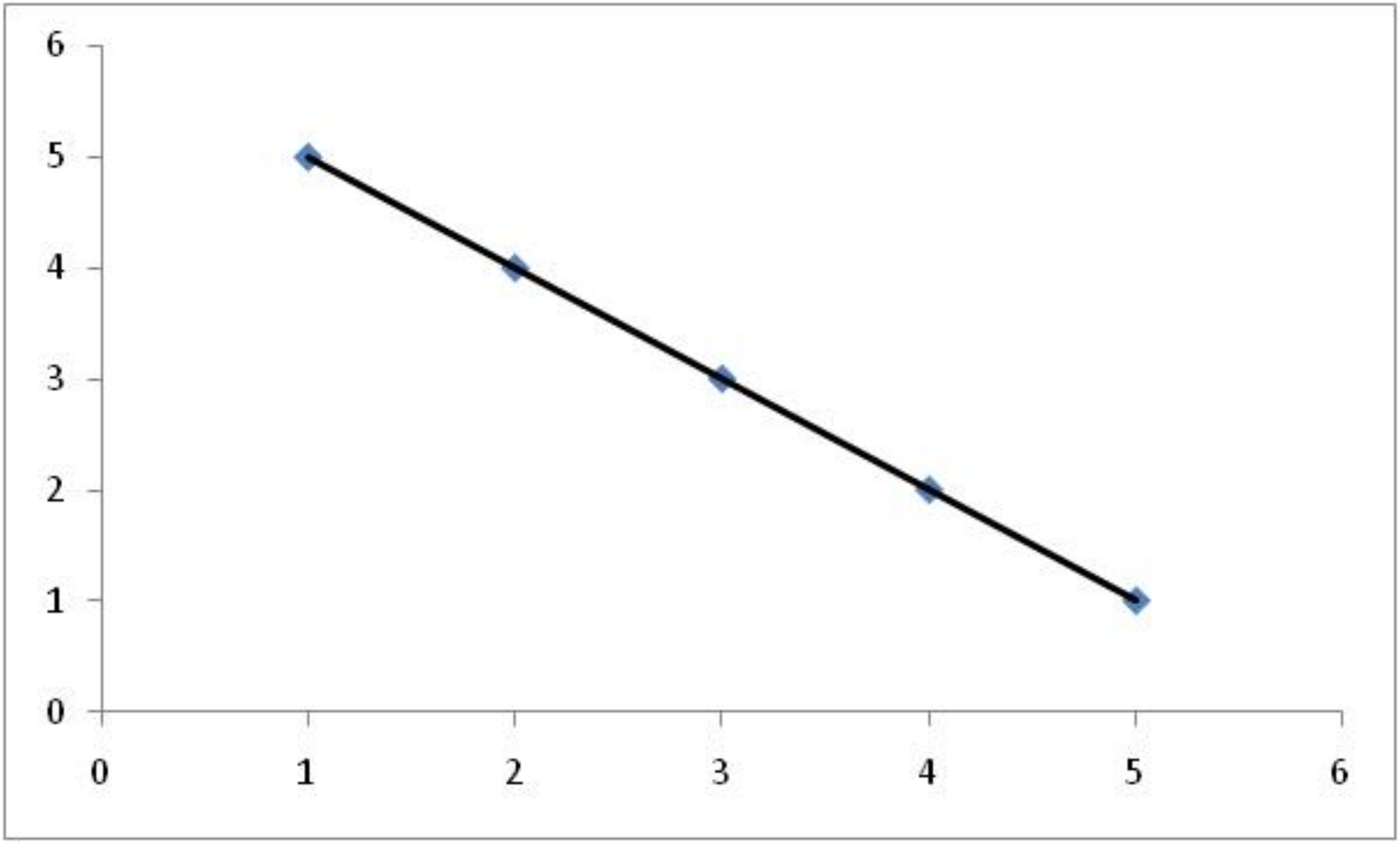


Regression: goodness of fit

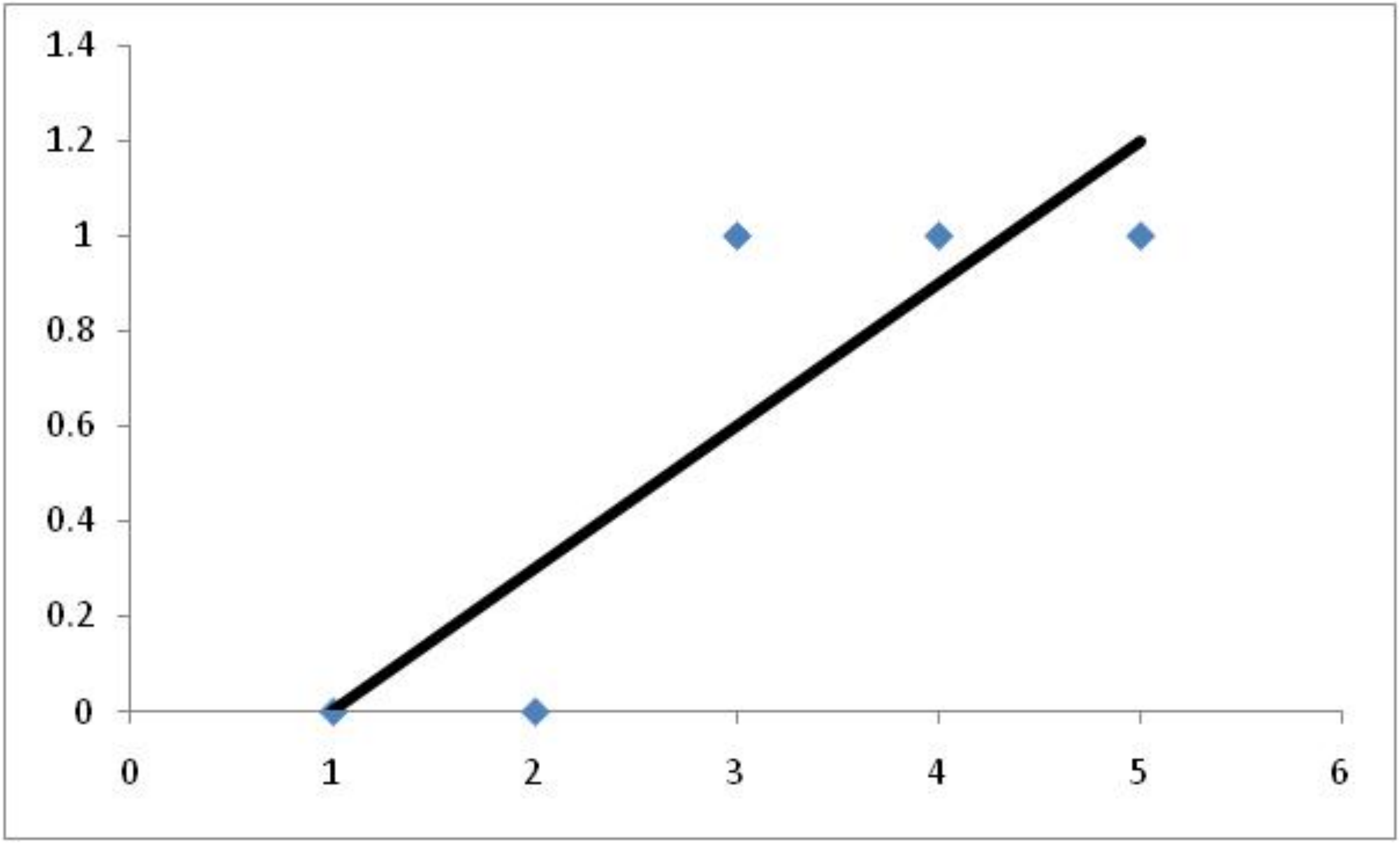
How well does the line fit the data?



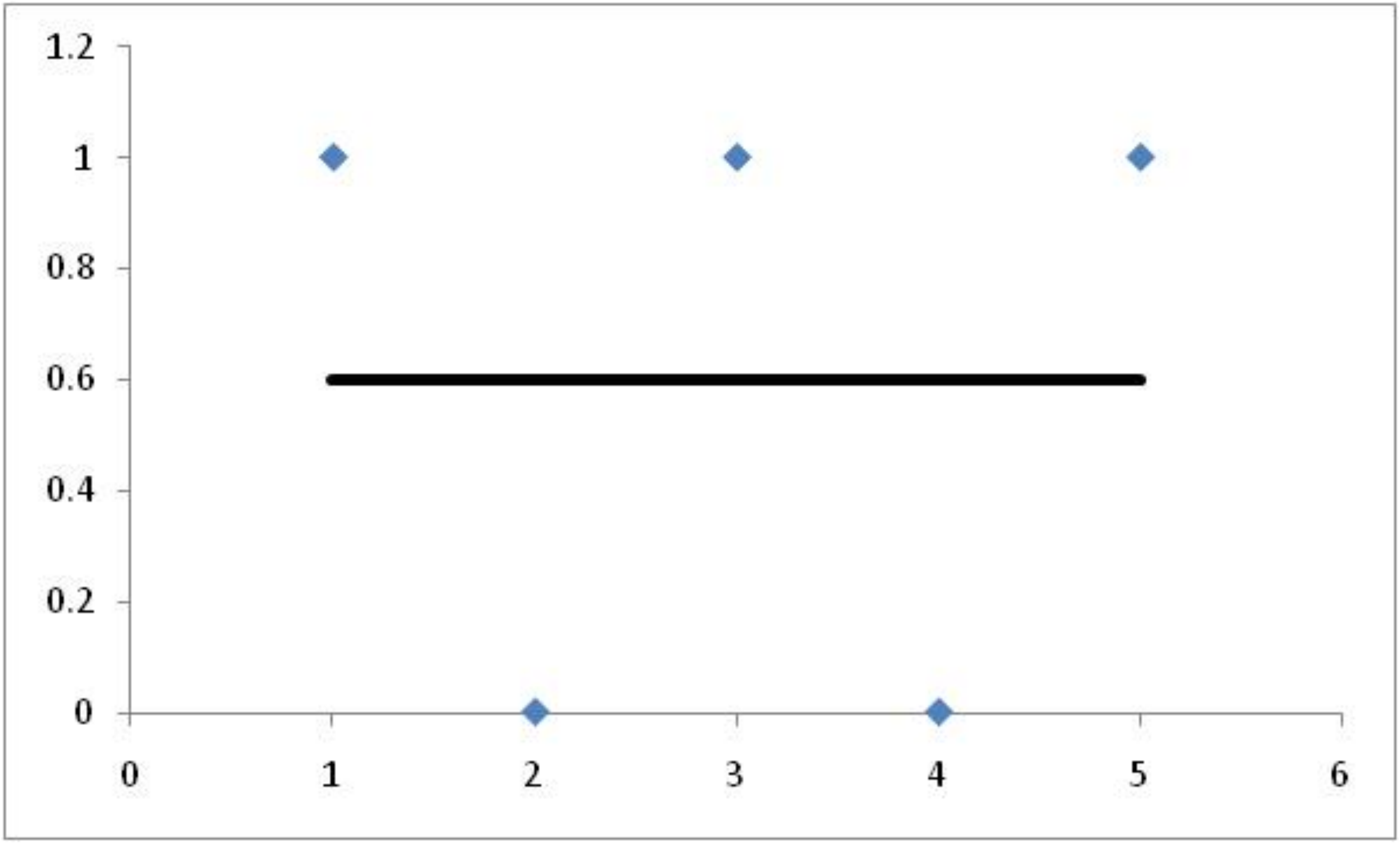
How well does the line fit the data?



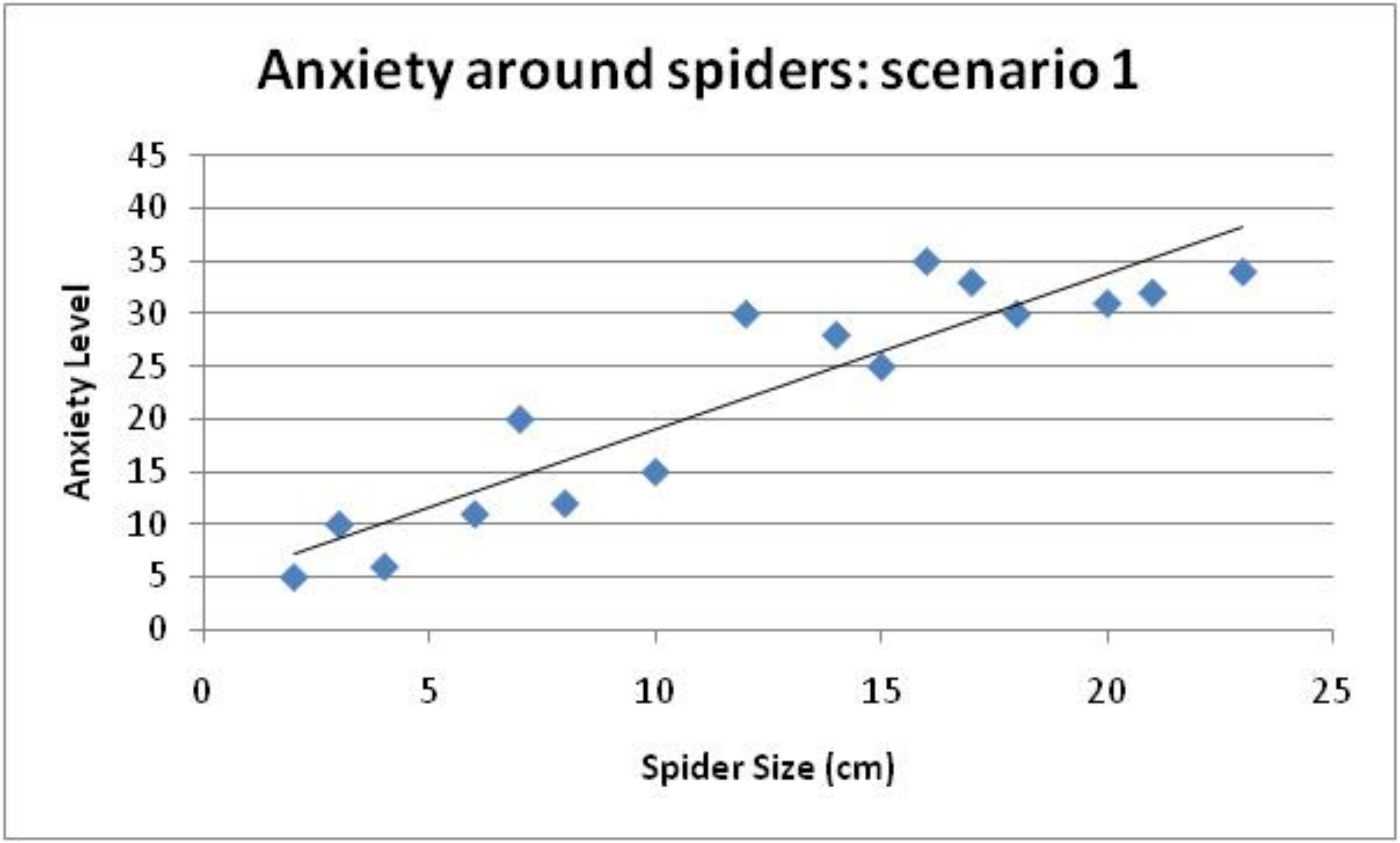
How well does the line fit the data?



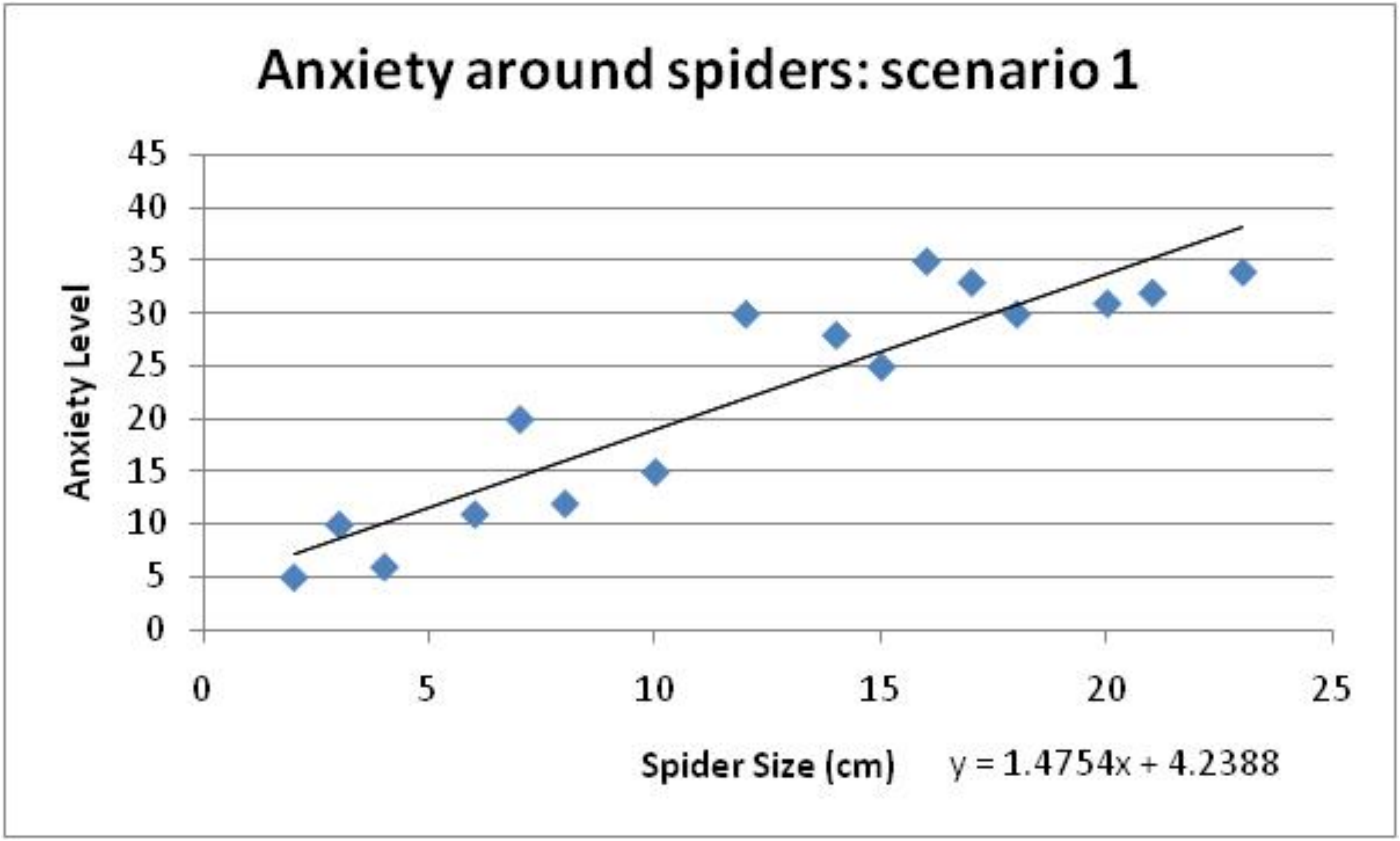
How well does the line fit the data?



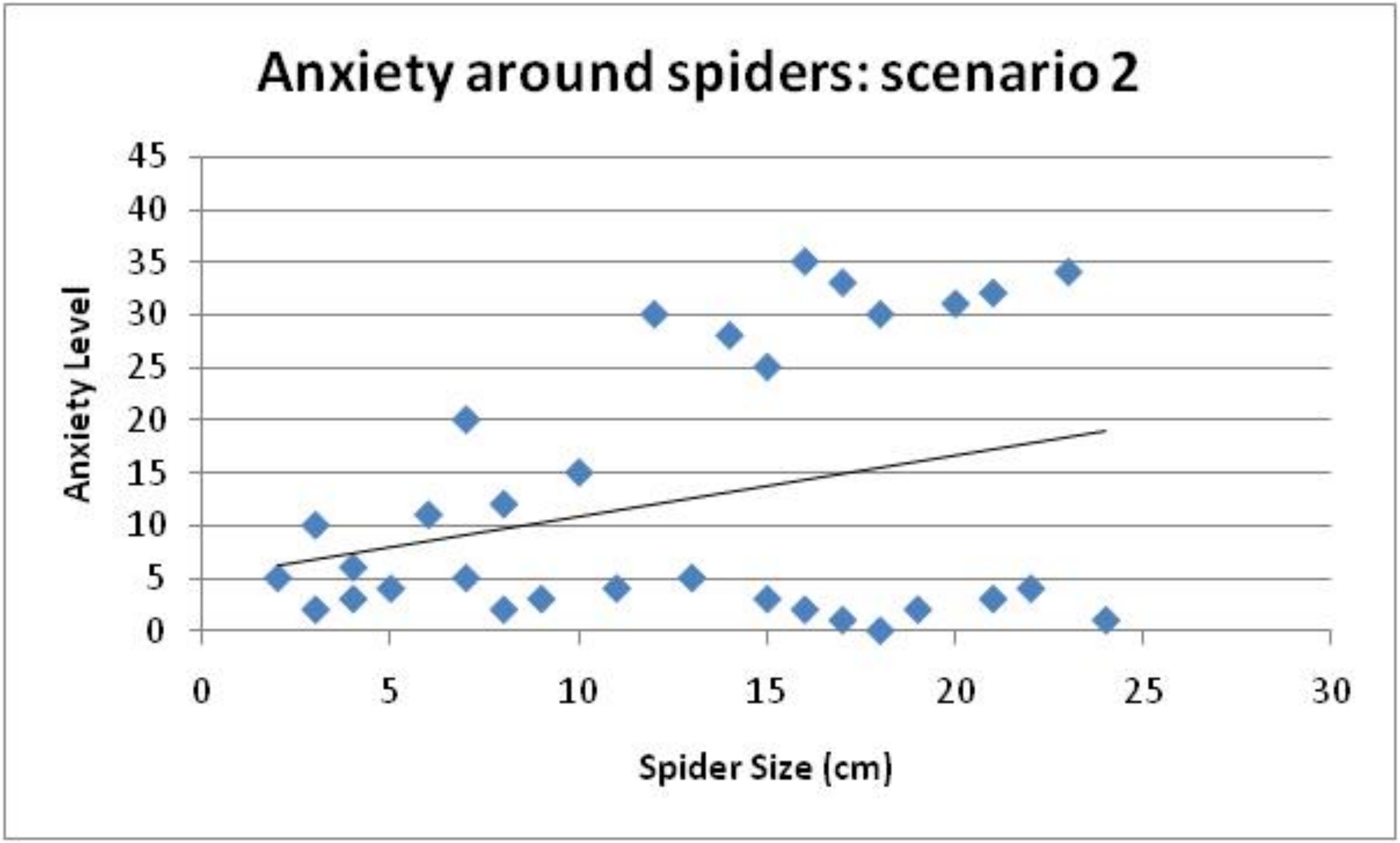
How well does the line fit the data?



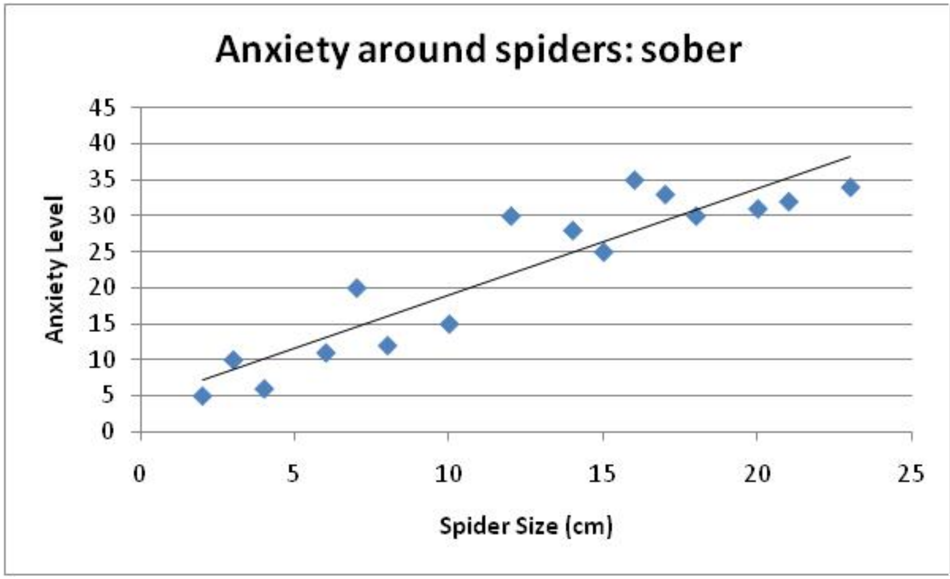
Simple Regression

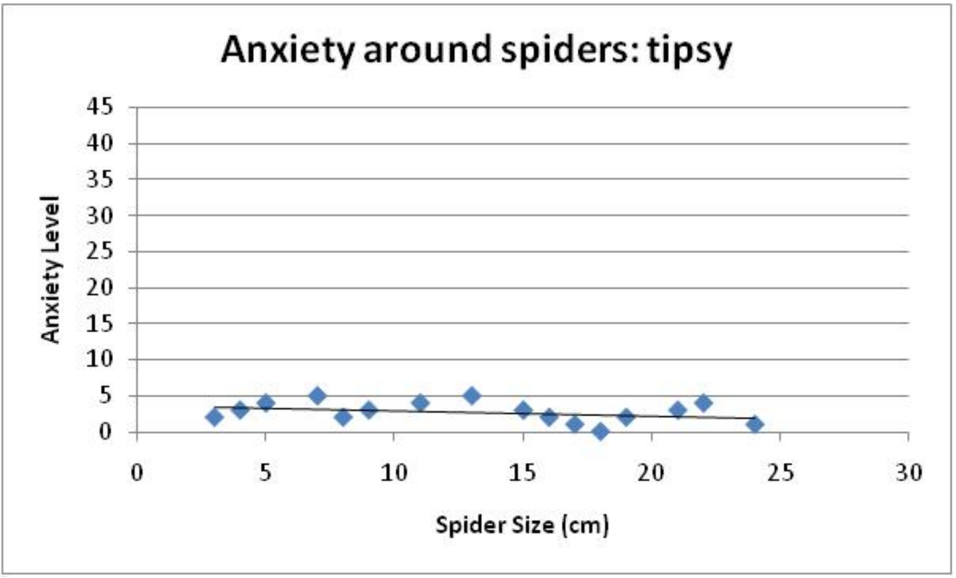


Another Spider/Anxiety Scenario

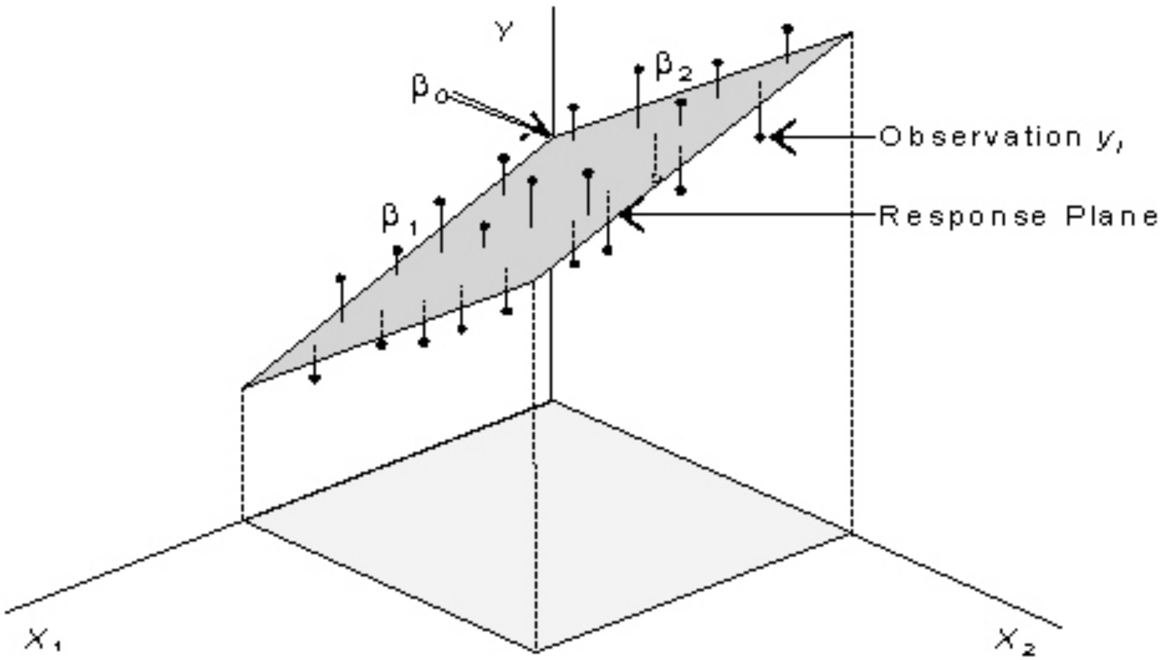


Another Spider/Anxiety Scenario





Multiple Regression: 2 Predictors



Source: http://www.sjsu.edu/faculty/gerstman/EpiInfo/cont-mult1.jpg