STA221

Neil Montgomery

Last edited: 2017-01-31 13:48

regression

linear models

Basic model: Output $=$ Input $+$ Noise

You've seen a few of these already; perhaps not expressed this way.

linear models

Basic model: Output = Input + Noise

You've seen a few of these already; perhaps not expressed this way.

The "one sample t-test" can be thought of a way to analyze data that can be modeled as:

$$Y_i = \mu + \varepsilon_i$$

where ε_i are independent $N(0, \sigma)$ and n is the sample size.

linear models

Basic model: Output = Input + Noise

You've seen a few of these already; perhaps not expressed this way.

The "one sample t-test" can be thought of a way to analyze data that can be modeled as:

$$Y_i = \mu + \varepsilon_i$$

where ε_i are independent $N(0, \sigma)$ and n is the sample size.

The "two sample t-test" could be modeled as:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where $i = 1, 2$ and the μ_i are the two population means. (There are a few ways to treat the ε_{ij}.)

several numerical variables

Suppose your dataset has a numerical variable we'll call y and other variable (typically also numerical) called x. Most datasets will have several!

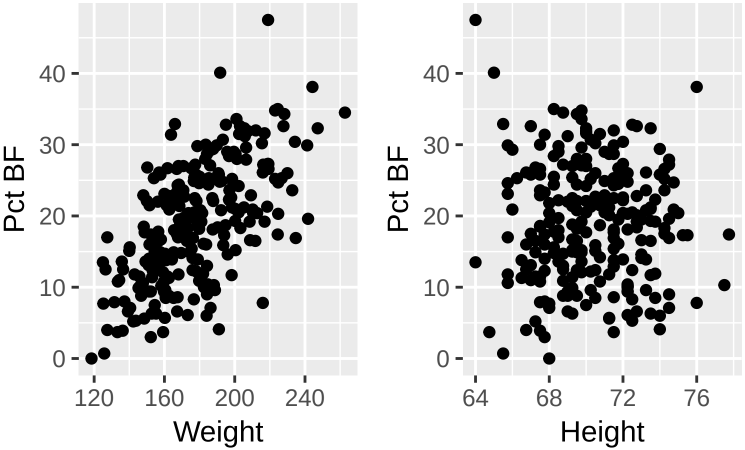
several numerical variables

Suppose your dataset has a numerical variable we'll call y and other variable (typically also numerical) called x. Most datasets will have several!

Let's consider the male body fat dataset that is discussed in the textbook (Chapter 24).

```
## # A tibble: 250 × 15
##   `Pct BF`   Age Weight Height  Neck Chest Abdomen    waist   Hip
##      <dbl> <int>  <dbl>  <dbl> <dbl> <dbl>   <dbl>    <dbl> <dbl>
## 1     12.3    23 154.25  67.75  36.2  93.1    85.2 33.54331  94.5
## 2      6.1    22 173.25  72.25  38.5  93.6    83.0 32.67717  98.7
## 3     25.3    22 154.00  66.25  34.0  95.8    87.9 34.60630  99.2
## 4     10.4    26 184.75  72.25  37.4 101.8    86.4 34.01575 101.2
## 5     28.7    24 184.25  71.25  34.4  97.3   100.0 39.37008 101.9
## # ... with 245 more rows, and 6 more variables: Thigh <dbl>,
## #   Knee <dbl>, Ankle <dbl>, Bicep <dbl>, Forearm <dbl>, Wrist <dbl>
```

body fat EDA


Pct BF
40
30
20
10
0
120
160
200
240
Weight
Pct BF
40
30
20
10
0
64
68
72
76
Height

linear model for two numerical variables

When there is a linear relationship between two variables, we might propose a linear model such as:

linear model for two numerical variables

When there is a linear relationship between two variables, we might propose a linear model such as:

$\texttt{Pct BF} = \texttt{Weight} + \text{noise}$

linear model for two numerical variables

When there is a linear relationship between two variables, we might propose a linear model such as:

$\texttt{Pct BF} = \texttt{Weight} + \text{noise}$

$\texttt{Pct BF} = \texttt{Height} + \text{noise}$

linear model for two numerical variables

When there is a linear relationship between two variables, we might propose a linear model such as:

$\texttt{Pct BF} = \texttt{Weight} + \text{noise}$

$\texttt{Pct BF} = \texttt{Height} + \text{noise}$

In general:

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

where y and x are the variables and ε is the random noise.

linear model for two numerical variables

When there is a linear relationship between two variables, we might propose a linear model such as:

$\texttt{Pct BF} = \texttt{Weight} + \text{noise}$

$\texttt{Pct BF} = \texttt{Height} + \text{noise}$

In general:

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

where y and x are the variables and ε is the random noise.

When there are only two variables this is called a *simple regression model.*

linear model for two numerical variables

When there is a linear relationship between two variables, we might propose a linear model such as:

`Pct BF` = `Weight` + noise

`Pct BF` = `Height` + noise

In general:

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

where y and x are the variables and ε is the random noise.

When there are only two variables this is called a *simple regression model.*

The *parameter* β_1 is the slope of the line and is of primary interest. (The parameter β_0 is the y-intercept and not normally of any interest.)

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

y can be called the "output" variable and x can be called the "input" variable. I think these are the best names.

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

y can be called the "output" variable and x can be called the "input" variable. I think these are the best names.

I don't mind called them "response" and "predictor".

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

y can be called the "output" variable and x can be called the "input" variable. I think these are the best names.

I don't mind called them "response" and "predictor".

I hate "dependent" and "independent" variables. These words are already being used by an important concept in probability.

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

y can be called the "output" variable and x can be called the "input" variable. I think these are the best names.

I don't mind called them "response" and "predictor".

I hate "dependent" and "independent" variables. These words are already being used by an important concept in probability.

Think of the model from the inside and move out. It starts with x, which can be anything.

- It doesn't have to be random.

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

y can be called the "output" variable and x can be called the "input" variable. I think these are the best names.

I don't mind called them "response" and "predictor".

I hate "dependent" and "independent" variables. These words are already being used by an important concept in probability.

Think of the model from the inside and move out. It starts with x, which can be anything.

- It doesn't have to be random.
- It could be a pre-specified grid of values.

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

y can be called the "output" variable and x can be called the "input" variable. I think these are the best names.

I don't mind called them "response" and "predictor".

I hate "dependent" and "independent" variables. These words are already being used by an important concept in probability.

Think of the model from the inside and move out. It starts with x, which can be anything.

- It doesn't have to be random.
- It could be a pre-specified grid of values.
- The "grid" could consist of as few as two values!

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

This line is the basis of the relationship between input and output.

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

This line is the basis of the relationship between input and output.

Finally, to this line we add some random noise ε to get the final model:

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

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

This line is the basis of the relationship between input and output.

Finally, to this line we add some random noise ε to get the final model:

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

For the moment we'll put these requirements on the random noise:

- the noise has constant variation

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

This line is the basis of the relationship between input and output.

Finally, to this line we add some random noise ε to get the final model:

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

For the moment we'll put these requirements on the random noise:

- the noise has constant variation
- the noise for each observation is independent

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

This line is the basis of the relationship between input and output.

Finally, to this line we add some random noise ε to get the final model:

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

For the moment we'll put these requirements on the random noise:

- the noise has constant variation
- the noise for each observation is independent

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

This line is the basis of the relationship between input and output.

Finally, to this line we add some random noise ε to get the final model:

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

For the moment we'll put these requirements on the random noise:

- the noise has constant variation
- the noise for each observation is independent

We'll add another requirement when the time comes.

estimating the slope (and intercept)

Now we proceed through the typical steps of a data analysis (the χ^2 procedures were an exception!)

- We have a model with unknown parameters.

estimating the slope (and intercept)

Now we proceed through the typical steps of a data analysis (the χ^2 procedures were an exception!)

- We have a model with unknown parameters.
- So we gather data, and use the data to estimate the paramaters.

estimating the slope (and intercept)

Now we proceed through the typical steps of a data analysis (the χ^2 procedures were an exception!)

- We have a model with unknown parameters.
- So we gather data, and use the data to estimate the paramaters.
- Use probability to make inferences using these estimates.

estimating the slope (and intercept)

Now we proceed through the typical steps of a data analysis (the χ^2 procedures were an exception!)

- We have a model with unknown parameters.
- So we gather data, and use the data to estimate the paramaters.
- Use probability to make inferences using these estimates.

estimating the slope (and intercept)

Now we proceed through the typical steps of a data analysis (the χ^2 procedures were an exception!)

- We have a model with unknown parameters.
- So we gather data, and use the data to estimate the paramaters.
- Use probability to make inferences using these estimates.

The classic method of regression parameter estimation given data is called *least squares regression*.

- The data come in pairs $(y_1, x_1), (y_2, x_2), \ldots (y_n, x_n)$.

estimating the slope (and intercept)

Now we proceed through the typical steps of a data analysis (the χ^2 procedures were an exception!)

- We have a model with unknown parameters.
- So we gather data, and use the data to estimate the paramaters.
- Use probability to make inferences using these estimates.

The classic method of regression parameter estimation given data is called *least squares regression*.

- The data come in pairs $(y_1, x_1), (y_2, x_2), \ldots (y_n, x_n)$.
- For any "candidate" slope b_0^* and intercept b_1^* we could construct the set of "predictions" $\hat{y}_i = b_0^* + b_1^* x_i$ and their "residuals" $\varepsilon_i = y_i - \hat{y}_i$

more least squares

Here's the actual "least squares" part...

It is possible to find the unique slope and intercept that makes this sum of squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0^* + b_1^* y_i))^2$$

as small as possible.

We'll call the unique intercept and slope b_0 and b_1, respectively.

more least squares

Here's the actual "least squares" part...

It is possible to find the unique slope and intercept that makes this sum of squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0^* + b_1^* y_i))^2$$

as small as possible.

We'll call the unique intercept and slope b_0 and b_1, respectively.

(More common to call them $\hat{\beta}_0$ and $\hat{\beta}_1$.)

more least squares

Here's the actual "least squares" part...

It is possible to find the unique slope and intercept that makes this sum of squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0^* + b_1^* y_i))^2$$

as small as possible.

We'll call the unique intercept and slope b_0 and b_1, respectively.

(More common to call them $\hat{\beta}_0$ and $\hat{\beta}_1$.)

The formula for the slope estimator b_1 turns out to be (corrected from class!):

$$b_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

more least squares

Here's the actual "least squares" part...

It is possible to find the unique slope and intercept that makes this sum of squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0^* + b_1^* y_i))^2$$

as small as possible.

We'll call the unique intercept and slope b_0 and b_1, respectively.

(More common to call them $\hat{\beta}_0$ and $\hat{\beta}_1$.)

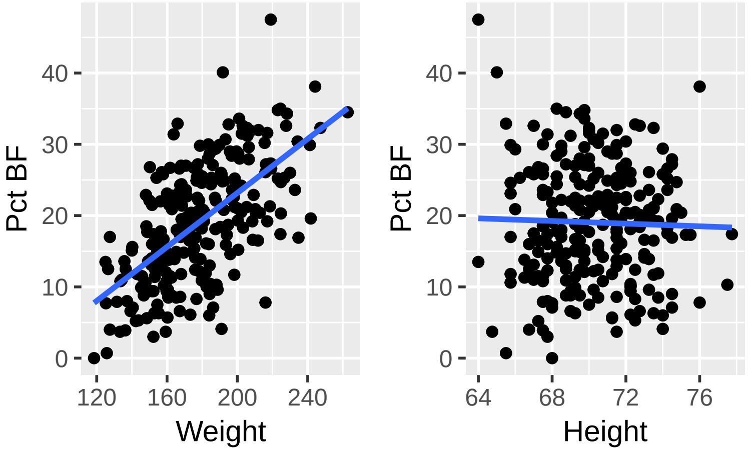
The formula for the slope estimator b_1 turns out to be (corrected from class!):

$$b_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

The formula for the intercept is $b_0 = \overline{y} - b_1 \overline{x}$

body fat examples

Here are the plots with the least squares regression lines added:



bodyfat calculation examples

Obviously don't do these by hand! Here is basic R regression output:

```
##
## Call:
## lm(formula = `Pct BF` ~ Weight, data = bodyfat)
##
## Coefficients:
## (Intercept)       Weight
##    -14.6931       0.1894
```

```
##
## Call:
## lm(formula = `Pct BF` ~ Height, data = bodyfat)
##
## Coefficients:
## (Intercept)       Height
##    25.58078     -0.09316
```