Introductory Statistics Lectures

Linear regression

How to mathematically model a linear relationship and make predictions.

Anthony Tanbakuchi
Department of Mathematics
Pima Community College

(Compile date: Mon Apr 28 20:50:28 2008)

Contents

1 Linear regression

1.1 Introduction

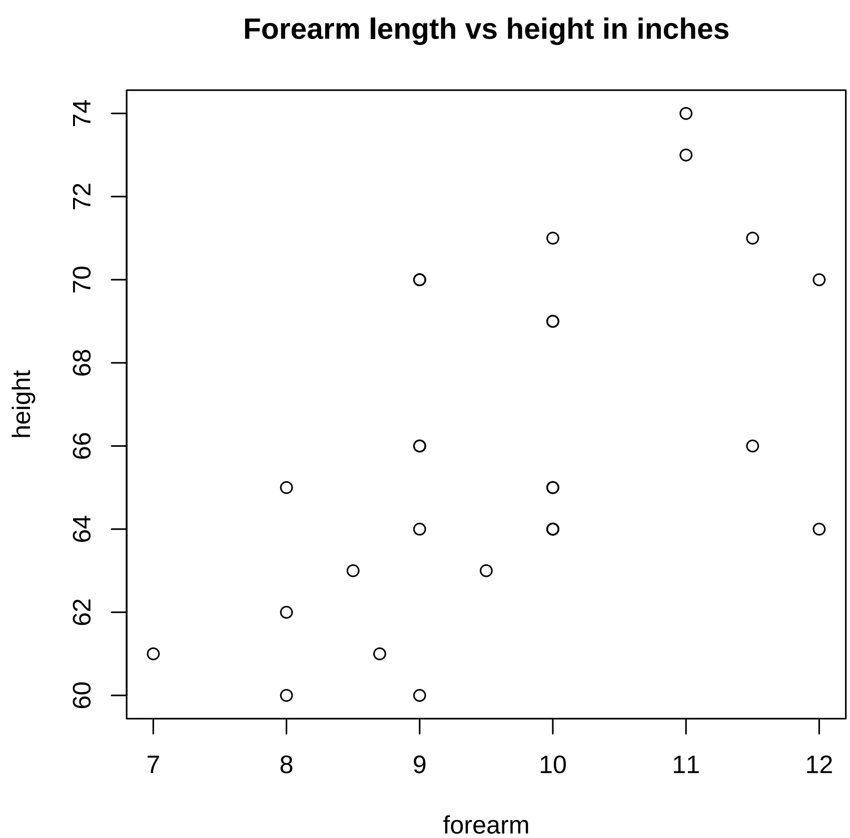
Motivation

Example 1. Previously, we saw that there was a significant linear relationship between a an individual's height and their forearm length. Since a linear relationship exists, we would like to be able to predict an individual's height if their forearm length is 9 in. Can we mathematically model the relationship using the class data?

```
R: load("ClassData.RData")
R: attach(class.data)
```

What's the best line through the data?

```
R: plot(forearm, height, main = "Forearm length vs height in ↩
    inches")
```



Definition 1.1 Deterministic model.

A model that can **exactly** predict the value of a variable. (algebra)

Example: The area of a circle can be determined exactly from it's radius: $A = \pi r^2$.

Definition 1.2 Probabilistic model.

A model where one variable an be used to **approximate** the value of another variable. More specifically, one variable is not completely determined by the other variable.

Example: Forearm length of an individual can be used to estimate the approximate height of an individual but not an exact height.

Modeling a linear relationship

Equation of a line: algebra
Recall
$$y = mx + b \tag{1}$$

- x is the **independent** variable.
- y is the **dependent variable**. (Since y depends on x.)
- b is the y-intercept.
- m is the slope

Equation of a line: statistics
We will write the equation of a line as:
$$\hat{y} = b_0 + b_1 x$$

- x is the **predictor variable**.
- $\hat{y}$ is the **response variable**.
- b_0 is the y-intercept
- b_1 is the slope

b_0 and b_1 are sample statistics that we use to estimate the population parameters β_0 and β_1.

RESIDUAL ϵ. Definition 1.3

The residual is the "error" in the regression equation prediction for the sample data. For each (x_i, y_i) observed sample data, we can plug x_i into the regression equation and estimate $\hat{y}_i$. The residual is the difference of the **observed** y_i from the **predicted** $\hat{y}_i$.

$$\begin{aligned} \epsilon_i &= y_i - \hat{y}_i && (2) \\ &= (\text{observed } y) - (\text{predicted } y) && (3) \end{aligned}$$

STEPS FOR REGRESSION

Use the following steps to model a linear relationship between two quantitative variables:
1. Determine which variable is the predictor variable (x) and which variable is the response variable (y).
2. Make a scatter plot of the data to determine if the relationship is linear.
3. Determine if the linear correlation coefficient is significant.
4. Write the model and determine the coefficients (b_0 and b_1).
5. Plot the regression line on the data.
6. Check the residuals for any patterns.

1.2 Simple Linear regression

What is a best fit line?

LEAST-SQUARES PROPERTY. Definition 1.4

We will define the "best" fit line to be the line that minimizes the squared residuals. Thus, the best line results in the **smallest possible** sum of squared error (SSE):

$$\text{SSE} = \sum \epsilon_i^2 \tag{4}$$

The linear regression equation

$$\boxed{\hat{y} = b_0 + b_1 x} \tag{5}$$

Where b_1 and b_0 satisfying the least-squares property are:

$$\boxed{b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}} \tag{6}$$

$$\boxed{b_0 = \bar{y} - b_1 \bar{x}} \tag{7}$$

Finding the regression equation for our example

Example 2. Lets find the regression equation for forearm and height data. The forearm length will be the **predictor** variable (x) and the height will be the **response variable** (y). We need to find b_0 and b_1.

Define needed variables:

```
R: x = forearm
R: y = height
R: x.bar = mean(x)
R: y.bar = mean(y)
```

Find the slope using equation 6:

```
R: b1 = sum((x - x.bar) * (y - y.bar))/sum((x - x.bar)^2)
R: b1
[1] 1.7726
```

Find the y-intercept using equation 7:

```
R: b0 = y.bar - b1 * x.bar
R: b0
[1] 48.811
```

Thus our linear model for this relationship is:

$$\hat{y} = 48.8 + 1.77x$$

MAKING PREDICTIONS

Example 3. Using the results from the previous example, predict the height of an individual if their forearm length is 9 inches.

Use our fitted regression equation and plug in 9 in.

```
R: y.hat = b0 + b1 * 9
R: y.hat
[1] 64.764
```

Thus, the best **point estimate** prediction for the height of an individual with a forearm length of 9 inches is 64.8 inches.

Cautions when making predictions

- Stay within the scope of the data. Don't predict outside the range of sample x values.
- Ensure your model is applicable for what you wish to predict. Is it the same population? Is the data current?

1.3 Regression using R

Rather than typing in the equations for b_0 and b_1 each time, R can calculate them for us:

LINEAR REGRESSION:

```
results=lm(model)
results
plot(x, y)
abline(results)
plot(x, results$resid)
```

R COMMAND

Various models in R:

MODEL TYPE	EQUATION	R MODEL
lin 1 indep var	$y = b_0 + b_1x_1$	`y ~ x1`
...0 intercept	$y = 0 + b_1x_1$	`y ~ 0+x1`
lin 2 indep vars	$y = b_0 + b_1x_1 + b_2x_2$	`y ~ x1+x2`
...inteaction	$y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2$	`y ~ x1+x2+x1*x2`

For simple linear regression use a model: `y ~ x` to indicate that y is linearly related to x. Both x and y are **ordered vectors** of data. Output shows regression coefficients, plots the data with the regression line, and plot residuals.

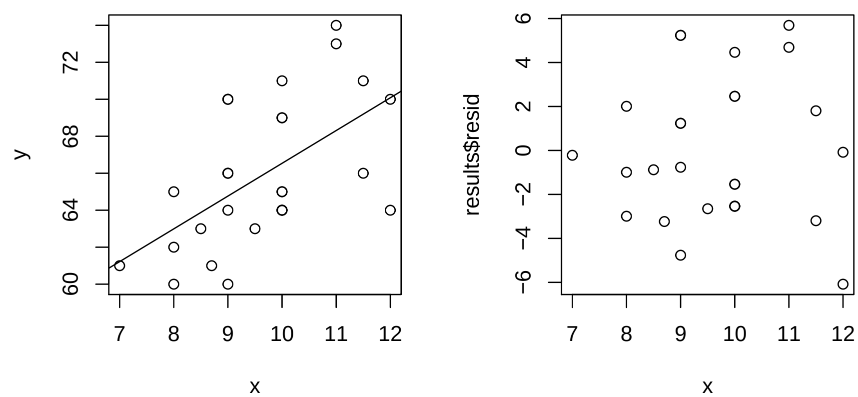
```
R: x = forearm
R: y = height
R: results = lm(y ~ x)
R: results
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      48.81         1.77
```

Plotting the regression equation on the data to check model

Use the following commands:

```
R: par(mfrow = c(1, 2))
R: plot(x, y)
```

```
R:  abline(results)
R:  plot(x, results$resid)
```



RESIDUAL PLOTS

TODO!

1.4 Prediction intervals

R Command

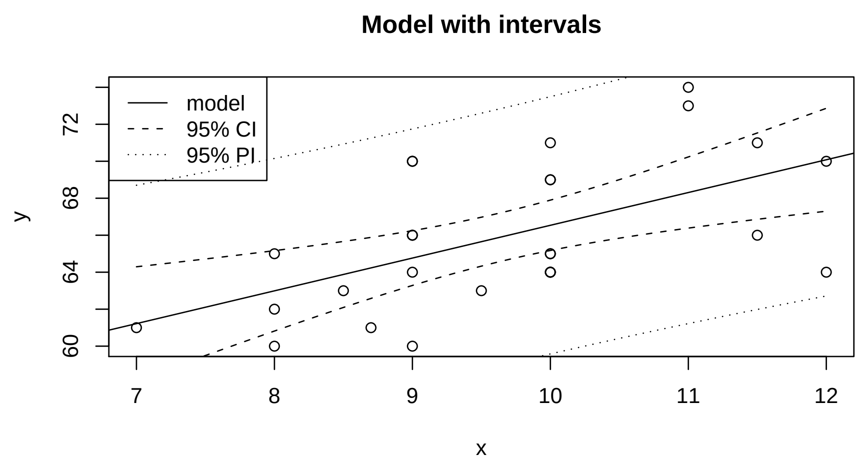
Prediction intervals:
`predict(results, newdata=data.frame(x=9), int="pred")`
Make point estimate and prediction interval for $x = 9$ using regression model stored in `results` .

Example 4. To make a prediction for a forearm length of 9 inches using the previous model in `results` :

```
R:  res = predict(results, newdata = data.frame(x = 9),
+      int = "pred")
R:  res
         fit    lwr    upr
[1,] 64.764 57.782 71.746
```

The best point estimate for the individual's height (in inches) is 64.8. The 95% prediction interval for the individual's height is (57.8, 71.7).

Prediction Intervals & Confidence Intervals



1.5 Multiple Regression

1.6 Summary

Linear Regression and Predictions
Requirements: (1) linear relationship (2) residuals are random (independent), have constant variance across x and are normally distributed.

1. Determine **predictor variable** (x) and **response variable** (y).
2. Check for linear relationship: `plot(x,y)` (otherwise stop!)
3. Check for influential points.
4. Check for statistically significant correlation: `cor.test(x,y)`
   If a significant relation **does not exist**, the best predictor for **any** x is $\bar{y}$.
5. Find the regression equation: `results=lm(y ~ x)` .
6. Plot the line on the data: `plot(x, y); abline(results)`
7. To predict x=10 with a 95% prediction interval: `predict(results, newdata=data.frame(x=10), int="pred")`
   Don't predict outside of sample data x values!

1.7 Additional Examples

Use Data Set 1 in Appendix B:

Example 5. Find a linear model to predict the leg length (cm) of men based on their height (in). Then predict the leg length of a 68 in male. Also, how much variation does the model explain?

Example 6. Find a linear model to predict the cholesterol level (mg) of men based on their weight (lbs) . Then predict the cholesterol of a man weighing 200 lbs.