Overview of Regression Analysis

Sunday May 12, 2019 at 14:45

- Data
    - Read the data
    - View the relevant data
- Linear Model
    - Y-intercept and slope parameters
    - Residuals
    - Estimate the model
- Sample Estimates vs Population Values
    - Inference
    - Forms of inference
    - Standard errors
    - Mechanics
    - Interpretation
- Forecasting
    - Purpose of regression analysis
    - Forecasted value and prediction interval

```
library(lessR)
```

```
##
## lessR 3.8.5     feedback: gerbing@pdx.edu     web: lessRstats.com/new
## ---------------------------------------------------------------------
## 1. d <- Read("")           Read text, Excel, SPSS, SAS or R data file
##                            d: default data frame, no need for data=
## 2. l <- Read("", var_labels=TRUE)   Read variable labels into l,
##                            required name for data frame of labels
## 3. Help()                  Get help, and, e.g., Help(Read)
## 4. hs(), bc(), or ca()     All histograms, all bar charts, or both
## 5. Plot(X) or Plot(X,Y)    For continuous and categorical variables
## 6. by1= , by2=             Trellis graphics, a plot for each by1, by2
## 7. reg(Y ~ X, Rmd="eg")    Regression with full interpretative output
## 8. style("gray")           Grayscale theme, + many others available
##    style(show=TRUE)        all color/style options and current values
## 9. getColors()             create many styles of color palettes
##
## lessR parameter names now include _'s. Names with a period are
## deprecated, but still work. Ex: bin_width instead of bin.width.
```

Data

Read the data

```
#d <- Read("http://lessRstats.com/data/bodyfat10.csv")
d <- Read("~/Dropbox/511Stuff/BookNew/Ch08/data/bodyfat10.csv")
```

```
## Data Types
## -------------------------------------------------------------
## integer: Numeric data values, integers only
## double: Numeric data values with decimal digits
## -------------------------------------------------------------
##
##     Variable                   Missing  Unique
##         Name      Type  Values  Values  Values   First and last values
## ------------------------------------------------------------------------
--------
##  1         ID   integer     10       0      10   228  209  105 ... 156  173  163
##  2        BF1    double     10       0      10   16.1  19.3  17.2 ... 22  13.4  21.
4
##  3        BF2    double     10       0      10   16.1  19.5  17.3 ... 22.5  13.1  2
1.8
##  4    Density    double     10       0      10   1.062  1.0543 ... 1.0689  1.0492
##  5        Age   integer     10       0       8   57  49  43 ... 31  37  35
##  6         Wt    double     10       0      10   182.25  168.25 ... 151  166.25
##  7         Ht    double     10       0       8   71.75  71.75  75.5 ... 71.5  67  6
8
##  8        Adi    double     10       0       8   24.9  23  24 ... 24.4  23.7  25.3
##  9        FFW    double     10       0      10   152.9  135.9 ... 130.8  130.7
## 10       Neck    double     10       0       9   39.4  38.3  38.5 ... 36.2  35.3  3
8.5
## 11      Chest    double     10       0      10   103.4  98.3  110.1 ... 101.1  92.6
99.1
## 12        Abd    double     10       0      10   96.7  89.7  88.7 ... 92.4  83.2  9
0.4
## 13        Hip    double     10       0      10   100.7  99.1  102.1 ... 99.3  96.4
95.6
## 14      Thigh    double     10       0      10   59.3  56.3  57.5 ... 59.4  60  55.
5
## 15       Knee    double     10       0       8   38.6  38.8  40 ... 39  38.1  34.2
## 16      Ankle    double     10       0       9   22.8  23  24.8 ... 24.6  22  21.9
## 17     Biceps    double     10       0      10   31.8  29.5  35.1 ... 30.1  31.5  3
0.2
## 18    Forearm    double     10       0      10   29.1  27.9  30.7 ... 28.2  26.6  2
8.7
## 19      Wrist    double     10       0       7   19  18.6  19.2 ... 18.2  16.7  17.
7
## ------------------------------------------------------------------------
--------
```

We wish to use height to predict weight. The variable names in the data table for these variables are Ht and Wt. So Ht is the X variable, or predictor variable, and Wt is the Y variable, the response variable.

View the relevant data

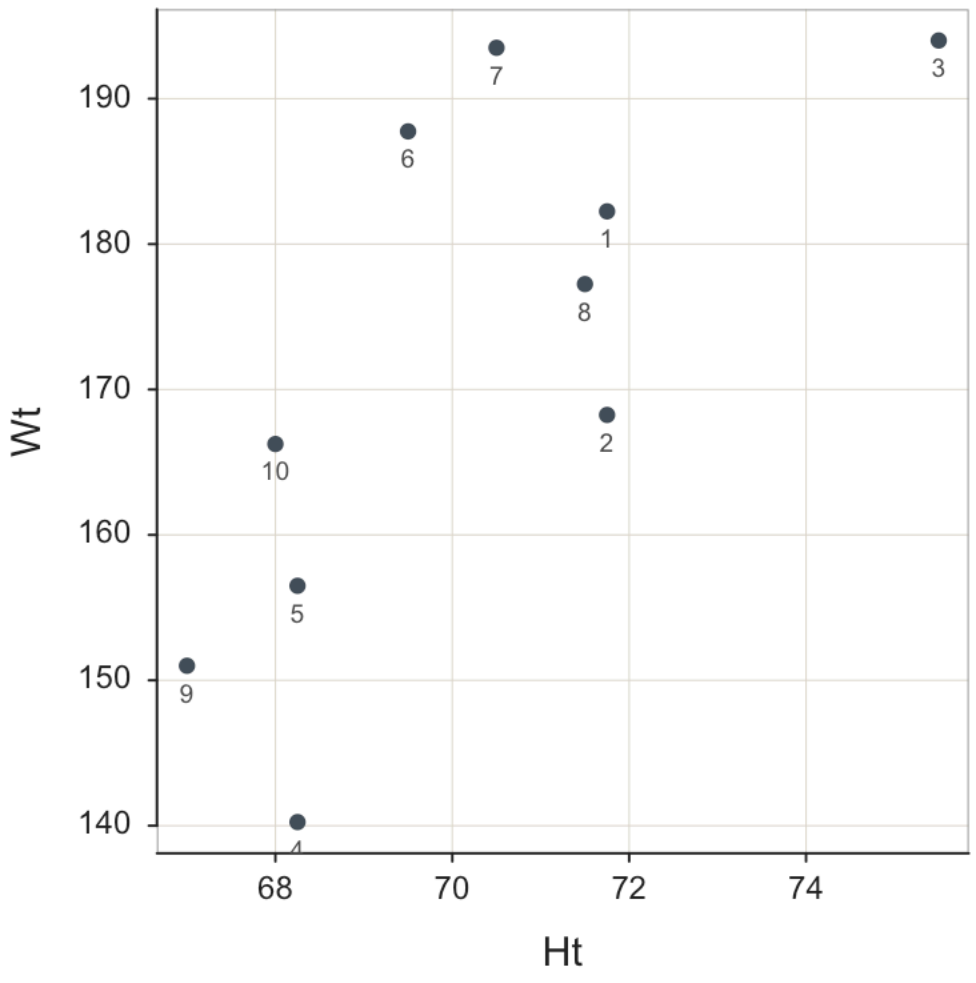
Of the many variables in the data table, here we view just the two of interest, Ht and Wt. To limit the display to only two columns of `d`, use the standard R subset notation of **[rows,cols]** to extract all the rows of `d` and just the two columns name Ht and Wt. Because we list more than a single variable in the subset, need to define a vector of variables with the R **c()** function.

```
d[,c("Ht", "Wt")]
```

```
##        Ht     Wt
## 1   71.75 182.25
## 2   71.75 168.25
## 3   75.50 194.00
## 4   68.25 140.25
## 5   68.25 156.50
## 6   69.50 187.75
## 7   70.50 193.50
## 8   71.50 177.25
## 9   67.00 151.00
## 10 68.00 166.25
```

Each set of paired Ht and Wt values in each row of the data table corresponds to a single point in the scatter plot.

```
Plot(Ht, Wt, add="labels", quiet=TRUE)
```



Linear Model

Y-intercept and slope parameters

Refer to each pair of data points in general as $< X, Y >$, or, more specifically for this analysis, $< X_{Ht}, Y_{Wt} >$. To forecast unknown values of Wt from Ht, such as a task faced by a garment manufacturing firm that seeks to find optimal size patterns, we seek a linear model that summarizes the relation between Ht and Wt.

$$\hat{Y} = b_0 + b_1(X)$$

The regression analysis will estimate the value of the two parameters of the linear model: $b_0 and b_1$. That is, the regression analysis identifies a line through the scatterplot according to the basic definitions of those parameters:
b_0: y-intercept, where the line crosses the y-axis
b_1: slope coefficient, the angle of the line, how much Y increases for a unit increase in X

The meaning of the slope coefficient contributes much to the understanding of the relationship between the variables Y and X. We can estimate how much Y will change as X changes. For this example, how much – on average – does weight increase for each additional inch of height.

Note, however, that the estimated coefficients that describe the regression line, b_0 and b_1, are descriptive statistics. That is, they describe how X relates to Y in this particular sample only.

The issue is that we do not know the true regression model that generated the data. Take another sample and a different regression line results according to *different* estimates of b_0 and b_1. The sample regression line *would* randomly fluctuate from random sample to random sample even though only one sample is typically

taken and only one regression line estimated. That is why a confidence interval for the value of the slope coefficient b_1 is needed.

Residuals

The line has to satisfy criterion of fit, the best-fitting line according to that fit measure. The scatterplot is not a line, hence the word: scatter. So the configuration of points that represent data are not perfectly summarized by a line. How to choose the line? The choice is based on the error, the residual,

$$e = Y - \hat{Y}$$

The actual data value, Y, consists of what is *explained* by the model, what lies on the line, and the residual, e, for what is *unexplained*, the distance from the line.

$$Y = b_0 + b_1(X) + e$$

The model is estimated, that is, the coefficients b_0 and b_1 are chosen, such that the resulting model minimizes the sum of the squared residuals, the *least-squares criterion*.

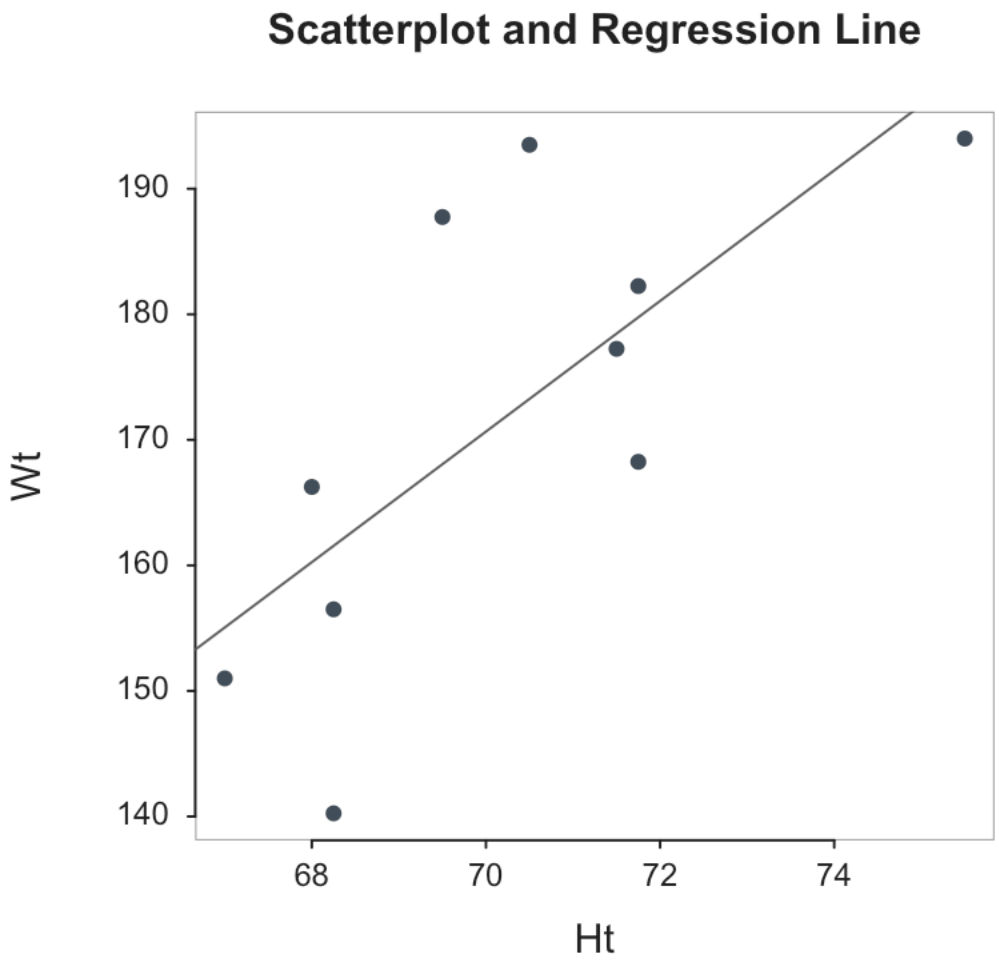
$$\text{choose } b_0 \text{ and } b_1 \text{ to minimize: } \Sigma e^2$$

Any other value of b_0 and/or b_1 results in a sum of squared errors, Σe^2, that is larger than the value obtained with the least-squares coefficients.

Estimate the model

Here we use the brief form of the `lessR` function `Regression()` to estimate the model and provide several other analyses.

```
reg.brief(Wt ~ Ht)
```


Scatterplot and Regression Line
Wt
190
180
170
160
150
140
68
70
72
74
Ht

```
##
##
##   BACKGROUND
##
## Data Frame:  d
##
## Response Variable: Wt
## Predictor Variable: Ht
##
## Number of cases (rows) of data:  10
## Number of cases retained for analysis:  10
##
##
##   BASIC ANALYSIS
##
##              Estimate    Std Err  t-value  p-value   Lower 95%   Upper 95%
## (Intercept)  -193.452    126.340   -1.531    0.164    -484.794      97.889
##          Ht     5.202      1.799    2.892    0.020       1.054       9.349
##
##
## Standard deviation of residuals:  13.681 for 8 degrees of freedom
##
## R-squared:  0.511    Adjusted R-squared:  0.450    PRESS R-squared:  0.307
##
## Null hypothesis that all population slope coefficients are 0:
##   F-statistic: 8.363     df: 1 and 8     p-value:  0.020
##
##
##             df    Sum Sq   Mean Sq   F-value   p-value
## Model        1  1565.226  1565.226     8.363     0.020
## Residuals    8  1497.249   187.156
## Wt           9  3062.475   340.275
##
##
##   RELATIONS AMONG THE VARIABLES
##
##   RESIDUALS AND INFLUENCE
##
##   FORECASTING ERROR
```

Sample Estimates vs Population Values

Inference

We want population values, but only have sample estimates. This is the driving issue of inferential statistics. A sample mean, the estimated mean, is not the corresponding population mean – it is an estimate. Same for the linear model parameters estimated by the least-squares regression analysis. The sample slope coefficient is not the corresponding population coefficient.

Statistical inference: Generalize from a sample estimate to the corresponding population value.

Sample estimates describe data, and inferential analysis generalizes to the population. Every statistical estimate should be presented in conjunction with the corresponding inferential analysis.

In statistics the tradition is to express statistical estimates in terms of roman letters and corresponding population values in the Greek version of the roman letter. The sample estimated slope coefficient is b_1. Designate the corresponding, unknown population value as β_1, which uses the Greek letter beta.

Write the population version of the model we estimate from the data with beta's instead of b's.

$$Y = \beta_0 + \beta_1(X) + e$$

Not only do we know the correct model that actually generated the data, even if we have the model form correctly specified we do not know the values of the regression coefficients, population Y-intercept and slope, from the corresponding linear model.

Forms of inference

There are two forms of (classical) inferential analysis that address inferring the value of an unknown population value, such as β_1:

1. **Hypothesis test**: Plausibility of a specified hypothesized value,
   such as the null hypothesis that $\beta_1 = 0$
2. **Confidence interval**: Range of plausible values at a specified probability that likely contain the true population value,
   such as 95% probability that the true value of β_1 is between 1.05 and 9.35

Ultimately our results concern the population. All conclusions regarding the inferential analyses are in terms of the population.

Standard errors

The key conceptual issue of inferential analysis is the standard error.

Standard error: Standard deviation of a statistic (usually across hypothetical random samples all of the same size)

Why do we care? If the standard error is small, then the sample estimate does not vary much across samples, so any one sample estimate is likely close to the true population value. How much variation? If normal, which is usual, then the two standard error (deviations) rule of 95% variation applies.

Hypothesis test: The hypothesized value is considered plausible if the sample value lies within two standard errors of the hypothesized value, the t-statistic.

Confidence interval: The confidence interval spans two standard errors on either side of the sample value.

Both analyses always are consistent with each other. A rejected hypothesized value will not lie in the confidence interval. A non-rejected hypothesized value is plausible so lies within the confidence interval.

Mechanics

The hypothesis test provides evidence as to the plausibility of the null hypothesis. For the population slope coefficient, is there a relation between X and Y. If not, then as the value of X increases, the value of Y could increase or decrease as an overall pattern.

Assume the null hypothesis of $\beta_1 = 0$ is true. Note you are making an assumption, so *all* conclusions start with this assumption. If the null is false, then expect a sample value, b_1, to be "far" from zero, specifically more two standard errors. The probability of such a result, given the assumption, is called the p-value. If the p-value is less than 5%, then likely there is a relationship.

Interpretation

The interpretations are how research and analysis results are communicated. This is how you will talk to your "boss", that is, the people who hired you and tasked you with this analysis.

1. Communicate the results in terms of population values.
2. Apply the results specifically to the analysis being conducted, do *not* just repeat general definitions.
3. Apply the results in terms of the meaning of the underlying parameter, not just a verbal restatement of the statistical result.

Hypothesis test

The Hypothesis test tells us if a relationship exists, from which we can infer the direction, + or -. If a relationship is plausible, that is, $\beta \neq 0$, then simply state that X is related to Y, using the actual variable names, not the generic X and Y. Moreover, state the qualitative nature of the relationship, + or -. Your "boss" does not just want to know if there is a relationship, but also in what direction.

Interpret HT of β: As Height increases, on average, Weight increases.

Use statistical jargon like "null hypothesis" and "p-value" for statistical reasoning, but no jargon for interpreting the results.

Confidence interval

The confidence interval is of particular interest if a relationship is plausible because the confidence interval provides an estimate of the plausible size of the relationship. Interpret this confidence interval in terms of the meaning of the slope coefficient.

Interpret CI of β: At the 95% level of confidence, for each one inch increase in Height, on average, weight likely increases somewhere from slightly more than 1 pound to 9 1/3 pounds.

The sample slope coefficient b_i describes this relationship for the sample. We want this analysis for the unknown population value β_1. We do not know its value, but we have a probability (such as 95%) that the value lies within the specified range.

Forecasting

Purpose of regression analysis

Conduct a regression analysis to achieve some combination of two goals.

1. Understand the relations among predictor variables X and response Y

such as how cargo weight impacts MPG
2. Predict (forecast) unknown values of Y from predictor variables X

Usually we prefer to gain knowledge for both objectives for any given analysis, though any particular analysis may focus more on of the goals. We wish, for example, to predict MPG given cargo weight, but also good to understand the nature of the relationship, how much MPG changes as as cargo weight increases.

The interpretation of the slope coefficient focuses on the first objective. Generating the forecasted value and an accompanying prediction interval attains the second objective. In both cases we follow the rule to report a statistical estimate with a band of uncertainty: the *confidence interval* for a statistic such as a slope coefficient, and a *prediction interval* for a single data value such as a forecasted value.

Forecasted value and prediction interval

A key consideration in assessing how well the model forecasts is to realize that the estimation of the model coefficients, intercept and slope, with an algorithm such as least-squared residuals, is optimal for the data on which the estimates were computed, the *training data*. The model generally does a better job of fitting its computed $\hat{Y}$'s to the same data than it does on new data because chooses the model by minimizing the sum of squared errors, Σe^2, for the training data.

Performance generally weakens when the model is applied to new data, often called *testing data*. If forecasting performance decreases much, the estimated model has modeled statistical noise in the training data so that the relations specified by the model do not generalize to different data.

Modeling statistical, random noise in the data, sampling error, that does not repeat from sample to sample is called *overfitting* the training data. One contribution to overfitting is sample size. Small samples are particularly susceptible to overfitting. The least-squares algorithm "works hard" to minimize Σe^2, but in doing so takes advantage of chance to produce an outcome that is optimized only to that particular sample.

The prevention of overfitting is a prime consideration to building robust and accurate forecasting models. Compare the following two ways to apply the regression model, somewhat confusing because the same notation applies to the same computation, but different concepts.

fitted $\hat{Y}$ = model $b_0 + b_1(X)$ applied to *training data*
forecasted $\hat{Y}$ = model $b_0 + b_1(X)$ applied to *testing data*

As an example, we return to our simple Ht-Wt analysis. Here re-run the analysis to obtain forecasts and prediction intervals for two different heights, neither of which are in the original data table: 67.5 inches and 71.0 inches.

This analysis requires the full `Regression()` output instead of the brief form illustrated previously. In this example, turn off the default visualizations, and save the output to an object called `r` and then retrieve just the forecasting part so we can focus on just that part of the analysis. Find the names of all the pieces in the R object `r` from `names(r)`.

```
r <- reg(Wt ~ Ht, X1.new=c(67.5, 71.0), graphics=FALSE)
r$out_predict
```

```
## Data, Predicted, Standard Error of Forecast, 95% Prediction Intervals
##    [sorted by lower bound of prediction interval]
## -------------------------------------------------
##         Ht  Wt    pred     sf  pi:lwr  pi:upr  width
##   1 67.500     157.656 15.148 122.725 192.587 69.862
##   2 71.000     175.861 14.420 142.608 209.114 66.506
```

Interpret Prediction Interval for Height=67.5 inches: With 95% confidence, the forecasted weight for an adult male who is 67.5 inches tall would vary from 122.7 lbs to 192.6 lbs.

The prediction interval is almost 70 lbs wide. Note that this prediction interval is so wide as to be meaningless (not surprising given the small sample size and wide variability of the data around each point on the regression line). And that is the purpose of the prediction interval, to communicate the accuracy of the forecast.

A single forecasted value can always be obtained, in this case, $\hat{Y} = 157.66$ lbs for a height of 67.5 inches. But the actual value of Y when obtained would not equal exactly its forecasted value, $\hat{Y}$. Analysis of the size of the prediction interval provides much insight into the usefulness of the forecast.

Want to decrease the size of the prediction interval? The means to accomplish this objective, given the large variability of the data at each value of X, would be to bring additional information to the model. One way to add information is to add additional predictor variables that are related to Y, weight, but not so related to each other. That way each additional predictor variable contributes new information to the understanding of the value of Y. This analysis is called *multiple regression*.