Chapter 9
Assessing Relationships with Correlation and Regression

Section 9.1
Multiple Regression

School of Business Administration
Portland State University

9.1a
The Model

Goals of a Regression Analysis

Need unique and relevant predictor variables

- **Regression model**: Relate one or more predictor variables to a response variable y, such as, for one predictor, $\hat{y} = b_0 + b_1 x_1$
- The analysis of a regression model has two primary purposes
  - Forecast the value of y, $\hat{y}$?: From the value of each predictor variable forecast (predict) unknown values of response y
  - Explain the value of y such as with the slope coefficients, b_j: Show relationship of each predictor variable with y, *with the values of all other predictor variables held constant*
- To enhance these goals, add predictor variables to the model
- Choose predictor variables that satisfy the following conditions
  - New Information: A proposed predictor variable is relatively uncorrelated with the predictor variables already in the model
  - Relevant Information: A proposed predictor variable correlates with y

Multiple Regression: Organization of the Data

Data into rows and columns

	X1	X2	y
1	19	9.47	1.45
2	16	8.29	1.59
⋮			
50	15	7.88	1.13

Table: Excerpt from data table for multiple regression with two predictor variables and 50 rows of data

- **Observation** or case: Data for a single row, that is, for a single person or company or whatever the unit of analysis
- One column for each predictor variable, X_j, here $j = 1, 2$
- One column for the response variable, y

The Multiple Regression Model

As many predictors as desired, usually up to six or so

- **Multiple regression model**: The fitted value of y as a function, here linear, for a set of m predictor variables

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m$$

- **Partial slope coefficient**: The slope coefficient, b_j, for the j^{th} predictor variable from a multiple regression model
- **Key Concept**: The value of each b_j changes depending on the other predictor variables that are in the model
- As an example, consider Age, Height and Weight for a sample of 100 men and two different regression models
  - $\hat{y}_{Wt} = -182.59 + 5.18 x_{Ht}$
  - $\hat{y}_{Wt} = -209.82 + 5.43 x_{Ht} + 0.21 x_{Age}$
  - $b_{Ht} = 5.18$ from 1^{st} model $\neq b_{Ht} = 5.43$ from 2^{nd} model

9.1b
Statistical Control

Assessing the Relation of Each Predictor Variable with y

The next best thing to experimental control

- How does one variable x causally impacts the response variable y?
- **Key Concept**: The partial slope coefficient, b_j, is the *average* change in y for each unit increase in x_j *with the values of all other predictor variables held constant*, i.e., controlled
- **Ceteris Paribus** (with other things the same): Partial slope coefficient b_j isolates the relation of predictor x_j to response y with effects of the remaining predictor variables held constant
- Holding the values of the other variables constant mimics the equivalence attained with experimental control
- To assess the effect of
  - house selling price according to size in square feet, control by the age of the house
  - salaries by gender, control by years of work experience

Control of Confounding Variables and Casual Analysis

The next best thing to experimental control

- **Key Concept**: To demonstrate causality requires controlling for confounding variables that lead to spurious relationships
- *Experimental control* with randomization and manipulation is always preferred, but not always possible
  - cannot manipulate Gender to study impact on Salary
  - cannot manipulate the MBA program a student attends to study impact on Salary
- **Statistical control**: Observe the relation between two variables y and x_1 with the value of one or more other potential confounding variables, $x_2, x_3 \ldots$ explicitly held constant
- Accomplish statistical control with some version of multiple regression, that is, with two or more predictor variables

Warning: Regression Not Necessarily Causal Analysis

Regression reflects causality only in special circumstances

- Many people are seduced by the mathematical expression of a regression model into falsely concluding that the equation expresses a causal relationship
- The seduction is understandable
  - enter a value of X into the equation, out comes a value of y
  - enter a different value of X, out comes a different value of y
- **Key Concept**: Mathematical manipulation of values in an equation does not necessarily correspond to a causal effect in the real world
- The estimated regression model depends on the correlations of one or more X variables with each other and with y, and these correlations could be spurious

Ex: Regression Not Necessarily Causal Analysis

Regression reflects causality only in special circumstances

- Consider correlation of the Number of Fire Trucks and Economic Damage . . .
  - the two variables are correlated
  - correlation is spurious (not causal), due to a common factor that causes both variables: Severity of the Fire
  - Could regress Economic Damage onto Number of Fire Trucks, and find a statistically significant slope coefficient
  - but, more fire trucks does *not* cause more damage
- The slope coefficient b_1 shows how damage increases with more fire trucks, understanding that b_1 provides information regarding the correlation, but not causality

Ex: Regression Not Necessarily Causal Analysis

Regression reflects causality only in special circumstances

- Estimate, from the data gather at many different fires, Economic Damage (D) in \$1000's as a function of the Number of Fire Trucks (F) at the fire
- Consider the following *hypothetical* result, $b_0 = 10$ and $b_1 = 85$

  $\hat{y}_D = 10 + 85x_F$
- The corresponding slope coefficient $b_1 = 85$ specifies that, on average, the Damage increases by \$85,000 for each additional Fire Truck present at the fire
- Yet the positive slope coefficient does not imply causality because Fire Trucks do not cause damage, in fact, their presence minimizes the damage
- Control for the effect of Severity of the Fire, include it as a second predictor variable, and the effect of Number of Fire trucks will disappear with b decreasing from 85 to $b \approx 0$

9.1c
Analysis

I: Overview of Analysis, Input

Criterion for Estimating the Multiple Regression Model

As many predictors as desired, usually up to six or so

- For one predictor variable, or many, implement the OLS algorithm via computer to choose the values of the b_j's from the training data that together minimize the sum of squared residuals, the training errors

  $$\sum(y_i - \hat{y}_i)^2 = \sum e_i^2$$

  across all of the observations for that particular data set
- Can calculate $\hat{y}$ from the values of the predictor variables for each observation, that is, for each row of the data table
- Can then calculate the residual error term, $y_i - \hat{y}_i = e_i$, for each observation

Inferential Analysis for the Slope Coefficient

What is the value of each population slope coefficient?

- Everything done so far in this discussion of regression, such as obtaining the sample slope coefficients, b_j, has been in terms of descriptive statistics only
- Is there a relationship *in the population* between the j^{th} predictor variable, x_j, and the response variable y, with the values of all other variables in the model held constant?
- The *population* regression model for a set of m x_j's is

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$$

- The focus of the inferential test for each β_j is on 0
  - If $\beta_j < 0$, a negative relationship
  - If $\beta_j = 0$, no relationship
  - If $\beta_j > 0$, a positive relationship

R: Multiple Regression

Simple extension of regression with one predictor variable

- The `lessR` function `Regression` provides regression analysis, here illustrated for response variable y and three predictor variables, x1, x2 and x3

```
> Regression(y ~ x1 + x2 + x3)
```

- The `reg` output is extensive
  - Partial slope coefficients b_0, b_1, b_2 and b_3 and corresponding hypothesis tests and confidence intervals
  - Correlations among all the variables in the model
  - If one predictor variable, scatter plot with regression line
  - If multiple predictors, scatter plot matrix and a version of R^2 for all possible models from the predictor variables
  - Forecasting error intervals
  - Residuals and influence statistics (defined later)

Example Data Set

Input for the regression analysis

- Data table: `http://lessRstats.com/data/bodyfat10.csv`
- The data set consists of some body measurements of 10 men
- There are many variables, of which some values for Weight, Height and Age are listed below

```
       Wt    Ht Age
1  182.25 71.75  57
2  168.25 71.75  49
...
10 166.25 68.00  35
```

- To account for Weight in terms of Height and Age, run the analysis with: `> Regression(Wt ~ Ht + Age)`

II: Output, Interpret Estimates

R: `Regression` Output, General Form of the Estimates

R and Excel output for this basic analysis are similar

- For a basic regression analysis, equivalent to that provided by Excel, use reg.brief in place of Regression
- Basic output is presented as a table, illustrated on the next slide, with one row for each b_j, starting with the intercept, b_0
  - The first column is a list that begins with the intercept followed by each of the predictor variables
  - The second column contains each estimate b_j
  - The third column lists the standard error of each b_j
  - The next two columns list the corresponding t-values and p-values for each $H_0 : \beta_j = 0$
  - The last two columns are the lower and upper bounds for the corresponding 95% confidence intervals for each β_j

R: `Regression` Output, The Estimates and Inference

R and Excel output for this basic analysis are similar

- The output displays in one block of lines, but here listed in two blocks to fit on the page

```
             Estimate    Std Err  t-value  p-value
(Intercept) -191.634    138.723   -1.381    0.210
         Ht     5.155      2.089    2.468    0.043
        Age     0.036      0.630    0.057    0.956
```

```
             Lower 95%   Upper 95%
(Intercept)    -519.661     136.393
         Ht       0.216      10.094
        Age      -1.453       1.525
```

The Estimated OLS Model

Rewrite the tabular output as a standard linear equation

- One task to perform after obtaining the analysis is to write the model that has been estimated
- That is, obtain the specific values for the estimated coefficients in the regression model, b_0, b_{Ht}, b_{Age}
- In this example, the model, as read from the R output, follows

$$\hat{y}_{Wt} = -191.63 + 5.15x_{Ht} + 0.04x_{Age}$$

- Now, the questions of interest involve the interpretation and inferential analyses, particularly of the slope coefficients
- Including Age in the model provides a statistical control for Age when assessing the relation of Height to Weight
- The resulting partial slope coefficient for Height indicates the relation of Height to Weight without any variation in Age

R: `Regression` Output, 1st Partial Slope Coefficient

Conclusions for Height

- **Descriptive Result**: $b_{Ht} = 5.15$, so for these measurements of 10 men, each increase in Height of one inch leads to an average increase of 5.15 lbs, for all men of the same Age
- Further, the p-value that evaluates the null hypothesis that $H_0 : \beta_{Ht} = 0$ is small, p-value < 0.043, so reject H_0
- **Interpretation** of Hypothesis Test: A relation of Height and Weight is detected for men of the same Age
- Consistent with this result, the 95% confidence interval only contains positive values, from 0.22 lbs to 10.09 lbs
- **Interpretation** of Confidence Interval: At the 95% level of confidence, on average, associate each 1 inch increase of Height with an increase in Weight somewhere from 0.22 lbs to 10.09 lbs, for all men of the same Age

R: `Regression` Output, 2nd Partial Slope Coefficient

Conclusions for Age

- **Descriptive Result**: $b_{Age} = 0.04$, so for these measurements of 10 men, each increase in Age of one year leads to an average increase of 0.04 lbs, for all men of the same Height
- However, the p-value that evaluates the null hypothesis that $H_0 : \beta_{Age} = 0$ is large, p-value $= 0.956 > \alpha = 0.05$, no reject
- **Interpretation** of Hypothesis Test: No relation of Age and Weight detected for men of the same Height
- Consistent with this result, the 95% confidence interval, which is from -1.45 lbs to 1.53 lbs, contains 0
- **Interpretation** of Confidence Interval: Conclude that the relation of Age on Weight is not detected, at least when all men are of the same Height
- Perhaps there is an effect, but the effect is reasonably small and this analysis did not have enough power to detect it

III: Assess Model Fit

R: Regression Output, Model Fit Index, s_e

Analyze the size of the residuals

- To analyze the efficacy of the *entire* model, not just the individual partial slope coefficients, examine the fit indices

```
Standard deviation of residuals:
  14.62 for 7 degrees of freedom
If normal, the approximate 95% range of residuals about each
  fitted value is 2*t-cutoff*14.62165, with a 95% interval
  t-cutoff of 2.365
95% range of variation: 69.15
```

- The $s_e = 14.62$ lbs for $df = n - 3 = 10 - 3 = 7$
- **Descriptive Result**: For a sample of data in which the model minimizes the sum of squared residuals of Weight about the fitted value for specific values of Height and Weight, 95% of the fitted Weights are estimated to span a range of 69.15 lbs, much too large to be of practical use

R: Regression Output, Model Fit Index, R^2

Compare the size of the residuals to those of the null model

```
R-squared:  0.511     Adjusted R-squared  0.372

Null hypothesis that population R-squared=0
  F-statistic: 3.662     df: 2 and 7     p-value:  0.082
```

- $R^2 = 0.51$, so in this sample a reduction in training error moving from the null model of Weight to the current model with predictor variables Height and Age
- **Statistical Decision**: For the test of the null hypothesis, H_0 : population $R^2 = 0$, the p-value is small, 0.082, so do not reject the null hypothesis that the population $R^2 = 0$ and conclude that no difference of R^2 is detected from 0
- **Interpretation**: Using Height to account for Weight, with Age held constant, does not significantly reduce the residuals than a model without these two predictor variables

Compare Fit for 1- and 2-Predictor Variable Models

More predictor variables, better fit

- The addition of a predictor variable to a model for a given data set increases the fit of the model
  - The sum of squared residuals, $\sum e_i^2$ always goes down
  - R^2 goes up
- In this example, moving from Height to Height and Age in a model for Weight
  - $\sum e_i^2$ dropped from 1497.2 lbs to 1496.5 lbs
  - R^2 rose from 0.5111 to 0.5113
- Fit will always increase, but the increase will be trivial, as in this example, unless the new predictor variable
  - adds new information to the model: it does not correlate much with the predictor variables in the model
  - is relevant: it correlates with the response variable

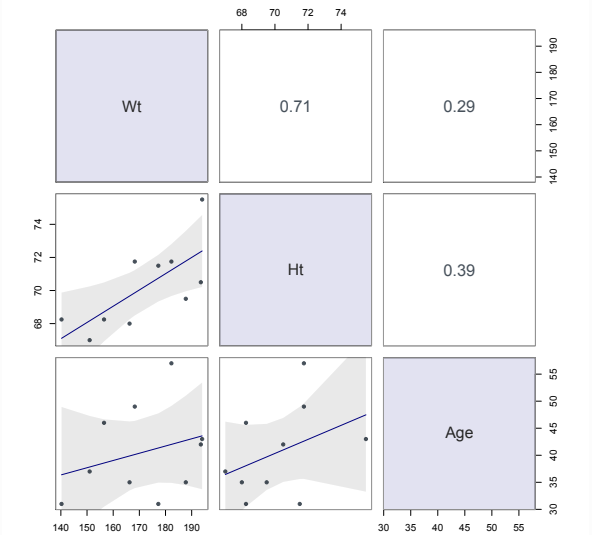
Scatterplot Matrix

Plot of multiple scatterplots on one graph

- As previously discussed, a scatterplot is a graph of the relation between two numeric variables
- With multiple regression, there are more than two variables, and, of course, when there are more than two variables, there is more than one scatterplot
  - For example, for three variables there is a scatterplot between Variables 1 and 2, 1 and 3 and 2 and 3
- **Scatterplot matrix**: The display of a scatterplot for each pair of variables, all on one graph

- The scatterplot matrix of all the variables in the model is automatically displayed by the less `Regression` function

R: `Regression` Output, Scatterplot Matrix



Forecasts from New Data

Have R calculate a forecast for any values of the predictors

- By default, the prediction interval is provided for each set of values for the predictor variables in the data
- One of these prediction intervals is applicable to a forecast from new data if the values of the predictor variables in the new data matches the given value of the predictor variables in the original data
- To have `Regression` provide prediction intervals for new data, specify the new data values for the predictor variables using the options `X1.new`, `X2.new` ... up to `X5.new`
  - Can specify the values individually with the combine function, such as `X1.new=c(2.3,4.1)`
  - Can specify the values systematically with the sequence function, such as `X1.new=seq(0,4,0.25)`, which specifies a range of values from 0 to 4 in intervals of .25

`Regression` Output: Forecasts from New Data

- Specify new values of X, Ht: 64, 65 and 65 inches, and two new values of Age: 40 and 50

```
> reg(Wt ~ Ht + Age, X1.new=c(64:66), X2.new=c(40,50))

  Data, Fitted Values, Confidence and Prediction Intervals
     [sorted by lower bound of prediction interval]
 ------------------------------------------------------------
  Ht   Age Wt fitted ci:lwr ci:upr pi:lwr pi:upr  width
  64 50.00    140.08 100.22 179.94  87.31 192.84 105.53
  64 40.00    139.72 107.52 171.92  92.47 186.96  94.49
  65 50.00    145.23 109.79 180.68  95.72 194.75  99.03
  65 40.00    144.87 117.27 172.48 100.63 189.12  88.49
  66 50.00    150.39 119.19 181.58 103.82 196.95  93.13
  66 40.00    150.03 126.87 173.18 108.42 191.64  83.23
```

- Ht and Wt is for new data with unknown values of Wt for the estimation algorithm, so the values for Wt are empty in the output

- The End