3.2 Least–Squares Regression

Idea:

Goal:
Minimize the sum of the squared vertical distances between y_i and $\hat{y}_i$ for $i = 1, \ldots, n$, i.e.,

$$\text{minimize } \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \text{ minimize} \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$$

The "best" choice of all possible lines thet meets the criteria above is the **least squares line** $\hat{y} = b_0 + b_1 x$ where the **slope** b_1 is calculated as

$$b_1 = \frac{SS(xy)}{SS(x)}$$

and the **y–intercept** b_0 is calculated as

$$b_0 = \overline{y} - b_1 \overline{x}.$$

Note:
The textbook uses a different formula to calculate b_1. However, both formulas are equivalent. No matter which formula you use, you will get the same numerical results. Once again, using $SS(xy)$ and $SS(x)$ has some numerical advantages compared to the textbook version.

Example: "The Cost of Victories"

i	Payroll (x_i)	Wins (y_i)	x_i^2	y_i^2	$x_i y_i$
1					
2					
3					
4					
5					
6					
7					
8					
	$\sum_{i=1}^{n} x_i =$	$\sum_{i=1}^{n} y_i =$	$\sum_{i=1}^{n} x_i^2 =$	$\sum_{i=1}^{n} y_i^2 =$	$\sum_{i=1}^{n} x_i y_i =$

Then,

$$SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} =$$

$$SS(y) = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} =$$

$$SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} =$$

Finally, $b_1 = \frac{SS(xy)}{SS(x)} =$

$b_0 = \overline{y} - b_1\overline{x} =$

The least squares regression line (or equation) is

$$\hat{y} =$$

Interpretation:
For every additional million on the payroll, we can expect an average of 0.318 additional wins.

Prediction:
We can use the least squares regression equation

$$\hat{y} = 5.423 + 0.318x$$

to **predict** the expected number of wins:

x	$\hat{y}$
20	wins predicted
40	wins predicted
60	wins predicted
0	wins predicted ??

Note:
Be careful when predicting — prediction is only meaningful where we have data (and a little bit beyond). The behavior of a relationship (linear or of any other type) might change dramatically in a region where we have no data. Extrapolation should be avoided whenever possible. For the "Cost of Victories" data set, it is nonsense to predict about 5 wins in case of a payroll of \$ 0.00.

3.3 Correlation

Pearson's Correlation Coefficient r is a measure that indicates how "close" the data points fall to a straight line. It is calculatd as

$$r = \frac{SS(xy)}{\sqrt{SS(x)}\sqrt{SS(y)}}.$$

Note:

(i) It always holds that $-1 \leq r \leq 1$.

(ii) r is cloas to $+1$ when the data points fall close to a straight *rising* line.

(iii) r is close to -1 when the data points fall close to a straight *falling* line.

(iv) $r = +1$ when all data points fall *exactly* on a straight *rising* line.

(v) $r = -1$ when all data points fall *exactly* on a straight *falling* line.

(vi) r is close to 0 when there is *little* (or none) linear relationship between x and y.

Example: "The Cost of Victories"

$$r = \frac{SS(xy)}{\sqrt{SS(x)}\sqrt{SS(y)}} =$$

Conclusion:
There is some positive linear relationship between payroll and wins. But fortunately, r is not too close to $+1$, i.e., the relationship is not very strong. Where would the excitement in sports be if the underdog could not beat the favorite team any longer?

3.4 Residuals (or Errors)

The differences between the **observed values** y_i and the **predicted values** $\hat{y}_i$ are called **residuals** (or **errors**) e_i, i.e.,

$$e_i = y_i - \hat{y}_i.$$

Example: "The Cost of Victories"

i	Payroll (x_i)	Wins (y_i)	Predicted Wins ($\hat{y}_i$)	e_i	e_i^2
1	31	8			
2	49	17			
3	56	20			
4	33	14			
5	55	24			
6	60	29			
7	45	24			
8	22	19			
				$\sum_{i=1}^{n} e_i =$	$\sum_{i=1}^{n} e_i^2 =$

Note:
$\sum_{i=1}^{n} e_i = 0$, i.e., the sums of the residuals is always equal to 0 when we fit a least squares regression line to our data. Small deviations from 0 are possible with real data due to rounding errors. However, when your residuals sum up to something completely different from 0, something went wrong in your calculations.

Residual Plots:
Plot your e_i's versus the x_i's first and see if there is an overall pattern. Also plot residuals against additional variables such as the number of each observation (i.e., 1 through n), the predicted values, **lurking variables** which often are known divisions in the data (e.g., different days on which the data has been obtained, different persons that called the data, etc.), and so on. When we see any regular pattern in a residual plot, something is wrong with our statistical analysis. In the situation of a least squares regrssion line this means that the fitted line does *not* fully describe the relationship between x and y. Patterns that can frequently be seen in residual plots are shown in the Vardeman handout. A possible explanation for each of these patterns is given.

Outliers versus Influential Observations:

An **outlier** is a point outside the overall pattern in a scatterplot (or any other statistical display). It a scatterplot, this can be a point that is unusually large (or small) in x and/or in y direction.

An **influential point** in a scatterplot is a point that, if removed, would cause the regression line fitted to the data to change considerably. Influential points are most often outliers in the x direction (but rarely in the y direction). They tear the regression line towards them so they typically cannot be detected in a residual plot.

The effect that an individual point can have to the regression line can be seen at

`http://www.stat.sc.edu/~west/javahtml/Regression.html`

Association and Causation:

Even if an explanatory variable x and response variable y are highly correlated (i.e., associated), a change in x does not necessarily cause a change in y. There may be an extra (unknown) variable which is more relevant to the outcome of y, but we can't measure it or do not even know about it.

Example:

German car insurance companies found out that cars that are parked in a garage get less frequently involved in accidents than cars that are parked outside a garage. This is the association — but does a car that is parked get involved in any accident at all? Typically, only cars that are driven have accidents ...

However, there is a reason why insurance companies ask the question whether people park their car in a garage before they determine the insurance rate. The question that relates to one of the main reasons for accidents cannot be asked in a questionnaire, i.e., whether drivers drive reckless. Garage space is very expensive in Germany. Often, only those people that care a lot for their car want to pay the extra money for a garage. Those people usually also take more care when driving, e.g., are more careful when entering or departing a parking lot, wait a bit longer before turning at intersections, etc.

So, there is an association between garage and accidents. However, the number of accidents will remain the same even if everybody gets a free garage since the main cause for accidents is the driving style. And this is unobservable or cannot easily be determined.

3.5 Relations in Categorical Data

Example: "Years of School Completed, by Age"
(from Moore/McCabe, Table 2.14, Page 194)

This data is represented in a **Two–way table**:

Row Variable: Education
Column Variable: Age Group

- We can look at the distribution of each variable separately.
  - These distributions show how often each outcome occurs.
  - The row and column totals give us the info about the distribution of each variable separately.
  - Since we read these numbers from the margins of the table, these distributions are called **marginal distributions**.
- Instead of working with numbers or counts, we usually work with percentages. The following relationship holds:
  - counts $\longrightarrow$ frequency
  - percents $\longrightarrow$ relative frequency

  We can present each of the marginal distributions in a bar chart (see Figure 2.38 for "Years of School")
- If we want to analyze relationships, we have to calculate these numbers for each category separately, i.e., we have to determine the **conditional distributions**. We use the term "conditional" because this distribution refers only to those people that fall into a particular age group (if we condition by "Age Group") or into a particular education level (if we condition by "Education").
  - The conditional distributions can be easily distributed by statistical software, e.g, SAS (see Figure 2.39).
  - A possible graphical representation is the use of multiple bar charts (see Figure 2.39) or mosaic plots (see Homework).

Mosaic Plots:

- another way to present multidimensional (≥ 2 dimensions) categorical data
- famous example: "Titanic Data"
  see Handout from Schmelzer (1997): "Processing Text Information in the Highly Interactive XploRe Environment", Draft.
- other example: Medical Data

Variable	Categories
cause of death	cancer/other
sex	male/female
age	under 65/over 65
smoker	yes/no
...	...

- typically, each variable has 2 to 6 categories (or factors)
- it is easy to distinguish about up to 64 or 70 cells in a Mosaic Plot — if we have more cells, it is often very difficult to see something in a Mosaic Plot