Chapter 8

Regression methods

Reading:


8.1 Pearson’s sample correlation

Suppose we have the following data collected in pairs:

\((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\).

Let

\[ S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad \text{and} \quad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \]

Then (Pearson’s) sample correlation is

\[ r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}. \]

It measures the linear association between \(x\) and \(y\), and has the following properties:

- \(-1 \leq r \leq 1\).
- \(r\) does not depend on the labeling of \(x\) and \(y\).
- \(r\) is independent of the units of measurement.
• $r$ is sensitive to outliers.

We use the cor function in R to calculate the sample correlation.

### 8.1.1 An example

Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting and exercise. Data was collected on the lean body mass and metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass given in kilograms is a person’s weight leaving out all fat. Metabolic rate is a measured in calories burned per 24 hours, the same calories used to describe the energy content of foods. The researchers believe the lean body mass is an important influence on metabolic rate.

Do/answer the following using R:

(a) Read in the data from the class website. Store it as a data frame called metabolic.

(b) Make a scatterplot of the data, using different colors and/or symbols for women and men. Add a legend to the graph using the legend function. See Figure 8.1 for an example.

(c) Do you think that the correlation will be the same for men and women? Why or why not?

(d) Calculate $r$ for women alone and also for men alone using the cor function.

(e) Calculate the mean body weight for women and the men. Does the fact that men are heavier than the women on the average influence the correlations?

(f) Would the correlations differ if we changed from measuring the body mass in kilograms to the body mass in pounds?

### 8.2 Simple linear regression

Now suppose we want to explain $y$ (the response or dependent variable) in terms of $x$ (the explanatory, independent or predictor variable). We wish to regress $y$ on $x$.

**Data:** $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Consider the simplest model for these data – fitting a straight line of the form

$$Y_i = \alpha + \beta x_i + \epsilon_i.$$
Here $\alpha$ and $\beta$ are the **parameters** of the model. We assume the errors $\{\epsilon_i : i = 1, \ldots, n\}$ are a set of independent random variables with $\epsilon_i \sim N(0, \sigma^2)$, for each $i$. The error in fitting this line to the data is called the **residual sum of squares**:

$$\text{RSS} = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2.$$  

We wish to find the values of $\alpha$ and $\beta$ which minimize this error.

**Game plan**: simplify, differentiate (with respect to $\alpha$ and $\beta$) and set equal to zero. Solve the resulting equations.

**Exercise**: Show that the estimate of the **slope** of the line is

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = r \left( \frac{s_y}{s_x} \right),$$

and the estimate of the **intercept** is given by

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}.$$
where, e.g., $s_x$ is the standard deviation of the $x$ values.

**Exercise:** Using the summary statistics calculated for the female metabolic rates, fit a straight line model to relate Mass and Rate using the equations above.

### 8.2.1 Fitting the straight line model directly in R

Let $Y_i$ denote the metabolic rate and let $x_i$ be the mass for female $i$ ($i = 1, \ldots, n = 12$). We fit the model:

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

where $\{\epsilon_i : i = 1, \ldots, n\}$ is a sequence of independent and identically distributed (iid) $N(0, \sigma^2)$ random variables.

To fit the model in R we use

```r
## fit a linear regression model that predicts the mean metabolic rate
## for the females using the mass as an explanatory variable.
female.model <- lm(metabolic$Rate[is.female] ~ metabolic$Mass[is.female])
```

A shorter form of this command is

```r
## refit the command using the 'data' option to select the data frame.
female.model <- lm(Rate[is.female] ~ Mass[is.female], data=metabolic)
```

Or

```r
## refit the command using the 'data' option to select the data frame.
female.model <- lm(Rate ~ Mass, data=metabolic, subset=is.female)
```

To summarize the model we use

```r
## summarize the statistical model.
summary(female.model)
```
Here is the model output:

Call:
\texttt{lm(formula = Rate \sim Mass, data = metabolic, subset = is.female)}

Residuals:
\begin{tabular}{cccccc}
Min & 1Q & Median & 3Q & Max \\
-95.87 & -82.28 & 16.28 & 33.62 & 207.74 \\
\end{tabular}

Coefficients:
\begin{tabular}{cccccc}
Estimate & Std. Error & t value & Pr(>|t|) \\
(Intercept) & 201.162 & 181.701 & 1.107 & 0.294169 \\
Mass & 24.026 & 4.174 & 5.756 & 0.000184 *** \\
\end{tabular}

---
Signif. codes: 0 ***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 95.08 on 10 degrees of freedom
Multiple R-squared: 0.7682, Adjusted R-squared: 0.745
F-statistic: 33.13 on 1 and 10 DF, p-value: 0.0001836

Thus the mean metabolic rate, $E(Y)$, at a mass $x$ is given by

$$E(Y) = 201.2 + 24.0 \times x.$$  

In this model the intercept is insignificant (p-value of 0.29), but the slope parameter is significant (p-value < 0.001). It usually does not matter if the intercept is significant or not – we include it in the model anyway (Why?)

### 8.3 Model validation – assessing the model fit

#### 8.3.1 What does $R^2$ measure?

As well as measuring the squared correlation between $x$ and $y$, we can show that

$$R^2 = \frac{\text{variance of predicted values}}{\text{variance of observed values}}.$$  

i.e., $R^2$ measures the proportion of the variability in the response accounted for by the variability in the model – it is also known as the coefficient of determination. The closer $R^2$ is to one, the better the model explains the variation of the $Y$ values.

**Exercise:** Using the \texttt{cor} function, check the value of $R^2$ from the model summary above.
8.3.2 Fitted values and residual

The fitted/predicted values are

\[ \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i, \quad i = 1, \ldots, n, \]

and the residuals are

\[ \hat{\epsilon}_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n. \]

The residuals \( \hat{\epsilon}_i \) should have the same statistical properties as the errors \( \epsilon_i \):

- They have mean zero.
- They have constant variance \( \sigma^2 \).
- They are drawn from a normal distribution.

In the upcoming sections, we show how to check these assumptions.

8.3.3 Obtaining the residuals and fitted values

Try the following code:

```r
## calculate the residuals and fitted values.
resids <- resid(female.model)
fits <- fitted(female.model)
```

8.3.4 Checking model fit via residual plots

Normal plot and histogram of residuals:

- If residuals come from a normal distribution, points on normal plot should follow a straight line, and histogram should look bell shaped.

Residuals versus fits:

- Residuals should be randomly scattered around zero.
- Large residuals may indicate an outlier or influential data point.
• **Nonlinear** patterns or changes in **variability** indicate problems with the model fit.

Observations number versus residual:

• **Patterns** across the observation numbers may indicate measurement problems.

**Exercise:** Produce all of the above residuals plots in R for the regression model relating mass to the mean metabolic rate for the females.

### 8.3.5 An estimate of $\sigma^2$

The residuals are an **estimate** of the errors in the model. The **variability** of the errors is the **Residual Error SS** calculated as

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2.$$  

Now RSS has $n - 2$ degrees of freedom (**Residual Error df**) and its mean square (**Residual Error MS**) is

$$\hat{\sigma}^2 = \frac{RSS}{n-2}.$$  

In the model summary for the `lm` function, the estimate of $\sigma$ is denoted **Residual standard error**.

**Exercise:** What is the estimate of $\sigma$ in the regression model relating mass to the mean metabolic rate for the females.

### 8.4 Outliers and influential points

An **outlier** is an observation which lies outside the pattern of the rest of the data. Not all outliers of a model have large residuals!

An **influential point** is an observation which affects the regression results if that value is removed from the dataset. This can be a bigger problem for your model. How can we test whether an observation is influential?

**Influence example:** Compare the two models and the $R^2$ value for the female metabolic rate, with and without observation number 4.