STAT 373 Online Fall 2020

Lin Xuntian

September 2020

Contents

1 Simple Linear Regression

1.1 Introduction to Regression Modelling

Understanding and quantifying variability in data is the backbone of statistics and statistical inference. In regression modelling, we attempt to explain, or account for, variation in a response variable, y, by using a statistical model to describe the relationship between the response and one or more explanatory variables, x_1, x_2, ... We can then use the model to learn and answer questions about relationships between the explanatory variables and the response, and/or predict the value of the response for a given set of explanatory variables.

Consider the following examples:

Example 1

An auditor wishes to determine whether the cost of overhead claimed by offices in a certain group is consistent with the office's attributes, including size, age, number of clients, number of employees, and the cost of living index of the city in which the office is located. To this end, the auditor creates a regression model to describe the relationship between these attributes and the (claimed) overhead in order to estimate the expected overhead for each office. The auditor can then investigate any claim for which a large discrepancy exists between the observed overhead and the expected overhead estimated from the model.

Example 2

Is there systemic gender inequity in the salaries of Waterloo faculty members?

Solution: To answer this question, a Waterloo working committee obtained information on each faculty member, including rank, academic unit, years of service, gender, and annual salary, and fit a regression model to the data. Based on the model, they found that, after accounting for rank, academic unit, years of service, and several other variables, males were getting paid significantly more than females, on average. The results from this regression analysis resulted in an immediate increase of $2905 to the annual salaries of all female faculty members.

Example 3

Before listing a house, a realtor wishes to estimate its market value based on recent selling prices of homes in the area. Information is obtained on attributes of these homes that may help to account for selling price, such as size, lot size, number of rooms, number of bathrooms, number of stories, whether the house has a garage, etc., and a regression model is created to describe the relationship between selling price and these variables. The realtor can now use the model to estimate market value and predict selling price of the house, based on its attributes.

Note that in each of these examples, the objective for fitting a regression model is different, illustrating the power and usefulness of regression modelling. In the first example, the investigator wishes to detect discrepancies between an office's (claimed) overhead and the expected overhead for that office estimated from the regression model. In the second example, investigators wish to determine whether there is a relationship between gender and salary (i.e. whether there is a difference in mean salaries between male and female faculty members) after accounting for potentially confounding explanatory variables such as academic unit, rank, etc. Whereas in the last example, the objective was to predict the value of the response (market value) for a given set of explanatory variates.

We will be looking at each of these examples in more detail throughout the course. First, however, we will begin with a review of **simple linear regression**, in which we model the relationship between a single explanatory variable and the response.

1.2 Graphical and Numerical Summaries for Bivariate Data

With bivariate data, $\{x, y\}$, such as we have here, a scatterplot is an essential tool in visualizing and understanding the nature and strength of the relationship between an explanatory variable and a response.

Definition

A quantitative measure of the strength of a linear relationship between two variables is given by the correlation coefficient, r, defined as

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Properties of r:

1. $-1 \leqslant r \leqslant 1$, where the closer r is to 1 (-1), the stronger the positive (negative) relationship.
2. r is unitless. (Note that the units of the numerator and denominator will cancel). We can thus compare the relative strength of linear relationships across different scales and datasets.

1.3 The Simple Linear Regression Models

Consider again the scatterplot of overhead, y, vs. office size, x, for the audit data

1. a **deterministic** component, that describes the variation in accounted for by the functional form of the underlying relationship between y and x. Based on the scatterplot, the deterministic component can be adequately described by the linear function $\mu = \beta_0 + \beta_1 x$, where μ is the mean value of y for a given value of x.
2. an **error** term, denoted by the random variable ϵ, that describes the random variation in y not accounted for by the underlying relationship with x.

Definition

Incorporating both the deterministic and error components into our model yields the **simple linear regression** (SLR) model, expressed as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

where

- β_0 denotes the intercept parameter
- β_1 denotes the slope parameter
- the index i denotes the observation number (e.g. $\{x_3, y_3\}$ denotes the size and overhead associated with the third office in the dataset).

We will see in future lessons that, in order to derive the distributions of estimators for statistical inference procedures (i.e. confidence intervals and hypothesis tests for model parameters), we require certain distributional assumptions about the error random variable, ϵ.

In linear regression, we typically assume that the errors, ϵ_i , follow a normal distribution, with mean = 0, and variance denoted by σ^2. We also must assume that the errors are independent (recall for a normal random variable, independent errors $\Leftrightarrow \text{Cov}(\epsilon_j, \epsilon_j) = 0,\ j \neq k$).

Definition

Incorporating these assumptions into our SLR model yields the **normal** model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, \sigma^2) \text{ ind.} \quad i = 1, \ldots, n$$

Assumptions of the Normal Model

1. the functional form (e.g. linear) of the relationship between and is correctly specified by the deterministic component of the model
2. the errors follow a normal distribution
3. errors have a constant variance, denoted by σ^2(this property is sometimes referred to as **homoskedasticity**)
4. the errors are independent

For the normal model to be an appropriate model to use in investigating the relationship between y and x, these assumptions must hold. Otherwise, our model will be inappropriate and any conclusions we obtain from our regression analysis will be invalid. We will be examining these model assumptions in more detail in later sections.

1.4 Least Squares Estimation of Model Parameters

Definition

Solving these normal equations yields the **least squares estimates**

$$\begin{aligned}\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \\ \hat{\beta}_1 &= \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}\end{aligned}$$

1.5 The Fitted Model

Definition

The fitted model, or **fitted line**, for the SLR model is expressed as $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$, where $\hat{\mu}$ is the estimated mean value of the response y for a given value of x.

Note that the fitted model is sometimes expressed in terms of the predicted value of the response, $\hat{\mu} = \hat{y}_0 + \hat{\beta}_1 x$. While $\hat{\mu}$ and $\hat{y}$ are identical in terms of the value they represent, there are subtle differences in their interpretation that we will discuss in a later section.

Definition

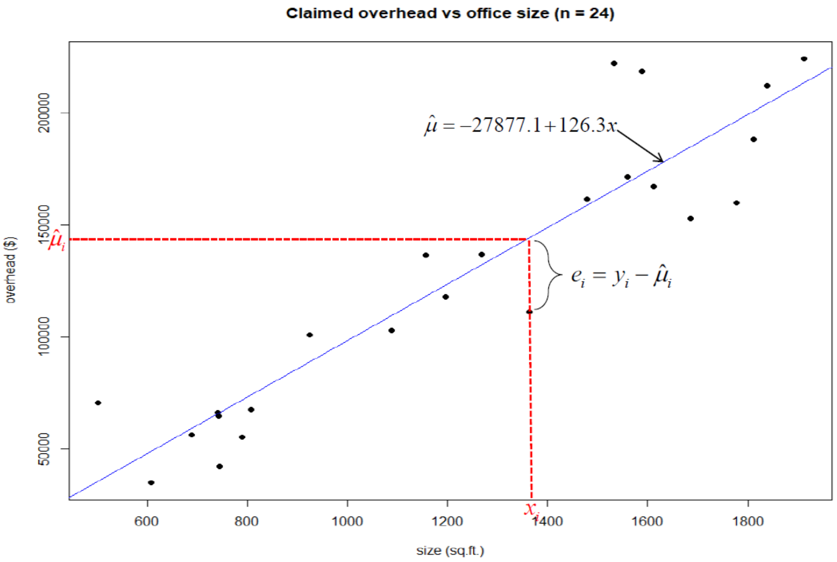
The **fitted residual** of the i^{th} observation, e_i, is the difference between the observed response, y_i, and the **fitted value**, $\hat{\mu}_i$, defined as

$$e_i = y_i - \hat{\mu}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

We will see in later sections that much of our statistical analysis from a regression model relies on the calculated value of the **sum of squares of the residuals**, $\sum e_i^2$.

Notes on the residuals:

Figure 1: From the audit data



- Understand the distinction between the residual, e_i, and the error, $e_i = y_i - \hat{\mu}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$.

  The error is the random variable, on which we impose certain distributional assumptions, we use to model the random variation in the response for a given value of x.

  The residual, $e_i = y - \hat{\mu}_i$, is the difference between the response and the estimated mean response, the values of which we calculate from the fitted line. We can think of the residuals as estimates of the errors.

- By taking the partial derivative with respect to each parameter and setting $= 0$ in our least squares estimation procedure, we have imposed two constraints on our residuals:

$$\sum e_i = 0$$
$$\sum x_i e_i = 0$$

  These constraints allow us to compute the remaining two residuals from $n-2$ observations. Thus, we say that the fitted model is associated with $n-2$ **degrees of freedom**.

1.6 Least Squares Estimation of σ^2

Recall the normal model given by

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2) \text{ ind.} \quad i = 1, \ldots, n$$

Inference for model parameters requires not only the estimation of β_0 and β_1, but also on the estimation of the error variance, σ^2.

Definition

In any least squares regression model, this is obtained by dividing the sum of squares of the residuals by the degrees of freedom, giving the least squares estimate as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{\mu}_i)^2}{n-2}$$

Note that $\hat{\sigma}^2$ is an **unbiased** estimate of σ^2 (i.e. $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$).

The **residual standard error** is the square root of the estimated variance, given by $\hat{\sigma} = \sqrt{\dfrac{\sum e_i^2}{n-2}}$

The residual standard error can be interpreted as the estimated standard deviation of the errors, and is a measure of the random variation in the response for a given value of x. The smaller the value of the residual standard error, the more variation in the response is explained by the relationship with x, and the better the fit of the model.

1.7 Interpretation of Parameter Estimates

Example 4

Let $\hat{\mu}_{x_0}$ be the estimated mean response at $x = x_0$, and let $\hat{\mu}_{x_0+1}$ be the estimated mean response at $x = x_0 + 1$.

Then

$$\begin{aligned} \hat{\mu}_{x_0+1} &= \hat{\beta}_0 + \hat{\beta}_1(x_0 + 1) \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_1 \\ &= \hat{\mu}_{x_0} + \hat{\beta}_1 \end{aligned}$$

Thus we can see that, in general, $\hat{\beta}_1$ can be interpreted as

The estimated mean change in the response, y, associated with a change of one unit in x.

Example 5

Note that for $x = 0$, the estimated mean response reduces to $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1(0) = \hat{\beta}$.

Thus, $\hat{\beta}_0$ may be interpreted in certain situations as the estimated mean value y of at $x = 0$.

However, there is an important caveat: This interpretation may be nonsensical or meaningless in cases where $x = 0$ is not a relevant value, or where $x = 0$ is not in the range of values used in the fit of the model.

This serves as an important reminder:

Never extrapolate results to values of outside the range used to fit the model.

Example 6

Recall the least squares estimate of the standard deviation of the errors, σ, called the residual standard error and given by:

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

where $e_i = y_i - \hat{\mu}_i$ are the residuals of the fitted model.

Note that the residual standard error is similar to the (sample) standard deviation of the residuals (only with $n - 2$ degrees of freedom instead of $n - 1$), and is thus **a measure of the variability of the response about the fitted line**. The smaller the residual standard error, the closer the data are to the fitted line, and the better the fit of the model.

Similar to a standard deviation, the residual standard error can be roughly interpreted as a typical or 'standard' distance (or absolute difference) between the response, y_i, and the fitted value, $\hat{\mu}_i$.

1.8 Inference for the Slope Parameter

Definition

A $(1-\alpha)100\%$ confidence interval for β_1 is of the form:

$$\hat{\beta}_1 \pm t_{n-2,\,1-\alpha/2}\,\mathrm{SE}(\hat{\beta}_1)$$

Notes:

- $t_{n-2,\,1-\alpha/2}$ denotes the critical value from a t_{n-2} distribution corresponding to confidence level $(1-\alpha)100\%$. (Be sure you know how to obtain this value for a given confidence level from both R and the posted t-tables)
- $t_{n-2,\,1-\alpha/2}\,\mathrm{SE}(\hat{\beta}_1)$ is called the **margin of error** of the interval. It can be thought of as the bound on the difference between the value of the estimate and the actual (unknown) value of the parameter for the given confidence level.
- it should be obvious, both intuitively and from the form of the confidence interval that
  - the higher the confidence level, the wider the interval
  - the larger the standard error, $\mathrm{SE}(\hat{\beta}_1)$, the wider the interval

Example 7

Provide a 95% confidence interval for β_1 from the audit SLR model.

Solution: We need to obtain the interval:

$$\hat{\beta}_1 \pm t_{22,\,0.975}\,\mathrm{SE}(\hat{\beta}_1)$$

These values give us a 95% confidence interval for β_1 of

$$\begin{aligned} &= 126.33 \pm 2.074(10.88) \\ &= 126.33 \pm 22.57 \\ &= (103.76, 148.90) \end{aligned}$$

Example 8

How can we interpret this interval in the context of the study?

Solution: We are 95% confident that for every additional increase of one square foot in office size, the mean increase in overhead is between \$103.76 and \$148.90.

Example 9

What conclusions can be drawn from our confidence interval about whether a relationship exists between overhead and office size?

Solution: Since $\beta_1 = 0$ is not in the interval, $(103.76, 148.90)$, we conclude that there is a **significant** positive relationship between overhead and office size.

(Had 0 been in the interval, then 0 would be considered a plausible value for and we would thus conclude that there was **no significant relationship** between overhead and office size)

Summary (Hypothesis test for slope parameter β_1)

1. Present the null and alternative hypotheses H_0: $\beta_1 = 0$
2. Calculate the value of the test statistic $\dfrac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} \sim t_{n-2}$, under H_0, $\beta_1 = 0$
3. Obtain p-value $= \mathrm{P}(|T| \geqslant |t|) = 2\,\mathrm{P}(T \geqslant |t|)$, where $T \sim t_{22}$
4. Conclusion in context of study

Types of errors in hypothesis testing

Note that whenever we draw a conclusion from a hypothesis test regarding the significance of the parameter, we could be in error, since we are drawing conclusion based on a probabilistic criteria.

For example, we may reject H_0: $\beta_1 = 0$ and conclude there is a significant relationship, when no relationship exists ($\beta_1 = 0$). Conversely, we may accept H_0: $\beta_1 = 0$ when, in fact, $\beta_1 \neq 0$ and a relationship exists.

The possibility that we could have made one of these errors should always be kept in mind when drawing conclusions from a hypothesis test (as well as from a confidence interval). These two errors are called:

- **Type I error:** Rejecting the null hypothesis when it is true
- **Type II error:** Accepting (i.e. not rejecting the null hypothesis) when it is true.

Note that for any hypothesis test, P(Type I error) = .05 (Convince yourself of this. It will help in your understanding of p-values)

Two-sided vs one-sided tests

By default, we will use a two-sided alternative hypothesis when performing hypothesis tests, since, in most cases, we are concerned with discovering significant relationships in either direction (positive or negative), and have little or no prior reliable knowledge of the possible direction of the relationship. If a one-sided alternative seems appropriate, it will be specified.

Note that hypothesis tests for which a one-sided alternative (e.g. H_a: $\beta_1 > 0$, or H_a: $\beta_1 < 0$) is appropriate yield a p-value $= \mathrm{P}(t_{n-2} > |t|)$, half the p-value that one would obtain with the two-sided alternative, H_a: $\beta_1 \neq 0$.

The relationship between confidence interval and hypothesis tests

Note that the conclusions (i.e. whether not a significant relationship exists) drawn from a 95% confidence interval for will always be consistent with conclusions drawn from a test of H_0: $\beta_1 = 0$

- If the 95% confidence interval contains 0, then a (two-sided) test of H_0: $\beta_1 = 0$ would yield a p-value $\geqslant .05$
- If the 95% confidence interval does not contain 0, then a (two-sided) test of H_0: $\beta_1 = 0$ would yield a p-value $< .05$