Chapter 11: Linear regression models

October 21, 2018

- We are often interested in understanding the *relationship* between two or more variables.
- Want to model a functional relationship between a "predictor" (input, independent variable) and a "response" variable (output, dependent variable, etc.).
- But real world is noisy, no $f = ma$ (Force = mass $\times$ acceleration). We have observation noise, weak relationship, etc.

Examples:

- How is the *sales price* of a house related to its size, number of rooms and property tax?
- How does the probability of *surviving* a particular surgery change as a function of the patient's age and general health condition?
- How does the *weight* of an individual depend on his/her height?

---

1 Method of least squares

Suppose that we have n data points $(x_1, Y_1), \ldots, (x_n, Y_n)$. We want to predict Y given a value of x.

- Y_i is the value of the *response* variable for the i-th observation.
- x_i is the value of the *predictor* variable for the i-th observation.

- **Scatter plot:** Plot the data and try to visualize the relationship.

- Suppose that we think that Y is a *linear* function (actually here a more appropriate term is "affine") of x, i.e.,

$$Y_i \approx \beta_0 + \beta_1 x_i,$$

  and we want to find the "best" such linear function.

- For the correct parameter values β_0 and β_1, the *deviation* of the observed values to its expected value, i.e.,

$$Y_i - \beta_0 - \beta_1 x_i,$$

  should be *small*.

- We try to *minimize* the sum of the n squared deviations, i.e., we can try to minimize

$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2$$

  as a function of b_0 and b_1. In other words, we want to minimize the sum of the squares of the vertical deviations of all the points from the line.

- The least squares estimators can be found by differentiating Q with respect to b_0 and b_1 and setting the partial derivatives equal to 0.

- Find b_0 and b_1 that solve:

$$\begin{aligned} \frac{\partial Q}{\partial b_0} &= -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 x_i) = 0 \\ \frac{\partial Q}{\partial b_1} &= -2\sum_{i=1}^{n} x_i(Y_i - b_0 - b_1 x_i) = 0. \end{aligned}$$

1.1 Normal equations

- The values of b_0 and b_1 that minimize Q are given by the solution to the *normal equations*:

$$\sum_{i=1}^{n} Y_i = n b_0 + b_1 \sum_{i=1}^{n} x_i \tag{1}$$

$$\sum_{i=1}^{n} x_i Y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2. \tag{2}$$

- Solving the normal equations gives us the following point estimates:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}, \tag{3}$$

$$b_0 = \bar{Y} - b_1 \bar{x}, \tag{4}$$

where $\bar{x} = \sum_{i=1}^n x_i/n$ and $\bar{Y} = \sum_{i=1}^n Y_i/n$.

In general, if we can parametrize the form of the functional dependence between Y and x in a linear fashion (linear in the parameters), then the method of least squares can be used to estimate the function. For example,

$$Y_i \approx \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

is still linear in the parameters.

2 Simple linear regression

The model for *simple linear regression* can be stated as follows:

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

- β_0, β_1 and σ^2 are *unknown* parameters.
- ϵ_i is a *random error* term whose distribution is unspecified:

$$\mathbb{E}(\epsilon_i) = 0, \qquad \text{Var}(\epsilon_i) = \sigma^2, \qquad \text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \text{for } i \neq j.$$

- x_i's will be treated as known *constants*. Even if the x_i's are random, we condition on the predictors and want to understand the *conditional distribution* of Y given X.
- **Regression function**: Conditional *mean* on Y given x, i.e.,

$$m(x) := \mathbb{E}(Y|x) = \beta_0 + \beta_1 x.$$

- The regression function shows how the mean of Y changes as a *function* of x.
- $\mathbb{E}(Y_i) = \mathbb{E}(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i$
- $\text{Var}(Y_i) = \text{Var}(\beta_0 + \beta_1 x_i + \epsilon_i) = \text{Var}(\epsilon_i) = \sigma^2$.

2.1 Interpretation

- The slope β_1 has units "y-units per x-units".
  - For every 1 inch increase in height, the model predicts a β_1 *pounds increase* in the mean weight.
- The intercept term β_0 is not always meaningful.
- The model is *only valid* for values of the explanatory variable in the domain of the data.

2.2 Estimation

- After formulating the model we use the observed data to *estimate* the *unknown* parameters.
- Three unknown parameters: β_0, β_1 and σ^2.
- We are interested in finding the estimates of these parameters that *best fit* the data.
- Question: *Best* in what sense?

2.2.1 Estimated regression function

- The *least squares* estimators of β_0 and β_1 are those values b_0 and b_1 that minimize:

$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2.$$

- Solving the normal equations gives us the following point estimates:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{5}$$

$$\hat{\beta}_0 = \bar{Y} - b_1 \bar{x}, \tag{6}$$

  where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $\bar{Y} = \sum_{i=1}^{n} Y_i/n$.

- We estimate the regression function:

$$\mathbb{E}(Y) = \beta_0 + \beta_1 x$$

  using

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

- The term
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad i = 1, \ldots, n,$$
is called the *fitted* or *predicted* value for the i-th observation, while Y_i is the observed value.

- The *residual*, denoted e_i, is the difference between the observed and the predicted value of Y_i, i.e.,
$$e_i = Y_i - \hat{Y}_i.$$

- The residuals show how far the individual data points fall from the regression function.

2.2.2 Properties

1. The sum of the residuals $\sum_{i=1}^n e_i$ is zero.

2. The sum of the squared residuals is a minimum.

3. The sum of the observed values equal the sum of the predicted values, i.e., $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i$.

4. The following sums of weighted residuals are equal to zero:
$$\sum_{i=1}^n x_i e_i = 0 \qquad \sum_{i=1}^n e_i = 0.$$

5. The regression line always passes through the point $(\bar{x}, \bar{Y})$.

2.2.3 Estimation of σ^2

- Recall: $\sigma^2 = \mathrm{Var}\,(\epsilon_i)$.

- We might have used $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (\epsilon_i - \bar{\epsilon})^2}{n-1}$. But ϵ_i's are not *observed*!

- Idea: Use e_i's, i.e., $s^2 = \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$.

- The divisor $n-2$ in s^2 is the number of *degrees of freedom* associated with the estimate.

- To obtain s^2, the two parameters β_0 and β_1 must first be estimated, which results in a loss of *two* degrees of freedom.

- Using $n-2$ makes s^2 an *unbiased* estimator of σ^2, i.e., $\mathbb{E}(s^2) = \sigma^2$.

2.2.4 Gauss-Markov theorem

The least squares estimators $\hat{\beta}_0, \hat{\beta}_1$ are *unbiased* (why?), i.e.,

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \qquad \mathbb{E}(\hat{\beta}_1) = \beta_1.$$

A *linear estimator* of β_j $(j = 0, 1)$ is an estimator of the form

$$\tilde{\beta}_j = \sum_{i=1}^{n} c_i Y_i,$$

where the coefficients $c_1, \ldots, c_n$ are only allowed to depend on x_i.

Note that $\hat{\beta}_0, \hat{\beta}_1$ are linear estimators (show this!).

Result: No matter what the distribution of the error terms ϵ_i, the least squares method provides *unbiased* point estimates that have *minimum* variance among all *unbiased linear estimators*.

The Gauss-Markov theorem states that in a linear regression model in which the errors have *expectation zero* and are *uncorrelated* and have *equal variances*, the *best linear unbiased estimator* (BLUE) of the coefficients is given by the *ordinary least squares estimators*.

2.3 Normal simple linear regression

To perform *inference* we need to make assumptions regarding the distribution of ϵ_i.

We often assume that ϵ_i's are *normally* distributed.

The *normal error* version of the model for simple linear regression can be written:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

Here ϵ_i's are independent $N(0, \sigma^2)$, σ^2 unknown.

Hence, Y_i's are independent normal random variables with mean $\beta_0 + \beta_1 x_i$ and variance σ^2.

Picture?

2.3.1 Maximum likelihood estimation

When the probability distribution of Y_i is *specified*, the estimates can be obtained using the method of *maximum likelihood*.

This method chooses as estimates those values of the parameter that are most *consistent* with the observed data.

The *likelihood* is the *joint density* of the Y_i's viewed as a function of the unknown parameters, which we denote $L(\beta_0, \beta_1, \sigma^2)$.

Since the Y_i's are *independent* this is simply the *product* of the density of individual Y_i's.

We seek the values of β_0, β_1 and σ^2 that maximize $L(\beta_0, \beta_1, \sigma^2)$ for the given x and Y values in the sample.

According to our model:

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

The likelihood function for the n independent observations $Y_1, \ldots, Y_n$ is given by

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 x_i)^2\right\} \qquad (7) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2\right\}. \end{aligned}$$

The value of $(\beta_0, \beta_1, \sigma^2)$ that maximizes the likelihood function are called *maximum likelihood estimates* (MLEs).

The MLE of β_0 and β_1 are *identical* to the ones obtained using the method of *least squares*, i.e.,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{S_x^2},$$

where $S_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

The MLE of σ^2: $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}$.

2.4 Inference

Our model describes the *linear* relationship between the two variables x and Y.

Different samples from the same population will produce different point estimates of β_0 and β_1.

Hence, $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables with sampling distributions that describe *what values* they can take and *how often* they take them.

Hypothesis tests about β_0 and β_1 can be constructed using these distributions.

The next step is to perform *inference*, including:

- Tests and confidence intervals for the *slope* and intercept.
- Confidence intervals for the *mean response*.
- *Prediction* intervals for new observations.

Theorem 1. *Under the assumptions of the normal linear model,*

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{S_x^2} & -\frac{\bar{x}}{S_x^2} \\ -\frac{\bar{x}}{S_x^2} & \frac{1}{S_x^2} \end{pmatrix}\right)$$

where $S_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$. *Also, if* $n \geq 3$, $\hat{\sigma}^2$ *is independent of* $(\hat{\beta}_0, \hat{\beta}_1)$ *and* $n\hat{\sigma}^2/\sigma^2$ *has a* χ^2*-distribution with* $n-2$ *degrees of freedom.*

Note that if the x_i's are random, the above theorem is still valid if we condition on the values of the predictor x_i's.

Exercise: Compute the variances and covariance of $\hat{\beta}_0, \hat{\beta}_1$.

2.4.1 Inference about β_1

We often want to perform tests about the *slope*:

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0.$$

Under the null hypothesis there is *no linear relationship* between Y and x – the *means* of probability distributions of Y are equal at all levels of x, i.e., $\mathbb{E}(Y|x) = \beta_0$, for all x.

The *sampling distribution* of $\hat{\beta}_1$ is given by

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_x^2}\right).$$

Need to show that: $\hat{\beta}_1$ is normally distributed,

$$\mathbb{E}(\hat{\beta}_1) = \beta_1, \qquad \text{Var}\,(\hat{\beta}_1) = \frac{\sigma^2}{S_x^2}.$$

Result: When $Z_1, \ldots, Z_k$ are *independent* normal random variables, the linear combination

$$a_1 Z_1 + \ldots + a_k Z_k$$

is also *normally* distributed.

Since $\hat{\beta}_1$ is a linear combination of the Y_i's and each Y_i is an *independent normally* distributed random variable, then $\hat{\beta}_1$ is also normally distributed.

We can write $\hat{\beta}_1 = \sum_{i=1}^n w_i Y_i$ where

$$w_i = \frac{x_i - \bar{x}}{S_x^2}, \qquad \text{for } i = 1, \dots, n.$$

Thus,

$$\sum_{i=1}^n w_i = 0, \quad \sum_{i=1}^n x_i w_i = 1, \quad \sum_{i=1}^n w_i^2 = \frac{1}{S_x^2}.$$

- **Variance for the estimated slope:** There are *three* aspects of the scatter plot that affect the variance of the regression slope:
  - The *spread* around the *regression line* (σ^2) – less scatter around the line means the slope will be more consistent from sample to sample.
  - The *spread* of the *x values* ($\sum_{i=1}^n (x_i - \bar{x})^2/n$) – a large variance of x provides a more stable regression.
  - The *sample size* n – having a larger sample size n, gives more consistent estimates.
- **Estimated variance:** When σ^2 is *unknown* we replace it with the

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

  Plugging this into the equation for Var $(\hat{\beta}_1)$ we get

$$se^2(\hat{\beta}_1) = \frac{\tilde{\sigma}^2}{S_x^2}.$$

  Recall: *Standard error* se$(\hat{\theta})$ of an estimator $\hat{\theta}$ is used to refer to an *estimate* of its *standard deviation.*

  Result: For the normal error regression model:

$$\frac{SSE}{\sigma^2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2_{n-2},$$

  and is *independent* of $\hat{\beta}_0$ and $\hat{\beta}_1$.

- (**Studentized statistic:**) Since $\hat{\beta}_1$ is *normally* distributed, the standardized statistic:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathrm{Var}\,(\hat{\beta}_1)}} \sim N(0,1).$$

  If we replace $\mathrm{Var}\,(\hat{\beta}_1)$ by its estimate we get the *studentized* statistic:

$$\frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}(\hat{\beta}_1)} \sim t_{n-2}.$$

  Recall: Suppose that $Z \sim N(0,1)$ and $W \sim \chi^2_p$ where Z and W are independent. Then,

$$\frac{Z}{\sqrt{W/p}} \sim t_p,$$

  the *t-distribution* with p *degrees of freedom.*

- **Hypothesis testing:** To test

$$H_0 : \beta_1 = 0 \qquad \text{versus} \qquad H_a : \beta_1 \neq 0$$

  use the *test-statistic*

$$T = \frac{\hat{\beta}_1}{\mathrm{se}(\hat{\beta}_1)}.$$

  We reject H_0 when the observed value of $|T|$ i.e., $|t_{obs}|$, is *large*!

  Thus, given *level* $(1-\alpha)$, we reject H_0 if

$$|t_{obs}| > t_{1-\alpha/2,n-2}$$

  where $t_{1-\alpha/2,n-2}$ denotes the $(1-\alpha/2)$-quantile of the t_{n-2}-distribution, i.e.,

$$1 - \frac{\alpha}{2} = \mathbb{P}(T \leq t_{1-\alpha/2,n-2}).$$

- ***P*-value:** p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

  The p-value depends on H_1 (one-sided/two-sided).

  In our case, we compute p-values using a t_{n-2}-distribution. Thus,

$$p\text{-value} = \mathbb{P}_{H_0}(|T| > |t_{obs}|).$$

  If we know the p-value then we can decide to accept/reject H_0 (versus H_1) at any given α.

- **Confidence interval:** A *confidence interval* (CI) is a kind of *interval estimator* of a population parameter and is used to indicate the reliability of an estimator.

  Using the sampling distribution of $\hat{\beta}_1$ we can make the following probability statement:

$$\begin{aligned}
\mathbb{P}\left(t_{\alpha/2,n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}(\hat{\beta}_1)} \leq t_{1-\alpha/2,n-2}\right) &= 1-\alpha \\
\mathbb{P}\left(\hat{\beta}_1 - t_{1-\alpha/2,n-2}\mathrm{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 - t_{\alpha/2,n-2}\mathrm{se}(\hat{\beta}_1)\right) &= 1-\alpha.
\end{aligned}$$

  Thus, a $(1-\alpha)$ confidence interval for β_1 is

$$\left[\hat{\beta}_1 - t_{1-\alpha/2,n-2} \cdot se(\hat{\beta}_1), \hat{\beta}_1 + t_{1-\alpha/2,n-2} \cdot se(\hat{\beta}_1)\right]$$

  as $t_{1-\alpha/2,n-2} = -t_{\alpha/2,n-2}$.

2.4.2 Sampling distribution of $\hat{\beta}_0$

The *sampling distribution* of $\hat{\beta}_0$ is

$$N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}\right)\right).$$

Verify at home using the same procedure as used for $\hat{\beta}_1$.

Hypothesis testing: In general, let c_0, c_1 and c_* be specified numbers, where at least one of c_0 and c_1 is nonzero. Suppose that we are interested in testing the following hypotheses:

$$H_0 : c_o\beta_0 + c_1\beta_1 = c_*, \qquad \text{versus} \qquad H_0 : c_o\beta_0 + c_1\beta_1 \neq c_*. \tag{8}$$

We should use a scalar multiple of

$$c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*$$

as the test statistic. Specifically, we use

$$U_{01} = \left[\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{S_x^2}\right]^{-1/2}\left(\frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*}{\tilde{\sigma}}\right),$$

where

$$\tilde{\sigma}^2 = \frac{S^2}{n-2}, \qquad S^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} e_i^2.$$

Note that $\tilde{\sigma}^2$ is an unbiased estimator of σ^2.

For each $\alpha \in (0,1)$, a level α test of the hypothesis (8) is to reject H_0 if

$$|U_{01}| > T_{n-2}^{-1}\left(1 - \frac{\alpha}{2}\right).$$

The above result follows from the fact that $c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*$ is normally distributed with mean $c_0\beta_0 + c_1\beta_1 - c_*$ and variance

$$\begin{aligned}
\mathrm{Var}\,(c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*) &= c_0^2\mathrm{Var}\,(\hat{\beta}_0) + c_1^2\mathrm{Var}\,(\hat{\beta}_1) + 2c_0c_1\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\
&= c_0^2\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}\right) + c_1^2\sigma^2\frac{1}{S_x^2} - 2c_0c_1\frac{\sigma^2\bar{x}}{S_x^2} \\
&= \sigma^2\left[\frac{c_0^2}{n} + \frac{c_0^2\bar{x}^2}{S_x^2} - 2c_0c_1\frac{\bar{x}}{S_x^2} + c_1^2\frac{1}{S_x^2}\right] \\
&= \sigma^2\left[\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{S_x^2}\right].
\end{aligned}$$

Confidence interval: We can give a $1-\alpha$ confidence interval for the parameter $c_0\beta_0 + c_1\beta_1$ as

$$c_0\hat{\beta}_0 + c_1\hat{\beta}_1 \mp \tilde{\sigma}\left[\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{S_x^2}\right]^{1/2} T_{n-2}^{-1}\left(1 - \frac{\alpha}{2}\right).$$

2.4.3 Mean response

We often want to estimate the *mean* of the probability distribution of Y for some value of x.

- The *point estimator* of the mean response

$$\mathbb{E}(Y|x_h) = \beta_0 + \beta_1 x_h$$

  when $x = x_h$ is given by

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h.$$

  Need to:

  - Show that $\hat{Y}_h$ is *normally* distributed.
  - Find $\mathbb{E}(\hat{Y}_h)$.
  - Find $\mathrm{Var}\,(\hat{Y}_h)$.

- The sampling distribution of $\hat{Y}_h$ is given by

$$\hat{Y}_h \sim N\left(\beta_0 + \beta_1 x_h, \sigma^2\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_x^2}\right)\right).$$

Normality:

Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are *linear combinations* of independent normal random variables Y_i.

Hence, $\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$ is also a linear combination of independent normally distributed random variables.

Thus, $\hat{Y}_h$ is also normally distributed.

Mean and variance of $\hat{Y}_h$:

Find the expected value of $\hat{Y}_h$:

$$\mathbb{E}(\hat{Y}_h) = \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_h) = \mathbb{E}(\hat{\beta}_0) + \mathbb{E}(\hat{\beta}_1) x_h = \beta_0 + \beta_1 x_h.$$

Note that $\hat{Y}_h = \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_h = \bar{Y} + \hat{\beta}_1 (x_h - \bar{x})$.

Note that $\hat{\beta}_1$ and $\bar{Y}$ are *uncorrelated*:

$$\mathrm{Cov}\left(\sum_{i=1}^n w_i Y_i, \sum_{i=1}^n \frac{1}{n} Y_i\right) = \sum_{i=1}^n \frac{w_i}{n}\sigma^2 = \frac{\sigma^2}{n}\sum_{i=1}^n w_i = 0.$$

Therefore,

$$\begin{aligned} \mathrm{Var}\,(\hat{Y}_h) &= \mathrm{Var}\,(\bar{Y}) + (x_h - \bar{x})^2 \mathrm{Var}\,(\hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + (x_h - \bar{x})^2 \frac{\sigma^2}{S_x^2}. \end{aligned}$$

When we do not know σ^2 we estimate it using $\tilde{\sigma}^2$. Thus, the *estimated variance* of $\hat{Y}_h$ is given by

$$\mathrm{se}^2(\hat{Y}_h) = \tilde{\sigma}^2 \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_x^2}\right).$$

The variance of $\hat{Y}_h$ is *smallest* when $x_h = \bar{x}$.

When $x_h = 0$, the variance of reduces to the variance of $\hat{\beta}_0$.

- The sampling distribution for the studentized statistic:

$$\frac{\hat{Y}_h - \mathbb{E}(\hat{Y}_h)}{\mathrm{se}(\hat{Y}_h)} \sim t_{n-2}.$$

All inference regarding $\mathbb{E}(\hat{Y}_h)$ are carried out using the t-distribution. A $(1-\alpha)$ CI for the *mean response* when $x = x_h$ is

$$\hat{Y}_h \mp t_{1-\alpha/2, n-2}\, \mathrm{se}(\hat{Y}_h).$$

2.4.4 Prediction interval

A CI for a *future* observation is called a *prediction interval*.

Consider the prediction of a new observation Y corresponding to a given level x of the predictor.

Suppose $x = x_h$ and the new observation is denoted $Y_{h(new)}$.

Note that $\mathbb{E}(\hat{Y}_h)$ is the *mean* of the distribution of $Y|X = x_h$.

$Y_{h(new)}$ represents the prediction of an *individual outcome* drawn from the distribution of $Y|X = x_h$, i.e.,

$$Y_{h(new)} = \beta_0 + \beta_1 x_h + \epsilon_{new},$$

where ϵ_{new} is independent of our data.

- The *point estimate* will be the *same* for both.

  However, the variance is *larger* when predicting an individual outcome due to the *additional variation* of an individual about the mean.

- When constructing prediction limits for $Y_{h(new)}$ we must take into consideration two sources of variation:

  - Variation in the *mean of Y*.
  - Variation around the mean.

- The *sampling* distribution of the studentized statistic:

$$\frac{Y_{h(new)} - \hat{Y}_h}{\text{se}(Y_{h(new)} - \hat{Y}_h)} \sim t_{n-2}.$$

  All inference regarding $Y_{h(new)}$ are carried out using the t-distribution:

$$\text{Var}\left(Y_{h(new)} - \hat{Y}_h\right) = \text{Var}\left(Y_{h(new)}\right) + \text{Var}\left(\hat{Y}_h\right) = \sigma^2 \left\{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_x^2}\right\}.$$

  Thus, $\text{se}_{pred} = \text{se}(Y_{h(new)} - \hat{Y}_h) = \tilde{\sigma}^2 \left\{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_x^2}\right\}$.

  Using this result, $(1 - \alpha)$ *prediction interval* for a new observation $Y_{h(new)}$ is

$$\hat{Y}_h \mp t_{1-\alpha/2, n-2}\ \text{se}_{pred}.$$

2.4.5 Inference about both β_0 and β_1 simultaneously

Suppose that β_0^* and β_1^* are given numbers and we are interested in testing the following hypothesis:

$$H_0 : \beta_0 = \beta_0^* \text{ and } \beta_1 = \beta_1^* \qquad \text{versus} \qquad H_1 : \text{at least one is different} \tag{9}$$

We shall derive the likelihood ratio test for (9).

The likelihood function (7), when maximized under the unconstrained space yields the MLEs $\hat{\beta}_1, \hat{\beta}_1, \hat{\sigma}^2$.

Under the constrained space, β_0 and β_1 are fixed at β_0^* and β_1^*, and so

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0^* - \beta_1^* x_i)^2.$$

The likelihood statistic reduces to

$$\Lambda(\mathbf{Y}, \mathbf{x}) = \frac{\sup_{\sigma^2} L(\beta_0^*, \beta_1^*, \sigma^2)}{\sup_{\beta_0, \beta_1, \sigma^2} L(\beta_0, \beta_1, \sigma^2)} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} = \left[\frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^{n}(Y_i - \beta_0^* - \beta_1^* x_i)^2}\right]^{n/2}.$$

The LRT procedure specifies rejecting H_0 when

$$\Lambda(\mathbf{Y}, \mathbf{x}) \le k,$$

for some k, chosen given the level condition.

Exercise: Show that

$$\sum_{i=1}^{n}(Y_i - \beta_0^* - \beta_1^* x_i)^2 = S^2 + Q^2,$$

where

$$\begin{aligned} S^2 &= \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ Q^2 &= n(\hat{\beta}_0 - \beta_0^*)^2 + \left(\sum_{i=1}^{n} x_i^2\right)(\hat{\beta}_1 - \beta_1^*)^2 + 2n\bar{x}(\hat{\beta}_0 - \beta_0^*)(\hat{\beta}_1 - \beta_1^*). \end{aligned}$$

Thus,

$$\Lambda(\mathbf{Y}, \mathbf{x}) = \left[\frac{S^2}{S^2 + Q^2}\right]^{n/2} = \left[1 + \frac{Q^2}{S^2}\right]^{-n/2}.$$

It can be seen that this is equivalent to rejecting H_0 when $Q^2/S^2 \ge k'$ which is equivalent to

$$U^2 := \frac{\frac{1}{2}Q^2}{\tilde{\sigma}^2} \ge \gamma.$$

Exercise: Show that, under H_0, $\frac{Q^2}{\sigma^2} \sim \chi^2_2$. Also show that Q^2 and S^2 are independent.

We know that $S^2/\sigma^2 \sim \chi^2_{n-2}$. Thus, under H_0,

$$U^2 \sim F_{2,n-2},$$

and thus $\gamma = F^{-1}_{2,n-2}(1-\alpha)$.

3 Linear models with normal errors

3.1 Basic theory

This section concerns models for independent responses of the form

$$Y_i \sim N(\mu_i, \sigma^2), \qquad \text{where} \qquad \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$$

for some known vector of explanatory variables $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{ip})$ and *unknown* parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, where $p < n$.

This is the linear model and is usually written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(in vector notation) where

$$\mathbf{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X}_{n\times p} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix}, \quad \boldsymbol{\beta}_{p\times 1} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon}_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

Sometimes this is written in the more compact notation

$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}),$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.

It is usual to assume that the $n \times p$ matrix $\mathbf{X}$ has full rank p.

3.2 Maximum likelihood estimation

The log–likelihood (up to a constant term) for $(\boldsymbol{\beta}, \sigma^2)$ is

$$\begin{aligned}
\ell(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2 \\
&= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2.
\end{aligned}$$

An MLE $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ satisfies

$$\begin{aligned}
0 &= \frac{\partial}{\partial\beta_j}\ell(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}x_{ij}(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}), \quad \text{for } j = 1, \ldots, p, \\
\text{i.e.,} \quad & \sum_{i=1}^{n}x_{ij}\mathbf{x}_i^\top\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n}x_{ij}y_i \qquad \text{for } j = 1, \ldots, p,
\end{aligned}$$

so

$$(\mathbf{X}^\top\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{Y}.$$

Since $\mathbf{X}^\top\mathbf{X}$ is non-singular if $\mathbf{X}$ has rank p, we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

The **least squares estimator** of β minimizes

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Check that this estimator coincides with the MLE when the errors are normally distributed.

Thus the estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ may be justified even when the normality assumption is uncertain.

Theorem 2. *We have*

1.

$$\hat{\beta} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}), \tag{10}$$

2.

$$\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}})^2$$

and that $\hat{\sigma}^2 \sim \frac{\sigma^2}{n}\chi^2_{n-p}$.

3. Show that $\hat{\boldsymbol{\beta}}$ *and* $\hat{\sigma}^2$ *are independent.*

Recall: Suppose that $\mathbf{U}$ is an n-dimensional random vector for which the mean vector $\mathbb{E}(\mathbf{U})$ and the covariance matrix $\mathrm{Cov}(\mathbf{U})$ exist. Suppose that $\mathbf{A}$ is a $q \times n$ matrix whose elements are constants. Let $\mathbf{V} = \mathbf{A}\mathbf{U}$. Then

$$\mathbb{E}(\mathbf{V}) = \mathbf{A}\mathbb{E}(\mathbf{U}) \qquad \text{and} \qquad \mathrm{Cov}(\mathbf{V}) = \mathbf{A}\mathrm{Cov}(\mathbf{U})\mathbf{A}^\top.$$

Proof of 1: The MLE of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$, and we have that the model can be written in vector notation as $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

Let $\mathbf{M} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ so that $\mathbf{M}\mathbf{Y} = \hat{\boldsymbol{\beta}}$. Therefore,

$$\mathbf{M}\mathbf{Y} \sim N_p(\mathbf{M}\mathbf{X}\boldsymbol{\beta}, \mathbf{M}(\sigma^2\mathbf{I})\mathbf{M}^\top).$$

We have that

$$\begin{aligned} \mathbf{M}\mathbf{X}\boldsymbol{\beta} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned} \qquad \text{and} \qquad \begin{aligned} \mathbf{M}\mathbf{M}^\top &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1} \end{aligned}$$

since $\mathbf{X}^\top\mathbf{X}$ is symmetric, and then so is it's inverse.

Therefore,

$$\hat{\boldsymbol{\beta}} = \mathbf{M}\mathbf{Y} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}).$$

These results can be used to obtain an exact $(1-\alpha)$-level confidence region for $\boldsymbol{\beta}$: the distribution of $\hat{\boldsymbol{\beta}}$ implies that

$$\frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p.$$

Let

$$\tilde{\sigma}^2 = \frac{1}{n-p}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \sim \frac{\sigma^2}{n-p}\chi^2_{n-p},$$

so that $\hat{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$ are still independent.

Then, letting $F_{p,n-p}(\alpha)$ denote the upper α-point of the $F_{p,n-p}$ distribution,

$$1-\alpha = \mathbb{P}_{\boldsymbol{\beta},\sigma^2}\left(\frac{\frac{1}{p}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\tilde{\sigma}^2} \le F_{p,n-p}(\alpha)\right).$$

Thus,

$$\left\{\boldsymbol{\beta}\in\mathbb{R}^p : \frac{\frac{1}{p}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\tilde{\sigma}^2} \le F_{p,n-p}(\alpha)\right\}$$

is a $(1-\alpha)$-level confidence set for $\boldsymbol{\beta}$.

3.2.1 Projections and orthogonality

The *fitted values* $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ under the model satisfy

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{Y} \equiv \mathbf{PY},$$

say, where $\mathbf{P}$ is an *orthogonal projection* matrix (i.e., $\mathbf{P}=\mathbf{P}^\top$ and $\mathbf{P}^{\mathbf{2}}=\mathbf{P}$) onto the column space of $\mathbf{X}$.

Since $\mathbf{P}^{\mathbf{2}}=\mathbf{P}$, all of the eigenvalues of $\mathbf{P}$ are either 0 or 1 (Why?).

Therefore,

$$\text{rank}(\mathbf{P}) = \text{tr}(\mathbf{P}) = \text{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top) = \text{tr}((\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{X}) = \text{tr}(\mathbf{I}_p) = p$$

by the *cyclic property* of the trace operation.

Some authors denote $\mathbf{P}$ by $\mathbf{H}$, and call it the hat matrix because it “puts the hat on $\mathbf{Y}$”. In fact, $\mathbf{P}$ is an orthogonal projection. Note that in the standard linear model above we may express the **fitted** values

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

as $\hat{\mathbf{Y}} = \mathbf{PY}$.

---

1. Show that $\mathbf{P}$ represents an orthogonal projection.
2. Show that $\mathbf{P}$ and $\mathbf{I}-\mathbf{P}$ are positive semi-definite.
3. Show that $\mathbf{I}-\mathbf{P}$ has rank $n-p$ and $\mathbf{P}$ has rank p.

Solution: To see that $\mathbf{P}$ represents a projection, notice that $\mathbf{X}^\top\mathbf{X}$ is symmetric, so its inverse is also, so

$$\mathbf{P}^\top = \{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\}^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{P}$$

and

$$\mathbf{P}^2 = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{P}.$$

To see that $\mathbf{P}$ is an orthogonal projection, we must show that $\mathbf{PY}$ and $\mathbf{Y} - \mathbf{PY}$ are orthogonal. But from the results above,

$$(\mathbf{PY})^\top(\mathbf{Y} - \mathbf{PY}) = \mathbf{Y}^\top\mathbf{P}^\top(\mathbf{Y} - \mathbf{PY}) = \mathbf{Y}^\top\mathbf{PY} - \mathbf{Y}^\top\mathbf{PY} = \mathbf{0}.$$

$\mathbf{I} - \mathbf{P}$ is positive semi-definite since

$$\mathbf{x}^\top(\mathbf{I} - \mathbf{P})\mathbf{x} = \mathbf{x}^\top(\mathbf{I} - \mathbf{P})^\top(\mathbf{I} - \mathbf{P})\mathbf{x} = \|\mathbf{x} - \mathbf{Px}\|^{\mathbf{2}} \geq \mathbf{0}.$$

Similarly, $\mathbf{P}$ is positive semi-definite.

Cochran's theorem: Let $\mathbf{Z} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, and let $\mathbf{A_1}, \ldots, \mathbf{A_k}$ be $n \times n$ positive semi–definite matrices with $\text{rank}(\mathbf{A}_i) = r_i$, such that

$$\|\mathbf{Z}\|^2 = \mathbf{Z}^\top\mathbf{A}_1\mathbf{Z} + \ldots + \mathbf{Z}^\top\mathbf{A}_k\mathbf{Z}.$$

If $r_1 + \cdots + r_k = n$, then $\mathbf{Z}^\top\mathbf{A}_1\mathbf{Z}, \ldots, \mathbf{Z}^\top\mathbf{A}_k\mathbf{Z}$ are independent, and

$$\frac{\mathbf{Z}^\top\mathbf{A}_i\mathbf{Z}}{\sigma^2} \sim \chi^2_{r_i}, \quad i = 1, \ldots, k.$$

Problem 2: In the standard linear model above, find the maximum likelihood estimator $\hat{\sigma}^2$ of σ^2, and use Cochran's theorem to find its distribution.

Solution: Differentiating the log-likelihood

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

we see that an MLE $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ satisfies

$$0 = \frac{\partial\ell}{\partial\sigma^2}\bigg|_{(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2,$$

so

$$\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \equiv \frac{1}{n}\|\mathbf{Y} - \mathbf{PY}\|^2,$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top$. Observe that

$$\|\mathbf{Y} - \mathbf{P}\mathbf{Y}\|^{\mathbf{2}} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{P})^\top(\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{P})\mathbf{Y},$$

and from the previous question we know that $\mathbf{I} - \mathbf{P}$ and $\mathbf{P}$ are positive semi-definite and of rank $n - p$ and p, respectively. We cannot apply Cochran's theorem directly since $\mathbf{Y}$ does not have mean zero. However, $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ does have mean zero and

$$\begin{aligned}(\mathbf{Y}-&\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{I} - \mathbf{P})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}^\top(\mathbf{I} - \mathbf{P})\mathbf{Y} - \mathbf{2}\boldsymbol{\beta}^\top\mathbf{X}^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top)\mathbf{Y} + \boldsymbol{\beta}^\top\mathbf{X}^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top)\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{Y}^\top(\mathbf{I} - \mathbf{P})\mathbf{Y}.\end{aligned}$$

Since

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^{\mathbf{2}} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{I} - \mathbf{P})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top\mathbf{P}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

we may therefore apply Cochran's theorem to deduce that

$$\mathbf{Y}^\top(\mathbf{I} - \mathbf{P})\mathbf{Y} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{I} - \mathbf{P})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \sim \sigma^2\chi^2_{n-p},$$

and hence

$$\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{P}\mathbf{Y}\|^{\mathbf{2}} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{I} - \mathbf{P})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \sim \frac{\sigma^2}{n}\chi^2_{n-p}.$$

3.2.2 Testing hypotheis

Suppose that we want to test

$$H_0 : \beta_j = \beta_j^* \qquad \text{versus} \qquad H_0 : \beta_j \neq \beta_j^*$$

for some $j \in \{1, \ldots, p\}$, where β_j^* is a fixed number. We know that

$$\hat{\beta}_j \sim N(\beta_j, \zeta_{jj}\sigma^2),$$

where $(\mathbf{X}^\top\mathbf{X})^{-1} = ((\zeta_{ij}))_{p\times p}$. Thus, we know that

$$T = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\tilde{\sigma}^2\zeta_{jj}}} \sim t_{n-p} \text{ under } H_0,$$

where we have used Theorem 2.

3.3 Testing for a component of $\boldsymbol{\beta}$ – not included in the final exam

Now partition $\mathbf{X}$ and $\boldsymbol{\beta}$ as

$$\underbrace{\mathbf{X}}_{n\times p} = (\underbrace{\mathbf{X}_0}_{n\times p_0} \quad \underbrace{\mathbf{X}_1}_{n\times(p-p_0)}) \qquad \text{and} \qquad \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \end{pmatrix} \begin{array}{l} \updownarrow p_0 \\ \updownarrow p-p_0 \end{array}.$$

Suppose that we are interested in testing

$$H_0 : \boldsymbol{\beta}_1 = 0, \qquad \text{against} \qquad H_1 : \boldsymbol{\beta}_1 \neq 0.$$

Then, under H_0, the MLEs of $\boldsymbol{\beta}_0$ and σ^2 are

$$\hat{\hat{\boldsymbol{\beta}}}_0 = (\mathbf{X}_\mathbf{0}^\top \mathbf{X}_\mathbf{0})^{-\mathbf{1}} \mathbf{X}_\mathbf{0}^\top \mathbf{Y}, \qquad \hat{\hat{\sigma}}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}_0 \hat{\hat{\boldsymbol{\beta}}}_0\|^2.$$

$\hat{\hat{\beta}}_0$ and $\hat{\hat{\sigma}}^2$ are independent. The fitted values under H_0 are

$$\hat{\hat{\mathbf{Y}}} = \mathbf{X}_0 \hat{\hat{\boldsymbol{\beta}}}_0 = \mathbf{X}_0(\mathbf{X}_0^\top \mathbf{X}_0)^{-1}\mathbf{X}_0^\top \mathbf{Y} = \mathbf{P}_0 \mathbf{Y}$$

where $P_0 = \mathbf{X}_0(\mathbf{X}_0^\top \mathbf{X}_0)^{-1}\mathbf{X}_0^\top$ is an orthogonal projection matrix of rank p_0.

The likelihood ratio statistic is

$$\begin{aligned} -2\log\Lambda &= 2\left\{-\frac{n}{2}\log\left(\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2\right) - \frac{n}{2} + \frac{n}{2}\log\left(\|\mathbf{Y} - \mathbf{X}_0\hat{\hat{\boldsymbol{\beta}}}_0\|^2\right) + \frac{n}{2}\right\} \\ &= n\log\left(\frac{\|\mathbf{Y} - \mathbf{X}_0\hat{\hat{\boldsymbol{\beta}}}_0\|^2}{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}\right) = n\log\left(\frac{\|\mathbf{Y} - \mathbf{P_0 Y}\|^2}{\|\mathbf{Y} - \mathbf{PY}\|^2}\right). \end{aligned}$$

We therefore reject H_0 if the ratio of the residual sum of squares under H_0 to the residual sum of squares under H_1 is large.

Rather than use Wilks' theorem to obtain the asymptotic "null distribution" of the test statistic [which anyway depends on unknown σ^2], we can work out the exact distribution in this case.

Since $(\mathbf{Y} - \mathbf{PY})^\top(\mathbf{PY} - \mathbf{P_0 Y}) = \mathbf{0}$, Pythagorean theorem gives that

$$\|\mathbf{Y} - \mathbf{PY}\|^2 + \|\mathbf{PY} - \mathbf{P_0 Y}\|^2 = \|\mathbf{Y} - \mathbf{P_0 Y}\|^2. \tag{11}$$

Using (11),

$$\frac{\|\mathbf{Y}-\mathbf{P_0Y}\|^2}{\|\mathbf{Y}-\mathbf{PY}\|^2} = \frac{\|\mathbf{Y}-\mathbf{PY}\|^2}{\|\mathbf{Y}-\mathbf{PY}\|^2} + \frac{\|\mathbf{PY}-\mathbf{P_0Y}\|^2}{\|\mathbf{Y}-\mathbf{PY}\|^2}$$
$$= 1 + \frac{\|\mathbf{PY}-\mathbf{P_0Y}\|^2}{\|\mathbf{Y}-\mathbf{PY}\|^2}.$$

Consider the decomposition:

$$\|\mathbf{Y}\|^2 = \|\mathbf{Y}-\mathbf{PY}\|^2 + \|\mathbf{PY}-\mathbf{P_0Y}\|^2 + \|\mathbf{P_0Y}\|^2$$

and a similar one for $\mathbf{Z} = \mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0$.

Under H_0, $\mathbf{Z} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. This allows the use of Cochran's theorem to ultimately conclude that $\|\mathbf{PY}-\mathbf{P_0Y}\|^2$ and $\|\mathbf{Y}-\mathbf{PY}\|^2$ are independent $\sigma^2\chi^2_{p-p_0}$ and $\sigma^2\chi^2_{n-p}$ random variables, respectively.

Exercise: Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are partitioned as $\mathbf{X} = (\mathbf{X}_0 | \, \mathbf{X}_1)$ and $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_0^T | \, \boldsymbol{\beta}_1^T)$ respectively (where $\boldsymbol{\beta}_0$ has p_0 components and $\boldsymbol{\beta}_1$ has $p - p_0$ components).

1. Show that
$$\|\mathbf{Y}\|^2 = \|\mathbf{P_0Y}\|^2 + \|(\mathbf{P}-\mathbf{P_0})\mathbf{Y}\|^2 + \|\mathbf{Y}-\mathbf{PY}\|^2.$$

2. Recall that the likelihood ratio statistic for testing
$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0} \qquad \text{against} \qquad H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}$$
is a strictly increasing function of $\|(\mathbf{P}-\mathbf{P_0})\mathbf{Y}\|^2/\|\mathbf{Y}-\mathbf{PY}\|^2$.
Use Cochran's theorem to find the joint distribution of $\|(\mathbf{P}-\mathbf{P}_0)\mathbf{Y}\|^2$ and $\|\mathbf{Y}-\mathbf{PY}\|^2$ under H_0. How would you perform the hypothesis test?
[Hint: rank$(\boldsymbol{P}) = p$*, and* rank$(\mathbf{I}-\mathbf{P}) = n - p$*. Similar arguments give that* rank$(\mathbf{P_0}) = p_0$.

Solution:

1. Recall that since $(\mathbf{Y}-\mathbf{PY})^\top(\mathbf{PY}-\mathbf{P}_0\mathbf{Y}) = 0$ Pythagorean theorem gives that
$$\begin{aligned}\|\mathbf{Y}-\mathbf{PY}\|^2 + \|\mathbf{PY}-\mathbf{P}_0\mathbf{Y}\|^2 &= \|\mathbf{Y}-\mathbf{P}_0\mathbf{Y}\|^2 \\ &= (\mathbf{Y}-\mathbf{P}_0\mathbf{Y})^\top(\mathbf{Y}-\mathbf{P}_0\mathbf{Y}) \\ &= \mathbf{Y}^\top\mathbf{Y} - 2\mathbf{Y}^\top\mathbf{P}_0\mathbf{Y} + \mathbf{Y}^\top\mathbf{P}_0^\top\mathbf{P}_0\mathbf{Y} \\ &= \mathbf{Y}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{P}_0\mathbf{P}_0^\top\mathbf{Y} \\ &= \|\mathbf{Y}\|^2 - \|\mathbf{P}_0\mathbf{Y}\|^2\end{aligned}$$
giving that
$$\|\mathbf{Y}-\mathbf{PY}\|^2 + \|\mathbf{PY}-\mathbf{P}_0\mathbf{Y}\|^2 + \|\mathbf{P}_0\mathbf{Y}\|^2 = \|\mathbf{Y}\|^2$$
as desired.

2. Under H_0, the response vector $\mathbf{Y}$ has mean $\mathbf{X}_0\boldsymbol{\beta}_0$, and so $\mathbf{Z} = \mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0$ satisfies

$$\begin{aligned}\|\mathbf{Z}\|^2 &= \|\mathbf{Z} - \mathbf{P}\mathbf{Z}\|^2 + \|\mathbf{P}\mathbf{Z} - \mathbf{P}_0\mathbf{Z}\|^2 + \|\mathbf{P}_0\mathbf{Z}\|^2 \\ &= \mathbf{Z}^\top\mathbf{Z} - 2\mathbf{Z}^\top\mathbf{P}\mathbf{Z} + \mathbf{Z}^\top\mathbf{P}^\top\mathbf{P}\mathbf{Z} + \mathbf{Z}^\top(\mathbf{P} - \mathbf{P}_0)^\top(\mathbf{P} - \mathbf{P}_0)\mathbf{Z} + \mathbf{Z}^\top\mathbf{P}_0^\top\mathbf{P}_0\mathbf{Z} \\ &= \mathbf{Z}^\top(\mathbf{I} - \mathbf{P})\mathbf{Z} + \mathbf{Z}^\top(\mathbf{P} - \mathbf{P}_0)\mathbf{Z} + \mathbf{Z}^\top\mathbf{P}_0\mathbf{Z}.\end{aligned}$$

But

$$\begin{aligned}\mathbf{Z}^\top(\mathbf{P} - \mathbf{P}_0)\mathbf{Z} &= (\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)^\top(\mathbf{P} - \mathbf{P}_0)(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0) \\ &= \mathbf{Y}^\top(\mathbf{P} - \mathbf{P}_0)\mathbf{Y} - 2\boldsymbol{\beta}_0^\top\mathbf{X}_0^\top(\mathbf{P} - \mathbf{P}_0)\mathbf{Y} + \boldsymbol{\beta}_0^\top\mathbf{X}_0^\top(\mathbf{P} - \mathbf{P}_0)\mathbf{X}_0\boldsymbol{\beta}_0.\end{aligned}$$

Since $\mathbf{X}_0\boldsymbol{\beta}_0 \in U_0$ and $(\mathbf{P} - \mathbf{P}_0)\mathbf{Y} \in U_0^\perp$, and U_0 and $U_0^\perp$ are mutually orthogonal, and moreover $\mathbf{P}\mathbf{X}_0\boldsymbol{\beta}_0 = \mathbf{P}_0\mathbf{X}_0\boldsymbol{\beta}_0 = \mathbf{X}_0\boldsymbol{\beta}_0$, this gives

$$\mathbf{Z}^\top(\mathbf{P} - \mathbf{P}_0)\mathbf{Z} = \mathbf{Y}^\top(\mathbf{P} - \mathbf{P}_0)\mathbf{Y},$$

Similarly,

$$\begin{aligned}\mathbf{Z}^\top(\mathbf{I} - \mathbf{P})\mathbf{Z} &= (\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)^\top(\mathbf{I} - \mathbf{P})(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0) \\ &= \mathbf{Y}^\top(\mathbf{I} - \mathbf{P})\mathbf{Y} - 2\boldsymbol{\beta}_0^\top\mathbf{X}_0^\top(\mathbf{I} - \mathbf{P})\mathbf{Y} + \boldsymbol{\beta}_0^\top\mathbf{X}_0^\top(\mathbf{I} - \mathbf{P})\mathbf{X}_0\boldsymbol{\beta}_0 \\ &= \mathbf{Y}^\top(\mathbf{I} - \mathbf{P})\mathbf{Y},\end{aligned}$$

since $\mathbf{X}_0\boldsymbol{\beta}_0 \in U_0$ and $(\mathbf{I} - \mathbf{P})\mathbf{Y} \in U^\perp \subseteq U_0^\perp$, while $(\mathbf{I} - \mathbf{P})\mathbf{X}_0\boldsymbol{\beta}_0 = \mathbf{X}_0\boldsymbol{\beta}_0 - \mathbf{X}_0\boldsymbol{\beta}_0 = 0$. Since

$$\text{rank}(\mathbf{I} - \mathbf{P}) + \text{rank}(\mathbf{P} - \mathbf{P}_0) + \text{rank}(\mathbf{P}_0) = n - p + p - p_0 + p_0 = n$$

we may therefore apply Cochran's theorem to deduce that under H_0, $\|(\mathbf{P} - \mathbf{P}_0)\mathbf{Y}\|^2$ and $\|\mathbf{Y} - \mathbf{P}\mathbf{Y}\|^2$ are independent with

$$\|(\mathbf{P} - \mathbf{P}_0)\mathbf{Y}\|^2 = \mathbf{Y}^\top(\mathbf{P} - \mathbf{P}_0)\mathbf{Y} = \mathbf{Z}^\top(\mathbf{P} - \mathbf{P}_0)\mathbf{Z} \sim \sigma^2\chi^2_{p-p_0},$$

and

$$\|(\mathbf{I} - \mathbf{P})\mathbf{Y}\|^2 = \mathbf{Y}^\top(\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{Z}^\top(\mathbf{I} - \mathbf{P})\mathbf{Z} \sim \sigma^2\chi^2_{n-p}.$$

It follows that under H_0,

$$F = \frac{\frac{1}{p-p_0}\|(\mathbf{P} - \mathbf{P}_0)\mathbf{Y}\|^2}{\frac{1}{n-p}\|(\mathbf{I} - \mathbf{P})\mathbf{Y}\|^2} \sim F_{p-p_0,n-p},$$

so we may reject H_0 if $F > F_{p-p_0,n-p}(\alpha)$, where $F_{p-p_0,n-p}(\alpha)$ is the upper α-point of the $F_{p-p_0,n-p}$ distribution.

Thus under H_0,

$$F = \frac{\frac{1}{p-p_0}\|\mathbf{PY} - \mathbf{P_0Y}\|^2}{\frac{1}{n-p}\|\mathbf{Y} - \mathbf{PY}\|^2} \sim F_{p-p_0,n-p}.$$

When $\mathbf{X}_0$ has one less column than $\mathbf{X}$, say column k, we can leverage the normality of the MLE $\hat{\beta}_k$ in (10) to perform a t-test based on the statistic

$$T = \frac{\hat{\beta}_k}{\sqrt{\tilde{\sigma}^2 \text{diag}[(\mathbf{X}^\top \mathbf{X})^{-1}]_k}} \sim t_{n-p} \text{ under } H_0 \quad \text{[i.e., } \beta_k = 0\text{]}.$$

[This is what R uses, though the more general F–statistic can also be used in this case.]

The above theory also shows that under H_1, $\frac{1}{n-p}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2$ is an unbiased estimator of σ^2. This is usually used in preference to the MLE, $\hat{\sigma}^2$.

Example:

1. Multiple linear regression:

   For countries $i = 1, \ldots, n$, consider how the fertility rate Y_i (births per 1000 females in a particular year) depends on

   - the gross domestic product per capita x_{i1}
   - and the percentage of urban dwellers x_{i2}.

   The model

   $$\log Y_i = \beta_0 + \beta_1 \log x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad i = 1, \ldots, n$$

   with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, is of linear model form $Y = X\beta + \varepsilon$ with

   $$Y = \begin{pmatrix} \log Y_1 \\ \vdots \\ \log Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & \log x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & \log x_{n1} & x_{n2} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

   On the original scale of the response, this model becomes

   $$Y = \exp(\beta_0) \exp(\beta_1 \log x_1) \exp(\beta_2 x_2) \varepsilon$$

   Notice how the possibility of transforming variables greatly increases the flexibility of the linear model. [But see how using a log response assumes that the errors enter multiplicatively.]

4 One-way analysis of variance (ANOVA)

Consider measuring yields of plants under a control condition and $J-1$ different treatment conditions.

The explanatory variable (factor) has J levels, and the response variables at level j are $Y_{j1}, \ldots, Y_{jn_j}$. The model that the responses are independent with

$$Y_{jk} \sim N(\mu_j, \sigma^2), \quad j = 1, \ldots, J; \;\; k = 1, \ldots, n_j$$

is of linear model form, with

$$Y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{J1} \\ \vdots \\ Y_{Jn_J} \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ & & \vdots & & \\ 0 & \cdots & \cdots & 0 & 1 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} n_1 \\ \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} n_2 \\ \\ \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} n_J \end{matrix} \qquad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{pmatrix}.$$

An alternative parameterization, emphasizing the differences between treatments, is

$$Y_{jk} = \mu + \alpha_j + \varepsilon_{jk}, \quad j = 1, \ldots, J; \;\; k = 1, \ldots, n_j$$

where

- μ is the baseline or mean effect
- α_j is the effect of the j^{th} treatment (or the control $j = 1$).

Notice that the parameter vector $(\mu, \alpha_1, \alpha_2, \ldots, \alpha_J)^\top$ is not identifiable, since replacing μ with $\mu + 10$ and α_j by $\alpha_j - 10$ gives the same model. Either a

- corner point constraint $\alpha_1 = 0$ is used to emphasise the differences from the control, or the

- sum–to–zero constraint $\sum_{j=1}^{J} n_j\alpha_j = 0$

can be used to make the model identifiable. R uses corner point constraints.

If $n_j = K$, say, for all j, the data are said to be balanced.

We are usually interested in comparing the null model

$$H_0 : Y_{jk} = \mu + \varepsilon_{jk}$$

with that given above, which we call H_1, i.e., we wish to test whether the treatment conditions have an effect on the plant yield:

$$H_0 : \alpha = 0, \text{ where } \alpha = (\alpha_1, \ldots, \alpha_J), \qquad \text{against} \qquad H_1 : \alpha \neq 0.$$

Check that the MLE fitted values are

$$\hat{Y}_{jk} = \bar{Y}_j \equiv \frac{1}{n_j}\sum_{k=1}^{n_j} Y_{jk}$$

under H_1, whatever parameterization is chosen, and are

$$\hat{\hat{Y}}_{jk} = \bar{Y} \equiv \frac{1}{n}\sum_{j=1}^{J} n_j\bar{Y}_j, \qquad \text{where } n = \sum_{j=1}^{J} n_j,$$

under H_0.

Theorem 3. *(Partitioning the sum of squares) We have*

$$SS_{total} = SS_{within} + SS_{between},$$

where

$$SS_{total} = \sum_{j=1}^{J}\sum_{k=1}^{n_j}(Y_{jk}-\bar{Y})^2, \qquad SS_{within} = \sum_{j=1}^{J}\sum_{k=1}^{n_j}(Y_{jk}-\bar{Y}_j)^2, \qquad SS_{between} = \sum_{j=1}^{J} n_j(\bar{Y}_j-\bar{Y})^2.$$

Furthermore, SS_{within} has $\sigma^2\chi^2$-distribution with $(n-J)$ degrees of freedom and is independent of $SS_{between}$. Also, under H_0, $SS_{between} \sim \sigma^2\chi^2_{J-1}$.

Our linear model theory says that we should test H_0 by referring

$$F = \frac{\frac{1}{J-1}\sum_{j=1}^{J} n_j(\bar{Y}_j-\bar{Y})^2}{\frac{1}{n-J}\sum_{j=1}^{J}\sum_{k=1}^{n_j}(Y_{jk}-\bar{Y}_j)^2} \equiv \frac{\frac{1}{J-1}S_2}{\frac{1}{n-J}S_1}$$

to $F_{J-1,n-J}$, where S_1 is the "within groups" sum of squares and S_2 is the "between groups" sum of squares. We have the following ANOVA table.

Source of variation	Degrees of freedom	Sum of squares	F–statistic
Between groups	$J-1$	S_2	$F = \dfrac{\frac{1}{J-1}S_2}{\frac{1}{n-J}S_1}$
Within groups	$n-J$	S_1	
Total	$n-1$	$S_1 + S_2 = \sum_{j=1}^{J}\sum_{k=1}^{n_j}(Y_{jk} - \bar{Y})^2$	