STA 114: Statistics

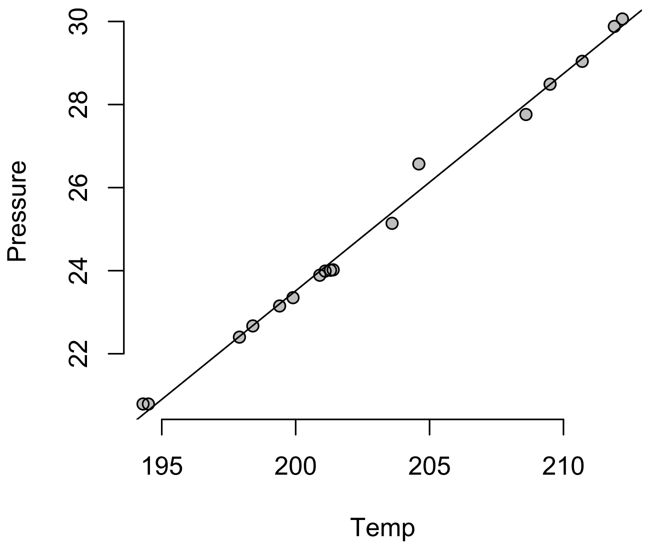
Notes 21. Linear Regression

Introduction

A most widely used statistical analysis is regression, where one tries to explain a response variable Y by an explanatory variable X based on paired data $(X_1, Y_1), \cdots, (X_n, Y_n)$. The most common way to model the dependence of Y on X is to look for a linear relationship with additional noise,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

with $\epsilon_1, \cdots, \epsilon_n$ taken to be independent and identically distributed random variables with mean 0 and variance σ^2. The unknown quantities are $(\beta_0, \beta_1, \sigma^2)$. Many pairs of natural measurements exhibit such linear relationships. The figure below shows atmospheric pressure (in inches of Mercury) against boiling point of water (in degrees F) based on 17 pairs of observations. Although water's boiling point and atmospheric pressure should have a precise physical relationship, there would always be some deviation in actual measurements due to factors that are hard to control.



Least squares line

The straight line you see in the above figure is the line that "best fits" the data. This is found as follows. For any line $y = b_0 + b_1 x$, we can find the "residuals" $e_i = y_i - b_0 - b_1 x_i$ if we tried to explain the observed values of Y by those of X using this line. The total deviation can be measured by the sum of squares of the residuals $d(b_0, b_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$. We could find b_0 and b_1 that minimize $d(b_0, b_1)$. This is easily done by using calculus. We set $\frac{\partial}{\partial b_0} d(b_0, b_1) = 0$, $\frac{\partial}{\partial b_1} d(b_0, b_1) = 0$ and solve for b_0, b_1. In particular,

$$0 = \frac{\partial}{\partial b_0} d(b_0, b_1) = \sum_{i=1}^n 2(y_i - b_0 - b_1 x_i)(-1) = -2n(\bar{y} - b_0 - b_1 \bar{x})$$

$$0 = \frac{\partial}{\partial b_1} d(b_0, b_1) = \sum_{i=1} 2(y_i - b_0 - b_1 x_i)(-x_i) = -2(\sum_{i=1}^n x_i y_i - n b_0 - b_1 \sum_{i=1}^n x_i^2).$$

These are two linear equations in two unknowns b_0, b_1. The solutions are:

$$\hat{b}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}.$$

Let's denote $s_{xy} = \frac{1}{n-1}(y_i - \bar{y})(x_i - \bar{x})$ [this is the sample covariance between Y and X]. Then using our old notation $s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ we can write the least squares solution as $\hat{b}_0 = \bar{y} - s_{xy}\bar{x}/s_x^2$, $\hat{b}_1 = s_{xy}/s_x^2$.

The method of least squares was used by physicists working on astrological measurements in the early 18th century. A statistical framework was developed much later. The main import of the statistical development, as usual, has been to incorporate a notion of uncertainty.

Statistical analysis of simple linear regression

To put the linear regression relationship $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ into a statistical model, we need a distribution on the ϵ_i's. The most common choice is a normal distribution $\mathsf{Normal}(0, \sigma^2)$. This can be justified as follows: the additional factors that give rise to the noise term are many in number and act independently of each other, each making a little contribution. By the central limit theorem the aggregate of such numerous, independent, small contributions should behave like a normal variable. The mean is fixed at zero because any non-zero mean can be absorbed in the intercept β_0.

So our statistical model is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $\epsilon_i \overset{\text{IID}}{\sim} \mathsf{Normal}(0, \sigma^2)$ with model parameters $\beta_0 \in (-\infty, \infty)$, $\beta_1 \in (-\infty, \infty)$, and $\sigma^2 > 0$. We also need to assume that the error terms ϵ_i's are independent of the explanatory variables X_i (because the errors account for additional factors beyond the explanatory variable). Then, the log-likelihood function in the model parameters is given by,

$$\ell_{x,y}(\beta_0, \beta_1, \sigma^2) = \text{const} - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}.$$

To find the MLE, first notice that for every σ^2, the log-likelihood function is maximized at (β_0, β_1) equal to the least squares solutions $(\bar{y} - s_{xy}\bar{x}/s_x^2, s_{xy}/s_x^2)$, and so we must have $\hat{\beta}_{1,\text{MLE}}(x, y) = \bar{y} - s_{xy}\bar{x}/s_x^2$, $\hat{\beta}_{2,\text{MLE}}(x, y) = s_{xy}/s_x^2$. To shorten notations we'll just write $\hat{\beta}_0$ and $\hat{\beta}_1$ for $\hat{\beta}_{0,\text{MLE}}(x, y)$ and $\hat{\beta}_{1,\text{MLE}}(x, y)$ respectively.

Consequently, the MLE of σ^2 is found by setting $\frac{\partial}{\partial \sigma^2}\ell_{x,y}(\hat{\beta}_0, \hat{\beta}_1, \sigma^2) = 0$. This given,

$$\hat{\sigma}^2_{\text{MLE}}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

It is more common to estimate σ^2 by

$$\hat{\sigma}^2 = s^2_{y|x} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

where $n-2$ indicates that two unknown quantities (β_0 and β_1) were to be estimated to define the residuals.

Sampling theory

To construct confidence intervals and test procedures for the model parameters, we'll first need to look at their sampling theory. We'll explore these treating the explanatory variable as non-random, i.e., as if we picked and fixed the observations $x_1, \cdots, x_n$. You can interpret this as being the conditional sampling theory of the estimated model parameters given the observed explanatory variables.

It turns out that when $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with $\epsilon_i \overset{\text{IID}}{\sim} \mathsf{Normal}(0, \sigma^2)$, for any two numbers a and b,

1. $a\hat{\beta}_0 + b\hat{\beta}_1 \sim \mathsf{Normal}\left(a\beta_0 + b\beta_1, \sigma^2\left\{\frac{a^2}{n} + \frac{(a\bar{x}-b)^2}{(n-1)s_x^2}\right\}\right)$.
2. $(n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-2)$
3. and these two random variables are mutually independent.

A proof of this is given at the end of this handout. It is only for those who are curious to know why this happens. You may ignore it if you're not interested.

Two special cases of the above result would be important to us. First, with $a = 1$ and $b = 0$ the result gives $\hat{\beta}_0 \sim \mathsf{Normal}(\beta_0, \sigma^2\{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\})$ and is independent of $\hat{\sigma}^2$. Second, with $a = 0$ and $b = 1$ we have $\hat{\beta}_1 \sim \mathsf{Normal}(\beta_1, \sigma^2\frac{1}{(n-1)s_x^2})$ and is independent of $\hat{\sigma}^2$.

ML confidence interval

We shall only look at confidence intervals for parameters that can be written as $\eta = a\beta_0 + b\beta_1$ for some real numbers a and b (again, the interesting cases would be $(a = 1, b = 0)$ corresponding to β_0 and $(a = 0, b = 1)$ corresponding to β_1). From the result above it follows that $\hat{\eta} = a\hat{\beta}_0 + b\hat{\beta}_1 \sim \mathsf{Normal}(\eta, \sigma^2_\eta)$ where $\sigma^2_\eta = \sigma^2\left[\frac{a^2}{n} + \frac{(a\bar{x}-b)^2}{(n-1)s_x^2}\right]$ and $\hat{\eta}$ is independent of $\hat{\sigma}^2$.

Let $\hat{\sigma}^2_\eta = \hat{\sigma}^2\left[\frac{a^2}{n} + \frac{(a\bar{x}-b)^2}{(n-1)s_x^2}\right]$. A bit of algebra shows that

$$\frac{\hat{\eta} - \eta}{\hat{\sigma}_\eta} \sim t(n-2).$$

Therefore a $100(1-\alpha)\%$ confidence interval for η is given by

$$B(x,y) = \hat{\eta} \mp \hat{\sigma}_\eta z_{n-2}(\alpha).$$

This confidence interval is also the ML confidence interval, i.e., it matches $\{\ell^*_{x,y}(\eta) \ge \max_\eta \ell^*_{x,y}(\eta) - c^2/2\}$ for some $c > 0$ where $\ell^*_{x,y}(\eta)$ is the pofile log-likelihood of η, but we would not pursue a proof here.

Hypotheses testing

Again we restrict ourselves to parameters that can be written as $\eta = a\beta_0 + b\beta_1$. To test $H_0 : \eta = \eta_0$ against $H_1 : \eta \neq \eta_1$, the size-α ML test is the one that rejects H_0 whenever η_0 is outside the $100(1-\alpha)\%$ ML confidence interval $\hat{\eta} \mp \hat{\sigma}_\eta z_{n-2}(\alpha)$. The p-value based on these tests is precisely the α for which η_0 is just on the boundary of this interval.

Of particular interest is testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. The null hypothesis says that X offers no explanation of Y within the linear relationship framework. This can be dealt as above with $a = 0$ and $b = 1$ and thus checking whether 0 is outside the interval $\hat{\beta}_1 \mp \hat{\sigma} z_{n-2}(\alpha)/\sqrt{(n-1)s_x^2}$.

Prediction

Suppose you want to predict the value Y^* of Y corresponding to an $X = x^*$ that is known to you. The model says $Y^* = \beta_0 + \beta_1 x^* + \epsilon^*$ where $\epsilon^* \sim \mathsf{Normal}(0, \sigma^2)$ and is independent of $\epsilon_1, \cdots, \epsilon_n$. If we knew β_0, β_1, σ, we could give the interval $\beta_0 + \beta_1 x^* \mp \sigma z(\alpha)$, taking into account the variability σ^2 of ϵ^*. Our best guess for $\beta_0 + \beta_1 x^*$ is the fitted value $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$ (the point on the least squares line with x-coordinate x^*) with additional associated variability $\sigma^2\{\frac{1}{n} + \frac{(\bar{x}-x^*)^2}{(n-1)s_x^2}\}$ [follows from the result with $a = 1$, $b = x^*$]. Also, we have an estimate $\hat{\sigma}^2$ for σ^2. Putting all these together, it follows that the predictive interval

$$\hat{\beta}_0 + \hat{\beta}_1 x^* + \hat{\sigma}\left[1 + \frac{1}{n} + \frac{(\bar{x}-x^*)^2}{(n-1)s_x^2}\right]^{\frac{1}{2}} \cdot z_{n-2}(\alpha)$$

has a $100(1-\alpha)\%$ coverage.

Testing goodness of fit

Once we fit the model, we can construct our residuals $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, $i = 1, \cdots, n$. The histogram of these residuals should match the $\mathsf{Normal}(0, \sigma^2)$ pdf, with σ^2 estimated by $\hat{\sigma}^2$. So we could carry out a Pearson's goodness-of-fit test with these residuals in the same manner we did for iid normal data. The only difference would be that the p-value range would now be given by $1 - F_{k-1}(Q(x))$ to $1 - F_{k-4}((Q(x))$ to reflect that 3 parameters are being estimated (β_0, β_1 and σ).

Technical details (ignore if you're not interested)

To dig into the sampling distribution of $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\sigma}^2$, we first need to recall an important result we had seen a long time back (Notes 7 to be precise). We saw that if $Z_1, \cdots, Z_n \overset{\text{IID}}{\sim}$

$\mathsf{Normal}(0,1)$ and $W_1, \cdots, W_n$ relate to $Z_1, \cdots, Z_n$ through the system of equations:

$$\begin{aligned} W_1 &= a_{11}Z_1 + \cdots + a_{1n}Z_n \\ W_2 &= a_{21}Z_1 + \cdots + a_{2n}Z_n \\ &\vdots \\ W_n &= a_{n1}Z_1 + \cdots + a_{nn}Z_n \end{aligned}$$

then $W_1, \cdots, W_n$ are also independent $\mathsf{Normal}(0,1)$ variables with $W_1^2 + \cdots + W_n^2 = Z_1^2 + \cdots + Z_n^2$ provided the coefficients on each row form a vector of unit length and these vectors are orthogonal to each other. We had used this result to show that if $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(0,1)$ then $\bar{X} \sim \mathsf{Normal}(0, 1/\sqrt{n})$ and $\sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2(n-1)$ and these two are independent of each other.

We will pursue a similar track here. First we note that when $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with $\epsilon_i \overset{\text{IID}}{\sim} \mathsf{Normal}(0, \sigma^2)$,

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 \bar{x} + \bar{\epsilon}$$

and with $u_i = (x_i - \bar{x}) / \sum_{i=1}^n (x_i - \bar{x})^2$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n u_i Y_i = \sum_{i=1}^n u_i (\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_1 + \sum_{i=1}^n u_i \epsilon_i$$

because $\sum_{i=1}^n u_i = 0$ and $\sum_{i=1}^n u_i x_i = 1$.

Define $\zeta_1 = \epsilon_1/\sigma, \cdots, \zeta_n = \epsilon_n/\sigma$, then $\zeta_1, \cdots, \zeta_n \overset{\text{IID}}{\sim} \mathsf{Normal}(0,1)$ and from calculations above $\bar{Y} = \beta_0 + \beta_1 \bar{x} + \sigma \bar{\zeta}$ and $\hat{\beta}_1 = \beta_1 + \sigma \sum_{i=1}^n u_i \zeta_i$. Now get $W_1, \cdots, W_n$ as

$$\begin{aligned} \sqrt{n}\,\bar{\zeta} = W_1 &= \frac{1}{\sqrt{n}}\zeta_1 + \cdots + \frac{1}{\sqrt{n}}\zeta_n \\ \left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^{1/2} \sum_{i=1}^n u_i \zeta_i = W_2 &= \frac{x_1 - \bar{x}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}\zeta_1 + \cdots + \frac{x_n - \bar{x}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}\zeta_n \\ W_3 &= a_{31}\zeta_1 + \cdots + a_{n1}\zeta_n \\ &\vdots \\ W_n &= a_{n1}\zeta_1 + \cdots + a_{nn}\zeta_n \end{aligned}$$

where the rows are unit length and mutually orthogonal (the first two rows are so by design, and so can be extended to a set of n rows by what is known as Gram-Schmidt orthogonalization). So we have $W_i \overset{\text{IID}}{\sim} \mathsf{Normal}(0,1)$ and $W_1^2 + \cdots + W_n^2 = \zeta_1^2 + \cdots + \zeta_n^2$.

This leads to a rich collection of results. First we see that $\hat{\beta}_1 = \beta_1 + \sigma W_2 / \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$ with $W_2 \sim \mathsf{Normal}(0,1)$. Hence we must have

$$\hat{\beta}_1 \sim \mathsf{Normal}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

Next,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} = \beta_0 + \sigma \left[\frac{W_1}{\sqrt{n}} - \frac{\bar{x} W_2}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]$$

with W_1, W_2 independent $\mathsf{Normal}(0,1)$. Therefore,

$$\hat{\beta}_0 \sim \mathsf{Normal}\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \right).$$

Furthermore, for any two numbers a and b

$$\begin{aligned} a\hat{\beta}_0 + b\hat{\beta}_1 &= a\beta_1 + b\beta_0 + \sigma \left[\frac{aW_1}{\sqrt{n}} - \frac{(a\bar{x} - b) W_2}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \right] \\ &\sim \mathsf{Normal}\left(a\beta_0 + b\beta_1, \sigma^2 \left\{ \frac{a^2}{n} + \frac{(a\bar{x} - b)^2}{(n-1)s_x^2} \right\} \right). \end{aligned}$$

Next we look at $\hat{\sigma}^2$. First note that,

$$Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = \epsilon_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i = \sigma \left[\zeta_i - \frac{W_1}{\sqrt{n}} - \frac{(x_i - \bar{x}) W_2}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]$$

which leads to the following identity (with a bit of algebra that you can ignore)

$$\frac{\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sigma^2} = \sum_{i=1}^n \zeta_i^2 - W_1^2 - W_2^2 = W_3^2 + \cdots + W_n^2$$

because $\sum_{i=1}^n \zeta_i^2 = \sum_{i=1}^n W_i^2$. But $W_3^2 + \cdots + W_n^2$ is the sum of $n-2$ independent $\mathsf{Normal}(0,1)$ variables, so

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sigma^2} \sim \chi^2(n-2)$$

and because $W_3, \cdots, W_n$ are independent of W_1, W_2, we can conclude $\hat{\sigma}^2$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$.