Regression Methods in Statistics

Lawrence Arscott

February 12, 2024

A brief and informal treatment of statistical regression methods to accompany R code found here. Assumes some basics in statistics, such as an understanding of the expectation value of a probability distribution and its properties. For further details, see for example [1].

1 Basics

1.1 Definitions

Definition 1.1 (Variance)**.**
The variance of a probability distribution is the expected squared deviation of an observation.

$$\begin{aligned} \mathrm{Var}(x) &= E\left((x - E(x))^2\right) \\ &= E(x^2) - E(x)^2 \end{aligned} \tag{1}$$

Remark 1.1 (Sample variance)**.**
*When estimating population variance from a sample, multiply the variance of the sample by $\frac{n}{n-1}$, where n is sample size. This estimation of the population variance using the sample is usually referred to as the **sample variance**. See statement A.1 in appendix for proof.*

Definition 1.2 (Covariance)**.**

$$\begin{aligned} \mathrm{Cov}(x, y) &= E\big((x - E(x)) \cdot (y - E(y))\big) \\ &= E(xy) - E(x) \cdot E(y) \end{aligned}$$

- Note $\mathrm{Cov}(x, x) = \mathrm{Var(x)}$

1.2 Properties

Variance of sums of variables:

$$\begin{aligned}\operatorname{Var}\left(\sum_i x_i\right) &= E\left(\left(\sum_i x_i\right)^2\right) - E\left(\sum_i x_i\right)^2 \\ &= \sum_i \sum_j E(x_i x_j) - \left(\sum_i E(x_i)\right)^2 \\ &= \sum_i \sum_j E(x_i x_j) - \sum_i \sum_j E(x_i)E(x_j) \\ &= \sum_i \sum_j \operatorname{Cov}(x_i, x_j)\end{aligned}$$

- thus for variables that do not correlate, $\operatorname{Var}\left(\sum_i x_i\right) = \sum_i \operatorname{Var}(x_i)$

Variance of product of **independent** variables:

$$\operatorname{Var}(xy) = \operatorname{Var}(x)\operatorname{Var}(y) + \operatorname{Var}(x)E(y)^2 + E(x)^2\operatorname{Var}(y)$$

Distributivity of covariance:

$$\operatorname{Cov}(x, y+z) = \operatorname{Cov}(x,y) + \operatorname{Cov}(x,z)$$

2 Regression

2.1 Least squares approach: simple linear regression

We have a set of data points (x_i, y_i). We search for an underlying linear relationship $y = \alpha + \beta x$ and assume the y_i are observed with a small error term ϵ_i:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

We derive estimates $\hat{\alpha}, \hat{\beta}$ of α, β by minimising the resulting sum of squared errors:

$$\begin{aligned} f &= \sum_i (\epsilon_i)^2 \\ &= \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \end{aligned}$$

We assume:

- The errors do not correlate
- The errors are normally distributed with mean 0 and some variance σ^2:

$$\epsilon_i \sim N(0, \sigma^2)$$

Through considering partial derivatives, we get

$$\begin{aligned} \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\ &= S_{xy}/S_{xx} \end{aligned} \tag{2}$$

Note since $\sum_i (x_i - \bar{x} = 0)$, we may simplify to $\hat{\beta} = \dfrac{\sum_i (x_i - \bar{x}) y_i}{S_{xx}}$ which we use later.

From our assumptions, it may be derived that our estimators $\hat{\alpha}$ and $\hat{\beta}$ are normally distributed. We know go about finding expectation value and variance.

2.1.1 Expectation value for, and variance of, beta-hat

Here we assume the x_i to be known fixed constants.

$$\begin{aligned} E(\hat{\beta}) &= E\left(\frac{\sum_i (x_i - \bar{x}) y_i}{S_{xx}}\right) \\ &= \frac{1}{S_{xx}} \sum_i (x_i - \bar{x}) E(y_i) \\ &= \frac{1}{S_{xx}} \sum_i (x_i - \bar{x})(\alpha + \beta x_i) \\ &= \frac{1}{S_{xx}} \sum_i (x_i - \bar{x}) \beta x_i \\ &= \beta \end{aligned}$$

since $\sum_i (x_i - \bar{x}) x_i = S_{xx}$.
Hence $\hat{\beta}$ is an **unbiased estimator** of β.

We now turn to the variance in our estimation. Recall σ^2 to be the variance in the errors in observing the y_i.

$$\begin{aligned}
\operatorname{Var}(\hat{\beta}) &= \operatorname{Var}\left(\frac{\sum_i (x_i - \bar{x}) y_i}{S_{xx}}\right) \\
&= \frac{1}{(S_{xx})^2} \sum_i (x_i - \bar{x})^2 \operatorname{Var}(y_i) \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

where we used the fact the errors in observing the y_i are not correlated.

Hence $\hat{\beta}$ follows the normal distribution

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right) \tag{3}$$

2.1.2 Expectation value for, and variance of, alpha-hat

It follows from eq.2 of the last section that

$$E(\hat{\alpha}) = \alpha$$

hence $\hat{\alpha}$ is an unbiased estimator of α.

Next we consider $\operatorname{Var}(\hat{\alpha}) = \operatorname{Var}(\bar{y} - \hat{\beta}\bar{x})$. We start by considering $\operatorname{Cov}(\bar{y}, \hat{\beta})$:

$$\begin{aligned}
\operatorname{Cov}(\bar{y}, \hat{\beta}) &= \operatorname{Cov}\left(1/n \sum_i y_i, \frac{\sum_j (x_j - \bar{x}) y_j}{S_{xx}}\right) \\
&= \frac{1}{nS_{xx}} \sum_j (x_j - \bar{x}) \operatorname{Cov}\left(\sum_i y_i, y_j\right) \\
&= \frac{1}{nS_{xx}} \sum_i \sum_j (x_j - \bar{x}) \operatorname{Cov}(y_i, y_j) \\
&= \frac{\sigma^2}{nS_{xx}} \sum_i (x_i - \bar{x}) \\
&= 0
\end{aligned}$$

Hence

$$\begin{aligned}
\operatorname{Var}(\hat{\alpha}) &= \operatorname{Var}(\bar{y}) + \bar{x}^2 \operatorname{Var}(\hat{\beta}) \\
&= \sigma^2 \left(\frac{1}{n} - \frac{\bar{x}^2}{S_{xx}}\right)
\end{aligned}$$

2.1.3 Estimation of error variance

It can be shown that an unbiased estimate for $\text{Var}(\epsilon_i) = \sigma^2$ is given by:

$$\widehat{\sigma^2} = S^2 = \frac{\sum_i (\epsilon_i)^2}{n-2} \tag{4}$$

Eq. 4 can then be used to estimate $\text{Var}(\hat{\alpha})$ and $\text{Var}(\hat{\beta})$:

$$\widehat{\text{Var}}(\hat{\alpha}) = S^2 \left(\frac{1}{n} - \frac{\bar{x}^2}{S_{xx}} \right)$$
$$\widehat{\text{Var}}(\hat{\beta}) = \frac{S^2}{S_{xx}}$$

2.1.4 Hypothesis testing

Say we wish to test $H_0 : \beta = 0$.
It is tempting to hypothesis test using the normal distribution (eq. 3), substituting our unbiased estimator S^2 for σ^2. We must however take into account the whole spread of possible underlying σ^2 weighted by how likely they are to produce the observation S^2. In fact, the discrepancy $S^2 - \sigma^2$ varies proportionally to a *chi-squared distribution* with $n - 2$ degrees of freedom. Taking this into account leads us to construct the *t-value*:

$$t = \frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} \tag{5}$$

where $s_{\hat{\beta}} = \sqrt{\widehat{\text{Var}}(\hat{\beta})}$ is the (unbiased) estimator for the standard deviation of $\hat{\beta}$.

t is then distributed according to a *t-distribution* with $n - 2$ degrees of freedom:

$$t \sim t_{n-2}$$

We won't go into details of the distribution's derivation, and simply remember that it suitably tests H_0.

2.2 Multiple linear regression

The extension to multiple control variables, say some data set $(\underline{x}_i, y_i)$ is natural. It is essentially the same minimisation problem:

$$f = \sum_i (y_i - \hat{\alpha} - \underline{\hat{\beta}} \cdot \underline{x}_i)^2$$

We easily get

$$\hat{\alpha} = \bar{y} - \underline{\hat{\beta}} \cdot \underline{\bar{x}}$$

while $\hat{\beta}$ requires further thought.
We should be convinced we can recover a linear regression on $(\underline{x}_i, y_i)$ from a translated and scaled data set

$\left(\frac{x_i - A}{c_1}, \frac{y_i - B}{c_2}\right)$. With this in mind, we simplify things by setting means to 0 and variances to 1. Note this drops $\hat{\alpha}$, and we are searching for an optimal relation

$$\underline{y} = X\underline{\hat{\beta}}$$

where y_i corresponds to observation i, and entry X_{ij} in the matrix X is control variable j for observation i. We are looking to minimise

$$f = (\underline{y} - X\underline{\hat{\beta}}) \cdot (\underline{y} - X\underline{\hat{\beta}})$$

Taking partial derivatives yields a minimum when

$$X^T X\underline{\hat{\beta}} = X^T \underline{y}$$

(proof given as statement A.2 in appendix)
For further details of the derivation, see [2]

A Proofs

Statement A.1 (Unbiased estimator of population variance)**.**
Assume a normally distributed population with mean μ and variance σ^2. Consider a sample $\{x_i\}_{i=1}^n$. We show $s^2 = \dfrac{\sum_i (x_i - \bar{x})^2}{n-1}$ is an unbiased estimator for σ^2.

Proof. We will use

$$\begin{aligned} E(x^2) &= \mathrm{Var}(x) + (E(x))^2 \\ &= \sigma^2 + \mu^2 \end{aligned}$$

and

$$\begin{aligned} E(\bar{x}^2) &= \mathrm{Var}(\bar{x}) - (E(\bar{x}))^2 \\ &= \sigma^2/n - \mu^2 \end{aligned}$$

to get

$$\begin{aligned} E(s^2) &= \frac{1}{n-1} E\left(\sum_i (x_i - \bar{x})^2\right) \\ &= \frac{1}{n-1} E\left(\sum_i (x_i)^2 - 2\bar{x}\sum_i x_i + n\bar{x}^2\right) \\ &= \frac{1}{n-1} E\left(\sum_i (x_i)^2 - 2\bar{x}\sum_i x_i + n\bar{x}^2\right) \\ &= \frac{1}{n-1} \left(\sum_i E(x_i^2) - nE(\bar{x}^2)\right) \\ &= \frac{1}{n-1} \left(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right) \\ &= \sigma^2 \end{aligned}$$

□

Statement A.2. *Setting the derivative of $f = (\underline{y} - X\underline{\hat{\beta}}) \cdot (\underline{y} - X\underline{\hat{\beta}})$ w.r.t. $\underline{\hat{\beta}}$ to 0 yields*

$$X^T X \underline{\hat{\beta}} = X^T \underline{y}$$

Proof. We use the Einstein summation convention. We have:

$$\begin{aligned} f &= (\underline{y} - X\underline{\hat{\beta}})_i (\underline{y} - X\underline{\hat{\beta}})_i \\ &= (y_i X_{ij} \hat{\beta}_j)(y_i X_{ik} \hat{\beta}_k) \\ &= y_i y_i - 2X_{ik}\hat{\beta}_k y_i + X_{ij} X_{ik} \hat{\beta}_j \hat{\beta}_k \end{aligned}$$

Then taking the partial derivative:

$$\frac{\partial f}{\partial \hat{\beta}_m} = -2X_{im} y_i + 2X_{im} X_{ik} \hat{\beta}_k$$

Setting this to 0, we get for all m:

$$X_{im} y_i = X_{im} X_{ik} \hat{\beta}_k$$

or indeed

$$X^T X \underline{\hat{\beta}} = X^T \underline{y}$$

$\square$

References

[1] Simon J. Sheather. *Modern approach to regression with R (Springer texts in statistics)*. Springer, 2009. ISBN: 978-0-387-09607-0.

[2] Marco Taboga. *Linear regression with standardized variables*. 2021. URL: `https://www.statlect.com/fundamentals-of-statistics/linear-regression-with-standardized-variables`.