Inference in Linear Regression

Guy Lebanon

May 17, 2010

One of the powerful aspects of linear regression[1] is that the distribution of the least square estimator $\hat{\beta}$ is given in closed form. This leads to a wide range of inference tools and analytical results. We start with a basic result that will be needed later.

Proposition 1. *Let X be a random vector with mean μ and variance Σ. Then $\mathsf{E}(X^\top AX) = tr(A\Sigma)+\mu^\top A\mu$.*

Proof.

$$\begin{aligned}\mathsf{E}(X^\top AX) = \mathrm{tr}(\mathsf{E}(X^\top AX) = \mathsf{E}(\mathrm{tr}(X^\top AX) = \mathsf{E}(\mathrm{tr}(AXX^\top) = \mathrm{tr}(\mathsf{E}(AXX^\top)) = \mathrm{tr}(A\mathsf{E}(XX^\top))\\ = \mathrm{tr}(A(\mathsf{Var}(X) + \mu\mu^\top) = \mathrm{tr}(A\Sigma) + \mathrm{tr}(A\mu\mu^\top) = \mathrm{tr}(A\Sigma) + \mu^\top A\mu.\end{aligned}$$

□

The quality of fit of the least squares predictor to the training data is often measured through the concepts of the residual vector $\mathbf{e} \stackrel{\text{def}}{=} \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$ whose distribution is $\mathbf{e} \sim N(0, \sigma^2(\mathbf{I} - \mathbf{H}))$ since it is a linear function of $\mathbf{Y}$ and

$$\mathsf{Var}(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\mathsf{Var}(\mathbf{Y})(\mathbf{I} - \mathbf{H})^\top = (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})^\top = \sigma^2(\mathbf{I} - \mathbf{H}).$$

The residual sum of squares (RSS) measures the overall prediction distortion of $\hat{\beta}$ on the training set

$$\begin{aligned}\mathrm{RSS} = \mathbf{e}^\top\mathbf{e} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{H})^\top(\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{H})^2\mathbf{Y} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{H})\mathbf{Y} \quad \text{or alternatively}\\ = \mathbf{e}^\top\mathbf{e} = (\mathbf{Y} - \mathbf{X}\hat{\beta})^\top(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{Y}^\top\mathbf{Y} - 2\hat{\beta}^\top\mathbf{X}^\top\mathbf{Y} + \hat{\beta}^\top\mathbf{X}^\top\mathbf{X}\hat{\beta} = \mathbf{Y}^\top\mathbf{Y} - \hat{\beta}^\top\mathbf{X}^\top\mathbf{Y} + \hat{\beta}^\top(\mathbf{X}^\top\mathbf{X}\hat{\beta} - \mathbf{X}^\top\mathbf{Y})\\ = \mathbf{Y}^\top\mathbf{Y} - \hat{\beta}^\top\mathbf{X}^\top\mathbf{Y}.\end{aligned}$$

Proposition 2. *Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be a matrix of rank p. Then $S^2 \stackrel{\text{def}}{=} RSS/(n-p)$ is an unbiased estimator of σ^2 which is independent of $\hat{\beta}$. Furthermore, $RSS/((n-p)\sigma^2) = S^2/\sigma^2 \sim \chi^2_{n-p}$.*

Proof. Use Theorem 1, the fact that the trace of a projection matrix is its rank, and the fact that $\mathsf{E}(\mathbf{Y})$ is already projected by $\mathbf{H}$ to obtain

$$\begin{aligned}\mathsf{E}(\mathrm{RSS}) = \mathsf{E}(\mathbf{Y}^\top(\mathbf{I} - \mathbf{H})\mathbf{Y}) = \sigma^2\mathrm{tr}(\mathbf{I} - \mathbf{H}) + \mathsf{E}(\boldsymbol{Y})^\top(\mathbf{I} - \mathbf{H})\mathsf{E}(\mathbf{Y}) = \boldsymbol{\sigma^2}\mathrm{tr}(\mathbf{I} - \mathbf{H}) = \boldsymbol{\sigma^2}(\mathbf{n} - \mathbf{p}).\\ \mathsf{Cov}(\hat{\beta}, \mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathsf{Cov}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}, (\mathbf{I} - \mathbf{H})\mathbf{Y}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathsf{Var}(\mathbf{Y})(\mathbf{I} - \mathbf{H}) = 0\end{aligned}$$

and therefore $\hat{\beta}$ is independent of $\mathbf{e}$ (both are normal) and also of $\mathrm{RSS} = \mathbf{e}^\top\mathbf{e}$. Finally,

$$\mathrm{RSS} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{H})\mathbf{Y} = (\mathbf{Y} - \mathbf{X}\beta)^\top(\mathbf{I} - \mathbf{H})(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{e}^\top(\mathbf{I} - \mathbf{H})\mathbf{e}$$

which is a quadratic form with a rank $n - p$ matrix and therefore correspond to χ^2_{n-p} distribution. □

[1] Please read the companion note on linear regression to familiarize with definitions and notation

Since $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$, we can use pivotal quantities to obtain small sample (t-value) confidence intervals or hypothesis tests, for example the t-statistic

$$\sqrt{n-p}\frac{\beta_j - \hat{\beta}_j}{\sqrt{\text{RSS}}\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}} = \frac{\beta_j - \hat{\beta}_j}{\sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}}\frac{\sqrt{n-p}\sigma}{\sqrt{\text{RSS}}} \sim t_{n-p} \tag{1}$$

can be used to perform inference on the marginal value of β_j. It is important to realize that confidence intervals or hypothesis tests such as the ones based on (1) should be interpreted with respect to fixed $\mathbf{X}$. In other words, the randomness reflected in the confidence intervals will be due to the response variables $\mathbf{Y}$ while maintaining the same observed $\mathbf{X}$ over and over again.

Proposition 3. *If $Y \sim N(\mu, \Sigma)$ where $\Sigma \in \mathbb{R}^{n\times n}$ is positive definite , then $Q = (Y-\mu)^\top\Sigma^{-1}(Y-\mu) \sim \chi^2_n$.*

Proof. Standartizing by $Y = \Sigma^{1/2}Z + \mu$, $Z \sim N(0, I)$ we have $Q = Z^\top\Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2}Z = Z^\top Z = \sum_{i=1}^n Z_i$. □

As a result, we have that if $\mathbf{X}$ is a $n \times p$ matrix of full rank

$$\sigma^{-2}(\hat{\beta}-\beta)^\top\mathbf{X}^\top\mathbf{X}(\hat{\beta}-\beta) = (\hat{\beta}-\beta)^\top\mathsf{Var}\,(\hat{\beta})^{-1}(\hat{\beta}-\beta) \sim \chi^2_p.$$