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 \(\mu\) and variance \(\Sigma\). Then \(E(X^\top AX) = tr(A\Sigma) + \mu^\top A\mu\).

**Proof.**

\[
E(X^\top AX) = E(tr(X^\top AX)) = E(tr(AXX^\top)) = tr(AE(XX^\top)) = tr(A\Sigma) + tr(A\mu\mu^\top) = tr(A\Sigma) + \mu^\top A\mu.
\]

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

\[
\text{Var}(e) = (I - H)\text{Var}(Y)(I - H)^\top = (I - H)\sigma^2(I - H)^\top = \sigma^2(I - H).
\]

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

\[
\text{RSS} = e^\top e = Y^\top (I - H)^\top (I - H)Y = Y^\top (I - H)^2Y = Y^\top (I - H)Y \quad \text{or alternatively}
\]

\[
= e^\top e = (Y - X\hat{\beta})^\top (Y - X\hat{\beta}) = Y^\top Y - 2\hat{\beta}^\top X^\top Y + \hat{\beta}^\top X^\top X\hat{\beta} = Y^\top Y - \hat{\beta}^\top X^\top Y + \hat{\beta}^\top (X^\top X\hat{\beta} - X^\top Y)
\]

\[
= Y^\top Y - \hat{\beta}^\top X^\top Y.
\]

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

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

\[
E(\text{RSS}) = E(Y^\top (I - H)Y) = \sigma^2 tr(I - H) + E(Y)^\top (I - H)E(Y) = \sigma^2 tr(I - H) = \sigma^2(n - p).
\]

\[
\text{Cov}((\hat{\beta}, Y - X\hat{\beta})) = \text{Cov}((X^\top X)^{-1}X^\top Y, (I - H)Y) = (X^\top X)^{-1}X^\top \text{Var}(Y)(I - H) = 0
\]

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

\[
\text{RSS} = Y^\top (I - H)Y = (Y - X\beta)^\top (I - H)(Y - X\beta) = e^\top (I - H)e
\]

which is a quadratic form with a rank \(n - p\) matrix and therefore correspond to \(\chi^2_{n-p}\) distribution. \(\square\)

\(^1\)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

$$
\frac{\sqrt{n-p}}{\sqrt{\text{RSS} \sqrt{\text{Var}(\hat{\beta})}}} \sqrt{n-p} \sim t_{n-p} \quad (1)
$$

can be used to perform inference on the marginal value of $\beta_j$. 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 $\mathbf{Y} \sim N(\mu, \Sigma)$ where $\Sigma \in \mathbb{R}^{n \times n}$ is positive definite, then $Q = (\mathbf{Y} - \mu)^\top \Sigma^{-1} (\mathbf{Y} - \mu) \sim \chi^2_n$.

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

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

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