Regression Cheat Sheet

Kevin Penner

1 Notation

1. Population Regression: a (multiple) linear regression that describes a population; denoted $y = X\beta + \epsilon$ where

   - $y = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T$ is $n \times 1$ column vector of response variables; y is observed
   - X is $n \times p$ matrix of explanatory (independent) variables (n observations, p variables, $n > p$); X is observed
   - β is $p \times 1$ column vector of regression parameters
   - $\epsilon = \begin{bmatrix} \epsilon_1 & \epsilon_2 & \dots & \epsilon_n \end{bmatrix}^T$ is $n \times 1$ column vector of errors; ϵ is unobserved

2. Estimated Regression: least squares estimate of population regression; estimated regression might not equal the population regression due to measurement error, missing data, nonrandom sample, etc.; denoted $y = Xb + e$ where

   - y is $n \times 1$ column vector of response variables as above
   - X is $n \times p$ matrix of explanatory variables as above
   - $b = \begin{bmatrix} b_1 & b_2 & \dots & b_p \end{bmatrix}^T$ is an estimate of β
   - $e = \begin{bmatrix} e_1 & e_2 & \dots & e_n \end{bmatrix}^T$ is residual vector

2 Regression Model in Matrix Form

- Denote the $(i, j)^{th}$ element of the explanatory matrix X in the following way:

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1p} \\ \vdots & & \dots & \vdots \\ 1 & x_{n2} & \dots & x_{np} \end{bmatrix}.$$

  The column of ones corresponds to the constant term in the regression.

- Estimate β with b by minimizing sum of squared residuals (SSR), where

$$SSR(g) = \sum_{k=1}^{n} (y_k - x_k g)^2 \tag{1}$$

and g is any $p \times 1$ parameter vector.

- From a calculus point of view, for b to minimize SSR, b must satisfy the first order condition

$$\frac{\partial \ SSR(b)}{\partial \ g} = 0. \tag{2}$$

Using the fact that the derivative of each term in (1) is $-2(y_k - x_k g)x_k$, (2) is equivalent to

$$\sum_{k=1}^{n} x_k^T (y_k - x_k b) = X'e = 0 \text{ (vector)}. \tag{3}$$

Since the first column of X is all ones, (3) implies

$$\sum_{k=1}^{n} \underbrace{(y_k - b_1 - b_2 x_{k2} - \ldots - b_p x_{kp})}_{e_k} = 0,$$

i.e. the residuals always sum to 0 when an intercept is included in the equation.

3 Assumptions to make least squares estimators unbiased estimators of population parameters

Assumption 1: (Linear in parameters) The model can be written in the form $y = X\beta + \epsilon$.

Assumption 2: (Zero conditional mean) Conditional on the entire matrix X, each error ϵ_i has zero mean: $E[\epsilon|X] = 0$ (vector).

Assumption 3: (No perfect collinearity) In the sample, none of the independent variables is constant, and there are no *exact* linear relationships among the independent variables: X has rank p. Thus, $X^T X$ is nonsingular.

Assumption 4: (Homoskedasticity and no serial correlation)

1. $Var[\epsilon_i|X] = \sigma^2, \ i = 1, \ldots, n$. This is homoskedasticity: the variance of ϵ_i cannot depend on any element of X, and the variance must be constant across observations.
2. $Cov[\epsilon_i, \epsilon_j|X] = 0, \ i \neq j$.

Thus, $Var[\epsilon|X] = \sigma^2 I_n$.

Under these assumptions, we say b is the *best linear unbiased estimator*. Furthermore, under these assumptions, the unbiased estimator of the error variance σ^2 can be written

$$s^2 = e^T e/(n-p). \tag{4}$$

Before we prove this, we briefly discuss the "hat matrix" H that has leverage values on its diagonal:

- $H = X(X^T X)^{-1} X^T$
- If we take the SVD
$$X = U \begin{bmatrix} \Sigma_p \\ 0_{n-p} \end{bmatrix} V^T$$
where U is $n \times n$ orthogonal, Σ_p has positive diagonal (by Assumption 3), and V^T is $p \times p$ orthogonal, then
$$H = U \begin{bmatrix} I_p & \\ & 0_{n-p} \end{bmatrix} U^T$$
- $\hat{y} = Hy$, where $\hat{y}$ is a vector of fitted values.

With this in mind, we prove the unbiasedness of (4):

Proof.

$$\begin{aligned} e &= y - \hat{y} \\ &= (I-H)y \\ &= (I-H)X\beta + (I-H)\epsilon \\ &= (I-H)\epsilon. \end{aligned} \tag{5}$$

Since $(I-H)$ is symmetric and idempotent,

$$\begin{aligned} e^T e &= \epsilon^T (I-H)^T (I-H)\epsilon \\ &= \epsilon^T (I-H)\epsilon \\ &= trace(\epsilon^T (I-H)\epsilon) \text{ since scalar.} \end{aligned} \tag{6}$$

Thus,

$$\begin{aligned} E[\epsilon^T (I-H)\epsilon|X] &= E[tr(\epsilon^T (I-H)\epsilon)|X] \\ &= E[tr((I-H)\epsilon\epsilon^T)|X] \\ &= tr(E[(I-H)\epsilon\epsilon^T|X]) \\ &= tr((I-H)E[\epsilon\epsilon^T|X]) \\ &\quad \text{since } (I-H) \text{ is non-random} \\ &= tr((I-H)\sigma^2 I_n) \\ &= \sigma^2(n-p) \\ &\quad \text{since the trace of idempotent matrix is its rank.} \end{aligned} \tag{7}$$

Rearranging (7) using (6), we get

$$E[\epsilon^T(I-H)\epsilon|X]/(n-p) = \sigma^2 = E[e^Te/(n-p)|X]. \tag{8}$$

□

4 Leverage and residuals

Using (4), we can derive a relationship between leverage and residuals. Denote the variance-covariance matrix of e as $V(e)$. From (5), we have $e = (I-H)\epsilon$, so

$$\begin{aligned} e - E[e] &= (I-H)\epsilon - E[(I-H)\epsilon] \\ &= (I-H)\epsilon - (I-H)E[\epsilon] \\ &\quad \text{since } (I-H) \text{ is non-random} \\ &= (I-H)\epsilon \\ &\quad \text{since } E[\epsilon] = 0 \text{ by Assumption 2.} \end{aligned} \tag{9}$$

Then

$$\begin{aligned} V(e) &= E[(e-E[e])(e-E[e])^T] \\ &= (I-H)E[\epsilon\epsilon^T](I-H)^T \\ &= (I-H)I\sigma^2(I-H)^T \\ &\quad \text{since } Var[\epsilon] = E[\epsilon\epsilon^T] \text{ by Assumptions 2 and 4} \\ &= (I-H-H+H^2)\sigma^2 \\ &= (I-H)\sigma^2 \text{ since } H^2 = H, \end{aligned}$$

so the variance of an individual residual e_i, denoted $Var(e_i)$, is the i^{th} diagonal element of $V(e)$, which is $(1-h_{ii})\sigma^2$.

Since we do not know σ^2, we estimate the variance of e_i using s^2 from (4):

$$\begin{aligned} Var(e_i) &= (1-h_{ii})s^2 \\ &= (1-h_{ii})e^Te/(n-p) \\ &\leq (1-h_{ii})e^Te. \end{aligned}$$

Finally, from (5) and (9), we see $e - E[e] = (I-H)\epsilon = e$, so an alternate expression for $V(e) = E[(e-E[e])(e-E[e])^T]$ is $V(e) = ee^T$, so $Var(e_i) = e_i^2$.