STAT 151A: Interpretation of $\widehat{\beta}_j$

Billy Fang

Let

$$X = \begin{bmatrix} 1 & | & & | \\ \vdots & x_1 & \cdots & x_p \\ 1 & | & & | \end{bmatrix}$$

be our design matrix and assume $X^\top X$ is invertible.

Let $\widetilde{X}$ be matrix obtained by removing column x_1 from X. Let $H = X(X^\top X)^{-1}X^\top$ be the projection onto $C(X)$, and let $\widetilde{H} = \widetilde{X}(\widetilde{X}^\top \widetilde{X})^{-1}\widetilde{X}^\top$ be the projection onto $C(\widetilde{X})$.

Let $\widehat{\beta} = (X^\top X)^{-1}X^\top y$ be the least squares coefficients of regressing y onto X, and let $\widehat{y} = Hy = X\widehat{\beta}$ be the fitted values.

Similarly, $\widetilde{\widehat{y}} := \widetilde{H}y$ and $\widehat{x}_1 := \widetilde{H}x_1$ are the result of regressing y and x_1 respectively onto the columns of $\widetilde{X}$.

1 $\widehat{\beta}_1$ as the slope coefficient of a simple regression of residuals on residuals

Your lecture notes (Section 1.3 "Interpretation of $\widehat{\beta}$" in "Multiple Regression II") claim the following.

Proposition 1.1. $\widehat{\beta}_1$ is the slope coefficient from a simple regression of the residuals $y - \widetilde{\widehat{y}}$ onto the residuals $x_1 - \widehat{x}_1$.

[Note that this result can easily be modified to a statement about $\widehat{\beta}_j$ for some other j.]

Proof (optional). First note that $H\widetilde{H} = \widetilde{H}$ because $C(\widetilde{X}) \subseteq C(X)$. Therefore $C(H\widetilde{H}) = C(\widetilde{H}) = C(\widetilde{H}) \cap C(H)$, so (by HW1 Q3) we have

$$\widetilde{H}H = H\widetilde{H} = \widetilde{H}. \tag{1}$$

Moreover,

$$\begin{aligned}
\widetilde{\widehat{y}} &:= \widetilde{H}y \\
&= \widetilde{H}Hy && \text{using (1)} \\
&= \widetilde{H}\widehat{y} \\
&= \widetilde{H}(\widehat{\beta}_0\vec{1} + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_p x_p) \\
&= \widehat{\beta}_0\vec{1} + \widehat{\beta}_1\widetilde{H}x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_p x_p \\
&= \widehat{y} - \widehat{\beta}_1 x_1 + \widehat{\beta}_1\widetilde{H}x_1.
\end{aligned}$$

The second-to-last equality comes from distributing $\widetilde{H}$ over the sum and noting that $\widetilde{H}\vec{1} = \vec{1}$ and $\widetilde{H}x_j = x_j$ for all $j \neq 1$.

Therefore,

$$y - \widetilde{\widehat{y}} = (y - \widehat{y}) + \widehat{\beta}_1(x_1 - \widehat{x}_1)$$

A simple regression of $y - \widetilde{\widehat{y}}$ onto $x_1 - \widehat{x}_1$ would project $y - \widetilde{\widehat{y}}$ onto the span of $\vec{1}$ and $x_1 - \widehat{x}_1$, which is a subspace of $C(X)$ since $\vec{1}, x_1, \widehat{x}_1 \in C(X)$. Let $\widetilde{\widetilde{H}}$ denote the projection matrix onto this space. Since $y - \widehat{y} \in C(X)^\perp$, the fitted values from this simple regression can be written as

$$\widetilde{\widetilde{H}}(y - \widetilde{\widehat{y}}) = \widehat{\beta}_1(x_1 - \widehat{x}_1) = 0 \cdot \vec{1} + \widehat{\beta}_1(x_1 - \widehat{x}_1).$$

Thus $\widehat{\beta}_1$ is the slope coefficient in this simple regression. □

Some of the ingredients of the proof lead directly to the variance result below.

It may be hard to grasp the intuition behind this result. Drawing a geometric picture of projecting y onto some subspace $C(X)$ and a smaller subspace $C(\widetilde{X})$ may be helpful.

Alternatively, an extremely hand-wavy explanation is as follows. The residuals $y - \widetilde{\widehat{y}}$ represents the remaining "information" in the response variable y that was not explained by variables $x_2, \ldots, x_p$. Similarly, the residuals $x_1 - \widehat{x}_1$ represents the remaining "information" in the explanatory variable x_1 that was not explained by the other variables $x_2, \ldots, x_p$. Then $\widehat{\beta}_1$ is related to how much of the "remaining information in y" is explained by the "remaining information in x_1," via a simple regression. Again, this is completely non-rigorous.

2 The variance of $\widehat{\beta}_1$

The same lecture notes (and page 113 of the textbook) also claim the following.

Proposition 2.1.

$$\mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_{i1} - \widehat{x}_{i1})^2} = \frac{1}{1 - R_1^2} \cdot \frac{\sigma^2}{\sum_{i=1}^n (x_{i1} - \overline{x}_1)^2}$$

[Again, this is easily modified to get an expression for the variance of $\widehat{\beta}_j$ for some other j.]

Proof (optional). To prove the first equality, we use the fact that $\widehat{\beta}_1(x_1 - \widehat{x}_1) = \widehat{y} - \widetilde{\widehat{y}}$ (see previous proof). Recall (from lecture notes or Lab 3) also that $(I - H)y = (I - H)\epsilon$ and $(I - \widetilde{H})y = (I - \widetilde{H})\epsilon$, which together imply $(H - \widetilde{H})y = (H - \widetilde{H})\epsilon$.

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}_1)\|x_1 - \widehat{x}_1\|^2 &= \mathrm{Var}\left[(\widehat{\beta}_1(x_1 - \widehat{x}_1))^\top (\widehat{\beta}_1(x_1 - \widehat{x}_1))\right] \\
&= \mathrm{Var}[(\widehat{y} - \widetilde{\widehat{y}})^\top (\widehat{y} - \widetilde{\widehat{y}})] \\
&= \mathrm{Var}[y^\top (H - \widetilde{H})^\top (H - \widetilde{H}) y] \\
&= \mathrm{Var}[\epsilon^\top (H - \widetilde{H})^\top (H - \widetilde{H}) \epsilon] \\
&= \mathrm{Var}[\epsilon^\top (H - \widetilde{H}) \epsilon] \\
&= \sigma^2 \,\mathrm{tr}(H - \widetilde{H}) && \text{see lecture notes or Lab 3} \\
&= \sigma^2 && \mathrm{tr}(H - \widetilde{H}) = \mathrm{tr}(H) - \mathrm{tr}(\widetilde{H}) = (p+1) - p
\end{aligned}$$

To prove the second equality, it suffices to check the denominators are equal, i.e.

$$(1 - R_1^2)\|x_1 - \overline{x}_1\|^2 = \|x_1 - \widehat{x}_1\|^2.$$

This follows immediately from $\left(1 - \frac{\mathrm{RegSS}}{\mathrm{TSS}}\right)\mathrm{TSS} = \mathrm{RSS}$, where all the SS quantities are for the regression of x_1 onto the columns of $\widetilde{X}$. □

See your lecture notes and page 113 of the textbook for how to interpret this result. Recall that $\frac{\sigma^2}{\sum_{i=1}^n (x_{i1} - \overline{x}_1)^2}$ is the variance of the slope coefficient in *simple* regression of y onto x_1. The above result shows that when you do *multiple* regression with x_1 along with other variables, then the corresponding slope coefficient $\widehat{\beta}_1$ for x_1 is the same, but multiplied by the variance inflation factor $\frac{1}{1 - R_1^2}$, which is large if x_1 is very correlated with the other variables.

Note that the other formula $\mathrm{Var}(\widehat{\beta}_1) = \sigma^2 (X^\top X)^{-1}_{1,1}$ is therefore equal to the above. The reason why we used this formula more often is because it does not involve this extra regression (of x_1 onto the other variables). But the formulas in the proposition are useful for interpretation, as noted in the previous paragraph.