Statistics 312/612, fall 2016
Homework # 10
Due: Friday 9 December

Please attempt this homework by yourself, with no help from others. Please cite explicitly any sources that you use.

Consider again the usual least squares fit obtained by choosing the vector $b \in \mathbb{R}^p$ to minimize $\|y - Xb\|^2$, where $y \in \mathbb{R}^n$ and X is a given $n \times p$ matrix, not necessarily of full rank.

Suppose we are worried about the observation y_i. For concreteness take $i = 1$ and partition the matrices as

$$y = \begin{pmatrix} y_1 \\ Y \end{pmatrix} \qquad \text{and} \qquad X = \begin{pmatrix} w_1^T \\ W \end{pmatrix} \qquad \text{where } W := \begin{pmatrix} w_2^T \\ \vdots \\ w_n^T \end{pmatrix},$$

where Y is the $(n-1) \times 1$ vector $[y_2, \dots, y_n]'$ and W is the $(n-1) \times p$ matrix obtained by deleting the first row, w_1^T, from X.

Let $\mathcal{X}$ denote the subspace of $\mathbb{R}^n$ spanned by the columns of X. Let H denote the hat matrix for orthogonal projection onto $\mathcal{X}$. Let $\mathcal{W}$ denote the subspace of $\mathbb{R}^p$ spanned by $\{w_i : 2 \le i \le n\}$.

The least squares estimator $\widehat{b}$ is defined to minimize $\|y - Xb\|^2$. If $\text{rank}(X) = k < p$ then $\widehat{b}$ is not unique, but all solutions give the same fitted value $X\widehat{b} = \widehat{y} = Hy$, where H denotes the hat matrix for orthogonal projection onto $\mathcal{X}$. Similarly, the $\widehat{B}$ that minimizes $\|Y - Wb\|^2$ need not be unique but all solutions give the same $\widehat{Y} = W\widehat{B}$.

There are various diagnostic procedures that try to detect bad violations of the normality assumption. This Homework describes three seemingly different diagnostics that turn out to be almost equivalent.

[1] Write e_1 for the unit vector with 1 in the first position.

(i) (10 points) Show that

$$\|y - Xb - e_1 c\|^2 = (y_1 - w_1^T b - c)^2 + \|Y - Wb\|^2$$

and that the left-hand side is minimized by choosing b equal to any $\widehat{B}$ that minimizes $\|Y - Wb\|^2$ and then choosing $\widehat{c}$ appropriately. Find $\widehat{c}$.

Solution: For the decomposition, take the squared length of

$$y - Xb - ce_1 = \begin{pmatrix} y_1 - w_1^T b - c \\ Y - Wb \end{pmatrix}$$

The least squares $\widehat{B}$ minimizes the $\|Y - Wb\|^2$ contribution then $\widehat{c} = y_1 - w_1^T \widehat{B}$ minimizes the first contribution.

(ii) (5 points) Explain why $\widehat{c}$ takes the same value for all choices of $\widehat{B}$ in (i) if and only if w_1 lies in $\mathcal{W}$. Hint: overparametrized handout.

Solution: The columns of the $(n-1) \times p$ matrix W span a subspace $\text{span}(W)$ of $\mathbb{R}^{n-1}$. The columns of the $p \times (n-1)$ matrix W^T span a subspace $\mathcal{W} = \text{span}(W^T)$ of $\mathbb{R}^p$. The projection of Y onto $\text{span}(W)$ is $\widehat{Y} = W\widehat{B}$, where $\widehat{B}$ denotes any minimizer of $\|Y - Wb\|^2$.

If W has rank ℓ with singular value decomposition $W = U\Lambda V^T = \sum_{i\le\ell} \lambda_i u_i v_i^T$, where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_\ell > 0$, then

$$\widehat{Y} = \sum\nolimits_{i\le\ell} \langle Y, u_i\rangle u_i = UU^T Y.$$

The $p\times 1$ vectors $\{v_i : 1 \le i \le \ell\}$ provide an onb for $\mathcal{W}$ and $\{v_i : \ell < i \le p\}$ provide an onb for $\mathcal{W}^\perp$. If $b = \sum_{i\le p} t_i v_i$ then $Wb = \widehat{Y}$ iff

$$t_i = \begin{cases} \langle Y, u_i\rangle/\lambda_i & \text{for } 1 \le i \le \ell \\ \text{unconstrained} & \text{for } \ell < i \le p \end{cases}.$$

That is, $W\widehat{B} = \widehat{Y}$ iff

$$\widehat{B} = V\Lambda^{-1}U^T Y + g \qquad \text{with } g \in \mathcal{W}^\perp.$$

The difference $y_1 - w_1^T\widehat{B}$ takes the same value for every choice of $\widehat{B}$ iff $w_1^T g$ takes the same value for every g in $\mathcal{W}^\perp$. That happens iff $w_1^T g = 0$ for every g in $\mathcal{W}^\perp$, which is true only when $w_1 \perp \mathcal{W}^\perp$, that is, when $w_1 \in \mathcal{W}$.

Many of you confused the space $\mathcal{W}$ with the column space of W. Some of you even asserted that $\mathcal{W}$ (a subspace of $\mathbb{R}^p$) was the same as $\mathcal{X}$ (a subspace of $\mathbb{R}^n$) if $w_1 \in \mathcal{W}$.

(iii) (5 points) If $e_1 \in \mathcal{X}$ show that $H_{11} = 1$. *(Here H_{11} denotes the element $H[1,1]$.)*

(iv) (5 points) If $e_1 \in \mathcal{X}$ show that $w_1 \notin \mathcal{W}$. Hint: $w_1 = X^T e_1$.

(v) (5 points) If $e_1 \notin \mathcal{X}$ show that $H_{11} < 1$ and $w_1 \in \mathcal{W}$. Hint: $(I - H)e_1 \perp \mathcal{X}$.

SOLUTION: Split e_1 into orthogonal components in $\mathcal{X}$ and $\mathcal{X}^\perp$,

$$e_1 = He_1 + (I_n - H)e_1 = Xd + z$$

for some d in $\mathbb{R}^p$. By the orthogonality

$$1 = \|e_1\|^2 = \|He_1\|^2 + \|z\|^2 = H_{11} + \|z\|^2.$$

Here I have used the fact that $H = H^T = H^2$ to simplify $\|He_1\|^2 = e_1^T H^T He_1$ to $e_1^T He_1 = H_{11}$.

Thus $H_{11} = 1$ if and only if $z = 0$, which is equivalent to $e_1 \in \mathcal{X}$.

Similarly

$$z_1 = \langle z, e_1\rangle = \langle z, Xd\rangle + \langle z, z\rangle = \|z\|^2,$$

which shows that $z = 0$ if and only if $z_1 = 0$.

If $e_1 \in \mathcal{X}$ then $e_1 = He_1$ and $z = 0$, so that

$$\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = e_1 = Xd = \begin{pmatrix} w_1^T d \\ w_2^T d \\ \vdots \\ w_n^T d \end{pmatrix}$$

The vector d is orthogonal to $w_2, \dots, w_n$ but not to w_1. The vector w_1 cannot be a linear combination of $w_2, \dots, w_n$.

If $e_1 \notin \mathcal{X}$, then $z \ne 0$ and $z_1 = \|z\|^2 > 0$. Also, because z is orthogonal to $\mathcal{X}$,

$$0 = z^T X = z_1 w_1^T + z_2 w_2^T + \cdots + z_n w_n^T$$

which rearranges to $w_1 = (z_2 w_2 + \cdots + z_n w_n)/z_1 \in \mathcal{W}$.

From now on assume $e_1 \notin \mathcal{X}$. Write $\widetilde{\mathcal{X}}$ for the subspace spanned by the e_1 and the columns of X. Write κ for $\sqrt{1 - H_{11}}$, the length of the vector $z := (I - H)e_1$.

SOLUTION: If $e_1 \notin \mathcal{X}$ there are three orthogonal subspaces of $\mathbb{R}^n$ that enter the solution: the k-dimensional subspace $\mathcal{X}$ spanned by the columns of X; the 1-dimensional subspace $\mathcal{Q}$ spanned by q_0, the unit vector for which

$$e_1 = z + He_1 = \kappa q_0 + He_1;$$

and the $(n - k - 1)$-dimensional subspace $\widetilde{\mathcal{X}}^{\perp}$ that is orthogonal to $\widetilde{\mathcal{X}}$, which is spanned by the columns of X and e_1.

[2] Define $q_0 := z/\kappa$. Let $\{q_j : 1 \le j \le k\}$ be an onb for $\mathcal{X}$.

(i) (5 points) Explain why $\{q_j : 0 \le j \le k\}$ is an onb for $\widetilde{\mathcal{X}}$.

SOLUTION: We need to show

$$\widetilde{\mathcal{X}} = \text{SPAN}(X, e_1) = \text{SPAN}(X, q_0) = \text{SPAN}(q_1, \dots, q_k, q_0).$$

For $\text{SPAN}(X, e_1) \subseteq \text{SPAN}(X, q_0)$ note that $e_1 = \kappa q_0 + He_1$ and $He_1 \in \mathcal{X}$. For the other inclusion note that $q_0 = (e_1 - He_1)/\kappa$, a linear combination of e_1 and He_1.

(ii) (5 points) Show that $\widetilde{H} = H + q_0 q_0^T$ is the hat matrix for orthogonal projection onto $\widetilde{\mathcal{X}}$.

SOLUTION: The matrix for orthogonal projection onto $\mathcal{Q}$ is $H_0 = q_0 q_0^T$. The matrix for orthogonal projection onto $\mathcal{X}$ is $H = \sum_{1 \le j \le k} q_j q_j^T$. The matrix $\widetilde{H}$ for orthogonal projection onto $\widetilde{\mathcal{X}}$ is just the sum of $H + H_0 = \sum_{0 \le j \le k} q_j q_j^T$. The term $q_j q_j^T$ is the matrix for orthogonal projection onto the subspace spanned by q_j.

(iii) (10 points) Show that the component of y in the q_0 direction equals $\widehat{c}z$.

SOLUTION: The vector y is a sum of three orthogonal components $Hy + H_0 y + (I_n - \widetilde{H})y$. The projection onto $\widetilde{\mathcal{X}}$ equals $\widetilde{y} = (H + H_0)y = X\widehat{B} + \widehat{c}e_1$. Projection of $\widetilde{y}$ orthogonal to $\mathcal{X}$ kills off the Hy component, leaving

$$H_0 y = (I_n - H)\widetilde{y} = (I_n - H)(X\widehat{B} + \widehat{c}e_1) = \widehat{c}(I_n - H)e_1 = \widehat{c}z = \widehat{c}\kappa q_0.$$

(iv) (10 points) If $y \sim N(\mu, \sigma^2 I_n)$ with $\mu \in \mathcal{X}$, show that $\widehat{c} \sim N(0, \sigma^2/\kappa^2)$.

SOLUTION: Write y as $\mu + \sigma\xi$ where $\mu \in \mathcal{X}$ and $\xi \sim N(0, I_n)$. Remember that $q_0^T \xi \sim N(0, 1)$. Then

$$\kappa\widehat{c}q_0 = H_0(\mu + \sigma\xi) = \sigma H_0 \xi = \sigma q_0 q_0^T \xi.$$

The μ disappears because $q_0 \perp \mathcal{X}$. Thus $\widehat{c} = \sigma q_0^T \xi/\kappa \sim N(0, \sigma^2/\kappa^2)$.

[3] Define $\widehat{\sigma}^2 = \|y - X\widehat{b}\|^2/(n - k)$ and $\widehat{S}^2 = \|Y - W\widehat{B}\|^2/(n - k - 1)$. Suppose $y \sim N(\mu, \sigma^2 I_n)$ with $\mu \in \mathcal{X}$.

(i) (10 points) Show that the statistic

$$\text{ESR}_1 := \kappa\widehat{c}/\widehat{S} = q_0' y/\widehat{S}$$

has a t_{n-k-1} distribution.

SOLUTION: It helps to summarize what we know about the various subspaces and components before attempting the remaining questions. Most importantly it is vital to distinguish between projections onto $\mathcal{X}$ and projections onto $\widetilde{\mathcal{X}}$: It would cause great trouble if we used $\widehat{y}$ to denote both Hy and $\widetilde{H}y$. Once again write y as $\mu + \sigma\xi$ where $\mu \in \mathcal{X}$ and $\xi \sim N(0, I_n)$.

$$
\begin{aligned}
\widehat{y} &= Hy = X\widehat{b} = \mu + \sigma H\xi \\
\widetilde{y} &= \widetilde{H}y = X\widehat{B} + \widehat{c}e_1 = \widehat{y} + H_0 y = \mu + \sigma(H + H_0)\xi \\
H_0 y &= \widehat{c}z = \sigma H_0 \xi \qquad \text{so that } \kappa\widehat{c} = \sigma q_0^T \xi \\
\widehat{Y} &= \text{component of } Y \text{ in } \textsc{span}(W) \\
R &= Y - \widehat{Y} \\
r &= y - \widehat{y} = (I_n - H)y = \sigma(I_n - H)\xi = \sigma H_0\xi + \sigma(I_n - \widetilde{H})\xi \\
\widetilde{r} &= (I_n - \widetilde{H})y = \sigma(I_n - \widetilde{H})\xi
\end{aligned}
$$

The solution to Problem [1] provides another connection between $\widetilde{r}$ and the residual $R = Y - W\widehat{B}$ for the least squares problem with the first row removed:

$$
\widetilde{r} = y - X\widehat{B} - \widehat{c}e_1 = \begin{pmatrix} y_1 - w_1^T\widehat{B} - \widehat{c} \\ Y - W\widehat{b} \end{pmatrix} = \begin{pmatrix} 0 \\ R \end{pmatrix}.
$$

Consequently,

<1> $$\sigma^2\|(I_n - \widetilde{H})\xi\|^2 = \|\widetilde{r}\|^2 = \|R\|^2 = \|Y - W\widehat{B}\|^2 = (n-k-1)\widehat{S}^2$$

and, by orthogonality of $\mathcal{Q}$ and $\widetilde{\mathcal{X}}^{\perp}$,

<2> $$(n-k)\widehat{\sigma}^2 = \|r\|^2 = \|H_0 y + \widetilde{r}\|^2 = \|H_0 y\|^2 + \|\widetilde{r}\|^2 = \sigma^2\|H_0\xi\|^2 + \|R\|^2.$$

subspace:	$\mathcal{X}$	$\mathcal{Q} = \textsc{span}(q_0)$	$\widetilde{\mathcal{X}}^{\perp}$
dimension:	k	1	$n-k-1$
	$\widetilde{\mathcal{X}}$ (spanning $\mathcal{X}$ and $\mathcal{Q}$)		
		$\mathcal{X}^{\perp}$ (spanning $\mathcal{Q}$ and $\widetilde{\mathcal{X}}^{\perp}$)	
component of y:	$\widehat{y} = Hy = X\widehat{b}$	$q_0 q_0^T y = \widehat{c}z$	$\widetilde{r} = (I_n - \widetilde{H})y$
	$\widetilde{y} = \widetilde{H}y = X\widehat{B} + \widehat{c}e_1$ (spanning $\mathcal{X}$ and $\mathcal{Q}$)		
		$r = (I_n - H)y$ (spanning $\mathcal{Q}$ and $\widetilde{\mathcal{X}}^{\perp}$)	
sum of squares:		$(q_0^T y)^2 = \widehat{c}^2\kappa^2$	$\|\widetilde{r}\|^2 = (n-k-1)\widehat{S}^2$
		$\|r\|^2 = ((n-k)\widehat{\sigma}^2$ (spanning $\mathcal{Q}$ and $\widetilde{\mathcal{X}}^{\perp}$)	

Solution: Note that

$$
\text{ESR}_1 = \frac{\sigma q_0^T \xi}{\sqrt{\sigma^2\|(I_n - \widetilde{H})\xi\|^2/(n-k-1)}}.
$$

By orthogonality of q_0 and $\widetilde{\mathcal{X}}^{\perp}$, the numerator and denominator are independent, with $q_0^T\xi \sim N(0,1)$ and $\|(I_n - \widetilde{H})\xi\|^2 \sim \chi^2_{n-k-1}$, which is precisely the way to get a t_{n-k-1} distribution.

(ii) (extra credit) Define

$$\mathrm{ISR}_1 := \kappa\widehat{c}/\widehat{\sigma} = q_0' y/\widehat{\sigma}.$$

Show that ISR_1 is a monotonely increasing function of ESR_1.

SOLUTION: Temporarily write η for $q_0^T y = \widehat{c}\kappa = \sigma q_0^T \xi$. Note that $\mathrm{ESR}_1 = \eta/\widehat{S}$. From <2>,

$$(n-k)\widehat{\sigma}^2 = \|r\|^2 = \eta^2 + \|\widetilde{r}\|^2 = \eta^2 + (n-k-1)\widehat{S}^2.$$

Thus

$$\mathrm{ISR}_1 = \eta/\widehat{\sigma} = \frac{\eta\sqrt{n-k}}{\sqrt{\eta^2 + (n-k-1)\widehat{S}^2}} = \frac{\mathrm{ESR}_1\sqrt{n-k}}{\sqrt{\mathrm{ESR}_1^2 + (n-k-1)}}$$

For each positive constant C, the function $t \mapsto t/\sqrt{t^2 + C}$ is strictly increasing: it has derivative $C(t^2+p)^{-3/2}$.

[4] (extra credit) Define

$$\mathcal{D}_1 = \frac{\|X\widehat{b} - X\widehat{B}\|^2}{k\widehat{\sigma}^2}.$$

Show that $\mathcal{D}_1$ is a monotonely increasing function of $|\mathrm{ISR}_1|$.

SOLUTION: From the decomposition of y in the table,

$$X\widehat{b} - X\widehat{B} = \widehat{y} - (\widetilde{y} - \widehat{c}e_1) = \widehat{c}e_1 - \widehat{c}z = \widehat{c}He_1$$

so that

$$\|X\widehat{b} - X\widehat{B}\|/\widehat{\sigma} = |\widehat{c}|\sqrt{H_{11}}/\widehat{\sigma} = |\mathrm{ISR}_1|\sqrt{H_{11}}/\kappa.$$

Some of you noticed that $\mathcal{D}_1$ is identically 0 if $H_{11} = 0$, which happens iff $He_1 = 0$, that is, $e_1 \perp \mathcal{X}$. Equivalently $w_1^T = e_1^T X = 0$. In that case the model asserts that $y_1 \sim N(0, \sigma^2)$. Clearly y_1 is then of no use for estimating the β for which $\mu = X\beta$. The calculations give $\kappa = 1$ and $z = q_0 = e_1$ and $\widehat{c} = \eta = y_1$. The statistic ESR_1 simplifies to $y_1/\widehat{S}$. Despite what

https://en.wikipedia.org/wiki/Cook's_distance

says, there is no way to standardize a random variable that is identically zero to give it an F distribution.