Stat 151a Linear Models
Homework 3 Solutions

October 14, 2015

1. For an orthonormal set $u_1, \ldots, u_n$ we will freely use the observation (which you should be familiar with by now) that

$$(u_i, u_j) = u_i^T u_j = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

   (a) Since $u_1, \ldots, u_n$ is an orthonormal basis of $\mathbb{R}^n$, we can write any vector y as a linear combination of the elements of this basis. That is $y = \sum_i c_i u_i$. Now let us take inner product with a particular u_j,

$$(y, u_j) = (\sum_i c_i u_i, u_j) = c_j$$

   implying that for each j, $c_j = (y, u_j)$. Plugging this back in $y = \sum_i (y, u_i) u_i$ as asserted. Note that this is the unique expansion of y in terms of u_i's

   (b) We shall use the following fact: If a matrix A satisfies $Ax = x$ for all vectors x then $A = I$. Now applying this to $A = \sum_i u_i u_i^T$ and expanding x in terms of $u_1, \ldots, u_n$ as $x = sum_i c_i u_i$, we get

$$Ax = (sum_i u_i u_i^T)(\sum_j c_j u_j) = \sum_i u_i c_i = x$$

   which holds for any x. Thus, using the above mentioned fact we have proved that $\sum_i u_i u_i^T = I$

   (c)

$$\| \sum_i c_i u_i \|^2 = (\sum_i c_i u_i, \sum_j c_j u_j) = \sum_i c_i^2$$

   (d) The projection onto a space $\mathcal{S}$ is defined as follows,

$$\Pi_{\mathcal{S}} y = \arg \min_{x \in \mathcal{S}} \|y - x\|^2$$

   In our case, $\mathcal{S} = \{\sum_{i=1}^r c_i u_i | c_i \in \mathbb{R}\}$ so that

$$\begin{aligned} \Pi_{\mathcal{S}} y &= \arg \min_{\sum_{i=1}^r c_i u_i} \| \sum_{i=1}^n (y, u_i) u_i - \sum_{i=1}^r c_i u_i \|^2 \\ &= \arg \min_{\sum_{i=1}^r c_i u_i} \| \sum_{i=1}^r \{(y, u_i) - c_i\} u_i + \sum_{i=r+1}^n (y, u_i) u_i \|^2 \\ &= \arg \min_{\sum_{i=1}^r c_i u_i} \sum_{i=1}^r \{(y, u_i) - c_i\}^2 + \sum_{i=r+1}^n (y, u_i)^2 \\ &= \sum_{i=1}^r (y, u_i) u_i \end{aligned}$$

The orthogonal complement of a space $\mathcal{S}$ is the set of vectors which are orthogonal to all vectors in $\mathcal{S}$. In our case, this can be written as

$$\mathcal{S}^{\perp} = \{x | (x, u_i) = 0 \quad \forall i = 1, \ldots, r\}$$

Since any x can be written as $\sum_{i=1}^{n}(x, u_i)u_i$ it is now obvious that

$$\mathcal{S}^{\perp} = \left\{ \sum_{i=r+1}^{n} c_i u_i | c_i \in \mathbb{R} \right\} = \text{span}\{u_{r+1}, \ldots, u_n\}$$

Applying to the previous problem, we get

$$\Pi_{\mathcal{S}^{\perp}} y = \sum_{i=r+1}^{n} (y, u_i)u_i$$

(e) This follows from applying problem (c) to the answers from problem (d)

2. (a) The design matrix X is as follows

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

The fact that not all x_i are equal means that $(x_1, \ldots, x_n) \notin \text{span}\{\mathbf{1}\}$. Thus, the two columns of X are not linearly dependent, so X is of full rank ($\text{rank}(X) = 2$). Thus all linear functions of β, in particular β^0 and β_1 are both estimable.

(b) Recall that in simple linear regression, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. The equation for the regression line is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

and the point $(\bar{x}, \bar{y})$ is on this line if it satisfies this equation, which is immediate by plugging in our expression for $\hat{\beta}_0$

(c) If $\bar{x} = 0$, we have the following estimates

$$\hat{\beta}_0 = \bar{y} = \beta_0 + \bar{\epsilon}, \qquad \hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)} = \beta_1 + \frac{x^T f}{Var(x)}$$

where $f = (\epsilon_i - \bar{\epsilon})$. To verify, that $\hat{\beta}_0$ and $\hat{\beta}_1$ are uncorrelated we simply need to show that $\bar{\epsilon}$ and f are uncorrelated. Using the fact that if $X \sim N(\mu, \Sigma)$ then $Cov(AX, BX) = A\Sigma B^T$,

$$Cov(\bar{\epsilon}, f) = Cov(\mathbf{1}^T/n\epsilon, (I - \mathbf{1}\mathbf{1}^T/n)\epsilon) = \mathbf{1}^T/n(I - \mathbf{1}\mathbf{1}^T/n) = 0$$

(d) For normal random variables, uncorrelated implies independent. Using, the result from previous problem, we are done.

3. (a) Let $x_{i\cdot}$ denote the i^{th} row of the design matrix X. We can write

$$\beta_0 + \beta_1 \bar{x}_1 + \cdots + \beta_p \bar{x}_p = (1, \bar{x}_1, \ldots, \bar{x}_p)^T \beta$$

which means that $\beta_0 + \beta_1 \bar{x}_1 + \cdots + \beta_p \bar{x}_p$ is estimable iff $(1, \bar{x}_1, \ldots, \bar{x}_p)$ is in the row span of X. But,

$$(1, \bar{x}_1, \ldots, \bar{x}_p) = \sum_i \frac{1}{n} x_{i\cdot} \in \mathcal{R}(X)$$

So, $\beta_0 + \beta_1 \bar{x}_1 + \cdots + \beta_p \bar{x}_p$ is estimable.

(b) In the same vein as Problem 2(b), we shall see that the least squares estimate of $\beta_0 + \beta_1\bar{x}_1 + \cdots + \beta_p\bar{x}_p$ is $\bar{y}$. One way to see this is

$$\begin{aligned}
&\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \cdots + \hat{\beta}_p\bar{x}_p \\
&= \frac{1}{n}\sum_i \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \\
&= \frac{1}{n}\sum_i \hat{y}_i = \bar{\hat{y}}
\end{aligned}$$

But since $y = \hat{y} + e$ and $\bar{e} = 0$, we have $\bar{\hat{y}} = \bar{y}$

(c)

$$Var(\bar{y}) = Var(\bar{\epsilon}) = \frac{\sigma^2}{n}$$

which can be estimated as $\hat{\sigma^2}/n$ where

$$\hat{\sigma}^2 = \frac{1}{n-r}\sum_i (y_i - \hat{y}_i)^2$$

4. (a) Since `WEIGHT + 3*HEIGHT` is a linear combination of `WEIGHT` and `HEIGHT`, the effect of the last feature is not distinguishable from the first two. Thus, producing `NA`s. In particular, notice that among $\beta_2, \beta_3, \beta_4$ only the linear combinations $\beta_2 + \beta_4$ and $\beta_3 + 3\beta_4$ are estimable

   (b) As discussed earlier, `0.24341` is the estimate of $\beta_2 + \beta_4$, not of β_2. β_2 is not even estimable.

   (c) Since the column space of the features did not change in the two models, and the residuals are simply the projection of the responses onto the orthogonal of the column space of features, the residuals are actually exactly the same in both models. Thus, the RSS can be read of from the previous output. It is $\hat{\sigma}^2 \times (n - rank(X)) =$ `5.142^2*247 = 6530.721`. $\hat{\sigma}$ is the residual standard error in the output. $n - rank(X)$ is the degrees of freedom in the output

   (d) Similarly, we can read off the RSS in both models as RSS(m) = `5.696^2*249 = 8078.66` and RSS(M) = `6530.721`. To calculate the F-statistic we need

$$\frac{(RSS(m) - RSS(M)/(df_M - df_m)}{RSS(M)/df_M} = \frac{773.9695}{5.142^2} = 29.27249$$

   The associated p-value is `1`, indicating very strong evidence that the deleted features were pertinent to the regression.

5. (a) Recall that $Var(\beta_0)$ is estimated as $\hat{\sigma}^2(X^TX)^{-1}_{11}$. In the output, we can see $(X^TX)^{-1}_{11}$ from the matrix and $Var(\beta_0)$ as square of the standard error of intercept term. Plugging these in

$$\text{Residual standard error } = \hat{\sigma} = 5.328712$$

   From the same manipulation on $\hat{\beta}_{weight}$, we get

$$\text{Standard error (weight) } = \sqrt{\hat{\sigma}^2 * (X^TX)^{-1}_{33}} = 0.0273406$$

   From the definition of t-value,

$$\text{t-value (weight) } = \frac{\text{estimate (weight)}}{\text{standard error (weight)}} = 4.525504$$

   Similarly,

$$\text{estimate (thigh) } = \text{t-value (thigh)} \times \text{standard error (thigh)} = 0.3654269$$

And,

$$(X^TX)^{-1}_{55} = \frac{(\text{standard error (thigh)})^2}{(\text{residual standard error})^2} = 0.0007873251$$

Finally, we can gather TSS from R^2 in output. Since $R^2 = 1 - RSS/TSS$,

$$TSS = RSS/(1 - R^2) = 7013.607/(1 - 0.5349) = 15079.78$$

getting an F-statistic

$$\text{F statistic} = \frac{(TSS - RSS)/(rank - 1)}{RSS/(n - rank)} = 67.12994$$

(b) Our estimate of the bodyfat percentage at the new design points is

$$\text{fit (confidence interval)} = x^T_{new}\hat{\beta} = \hat{\beta}_{intercept} + 30\hat{\beta}_{age} + 180\hat{\beta}_{weight} + 72\hat{\beta}_{height} + 60\hat{\beta}_{thigh} = 15.61978$$

Since a confidence interval constructed from t-distribution is symmetric around its fit value and we know the lower cutoff the upper cutoff can be calculated as

$$\text{upr (confidence interval)} = \text{fit} + (\text{fit} - \text{lwr}) = 16.6723$$

The prediction interval is centered at exactly the same point as the confidence interval. Further, from the definition notice that the prediction interval can be arrived at by inflating the confidence interval by $t_{247;0.975}\hat{\sigma}$ where $t_{247;0.975}$ is the 0.975-th quantile of t_{247}. So finally,

$$\begin{aligned}
\text{fit (prediction interval)} &= 15.61978 \\
\text{lwr (prediction interval)} &= \text{lwr (confidence interval)} - t_{247;0.975}\hat{\sigma} = 4.123176 \\
\text{upr (prediction interval)} &= \text{lwr (confidence interval)} + t_{247;0.975}\hat{\sigma} = 27.11638
\end{aligned}$$

(c) Using all the facts we learnt in the previous problems,

$$\begin{aligned}
&\text{Res. df of model 1 is } 248 \\
&\text{Res. df of model 2 is } 247 \\
&\text{Df (difference between the two residual degrees of freedom) is } 1 \\
&\text{RSS of model 2 is } 7013.607 \\
&\text{Sum of Sq (numerator of F-stat * Df) is } 46.008 \\
&\text{RSS of model 1} = \text{RSS of model 2} + \text{Sum of Sq} = 7059.58
\end{aligned}$$

6. (a) False. Think of data generated from $y = 2x$ with no error. The slope of the regression line is 2

(b) True. The slope of the regression line of y on x can be written as

$$\hat{b} = \frac{Cor(x, y)SD(y)}{SD(x)}$$

When the data are standardized, $SD(y) = SD(x) = 1$ so that $\hat{b} = Cor(x, y)$. Since correlation is always between -1 and 1 the slope can never be larger than 1, or smaller than -1 for that matter

(c) True. We have seen that residual standard error of a bigger model is always lesser or equal to that of a smaller model.

(d) False. We have seen that $Var(\hat{e}_i) = h_{ii}$, the i^{th} diagonal entry of the hat matrix. There is no reason why $h_{11} = \cdots = h_{nn}$

(e) True. $c^T\beta$ is estimable iff $c \in \mathcal{R}(X)$. If X^TX is invertible then X is of full column rank so $\mathcal{R}(X) = \mathbb{R}^p$

(f) False. It is always true that $y = \hat{y} + \hat{e} = Hy + (I - H)y$. Since $H(I - H) = 0$, the fitted values and the residuals are always orthogonal

(g) False. From the above decomposition, $Cov(\hat{y}, \hat{e}) = Cov(Hy, (I - H)y) = \sigma^2 HI(I - H) = 0$

(h) True. If the normality assumption is violated, uncorrelated does not imply independent. Exercise: Think of an example of two dependent random variables which are uncorrelated

(i) False. This only validates that y is dependent on X. It is entirely possible that a carefully chosen non-linear model will be better at explaining the dependency

(j) False. β_2 is an unknown quantity. One interpretation of p-value is the chance of incorrectly rejecting a true hypothesis. Thus in this example there is 0.005 chance of β_2 being small

(k) True. The p-value is based on the value of $\hat{\beta}_2$. Since the p-value is small, $|\hat{\beta}_2|$ must be big