1 Notes on R^2

The book very briefly mentions R^2 as a goodness of fit metric (in chapters 9 and 10). There it is called the "proportional decrease in model error" and defined as

$$R^2 = \frac{MSE_0 - MSE}{MSE_0} = 1 - \frac{MSE}{MSE_0}$$

where $MSE_0 = Ave_x(\overline{y} - \mu(x))^2$ and $MSE = Ave_x(\widehat{f}(x) - \mu(x))^2$ where "$\mu(x)$ is the true mean of Y". I guess this is sufficiently similar to the more familiar definition on Wikipedia:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where $SS_{res} = \sum_i (y_i - f_i)^2$ and $SS_{tot} = \sum_i (y_i - \overline{y})^2$. (It seems like the translation should be $\overline{y}$ agrees in both cases and is the mean of the data, $\widehat{f}(x)$ translates to $f_i = f(x_i)$, and $\mu(x)$ translates to $y_i = \mu(x_i)$ is the true value. x_i and x range over the data in both cases; in the latter it is explicit in the sum and in the former is it implicit in the average.)

Assume that $\overline{y} = 0$ (potentially by centering the outputs if necessary). Then the R^2 coefficient simplifies to

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i y_i^2} = 1 - \frac{< y - X\beta, y - X\beta >}{< y, y >}.$$

In terms of Euclidean distances, $< y, y >$ is the length of the output vector while $< y - X\beta, y - X\beta >$ is the length of the residual vector. By least squares, the residual vector $y - X\beta$ is orthogonal to any linear combination of the columns of X and in particular is orthogonal to the vector $X\beta$. (If not, then we could make the residuals even smaller, contradicting that β was chosen to be a least squares solution.) Thus, the triangle formed by $y, y - X\beta, X\beta$ is a right triangle so the side lengths satisfy $< y, y >=< X\beta, X\beta > + < y - X\beta, y - X\beta >$. Thus we can rewrite

$$R^2 = \frac{< X\beta, X\beta >}{< y, y >} = \cos(\theta)^2,$$

where θ is the angle between $X\beta$ and y. Thus when β is chosen according to a least squares fit, we can interpret R^2 as the square of the cosine of the angle between the outputs y and the predictions $X\beta$. The closer to 1, the better the fit (and conversely the closer to 0, the worse the fit). Note that we should not have negative R^2 coefficients (when using least squares) since we could just take $-\beta$ instead.

We can ask: given two least squares fits $y \sim X_1\beta_1$ and $y \sim X_2\beta_2$ with R^2 coefficients R_1, R_2, what is the range of possible R^2 coefficients of the least squares fit of y on the combined features (X_1, X_2)? In general, if the column space of X_1 is contained in the column space of X_2 (or vice versa), then the resulting R^2 will simply be $\max(R_1, R_2)$ – the fit cannot get worse by adding more data; at worst the additional features can be entirely redundant. Note in the special case that X_1 and X_2 have the same number of columns and the columns of X_1 are linearly independent, then the only way for the column span of X_2 to contain that of X_1 is if they have the same column spans, in which case we must also have $R_1 = R_2$. However, even if this is not the case, it is possible that the orthogonal (to the column space of X_2) part of the column space of X_1 is orthogonal to y (and thus useless for predicting y), so this does not change the conclusion that the minimum possible R^2 of the combined data is $\max(R_1, R_2)$.

On the other hand, if the column spaces of X_1 and X_2 are orthogonal, then by definition of least squares, the resulting fit will be $y \sim X_1\beta_1 + X_2\beta_2$ and the R^2 coefficient will be $R_1 + R_2$ by direct computation. Thus, the range for the resulting R^2 is $[\max(R_1, R_2), \min(1, R_1 + R_2)]$ since the fit cannot be better than perfect.

2 Exercises

1. The F statistic is $(RSS_0 - RSS_1)/(RSS_1/(N-p-1))$ since the two models differ only by a single variable (RSS_0 refers to the RSS for the smaller model while RSS_1 refers to the RSS for the larger model). By definition, $RSS_1/(N-p-1) = \widehat{\sigma}^2$ so it remains to show that $RSS_0 - RSS_1 = \widehat{\beta}_j^2/v_j$. For each $i = 0, 1$, the RSS can be rewritten as $RSS_i = (Y - X_i(X_i^TX_i)^{-1}X_i^TY)^T(Y - X_i(X_i^TX_i)^{-1}X_i^TY) = Y^T(I - X_i(X_i^TX_i)^{-1}X_i^T)^2Y = Y^T(I - X_i(X_i^TX_i)^{-1}X_i^T)^2Y$ which can be seen directly via computation or geometrically using the fact that $I - X_i(X_i^TX_i)^{-1}X_i^T$ is a projection matrix, where X_1 is the full data matrix and X_0 is the data matrix with feature j removed.

   Then $RSS_0 - RSS_1 = Y^T(X_1(X_1^TX_1)^{-1}X_1^T - X_0(X_0^TX_0)^{-1}X_0^T)Y$. Let $P_i = X_i(X_i^TX_i)^{-1}X_i^T$, which is the projection to the column space of X_i. Since the column space of X_0 is contained within the column space of X_1, the difference $P_1 - P_0$ is another projection operator. Thus we may interpret the difference as $RSS_0 - RSS_1 = ||(P_1 - P_0)Y||^2 = ||Y_1 - Y_0||^2$, where Y_i refers to the projection of Y to the column space of X_i. By Gram-Schmidt, we can express $Y_1 - Y_0 = \widehat{\beta}_j z_j$ where z_j is the residual of the regression of x_j on the remaining features $x_{i \neq j}$, so $||Y_1 - Y_0||^2 = \widehat{\beta}_j^2||z_j||^2$.

   The residuals of the linear regression of each column of X on the remaining columns can be realized by $X(X^TX)^{-1}$, since the columns of $X(X^TX)^{-1}$ are linear combinations of the columns of X, so they have the same column space (assuming $(X^TX)^{-1}$ is full rank, i.e. that X is full column rank) and since $X^T(X(X^TX)^{-1}) = I$. In particular, z_j can be realized as the j^{th} column of $X(X^TX)^{-1}$, so $z_j^Tz_j$ is the (j, j) entry of $((X^TX)^{-1}X^T)(X(X^TX)^{-1}) = (X^TX)^{-1}$, which is exactly v_j. Thus the F score for dropping a single feature is $\widehat{\beta}_j^2/(\widehat{\sigma^2}v_j) = z_j^2$ which is exactly the square of the corresponding z-score.

2. For this problem, we will assume that the underlying model is actually cubic, so that $y = x^T\beta + \varepsilon$ for each observation, where $x = (1, x, x^2, x^3)$ is the vector of powers of x up to degree 3 and $\varepsilon \sim N(0, \sigma^2)$ and the errors between instances are uncorrelated (hence independent by the normality assumption). Under these assumptions, $\widehat{\beta} = (X^TX)^{-1}X^TY = (X^TX)^{-1}X^T(X\beta + E) = \beta + (X^TX)^{-1}X^TE \sim N(\beta, (X^TX)^{-1}\sigma^2)$.

   To get pointwise variance estimates, we conclude from above that for each $a = (1, x_0, x_0^2, x_0^3)$, $a^T\widehat{\beta} \sim N(a^T\beta, a^T(X^TX)^{-1}a\sigma^2)$. Thus we can get an approximate 95% confidence interval for $a^T\beta$ using $(a^T\widehat{\beta} - 1.96\sqrt{a^T(X^TX)^{-1}a\sigma^2}, a^T\widehat{\beta} + 1.96\sqrt{a^T(X^TX)^{-1}a\sigma^2})$.

To get confidence set variance estimates, we conclude from above that $(\widehat{\beta}-\beta)^T(X^TX/\sigma^2)(\widehat{\beta}-\beta) \sim \chi_p^2$ (using the fact that for $A \sim N(B,C), S^{-1}(A-B) \sim N(0,I)$ where $C = SS^T$ is the Cholesky decomposition of a symmetric positive definite matrix C or any other square root decomposition of such a matrix and for $Z \sim N(0,I)$, $Z^TZ \sim \chi_p^2$ since it is a sum of squares of iid standard normal random variables). Thus an approximate 95% confidence set for β can be obtained as $C = \{\beta \mid (\widehat{\beta}-\beta)^T(X^TX/\sigma^2)(\widehat{\beta}-\beta) \leq (\chi_{p+1}^2)^{.05}\}$. This provides a 95% confidence interval for $a^T\beta$ via a^TC_β. To give an explicit confidence interval for $a^T\beta$, noting that $a^T\beta$ is linear in β and C_β is convex (it is an ellipsoid), we can use tools from convex optimization to compute the interval $(\min a^T\beta, \max a^T\beta)$ where the min and max are both computed over C_β. To demonstrate one of these computations (the other is entirely analogous), we formulate it as:

$$\max a^T\beta s.t.(\widehat{\beta}-\beta)^T(X^TX/9.487729\sigma^2)(\widehat{\beta}-\beta) \leq 1.$$

We solve this by using the Lagrangian form:

$$\max a^T\beta + \lambda(1-(\widehat{\beta}-\beta)^T(X^TX/9.487729\sigma^2)(\widehat{\beta}-\beta))s.t.\lambda \geq 0$$

and taking a derivative with respect to β and equating with 0:

$$a + 2\lambda((X^TX/9.487729\sigma^2)(\widehat{\beta}-\beta)) = 0 \implies \beta = \widehat{\beta} + (X^TX)^{-1}9.487729\sigma^2 a/(2\lambda).$$

By convexity, the optimal solution occurs on the boundary of the ellipsoid (i.e. obtaining equality in the constraint of being contained in the ellipsoid), so to solve for λ above, we plug back in:

$$((X^TX)^{-1}9.487729\sigma^2a/(2\lambda))^T(X^TX/9.487729\sigma^2)((X^TX)^{-1}9.487729\sigma^2a/(2\lambda)) = 1$$

$$\sqrt{9.487729\sigma^2/4 * a^T(X^TX)^{-1}a} = \lambda$$

finally yielding the maximum/minimum $a^T\beta$

$$a^T\beta = a^T\widehat{\beta} \pm \sqrt{9.487729\sigma^2a^T(X^TX)^{-1}a} \approx a^T\widehat{\beta} \pm 3.080216\sqrt{a^T(X^TX)^{-1}a\sigma^2}.$$

This is the same as if we had used ~ 3 standard deviations to produce the previous confidence interval, so the confidence set approach yields a wider band. (I guess this is wider due to the dimension: the ratio of the lengths of the intervals (confidence set / point estimate) scales (linearly) with the dimension due to computing the confidence set in p-dimensional space requiring the chi square distribution with p degrees of freedom.)

3. (a) Let c^Ty be another linear unbiased estimate of $a^T\beta$. Then we want to show that $\sigma^2c^Tc = Var(c^Ty) \geq Var(a^T\widehat{\beta}) = \sigma^2a^T(X^TX)^{-1}a$ or equivalently that $c^Tc \geq a^T(X^TX)^{-1}a$. The unbiased condition means that $c^TX\beta = E(c^Ty) = a^T\beta$. Since this must hold for any possible β, it must be the case that $c^TX = a^T$. Thus, $a^T(X^TX)^{-1}a = c^TX(X^TX)^{-1}X^Tc$, so it remains to show that $c^Tc \geq c^TX(X^TX)^{-1}X^Tc \iff c^T(I-X(X^TX)^{-1}X^T)c \geq 0$. The matrix $I-X(X^TX)^{-1}X^T$ is a projection matrix, so $c^T(I-X(X^TX)^{-1}X^T)c = ||(I-X(X^TX)^{-1}X^T)c||^2 \geq 0$ as desired.

(b) Let Cy be another unbiased linear estimate of β. Then we want to show that $Var(Cy) = CC^T\sigma^2 \geq Var(\widehat{\beta}) = (X^TX)^{-1}\sigma^2 \iff CC^T \geq (X^TX)^{-1}$. The unbiased assumption means that $CX\beta = E(Cy) = \beta$ for any possible beta, so that $CX = I$. (Note that this does not imply that C or X is invertible. For example, we could have $C = (X^TX)^{-1}X^T$ or many other choices differing by vectors orthogonal to the column space of X.) Thus, we need to establish that $CC^T \geq (X^TX)^{-1} = CX(X^TX)^{-1}X^TC^T$. In other words, we need to show that $CC^T - CX(X^TX)^{-1}X^TC^T = C(I - X(X^TX)^{-1}X^T)C^T$ is positive semidefinite. Analogously to before, the matrix $I - X(X^TX)^{-1}X^T$ is a projection matrix, so for any vector v, $v^TC(I - X(X^TX)^{-1}X^T)C^Tv = ||(I - X(X^TX)^{-1}X^T)C^Tv||^2 \geq 0$.

4. From one pass of the Gram-Schmidt procedure, we obtain a QR decomposition $X = QR$ where the columns of Q are the Gram-Schmidt orthonormalizations of the columns of X and R is an upper triangular matrix writing the columns of X as linear combinations of the columns of Q. Then $\widehat{\beta} = (X^TX)^{-1}X^TY = (R^TQ^TQR)^{-1}R^TQ^TY = (R^TR)^{-1}R^TQ^TY = R^{-1}Q^TY \implies R\widehat{\beta} = Q^TY$ which is straightforward to solve since R is upper triangular.

   At best during the Gram Schmidt procedure we can determine the coefficients $\tilde{\beta}_j =< z_j, y > / < z_j, z_j >$ where z_j is the residual of the linear regression of x_j on $x_1, \ldots, x_{j-1}$. Only the final coefficient $\tilde{\beta}_p$ can be used directly in $\widehat{\beta}_p$. The remaining ones must be solved for by back-substitution using the R matrix, so I don't see a way to extract the $\widehat{\beta}$ coefficients without at least the work of solving $R\widehat{\beta} = Q^TY$.

5. We can rewrite the expression being minimized as $\sum_i(y_i - (\beta_0^c - \sum_j \overline{x}_j\beta_j^c) - \sum_j x_{ij}\beta_j^c)^2 + \lambda\sum_j(\beta_j^c)^2$. Making the substitution $\beta_0 = (\beta_0^c - \sum_j \overline{x}_j\beta_j^c)$ and $\beta_i = \beta_i^c$ for $i > 0$, we obtain the expression $\sum_i(y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 + \lambda\sum_j \beta_j^2$ which is equivalent to the ridge regression problem. Thus the two problems have equivalent solutions with the transformation as indicated above (its inverse is $\beta_0^c = \beta_0 + \sum_j \overline{x}_j\beta_j$ and $\beta_i^c = \beta_i$ for $i > 0$). Importantly, note that β_0 and β_0^c do not appear in the regularization term so the equivalence results from the fact that we only transform β_0 and β_0^c. We can view this new regression problem as arising from the centered version of X, so the solution is the same as the usual ridge regression solution $(X_c^TX_c + \lambda I)^{-1}X_c^TY$.

   The exact same transformation as above yields the analogous result in LASSO regression. More generally, lasso and ridge regression are insensitive to translations in the input data features (the columns of X) and achieve the same minimization of the RSS+regularization function (with the result of changing the intercept term β_0).

6. Assuming a Gaussian prior $\beta \sim N(0, \tau I)$ and a conditional Gaussian sampling model $y|\beta \sim N(X\beta, \sigma^2 I)$, we would like to update our prior for β given an observation of y. Using Bayes theorem, we need to determine $f(\beta|y) = f(y|\beta)f(\beta)/f(y)$. Since y is assumed fixed here, we can treat $f(y)$ as a normalizing constant and determine the posterior distribution of $\beta|y$ using the format of $f(y|\beta)$ and $f(\beta)$.

$$f(\beta|y) = f(y|\beta)f(\beta)/f(y) \propto f(y|\beta)f(\beta) \propto \exp(-1/2(y - X\beta)^T\sigma^{-2}(y - X\beta))\exp(-1/2\beta^T\tau^{-1}\beta)$$

$$= \exp(-1/2\sigma^{-2}((y - X\beta)^T(y - X\beta) + \sigma^2\tau^{-1}\beta^T\beta)).$$

Without simplifying further (to identify the mean and variance explicitly), it is clear from the format (exponential of negative quadratic form) exactly agrees with a Gaussian distribution, so the posterior distribution of $\beta|y$ is Gaussian. The log likelihood of the pdf of the posterior distribution is equal to (where C is the constant of proportionality discarded above)

$$C - \sigma^{-2}/2((y - X\beta)^T(y - X\beta) + \sigma^2\tau^{-1}\beta^T\beta)$$

and maximizing this expression is equivalent to minimizing the right hand term

$$((y - X\beta)^T(y - X\beta) + \sigma^2\tau^{-1}\beta^T\beta)$$

which is exactly the ridge regression problem. Thus the regularization parameter λ should be taken to be $\sigma^2\tau^{-1}$.

7. This is exactly equivalent to the previous problem. Note however that the negative posterior log-likelihood of β minus a constant is propotional to the desired expression, rather than the negative posterior log likelihood itself being proportional to the desired expression.

8. For any vector v of length N, the projection onto the all 1s vector is given by $< v, 1 > / < 1, 1 > 1 = (\sum_i v_i)/N1 = \overline{v}1$, so the centered version of X can be obtained by replacing each column of X by the residual after linear regression onto the all 1s vector: $x - \overline{x}1$. Thus, the column space of $\tilde{X}$ is the span of all the centered columns of X. By definition of SVD, this is furthermore equal to the column space of U.

   Now, let $v = \sum_i c_i(x_i - \overline{x_i}1)$ be in the column space of U. Note that $< v, 1 >= \sum_i c_i < x_i, 1 > -\overline{x_i} < 1, 1 >= 0$ so v is in the column span of X and is orthogonal to the all 1s vector. Since the columns of Q form an orthonormal basis for the column space of X and the first column of Q is (a rescaled) all 1s vector, this means that v can be written as a linear combination of the columns of Q_2. Next let v be in the column space of Q_2. Then $v = \sum_{i>1} c_i z_i$. Since $z_i = x_i - \sum_j < x_i, z_j > z_j = (x_i - < x_i, z_1 > z_1) - \sum_{j>1} < x_i, z_j > z_j$, we can rewrite v in terms of the vectors $(x_i - < x_i, z_1 > z_1)$ which are exactly the centered columns of X. Thus v is contained in the column space of $\tilde{X}$ so Q_2 and U have the same column spaces.

   The SVD and QR decompositions are not unique and depend on the choice of algorithm used to compute it. Thus in general unless the algorithms are designed to have some compatibility, they will generally never agree up to sign flips. In the QR decomposition of X, we view Q as the orthonormalization of the columns of X according to the Gram-Schmidt process. In the SVD decomposition of X, we view V as the eigenvectors of X^TX and U as the eigenvectors of XX^T. If the eigenspaces of XX^T have multiplicity, then the choice of eigenvector decomposition will be horribly non-unique and there will be essentially no chance that U and Q_2 will coincide (unless compatibility is enforced). Otherwise, the only way to force agreement up to sign is if the columns of X come from the action of a triangular matrix on the columns of U.

9. The variable which will most reduce the RSS is the one whose correlation with r is highest (while also minimizing the correlation with the variables already in X_1). To

compute this, we would replace each column of X_2 with its (normalized) residual after linear regression onto the columns of X_1. Then the variable which will most reduce the RSS is the one whose correlation $< v, r >$ with the current residual r is largest in magnitude. To see why this is the case, note that adding this vector will exactly reduce the residual by $< v, r > / < v, v > v =< v, r > (v/ < v, v >)$ by construction. To describe the full algorithm, we give the full update step here:

(a) Assume that the columns of X_2 are already orthogonal to the columns of X_1 and furthermore that they are already normalized as well.

(b) Determine the largest entry of $(r^T X_2)$ and add the corresponding (unprocessed) predictor from X_2 to X_1, using Gram-Schmidt to extend the QR decomposition to include this newly added variable by extending R (note that we do not need to do any work in extending Q since the columns of X_2 are already orthonormal with respect to the columns of X_1 and hence of Q).

(c) Since the columns of X_2 were already orthogonal to the columns of X_1, simply subtract the projection of the remaining columns onto the newly added features and renormalize to set up X_2 for the next step.

10. This can be directly computed by the F statistic, where we vary the "smaller model" over all possible models obtained by removing a single variable. Since the denominators will all be the same, each F statistic will be directly proportional (by the same constant) to the increase in RSS caused by dropping the corresponding variable, so we should select the variable with smallest F statistic. By exercise 1, this is equivalent to selecting the variable with the smallest squared z-score. Note that this is equivalent to selecting the variable whose z-score is closest to 0 (i.e. having the smallest absolute z-score), since the z-score could be negative in principle.

11. The multivariate RSS can be written as $Tr((Y - f(X))\Sigma^{-1}(Y - f(X))^T)$. (One can view this as coming from independent observations $y_i \sim N(f(x_i), \Sigma)$, whose joint distribution is thus proportional to $\exp(-1/2 \sum_i (y_i - f(x_i))^T \Sigma^{-1}(y - f(x_i)))$ and thus whose log likelihood is maximized when $\sum_i (y_i - f(x_i))^T \Sigma^{-1}(y - f(x_i))$ is minimized.)

    Choosing $f(X) = XB$ for linear regression, we seek to optimize $Tr((Y - XB)\Sigma^{-1}(Y - XB)^T)$ with respect to B. To do this, take the derivative with respect to B and set equal to 0:

$$\frac{d}{dB} Tr((Y - XB)\Sigma^{-1}(Y - XB)^T) = \frac{d}{dB} Tr(Y\Sigma^{-1}Y^T - Y\Sigma^{-1}B^T X^T - XB\Sigma^{-1}Y^T + XB\Sigma^{-1}B^T X^T)$$

    Using the matrix cookbook for trace derivative formulas:

$$= -2X^T Y\Sigma^{-1} + 2X^T XB\Sigma^{-1}$$

    Equating with zero yields

$$-2X^T Y\Sigma^{-1} + 2X^T XB\Sigma^{-1} = 0 \implies X^T Y = X^T XB \implies B = (X^T X)^{-1} X^T Y.$$

    If Σ depends on the particular training observation (i.e. the variance of y_i is Σ_i), then maximizing the log-likelihood still corresponds to minimizing a quadratic form in B.

However, we no longer have an obvious trace formula and it is unclear whether a closed form solution can be derived since we can no longer collect all the training data into one matrix due to the difference covariance matrices.

12. Consider the ordinary least squares problem with the modified dataset (X, Y): min $\sum_{i=1}^{N+p}(y_i - x_i^T\beta)^2$. Splitting the sum into the first N and final p terms, we find $\sum_{i=1}^{N}(y_i - x_i^T\beta)^2 + \sum_{i=1}^{p}(0 - \sqrt{\lambda}\beta_i)^2 = \sum_{i=1}^{N}(y_i - x_i^T\beta)^2 + \sum_{i=1}^{p}\lambda\beta_i^2$ which is exactly the ridge regression minimization problem.

13. $\widehat{y}_{(M)}^{pcr} = \overline{y}1 + \sum_{m=1}^{M}\widehat{\theta}_m z_m = \overline{y}1 + X(\sum_{m=1}^{M}\widehat{\theta}_m v_m)$ so $\widehat{\beta}^{pcr}(M) = \sum_{m=1}^{M}\widehat{\theta}_m v_m$. When $M = p$, we can write $\widehat{\theta}$ in terms of X, its SVD, and Y: $\sum_{m=1}^{p}\widehat{\theta}_m v_m = V\widehat{\theta}$ where $\widehat{\theta}$ is the column vector with entries $\widehat{\theta}_i$. Analyzing $\widehat{\theta}$ further:

$$\widehat{\theta} = (< z_i, y > / < z_i, z_i >)_i = (v_i^T X^T y / v_i^T X^T X v_i)_i = (v_i^T / v_i^T X^T X v_i)_i X^T y.$$

where each vector $v_i^T / v_i^T X^T X v_i$ is a row vector. Writing $X = UDV^T$, we have $X^T X = VD^2V^T$ and thus $v_i^T V D^2 V^T v_i = e_i^T D^2 e_i = (D^2)_{ii}$. Thus $\widehat{\theta} = D^{-2}V^T X^T y$ so $V\widehat{\theta} = VD^{-2}V^T X^T y = (X^T X)^{-1}X^T y = \widehat{\beta}^{ls}$.

14. According to step 2(a), $z_1 = \sum_j \widehat{\varphi}_{1j} x_j$, where $\widehat{\varphi_{1j}} =< x_j, y >$. Then according to step 2(d), $x_i^{(1)} = x_i - < z_1, x_i > / < z_1, z_1 > z_1$. Since the x_i are orthonormal, $< z_1, x_i > =< \sum_j \widehat{\varphi}_{1j} x_j, x_i > = \widehat{\varphi}_{1i}$ and $< z_1, z_1 > = \sum_j \varphi_{1j}^2$, so $x_i^{(1)} = x_i - \widehat{\varphi}_{1i} / \sum_j \varphi_{1j}^2 z_1$.

    Then on the next iteration, $\widehat{\varphi}_{2i} =< x_i^{(1)}, y > =< x_i - \widehat{\varphi}_{1i} / \sum_j \varphi_{1j}^2 z_1, y > =< x_i, y > - \widehat{\varphi}_{1i} / \sum_j \widehat{\varphi}_{1j}^2 < z_1, y > = \widehat{\varphi}_{1i} - \widehat{\varphi}_{1i} / \sum_j \widehat{\varphi}_{1j}^2 < \sum_j \widehat{\varphi}_{1j} x_j, y > = 0$. Thus $z_2 = 0$ and this will continue forward in every future step, so PLS terminates after one step in the orthogonal case.

15. $Corr^2(y, X\alpha)Var(X\alpha) = (Cov^2(y, X\alpha)/Var(y)Var(X\alpha))(Var(X\alpha)) = Cov^2(y, X\alpha)/Var(y)$. Since $Var(y)$ is constant in the problem, the maximization problem is equivalent to maximizing $Cov(y, X\alpha) = y^T X\alpha$ over α. Thus the most correlated choice is $\alpha = X^T y$. The condition that $\alpha^T S\widehat{\varphi}_l = 0$ enforces orthogonality with previously chosen PLS directions, but also means that we cannot choose $\alpha = X^T y$ each time for successive PLS directions. Instead, the best choice will be the (unit vector) direction most correlated with $X^T y$, which is exactly $(X^{(j)})^T y$ where $X^{(j)}$ is the orthogonalization of X with respect to the previously chosen PLS directions $z_1, \ldots, z_{j-1}$. To see this, note that $X^T y$ has an orthogonal decomposition $\sum_{i=1}^{j-1} c_i\widehat{\varphi}_i + (X^{(j)})^T y$. The condition that $\alpha S\widehat{\varphi}_l = 0$ forces that α must be orthogonal to $\widehat{phi}_1, \ldots, \widehat{\varphi}_{j-1}$ so the best choice for $alpha$ is $(X^{(j)})^T y$.

16. Note that this table is only for the orthogonal case, where $X^T X = I_p$. In this case, note that $(X^T X)^{-1} = I$ as well, so $\sqrt{v_j} = 1$ in the z-score. Thus, the regression coefficients $\widehat{\beta}_j$ are all directly proportional via the same fixed constant $1/\widehat{\sigma}$ to the corresponding z-scores. Thus since the regression coefficients are independent in the orthogonal case, the best subset of size M consists of the M features with largest absolute regression coefficients, which agrees with the given formula.

For ridge regression, the closed form solution is $(X^TX + \lambda I)^{-1}X^TY$. When X is orthogonal, $X^TX = I$ so this reduces to $((1+\lambda)I)^{-1}X^TY$. In this case, the least squares solution is $\widehat{\beta} = X^TY$, so the ridge regression coefficients are $\widehat{\beta}/(1+\lambda)$.

For lasso regression, the analysis is more careful. Let $y = X\widehat{\beta} + r$ where $\widehat{\beta}$ is the least squares regression of y on X and thus r is orthogonal to the columns of X. Then the lasso problem seeks to minimize $||y - X\beta||_2^2 + \lambda||\beta||_1 = ||X\widehat{\beta} + r - X\beta||_2^2 + \lambda||\beta||_1 = ||\widehat{\beta} - \beta||_2^2 + ||r||_2^2 + \lambda||\beta|||_1$. Since $||r||$ does not depend on the minimization parameter β, the minimization problem can be equivalently written as $\min_\beta ||\widehat{\beta} - \beta||_2^2 + \lambda||\beta||_1 = \sum_i(\widehat{\beta}_i - \beta_i)^2 + \lambda|\beta_i|$. When $\beta_j \neq 0$, then $d/d\beta_j|\beta_j| = sign(\beta_j)$ so $d/d\beta_i = -2(\widehat{\beta}_i - \beta_i) + sign(\beta_i)\lambda$ and setting this equal to zero yields $\beta_i = \widehat{\beta}_i - sign(\beta_i)\lambda/2$. Note that we need to use fancier methods for when $\beta_i = 0$ (i.e. complementary slackness KKT conditions – in these cases, we obtain $-\widehat{\beta}_i + s\lambda/2 = 0$ for a slack variable $s \in [-1, 1]$ and this is achievable when $-\lambda/2 \leq \widehat{\beta}_i \leq \lambda/2$).

17. (coding problem)

18. (conjugate gradient algorithm, partial sol) Conjugate gradient algorithms perform a procedure analogous to Gram-Schmidt, where instead of using the usual inner product, we use one determined by a positive definite matrix $< u, v >= u^TAv$ to iteratively compute an approximate solution to a linear system $Ax = b$. This is analogous to the way in which we attempt to find a solution to $y = X\beta$. However, since a solution may not exist (and more problematic for conjugate gradient X will generally fail to be positive definite or even square or symmetric) we can instead search for a solution to $X^Ty = X^TX\beta$. The usual least squares approach would yield $\widehat{beta} = (X^TX)^{-1}X^Ty$ while the conjugate gradient method with $m \leq p$ iterations would yield $\widehat{\beta}^{pls}$.

19. $\widehat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^TY$ so $||\widehat{\beta}^{ridge}|| = Y^TX(X^TX + \lambda I)^{-1}(X^TX + \lambda I)^{-1}X^TY$. Using an SVD decomposition $X = UDV^T$, this simplifies to $Y^TUDV^T(VD^2V^T + \lambda I)^{-1}VV^T(VD^2V^T + \lambda I)^{-1}VDU^TY = Y^TUD^2(D^2 + \lambda I)^{-2}U^TY = ||D(D^2 + \lambda I)^{-1}U^TY||_2^2$. Writing $v = U^TY$, we get $||D(D^2 + \lambda I)^{-1}v||_2^2 = \sum_i d_{ii}/(d_{ii}^2 + \lambda)v_i^2$. As λ decreases to 0, the term $d_{ii}/(d_{ii}^2 + \lambda)$ increases to $1/d_{ii}$ and hence the 2-norm of the ridge regression estimate increases.

    A more conceptual way to argue this is geometrically. When solving the minimization problem $\min_\beta RSS(\beta)$ subject to $||\beta||_i \leq t$ for $i = 1, 2$, we seek to find the point where a level set of $RSS(\beta)$ intersects the penalization constraint region $||\beta||_i \leq t$ in a single point. As $\lambda \to 0$, t increases (in practice t will stop growing at some finite value when the constraint region includes the least squares estimate), and in particular the first point where the now larger constraint region intersects a level set of $RSS(\beta)$ in a single point is further from the origin by definition since it is on the boundary of a strict superset of the previous constraint region. Thus $||\beta||_i$ will increase (note that we must use the appropriate norm, i.e. 2 norm for ridge and 1 norm for lasso, otherwise it is possible e.g. for the lasso that the 2-norm of β will decrease even if the 1-norm increases).

[unsure, but seems plausible] It seems possible in principle that the norm of the pls regression coefficient will decrease even if the number of steps is increased. Note that $\widehat{\beta}^{pls}(m) = \sum_i \widehat{\theta}_i \widehat{\varphi}_i$ in the PLS algorithm. For example, consider the case when the columns of X are highly correlated with each other. Then the first step of PLS, which fits each column using its univariate regression coefficient will drastically overshoot y, resulting in a regression estimate larger than the least squares estimate, which would have been achieved after p steps, so the size must decrease somewhere.

20. (canonical correlation analysis)

21. (reduced-rank regression and canonical vectors)

22. (generalizing 21)

23. (a)

$$1/N| < x_j, y - u(\alpha) > | = 1/N| < x_j, y > -\alpha < x_j, X\widehat{\beta} > |$$
$$= 1/N| < x_j, y > -\alpha < x_j, X(X^TX)^{-1}X^Ty > |.$$

In the second term on the right, note that

$$< x_j, X(X^TX)^{-1}X^Ty >= x_j^TX(X^TX)^{-1}X^Ty = (X^TX)_j(X^TX)^{-1}X^Ty = e_j^TX^Ty = x_j^Ty.$$

Thus we obtain

$$= 1/N| < x_j, y > -\alpha < x_j, y > | = (1-\alpha)1/N| < x_j, y > | = (1-\alpha)\lambda.$$

(b) The absolute correlation is

$$1/N| < x_j, y - u(\alpha) > |/\sqrt{1/N < y - u(\alpha), y - u(\alpha) >}$$
$$= (1-\alpha)\lambda/\sqrt{1/N < y - u(\alpha), y - u(\alpha) >}$$

since each variable in X has variance 1, so we need to figure out the inner product in the denominator.

$$< y - \alpha X\beta, y - \alpha X\beta >=< \alpha y + (1-\alpha)y - \alpha X\beta, \alpha y + (1-\alpha)y - \alpha X\beta >$$
$$= (1-\alpha)^2N + 2\alpha(1-\alpha) < y, y - X\beta > +\alpha^2 < y - X\beta, y - X\beta >$$
$$= (1-\alpha)^2N + 2\alpha(1-\alpha) < y - X\beta + X\beta, y - X\beta > +\alpha^2RSS.$$

Since $\widehat{\beta}$ is the least squares solution, the residual $y - X\beta$ is orthogonal to all columns of X and in particular to any linear combination of the columns of X i.e. to $X\beta$. Thus

$$= (1-\alpha)^2N + 2\alpha(1-\alpha)RSS + \alpha^2RSS = (1-\alpha)^2N + \alpha(2-\alpha)RSS.$$

Plugging this back in yields

$$= (1-\alpha)\lambda/\sqrt{(1-\alpha)^2 + \alpha(2-\alpha)/NRSS}.$$

This, as $\alpha \to 1$, the numerator converges to 0 while the denominator converges to $\sqrt{RSS/N}$ which is fixed and nonzero (assuming y is not a linear combinations of the columns of X) so $\lambda(\alpha) \to 0$.

(c) Assuming that the absolute correlations start off tied, parts (a) and (b) prove that the absolute correlations remain tied and decrease monotonically as we progress towards the least squares solution. Thus, since the LAR algorithm guarantees that the variables in the active set have tied absolute correlations, parts (a) and (b) prove that LAR keeps the correlations tied and monotonically decreasing.

24. In order to show this, it suffices to show that the LAR direction has equal inner product (up to sign) with each predictor in A_k. For this, we compute

$$< x_j, X_{A_k}(X_{A_k}^T X_{A_k})^{-1} X_{A_k}^T r_k >= x_j^T X_{A_k}(X_{A_k}^T X_{A_k})^{-1} X_{A_k}^T r_k$$

Note that x_j^T is the j^{th} column/row in $X_{A_k}/X_{A_k}^T$ so $x_j^T X_{A_k}$ is the j^{th} row of $X_{A_k}^T X_{A_k}$ so $(x_j^T X_{A_k})(X_{A_k}^T X_{A_k})^{-1} = e_j$ the standard basis vector with a 1 in position j. Thus

$$= e_j^T X_{A_k}^T r_k = x_j^T r_k =< x_j, r_k >$$

and by definition of the LAR algorithm, all the $< x_j, r_k >$ are equal in magnitude and hence all predictors make an equal angle with the LAR direction.

25. At the beginning of the k^{th} step, there are k active variables having tied absolute correlations $1/N| < x_j, y > | = \lambda$ and the remaining $p - k$ variables have absolute correlations less than λ so $1/N| < x_i, y > | < \lambda$. As computed in problem 23, the absolute correlations

$$(1-\alpha)\lambda/\sqrt{(1-\alpha)^2 + \alpha(2-\alpha)RSS/N}$$

of the active variables with the current residual remain tied and decrease monotonically to 0 as we move towards the least squares solution (i.e. as α goes from 0 to 1).

For an inactive variable x, the absolute correlation with the residual (note that X refers only to the active variables)

$$1/N| < x, y - \alpha X\beta > |/\sqrt{(1-\alpha)^2 + \alpha(2-\alpha)/NRSS}$$

$$= 1/N| < x, y > -\alpha < x, X\beta > |/\sqrt{(1-\alpha)^2 + \alpha(2-\alpha)/NRSS}$$

begins smaller than λ when $\alpha = 0$ since in that case it's equal to $1/N| < x, y > | < \lambda$ by assumption. Since both correlations have the same denominator, we will ignore it and compare the numerators directly.

$$(1-\alpha)\lambda \qquad vs. \qquad 1/N| < x, y > -\alpha < x, X\beta > |$$

To proceed with the right hand side, write $x = \sum_i c_i x_i + b$ where x_i are the columns of X and b is orthogonal to the column space of X. Then

$$< x, X\beta >= (\sum_i c_i x_i + b)^T X(X^T X)^{-1} X^T y = \sum_i c_i x_i^T X(X^T X)^{-1} X^T y$$

$$= \sum_i c_i x_i^T y = \sum_i c_i < x_i, y > \Longrightarrow$$

$$1/N| < x, y > -\alpha < x, X\beta > | = 1/N| < \sum_i c_i x_i + b, y > -\alpha \sum_i c_i < x_i, y > |$$

$$= 1/N|(1-\alpha)\sum_i c_i < x_i, y > + < b, y > |.$$

When $\alpha \to 1$, this becomes $1/N| < b, y > |$ (which is positive unless the variable x is perfectly correlated with some linear combination of the current active features or otherwise perfectly uncorrelated with y – in either case, it would be useless for further prediction and never enter the active set). Since

(a) initially, the absolute correlations of the active variables with the current residual are all tied and strictly larger than the absolute correlation of any inactive variable with the current residual,

(b) the absolute correlations of the active variables with the current residual decrease monotonically to 0, and

(c) the absolute correlations of the inactive variables with the current residual do not go to 0,

there is some first point $\alpha_i \in [0, 1]$ for each inactive variable x_i where the absolute correlation of x_i with the current residual is tied with the absolute correlations of all the active variables with the current residual. Then letting $\alpha_* = \min_i \alpha_i$, we see that x_* will be the next variable to enter the active set at step $k+1$ at α_*, which we can find explicitly by solving

$$(1-\alpha)\lambda = |(1-\alpha)\sum_i c_i * sign(< x_i, y >)\lambda + < b, y > /N|$$

26. In exercise 9, we showed that at each step forward stepwise regression adds the variable most correlated with the residual (after "linearly adjusting" for active variables currently in the model), i.e. for which $|Cor(x_{j,A}, r)|$ is maximized. Then forward stepwise proceeds entirely to the least squares solution before repeating the process.

    On the other hand, LAR adds the variable most correlated with the current residual when it becomes tied with the correlations of the active variables with the current residual, without regard for "linearly adjusting" for the active variables currently in the model. This is because LAR does not (generally) proceed all the way to the least squares solution, so the current residual at any given time will still be correlated with the variables in the model, unlike in forward stepwise regression.

    For an example, of this, consider a regression problem where two features are (almost) perfectly correlated and both very correlated with the output y. Then forward stepwise regression will add one of the variables early on since it is very correlated with the residual. However, after having the first one, the second one contributes very little to decreasing the residual sum of squares so will not be added until much later. On the other hand, both variables will have high correlation with the current residuals so LAR will instead add them close together.

27. (a) To obtain the Lagrange dual function, we need to add one term per constraint not yet represented in the objective. Thus we create variables $\lambda_j^{\pm}$ and define the dual function

$$L(\beta) + \lambda \sum_j \beta_j^+ + \beta_j^- - \sum_j \lambda_j^+ \beta_j^+ - \lambda_j^- \beta_j^- .$$

The KKT optimality (stationarity) conditions arise from taking derivatives with respect to the optimization variables $\beta_j^{\pm}$:

$$d/d\beta_j^+ : \nabla L(\beta)_j + \lambda - \lambda_j^+ = 0$$

$$d/d\beta_j^- : -\nabla L(\beta)_j + \lambda - \lambda_j^- = 0$$

as well as the complementary slackness conditions corresponding to the constraints:

$$\lambda_j^+ \beta_j^+ = 0$$
$$\lambda_j^- \beta_j^- = 0$$

and the (primal+dual) feasibility (nonnegativity) constraints $\lambda_j^{\pm} \geq 0$ and $\beta_j^{\pm} \geq 0$.

(b) From the first two conditions and the nonnegativity constraints,

$$\nabla L(\beta)_j = -\lambda + \lambda_j^+ \geq -\lambda$$

$$\nabla L(\beta_j) = \lambda - \lambda_j^- \leq \lambda$$

so $-\lambda \leq \nabla L(\beta)_j \leq \lambda \implies |\nabla L(\beta)_j| \leq \lambda$.

First, either $\lambda = 0$ or $\lambda > 0$. If $\lambda = 0$, then since $|\nabla L(\beta)_j| \leq \lambda = 0$, it must be that $|\nabla L(\beta)_j| = 0$. Otherwise, suppose $\lambda > 0$. If $\beta_j^{\pm} > 0$, then the complementary slackness constraints imply $\lambda_j^{\pm} = 0$ and hence that $\pm\nabla L(\beta)_j = -\lambda$. Thus in the other constraint $\mp L(\beta)_j + \lambda - \lambda_j^{\mp} = 0$ we have $2\lambda = \lambda_j^{\mp}$ so in particular $\lambda_j^{\mp} \neq 0$ and hence by the complementary slackness constraints we must have $\beta_j^{\mp} = 0$. (In principle, it seems possible that $\beta_j^{\pm} = 0$, which does not fall into one of the situations above. This would correspond to a variable being unused and means either that adding in that variable would only serve to increase the loss or otherwise that the reduction in loss is offset by λ.)

$L(\beta) = (Y - X\beta)^T(Y - X\beta)$. Taking a gradient with respect to β yields $(-2X^T)(Y - X\beta)$. From the results above, we compute that the gradient is also $-\lambda * sign(\beta)$. Equating these two vectors coordinatewise, we obtain $-2x_j^T(Y - X\beta) = -\lambda * sign(\beta_j) \implies \lambda = sign(\beta_j) * 2x_j^T(Y - X\beta)$ so λ is directly proportional (through a factor of $2sign(\beta_j)$) to the correlation between x_j and the current residuals.

(c) From the relation between λ and the residuals computed in the previous problem, we established

$$(-2X^T)(Y - X\beta) = -\lambda sign(\beta)$$

and rearranging gives a direct expression for β in terms of λ:

$$\beta = (X^T X)^{-1}(X^T Y - \lambda/2 sign(\beta)).$$

Plugging this into the expression given $\beta(\lambda) = \beta(\lambda_0) - (\lambda - \lambda_0)\gamma_0$ and solving for γ_0 yields:

$$(X^TX)^{-1}(X^TY - \lambda/2sign(\beta)) = (X^TX)^{-1}(X^TY - \lambda_0/2sign(\beta)) - (\lambda - \lambda_0)\gamma_0$$

$$(X^TX)^{-1}sign(\beta)/2 = \gamma_0.$$

Note that since λ does not appear in the left hand side and all of the active variables are assumed to remain the same (i.e. no nonzero coefficient becomes 0 and no zero coefficient becomes nonzero), the vector $sign(\beta)$ and the matrix $(X^TX)^{-1}$ are constant for $\lambda \in [\lambda_1, \lambda_0]$ so the LHS is constant for all $\lambda \in [\lambda_1, \lambda_0]$. Thus the lasso solution path is linear for $\lambda \in [\lambda_1, \lambda_0]$ and more generally is piecewise linear as a function of λ.

28. With the additional variable, the lasso regression problem becomes

$$\min_{\beta^*} ||Y - X_1\beta_1 + \cdots + X_j(\beta_j + \beta_j^*) + \cdots + X_p\beta_p||^2 + \lambda(|\beta_1| + \cdots + |\beta_j| + |\beta_j^*| + \cdots + |\beta_p|).$$

Suppose that a solution β to this problem has $sign(\beta_j)sign(\beta_j^*) < 0$. This is a contradiction since the value could be decreased by setting $\widehat{\beta}_j = \beta_j + \beta_j^*$ and $\widehat{\beta}_j^* = 0$ while keeping the remaining values identical $\widehat{\beta}_i = \beta_i$, since this would not impact the RSS term while it would decrease the penalization term. Thus it must be the case that $sign(\beta_j)sign(\beta_j^*) \geq 0$ for any solution to the minimization problem above so that $|\beta_j| + |\beta_j^*| = |\beta_j + \beta_j^*|$ for any solution to the problem above. Thus it is equivalent to minimize

$$\min_{\beta^*} ||(Y - X_1\beta_1 + \cdots + X_j(\beta_j + \beta_j^*) + \cdots + X_p\beta_p)||^2 + \lambda(|\beta_1| + \cdots + |\beta_j + \beta_j^*| + \cdots + |\beta_p|)$$

which is exactly the original lasso regression problem before adding the new variable. Thus the set of solutions for $\widehat{\beta}_j$ and $\widehat{\beta}_j^*$ is all possible values such that $\widehat{\beta}_j + \widehat{\beta}_j^* = a$ and $\widehat{\beta}_j$ and $\widehat{\beta}_j^*$ do not have opposing signs. Note that the same argument generalizes directly to any number of copies of X_j, showing that the new set of solutions is any such that $\sum_i \beta_{j,i} = a$ and all the $\beta_{j,i}$ have the same sign (or are 0).

29. With the additional variables, the ridge regression problem becomes

$$\min_{\beta^*} ||Y - X_1\beta_1 + \cdots + X_j(\beta_{j,1} + \cdots + \beta_{j,m}) + \cdots + X_p\beta_p||^2$$

$$+\lambda(\beta_1^2 + \cdots + (\beta_{j,1}^2 + \cdots + (\beta_{j,m}^*)^2) + \cdots + \beta_p^2)$$

Suppose that a solution β to this problem has $\beta_{j,i} \neq \beta_{i,k}$. Then the value could be decreased by setting $\beta_{j,i} = \beta_{j,k} = (\beta_{j,i} + \beta_{j,k})/2$ since

$$(\beta_{j,i}^2 + \beta_{j,k}^2) - 2(\frac{\beta_{j,i} + \beta_{j,k}}{2})^2 = \beta_{j,i}^2/2 + \beta_{j,k}^2/2 - \beta_{j,i}\beta_{j,k} = \frac{1}{2}(\beta_{j,i} - \beta_{j,k})^2 > 0.$$

Thus it must be the case that $\beta_{j,1} = \cdots = \beta_{j,m}$ for any solution to the ridge regression minimization problem above.

Note that unlike in the lasso regression problem, this change can affect the entire ridge regression fit. This is due to the fact that if we kept $\sum_i \beta_{j,i} = a \implies \beta_{j,i} = a/m$, then $\sum_i \beta_{j,i}^2 = a^2/m$ so the contribution to the RSS has not changed while the contribution to the penalty term has decreased by a factor of m. As a result, it may be the case that we can find a more optimal solution by rebalancing all the regression coefficients. Geometrically, this corresponds to shrinking the constraint region determined by $||\beta||_2^2 \leq t$ in the j^{th} coordinate direction by a factor of m, and this can clearly impact the entire ridge regression fit.

To see this explicitly, note that since all the $\beta_{j,i}$ must be identical, we can set $\beta_j = \sum_i \beta_{j,i} = m\beta_{j,i}$ for any i and hence $\beta_{j,i} = \beta_j/m$. Thus $\sum_i \beta_{j,i}^2 = \beta_j^2/m$ and we can rewrite the ridge regression problem as

$$\min_\beta (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T D\beta$$

where $D = diag(1, \ldots, d_{jj} = 1/m, \ldots 1)$. Then solving this analytically yields $\beta = (X^TX + \lambda D)^{-1}X^TY$. (The β derivative is $(-2X^T)(Y - X\beta) + 2\lambda D\beta$ and setting this equal to 0 yields $X^TY = (X^TX + \lambda D)\beta \implies \beta = (X^TX + \lambda D)^{-1}X^TY$.) In the case when Y is regressed on a single variable X which is repeated m times, this yields $\beta_1 = (x^Ty)/(x^Tx + \lambda/m)$. From here, we need to multiply by $1/m$ to extract the values of the original (untransformed) coefficients $\beta_{1,i} = \beta_1/m = (x^Ty)/(mx^Tx + \lambda)$.

note There are several related questions one could ask about duplicating data in linear regression. Note that if we duplicate a variable in ordinary least squares, the result will no longer have a well-defined solution, since X^TX will not be full rank. In fact, X^TX has full rank if and only if the columns of X are linearly independent.

In LAR, it seems like the duplicated predictors will behave identically so there will be a tie when they are supposed to enter the active set. At this point, the algorithm presented for computing the LAR coefficient profile will fail since we can no longer compute the least squares solution and make progress towards the least squares solution until an inactive predictor's correlation becomes tied.

Another possibility is duplicating the data (whereas duplicating a variable adds repeated columns to X, duplicating the data would result in adding repeated rows to X). If we fully duplicate the data m times, then the effect is that the RSS is scaled by m. In ordinary least squares, this has no impact on the solution (though it does impact our estimate of the variance by making it artificially small). In lasso and ridge regression, it changes the relative importance of the regularization term, and we can view this as scaling down λ by a factor of m. Intuitively, this makes sense since we are prioritizing the RSS by scaling it by m, so the relative importance of the regularization term decreases.

30. This is equivalent to exercise 12. We can add p data points of the form $(x_i = \sqrt{\lambda a \alpha} e_i, y = 0)$ to the problem, leading to the new problem

$$\min_\beta ||y - X\beta||^2 + \lambda(1 - \alpha)||\beta||_1.$$