Quiz #4 – Solutions

1. Which of the three models with *k* predictors (*k* is any whole number between 0 and 10) has the smallest training RSS? (Select the single best answer.)

a) Best subset.
b) Forward stepwise.
c) Backward stepwise.
d) There is no way to know without further information.

Solution: A

2. Which of the three models with *k* predictors has the smallest test RSS? (Select the single best answer.)

a) Best subset.
b) Forward stepwise.
c) Backward stepwise.
d) There is no way to know without further information.

Solution: D

3. True or False? The predictors in the *k*-variable model identified by forward stepwise are a subset of the predictors in the (*k*+1)-variable model identified by forward stepwise selection.

Solution: True

4. True or False? The predictors in the *k*-variable model identified by backward stepwise are a subset of the predictors in the (*k*+1)-variable model identified by backward stepwise selection.

Solution: True

5. True or False? The predictors in the *k*-variable model identified by backward stepwise are a subset of the predictors in the (*k*+1)-variable model identified by forward stepwise selection.

Solution: False

6. True or False? The predictors in the *k*-variable model identified by forward stepwise are a subset of the predictors in the (*k*+1)-variable model identified by backward stepwise selection.

Solution: False

7. True or False? The predictors in the *k*-variable model identified by best subset are a subset of the predictors in the (*k*+1)-variable model identified by best subset selection.

Solution: False

8. Determine which of the following methods are appropriate for selecting a single best model from among the 11 models obtained for each approach. (Select all that apply.)

a) Choose the model with lowest training RSS value.
b) Choose the model with lowest training C_p value.
c) Choose the model with lowest training BIC value.
d) Choose the model with lowest training adjusted R^2 value.
e) Choose the model with lowest validation set or cross-validation MSE.

Solution: B, C, E

9. Consider fitting a least squares regression model to 3 predictor variables, X_1, X_2, and X_3. Which of the following statements are true? (Select all that apply.)

a) Shrinkage methods like ridge regression and the lasso try to trade-off some bias for reduced variance so that mean squared error is reduced.
b) If we fit a ridge regression model to X_1, X_2, and X_3, each model coefficient estimate is shrunk to somewhere between zero and the corresponding least squares model coefficient estimate.
c) If we fit a lasso model to X_1, X_2, and X_3, some model coefficient estimates can be shrunk all the way to zero.

Solution: A, C

There are two true statements: shrinkage methods trade-off bias for variance and lasso models can shrink coefficient estimates all the way to zero. However, it is not true that ridge regression shrinks each estimate to between zero and the corresponding least squares estimate. Figure 6.4 on page 216 in the textbook shows that this is not true since ridge estimates can have a different sign from the least squares estimate at certain points in the shrinkage path. This is an important point to appreciate because it can sometimes make ridge regression models challenging to interpret.

10. Which of the following statements about regression shrinkage methods are true? (Select all that apply.)

a) Using too large a value of λ can cause your model to underfit the data.

b) Using too large a value of λ can cause your model to overfit the data.
c) Using a very large value of λ cannot hurt the performance of your model; the only reason we do not set λ to be too large is to avoid numerical problems.

Solution: A

There was only one true statement for this question, that setting lambda too high can lead to underfitting.

11. You are training a ridge regression model. Which of the following statements are true? (Select all that apply.)

a) Adding many new predictors to the model helps prevent overfitting on the training set.
b) Adding many new predictors to the model makes it more likely to overfit the training set.
c) Shrinking the model coefficient estimates always results in equal or better performance on the training set.
d) Shrinking the model coefficient estimates always results in equal or better performance on observations not in the training set, i.e., the test set.

Solution: B

There was only one true statement for this question that too many predictors can lead to overfitting. It is not true that shrinking coefficient estimates always results in equal or better performance on the training set. The least squares estimates are optimal (in terms of smallest RSS) for the training set, so shrinking them can only worsen performance. The point of shrinking the estimates is to try to improve performance on the test set. This is a common theme of the course - using methods that attempt to limit overfitting the training data so that the model generalizes to better predict test data.

12. Suppose you ran ridge regression twice, once with $\lambda=0$, and once with $\lambda=1$. One of the times, you got coefficient estimates $\beta=(81.47,12.69)$, and the other time you got $\beta=(13.01,0.91)$. However, you forgot which value of λ corresponds to which value of β. Which one do you think corresponds to $\lambda=1$?

a) $\beta=(81.47, 12.69)$
b) $\beta=(13.01, 0.91)$

Solution: B

The smaller beta-coefficient estimates corresponds to $\lambda=1$ (more shrinkage) versus $\lambda=0$ (no shrinkage).

13. Ridge regression, relative to least squares, is:

a) More flexible (with respect to the possible values that the regression coefficient estimates can take) and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
b) More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.
c) Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
d) Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

Solution: C

Ridge regression is less flexible than least squares (since they restrict the size of the regression coefficient estimates), leading to decreased variance but increased bias.

14. The lasso, relative to least squares, is:

a) More flexible (with respect to the possible values that the regression coefficient estimates can take) and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
b) More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.
c) Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
d) Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

Solution: C

The lasso is less flexible than least squares (since they restrict the size of the regression coefficient estimates), leading to decreased variance but increased bias.

15. True or False? Since the lasso performs variable selection by shrinking some regression coefficient estimates to zero, it always leads to better prediction accuracy than ridge regression.

Solution: False

Although the lasso produces simpler, more interpretable models involving only a subset of the predictors, it may or may not lead to better prediction accuracy than ridge regression, depending on the nature of the dataset (see page 223-4).

16. Relative to multiple linear regression using least squares, what are the typical properties of the coefficient estimates for dimension reduction methods like principal components regression and partial least squares when used in situations where the number of predictors, p, is large relative the number of observations, n?

a) Reduced bias and reduced variance.
b) Reduced bias and increased variance.
c) Increased bias and reduced variance.
d) Increased bias and increased variance.

Solution: C

Since the coefficients for dimension reduction methods are constrained, they can be biased. However, if a small number of principal components or partial least squares directions are used, then the variance can be reduced (see page 230).

17. Match the methods to the types of technique.

a) Principal components analysis – Unsupervised Learning
b) Principal components regression – Supervised Learning

Solution: PCA is a tool for unsupervised learning, whereas PCR applies PCA in a supervised setting. In PCR, each principal component is associated with Y as defined in Section 2.1.4 as supervised learning. Though the principal components are not created based on the relationship with Y, they are ultimately used in a regression of Y on Z.

18. True or False? In a setting with p predictors, principal components regression using p principal components is identical to fitting a multiple linear regression model using least squares to all p predictors.

Solution: True

Equation (6.18) on page 229/230 shows that in a setting with p predictors, principal components regression using p principal components is identical to fitting a multiple linear regression model using least squares to all p predictors.

19. Consider the fitted or predicted values for multiple linear regression, the first principal component, and the first partial least squares direction. Each can be written as a linear combination of the predictors, where each method differs in how it computes the predictor weights. Match each method to the description of how the predictor weights are computed.

a) Multiple linear regression fitted values - The predictor weights minimize the sum of squared differences between the response and the linear combination of the predictors.

b) First principal component - The predictor weights maximize the variance of the linear combination of the predictors.
c) First partial least squares direction - The weight for each predictor is equal to the slope coefficient from a simple linear regression of the response onto that predictor.

Solution: Multiple linear regression fitted values have predictor weights that minimize the sum of squared differences between the response and the linear combination of the predictors (i.e., least squares). The first principal component has predictor weights that maximize the variance of the linear combination of the predictors (see page 231). The first partial least squares direction has a weight for each predictor equal to the slope coefficient from a simple linear regression of the response onto that predictor (see page 237).

20. Select the true statements from the following. (Select all that apply.)

a) Principal components regression and partial least squares are both dimension reduction methods.
b) Principal components regression and partial least squares both identify linear combinations of predictors in a supervised way.
c) Principal components regression and partial least squares are both less flexible than least squares since they constrain the estimated regression coefficients.
d) Principal components regression and partial least squares are both variable selection methods.

Solution: A, C

A and C are correct statements. B is incorrect because PCR does not identify linear combinations of predictors in a supervised way. (see page 237). D is incorrect because neither PCR nor PLS are variable selection methods (see pages 235/6).

21. Which of the following is the most appropriate way to select the number of principal components to use in PCR?

a) Use the first two principal components so that we can display them with the response variable in a 3D scatterplot.
b) Choose enough components to explain a certain proportion of the variance in the predictor data, say 90%.
c) Minimize mean squared error in the training set.
d) Maximize R^2 in an independent test set.

Solution: D

Answers A and B are completely arbitrary, whereas C is a big no-no no matter what the method.

22. Select the true statements from the following in the context of high-dimensional data in which the number of predictors, p, is larger than the number of observations, n. (Select all that apply.)

a) Multiple linear regression using least squares overfits the data and should not be used.
b) Models that constrain the regression coefficients such as principal components regression can avoid overfitting.
c) Adding predictors to a high-dimensional analysis tends to reduce test set error, whether or not the predictors are associated with the response.
d) Multicollinearity issues are less of an issue in high-dimensional data because we have so many predictors to select from.

Solution: A, B

Answers A and B are both true (see section 6.4). C is incorrect since adding predictors that are not associated with the response leads to a deterioration in the fitted model (the curse of dimensionality). D is incorrect because multicollinarity issues tend to be more extreme in high-dimensional settings (see page 243).