STAT 154: Study Guide/Practice Problems for Midterm 1

Johnny Hong

October 11, 2017

1 True/False

Indicate True or False for each of the following statements. Justify your answers.

1. The principal components are always orthogonal to each other.
2. PCA relies on an eigendecomposition of the data matrix X.
3. When doing PCA, we should always keep only the first two PCs.
4. Johnny ran a PCA based on the correlation matrix. Gaston ran a PCA based on the covariance matrix. They must obtain the same PCs (up to a sign difference).
5. Each of the PCs is a linear combination of the columns of X.
6. For a simple linear regression, the correlation coefficent always has the same sign as the OLS estimate of the slope.
7. Suppose $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ have the same sample averages and the same sample standard deviations. Let $\hat{\alpha}_0$ and $\hat{\alpha}_1$ be the least squares estimates of the intercept and the slope when regressing y on x. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be those when regressing x on y. Then $\hat{\alpha}_0$ must be the same as $\hat{\beta}_0$ and $\hat{\alpha}_1$ must be the same as $\hat{\beta}_1$.
8. Suppose the sum of $\{x_i\}_{i=1}^n$ is 0. Then the y-intercept of the least squares line must be the sample average of y.

9. If X is $n \times n$ with rank n, then the vector of fitted values $\hat{y}$ is exactly the same as the vector of observed values y.

10. Multicollinearity leads to bias in the OLS estimates.

Assume the standard simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with ϵ_i iid $N(0, \sigma^2)$ and all the x_i's being fixed for the remaining T/F questions. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the least-squares estimators of β_0 and β_1 respectively.

1. The sum of residuals is **always** 0.

2. Let Y be a new response and x be the associated predictor. The prediction interval for Y is **always** wider than the confidence interval for $\beta_0 + \beta_1 x$.

3. Suppose I have the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ after fitting the model and now I want to generate a least-squares line with x being the response and Y being the predictor. The slope of this new least-squares line is $\frac{1}{\hat{\beta}_1}$.

4. If $\bar{x} = 0$, then $\hat{\beta}_0$ and $\hat{\beta}_1$ are uncorrelated.

2 Mathematical Questions

1. Consider the standard linear regression model $y = X\beta + \epsilon$, where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix, and $\epsilon \sim N(0, I_n)$ is the noise vector. Assume $n > p$ and X has full column rank.

   (a) What is the hat matrix H? Express your answer in terms of the matrix X.

   (b) Prove that H is symmetric.

   (c) Prove that H is idempotent, meaning that $H^2 = H$.

   (d) Recall that H is a projection matrix. On what space does it project?

   (e) Prove that the sum of the diagonal entries of H is p.

   (f) Show that for all $i = 1, ..., n$,

   $$H_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}, \tag{1}$$

   where $\hat{y}_i$ is the fitted value for the ith observation.

(g) Recall that H_{ii} is referred as the leverage score for the ith observation. Using Equation (1), explain how leverage scores can be viewed as measures of self-sensitivity.

2. Suppose I have two vectors of length 100 x and y in R and I fit the simple linear regression model using `lm(y x)`.

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.88399 -0.42440 -0.06309  0.33691  1.33428 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.10515    -------    0.93    0.355    
x            3.01347    0.08893   ------   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5315 on ----- degrees of freedom
Multiple R-squared:  0.9214,Adjusted R-squared:  0.9206 
F-statistic:  1148 on 1 and ----- DF,  p-value: < 2.2e-16
```

(a) Find the estimated standard error of the estimated intercept $\hat{\beta}_0$.

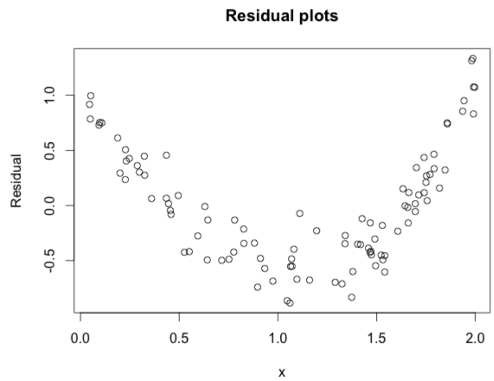
(b) What should be the t value for $\hat{\beta}_1$ in the above output?

(c) Test $H_0 : \beta_0 = 0$ versus $H_0 : \beta_0 \neq 0$ at 5% level.

(d) Construct a (two-sided) 95% confidence interval for β_1.

(e) Use the confidence interval obtained above to test $H_0 : \beta_1 = 4$ versus $H_1 : \beta_1 \neq 4$.

(f) Suppose we get the following residual plot:



Comment.

(g) Suggest a graphical technique to check the normality assumption.

3 Concept questions

1. What is supervised learning? What is unsupervised learning?

2. Give an example of an unsupervised learning technique.

3. Give an example of a supervised learning technique.

4. Explain why in-sample mean squared error is not a good metric to assess a model's predictive power for a regression problem.

5. What does Gauss-Markov theorem tell us about the OLS estimators?

6. Explain the idea of overfitting.

7. Sometimes we will split the dataset into three parts: the training set, the validation set, and the test set. Explain the use of each.

8. In the context of OLS, explain what multicolinearity is and why it is an issue.

9. After fitting a linear regression model using `lm()`, I found that the R^2 is close to 0. Should I conclude that there is no relationship between the predictor(s) and the response? Why or why not?

10. Give a method to select the number of components in principal component regression.

11. Consider ridge regression with regularization parameter λ. Sketch the general pattern you expect to see in each of the following plots and briefly explain the rationale for the plots:

    (a) squared bias vs λ

    (b) variance vs λ

    (c) MSE vs λ

4 Coding questions

1. Let $\mathbf{y}$ be the response vector and $\mathbf{X}$ be the design matrix. Suppose I want to estimate the predictive power of the OLS model. Write pseudocode for computing the 10-fold cross-validation MSE.

2. (Ex. 3.12 from ESL)

    (a) Suppose $\mathbf{X}$ is mean-centered. Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented dataset. We augment $\mathbf{X}$ with p additional rows $\sqrt{\lambda}\mathbf{I}$ and augment $\mathbf{y}$ with p zeros.

    (b) Using the idea in part a), implement a function called `ridge()` that returns the ridge regression estimates. The function should take three arguments:

    - y: a response vector of length n
    - X: a $n \times p$ predictor matrix (you can assume that X is already mean-centered)
    - lambda: a postive scalar, representing the regularization parameter

    You are allowed to use `lm()` in your function.