COLUMBIA UNIVERSITY
Statistical Inference and Modeling
Quiz 4

Prof. Marco Avella
TA: Ian Kinsella, Shun Xu
April 4, 2019

Name:

UNI:

You have 20 minutes to answer the following 10 questions. Good luck!

Question 1

Assume that we have an i.i.d sample of pairs $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ where $\mathbf{X}_i \in \mathbb{R}^p$. Consider the linear model

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \;\; i = 1, \ldots, n \tag{1}$$

where ε_i, are i.i.d. with $\mathbb{E}[\varepsilon_i] = 0$, $\text{var}(\varepsilon_i) = \sigma^2$ and $\text{cov}(\mathbf{X}_i, \varepsilon_i) = \mathbf{0}$. Let $\hat{\boldsymbol{\beta}}$ be the least squares estimator. Given a fixed design matrix $\mathbf{X}$, which of the following statements is NOT correct?

(a) If $\sigma^2 = 1$, then $\text{var}\Big(\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2\Big) = 2(n-p)$

(b) $\text{tr}\{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\} = p$

[✓ (c)] $\hat{\sigma}_0^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2$ is an unbiased and consistent estimator of σ^2.

(d) $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

Question 2

Assume model (1). Further assume that errors are normally distributed and that $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$. Which of the following statements is NOT correct?

[✓ (a)] $\hat{\boldsymbol{\beta}} = n\mathbf{X}^T\mathbf{Y}$.

(b) $\hat{\beta}_1$ and $\hat{\beta}_2$ are uncorrelated

(c) $\hat{\beta}_j \sim N(\beta_j, \frac{1}{n}\sigma^2)$ for all $j = 1, \ldots, p$

(d) Under $H_0 : \beta_1 = \beta_2$ we have that $\sqrt{n}(\hat{\beta}_1 - \hat{\beta}_2) \sim N(0, 2\sigma^2)$.

Question 3

Consider the model (1) with normal errors and the ridge regression estimator $\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$. Given a fixed design matrix $\mathbf{X}$, which of the following statements is NOT correct?

(a) $\text{var}(\hat{\boldsymbol{\beta}}_\lambda) = \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}$

(b) $\hat{\boldsymbol{\beta}}_\lambda$ is normally distributed.

[✓ (c)] $\hat{\boldsymbol{\beta}}_\lambda$ always has a better MSE than the least squares estimator.

(d) $\hat{\boldsymbol{\beta}}_\lambda$ is a consistent estimator if the tuning parameter is such that $\lambda \to 0$ as $n \to \infty$

Question 4

Consider the partition $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ and the restricted least squares estimator $\hat{\boldsymbol{\beta}}_1^R = (\mathbf{X}_1\mathbf{X}_1)^{-1}\mathbf{X}_1\mathbf{Y}$.
Under $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$, which of the following statements is NOT correct?

(a) $\hat{\boldsymbol{\beta}}_1^R$ is an unbiased estimator of $\boldsymbol{\beta}_1$.

(b) $\mathbb{E}[\hat{\boldsymbol{\beta}}_2] = \mathbf{0}$.

[✓ (c)] The likelihood ratio statistic follows an exact chi squared distribution.

(d) $\|\mathbf{Y} - \mathbf{X}^T\hat{\boldsymbol{\beta}}_1^R\|_2^2 - \|\mathbf{Y} - \mathbf{X}^T\hat{\boldsymbol{\beta}}\|_2^2$ and $\|\mathbf{Y} - \mathbf{X}^T\hat{\boldsymbol{\beta}}\|_2^2$ are independent.

Question 5

Which of the following statements is NOT correct.

[✓ (a)] No estimator can have a negative R^2.

(b) $R^2/(1 - R^2)$ is proportional to an F statistic with $p - 1$ and $n - p$ degrees of freedom.

(c) The smaller the adjusted R^2 the poorer the fit of the data.

(d) Adding more variables to a least squares fittinig procedure will increase the R^2.

Question 6

A least squares fit in R gave the following output:

```
Call:
lm(formula = sal ~ lag + trend + dis, data = salinity)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6613 -0.8242  0.2222  0.6459  2.7537

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.69427    3.13026   3.097  0.00492 **
lag          0.77692    0.08597   9.038 3.41e-09 ***
trend       -0.02835    0.16087  -0.176  0.86160
dis         -0.29903    0.10712  -2.792  0.01013 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*‘ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.327 on 24 degrees of freedom
Multiple R-squared:  0.8273,Adjusted R-squared:  0.8057
F-statistic: 38.32 on 3 and 24 DF,  p-value: 2.605e-09
```

$= \hat{\sigma} = \sqrt{\frac{1}{n-p} \|Y - X\hat{\beta}\|^2}$

Assume that $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \mathbf{I}_n\sigma^2)$ in the following queations. Assume we wish to fit a smaller model using only two of the three predictors:

1. Based on only the above output, which two predictors would you keep in the simplified model?

   lag and dis based on p-values

2. Using only information from the regression output of both fits, what test could you use to test the null hypothesis that the left out predictor is 0?

   Could do a LRT to test $H_0: \beta_{trend} = 0$
   $\Rightarrow$ F test

3. Give a formula for the above test in terms of values that could be found in the regression outputs.

   $$F = \frac{n-p}{p-q} \frac{\|\hat{\varepsilon}^R\|^2 - \|\hat{\varepsilon}\|^2}{\|\hat{\varepsilon}\|} = 24 \frac{25 \cdot \hat{\sigma}_R^2 - 24\hat{\sigma}^2}{24\hat{\sigma}^2}$$

Next, assume that we wish to test the null hypothesis $H_0 : \sigma^2 = 1$ against $H_a : \sigma^2 > 1$ for the original fit (i.e. using the provided regression output):

4. Which quantity found in the output could be best used as a test statistic (up to a scaling by a constant) for the null hypothesis?

   $\hat{\sigma}^2 = 1.327^2$

5. Give a formula for your test statistic and state its distribution.

$$(n-p)\hat{\sigma}^2 = \|Y - X\hat{\beta}\|^2 \sim \sigma^2 \chi^2_{n-p}, \quad \mathbb{P}(\chi^2_{n-p} > 24 \cdot 1{,}327^2)$$