Linear Regression Models
Statistics GR5205/GU4205 — Fall 2020

Homework 3

The following problems are due on Monday, Oct 19th, 11:59pm.

1. **Problem 2.14 & 2.24 in KNN**

   Continue with the *Copier maintenance* data.

   (a) Obtain a 95% confidence interval for the mean service time on calls in which six copiers are serviced. Interpret your interval.

   (b) Obtain a 95% prediction interval for the service time on the next call in which six copiers are serviced. How does this interval compare to that in part (a)?

   (c) Management wishes to estimate the expected service time *per copier* on calls in which six copiers are serviced. Obtain and interpret an appropriate 95% confidence interval.

   (d) Set up the basic ANOVA table: sums of squares, degrees of freedom, and mean squares for regression and error.

   (e) Conduct an F-test to determine whether or not there is linear association between time spent and number of copiers serviced. Clearly state your null and alternative hypotheses. Report and interpret the p-value for your test.

   (f) By how much, relatively, is the total variation in number of minutes spent on a call reduced when the number of copiers serviced is introduced in to the analysis? Is this a relatively small or large reduction? What is the name of this measure?

2. **Problem 6.15 in KNN**

A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (X_1, in years), severity of illness (X_2, an index), and anxiety level (X_3, an index). The administrator randomly selected 46 patients and collected the data in the file `patient_satisfaction.txt`, where larger values of Y, X_2, and X_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety.

(a) Obtain the scatterplot matrix, and comment on the relationships among the variables.

(b) Fit the multiple regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

for $i = 1, \dots, n$, where

$$\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \text{ are iid } N(0, \sigma^2) .$$

Carefully interpret the value of b_2, the least squares estimate of β_2.

(c) Plot the residuals against $\hat{Y}$ and each of the predictor variables. Do these plots suggest any model assumptions may be violated?

(d) Prepare a Q-Q plot of the residuals. What does this plot reveal?

3. **Problems 6.16 and 6.17 in KNN)**

Continue with the *Patient satisfaction* data. Assume the regression model defined in Q2 part (b) is appropriate.

(a) Test whether there is a regression relation between Y and X_1, X_2, X_3 using F test. Clearly state your null and alternative hypotheses, report a p-value, and interpret the results.

(b) Obtain joint interval estimates of β_1, β_2, and β_3, using a 90% family confidence coefficient. Interpret your results.

(c) Calculate R^2. What does it indicate here?

(d) Obtain a 95% confidence interval for the mean satisfaction when $X_{h1} = 35$, $X_{h2} = 45$, and $X_{h3} = 2.2$. Interpret your interval.

(e) Obtain a 95% prediction interval for a new patient's satisfaction when $X_{h1} = 35$, $X_{h2} = 45$, and $X_{h3} = 2.2$. Interpret your interval.

4. **Understanding R^2.** Consider the 3-dim multivariate normal distribution:

$$\begin{bmatrix} X_1 \\ X_2 \\ \epsilon \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x^2 & 0 \\ \rho\sigma_x^2 & \sigma_x^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} \right).$$

Let

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon.$$

(a) What is the distribution of Y? What is the mean and variance of Y?

(b) Define

$$r_1 = \frac{\mathrm{Cov}(X_1, Y)}{\sqrt{\mathrm{Var}[X_1]\mathrm{Var}[Y]}}.$$

Calculate r_1 in terms of $\beta_0, \beta_1, \beta_2, \mu_1, \mu_2, \sigma^2, \sigma_x^2, \rho$.

(c) Now define

$$r = \frac{\mathrm{Cov}(X_1\beta_1 + X_2\beta_2, Y)}{\sqrt{\mathrm{Var}[X_1\beta_1 + X_2\beta_2]\mathrm{Var}[Y]}}.$$

Calculate r in terms of $\beta_0, \beta_1, \beta_2, \mu_1, \mu_2, \sigma^2, \sigma_x^2, \rho$.

(d) Conclude $r_1^2 \leq r^2$.

Comments:

(a) Recall that $R^2 = 1 - \frac{RSS}{SS_{\text{total}}}$, it is an estimator for the r^2 where r is defined in (b).

(b) Adding in more variables into your model will never decrease r^2.

5. (Bonus) **Independence of LS estimator $\widehat{\beta}$ and the Unbiased Estimator $\widehat{\sigma}^2$**

A very good reference for multivariate normal (Gaussian) distribution is the following:

https://www.statlect.com/probability-distributions/normal-distribution-linear-combinations

Fix $p \geq 2$. Given $x_1, \ldots, x_n$ and write

$$x = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_1^{(p-1)} \\ 1 & x_2^{(1)} & \cdots & x_2^{(p-1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1, & x_n^{(1)} & \cdots & x_n^{(p-1)} \end{bmatrix} = \begin{bmatrix} 1 & x_1^\top \\ 1 & x_2^\top \\ \vdots & \vdots \\ , & x_n^\top \end{bmatrix},$$

where $x \in \mathbb{R}^{n \times p}$ has rank p, notice that this already implies $n \geq p$. Moreover, $\text{rank}(x^\top x) = \text{rank}(x) = p$, which implies that $x^\top x$ is invertible. Suppose for $\beta \in \mathbb{R}^{p \times 1}$, the following model holds:

$$Y = x\beta + \epsilon, \qquad \text{where } \epsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right).$$

We are going to show that the MSE minimizer $\widehat{\beta}$ is independent of the MSE $Q(\widehat{\beta}) = \|Y - \widehat{Y}\|^2$ in the following steps:

(a) Let

$$A = \begin{bmatrix} I_p \\ 0 \end{bmatrix},$$

which is an $n \times p$ matrix with the first p rows forms an p-dim identity matrix, and the rest $n-p$ rows are all 0. Therefore $\text{rank}(A) = \text{rank}(A^\top) = \text{rank}(A^\top A) = \text{rank}(I_p) = p$. Then, what is the distribution of $A^\top \epsilon$?

(b) Now define

$$A_\perp = \begin{bmatrix} 0 \\ I_{n-p} \end{bmatrix},$$

then

$$I_n = \begin{bmatrix} I_p & 0 \\ 0 & I_{n-p} \end{bmatrix} = \begin{bmatrix} A^\top A & 0 \\ 0 & A_\perp^\top A_\perp \end{bmatrix}.$$

Show that $A^\top \epsilon$ and $A_\perp^\top \epsilon$ are independent random vectors, and verify that

$$\|A_\perp^\top \epsilon\|^2 = \|\epsilon\|^2 - \|A^\top \epsilon\|^2 \sim \sigma^2 \chi^2(n-p).$$

(c) We know that $(x^\top x)^{-1}$ is a $p \times p$ symmetric positive definite matrix, according to our linear algebra knowledge, there exists a unique $p \times p$ symmetric positive definite matrix N such that $N^2 = (x^\top x)^{-1}$. We usually write $N = (x^\top x)^{-1/2}$, and we have:

$$N^{-1} = (x^\top x)^{1/2}, \qquad N^{-2} = \left(N^{-1}\right)^2 = x^\top x.$$

Verify that for $\tilde{A} = x(x^\top x)^{-1/2} \in \mathbb{R}^{n \times p}$, $\tilde{A}^\top \tilde{A} = I_p$. What is the distribution of $\tilde{A}^\top \epsilon$?

(d) Like we define $A_\perp$ in (c), we can still construct a matrix $\tilde{A}_\perp \in \mathbb{R}^{n \times (n-p)}$ such that

$$\tilde{A}_\perp^\top \tilde{A}_\perp = I_{n-p}, \qquad \tilde{A}^\top \tilde{A}_\perp = 0.$$

Show that $\tilde{A}^\top \epsilon$ and $\tilde{A}_\perp^\top \epsilon$ are independent random vectors. Verify that

$$\|\tilde{A}_\perp^\top \epsilon\|^2 = \|\epsilon\|^2 - \|\tilde{A}^\top \epsilon\|^2 = \epsilon^\top \left(I_n - x(x^\top x)^{-1} x^\top\right) \epsilon \sim \sigma^2 \chi^2(n-p).$$

(e) Verify that $\widehat{\beta} = \beta + x^\top \epsilon$ and $Q(\widehat{\beta}) = \epsilon^\top \left(I_n - x(x^\top x)^{-1} x^\top\right) \epsilon$. What is the distribution of $\widehat{\beta}$ and $Q(\widehat{\beta})$? Show that they are independent using the result from (d).