Stats 413 Fall 2019

Exam I

Total Points: 50

2:30 pm – 3:50pm. Wednesday, Oct 9

Name: Pete MacDonald

ID: ____________________

Instruction: Answer each question in the space provided. You may continue the answer on the back of the sheet, but please do not continue the answer for any question on the sheet for a different question. If you need more space, please use extra blank sheets provided and label them clearly.

SOLUTION

1. [22 pts] Consider a multiple linear regression model $Y_i = \beta_0+\beta_1 X_{i,1}+\beta_2 X_{i,2}+\beta_3 X_{i,3}+e_i, i = 1, \dots, 100$. We get the following R-output (values are rounded to two digits).

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
   Min     1Q Median     3Q    Max
 -4.50  -1.35   0.05   1.18   4.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.68       0.20     3.5   8e-04 ***
x1            -0.38       0.20    -1.9    0.06 .
x2             1.40       0.19     7.2   1e-10 ***
x3             0.98       0.21     4.6   1e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.9 on 96 degrees of freedom
Multiple R-squared:  0.44,  Adjusted R-squared:  0.42
F-statistic:  25 on 3 and 96 DF,  p-value: 3.9e-12
```

(a) [4 pts] For the considered linear regression model, write out the corresponding mean function. In addition, specify what assumptions are made for the statistical errors in this model.

- $E[Y \mid X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.
- assumptions on errors:

  $E[e_i \mid X] = 0$ e_i's are independent.

  $\mathrm{Var}(e_i \mid X) = \sigma^2$ $(\mathrm{Cov}(\vec{e}) = \sigma^2 I_n)$

(b) [5 pts] What is the ordinary least square estimate of β_3, i.e., $\hat{\beta}_3$? How to interpret this value? What is the value of the standard error of $\hat{\beta}_3$? How to interpret this value? Give a 95% confidence interval of β_3 (suppose the 97.5% quantile of the used t-distribution is 2).

- $\hat{\beta}_3 = 0.98$. expected change in Y for a 1 unit increase in X_3, holding X_1 & X_2 constant
- $se(\hat{\beta}_3) = 0.21$. estimate of standard deviation of the estimator $\hat{\beta}_3$.
- 95% CI for β_3: $[0.98 - 2 \cdot 0.21, 0.98 + 2 \cdot 0.21] = [0.56, 1.4]$

(c) [1 pt] Calculate the predicted value of the response of a new observation with $X_1^* = 1, X_2^* = 1, X_3^* = 0$.

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 X_1^* + \hat{\beta}_2 X_2^* + \hat{\beta}_3 X_3^*$$
$$= 0.68 + (-0.38) \cdot 1 + 1.4 \cdot 1 + 0$$
$$= 1.7$$

(d) [3 pts] What are the values of R^2 and adjusted R^2 for the above regression? How shall we interpret the value of R^2?

$R^2 = 0.44$, $R^2_{adj} = 0.42$

R^2 is the proportion of variance explained by the linear relationship w/ X_1, X_2, X_3.

(e) [3 pts] For the F-statistic in the last line of the R-output, write out the corresponding mean functions under the null and alternative hypotheses. What distribution does the corresponding test statistic follow under the null hypothesis? Give your conclusion for this test.

- $H_0: E[Y|X] = \beta_0$, $H_a: E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.
- Under H_0, the F-statistic follows an $F_{3,96}$ distribution
- $p = 3.9 \times 10^{-12} < 0.05$, so we reject H_0.

(f) [6 pts] We have the following ANOVA (type II) table for the regression $Y \sim X_1 + X_2 + X_3$.

```
Anova Table (Type II tests)

Response: y
          Sum Sq Df F value    Pr(>F)
x1         13.37  1  3.5426   0.06284
x2        198.27  1 52.5220 1.073e-10
x3         80.89  1 21.4266 1.152e-05
Residuals 362.40 96
```

For each F-test in this ANOVA table, specify the mean functions under the null and alternative hypotheses. What distribution does the corresponding test statistic follow under each null hypothesis? Based on the p-values, what are your conclusions about these tests?

- row X_1: $H_0: E[Y|X] = \beta_0 + \beta_2 X_2 + \beta_3 X_3$
  $H_a: E[Y|X] = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_1 X_1$.
  $F \sim F_{1,96}$. $p > 0.05$, do not reject H_0.

- row X_2: $H_0: E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_3 X_3$
  $H_a: E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_2 X_2$
  $F \sim F_{1,96}$. $p < 0.05$, reject H_0.

- row X_3: $H_0: E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ $F \sim F_{1,96}$
  $H_a: E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$, $p < 0.05$, reject H_0.

2. [18 pts] Consider a multiple linear regression with n observations and p predictors. $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$, where $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is the $n \times 1$ vector of responses; $\mathbf{X}$ is the $n \times (p+1)$ matrix of predictors, including a column of 1's for the intercept; and $\mathbf{e} = (e_1, \ldots, e_n)^\top$ is the $n \times 1$ vector of statistical errors. Let $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$.

(a) [5 pts] Derive the formula of the ordinary least square (OLS) estimator $\widehat{\beta}$ (please show details).

(b) [4 pts] Show that $\widehat{\beta}$ is an unbiased estimator of β.

(c) [4 pts] Let $\widehat{\mathbf{Y}} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^\top$ be the fitted values from the OLS estimation. Prove that the sample mean of $\{\widehat{Y}_1, \ldots, \widehat{Y}_n\}$ is the same to the sample mean of $\{Y_1, \ldots, Y_n\}$, i.e., $n^{-1} \sum_{i=1}^n \widehat{Y}_i = \bar{Y}$.

(d) [5 pts] Prove that TSS = RSS + SS_{reg}, where TSS is the total sum of squares (i.e. $\sum_{i=1}^n (Y_i - \bar{Y})^2$), RSS is the residual sum of squares (i.e. $\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$), and SS_{reg} is the regression sum of squares (i.e. $\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2$).

(a) $\hat{\beta}$ minimizes $\underbrace{RSS}_{J(\beta)=} = \sum_{i=1}^{n} (y_i - x_i^T\beta) = (y - X\beta)^T(y - X\beta) = \cdots$

See also lecture 4

$$J(\beta) = Tr((y - X\beta)^T(y - X\beta))$$

$$= Tr(y^Ty) - 2Tr(y^TX\beta) + Tr(\beta^TX^TX\beta)$$

$$\nabla_\beta J(\beta) = -2(y^TX)^T + 2X^TX\beta$$

set $= 0$

$$0 = -2X^Ty + 2X^TX\hat{\beta} \implies \hat{\beta} = (X^TX)^{-1}X^Ty \quad (\text{if } (X^TX)^{-1} \text{ exist})$$

(b) $\hat{\beta} = (X^TX)^{-1}X^Ty \qquad y = X\beta + e.$

$$= (X^TX)^{-1}X^T(X\beta + e)$$

$$= \beta + (X^TX)^{-1}X^Te.$$

$$E[\hat{\beta}] = E[\beta + (X^TX)^{-1}X^Te] = \beta + \underbrace{(X^TX)^{-1}X^TE[e]}_{=0} \qquad E[e] = 0.$$

$$= \beta.$$

thus, $\hat{\beta}$ is unbiased for β.

[Blank page]

(c) equivalently show $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$

$$\text{LHS} = \sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} (y_i + \hat{e}_i) = \sum_{i=1}^{n} y_i + \underbrace{\sum_{i=1}^{n} \hat{e}_i}_{=0}$$

$\bar{\hat{e}} = 0$ from normal equations.

$$= \sum_{i=1}^{n} y_i$$

normal equations: $X^T(y - X\hat{\beta}) = 0$, so from first column of X,

$$\sum_{i=1}^{n} (y_i - x_i^T\hat{\beta}) = 0.$$

$$\Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \underbrace{x_i^T\hat{\beta}}_{\hat{y}_i} = \sum_{i=1}^{n} \hat{y}_i$$

(d). $\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \cdots$

$$\cdots = \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{\text{RSS}} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{\text{SSreg.}} + \overset{(*)}{2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}$$

$y_i - \hat{y}_i = \hat{e}_i$

$\hat{y}_i = x_i^T\hat{\beta}$

3. [10 pts] Consider a simple linear regression model. Give the definition form of R^2. Prove that R^2 is equal to the square of the correlation between the predictor and the response.

- $$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{(using 2(d))}.$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\rho^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= \frac{\sum_{i=1}^{n}(\hat{\beta}_1(x_i - \bar{x}))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

$$= \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} = \rho^2.$$