Practice midterm 2

Instructions for midterm 2:

1. Calculators will be provided in Midterm 2. You are encouraged to bring your own calculator.
2. It is closed-book and closed-note, but you are allowed to bring in two sides of a US Letter size paper with notes on it. Cell phones, mobile devices, tablets, laptops and other electronic devices with wireless capability must be turned off and put away.

1. See 16_Nov_2_Notes.pdf Pages 10–12. Read and interpret results from R lm().
   (a) How to calculate t value of a coefficient from Estimate and Std. Error?
   (b) What is "Residual standard error"?
   (c) How can we construct a two-sided t confidence interval based on the R output table?
   (d) What is the formula of Multiple R-squared and Adjusted R-squared?
   (e) For the F-statistic in the output, which null hypothesis it is constructed for?
2. How to detect outliers, high leverage, and influential points?
3. Read and interpret different diagnostic plots. See, e.g., Notes_Nov_14 Page 3.
4. How to identify potential issues from residual plots? See Notes_Nov_7 Page 2.
5. Differences between intervals of two types of predictions. What are the two types of predicted values? How are the variances derived? See Notes_Nov_2 Page 14.
6. How are t-tests and confidence intervals constructed? (Test a single coefficient β_i and a linear combination of multiple coefficients $a^\top \beta$.) See Notes_Oct_31.

Practice midterm 2

Instructions for midterm 2:

1. Calculators will be provided in Midterm 2. You are encouraged to bring your own calculator.

2. It is closed-book and closed-note, but you are allowed to bring in two sides of a US Letter size paper with notes on it. Cell phones, mobile devices, tablets, laptops and other electronic devices with wireless capability must be turned off and put away.

Sample question in discussion section last week.

1. See 16_Nov_2_Notes.pdf Pages 10–12. Read and interpret results from R lm().

   (a) How to calculate t value of a coefficient from Estimate and Std. Error?

   (b) What is "Residual standard error"?

   (c) How can we construct a two-sided t confidence interval based on the R output table?

   (d) What is the formula of Multiple R-squared and Adjusted R-squared?

   (e) For the F-statistic in the output, which null hypothesis it is constructed for?

2. How to detect outliers, high leverage, and influential points?

3. Read and interpret different diagnostic plots. See, e.g., Notes_Nov_14 Page 3. ✓

4. How to identify potential issues from residual plots? See Notes_Nov_7 Page 2.

5. Differences between intervals of two types of predictions. What are the two types of predicted values? How are the variances derived? See Notes_Nov_2 Page 14.

6. How are t-tests and confidence intervals constructed? (Test a single coefficient β_i and a linear combination of multiple coefficients $a^\top \beta$.) See Notes_Oct_31.

4. Hypothesis Testing

[Test 1] (1) Likelihood function

(2) MLE under full model and constraints $A\beta = c$

(3) Likelihood ratio test Λ $= \dfrac{ML\ (\text{full model})}{ML\ (\text{constrained model})}$

ML: Maximum likelihood $A\beta = c$

$$\Lambda = \frac{L(\hat\beta, \hat\sigma^2)}{L(\hat\beta_H, \hat\sigma_H^2)} = \left(\frac{\hat\sigma_H^2}{\hat\sigma^2}\right)^{\frac{n}{2}}$$

$$\Downarrow$$

(4) Pivotal statistic (Λ after monotone transformation)

$$\text{F-stat} = \frac{n-p}{q}\left(\frac{\hat\sigma_H^2}{\hat\sigma^2} - 1\right) \quad \text{(LRT after transformation)} \quad (1)$$

$$= \frac{(RSS_H - RSS)/q}{RSS/(n-p)} = \frac{(RSS_H - RSS)\ /\ (df.H - df.F)}{RSS\ /\ df.F} \quad (2)$$

$$\checkmark = \frac{(A\hat\beta - c)^T \{\mathrm{var}(A\hat\beta - c)\}^{-1}(A\hat\beta - c)/q}{\hat\sigma^2/\sigma^2} \quad \text{with } \hat\sigma^2 = \frac{RSS}{n-p} \quad (3)$$

✓ Quadratic

$E(\hat\sigma^2) = \sigma^2$

$\sim F_{q,n-p}$ under Ho (F-distribution with d.o.f q and $n-p$)

What is q . What is $n-p$?

Quadratic $f(x) = x^2$

$f(x) = x^T x$

$x = A\hat\beta - c$

$x^T W x$

$w = \mathrm{cov}(A\hat\beta - c)$

$A\beta - c = 0$

$x = 0 \Rightarrow x^T x = 0 \Rightarrow$

(5) Interpretations of F-test for a general linear hypothesis

(1) Pivotal statistic transformed from LRT

(2) Relative change of residual sum of squares (RSS)

(3) Quadratic form of $A\hat\beta - c$ (If $A\beta = c$, $A\hat\beta - c$ should be small.)

[Test 2] F-distribution

(1) Definition

(2) F - stat ~ $F_{q,\,n-p}$ under H_0 ⇒ p-value & test

(3) R: (1) Fit full model (2) Fit constrained model (3) anova

[Test 3] t-distribution (squared t ⇔ F-test with $q=1$) ✓

(1) Test one linear constraint: $H_0: a^T\beta - c = 0$ 1×1

☆ $$T_{stat} = \frac{(a^T\hat\beta - c)\,/\,\sqrt{var(a^T\hat\beta - c)}}{\hat\sigma/\sigma} = (a^T\hat\beta - c)\,/\,\sqrt{\widehat{var}(a^T\hat\beta - c)}$$

$\left[\widehat{var}(a^T\hat\beta - c) = \hat\sigma^2 a^T (X^TX)^{-1} a \quad \left(\text{replacing } \sigma \text{ in } var(a^T\hat\beta - c) \text{ by } \hat\sigma\right)\right]$

(2) Single coefficient $H_0: \beta_i = 0$

$$T_{stat} = \frac{\hat\beta_i}{S.E.(\hat\beta_i)}$$

R ourput from (ml·)

$\beta_i - c = 0$

(3) t - test ⇒ confidence intervals ✓

F - test ⇒ confidence regions ✓

[Test 4] R^2: multiple coefficient of determination

(1) 3 equivalent definitions

$\bar{Y}$ as the fitted values with only intercept no covariate

$$R^2 = \underset{\checkmark}{corr^2(\hat{Y}, Y)} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{TSS} = 1 - \frac{RSS}{TSS}$$

with $TSS = \underset{\checkmark}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$, $RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

(2) Decomposition of variances

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$$

$$SS_{Total} = SS_{Error} + SS_{Regression}$$

$$SST \qquad SSE \qquad SSR$$

$$\left(\text{Also } RSS = \sum_{i=1}^{n}\hat{\varepsilon}_i^2\right)$$

Then $R^2 = \dfrac{SSR}{SST} = 1 - \dfrac{SSE}{SST}$

$$\Downarrow \qquad\qquad \Downarrow$$

$$\star\ \frac{\text{variation in regression}}{\text{total variation}} \qquad \star\ 1-\frac{\text{variation in error ✓}}{\text{total variation ✓}}$$

(3) Adjusted R^2:

$$R^2_{Adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

⇒ RSS ful model

(with degree corrected)

$\dfrac{SS}{df} \rightarrow$ Mean SS

output in the lm( ) function

$E\left(\dfrac{SSE}{n-p}\right) = E\left(\dfrac{RSS}{n-p}\right) = \sigma^2$ ✓

[Test 5] Prediction intervals

$Y \sim 1+X$ SST SSE

(1) Confidence intervals for mean prediction. ✓

$Y \sim 1+X+Z$ SST change

(2) Prediction intervals for a future response. ✓

SSE ↓

$$V_p = V_c + \sigma^2$$

adding covariates introduce parameters

$n-2 \Rightarrow n-3$

5. Diagnostics and Remedies

[Diag 1] 1. Three assumptions on error terms (residual plots)

2. Unusual observations. outlier, leverage, influential ✓

(Definitions, detection methods)

3. Model structure nonlinear

[Diag 2]

1. Issues of predictors (multicollinearity, VIF)
2. Remedies of error assumptions (GLS and WLS)
3. Measurement errors