STP 530 Midterm Review Notes

Note: These review notes do not serve as the sole guideline for midterm exam topic coverage. Students should refer to all course materials to date to prepare for the exam.

1 Various models covered so far

- Simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

With the assumption ε i.i.d. $\sim \mathrm{N}(0, \sigma^2)$, we have

$$E\{Y_i\} = \beta_0 + \beta_1 X_i$$

The fitted model is expressed by

$$\hat{Y}_i = b_0 + b_1 X_i$$

- Multiple regression models (additive)

$$E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$$

2 Interpretation of model coefficients

For each model above, you should be able to precisely interpret each model coefficient within the context of the problem.

3 Matrix notation

All the above models can be expressed with a unified matrix notation framework.

- Matrix notation for all the above

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

where

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}_{n\times p} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}_{p\times 1} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

and n is the number of subjects in the dataset and p is the number of unknown parameters in the model.

4 Parameter estimation

Least square estimation: minimize the sum of squared errors (SSE).

$$\text{SSE} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Algebraic solution using calculus (demonstrated with simple linear regression)

$$\text{SSE} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Set

$$\frac{\partial \text{SSE}}{\partial \beta_0} = 0 \qquad \text{and} \qquad \frac{\partial \text{SSE}}{\partial \beta_1} = 0$$

Then solve for $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$b_1 = \frac{Cov\{X, Y\}}{Var\{X\}}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Linear algebra solution

In the linear algebra framework, SSE is equal to $||\boldsymbol{\varepsilon}||^2 = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2$.

Minimizing SSE is equivalent to minimizing the length of the $\boldsymbol{\varepsilon}$ vector, which is $||\boldsymbol{\varepsilon}||$. Therefore, $\hat{\mathbf{Y}}$ must be the *projection* of $\mathbf{Y}$ onto the space spanned by the columns of $\mathbf{X}$, resulting in the orthogonal relationship:

$$\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}.$$

Thus,

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{b}) = \mathbf{0}.$$

and

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

5 Inferences

The above solutions are only estimates of the true parameter values. These estimates vary from sample by sample, and we need to conduct statistical inferences to draw more informative conclusions.

Two major forms of statistical inferences:

- Hypothesis tests
  - The t-tests can be conducted to see if each individual slope is significantly different from 0 or another specific value. One-sided t-tests can also be conducted to see if a slope is significantly bigger or smaller than 0 or another specific value.
  - The F-test of global model utility) can be conducted to see if the whole model is useful at all.
  - The F-test of extra sum of squares can be conducted to see if the last predictor in the model should be kept or dropped from the model.
  - The F-test of the linear testing approach can be conducted to see if the full model fits the data significantly better than the reduced model.
- Confidence intervals
  - Confidence interval of the model coefficient estimates.
  - Confidence interval of the mean of Y distribution given X values.
  - Prediction interval of the Y value of an individual case given X values.

5.1 Assumption

Central to all inferences is the assumption on the error term ε, which is $\varepsilon \; iid \sim N(0, \sigma^2)$

Note: This applies to all ordinary linear regression models, including models with higher-order terms, interaction terms, and categorical predictors.

σ^2 is crucial for all inferences but is never known. It can be estimated by s^2:

$$s^2 = \frac{\text{SSE}}{n-p} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-p}$$

5.2 Confidence interval

The structure of confidence intervals:

$$\begin{aligned}\text{Confidence Interval} &= \text{Point Estimate} \pm \text{Margin of Error} \\ &= \text{Point Estimate} \pm \text{Distribution Multiplier} * \text{Standard error of estimate}\end{aligned}$$

Standard errors:

$$s\{b_1\} = s\sqrt{\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$s\{\hat{Y}_h\}_{CI} = s\sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$s\{\hat{Y}_h\}_{pred} = s\sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

With matrix notation:

$$\boldsymbol{s}^2\{\mathbf{b}\} = s^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{\mathbf{e}'\mathbf{e}}{n-p}(\mathbf{X}'\mathbf{X})^{-1} \qquad \text{(where } \underset{n\times 1}{\mathbf{e}} = [e_1, e_2, \cdots, e_n]' = \mathbf{Y} - \hat{\mathbf{Y}}.\text{)}$$

$$s^2\{\hat{Y}_h\}_{\text{CI}} = s^2\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h$$

$$s^2\{\hat{Y}_h\}_{\text{pred}} = s^2(1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$$

5.3 Hypothesis testing

The two-sided t-test of individual model coefficient:

1. Assumptions: ε i.i.d. $\sim \text{N}(0, \sigma^2)$
2. Hypotheses: $H_0 : \beta_k = 0 \qquad H_1 : \beta_k \neq 0$
3. Test-statistic:
$$t = \frac{b_k - 0}{s\{b_k\}}$$
4. P-value: The two-tail probability in the sampling distribution — If the true slope $\beta_k = 0$, what is the probability that we get a b_k this far from 0 or even more extreme?
5. Conclusion: Reject H_0 if p-value is less than a pre-determined significance level (typically 0.05) and conclude that the true population slope is not 0, in other words, there is linear association between X and Y.

The F-test of global model utility:

1. Assumptions: ε i.i.d. $\sim \mathrm{N}(0, \sigma^2)$
2. Hypotheses: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$
   H_1 : not all $\beta_k (k = 1, \cdots, p-1)$ equal zero
3. Test-statistic:
$$F = \frac{MSR}{MSE}$$
4. P-value: The right-tail probability in the F distribution.
5. Conclusion: Reject H_0 if p-value is less than a pre-determined significance level (typically 0.05) and conclude that not all population slopes in this model are 0, in other words, there is linear association between y and at least one predictor in the model.

The F-test of the linear testing approach for comparing two nested models:

- The **full model** (an example): $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- The **reduced (a.k.a., restricted) model** (examples):

$\beta_2 = 0$	which renders	$E\{Y\} = \beta_0 + \beta_1 X_1$
$\beta_1 = \beta_2 = 0$	which renders	$E\{Y\} = \beta_0$
$\beta_1 = \beta_2$	which renders	$E\{Y\} = \beta_0 + \beta_R (X_1 + X_2)$
$\beta_1 = 3$	which renders	$E\{Y\} - 3X_1 = \beta_0 + \beta_2 X_2$

1. Assumptions: ε i.i.d. $\sim \mathrm{N}(0, \sigma^2)$
2. Hypotheses: H_0 : The parameter constraints hold.
   H_1 : Some part of the parameter constraints does not hold.
3. Test-statistic:
$$F^* = \frac{\dfrac{SSE(R) - SSE(F)}{df_R - df_F}}{\dfrac{SSE(F)}{df_F}}$$
   Note: The degree of freedom of the SSE of either model is $(n - p)$, where p is the total number of estimated parameter in the respective model.
4. P-value: The right-tail probability in the F distribution $F(df_R - df_F, df_F)$
5. Conclusion: Reject H_0 if p-value is less than a pre-determined significance level (typically 0.05) and conclude that the full model fits the data significantly better than the reduced model.

6 ANOVA and R^2s

6.1 ANOVA

Total Sum of Squares: $SSTO = \sum(Y_i - \bar{Y})^2$

Error Sum of Squares: $SSE = \sum(Y_i - \hat{Y}_i)^2$

Regression Sum of Squares: $SSR = \sum(\hat{Y}_i - \bar{Y})^2$

When the regression model is estimated by the least square method,

$$SSTO = SSE + SSR$$

The corresponding partitioning of degrees of freedom is:

$$n - 1 = (n - p) + (p - 1)$$

6.2 R^2s

The **original** R^2:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Interpretation: About $R^2 100\%$ of *variation* in Y is explained by the model..

Limitation of R^2: Adding more X variables to the regression model will always increase R^2 value (or hold it equal, theoretically speaking). So if you build a model based on R^2 you will always end up with the most complex model.

The **adjusted** R^2 strikes a balance between prediction accuracy and model parsimony:

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}} = 1 - \frac{MSE}{MSTO}$$

Interpretation: About $R^2 100\%$ of *variance* in Y is explained by the model..

The **Coefficient of partial determination** measures the proportionate reduction in the variation in Y remaining after a predictor is added to an existing model.

$$R^2_{Y\,2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)}$$

Interpretation: By adding X_2 to the model where X_1 already exists, the variation in Y unexplained by the model reduces by $R^2_{Y\,2|1} 100\%$.

7 Diagnostics

- (Not Tested) Detect multicollinearity. Examine correlation matrix, scatterplot matrix, VIF values. Remove some of the predictors so that the remaining predictors are not linearly dependent. For polynomial models, code the independent variables.
- Detect lack of linear fit. Plot residuals against each X and against $\hat{Y}$. Add polynomial terms if non-linear trends are found.
- Detect unequal error variances (heteroscedasticity). Plot residuals against $\hat{Y}$. Transform Y if heteroscedasticity pattern is strong.
- Check the normal distribution of the residuals. Examine the Q-Q plot of the residuals. Transform Y if the points severely depart from the reference line.
- Identify outliers. Examine standardized residuals. Values beyond -3 or 3 are deemed large residuals (outliers).