Summary of Topics for Midterm Exam #2
STA 371G, Fall 2017

Listed below are the major topics covered in class that are likely to be in Midterm Exam #2:

- Mean (expectation), variance and standard deviation of a discrete random variable.

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i P(X = x_i), \quad \text{Var}[X] = \sum_{i=1}^{n} (x_i - \mathbb{E}[X])^2 P(X = x_i), \quad \text{sd}[X] = \sqrt{\text{Var}[X]}$$

- Normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$, where μ is the mean, σ^2 is the variance, and σ is the standard deviation.
  - Probability density function: area under the curve represents probability.
  - Standard normal distribution $Z \sim \mathcal{N}(0, 1)$.
  - Standardizing a normal random variable $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.
  - $P(X < x) = P(\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}) = P(Z < \frac{x-\mu}{\sigma})$.
  - $P(-1 < Z < 1) \approx 0.68$; $P(\mu - \sigma < X < \mu + \sigma) \approx 0.68$.
  - $P(-2 < Z < 2) \approx 0.95$; $P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95$.

- Simple Linear Regression
  - Least squares estimation: given n observations $(x_1, y_1), \cdots, (x_n, y_n)$, we estimate the intercept b_0 and slope b_1 by finding a straight line $\hat{y}_i = b_0 + b_1 x_i$ that minimizes the sum of squared residuals (SSE)

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_i)\right]^2.$$

  - Sample means of X and Y

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

  - Sample covariance

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

  - Sample correlation

$$r_{xy} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{s_x^2 s_y^2}} = \frac{\text{Cov}(X, Y)}{s_x s_y}.$$

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}, \quad s_x = \sqrt{s_x^2}$$

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}, \quad s_y = \sqrt{s_y^2}$$

- Interpreting covariance, correlation and regression coefficients.
- SST = SSR + SSE
- Coefficient of determination:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = r_{xy}^2$$

- Regression assumptions and statistical model.

$$Y = \beta_0 + \beta_1 X + \epsilon,\ \ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \ \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

  Assuming β_0, β_1 and σ^2 are known, given x_i, the 95% prediction interval of y_i is

$$(\beta_0 + \beta_1 x_i) \pm 2\sigma.$$

- We estimate σ with the regression standard error s as

$$s = \sqrt{\frac{\sum_{i=1}^n e^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}.$$

- Approximately we have $b_1 \sim \mathcal{N}(\beta_1, s_{b_1}^2)$ and $b_0 \sim \mathcal{N}(\beta_0, s_{b_0}^2)$, where the standard errors of b_1 and b_0 are

$$s_{b_1} = \sqrt{\frac{s^2}{(n-1)s_x^2}}, \quad s_{b_0} = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2}\right)}.$$

  Thus approximately we have the 95% confidence intervals for β_1 and β_0 as

$$b_1 \pm 2s_{b_1}, \quad b_0 \pm 2s_{b_0}.$$

- Hypothesis testing:
  - We test the null hypothesis $H_0 : \beta_1 = \beta_1^0$ versus the alternative $H_1 : \beta_1 \neq \beta_1^0$.
  - The t-stat $t = \frac{b_1 - \beta_1^0}{s_{b_1}}$ measures the number of standard errors the estimate b_1 is from the proposed value β_1^0.
  - The p-value provides a measure of how weird your estimate b_1 is if the null hypothesis is true.
  - We usually reject the null hypothesis if $|t| > 2$, $p < 0.05$, or β_1^0 is not within the 95% confidence interval $(b_1 - 2s_{b_1}, b_1 + 2s_{b_1})$.
- Forecasting:
  - Given X_f, the 95% plug-in prediction interval of Y_f is $(b_0 + b_1 X_f) \pm 2s$.
  - A large predictive error variance (high uncertainty) comes from a large s, a small n, a small s_x and a large difference between X_f and $\bar{X}$.

- Multiple Linear Regression
  - Statistical model:

    $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

    $$Y \mid X_1 \ldots X_p \sim \mathcal{N}(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \sigma^2)$$

  - Interpretation of regression coefficients.
  - Fitted values: $\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}$
  - Least squares estimation: find b_0, b_1, $\cdots$, b_p that minimize the sum of squared residuals $\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$.
  - Regression standard error:

    $$s = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-p-1}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p-1}}.$$

  - $\bar{e} = 0$, $\mathrm{Corr}(X_j, e) = 0$, $\mathrm{Corr}(\hat{Y}, e) = 0$
  - $R^2 = \left(\mathrm{Corr}(Y, \hat{Y})\right)^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
  - Approximately we have $b_j \sim \mathcal{N}(\beta_j, s_{b_j}^2)$.
    - ∗ 95% confidence interval for β_j: $b_j \pm 2s_{b_j}$
    - ∗ t-stat: $t_j = \frac{b_j - \beta_j^0}{s_{b_j}}$.
    - ∗ $H_0 : \beta_j = \beta_j^0$ versus $H_1 : \beta_j \neq \beta_j^0$. Reject H_0 if $|t_j| > 2$, p-value < 0.05, or β_j^0 is not within $(b_j - 2s_{b_j}, b_j + 2s_{b_j})$
  - F-test of overall significance.
    - ∗ $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ versus H_1: at least one $\beta_j \neq 0$.
    - ∗ $f = \frac{R^2/p}{(1-R^2)/(n-p-1)} = \frac{SSR/p}{SSE/(n-p-1)}$
    - ∗ If H_0 is true, then $f > 4$ is very significant in general.
    - ∗ If f is large (the corresponding p-value is small), we reject H_0.
  - Understanding multiple linear regression
    - ∗ Correlation is not causation
    - ∗ Multiple linear regression allows us to control all important variables by including them into the regression model
    - ∗ Dependencies between the explanatory variables (X's) will affect our interpretation of regression coefficients
    - ∗ Dependencies between the explanatory variables (X's) will inflate the standard errors of regression coefficients

    $$s_{b_j}^2 = \frac{s^2}{\text{variation in } X_j \text{ not associated with other } X\text{'s}}$$

- Dummy Variables and Interactions
  - Dummy variables
    - Gender: Male, Female; Education level: High-school, Bachelor, Master, Doctor; Month: Jan, Feb, $\cdots$, Dec
    - A variable of n categories can be included into multiple linear regression using C dummy variables, where $1 \leq C \leq n-1$
    - Representing a variable of n categories with n dummy variables will lead to the problem of "perfect multicollinearity"
    - Interpretation: the same slope but different intercepts
  - Interactions
    - Interpretation: different intercepts and slopes
- Diagnostics and Transformations
  - Diagnostics
    - Model assumptions:
      - Statistical model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

      - The mean of Y is a linear combination of the X's
      - The errors ϵ_i (deviations from the true mean) are independent, and identically normally distributed as $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
    - Understanding the consequences of violating the model assumptions
    - Detecting and explaining common model assumption violations using the residual plots.
  - Modeling non-linearity with polynomial regression
    - Statistical model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_m X^m + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

    - We can always increase m if necessary, but $m = 2$ is usually sufficient.
    - Be very careful about over-fitting and doing prediction outside the data range, especially if m is large.
    - For $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$, the marginal effect of X on Y is

$$\frac{\partial \mathbb{E}[Y|X]}{\partial X} = \beta_1 + 2\beta_2 X,$$

      which means the slope is a function of X (no longer a constant).
  - Handling non-constant variance with Log-Log transformation
    - Statistical model:

$$\log(Y) = \beta_0 + \beta_1 \log(X) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$Y = e^{\beta_0} X^{\beta_1} e^{\epsilon}, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

    - Interpretation: about β_1% change in Y per 1% change in X.
    - Example: price elasticity
    - 95% plug-in prediction interval of $\log(Y)$

$$(\beta_0 + \beta_1 \log(X)) \pm 2s$$

    - 95% plug-in prediction interval of Y

$$\left(e^{\beta_0+\beta_1 \log(X)-2s},\ e^{\beta_0+\beta_1 \log(X)+2s}\right) = \left(e^{\beta_0-2s}X^{\beta_1},\ e^{\beta_0+2s}X^{\beta_1}\right)$$

  - Log transformation of Y
    - Statistical model:

$$\log(Y) = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$Y = e^{\beta_0} e^{\beta_1 X} e^{\epsilon}, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

    - Interpretation: about $(100\beta_1)$% change in Y per unit change in X (if β_1 is small).
    - Example: exponential growth

- Time Series
  - Trend, seasonal, cyclical, and random components of a time series
  - Fitting a trend
    - Linear trend:

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

    - Exponential trend:
      - Model: $\log(Y_t) = \beta_0 + \beta_1 t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$
      - Interpretation: Y_t increases by about $(100\beta_1)$% per unit time increase.
    - Modeling non-linearity by adding t^2 into the regression model: the slope changes as time changes.
    - 95% plug-in prediction interval
  - Autoregressive models
    - Random walk model: $Y_t = \beta_0 + Y_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$
    - Autoregressive model of order 1 (AR(1)):

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

    - Autocorrelation of residuals: $\mathrm{Corr}(e_t, e_{t-1})$
    - Trend+AR(1):

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

    - Logtransformation + trend + AR(1):

$$\log(Y_t) = \beta_0 + \beta_1 \log(Y_{t-1}) + \beta_2 t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

- Modeling seasonality
  - Using no more than 11 dummy variables for 12 months; using no more than 3 dummy variables for 4 quarters
  - Seasonal model:

    $$Y_t = \beta_0 + \beta_1 Jan + \cdots + \beta_{11} Nov + \epsilon_t$$

  - Seasonal + AR(1) + linear trend:

    $$Y_t = \beta_0 + \beta_1 Jan + \cdots + \beta_{11} Nov + \beta_{12} Y_{t-1} + \beta_{13} t + \epsilon_t$$

    - Model for t in December: $Y_t = \beta_0 + \beta_{12} Y_{t-1} + \beta_{13} t + \epsilon_t$
    - Model for t in Jan: $Y_t = (\beta_0 + \beta_1) + \beta_{12} Y_{t-1} + \beta_{13} t + \epsilon_t$
    - Model for t in October: $Y_t = (\beta_0 + \beta_{10}) + \beta_{12} Y_{t-1} + \beta_{13} t + \epsilon_t$
  - Logtransformation + Seasonal + AR(1) + trend

    $$\log(Y_t) = \beta_0 + \beta_1 Jan + \cdots + \beta_{11} Nov + \beta_{12} \log(Y_{t-1}) + \beta_{13} t + \epsilon_t$$

- Diagnose the residual plot of a time series regression model:
  - Are there any clear temporal patterns?
  - Are the residuals autocorrelated?
  - What kind of model assumptions have been violated?
- Understand when and how to include log transformation, non-linearity, dummy variables, interactions, AR(1) and trend to improve a time series regression model.