SD 7840: Applied Regression Analysis

Darshan Patel

Spring 2019

Contents

1 Introduction **2**

2 Simple Linear Regression **3**

3 Multiple Regression **10**

4 Qualitative Variables and Interaction Terms **15**

5 Variable Selection Methods **16**

6 Regression Diagnostics **22**

7 Multicollinearity and Autocorrelation **25**

8 Transformations **30**

9 Logistic Regression **32**

10 Multiple Logistic Regression **40**

11 Experimental Design and One-Way ANOVA **45**

1 Introduction

- A variable describes a characteristic of an individual/object
- The distribution of a variable describes what values the variables takes and how often it takes these values in the population
- Four Regression Models

model	response variable	explanatory variable(s)
simple linear regression	numerical	1 numerical variable
multiple regression	numerical	numerical and/or categorical variables
logistic regression	categorical variables with two levels	numerical and/or categorical variables
one-way ANOVA	numerical	1 categorical variable

- Univariate measures are means (measure of center), standard deviation (measure of spread) as well as visual description of histograms
- One bivariate measure is correlation which measures the strength of the linear relationship between two numerical variables
- Pearson's Correlation
  - Notation: population correlation ρ, sample correlation r
  - To estimate ρ using data:

$$r = \text{Cor}[X, Y] = \text{Cor}[Y, X] = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

    where $\bar{x}$ and $\bar{y}$ are the sample means and s_x and s_y are sample standard deviations
  - Correlation measures the strength of a linear relationship - if the relationship between X and Y cannot be described by a line, then correlation is meaningless
  - Correlation does not imply causation - if X and Y have a high correlation, it does not tell anything about whether X causes Y or whether Y causes X
  - Pearson's correlation is bounded on $[-1, 1]$
  - If X and Y have a linear relationships and as X increases, Y increases, then there is a positive trend and a positive r value
  - If X increases and Y decreases, then there is a negative trend and a negative r value
  - If there is no trend, $r \approx 0$
  - Note that r has no units

- Simple Linear Regression Model

$$Y = \text{y-intercept} + \text{slope}X + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$$

- Multiple Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

- Logistic Regression

  Let $Y = \begin{cases} 0 \\ 1 \end{cases}$; then

$$\mathbb{P}(Y = 1) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

  where log odds is

$$\log\left(\frac{\mathbb{P}(Y = 1)}{1 - \mathbb{P}(Y = 1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- One-Way ANOVA

$$y_{ij} = \mu_j + \varepsilon_{ij}$$

  where y_{ij} is the response for the ith observation in group j and μ_j is the mean of the jth group

2 Simple Linear Regression

- Population Correlation ρ is estimated by the sample correlation r

$$r = \text{Cov}[X, Y] = \text{Cov}[Y, X] = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

  where $\bar{x}$ and $\bar{y}$ are the sample means and s_x and s_y are sample standard deviations

- Note that r is bounded on $[-1, 1]$

- Correlation measures the strength of a linear association
  - If the relationship between X and Y cannot be described by a line, then correlation is meaningless
  - Check a scatterplot to see if X and Y have a linear relationship before using the statistic r

- Correlation does not imply causation; if X and Y have a high correlation, it does not say anything about whether X causes Y or whether Y causes X

- Population Model

$$Y = \text{E}[Y \mid X] = \beta_0 + \beta_1 X + \varepsilon$$

  Here x is the explanatory/independent variable whereas Y is the response/dependent variable

- Estimated Model
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
for $i = 1, 2, \ldots, n$

- Error ε is estimated by the residual, e
$$e_i = \text{observed } y - \text{ predicted } y = y_i - \hat{y}_i$$

- The model is fit by determining which values of $\hat{\beta}_0$ and $\hat{\beta}_1$ create the smallest SSE (sum of squared errors) and where the sum of the residuals is equal to 0
$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \Big)^2$$
This method is called "Ordinary Least Squares" or OLS

- Using the OLS method,
$$\hat{\beta}_1 = r \frac{s_y}{s_x}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
where s_x, s_y are the sample standard deviations of the x and y variables respectively, $\bar{x}$ and $\bar{y}$ are the corresponding sample means and r is the sample correlation (note that correlation, r has no units)

- The estimated slope is $\hat{\beta}_1$
  - One unit increase in x is associated with a $\hat{\beta}_1$ change in y
  - Measurement units of slope = units of y / units of x

- The estimated y-intercept is $\hat{\beta}_0$
  - This is the y-value when $x = 0$ (i.e., where the line crosses the y-axis)
  - Measurement units of y-intercept = units of y
  - The y-intercept may not always be meaningful; look at contextual interpretation, is 0 within the range of observed x-values?

- The points that are always on the regression line is $(\bar{x}, \bar{y})$ and $(0, \hat{\beta}_0)$

- R^2 is the coefficient of determination, which can be denoted as r^2 for simple linear regression

- The standard deviation of y is
$$s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\text{SST}}{n-1}}$$
where SST is the total variation for the variable y

- Note that the total variation of the residuals, e, is SSE

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\Big)^2$$

- Then the coefficient of determination can be rewritten as

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

  where $\frac{\text{SSE}}{\text{SST}}$ is the fraction of variation in y that is not explained by x using the regression model

- Therefore, R^2 is the fraction of variability in y can can be explained by x using the regression model

- Note that R^2 is bounded on $[0, 1]$
  - If $R^2 = 0$, SSE = SST and so, x is useless in explaining y, all $\hat{y} = \bar{y}$
  - If $R^2 = 1$, SSE = 0 and so, x explains y perfectly, all residuals are 0

- Regression Modeling Steps
  1. Hypothesize model form to explain y
  2. Collect sample data
  3. Clean data and do explanatory data analysis
  4. Use sample data to estimate the unknown model parameter (e.g., $\beta_0, \beta_1, \ldots, \beta_p$)
  5. Specify the probability distribution of the error term and estimate any unknown parameters of this distribution (most often, $\varepsilon \sim N(0, \sigma^2)$ is assumed to estimate σ^2); also, check the validity of each assumption made about this probability distribution and the model
  6. Statistically check the usefulness of the model
  7. When satisfied that the model is useful, use it for prediction, estimation, etc.

- Recall:
  - True Model: $Y = \beta_0 + \beta_1 X + \varepsilon$
  - Estimated Model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
  - Slope Estimate: $\hat{\beta}_1 = r\frac{s_y}{s_x}$
  - y-intercept Estimate: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_i$
  - Residual (estimated error): $e_i = y_i - \hat{y}_i$

- Assumptions for Simple Linear Regression: we want to estimate the following model based on the OLS method:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

  - x values are fixed and are measured without error
  - x and y are linearly related
  - errors are independent of each other
  - the points are evenly distributed above/below the regression line
    * that is, σ^2 is a good estimate of he variability of Y around the regression line for all values of X
    * constant variance assumption, or homoscedasticity
  - errors have a mean of 0 (OLS ensures this)
  - the errors are normally distributed

- Essentially, $\varepsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$
- Residual Plot
  - Check for linearity and constant variance using a residual plot by plotting e_i values against x_i values
  - Note that the OLS method ensures the mean of the residuals is 0
  - Look out for
    * pattern $\rightarrow$ linearity assumption violated
    * increasing/decreasing spread of points as x increases $\rightarrow$ constant variance assumption violated
    * individual points with large residuals, often are outliers, influential points or leverage points
  - If there is a pattern in the plot, then either the linearity or constant variance assumptions have been violated
- If the regression assumptions are satisfied, given a specific value for x, y is normally distributed with mean $\beta_0 + \beta_1 x$ (this implies $\mathrm{E}[\varepsilon] = 0$) and standard deviation σ; this distribution should be applicable to the whole regression line
- Normal Quantile Plots: ε has a normal distribution
  - Assumption Check: evaluate whether ε has a normal distribution by looking at a normal quantile plot of the residuals, $e \rightarrow$ points should form a straight line
  - Intuitively, the theoretical quantiles are on the x-axis while the empirical (sample) quantiles are on the y-axis
  - Interpretation: if the points on the plot form a straight line, the data may come from a normal distribution
    * A left-skewed distribution has a downward facing distribution for the quantile plot
    * A right-skewed distribution has an upward facing distribution for the quantile plot

    - Heavy-tailed distributions: more probabilities at the tails (extremes) than normal distribution (some examples include t, Cauchy, Pareto)

- For the normality assumption to be satisfied, the points should form a straight line in the normal quantile plot of the residuals
- When the variance is not constant, $\hat{\sigma}$ does not describe the variability of the residuals at all values of x (high/low) and so the regression assumption is violated
- Estimating and Interpreting σ^2
  - Regression Assumption: $\varepsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$
  - Estimate of σ denoted as $\hat{\sigma}$:

$$\hat{\sigma} = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{n-2}}$$

  - $\hat{\sigma}$ is also called the RMSE - room mean squared error
  - General Interpretation: since ε is assumed to be normally distributed, we expect approximately 95% of the observed Y values to be within $\pm 2\hat{\sigma}$ of the regression line
  - Units of $\hat{\sigma}$ is the units of the Y variable
- If constant variance and normality assumptions hold, ± 2 RMSE above and below the regression line should contain roughly 95% of the observations
- Approximating RMSE ($\hat{\sigma}$)
  - $\hat{\sigma}$ is the standard deviation of y around the regression line
  - s_y is the standard deviation of y around the sample means
  - For large sample sizes,

$$\hat{\sigma} \approx s_y \sqrt{1 - r^2}$$

  - Properties: $s_y \geq 0$, $-1 \leq r \leq 1$, $0 \leq r^2 \leq 1$
  - If $r^2 = 1$, then $\hat{\sigma} = 0$ and all variation in y is explained by x
  - if $r^2 = 0$, then $\hat{\sigma} = 1$ and none of the variation in y is explained by x
- Sampling Distribution of $\hat{\beta}_1$
  - Assumption: $\varepsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$
  - $\hat{\beta}_1$ is an estimate of β_1 and is a statistic
  - $\hat{\beta}_1$ is a random variable with a normal distribution of mean β and standard deviation

$$\text{SD}[\hat{\beta}_1] = \frac{\sigma}{\sqrt{\sum (x - \bar{x})^2}}$$

    where σ is the same standard deviation of the errors from $\varepsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$

  - $\sum(x-\bar{x})^2$ is the total variation in x

$$s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

    This is called the sampling distribution of $\hat{\beta}_1$
  - $\hat{\beta}_1 = r\frac{s_y}{s_x}$ is an unbiased estimator of β_1 ($\mathrm{E}[\hat{\beta}_1] = \beta_1$)
  - In practice, σ is unknown and must be estimated by $\hat{\sigma}$, or RMSE
  - Estimate $\mathrm{SD}[\hat{\beta}_1]$ with the standard error of $\hat{\beta}_1$

$$\mathrm{SE}[\hat{\beta}_1] = \frac{\hat{\sigma}}{\sqrt{\sum(x-\bar{x})^2}}$$

  - Therefore when running hypothesis tests for β_1, use the t-distribution instead of the normal distribution since $\mathrm{SD}[\hat{\beta}_1]$ is estimated by $\mathrm{SE}[\hat{\beta}_1]$
  - This theory can also be used to run hypothesis tests using the slope and constructing confidence intervals for the slope

- A small RMSE and large $s_x\sqrt{n-1} = \sqrt{\sum(x-\bar{x})^2}$ is good
- t-Test for the Population Slope, β_1
  - Hypotheses:

$$H_0 : \beta_1 = \beta_1^0$$
$$H_1 : \beta_1 \neq \beta_1^0,\ \beta_1 < \beta_1^0, \text{ or } \beta_1 > \beta_1^0$$

  - Choose significance level α
  - Collect data and check that the regression assumptions hold
  - Compute the test statistic assuming the null hypothesis is true

$$t_1 = \frac{\hat{\beta}_1 - \beta_1^0}{\mathrm{SE}[\hat{\beta}_1]}$$

    which estimates how many standard units $\hat{\beta}_1$ is from the hypothesized β_1^0 and $SE[\hat{\beta}_1]$ is the standard deviation of slope estimate
  - t_1 has a $t(n-2)$ distribution
  - Compute p-value
  - Reject H_0 only if p-value$< \alpha$
- Confidence Interval for Population Slope
  - For a $(1-\alpha)\times 100\%$ confidence interval for the population slope

$$\underbrace{\hat{\beta}_1}_{\text{point estimate}} \pm \underbrace{t_{(n-2,\alpha/2)} \times \mathrm{SE}[\hat{\beta}_1]}_{\text{margin of error}}$$

  - $t_{(n-2,\alpha/2)}$ is the critical value, the value on the t-distribution with $n-2$ degrees of freedom where the cumulative distribution value is $1-\frac{\alpha}{2}$

- In-sample prediction: obtain predicted values from observations used to fit the model
- Interpolation: prediction of y from an x which is in the range of the data
- Extrapolation: prediction of y from an x which is outside the range of the data
- Extrapolating is dangerous because it is unknown whether the linear relationship holds in regions of no data
- When interpreting the y-intercept, check if you are interpolating/extrapolating
- Out-of-sample prediction: predictions for observations not used to fit the model (i.e., new x values)
- Average vs. Individual Responses
  - Example: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ where $\hat{y}$ is a random quantity
  - $\hat{y}$ is the estimate of the average value for y given the value of x_0
    - can construct a confidence interval for the average estimate when $x = x_0$
    - Interpretation: there is a $C\%$ chance that the confidence interval for $\hat{y}$ covers the true average y given that particular value of x_0
  - For a specific estimate of y given an x_0, expect more uncertainty a wider interval
    - can construct a prediction interval for a single observation when $x = x_0$
    - Interpretation: there is a $C\%$ chance that the prediction interval for $\hat{y}$ covers the true value of y given that particular value of x_0
  - Since the averages are less variable than single data points, confidence intervals are narrower than prediction intervals
  - Visualizing Confidence and Prediction Intervals for Prediction
    - Confidence Interval for prediction: examining average value for y given x variable values
    - Prediction Interval for prediction: examining values for a new/single y given x variable values
    - Averages are less variable than single data points and so confidence intervals are narrower than prediction intervals
    - In-sample predictive performance is almost always better than out-of-sample performance
    - When reporting modeling results, always include the following:
      - data cleaning steps
      - list all variables in the data set, not just the ones in the model(s)
      - summary statistics of variables in the model

  - which (if any) observations were removed and why
  - model fitting process
  - final model, assumption checks, hypothesis test (for hypothesis tests, always include the following: hypotheses, test statistic, degrees of freedom (if applicable), p-value, and test conclusion)

- Summary: Regression Diagnostics for Simple Linear Regression
  - x's are fixed and measured without error - how is data explored and processed?
  - x and y are linearly related - residuals vs. x variable plot $\rightarrow$ check for trends/curves
  - constant variance (homoscedasticity) - residuals vs. x variable plot $\rightarrow$ check for fanning, different variability of residuals for various x values
  - the errors are normally distributed - check normal quantile plot of residuals
  - errors are independent of each other - data collected sequentially? data collected in groups?
- Violations o assumptions can affect standard error estimates, test statistics, p-values, confidence intervals, prediction intervals, etc
- Regression is fairly robust to violations of the constant variance and normality assumptions, i.e., can still proceed with inferential statistics with moderate violations of assumptions; regression is not robust to violations of other assumptions
- What to do with violations of assumptions? Potential actions:
  - transform
  - use weighted least squares, generalized least squares
  - instrumental variables to correct for measurement errors
  - other types of models (e.g., Poisson regression, nonlinear models, etc.)

3 Multiple Regression

- In simple linear regression, estimate $\beta_0, \beta_1, \sigma^2$ using $Y = \beta_0 + \beta_1 X + \varepsilon$
- In multiple regression, estimate $\beta_0, \beta_1, \beta_2, \ldots, \beta_p, \sigma^2$, where p is the number of explanatory variable, using

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- Estimate the coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ values such that the SSE is minimized

$$\begin{aligned} \text{SSE} &= \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 \\ &= \sum_{j=1}^{n} (y_j - (\hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + \cdots + \hat{\beta}_p x_{pj}))^2 \end{aligned}$$

- Least Squares Method
  - The population model in matrix notation is $y = X\beta + \varepsilon$ where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix}$$

  - X is called a design matrix and has $n \times (p+1)$ dimensions
  - Then compute

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

    where X^T is the transpose of X

- Model Assumptions: let the population model be $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$ and the sample (estimated) model be $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$
  - x values are fixed and are measured without error
  - variables are linearly related
  - $\mathrm{E}[\varepsilon] = 0$ and so $\mathrm{E}[Y \mid X_1, \dots, X_p] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$, i.e. the mean of the error is 0 and so the mean of the Y given the $X_1, \dots, X_p$ values is its value on the population regression "line"
  - $\mathrm{Var}[\varepsilon] = \sigma^2$
  - ε is normally distributed
  - ε values are independent, i.e. the error from one value of Y is not dependent on the error from any other value of Y
  - Extra Check: X variables are not too highly correlated (collinearity/multicollinearity)
- Model Evaluation: given that the model assumptions are satisfied, how do you determine how good the model is?
  - Estimate variability of response variable around the regression line using RMSE
  - Compute fraction of variability in response variable which can be explained by the regression model using R^2 or adjusted R^2
  - Test all slopes at once using overall F-test
  - Test each slope separately using individual t-tests
  - Compare nested models using partial F-tests

- Let the variance of Y, or average variability of Y be defined as

$$s_y^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

- Then the total variability of Y is

$$\text{SST} = \sum(y_i - \bar{y})^2$$

- The goal of regression is to explain some of this variability in Y using the explanatory variables

- The SST can be rewritten as follows:

$$\text{SST} = \text{SSR} + \text{SSE}$$
$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  where SSR is the regression sum of squares (explained variation) and SSE is the error sum of squares (unexplained variation)

- Then s_y^2 becomes

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\text{SST}}{n-1}$$

- The error variance estimate is given by

$$\hat{\sigma} = \text{RMSE} = \sqrt{\frac{\text{SSE}}{n-(p+1)}}$$

- If regression assumptions hold, expect about 95% of the Y values to be within ± 2 RMSE values from the regression line

- Compute R^2 as follows:

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

  Same interpretation as before: $(100 \times R^2)\%$ of the variation in Y can be explained by the X's (i.e., the regression model)

- Every time a variable is added to the regression, the R^2 value will stay the same or increase, even if the variable is not helpful

- R^2_{adjust} is more conservative than R^2, essentially adjusting for more variables vs. sample size

- The formula for the adjust multiple coefficient of determination is as follows

$$\begin{aligned} R^2_{\text{adjusted}} &= 1 - \left(\frac{n-1}{n-(p+1)}\right)\left(\frac{\text{SSE}}{\text{SST}}\right) \\ &= 1 - \left(\frac{n-1}{n-p-1}\right)\left(1-R^2\right) \end{aligned}$$

- Note that $R^2_{\text{adjusted}} \leq R^2$ and that for every poor models, R^2_{adjusted} could be negative; if sample size n is not much bigger than $p+1$, then R^2 will not be very informative
- Overall F-Test
  - Hypotheses:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ H_A &: \text{ at least one } \beta_i \text{ is not zero} \end{aligned}$$

    Note that the y-intercept is not included
  - The significance level is α
  - Collect sample of size n and compute SST and SSE
  - The test statistic is as follows:

$$F_c = \frac{(\text{SST}-\text{SSE})/p}{\text{SSE}/(n-(p+1))} = \frac{\text{SSR}/p}{\text{SSE}/(n-(p+1))} = \frac{\text{MSR}}{\text{MSE}}$$

    which has a F distribution with p and $n-(p+1)$ degrees of freedom
  - Reject H_0 if p-value $= P(F > F_c) < \alpha$
- ANOVA (Analysis of Variance) Table

source	degrees of freedom	sums of squares	mean square (variance)	F
regression model	p	SSR	$\text{MSR} = \frac{\text{SSR}}{p}$	$F_C = \frac{\text{MSR}}{\text{MSE}}$
error	$n-p-1$	SSE	$\text{MSE} = \frac{\text{SSE}}{n-p-1}$	
total	$n-1$	SST		

- F Distribution
  - continuous, positive distribution with density function:

$$f(x) = \frac{\sqrt{\frac{(jx)^j k^k}{(jx+k)^{j+k}}}}{x\text{B}\left(\frac{j}{2},\frac{k}{2}\right)}$$

    where $j > 0$, $k > 0$ and B is the beta function
  - useful in regression for analysis of variance (ANOVA)

- In simple linear regression only, the two-sided t-test for slope, the overall F-test and the two-sided t-test for population correlation ρ are equivalent hypothesis tests
- t-test for β_i
  - The significance level is α
  - Hypotheses:
$$H_0 : \beta_i = \beta_i^*$$
$$H_A : \beta_i \neq \beta_i^* \text{ or } \beta_i < \beta_i^* \text{ or } \beta_i > \beta_i^*$$
  - Collect sample of size n and compute $\hat{\beta}_i$ and the standard error of $\hat{\beta}_i$, $\text{SE}[\hat{\beta}_i]$
  - Compute the test statistic:
$$t_i = \frac{\hat{\beta}_i - \beta_i^*}{\text{SE}[\hat{\beta}_i]}$$
    which has a t distribution with $n - (p + 1)$ degrees of freedom
  - Compute the p-value and reject H_0 if p-value $< \alpha$
  - The confidence interval is
$$\hat{\beta}_i \pm t_{\alpha/2}\text{SE}[\hat{\beta}_i]$$
  - Note: t-tests in multiple regression are testing a slope assuming ALL of the other variables are already in the model
- Models can be compared formally if they are nested, meaning that the explanatory variables in one model are a subset of the explanatory variables in the other model, using a partial F-test (note: this will only work if the same observations are used in both models)
- Partial F-test
  - Models
    $\text{E}[Y \mid X] = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q$ (reduced model, RM)
    $\text{E}[Y \mid X] = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + \beta_{q+1} X_{q+1} + \cdots + \beta_p X_p$ (full model, FM)
  - The significance level is α
  - Hypotheses (testing $p - q$ slopes):
$$H_0 : \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$$
$$H_A : \text{ at least one slope is not zero}$$
  - Collect sample of size n (same observations for both models)
  - Compute the test statistic:
$$F_c = \frac{(\text{SSE}_{\text{RM}} - \text{SSE}_{\text{FM}})/(p - q)}{\text{SSE}_{\text{FM}}/(n - (p + 1))}$$
    which has a F distribution with $p - q$ and $n - (p + 1)$ degrees of freedom and where RM denotes the reduced model and FM denotes the full model

  - Reject H_0 if p-value $= P(F > F_c) < \alpha$

- Note that when explanatory variables are arranged differently in linear models and its output is compared with each other, the coefficients will be the same; in the ANOVA test for both, the SSE, SSR will be the same for both models and so the overall F-test statistic will be the same for all models, however the individual sums of squares for individual variables will differ because of how they are calculated (sequential sums of squares)

- Predicting "in the range of the xs" becomes more complicated in multiple regression; make sure that all predictions are in the correct range for all xs simultaneously and is feasible

- Confidence and prediction intervals have the same interpretations as in simple linear regression; as in simple linear regression, averages are less variable than single data points and so prediction intervals are always wider than confidence intervals

- When two models have similar predictive abilities, choose the model with the fewer number of explanatory variables or remove variables which are not statistically significant (i.e., t-test for slope is not statistically significant)

- n vs. p

  - Let n be the sample size (i.e., number of usable observations in the data set), p be the number of explanatory variables available in the data set and k be the number of explanatory variables in the final model
  - Note that $k \leq p$
  - Ideally, $n >> p$, or, n is much larger than p (and therefore k)
  - This means there is a lot of data to estimate each of the slope coefficients
  - The final model must have $n > k$, otherwise regression model cannot be fit
  - Potential issues: $n >> p$ but p is quite large, or $p > n$; in both cases, it is difficult to sift through so many explanatory variables

- How to choose among large p?

  - Use mechanical methods (e.g., stepwise regression) to choose k variables from the p available
  - This risks including explanatory variables which are irrelevant because we will check many models when using mechanical methods

4 Qualitative Variables and Interaction Terms

- Explanatory variables can be nominal categorical variables and/or combinations of variables called interaction terms

- Category variables of k categories (i.e., levels) rare encoded into $k-1$ numerical variables, or indicator (dummy) variables
- The y-intercept has a more nuanced meaning when categorical variables are included
- Use partial F-test for testing if a categorical variable slope is zero or not
- There are multiple y-intercepts when using categorical variables and interactions, interpret output wisely and add slope to intercept when applicable
- A partial F-test comparing a model with an interaction term and without the interaction would be equivalent to the t-test for the slope of the interaction since the interaction term is just a dummy variable
- Running a linear regression on interaction terms create multiple parallel equations where the slope coefficients are equivalent and the y-intercept depends on the values of the interaction term
- A partial F-test comparing the model with or without a categorical variable is equivalent to the t-tests for slopes because each variable is comprised of a single dummy variable
- If an interaction term is included in the mode, the lower order terms must be included in the model even if they are not statistically significant

5 Variable Selection Methods

- Purpose of Building a Model
  - Description and model building
    - "science" - trying to understand the main relationships with the response variable
    - focus: parsimonious model (prefer smaller model)
  - Estimation and prediction
    - want to use the model to predict new observations
    - focus: reduce RMSE even if it means extra explanatory variables are needed
  - Control
    - more "science" - trying to understand the main relationships with the response variable in a causal way (note: this generally needs an experimental set up)
    - focus: accurate estimate of slopes, small standard errors for slope estimates
- If there are q possible explanatory variables, either all q variables are in the true model or only p of the q variables are in the true model where $p < q$

- Including all q variables in the model when the true model needs all q variables or including only p variables in the model when the true model needs those p variables is good
- If all q variables are included when only p are needed, then there is loss of precision in the estimated slopes and predictions
- If only p variables are included when all q are needed, then the estimated slopes are biased and RMSE is biased upward
- Models usually fit better on the data used to estimate the model as opposed to new data
- Simulate the existence of new data by dividing data into two groups:
  - training set - data used to explore and fit the model
  - test set - data used to validate the final model
- There is no such thing as a "best set" of variables; what is considered a good model depends on the purpose of building the model in the first place as well
- Mechanical Variable Screen Procedures
  - Forward Selection
    1. Start with a regression with the most significant explanatory variable
    2. Among the remaining variables, choose the most significant explanatory variable and add it to the model
    3. Among the remaining variables, choose the most significant explanatory variable and add it to the model, etc.
    4. Stop adding variables when (a) there are no more variables or(b) no more variables can be added based on a pre-set criterion (e.g., stop adding variables when no p-value is below $\alpha = 0.05$)
  - Backward Elimination: similar to forward selection but start with all explanatory variables and remove variables one by one based on which is the least significant
  - All-Possible Regressions: check all possible regressions (e.g., all 1-variable models, all 2-variable models, etc.) and choose model based on some criterion (note: if there are k variables, then there 2^k possible models, including the y-intercept only model)
  - Best Subset Regression
    1. Pick best regression among all regressions with one variable based on some criterion
    2. Pick best regression among all regressions with two variables based on some criterion
    3. Continue doing this; then, choose among these "best" regressions based on that criterion

- Common Variable Selection Criteria
  - p-values and α
  - R^2_{adj}, MSE (but generally better to use RMSE as it is easier to interpret)
  - Mallow's C_p, BIC, AIC
- These variable selection methods and criteria will not always yield the same model as "best"
- R^2_{adj} and RMSE
  - The adjusted R^2 can be rewritten as

$$\begin{aligned} R^2_{\text{adj}} &= 1 - \left(\frac{n-1}{n-(p+1)}\right)(1-R^2) \\ &= 1 - (n-1)\left(\frac{\text{MSE}}{\text{SST}}\right) \end{aligned}$$

    Then as MSE decreases, R^2_{adj} increases since SST is the same across models
  - Choose the model with the highest R^2_{adj} or the lowest MSE (or lowest RMSE)
  - Some people have argued that the penalty for including additional explanatory variables is too small in the R^2_{adj} formula
  - Note: RMSE $= \sqrt{\text{MSE}}$
  - Depending on the scale of the y variable, MSE may be very large and is in squared units, so it's generally easier to compare RMSEs across models instead
- Mallow's C_p Criterion
  - Use C_p to compare tow nested models (measures bias)
  - Formula:

$$C_p = \frac{\text{SSE}_p}{\text{MSE}_q} + 2(p+1) - n$$

    where q and p are the number of explanatory variables in the complete and subset models respectively
  - Want C_p low and to be close to $p+1$
  - When C_p is close to $p+1$, there is low bias
- AIC - Akaike Inforation Criterion
  - AIC tries to balance accuracy (fit) and simplicity (smaller number of variables, parsimony)
  - Advantage: unlike with F-tests, non-nested models can be compared; but, just like with partial F-tests, the same observations must be used in the two models being compared

  - Formula:
$$\text{AIC} = n \log\left(\frac{\text{SSE}_p}{n}\right) + 2(p+1)$$
where log is natural logarithm and p is the number of explanatory variables
  - Interpretation: As AIC decreases, the model is more preferred
  - Therefore, for two models with similar SSE values, AIC penalizes the model with the larger number of variables

- BIC - Bayes Information Criterion
  - Formula:
$$\text{BIC} = n \log\left(\frac{\text{SSE}_p}{n}\right) + (p+1)\log n$$
where log is natural logarithm and p is the number of explanatory variables
  - The only difference between AIC and BIC is that the penalty for more explanatory variables is larger; the $2(p+1)$ in AIC is replaced by $(p+1)\log n$ in BIC; $2 < \log n$ as long as $n \geq 8$
  - This larger penalty term helps reduce the possibility of overfitting
  - Interpretation: As BIC decreases, the model is more preferred
  - Rule of thumb: BIC difference should be larger than 2 between two models to be considered a substantial difference between the models

- External Model Validation
  - Models generally fit sample data better than new data
  - If the differences are substantial, then there is chance of overfitting has occurred: fitted to the idiosyncrasies of the data, as opposed to the general process from which the data is sampled
  - Validation methods:
    * examining $\hat{y}$ - big or small residuals?
    * examining $\beta_1, \ldots, \beta_k$ - do the signs (i.e., $+/-$) make sense? do the interpretations of the partial slopes make sense?
    * validating with new data
    * data splitting
    * jackknife (resampling) methods, cross-validation

- Data Splitting
  - Sample size of training data is n
  - Denote new (or test set) data: $n+1$, $n+2$, $\ldots$, $n+m$
  - Let $\hat{y}_i$ be the predictions for the new (or test) data

- Compute
$$R^2_{\text{prediction}} = 1 - \left\{ \frac{\sum_{i=n+1}^{n+m}(y_i - \hat{y}_i)^2}{\sum_{i=n+1}^{n+m}(y_i - \bar{y})^2} \right\}$$
where $\bar{y}$ is he mean of the training set response variable
- Compare $R^2_{\text{prediction}}$ with R^2 from the original model; if there is a big drop, then there is a problem
- Compute
$$\text{RMSE}_{\text{prediction}} = \sqrt{\frac{\sum_{i=n+1}^{n+m}(y_i - \hat{y}_i)^2}{m - (p+1)}}$$
where p is the number of explanatory variables in the model
- Compare $\text{RMSE}_{\text{prediction}}$ with RMSE from the original model

- Jackknife (Resampling) Methods: if the data set is too small to split, use the jackknife:
  1. Fit model with the first observation removed
  2. Predict y value for the first observation and call it $\hat{y}_{(1)}$ instead of $\hat{y}_1$
  3. Fit model with second observation excluded but add the first observation back in
  4. Predict y value for the second observation; call it $\hat{y}_{(2)}$ instead of $\hat{y}_2$
  5. Repeat until there are n predictions $\hat{y}_{(1)}, \ldots, \hat{y}_{(n)}$
  6. Compute statistics:
$$R^2_{\text{jackknife}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{(i)})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{PRESS}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$
$$\text{RMSE}_{\text{jackknife}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_{(i)})^2}{n - (p+1)}} = \sqrt{\frac{\text{PRESS}}{n - (p+1)}}$$
where $\bar{y}$ is the mean of the training dataset response variable and p is the number of explanatory variables in the model

- PRESS is usually larger than SSE from the fitted model and so $R^2_{\text{jackknife}}$ is usually smaller than R^2 and $\text{RMSE}_{\text{jackknife}}$ is usually larger than RMSE

- PRESS Criterion - related to the jackknife model validation methods; for a given model:
  1. Fit model with the first observation removed
  2. Predict y value for the first observation and call it $\hat{y}_{(1)}$ instead of $\hat{y}_1$
  3. Fit model with second observation excluded but add the first observation back in
  4. Predict y value for the second observation; call it $\hat{y}_{(2)}$ instead of $\hat{y}_2$
  5. Repeat until there are n predictions $\hat{y}_{(1)}, \ldots, \hat{y}_{(n)}$

6. Compute PRESS statistic:

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2$$

7. Choose model with lowest PRESS statistic

- Note that PRESS is usually larger than SSE $= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- This is slightly different from the MSE method because predictions are computed on data that are not used to fit the model
- How to choose among a large number of variables?
  - Especially when using variable screening procedures, possibly hundreds of t-tests are carried out, checking if slopes are significant for each combination of variables
  - When α is set for these t-tests, then

$$P(\text{Type I error}) = \alpha \text{ for each test}$$

  - Type I error - reject null hypothesis when it is true
  - In a t-test for slopes, a Type I error is saying that the slope is significant when it is actually not
  - Making such an error means that that explanatory variables that aren't actually significant is including in the model - just significant by chance
  - When many hypothesis tests are done, especially when they are related to each other, the probability of a Type I error across all tests combined is more than α
- Bonferroni Correction
  - Usual method: determine whether explanatory variable x is significant if the p-value of the t-test is less than α
  - Bonferroni correction: for each hypothesis test, compare p-value instead to α/h where h is the total number of t-tests
  - The result of using the Bonferroni correction is that fewer variables may be included in the model which are most significant
  - The Bonferroni correction is conservative (i.e, more hypotheses can be rejected than this algorithm permits) but it is very easy to implement
- Note that when using mechanized processes such as forward stepwise, transformations and interaction terms are missed; there are also many models being tried and so there are many hypothesis tests
- These processes should only be used for selecting variables, not determining the final model - the usual steps, including checking assumptions and looking at plots, should be used to determine the final model

6 Regression Diagnostics

- When a new variable z is added to a model, one of the following occurs:
  - Scenario 1: t-test for variable z is not significant; slopes for other variables already in the model do not change much; thus in most cases, leave out variable z from the model
  - Scenario 2: t test for variable z is significant, slopes for other variables in the model change a lot; check for collinearity/multicollinearity among the explanatory variables
  - Scenario 3: t-test for variable z is significant, slopes for other variables in the model do not change a lot; keep variable z in the model
  - Scenario 4: t-test for variable z is not significant; slopes for other variables in the model change a lot; most likely collinearity/multicollinearity
- Getting to the final model is a trial and error process and then there is a need to check assumptions
- Multiple Regression Assumptions: let the population model be $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$ where y is linearly related to the x variables and the sample (estimated) model is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$; then
  - x variables are fixed and measured without error
  - $\mathrm{E}[\varepsilon] = 0$, then $\mathrm{E}[Y \mid X] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, i.e. the mean of the error term is 0 and so the mean of Y given the $X_1, \ldots, X_p$ values is its value on the population regression "line"
  - $\mathrm{Var}[\varepsilon] = \sigma^2$, i.e. the variance of the error term is constant regardless of the $X_1, \ldots, X_p$ values and is denoted by σ^2
  - ε's are normally distributed
  - ε's are independent
  - x variables are not too highly correlated (collinearity / multicollinearity)
- Given the population model and fitted model, where the ε's are the errors, the errors ε can be estimated by computing the residuals, e

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= \text{observed } y - \text{ predicted } y \\ &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi}) \end{aligned}$$

- A residual can be computed for each observation in the data set: $e_1, \ldots, e_n$
- Leverage Values
  - Recall that $\hat{y} = X\hat{\beta}$ where the y-intercept and slopes are determined by $\hat{\beta} = (X^\top X)^{-1} X^\top y$

- Then by plugging in for $\hat{\beta}$

$$\hat{y} = \underbrace{X(X^\top X)^{-1}X^\top}_{H} y$$

where $H = X(X^\top X)^{-1}X^\top$ is called the hat matrix or the projection matrix and is an $n \times n$ matrix

- The ith row of H, denoted as $p_{i1}, \ldots, p_{in}$, are functions of the explanatory variables only, not the response
- H can be used as weights to rewrite $\hat{y}_i$ for the ith observation as

$$\hat{y}_i = p_{i1}y_1 + p_{i2}y_2 + \cdots + p_{in}y_n$$

for $i = 1, \ldots, n$

- The diagonal values of H are called the leverage values: $p_{11}, \ldots, p_{nn}$

- Standardized Residuals
  - Another way to express those leverage values is

$$p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

p_{ii} will be large if x_i is far away from its mean $\bar{x}$, signifying a high leverage point

  - In addition,

$$\text{Var}[e_i] = \sigma^2(1 - p_{ii})$$

where σ^2 is the variability of the error term ε but the variance of the statistic e_i is calculated here

  - To get back to equal variance, compute the standardized residuals

$$z_i = \frac{e_i - 0}{\sigma\sqrt{1 - p_{ii}}}$$

  - σ is unknown; therefore use $\hat{\sigma}$, or RMSE, to compute the internally studentized residual

$$r_I = \frac{e_i - 0}{\hat{\sigma}\sqrt{1 - p_{ii}}}$$

  - r_i is commonly referred to as the standardized residual; for large samples, r_i values should be approximately $N(0, 1)$ (assuming the regression assumptions hold)
  - This means that the empirical rule for normal distributions can be used to determine whether a residual is large or small

- Linearity: Two Types of Standardized Residual Plots

  - Plot fitted value $\hat{y}$ on the x-axis and the standardized residual r_i on the y-axis and look for: trends (i.e., non-linearity), dramatic changes in variability across $\hat{y}$ values (i.e., non-constant variance), outliers, more than 5% of the standardized residuals outside of ± 2 from 0
  - For each explanatory variable, plot x values on the x-axis and the standardized residual r_i on the y-axis and look for: trends (check for linearity), dramatic changes in variability across x values (check for non-constant variance), more than 5% of the standardized residuals outside of ± 2

- Constant Variance
  - Assumption: ε's have constant variance across the regression "line", σ^2 - this property is called homoscedastiticy (i.e., constant variance)
  - If this property is not true, then varying residual variance is calld heteroscedasticity (i.e., non-constant variance)
  - Check: standardized residuals r_i vs $\hat{y}_i$ plot; for more detailed information, check each r_i vs each x variable plot
  - Regression is robust to violations of this assumption; rule of thumb: widest variation is no more than twice the narrowest variation; however, if severe, interpretation of p-values, hypothesis tests, confidence intervals are incorrect
- In a standard residual vs $\hat{y}$ plot, look for changes in vertical variability as $\hat{y}$ increases
- In a $\sqrt{|\text{standard residual}|}$ vs $\hat{y}$ plot, or scale-location plot, look whether the red line is flat is flat or not
- Interpreting a Normal Quantile Plot
  - Assumption Check: evaluate whether ε has a normal distribution by looked at a normal quantile plot of the standardized residuals r_I - points should form a straight line
  - If all points fall on a straight line, then the residuals are normally distributed
  - Consequences of Non-normality: affects Type I error rates for statistical tests and confidence intervals for slope coefficients, i.e., $P(\text{Type I}) \neq \alpha$
  - Regression is fairly robust against violations of normality assumption
  - Be careful with heavy-tailed errors; results will be more sensitive to those extreme data points
- If the plot appears S-shaped on the normal quantile plot, then the distribution of residuals is heavy-tailed
- If there is non-constant variance or non-normality, try transforming the data or doing weighted least squares
- Outliers, Leverage, Influential Points

  - Outlier (univariate): unusual value for a given variable (x, y, etc.), e.g., points beyond the whiskers in a box plot
  - Outliers in the regression context: an observation with an unusual y value given the x value; tends to have large residuals
  - Leverage: point with an unusual x value, i.e., x value far away from its mean
  - Influential Point: point which is both a regression outlier and a leverage point; observation has a large influence on regression estimates

- To check for outliers, look at box plots where an outlier lies outside $1.5 \times$ IQR where the interquartile range (IQR) is the 75th percentile - 25th percentile

- Cook's Distance

  - Cook's Distance calculation

$$C_i = \frac{r_i^2}{p+1} \cdot \frac{p_{ii}}{1-p_{ii}}$$

    where p_{ii} is the leverage value for the ith observation and $\frac{p_{ii}}{1-p_{ii}}$ is called the potential function, r_i is the standardized residual of the ith observation and p is the number of explanatory variables

  - Interpretation: Cook's distance for observation i measures the difference between the β_j's from the model with all data points and the β_j's when the ith data point is removed
  - Rule of Thumb: a point is influential if C_i for the ih point is above the 50th percentile value of an F distribution with $p+1$ and $n-(p+1)$ degrees of freedom
  - More Practical Rule of Thumb: a point is influential if C_i is larger than 1 and it sticks out in the plots

- Other Diagnostic Measure and Plots include:

  - Welsch and Kuh Measure (DFITS, very similar to Cook's distance)
  - Hadi's influence measure
  - potential-residual plot (related to leverage-residual plot)
  - added-variable plot
  - residual plus component plot

7 Multicollinearity and Autocorrelation

- Multiple Regression Assumptions - let the population model be $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$ where y is linear related to the x variables, and the sample (estimated) model is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$, then

  - the x variables are fixed and measured without error

  - $\mathrm{E}[\varepsilon] = 0$, then $\mathrm{E}[Y \mid X] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, i.e. the mean of the error term is 0 and so the mean of Y given the $X_1, \ldots, X_p$ values is its value on the population regression "line"
  - $\mathrm{Var}[\varepsilon] = \sigma^2$, i.e., the variance of the error term is constant regardless of the $X_1, \ldots, X_p$ values and is denoted by σ^2
  - ε's are normally distributed
  - x variables are not too highly correlated (collinearity/multicollinearity)

- Multicollinearity - two or more explanatory variables which are moderate/highly correlated with each other
- Note: If it's just two explanatory variables which are correlated, that is simply called collinearity
- Collinearity/multicollinearity is an issue only if the relationships are very strong
- Some problems caused by severe multicollinearity:
  - Difficulty interpreting the partial slope estimate ($\hat{\beta}_j$)
  - Inflated $\mathrm{SE}[\hat{\beta}_j]$ values
  - Effects on the signs of parameters
  - Potential issues with prediction including when constructing prediction and confidence intervals
  - Rounding errors when estimating $\hat{\beta}_j$'s, $\mathrm{SE}[\hat{\beta}_j]$'s, etc. (i.e., computer errors)
- Detecting Multicollinearity: if one of the following occurs, it may indicate the presence of multicollinearity:
  - significant correlations between pairs of explanatory variables - look at the scatterplot and correlation matrices
  - t-tests not significant at all (or nearly all) individual β estimates, but the overall F-test is significant
  - opposite signs, from what is expected, in the estimated β values
  - variance inflation factor, VIF, > 10
- Solutions for Reducing the Effects of Multicollinearity:
  - drop one or more explanatory variables from the final model (i.e., get rid of redundant variables)
  - combine variables, possibly using principle components
  - use ridge regression to reduce issues related to rounding error
  - conduct an experiment to determine causal relationships among variables
- Variance Inflation Factor

- Variance Inflation Factor (VIF) measures how much the variance of $\hat{\beta}_j$ is increased (inflated) by the other x variables in the regression
- For each jth explanatory variable,

$$\text{SE}[\hat{\beta}_j]^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\left(\frac{1}{1 - R_j^2}\right)$$

  where
  - $\text{SE}[\hat{\beta}_j]$ is the standard error of the estimated partial slope j
  - $\hat{\sigma}$ is the standard deviation of the errors (RMSE) and so $\hat{\sigma}^2$ is the variance of the errors (MSE)
  - $\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the total variability of the $j$the explanatory variable
  - R_j^2 is the coefficient of determination when regression variable j against the other explanatory variables in the regression (excluding response variable)
- Then

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

  is the amount that $\text{SE}[\hat{\beta}_j]^2$ is increased because of the other x variables int he regression
- The VIF value should be computed for each numerical explanatory (i.e., x) variable in the model
- Steps:
  1. Say $x_1, \ldots, x_p$ are used in a regression model where the response is y (note: y is irrelevant in computing VIF)
  2. Fit model

$$\hat{x}_j = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{j-1} x_{j-1} + \alpha_{j+1} x_{j+1} + \cdots + \alpha_p x_p$$

  3. Compute the R^2 value for the regression above; this is R_j^2
  4. Then compute

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

- Recall that the R^2 is the fraction of variation in the response variable which can be explained by the explanatory variables
- If one explanatory variable is regressed against the others, then R_j^2 would be interpreted as the fraction of variation in x_j that can be explained by the other explanatory variables in the model
- Example: if all of the x variables are uncorrelated with each other, R_j^2 would be 0 and $\text{VIF}_j = 1$, meaning no inflation of variance
- General Rule: multicollinearity is severe for variable j if $\text{VIF}_j > 10$; moderate if $\text{VIF}_j > 5$

  - NoteL some software packages compute tolerance instead of VIF, where tolerance = $\frac{1}{\text{VIF}}$

- Independence of Errors
  - relevant for time series data - e.g. data collected hourly, daily, monthly, etc.; today's data value may be related to yesterday's data value (autocorrelation / serial correlation)
  - Plot residuals by order of observations
  - Hypothesis test: Durbin-Watson test for residual correlation - most common is "greater" (test for positive autocorrelation)
  - Effects of autocorrelation: standard errors of parameter estimates are underestimated, RMSE value too low, interpretation of p-values and confidence intervals are incorrect
  - If residuals are correlated, using consecutive differences, or the previous value in the regression may help (but generally, it's better to look at a time series method)

- Autocorrelated Residuals
  - Let the fitted model be $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$, the residuals from the fitted model be $e_1, e_2, \ldots, e_n$
  - If the order the data was corrected matters (e.g., time series), consecutive residuals can be paired as follows:

"x"	"y"
e_1	e_2
e_2	e_3
e_3	e_4
e_4	e_5
$\vdots$	$\vdots$
e_t	e_{t-1}
$\vdots$	$\vdots$
e_{n-1}	e_n

  - Next step: check to see if neighboring residuals have the following relationship (e.g., first-order autoregressive serial correlation), where ω_t is an error term and $|\rho| < 1$:

$$\varepsilon_t = \rho\varepsilon_{t-1} + \omega_t$$

- Durbin-Watson Test for Residual Correlation - testing for positive autocorrelation
  - Hypotheses:

$$H_0: \text{ no residual correlation, } \rho = 0$$
$$H_A: \text{ positive residual correlation, } \rho > 0$$

  - Let the significance level be α

  - Collect data and assume the residuals are normally distributed
  - Durbin-Watson test statistic:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_i^2}$$

  - Conclusion: get two "critical values", d_L and d_U, for each α level to which the test statistic d is compared against
    * if $d < d_L$: reject H_0
    * if $d > d_U$: fail to reject H_0
    * if $d_L \leq d \leq d_U$: cannot make a conclusion

- Interpreting the Durbin-Watson d Statistic
  - d statistic:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_i^2} \approx 2 - 2\hat{\rho} = 2(1 - \hat{\rho})$$

  - If e_t are uncorrelated, then $\hat{\rho}$ will be close to 0 and so d will be close to 2
  - If e_t are highly, positively correlated, then $\hat{\rho}$ would be close to 1 and so d will be close to 0
  - If e_t are highly, negatively correlated, then $\hat{\rho}$ would be close to -1 and so d will be close to 4
  - Therefore the range of d is 0 to 4

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \underbrace{\frac{\sum_{t=2}^{n} e_t^2}{\sum_{t=1}^{n} e_t^2}}_{\approx 1} + \underbrace{\frac{\sum_{t=2}^{n} e_{t-1}^2}{\sum_{t=1}^{n} e_t^2}}_{\approx 1} - 2\underbrace{\frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}}_{\approx \hat{\rho}}$$

    where $\hat{\rho}$, in a first order setting, equals the correlation between consecutive pairs of residuals
  - Note: for a test of negative autocorrelation, replace d with $4 - d$ in the tes
  - The Durbin-Watson test is for lag 1 autocorrelation - e_t compared with e_{t-1}; it will not catch autocorrelation at higher laps even though the independence assumption would be violated

- Let the model be $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ and $\text{VIF}_j = \frac{1}{1-R_j^2}$; when there are only two explanatory variables, regress one explanatory variable against the other, which is a simple linear regression; then R_j^2 is simply the square of the sample correlation, r, and $\text{VIF}_1 = \text{VIF}_2$, because correlation of x_1 and x_2 is the same as the correlation between x_2 and x_1; therefore, for models like this, the relationship between correlation and VIF can be displayed visually by plotting correlation r against VIF

8 Transformations

- If relationships between x and y is not linear, as seen in the scatterplots and correlations between the explanatory and response variables, it can be made linear using a transformation
- Another use for transformations is to reduce heteroscedasticity
- Some examples of common transformations
  - y vs. $\log x$
  - $\log y$ vs. x
  - $\log y$ vs. $\log x$
  - centering, standardizing variables
- Choosing the right transformation is a mixture between subject-area knowledge, trial and error and looking at scatterplots
- Other variance stabilizing transformations include $\sqrt{y}$, $\sin^{-1}\sqrt{y}$, and more
- $\log x$ Model
  - models of the form
$$y = \beta_0 + \beta_1 \log x + \varepsilon$$
  - This transformation cannot be used directly if x values are not positive
  - Note: Additional explanatory variables can be included and have a log explanatory variable in a larger regression mode; however the interpretation below should be applied to the log explanatory terms only
  - Slope interpretation for x: a 1% increase in x is associated with an $\approx \hat{\beta}_1/100$ change in y
  - y-intercept interpretation: since $\log 1 = 0$, the y-intercept is the value of y when $x = 1$ - check whether this is meaningful within the context of the data
- $\log y$ Model
  - Multiplicative model:
$$y = e^{\beta_0+\beta_1 x}\varepsilon$$
$$\log y = \beta_0 + \beta_1 x + \underbrace{\log \varepsilon}_{\text{error}}$$
  - Additional explanatory variables can be included; y must be a positive variable to use the log transformation
  - Interpretation of slope: a 1 unit increase in x is associated with a change in y by $\approx (e^{\beta_i} - 1) \times 100\%$
- $\log x$ and $\log y$ Models

  - Modes of the form

$$y = \alpha x^{\beta_1} \varepsilon$$
$$\log y = \underbrace{\beta_0}_{\log \alpha} + \beta_1 \log x + \underbrace{\log \varepsilon}_{\text{error}}$$

  - This transformation cannot be used directly if x and y values are not positive
  - Additional explanatory variables can be included and have a log explanatory variable in a larger regression model; however the interpretation below should be applied to the log explanatory terms only
  - In a $\log x$ and $\log y$ model, a 1% increase in x is associated with a % change in y on the average; in other words, a 1% increase in x is associated with an $\approx \beta_1\%$ change in y
  - The y-intercept is the $\log y$ value when $x = 1$ (and so $\log x = 0$); this should still be checked if it's meaningful in the context of the data

- Jensen's Inequality

  - Let y be a random variable with expected value $\mathrm{E}[y]$ and $f(y)$ be a function of the random variable y with expected value $\mathrm{E}[f(y)]$, then if $f(\cdot)$ is a convex function,

$$f(\mathrm{E}[y]) \leq \mathrm{E}[f(y)]$$

    an if $f(\cdot)$ is a concave function, like $\log(\cdot)$, then

$$f(\mathrm{E}[y]) \leq \mathrm{E}[f(y)]$$

  - The goal of regression is to find a model of $\mathrm{E}[y \mid x]$, or the expected value of y given x
  - When the response y is transformed using the natural log, the model being fitted is $\mathrm{E}[\log y \mid x]$, or the expected value of $\log y$ given x
  - Predictions from such a regression model are in terms of average $\log y$ given an x but what is really desired is a prediction of average y given an x
  - Due to Jensen's inequality, generally speaking,

$$\mathrm{E}[\log y] \neq \log \mathrm{E}[y]$$

    * In words, this means that the average of the variable $\log y$ is not equal to taking the log of the average value of y
    * Therefore, $\mathrm{E}[y] \neq \exp(\mathrm{E}[\log y])$, which means that if $\mathrm{E}[\log y]$ is modeled, average y is not obtained by simply taking the inverse transformation, $\exp(\cdot)$
  - The solution for making predictions for y when the regression model is fit on $\log y$ is as follows

$$\hat{y} = \exp\left\{\hat{\log y} + \frac{(n-p-1)\mathrm{RMSE}^2}{2n}\right\} = \exp\left\{\hat{\log y} + \frac{\mathrm{SSE}}{2n}\right\}$$

    where $\hat{\log y}$ is the prediction from the fitted regression model and the adjustment $\frac{(n-p-1)\mathrm{RMSE}^2}{2n}$ accounts for the Jensen's inequality issue

  - Another method of fitting a model is called maximum likelihood estimation (MLE)
    * For simple and multiple regression, the MLE estimates of the y-intercept and slopes $(\hat{\beta}_0, \ldots, \hat{\beta}_p)$ is the same as the OLS estimate of the y-intercept and slopes; but, MLE makes specific use of $\varepsilon \sim N(0, \sigma)$ assumption
    * However, the estimate of the error variance is not the same

$$\sigma^2_{\text{MLE}} = \frac{(n-p-1)\sigma^2_{\text{OLS}}}{n} = \frac{\text{SSE}}{n}$$

    * The $\frac{(n-p-1)\text{RMSE}^2}{2n}$ adjustment is based on properties of MLEs
  - Similar adjustments can be made when interpreting slopes, etc. for $\log y$ models

- Centering and Scaling Variables
  - Let x be a variable (explanatory or response), $\bar{x}$ its sample mean and s_x its sample standard deviation
  - Centering a variable: $x - \bar{x}$ where $\bar{x}$ is the sample mean of that variable
    * The mean of $x - \bar{x}$ is now 0 but standard deviation is still s_x
    * Units of transformed variable: units of x
  - Scaling (standardizing) a variable: $\frac{x-\bar{x}}{s_x}$ where s_x is the standard deviation of the variable
    * The mean of $\frac{x-\bar{x}}{s_x}$ is now 0 and the standard deviation is now 1
    * Units of transformed variable: no units; interpret in terms of standard deviations away from mean
  - Scaling makes the variables unit-less; if the explanatory variables are on very different scales, then scaling will help with interpreting partial slopes
  - Some texts suggest centering a variable in an interaction term reduces multicollinearity, but there is mixed evidence on this front
  - Like all transformations, centering/scaling should not be an automatic step; if should only be done if it's helpful

9 Logistic Regression

- In multiple regression, the response y is numerical and the explanatory variables are numerical or categorical
- In logistic regression, the response y is a binary categorical variable (i.e., two categories) while the explanatory variables are numerical or categorical
- A one-way ANOVA model has a response y which is a numerical value and one (nominal) categorical explanatory variable

- In a logistic regression model, the binary response is recorded as

$$y = \begin{cases} 1 \\ 0 \end{cases}$$

- $\pi = P(y = 1)$ is the probability of success for response variable y given explanatory variable values
- The odds of success is the probability of success / probability of failure

$$\text{Odds} = \frac{P(y=1)}{P(y=0)} = \frac{\pi}{1-\pi}$$

- The log odds is

$$\text{Log odds} = \log\left(\frac{\pi}{1-\pi}\right)$$

- The use of log odds instead of odds allow comparison of the range of y-axis
- When
  - odds $= 1$, then success is equally likely to failure
  - odds > 1, then success is more likely than failure
  - odds < 1, then success is less likely than failure
- The logistic regression model fitted is

$$\text{logit}[\pi] = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

  where the middle term is called the logit function (i.e., log odds), a type of link function
- The model can be rewritten in terms of π

$$\pi = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}$$

- The model for $1 - \pi$ is

$$1 - \pi = 1 - \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}} = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}$$

- The logistic regression model is fit using maximum likelihood estimation
- Bernoulli Trials
  - A Bernoulli trial is a random process with binary outcome and a probability π of success

  - X is a Bernoulli random variable with potential outcomes: $X = 0$ or failure, $X = 1$ or success
  - X is a discrete random variable
  - The probability mass function for X is

$$P(X = k) = \pi^k (1 - \pi)^{1-k}$$

    where $k = 0$ or $k = 1$ as those are the only possible outcomes for X

- The Bernoulli trial is similar to a logistic regression model where each observation in the logistic regression model is a Bernoulli trial
- For each observation (i.e., Bernoulli trial), the probability mass function is

$$P(Y = y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

  where $y_i = 0$ or $y_i = 1$

- The next step is to determine the distribution of all observations together; this is the joint distribution
- If each observation is assumed to be independent of the others, then the joint distribution is simply the product of the individual probability mass functions:

$$\prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

- Now find a function for $\pi_i = P(Y = 1 \mid X = x)$, the probability of success given the values of explanatory variables
- A function is needed that describes π_i as a function of x_i
  - $\pi(x_i)$ is a linear function of x which cant work because probabilities must be between 0 and 1 since it is a probability
  - $\log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)$ is a linear function of x which is better because the function can be any real number
  - This function is called the logit function

$$\log\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \beta_0 + \beta_1 x_i$$

    and so

$$\pi(x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}$$

- The likelihood function is a function of parameters given data:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}$$
$$= \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-(\beta_0+\beta_1 x_i)}}\right)^{y_i} \left(1 - \frac{1}{1 + e^{-(\beta_0+\beta_1 x_i)}}\right)^{1-y_i}$$

- Maximum Likelihood Estimation
  - The logistic regression fit is done using a technique called maximum likelihood estimation (MLE)
  - The likelihood function is a function of parameters given data
  - Find values of parameters (e.g., β_0, β_1) that makes the likelihood function value the largest (i.e., maximum)
    1. Take log of likelihood function, turning products into sums and gets rid of e
    2. Take partial derivatives of log-likelihood function with respect to each parameter (e.g., $\frac{\partial \log L}{\partial \beta_0}$ and $\frac{\partial \log L}{\partial \beta_1}$)
    3. Set partial derivatives to zero and solve for parameter - done numerically for logistic regression as there is no closed form solution
  - The maximum likelihood estimates of β_0 and β_1 have the following desirable properties as $n \to \infty$:
    * Asymptotically normal - the estimates become more normally distributed
    * Consistent - estimate converges to population value
    * Efficient - there is no other consistent estimator which has a better asymptotic mean squared error
    * These properties allow hypothesis tests to work

- Logistic Regression Assumptions
  - Explanatory variables are measured without error
  - The model is correctly specified (no extraneous variables, all important variables included, etc.)
  - The outcomes are not completely linearly separable i.e., knowing x values cannot completely determine whether $y = 0$ or $y = 1$; makes β estimates unstable
  - No outliers, etc - compare slope values, etc. when each observation is removed one at a time; look at scatterplots, box plots of data to search for outliers; look at approximated leverage values, Cook's distance, etc
  - Observations are independent - check data collection process
  - Collinearity / multicollineariy - applicable if there are multiple explanatory variables; if there is perfect collinearity or multicollinearity, a logistic regression model cannot be fitted; if there's a high level, it makes β estimates imprecise; use VIF and look at scatterplot matrices of explanatory variables

  - Sample size, n - requires more observations than usual regression, especially if one of the categories occurs rarely; rule of thumb: at least 10 observations for each outcome (0/1) per predictor in the model

- Interpretation of β Parameters in the Logistic Model
  - Model:
$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$
  - $\pi = P(y = 1)$ and $\log\left(\frac{\pi}{1-\pi}\right)$ is the log odds
  - $\hat{\beta}_0$, or $e^{\hat{\beta}_0}$ is the odds of success when all x variables are 0
  - $\hat{\beta}_j$ where $j > 0$ is the change in log-odds for every 1 unit increase in x_j holding all other x's fixed, but this is not a very intuitive interpretation
  - Meaningful interpretation of $\hat{\beta}_j$:
$$(e^{\hat{\beta}_j} - 1) \times 100\%$$
    is the percentage change in odds $\frac{\pi}{1-\pi}$ for every unit increase in x_j holding all other x's fixed

- Log-Likelihood Test for Overall Model
  - Hypotheses
$$H_0 : \beta_1 = \cdots = \beta_p = 0$$
$$H_A : \text{ at least one of the slopes is not 0}$$
  - Set α level and fit the model
  - The test statistic is
$$G = -2\log\left(\frac{L_{H_0}}{L_{\text{model}}}\right) = -2\log(\log L_{H_0} - \log L_{\text{model}})$$
    where L is the likelihood function; G has a χ^2 distribution with p degrees of freedom, L_{H_0} is the likelihood if all slopes are zero (i.e., model n the null hypothesis) and L_{model} is the likelihood for the fitted model
  - In R, it is
$$\text{G} = \text{null deviance} - \text{residual deviance}$$
    with *df* = null deviation *df* - residual deviance *df*
  - Reject null hypothesis if $P(\chi^2 > G) < \alpha$

- The log-likelihood test for overall model is analogous to the overall F-test in regular regression

- χ^2 Distribution

  - The probability density function is

$$f(x) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

    where

$$\Gamma(z) = \int_0^\infty y^{z-1} e^{-y}\, dy$$

  - Here $x \geq 0$ and the distribution is continuous
  - The parameter k is degrees of freedom
  - The square of $N(0,1)$ random variable is a χ^2 distribution with 1 degree of freedom
  - The sum of independent χ^2 random variables is also χ^2 with degrees of freedom equal to the sum of the degrees of freedom of the individual variables

- Log-likelihood test requires null deviance and residual deviance, bu since the logistic model models the the log odds and not the 0s and 1s directly, there is a different notion of residuals
- There are two definitions of residuals: Pearson residuals and deviance residuals
- Deviance Residual

$$d_i = s_i \sqrt{-2(\log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i))}$$

  where

$$s_i = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = 0 \end{cases}$$

  The deviance residual measures the contribution of each point to the likelihood equation

- Null Deviance and Residual Deviance
  - Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$
$$H_A : \text{ at least one of the slopes is not } 0$$

  - Null deviance: deviance computed for the y-intercept only model
  - Residual deviance: deviance computed for the fitted model

- z-Tests for Slopes
  - Hypotheses:

$$H_0 : \beta_j = 0$$
$$H_A : \beta_j \neq 0$$

  - Set α and fit the model

  - Test statistic:
$$z = \frac{\hat{\beta}_j - 0}{\text{se}_{\hat{\beta}_j}}$$
where the test statistic z has a standard normal distribution
  - The p-value is for the two-sided test: $2P(Z > |z|)$
  - Notes: estimates computed using maximum likelihood become normally distributed as n increases; this test is different from the likelihood test for the overall model

- For logistic regression, leverage values are computed differently because multiple regression has a linear setup and logistic regression does not
- To compute standardized deviance residuals, calculate
$$d_i^* = \frac{d_i}{\sqrt{1 - H_{ii}}}$$
where H_{ii} is the leverage
- The fitted values from logistic regression model are $\hat{\pi}$, the probability of success, not log odds
- Recall that the response variable y is originally categorical; estimated probability for an observation can be converted into categories
  - Recall: $\pi = P(y = 1)$
  - Choose a threshold π^*
  - if estimated probability $< \pi^*$, then classify the observation as $\hat{y} = 0$
  - if estimated probability $> \pi^*$, then classify the observation as $\hat{y} = 1$
  - Note: usually if estimated probability $= \pi^*$, then classify as $\hat{y} = 1$
- Evaluating Classifications

	1	0
1	TP = true positive	FP = false positive
0	FN = false negative	TN = true negative
	P=positive	N=negative

where vertically is category in data y and horizontally is predicted category $\hat{y}$
  - Correct decision - true positive, true negative
  - Incorrect decision - false positive, false negative
  - False positive - Type I error
  - False negative - Type II error
  - TP, TN, FP and FN all depend on the threshold value

- As false positive decreases, false negatives increases as the threshold value is moved
- Accuracy: the probability a correct prediction is made

$$\text{accuracy} = P(\text{correct prediction}) = \frac{TP + TN}{P + N} = \frac{TP + TN}{n}$$

  where $P + N = n$ is the sample size; accuracy rate changes with threshold chosen
- Accuracy is too general; FN and FP have different repercussions
- Sensitivity and specificity are conditional probabilities
  - Sensitivity: probability that an observation is classified as 1 given that the true category is 1

$$\text{sensitivity} = P(\hat{y} = 1 \mid y = 1) = \frac{TP}{P}$$

  - High sensitivity means low Type II error rate (i.e., false negatives)
  - Specificity: probability that an observation is classified as 0 given that the true category is 0

$$\text{specificity} = P(\hat{y} = 0 \mid y = 0) = \frac{TN}{N}$$

  - High specificity means low Type I error rate (i.e., false positives)
  - Rates changes based on the classification threshold
  - Ideally, both should be high but as sensitivity increases, specificity decreases
- ROC Curves
  - ROC = receiver operating characteristic
  - Allows seeing how choosing different thresholds perform
  - It plots the true positive rate (y-axis) vs false positive rate (x-axis) at different threshold values
  - Compare the curve against the 45° line
  - Note: false positive rate = 1− specificity
- AUC is the area under the ROC curve; ideally, it should be close to 1
- When AUC = 1, the observations are perfectly classified, i.e., almost completely non-overlapping distributions of outcomes
- When AUC = 0.5, it is just as good as random guessing, i.e., nearly completely overlapping distributions of outcomes
- Rule of thumb for interpreting AUC curves:
  - $0.9 <$ AUC < 1: excellent

- $0.8 < \text{AUC} < 0.9$: good
- $0.7 < \text{AUC} < 0.8$: fair
- $0.6 < \text{AUC} < 0.7$: poor
- $0.5 < \text{AUC} < 0.6$: fail
- $\text{AUC} < 0.5$: worse than guessing

10 Multiple Logistic Regression

- Multiple logistic regression models a binary response variable with variables that can be categorical, interaction terms, etc
- Multicollinearity can be an issue with logistic regression - check scatterplots and/or VIF
- The largest issue with multiple logistic regression is sample size; a rule of thumb is to have at least 10 observations for each outcome (0/1) per predictor in the model
- Odds and Log Odds
  - Binary response:
$$y = \begin{cases} 1 & \text{success} \\ 0 & \text{failure} \end{cases}$$
  - $\pi = P(y = 1)$, the probability of "success" for response variable y given explanatory variable values
  - Odds of success:
$$\text{Odds} = \frac{\text{probability of success}}{\text{probability of failure}} = \frac{P(y=1)}{P(y=0)} = \frac{\pi}{1-\pi}$$
    - If odds $= 1$, success is equally likely to failure
    - if odds > 1, success is more likely than failure
    - if odds < 1, success is less likely than failure
- Logistic Regression
  - Odds are always positive, log odds can be any real number, thus it is better to construct models using log odds
  - Logistic regression model:
$$\text{logit}[\pi] = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$
    The middle term is called the logit function (i.e., log odds), a type of link function

  - The model can be rewritten in terms of π

$$\pi = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

  - The model for $1-\pi$ is

$$1-\pi = 1 - \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

  - The logistic regression model is fit using maximum likelihood estimation

- Logistic Regression Assumptions
  - explanatory variables are measured without error
  - model is correctly specified (no extraneous variables, all important variables included, etc.)
  - outcomes are not completely linearly separable
    - i.e., knowing x values cannot completely determine whether $y = 0$ or $y = 1$
    - makes β estimates unstable
  - no outliers, etc
    - compare slope values, etc. when each observation is removed one at a time
    - look at scatterplots, box plots of data to search for outliers
    - look at approximated leverage values, Cook's distance, etc
  - observations are independent - check data collection process
  - collinearity / multicollinearity (applicable if there are multiple explanatory variables)
    - perfect collinearity / multicollinearity - cannot fit logistic regression
    - high level of collinearity / multicollinearity - makes β estimates imprecise
    - use VIF and look at scatterplot matrices of explanatory variables; VIF > 10 for an explanatory variable indicates a collinearity / multicollinearity issue
  - sample size, n - requires more observations than usual regression, especially if one of the categories occurs rarely; rule of thumb: at least 10 observations for each outcome (0/1) per predictor in the model

- Log-Likelihood Test for Overall Model
  - Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$
$$H_A : \text{ at least one of the slopes is not 0}$$

  - Set α level and fit the model

- Test statistic:

$$G = -2\log(L_{H_0}/L_{\text{model}}) = -2(\log L_{H_0} - \log L_{\text{model}})$$

where L is the likelihood function, G has a χ^2 distribution with p degrees of freedom, L_{H_0} is the likelihood if all slopes are zero (i.e., model in the null hypothesis) and L_{model} is the likelihood for the fitted model

- Reject null hypothesis if $P(\chi^2 > G) < \alpha$

- Log-Likelihood Test for Comparing Nested Models
  - Nested models:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g \text{ (reduced model)}$$
$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g + \beta_{g+1} x_{g+1} + \cdots + \beta_k x_k \text{ (complete model)}$$

  - Set significance level α
  - Hypotheses (testing $k - g$ slopes):

$$H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$$
$$H_A : \text{ at least one slope is not } 0$$

  - sample size of n (same observations for both models)
  - Test statistic:

$$G = \text{residual deviance for reduced model} - \text{residual deviance for complete model}$$

which has a χ^2 distribution with $k - g$ degrees of freedom

  - Reject H_0 if p-value $= P(\chi^2 > G) < \alpha$

- z-Tests for Slopes
  - Hypotheses:

$$H_0 : \beta_j = 0$$
$$H_A : \beta_j \neq 0$$

  - Set significance level α
  - Test statistic

$$z = \frac{\hat{\beta}_j - 0}{\text{se}_{\hat{\beta}_j}}$$

This test statistic has a standard normal distribution

  - p-value for two-sided test: $2P(Z > |z|)$
  - Reject H_0 if p-value $< \alpha$

- Fitted values from logistic regression model are the predicted probability of success (i.e., $\hat{P}(y = 1 \mid x)$)
- The estimated probability can be converted into categories for observations
  - $\pi = P(y = 1)$
  - Choose a threshold π^*
  - If estimated probability $< \pi^*$, classify the observation as $\hat{y} = 0$
  - If estimated probability $> \pi^*$, classify the observation as $\hat{y} = 1$
  - Note: usually if estimated probability $= \pi^*$, classify the observation as $\hat{y} = 1$
- Evaluating Classifications

	1	0
1	TP = true positive	FP = false positive
0	FN = false negative	TN = true negative
	P=positive	N=negative

  where vertically is category in data y and horizontally is predicted category $\hat{y}$
  - Correct decision - true positive, true negative
  - Incorrect decision - false positive, false negative
  - False positive - Type I error
  - False negative - Type II error
  - TP, TN, FP and FN all depend on the threshold value
- Accuracy: the probability a correct prediction is made

$$\text{accuracy} = P(\text{correct prediction}) = \frac{TP + TN}{P + N} = \frac{TP + TN}{n}$$

- Accuracy is too general; FN and FP have different repercussions
- Sensitivity: probability that an observation is classified as 1 ($\hat{y}$) given that the true category is 1 (y)

$$\text{sensitivity} = P(\hat{y} = 1 \mid y = 1) = \frac{TP}{P}$$

- High sensitivity means low Type II error rate (i.e., false negatives)
- Specificity: probability that an observation is classified as 0 ($\hat{y}$) given that the true category is 0 (y)

$$\text{specificity} = P(\hat{y} = 0 \mid y = 0) = \frac{TN}{N}$$

- High specificity means low Type I error rate (i.e., false positives)

- Ideally, both should be high but as sensitivity increases, specificity decreases
- ROC = receiver operating characteristic - a plot of the true positive rate (y-axis) vs. false positive rate (x-axis) at different threshold values); compare the curve against the 45° line
- Note: false positive rate = 1− specificity
- A 50% threshold is standard (but somewhat arbitrary) and the classification rates can be improved on by choosing a different threshold
- ROC curves show how well the models do at predictions at different thresholds
- A common definition of "best" thresholds is the one which maximizes the sum of sensitivity and specificity, but this assumes that FNs and FPs are equally bad (which is not always the case)
- AUC = area under the ROC curve - overall measure of quality for fitted model; it should be close to 1 (i.e., area under 1×1 square is 1)
- If AUC = 1, observations are perfectly classified; if AUC = 0.5, observations are classified as good as randomly guessing
- Rule of thumb for interpreting AUC curves:
  - $0.9 < \text{AUC} < 1$: excellent
  - $0.8 < \text{AUC} < 0.9$: good
  - $0.7 < \text{AUC} < 0.8$: fair
  - $0.6 < \text{AUC} < 0.7$: poor
  - $0.5 < \text{AUC} < 0.6$: fail
  - $\text{AUC} < 0.5$: worse than guessing
- Ways of defining the "best threshold": choose the threshold which has the largest sum of sensitivity and specificity, choose the threshold which has the largest AUC
- When dealing with missing data, probabilities cannot be predicted and thus not classify; accuracy, sensitivity and specificity can be recalculated taking this into account but predictive performance will be lower
- Other methods such as classification and regression trees (CART), random forests, etc. are able to make predictions for missing data but that does not mean the reasons why observations are missing can be ignored, nor can the (possibly unknown) effects / biases those missing observations have on the resulting model can be ignored
- Regardless of whether the missing data are included or excluded, classifications will still generally be better for the observations used to fit the model

- Similar to multiple regression, techniques like splitting data into training and test sets (need large n) and jackknife / cross - validation to mimic training and test sets can be used to do in-sample and out-of-sample classifications
- When there are a lot of explanatory variables to choose from, use forward/backward/ etc. stepwise techniques to help narrow the list of potentially useful variable; just like in multiple regression, these should help in determining which variables to look at closer and not which variables should be in the final model
- Getting to the final model, especially when there are a lot of explanatory variables requires doing many hypothesis tests; be aware that, as a result, variables that look significant may get included, but actually aren't; the Bonferroni correction is helpful to reduce this problem

11 Experimental Design and One-Way ANOVA

- Experiments, if carried out correctly, can help determine causality; but correlation does not imply causation
- By controlling the explanatory variables, the effect of the variables on the response can be determined
- There are many types of experimental design: completely randomized, randomized block, Latin square/cub, incomplete block, complete factorial, and more
- This section will be on balanced completely randomized designs, analyzed using a one-way ANOVA model
- Main Principles of Experimental Design:
  - Control the effects of lurking / confounding variables on the response through the design of the experiment
  - Randomize: use chance to assign subjects to treatment groups
  - Replicate: repeat with as many experimental units as possible
- Following these principles reduces biases
- Keywords
  - experiment: process of collecting data under a controlled setting
  - design of the experiment: method of collecting the data
  - response variable: variable measured in the experiment
  - experimental unit: item/person upon which the response variable is measured (not the response variable itself)
  - factors: explanatory variables in the experiment

  - level: intensities of the explanatory variable (within a factor)
  - treatment: combination of levels in an experiment
  - lurking variable: variables that affect the results but which were not included in the study

- Designing an Experiment
  - Choose factors to be included, identify the response variable(s)
  - Choose the treatments (i.e., factor-level combinations)
  - Determine sample size (usually based on some combination of desired standard error and time and cost constraints)
  - Determine how treatments will be assigned to the experimental units (i.e., design of experiment)

- One-Way Model with a Completely Randomized Design
  - response variable y - numerical variable
  - explanatory variable x - qualitative with p categories (called treatments)
  - Goal: want to compare average of response variable y across p treatments
  - have 1 factor, with p levels
  - sample size n
  - randomly assign each of the n experimental units to one of the p treatments

- Balanced and Unbalanced Designs
  - Balanced design: equal sample sizes for each treatment group
  - Unbalanced design: unequal sample sizes for each treatment group (this means that there is more information about some groups, with higher sample sizes, than other groups, with smaller sample sizes)
  - Unbalanced designs can cause a lot of problems especially with more complex experimental design schemes
  - It is possible to start off with a balanced design but because of nonresponse and/or test subjects dropping out before the end of the study, the resulting data becomes unbalanced

- Suppose there are two populations, each with a sample from which a parameter is measured
  - Goal: compare parameter A with parameter B using statistic A and statistic B respectively
  - Hypotheses:

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

- Two populations have the same variance; compute the test statistic assuming the null hypothesis is true

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  which has a t distribution with $n)1 + n_2 - 2$ degrees of freedom

- The pooled variance is

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

- Assumptions: independent SRSs, $\sigma_1 = \sigma^2$, hypothesis test is exact if populations are normally distributed, approximated for large n
- Rule of thumb: if the larger sample standard deviation is no more than twice the smaller standard deviation, it is generally ok to assume that $\sigma_1 = \sigma_2$
- Note: the $n_1 = n_2$ case is more robust against both assumptions
- The $(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- Now, if the two populations have different variances, then the test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- The test statistic comes from an approximate t-distribution with the Satterthwaite approximation for degrees of freedom:

$$\text{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

- Assumptions: independent SRSs, hypothesis test is exact if populations are normally distributed, otherwise approximate for large n
- Note: the case where $n_1 = n_2$ is more robust against normality

- One-Way ANOVA Model
  - Regression Model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon$$

  - This model can be rewritten as

$$y_{ij} = \mu_j + \varepsilon_{ij}$$

    where j is the treatment ($j = 1, \ldots, p$) and i is the ith experimental unit in the jth treatment group
  - μ_j is the mean of the response variable in the jth treatment group; y_{ij} is the response variable value of the ith experimental unit in the jth treatment group; ε_{ij} is the random error for the ith experimental unit in the jth treatment group
  - The estimate of μ_j goes to $\bar{y}_j$, the mean of the response values in the jth treatment group
  - Note: think of the model as $\beta_0 = \mu_0$, $\beta_0 + \beta_1 = \mu_1$, $\beta_0 + \beta_2 = \mu_2$, and so foth; so group p would be the dropped category when converting categorical variables into dummy variables

- Completely Randomized Design
  - Regression model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

  - There are two treatments, so one dummy (indicator) variable:

$$x = \begin{cases} 1 & \text{treatment 1} \\ 0 & \text{treatment 2} \end{cases}$$

  - Converting the regression model notation into the one-way ANOVA model notation:

$$y_{ij} = \mu_j + \varepsilon_{ij}$$

    where μ_1 is the average response for one treatment and μ_2 is ths average response for the other treatment
  - One-Way ANOVA Model Assumptions
    * independent observations - completely randomized design takes care of this assumption, assuming observations are a simple random sample from the population to begin with
    * Assume ε_{ij} are normally distributed with mean 0 and standard deviation σ, i.e., $\varepsilon_{ij} \sim N(0, \sigma)$
    * To check assumptions, normality - check normal quantile plot of residuals, constant variance - satisfied if largest standard deviation is less than twice the smallest standard deviation; if violated, try transforming the data; this assumption check becomes more difficult with an unbalanced layout
    * σ estimate is s

- One-Way ANOVA F-test for Balanced, Completely Randomized Designs

- Specify the null ($H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$) and alternate hypothesis (H_A: not all of the μ_j are equal; i.e., at least two means are different)
- Choose significance level α
- Collect data: p independent SRSs of sizes $n_1 + n_2 + \cdots + n_p = n$, estimate parameters $\mu_1, \mu_2, \ldots, \mu_p$ by $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_p$ and standard deviation s_j for each group j
- Compute test statistic assuming the null hypothesis is true

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{\frac{\sum_{i=1}^{p} n_j(\bar{y}_j - \bar{y})^2}{p-1}}{\frac{\sum_{j=1}^{p}(n_j-1)s_j^2}{n-p}}$$

  with an $F(p-1, n-p)$ distribution
- Compute the p-value and reject H_0 if p-value $< \alpha$
- Assumptions: independent observations, normally distributed measurements in each group with the same population standard deviation

- One-Way ANOVA Analysis: SST = SSG + SSE

source	df	sum of squares	mean square (variance)	F
groups	$p-1$	$\text{SSG} = \sum_{j=1}^{p} n_j(\bar{y}_j - \bar{y})^2$	$\text{MSG} = \frac{\text{SSG}}{p-1}$	$F = \frac{\text{MSG}}{\text{MSE}}$
error	$n-p$	$\text{SSE} = \sum_{j=1}^{p}(n_j-1)s_j^2$	$\text{MSE} = \frac{\text{SSE}}{n-p}$	-
total	$n-1$	$\text{SST} = \sum_{i,j}(y_{ij} - \bar{y})^2$	-	-

  Note: $n_1 + n_2 + \cdots + n_p = n$ and s_j is the treatment group sample standard deviation

- Degrees of Freedom
  - For linear regression:

source	df
regression	k
error	$n-(k+1)$
total	$n-1$

  - For one-way ANOVA:

source	df
groups	$p-1$
error	$n_1 + n_2 + \cdots + n_p - p = n - p$
total	$n_1 + n_2 + \cdots + n_p - 1 = n - 1$

  - In regression, k = number of explanatory variables = number of slopes
  - In one-way ANOVA, p = number of groups; note that if a categorical variable has p types, then $p-1$ dummy variables are needed, leading to $p-1$ slopes

- Sum of Squares

  - For linear regression:

source	sums of squares
regression	$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
error	$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
total	$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

  - For one-way ANOVA:

source	sum of squares
groups	$\text{SSG} = \sum_{j=1}^{p} n_j(\bar{y}_j - \bar{y})^2$
error	$\text{SSE} = \sum_{j=1}^{p}(n_j - 1)s_j^2$
total	$\text{SST} = \sum_{i,j}(y_{ij} - \bar{y})^2$

  - $\hat{y}_i$ in regression is $\bar{y}_j$ in one-way ANOVA
  - In each group jm there is only one prediction: $\bar{y}_j$
  - Note that

$$s_j^2 = \frac{\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}{n_j - 1}$$

- Comparing Means
  - One-way ANOVA F-test is used for testing all means at once (like an overall F-test in multiple regression), but i is not useful to see which treatment group means are significantly different from each other (i.e., pairwise difference)
  - If there are treatment pairs we are particularly interested in when designing the experiment, construct pairwise confidence intervals
    - Confidence interval for a single treatment j:

$$\bar{y}_j \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n_j}}\right)$$

    - Pairwise confidence interval - difference between two treatments j and j^*:

$$(\bar{y}_j - \bar{y}_{j^*}) \pm t_{\alpha/2} s \sqrt{\frac{1}{n_j} + \frac{1}{n_{j^*}}}$$

      where $t_{\alpha/2}$ is the upper tail value on the t-distribution with probability $\alpha/2$ with $n - p$ degrees of freedoms and s is the pooled standard deviation

- Pooled Standard Deviation
  - One of the one-way ANOVA assumptions is that the population standard deviation σ is the same across all p groups
  - This means that the sample standard deviation of each group s_j is an estimate of the same population standard deviation value

  - A better estimate of σ is calculated by combining data across the p groups, forming the pooled standard deviation
  - σ estimated by the pooled standard deviation

$$s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_p-1)s_p^2}{(n_1-1)+(n_2-1)+\cdots+(n_p-1)}}$$
$$= \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_p-1)s_p^2}{n-p}}$$

    if $n_1 + n_2 + \cdots + n_p = n$
  - s is a weighted average of the treatment group standard deviations (basically weighed by sample size)

- Multiple Comparisons
  - Type I error - reject null hypothesis when it is true
  - $P(\text{Type I error}) = \alpha$ for each hypothesis test
  - When many tests are done at once, the combined probability of a Type I error is higher than α across all tests
  - In multiple regression, the Bonferroni correction was used to check for the significance of individual explanatory variables
  - If all pairwise confidence intervals of differences were constructed, the same problem would occur
  - Solution: post-hoc comparisons of all pairwise treatment means: Bonferroni's Method, Scheffe's Method, Tukey's Method
  - These methods will give an experimentwise Type I error rate of α as opposed to a comparisonwise rate of α

- Tukey's Method
  - There are two versions: balanced and unbalanced - most common method used
  - The unbalanced version is approximate
  - This method is also called the Tukey-Kramer method or Tukey's HSD (honestly significant difference)

- Tukey's Multiple Comparisons Procedure (balanced)
  - Set α level (i.e., Type I error rate)
  - Assume independent observations, normality and equal σ across groups
  - Hypotheses:

$$H_0 : \mu_j = \mu_{j'}$$
$$\mu_j \neq \mu_{j'}$$

  - Pairwise confidence interval

$$\bar{y}_j - \bar{y}_{j'} \pm q_\alpha(p, n-p)\frac{s}{\sqrt{n'}}$$

    * $j \neq j'$, $\bar{y}_j - \bar{y}_{j'}$ is a pairwise difference in means
    * p is the number of treatments
    * s = RMSE, from one-way ANOVA model
    * n' - number of observations within a treatment
    * $q_\alpha(p, (n'p) - p)$ - critical value of the studentized range distribution, degrees of freedom is $(n'p) - p$ (same as RMSE degrees of freedom)
  - If 0 is in the interval, then there is no a significant difference between groups j and j'; in other words, significantly different pairs have confidence intervals which do not include 0
  - Pairwise test statistic:

$$\frac{\bar{y}_j - \bar{y}_{j'}}{s/\sqrt{n'}}$$

  - Pairwise p-value: probability of being greater than the absolute value of the test statistic on a studentized range distribution for a level α test, p treatment groups and degrees of freedom $(n'p) - p$

- When to use p-value Adjustments
  - Before data is collected, if here are only one or two pairs of differences, adjustments are not needed because the total number of hypothesis tests being done is low
  - Post-hoc test: need to be conservative when checking all pairs - use Tukey's HSD (or Bonferroni, etc.) - these adjustments are there to reduce errors

- Order Statistics
  - Order statistics is a subfield of statistics and is the study of the statistical properties of ordered/ranked data; minimum, maximum, median, quantiles are all examples of statistics based on ordered/ranked data
  - n independently and identically distributed observations: $x_1, \ldots, x_n$
  - Let $x_{1:n}$ denote the ith order statistic, this means
    * $x_{1:n}$ denotes the smallest observation in the data set $\rightarrow$ first order statistic (i.e., minimum)
    * $x_{2:n}$ denotes the second smallest observation in the data set $\rightarrow$ second order statistic
    * This keeps going on
    * $x_{n:n}$ denotes the largest observation in the data set $\rightarrow$ nth order statistic (i.e., maximum)

  - Let $x_1, \ldots, x_n$ be iid random variables drawn from a continuous distribution with pdf $f(x)$ and cdf $F(x)$, then the pdf of the ith order statistic, $x_{i:n}$ is

$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!}(F(x))^{i-1}(1-F(x))^{n-i}f(x)$$

- Studentized Range Distribution
  - Recall: the null hypothesis for the Tukey tests are that there are no difference in means between the two groups
  - Draw n' independent observations from p independent populations which have the same distribution, $N(\mu, \sigma)$:
    * $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p$ - p sample means
    * $\bar{x}_{1:p}$ - smallest sample mean
    * $\bar{x}_{p:p}$ - largest sample mean
  - Studentized range distribution is the distribution of the following random variable

$$q = \frac{\bar{x}_{p:p} - \bar{x}_{1:p}}{s/\sqrt{n'}}$$

    where s is the pooled standard deviation
  - The distribution of q is defined by p, the number of groups, and the degrees of freedoms, $n - p$, where n is the total sample size in the balanced case: $n = n'p$