Response: **Random** Variable...varies w/ changes to predictor
Predicting: **Fixed** Variable...Does not change w/ response

y Actual
$\bar{y}$ Mean
$\hat{y}$ Estimated

Simple Linear Regression

Uses of Regression:
1. Prediction 2. Modelling 3. Testing

$$Y = \beta_0 + \beta_1 x + \epsilon \quad \leftarrow \text{Deviance from Linear Model}$$

Assumptions:
1. Linearity/Mean Zero Assumption
2. Constant Variance Assumption
3. Independence (random variables)
4. Normal distribution of deviance

Notes on SLR:
- Model parameters (beta_0, beta_1, sig^2) are always unknown (estimated)
- Minimize sum of squared errors

Fitted Values and Residuals:

Fitted Values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1$
Residuals: $r_i = \hat{\epsilon}_i = y_i - \hat{y}_i$

MSE (Estimator for sig^2)
$MSE = \frac{\sum_{i=1}^{n} r_i^2}{n-2} = \frac{SSE}{n-2}$

$Std.Err. = \frac{Estimate}{t-value}$

Variance Sampling Distribution:

$$\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n-2} \sim \chi^2_{n-2} \quad \text{n-2 DOF}$$

Sample Variance Estimation:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{n-1 DOF}$$

Model Interpretation: MAKE SURE DISTINGUISH BETWEEN PREDICTION AND EXTRAPOLATION

Sampling distribution of beta_1:
- Do not know the true variance... aproximate using MSE.
- This changes sampling distribution to t-distribution with n-2 df

$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{S_{XX}}}}$ ~ t_(n-2)

Confidence interval calculation:

$$\left(\beta_1 - \left(t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\text{MSE}}{S_{XX}}}\right), \beta_1 + \left(t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\text{MSE}}{S_{XX}}}\right)\right) \qquad t_{\frac{\alpha}{2},n-2} \leftarrow \text{t-critical point}$$

$$t-value = \frac{\hat{\beta}_1 - 1}{\sqrt{\frac{\sigma^2}{S_{XX}}}} = \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)}$$

Testing statistical significant:

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

Reject null if |t-value| is large or p-value is small
p-value: How rejectable is the null hypothesis?
Testing Assumptions
- Linearity Assumption (Residuals vs X)
- Constant Variance (Residuals vs fitted values)
- Independence (No clusters of residuals)
- Normality (Straight normality plot -- q-q OR histogram) -error terms are normal

Testing statistically positive:

$H_0 : \beta_1 \leq 0$
$H_1 : \beta_1 > 0$

Interested in left side of t-tail

P-value = P(T_{n-2} > t-value)
P-value = 1- pt(tvalue, dof)

Variable Transformations (Needed if relationship between X and Y is not linear)
- Box-Cox Transformation
  - Transform response variable from y to y^* $\quad y^* = y^\lambda$
  - Normality or constant variance assumption does not hold

1 (don't transform)
1/2 (square root)
0 (ln)
-1 (1/y)

Outliers (May or may not impact the fit of the regression)
- Leverage points --> Data points far from the mean of the x's
- Influential points --> Far from the mean of either or both of the x's and y's (influence fit)

Check for outliers using standardized residuals
- Greater than 1 --> Large
- Greater than 2 --> Extremely Large

$$r_i^* = \frac{y_i - \hat{y}_i}{\sqrt{MSE}}$$

Coefficient of Determination (r) --> % of variation in Y that can be explained by X

$$R^2 = 1 - \frac{SSE}{SST} \qquad SSE = \sum_{i=1}^{n} r_i^2 \qquad SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Correlation Coefficient = Coefficient of variation (determination)

$$\rho = cor(X,Y) = \hat{\beta}_1 \sqrt{\frac{S_{XX}}{S_{YY}}}$$

ANOVA (Analysis of Variances)

Objectives
1. Analysis of variablilty in the data
2. Testing for equal means $\quad H_0 : \mu_1 = \mu_2 = ... = \mu_k \quad H_A$: Some means are different
3. Estimation of simultaneous confidence intervals for the mean $\mu_i - \mu_j$ for i and j=1,...,k

Assumptions
- Constant Variance Assumption
- Independence Assumption: $\{\sigma_{1j}, ..., \sigma_{kj}\}$ are independent random variables
- Normality Assumption: Residuals approximate to normal distribution

Pooled Variance Estimator: $S^2_{pool} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}{N-k}$ DOF where N = total number of samples
MSE
Sum of Squared Errors (SSE): $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$

Sampling distribution of pooled variance is chi-square dist. with N-k DOF

Confidence Interval: $\left(\hat{\mu}_i - t_{\alpha/2,\,N-k}\sqrt{\text{MSE}/n_i}, \hat{\mu}_i + t_{\alpha/2,\,N-k}\sqrt{\text{MSE}/n_i}\right)$

SST: DOF = N-1, only lose one DOF because estimating one mean instead of k w/ MSE
Notes:
- MSE = SSE/(N-k) = within-group variability
- MSST_R = SST_R/(k-1) = between-group variability
- ANOVA = Compare between to within variability
- F = between group variability / within-group variability

Reject null if F_0 > F_a(k-1, N-k) -- one sided test; reject null if p-value is less than alpha
- Must test for assumptions prior to running ANOVA. Plot residuals, q-q, histogram of residuals
- Can also perform transformation if not normal and histogram is not normal

k groups --> k+1 parameters to estimate (k group means + pooled variance estimator)

Multiple Linear Regression (chi-squared, n-p-1 dof) ← Assuming errors are normally distributed and intercept (n-p w/o)
Same assumptions as linear regresion
Includes quantitative and qualitative predictors
# of predictors: If quantitative w/ intercept then n+2, qualitative w/ intercept then k-1

Types of Variables:
Controlling - Control for bias selection in the sample. Capture more meaningful relationships
Explanatory - Explain variability in response variable
Predictive - Best to predict variability in response regardless of explanatory power

Variance estimate = (Standard error)^2

Variability Source	DF	Sum of Squares	Mean SS	F-Statistic
Regression	p	SSReg	SSReg / p	MSSReg / MSE
Residual	n-p-1	SSE	SSE / (n-p-1)	
Total	n-1	SST		

MLR is generalization of SLR and ANOVA

Confidence Intervals: Not normally distributed --> t-distribution wiht n-p-1 DF
A function of estimate, t-critical point and standard error of estimate

Uncertainty in new prediction:
- Uncertainty in parameter estimates
- Uncertainty in newness of data

Marginal versus Conditional:
- Marginal: Simple linear regression captures the association of a predicting variables to the response varaible marginally, i.e., w/o consideration of other factors
- Conditional: MLR captures association of a predicting variable to response conditionally i.e., conditional of all other predicting variables in the model

Parameter estimates normally distributed amongst actual with variance because errors are assumed to be normally distributed about 0 with deviation of sigma squared. Unbiased regardless of the distribution of the data

Confidence Interval Estimation $\hat{\beta}_j \pm (t_{\frac{\alpha}{2},n-p-1})(SE(\hat{\beta}_j))$

To see if \beta_j is statistically significant..check if 0 is in conf int

Use t-test for statistical significance given all other predicting parameters

$H_0 : \beta_j = 0 \quad H_a : \beta_j \neq 0$

Reject null hypothesis (p-value < alpha) = parameter is statistically significant

Test for subset of coefficients using a partial F-test. Ha is that at least one coefficient is non-0
Reject null hypothesis if p-value is less than alpha or F-Statistic > F-critical point

Number of parameters = Num of Quant Predictors + Intercept + error term + (k-1 dummy var)

Testing Subset of Model
- Perform F-test using full model and subset of model
- If p-value is small reject null hypothesis and conclude at least one of the excluded predictors is statistically non-zero

Predictions
- Prediction is not the same as regression line estimation
- Two uncertanties in predictions (parameter deviances and uncertain new observations)
- Note: Prediciton **line** is the same as the estimated regression **line**

Outliers
- Cooks Distance (how much fitted values change when ith obs removed)
  - Rule of Thumb Di > 4/n, Di > 1d

Residual Analysis
- Constant Variance & Uncorrelated errors
  - Response/fitted values vs residuals
- Linearity
  - Predicting Variables vs residuals
- Normality
  - Histogram and QQ plot of std residuals

Testing Full Model
- Use F test to evaluate the overall regression

Multicollinearity (test using VIF < max(10, 1/(1-R2_m)
- High VIF -> Collinearity
- Symptoms
  - Estimated coefficients are unstable
  - Standard error of coeff is artificially large
  - F-statistic is significant but indivual t-tests are not
  - No inflation/deflation of R2
  - *columns* in design matrix are linearly depend.

$VIF_j = \frac{1}{1-R_j^2}$

Adjusted R2 vs R2
- Adjusted R2 is used for comparing different models, not typically used as a measure of the amount of variance explained by an individual model (this is the R2), always <= R2
- Adjusted R2 is a modified version of R-squared that has been adjusted for num of predictors in the model

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

n: number of data points
k: number of variables excluding the constant

Logistic Regression
- Model the probability of success given teh predictors. No error term!

Assumptions
- Linearity $\quad g(p(X_1, ..., X_p)) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$
- Independence: Y1,...,Yn are independent random variables
- Logit Link (log odds) Function: $\quad g(p) = ln(\frac{p}{1-p})$

Interpretation
- log odds : $\quad ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x$
- exponential of the logit function : $\quad \frac{p(x)}{1-p(x)} = e^{\beta_0 + \beta_1 x}$
- odds ratio at X=a versus X = b : $\quad \frac{e^{\beta_0 + \beta_1 a}}{e^{\beta_0 + \beta_1 b}} = e^{\beta_1 (a-b)}$

Regression coefficient is interpreted as the log of the odds ratio for an increase of one unit in the predicting variable

If X is a dummy var, interpret as the log of odds ratio of one category vs baseline