Regression

What assumptions about our data and model do we make?

E[εi] = 0, Var(εi) = σ2, and Cov(εi,εj) = 0. εi ∼N(0, σ2)

Or, using the acronym from class: (L)inearity, (I)ndependence, (N)ormality, (E)qual variance / homoscedasticity

Ordinary Least Squares : (parameter estimation) minimizing the Sum of Squared Errors (SSE) or Residual Sum of Squares with respect to β0 and β1

• To model or predict the values of an outcome of interest.

• To explore and quantify the relationship between an exposure(s) and an outcome.

**Regression Diagnostics**

1. **Linearity.** Plot the residuals against the predicted value, Yˆ. The residuals should be scattered randomly above and below the zero line, with no discernible pattern.

2. **Normality.** Look at a histogram of the residuals. If the normality assumption holds, this histogram should approximate a normal distribution—meaning it should be symmetric and centered around 0.

Anderson-Darling test of goodness-of-fit to a specified continuous univariate probability distribution.

3. **Equal Variance.** Again look at the plot of the residuals against the predicted value. The residuals should be just as “spread out” for large values of Yˆ as they are for small values of Yˆ. Breusch-Pagan test

Regress residuals squared on y\_hat, test whether the slope of y hat is 0.

**confidence interval**: for average response. The width of the CI for E(Y) increases as x moves away from the center(mean x). only accounts for the uncertainty in estimating the mean value.

**prediction interval**: on a specific observation. Accounts for the uncertainty in parameter estimation and the tendency of y to fluctuate from its mean value. Width also increases as x moves away from mean x.

A **confounder** is a variable that is associated with both our outcome of interest and the exposure, *but that is not a consequence of the exposure*. we’ll consider a variable to be a meaningful confounder of the relationship between X and Y if adjusting for it changes the estimated slope by 10% or more.

**effect modifier:** if the magnitude of the association between our predictor and our outcome varies across its different levels

**Diagnostics :** residual evaluation and outlier detection

**Residual plots**: Histogram of crude or standardized residuals Should appear *Normally* distributed; Histogram of studentized or jackknife residuals Will roughly approximate a t distribution (or Normal when N is large); Boxplot Residuals should be *symmetrical* without any residuals beyond the whiskers. Q-Q plot Residuals should follow a straight line; Scatter plot against of Y Should be a fluffy, patternless cloud(linearity, equal variance).

Outlier Detection : An influential point is an outlier that greatly a↵ects the slope of the regression line. Report inferences both with and without outliers(avoid overfitting). If things don’t change substantially, go with all data. The existence of many outliers is often an indication that either linear model assumptions are violated, there is a problem with data collection, or the data itself requires a different or more complex model.

The goal is to trace back to the potential sources of data contamination and eliminate them. If the proportion of outliers is not substantial and negligible, they can be dropped. If outliers are not caused by errors/they are true outliers by nature, there are several things we can do: drop the outliers if we can justify this operation, transform or standardize the variables, impute the erroneous values

**Leverage**: measures how far observation i’s covariates are from the overall covariate average. Not necessarily influential

**DF Beta** statistic: quantifies how much a coefficient changes if we delete a particular observation.

**DFFITS** : a measure of how influential a point is to the fit of a line.

**Cook’s Distance** : measures how much the regression parameters would change if observation i is deleted.

**Non-Linear Model Fitting:**

**Smoothing**: non-parametric model . Parametric models are models with consistent ”shape”. They require assumptions and in return allow you to summarize a data relationship in terms of parameters. Non-parametric models require fewer (if any) assumptions and are much more flexible, but may be hard to summarize. They are useful for both visualization and prediction.

LOWESS is a locally weighted running line smoother of Y over a neighborhood of x (which ”moves” down the x axis). The algorithm gives more weight to points in the ”middle” of the neighborhood.

**Categorized Covariate**: turn the covariate into a categorical variable, aka putting groups of similar values into ”bins”.

**Polynomial/Nonlinear Terms:** higher/fractional polynomial terms, piecewise

**Spline:** a piecewise parametric model which separates data along the x axis by knots. Constant, linear, cubic

**Generalized Additive Models (GAM):** outcome is modeled as a sum of functions of covariates. Semi-parametric; 4 df=linear trend+wriggly fit. The non-parametric ANOVA shows how much more square error an additional 3 degrees of freedom ac- counts for.

**Variable Selection**

In controlled settings, such as randomized clinical trials, variable subset selection is usually less critical. the expectation of balance amongst all covariates, including those we haven’t even measured. In contrast, observational studies typically collect a large number of potentially predictive variables without strong prior information of which covariates are important predictors of the outcome.

**What kinds of problems should we worry about when removing observations from a data set?:**

We worry when observations that are being dropped share a certain characteristic. This is usually very apparent in clinical trials, where, for example, the observations being dropped due to missingness could be those patients who are the worst o . If those patients all share a common treatment, that would introduce some bias in our crude result that we would have to consider how to adjust for.

Forward Selection starts with an intercept-only model and will add in the most significant covariate (according to p-value) as long as that covariate is significant at the slentry-level (here, 0.15). The process repeats until all possible covariates are added.

Backward Elimination starts with the full model and will remove the least significant covariate (according to p-value) as long as that covariate is not significant at the slstay-level (here, 0.15). The process repeats until all possible covariates are eliminated.

Stepwise Selection can start o either Forwards or Backwards (either with the full model or intercept-only model). It then completes each round by adding and eliminating covariates from the model according to the slentry and slstay levels.

searching throughout the entire subset space. We then choose the model that minimizes or maximizes certain criteria (AIC, Root Mean Squared Error, adjusted R-squre)

If X1\*X2 is included in the model and X1^2 also exists, do include the quadratic X1\*X2.

Violation of assumptions:

as long as our model is correctly specified (i.e. as long as the linearity assumption holds) and our errors (residuals) have mean 0, our estimates βˆ will be unbiased for the true β! This holds regardless of whether or not we have independence, normality of the residuals, or constant variance.

Relaxing assumptions: if linear and independent assumptions hold:

Unbiased point estimates, robust standard error method

The robust variance estimators are large sample variances, meaning that our t-tests and F-tests no longer hold exactly—we need to be concerned about this if the sample size is small. We also have the drawback that such measures as R2, AIC, BIC, etc. can no longer be calculated.

**Multicollinearity** exists when two or more of the predictors in a regression model are moderately or highly correlated.

There are two types of multicollinearity:

* **Structural multicollinearity** is a mathematical artifact caused by creating new predictors from other predictors — such as, creating the predictor *x*2from the predictor *x*.
* **Data-based multicollinearity**, on the other hand, isa result of a poorly designed experiment, reliance on purely observational data, or the inability to manipulate the system on which the data are collected.

Uncorrelated predictors:

* The estimated slope coefficients *b*1 and *b*2 are the same regardless of the model used.
* The standard errors *se*(*b*1) and *se*(*b*2) don't change much at all from model to model.
* The sum of squares *SSR*(*x*1) is the same as the sequential sum of squares *SSR*(*x*1|*x*2). The sum of squares *SSR*(*x*2) is the same as the sequential sum of squares *SSR*(*x*2|*x*1).

the effect on the response ascribed to a predictor doesn't depend on the other predictors in the model.

If the primary purpose of your regression analysis is to estimate a mean response *μ*Y or to predict a new response *y*, you don't have to worry much about multicollinearity.

When predictor variables are correlated, the marginal contribution of any one predictor variable in reducing the error sum of squares varies depending on which other variables are already in the model.

When predictor variables are correlated, the precision of the estimated regression coefficients decreases as more predictor variables are added to the model.

Some of the common methods used for detecting multicollinearity include:

* The analysis exhibits the signs of multicollinearity — such as, estimates of the coefficients vary from model to model.
* The *t*-tests for each of the individual slopes are non-significant (*P* > 0.05), but the overall *F*-test for testing all of the slopes are simultaneously 0 is significant (*P* < 0.05).
* The correlations among pairs of predictor variables are large.

**variance inflation factors** (*VIF*) :  quantifies how much the variance is inflated. 1/(1-Rk^2)

The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

Regress each explanatory variable on all other explanatory variables. Get the coefficient of determination of these models. The higher the multicollinearity the higher the VIF.

For data-based:

remove one or more of the violating predictors from the regression model.

Another way is to collect additional data under different experimental or observational conditions.

For structure-based:

**centering the predictors**." Centering a predictor merely entails subtracting the mean of the predictor values in the data set from each predictor value.

**extrapolation**

**Nonconstant Variance: variance-stabilizing transformation**(logarithm, square root, **weighted least squares**, A generalization of weighted least squares is to allow the regression errors to be correlated with one another in addition to having different variances.

 the regression depth of a hyperplane H is the smallest number of residuals that need to change sign to make H a **nonfit**.1 A regression hyperplane is called a nonfit if it can be rotated to horizontal (i.e., parallel to the axis of any of the predictor variables) without passing through any data points.

**Handle missing data:**

1. Listwise deletion: remove all observations with any missing variable (not substantial missingness, MCAR)
2. Imputation
3. Marginal mean imputation: the mean of x using all non-missing x
4. Conditional mean imputation: regress x on other covariates, predict the missing values with model

Underestimate standard errors, overestimate test statistics

1. Multiple imputation:

Regress the missing variable multiple times with the previously imputed data to update the full dataset. combine the estimates of model using m complete datasets

Multiple imputation has solved this problem by incorporating the uncertainty inherent in imputation. It has four steps:

1. Create *m* sets of imputations for the missing values using an imputation process with a random component.
2. The result is *m* full data sets. Each data set will have slightly different values for the imputed data because of the random component.
3. Analyze each completed data set. Each set of parameter estimates will differ slightly because the data differs slightly.
4. Combine results, calculating the variation in parameter estimates.
5. Maximum likelihood:

Assume multivariate normality of data.

1. Missing observation correlation: use the correlation between x and y, derive x based on y

**Handle high dimensionality:**

1. Subset selection (forward, backward, stepwise, best subset)
2. Regularization / shrinkage : cross-validation decides tuning parameter
3. Dimension reduction: pca, partial lease square(places the highest weight on the variables that are most strongly related to the response, regress the residuals after the first component on feature space to identify the second component, finally regress y on all components )

When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.  linear discriminant analysis is popular when we have more than two response classes.

Classical statistical parametric tests compare observed statistics to theoretical sampling distributions. Resampling a data-driven, not theory-driven methodology which is based upon repeated sampling within the same sample.   
  
Resampling refers to methods for doing one of these

* Estimating the precision of sample statistics (medians, variances, percentiles) by using subsets of available data (jackknifing) or drawing randomly with replacement from a set of data points (bootstrapping)
* Exchanging labels on data points when performing significance tests (permutation tests, also called exact tests, randomization tests, or re-randomization tests)
* Validating models by using random subsets (bootstrapping, cross validation)

The jackknife [estimator](https://en.wikipedia.org/wiki/Estimator) of a parameter is found by systematically leaving out each observation from a dataset and calculating the estimate and then finding the average of these calculations. Given a sample of size n, the jackknife estimate is found by aggregating the estimates of each n−1-sized sub-sample.

Selection bias, in general, is a problematic situation in which error is introduced due to a non-random population sample.

Hadoop consists of the *Hadoop Common* package, which provides file system and operating system level abstractions, a MapReduce engine (either MapReduce/MR1 or YARN/MR2)[[59]](https://en.wikipedia.org/wiki/Apache_Hadoop#cite_note-59) and the [Hadoop Distributed File System](https://en.wikipedia.org/wiki/Apache_Hadoop#Hadoop_distributed_file_system) (HDFS).

There are several statistical tests that can be used to assess whether data are likely from a normal distribution. The most popular are the Kolmogorov-Smirnov test, the Anderson-Darling test, and the Shapiro-Wilk test1.

**Newton–Raphson method**: is a method for finding successively better approximations to the [roots](https://en.wikipedia.org/wiki/Root_of_a_function) (or zeroes) of a [real](https://en.wikipedia.org/wiki/Real_number)-valued [function](https://en.wikipedia.org/wiki/Function_(mathematics)). The method starts with a function *f* defined over the [real numbers](https://en.wikipedia.org/wiki/Real_number) *x*, the function's [derivative](https://en.wikipedia.org/wiki/Derivative) *f ′*, and an initial guess *x*0 for a root of the function *f*. until a sufficiently accurate value is reached.

**Markov chain Monte Carlo** (**MCMC**) methods are a class of [algorithms](https://en.wikipedia.org/wiki/Algorithm) for sampling from a [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution) based on constructing a [Markov chain](https://en.wikipedia.org/wiki/Markov_chain) that has the desired distribution as its [equilibrium distribution](https://en.wikipedia.org/wiki/Markov_chain#Steady-state_analysis_and_limiting_distributions). The state of the chain after a number of steps is then used as a sample of the desired distribution. The quality of the sample improves as a function of the number of steps.

Gibbs: approximates a distribution fi by sampling alternatively from its full conditional distributions.

 The goal of multivariate testing is to determine which combination of variations performs the best out of all of the possible combinations.

which are tested concurrently to find the winning variation.

 a multivariate test can eliminate the need to run several sequential A/B tests on the same page with the same goal.

The most difficult challenge in executing multivariate tests is the amount of visitor traffic required to reach meaningful results.

Another challenge of multivariate testing is when one or more of the variables being tested do not have a measurable effect on the conversion goal. (effects offset)

Box-cox transformation:  a **power transform** is a family of functions that are applied to create a [monotonic transformation](https://en.wikipedia.org/wiki/Monotonic_function) of data using [power functions](https://en.wikipedia.org/wiki/Power_function). This is a useful [data transformation](https://en.wikipedia.org/wiki/Data_transformation_(statistics)) technique used to stabilize variance, make the data more [normal distribution](https://en.wikipedia.org/wiki/Normal_distribution)-like, improve the validity of measures of association such as the [Pearson correlation](https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient) between variables and for other data stabilization procedures.

Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing or stretching. Eigenvalue can be referred to as the strength of the transformation in the direction of eigenvector or the factor by which the compression occurs.

Simple random sample:

a simple random sample is a set of n objects in a [population](http://www.statisticshowto.com/what-is-a-population/)of N objects where all possible samples are equally likely to happen.

A simple random sample is chosen in such a way that every set of individuals has an equal chance to be in the selected sample.

Make a list all objects in the population-> assign a sequential number to all objects->

Use a random number generator to select the sample.

With random sampling, each object does not necessarily have an equal chance of being chosen.

There are two major categories of sampling methods ([figure 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5325924/figure/F1/)): 1; **probability sampling** methods where all subjects in the target population have equal chances to be selected in the sample [[1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5325924/#B1),[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5325924/#B2)] and 2**; non-probability sampling methods** where the sample population is selected in a non-systematic process that does not guarantee equal chances for each subject in the target population [[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5325924/#B2),[3](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5325924/#B3)]. Samples which were selected using probability sampling methods are more representatives of the target population.

***Simple random sampling***

when the whole population is accessible and the investigators have a list of all subjects in this target population.

***Stratified random sampling***

it requires the condition of sampling frame being available

 whole population is divided into homogeneous strata or subgroups according a demographic factor

 it allows obtaining samples from minority/under-represented populations.

***Systematic random sampling (Interval sampling)***

include the last patient from every 5 patients.

***Cluster sampling (Multistage sampling)***

It is used when creating a sampling frame is nearly impossible due to the large size of the population.

the population is divided by geographic location into clusters. A list of all clusters is made and investigators draw a random number of clusters to be included.

list all individuals within these clusters, and run another turn of random selection to get a final random sample exactly as simple random sampling.

Power:

Statistical Power is the probability that a statistical test will detect differences when they truly exist.

Sample size, significance level, estimated effect( the difference between the null and the alternative), the standard deviance of the population.

CI:

If repeated samples were taken and the 95% **confidence interval** was computed for each sample, 95% of the **intervals** would contain the population mean. A 95% **confidence interval** has a 0.95 probability of containing the population mean.

Ancova:

is a general linear model that includes both ANOVA (categorical) predictors and Regression (continuous) predictors.

Convex set:

A subset of Euclidean space such that a line drawn between any two points in the subset remains completely within the subset.

gradient descent

A technique to minimize [**loss**](https://developers.google.com/machine-learning/crash-course/glossary#loss) by computing the gradients of loss with respect to the model's parameters, conditioned on training data. Informally, gradient descent iteratively adjusts parameters, gradually finding the best combination of [**weights**](https://developers.google.com/machine-learning/crash-course/glossary#weight) and bias to minimize loss.

GLM: is a method to model a function of the dependent variable using the linear combination of predictors. The dependent variable can not be modeled directly because the error distribution is not normal distribution. GLM is able to model dependent variables whose distribution belongs to the exponential family.

GEE: generalized estimating equation is a way of modeling repeated measurements in the framework of GLM. We don’t restrict the data type of the response variable and take the covariance structure of responses into consideration. So apart from the link function and linear predictor setup, we also need to specify the covariance structure and acquire estimation of it.

Log-linear:

 when there is no clear distinction between response and explanatory variables, or there are more than two responses. This is a major difference between logistic models and log-linear models. In the former a response is identified, but no such special status is assigned to any variable in log-linear modelling.

assume discrete variables to be nominal, but these models can be adjusted to deal with ordinal and matched data.

The *N* = *I × J* counts in the cells are assumed to be independent observations from a Poisson random variable

chi-square test of independence, but now we want to model the cell counts with the *log-linear model of independence*and ask if this model fits well.