Lecture 2: Urban Analytics in Philadelphia

Suburbanism	Urbanism
Low population density	High population density
Safety by **separation** of commercial and residential	Safety through **natural surveillance** and proximity
Automobile-centered designs promote mobility	**Pedestrian**-centered designs promote human interaction
Economic success through efficient **mobility**	Economic success through **proximity** of people
Top down, large scale development **by appointed few** to maximize efficiency and accessibility	Organic development **driven by residents** that have a stake in their surroundings
Appears clean and **organized**	Appears **chaotic**

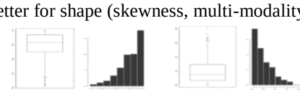
Lecture 3: Exploratory Data Analysis I

Categorical Variables: place an individual into one of several groups (gender, race, type of crime). Bar plots → x-axis = categories, y-axis = count. Can also use pie charts, but worse because you can't see the actual totals of each category and hard to see small categories.

Continuous Variables: take on numerical values across an entire range (height, income, population). Distribution describes what values a variable takes and how frequently these values occur. Center, spread, shape, outliers. Histograms are same idea as bar plots, but first divide the continuous range of values into bins. X-axis = bins.

Boxplots: Box contains the central 50% of the values, Median is the line with 50% of values on either side, Mean is the diamond, Whiskers contain most of the rest of distribution (except for outliers). Boxplots are useful for displaying center and spread, plus outliers. Histograms are better for shape (skewness, multi-modality—high frequency values).

Left Skewed: Mean < med.



Right Skewed: Mean > med.

Trimmed mean: chop top/bottom % from data then take the mean. Throwing away all data except middle value = median.

Standard Deviation: the variance is the average of the squared deviations of each observation. ---------------→ $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

Interquartile Range: IQR=Q2-Q1 → robust to outliers

Lecture 4: Exploratory Data Analysis II

Log Transformations: describes the magnitude of a value instead of the value itself → $\log_{10}(100)=2$ versus $\log_{10}(1000)=3$. Make skewed distributions look more symmetric.

Correlation: a value between -1 and 1 that provides both the sign and strength of a linear association between two variables. How "tight" the linear relationship is.

Lecture 5: Probability

Long run interpretation of probability: P(X=x) is the proportion of times that random variable X takes the value x over an infinitely long sequence of repetitions of an experiment or process.

Normal Distribution: denoted N(μ,s)

In Standard Normal distribution→ (μ=0,s=1).

68% of observations are between mean +/- 1 SD. 95% are between mean +/- 2 SD.

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Standardization: if X has non-standard distribution, convert Z into standard:

$$\boxed{Z = \frac{X-\mu}{\sigma}}$$

Sampling Distribution of Mean:

$$\mathrm{VAR}(\bar{X}) = \frac{\sigma^2}{n} \qquad \mathrm{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

As sample size increases, spread of the sample mean decreases.

Central Limit Theorem: if the sample size is large enough, then the sample mean has an approximately Normal distribution (no matter the shape of distribution of the original data).

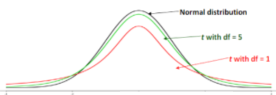
Lecture 6: Inference

Confidence Interval for a Population Mean: The 100C% confidence interval for a population mean is -------------→

$$\left(\bar{X} - t^*_{n-1} \cdot \frac{s}{\sqrt{n}} \quad , \quad \bar{X} + t^*_{n-1} \cdot \frac{s}{\sqrt{n}}\right)$$

(It comes from a t-distribution with n-1 degrees of freedom).

The smaller the degrees of freedom, the thicker the tails of the t distribution →



Null (H_0): usually an assumption that there is no difference or effect.

Alternative (H_a): states that there is a substantive difference or effect.

Q: "How many standard errors is our observed sample value from the hypothesized value?"

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

A: If p-value < α then reject null hypothesis.

Two-sided Testing: must *double the p-value* to account for probability in both tails of t-distribution.

Connection between Intervals and Tests: if our confidence level C is equal to 1- α then a two-sided hypothesis test rejects the null hypothesis (=) if our hypothesized value μ falls outside the CI.

	H_0 True	H_0 False
Reject H_0	Type I Error	✓
Don't Reject H_0	✓	Type II Error

Type 1 = False positive = α; Type 2 = false negative.

Bonferroni Correction: instead of using α-level of 0.05, we use an α-level of 0.05/*m* where *m* is the number of tests being done.

Lecture 7: Association between Categorical Variables

	Poor	Not Poor	Total
Black	0.072	0.377	0.449
Not Black	0.066	0.486	0.551
Total	0.137	0.863	1.000

There is an association between race and poverty if the probability of poverty changes depending on race. P(poor) = 0.137 and P(poor | black) = 0.072 / 0.449 = 0.160. **Independence rule**: P(poor, black) = P(poor) x P(black) = 0.137 x 0.449 = 0.062 → Expectation!

Chi-Squared Statistic: under null hypothesis of no association, X^2 follows distribution with (rows - 1) x (columns - 1) degrees of freedom.

$$X^2 = \sum_{\text{cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

Fisher's exact test is an alternative that is the better with small counts.

Cramer's V: Chi-squared statistic depends on # of observations and # of categories, making it difficult to compare the strength of association across datasets. If V = 0, no association.

$$V = \sqrt{\frac{X^2}{n \cdot \min(r-1, c-1)}}$$

Lecture 8: Comparison across Categories

Two-Sample Z-Test for Proportions:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\mathrm{SE}(\hat{p}_1 - \hat{p}_2)} \quad \text{where } \mathrm{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

*Don't forget to multiply the p-value by 2 if using a two-sided.

**Stick with normal distributions for proportions

Confidence Interval for Difference in Proportions:

$$\left(\hat{p}_1 - \hat{p}_2 \pm Z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$$

Note: 95% CI → Z* = 1.96

Two-Sample T-Test for Means: Note: df = min (n_1,n_2) -1

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\mathrm{SE}(\bar{X}_1 - \bar{X}_2)} \quad \text{where } \mathrm{SE}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Confidence Interval for Difference in Means -------→

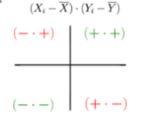
$$\left(\bar{X}_1 - \bar{X}_2 \pm t^*_{df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

Lecture 9: Continuous Variable Association

Covariance between Two Variables: The spread between two variables is captured by the covariance. The test statistic for this hypothesis test for correlation is = T. SE(r) will be given, too complicated.

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}) \cdot (Y_i - \bar{Y})$$

$$r = \mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{SD}(X)\ \mathrm{SD}(Y)} \qquad T = \frac{r - 0}{\mathrm{SE}(r)}$$



Equation of a Line: B_0 is the intercept; B_1 is the slope.

We want the line that has the smallest total residuals, but we must square the residuals and add them up. The line with the smallest total residuals is the least-squared line. S_y is the SD of y variable, etc.

$$\boxed{y_i = \beta_0 + \beta_1 x_i}$$

$$\mathrm{SSR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \qquad \boxed{b_1 = r \cdot \frac{s_y}{s_x} \qquad b_0 = \bar{y} - b_1 \bar{x}}$$

Lecture 10: Interpretation and Prediction with a Linear Model

The slope b_1 is the average change in the Y variable that is associated with a one unit change in the X variable. The intercept b_0 is the average value of the Y variable when the X variable is equal to zero. Caution: the intercept should only be interpreted if a value of X = 0 falls within the range of the observed data. Same goes for very large X values beyond the data set. Predictors depend on linear data.

Root Mean Square Error (RMSE): the average size of our prediction errors ---------------------------------→

$$\mathrm{RMSE} = \sqrt{\frac{\sum \hat{r}_i^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

R^2: square of the correlation, fraction of variation explained by model.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

squared errors from using model

squared errors from using mean

Lecture 11: Testing and Prediction with a SLR Model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim \mathrm{Normal}(0, \sigma_\epsilon^2)$$

Assumptions of SLR Model: 1. Independence between residuals. 2. Equal variances of residuals. 3. Normally distributed residuals around the "line". 4. Linearity between predictor and Y

Residual Standard Deviation -------→

$$\hat{\sigma}_\epsilon = \mathrm{RMSE} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

The true regression line does not equal the estimated LS regression line. We have uncertainty about the line and we need to incorporate this uncertainty into our predictions and make sure the linear relationship is significant. Run a

Hypothesis Test for a Slope:

$$T = \frac{b_1 - 0}{\mathrm{SE}(b_1)} \qquad \mathrm{SE}(b_1) = \frac{\mathrm{RMSE}}{\sqrt{n-1}} \times \frac{1}{s_x}$$

*We use t distribution with n-2 degrees of freedom.

Confidence Intervals for Regression Coefficients: The 100C% Confidence Interval for slope B_1 and intercept B_0

$$\left(b_1 \pm t^*_{n-2} \cdot \mathrm{SE}(b_1)\right) \qquad \left(b_0 \pm t^*_{n-2} \cdot \mathrm{SE}(b_0)\right)$$

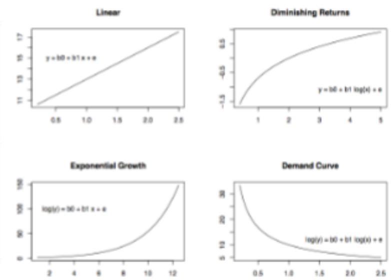
Predicting Intervals with Normal Distribution: we can report the uncertainty in a predicted value by calculating a 95% prediction interval for the value:

$$\left(\hat{y}_{\text{new}} \pm t^*_{n-2} \cdot \mathrm{SE}(\hat{y}_{\text{new}})\right) \qquad \mathrm{SE}(\hat{y}_{\text{new}}) = \mathrm{RMSE} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{(n-1) \cdot s_x^2}}$$

Lecture 12: Non-Linearity in the SLR Model

Residual Plot with Idealized Data: if relationship between X and Y is linear, residual plot should look like random scatter. If non-random → non-linear.

y on *x* Linear As *x* increases by 1, avg *y* increases by b_1	*y* on log(*x*) Diminishing returns As *x* increases by 1%, avg *y* increases by 0.01 b_1
log(*y*) on *x* Exponential growth As *x* increases by 1, avg *y* increases by 100 b_1%	log(*y*) on log(*x*) Demand curve As *x* increases by 1%, avg *y* increases by b_1%



Lecture 13: Violations of the SLR Model

Heteroskedasticity: fanning out pattern…variances are not equal. RMSE is too small for errors at large X levels and too large for errors at small X levels. Prediction bands too small for smaller X levels and too large for larger X levels.

Possible Solution: Re-scale the residuals (divide entire equation by X variable).

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \frac{\beta_1 x_i}{x_i} + \frac{\epsilon_i}{x_i}$$

Errors shouldn't fan out anymore because we are dividing by X

…then re-parameterize the equation.

Diagnosing Non-Normality: look at normal quantile plot of residuals. Very few quantiles should be outside the red bands. If our residuals are not normally distributed, then we cannot trust our 95% prediction bands.

Diagnosing Dependent Residuals: plot the residuals versus the rows of dataset. If rows are ordered by time, any systematic pattern in this plot would be evidence of autocorrelation / dependence over time between residuals.

Outliers: the most influential outliers are on the extremes of the X variable.

Lecture 15: Multiple Regression

Scatterplot Matrix: allows us to see not only relationship between Y variable & many X variables, but also relationship among X's themselves

Equation of a Line w/ Multiple Predictors: equation of plane (2 variables) or hyper-plane (3+ variables). $\boxed{y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i}$

Estimating the Best Fit Coefficients: best fit values b_0, b_1, and b_2 are the ones that minimize the sum of squared residuals (derived using calculus.)

Intercept Interpretation: "b_0 is the average value of the Y variable when *both* X variables equal zero" (Note: beware of extrapolations beyond data range.)

Slope Interpretation: b_k is the expected change in Y for a unit change in X_k, for a fixed value of all other predictors. Slopes from MRM also called *Partial Effects* (the partial effect b_k is the effect on X_k on Y having adjusted or controlled for the effect of the other predictors on Y.) Slope of SLR is *Marginal Effect*. Note: "effect" ≠ causal.

RMSE w/ Multiple Predictors: we define residuals (errors) the same way:

Adjustment to RMSE: number of observations – number of coefficients (K slopes + intercept) $y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_K x_{Ki})$ $\boxed{\text{RMSE} = \sqrt{\dfrac{\sum (y_i - \hat{y}_i)^2}{n-K-1}}}$

R^2 w/ Multiple Predictors: still the square of the correlation and interpreted as the proportion of the variation in Y explained by the model. R^2 cannot decrease when we add more predictors (even if extras are irrelevant.) Adjusted R^2 can decrease, so we use this to see if adding a variable is worth it.

Lecture 16: Inference and Diagnostic for MRM

MRM has same assumptions for the residuals (errors) as SLR (see Lecture 11).

Global Test of MRM: does the entire collection of predictor variables in the model provide significantly better predictions of the outcome? Null hypothesis is that none of the predictors have a significant relationship with Y:

H_0: $B_1=B_2=\ldots=B_K=0$ vs. H_a: at least one $B_k=0$ $\boxed{F = \dfrac{R^2}{1-R^2} \times \dfrac{n-K-1}{K}}$

F Statistic: test statistic for significantly improvement of entire model (larger values of R^2 give a large F statistic) → check p-value of F statistic in ANOVA table.

Testing Individual Predictor Variables: the test statistic for whether a specific predictor X_k should be included in the model (p-values found in Parameter Estimates) ---------------------------------→ $\boxed{T = \dfrac{b_k - 0}{SE(b_k)}}$

Note: each of the p-values are conditional on the other predictors on the model (X_1 is a significant predictor *on top of* X_2, vise versa, etc.)

Confidence Intervals of Partial Effects: (b_k +/- t^*_{n-K-l} * $SE(b_k)$) → df = n-K-1

For an approximate 95% CI use normal distribution where $t^*_{n-K-l} = 2$

Diagnostics: Actual vs. Predicted Values: the closer the points are to the line, the better the fit of the MRM. Residuals vs. Predicted Values: similar to SLR, we are looking for *patterns* in points that would indicate *non-linearity*. Leverage Plots: show the partial effects of each predictor. The slope of the fitted line in each leverage plot is the slope for that predictor in MRM. Good for revealing *high leverage points* that are hard to spot in marginal views of model fit.

Lecture 17: Collinearity

Collinearity occurs when the correlations among the predictors get large. Highly correlated predictors tend to change together, making it difficult to estimate the partial slope. The partial slope of X_1 attempts to measure how the average of Y changes when X_1 *alone* changes. If X_1 is highly correlated with X_2, we cannot separate the effect of X_1 from the effect of X_2. Both change together.

Effects of Collinearity: makes model difficult to interpret because the partial effects can change magnitude and/or sign as other predictors added/subtracted. Collinearity increases the standard errors for our coefficients because we are more uncertain about the partial effects → wider 95% CI, T-stat decreases.

Variation Inflation Factors (VIF): we can use how much the SE of b_k increases as a measure of the effect of collinearity on our model. If uncorrelated, VIF = 1. If correlated, VIF can be much larger than 1.

VIF of 2.8 means the variance of partial effect is 2.8x more than if there were no collinearity. $\boxed{\text{VIF}(X_k) = \left[\dfrac{SE(b_k)\text{ with correlated X's}}{SE(b_k)\text{ if X's uncorrelated}}\right]^2}$

Diagnosing Collinearity: look at Leverage Plots, if the points are "squished" into a more limited range then this is because the correlation limits the real variation. What to do? Live with it (no violation, just hard to interpret), remove redundant variables, or transform/combine predictors (e.g. poverty with income)

Lecture 18: Categorical Predictors

T-test under marginal SLR analysis does not account for confounding variables.

MRM w/ Poverty and Highway: since either highway=1 or highway=0, the partial

$$y_i = b_0 + b_1\,\text{highway}_i + b_2\,\text{poverty}_i \qquad y_i = \begin{cases} b_0 \quad\quad + b_2\,\text{poverty}_i & \text{if highway}_i = 0 \\ b_0 + b_1 + b_2\,\text{poverty}_i & \text{if highway}_i = 1 \end{cases}$$

effect b_1 of the highway predictor acts as an additive bump on *intercept*. Fitted lines are parallel because they only differ in terms of extra b_1 on intercept.

We can evaluate the *significance* of categorical predictor variable with:

1. **Effect test**: F-test of increase in R^2 due to new categorical predictor.
2. **Partial effect**: t-test of whether partial slope is different from zero.

Cautions: 1. Check which category JMP codes as baseline (X=0). 2. Ignore "Parameter Estimates", only look at "Indicator Function Parameterization."

Categorical Variables w/ >2 Categories: if a categorical variable only has M different categories then we represent it with a set of M-1 indicators: X_1=1 if highway, 0 if not highway. X_2=1 if major street, 0 if not major street. Note that one of the categories is always defined as the *baseline category* (where X=0)

Lecture 19: Interactions

Extending MRM to allow slope on one predictor to depend on value of another.

Diagnostic Plots for MRM: *residual plot* is examined in order to diagnose non-linearity (non-randomness) and heteroscedasticity (fanning out). *Side-by-side boxplots of residuals* between categories diagnoses heteroscedasticity (when variances of the residuals look different.)

Indicator variables are the most natural way to numerically code a categorical variable, but we are limited only to a change in the intercept, not slope.

Interaction Terms for a Change in Slope: multiply cat. indicator by other X…

$$y_i = b_0 + b_1\,\text{highway}_i + b_2\,\text{poverty}_i + b_3\,\text{highway}_i \times \text{poverty}_i \qquad y_i = \begin{cases} b_0 + b_1 + (b_2 + b_3)\cdot\text{poverty}_i & \text{if no Highway} \\ b_0 \quad + \quad b_2\cdot\text{poverty}_i & \text{if Highway} \end{cases}$$

The product of two predictor variables is called an *interaction term*. Since the categorical variable affects both the intercept and slope of other X, we must now interpret the partial effect of categorical *for a particular level of other X*.

Interaction Terms Between 2 Continuous Vars: it isn't easy to interpret interaction terms between two continuous variables, other than to say that we are allowing the slope on X_1 to change with X_2, and vice versa. Great way to add *non-linear relationships*. Interactions make model flexible.

Indicator Function Parameterization

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.27825	0.094458	45.29	<.0001*
poverty	1.9690916	0.242263	8.13	<.0001*
highway[no]	-0.611572	0.103031	-5.94	<.0001*
poverty*highway[no]	0.728436	0.270674	2.69	0.0072*

E.g. for neighborhoods with poverty = 0, the partial effect of **not** having a highway is 0.612 less log violent crimes

E.g. for neighborhoods with poverty = 1, the partial effect of **not** having a highway is (-0.612+0.728) = 0.116 more log violent crimes

Lecture 20: Model Building

Choosing the best set of predictors that balances prediction w/ interpretation?

Model Building Road Map: 1. Exploratory Analysis and SLR (univariate and bivariate summaries, transformations) 2. MRM with Manual Selection (adding new variables, removing insignificant variables from the model) 3. MRM Automated Selection (using automated stepwise to pick variables) 4. Diagnostics (residual plots, leverage plots) 5. Prediction of new observations, insight from partial effects.)

General Guidelines for Manual Model Building: starting point for a MRM has only main effects of predictor variables; only include main effects of predictors if they have significant partial effects; for any interaction term you are considering, make sure to include main effects for those predictor variables; only keep interaction terms that have significant partial effects.

Automated Model Building: incrementally build using "greedy algorithms".

Best to build manual then use automatic to make sure no predictors overlooked.

Stepwise Regression: add X which gives the largest R^2 improvement, remove the X which gives the smallest decrease in R^2. **Mixed Stepwise**: first perform an *add* step, then a *delete* step, then an *add step*, etc. until no more X's can be added or deleted.

Forward Selection: start with an empty model and only add X's, gradually building up a model. **Backward Elimination**: start with the full model and only remove X's, gradually reducing a model by elimination.

Stopping Criteria: continue Add Step if the p-value of X < "Prob to Enter" (~ 0.25). Continue Delete Step if the p-value of X > "Prob to Leave" (~0.10).

Lecture 21: Time Series

Modeling an outcome variable that is a sequence of observations over time.

Crimes in Philly Example: day=day of year, month=*categorical*.

Side-by-side boxplots of residuals can show that residuals vary by month.

Residual Autocorrelation: whenever a regression is performed with time series data, there is a possibility of sequential dependence between successive residuals e_t and e_{t-1} ("more crimes than expected on a particular day means it is more likely that there will be more crimes than expected the next day"). **Autocorrelation Diagnostic**: meandering pattern (staying positive for a while, then negative, etc.) of residuals over time is evidence of *autocorrelation*.

Autocorrelation Solution: create a *Lag Residuals* variable (represent the residual for day t with e_t then the lag residual for day t is e_{t-1}. This is just residuals shifted down by one row. First lag residual is missing (no lag time pt.). Plotting Residuals by Lag Residuals can show autocorrelation if linear. If there is *still* autocorrelation after first lag residual variable, add another, etc.

Durbin Watson Test: autocorrelation if p-value is below alpha level.

Lecture 22: Logistic Regression I

Binary categorical variable as *outcome* and continuous variable as *predictor*.

$Y_i = \begin{cases} 1 & \text{if increased crime} \\ 0 & \text{if decreased crime} \end{cases}$ $P(Y_i = 1) = \beta_0 + \beta_1 X_i$ Note: must transform so LHS is unrestricted → called *logistic link function*.

Estimation of Logistic Coefficients: JMP iteratively reweighted least squares.

Translating Log Odds into Probabilities: un-do the log odds transformation:

$$\boxed{\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 X_i} \quad \boxed{P(Y_i = 1) = \frac{e^{b_0 + b_1 X_i}}{1 + e^{b_0 + b_1 X_i}}} \quad \boxed{P(Y_i = 1) = \frac{e^{b_0}}{1 + e^{b_0}}}$$

Note: fitted line for log odds of Y=1 = fitted curve for probability of Y=1.

Intercept Interpretation: intercept b_0 is the log odds of "TRUE" when X=0.

Slope Interpretation: slope b_1 is the change in the log odds of Y=1 for a one unit change in the predictor variable. Compare P(Y=1|X=0) vs. P(Y=1|X=1). $P(Y_i = 1|X_i = 0) = \frac{e^{b_0}}{1 + e^{b_0}}$ vs. $P(Y_i = 1|X_i = 1) = \frac{e^{b_0 + b_1}}{1 + e^{b_0 + b_1}}$ **Hypothesis Testing**: how do we establish whether our relationship is significant? Is slope B_1 sig. diff. than 0? Follows *Chi-Square*: if p<0.05 → X is significant predictor of Y.

Lecture 23: Logistic Regression II

RMSE: compares P(Y=1) to actual value of Y=1 or 0. Lower is better.

Misclassification Rate: number of incorrect predictions when using "most likely" Y=1 or 0. Lower is better. **Confusion Matrix**: gives more detail about correct vs. incorrect predictions from logistic regression model. Look at false positives and false negatives. **Intercept Interpretation w/ Multiple Predictors**: b_0 is the log odds that Y=1 when *both* X variables equal zero (be careful when $X_{1 \text{ or } 2}$=0 are extrapolations.)

Slope Interpretation w/ Multiple Predictors: b_k is the expected change in the log odds of Y=1 for a unit change in X_k for a fixed value of all other X's. *Partial effect* b_1 is the effect of X_1 on the log odds of Y=1 having controlled for effect of X_2 on Y.

Lecture 24: Tree Models

When assumption that relationship between X's and Y is *linear* doesn't hold.

Step Function: one of the simplest non-linear relationships between X and Y. With data, the mean of Y changes between two different ranges of the X variable (e.g. X>0.65 vs. X<0.65). **Binary Decision Tree**: we create a tree that partitions our data based upon a binary split of the X variable. **Estimation**: How do we pick the correct predictor variables to split on, and how do we choose the correct split points? We want to create a binary split that makes the nodes (two groups created by split) as *pure* from each other as possible. **Pure**: Y values within each node are much more similar than the Y values between the nodes. Within node variance < between node variance. **Tree Building**: More complex/larger trees probably fit the data better but are more difficult to interpret. **Split Step**: scan through all predictor variables and all split points within a predictor variable and find the split that would give new nodes with the biggest improvement in *purity*. **Prune Step**: scan through all splits already in the tree and remove any splits that do not substantially improve *purity*. (Note: JMP output for purity is *LogWorth*. No more transformations since we don't care about linear relationships!) **Overfitting Data**: model predicts fitted data very well, but predicts future observations poorly because we have fit too complex of a model.

Validation: a way of using only part of a data set to estimate the tree and using the other part to assess the predictive ability of the model (~20% of data as hold out.)

Interpreting Tree Models: splits that appear earliest in the tree are most important. Predictors that are nested in the same branch of a tree are the tree version of *interactions* (two or more predictors combining to change the mean of the outcome variable.) Note: outcomes can be binary categorical outcomes or continuous.

Lecture 25: Urban Case Study

Very poor neighborhoods have significantly greater #'s of violent crimes than wealthier neighborhoods. Strong association between poverty and crime (according to **marginal analysis**, but ignores confounding variables.) Each neighborhood characteristic is predictive of violent crime (because they each have significant partial effects.) Poverty, income, comresprop and population are most strongly associated with violent crime. **Tree model** confirms that poverty, comresprop and population are most strongly associated with violent crime. **Time series** show that all crime is trending downwards in Philadelphia overall. Poverty and vacantprop are significant predictors of crime increase over time. **Beyond 102**: businesses open longer are associated with fewer crimes, crimes stay away from vacancies.