Linear Regression:

$correlation: r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{s_x})(\frac{y_i-\bar{y}}{s_y})$ where $mean = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $sd = s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

Variance of ϵ_i	Variance of $\hat{Y}_i$	Variance of Y_i	Avg sq of errors	Residual se
SSE/RSS $=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	SSR/ESS $=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y}_i)^2$	TSS$=\sum_{i=1}^{n}(Y_i-\bar{Y}_i)^2$	MSE$=\frac{1}{n-2}*RSS$	RMSE$=\sqrt{MSE}$

Predicted value: $\hat{y_i} = b_0 + b_1 x_i$ / Slope:$b_1 = r * \frac{S_y}{S_x}$ / Intercept:$b_0 = \bar{y} - b_1 * \bar{x}$/ Residual: $e_i = y_i - \hat{y_i}$

We choose the line to make SSE or RSSR, MSE, RMSE as small as possible.

Fraction of variance explained by X: $R^2 = \frac{TSS-SSE}{TSS} = 1 - \frac{SSE}{TSS}$

The larger R^2, the stronger the linear relationship; the more confident we are in our prediction. **Interpretation: the proportion of variation explained by the regression**(where the variation refers to sample variance of y)

R2 is not useful for deciding between regressions when

1.The response variables y are different due to transformation. 2. The data points are different due to the removal of outliers.

Model Diagnostics:

1. linearity: by checking whether the residuals follow a symmetric pattern with respect to h = 0.

2. homoscedasticity: by checking whether the residuals are evenly distributed within a band

3. normality: by looking look at the qqplot of the residuals, most of points close to the line. 4. independence

Model Diagnostics

- Make sure there are no gross violations of the model
  - Is the relationship between *x* and *y* *linear*?
  - Do the residuals show iid normal behavior (i.e., *independent, equal variance, normality*)?
  - Are there *outliers* that may distort the model fit?
- Crucial steps in checking a model
  - A *y vs. x scatterplot* should reveal a linear pattern, linear dependence.
  - A *Residual vs. x scatterplot* should reveal no meaningful pattern.
  - A *Residual vs. Predicted scatterplot* should reveal no meaningful pattern.
  - A *histogram and normal quantile plot* of the residuals should be consistent with the assumption of normality of the errors.

Interpretation for Models on Log-scale

- $y = a + bx$
  If x changes from x to $x+\delta$, the exact change in y is $b\delta$.
- $y = a + b\log_e x$
  If x changes from x to $x(1+p\%)$ (a $p\%$ change), the approximate change in y is $bp\%$.
- $\log_e y = a + bx$
  If x changes from x to $x+\delta$, the approximate change in y is y to $y(1+b\delta)$.
- $\log_e y = a + b\log_e x$
  If x changes from x to $x(1+p\%)$ (a $p\%$ change), the approximate change in y is y to $y(1+bp\%)$.

1. Is $\beta_1 = 0$? $\Rightarrow$ t-statistics
2. Are all $\beta_j = 0$? $\Rightarrow$ F-statistics
3. 95% Confidence interval of β_j? $\hat{\beta_j} \pm 2 * SE(\hat{\beta_j})$
4. 95% Confidence interval of $\hat{y}_i$? $\hat{y_i} \pm 2 * RMSE$

Collinearity: Substantial correlation among predictors. correlation between independent variables

Confounding: influences dependent and independent variable

Without interaction:

Interpretation Y $\sim$ X: for every one-unit increase in X, the average increase in Y is β_1

That the coefficient of X is β_1, significant at 0.05 level. This coefficient shows that X has a positive/negative **marginal effect** on Y. R2 of regressing Y on X is 'Multiple R-squared'.

Interpretation Y $\sim$ X1+X2: The coefficient of X1 in this model is β_1, significant at 0.05 level. The coefficient of X1 show that year has a positive/negative **partial effect** on Y, given a constant X2. Given a constant X2, the expectation of Y will increase/decrease β_1 unit if the X1 increases one unit.

With interaction: coef of interaction indicates diff in slopes btwn group

The OLS model is $Y = \beta_0 + \beta_1 * X1 + \beta_2 * X2 + \beta_3 * X1 * X2 + \epsilon_i$

Interpretation: The coefficient of interaction term is β_3, significant at 0.05 level. The marginal effect of X1 on Y is $\beta_1 + \beta_3 * X2$.

With categorical variable: coef of categ term indicates diff btwn parallel line

The OLS model is $Y = \beta_0 + \beta_1 * X1 + \beta_2 * Cy4 + \beta_3 * Cy5 + \epsilon_i$

The effect of cy in this model: cy4 has a significant estimator coefficient (β_2) at 0.01 level, which means that cy4 has a positive effect on Y compared with situation cyl is equal to 3/baseline. However, Cy5 don't own a significant estimator.

Least-Squares Estimation: Find $b_0, b_1, ..., b_k$ to minimize $SSE = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - .. - b_k x_{ik})^2$

T-test: is $\beta_1 = 0$? : $H_0 : \beta_1 = 0$ *vs.* $H_a : \beta_1 \neq 0$: $T = \frac{\beta_1 - null\ value}{SE}$ $df = n - 2$

F-test: for predicting Y/is all $\beta_j = 0$? $H_0 : \beta_k = 0$ for all k $F = \frac{\frac{TSS-SSE}{K}}{\frac{SSE}{n-K-1}} \sim F_{k,n-K-1}$

ANOVA Type 1: H_0=sub model is adequate, H_a=sub model is not adequate

P<0.05: reject H_0, use longer model; P>0.05: fail to reject/no evidence to against H_0, use shorter model

Anova Type 2: fit a full model, delete X with p-value >0.05 (one at time)

RMSE can be used to estimate σ/ **VIF** $= \frac{1}{1-R^2}$, x's uncorr, VIF=1

R^2 equals to the squared correlation(r^2) between y and $\hat{y}$=1-$\frac{RSS}{TSS}$/$R^2 adjusted = 1 - \frac{SSE/n-K-1}{TSS/(n-1)}$ n # of points, k # of x

95% CI for β_j: $\hat{\beta_j} \pm 2 * SE(\hat{\beta_j})$ / 95% CI for $\hat{y_i}$: $\hat{y_i} \pm 2 * RMSE$

x: independent variable, explanatory variable or predictor/ y dependent variable or response

Model Selection:

Prediction Accuracy: $p > n$, to control the variance(i,e variance is infinite) and enable model fitting; $n > p$ a lot of variability, overfitting and consequently poor predictions; $n >> p$ LS estimate have low variance, and will perform well on test observations.

Model Interpretability: by removing irrelevant features through setting the corresponding coefficients to be zero.

p ↑ – Flexibility/Accuracy ↑ – Interpretability ↓ – Test MSE ↑ – Train MSE uncertain – Test MSE<Train MSE

p ↑, RSS ↓ R^2 ↑, Variance ↑ bias ↓, Note: RSS and R^2 are **not suitable** for selecting the best model among models with **different number of predictors**. Target: choose a model with **low test error**, not a model with low training error.

1. Indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting. 2. Directly estimate the test error, using either a validation set approach or a cross-validation approach.

Three Classes of Model Selection Methods

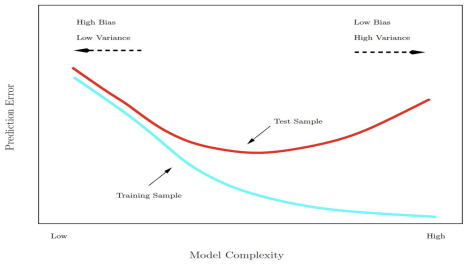
- *Subset Selection*. We identify a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.
- *Shrinkage*. We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as *regularization*) has the effect of reducing variance and can also perform variable selection.
- *Dimension Reduction*. We project the p predictors into a M-dimensional subspace, where $M < p$. This is achieved by computing M different *linear combinations*, or *projections*, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

Model Comparison Criteria

Adjustment to the Training Error

- Adjusted R^2 — Adjusted $R^2 = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}$
- Mallow's C_p — $C_p = \frac{1}{n}(\text{RSS} + 2p\hat{\sigma}^2)$
- Bayesian Information Criterion (BIC) — $BIC = \frac{1}{n}(\text{RSS} + \log(n)p\hat{\sigma}^2)$
- Akaike Information Criterion (AIC) — $AIC = -2\log L + 2 \cdot p$

Model Complexity & Prediction Error



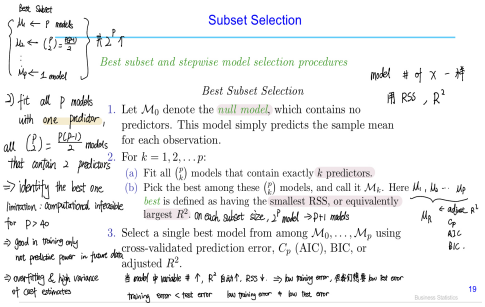
Choose **largest** Adjusted R^2, **smallest** BIC, Mallow's C_p, AIC/ FSS & BSS not guaranteed to find the best possible model (undirect estimate)

Best subset selection: suffers from computational limitations when p large, large search space lead to overfitting and high variance. total: 2^p models

Forward stepwise selection: a model containing no predictors, and adds predictors to the model, one-at-a-time, until all of the predictors are in the model. Choose variable gives the greatest additional improvement to the fit is added to the model.(n can <p), total: 1+p(p+1)/2 models

Backward stepwise selection: full least squares model containing all p predictors, then removes the least useful predictor(n need>p)

Subset Selection



Backward Stepwise Selection: Details

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   2.1 Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   2.2 Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Validation and Cross-Validation

- Each of the procedures returns a sequence of models $\mathcal{M}_k$ indexed by model size $k = 0, 1, 2, \ldots$. Our job here is to select $\hat{k}$. Once selected, we will return model $\mathcal{M}_{\hat{k}}$
- We compute the validation set error or the cross-validation error for each model $\mathcal{M}_k$ under consideration, and then select the k for which the resulting estimated test error is smallest.
- This procedure has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and *doesn't require an estimate of the error variance σ^2.*
- It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2.

Validation Set & Cross Validation problem: 1 Only a subset of the observations are used to fit the model.2 The validation estimate of the test error depend on the split of the raw data(highly variable). 3 The validation test error may tend to **overestimate** the test error for the model fit on the entire data set. **K-fold Cross-Validation (CV):** randomly dividing the set of observations into k groups, of approximately equal size. MSE is then computed on the observations in the held-out fold. Calculate the avg mse for k groups. select the k for which the resulting estimated test error is smallest.

- Let the K parts be $C_1, C_2, \ldots C_K$, where C_k denotes the indices of the observations in part k. There are n_k observations in part k: if N is a multiple of K, then $n_k = n/K$.
- Compute

$$\text{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \text{MSE}_k \left(= \frac{1}{k}\sum_{i=1}^{k} \text{MSE}_i \right)$$

  where $\text{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$, and $\hat{y}_i$ is the fit for observation i, obtained from the data with part k removed.
- Setting $K = n$ yields n-fold or *leave-one out cross-validation* (LOOCV).

Shrinkage Methods

- The subset selection methods use least squares to fit a linear model that contains a subset of the predictors.
- As an alternative, we can fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.
- Ridge regression and Lasso regression
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance.

Ridge Regression

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

- In contrast, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

  where $\lambda \geq 0$ is a *tuning parameter*, to be determined separately.

Shrinkage:

Ridge regression: As with least squares, ridge regression seeks coef estimates that fit the data well, by **making the RSS small.** When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates, variance high,bias 0. However, as $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero. $\lambda \uparrow$, variance $\downarrow$, bias $\uparrow$. MSE down then up

It will produce a different set of coefficient estimates, $\hat{\beta}_\lambda^R$, for each value of λ. As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias. Use CV to select λ

Ridge Regression: Scaling of Predictors

- The standard least squares coefficient estimates are *scale equivariant*: multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$. In other words, regardless of how the jth predictor is scaled, $X_j\hat{\beta}_j$ will remain the same.
- In contrast, the ridge regression coefficient estimates can change *substantially* when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.
- Therefore, it is best to apply ridge regression after *standardizing the predictors*, using the formula

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

The Bias-Variance Decomposition

- Assume that

$$Y = f(X) + \varepsilon$$

  where E(ε)=0 and Var(ε)=σ_ε^2.
- At an input point $X = x_0$, the expected squared prediction error is

$$\begin{aligned} \text{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\ &= \sigma_\varepsilon^2 + [\text{E}\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - \text{E}\hat{f}(x_0)]^2 \\ &= \sigma_\varepsilon^2 + \text{Bias}^2(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0)) \\ &= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}. \end{aligned}$$

- The more complex the model, the lower the bias but the higher the variance.

The Lasso

- Ridge regression does have one obvious disadvantage: unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model
- The *Lasso* is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

- In statistical parlance, the lasso uses an ℓ_1 (pronounced "ell 1") penalty instead of an ℓ_2 penalty. The ℓ_1 norm of a coefficient vector β is given by $\|\beta\|_1 = \sum |\beta_j|$.

Lasso: In lasso, the l_1 penalty has the effect of forcing some of the coeffcient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large. Hence, much like best subset selection, the lasso performs **variable selection**. Lasso yields sparse models: models that involve only a subset of the variables. lasso.1se/lasso.min/CV to select λ

Logistic Regression: $Y_i \sim Binomial(n_i, p_i)$ s curve(x-axis: x, y-axis: Probability), β_1 slope, $\beta_0 \uparrow$ shift left,$-\frac{\beta_0}{\beta_1} = threshold, where prob of succ = .5$

logistic function: $p(x) = P(Y=1|x) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$

odds: $\frac{p(x)}{1-p(x)} = \frac{P(Y=1|x)}{P(Y=0|x)} = e^{\beta_0+\beta_1 X}$ indicates how much more likely it is that an observation is a member of the target group rather than a member of the other group. i.e, Values of the odds close to 0 and ∞ indicate very low and very high probabilities of y, respectively.

log odds or logit: $logit(p_i) = log(\frac{p(x)}{1-p(x)}) = log(\frac{P(Y=1|x)}{P(Y=0|x)}) = \beta_0 + \beta_1 X$ We see that the logistic regression model has a logit that is linear in X.

Likelihood(Use MLE estimate β): $\mathcal{L}(\beta_0, \beta_1|Y, X) = \prod_{i=1}^n p(x_i)^{y_i}(1-p(x_i))^{1-y_i}$

$l(\beta_0, \beta_1|Y, X) = log[\prod_{i=1}^n p(x_i)^{y_i}(1-p(x_i))^{1-y_i}] = \sum_{i=1}^n (y_i * log(p(x_i)) + (1-y_i) * log(1-p(x_i)))$

Without dummy: $log(\frac{P(Y=1|x)}{P(Y=0|x)}) = \beta_0 + \beta_1 X$ an increase in X is associated with an increase in the probability of Y=success.

For every one-unit increase in X is associated with an increase in the log odds of Y by $\hat{\beta_1}$. Equivalently, it multiplies the odds by e^{β_1}

With dummy: $log(\frac{P(Y=1|x)}{P(Y=0|x)}) = \beta_0 + \beta_1 X_{=1}$ the log odds of Y is equal to default(success) will increase β_1 if X=1 **compared to** X=0. The intercept means the log odds of Y=success given X=0(baseline). Both of estimators are significant.

1. Intercept: The probability that a non student(X=0) will have a default(Y=success) is $p(x) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$

2. β_1: The X=1 group has $100\% * (e^{\beta_1} - 1)$ less/more odds of having default than the X=0 group.

Poisson:

Suppose that a random variable Y takes on non-negative integer values, Poisson i.e. $Y \in 0, 1, 2, ...$ If Y follows the Poisson distribution, then $P(Y=k) = \frac{e^{-\lambda} * \lambda^k}{k!} for\ k = 0, 1, 2....$ Here, $\lambda > 0$ is the expected value of Y , i.e. E(Y). It turns out that λ also equals the variance of Y , i.e. $\lambda = E(Y) = Var(Y)$. This means that if Y follows the Poisson distribution, then **the larger the mean of Y, the larger its variance.**

$log(\lambda(X_1, ..., X_p)) = \beta_0 + \beta_1 X_1 + ... + \beta_p * X_p / \lambda(X_1, ..., X_p) = e^{\beta_0+\beta_1 X_1+...+\beta_p * X_p}$

Likelihood: $l(\beta_0, \beta_1, ..., \beta_p) = \prod_{i=1}^n \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!}$ where $\lambda(x_i) = e^{\beta_0+\beta_1 X_{i1}+...+\beta_p * X_{ip}}$

Pros: fix the heteroscedasticity of the data calls into question the suitability of a linear regression model(mean-variance relationship)

Interpretation: An increase in X_j by one unit is associated with a change in $E(Y) = \lambda$ by a factor of $exp(\beta_j)$.

Example: a change in weather(X dummy) from clear(baseline) to cloudy skies is associated with a change in mean bike usage(Y) by a factor of $exp(-0.08/\beta) = 0.923$, i.e. on average, only 92.3% as many people will use bikes when it is cloudy relative to when it is clear(baseline).

offset: if we only observe Y, but fail to observe each Y_l, we can still treat Y as the observed count data, but with an additional offset term $\log n$.

$\log E[Y] = \log E[\sum_{l=1}^n Y_l] = \log \sum_{l=1}^n E[Y_l] = \log nE[Y_l] = \log n + X'\beta$.

Interpretation:

1. As we move from duration 0-4(baseline) to 5-9(X1) the log of the mean increases by almost one, which means that the number of CEB(Y) gets multiplied by $\exp\{0.9977/\beta_1\} = 2.71$.

2. By duration 25-29(X4), women in each category of residence and education(fixed others) have $\exp\{1.977/\beta_4\} = 7.22$ **times** as many children(Y) as they did at duration 0-4(baseline).

3. Women with secondary or higher education(X5) have 27% fewer kids $(1 - \exp\{-0.3096\} = 0.27)$, than women with no education(baseline) who live in the same type of place of residence(fixed others).

GLM:

Each approach uses predictors X1,...,Xp to predict a response Y. We assume that, conditional on X1,...,Xp , Y belongs to a certain family of distributions. For linear regression, we typically assume that Y follows a Gaussian or normal distribution. For logistic regression, we assume that Y follows a Bernoulli distribution. For Poisson regression, we assume that Y follows a Poisson distribution $E(Y|X_1, ..., X_p) = \lambda(X_1...X_p) = e^{\beta_0+\beta_1 * X_1+...+\beta_p * X_p}$

Link function, η, which applies a transformation to $E(Y|X_1, .., X_p)$ so that the transformed mean is a linear function of the predictors. That is, $\eta(E(Y|X_1, ..., X_p)) = \beta_0 + \beta_1 * X_1 + ... + \beta_p * X_p$.

The link functions for linear, logistic and Poisson regression are $\eta(\mu) = \mu, \eta(\mu) = log(\mu/(1-\mu))$, and $\eta(\mu) = log(\mu)$

Classification Terminology:

Specificity (true negative rate) $= Prob(\hat{Y} = 0|Y = 0) = \frac{TN}{TN+FP}$ / **Sensitivity (true positive rate)** $= Prob(\hat{Y} = 1|Y = 1) = \frac{TP}{TP+FN}$

False Positive Rate $= 1 - Specificity = P(\hat{Y} = 1|Y = 0) = \frac{FP}{FP+TN}$ / **True Positive Rate** $= \frac{TP}{TP+FN}$ / Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$

Missclassifications is: $\frac{FP+FN}{TP+TN+FP+FN}$/ Precision: $\frac{TP}{TP+FP}$/ **Type 1 error:** $\frac{FP}{N}$/ **Type 2 error:** $\frac{FN}{P}$

The ROC curve can simultaneously displaying the two types of errors for all possible thresholds. The overall performance of a classifier, summarized over all possible thresholds, is given by the AUC. An ideal ROC curve will hug the top left corner, so the **larger the AUC the better the classifier**.

The ROC plot: x-axis: FP rate (0-1) / Specificity(1-0), y-axis: True positive rate

Odds Ratio

- From $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$, we have $\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x} = e^{\beta_0} e^{\beta_1 x}$
- The odds ratio increases **multiplicatively** by e^{β_1} for every 1-unit increase in x
  - The odds at $X = x + 1$ are e^{β_1} times the odds at $X = x$
  - $\frac{odds(x+1)}{odds(x)} = e^{\beta_1}$
- Therefore, e^{β_1} **is an odds ratio!**
- e^{β_1} represents the change in the odds of the outcome (multiplicatively) by increasing x by 1 unit
  - If $\beta_1 > 0$, the odds and probability increase as x increases ($e^{\beta_1} > 1$)
  - If $\beta_1 < 0$, the odds and probability decrease as x increases ($e^{\beta_1} < 1$)
  - If $\beta_1 = 0$, the odds and probability are the same at all x levels ($e^{\beta_1=1}$)

Poisson Regression

- Three ingredients:
  - Likelihood of response = Poisson density
$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!},$$
  - Find expectation of y condition on x
$$\mathrm{E}(Y|X_1, \ldots, X_p) = \lambda(X_1, \ldots, X_p)$$
  - Link expectation with the linear function of predictors
$$\log(\lambda(X_1, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$
- Estimate β by maximizing the likelihood.

Poisson vs Linear Regression

- An increase in X_j by one unit is associated with a change in $E(Y) = \lambda = e^{X\beta}$ by a factor of $\exp(\beta_j)$.
  - By contrast, in linear regression, an increase in X_j by one unit is associated with an increase of β_j in $E(Y) = \mu = X\beta$.
- Poisson regression implicitly assumes that mean equals variance.
  - By contrast, linear regression assumes the variance takes on a constant value.
- Poisson regression gives non-negative predictions.
  - By contrast, linear regression predictions can be negative.

Linear Discriminant Analysis/LDA:

Choose a separating line that: 1. maximizes the similarity between members of the same group 2. minimizes the similarity between members belonging to different groups EX. $p(Y=k|X=x) = \frac{p(X=x|Y=k)*p(Y=k)}{\sum_k p(X=x|Y=k)*p(Y=k)}$ if $p(Y=1|X=x) > p(Y=2|X=x)$ belongs to class 1; We classify a new point according to which density is highest.

Assigning x to the class with the **largest discriminant score**: $\delta_k(x) = x * \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$, $\delta_k(x)$ is linear function of x

When K=2 classes, $\pi_1 = \pi_2 = 0.5$ the decision boundary is $x = \frac{\mu_1+\mu_2}{2}$

linDA: use $da.reg1\$functions$ to get coefficient(use each column). 1.A family is classified into the class of owners if the owner(Y_1) **function score is higher** than the nonowner(Y_2) function score

$\hat{\delta}(Nonnower|Income, Lot\_Size) = -51.42 + 0.3294 * Income + 4.682 * Lot\_Size / \hat{\delta}(Owner|Income, Lot\_Size) = -73.16 + 0.4296 * Income + 5.467 * Lot\_Size$

2. An alternative way is to compute **the probability of belonging to each of the classes** and assigning the record to the most likely class.

$P(Nonowner|Income, Lot\_Size) = \frac{\exp\{\hat{\delta}(Nonowner|Income,Lot\_Size)\}}{\exp\{\hat{\delta}(Owner|Income,Lot\_Size)\}+\exp\{\hat{\delta}(Nonowener|Income,Lot\_Size)\}}$

$P(Owner|Income, Lot\_Size) = \frac{\exp\{\hat{\delta}(Owner|Income,Lot\_Size)\}}{\exp\{\hat{\delta}(Owner|Income,Lot\_Size)\}+\exp\{\hat{\delta}(Nonowener|Income,Lot\_Size)\}}$ or $\frac{e^{da.reg1\$scores[,2]}}{e^{da.reg1\$scores[,1]}+e^{da.reg1\$scores[,2]}}$

QDA: Assume each class has its own covariance matrix. i.e. an observation from the kth class is of the form $X \sim N(\mu k, \sigma k)$, where Σk is a covariance matrix for the kth class. When there are p predictors, then estimating a covariance matrix requires estimating p(p+1)/2 parameters. QDA estimates a separate covariance matrix for each class, for a total of Kp(p+1)/2 parameters.

LDA is special QDA and NB; LDA is trying to approximate the Bayes classifier, which has the **lowest total error rate out of all classifiers**,regardless of the class from which the errors stem.→ lower threshold

p ↑ , LDA has low var/high bias, less flexible classifier than QDA, overfitting. QDA has high var/low bias, good for large training set and assumption for K classes cov matrix is untenable. Naive Bayes for large p or small n.

Bayes Theorem for Classification

Thomas Bayes was a famous mathematician whose name represents a big subfield of statistical and probablistic modeling. Here we focus on a simple result, known as Bayes theorem:

posterior (update) likelihood · prior (no info)

$$\Pr(Y=k|X=x) = \frac{\Pr(X=x|Y=k)\cdot \Pr(Y=k)}{\Pr(X=x)}$$

One writes this slightly differently for discriminant analysis:

$$\Pr(Y=k|X=x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}, \quad \text{where}$$

- $f_k(x) = \Pr(X=x|Y=k)$ is the *density* for X in class k. Here we will use normal densities for these, separately in each class.
- $\pi_k = \Pr(Y=k)$ is the marginal or *prior* probability for class k.

Linear Discriminant Analysis when $p = 1$

The Gaussian density has the form (one predictor)

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

Here μ_k is the mean, and σ_k^2 the variance (in class k). We will assume that all the $\sigma_k = \sigma$ are the same.

Plugging this into Bayes formula, we get a rather complex expression for $p_k(x) = \Pr(Y=k|X=x)$:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

Classification Terminology

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

TABLE 4.6. *Possible results when applying a classifier or diagnostic test to a population.*

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

TABLE 4.7. *Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.*

Other Forms of Discriminant Analysis

$$\Pr(Y=k|X=x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$$

When $f_k(x)$ are Gaussian densities, with the same covariance matrix $\mathbf{\Sigma}$ in each class, this leads to linear discriminant analysis. By altering the forms for $f_k(x)$, we get different classifiers.

- With Gaussians but different $\mathbf{\Sigma}_k$ in each class, we get *quadratic discriminant analysis*.
- With $f_k(x) = \prod_{j=1}^p f_{jk}(x_j)$ (conditional independence model) in each class we get *naive Bayes*. For Gaussian this means the $\mathbf{\Sigma}_k$ are diagonal.
- Many other forms, by proposing specific density models for $f_k(x)$, including nonparametric approaches.

Comparison of LDA, QDA, NB, Logistic

- LDA is a special case of QDA? True.
- Any linear classifier is a special case of NB? True.
- Discriminant Analysis is a type of discriminative learning? False.

	LDA	QDA	NB
Gaussian $f_{kj}(x_j)$?	Yes	Yes	Not necessarily
Diagonal Σ_k?	No	No	Yes
Shared $\Sigma_k = \Sigma$?	Yes	No	No

The First Principal Component

- Consider a set of features: $X_1, X_2, \ldots, X_p$ on n individuals. The first principal component (PC) is the linear combination $Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$ that has the largest variance

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^p \phi_{j1}x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^p \phi_{j1}^2 = 1.$$

- The entries $z_{11}, z_{21}, \ldots, z_{n1}$ are the PC scores.
- The PC loading vector: $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})^T$
- The second principal component Z_2 is the linear combination that has the maximal variance such that ϕ_1 and ϕ_2 are orthogonality.

Nominal and Ordinal:

The simplest decision criterion is whether that outcome is nominal (i.e., no ordering to the categories) or ordinal (i.e., the categories have an order).

Multinomial Regression: estimates a separate binary logistic regression model for each dummy variables. The result is M-1 binary logistic regression models. Each model conveys the effect of predictors on the probability of success in that category, in comparison to the reference category.

$P(Y=k|X=x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+..+\beta_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+..+\beta_{lp}x_p}}$, $log(\frac{P(Y=k|X=x)}{P(Y=K|X=x)}) = \beta_{k0} + \beta_{k1}x_1 + .. + \beta_{kp}x_p$ for k = 1,...,K-1

$P(Y=K|X=x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+..+\beta_{lp}x_p}}$ for K=K

Interpretation: for A.B.C 3 groups

β_{A0}: if we set B to be the baseline, then we can interpret β_{A0} as the log odds of A versus B, given that x1 = . . = xp = 0.

β_{Aj}:a one-unit increase in X_j is associated with a β_{Aj} increase in the log odds of A over B/ if X_j increases by one unit, then $\frac{P(Y=A|X=x)}{P(Y=B|X=s)}$ increases by $e^{\beta_{Aj}}$

Ordinal: Used to predict the dependent variable with 'ordered' multiple categories and independent variables. i.e, it is used to facilitate the interaction of dependent variables (having multiple ordered levels) with one or more independent variables.

Nominal Logistic Regression

- Choose one category as the reference category, say the 1st category
- Define the logits for the other categories as

$$\text{logit}(\pi_j) \equiv \log\left(\frac{\pi_j}{\pi_1}\right) = x^T\beta_j, \quad \text{for } j = 2, \ldots, J.$$

- Since the probabilities add up to 1, we have

$$\hat{\pi}_1 = \frac{1}{1+\sum_{j=2}^J \exp(x^T\hat{\beta}_j)}$$

$$\hat{\pi}_j = \frac{\exp(x_j^T\hat{\beta}_j)}{1+\sum_{j=2}^J \exp(x^T\hat{\beta}_j)}, \quad \text{for } j = 2, \ldots, J.$$

- Changing the reference category will change the above probabilities?

Ordinal Logistic Regression

- Cumulative logit model

$$\log\frac{\pi_1+\cdots+\pi_j}{\pi_{j+1}+\cdots+\pi_J} = x^T\beta_j.$$

- Special case: Proportional odds model

$$\log\frac{\pi_1+\cdots+\pi_j}{\pi_{j+1}+\cdots+\pi_J} = \beta_{0j}+\beta_1x_1+\cdots+\beta_{p-1}x_{p-1}.$$

The First Principal Component

- Consider a set of features: $X_1, X_2, \ldots, X_p$ on n individuals. The first principal component (PC) is the linear combination $Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$ that has the largest variance

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^p \phi_{j1}x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^p \phi_{j1}^2 = 1.$$

- The entries $z_{11}, z_{21}, \ldots, z_{n1}$ are the PC scores.
- The PC loading vector: $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})^T$
- The second principal component Z_2 is the linear combination that has the maximal variance such that ϕ_1 and ϕ_2 are orthogonality.

SVM:

P-dimensional Hyperplane: $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0$

Maximal Margin Hyperplane: the separating hyperplane that is farthest from the training observations (margin largest);extremely sensitive to a change in a single observation → over-fit the training data.

Margin/M: the smallest such distance is the minimal distance from the observations to the hyperplane;

Note: M represents the margin of our hyperplane, and the optimization problem chooses $\beta_0, \beta_1, ..., \beta_p$ to maximize M. Formula Constraints ensure that each observation is on the correct side of the hyperplane and at least a distance M from the hyperplane.

Maximal Margin classifier: classify a test observation based on which side of the maximal margin hyperplane it lies.

The Non-separable Case: If no separating hyperplane exists,that means no maximal margin classifier, or no solution with M > 0.

Support Vector Classifier: The generalization of the maximal margin classifier to the non-separable case; 1. Greater robustness to individual observations 2.Better classification of most of the training observations.

Soft margin: a hyperplane that almost separates the classes; 1. btwn margin & hyperplane: wrong side 2. outside margin: correct side 3. on the margin

Support vectors: Observations that lie directly on the margin, or on the wrong side of the margin for their class

C ↓ narrow margins: highly fit the data, ↓ bias ↑ variance

C ↓ wider margins, allow more violations, more Support vectors: less fit the data, ↑ bias ↓ variance

One-verses-One classification: approach constructs $\binom{K}{2}$ SVMs, each of which compares a pair of classes.

Construction of Maximal Margin Classifier

- Collect n training observations $X_1, X_2, \ldots, X_n$ of p dimensions, and associated class labels $y_1, y_2, \ldots, y_n \in \{-1,1\}$.

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{maximize}}\ M$$
$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M \ \ \forall\, i = 1, \ldots, n.$$

- The inequality constraint: each observation is on the correct side of the hyperplane, provided $M > 0$.
- The quality constraint: just a normalization, since rescaling β does not change the hyperplane.
- Maximal M: each observation is on the correct side and at least a distance M from the hyperplane.

The Non-separable Case: Soft Margin

- The non-separable case pursues:
  - Greater robustness to individual observations,
  - Better classification for most of the training observations.

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}}\ M$$
$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$
$$\epsilon_i \geq 0,\ \sum_{i=1}^{n} \epsilon_i \leq C,$$

- Slack variables ϵ_i allow observations to be on the wrong side.
- Cost C is a budget for the amount that the margin can be violated.
- "Support Vectors": observations that lie on the margin, or on the wrong side of the margin.

The Non-separable Case: Soft Margin

Interpretation of C

- $C = 0$, no budget for violations, $\epsilon_1 = \cdots \epsilon_n = 0$, the non-separable optimization works as the hard margin case.
- $C > 0$, no more than C observations can be on the wrong side of the hyperplane.
- C, as a nonnegative tuning parameter, can be chosen via cross-validation.

Interpretation of ϵ_i

- $\epsilon_i > 0$, the i^{th} observation is on the wrong side of the margin.
- $\epsilon_i > 1$, it is on the wrong side of the hyperplane.

"Support Vectors": observations that lie on the margin, or on the wrong side of the margin.

SVM Facts

1. The soft margin SVM optimization for the non-separable case has no solution with $M > 0$? False.
2. For observations that are not support vectors, they must be at least a distance M from the hyperplane? True.
3. If C is integer, it is not possible to have more than C observations on the wrong side of the margin? False.
4. Smaller C leads to wider margin, thus a classifier that is more biased but has lower variance? False.
5. SVM is very robust to outliers that are not support vectors? True.

SVMs with More than Two Classes

Two popular ideas:

- One-Versus-One Approach:
  - Construct $\binom{K}{2}$ SVMs for each pair of classes.
  - Assign a test observation to the class to which it was most frequently assigned in these $\binom{K}{2}$ pairwise classifications.
- One-Versus-All Approach:
  - Construct K SVMs, each time comparing one of the K classes to the remaining $K - 1$ classes.
  - Assign a test observation x to the class for which $\beta_{0k} + \beta_{1k} x_1 + \cdots + \beta_{pk} x_p$ is largest.

Summary

- SVM finds the classification hyperplane with maximal margin.
- Use CV to tune cost C, which controls bias-variance tradeoff.
- Equivalent to hinge loss + ridge penalty.
- Can be used for multi-class classification and regression.
- SVM with nonlinear kernels is beyond this scope of the course.

SVM Classification: $minimize_{\beta_0,\beta_1,..,\beta_p}\ \sum_{i=1}^{n} max[0, 1 - y_i f(x_i)] + \lambda \sum_{j=1}^{p} \beta_j^2$, where λ is a nonnegative tuning parameter (smallλ similar to small C)

λ large, $\beta_0, \beta_1, .., \beta_p$ small, more violations to the margin are tolerated, low variance and high bias classifier.

λ large, $\beta_0, \beta_1, .., \beta_p$ large, few violations t0 the margin will occur, high variance and low bias classifier.

SVM Regression: $minimize_{\beta_0,\beta_1,..,\beta_p}\ \sum_{i=1}^{n} max[0, |y_i - f(x_i)| - \epsilon] + \lambda \sum_{j=1}^{p} \beta_j^2$, ϵ-insensitive loss + ridge penalty

Comparisons about obs far from center:

SVM decision rule is based only on a potentially small subset of the training observations (the support vectors) means that it is **quite robust** to the behavior of observations that are far away from the hyperplane

LDA depends on the mean of all of the observations within each class & the within-class covariance matrix computed using all of the observations. It is **not robust** to any observations.

Logistic regression, unlike LDA, has very **low sensitivity** to observations far from the decision boundary

the support vector classifier and logistic regression are closely related.

Compare PCA & LDA:

Both methods are used to reduce the number of features in a dataset while retaining as much information as possible. PCA is an unsupervised learning algorithm while LDA is a supervised learning algorithm. **This means that PCA finds directions of maximum variance regardless of class labels while LDA finds directions of maximum class separability.**

PCA works by identifying the directions (components) that maximize the variance in a dataset. i.e, it seeks to find the linear combination of features that captures as much variance as possible. The first component is the one that captures the maximum variance, the second component is orthogonal to the first and captures the remaining variance, and so on. PCA is a useful technique for dimensionality reduction when your data has linear relationships between features – that is, when you can express one feature as a function of the other(s). So, we can use PCA to compress data while retaining most of the information content by choosing just the right number of features (components).

LDA Another linear transformation technique that is used for dimensionality reduction. A supervised learning method, which means it takes class labels into account when finding directions of maximum variance. This makes LDA particularly well-suited for classification tasks where you want to maximize class separability. LDA assumes that your data is centered around the origin and that your features are uncorrelated with one another. Once your data has been cleaned and transformed, you can fit an LDA model which will return a projected version of your data that has been reduced to the desired number of dimensions while maximizing class separability.