Need each point's leverage statistic. For simple linear regression leverage is given by (meaning one variable.)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

For multiple linear regression it's probably something like

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^{\circledast 2}}{\sum_{i=1}^{n}(x_i - \bar{x})^{\circledast 2}}$$

- $h_i \in [1/n, 1]$, average leverage is always $(p+1)/n$. So if some point has ~~high~~ significantly higher than $(p+1)/n$, we should be suspicious of high leverage
- ★ Outliers with high leverage can really screw up a model. (Plot standardized residuals vs. leverage to see).

2/9/23 Day 8

6. Collinearity

- Collinearity refers to the situation in which two or more predictor variables are closely related to one another.
- It can be difficult to separate the individual effects of collinear variables on the response.
- Collinearity increases the uncertainty of the coefficient estimates.
- Collinearity ⇒ increased standard error ⇒ $t = \hat{\beta}_j / SE(\hat{\beta}_j) \downarrow$ ⇒ failure to detect $\beta_j \neq 0$ ⇒ power of hypothesis test (probability of correctly detecting non-zero coefficient) decreases.

- You can detect collinearity between 2 variables by ~~detecting~~ inspecting the correlation matrix of the predictors.
- But, there is multicollinearity: when 3 or more variables are correlated even if all individual pairs are not.

Instead, we have a variance inflation factor (VIF). VIF is the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own.

$1 \leq VIF$

$VIF > 5$ or 10 is a problem

An expression for the VIF is

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

oh cool

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

If $R^2_{X_j|X_{-j}}$ is close to 1, then collinearity is present, so VIF will be large.

Solutions:

1) Drop one of the problematic variables
2) Combine the collinear predictors into a new variable (you could sum them or something

So, for the residuals, you could test the hypothesis that the best fit line to the residuals is 0, and if you can't reject, then proceed.

3.5: Comparison of Linear Regression w/ KNN Regression

KNN regression is a ~~parametric~~ nonparametric regression algorithm.

Given a value K and a prediction point x_0, KNN reg ids the K nearest predictor points (set N_0) and then estimates

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

Choose K according to bias-variance tradeoff

- Parametric forms outperform nonparametric if the parametric form is close to the true f.
- Note that in higher dimensions, KNN reg often performs worse than linear regression.
- This is because higher dimension effectively results in a decrease in sample size, the curse of dimensionality, where ~~neighbor~~ every point becomes more distant from its neighbors

Go back and do the lab.

Ch. 4: Classification

- "Qualitative" = "Categorical"

This chapter discusses 3 classifiers: logistic regression, linear discriminant analysis, and KNN

Later chapters discuss generalized additive models, trees, random forests, boosting, and support vector machines.

4.1: An Overview of Classification

Training observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$

4.2 Why not linear regression?

ex) Classify medical condition

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if drug overdose} \\ 3 & \text{if epileptic seizure} \end{cases}$$

4.3: Logistic Regression

- Models the probability that Y belongs to a particular category

4.3.1: The Logistic Model

(4.2) $$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$, the logistic function is $\frac{e^{\cdot}}{1+e^{\cdot}}$

To fit the model, we use maximum likelihood.

One can manipulate (4.2) to get

(4.3) $$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}$$

odds (Note that odds are better for horse betting).

Note that an increase by one unit in X will increase the odds by a factor of e^{β_1}.

4.3.2: Estimating the Regression Coefficients

- Maximum likelihood has good statistical properties, so this is the preferred way of estimating the coefficients.
- Basic intuition: choose β_0, β_1 so that $\hat{p}(x_i)$ matches the observed class for x_i as closely as possible, for all i.
- Likelihood function

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

Software can do it.

4.3.3: Making Predictions

- Just plug in the predictor values
- You can use dummy variables the same way as with linear regression.

4.3.4: Multiple Logistic Regression

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

$$\Leftrightarrow p(X) = \frac{e^{\sum_{i=0}^{p} \beta_i X_i}}{1 + e^{\sum_{i=0}^{p} \beta_i X_i}}$$

Use ML method to fit model.

Note that confounding occurs when other predictors are relevant, and so the confounding variables need to be included in the model.

4.3.5: Logistic Regression for >2 Response Classes

Could model
$P(Y=1|X)$ & $P(Y=2|X)$ and then
$P(Y=3|X) = 1 - P(Y=1|X) - P(Y=2|X)$
but ~~linear~~ discriminant analysis is more popular for this problem.

4.4: Linear Discriminant Analysis

- Approach: Model the distribution of the predictors X in each of the response classes. Then, use Bayes' theorem to flip these around into estimates for $P(Y=k|X=x)$.

Why use another method?

1) When classes are well separated, the parameter estimates in logistic regression are unstable. Not so for LDA
2) If n is small and the dsbn of X is approx normal in each class k, then LDA again more stable.

- LDA is more popular ~~works better~~ for multiple classes.

4.4.1) Using Bayes' Theorem for Classification

We want to classify an obs into one of $K \geq 2$ classes.

Let π_k represent the overall prior probability that a randomly chosen observation comes from the k-th class. The prior is the probability that the observation comes from class k.

Let $f_k(x) = \mathbb{P}(X = x \mid Y = k)$ denote the density function of X for an observation coming from class k.

Then Bayes' theorem states that

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

- π_k is the fraction of training data belonging to class k
- f_k is harder to estimate unless you assume a simple form.

4.4.2: Linear Discriminant Analysis for $p = 1$

- We have one predictor.
- We want to estimate $p_k(x)$ and class the observation into the most likely class.

Assuming

1) $f_k(x)$ is a normal density: $f_k(x) = [\sqrt{2\pi}\sigma_k]^{-1} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$
2) For now, $\sigma_1^2 = \cdots = \sigma_K^2 =: \sigma^2$

If we plug this into the estimator, we get

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

$$\log(p_k(x)) = \log(\pi_k \exp$$
$$= \log(\pi_k) - \frac{(x-\mu_k)^2}{2\sigma^2}$$

Taking the log of both sides and rearranging, we can show that this is equivalent to assigning x to the class k for which

$$\delta_k(x) := x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \quad \text{(go through :x)}$$

is largest

LDA estimates the Bayes classifier w/

$$\hat{\pi}_k = n_k/n$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \qquad \hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

So, based on the scratch work this is the simplification that makes the discriminant function linear.

4.4.3: Linear Discriminant Analysis for $p>1$

Assume $X = (X_1, \ldots, X_p)$ is drawn from a multivariate Gaussian distribution with class specific mean vectors and a common covariance matrix.

The multivariate Gaussian distribution

- Each predictor is μ, σ distributed, w/ some correlation.

We write this as $X \sim N(\mu, \Sigma)$, where Σ is the covariance matrix.

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

In the case of