CSCE 633: Machine Learning

Lecture 5: Linear Regression

Texas A&M University

9-4-19

Goals of this lecture

- Simple Linear Regression
- Multiple Linear Regression
- Convexity

Advertising Example

If n = 30

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

With $n = 30$ the t-statistic for the null hypothesis are around 2 and 2.75 respectively

We conclude $\beta_0 \neq 0$ and $\beta_1 \neq 0$

Optimal Coefficents: $\hat{\beta}_0$, $\hat{\beta}_1$

- $\boxed{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}}$
- $\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$

Important Questions to Ask

- Is there a relationship between budget and sales?
- If there is a relationship, how strong is it?
- Which of the three media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Multiple Linear Regression

- Advertising has more than just TV budget
- How do we account for them?

Multiple Linear Regression

- Advertising has more than just TV budget
- How do we account for them?
- 3 separate linear regressions?

Multiple Linear Regression

- Advertising has more than just TV budget
- How do we account for them?

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- β_j is the average effect on Y of a one unit change in X_j *holding all other parameters fixed*

$$sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper + \epsilon$$

Estimating Multiple Coefficients

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

- Again, a least squares approach
- $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2$
- Again, take the partial derivatives, set to 0, and solve. Complicated in this form
- Matrix form - later
- plenty of solvers to calculate this

Simple Regressions

Simple regression of `sales` on `radio`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	0.00115

Multiple Regressions

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Multiple Regressions

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

- newspaper budget acting as a surrogate for radio budget
- For example, shark attacks and ice cream sales related at a beach

More Important Questions

- Is at least one of the predictors $X_1, X_2, \cdots, X_p$ useful in predicting response Y?
- Do all predictors help explain Y? or only some?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict and how accurate is this prediction? F-statistic

More Important Questions

- Is at least one of the predictors $X_1, X_2, \cdots, X_p$ useful in predicting response Y?
- Do all predictors help explain Y? or only some?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict and how accurate is this prediction?

Variable Importance

- F-Statistic and p-values tell us at least one feature is related to response.
- Which one?

Variable Importance

- F-Statistic and p-values tell us at least one feature is related to response.
- Which one?
- Feature (Variable) Selection!

Variable Importance

- F-Statistic and p-values tell us at least one feature is related to response.
- Which one?
- Feature (Variable) Selection!
- Ideally, we would like to try a lot of sub models.
- p = 2, four models
- Pick best by some measure (AIC, BIC, Adjusted R^2)

Variable Importance

- F-Statistic and p-values tell us at least one feature is related to response.
- Which one?
- Feature (Variable) Selection!
- Ideally, we would like to try a lot of sub models.
- p = 2, four models
- Pick best by some measure (AIC, BIC, Adjusted R^2)
- But for p features, we have 2^p subsets

Forward (Greedy) Selection

- Start with the null model
- Fit p linear regressions of 1 variable
- Calculate RSS

Forward Selection

- Start with the null model
- Fit p linear regressions of 1 variable (problem for this: var2,3,4 in combination might be better than var1)
- Calculate RSS
- select the variable with lowest RSS
- repeat
- stop when some stopping criteria is met

Backward Selection

- Start with the full model
- calculate p-values
- remove the variable with largest p-value
- re-calculate
- repeat until some stopping criteria is met (for example, all remaining p-value $< \tau$)
- Cannot be used if $p > n$

Forward Backward (Mixed) Selection

- Start with no variables selected
- Add in a forward stepwise fashion
- But at east stage, check p-values
- If p-values for any variable become too large, remove them

More Important Questions

- Is at least one of the predictors $X_1, X_2, \cdots, X_p$ useful in predicting response Y?
- Do all predictors help explain Y? or only some?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict and how accurate is this prediction?

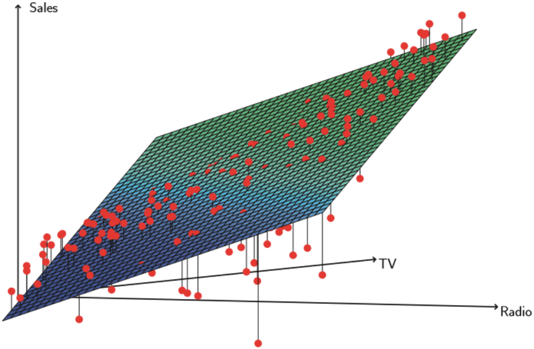
Model Fit

- Once the model with features selected is implemented, how do we measure fit?
- RSE and R^2 are the common measures
- R^2 is now $Cor(Y, \hat{Y})^2$
- However, more variables will still increase R^2 because you are fitting least squares
- RSE however, does not get better by just adding more features.
- In our advertising example, we eliminate newspaper from our model

More Important Questions

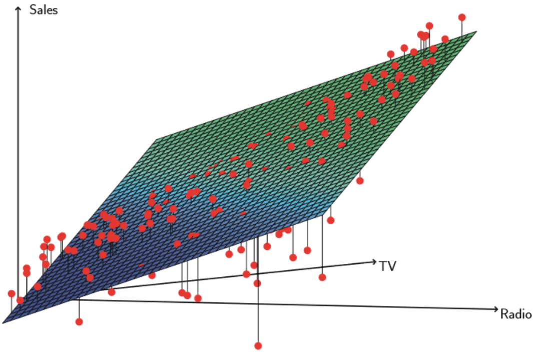
- Is at least one of the predictors $X_1, X_2, \cdots, X_p$ useful in predicting response Y?
- Do all predictors help explain Y? or only some?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict and how accurate is this prediction?

Residuals



- Positive residuals appear to fall along line balancing TV and Radio
- Negative residuals appear to fall outside of this range
- Indicates the combined interaction of TV and Radio could be important

Residuals



- Positive residuals appear to fall along line balancing TV and Radio
- Negative residuals appear to fall outside of this range
- Indicates the combined interaction of TV and Radio could be important
- But, before that, what about other kinds of predictors?

Credit Balance Example

```
> summary(Credit)
       ID            Income           Limit           Rating          Cards            Age          Education         Gender    Student 
 Min.   :  1.0   Min.   : 10.35   Min.   :  855   Min.   : 93.0   Min.   :1.000   Min.   :23.00   Min.   : 5.00    Male :193   No :360  
 1st Qu.:100.8   1st Qu.: 21.01   1st Qu.: 3088   1st Qu.:247.2   1st Qu.:2.000   1st Qu.:41.75   1st Qu.:11.00   Female:207   Yes: 40  
 Median :200.5   Median : 33.12   Median : 4622   Median :344.0   Median :3.000   Median :56.00   Median :14.00                         
 Mean   :200.5   Mean   : 45.22   Mean   : 4736   Mean   :354.9   Mean   :2.958   Mean   :55.67   Mean   :13.45                         
 3rd Qu.:300.2   3rd Qu.: 57.47   3rd Qu.: 5873   3rd Qu.:437.2   3rd Qu.:4.000   3rd Qu.:70.00   3rd Qu.:16.00                         
 Max.   :400.0   Max.   :186.63   Max.   :13913   Max.   :982.0   Max.   :9.000   Max.   :98.00   Max.   :20.00                         
 Married              Ethnicity      Balance       
 No :155   African American: 99   Min.   :   0.00  
 Yes:245   Asian           :102   1st Qu.:  68.75  
           Caucasian       :199   Median : 459.50  
                                  Mean   : 520.01  
                                  3rd Qu.: 863.00  
                                  Max.   :1999.00  
```

Credit Balance Example

```
> summary(Credit)
       ID             Income           Limit           Rating          Cards            Age          Education        Gender     Student
 Min.   :  1.0   Min.   : 10.35   Min.   :  855   Min.   : 93.0   Min.   :1.000   Min.   :23.00   Min.   : 5.00    Male :193   No :360
 1st Qu.:100.8   1st Qu.: 21.01   1st Qu.: 3088   1st Qu.:247.2   1st Qu.:2.000   1st Qu.:41.75   1st Qu.:11.00   Female:207   Yes: 40
 Median :200.5   Median : 33.12   Median : 4622   Median :344.0   Median :3.000   Median :56.00   Median :14.00
 Mean   :200.5   Mean   : 45.22   Mean   : 4736   Mean   :354.9   Mean   :2.958   Mean   :55.67   Mean   :13.45
 3rd Qu.:300.2   3rd Qu.: 57.47   3rd Qu.: 5873   3rd Qu.:437.2   3rd Qu.:4.000   3rd Qu.:70.00   3rd Qu.:16.00
 Max.   :400.0   Max.   :186.63   Max.   :13913   Max.   :982.0   Max.   :9.000   Max.   :98.00   Max.   :20.00
 Married              Ethnicity       Balance
 No :155   African American: 99   Min.   :   0.00
 Yes:245   Asian           :102   1st Qu.:  68.75
           Caucasian       :199   Median : 459.50
                                  Mean   : 520.01
                                  3rd Qu.: 863.00
                                  Max.   :1999.00
```

- Qualitative and quantitative
- What if we want to investigate the difference in balances between males and females?

Factors

- A categorical variable with multiple levels
- Take a factor of two levels - created indicator or dummy variables

Factors

- A categorical variable with multiple levels
- Take a factor of two levels - created indicator or dummy variables

$$x_i = \begin{cases} 1, & \text{if ith person is female} \\ 0, & \text{else} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if ith person is female} \\ \beta_0 + \epsilon_i, & \text{else} \end{cases}$$

Credit Balance: Factors

	Coefficient	Std. error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender[Female]	19.73	46.05	0.429	0.6690

- p-value for dummy variable is very high, what are β_0 and β_1?
- $0/1$ coding is arbitrary, no effect on regression fit, but does alter interpretation
- Could also code as $\{-1, +1\}$

Credit Balance: Factors

	Coefficient	Std. error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender[Female]	19.73	46.05	0.429	0.6690

- p-value for dummy variable is very high, what are β_0 and β_1?
- $0/1$ coding is arbitrary, no effect on regression fit, but does alter interpretation
- Could also code as $\{-1, +1\}$
- With that, β_0 is now the average credit balance independent of gender effect, while β_1 models the impact of gender.

Credit Balance: Factors

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       519.670     23.026  22.569   <2e-16 ***
gender_indicator    9.867     23.026   0.429    0.669
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- p-value for dummy variable is very high, what are β_0 and β_1?
- $0/1$ coding is arbitrary, no effect on regression fit, but does alter interpretation
- Could also code as $\{-1, +1\}$
- With that, β_0 is now the average credit balance independent of gender effect, while β_1 models the impact of gender.

Credit Balance Example

```
> summary(Credit)
       ID            Income            Limit           Rating          Cards            Age          Education        Gender      Student
 Min.   :  1.0   Min.   : 10.35   Min.   :  855   Min.   : 93.0   Min.   :1.000   Min.   :23.00   Min.   : 5.00    Male :193   No :360
 1st Qu.:100.8   1st Qu.: 21.01   1st Qu.: 3088   1st Qu.:247.2   1st Qu.:2.000   1st Qu.:41.75   1st Qu.:11.00   Female:207   Yes: 40
 Median :200.5   Median : 33.12   Median : 4622   Median :344.0   Median :3.000   Median :56.00   Median :14.00
 Mean   :200.5   Mean   : 45.22   Mean   : 4736   Mean   :354.9   Mean   :2.958   Mean   :55.67   Mean   :13.45
 3rd Qu.:300.2   3rd Qu.: 57.47   3rd Qu.: 5873   3rd Qu.:437.2   3rd Qu.:4.000   3rd Qu.:70.00   3rd Qu.:16.00
 Max.   :400.0   Max.   :186.63   Max.   :13913   Max.   :982.0   Max.   :9.000   Max.   :98.00   Max.   :20.00
 Married               Ethnicity      Balance
 No :155   African American: 99   Min.   :   0.00
 Yes:245   Asian           :102   1st Qu.:  68.75
           Caucasian       :199   Median : 459.50
                                  Mean   : 520.01
                                  3rd Qu.: 863.00
                                  Max.   :1999.00
```

- Qualitative and quantitative
- What if we want to investigate a factor with more levels?

Factors

- Take a factor of multiple levels - create multiple indicator or dummy variables

$$x_{i1} = \begin{cases} 1, & \text{if ith person is Asian} \\ 0, & \text{else} \end{cases}$$

$$x_{i2} = \begin{cases} 1, & \text{if ith person is Caucasian} \\ 0, & \text{else} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i =$$

$$\begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if ith person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i, & \text{if ith person is Caucasian} \\ \beta_0 + \epsilon_i, & \text{African American} \end{cases}$$

Factors

- Take a factor of multiple levels - create multiple indicator or dummy variables

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if ith person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i, & \text{if ith person is Caucasian} \\ \beta_0 + \epsilon_i, & \text{African American} \end{cases}$$

- β_0 average credit balance for African American
- β_1 Diff in average balance between Asian and African American
- β_2 Diff in average balance between Caucasian and African American
- Always 1 fewer dummy variable than level in factor.
- Level with no dummy variable is your baseline for comparison
- F-Statistic - reject hypothesis of no relationship between balance and ethnicity

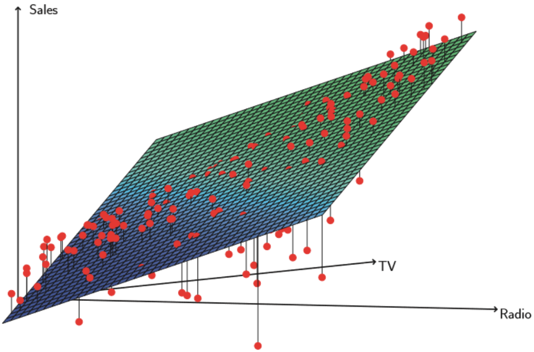
Factors

```
Residuals:
    Min      1Q  Median      3Q     Max 
-531.00 -457.08  -63.25  339.25 1480.50 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   531.00      46.32  11.464   <2e-16 ***
asian         -18.69      65.02  -0.287    0.774    
caucasian     -12.50      56.68  -0.221    0.826    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 460.9 on 397 degrees of freedom
Multiple R-squared:  0.0002188,	Adjusted R-squared:  -0.004818 
F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

Important Questions to Ask

- Is there a relationship between budget and sales?
- If there is a relationship, how strong is it?
- Which of the three media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Residuals



- Positive residuals appear to fall along line balancing TV and Radio
- Negative residuals appear to fall outside of this range
- Indicates the combined interaction of TV and Radio could be important
- But, before that, what about other kinds of predictors?

Extending Additive and Linear Assumptions on X and Y

- TV and Radio both associated with sales
- 1 unit increase in TV increases sales, independent of radio budget
- But what if radio budget improves effectiveness of TV?
- We say there is a synergy in marketing, we call this an interaction effect in machine learning

Interaction Terms

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Interaction Terms

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta_1} X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

- The effect of X_1 on Y is no longer a constant

Interaction Terms

$$sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \epsilon$$

$$\begin{aligned} sales &= \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 TVRadio + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 Radio) TV + \beta_2 Radio + \epsilon \\ &= \beta_0 + \tilde{\beta_1} TV + \beta_2 Radio + \epsilon \end{aligned}$$

- The effect of X_1 on Y is no longer a constant

Interaction Terms

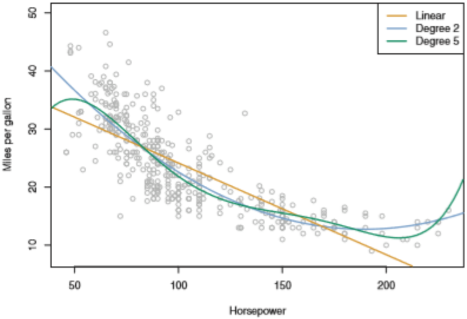
	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

- Superior p-values to the main effects model
- If the p-value of the interaction term is important? Do we keep the main effects terms in the model?

Final Important Questions

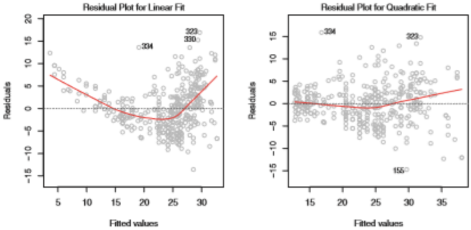
- What if the data relationship is not linear?
- What if the error terms are correlated?
- What if there is a non-constant variance in error terms?
- What about outlier points?
- What about high-leverage points?
- What about variables that are collinear?

Non-Linear (Polynomial) Regression



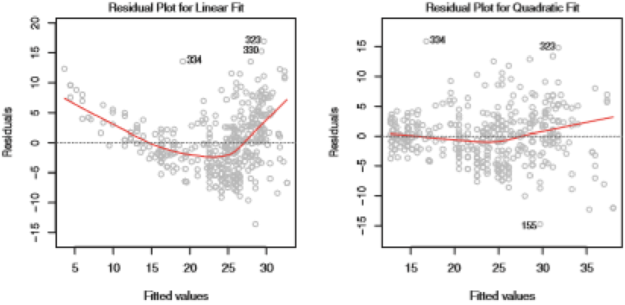
- $mpg = \beta_0 + \beta_1 HP + \beta_2 HP^2 + \epsilon$
- Still a linear model - so can solve with normal software
- But why not go to 3rd degree? 4th?
- Can I tell linearity after I build a model?

Non-Linear (Polynomial) Regression



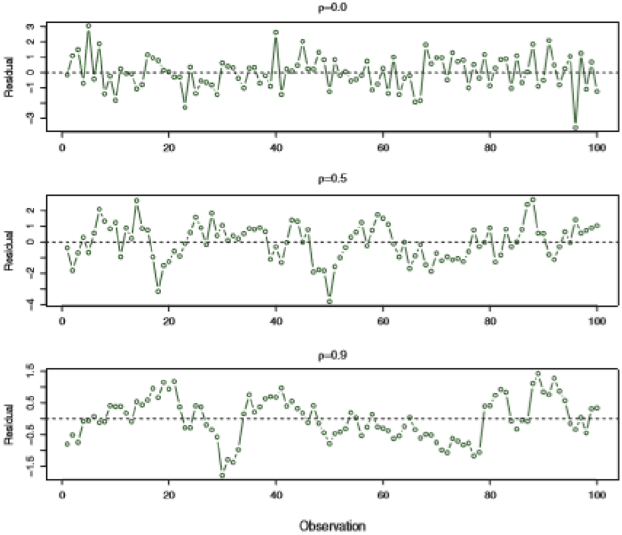
- Pattern in residuals usually indicates higher-order interactions

Non-Linear (Polynomial) Regression



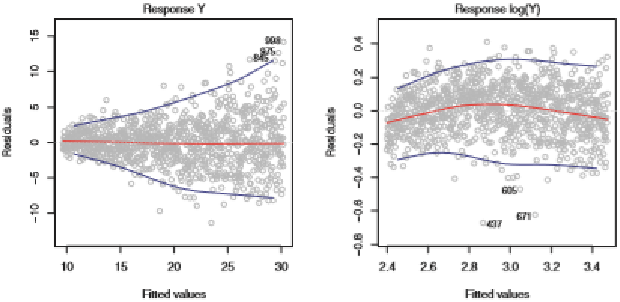
- Pattern in residuals usually indicates higher-order interactions

Correlated Residuals



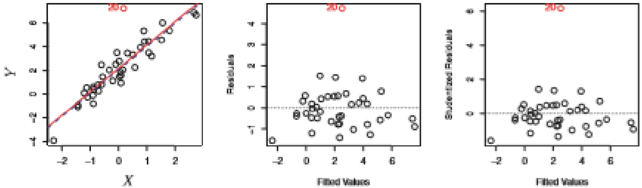
- Standard error underestimates if they are correlated
- Time series - error of near by terms often correlated

Variance Residuals



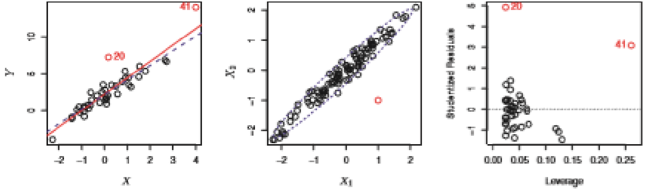
- $Var(\epsilon) = \sigma^2$ - funnel shape in plot
- Solve with a weighted least squares, where the weights are proportional to the inverse of class distribution. $\frac{\sigma^2}{n_i}$

Outliers



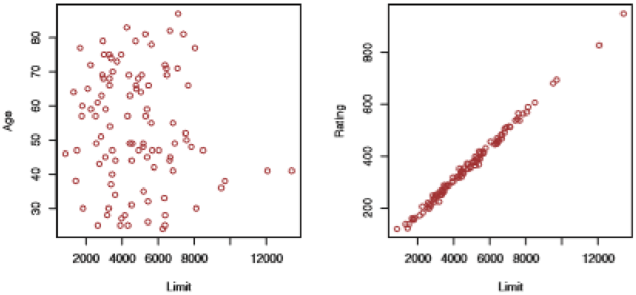
- Studentized residual - divide each residual by its standard error. Any value > 3 or < -3 is likely to be an outlier

High Leverage Points



- High Leverage Points are those with rare X_i values
- Create a leverage statistic $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$
- Always between $\frac{1}{n}$ and 1 - average is always $\frac{p+1}{n}$

Collinearity



- Power of hypothesis test is reduced because t-statistic divides β by standard error, which goes down with collinearity
- Variance Inflation Factor $= \frac{1}{1-R^2_{x_j|x_{-j}}}$, comparing regressions of all but j

Important Questions to Ask

- Is there a relationship between budget and sales?
- If there is a relationship, how strong is it?
- Which of the three media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Least Squares with Multiple Variables

$p(y|x, \theta) = N(y|\beta^T x, \sigma^2)$

Maximum Likelihood Estimation results in:
$\hat{\theta} = argmax_\theta \log p(D|\theta)$

Assume training data are independent and identically distributed - then the log-likelihood is:
$l(\theta) = \log p(D|\theta) = \sum_{i=1}^{n} \log p(y_i|x_i, \theta)$

Least Squares with Multiple Variables

$p(y|x, \theta) = N(y|\beta^T x, \sigma^2)$

Maximum Likelihood Estimation results in:
$\hat{\theta} = argmax_\theta \log p(D|\theta)$

Assume training data are independent and identically distributed - then the log-likelihood is:
$l(\theta) = \log p(D|\theta) = \sum_{i=1}^{n} \log p(y_i|x_i, \theta)$

Can equivalently minimize the negative log-likelihood:
$NLL(\theta) = -\sum_{i=1}^{n} \log p(y_i|x_i, \theta)$

Least Squares with Multiple Variables

The log-likelihood is:
$l(\theta) = \log p(D|\theta) = \sum_{i=1}^{n} \log p(y_i|x_i, \theta)$

Can insert our definition of the Gaussian into this formula to get:
$l(\theta) = \sum_{i=1}^{n} \log \left[(\frac{1}{2\pi\sigma^2})^2 exp(-\frac{1}{2\sigma^2}(y_i - \beta^T x_i)^2) \right]$

$= -\frac{1}{2\sigma^2} RSS - \frac{n}{2} \log(2\pi\sigma^2)$

RSS of β vector

- $RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 = \|e\|_2^2$
- Mean Squared Error (MSE) $= \frac{RSS}{n}$
- MLE for β is one that minimizes RSS (least squares)

Differentiation of NLL

- We re-write the NLL so it is easier to differentiate
- $NLL(\beta) = \frac{1}{2} - (y - X\beta)^T(y - X\beta) = \frac{1}{2}\beta^T(X^TX)\beta - \beta^T(X^Ty)$
- Where $X^TX = \sum_{i=1}^{n} x_i x_i^T$ is a $p \times p$ matrix - sum of squares

Differentiation of NLL

- We re-write the NLL so it is easier to differentiate
- $NLL(\beta) = \frac{1}{2} - (y - X\beta)^T(y - X\beta) = \frac{1}{2}\beta^T(X^TX)\beta - \beta^T(X^Ty)$
- Where $X^TX = \sum_{i=1}^{n} x_i x_i^T$ is a $p \times p$ matrix - sum of squares
- $X^Ty = \sum_{i=1}^{n} x_i y_i$

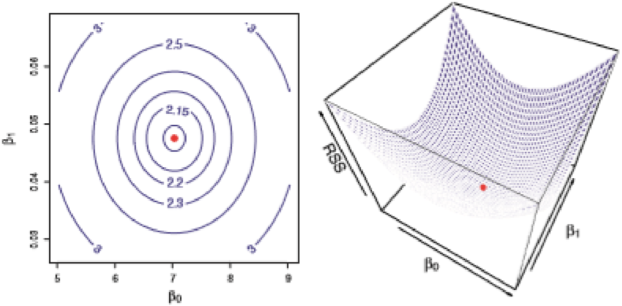
Differentiation with vectors

- $\frac{\partial(b^T a)}{\partial a} = b$
- $\frac{\partial(a^T Aa)}{\partial a} = (A + A^T)a$
- $\frac{\partial}{\partial a} tr(BA) = B^T$ where $tr(A) = \sum_i A_i i$ is the trace of the matrix
- $\frac{\partial}{\partial a} \log|A| = A^{-T} = (A^{-1})^T$
- $tr(ABC) = tr(CAB) = tr(BCA)$

Differentiation of NLL

- We re-write the NLL so it is easier to differentiate
- $NLL(\beta) = \frac{1}{2} - (y - X\beta)^T(y - X\beta) = \frac{1}{2}\beta^T(X^TX)\beta - \beta^T(X^Ty)$
- Where $X^TX = \sum_{i=1}^{n} x_i x_i^T$ is a $p \times p$ matrix - sum of squares
- $X^Ty = \sum_{i=1}^{n} x_i y_i$
- Gradient $g(\beta) = (X^TX\beta - X^Ty) = \sum_{i=1}^{n} x_i(\beta^T x_i - y_i)$
- Setting $= 0$, we get $X^TX\beta = X^Ty$
- So we get $\beta_{OLS} = (X^TX)^{-1}X^Ty$

Why does this work? Convexity



Convexity

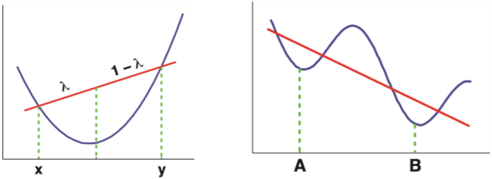
- Set S is convex if for any θ, $\theta' \in S$, there exists
- $\lambda\theta + (1-\lambda)\theta' \in S \forall \lambda \in [0, 1]$
- In practice - draw a line between two points in a Set, and it is convex if every point on the line still lies within the set

Convexity

- Set S is convex if for any θ, $\theta' \in S$, there exists
- $\lambda\theta + (1 - \lambda)\theta' \in S \forall \lambda \in [0, 1]$
- In practice - draw a line between two points in a Set, and it is convex if every point on the line still lies within the set

(a) (b)

Figure 7.4 (a) Illustration of a convex set. (b) Illustration of a nonconvex set.



(a) (b)

Convexity

- A function $f(\theta)$ is convex if its epigraph (the set of points above the function) defines a convex set.
- A function $f(\theta)$ is convex if it is defined on a convex set and if, for any θ, $\theta' \in S$, and for any $0 \leq \lambda \leq 1$
- $f(\lambda\theta + (1-\lambda)\theta') \leq \lambda f(\theta) + (1-\lambda)f(\theta')$
- If the inequality is strict, this is called strictly convex
- If $f(\theta)$ is concave, then $-f(\theta)$ is convex
- Second Derivative Test $\frac{\partial^2}{\partial\theta^2}f(\theta) > 0$ then f is convex

Convexity

- Question: Assume the following non-linear regression model.
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2$
- $RSS = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2))^2$
- Which of the following is true?
- A: We don't know if RSS has a global minimum with respect to β_0, β_1, β_2
- B: RSS has a local minimum with respect to β_0, β_1, β_2, which is dependent on the training data
- C: RSS has a local minimum with respect to β_0, β_1, β_2, which is also the global minimum

Convexity

- C: RSS has a local minimum with respect to β_0, β_1, β_2, which is also the global minimum
- Rename $\beta = (\beta_0, \beta_1, \beta_2)^T$, $z = (1, x, x^2)^T$
- $RSS = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2))^2$
- $= \sum_{i=1}^{n}(y_i - \beta^T z)^2$ which is convex

Takeaways and Next Time

- Ordinary Least Squares Optimization
- Linear Regression
- Convexity and Optimization
- Next Time: More Complex Regressions/Classifications and Regularization
- example and figure sources: James, Witten, Hastie, Tibshirani (ISLR)