

Machine Learning I

Lecture 7: Linear Regression

Nathaniel Bade

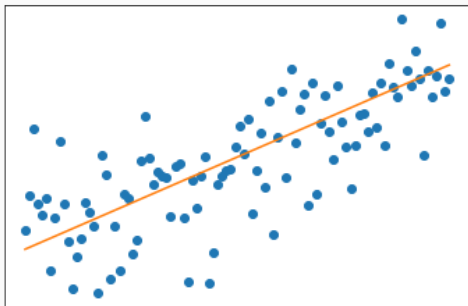
Northeastern University Department of Mathematics

Table of contents

1. Least Squares for Linear Models
2. Distribution of Parameters
3. Estimating Parameter Variance
4. Confidence Intervals for Coefficients
5. Fit, standard error and t statistic on example data
6. Dropping Variables by Statistical Significance
7. Subset Selection Methods

Least Squares for Linear Models

Linear Models: Definition

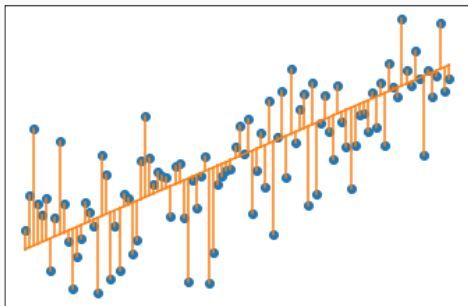


A **linear regression model** is a model of data fitting that assume that the regression function $E(Y|X)$ is linear in the inputs X_1, \dots, X_p ,

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

and uses the residual sum of squares $RSS(\beta)$ as a loss function. Here ϵ is a random variable with $E[\epsilon] = 0$.

Linear Models: Definition

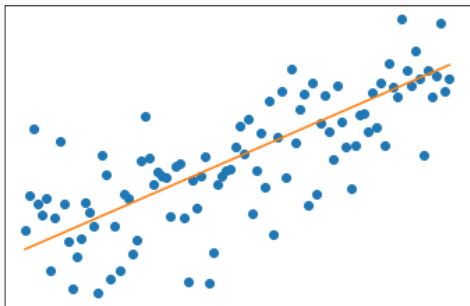


A **linear regression model** is a model of data fitting that assume that the regression function $E(Y|X)$ is linear in the inputs X_1, \dots, X_p ,

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

and uses the residual sum of squares $RSS(\beta)$ as a loss function. Here ϵ is a random variable with $E[\epsilon] = 0$.

Linear Models: Definition

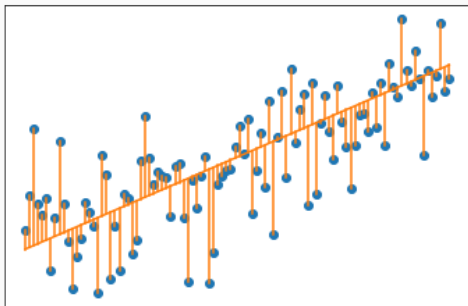


In vector form, we add a constant term $X = (1, X_1, \dots, X_p)$ and let β be a $p + 1$ vector including β_0 :

$$Y = \beta_0 + X_1\beta_1 + \dots + X_p\beta_p + \epsilon = X^T\beta + \epsilon,$$

where again we use the residual sum of squares $RSS(\beta)$ as a loss function.

Linear Models: ERM Solution

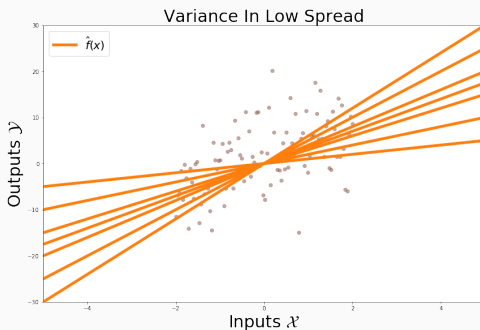


Given the $N \times (p + 1)$ matrix \mathbf{X} of N input data points x_i and a vector \mathbf{y} of the N corresponding labels y_i , we saw that

$$RSS(\beta) = \sum_{i=1}^N (y_i - x_i^T \beta)^2 = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

is minimized by $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$.

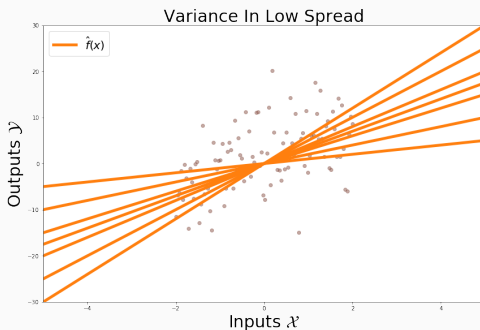
Linear Models: Bias-Complexity



We also have shown that for the fit linear function $\hat{f} = X^T \hat{\beta}$, the bias-complexity trade off can be written as a splitting of the pointwise error into a bias term and a variance term

$$Err(x_0) = E_{\mathcal{T}} [y - \hat{f}(x_0)]^2 = \sigma_{\epsilon}^2 + \text{Bias}^2(x_0) + \text{Var}(x_0)$$

Linear Models: Bias-Complexity

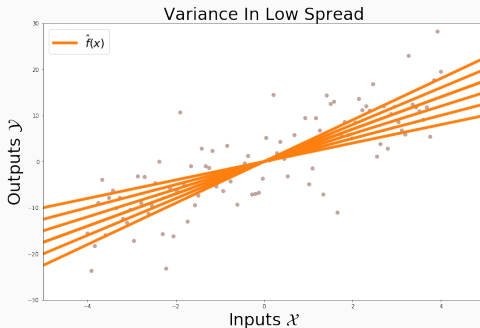


If the data is actually linear, we saw that the bias was 0 and the variance was given by and expression

$$\text{Var}(x_0) = \sigma_{\epsilon}^2 x_0^T E_X[(\mathbf{X}^T \mathbf{X})^{-1}] x_0.$$

which was in turn a function of variance of the inputs x_i .

Linear Models: Pointwise Variance

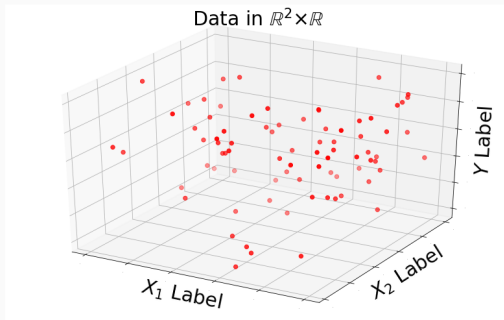


If the data is actually linear, we saw that the bias was 0 and the variance was given by and expression

$$\text{Var}(x_0) = \sigma_{\epsilon}^2 x_0^T E_X[(\mathbf{X}^T \mathbf{X})^{-1}] x_0.$$

which was proportional to the variance of the inputs x_i .

Linear Models: Pointwise Variance



When fitting multivariate (multi-feature) linear models, one of our main goals is to understand which parameters most directly control the labeling.

For a multilinear model, the coefficients β give some measure of correlation, but we still need to investigate what exactly these variables mean, what constitute expected and unexpected values.

Example Data

In the lecture we're going to use as an example a data set of undergraduate confidence in admission to graduate school. The data has 8 input features and 1 output feature

Data Sample:

Serial No.	GRE Score	TOEFL Score	University Rating
1	337	118	4
2	324	107	4
3	316	104	3

SOP	LOR	CGPA	Research	Chance of Admit
4.5	4.5	9.65	1	0.92
4.0	4.5	8.87	1	0.76
3.0	3.5	8.00	1	0.72

Example Data

The features of the data are

Serial No.

GRE Score: Out of 340.

TOEFL Score: Out of 120.

University Ranking: Out of 5.

SOP: Statement of purpose out strength of 5.

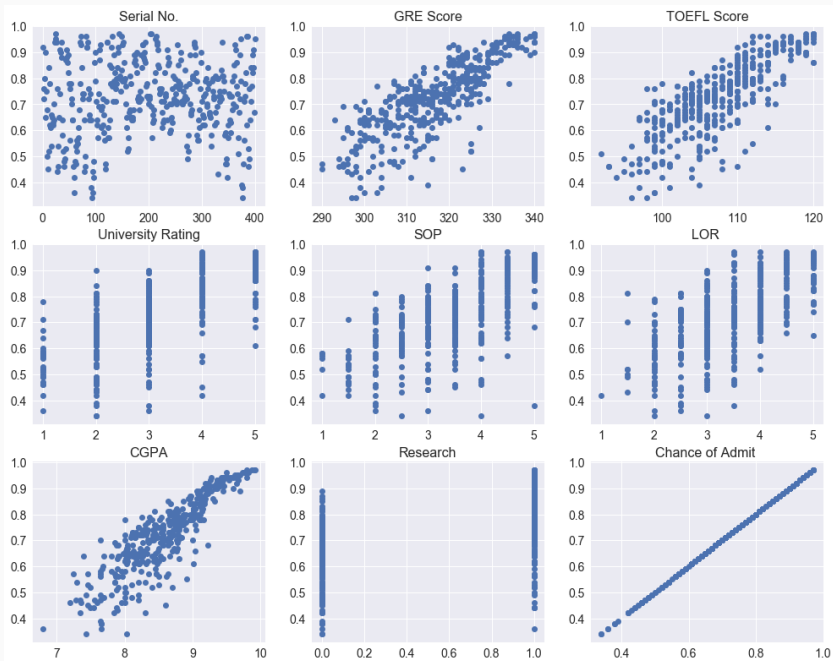
LOR: Letter of recommendation strength out of 5.

CGPA: Undergraduate GPA out of 10.

Research: Research experience 0 or 1.

Chance of Admit: Chance of admission.

Note: It is unclear how this data was collected, it is not authoritative.



Linear Models: Parameter Variance

The fit parameter $\hat{\beta}_i$ in the linear model

$$\hat{Y} = X^T \hat{\beta}$$

correspond roughly to the expected correlation between the i 'th feature and the chance of admission. Before we discuss the fit of the model, we will discuss the quantitative meaning of the parameters and understand what the data informs us an expected for the parameters $\hat{\beta}_i$.

In particular, we want to understand quantitatively when a parameter value $\hat{\beta}_i$ is statistically significant. Statistically, this mean we want expressions for the p -values and confidence intervals for the $\hat{\beta}_i$ that can be computed from the data set.

Mathematically, this boils down to understanding variance of $\hat{\beta}_i$ as a function of the input data \mathbf{X} and \mathbf{y} .

Distribution of Parameters

Linear Models: Parameter Variance

The derivation of the variance of the parameters $\text{Var}(\hat{\beta})$ has a similar derivation to the pointwise variance $\text{Var}(x_0)$ from Lecture 5.

We will perform the derivation in a slightly different way, using the properties of variance instead of the properties of expectation value. In particular, we will use the change in variance under the action of a linear transform.

Variance of Linear Transform

For vector ϵ of i.i.d. random variables each with variance σ_ϵ^2 , and a matrix A ,

$$\text{Var}(A\epsilon) = \sigma_\epsilon^2 A A^T.$$

This follow from the definition of variance:

$\text{Var}(A\epsilon) = E[(A\epsilon - E[A\epsilon])(A\epsilon - E[A\epsilon])^T]$	Definition of Var.
$= A E[(\epsilon - E[\epsilon])(\epsilon - E[\epsilon])^T] A^T$	A constant
$= A \text{Var}(\epsilon) A^T$	Definition of Var.
$= A(\sigma_\epsilon^2 I) A^T$	ϵ i.i.d.
$= \sigma_\epsilon^2 A A^T$	

The second the last line come from the fact that i.i.d. variables are uncorrelated, and so all off diagonal entries are 0.

Linear Models: Parameter Variance

The formula

$$\text{Var}(A\epsilon) = \sigma_\epsilon^2 AA^T. \quad (1)$$

allow us to make a direct derivation of the parameter variance. Assume that the input points are fixed and the variance comes only from the labels. Then

$$\begin{aligned} \text{Var}_\epsilon(\hat{\beta}) &= \text{Var}_\epsilon((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}) && \text{Definition of } \hat{\beta}, \\ &= \text{Var}_\epsilon((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X}\beta + \epsilon)) && \text{Definition of } \mathbf{y}, \\ &= \text{Var}_\epsilon(\beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon) && \text{Distribution,} \\ &= \sigma_\epsilon^2 ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T)^T && (1) \text{ Above,} \\ &= \sigma_\epsilon^2 (\mathbf{X}^T \mathbf{X})^{-1} && \text{Multiplication.} \end{aligned}$$

Linear Models: Parameter Variance

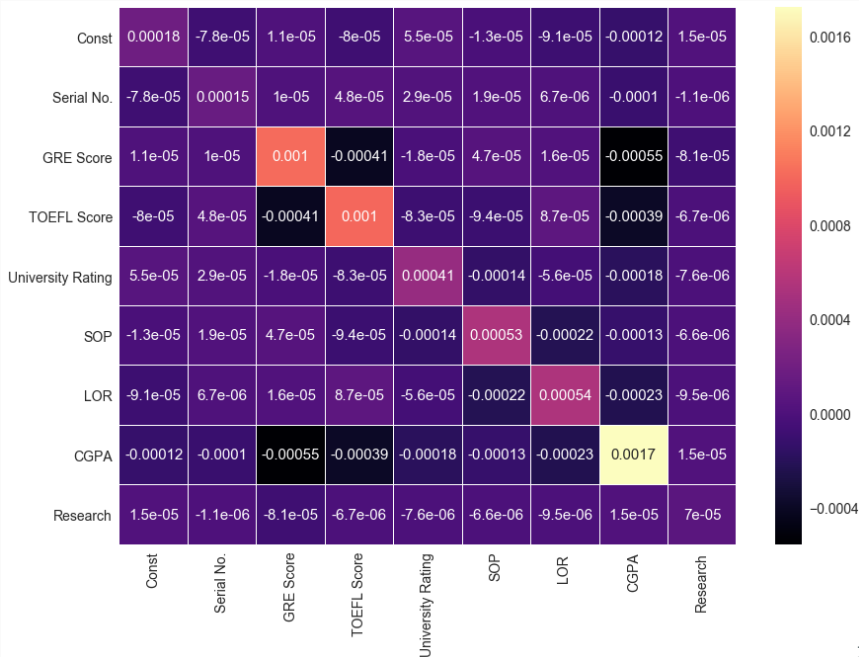
The variation of the parameters is then a $p + 1 \times p + 1$ matrix

$$\text{Var}(\hat{\beta}) = \sigma_{\epsilon}^2 (\mathbf{X}^T \mathbf{X})^{-1},$$

with each term scaled by the variance of the noise parameter (recall that \mathbf{X} contains a column of 1's).

Note, that unlike for the variance matrix of the i.i.d. variables ϵ in the proof, we expect there to be correlation between the β_i 's, and there will almost certainly be non-diagonal terms in the matrix.

For example, the variance in the parameters for the admission dataset is



Estimating Parameter Variance

Estimating Total Variance

In practice, the variance σ_ϵ^2 of the random variable isn't given to us but must be estimated from the data. The variance is typically estimated by

$$\hat{\sigma}^2 = \frac{1}{N - p - 1} \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \frac{RSS(\hat{\beta})}{N - p - 1}.$$

The factor of $N - p - 1$ in the denominator makes $\hat{\sigma}$ into an **unbiased estimator**, in the sense that

$$E[\hat{\sigma}^2] = \sigma^2.$$

Linear Models: Parameter Variance

The intuition here is roughly that just as the sample variance for one parameter is correct by

$$\bar{s} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}),$$

the sample variance for our $p+1$ parameters will be corrected by a $p+1$ shift.

$$\hat{\sigma}^2 = \frac{1}{N-p-1} \sum_{i=1}^N (y_i - \hat{y}_i)^2.$$

Since

$$\text{Var}_\epsilon(\hat{\beta}) = \sigma_\epsilon^2 (\mathbf{X}^T \mathbf{X})^{-1},$$

and the linear regression is an unbiased predictor, $E_{\mathcal{T}}[\hat{\beta}] = \beta_*$. If we assume that ϵ is normally distribution, that is $\epsilon \sim N(0, \sigma_\epsilon^2)$, then this means that

$$\hat{\beta} \sim N(\beta_*, \sigma_\epsilon^2 (\mathbf{X}^T \mathbf{X})^{-1}).$$

That is, the linear parameters are normally distributed around the true solution β_* with covariance matrix $\sigma_\epsilon^2 (\mathbf{X}^T \mathbf{X})^{-1}$.

Linear Models: Parameter Variance

The standard deviation for each coefficient can be read off the diagonal entries of the covariance matrix $\sigma_\epsilon^2 (\mathbf{X}^T \mathbf{X})^{-1}$,

$$\hat{\sigma}_i = \sqrt{\text{Var}(\hat{\beta}_i)} = \sigma_\epsilon \sqrt{(\mathbf{X}^T \mathbf{X})_{ii}^{-1}},$$

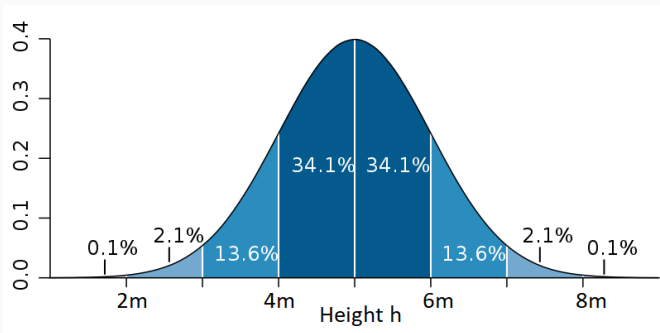
where $(\mathbf{X}^T \mathbf{X})_{ii}^{-1}$ is the i 'th diagonal entry of $(\mathbf{X}^T \mathbf{X})^{-1}$.

To test the hypothesis that $\beta_i = 0$, that is the i 'th feature has no bearing on the outcome, we write the standardized z score

Z-Score of β_i

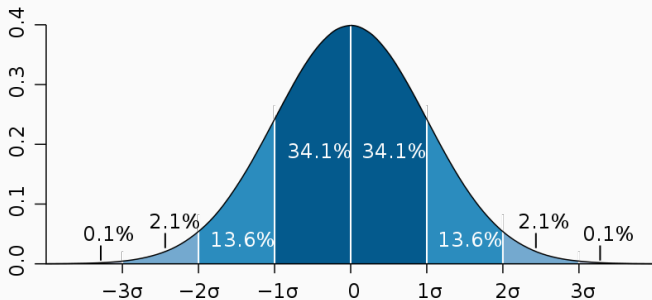
$$z_i \approx \frac{\hat{\beta}_i}{\hat{\sigma}_i} = \frac{\hat{\beta}_i}{\sigma_\epsilon \sqrt{(\mathbf{X}^T \mathbf{X})_{ii}^{-1}}}.$$

Z-scores



Note: The z-score rescales the values measured from normally distributed variable to be in units of standard deviation from the mean. The z-score gives a rough heuristic for the probability a value was arrived at by chance.

Z-scores



Note: The z-score rescales the values measured from normally distributed variable to be in units of standard deviation from the mean. The z-score gives a rough heuristic for the probability a value was arrived at by chance.

Evaluating Coefficients Based on Variance

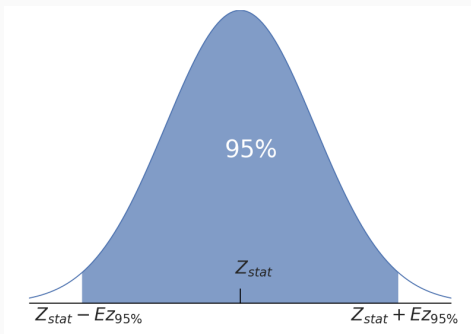
Note the the estimate in the z-score

$$\boxed{\text{Z-Score of } \beta_i} \quad z_i \approx \frac{\hat{\beta}_i}{\hat{\sigma}_i} = \frac{\hat{\beta}_i}{\sigma_\epsilon \sqrt{(\mathbf{X}^T \mathbf{X})_{ii}^{-1}}}$$

appears because z_i should really be drawn from a t distribution with $N - p - 1$ degrees of freedom. However, for $N \gg 1$ the t distribution is similar to a normal distribution, which practically may be easier to calculate.

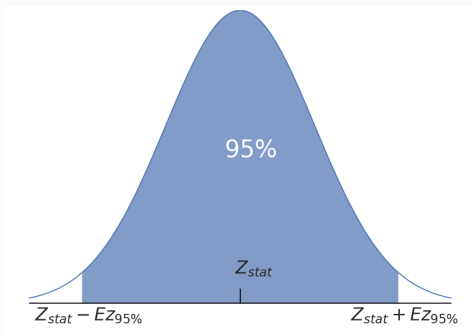
Confidence Intervals for Coefficients

Confidence Intervals for Coefficients



As a reminder, to construct a confidence interval, we assume the random variables are distributed according to a Gaussian. We then find $z_{a\%}$ by considering the number of standard deviations from the mean between which $a\%$ of the standard normal distribution is contained.

Confidence Intervals for Coefficients



If a statistic α is truly normally distributed, the confidence interval centered at α is the interval on which there is a $a\%$ chance the true value of Z lies:

$$\text{C.I.} : [\alpha - z_{a\%} E, \alpha + z_{a\%} E].$$

Here, $E = \alpha/n$ is the standard error of α . It is given by the standard deviation of α divided by the number of observations n .

Fit, standard error and t statistic on example data

Confidence Intervals for Coefficients

For the parameters β , the standard error is

$$E_i = \hat{\sigma}_\epsilon \sqrt{(\mathbf{X}^T \mathbf{X})_{ii}^{-1}}$$

and will allow us to compute the confidence intervals for the fit variables $\hat{\beta}$. As a reminder, to construct a confidence interval, we assume the random variables are distributed according to a Gaussian. The confidence interval for $\hat{\beta}_i$ is then

$$[\hat{\beta}_i - z_{a\%} \hat{\sigma}_\epsilon \sqrt{(\mathbf{X}^T \mathbf{X})_{ii}^{-1}}, \hat{\beta}_i + z_{a\%} \hat{\sigma}_\epsilon \sqrt{(\mathbf{X}^T \mathbf{X})_{ii}^{-1}}].$$

T-score For Application Data

Example: The fit coefficients for the admissions data with their standard errors.

Variable	coef	std err	t
const	0.3257	0.014	24.024
Serial No.	0.0722	0.012	5.884
GRE Score	0.0979	0.032	3.077
TOEFL Score	0.0994	0.032	3.144
University Rating	0.0433	0.020	2.133
SOP	0.0078	0.023	0.336
LOR	0.0663	0.023	2.860
CGPA	0.3077	0.041	7.418
Research	0.0217	0.008	2.587

T-score For Application Data

We find the highest significance for the constant β_0 , with the second highest being for GPA.

Variable	coef	std err	t
const	0.3257	0.014	24.024
Serial No.	0.0722	0.012	5.884
GRE Score	0.0979	0.032	3.077
TOEFL Score	0.0994	0.032	3.144
University Rating	0.0433	0.020	2.133
SOP	0.0078	0.023	0.336
LOR	0.0663	0.023	2.860
CGPA	0.3077	0.041	7.418
Research	0.0217	0.008	2.587

T-score For Application Data

Surprisingly, we also find a high significance for the coefficient of the Serial No. This is suspicious, and we see that the coefficient itself is small.

Variable	coef	std err	t
const	0.3257	0.014	24.024
Serial No.	0.0722	0.012	5.884
GRE Score	0.0979	0.032	3.077
TOEFL Score	0.0994	0.032	3.144
University Rating	0.0433	0.020	2.133
SOP	0.0078	0.023	0.336
LOR	0.0663	0.023	2.860
CGPA	0.3077	0.041	7.418
Research	0.0217	0.008	2.587

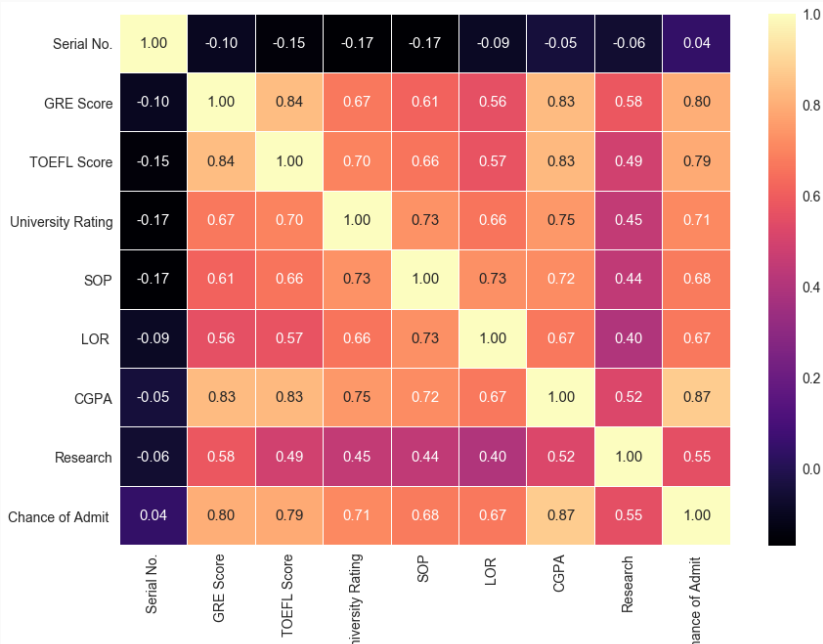
T-score For Application Data

Surprisingly, we also find a high significance for the coefficient of the Serial No. This is suspicious, and we see that the coefficient itself is small.

Variable	coef	std err	t
Serial No.	0.0722	0.012	5.884

This high significance mostly likely comes from the fact that the correlation is so low that the standard error is low. It means that the model may be over fitting, using a effectively random parameter to generate a tighter fit to the specific dataset. Such terms tend to lead to a higher variance in our models, and should be eliminated if possible. If we compute the correlation matrix we will see that Serial No. is effectively uncorrelated with the acceptance percentage.

Correlation



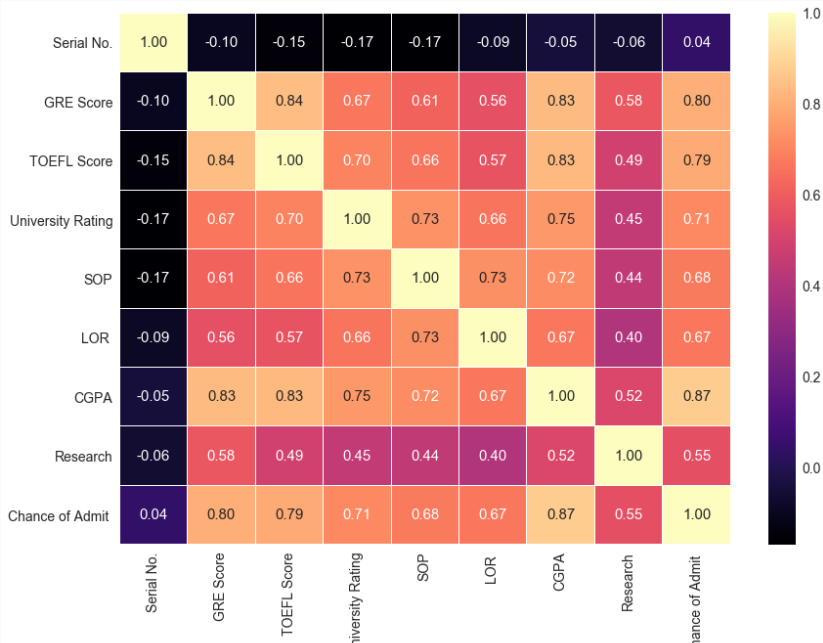
Lets breakdown what we're seeing on the previous slide:

Serial No. is basically uncorrelated with anything.

Admit is highly correlated with **CGPA**, **TOEFL Score** and **GRE Score**

Research has a lowish correlation with **Admit**, but also with everything else.

Correlation



T-score For Application Data

Going back to our fit, we find that **CGPA** has a highly significant coefficient, but **TOEFL** and **GRE** are much more similar to **research**. Why could this be?

Variable	coef	std err	t
const	0.3257	0.014	24.024
Serial No.	0.0722	0.012	5.884
GRE Score	0.0979	0.032	3.077
TOEFL Score	0.0994	0.032	3.144
University Rating	0.0433	0.020	2.133
SOP	0.0078	0.023	0.336
LOR	0.0663	0.023	2.860
CGPA	0.3077	0.041	7.418
Research	0.0217	0.008	2.587

T-score For Application Data

What about **SOP**? The correlation with **Admit** is .68 but the t value of the coefficient is utterly insignificant. Why could that be?

Variable	coef	std err	t
const	0.3257	0.014	24.024
Serial No.	0.0722	0.012	5.884
GRE Score	0.0979	0.032	3.077
TOEFL Score	0.0994	0.032	3.144
University Rating	0.0433	0.020	2.133
SOP	0.0078	0.023	0.336
LOR	0.0663	0.023	2.860
CGPA	0.3077	0.041	7.418
Research	0.0217	0.008	2.587

T-score For Application Data

The answer is probably internal correlation. The model gets the best fit by using **CGPA** to account for general trends and the other variables, highly correlated with it, to fine tune the fit.

Variable	coef	std err	t
const	0.3257	0.014	24.024
Serial No.	0.0722	0.012	5.884
GRE Score	0.0979	0.032	3.077
TOEFL Score	0.0994	0.032	3.144
University Rating	0.0433	0.020	2.133
SOP	0.0078	0.023	0.336
LOR	0.0663	0.023	2.860
CGPA	0.3077	0.041	7.418
Research	0.0217	0.008	2.587

This could be useful, or it lead to overfitting.

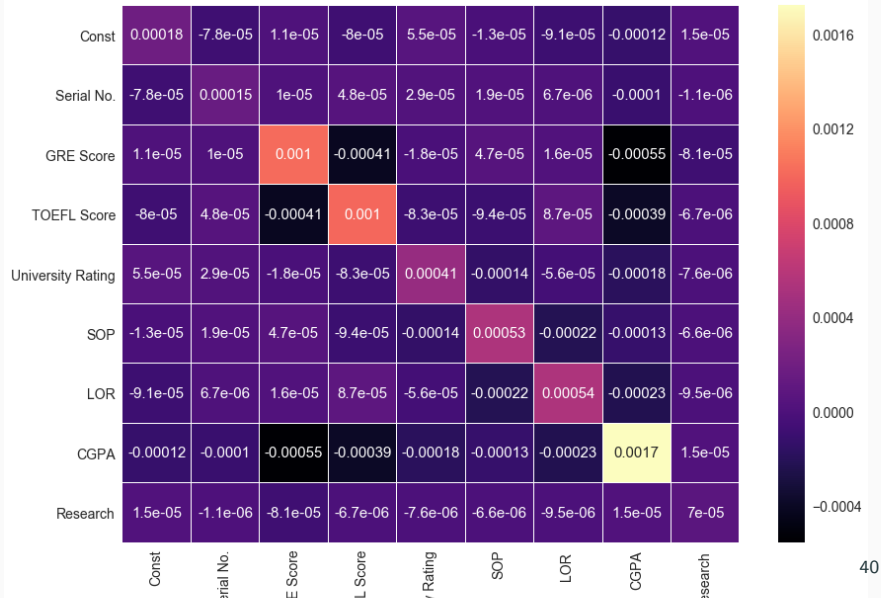
T-score For Application Data

For comparison, lets recall the variance matrix for the parameters

Variable	coef	std err	t
const	0.3257	0.014	24.024
Serial No.	0.0722	0.012	5.884
GRE Score	0.0979	0.032	3.077
TOEFL Score	0.0994	0.032	3.144
University Rating	0.0433	0.020	2.133
SOP	0.0078	0.023	0.336
LOR	0.0663	0.023	2.860
CGPA	0.3077	0.041	7.418
Research	0.0217	0.008	2.587

Variance for Admission

Variance Matrix $\text{Var}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}^T \mathbf{X})^{-1}$



These values are probably low due to the relatively large sample size and the small number of parameters. Note that we see that **CGPA** has a negative covariance with **GRE** and **TOEFL**.

This is what we would expect, since all three are highly correlated we would expect a lessening of fit by **CGPA** could be compensated for by a tightening of the fit on **GRE** or **TOEFL**.

Overfitting

One of the main challenges of predictive statistics is figuring out way to identify and reduce model variance.

A probability approach will use the F -statistic to compare significance of dropping groupings of variables from the model.

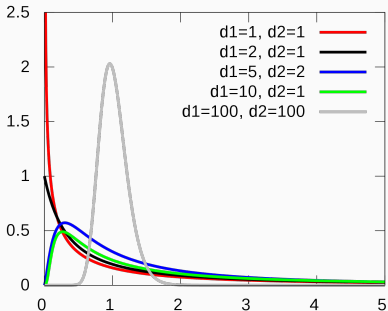
A more computational approach will iterate over all possible subsets of features, pulling out the smallest number of features that give a consistent error.

A more machine learning approach asks us to modify the loss function, attempting to punish the algorithm for incorrect coefficients.

We will explore all of these options in the lectures and labs, starting with a review of F -statistics.

Dropping Variables by Statistical Significance

F-statistic



The F_{d_1, d_2} distribution is used to compare the ratio of sample variances (normalized by the true variances) for normally distributed random variables, where d_1 and d_2 are the respective number of degrees of freedom in each sample. Formally,

$$F \sim \frac{s_1^2}{\sigma_1^2} / \frac{s_2^2}{\sigma_2^2}.$$

F-statistic

Consider two models, models 0 and 1, with model 0 fitting a subset of the parameters of model 1. Assume model 0 has p_0 parameters and model 1 $p_1 > p_0$ parameters.

The F -statistic measure the relative sum squared of residuals between the two models, which in turn allows us to use the F distribution to test the level of significance.

Computing RSS on all the test data, the F -statistic is defined as

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)} = .$$

Under the Gaussian assumption for ϵ and the null hypothesis that the smaller model (model 0) is correct, the F -statistic will have a

$$F \sim F_{p_1 - p_0, N - p_1 - 1}$$

distribution. We can then use the F distribution to assess the significance of removing a variable.

Exercise: Show that that z-scores z_j are the F statistics of dropping the single variable X_j from the model.

Under the Gaussian assumption for ϵ and the null hypothesis that the smaller model (model 0) is correct, the F -statistic will have a

$$F \sim F_{p_1 - p_0, N - p_1 - 1}$$

distribution. We can then use the F distribution to assess the significance of removing a variable

Subset Selection Methods

There are a few reasons why we might not be satisfied with the least squares estimate.

Prediction accuracy: Least squares estimates often have low bias at the cost of high variance. One is often able to reduce variance by selectively setting some coefficients to 0.

Interpretive power: With a large number of related parameters, one often wants to find a small number of predictors with strong, definite effects. You don't have to do quantum mechanics to build a car, and often when looking for the “big picture” we are willing to sacrifice some granularity.

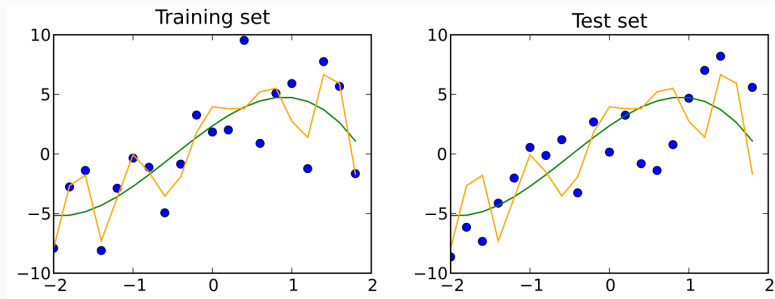
Best Subset Selection

In best subset selection, we perform a regression on every subset of variables of size k and pick the model with the smallest mean squared error. There are efficient algorithms (for example the *leaps and bounds* procedure) that make this possible for p up to 30 or 40, but the total number of subsets grows exponentially

$$\text{Pow}(\{1, \dots, k\}) = 2^k.$$

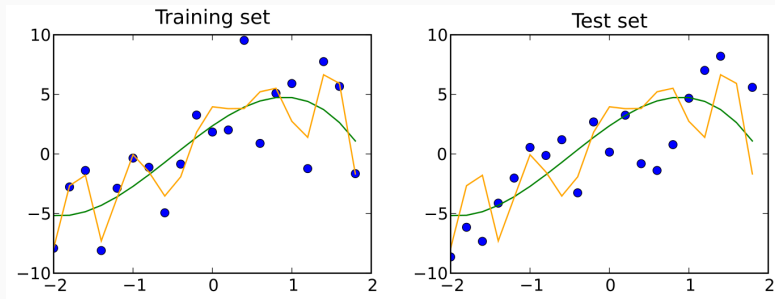
The question becomes how to choose k . There are many criteria, typically one uses the smallest model that minimizes the estimates of the expected prediction error.

Best Subset Selection



A common way to evaluate models is to randomly cut the available data into a training set and a test set. The exact percentages of data to use for training and test depend a bit on the problem, but an 80%-20% split between training and test data is common. Here, we see how low variance data fits better to the test set.

Best Subset Selection



In best subset selection, we train each of our models on a random 80% of the training data and then validate on the remaining 20%. If we have the computational power, we may do this for many random partitions of the data into training and test sets.

Best Subset Selection

For the Student Admissions data, we can perform best subset selection for subsets of all size k .

For $k = 1$, we are fitting a linear model to only one variable and picking out the best test regression.

Variable	MSE train	MSE test
CGPA	0.004629	0.005665
TOEFL Score	0.007271	0.008874
GRE Score	0.006701	0.009308
LOR	0.011041	0.011981
University Rating	0.009177	0.013495
SOP	0.010336	0.013964
Research	0.013240	0.017531
Serial No.	0.018787	0.026214

Best Subset Selection

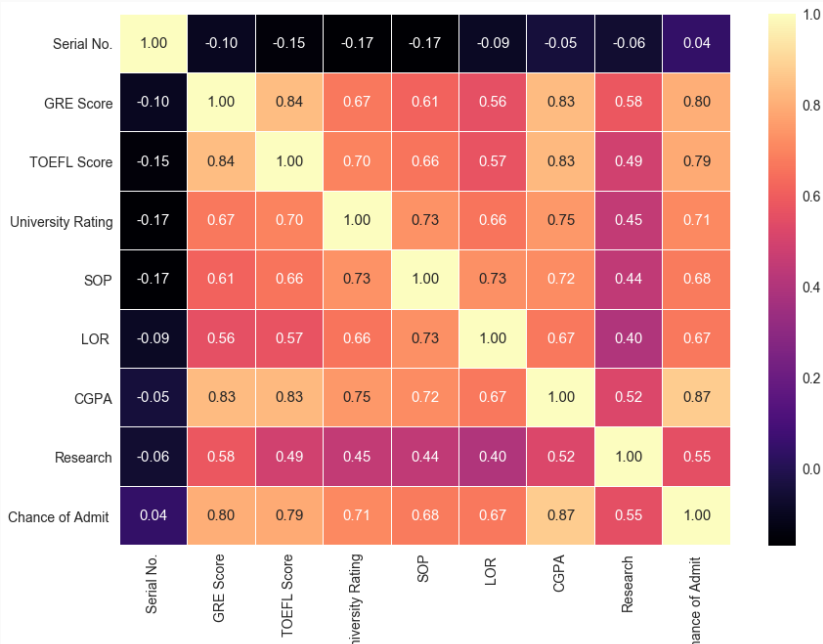
Another common human readable scoring mechanism is the r^2 statistic

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

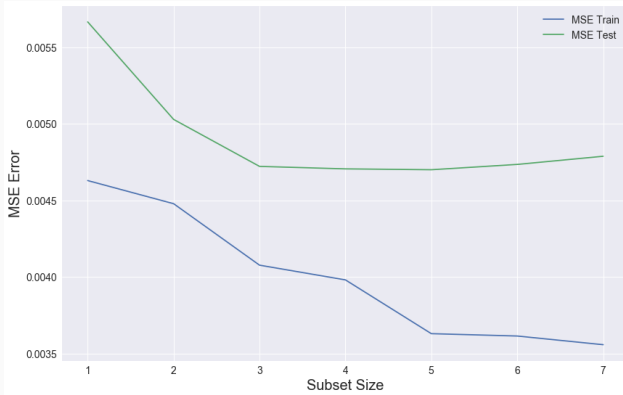
where $r^2 = 1$ is perfect correlation.

Variable	r^2 Train	r^2 Test
CGPA	0.780614	0.755038
GRE Score	0.639539	0.645389
TOEFL Score	0.656348	0.615228
University Rating	0.477414	0.514386
SOP	0.459257	0.453055
LOR	0.536031	0.415755
Research	0.321136	0.299415
Serial No.	-0.015145	0.005851

Correlation

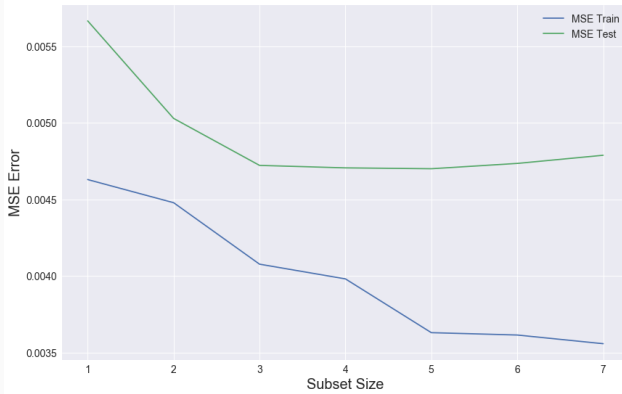


Best Subset Selection



For $k = 1$ the best subset regression matches up with the correlation matrix. As we increase k , the training error continues to drop, but the test error levels out and even starts to rise a bit for $k > 5$.

Best Subset Selection



It turns out the best subsets use **GRE**, **LOR** and **CGPA**. This may be a little surprising given that coefficient of **LOR** scored rather middling significance in the full model. Why could this be?

Forward Subset Selection

For enough features, the 2^p possible subsets of variables becomes computationally expensive. **Forward subset selection** tries to forge a good path through the forest of parameter choices.

The algorithm starts by fitting the intercept, and then sequentially adds into the model the variable that most improves the fit. The algorithm produces a nested sequence of models and is classified as a **greedy algorithm** since it makes local optimization choices at each steps instead of global ones.

Forward subset selection has a number of regression steps polynomial in the number of features, and can be made even more efficient with a bit of linear algebra.

Backward Subset Selection

Backward subset regression starts with the full model and sequentially deletes the predictor with the smallest z-score. It can only be used when the data set is larger than the number of features, ie when $N > p$.

Backward subset selection is particularly useful when you expect to have a large number of features contributing to the final fit.

Forward and Backward Subset Selection

There are two main advantages to forward and backward subset selection:

Computational: Forward/backward subset selection are more computationally efficient than performing an exhaustive search of all subsets of variables.

Stability: If we allow our model to select the best subset of the features, we must pay the price of a larger variance. As always, we may be able to afford this, but for small training sets we may be more interested in stability.

We will be implementing these methods over the next (few) labs.

This lecture was adapted primarily from Elements of Statistical Learning, Chapter 3.

For a good explanation of the proof that $\hat{\sigma}$ is an unbiased estimator, see

[https://stats.stackexchange.com/questions/20227/
why-is-rss-distributed-chi-square-times-n-p](https://stats.stackexchange.com/questions/20227/why-is-rss-distributed-chi-square-times-n-p)