Freie Universität Berlin

Data Science
Lecture 04

16.05.2014

Dr. Kashif Rasul
kashif @krasul #167

Shoaib Burq
sabman @sabman #167

Welcome to Data Science Lecture 4.

Last Time: Bias-Variance

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

The first fundamental concept of statistical learning is to get an intuition on the U shaped curve of the test MSE from the last lecture. This has to do with the fact that the MSE can be decomposed into three fundamental quantities:

Here the expected test MSE refers to the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets and and tested each at x_0. What this tells us is that in order to reduce the test error, we need a method with low variance and bias squared, both of which are positive.

Intuitively, variance is the amount $\hat{f}$ would change if we estimated it using different training data. So more flexible methods have higher variance.

The bias refers to the error introduced by approximating real-life problems, which are complicated, by a simple model. E.g. linear regression assumes a linear relationship between Y and X. Real-life is not like that. So more flexible models have a lower bias.

R^2

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

The R^2 provides the proportion of variance explained and takes a value from 0 to 1 and is independent of the scale of Y.

Here TSS is the total sum of squares (y_0-mean(Y)) + (y_1 - mean(Y)) + …TSS can be though of as the amount of variability inherent in the response before regression was performed. RSS is the measure of variability that is left unexplained after regression. So TSS-RSS is the amount of variability in the response that is explained (or removed) by regression.

A R^2 close to 1 indicates a large proportion of variability is explained in the response. A value near 0 means that the regression did not explain much of the variability in the response, which might happen if the linear model is wrong, or the inherent error is too high or both!

Multiple Linear Reg.

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:	Thu, 08 May 2014	Prob (F-statistic):	1.58e-96
Time:	13:56:39	Log-Likelihood:	-386.18
No. Observations:	200	AIC:	780.4
Df Residuals:	196	BIC:	793.6
Df Model:	3		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
Radio	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	454.

Instead of having multiple linear models for each feature, we could have a single model with multiple predictors. We can interpret the coefficients as the average effect on Y for a unit increase in the particular feature while holding all other predictors fixed. For e.g. for $\text{sales}=\beta_0+\beta_1\times\text{TV}+\beta_2\times\text{radio}+\beta_3\times\text{newspaper}+\varepsilon$.

Geometrically we are fitting a hyper-plane with the specific slopes given by the least square regression. There is an issue about correlated predictors here. What happens if two of the predictors are highly correlated? Then the variance of the coefficients increases. Interpretation also suffers, since we can't claim to keep one predictor constant since they will both move together.

So claims of causality should be avoided for observational data, since in reality predictors change together.

Here we see by a t statistic of -0.177 that is does not have an effect, or the p-value is close to 1 implying the null hypothesis. But we have to be careful how to interpret this. On it's own newspaper could be a good predictor, but in the presence of TV and radio, newspaper is not showing significance.

```
import scipy.stats as stats

corr = {}
corr['pearson'], _ =
stats.pearsonr(df.Radio,df.Newspaper)
corr['spearman'], _ =
stats.spearmanr(df.Radio,df.Newspaper)
corr['kendall'], _ =
stats.kendalltau(df.Radio,df.Newspaper)

print(corr)
```

```
{'kendall': 0.20707706351468924,
'spearman': 0.31697948906632362,
'pearson': 0.35410375076117517}
```


An in fact if we look at the correlations of the Radio and Newspaper, we see that they are correlated. So any effect of newspaper has been soaked up the Radio since they are correlated. So with Radio in the model newspaper is no longer needed. It doesn't tell us anything more, or improve the prediction, given we have measured the Radio data correctly.

Important Questions

- Is at least one of the predictors useful in predicting the response?
- Do all the predictors help to explain Y or only a subset?
- How well does the model fit the data?
- Given predictor values, what response values should we predict and how accurate is this?

The first three we can answer by looking at the model. The last question we will get to later, but we can use confidence intervals to quantify the uncertainty around the predictor.

Relationship

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)}$$

First question: Is There a Relationship Between the Response and Predictors? In the multiple regression case with p features, we need to ask if all the coefficients are zero. The alternative hypothesis is if at least one coefficient is non-zero. We perform this test via the F-statistic, where remember TSS is the total sum of squares measuring the amount of variability inherent in the model and RSS the variability unexplained.

So this ratio is the drop in training error divided by the features over the RSS divided by the number of parameter we fit, minus p and the 1 for the intercept. For our advertising data the F statistic is huge at 570. So this provides compelling evidence against the null hypothesis.

What if this ratio was closer to 1? Well the answer depends on n and p. When n is large an F value near 1 might still provide evidence against the H_0. For smaller n, a larger F is required.

These concepts actually break down though when p > n, and we will discuss this situation later.

Important Variables

- All (best) subset regression: compute with all possible subsets and choose between them
- Forward selection: fit null, then add the variable that results in lowest RSS and repeat…
- Backward selection: start with all, remove the one with largest p-value, and repeat…
- Mixed selection: combination of forward and back

Secondly, when we fit linear regression one of the important things we can do is to decide on which of the variables are important.

Best subset regression, the criterion that balances training error with model size. However we cannot examine all possible model, since there are 2^p of them! We need an automated approach that searches through them.

One way is forward selection. We begin with the null model with just the intercept. We then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS. Having done that, we add to that model the variable that result in the lowest RSS for the new two-variable model. We continue until for example all remaining variables have a p-value above some threshold.

Backward selection does this by removing.

Mixed selection combines the 2. Keep adding but drop at any point the p-value for one of the model rises above threshold. Continue until all variables have sufficiently low p-value.

Note backward cannot be used if $p > n$. Forward can include variables which might become redundant. Mixed fixes this. Later we will discuss more

Model Fit

- RSE
- R^2: fraction of variance explained
- Plot the data: reveal problems with the model

As we saw the two most common numerical measures of model fit are the RSE and the R or the fraction of variance explained.

In the presence of correlated features, there can be a synergy or interaction effect between the features, whereby combining them results in a bigger boost to the prediction than using any single one. We will see how to accommodate this synergistic effect through the use of interaction terms.

Qualitative Predictors

- Some predictors can be qualitative taking discrete sets of values
- Categorical predictors or factor variables
- E.g. gender, student status, marital status, ethnicity

Often some predictors are qualitative. For example looking at Credit data we have a number of qualitative features.

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

With only two levels we can create an indicator or dummy variable that takes 2 possible numerical values. We could code the females as 1 and males as -1. In terms of the regression model there would be no difference, but it does alter the interpretation of the coefficients. Then the intercept would be the overall average credit card balance, ignoring the gender, and β_1 how much say the females are above it.

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

When we have more than two levels, well we take a vaselike variable and add dummy variables for each of the other cases. So x_{i1} is 1 if the i^{th} person is Asian and 0 if not and x_{i2} is 1 if the i^{th} person is Caucasian and the baseline model is that the i^{th} person is African American since it does not have a parameter representing it.

The baseline determines what comparison we make in the model and not the RSE fit.

Interactions

- Assumed: increase of one unit of a feature is independent of the other features
- But suppose spending money on radio increases effectiveness of TV advertising
- In marketing this is known as a synergy effect or interaction

Let us now discuss extensions to the Linear model and talk first about interactions. The linear model made the assumption in the Advertising data that the average unit increase on say TV is always β_1 regardless on the amount spent on radio.

In this situation given a fixed budget spending half on radio and half on TV may increase sales more that if we allocated all to TV or radio.

So in our model the slope of radio should also increase as radio increases. So how do we deal with interactions?

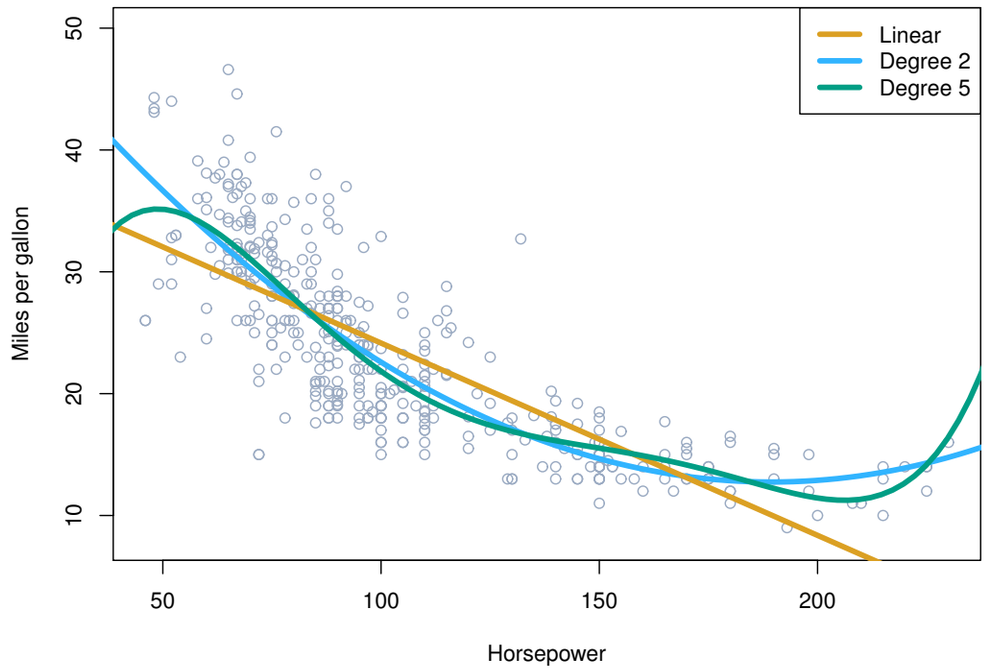
$$
\begin{aligned}
\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\
&= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon.
\end{aligned}
$$

Here we have a model where we have a term with the multiple of radio and tv together, which we can re-write slightly which makes it easier to interpret. And the β_3 is the interaction term, and if we look at it's p-value we will find it is significant. The R^2 for the model with interaction is 96.8% higher than the 89.7% without the interaction term.

Sometimes an interaction term has a very small p-value but the associated main effects do not. But if we include an interaction in a model, we should also include the main effects, even if the p-value are large. This is the hierarchy principle. We do this because without the main effects it's hard to interpret the model and so it is not a good idea.

Interactions between qualitative and quantitive dummy variables is easy to interpret too.

Non-linear effects



The other modification to the linear model is what to do if we have non-linear effects? Well for the Auto data if we plot the linear regression of Horsepower to predict milage, we see it does not fit quite well. We can fit a horsepower^2 term as well and another model with horsepower^5.

The polynomial variables are easy to do just like the dummy variable for categorical data, we can create a horsepower^2 and horsepower^5 feature and use it. It is still a linear model, since it's linear in the coefficients.

Potential Problems

- Non-linearity of the X-Y relationships
- Correlation of the error terms
- Non-constant variance of error terms
- Outliers
- High-leverage points
- Collinearity

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

An important assumption of the linear regression model is that the error terms are uncorrelated, e.g. the fact that $\varepsilon_i > 0$ provides no information about the sign of ε_{i+1}. If that is not the case, the confidence intervals have to be narrower, and the p-values lower, causing us to conclude a parameter is significant when it is not. Intuitively if means if the error terms are correlated, we may have an unwarranted sense of confidence in our model. In time-series data the errors are correlated.

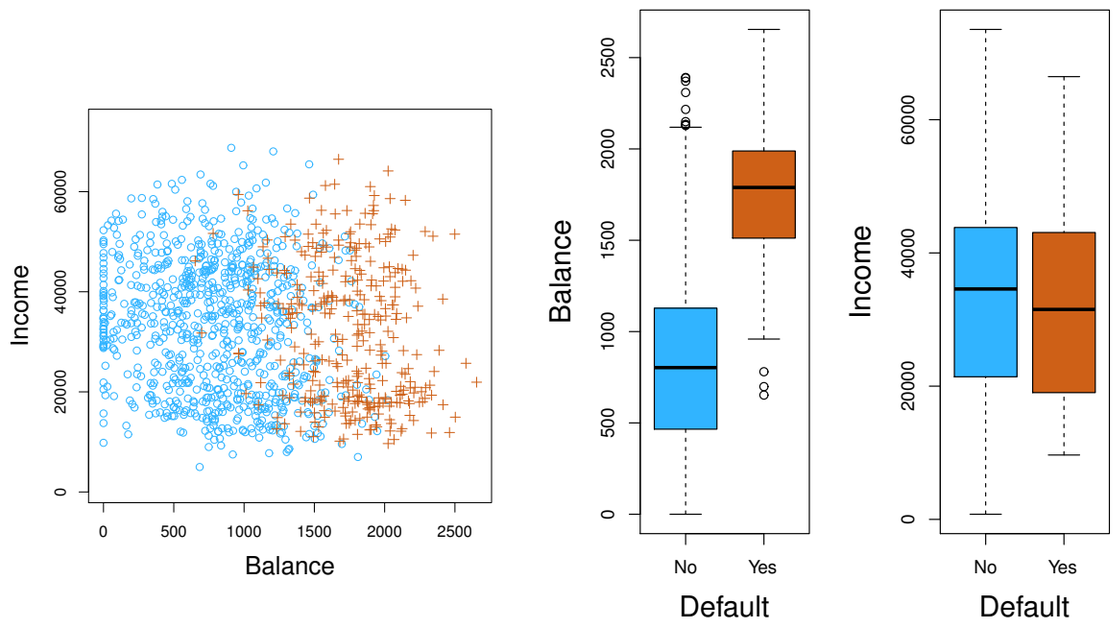
If the variance of the error is not constant something called heteroscedasticity, we might have to transform the Y to compensate this.

Outliers can be perhaps removed by plotting the residuals or studentized residuals.

High leverage points are responses with high predictors. We can use what's called a leverage statistic to quantify a predictor's leverage.

Finally co-linearity, results in a lower t-statistic.

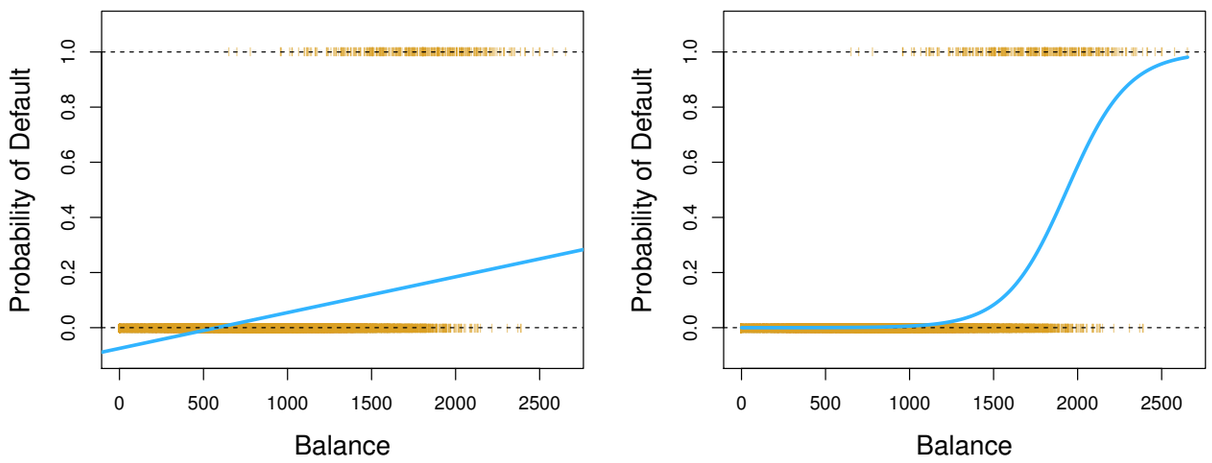
Classification



In many cases the response is qualitative. E.g. eye colour takes brown, green, blue etc. The process of predicting qualitative responses is called classification. E.g. give a transaction determine if it is fraudulent or not? So Y takes values from a discrete set.

So in the Default data set, we see the annual incomes and monthly credit card balances of a number of individuals on the left. The individuals who defaulted on their credit card payments are shown in orange, and those who did not in blue. On the right we have box plot of the balance and income as a function of default status.

We see a very pronounced relationship between the predictor and balance and status response. In most situations this not the case. Again in the black lines in the box plot is the median.



Pr(default=Yes | balance)

Why not use regression with a dummy response variable? Well for one this would imply an ordering of the response variable and the gap between the coding would probably not be natural. There is no natural way to convert a qualitative response variable with more than 2 levels into a quantitative one for linear regression.

Consider again the Default data set, where the response default falls into one of two categories, Yes or No. Here, linear regressions works but still has issues.

We are estimating probability of default given the balance, and with linear regression we end up with negative values! We need a model where the values lie between 0 and 1 which brings us to logistic regression.

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

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

$$\log \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X$$

In logistic regression we use the logistic function, the first equation, which always produces an S-shaped curve, between 0 and 1 regardless of the value of X.

With some manipulation we end up with the 2nd equation, called the odds. Taking the log of the odds or logit we arrive at the last equation which gives the logistic regression it's name. And this is interpretable too. Increasing X by one unit changes the log odds by β_1.

The way we estimate the coefficients now is via a general method called maximum likelihood. Intuitively we seek estimates of the coefficients such that the predicted probability $\hat{p}(x_i)$ of default for each i corresponds as closely as possible to the observed default status. We can formalise this in as an equation called the likelihood function and it's maximum are the coefficients we want.

In fact we can view linear regression as well in such a probabilistic setting and if we assume that the error is normally distributed, then maximising the likelihood function ends up implying that we minimise the least square error.

```python
from sklearn import linear_model

clf = linear_model.LinearRegression()
clf.fit(X_train, y_train)
clf.predict(X_test)
```


Let's talk a bit about Scikit learn: All objects within scikit-learn share a uniform common basic API consisting of three complementary interfaces: an estimator interface for building and fitting models, a predictor interface for making predictions and a transformer interface for converting data.

The predictor interface extends the notion of an estimator by adding a predict method that takes an array X_test and produces predictions for X_test, based on the learned parameters of the estimator. In the case of supervised learning estimators, this method typically returns the predicted labels or values computed by the model.

Most machine learning algorithms implemented in scikit-learn expect data to be stored in a two-dimensional array or matrix. The arrays can be either numpy arrays, or in some cases scipy.sparse matrices. The size of the array is expected to be [n_samples, n_features].

Check here: https://github.com/jakevdp/sklearn_pycon2014 for a good tutorial on Scikit learn, but worry not, we will do a scikit learn lab2 now!

http://bit.ly/data-science-4

Thank you that’s it for this lecture. Your feedback is greatly appreciated as it will help us to improve and drive the course in the direction you would like it to go.