Linear regression
Libraries
Using Statsmodels for teaching purposes since it has some nice characteristics for linear modeling. However, we recommend that you use scikit-learn

	Continuous	categorical
Supervised		
Unsupervised		

Questions about the advertising data
lets pretend you work for the company that manufactures and markets this widget. The company might ask you the following: On the basis of this data, how should we spend our advertising money in the future?
This general question might lead you to more specific questions:

1. Is there a relationship between ads and sales?
   Yes especially on tv and radio
2. How strong is that relationship?
   No as strong in newspaper a lot of noise, tv has nelson ratings
3. Which ad types contribute to sales?

4. What is the effect of each ad type of sales?
5. Given ad spending in a particular market, can sales be predicted?

   We want our model to reflect reality as much as possible
   Sum Squared residuals is also called the cost function the larger it is the larger the penalty we want to minimize the function as much as possible.

When line is flat the variance is just noise
Want to find b1 and b0

Black dots are the observed values of x and y
The blue line is our least squares line
The red lines are the residuals, which are the distances between the observed values and the least squares line.

How do the model coefficients relate to the leat squares line?

B0 is the intercept (the value of y when x=0)
B1 is the slope (the change in y divided by change in x)

Lets use Statsmodels to estimate the model coefficients for the advertising data:

Intercept = 7.032594
TV = 0.047537

Interpreting model coefficients
How do we interpret the TV coefficient (b1)?

A “unit” increase in TV as spending is associated with a 0.047437 “unit” increase in Sales.
Or more clearly: An additionally $1,000 spent on TV ads is associated with an increase in sales of 47.536 widgets.

Note that if an increase in TV ad spending was associated with a decrease in sales, b1 would be negative.

Using the model for prediction
Let’s say that there was a new market where the TV advertising spends was $50,000. What would we predict for the Sales in that market?

Y hat is our model
Y = b0 + b1x
Y = 7.032594 + 0.047537 X 50

Manually calculate the prediction
7.032594 + 0.047537 * 50
= 9.409444

thus, we would predict Sales of 9,409 widgets in that market
Of course we can also use Statsmodels to make the prediction:

you have to create a DataFrame since the Statsmodels formula interface expects it

Plotting the Least Squares Line
Let’s make predictions for the smallest and largest observed values of x, and then use the predicted values to plot the least squares line:

You can see the line is consistent with our data

Confidence in our model
Question: Is linear regression a jigh bias/low variance model, or a low variance.high bias model?

Answer: High bias/low variance. Under repeated sampling, the line will stay roughly in the same place (low variance), but the average of those models won’t do a great

job capturing the true relationship (high bias). Note that low variance is a useful characteristic when you don't have a lot of training data!
A closely related concept is confidence intervals. Statsmodels calculates 95% confidence intervals for our model coefficients, which are interpreted as follows: If the population from which this sample was drawn was sampled 100 times approximately 95% of those would contain the "true" coefficient.

Keep in mind that we only have a single sample of data, and not the entire population of data. The "true" coefficient is either within this interval or it isn't, but there's no way to actually know. We estimate the coefficient with the data we do have, and we show uncertainty about that estimate by giving a range that the coefficient is probably within.

Note that using 95% confidence intervals is just a convention. You can create 90% confidence intervals (which will be more narrow) 99% confidence intervals (which will be wider), or whatever intervals you like.

Hypothesis testing and p-values

Closely related to CI, is hypothesis testing. Generally speaking you start with a null hypothesis and an alternative hypothesis (that is oppsite the null). Then, you check whether the data supports rejecting the null hypothesis or failing to reject the null hypothesis.

Note that "failing to reject" the null is not the same as "accepting" the null hypothesis. The alternative hypothesis may indeed by true, except that you just don't have enough data to show that.

As it relates to model coefficients, here is the conventional hypothesis test:

Null hypothesis: There is no relationship between TV ads and Sales (and thus b1 equals zero)
Alternative hypothesis: There is a relationship between TV ads and Sales (and thus b1 is not equal to zero)

How do we test this hypothesis? Intuitively, we reject the null (and thus believe the alternative) if the 95% confidence interval does not include zero. Conversely, the p-value represents the probability that the coefficient is actually zero.

If the 95% confidence interval includes zero, the p-value for that coefficient will be greater that 0.05. If the 95% confidence interval does not include zero, the p-value will be less than 0.05 is one way to decide whether there is likely a relationship between the feature and the response. (Again, using 0.05 as the cutoff is just a convention.)

In this case, the p-value for TV is far less than 0.05, and so we believe that there is a relationship between TV ads and Sales.

Note that we generally ignore the p-value for the intercept.

How well does the model fit the data?
The most common way to evaluate the overall fit of a linear model is by the R-squared value. R-squared is the proportion of variance explained, meaning the proportion of variance in the observed data that is explained by the model, or the reduction in error over the null model. (The null model jst predicts the mean of the obderved response, and thus has an intercept and no slope.)

R-squared is between 0 and 1, and higher is better because it means that more variance is expanded by the model. Here's an example of what R-squared "looks like".

You can see that the blue line explains some of the variance in the data (R-squared =.54), the green line explains more of the variance (R-squared=.64), and the red line fits the training data even further (R-squared=0.66). (Does the red line look like its overfitting?)

Lets calculate the R-squared value for our simple linear model.

Multiple linear regressions
Simple linear regression can easily be extended to include multiple features. This is called multiple linear regressions

Each x represents a different feature, and each feature has its own coefficient. In this case:
Y = b0 + b1 X TV + b2 X Radio + b3 X Newspaper

Lets use Stats models to estimate these coeff

How do we interpret these coefficients? For a given amount of Radio and Newspaper ad spending, an increase of $1000 in TV ad spending is associated with an increase in Sales of 45.765 widgets.

A lot of the information we have been reviewing piece-by-piece is available in the model summary output:

What are a few key things we learn from this output?
TV and Radio have significant p-values, whereas Newspaper does not. Thus we reject the null hypothesis for TV and Radio (that there is no association between those features and Sales), and fail to reject the null hypothesis for Newspaper.

TV and Radio ad spending are both positively associated with Sales, whereas Newspaper ad spending is slightly negatively associated with Sales. (However, this is irrelevant since we have failed to reject the null hypothesis for Newspaper.)
This model has a higher R-squared (0.897) than the previous model, which means that this model provides a better fit to the data than a model that only includes TV.

Feature Selection
How do I decide which features to include in a linear model? Here's an idea:
Try different models, and only keep the predictors in the model if they have small p-values.
Check whether the R-squared values goes up when you add new predictors

What are the drawbacks to this approach?
Linear models rely upon a lot of assumptions (such as the features being independent), and if those assumptions are violates (which they usually are), R-squared and p-values are less reliable.
Using a p-value cutoff of 0.05 means that if you add 100 predictors to a model that are pure noise, 5 of them (on average) will still be counted as significant.
R-squared is susceptible to over fitting, and thus there is no guarantee that a model with a high R-squared value will generalize.