Linear Regression: Introduction to summary statistics

Output from statsmodels

OLS Regression Results

Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Budget	0.7846	0.133	5.901	0.000	0.520 1.049
Ones	4.44e+07	1.27e+07	3.504	0.001	1.92e+07 6.96e+07

Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

Target or dependent variable

OLS Regression Results

Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Budget	0.7846	0.133	5.901	0.000	0.520 1.049
Ones	4.44e+07	1.27e+07	3.504	0.001	1.92e+07 6.96e+07

Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

Model type: Ordinary Least Squares = linear regression

OLS Regression Results

Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Budget	0.7846	0.133	5.901	0.000	0.520 1.049
Ones	4.44e+07	1.27e+07	3.504	0.001	1.92e+07 6.96e+07

Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

OLS Regression Results

Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.
Df Model:	1		

m

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Budget	0.7846	0.133	5.901	0.000	0.520 1.049
Ones	4.44e+07	1.27e+07	3.504	0.001	1.92e+07 6.96e+07

Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

OLS Regression Results

Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.
Df Model:	1		

Residual degrees of freedom = Number of observations - number of parameters (including intercept)

OLS Regression Results

Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.
Df Model:	1		

Model degrees of freedom = Number of parameters – 1

(or # of features not including intercept)

OLS Regression Results

Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Budget	0.7846	0.133	5.901	0.000	0.520 1.049
Ones	4.44e+07	1.27e+07	3.504	0.001	1.92e+07 6.96e+07

Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

R^2

Some thoughts about R^2

OLS's most controversial metric

Revisiting the OLS cost function

- The OLS cost function minimizes the sum of squared residuals, also called the **sum of squared errors (SSE)**

$$\sum_{i=1}^{m}\left(y_{\beta}(x^{(i)}) - y_{obs}^{(i)}\right)^2$$

- We can also calculate the variance of the observed (actual) points, also called **the total sum of squares (SST)**

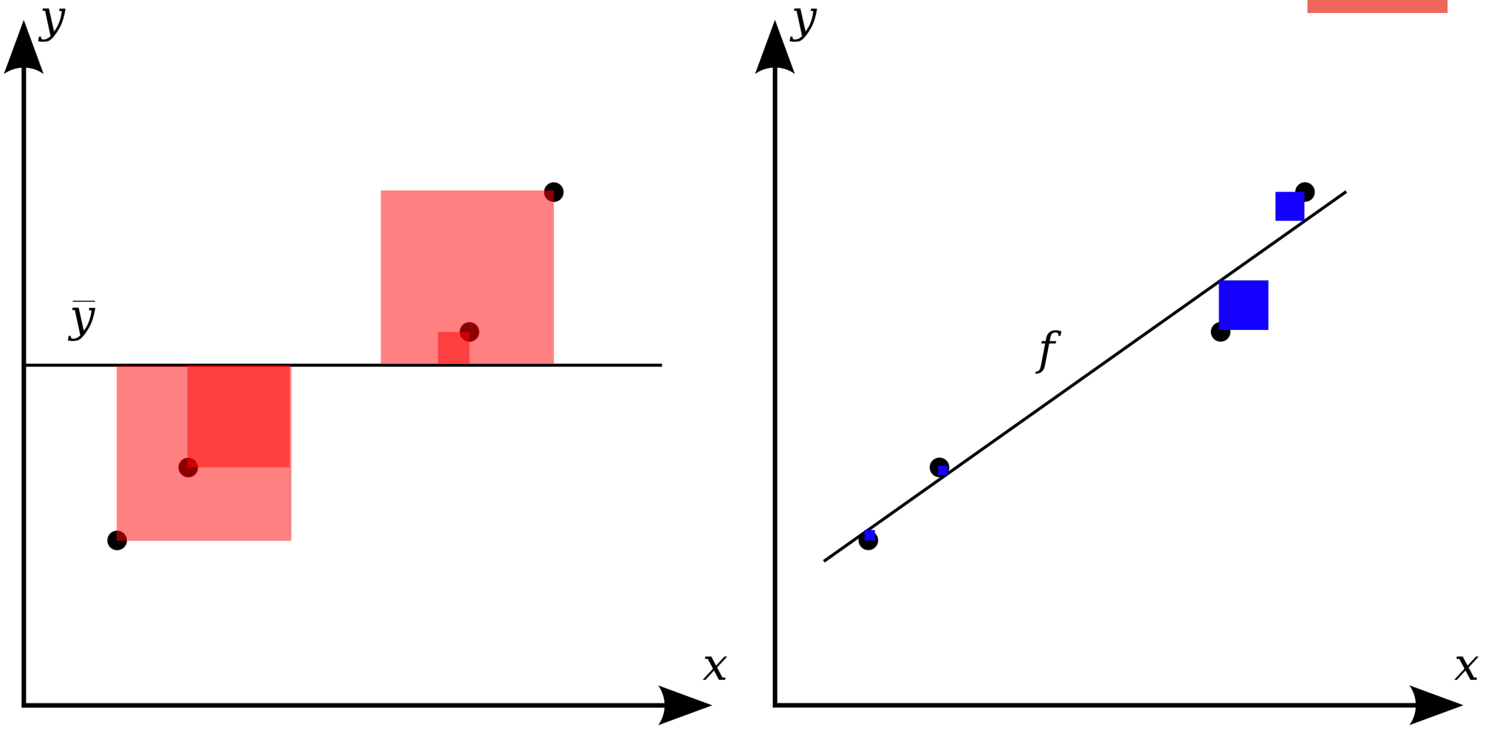
$$\sum_{i=1}^{m}\left(y_{obs} - y_{obs}^{(i)}\right)^2$$

Calculating the R^2 metric

- $$R^2 = 1 - \frac{SSE}{SST}$$

Calculating the R^2 metric

$$R^2 = 1 - \frac{SSE}{SST}$$



Source: Wikipedia page for R^2

"Distraction or nuisance?" – CMU stats prof on R^2

- There are some reasons we teach R^2 and reasons why it's popular
  - Easy to calculate and built-in to most stats packages
  - Does give us insight into how our model is behaving, *given some strong conditions*
    - Know that it doesn't measure "how well the model fits"
    - Use it to compare only models with the same number of features
- But there's been pushback against R^2 in the stats and science community
  - E.g. This CMU stats professor's notes that went viral (as much as such a thing can): http://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/10/lecture-10.pdf

Debunking myths about R^2 (from Prof Shalizi's lecture notes)

- 1) R^2 does not measure goodness of fit
  - It can be arbitrarily close to 1 for a wrong model and arbitrarily low for a correct model

Debunking myths about R^2 (from Prof Shalizi's lecture notes)

- 1) R^2 does not measure goodness of fit
  - It can be arbitrarily close to 1 for a wrong model and arbitrarily low for a correct model
- 2) R^2 is also pretty useless as a measure of predictability
  - R^2 says nothing about prediction error

Debunking myths about R^2 (from Prof Shalizi's lecture notes)

- 1) R^2 does not measure goodness of fit
  - It can be arbitrarily close to 1 for a wrong model and arbitrarily low for a correct model
- 2) R^2 is also pretty useless as a measure of predictability
  - R^2 says nothing about prediction error
- 3) R^2 cannot be compared across data sets

Debunking myths about R^2 (from Prof Shalizi's lecture notes)

- 1) R^2 does not measure goodness of fit
  - It can be arbitrarily close to 1 for a wrong model and arbitrarily low for a correct model
- 2) R^2 is also pretty useless as a measure of predictability
  - R^2 says nothing about prediction error
- 3) R^2 cannot be compared across data sets
- 4) R^2 cannot be compared between a model with untransformed Y and one with transformed Y , or between different transformations of Y

Debunking myths about R^2 (from Prof Shalizi's lecture notes)

- 1) R^2 does not measure goodness of fit
  - It can be arbitrarily close to 1 for a wrong model and arbitrarily low for a correct model
- 2) R^2 is also pretty useless as a measure of predictability
  - R^2 says nothing about prediction error
- 3) R^2 cannot be compared across data sets
- 4) R^2 cannot be compared between a model with untransformed Y and one with transformed Y , or between different transformations of Y
- 5) The one situation where R^2 can be compared is when different models are fit to the same data set with the same, untransformed response variable

Debunking myths about R^2 (from Prof Shalizi's lecture notes)

- 1) R^2 does not measure goodness of fit
  - It can be arbitrarily close to 1 for a wrong model and arbitrarily low for a correct model
- 2) R^2 is also pretty useless as a measure of predictability
  - R^2 says nothing about prediction error
- 3) R^2 cannot be compared across data sets
- 4) R^2 cannot be compared between a model with untransformed Y and one with transformed Y , or between different transformations of Y
- 5) The one situation where R^2 can be compared is when different models are fit to the same data set with the same, untransformed response variable
- 6) It is very common to say that R^2 is "the fraction of variance explained" by the regression
  - But if we regressed X on Y , we'd get exactly the same R^2 . This in itself should be enough to show that a high R^2 says nothing about explaining one variable by another

Adjusted R^2 can overcome *some* of the issues

$$\bar{R}^2 = 1 - \frac{SSE / df_e}{SST / df_t}$$

$df_e \rightarrow m - k - 1$

$df_t \rightarrow m - 1$

m= # points

k = # parameters

Adjusted R^2 can overcome *some* of the issues

$$\bar{R}^2 = 1 - \frac{SSE / df_e}{SST / df_t}$$

$df_e \rightarrow m - k - 1$

$df_t \rightarrow m - 1$

m= # points

k = # parameters

Use them with care…

More metrics

OLS Regression Results

Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.
Df Model:	1		

F-test

Null hypothesis:

This data can be modeled by setting all β values to zero
(the linear relationship we've found is purely due to chance)

OLS Regression Results

Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	$\boxed{\textbf{F-statistic:}}$	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.
Df Model:	1		

F-test

Prob (F-statistic):

The p-value for the whole model (i.e. probability of finding the observed-or more-extreme results when the null hypothesis (H_0) is true). If p-value <0.05, we **can** reject the null hypothesis (data is too extreme to fit this model just by chance). It doesn't guarantee the model is "true"!

OLS Regression Results

Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.
Df Model:	1		

Log likelihood

Likelihood is another cost function!

$$L(\beta_0, \beta_1) = p(y_{obs} | \beta_0, \beta_1)$$

For a given model with specific coefficients (assume the model is right), likelihood is the chance of seeing this real data. The model with maximum likelihood is the best fit!

OLS Regression Results

Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Budget	0.7846	0.133	5.901	0.000	0.520 1.049
Ones	4.44e+07	1.27e+07	3.504	0.001	1.92e+07 6.96e+07

t-test

Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

		coef	std err	t	P>|t|	[95.0% Conf. Int.]
β_1	**Budget**	0.7846	0.133	5.901	0.000	0.520 1.049
β_0	**Ones**	4.44e+07	1.27e+07	3.504	0.001	1.92e+07 6.96e+07

t-test

Null hypothesis:

This specific β value is zero (the data can be created by such a model, with the other β values intact)

P >|t|:

P-value for this test. If p-value < 0.05, we can reject the null hypothesis: This variable does contribute to this model (DOES or DOESN'T, not how much).

Normality test

Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

Null hypothesis:

ε is normally distributed, no skew, no excess kurtosis

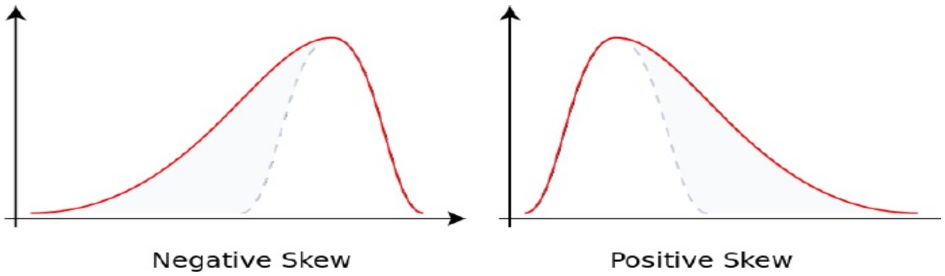
Prob(Omnibus):

The p-value for this test. If p-value < 0.05, we reject the null hypothesis: ε does not exactly follow the normal distribution that we assumed.

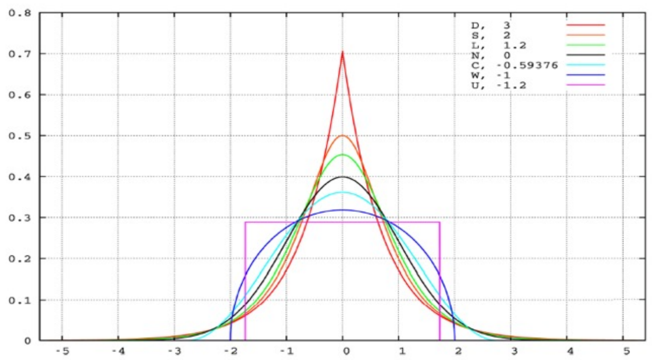
Skew & Kurtosis

Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

Skew (asymmetry)



Kurtosis (peakiness)



Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

Another normality test

Null hypothesis:

Again, ε is normally distributed. We are looking for skewness ~ 0, and kurtosis ~ 3. JB is an alternate to Omnibus and tests if those conditions are close enough to ideal to accept the Null.

Prob(JB):

The p-value for this test.

Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

An autocorrelation test

Null hypothesis: Errors are uncorrelated

The Durbin Watson test reports a test statistic, with a value from 0 to 4, where:

- 2 is no autocorrelation.
- 0 to <2 is positive autocorrelation (common in time series data).
- >2 to 4 is negative autocorrelation (less common in time series data).

Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

Sensitivity of prediction to small errors in input

Condition Number:

Given Mx=b, we can calculate the condition number :

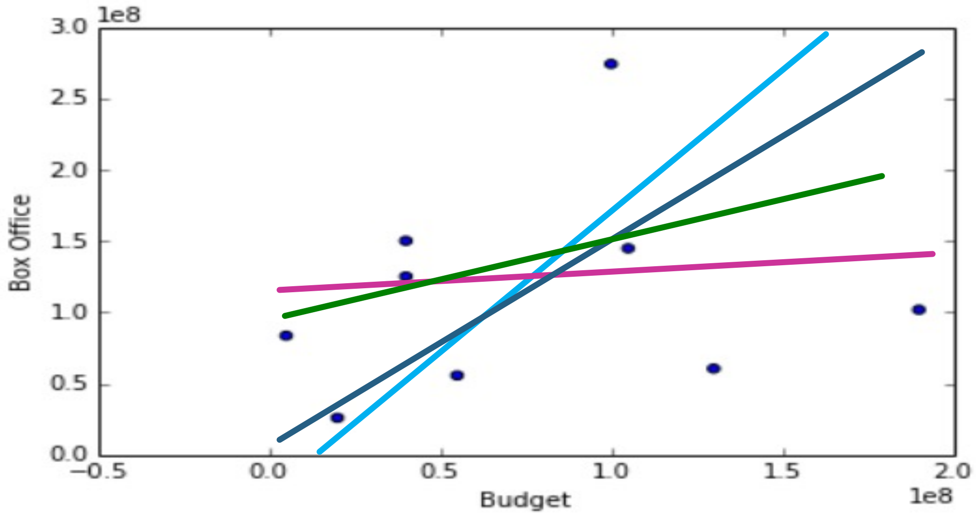
$$CN = \frac{|\lambda max(M)|}{|\lambda min(M)|}$$

Note: if the condition number is large, the data matrix is ill-posed (does not have a unique, well-defined solution). This means the solution is unstable and coefficients can easily change with new data.

Introduction to model selection

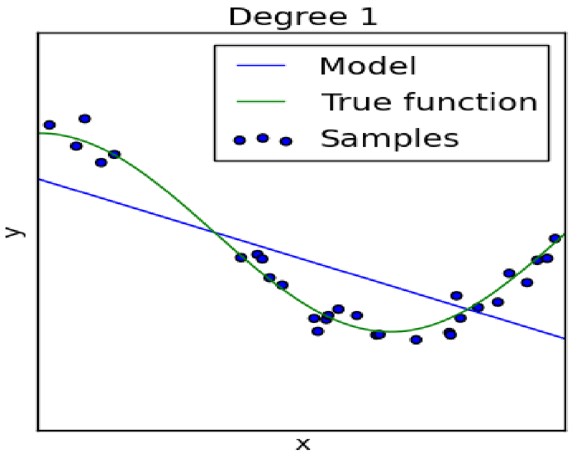
For models with the same number of features, easy:

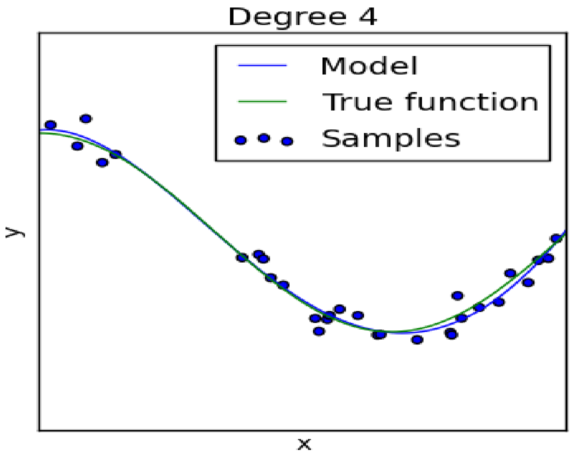
- Pick the model with the best cost function (highest log-likelihood)

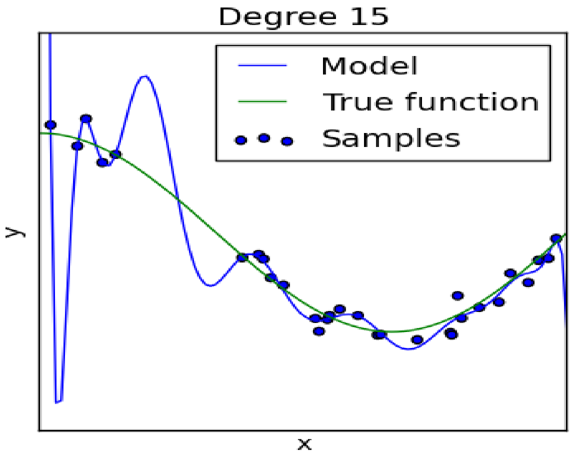


$$y_{\beta}(x) = \beta_0 + \beta_1 x + \varepsilon$$

Need to revisit over/under-fitting and model complexity





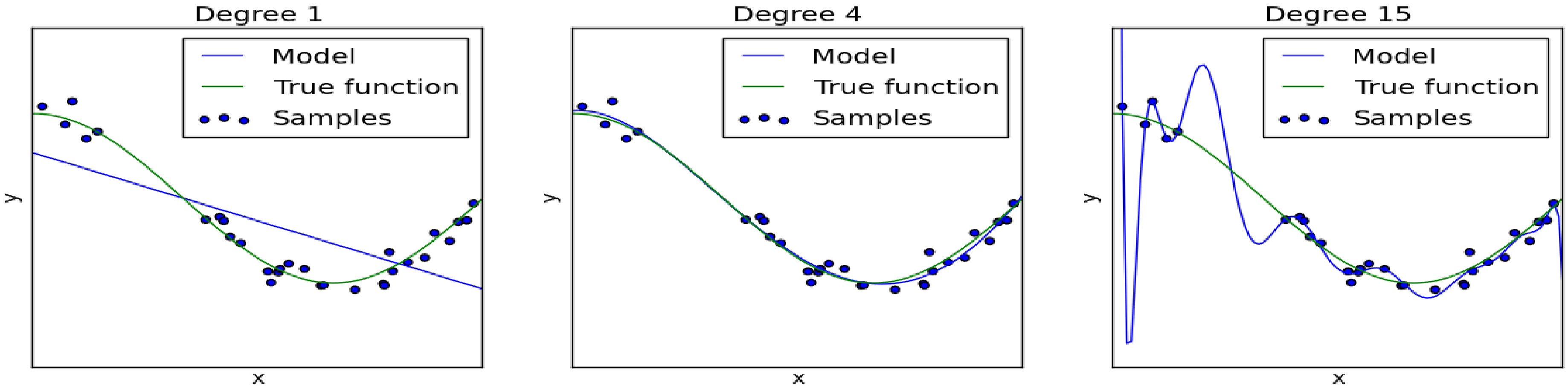


Under-fitting

Just Right

Overfitting

Need to revisit over/under-fitting and model complexity



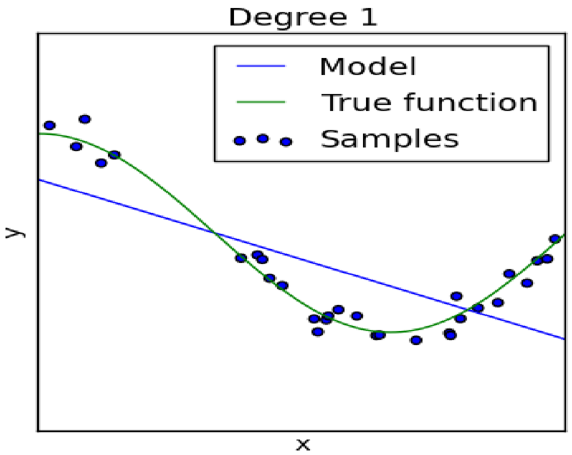
Process: Fit a training set then calculate MSE on your test set ~~(in sklearn: fit > predict > score)~~

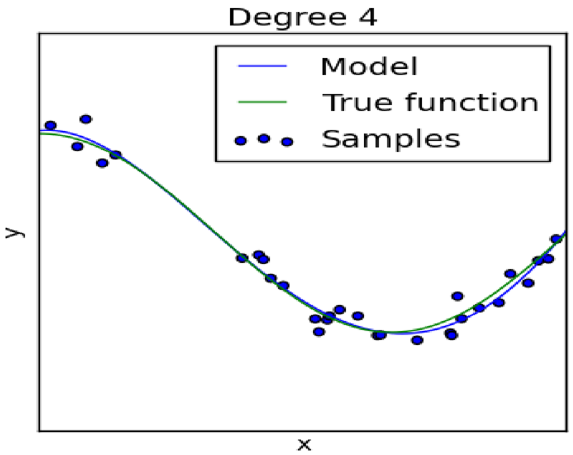
We can use adjusted R2 for models of different complexity

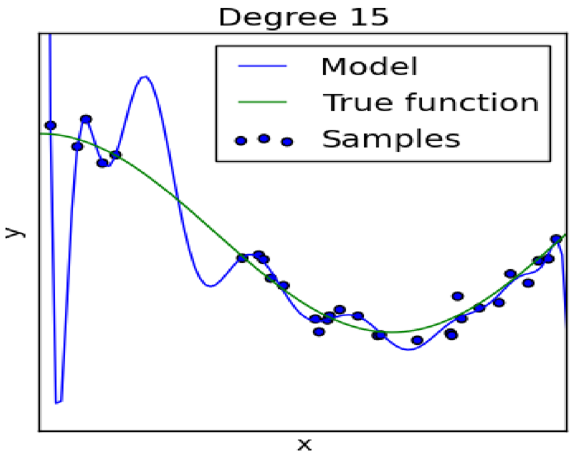
Low adj. R^2

Higher adj. R^2

Highest adj. R^2







OLS Regression Results

Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Budget	0.7846	0.133	5.901	0.000	0.520 1.049
Ones	4.44e+07	1.27e+07	3.504	0.001	1.92e+07 6.96e+07

Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

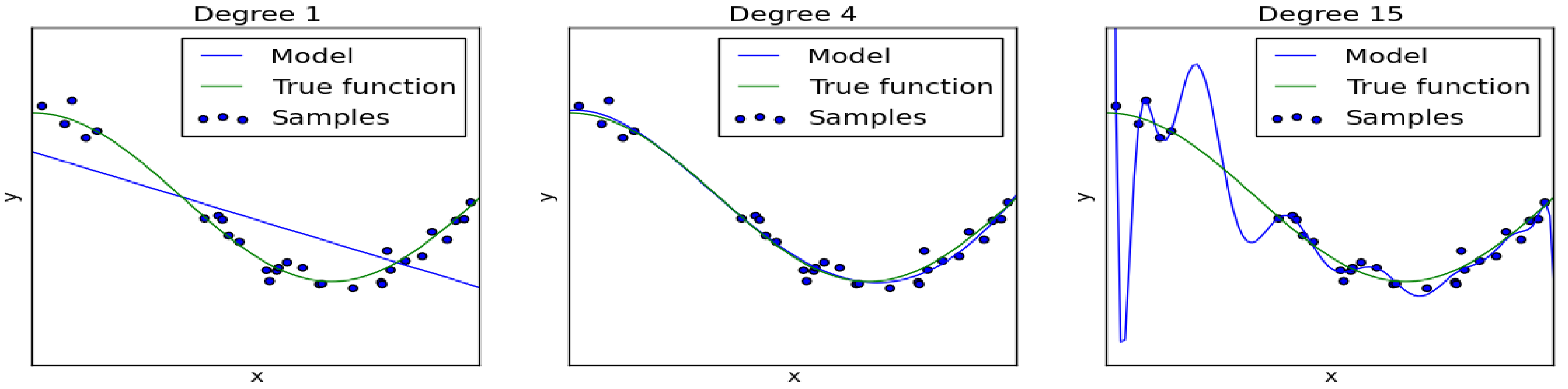
Akaike Information Criterion

AIC measures model parsimony (lower AIC is better)

$$AIC = 2k - 2\ln(L)$$

k = # parameters

L = Log likelihood



OLS Regression Results

Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Budget	0.7846	0.133	5.901	0.000	0.520 1.049
Ones	4.44e+07	1.27e+07	3.504	0.001	1.92e+07 6.96e+07

Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

Bayesian Information Criterion

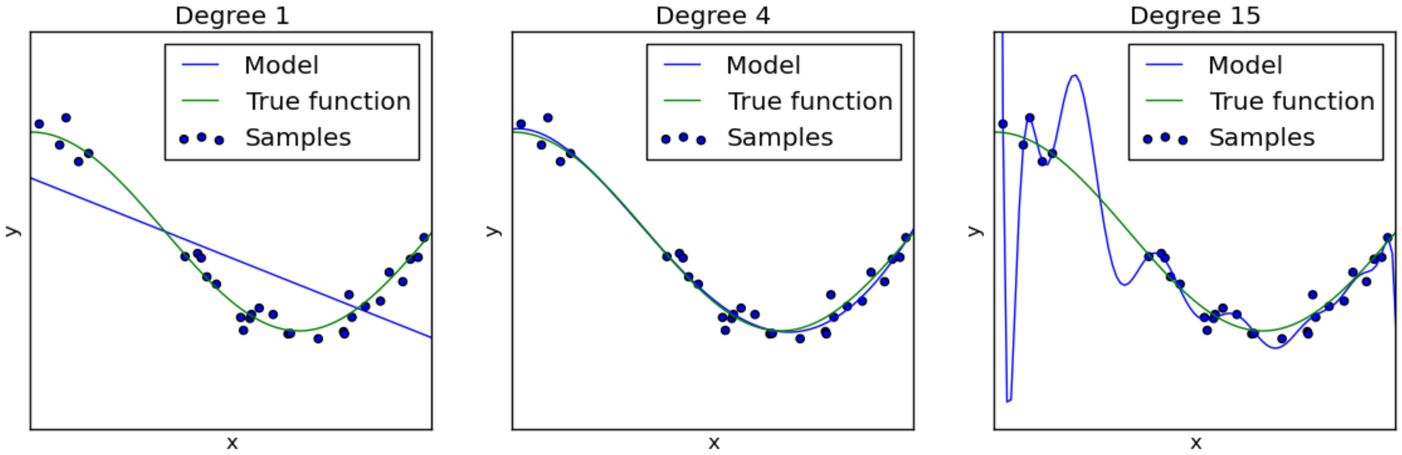
BIC is a penalized AIC, again lower is better

$$BIC = k \ln(m) - 2 \ln(L)$$

parameters

points

Log likelihood



My model isn't awesome enough yet!

What more can I do?

Try these and check test error (or AIC ,BIC, etc.):

1. Use a smaller set of features
2. Try adding polynomials
3. Check functional forms for each feature
4. Try including other features
5. Use more data (bigger training set)
6. Regularization (this week)
7. Try other model types (future lectures)