Linear Regression: What do these numbers mean?

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

y

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

Ordinary Least Squares

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		

R^2

	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

Best model minimizes

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

Sum of Squared Error
SSE

Variance of observed points (times m) is

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

Total Sum of Squares
SST

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

Randomness
left in the model

Variation in the data

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

Randomness left in the model

Variation in the data

SSE/SST is the portion of variation left unexplained by the model (handled by ε)

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

Randomness left in the model

Variation in the data

R^2 is the portion of variation explained by the model (R^2 is between 0 and 1)

(as long as the model has smaller residuals than the mean-only model)

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
(and the linear relationship we've found is purely due to chance)

Prob (F-statistic):

Is the p-value for this test. ie: it is the probability of finding the observed (or more extreme) results when the above null hypothesis (Ho) 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 mean 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 L

Likelihood is just a different cost function

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

For a given model (pair of β0 And β1 values),
Likelihood is the prob. Of getting exactly this set of observed y values

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

(and the data can be created by such a model (with the other β values intact)

P >|t|:

P-value for this test. Again 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:	$\boxed{39.749}$	**Durbin-Watson:**	0.674
Prob(Omnibus)	$\boxed{0.000}$	**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):

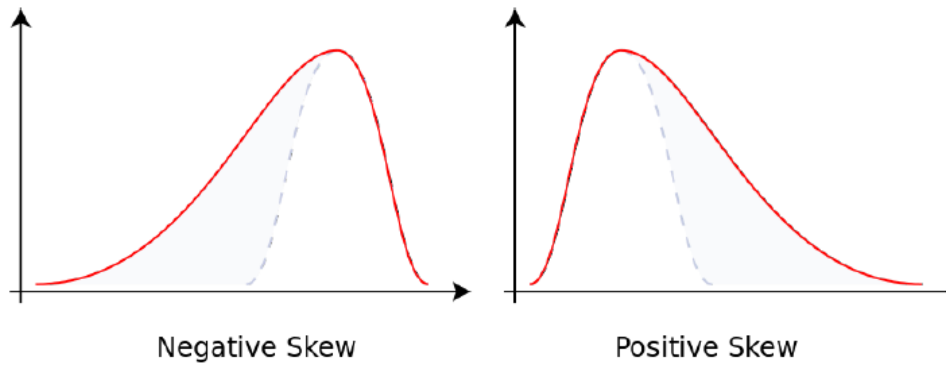
The p-value for this test. If p-value < 0.05, we reject the null hypothesis: ε does not exactly follow the normal distribution that we assumed.

We develop the normality test statistic :??

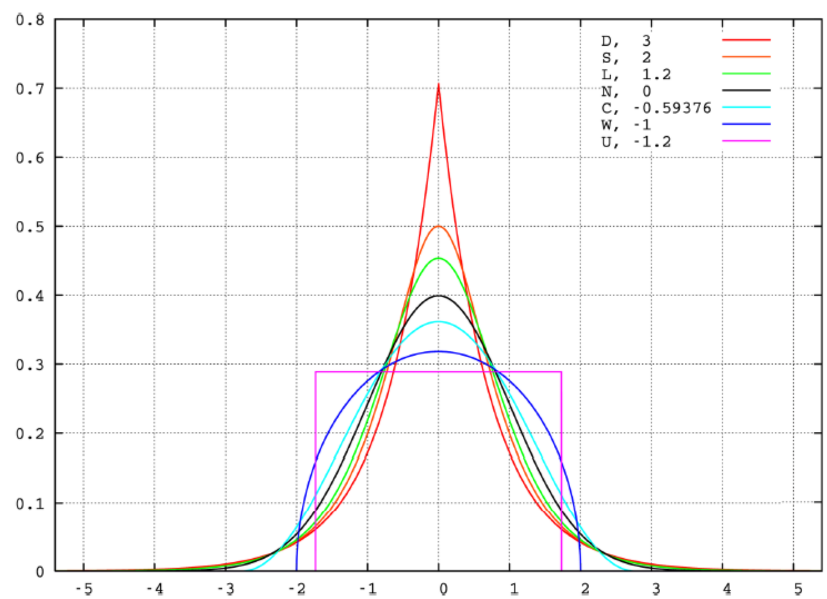
Skew & Kurtosis

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
(peakness)



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. Idea is : we are looking for a skewness coeff. ~ 0, and Kurtosis ~ 3. JB tests if those conditions are held against alternatives.

Prob(Omnibus):

The p-value for this test.

Omnibus:	39.749	**Durbin-Watson:**	$\boxed{0.674}$
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

Autocorrelation test

Null hypothesis:
Errors are uncorrelated

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

Sensitivity of prediction to small errors in input

Condition Number:

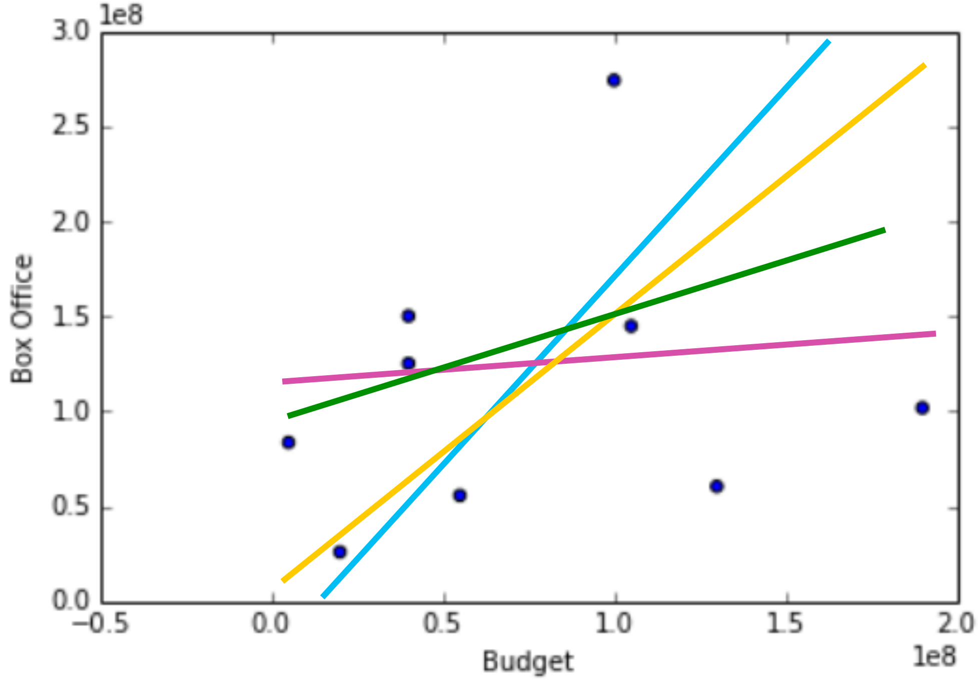
Given Mx=b, we can calculate the condition number :

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

Note that is the condition number becomes quite large, then this implies that the matrix is ill-posed (does not have a unique, well-defined solution). This may be due to multicollinear relationships between independent variables.

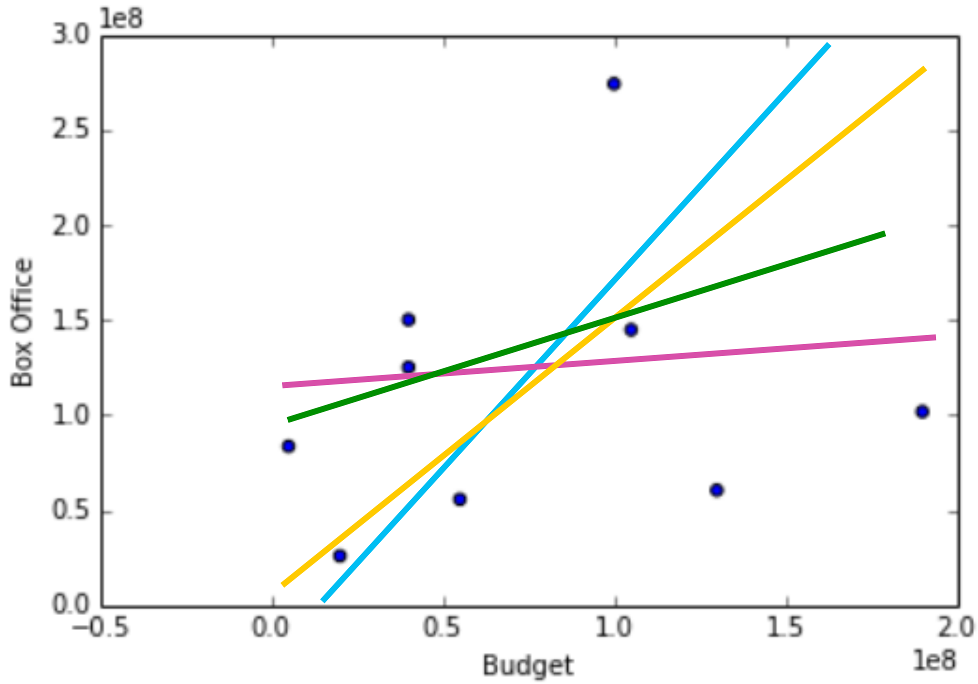
Model Selection I

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



For models with the same amount of parameters, easy:

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

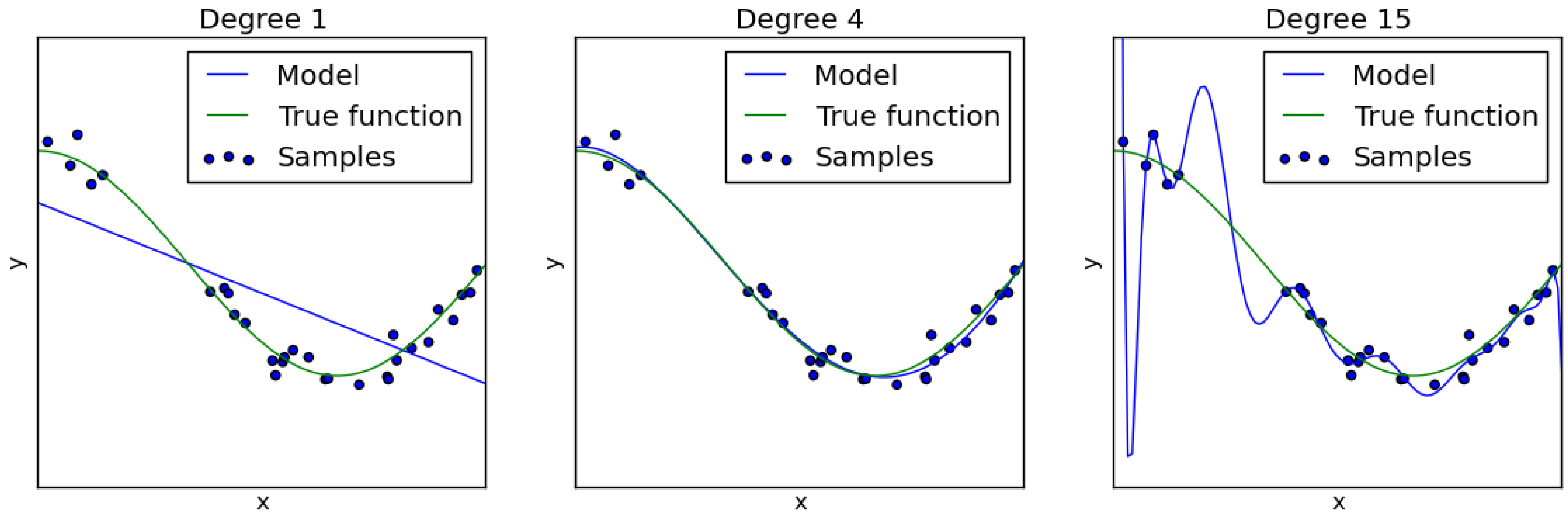


For models with the same amount of parameters, easy:

Take the one with the better cost function

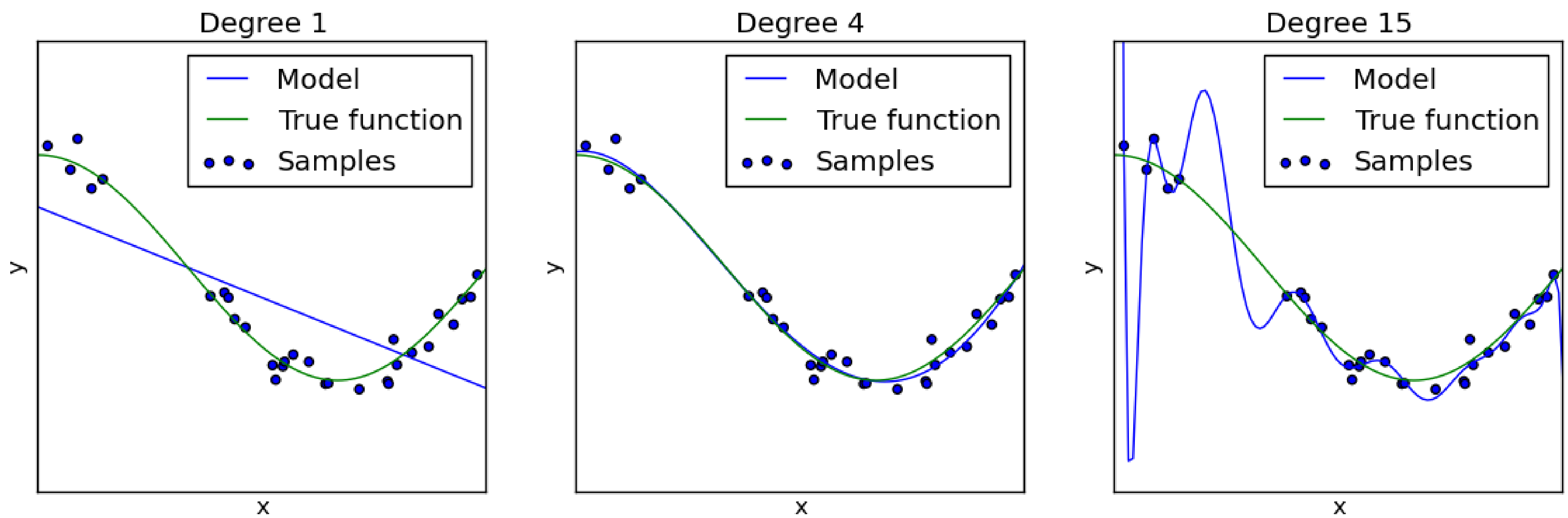
Log-Likelihood:	-1753.0

For models of different complexity: Beware under/overfitting



For models of different complexity: Beware under/overfitting

Underfitting — Just Right — Overfitting

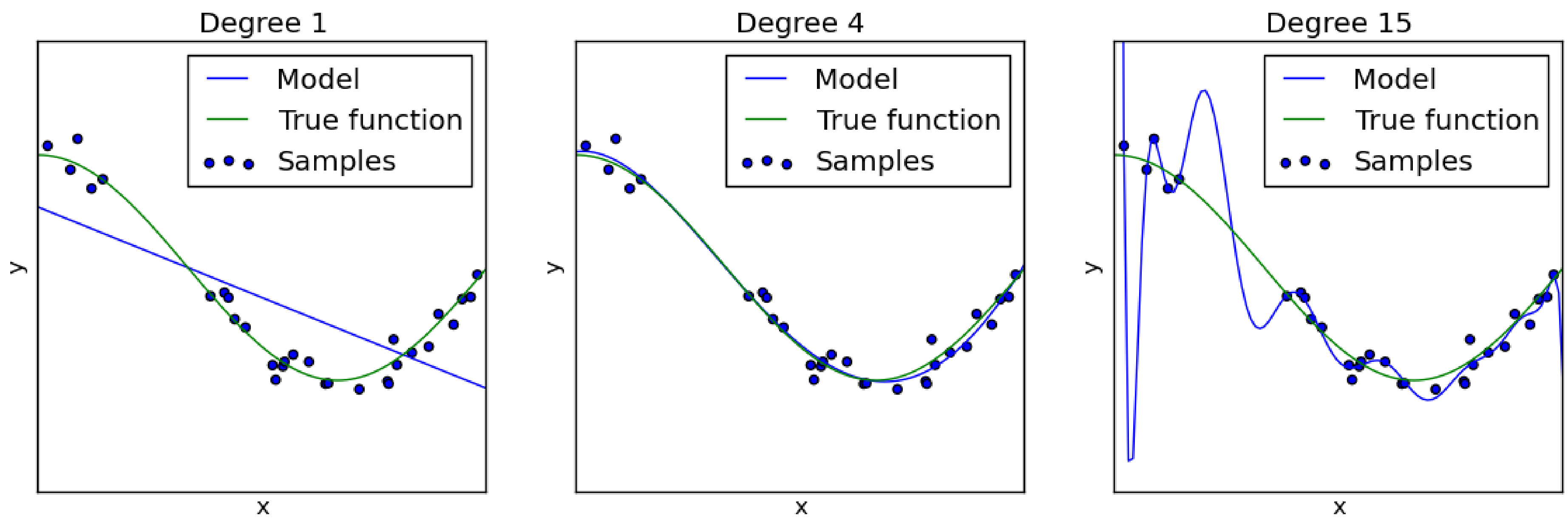


In machine learning, this is also called Bias/variance tradeoff

High bias
Low varience

Just Right

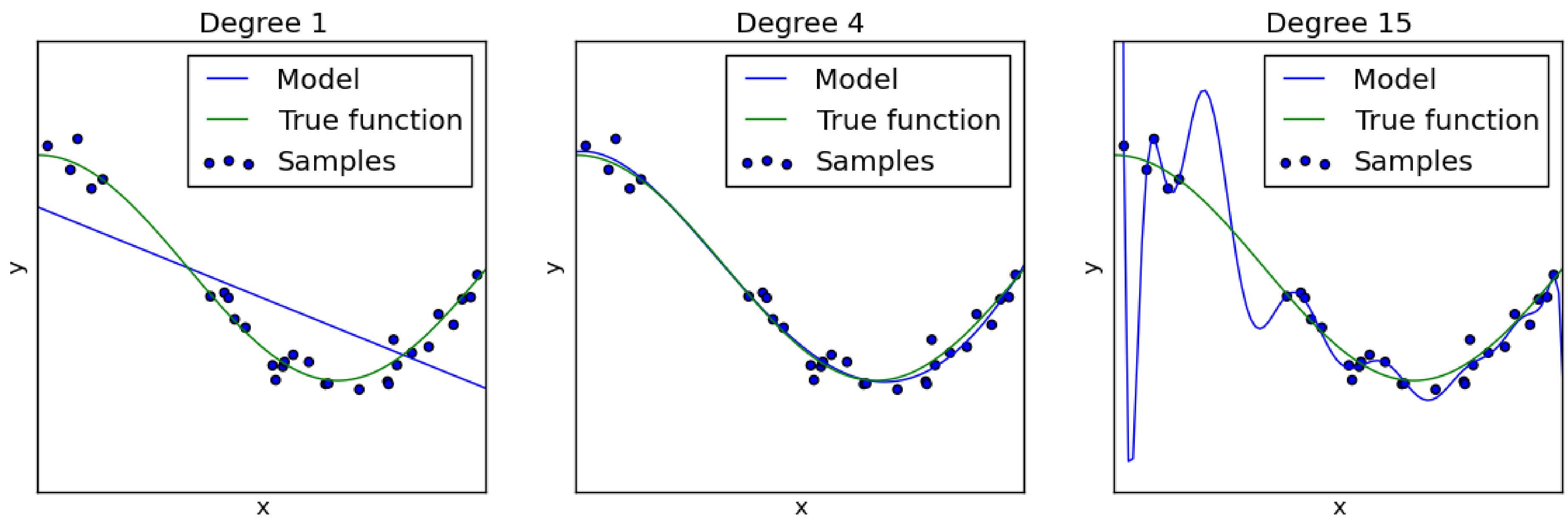
Low bias
High varience



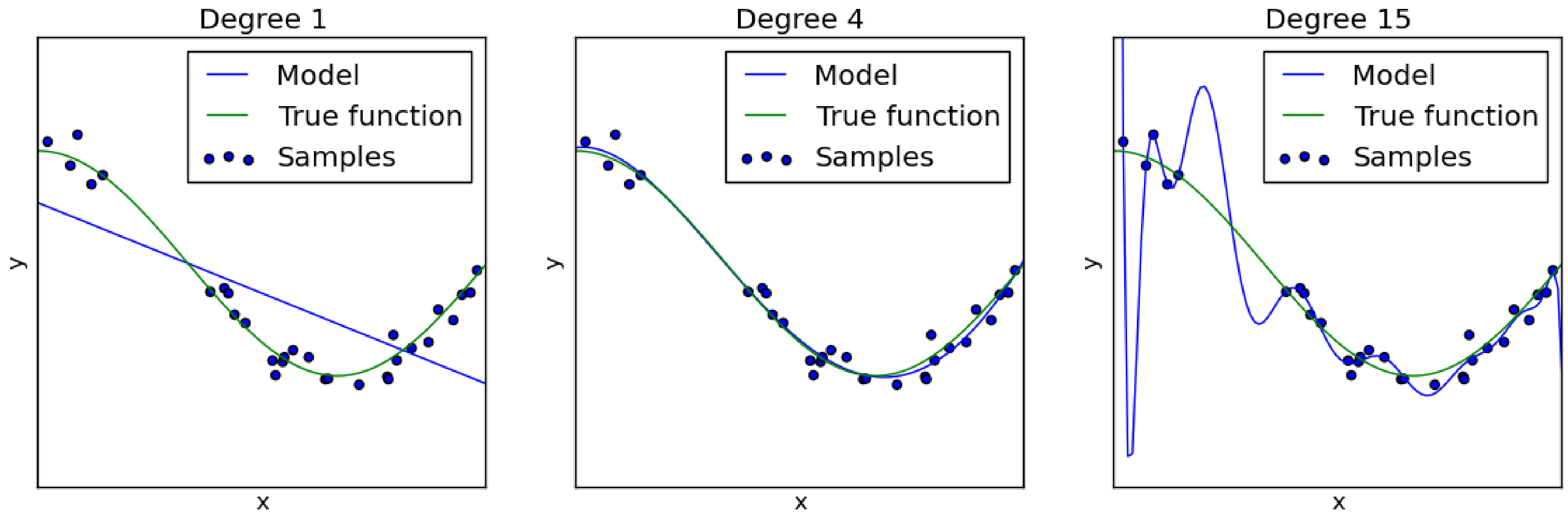
First training poorly,
predictions bad

Just Right

First training very well,
can’t generalize

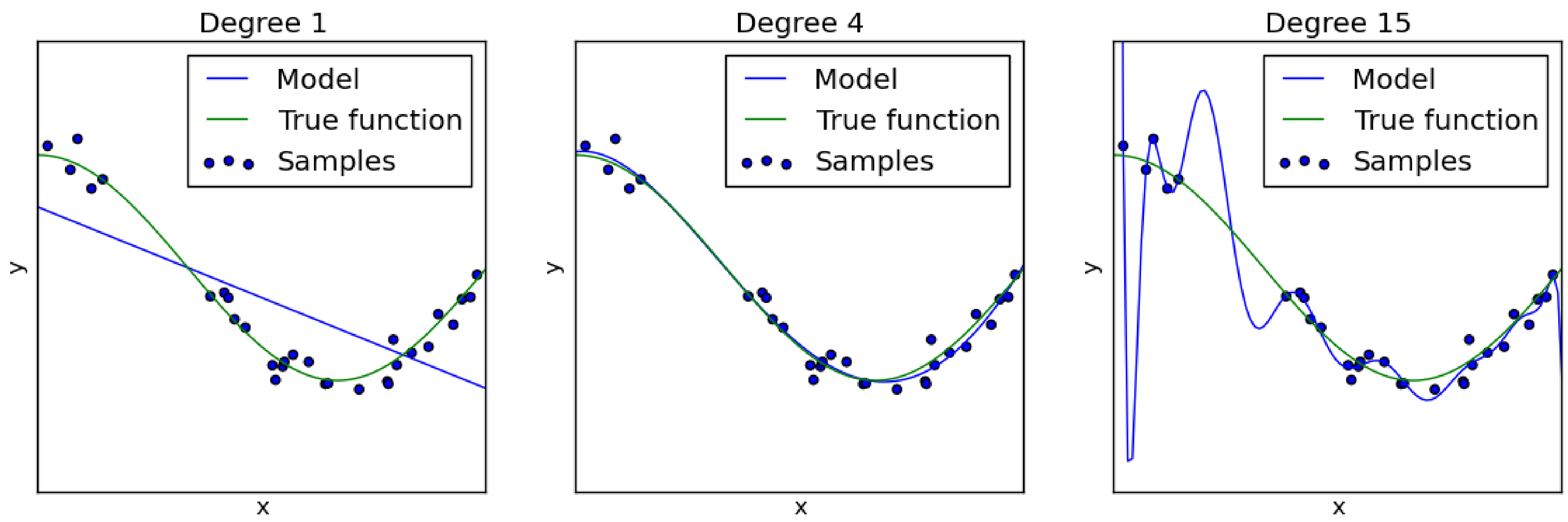


First and third will do poorly in the test set



Challenge: Fit a training set, calculate mean squared error on your test set (scikit learn)

There are a few metrics that try to measure this (without even looking at a test set yet)



OLS Regression Results

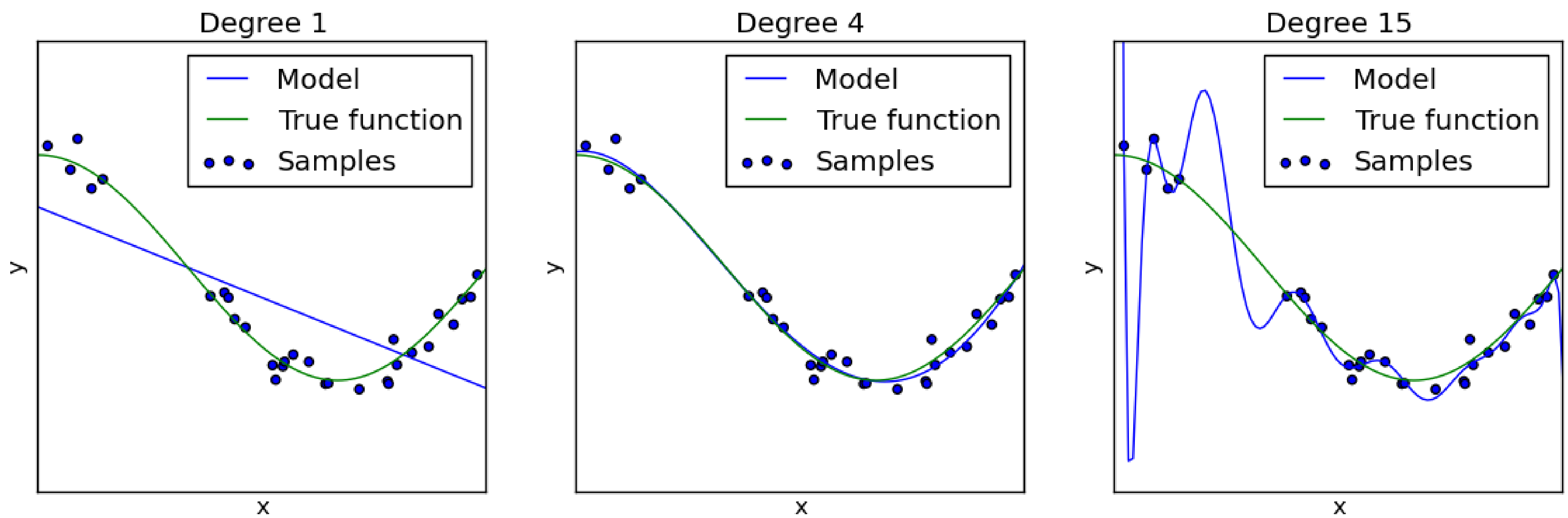
Dep. Variable:	DomesticTotalGross	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	34.82
Date:	Sun, 14 Sep 2014	**Prob (F-statistic):**	6.80e-08
Time:	21:59:46	**Log-Likelihood:**	-1738.1
No. Observations:	89	**AIC:**	3480.
Df Residuals:	87	**BIC:**	3485.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

Adjusted R^2

	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

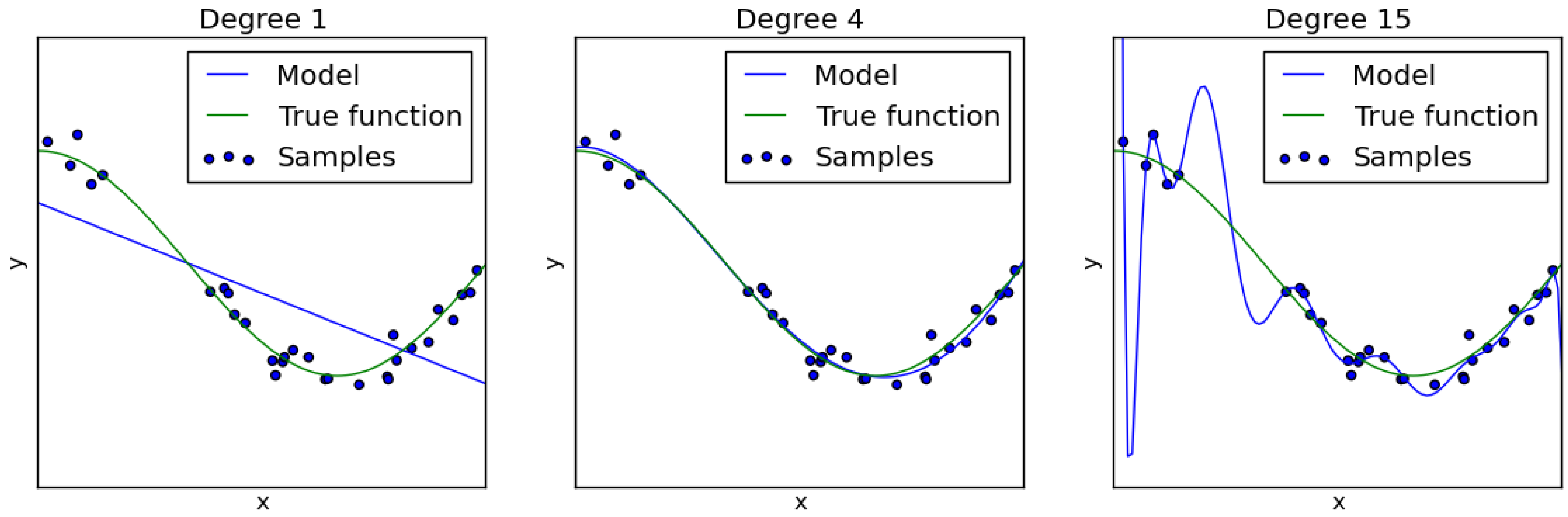
Omnibus:	39.749	**Durbin-Watson:**	0.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	99.441
Skew:	1.587	**Prob(JB):**	2.55e-22
Kurtosis:	7.091	**Cond. No.**	1.54e+08

Low R^2 Higher R^2 Highest R^2



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

Low R^2 | Higher R^2 | Highest R^2



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

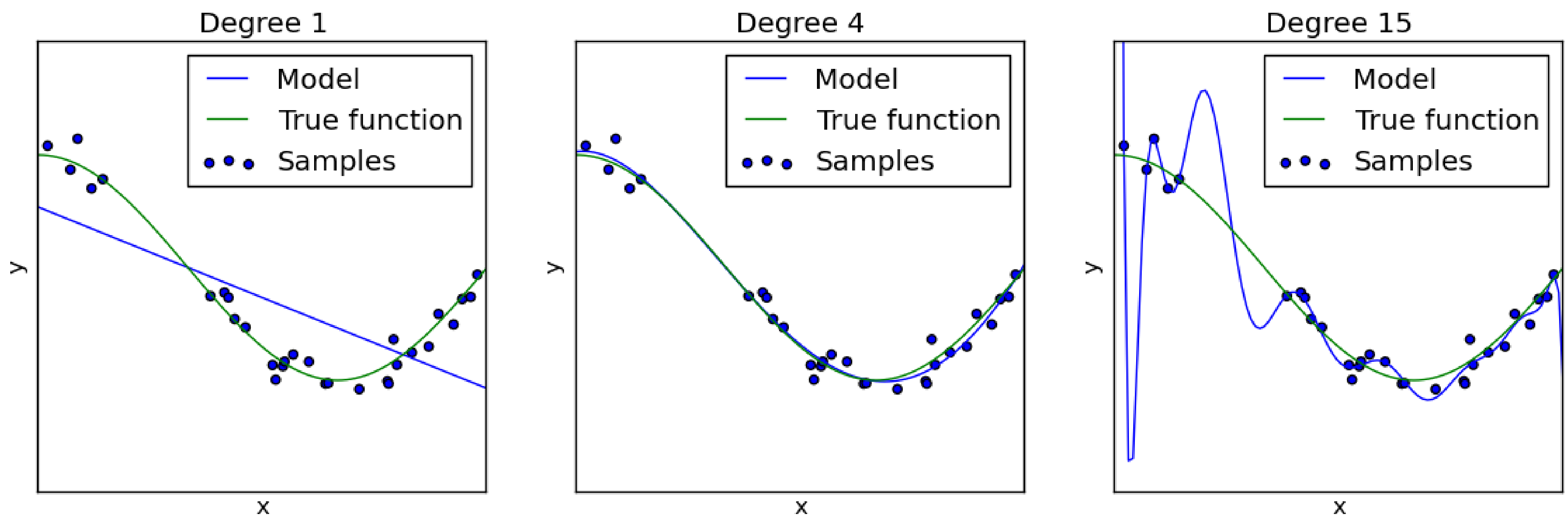
$m - k - 1$

$m - 1$

m= # points

k = # parameters

Low R^2 | Higher R^2 | Highest R^2



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

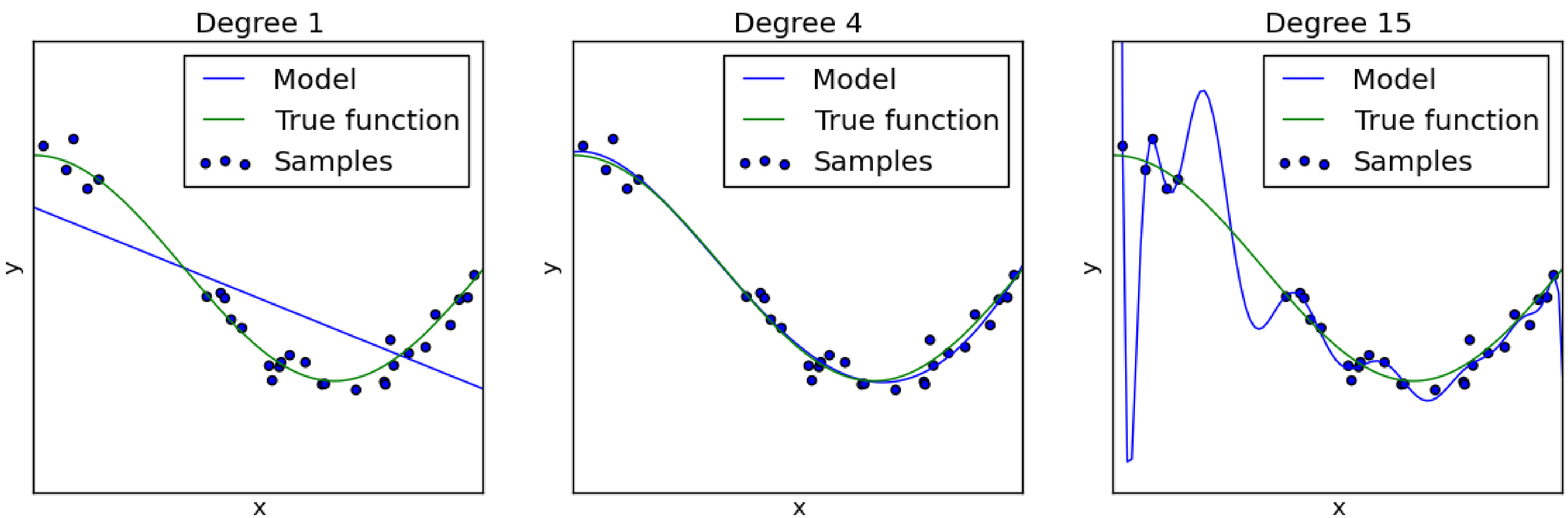
$df_e \longrightarrow m - k - 1$

$df_t \longrightarrow m - 1$

m= # points

k = # parameters

Low adj. R^2 | Max. adj R^2 | Low 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		

Akaike Information Criterion

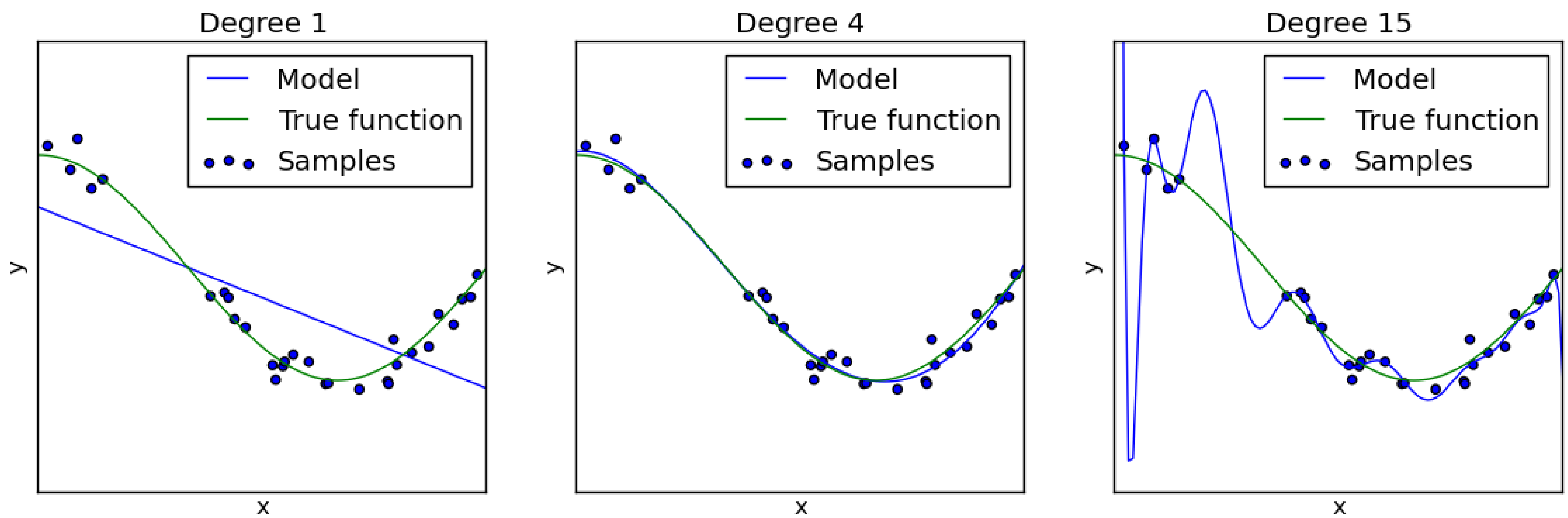
	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Budget	0.7846	0.133	5.901	0.000	0.520 1.049
Ones	4.44e+07	1.27e+07	3.504	0.001	1.92e+07 6.96e+07

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

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

parameters

Log likelihood



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

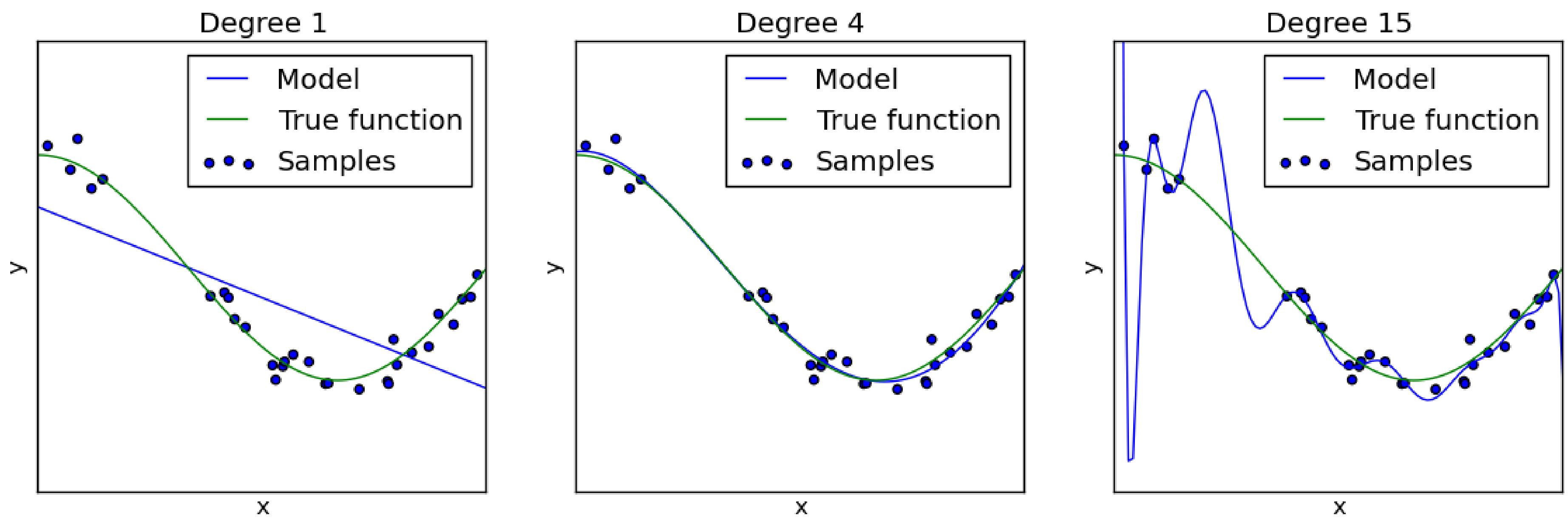
parameters

Log likelihood

Higher AIC

Min. AIC

Higher AIC



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		

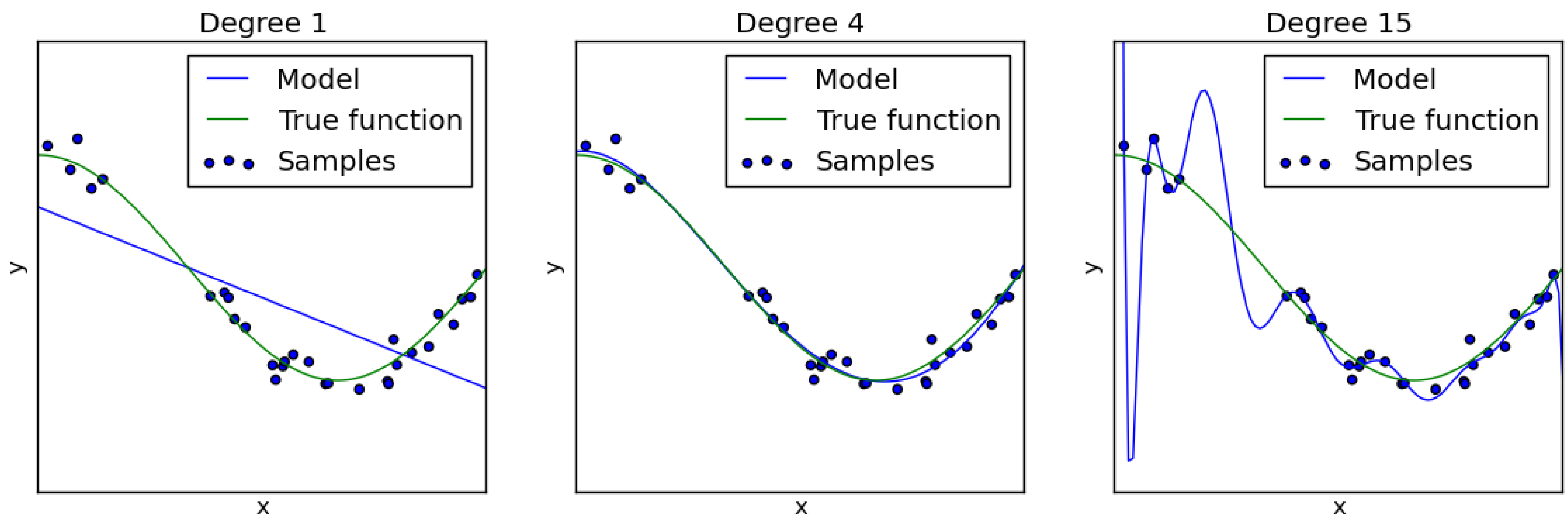
Bayesian Information Criterion

	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

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

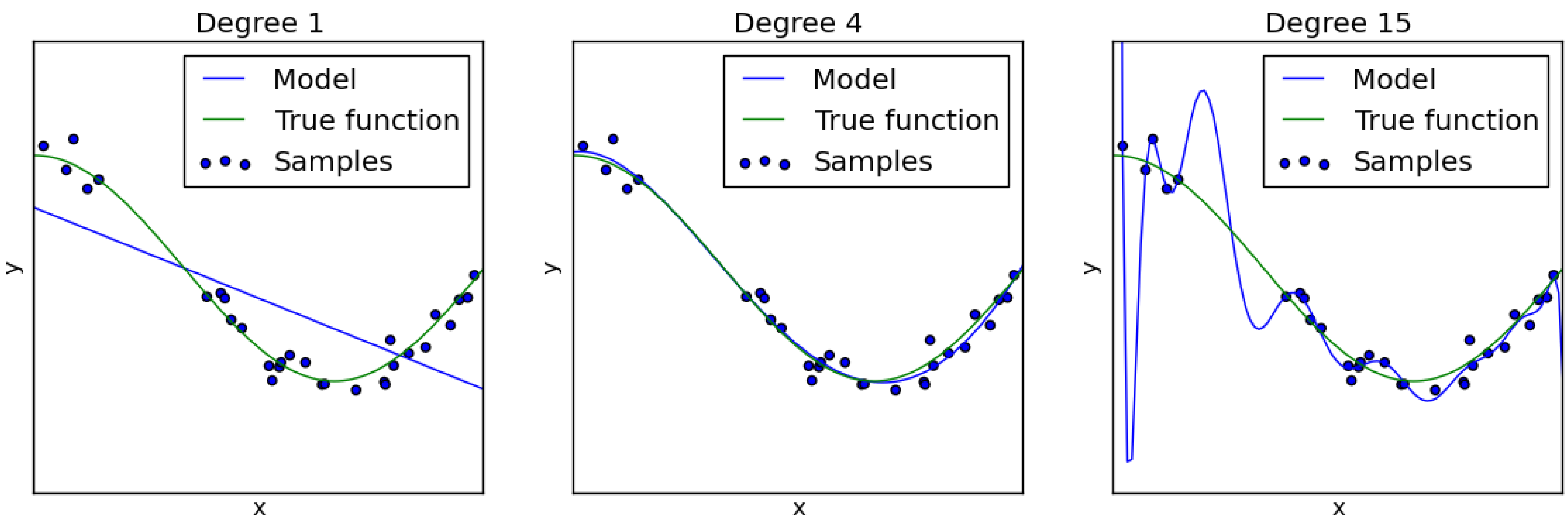
parameters # points Log likelihood



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

parameters   # points   Log likelihood

Higher BIC   Min. BIC   Higher BIC



My model is not awesome enough.

What do I do?

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

Use a smaller set of features
Try adding polynomials
Check functional forms for each feature
Try including other features
Use more data (bigger training set)
Regularization (tomorrow)
Try some other model (later)