Introduction to Regression and Model Fit, Part 2

Ivan Corneillet

Data Scientist

Learning Objectives

After this next lesson, you should be able to:

- How to conduct linear regression modeling
- Use interaction effects and binary categorical variables (also called dummy variables)
- Understand model complexity, underfitting, right fit, and overfitting
- Define regularization and error metrics for regression problems

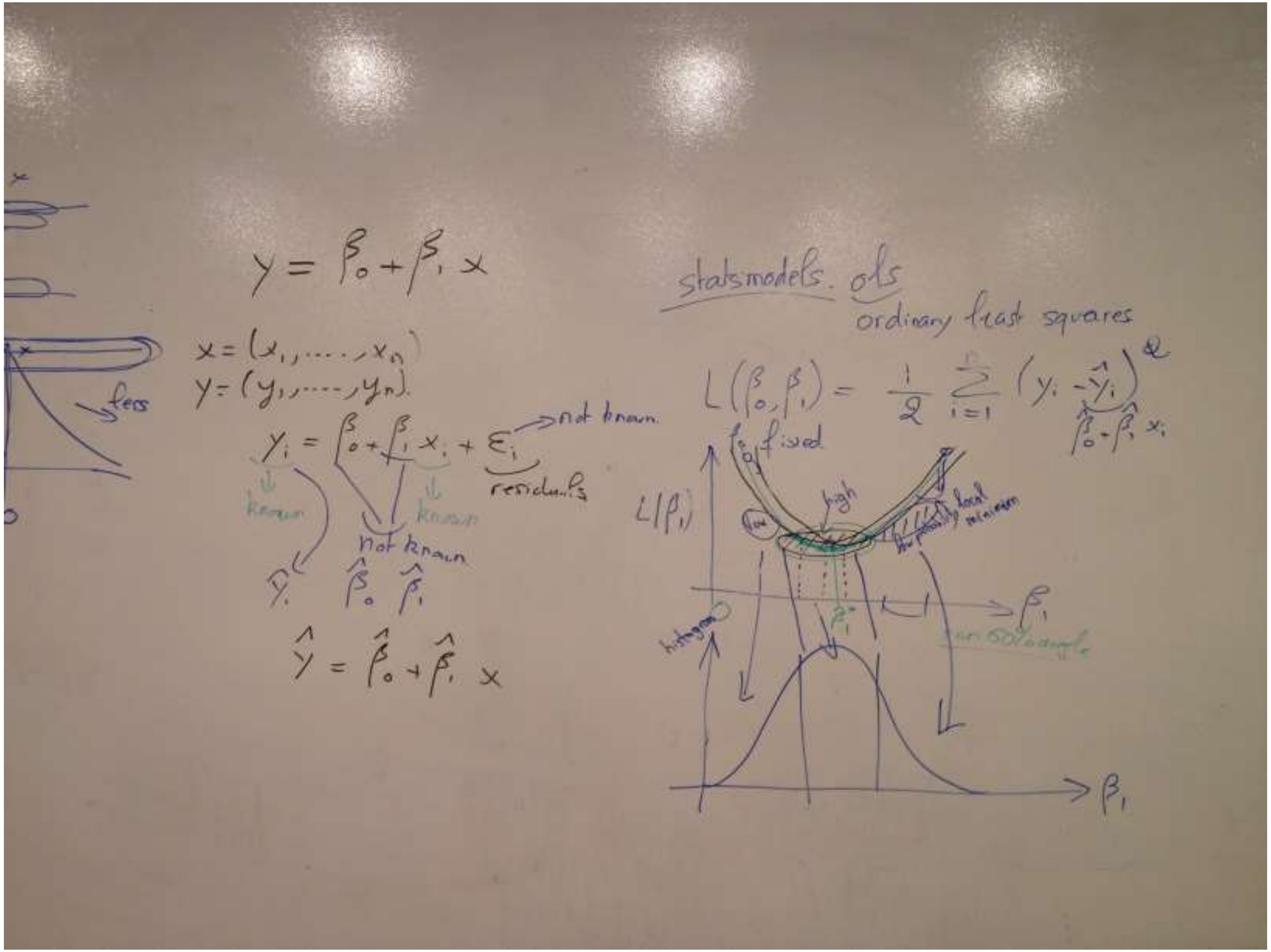
Announcements and Exit Tickets

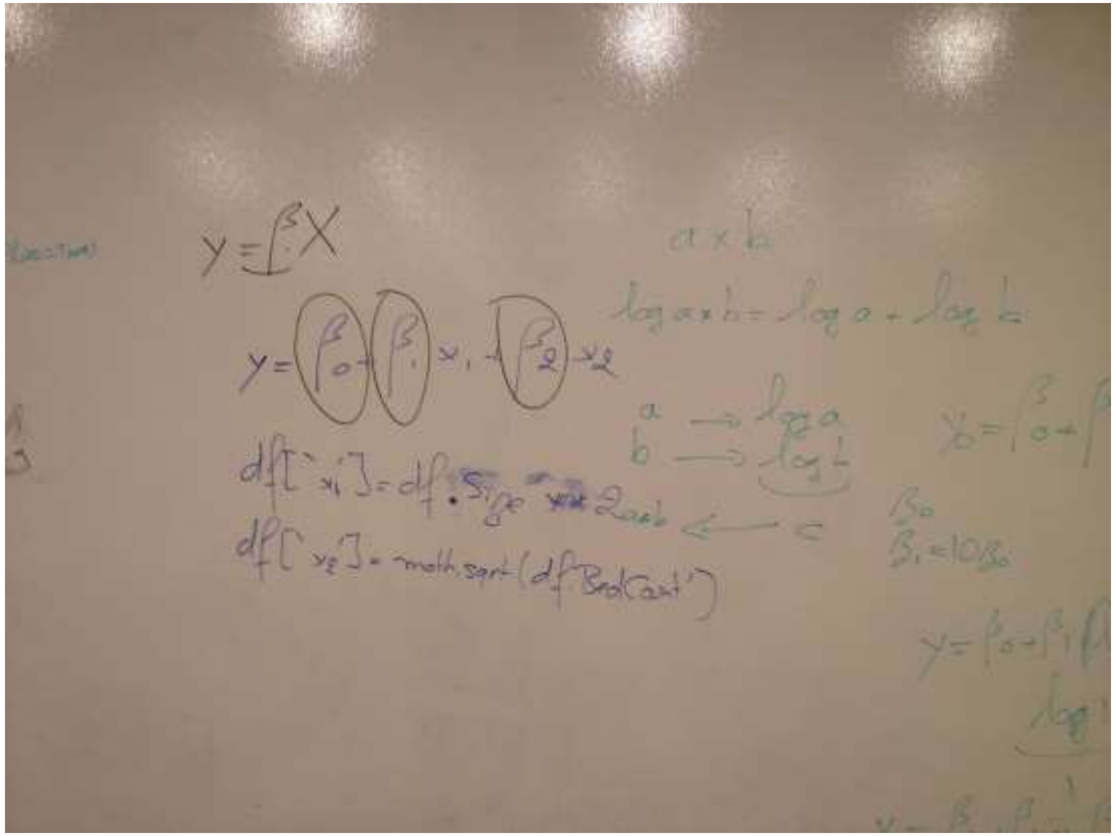
Q & A

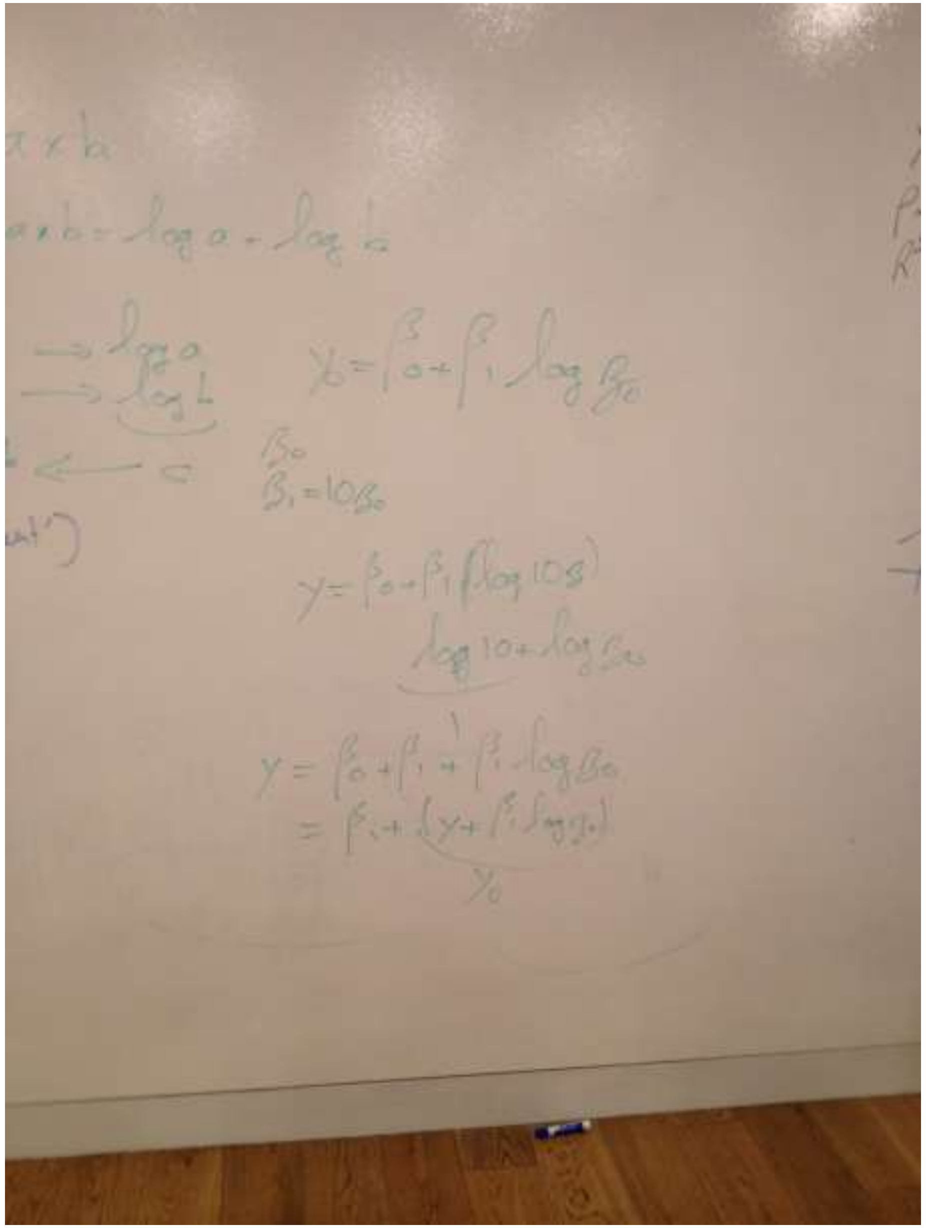
Review

Review

- Simple and Multiple Linear Regressions
- Common regression assumptions; how to check for them
- OLS (Ordinary Least Squares)
- How to interpret the model's parameters
- Variable Transformations
- Inference, Fit, R^2 (r-squared), and $\bar{R}^2$ (adjusted R^2)
- Multicollinearity

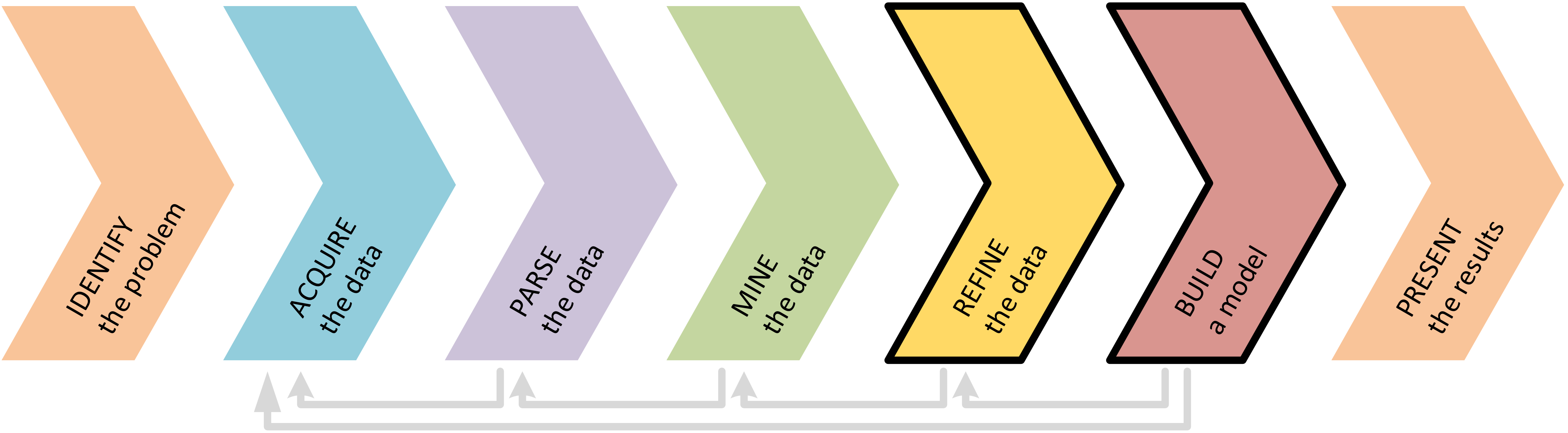

y = β0 + β1 x
x = (x1, ..., xn)
y = (y1, ..., yn)
yi = β0 + β1 xi + εi
not known
residuals
known
not known
ŷ = β̂0 + β̂1 x
statsmodels.ols
ordinary least squares
L(β0, β1) = 1/2 Σ (yi − ŷi)^2
fixed
high
local minimum
histogram
β1





Today

Today we keep our focus on the REFINE the data and BUILD a model steps but with (1) a focus on linear regression modeling and (2) what the inferential statistics tell us about the fit of these linear models



Today (cont.)

Research Design and Data Analysis	Research Design	Data Visualization in *pandas*	Descriptive Statistics for Exploratory Data Analysis Inferential Statistics for Model Fit	Exploratory Data Analysis in *pandas*
Foundations of Modeling	Linear Regression	Classification Models	Evaluating Model Fit	Presenting Insights from Data Models
Data Science in the Real World	Decision Trees and Random Forests	Time Series Data	Natural Language Processing	Databases

Here's what's happening today:

- Announcements and Exit Tickets
- Review
- ❺ Refine the Data and ❻ Build a Model | Linear Regression
  - F-statistic
  - Backward selection or "how to conduct linear regression modeling"
  - Linear Regression Modeling with sklearn (scikit-learn)
  - *statsmodels* vs. *sklearn*
  - Interaction Effects
  - Underfitting and overfitting; Training and generalization errors
  - (Complexity and Regularization)
  - Binary (dummy) categorical variables
- Lab – Introduction to Regression and Model Fit, Part 2
- Review

Linear Regression

Model's F-statistic

What β_i would make our multiple linear regression model useless?

- (the multiple linear regression model again)

$$y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k + \varepsilon$$

- Answer: If $\beta_0 = \beta_1 = \cdots = \beta_k = 0$, we don't have a model
  - ($y = 0$ isn't very exciting, is it?)

Model's F-statistic Hypothesis Test

- The *null hypothesis* (H_0) represents the status quo; that all β_i are zeros.

$$H_0: \beta_0 = \beta_1 = \cdots = \beta_k = 0$$

- The *alternate hypothesis* (H_a) represents the opposite of the null hypothesis (that at least one β_i is not zero) and holds true if H_0 is found to be false:

$$H_a: \exists i:\ \beta_i \neq 0$$

Linear Regression

Activity | Model's F-statistic

Activity | Model's F-statistic

EXERCISE

DIRECTIONS (10 minutes)

1. Using our Zillow dataset (`zillow-07-start.csv` in the `datasets` folder), run a simple linear regression between *SalePrice* (the *dependent* variable) and *Size* (the *independent* variable). Does the model has any predictive power? What F-value do you get? (You can choose to use today's codealong which setup the environment and loads the dataset for you)
2. Run another simple linear regression between *SalePrice* (the *dependent* variable) and *IsAStudio* (the *independent* variable). Answer the same questions: Does the model has any predictive power? What F-value do you get?
3. Using the F-distribution table, come up with a general criteria (assuming a reasonable sized dataset) to accept or reject the null hypothesis and make; also annotate when the model is useful and when it isn't
4. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions

Activity | Model’s F-statistic (cont.)

$SalePrice$ as a function of $Size$

Dep. Variable:	SalePrice	**R-squared:**	0.236
Model:	OLS	**Adj. R-squared:**	0.235
Method:	Least Squares	**F-statistic:**	297.4
Date:		**Prob (F-statistic):**	2.67e-58
Time:		**Log-Likelihood:**	-1687.9
No. Observations:	967	**AIC:**	3380.
Df Residuals:	965	**BIC:**	3390.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	0.1551	0.084	1.842	0.066	-0.010 0.320
Size	0.7497	0.043	17.246	0.000	0.664 0.835

Omnibus:	1842.865	**Durbin-Watson:**	1.704
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3398350.943
Skew:	13.502	**Prob(JB):**	0.00
Kurtosis:	292.162	**Cond. No.**	4.40

$SalePrice$ as a function of $IsAStudio$

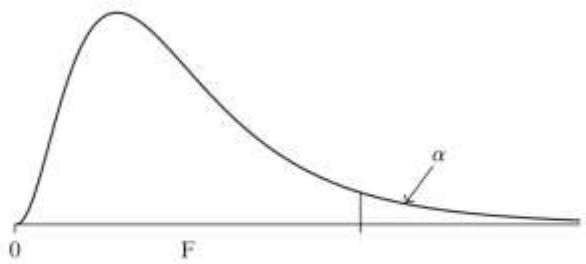
Dep. Variable:	SalePrice	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	-0.001
Method:	Least Squares	**F-statistic:**	0.07775
Date:		**Prob (F-statistic):**	0.780
Time:		**Log-Likelihood:**	-1847.4
No. Observations:	986	**AIC:**	3699.
Df Residuals:	984	**BIC:**	3709.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	1.3811	0.051	27.088	0.000	1.281 1.481
IsAStudio	0.0829	0.297	0.279	0.780	-0.501 0.666

Omnibus:	1682.807	**Durbin-Watson:**	1.488
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1342290.714
Skew:	10.942	**Prob(JB):**	0.00
Kurtosis:	182.425	**Cond. No.**	5.92

The F-distribution table ($\alpha = .05$) (note: $df_1 \cong k, df_2 = n$) (cont.)

F Distribution



	$\alpha = .05$									
	df_1									
df_2	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

Model's F-statistic ($\alpha = .05$) (cont.)

F-value	p-value	H_0 / H_a	Conclusion
≥ 4 (*) (*) (at least one variable and at least 120+ observations)	$\leq .05$	Found evidence that at least one $\beta_i \neq 0$: Reject H_0	At least one $\beta_i \neq 0$: The model is useful
< 4(*)	$> .05$	Did not find evidence that any $\beta_i \neq 0$: Fail to reject H_0	All $\beta_i = 0$: The model is useless (assume)

Linear Regression

Backward selection or "how to conduct linear regression modeling"

Two-step guidance on how to conduct linear regression modeling

❶ Model's significance

- Always start with the F-statistics for the whole model; only then check individual variables

❷ Regressors' significance

- Prefer to work solely with significant variables: if you observe insignificant variables you *usually* need to get rid of them and rerun your regression modeling without those
- Backward selection method
  - If you have insignificant variables, start dropping the most insignificant variable. If after removing that variable you still have insignificant variables, drop them one by one, until you are left with no insignificant variables

Linear Regression

Linear Regression Modeling with sklearn (scikit-learn)

Linear Modeling with *sklearn*

- When modeling with *sklearn* (scikit-learn), you'll use the following base principles:
  - All *sklearn* modeling classes are based on the base estimator `sklearn.base.BaseEstimator`
    - This means that all *sklearn* models take a similar form
    - All estimators take a matrix X (a *pandas* DataFrame), either sparse or dense
- Supervised estimators also take a vector y (the response) (a *pandas* Series)
- Estimators can be customized through setting the appropriate parameters

General format for *sklearn* model classes and methods

❶ `model = base_models.AnySKLearnObject()`

create an instance of an estimator class

❷ `model.fit(train_X, train_y)`

train your model; also called “fitting your data”

❸ `model.score(train_X, train_y)`

score your model using the training data using the default scoring method (recommended to use the metrics module in the future)

❹ `model.predict(test_X)`

predict your test data

❺ `model.score(test_X, test_y)`

score your model using your test data

❻ `model.predict(new_X)`

make predictions for a new set of data

Linear Regression

Codealong – Part A
Linear Regression Modeling with sklearn

Linear Regression

statsmodels vs. sklearn

statsmodels vs. *sklearn*

	Pros	Cons
statsmodels *(Takeaway: Use statsmodel for your modelling's inner-loop)*	❑ Does linear regression modelling very well ❑ Very convenient summary report about your model's fit: F-value and its p-value for the model. t-values, p-values, and confidence intervals for the coefficients ❑ Enable for quick iterations during the modeling phase	❑ Only does linear regression modelling ❑ Not very convenient when moving for modeling to predictions
sklearn *(Takeaway: Use sklearn to validate your model and then afterwards for production/prediction purpose)*	❑ Can be used to build a lot of different machine learning models with a very consistent programming interface (API) ❑ Lot of facilities (API) are available to validate your model (validation, cross-validation, etc. that we'll cover later in the course) ❑ Easy to use interface when moving over to prediction	❑ Doesn't provide a nice summary report for your linear regression model. E.g., no F-value for the entire model is reported and the p-values for the coefficients can be incorrect

Linear Regression

Back to our advertising dataset

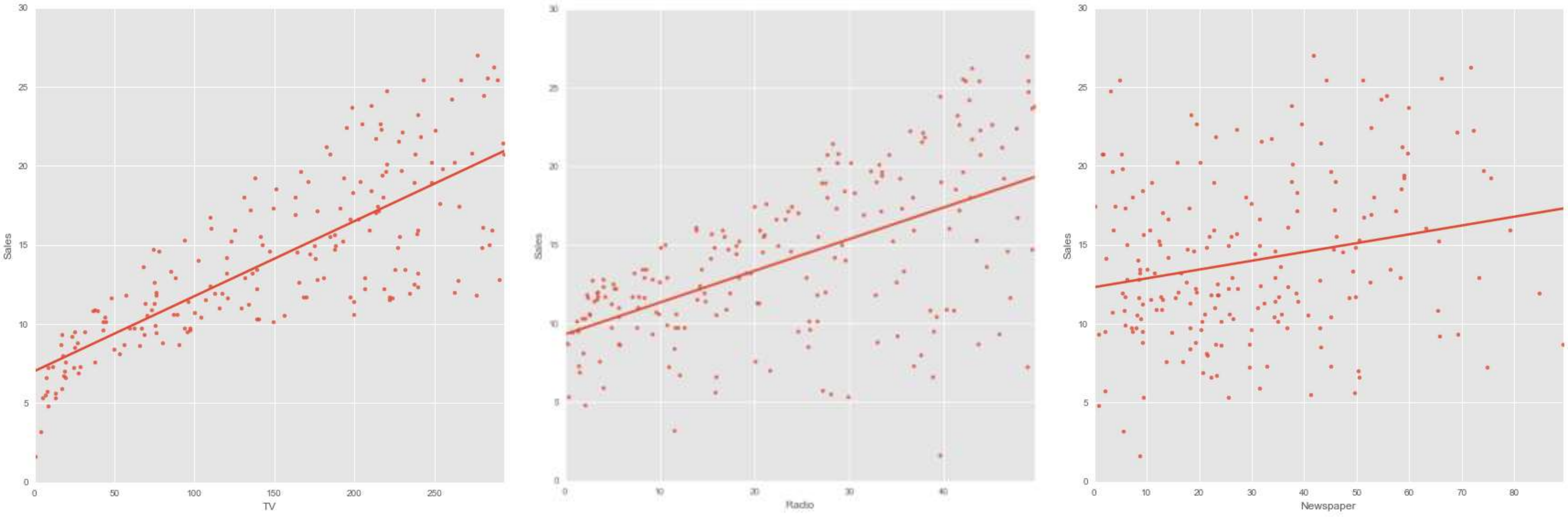
Source: An Introduction to Statistical Learning with Applications in R

Linear Regression

Simple Linear Regressions | $Sales \sim TV\ or\ Radio\ or\ Newspaper$

Is there a relationship between advertising budget and sales?

$Sales \sim TV$ $Sales \sim Radio$ $Sales \sim Newspaper$



Ordinary Least Squares

$Sales \sim TV$

Dep. Variable:	Sales	**R-squared:**	0.607
Model:	OLS	**Adj. R-squared:**	0.605
Method:	Least Squares	**F-statistic:**	302.8
Date:		**Prob (F-statistic):**	1.29e-41
Time:		**Log-Likelihood:**	-514.27
No. Observations:	198	**AIC:**	1033.
Df Residuals:	196	**BIC:**	1039.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	7.0306	0.462	15.219	0.000	6.120 7.942
TV	0.0474	0.003	17.400	0.000	0.042 0.053

Omnibus:	0.404	**Durbin-Watson:**	1.872
Prob(Omnibus):	0.817	**Jarque-Bera (JB):**	0.551
Skew:	-0.062	**Prob(JB):**	0.759
Kurtosis:	2.774	**Cond. No.**	338.

$Sales \sim Radio$

Dep. Variable:	Sales	**R-squared:**	0.333
Model:	OLS	**Adj. R-squared:**	0.329
Method:	Least Squares	**F-statistic:**	97.69
Date:		**Prob (F-statistic):**	5.99e-19
Time:		**Log-Likelihood:**	-566.70
No. Observations:	198	**AIC:**	1137.
Df Residuals:	196	**BIC:**	1144.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	9.3166	0.560	16.622	0.000	8.211 10.422
Radio	0.2016	0.020	9.884	0.000	0.161 0.242

Omnibus:	20.193	**Durbin-Watson:**	1.923
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	23.115
Skew:	-0.785	**Prob(JB):**	9.56e-06
Kurtosis:	3.582	**Cond. No.**	51.0

$Sales \sim Newspaper$

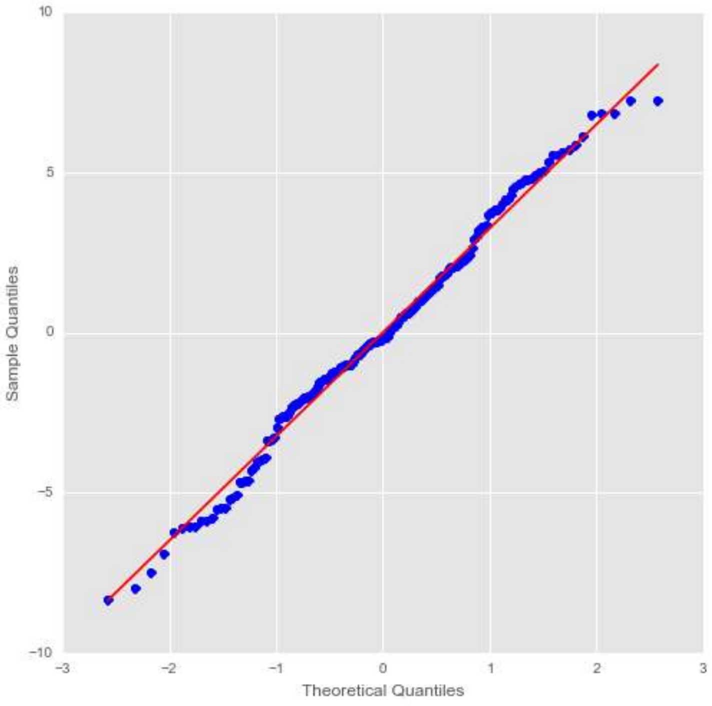
Dep. Variable:	Sales	**R-squared:**	0.048
Model:	OLS	**Adj. R-squared:**	0.043
Method:	Least Squares	**F-statistic:**	9.927
Date:		**Prob (F-statistic):**	0.00188
Time:		**Log-Likelihood:**	-601.84
No. Observations:	198	**AIC:**	1208.
Df Residuals:	196	**BIC:**	1214.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	12.3193	0.639	19.274	0.000	11.059 13.580
Newspaper	0.0558	0.018	3.151	0.002	0.021 0.091

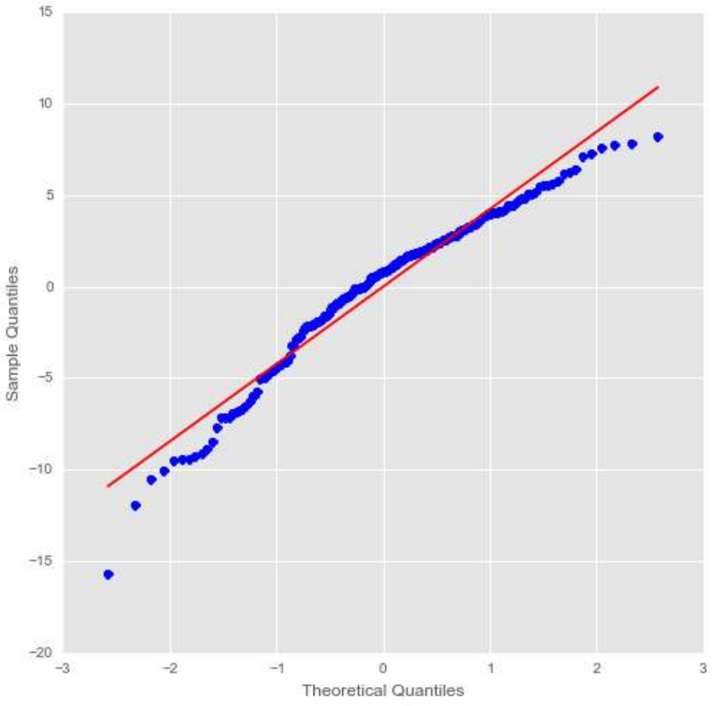
Omnibus:	5.835	**Durbin-Watson:**	1.916
Prob(Omnibus):	0.054	**Jarque-Bera (JB):**	5.303
Skew:	0.333	**Prob(JB):**	0.0706
Kurtosis:	2.555	**Cond. No.**	63.9

q-q plots of residuals. Are they normally distributed?

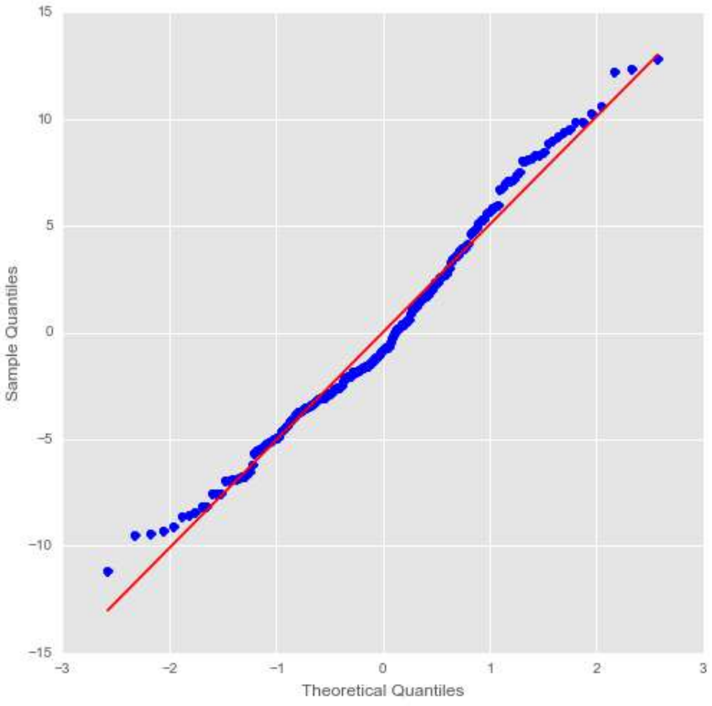
$Sales \sim TV$



$Sales \sim Radio$

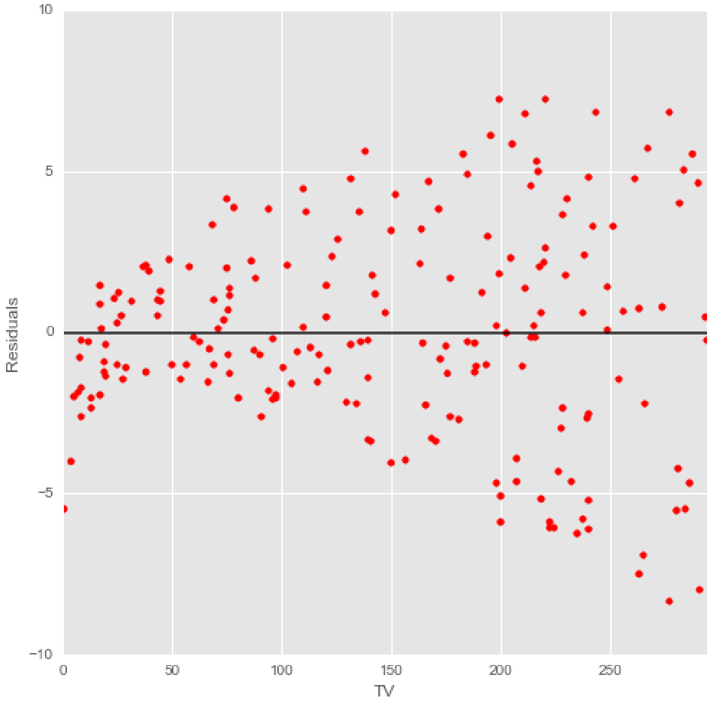


$Sales \sim Newspaper$

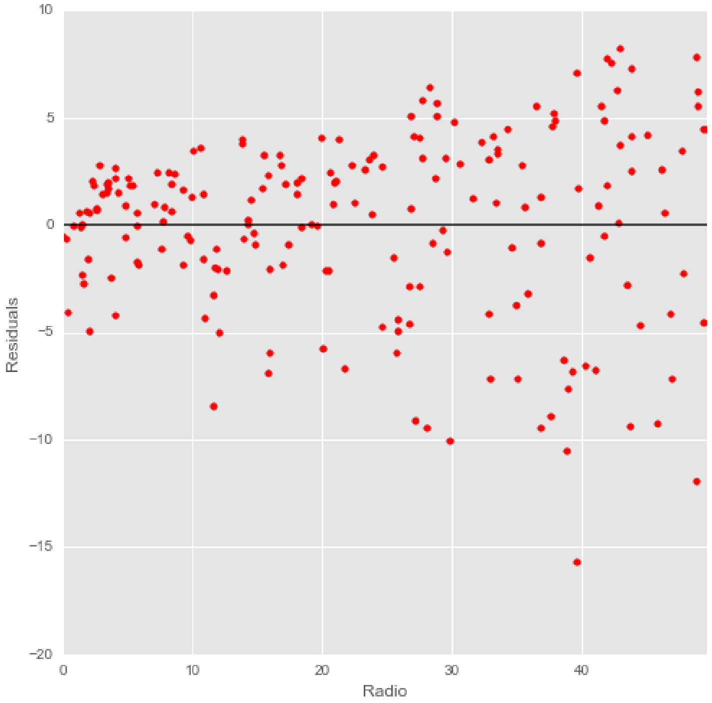


Scatterplots of residuals against advertising budget. Are they randomly distributed?

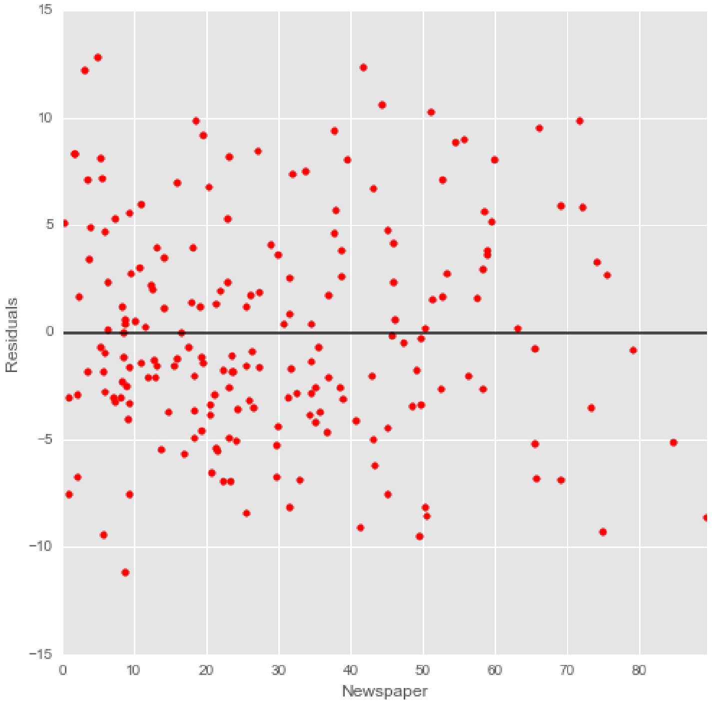
$Sales \sim TV$



$Sales \sim Radio$



$Sales \sim Newspaper$



Linear Regression

Multiple Linear Regression | $Sales \sim TV + Radio + Newspaper$

$Sales \sim TV + Radio + Newspaper$

Dep. Variable:	Sales	R-squared:	0.895
Model:	OLS	Adj. R-squared:	0.894
Method:	Least Squares	F-statistic:	553.5
Date:		Prob (F-statistic):	8.35e-95
Time:		Log-Likelihood:	-383.24
No. Observations:	198	AIC:	774.5
Df Residuals:	194	BIC:	787.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	2.9523	0.318	9.280	0.000	2.325 3.580
TV	0.0457	0.001	32.293	0.000	0.043 0.048
Radio	0.1886	0.009	21.772	0.000	0.171 0.206
Newspaper	-0.0012	0.006	-0.187	0.852	-0.014 0.011

Omnibus:	59.593	Durbin-Watson:	2.041
Prob(Omnibus):	0.000	Jarque-Bera (JB):	147.654
Skew:	-1.324	Prob(JB):	8.66e-33
Kurtosis:	6.299	Cond. No.	457.

Linear Regression

Multiple Linear Regression | $Sales \sim TV + Radio$

$Sales \sim TV + Radio$. Are we done yet?

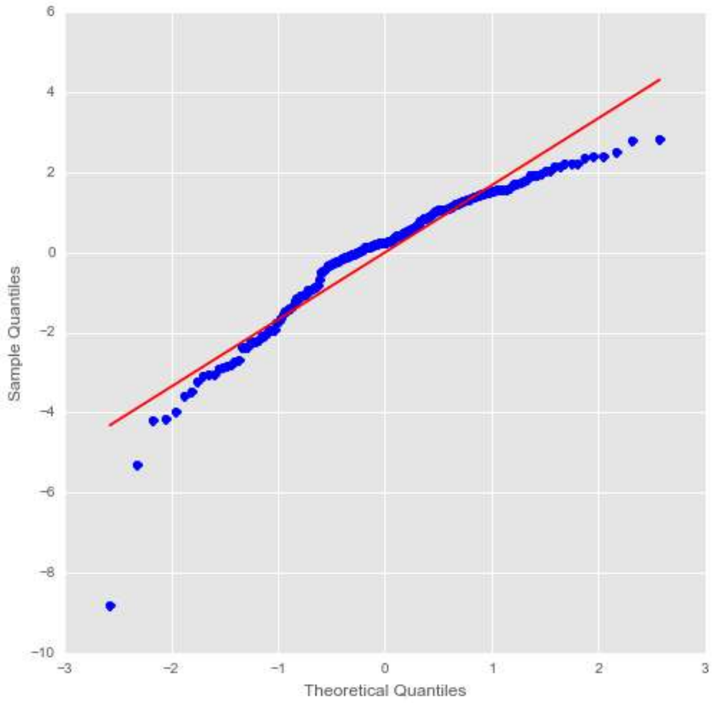
Dep. Variable:	Sales	**R-squared:**	0.895
Model:	OLS	**Adj. R-squared:**	0.894
Method:	Least Squares	**F-statistic:**	834.4
Date:		**Prob (F-statistic):**	2.60e-96
Time:		**Log-Likelihood:**	-383.26
No. Observations:	198	**AIC:**	772.5
Df Residuals:	195	**BIC:**	782.4
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	2.9315	0.297	9.861	0.000	2.345 3.518
TV	0.0457	0.001	32.385	0.000	0.043 0.048
Radio	0.1880	0.008	23.182	0.000	0.172 0.204

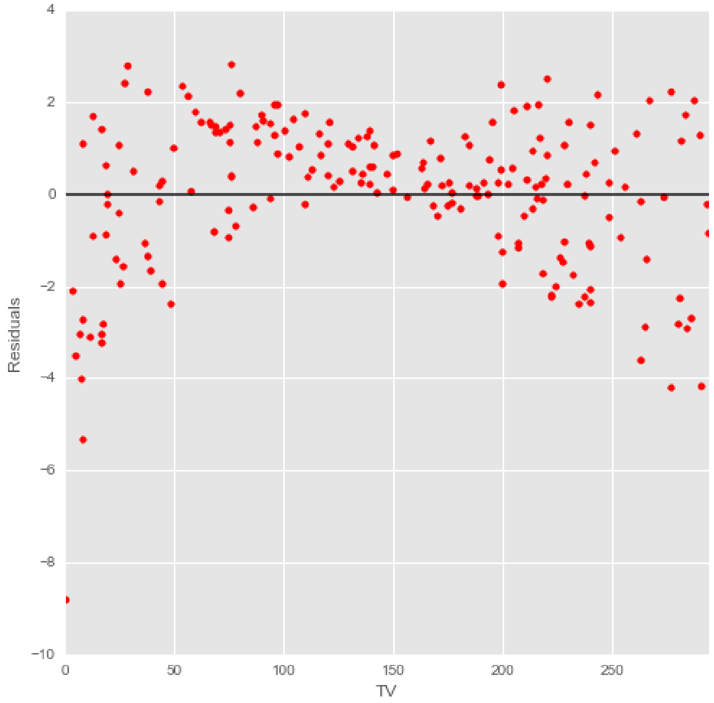
Omnibus:	59.228	**Durbin-Watson:**	2.038
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	145.127
Skew:	-1.321	**Prob(JB):**	3.06e-32
Kurtosis:	6.257	**Cond. No.**	423.

$Sales \sim TV + Radio$. What do you observe? Are we done yet?

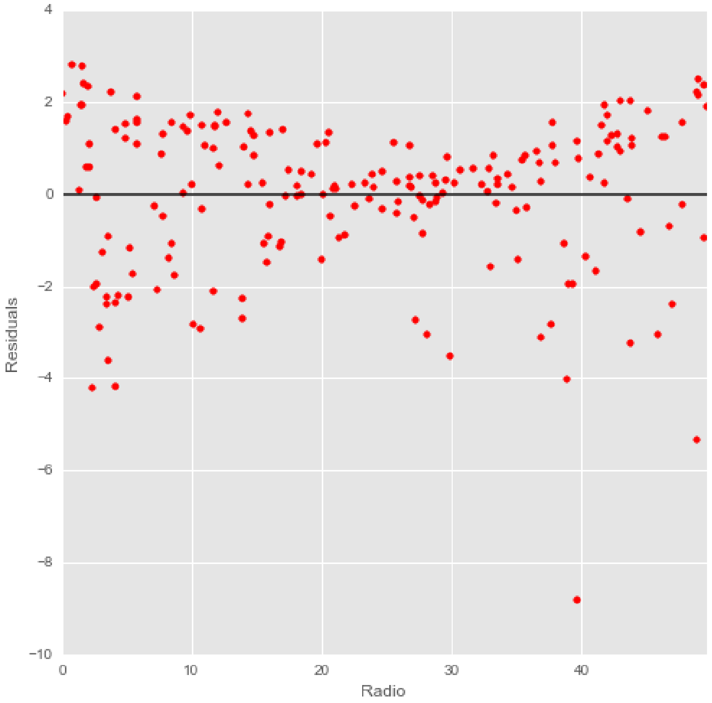
Residuals q-q plot



Residuals against TV



Residuals against $Radio$



$Sales \sim TV + Radio$

- $Sales = \underbrace{2.93}_{\widehat{\beta}_0} + \underbrace{.0457}_{\widehat{\beta}_1} \times TV + \underbrace{.188}_{\widehat{\beta}_2} \times Radio$

- This model assumes that the effect on sales of increasing one media (e.g., TV) is independent of the amount spent on the other media (e.g., $Radio$)

- More specifically, the model states that the average effect on sales of a one-unit increase (\$1,000) in TV is always ($\underbrace{.0457}_{\widehat{\beta}_1} \times \underbrace{.\$1{,}000}_{TV}$ = \$45.7), regardless of the amount spend on $Radio$

Linear Regression

Interaction Effects

Interaction effects

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising
  - the slope term for TV should increase as $Radio$ increases
- E.g., given a fixed budget of \$100,000, spending half on TV and half on radio may increase sales more than allocating the entire amount to either TV or radio
- This is known as a synergy effect in marketing; in statistics it is referred to as an interaction effect

Linear Regression

Codealong – Part B
Interaction Effects

Sales ~ TV + Radio + TV * Radio

Dep. Variable:	Sales	**R-squared:**	0.968
Model:	OLS	**Adj. R-squared:**	0.967
Method:	Least Squares	**F-statistic:**	1934.
Date:		**Prob (F-statistic):**	3.19e-144
Time:		**Log-Likelihood:**	-267.07
No. Observations:	198	**AIC:**	542.1
Df Residuals:	194	**BIC:**	555.3
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	6.7577	0.247	27.304	0.000	6.270 7.246
TV	0.0190	0.002	12.682	0.000	0.016 0.022
Radio	0.0276	0.009	3.089	0.002	0.010 0.045
TV:Radio	0.0011	5.27e-05	20.817	0.000	0.001 0.001

Omnibus:	126.182	**Durbin-Watson:**	2.241
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1151.060
Skew:	-2.306	**Prob(JB):**	1.12e-250
Kurtosis:	13.875	**Cond. No.**	1.78e+04

Interaction effects (cont.)

- $Sales = \underbrace{6.76}_{\widehat{\beta}'_0} + \underbrace{.0190}_{\widehat{\beta}'_1} \times TV + \underbrace{.0276}_{\widehat{\beta}'_2} \times Radio + \underbrace{.0011}_{\widehat{\beta}'_3} \times TV \times Radio$
- The interaction is important
  - β'_3 is statistically significant
  - R^2 with this model went up to 96.8% up from 89.5% for the model without interaction. This that $1 - \frac{1-.968}{1-.895} = .70 = 70\%$ of the unexplained variability in the previous model has been explained by the interaction term

Activity | Interaction effects

EXERCISE

DIRECTIONS (10 minutes)

1. Our TV budget is $50,000 that we consider increasing it by $5,000. What would be the corresponding increase in sales based on different levels of radio budget?
2. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions

Activity | Interaction effects (cont.)

EXERCISE

Radio budget	Model without interactions	Model with interactions
Formula	$\underbrace{.0457}_{\hat{\beta}_1} \times \Delta TV$	$\left(\underbrace{.0190}_{\hat{\beta}_1'} + \underbrace{.0011}_{\hat{\beta}_3'} \times Radio\right) \times \Delta TV$
$\$15{,}000$	$.0457 \times 5 = .228 = \$229$	$(.0190 + .0011 \times 15) \times 5$ $= .178 = \$178$
$\$10{,}000$	$\$229$	$(.0190 + .0011 \times 10) \times 5$ $= .150 = \$150$
$\$5{,}000$	$\$229$	$(.0190 + .0011 \times 5) \times 5$ $= .123 = \$123$

Hierarchy Principle

- Sometimes an interaction term $x_i \cdot x_j$ is significant, but one or both of its main effects (in this case x_i and/or x_j) are not

- The hierarchy principle
  - If we include an interaction in a model, we should also include the main effects, even if they aren't significant

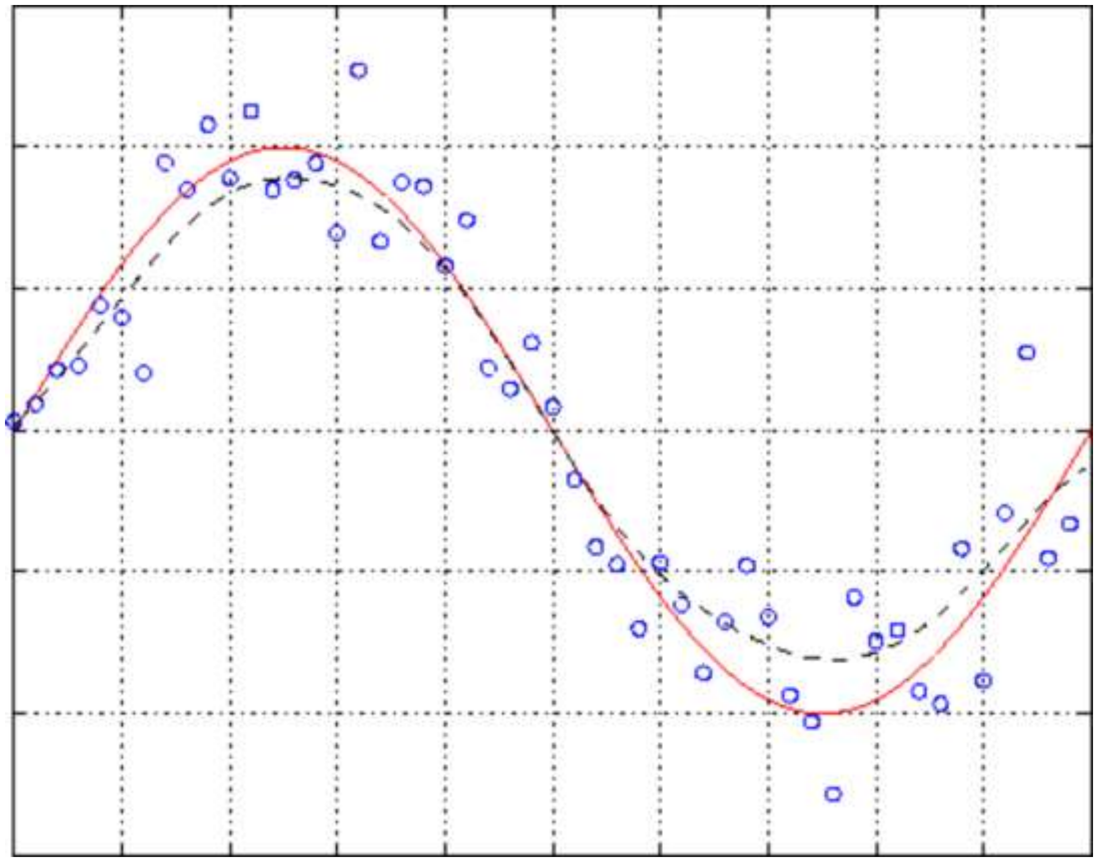
Linear Regression

Underfitting and overfitting

Training and generalization errors

Polynomial regressions

- Polynomial regressions ($y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \cdots + \beta_k \cdot x^k + \varepsilon$) allow us to fit very complex curves (nonlinear relationships) to the data
- *(For now, we will gloss over the multicollinearity issue we mentioned in the previous lecture)*



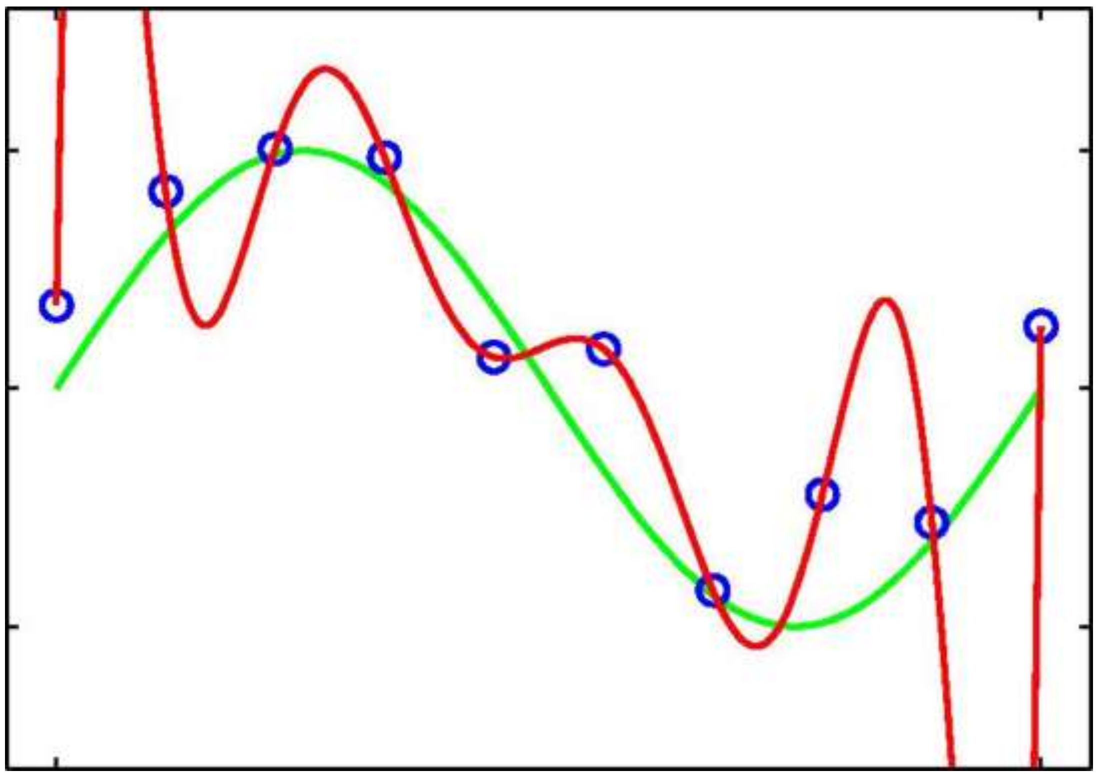
Training and generalization errors

- Training error
  - Error rate (e.g., $\|\varepsilon\|^2$ for OLS) derived from the training set ($x = [x_{i,j}]_{\substack{1 \le i \le n \\ 0 \le j \le k}}$) when estimating $\hat{\beta}$

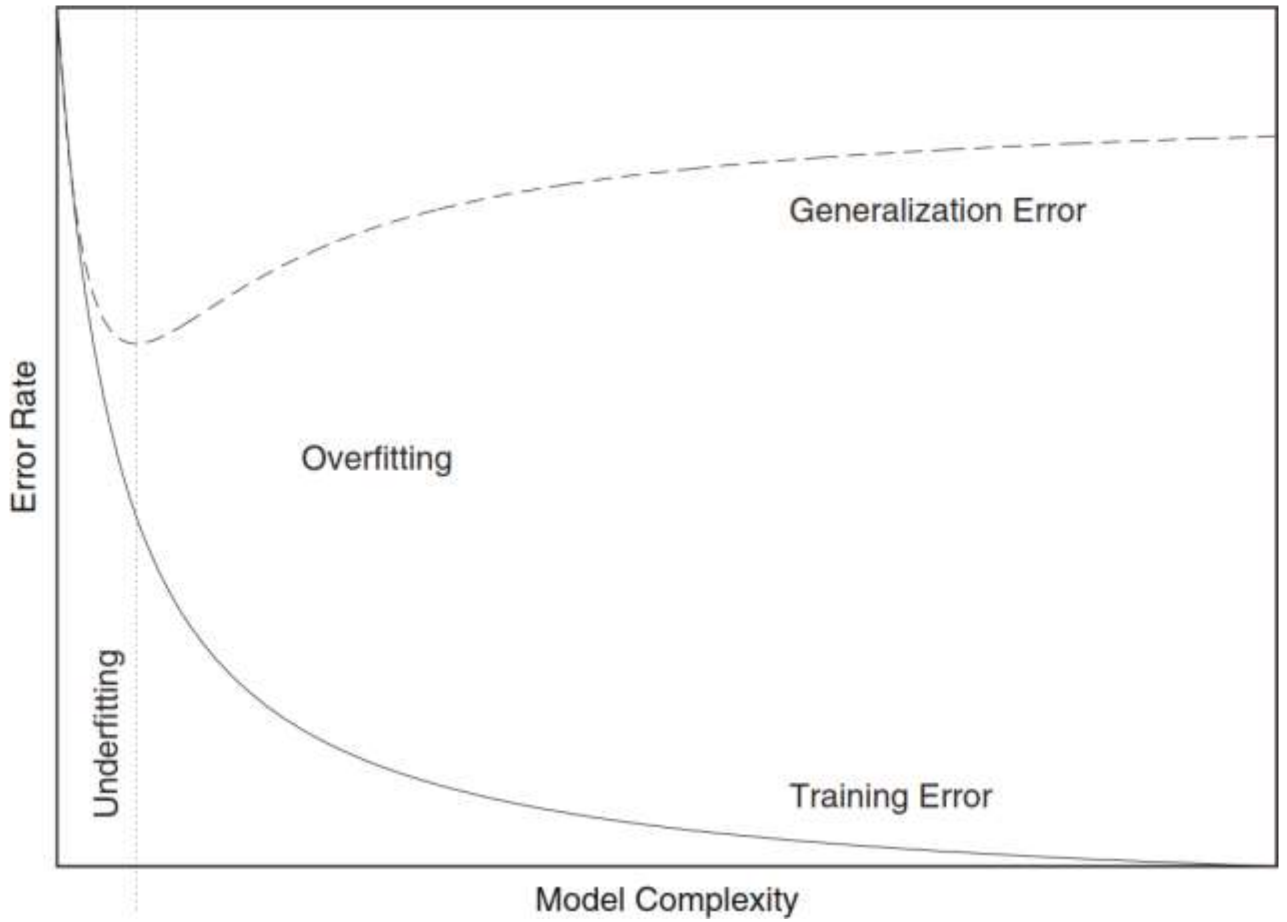
- Generalization error
  - Error rate when estimating $\hat{y}$ for unknown data points (data points that haven't been used to estimate $\hat{\beta}$)

How low can we push the training error?

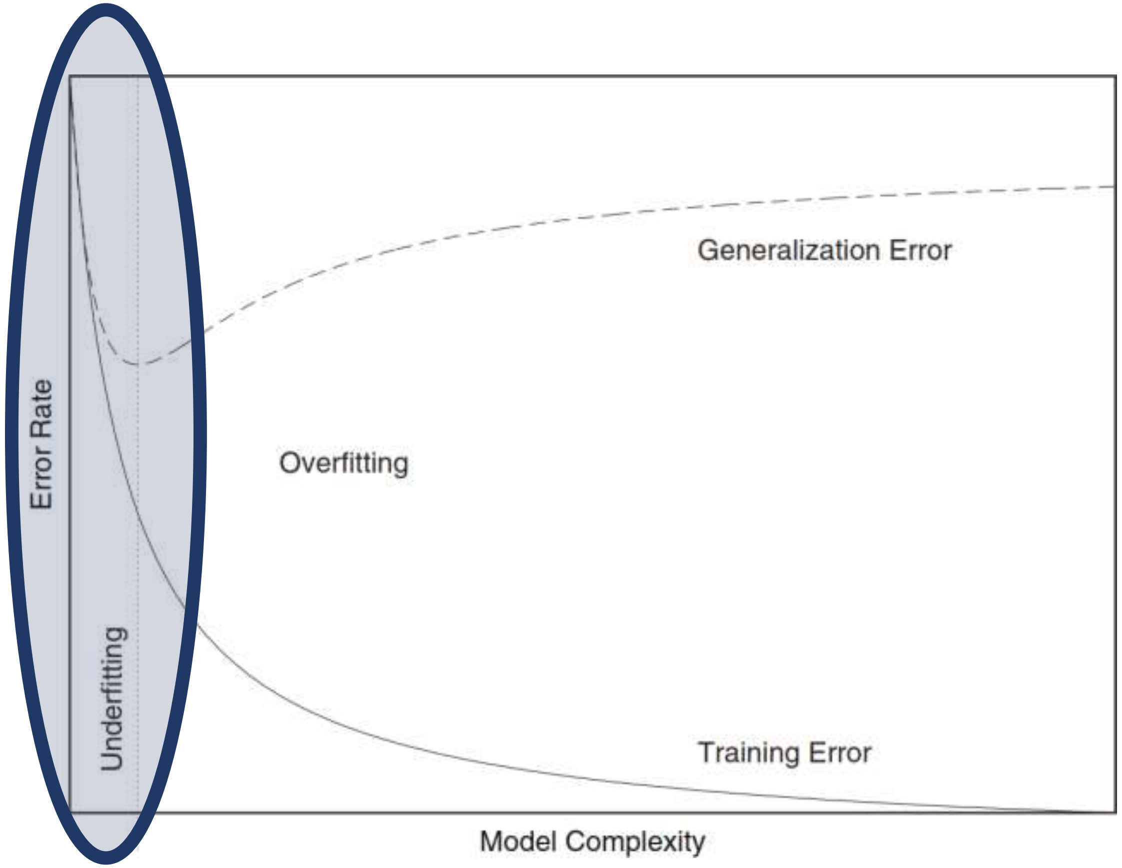
- Down to zero (effectively "memorizing" the entire training set)
- However, the model is now not only too complex but it will also not generalize well to data that was not used during training
  - This is called overfitting



Error rates, model complexity, and fit

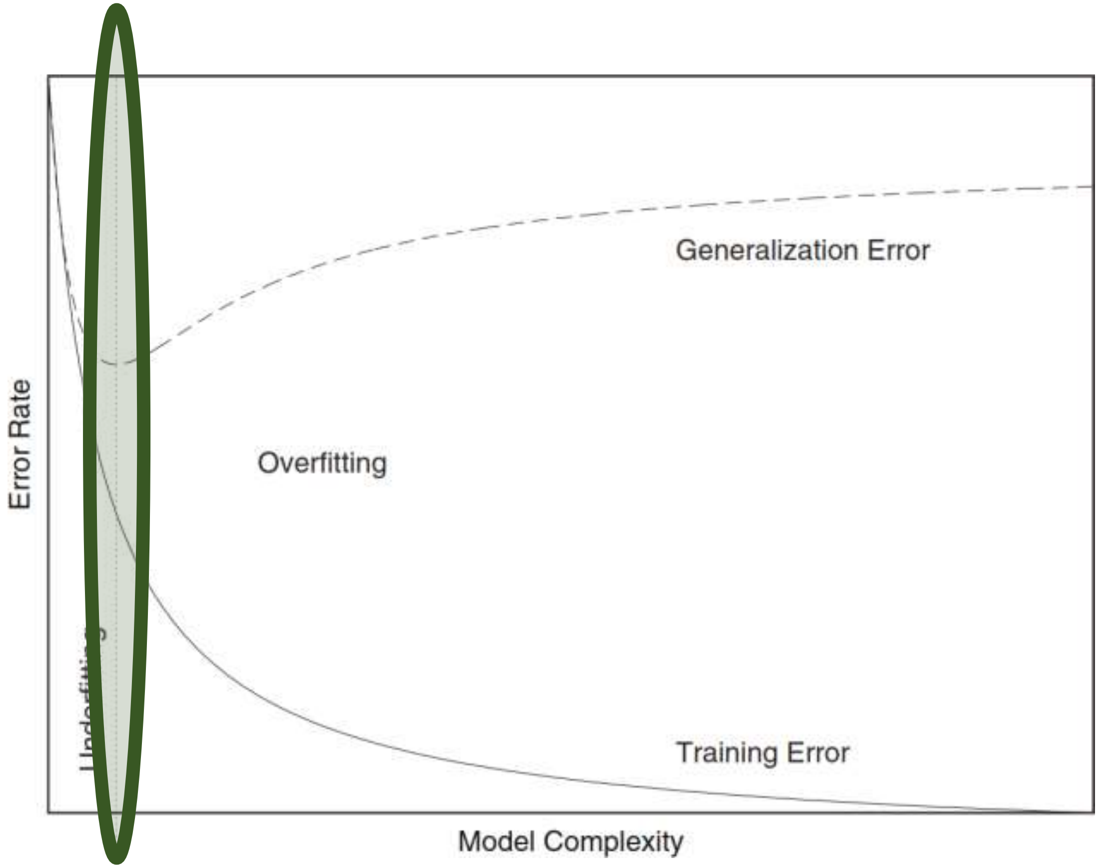


Error rates, model complexity, and fit (cont.)



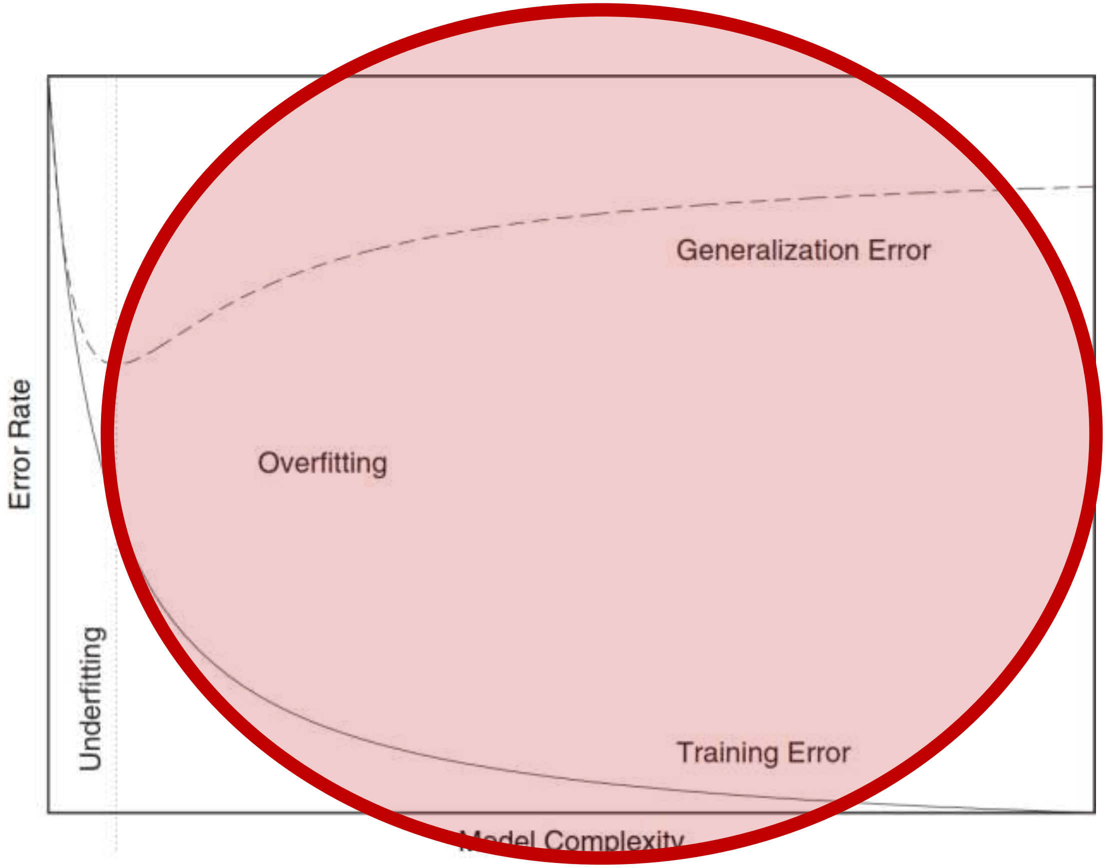
- Underfitting
  - Model is too simple and cannot represent the desired behavior very well
  - Both its training and generalization error are poor

Error rates, model complexity, and fit (cont.)



- Good fit
  - Model has the right level of complexity
  - It performs well on the training set (low training error) and generalize well to unknown data points (low generalization error)

Error rates, model complexity, and fit (cont.)



- Overfitting
  - Model is too complex
  - It performs very well on the training set (low training error) but does not generalize well to unknown data points (high generalization error)

Activity | Underfitting, good fit, and overfitting

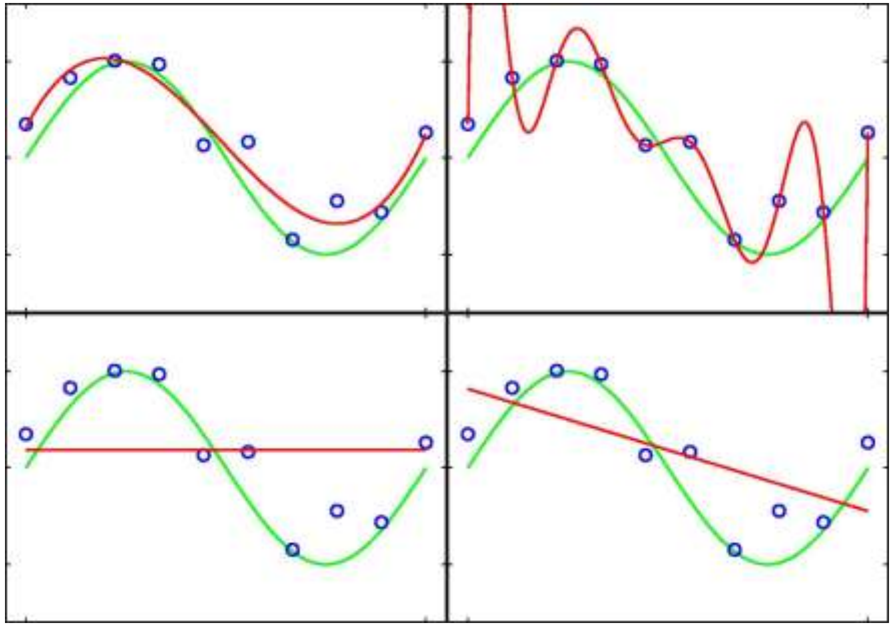
EXERCISE

DIRECTIONS (10 minutes)

1. Classify the following polynomial regressions according to their fit:
    1. Underfitting
    2. Good fit
    3. Overfitting
2. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions



Linear Regression

Complexity and Regularization

FYI | How do we define complexity?

- E.g., as a function of the size of the coefficients
  - $\|\beta\|_1 = \sum_{j=0}^{k} |\beta_j|$ (L1-norm)
  - $\|\beta\|_2^2 = \sum_{j=0}^{k} \beta_j^2$ (L2-norm)
  - (with $\beta = (\beta_0, \dots, \beta_k)$)

FYI | Regularization prevents overfitting by explicitly controlling model complexity

- These definitions of complexity lead to the following regularization techniques
  - $min\left(\underbrace{\|y - x \cdot \beta\|^2}_{OLS\ term} + \underbrace{\lambda\|\beta\|_1}_{regularization\ term}\right)$ (L1 regularization; a.k.a., Lasso)
  - $min(\|y - x \cdot \beta\|^2 + \lambda\|\beta\|_2^2)$ (L2 regularization; a.k.a., Ridge)
- (You will use the gradient descent technique discussed earlier to train your model)
- This formulation reflects the fact that there is a cost associated with regularization that we want to minimize

Linear Regression

Binary (a.k.a., Dummy) (Categorical) Variables

Back to the SF housing dataset and the issue of bed and bath counts

- So far, we've considered $BedCount$ and $BathCount$ as ratio variables
  - Namely that the price premium between a property with 1 bathroom and another with 2 bathrooms was the same between a property with 3 bathrooms and another with 4 bathrooms
- Does this make sense?

Dep. Variable:	SalePrice	**R-squared:**	0.137
Model:	OLS	**Adj. R-squared:**	0.136
Method:	Least Squares	**F-statistic:**	146.6
Date:		**Prob (F-statistic):**	1.94e-31
Time:		**Log-Likelihood:**	-1690.7
No. Observations:	929	**AIC:**	3385.
Df Residuals:	927	**BIC:**	3395.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	0.3401	0.099	3.434	0.001	0.146 0.535
BathCount	0.5242	0.043	12.109	0.000	0.439 0.609

Omnibus:	1692.623	**Durbin-Watson:**	1.582
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2167434.305
Skew:	12.317	**Prob(JB):**	0.00
Kurtosis:	238.345	**Cond. No.**	5.32

Back to the SF housing dataset and the issue of bed and bath counts

- Let's test this hypothesis and convert $BathCount$ to a nominal variable (indeed, we won't even assume an order) and then encode it to "dummy" categorical variables

m (# bathrooms)	$Bath = \begin{pmatrix} Bath_1, \\ Bath_2, \\ Bath_3, \\ Bath_4 \end{pmatrix}$ (encoding)
1	$(1, 0, 0, 0)$
2	$(0, 1, 0, 0)$
3	$(0, 0, 1, 0)$
4	$(0, 0, 0, 1)$

Activity | Binary (categorical) variables

EXERCISE

DIRECTIONS (10 minutes)

1. Complete the codealong by
    a. Run 4 regressions, one for each of the case highlighted in the handout (Each case only include 3 out of the 4 dummy variables we created) (we give you the first one…)
    b. What are the coefficients for the different βs?
    c. How do you interpret the βs?
    d. Why do we only need three dummy variables, not four?
2. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions

Activity | Binary (categorical) variables (cont.)

EXERCISE

$$SalePrice = \beta_1 + \beta_{1,2} \cdot Bath_2 + \beta_{1,3} \cdot Bath_3 + \beta_{1,4} \cdot Bath_4$$

(don't include $Bath_1$)

$$SalePrice = \beta_2 + \beta_{2,1} \cdot Bath_1 + \beta_{2,3} \cdot Bath_3 + \beta_{2,4} \cdot Bath_4$$

(don't include $Bath_2$)

$$SalePrice = \beta_3 + \beta_{3,1} \cdot Bath_1 + \beta_{3,2} \cdot Bath_2 + \beta_{3,4} \cdot Bath_4$$

(don't include $Bath_3$)

$$SalePrice = \beta_4 + \beta_{4,1} \cdot Bath_1 + \beta_{4,2} \cdot Bath_2 + \beta_{4,3} \cdot Bath_3$$

(don't include $Bath_4$)

Activity | Four linear regressions to run (cont.)

$$SalePrice = \beta_1 + \beta_{1,2} \cdot Bath_2 + \beta_{1,3} \cdot Bath_3 + \beta_{1,4} \cdot Bath_4$$

```
formula = 'SalePrice ~ Bath_2 + Bath_3 + Bath_4'
```

$$SalePrice = \beta_2 + \beta_{2,1} \cdot Bath_1 + \beta_{2,3} \cdot Bath_3 + \beta_{2,4} \cdot Bath_4$$

```
formula = 'SalePrice ~ Bath_1 + Bath_3 + Bath_4'
```

$$SalePrice = \beta_3 + \beta_{3,1} \cdot Bath_1 + \beta_{3,2} \cdot Bath_2 + \beta_{3,4} \cdot Bath_4$$

```
formula = 'SalePrice ~ Bath_1 + Bath_2 + Bath_4'
```

$$SalePrice = \beta_4 + \beta_{4,1} \cdot Bath_1 + \beta_{4,2} \cdot Bath_2 + \beta_{4,3} \cdot Bath_3$$

```
formula = 'SalePrice ~ Bath_1 + Bath_2 + Bath_3'
```

Activity | Four linear regressions to run (cont.)

Dep. Variable:	SalePrice	R-squared:	0.043
Model:	OLS	Adj. R-squared:	0.039
Method:	Least Squares	F-statistic:	11.78
Date:		Prob (F-statistic):	1.49e-07
Time:		Log-Likelihood:	-1314.2
No. Observations:	794	AIC:	2636.
Df Residuals:	790	BIC:	2655.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	0.9914	0.070	14.249	0.000	0.855 1.128
Bath_2	0.2831	0.099	2.855	0.004	0.088 0.478
Bath_3	0.4808	0.142	3.383	0.001	0.202 0.760
Bath_4	1.2120	0.232	5.231	0.000	0.757 1.667

Omnibus:	1817.972	Durbin-Watson:	1.867
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8069883.811
Skew:	19.917	Prob(JB):	0.00
Kurtosis:	495.280	Cond. No.	5.79

Dep. Variable:	SalePrice	R-squared:	0.043
Model:	OLS	Adj. R-squared:	0.039
Method:	Least Squares	F-statistic:	11.78
Date:		Prob (F-statistic):	1.49e-07
Time:		Log-Likelihood:	-1314.2
No. Observations:	794	AIC:	2636.
Df Residuals:	790	BIC:	2655.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	1.2745	0.071	18.040	0.000	1.136 1.413
Bath_1	-0.2831	0.099	-2.855	0.004	-0.478 -0.088
Bath_3	0.1977	0.143	1.386	0.166	-0.082 0.478
Bath_4	0.9290	0.232	4.003	0.000	0.473 1.384

Omnibus:	1817.972	Durbin-Watson:	1.867
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8069883.811
Skew:	19.917	Prob(JB):	0.00
Kurtosis:	495.280	Cond. No.	5.84

Dep. Variable:	SalePrice	R-squared:	0.043
Model:	OLS	Adj. R-squared:	0.039
Method:	Least Squares	F-statistic:	11.78
Date:		Prob (F-statistic):	1.49e-07
Time:		Log-Likelihood:	-1314.2
No. Observations:	794	AIC:	2636.
Df Residuals:	790	BIC:	2655.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	1.4722	0.124	11.881	0.000	1.229 1.715
Bath_1	-0.4808	0.142	-3.383	0.001	-0.760 -0.202
Bath_2	-0.1977	0.143	-1.386	0.166	-0.478 0.082
Bath_4	0.7313	0.253	2.886	0.004	0.234 1.229

Omnibus:	1817.972	Durbin-Watson:	1.867
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8069883.811
Skew:	19.917	Prob(JB):	0.00
Kurtosis:	495.280	Cond. No.	7.52

Dep. Variable:	SalePrice	R-squared:	0.043
Model:	OLS	Adj. R-squared:	0.039
Method:	Least Squares	F-statistic:	11.78
Date:		Prob (F-statistic):	1.49e-07
Time:		Log-Likelihood:	-1314.2
No. Observations:	794	AIC:	2636.
Df Residuals:	790	BIC:	2655.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	2.2035	0.221	9.969	0.000	1.770 2.637
Bath_1	-1.2120	0.232	-5.231	0.000	-1.667 -0.757
Bath_2	-0.9290	0.232	-4.003	0.000	-1.384 -0.473
Bath_3	-0.7313	0.253	-2.886	0.004	-1.229 -0.234

Omnibus:	1817.972	Durbin-Watson:	1.867
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8069883.811
Skew:	19.917	Prob(JB):	0.00
Kurtosis:	495.280	Cond. No.	11.7

Activity | What are the βs' coefficient? (cont.)

β_1 0.9914		$\beta_{1,2}$ 0.2831	$\beta_{1,3}$ 0.4808	$\beta_{1,4}$ 1.212
β_2 1.2745	$\beta_{2,1}$ –0.2831		$\beta_{2,3}$ 0.1977	$\beta_{2,4}$ 0.9290
β_3 1.4722	$\beta_{3,1}$ –0.4808	$\beta_{3,2}$ –0.1977		$\beta_{3,4}$ 0.7313
β_4 2.2025	$\beta_{4,1}$ –1.212	$\beta_{4,2}$ –0.9290	$\beta_{4,3}$ –0.7313	

Activity | What are the βs' coefficient? (cont.)

β_i	Value (Sale's price) of a property in SF with i bathrooms
$\beta_{i,j}$ when $j > i$	Increase of value for a property when increasing the number of bathrooms from i to j (while keeping the rest of the same…)
$\beta_{i,j}$ when $j < i$	Decrease of value for a property when decreasing the number of bathrooms from i to j (while keeping the rest of the same…)
$\beta_{i,j} = -\beta_{j,i}$	Going from i to *j* bathrooms has the opposite effect of going from *j* bathrooms to *i* bathrooms
$\beta_j = \beta_i + \beta_{i,j}$ for any i and j	E.g., $\beta_4 = \beta_1 + \beta_{1,4}$. I.e., the value of a 4 bathrooms can be derived from a 1 bedroom house and by increasing the number of bathrooms for 1 to 4
$\beta_{i,j} = \beta_{i,k} + \beta_{k,j}$ for any i, j and k	E.g., $\beta_{1,4} = \beta_{1,2} + \beta_{2,4}$. I.e., the increase in value from a 1 bathroom house to a 4 bathrooms house is identical to going from upgrading from 1 bathroom to 2 bathrooms and then from upgrading from 2 bathrooms to 4 bathrooms

Review

Review

- Linear Regressions
  - Simple and Multiple
  - Regression assumptions; how to check for them
- Variables
  - Variable Transformations; dummy categorical variables; Interaction effects and the hierarchy principle
  - How to interpret the model's parameters
- Inference and Fit
  - F-statistic
  - R^2 (r-squared), and $\bar{R}^2$ (adjusted R^2)
- Guidance on how to conduct linear regression modeling
  - Backward selection
- Estimating the βs and model complexity
  - OLS (Ordinary Least Squares)
  - Underfitting and overfitting, training and generalization errors, and regularization

Review

You should now be able to:

- How to conduct linear regression modeling
- Use interaction effects and dummy categorical variables
- Understand model complexity, underfitting, right fit, and overfitting
- Define regularization and error metrics for regression problems

DS

Q & A

Next Class

Introduction to Classification

Learning Objectives

After the next lesson, you should be able to:

- Define class label and classification
- Build a K-Nearest Neighbors using the *sklearn* library
- Evaluate and tune model by using metrics such as classification accuracy/error

Exit Ticket

Don't forget to fill out your exit ticket here