Feature Engineering and Polynomial Regression

Data Science Immersive

FLATIRON SCHOOL

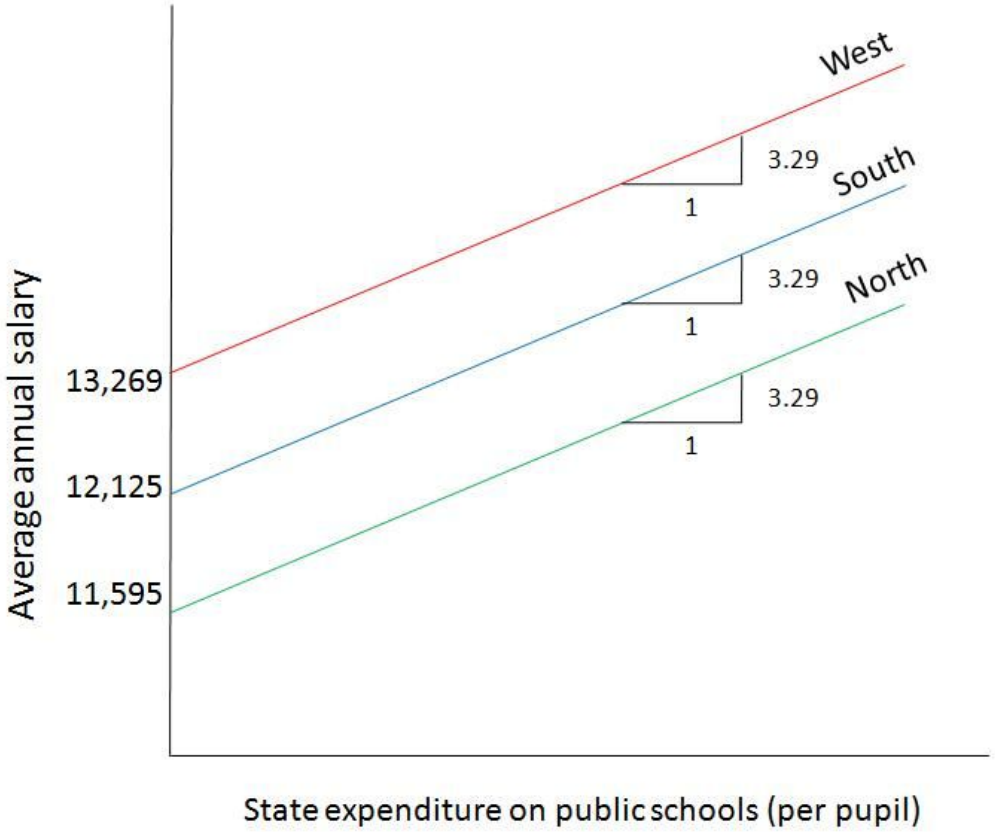
What We'll Cover

- Isolating the effect of certain variables
- Scaling variables - when, why, and how?
- Transforming variables
- Interpreting more complex regression results

Revisiting Multiple Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots$$

Categorical/Dummy Vars Refresher



Confusing Words

- Covariance - How much does this variable go up or down with another variable?

$$cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

- Correlation - standardized covariance - always takes a value in the interval (-1,1)

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

- Collinearity - correlation to a very high degree
- Multicollinearity - many-to-one or many-to-many correlation to a very high degree

Interaction Terms

$$ex1 : y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 x$$

$$ex2 : y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 x + \beta_4 d_1 d_2$$

$$ex3 : y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 x + \beta_4 d_1 d_2 + \beta_5 d_1 x + \beta_6 d_2 x + \beta_7 d_1 d_2 x$$

Example Regression Results

Method D (Raw variables and orthogonal products)

Predictor	Coefficient	Std. error	t	p	VIF	SEQ SS
Constant	43.134	7.516	5.74	0.000		
P1	0.76401	0.08419	9.08	0.000	1.3	10096.1
G	-20.846	1.732	-12.04	0.000	1.1	7908.0
K	2.015	1.902	1.06	0.293	1.2	116.7
S	8.358	1.853	4.51	0.000	1.2	1087.0
G.S	-22.819	4.108	-5.55	0.000	1.4	2129.0
G.K	-7.308	3.797	-1.92	0.058	1.1	295.6
S.K	2.208	4.852	0.46	0.650	1.6	62.2
G.S.K	1.542	9.862	0.16	0.876	1.6	11.0
P1.G	0.2262	0.1794	1.26	0.211	1.4	122.4
P1.K	0.1848	0.1777	1.04	0.302	1.4	51.9
P1.S	-0.1366	0.1826	-0.75	0.457	1.4	61.8
P1.G.S	0.0113	0.3811	0.03	0.976	1.4	12.6
P1.G.K	-0.4236	0.3813	-1.11	0.270	1.6	49.6
P1.S.K	-0.2912	0.4111	-0.71	0.481	1.4	30.1
P1.G.S.K	0.0772	0.8429	0.09	0.927	1.0	0.5

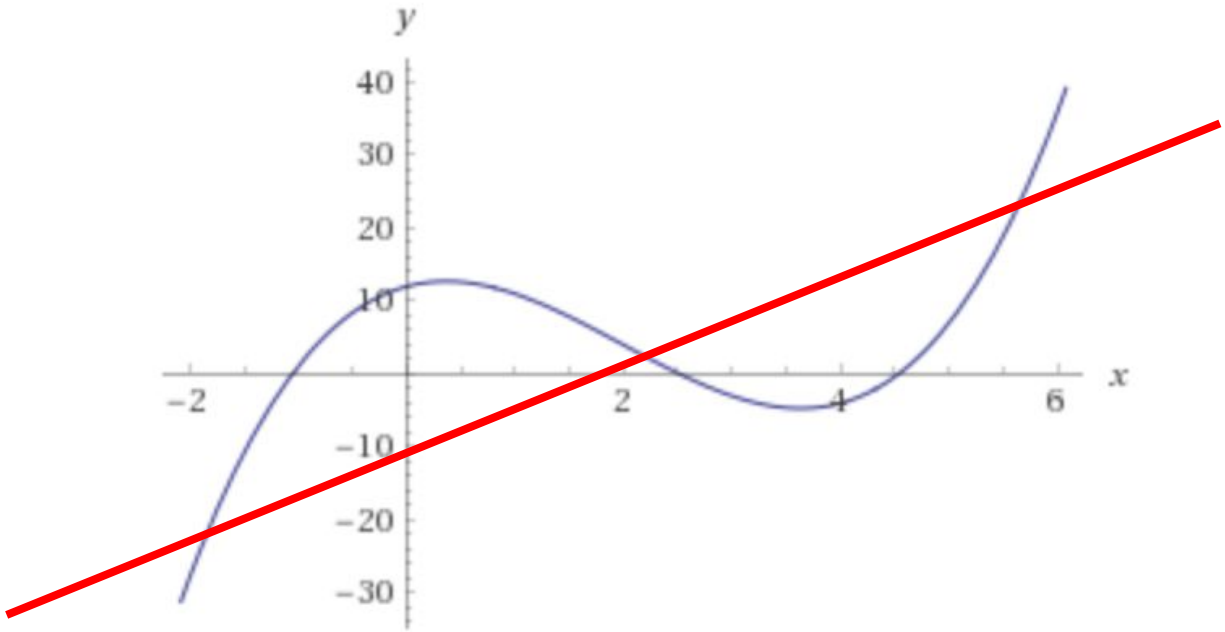
Linear Feature Scaling

	Bivariate	Model 1	Model 2	Model 3	Model 4
Work capacity relative to physical demands (very bad 0 to very good 10)	0.031**	0.028*	0.026*	0.030*	0.029*
Work capacity relative to psychological demands (very bad 0 to very good 10)	0.000	–0.018	–0.026+	–0.028*	–0.014
Working evenings or weekends (Never 1 to many times a week 5)	0.063***	0.057***	0.057***	0.057***	0.048***
Direct contact with customers/ clients (no/yes)	0.157**	0.100*	0.086+	0.080	0.097+
Freedom to decide work facets (scale)	0.020*	0.010	0.008	0.010	0.003
Support from management (no/yes)	0.076*	0.069*	0.046	0.051	0.018
Support from workmates (no/yes)	0.100*	0.009	0.005	–0.012	–0.021
Sports (practice) during working hours (no/yes)	0.156**	0.135*	0.133*	0.129*	0.095*

- Min/max scaling
- Mean Normalization
- Unit vector transformation

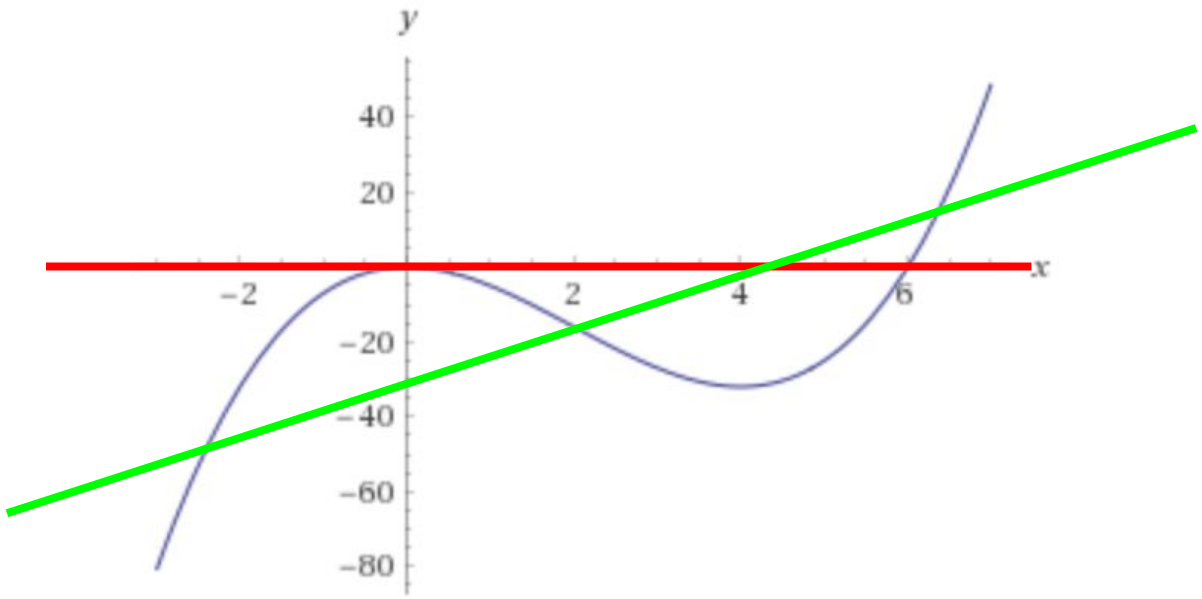
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Polynomial Regression



$$y = x^3 - 6x^2 + 4x + 12$$

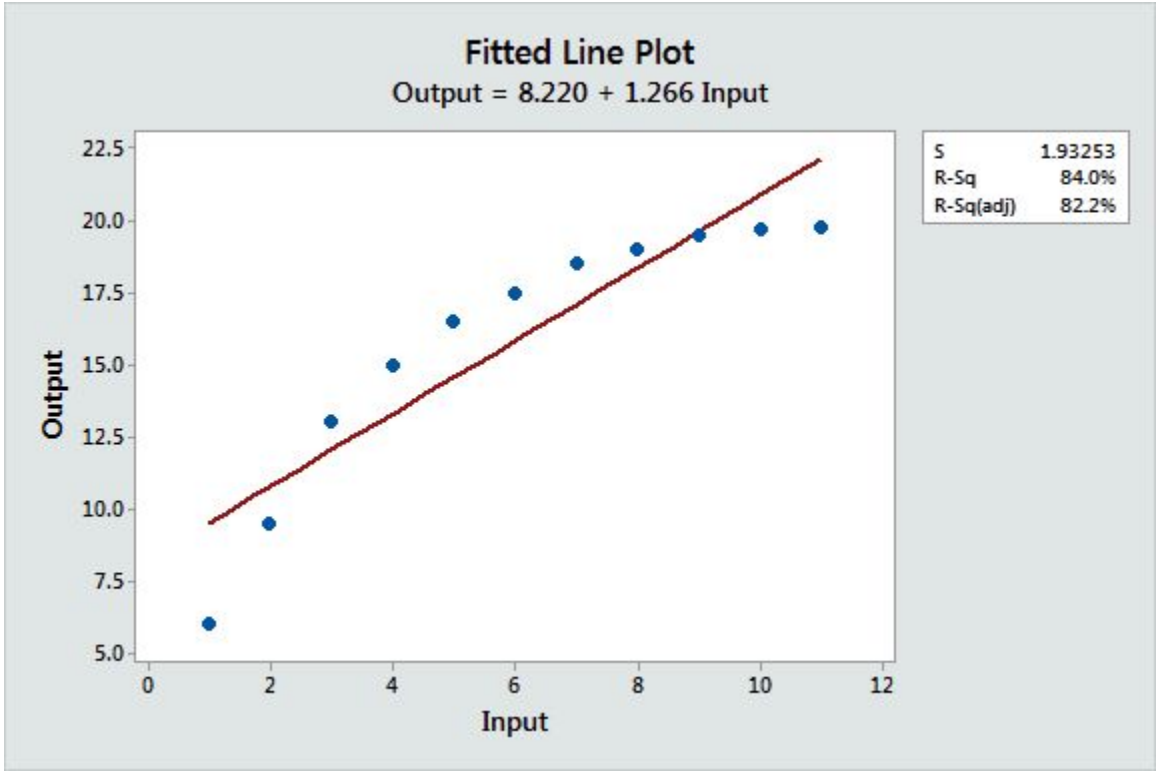
Polynomial Regression



$$y = x^3 - 6x^2$$

Polynomial Regression

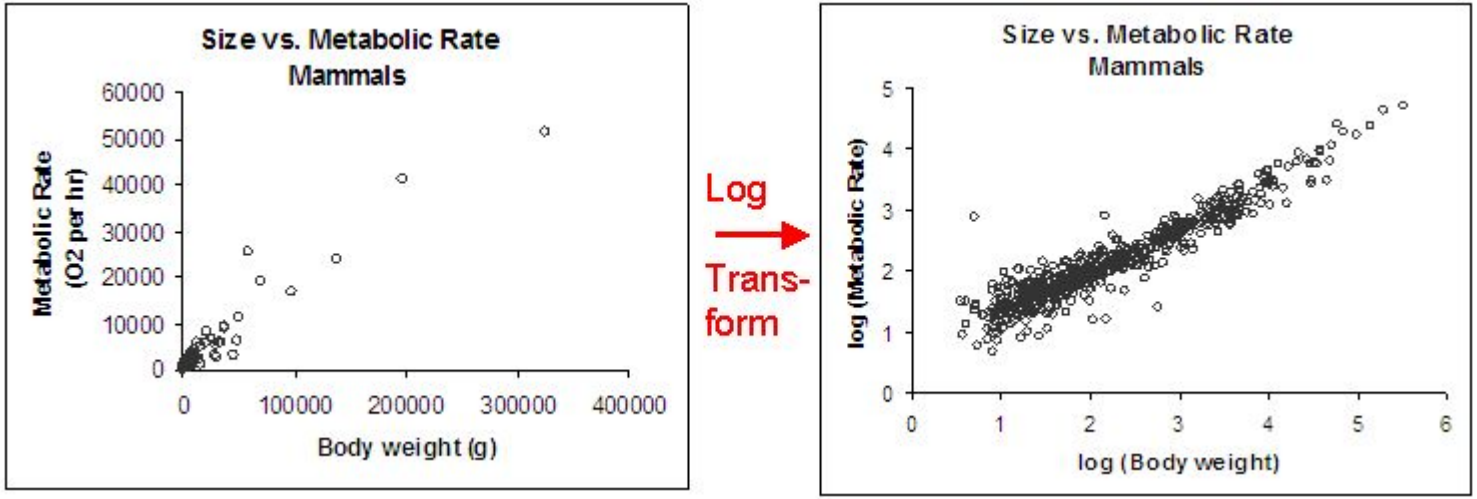
BAD:



Polynomial Regression

$$y_i = \beta_0 + \sum_{n=1}^{n_{max}} \beta_n x_i^n + \epsilon_i$$

Non-Linear, Non-Polynomial Transformations



Non-Linear, Non-Polynomial Transformations

Non-Linear, Non-Polynomial Transformations

- Log
- Absolute value
- Square root
- Exponent
- Logit/Probit for probability (more on this when we get to logistic regression)

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Tips for Transformations

- Make sure you understand how the transformation you applied changes the interpretation of your model $\frac{\delta y}{\delta x}$
- You can use more than one transformation
- There are no precise answers here - consult literature and use your intuition. Often times making decisions involves facing trade-offs. Visuals help too.
- Use several iterations of your model. No need to stick with just one. Often times researchers publish papers with 6 or so versions of the same regression.

Complex Regression Example

Table 2: Analysis on FDI and decisive factors in economic growth

	I	II	III	IV	V	VI
Constant	1.647101 (1.250171) (0.2198)	-16.18851** (-3.467541) (0.0179)	-29.36245 (-1.384852) (0.2383)	-40.68387 (-1.663888) (0.1947)	-32.86707 (-0.603803) (0.6073)	-101.2295 (-1.824503) (0.3192)
LNFGDP	3.738556** (3.804272) (0.0006)	11.57055** (4.367251) (0.0072)	22.06650 (1.324348) (0.2560)	26.63039 (1.522692) (0.2252)	22.18091 (0.661060) (0.5765)	50.39892 (1.740583) (0.3320)
H		0.922310* (2.108799) (0.0888)	2.407076 (1.015970) (0.3671)	4.118271 (1.379889) (0.2615)	2.914309 (0.368658) (0.7477)	7.643710 (1.221569) (0.4367)
H*FGDP			-0.406880 (-0.639158) (0.5575)	-0.563615 (-0.849564) (0.4580)	-0.417425 (-0.355612) (0.7561)	-1.333304 (-1.346708) (0.4066)
Yo				-0.002582 (-0.959532) (0.4081)	-0.001807 (-0.323815) (0.7768)	-0.002197 (-0.552850) (0.6785)
TGDP					0.012671 (0.171428) (0.8797)	0.043050 (0.776504) (0.5797)
EXR						0.130862 (1.719483) (0.3353)
R2	0.298571	0.833484	0.848915	0.884394	0.886068	0.971205
Adjusted R2	0.277941	0.766878	0.735601	0.730253	0.601239	0.798434
Prob(F-statistic)	0.000565	0.011315	0.040580	0.091451	0.260960	0.312116
Sample size	36	8	8	8	8	8

Working With Polynomial Regression