ISyE 6739 –Regression Supplementary Topics (Chapters 11 & 12)

Instructor: Kamran Paynabar
H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Tech

Kamran.paynabar@isye.gatech.edu
Office: Groseclose 436

Polynomial Regression

The linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is a general model that can be used to fit any relationship that is **linear in the unknown parameters** $\boldsymbol{\beta}$. This includes the important class of **polynomial regression models.** For example, the second-degree polynomial in one variable

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon \tag{12-46}$$

and the second-degree polynomial in two variables

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon \tag{12-47}$$

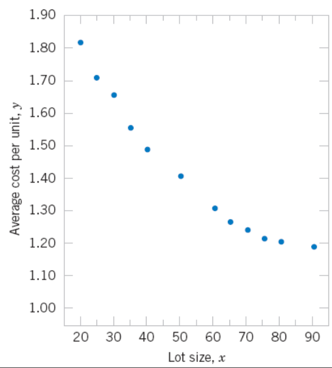
are linear regression models.

Example: Polynomial Regression

EXAMPLE 12-12 Airplane Sidewall Panels

Sidewall panels for the interior of an airplane are formed in a 1500-ton press. The unit manufacturing cost varies with the production lot size. The data shown below give the average cost per unit (in hundreds of dollars) for this product (y) and the production lot size (x). The scatter diagram, shown in Fig. 12-11, indicates that a second-order polynomial may be appropriate.

y	1.81	1.70	1.65	1.55	1.48	1.40
x	20	25	30	35	40	50
y	1.30	1.26	1.24	1.21	1.20	1.18
x	60	65	70	75	80	90



Example: Polynomial Regression

We will fit the model

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$$

The $\mathbf{y}$ vector, the model matrix $\mathbf{X}$ and the $\boldsymbol{\beta}$ vector are as follows:

$$\mathbf{y} = \begin{bmatrix} 1.81 \\ 1.70 \\ 1.65 \\ 1.55 \\ 1.48 \\ 1.40 \\ 1.30 \\ 1.26 \\ 1.24 \\ 1.21 \\ 1.20 \\ 1.18 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 20 & 400 \\ 1 & 25 & 625 \\ 1 & 30 & 900 \\ 1 & 35 & 1225 \\ 1 & 40 & 1600 \\ 1 & 50 & 2500 \\ 1 & 60 & 3600 \\ 1 & 65 & 4225 \\ 1 & 70 & 4900 \\ 1 & 75 & 5625 \\ 1 & 80 & 6400 \\ 1 & 90 & 8100 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_{11} \end{bmatrix}$$

Example: Polynomial Regression

Solving the normal equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ gives the fitted model

$$\hat{y} = 2.19826629 - 0.02252236x + 0.00012507x^2$$

Conclusions: The test for significance of regression is shown in Table 12-13. Since $f_0 = 1762.3$ is significant at 1%, we conclude that at least one of the parameters β_1 and β_{11} is not zero. Furthermore, the standard tests for model adequacy do not reveal any unusual behavior, and we would conclude that this is a reasonable model for the sidewall panel cost data.

Table 12-13 Test for Significance of Regression for the Second-Order Model in Example 12-12

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	0.52516	2	0.26258	1762.28	2.12E-12
Error	0.00134	9	0.00015		
Total	0.5265	11			

Regression with Categorical Predictors

- Many problems may involve qualitative or categorical variables.
- The usual method for the different levels of a qualitative variable is to use indicator variables.
- For example, to introduce the effect of two different operators into a regression model, we could define an indicator variable as follows:

$$x = \begin{cases} 0 \text{ if the observation is from operator 1} \\ 1 \text{ if the observation is from operator 2} \end{cases}$$

How about variables with 3 levels or more?

Example: Categorical Predictors

EXAMPLE 12-13 Surface Finish

A mechanical engineer is investigating the surface finish of metal parts produced on a lathe and its relationship to the speed (in revolutions per minute) of the lathe. The data are shown in Table 12-15. Note that the data have been collected using two different types of cutting tools. Since the type of cutting tool likely affects the surface finish, we will fit the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where Y is the surface finish, x_1 is the lathe speed in revolutions per minute, and x_2 is an indicator variable denoting the type of cutting tool used; that is,

$$x_2 = \begin{cases} 0, & \text{for tool type 302} \\ 1, & \text{for tool type 416} \end{cases}$$

Example: Categorical Predictors

Table 12-15 Surface Finish Data for Example 12-13

Observation Number, i	Surface Finish y_i	RPM	Type of Cutting Tool	Observation Number, i	Surface Finish y_i	RPM	Type of Cutting Tool
1	45.44	225	302	11	33.50	224	416
2	42.03	200	302	12	31.23	212	416
3	50.10	250	302	13	37.52	248	416
4	48.75	245	302	14	37.13	260	416
5	47.92	235	302	15	34.70	243	416
6	47.79	237	302	16	33.92	238	416
7	52.26	265	302	17	32.13	224	416
8	50.52	259	302	18	35.47	251	416
9	45.58	221	302	19	33.49	232	416
10	44.78	218	302	20	32.29	216	416

Example: Categorical Predictors

The parameters in this model may be easily interpreted. If $x_2 = 0$, the model becomes

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

which is a straight-line model with slope β_1 and intercept β_0. However, if $x_2 = 1$, the model becomes

$$Y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \epsilon = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon$$

which is a straight-line model with slope β_1 and intercept $\beta_0 + \beta_2$. Thus, the model $Y = \beta_0 + \beta_1 x + \beta_2 x_2 + \epsilon$ implies that surface finish is linearly related to lathe speed and that the slope β_1 does not depend on the type of cutting tool used. However, the type of cutting tool does affect the intercept, and β_2 indicates the change in the intercept associated with a change in tool type from 302 to 416.

The fitted model is

$$\hat{y} = 14.27620 + 0.14115x_1 - 13.28020x_2$$

$$\mathbf{X} = \begin{bmatrix} 1 & 225 & 0 \\ 1 & 200 & 0 \\ 1 & 250 & 0 \\ 1 & 245 & 0 \\ 1 & 235 & 0 \\ 1 & 237 & 0 \\ 1 & 265 & 0 \\ 1 & 259 & 0 \\ 1 & 221 & 0 \\ 1 & 218 & 0 \\ 1 & 224 & 1 \\ 1 & 212 & 1 \\ 1 & 248 & 1 \\ 1 & 260 & 1 \\ 1 & 243 & 1 \\ 1 & 238 & 1 \\ 1 & 224 & 1 \\ 1 & 251 & 1 \\ 1 & 232 & 1 \\ 1 & 216 & 1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 45.44 \\ 42.03 \\ 50.10 \\ 48.75 \\ 47.92 \\ 47.79 \\ 52.26 \\ 50.52 \\ 45.58 \\ 44.78 \\ 33.50 \\ 31.23 \\ 37.52 \\ 37.13 \\ 34.70 \\ 33.92 \\ 32.13 \\ 35.47 \\ 33.49 \\ 32.29 \end{bmatrix}$$

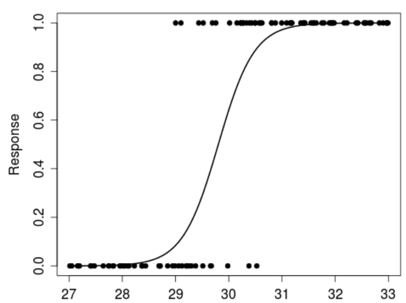
Logistic Regression

Assume Y is a binary variable following a Bernoulli distribution

$$Y \sim Bernoulli(p) \quad y = 0,1 \qquad E(y) = p$$

$$P(x) = p^y(1-p)^{1-y};\; y = 0,1 \qquad Var(y) = p(1-p)$$

We want to fit a regression model for y against x.



$$: \frac{1}{1 + e^{-(\beta_0 + x \cdot \beta)}}$$

Logistic Regression

In logistic regression E(Y | X) is be written as

$$E(y|x) = p(x) = \frac{1}{1+\exp[-(\beta_0+\beta_1 x)]}$$

The MLE method can be used to estimate the parameters of the model

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1-p(x_i))^{1-y_i}$$

The Log likelihood can be written as

$$\begin{aligned}
\ell(\beta_0, \beta) &= \sum_{i=1} y_i \log p(x_i) + (1-y_i)\log 1 - p(x_i) \\
&= \sum_{i=1}^{n} \log 1 - p(x_i) + \sum_{i=1}^{n} y_i \log \frac{p(x_i)}{1-p(x_i)} \\
&= \sum_{i=1}^{n} \log 1 - p(x_i) + \sum_{i=1}^{n} y_i(\beta_0 + x_i \cdot \beta) \\
&= \sum_{i=1}^{n} -\log 1 + e^{\beta_0 + x_i \cdot \beta} + \sum_{i=1}^{n} y_i(\beta_0 + x_i \cdot \beta)
\end{aligned}$$

No closed-form solutions. Numerical method is used for optimizing the log likelihood function.

Example: Logistic Regression

We will illustrate logistic regression using the data on launch temperature and O-ring failure for the 24 space shuttle launches prior to the *Challenger* disaster of January 1986. There are six O-rings used to seal field joints on the rocket motor assembly. The table below presents the launch temperatures. A 1 in the "O-Ring Failure" column indicates that at least one O-ring failure had occurred on that launch.

Temperature	O-Ring Failure	Temperature	O-Ring Failure	Temperature	O-Ring Failure
53	1	68	0	75	0
56	1	69	0	75	1
57	1	70	0	76	0
63	0	70	1	76	0
66	0	70	1	78	0
67	0	70	1	79	0
67	0	72	0	80	0
67	0	73	0	81	0

Example: Logistic Regression

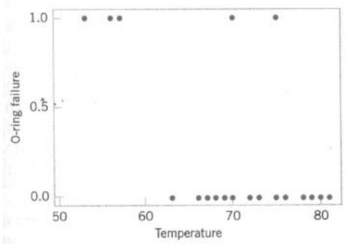


Figure 11-20 Scatter plot of O-ring failures versus launch temperature for 24 space shuttle flights.

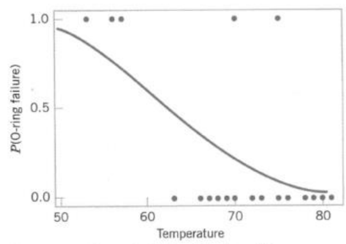


Figure 11-21 Probability of O-ring failure versus launch temperature (based on a logistic regression model).

Binary Logistic Regression: O-Ring Failure versus Temperature

Link Function: Logit

Response Information

Variable	Value	Count	
O-Ring F	1	7	(Event)
	0	17	
	Total	24	

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	Upper
Constant	10.875	5.703	1.91	0.057			
Temperat	−0.17132	0.08344	−2.05	0.040	0.84	0.72	0.99

Log-Likelihood = −11.515

Test that all slopes are zero: G = 5.944, DF = 1, P-Value = 0.015