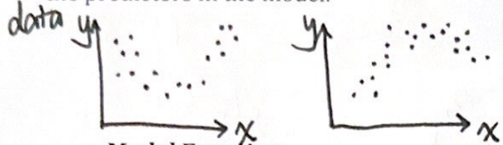
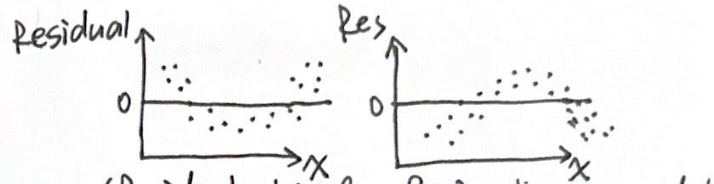
Quadratic Regression

A quadratic regression allows us to fit a parabola - a curve instead of a line. If we see a curved pattern in the data and/or the residuals, it may be helpful to add a quadratic term to the model x^2 where x is one of the predictors in the model.





(Residual plot after fitting linear model)

- **Model Equation:**

$$y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

- **Assumptions:**

$$\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$$

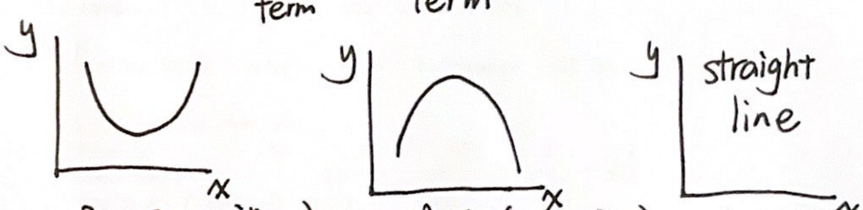
- **Fitted Equation:**

$$\hat{y} = a + b_1 X + b_2 X^2$$

constant — linear term — quadratic term

y: Quant Response
X: Quant Predictor
<u>Note</u>: $p = 2$, not 1 (X^2 ~~is~~ also counts as a predictor)

- **Graphs:**



$\beta_2 > 0$ (positive) $\quad$ $\beta_2 < 0$ (negative) $\quad$ $\beta_2 = 0$

- **Interpretation of Coefficients in the Fitted Model:**

a: constant term — Do not interpret

b_1: linear term — Do not interpret

b_2: quadratic term — Is β_2 sig diff from zero?
Look at p-val for β_2

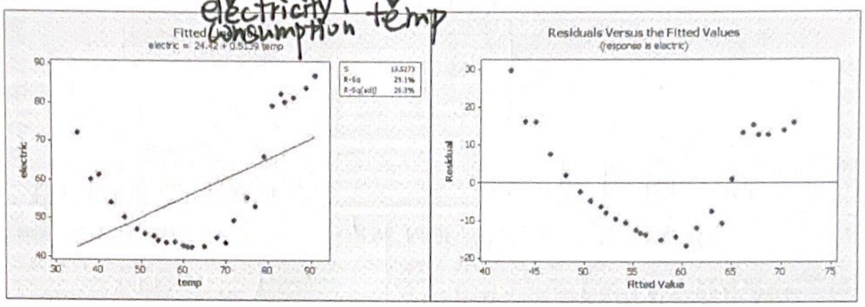
If small
↓
Look at the sign of b_2
"+" opens up
"−" opens down

If large
↓
β_2 NOT sig diff from zero, a straight line would be better.

Example: Suppose that we want to predict electricity consumption in dollars (y) based on average monthly high temperature (x) for one particular house. Data on $n=27$ months.

Linear Regression $y = \alpha + \beta_1 X + \varepsilon$

electricity consumption temp




```
Regression Analysis: electric versus temp
The regression equation is
electric = 24.4 + 0.514 temp

Predictor    Coef  SE Coef     T      P
Constant    24.42    10.57  2.31  0.029
temp       0.5139   0.1603  3.21  0.004

S = 13.5273   R-Sq = 29.1%   R-Sq(adj) = 26.3%

Analysis of Variance
Source           DF      SS      MS      F      P
Regression        1  1880.7  1880.7  10.28  0.004
Residual Error   25  4574.7   183.0
Total            26  6455.5

Unusual Observations
Obs  temp  electric    Fit  SE Fit  Residual  St Resid
  1  35.0     72.16  42.40    5.32     29.76     2.39R
R denotes an observation with a large standardized residual.
```

- Is temperature a good predictor of electrical consumption?

  Yes, p-val = 0.04 (t-test and ANOVA test p-vals agree with each other for SLR)

- Predict electrical consumption for months when the average high temperature is 50F.

  electric = 24.4 + 0.514 (50) = 50.1

- Is the model appropriate?

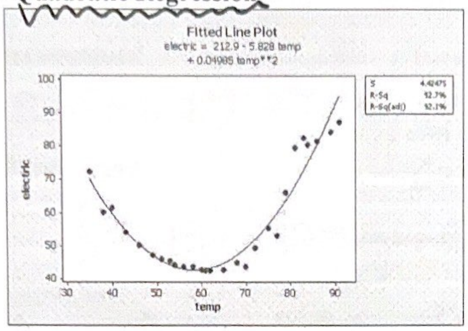
  Not appropriate.

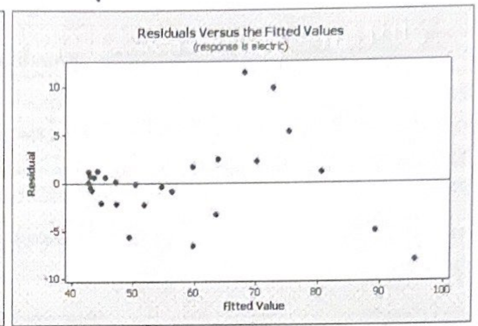
  Both graph of data and residual plot shows a curved pattern, which prompts us to add a quadratic ~~pattern~~ term x^2 (or temp^2) to the model.

electricity consumption temp temp^2

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Quadratic Regression





Residual plot pretty random (no clear pattern) Good

```
Regression Analysis: electric versus temp, temp2

The regression equation is
electric = 213 - 5.83 temp + 0.0499 temp**2
```

Predictor	Coef	SE Coef	T	P
Constant	212.93	13.47	~~15.81~~	~~0.000~~
temp	-5.8278	0.4411	~~-13.21~~	~~0.000~~
temp**2	0.049854	0.003443	14.48	0.000

S = 4.42475   R-Sq = 92.7%   R-Sq(adj) = $\boxed{92.1\%}$

Curve fits data much better than straight line

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	5985.6	2992.8	152.86	0.000
Residual Error	24	469.9	19.6		
Total	26	6455.5			

Source	DF	Seq SS
temp	1	1880.7
temp2	1	4104.8

Unusual Observations

Obs	temp	electric	Fit	SE Fit	Residual	St Resid
1	35.0	72.164	70.032	2.582	2.132	0.59 X
22	81.0	79.468	67.974	1.243	11.494	2.71R
23	83.0	82.469	72.671	1.369	9.798	2.33R
27	91.0	87.265	95.445	2.356	-8.180	-2.18R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

- **Model**

electricity consumption, temp, $temp^2$

$$y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

- **Assumptions**

$\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$ $\xrightarrow{\text{Check}}$
- Random errors/elect/months → Are months randomly selected? Don't know.
- Normal errors/elect → No outliers. look OK
- Constant variance of errors/elect → pts look evenly distributed around curve.

- **Fitted Equation**

$$\hat{y} = 213 - 5.83\,\underset{(X)}{temp} + 0.0499\,\underset{(X^2)}{temp^2}$$

- **ANOVA test**

Now that we have two predictors, ANOVA and t-tests are different. We do ANOVA test first b/c it is an overall test for the question: Are there any good predictor in model?

$H_0: \beta_1 = \beta_2 = 0$ H_a: at least one $\beta_i \neq 0$ $(i=1,2)$

TS: $F = 152.86$ p-val = 0.000 → Rej H_0, very strong evidence of at least one good predictor.

- **t tests**

α: constant — don't care

β_1: temp — don't care

β_2: $temp^2$ — Yes: $H_0: \beta_2 = 0$ $H_a: \beta_2 \neq 0$

TS: $t = 14.48$ p-val = 0.000 → Rej H_0, very strong evidence that the quadratic term $temp^2$ is a good predictor of elect consumption.

- **Is the Quadratic model better than the SLR model?**

<u>Yes</u>. ① p-val for $temp^2$ is tiny, so the newly-included quadratic term is a good predictor

② $R^2_{adj} = 26.3\%$ (linear) 92.1% (quadratic) [$R^2 = 29.1\%$ (linear) 92.7% (quadratic)]

- **Interpret R2 for the Quadratic model**

$$R^2 = \frac{SSR}{SST} = \frac{5985.6}{6455.5} = 92.7\%$$

<u>Interpretation</u>: 92.7% of variability in elect. consumption (y) is explained by the quadratic regression model (on temp and $temp^2$)

(Note: R^2 not longer equals r^2 b/c this is NOT SLR)

- **Compute the residual for observation 27.**

Recall: Residual = Obs y − Pred y

For observation 27, X = temp = 91.0 and y = elect = 87.265 (where to find them?)

$$\text{Pred } y = \hat{y}\Big|_{X=91.0} = 213 - 5.83(91) + 0.0499(91)^2 = 95.445$$

$$\text{Residual} = 87.265 - 95.445 = -8.180$$

[The actual elect. consumption in that month was \$8.18 <u>less</u> than the predicted]