Polynomial Regression

Tutorial 9: Lecture 21

Ravleen Bajaj

Today:

- What is Polynomial Regression?
- Motivation: Polynomial Regression
- Example 1: cars
- Example 2: yield
- Maximizer
- Key Takeaways

Polynomial Regression

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i$$

for some p = 0, 1, 2, ...

- ϵ_i's are independent and $\epsilon_i \sim N(0, \sigma^2)$.
- The mean of the response is a polynomial function of the predictor(s).
- p is called the degree of the polynomial
- By convention, all lower-order terms are included in the model

Motivation: Polynomial Regression

- When a car's speed doubles, its stopping distance **more than doubles**.
- Physics: Kinetic Energy $= \frac{1}{2}mv^2$
- Stopping distance is roughly proportional to v^2.
- A simple linear model under-predicts the distance at higher speeds.

Key Question Can a linear model capture the true (quadratic) relationship between speed and mean stopping distance?

The Cars Dataset

```
data(cars)
head(cars)
```

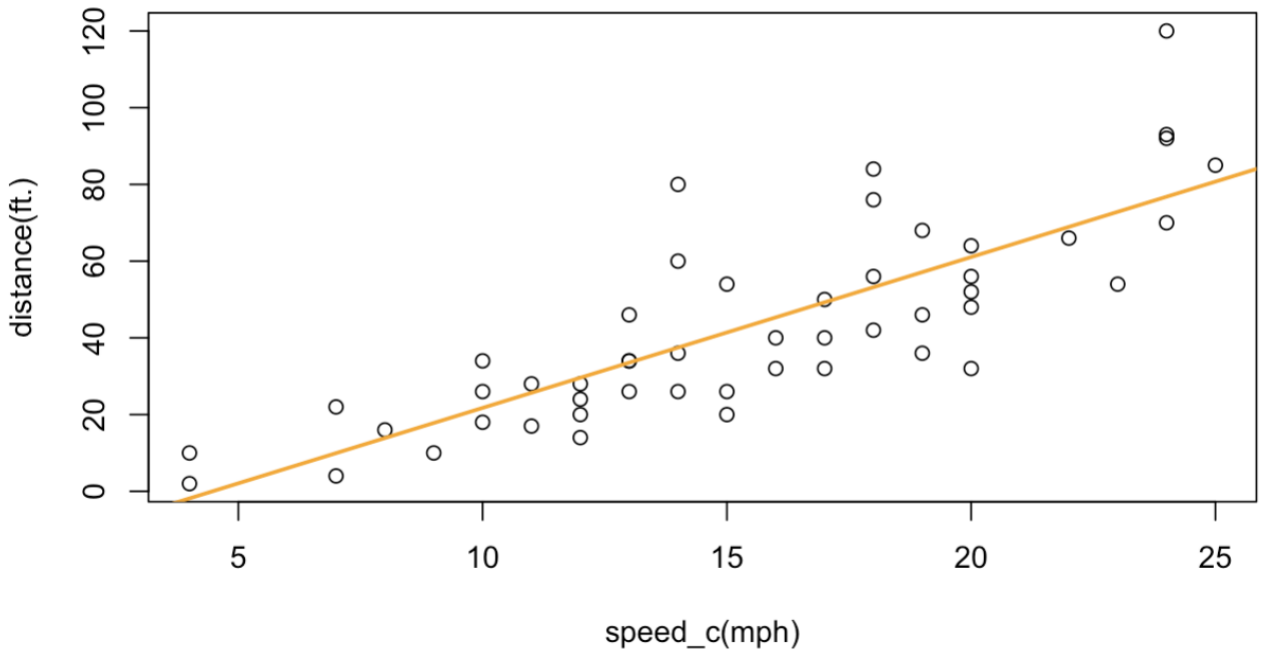
Speed	Distance
4	2
4	10
7	4
7	22
8	16
9	10

- `speed`: car speed (mph)
- `dist`: stopping distance (ft)

Fit a Linear Model

```
m_linear <- lm(dist ~ speed, data = cars)
plot(cars)
abline(m_linear, col = "red", lwd = 2)
```

- The slope represents the estimated mean increase in stopping distance per unit speed.



Add a Quadratic Term

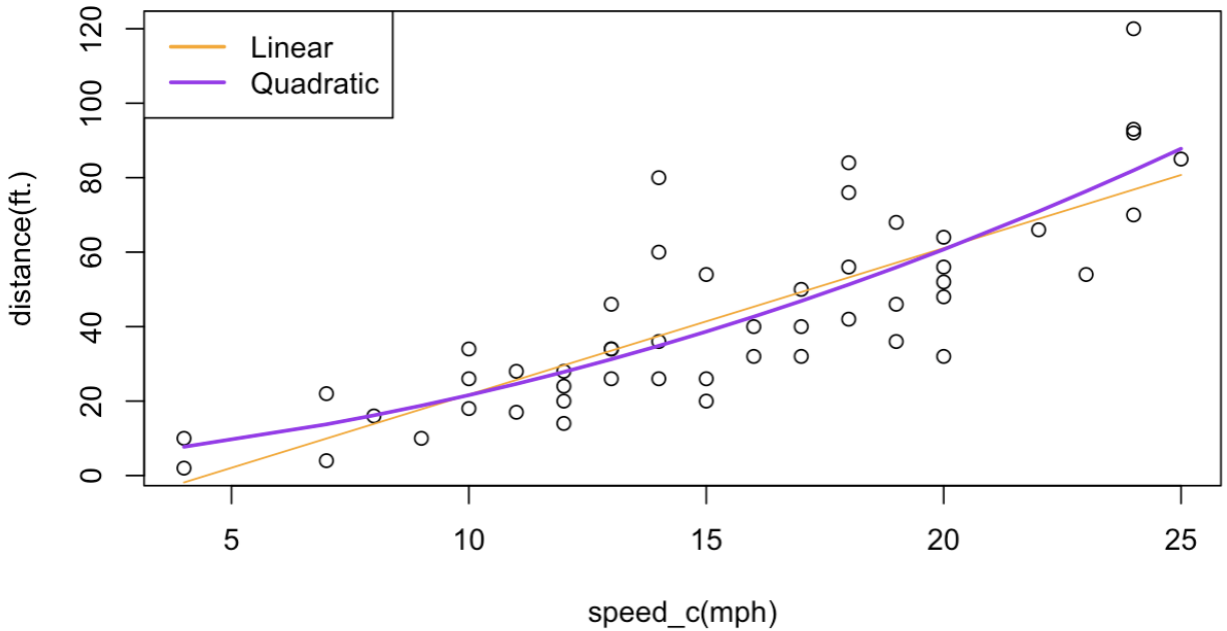
```
m_quadratic <- lm(dist ~ speed + I(speed^2), data = cars)
summary(m_quadratic)
```

- The speed^2 term captures acceleration.

```
Call:
lm(formula = dist ~ speed + I(speed^2), data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-28.720  -9.184  -3.188   4.628  45.152

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.47014   14.81716   0.167    0.868
speed        0.91329    2.03422   0.449    0.656
I(speed^2)   0.09996    0.06597   1.515    0.136

Residual standard error: 15.18 on 47 degrees of freedom
Multiple R-squared:  0.6673,    Adjusted R-squared:  0.6532
F-statistic: 47.14 on 2 and 47 DF,  p-value: 5.852e-12

    speed I(speed^2)
 24.61489   24.61489
```

Add a Quadratic Term

- The plots appear identical under both models, but we know based on underlying physics and statistical significance that the quadratic term is significant.
- Coefficient of the quadratic term determines the curvature.



Multicollinearity: Speed vs Speed^2

- The predictors `speed` and `speed`^{2} are highly correlated.
- This inflates standard errors and makes coefficient estimates unstable.

Variance Inflation Factor (VIF)

```
vif(m_quadratic)
speed       I(speed^2)
24.61489    24.61489
```

High VIF values indicate multicollinearity.

Centring to Reduce Collinearity

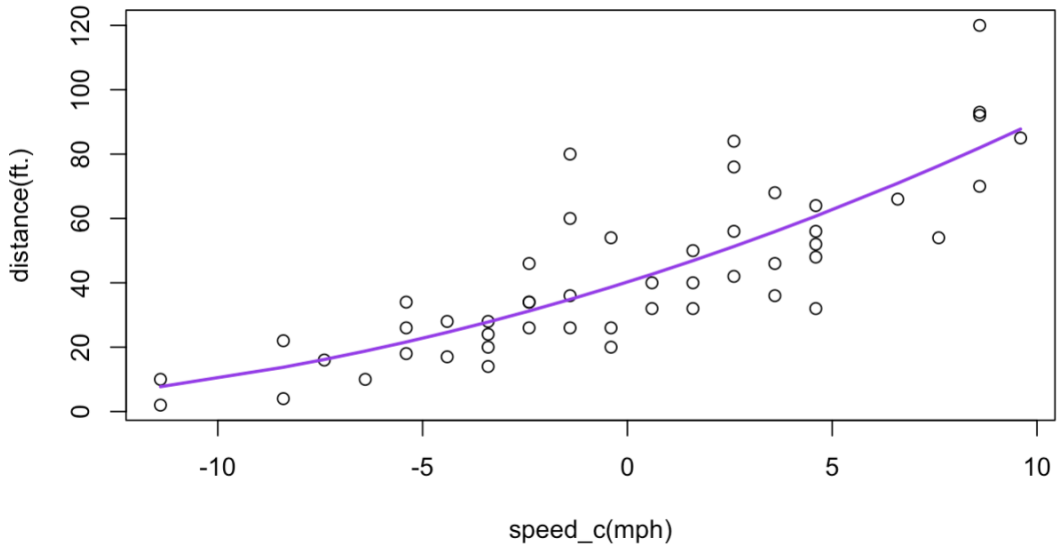
```
cars$speed_c <- scale(cars$speed, center = TRUE,
scale = FALSE)
m_centered <- lm(dist ~ speed_c + I(speed_c^2), data = cars)
```

- **Centering:** Subtract the average x from each measurement and use this adjusted predictor instead of x in the quadratic model.
- Centering reduces correlation between x_c and x_c^2.
- We now use this adjusted predictor instead of speed in the quadratic model.

Centring to Reduce Collinearity

```
vif(m_centered)
     speed_c       I(speed_c^2)
     1.009211      1.009211
```

- VIFs decrease!

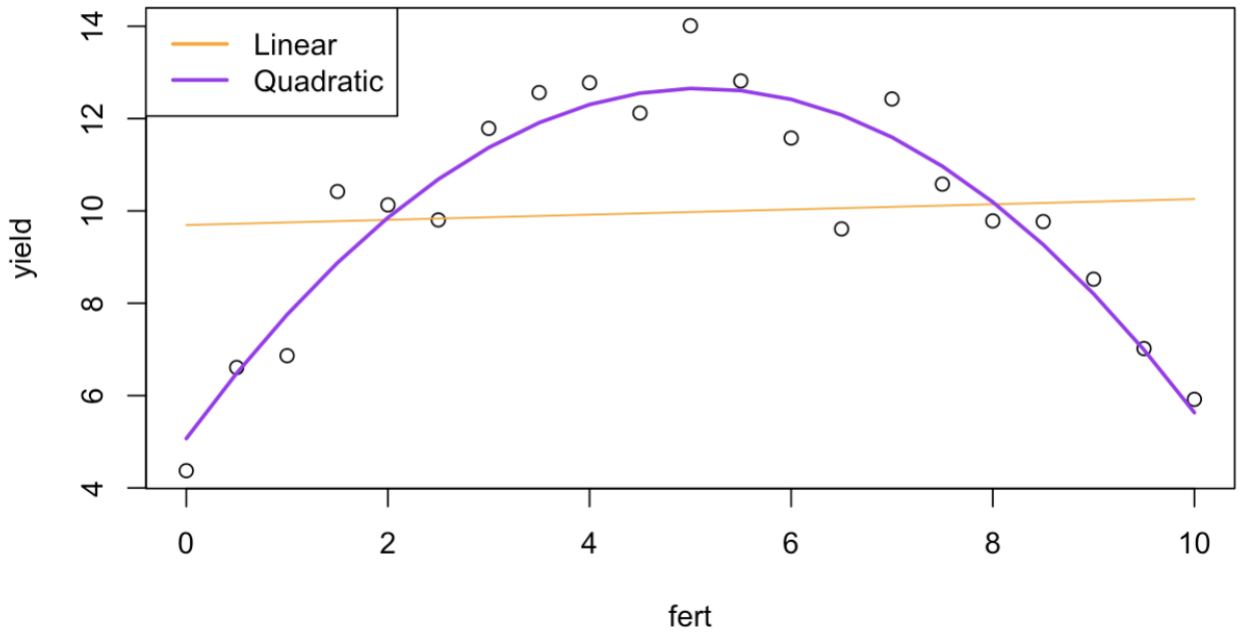


An Inverted Parabola Example: Fertilizer vs Crop Yield

- Too little fertilizer $\rightarrow$ low yield.
- Too much fertilizer $\rightarrow$ toxicity or waste.
- Relationship follows a **concave quadratic curve**.

```
set.seed(1)
fert <- seq(0, 10, by = 0.5)
yield <- 5 + 3*fert - 0.3*fert^2 + rnorm(length(fert), 0, 1)
plot(fert, yield)
l_yield <- lm(yield ~ fert)
m_yield <- lm(yield ~ fert + I(fert^2))

lines(fert, fitted(l_yield), col = "orange")
lines(fert, fitted(m_yield), col = "purple", lwd = 2)
legend("topleft", c("Linear", "Quadratic"), col = c("orange",
```

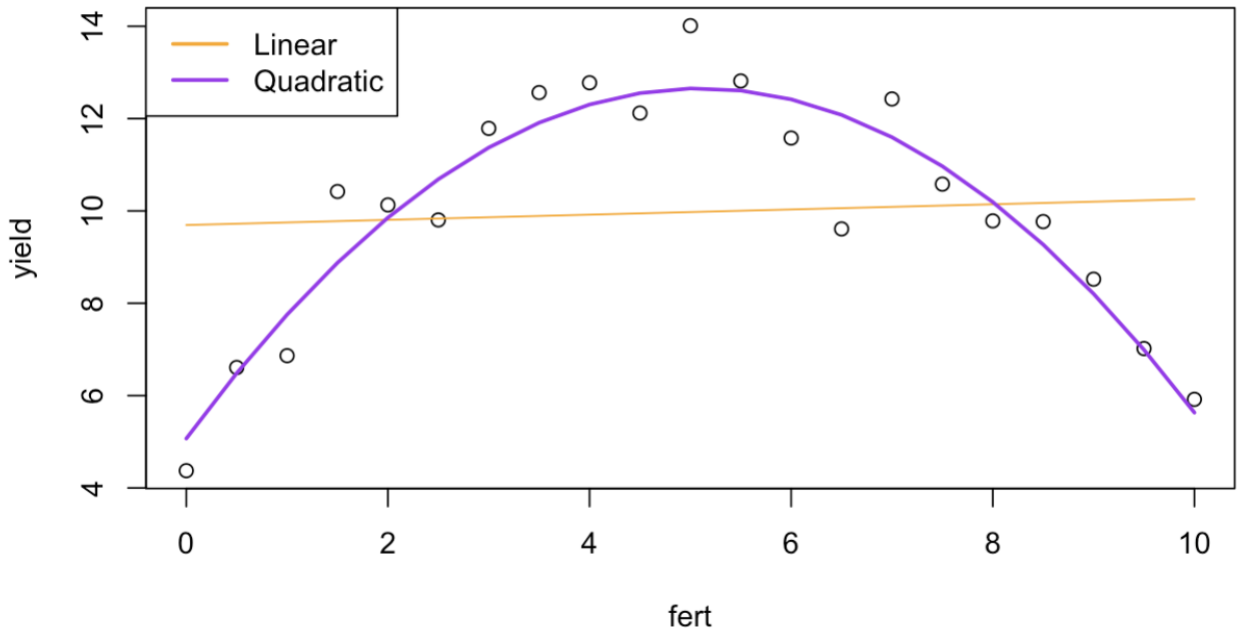
Inverted Parabola Example: Fertilizer vs Crop Yield




```
vif(m_yield)
fert       I(fert^2)
14.72998   14.72998
```

- VIFs are high.

Inverted Parabola Example: Fertilizer vs Crop Yield

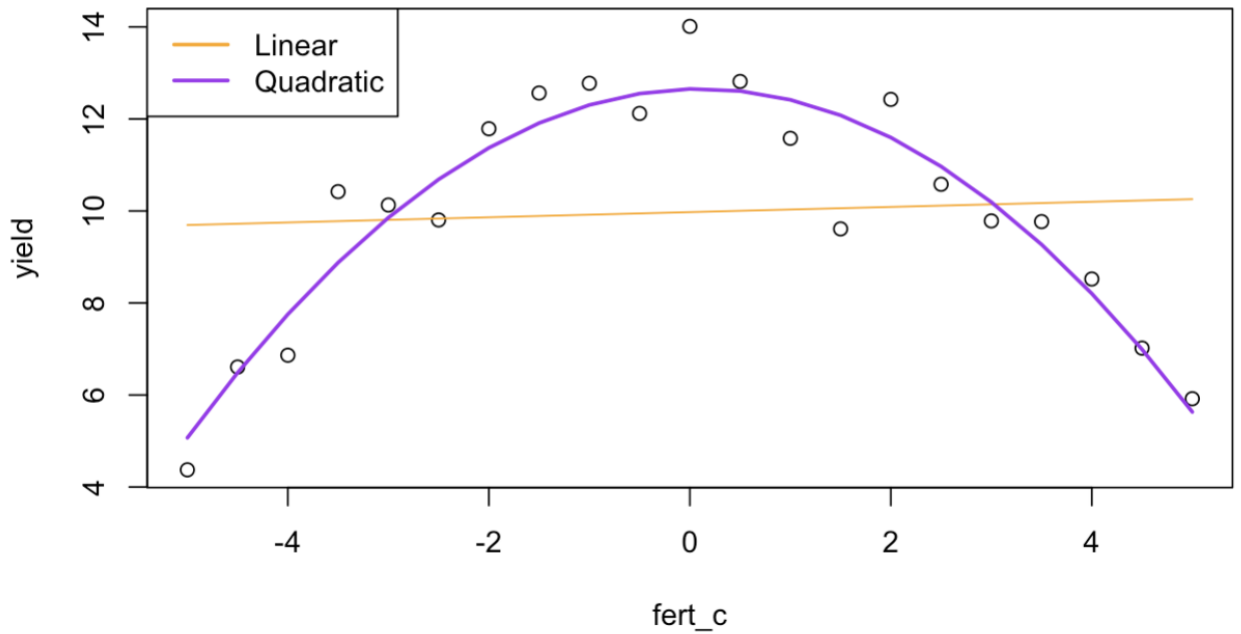



```
vif(m_yield)
fert       I(fert^2)
14.72998   14.72998
```

- VIFs are high.

Inverted Parabola Example: Fertilizer vs Crop Yield

```
fert_c <- fert - mean(fert)
m_centered2 <- lm(yield ~ fert_c + I(fert_c^2))
plot(fert_c, yield)
lines(fert_c, fitted(l_yield), col = "orange")
lines(fert_c, fitted(m_centered2), col = "purple", lwd = 2)
legend("topleft", c("Linear", "Quadratic"), col = c("orange",
```



Inverted Parabola Example: Fertilizer vs Crop Yield

```
vif(m_centered2)
fert_c      I(fert_c^2)
1           1
```

- VIFs decrease!

```
F1 <- matrix(fert_c, nrow=21, ncol =1)
F2 <- matrix(fert_c^2, nrow=1, ncol =21)

F <- F2 %*% F1
print(F)

0
```

Columns of X are orthogonal now, as implied by VIFs being equal to 1.

Finding the Maximizer

For a fitted quadratic model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

The value of x that maximizes $\hat{y}$ is:

$$x^* = -\frac{\hat{\beta}_1}{2\hat{\beta}_2}$$

Example Estimated centered fertilizer amount that gives maximum yield:

```
coef(m_centered2)[2] / (2 * coef(m_centered2)[3])
-0.09620358
```

Key Takeaways

- Quadratic polynomials capture curvature missed by linear regression models.
- Consider centering if multicollinearity is causing a problem.
- Use VIF to diagnose multicollinearity.
- We can find the maximum value of a fitted quadratic function and the value of the predictor where this maximum is achieved.

Thank You!

Questions?