Lecture 11:
Polynomial Regression
STAT 432, Spring 2021

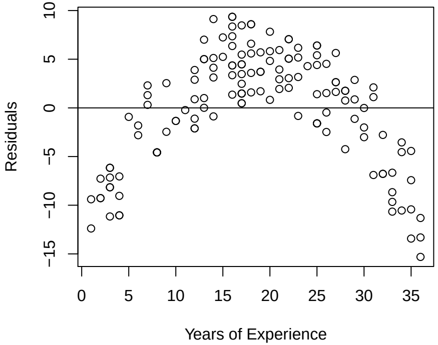
Salary Data Set

- For this example we consider a salary data set with $n = 143$ observations and two variables.
- We want to develop a regression model between y, salary (in thousands of dollars), and x, the number of years of experience. We are interested in using the model to make predictions and prediction intervals.
- Since the variables have a nonlinear, quadratic association we consider a polynomial regression model.

```
> profsalary <- read.csv("https://ericwfox.github.io/data/profsalary.csv")
> lm1 <- lm(Salary ~ Experience, data = profsalary)
> plot(Salary ~ Experience, data = profsalary,
       ylab = "Salary", xlab = "Years of Experience")
> abline(lm1)
```

Fitting a straight line obviously does not capture the trend in the data.

```
> plot(profsalary$Experience, resid(lm1),
      xlab = "Years of Experience", ylab = "Residuals")
> abline(h=0)
```



Quadratic Polynomial Regression Model

Since a quadratic relationship is evident, we consider the following polynomial regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

where y = salary, x = years of experience, and $\epsilon \sim N(0, \sigma^2)$ is the random error.

```
> lm2 <- lm(Salary ~ Experience + I(Experience^2), data=profsalary)
> summary(lm2)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     34.720498   0.828724   41.90   <2e-16 ***
Experience       2.872275   0.095697   30.01   <2e-16 ***
I(Experience^2) -0.053316   0.002477  -21.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.817 on 140 degrees of freedom
Multiple R-squared:  0.9247,Adjusted R-squared:  0.9236
F-statistic: 859.3 on 2 and 140 DF,  p-value: < 2.2e-16
```

Fitted quadratic regression model:

$$\hat{y} = 34.720 + 2.872x - 0.053x^2$$

Prediction when $x = 10$:

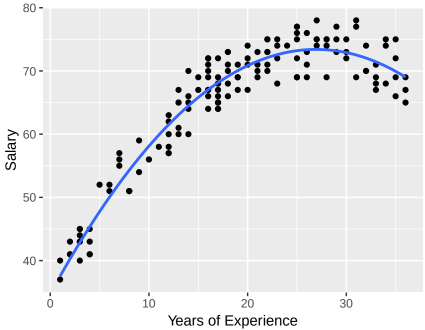
$$\hat{y} = 34.720 + 2.872(10) - 0.053(10^2) = 58.14$$

Using R:

```
> new_x <- data.frame(Experience = 10)
> predict(lm2, newdata = new_x, interval = "prediction")
       fit      lwr      upr
1 58.11164 52.50481 63.71847
```

The graphics package ggplot2 provides a convenient way to visualize the quadratic regression model.

```
library(ggplot2)
ggplot(profsalary, aes(x = Experience, y = Salary)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), se=F) +
  xlab("Years of Experience") + ylab("Salary")
```



The residual plot for the quadratic regression model shows no trend and the points are randomly scattered around zero. Thus, the conditions for regression appear satisfied.

```
> plot(profsalary$Experience, resid(lm2),
      xlab = "Years of Experience", ylab = "Residuals")
> abline(h=0)
```

Polynomial Regression (in general)

In general, a polynomial regression model of degree p can be written as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \epsilon$$

- One way to choose p is to keep adding terms until the added (highest order) term is no longer significant.
- It is recommended to keep all lower order terms in the model, even if they are not statistically significant. For example, if we fit a quadratic model, then we should keep the x term in the model.

We can consider a third degree (cubic) polynomial regression model for the salary data set.

```
> lm3 <- lm(Salary ~ Experience + I(Experience^2) + I(Experience^3), data=profsalary)
> summary(lm3)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     35.8319735  1.1579671  30.944   <2e-16 ***
Experience       2.5406068  0.2602417   9.762   <2e-16 ***
I(Experience^2) -0.0313332  0.0162368  -1.930   0.0557 .
I(Experience^3) -0.0003957  0.0002888  -1.370   0.1730
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.808 on 139 degrees of freedom
Multiple R-squared:  0.9257,Adjusted R-squared:  0.9241
F-statistic: 577.1 on 3 and 139 DF,  p-value: < 2.2e-16
```

However, the coefficient for the cubic term is not significant (p-value > 0.05), so there is no improvement over the quadratic model.