Polynomial Regression. Years of Experience working and Salary

Diego Ramos Crespo

March 31, 2025

1 Simple Linear Regression

Professional organizations frequently conduct surveys to analyze salary trends based on years of experience. The wage curve, which represents this relationship, is particularly useful for professionals evaluating their salary standing and for human resources departments making salary decisions.

In this section, we construct a simple linear regression model to estimate the salary (Y) as a function of years of experience (X). The regression equation is given by:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1.0.1}$$

where β_0 represents the intercept, β_1 is the slope coefficient, and ε is the error term.

To fit the model, we use the ordinary least squares (OLS) method in R. The following function computes the regression model:

```
# Fit simple linear regression model
lm_salary <- lm(y_salary ~ x_experience)

# Display model summary
summary(lm_salary)
```

The estimated coefficients can be extracted using the summary function. To visualize the regression line, we plot the data along with the fitted line as shown in Figure 1.

Figure 1: Simple Linear Regression.

The residuals vs. fitted values plot helps assess homoscedasticity, while the normal Q-Q plot checks the normality of residuals. When analyzing the diagnostics for the simple linear in Figure 2 model, several important behaviors can be observed in the diagnostic plots:

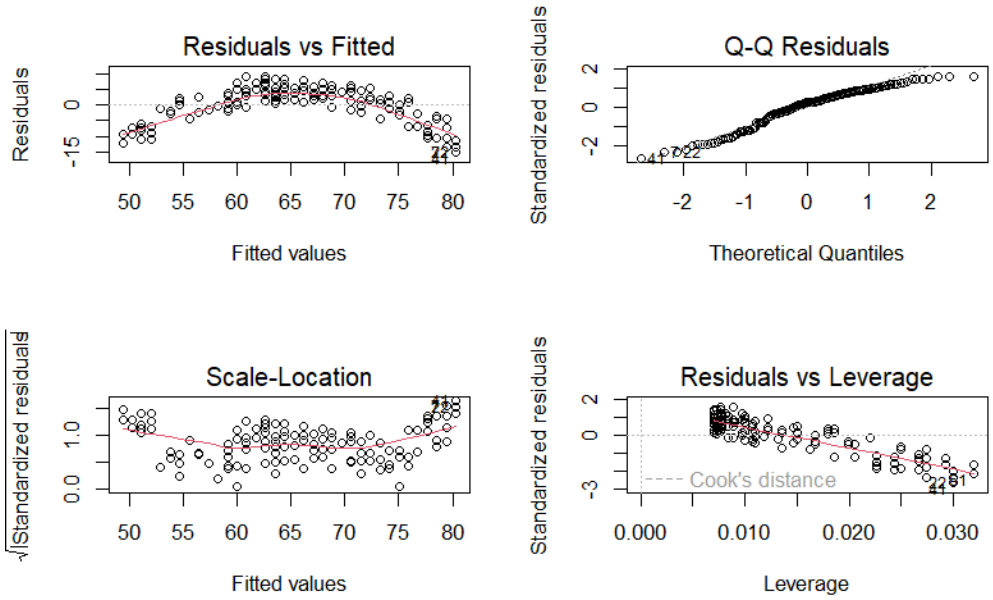

Figure 2: Quadratic Regression.

- **QQ Plot of Residuals**: The QQ plot of the residuals shows that they approximately follow a normal distribution. There are no large deviations from the straight line, suggesting that the assumption of normality for the residuals is reasonably well met.

- **Residuals vs Fitted Values**: In the residuals vs fitted values plot, the residuals do not appear to follow a clear linear pattern, indicating that the model might not fully capture the data and there may be some unaccounted non-linearity.

- **Scale-Location Plot**: In this plot, the points appear to be more evenly scattered, suggesting that the variance of the residuals is constant (homoscedasticity), although there are small variations.

- **Cook's Distance**: The Cook's distance plot shows some points above the typical threshold, indicating that these points may have a significant influence on the model fit. These points should be carefully considered for potential outliers or influential observations.

However, the model appears not to fit the data properly, therefore, we should look for another model. This different model is considered in the next section.

2 Polynomial Regression

In this part of the analysis, we extend the initial model by incorporating a new predictor, X_2, which represents the square of the years of experience variable, X. This quadratic term allows us to account for potential non-linearities in the relationship between salary and experience. The new model can be expressed as:

$$\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X + \widehat{\beta_2} X^2, \tag{2.0.1}$$

where β_0 is the intercept, β_1 is the coefficient for years of experience, β_2 is the coefficient for the quadratic term, and ε represents the error term. The inclusion of the quadratic term enables the model to capture curved relationships, as opposed to the linear assumption in the initial model.

2.1 Quadratic regression

To fit this model in R, we use the 'poly()' function, which allows us to include polynomial terms. In this case, we set the degree to 2 to capture the quadratic relationship. The model is then fit to the data, and the summary of the model provides the estimated coefficients for the intercept, linear term, and quadratic term.

With the new model fitted, we visualize the relationship between salary and experience by plotting the data along with the fitted quadratic regression line. The quadratic model introduces a curve that better reflects the shape of the data, which may be more appropriate if the relationship between salary and experience is non-linear. This new regression 'line' is displayed in Figure (1).

Next, we assess the model using diagnostic plots. These plots help us evaluate the assumptions of the regression model, such as linearity, homoscedasticity (constant variance of residuals), and the normality of residuals. For instance, the residuals vs. fitted values plot is used to check if the variance of the residuals is constant across all levels of the predictor variable, and the normal Q-Q plot helps us determine if the residuals follow a normal distribution.

When analyzing the diagnostics for the model, the following observations can be made based on the plots:

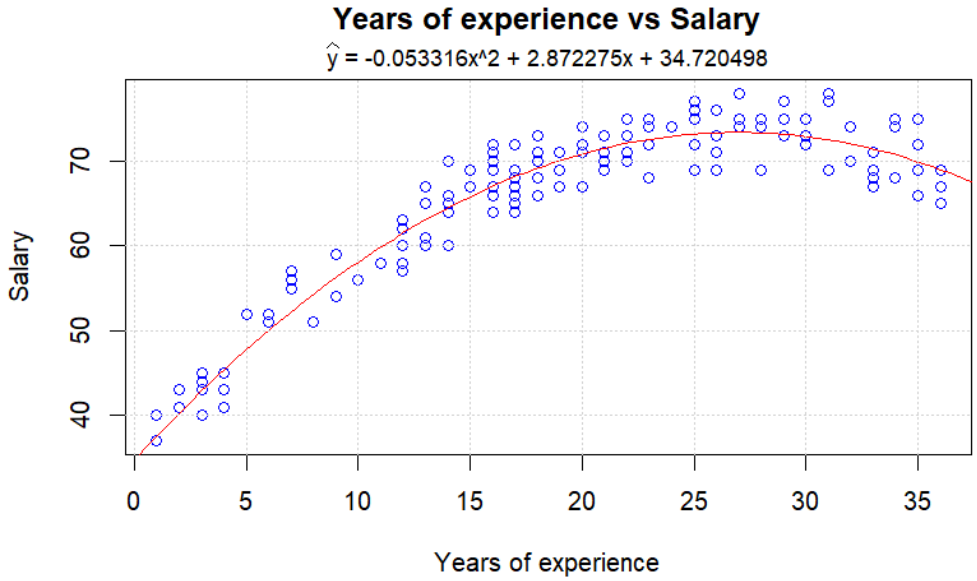

Figure 3: Quadratic Regression.

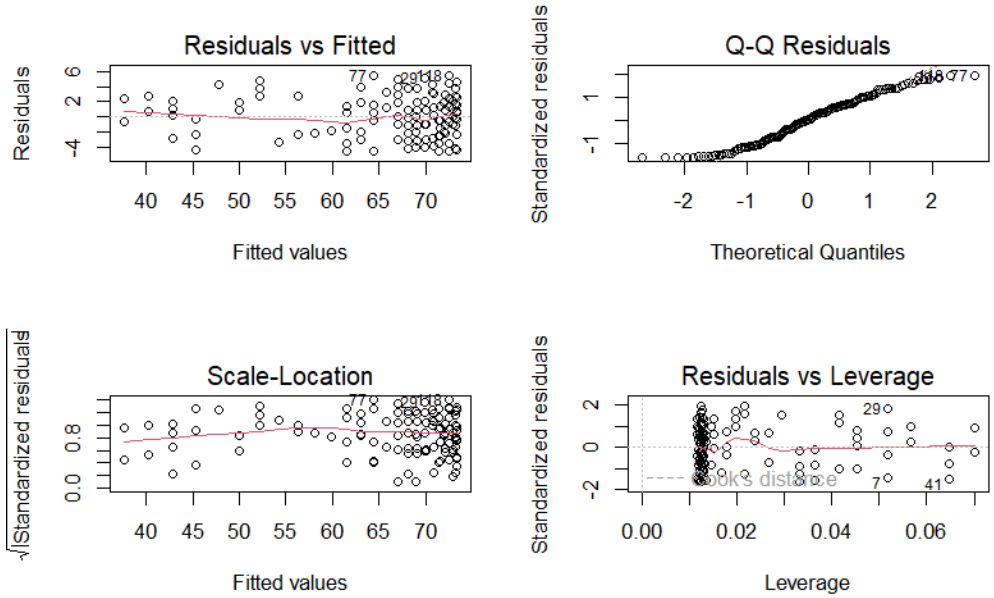

Figure 4: Quadratic Regression.

- **QQ Plot of Residuals**: The QQ plot of the residuals shows that they approximately follow a normal distribution, with no significant deviations from the straight line. This indicates that the residuals behave normally, satisfying the assumption of normality.

- **Residuals vs Fitted Values**: In this plot, the residuals do not appear to follow a clear linear pattern, suggesting the possibility of some unaccounted non-linearity in the data. The plot is somewhat skewed to the right.
- **Scale-Location Plot**: This plot shows that the residuals are somewhat dispersed, indicating that there might be some variance in the residuals. It is not perfectly uniform, with the points leaning towards the right side.
- **Residuals vs Leverage Plot**: In this plot, the points are concentrated on the left side, indicating that there may be influential points or outliers with low leverage. These points need to be examined for their influence on the model's fit.
- **Cook's Distance Plot**: In this plot, the points are generally distributed and appear to be well within acceptable limits, although there are a few points that might have some influence on the model, as indicated by Cook's distance.

Finally, we compare the performance of the quadratic model with the original linear model to determine if the inclusion of the quadratic term improves the fit. The first model is a simple linear regression, where we predict the salary (Y) based on years of experience (X). The output for the simple linear regression model is as follows:

```
Call:
lm(formula = y_salary ~ x_experience)

Residuals:
    Min      1Q  Median      3Q     Max
-15.310  -3.893   1.408   4.442   9.359

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.50593    1.08810   44.58   <2e-16 ***
x_experience  0.88345    0.05158   17.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 5.828 on 141 degrees of freedom
Multiple R-squared:  0.6754, Adjusted R-squared:  0.6731
F-statistic: 293.3 on 1 and 141 DF,  p-value: < 2.2e-16
```

The simple linear regression model explains 67.54% of the variance in salary, as indicated by the multiple R^2 value of 0.6754. The model is highly significant with a p-value less than 2.2×10^{-16}, suggesting that the relationship between years of experience and salary is strong. However, the residual standard error of 5.828 indicates that there is still some unexplained variability in the salary.

The second model is a multiple linear regression model that includes both the linear term and the quadratic term for years of experience. The output for the multiple linear regression model with the quadratic term is as follows:

```
Call:
lm(formula = Salary ~ Experience + I(Experience^2), data = profsalary_dts)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5786 -2.3573  0.0957  2.0171  5.5176

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
```

The multiple linear regression model with the quadratic term explains 92.47% of the variance in salary, as indicated by the multiple R^2 value of 0.9247. This is a significant improvement over

the simple linear regression model, which had an R^2 of 0.6754. The residual standard error has also decreased to 2.817, indicating a better fit of the model to the data.

Comparing the two models, we can conclude that the multiple linear regression model with the quadratic term provides a much better fit to the data than the simple linear regression model. The R^2 value of 0.9247 in the quadratic model is substantially higher than the R^2 value of 0.6754 in the simple model, indicating that the inclusion of the quadratic term improves the model's explanatory power.

Furthermore, the residual standard error has decreased from 5.828 in the simple linear regression model to 2.817 in the multiple linear regression model, suggesting that the quadratic model better captures the variability in salary data. This improvement in model performance justifies the inclusion of the quadratic term, as it allows for a more flexible and accurate representation of the relationship between salary and years of experience.

Both models are highly significant, with p-values less than 2.2×10^{-16} for all coefficients, suggesting that both models provide a statistically significant fit to the data. However, the quadratic model is clearly the superior model in this case due to its improved power in explaining and reducing the error.

2.2 ANOVA Table

To assess the significance of the multiple regression model, we perform an Analysis of Variance (ANOVA), which is shown in Table 1. The ANOVA table summarizes the variability in the data and helps determine whether the independent variables significantly explain the variation in the dependent variable.

Source	Sum of Squares	DoF	Mean Square	F-ratio	p-value
Regression	13640.858	2	6820.429	859.3178	2.43×10^{-79}
Residual	1111.184	140	7.937		
Total	14751.972	142			

Table 1: ANOVA Table for the Multiple Regression Model

To determine whether the regression model is statistically significant at a 99% confidence level, we conduct the following hypothesis test:

- **Null hypothesis** (H_0): The regression model is not significant, meaning that the independent

variables do not explain a significant portion of the variation in the dependent variable.

- **Alternative hypothesis** (H_1): The regression model is significant, meaning that at least one of the independent variables significantly contributes to explaining the variation in the dependent variable.

The test statistic used in ANOVA is the F-ratio, computed as:

$$F = \frac{\text{Mean Square Regression}}{\text{Mean Square Residual}} = \frac{6820.429}{7.937} = 859.3178$$

The corresponding p-value is 2.43×10^{-79}, which is extremely small and well below the 0.01 significance level.

Since the p-value is much smaller than 0.01, we reject the null hypothesis at the 99% confidence level. This confirms that the multiple regression model is statistically significant and explains a big proportion of the variability in the dependent variable. Therefore, the independent variables collectively have a significant effect on the response variable.

2.3 CI's for Coefficients

In order to assess the precision of the estimated regression coefficients, we construct individual confidence intervals at the 99% confidence level.

The confidence interval for each coefficient is computed using the formula:

$$\hat{\beta}_j \pm t_{\alpha/2,n-p} \cdot \text{SE}(\hat{\beta}_j)$$

where:

- $\hat{\beta}_j$ is the estimated coefficient,
- $t_{\alpha/2,n-p}$ is the critical value from the t-distribution with $n - p$ degrees of freedom,
- $\text{SE}(\hat{\beta}_j)$ is the standard error of the coefficient estimate.

For a 99% confidence level, the significance level is $\alpha = 0.01$, and the corresponding critical value from the t-distribution is:

$$t_{0.005,140} = 2.6114$$

Using the formula above, we obtain the following confidence displayed in Table 2 intervals for the regression coefficients:

Coefficient	Lower Bound	Upper Bound
$\widehat{\beta_0}$ (Intercept)	32.5564	36.8846
$\widehat{\beta_1}$ (Experience)	2.6224	3.1222
$\widehat{\beta_2}$ (Experience^2)	-0.0598	-0.0468

Table 2: 99% Confidence Intervals for Model Coefficients

- The confidence interval for $\widehat{\beta_0}$ (intercept) suggests that, when experience is zero, the expected salary falls between 32.56 and 36.88 (units depend on the dataset).
- The confidence interval for $\widehat{\beta_1}$ indicates that each additional year of experience is associated with an increase in salary between 2.62 and 3.12.
- The confidence interval for $\widehat{\beta_2}$ is entirely negative, confirming the presence of a quadratic effect.

Since none of the confidence intervals contain zero, we conclude that all three coefficients are statistically significant at the 99% confidence level.

2.4 Bonferroni Correction

When constructing multiple confidence intervals simultaneously, the probability of at least one interval not containing the true parameter value increases. To maintain an overall confidence level of 99%, we apply the Bonferroni correction.

The Bonferroni correction adjusts the significance level to account for multiple comparisons by dividing the overall significance level α by the number of parameters m. The adjusted significance level is:

$$\alpha' = \frac{\alpha}{m} = \frac{0.01}{3} = 0.0033$$

The corresponding critical value from the t-distribution with 140 degrees of freedom is:

$$t_{0.0033/2,140} = 2.948$$

The Bonferroni-adjusted confidence intervals for each coefficient are calculated using:

$$\hat{\beta}_j \pm t_{\alpha'/2,n-p} \cdot \text{SE}(\hat{\beta}_j)$$

where:

- $\hat{\beta}_j$ is the estimated coefficient,
- $t_{\alpha'/2,n-p}$ is the critical value from the t-distribution,
- $\text{SE}(\hat{\beta}_j)$ is the standard error of the coefficient.

Coefficient	Lower Bound	Upper Bound
$\widehat{\beta_0}$ (Intercept)	32.2453	37.1957
$\widehat{\beta_1}$ (Experience)	2.5865	3.1581
$\widehat{\beta_2}$ (Experience^2)	-0.0607	-0.0459

Table 3: Bonferroni-Adjusted 99% Confidence Intervals for Model Coefficients

- The Bonferroni confidence interval for $\widehat{\beta_0}$ is wider than the standard confidence interval, reflecting the conservative nature of this method to ensure an overall confidence level of 99%.
- The confidence interval for $\widehat{\beta_1}$ suggests that each additional year of experience increases salary between 2.59 and 3.16, adjusting for multiple comparisons.
- The confidence interval for $\widehat{\beta_2}$ remains negative, supporting the quadratic effect where salary increases with experience but at a diminishing rate.

Since none of these intervals contain zero, all coefficients remain statistically significant even under the stricter Bonferroni adjustment.

2.5 CI's for mean and prediction

To estimate the salary for a worker with 18 years of experience, we compute both a point estimate and confidence intervals for the mean and prediction.

The point estimate of the salary is given by:

$$\hat{y} = 32.7205 + (2.8723 \times 18) + (-0.0533 \times 18^2)$$

where $X_{\text{new}} = 18$. Substituting the estimated coefficients we have that:

$$\hat{y} = 69.14706$$

The standard error for the mean estimate is:

$$SE_{\text{mean}} = s\sqrt{\frac{1}{n} + \frac{(X_{\text{new}} - \bar{X})^2}{S_{xx}}}$$

where:

- $s = \sqrt{7.9}$ is the residual standard error,
- $n = 143$ is the sample size,
- $\bar{X}$ is the mean of the experience variable,
- $S_{xx} = \sum(X_i - \bar{X})^2$ is the sum of squares for the experience variable.

The critical value for a 90% confidence interval is:

$$t_{\alpha/2,140} = t_{0.05,140} = 1.655$$

Thus, the 90% confidence interval for the mean salary is:

$$[68.75627, 69.53786]$$

The standard error for the prediction interval is:

$$SE_{\text{pred}} = s\sqrt{1 + \frac{1}{n} + \frac{(X_{\text{new}} - \bar{X})^2}{S_{xx}}}$$

Using the same critical value, the 90% prediction interval for an individual salary is:

$$[64.47671, 73.81742]$$

- The confidence interval for the mean salary indicates that, on average, workers with 18 years of experience earn between \$68,756.27 and \$69,537.86.

- The wider prediction interval reflects the variability in individual salaries, suggesting that a randomly chosen worker with 18 years of experience is likely to have a salary between \$64,476.71 and \$73,817.42.

3 Cubic Model

To evaluate whether adding a cubic term (X^3) improves the model, we fit a regression of the form:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \hat{\beta}_3 X^3 \tag{3.0.1}$$

where:

- $\hat{\beta}_0 = 35.832$ (intercept)
- $\hat{\beta}_1 = 2.541$ (linear term)
- $\hat{\beta}_2 = -0.031$ (quadratic term)
- $\hat{\beta}_3 = -0.0004$ (cubic term)

The results indicate that while the linear and quadratic terms are statistically significant (with p-values < 0.05), the cubic term is not ($p = 0.173$). This suggests that adding the cubic term does not significantly improve the model. The following Figure 5, shows the new model using the cubic term, however, it does not seems to improve much.

To formally assess whether the cubic model is superior to the quadratic model, we perform an F-test comparing:

- Model 1: $Y = \beta_0 + \beta_1 X + \beta_2 X^2$
- Model 2: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$

The residual sum of squares (RSS) for each model is:

- Model 1: RSS = 1111.2
- Model 2: RSS = 1096.4

The F-statistic for testing the added complexity of Model 2 is:

$$F = \frac{(\text{RSS}_1 - \text{RSS}_2)/(df_1 - df_2)}{\text{RSS}_2/df_2} = \frac{(1111.2 - 1096.4)/1}{1096.4/139} = 1.876 \tag{3.0.2}$$

Figure 5: Cubic Regression.

The corresponding p-value is 0.173, which is not significant at the typical $\alpha = 0.05$ level. This confirms that adding the cubic term does not provide a statistically significant improvement to the model.

Although the cubic model has a slightly higher R^2 value ($R^2 = 0.9257$ vs. $R^2 = 0.9241$ for the quadratic model), the increase is negligible, and the cubic term is not statistically significant. Based on the principle of parsimony, we conclude that the quadratic model is preferable, as it provides a similar level of explanatory power with fewer parameters.