Chapter 3: Linear Regression

Linear regression is one of the simplest and most useful techniques for modeling the relationship between a quantitative response and one or more predictors. This chapter covers both simple and multiple linear regression.

1. Simple Linear Regression

We model the relationship between a single predictor X and the response Y as:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- β_0: intercept
- β_1 : slope
- ε: error term (captures variability in Y not explained by X)

The goal is to estimate β_0 and β_1 such that the model fits the data well.

Estimate Coefficients

We estimate the coefficients by minimizing the **Residual Sum of Squares (RSS):**

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The solution gives:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\text{and}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} \bar{x}$$

Standard Error

We can use the standard Error to get confidence intervals. The 95% confidence intervals are:

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

$$\hat{\beta}_2 \pm 2 \cdot SE(\hat{\beta}_2)$$

Hypothesis Tests

Null hypothesis

H_0: There is no relationship between X and Y

Alternative Hypothesis

H_a: There is some relationship

Mathematically:

$$H_0: \beta_1 = 0$$

$$H_0: \beta_1 \neq 0$$

t static:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

it measure the number of standard deviations $\hat{\beta}$ is away from 0.

p-value:

- probability of observing any number equal to $|t|$ or larger in absolute value, assuming $\beta_1 = 0$
- A small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance.
- What is small? Usually less than 5% (p<0.05)

Assessing Model Accuracy

Residual Standard Error

An estimate of the stdev(ε)

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i}^{n}(y_i - \hat{y}_i)^2}$$

- Measure of **lack of fit**
- divide it by mean value to get an idea!

R^2 *Static*

Proportion of variance; independent of Y

$$R^2 = 1 - \frac{RSS}{TSS}$$

- measures proportion of variability in Y that can be explained using X
- it goes from 0 to 1. we want it to be close to 1
- How close? It depends on the problem. In physics problems it has to be very close. But in social sciences it can be ok even at 0.6

2. Multiple Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Estimate Coefficients

- Minimize RSS
- Remember that correlation does not mean causation. More on that later.

Fundamental Questions

Is there a relationship between response and predictors?

$$H_0: \beta_1 = \beta_2 = \cdots \beta_p = 0$$

$$H_a: \text{at leaste one } \beta_j \neq 0$$

- F-Static: $F = \frac{(TSS-RSS)/p}{RSS/(n-p-1)}$
- If F > 1 we can reject null hypothesis
- We should also compute p-value

Deciding on important Variables

Which variables are actually important? How can we make sure that the variables are the cause of the response and not just a correlation?

A. **Forward Selection:** start with null model. We then fit p simple linear regressions and add the one with the lowest RSS. Then add the one with the lowest RSS for two-var model. Continue until some stopping rule

B. **Backward Selection:** Start with all variables, and remove the one with the largest p-value. The new model is fit, and save. Continue until stopping rule. (works only for p<n)
C. **Mixed:** start w/o variables. Add best fit. Keep going, if the p-vale goes above a threshold, remove that variable. Continue until we get a small enough p-value.

Model Fit

- We use R^2 and/or RSE
- Plot them to get a better idea

Predictions

There are 3 uncertainties associated with our model.

A. The coefficients are only estimated and the least square plane is an estimate of the true population regression plane. We compute coefficient intervals to tackle that issue.
B. Assuming linearity is almost always an approximation. We ignore it for now.
C. Irreducible error ε. We use prediction intervals for that.

3. Other Considerations

Qualitative data

- Predictors with two levels:$x_i = \begin{cases} 1, if\ someting \\ 0, if\ not\ something \end{cases}$
- That gives us: $y_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i \\ \beta_0 + \varepsilon_i \end{cases}$
- For more levels, create more dummy variables

Extensions

Regression assumes that the relationship between predictors and response is additive and linear.

- **Additive**: the association between X_j and Y doesn't depend on other predictions of X_i
- **Linear:** the change in Y is associated with one-unit change in X_i is constant, regardless of the value of X_j

Removing Additive Assumption

Synergy or interaction effect

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Non-linear relationships

Use polynomial regression

Potential Problems

When we ft a linear regression model to a particular data set, many problems may occur. Most common among these are the following.

Non-linearity

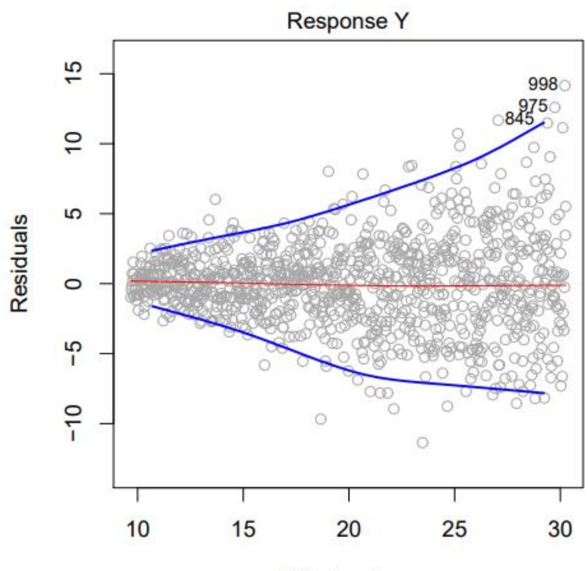
- Use residual plots
- Poly regression

Correlation of error terms

- ε's are corelated, not good
- plot time series residuals, and if the adjacent residuals have similar values then we have a problem

Non-constant variance of error terms

- assuming that $\mathrm{var}(\varepsilon_i) = \sigma^2$
- Sometimes though the variance of the error may increase with the values of the response. That is called **heteroscedasticity**



-
- Use weighted R^2

Outliers

- Plot the data
- Remove it
- If you are not sure, plot studentized residuals. That is divide ε_i by $SE(\varepsilon_i)$. if more than 3 then remove it.

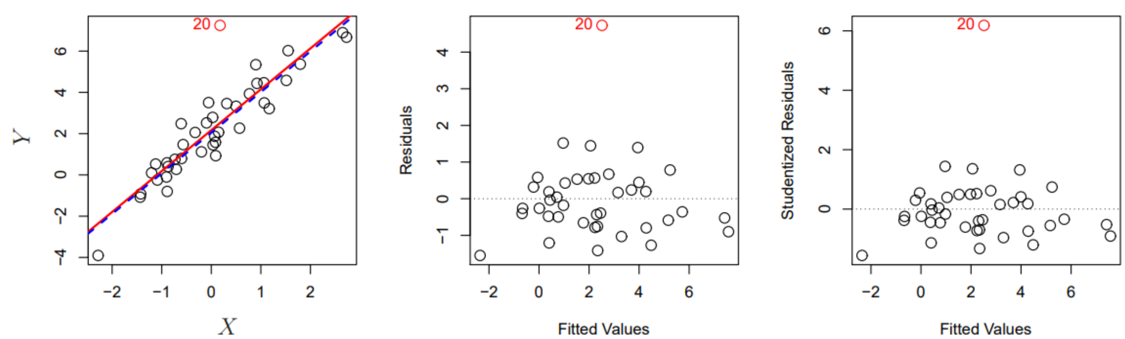


FIGURE 3.12. Left: *The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue.* Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of 6; typically we expect values between* -3 *and* 3.

High Leverage Points

- When the value of an observation is a lot higher than the rest.

- Leverage statistic: $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_{i'} - \bar{x})^2}$, and $\frac{1}{n} \leq h_i \leq 1$
- If a point has , $h_i > \frac{p+1}{n}$ that is a high leverage point

Collinearity

- Two or more predictors are closely related
- Look at correlation matrix of predictors
- For multicollinearity compute Variance Inflation Factor (VIF):

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

- VIF=1 means no collinearity
- Drop the problematic predictor

KNN Regression

Carefully explain the differences between the KNN classifier and KNN regression methods.

KNN regression: Predicts a **quantitative** output by averaging the responses of the K nearest neighbors.

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \aleph} y_i$$

KNN classification: Predicts a **qualitative** output by majority vote among the K nearest neighbors.

$$\hat{y}(x_0) = \text{most frequent class among } K \text{ nearest neighbors}$$

- Larger K => smoother fit, more bias
- Smaller K => low bias, large variance
- Usually worse than linear regression

An Introduction to Statistical Learning with Applications in Python
Notes: Ioannis Mastoras
May 2025