Linear Regression

Intermediate Data Analytics

Dr. Alex Marsella

2024-02-26

Introduction to Linear Regression

- Linear regression is a fundamental statistical method used for inference and prediction.
- Describes the relationship between two variables (two features of the world.)
- Useful for forecasting and finding out how much one variable affects another.
- Different from "correlation" in that it allows us to measure changes in the units of the variables of interest.
  - e.g. "When X increases by 1, we expect to see Y increase by 1.7"

From Correlation to Regression

- Recall correlation: Measures the strength and direction of a linear relationship between two variables.
- Regression builds on this: Not just describing the relationship, but modeling it to predict outcomes or infer causes.
- Single variable (simple) linear regression involves two variables: one independent (predictor, X) and one dependent (outcome, Y).
- Multiple regression involves more than one independent variables (more than one X determining Y).

The Linear Regression Model

- The equation of a line: $Y = \beta_0 + \beta_1 X + \epsilon$
  - Y: Dependent variable (what we're trying to predict)
  - X: Independent variable (predictor)
  - β_0: Intercept (value of Y when X = 0)
  - β_1: Slope (change in Y for a one-unit change in X)
  - ϵ: Error term (difference between observed and predicted values)

The "Error" or "Residual"

- Error, also called "residual"
  - ϵ_i for observation i, is equal to $y_i - \hat{y}_i$ where y_i is the actual value and $\hat{y}_i$ is the predicted value.
- Don't get mixed up, Error = Actual - Predicted, not the other way around!

Sum Squared of Errors (SSE)

- SSE measures the total deviation of the response values from the fit line.
- Formula: $SSE = \Sigma(y_i - \hat{y}_i)^2$ where y_i is the actual value and $\hat{y}_i$ is the predicted value.
- Minimizing SSE helps in finding the best-fitting line.

Simulating Data: Crime vs. Temperature in Chicago

- Assume a positive relationship between temperature and crime rates.
- We'll simulate data for a basic analysis:
  - `Temperature`: Predictor (X).
  - `Crime Rate`: Dependent variable (Y).

Plotting and Analysis

- We'll plot our simulated data.
- Fit a linear regression model.
- Visualize the line of best fit with error bars.

Simulating some Crime Data

- I am hardcoding a relationship where each degree of temperature increases crime rate by 1.5.
- I am adding a normally distributed "error" to that (noise) that has mean of 0 and sd of 5.

```
set.seed(123)  # for reproducibility
# drawing 100 temperatures in
# Fahrenheit from a uniform
# distribution
temperature <- runif(100, min = 30, max = 100)
crime_rate <- 1.5 * temperature + rnorm(100,
    mean = 0, sd = 5)  # Simulated crime rate
# that `+rnorm` is me simulating an
# error term
chicago <- data.frame(temperature, crime_rate)
```

Fitting a model.

The theoretical model:

- $CrimeRate = \beta_0 + \beta_1 * Temperature + \epsilon$

The fitted model using our sample data:

- $PredictedCrimeRate = \hat{\beta}_0 + \hat{\beta}_1 * Temperature$
- We *assume* the error is, on average, equal to 0. (I did hardcode it that way.)
  - If it's not, we have to use more advanced techniques.

```
# Fit a linear model
model <- lm(crime_rate ~ temperature, data = chicago)
```

Viewing the results

- $PredictedCrimeRate = .14770 + 1.49358 * Temperature$

Imagine temperature is 50, then we predict:

- $PredictedCrimeRate = .14770 + 1.49358 * 50 = 76.156$

```r
summary(model)
```

```
##
## Call:
## lm(formula = crime_rate ~ temperature, data = chicago)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1899  -3.0661  -0.0987   2.9817  11.0861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14770    1.65702   0.089    0.929
## temperature  1.49358    0.02442  61.173   <2e-16 ***
```

Cleaner way to observe the regression coefficients

```
model$coefficients
```

```
## (Intercept) temperature
##   0.1476965   1.4935835
```

Predicting crime?

- The “predict” command will automatically do the previous calculation for all observations
- It plugs in values of temperature from each observation to the model.
- We can then calculate our “error” term for each observation by subtracting predicted from actual.

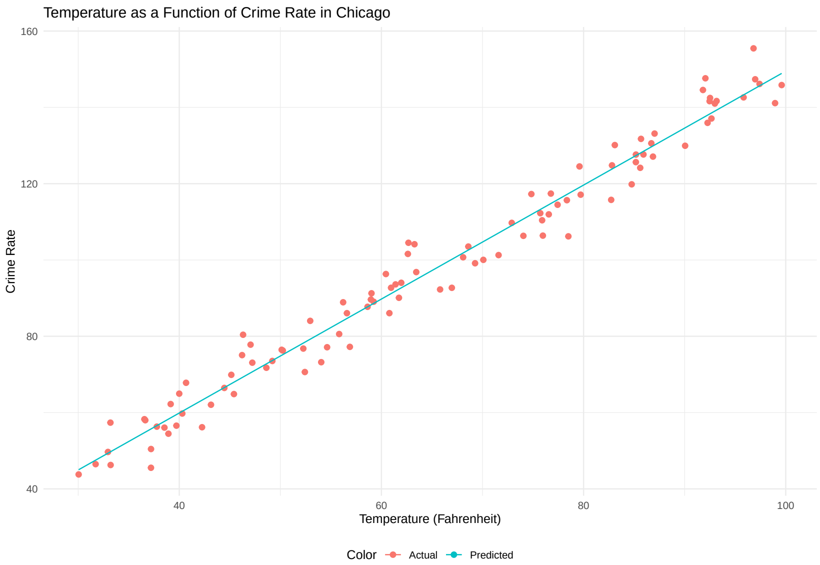
```
# Add predictions to the dataset
chicago$predicted_crime_rate <- predict(model,
    data = chicago)

# Calculate residuals (errors)
chicago$residuals <- with(chicago, crime_rate -
    predicted_crime_rate)
```

Plotting it

- NOTE: Us saying `geom_line(aes(y=predicted)` is the same as us doing `geom_smooth(method=lm)`
  - We ran the `lm`, the "linear model", already!

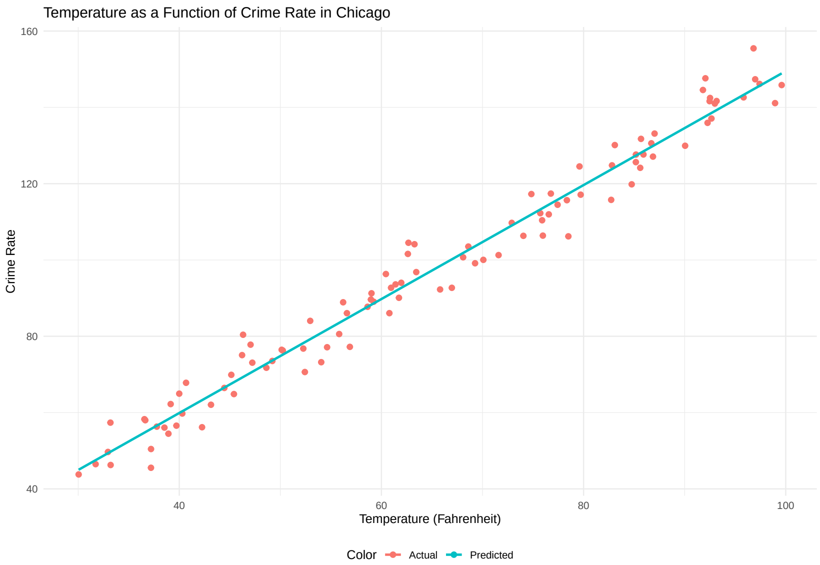
```
plot1 <- ggplot(chicago, aes(x = temperature, y = crime_rat
  geom_point(aes(color = "Actual"), size = 2) + # Actual da
  geom_line(aes(y = predicted_crime_rate, color = "Predicte
  labs(title = "Temperature as a Function of Crime Rate in
       x = "Temperature (Fahrenheit)", y = "Crime Rate", co
  theme_minimal() +
  theme(legend.position = "bottom")
```


Temperature as a Function of Crime Rate in Chicago
160
120
80
40
Crime Rate
40
60
80
100
Temperature (Fahrenheit)
Color
Actual
Predicted

Let me prove it to you

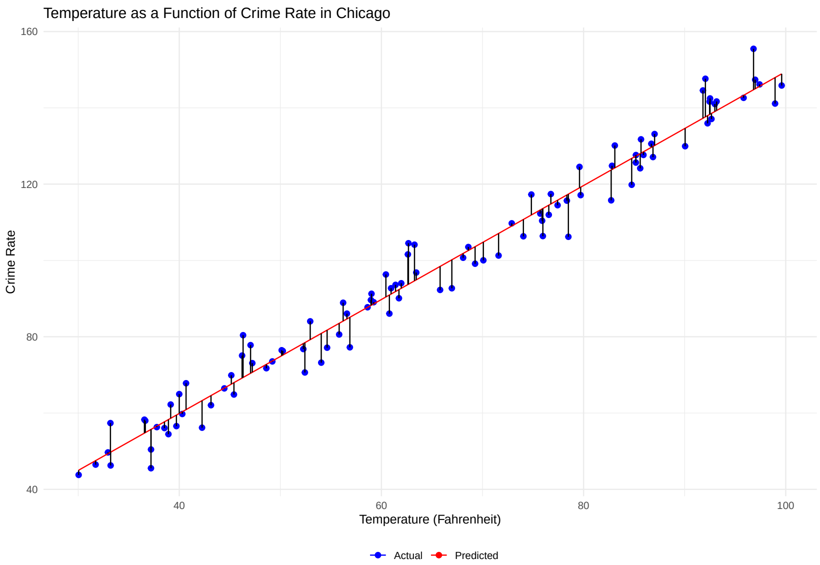
- Notice how the little message it gives when you run it literally says "using formula = y ~ x"

```
plot2 <- ggplot(chicago, aes(x = temperature, y = crime_rat
  geom_point(aes(color = "Actual"), size = 2) + # Actual da
  geom_smooth(method="lm", aes(color="Predicted"), se=FALSE
  labs(title = "Temperature as a Function of Crime Rate in
       x = "Temperature (Fahrenheit)", y = "Crime Rate", co
  theme_minimal() +
  theme(legend.position = "bottom")
```


Temperature as a Function of Crime Rate in Chicago
160
120
80
40
Crime Rate
40
60
80
100
Temperature (Fahrenheit)
Color
Actual
Predicted

Plotting with errors

```
plot_errors <- ggplot(chicago, aes(x = temperature, y = cri
  geom_point(aes(color = "Actual"), size = 2) + # Actual da
  geom_line(aes(y = predicted_crime_rate, color = "Predicte
    geom_segment(aes(xend = temperature, yend = predicted_c
  scale_color_manual("", breaks = c("Actual", "Predicted"),
  labs(title = "Temperature as a Function of Crime Rate in
       x = "Temperature (Fahrenheit)", y = "Crime Rate", co
  theme_minimal() +
  theme(legend.position = "bottom")
```


Temperature as a Function of Crime Rate in Chicago
Crime Rate
160
120
80
40
40
60
80
100
Temperature (Fahrenheit)
Actual
Predicted

Understanding the Model Output

```
##
## Call:
## lm(formula = crime_rate ~ temperature, data = chicago)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1899  -3.0661  -0.0987   2.9817  11.0861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14770    1.65702   0.089    0.929
## temperature  1.49358    0.02442  61.173   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
##
## Residual standard error: 4.846 on 98 degrees of freedom
## Multiple R-squared:  0.9745, Adjusted R-squared:  0.974
## F-statistic:  3742 on 1 and 98 DF,  p-value: < 2.2e-16
```

Statistical Significance

- Each coefficient in the regression model is hypothesis tested
  - $H_0 : \beta_k = 0$
  - $H_1 : \beta_k \neq 0$
- If the sample coefficient $\hat{\beta}_k$ is large enough (in absolute value) relative to its standard error, the odds of finding a coefficient that large if the true population coefficient β_k were 0, is very small
- p-value (Pr(>|t|)) measures that probability
  - answers the question "If the null is true, $\beta_k = 0$, what is the probability of finding a sample coefficient $\hat{\beta}_k$ this large in magnitude or larger?
  - p-value less than 0.05 is usually our standard for statistical significance.

Interpretations for your reference

Intercept

- What would we expect Y to be if X $=$ 0?
- Can't be interpreted if we never observe X $=$ 0 in our data.

R-squared

- Answers the question "how much of the variation in Y can be explained by variation in our X variables?"

F-statistic

- H_0: β_k for all k $= 0$
- H_1: At least one $\beta_k \neq= 0$
- Answers the question "is at least one of our X variables statistically significant?"
  - If p-value is small, then yes.

On Wednesday

- We will do some hands-on coding and exercises with regression.
- If you feel uncomfortable with some of the statistics here, you need to review.