# Estimating the Coefficients and Fitting Linear Models: Takeaways



## Syntax

### EVALUATING BIVARIATE RELATIONSHIPS

- Fitting a bivariate linear regression model in R:

```
lm_fit <- lm(response ~ predictor, data = df)
```

- Function for manually estimating the slope:

```
slope <- function(predictor, response){
  mean_predictor <- mean(predictor)
  mean_response <- mean(response)
  numerator <- sum((predictor - mean_predictor) * (response - mean_response))
  denominator <- sum((predictor - mean_predictor)^2)
  beta_1 <- numerator / denominator
  beta_1
}
```

- Function for manually estimating the intercept (requires slope estimate as function argument):

```
intercept <- function(predictor, response, slope){
  beta_0 <- mean(response) - (slope * mean(predictor))
  beta_0
}
```

- Checking equality:

```
dplyr::near(vector_1, vector_2)
```

- Manually adding predictions variable to dataframe:

```
df <- df %>%
  mutate(predictions = coef(lm_fit)[[1]] + coef(lm_fit)[[2]] * predictor)
```

- Manually adding residuals variable to dataframe:

```
df <- df %>%
  mutate(residuals = response - predictions)
```

- Manually estimating the residual sum of squares (RSS):

```
df <- df %>%
  mutate(resid_squared = residuals^2)
RSS <- df %>%
  summarize(RSS = sum(resid_squared)) %>%
  pull()
```

- R functions useful for accessing linear model outputs:

```
coef(lm_fit) # Intercept and slope coefficients
fitted(lm_fit) # Predictions
resid(lm_fit) # Residuals
deviance(lm_fit) # Residual sum of squares
```

## Equations

- The mathematical equation for bivariate linear regression:
  - $Y = \beta_0 + \beta_1 X + \epsilon$, where:
  - $Y$ indicates the prediction $Y$ based on $X$.
  - $\beta_0$ (the intercept) refers to the value on the y-axis where the value on the x-axis is equal to $0$.
  - $\beta_1$ (the slope) is the change in $Y$ for every single unit change in $X$.
  - $X$ is the predictor, or independent variable.
  - $\epsilon$ is the error term that captures everything that is missed by the model.
- The mathematical formula for making future predictions using linear model coefficients:
  - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, where:
  - $\hat{y}$ indicates a prediction of $y$ for any given value of $x$.
  - $\hat{\beta}_0$ (the intercept) refers to the estimated value on the y-axis where the value on the x-axis is equal to $0$.
  - $\hat{\beta}_1$ (the slope) is the estimated change in $\hat{y}$ for every single unit change in $x$.
  - $x$ is the value of the predictor variable.

- The mathematical formula for estimating slope:
  - $slope = \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, where:
  - $\sum_{i=1}^{n}$ is the sum of...
  - $x_i$ is the value of the $i$th predictor variable.
  - $y_i$ is the value of the $i$th response variable.
  - $\bar{x}$ is the average value of the predictor variables.
  - $\bar{y}$ is the average value of the response variables.
- The mathematical formula for estimating intercept:
  - $intercept = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

## Concepts

- **Bivariate linear regression** can be performed between pairs of quantitative data measured on an interval or ratio scale. Bivariate linear regression describes the relationship between a response variable of interest and a predictor variable. This method helps us to separate the signal (what the predictor variable tells us about the response variable) from the noise (what the predictor variable can't tell us about the response variable).
- **Slope** ($\hat{\beta}_1$) is the estimated change in $\hat{y}$ for every single unit change in $x$.
- **Intercept** ($\hat{\beta}_0$) refers to the estimated value on the y-axis where the value on the x-axis is equal to 0.
- **Residual** is the difference between the actual values and the model predictions for the response variable.
- The sum of the squared residuals is known as the **residual sum of squares** (RSS). The **least squares** criterion is the most common method for fitting a linear regression model. The lm() algorithm in R selects values for $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize $RSS$.

## Resources

- Wikipedia entry on linear regression.
- Wikipedia entry on regression analysis.
- Wikipedia entry on least squares.
- Wikipedia entry on residual sum of squares (RSS).