Supervised Learning: Linear Regression

06-22-2023

Model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

for i = 1, 2, . . . , n and where:

$$\epsilon_i \sim N(0, \sigma^2)$$

Simple Linear Regression Estimation:

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

* average value for Y given the value for X

- averaging out the error (ϵ has a mean of 0)

How to Calculate our Coefficient Estimates?

Ordinary least squares (OLS) finds the coefficient estimates by minimizing to residual sum of squares (RSS)

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Connection to Covariance and Correlation

Covariance = joint variability of two variables

Correlation = normalized form of the covariance, ranges from -1 to 1

Gapminder Data

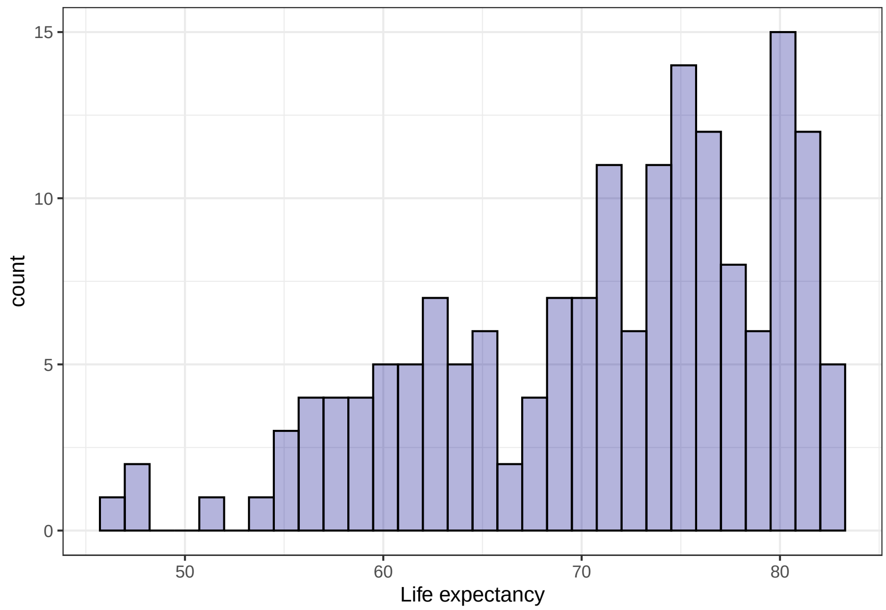
```
gapminder <- as_tibble(gapminder)
clean_gapminder <- gapminder %>%
  filter(year == 2011, !is.na(gdp)) %>%
  mutate(log_gdp = log(gdp))

clean_gapminder
```

```
## # A tibble: 168 x 10
##    country    year infant_mortality life_expectancy fertility population     gdp
##    <fct>     <int>            <dbl>           <dbl>     <dbl>      <dbl>   <dbl>
##  1 Albania    2011             14.3            77.4      1.75    2886010 6.32e 9
##  2 Algeria    2011             22.8            76.1      2.83   36717132 8.11e10
##  3 Angola     2011            107.             58.1      6.1    21942296 2.70e10
##  4 Antigua ~  2011              7.2            75.9      2.12      88152 8.02e 8
##  5 Argentina  2011             12.7            76        2.2    41655616 4.73e11
```

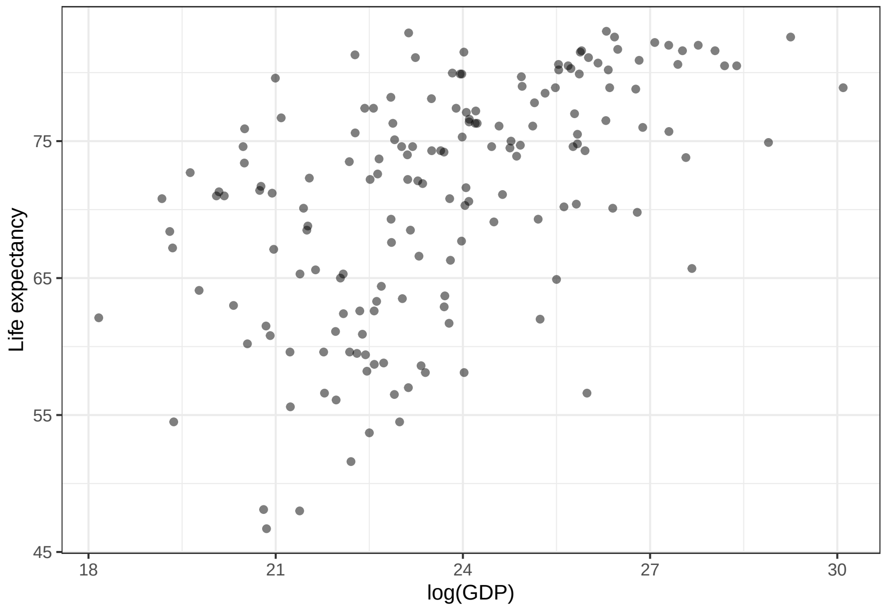
```
##   6 Armenia     2011              15.3             73.5      1.5      2967984 4.29e 9
##   7 Australia   2011               3.8             82.2      1.88    22542371 5.73e11
##   8 Austria     2011               3.4             80.7      1.44     8423559 2.31e11
##   9 Azerbaij~   2011              32.5             70.8      1.96     9227512 2.14e10
## 10 Bahamas     2011              11.1             72.6      1.9       366711 6.76e 9
## # i 158 more rows
## # i 3 more variables: continent <fct>, region <fct>, log_gdp <dbl>
```

Modeling Life Expectancy

```r
clean_gapminder %>%
  ggplot(aes(x = life_expectancy)) +
  geom_histogram(color = "black", fill = "darkblue", alpha = 0.3) +
  theme_bw() +
  labs(x = "Life expectancy")
```




```r
gdp_plot <- clean_gapminder %>%
  ggplot(aes(x = log_gdp, y = life_expectancy)) +
  geom_point(alpha = 0.5) +
  theme_bw() +  labs(x = "log(GDP)", y = "Life expectancy")

gdp_plot
```




```r
init_lm <- lm(life_expectancy ~ log_gdp, data = clean_gapminder)

summary(init_lm)
```

```
##
## Call:
## lm(formula = life_expectancy ~ log_gdp, data = clean_gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.901  -4.781   1.879   5.335  13.962
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   24.174      5.758   4.198 4.38e-05 ***
## log_gdp        1.975      0.242   8.161 7.87e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.216 on 166 degrees of freedom
## Multiple R-squared:  0.2864, Adjusted R-squared:  0.2821
## F-statistic: 66.61 on 1 and 166 DF,  p-value: 7.865e-14
```

Inference with OLS

p-values: estimated probability of observing the t-value or more extreme given the null hypothesis that $\beta = 0$ is true.

When the p-value < coefficient threshold of $\alpha = 0.05$, **sufficient evidence to reject the null hypothesis that the coefficient is zero.**

Typically, t-values with an absolute value greater than 2 indicate a **significant** relationship at $\alpha = 0.05$. I.e., there is a **significant** association between `life_expectancy` and `log_gdp`.

P-value Caveats

- If the true value of the coefficient is $\beta = 0$, the p-value is sampled from a **uniform(0,1) distribution**. So, it is just as likely to have a p-value of 0.45 as 0.84 or 0.999 or 0.000001.

Hence, we only reject for low α values like 0.05

- Controlling the Type 1 error rate at $\alpha = 0.05$, i.e., the probability of a **false positive** mistake
- 5% chance that you will conclude there is a significant association between x and y *even when there is none.*

Also, remember $SE = \frac{\sigma}{\sqrt{n}}$

- As n gets large **standard error goes to zero** and *all* predictors are eventually deemed significant
- While the p-values might be informative, we will explore other approaches to determine which subset of predictors to include (e.g., holdout performance)

Multiple R-squared

R-squared estimates the **proportion of variance** in Y explained by X.

```
with(clean_gapminder, cor(log_gdp, life_expectancy))^2
```

```
## [1] 0.2863522
```

Equivalently:

```
var(predict(init_lm)) / var(clean_gapminder$life_expectancy)
```

```
## [1] 0.2863522
```

Generating Predictions

```
train_preds <- predict(init_lm)
head(train_preds)
```

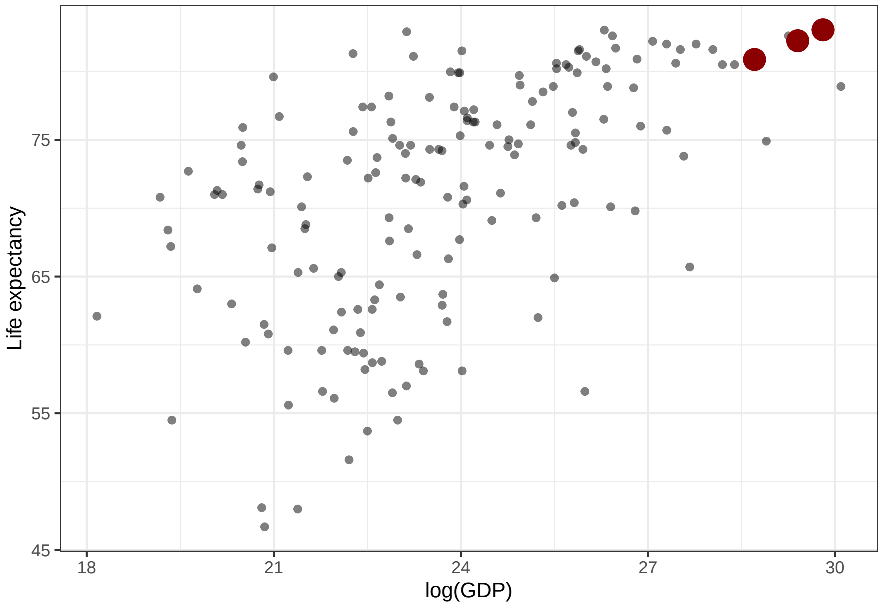
```
##        1        2        3        4        5        6
## 68.74401 73.78465 71.61243 64.66585 77.26605 67.97876
```

```
## also could do: head(init_lm$fitted.values)
```

Predictions for New Data

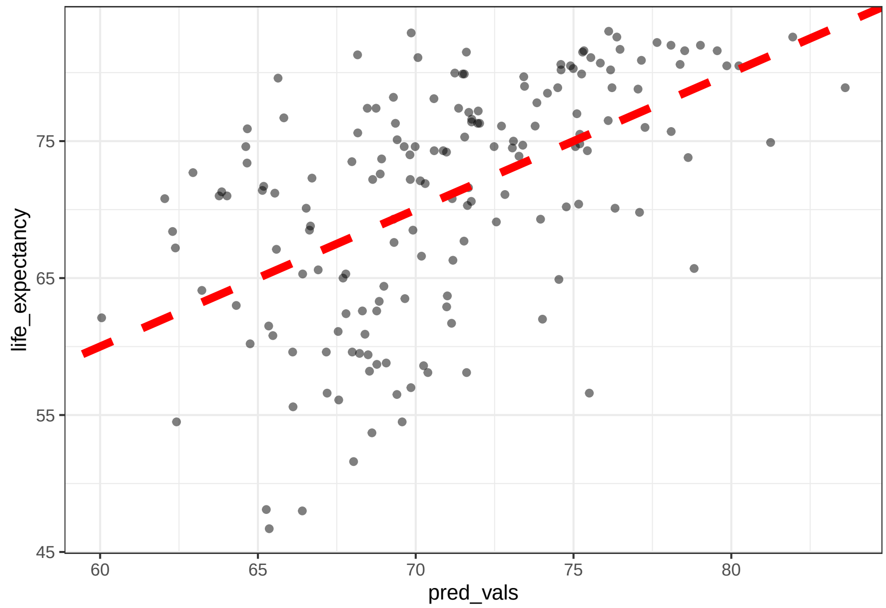
```
us_data <- clean_gapminder %>%
  filter(country == "United States")

new_us_data <- us_data %>%
  dplyr::select(country, gdp) %>%
```

```
  slice(rep(1, 3)) %>%
  mutate(adj_factor = c(0.25, 0.5, 0.75), log_gdp = log(gdp * adj_factor))

new_us_data$pred_life_exp <- predict(init_lm, newdata = new_us_data)

gdp_plot +
  geom_point(data = new_us_data, aes(x = log_gdp, y = pred_life_exp), color = "darkred", size = 5)
```



Observed Values Against Predictions

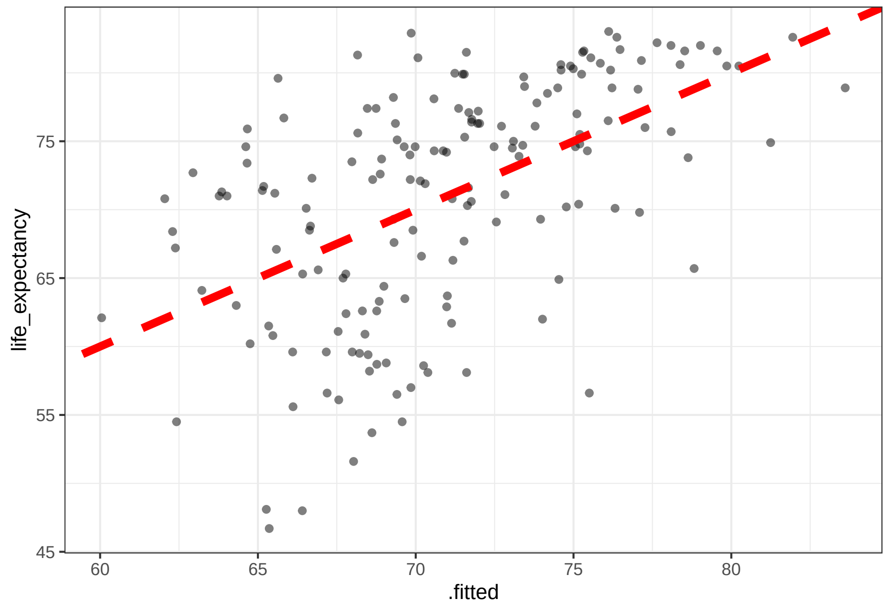
```
clean_gapminder %>%  mutate(pred_vals = predict(init_lm)) %>%
  ggplot(aes(x = pred_vals, y = life_expectancy)) +
  geom_point(alpha = 0.5) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red", size = 2) +
  theme_bw()
```



- “Perfect” model will follow the **diagonal**

With `broom` package:

```
clean_gapminder <-   broom::augment(init_lm, clean_gapminder)

clean_gapminder %>%
  ggplot(aes(x = .fitted, y = life_expectancy)) +
  geom_point(alpha = 0.5) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red", size = 2) +
  theme_bw()
```

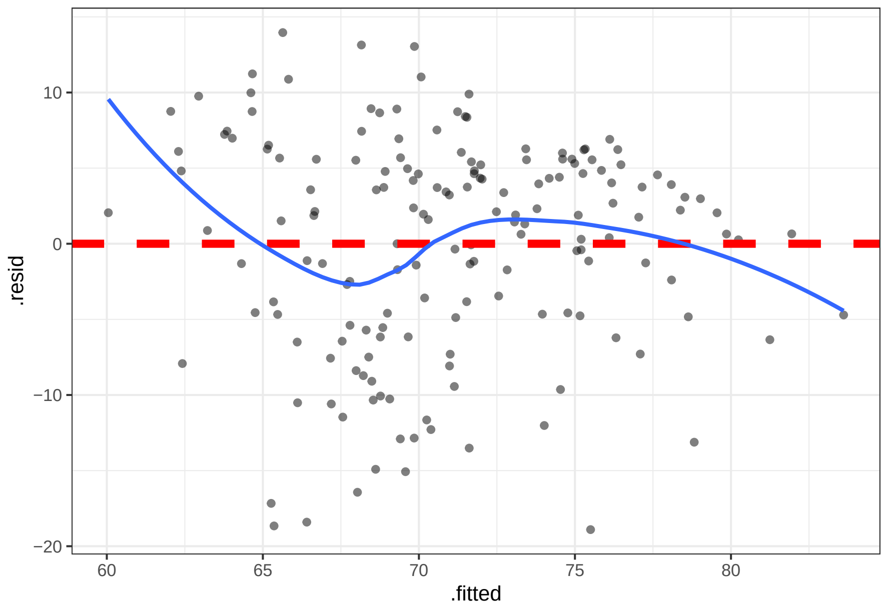


Residuals Against Predicted Values

- Residuals = observed - predicted
- Conditional on the predicted values, the residuals should have a mean of zero

- Residuals should NOT display any pattern

```
clean_gapminder %>%  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red", size = 2) +
  # To plot the residual mean
  geom_smooth(se = FALSE) +
  theme_bw()
```



Multiple Regression

Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$

where number of observations is greater than number of parameters being estimated.

```
multiple_lm <- lm(life_expectancy ~ log_gdp + fertility, data = clean_gapminder)
```

Use the adjusted R-squared when including multiple variables

- Adjusts for the number of parameters and number of observations being estimated by the model
- Adding more variables **will always increase** the Multiple R-squared

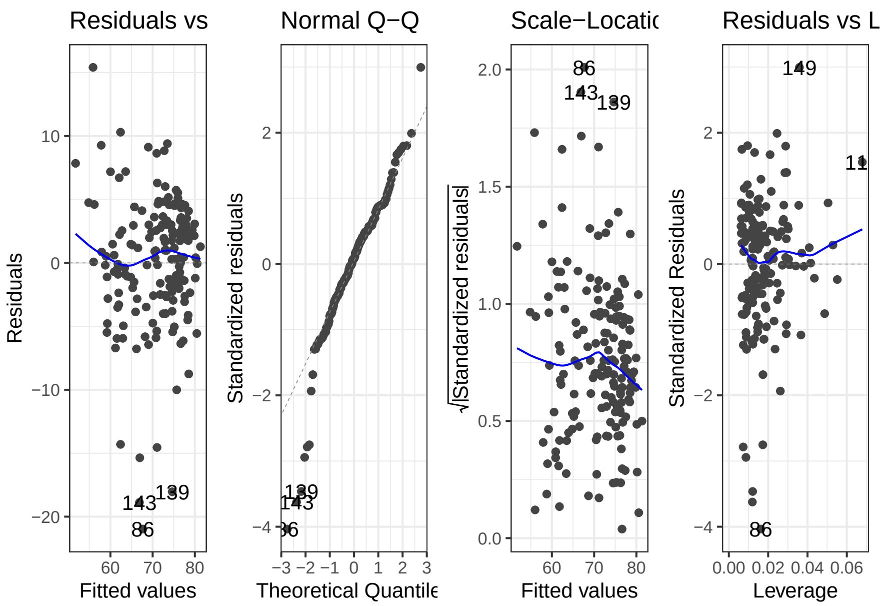
By assuming $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, what we really mean is:

$$Y \overset{\text{iid}}{\sim} N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p, \sigma^2)$$

Unbiased estimate: $\hat{\sigma}^2 = \frac{RSS}{n-(p+1)}$, degrees of freedom n - (p + 1). I.e., data supplies us with n degrees of freedom and we used up p + 1.

Checking the Assumptions about Normality with 'ggfortify'

```
autoplot(multiple_lm, ncol = 4) +
  theme_bw()
```



- standardized residuals = residuals/sd(residuals) which is equivalent to `.std.resid` from `augment()`.