Assessing the Accuracy of the Model: Takeaways

Syntax

EVALUATING BIVARIATE RELATIONSHIPS

- Visualize distribution of residuals with a histogram:

```
library(ggplot2)
residuals_df <- data.frame(lm_fit$residuals)
ggplot(data = residuals_df,
    aes(x = lm_fit.residuals)) +
    geom_histogram()
```

- View linear model summary:

```
summary(lm_fit)
```

- Manually estimating the t-statistic:

```
(lm_fit$coefficients[[2]] - 0)/ coef(summary(lm_fit))[, 2][[2]]
```

- Extract the p-value from a bi-variate linear model summary:

```
p_value <- coef(summary(lm_fit))[, 4][[2]]
```

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

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

- Extract RSS from model output:

```
RSS <- deviance(lm_fit)
```

- Manually estimate the residual standard error (RSE):

```
RSE <- sqrt(RSS / (nrow(df) - 2))
```

- Extract RSE from model output:

```
RSE <- sigma(lm_fit)
```

- Manually estimate total sum of squares (TSS):

```
TSS <- sum((df$response - mean(df$response))^2)
```

- Manually estimate r-squared:

```
r_squared <- 1 - RSS/TSS
```

- Extract r-squared value from linear model object:

```
r_squared <- summary(lm_fit)$r.squared
```

- Extract adjusted r-squared from linear model object:

```
adj_r_squared <- summary(lm_fit)$adj.r.squared
```

Equations

- Mathematical equation for hypothesis test:
    - Null hypothesis =
    - Alternative hypothesis =
- Mathematical equation for the t-statistic:
    -
- A 95% confidence interval for the intercept is *approximately* equal to:
    -

- Alternatively:
    -
- And the 95% confidence interval for the slope *approximately* equals:
    -
- Residual sum of squares (RSS):
    -
- Residual standard error is:
    -
- Total sum of squares (TSS):
    -
- R-squared:
    -

Concepts

- **Null hypothesis :** there is no relationship between predictor variable and the response variable.
- **Alternative hypothesis :** there is a relationship between predictor variable and the response variable.
- **t-statistic:** is the number of standard deviations that is from 0.
- **p-value:** is the probability of observing any value equal-to or larger than if the null hypothesis is true. A smaller p-value is better.
- **Confidence interval:** a confidence interval of 95% means that there is a 95% probability that the true unknown value of the coefficient will fall within the specified range.
- **Residual standard error ( ):** represents the average amount that our response variable measurements deviate from the true regression line. The is an estimate of the standard deviation of .
- **R-squared ( ):** a measure of the proportion of the variability in the response variable that can be explained by the predictor variable. The value falls between 0 and 1.

Resources

- Dataquest blog post on linear regression for predictive modeling in R.
- Dataquest blog post on linear regression error metrics.
- Wikipedia entry on the null hypothesis.
- Wikipedia entry on the t-statistic.
- Wikipedia entry on the t-distribution.

- Wikipedia entry on the coefficient of determination (r-squared).
- Wikipedia entry on the total sum of squares.
- An Introduction to Statistical Learning with Applications in R by James et al.