Residual Standard Error and R^2

Summary

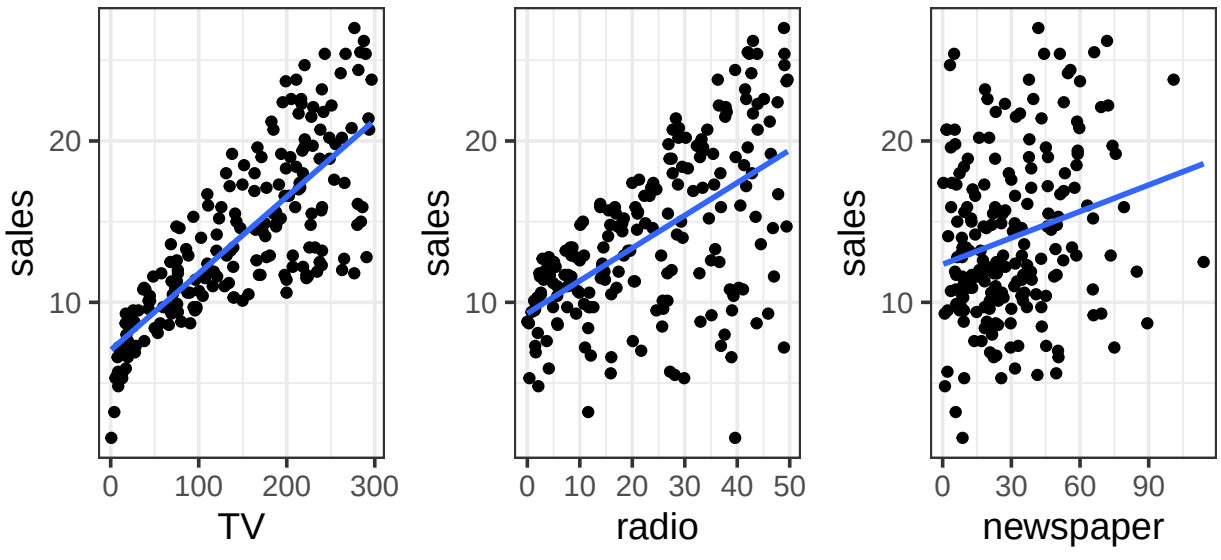
- We want to measure how useful a linear model is for predicting the response variable.
- The residual standard error is the standard deviation of the residuals
  - Smaller residual standard error means predictions are better
- The R^2 is the square of the correlation coefficient r
  - Larger R^2 means the model is better
  - Can also be interpreted as "proportion of variation in the response variable accounted for by the linear model" - see later statistics classes or the book for why.

Example data set

We have a data set with observations of four variables measuring advertising budgets and sales for a product in each of 200 markets:

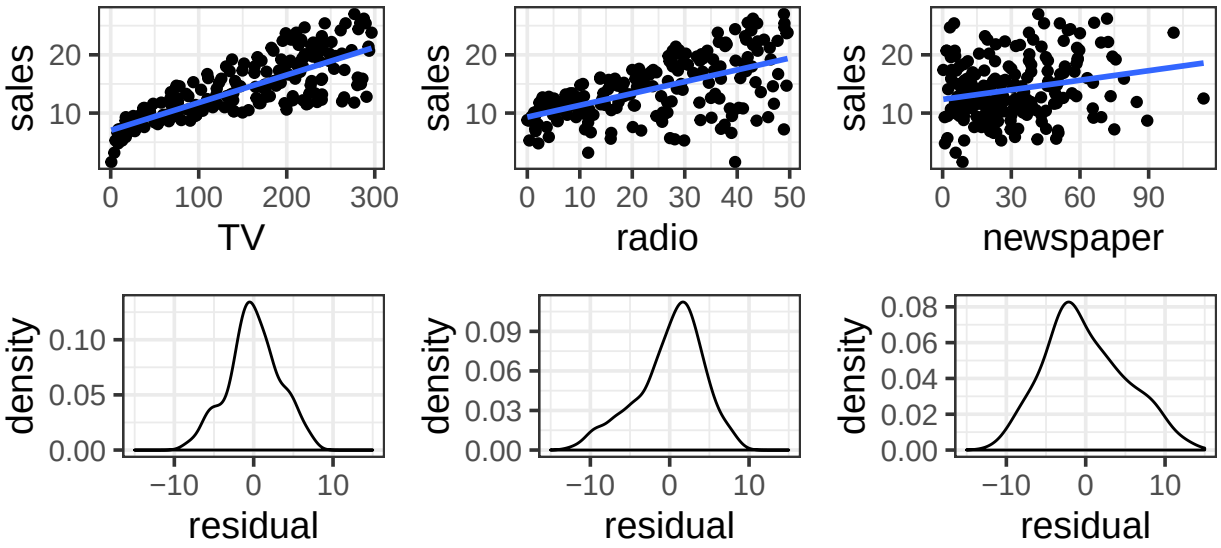
- `sales` is a measure of sales volume in thousands of units
- `TV` is TV advertising budget in thousands of dollars
- `radio` is radio advertising budget in thousands of dollars
- `newspaper` is newspaper advertising budget in thousands of dollars

Below are plots displaying three separate simple linear regression fits to the data. In all three plots/models, `sales` is the response variable; the explanatory variable is different for each model.



Residual Standard Error

- Residuals: $e_i = y_i - \hat{y}_i$ (vertical distance between point and line)
- Smaller residuals mean the predictions were better.
- Measure "size" of residuals with the standard deviation. For reasons discussed later, call this the residual standard error.



Residual standard errors:

- For regression on TV: 3.26
- For regression on radio: 4.28
- For regression on newspaper: 5.09

Recall that:

- If a variable follows an approximately normal distribution, about 95% of observations are within 2 standard deviations of the mean.
- The mean of the residuals is 0.

1. What is the interpretation of the residual standard deviation for the regression using TV as the explanatory variable, based on the "2 standard deviations rule"?

2. What is the interpretation of the residual standard deviation for the regression using radio as the explanatory variable, based on the "2 standard deviations rule"?

3. What is the interpretation of the residual standard deviation for the regression using newspaper as the explanatory variable, based on the "2 standard deviations rule"?

R^2

Square of the correlation coefficient r: between 0 and 1, closer to 1 is better.

- For regression on TV: 0.61
- For regression on radio: 0.33
- For regression on newspaper: 0.05

```
Advertising %>% select(TV, sales) %>% cor()
```

```
##                TV     sales
## TV     1.0000000 0.7822244
## sales 0.7822244 1.0000000
```

```
(0.7822244)^2
```

```
## [1] 0.611875
```

Finding things in the R output

```
fit <- lm(sales ~ TV, data = Advertising)
summary(fit)
```

```
##
## Call:
## lm(formula = sales ~ TV, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```