Linear regression (review)

Outline

- Linear regression formulation
- Assessing the accuracy of the coefficient estimates
- Assessing the accuracy of the model
- Predictions with the OLS model
- Potential problems
- Exercises

Linear regression formulation

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

- Epsilon comprises errors such as (i) missing variables that cause variation in the response Y, (ii) noise, and (iii) non-linear true relationship between the response and the predictors.

- OLS assumptions
  - The relationship is linear
  - No 2 predictors are perfectly correlated
  - Epsilon has zero mean correlated
  - Epsilon is uncorrelated with all predictors
  - Observations of epsilon are uncorrelated with each other
  - Epsilon has a constant variance
  - Epsilon is normally distributed (optional)

Linear regression formulation

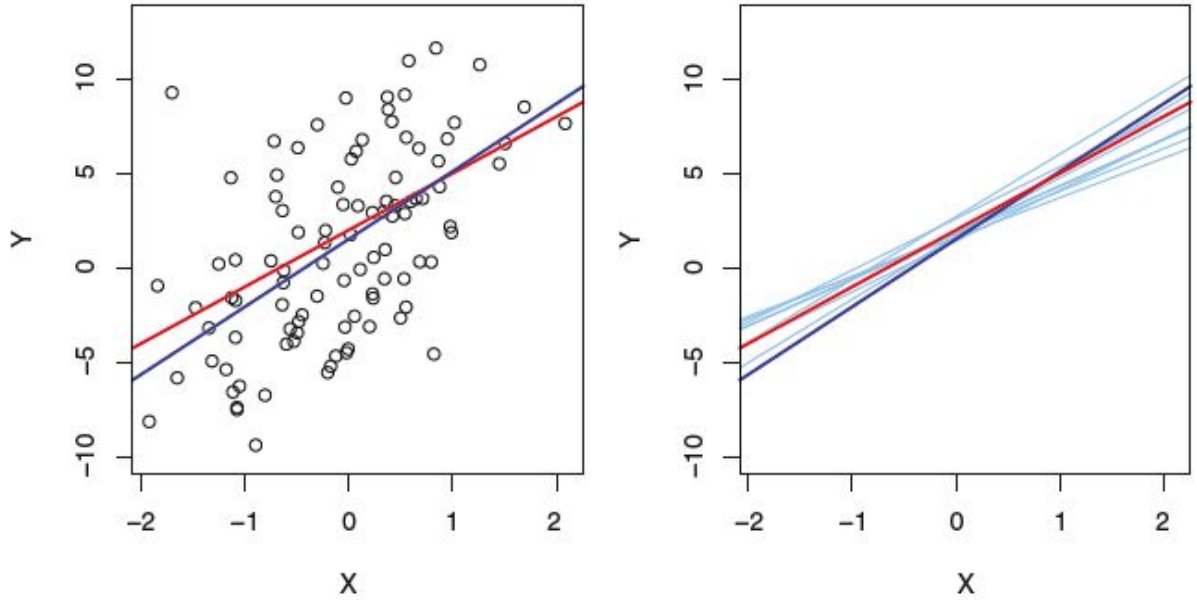
- OLS yields coefficients (beta values) that minimises the residual sum of squares, defined as

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

- Epsilon is often/always unknown, best approximated by the residual $(y_i - \hat{y}_i)$

Assessing the accuracy of the coefficient estimates

Sample regression model and population regression model



Assessing the accuracy of the coefficient estimates

The 95% confidence interval of coefficients

For simple linear regression: $\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$ $\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$.

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sigma^2 = \mathrm{Var}(\epsilon)$$

The variance of the error term and is approximated by square of the residual standard error (RSE)

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}}$$

[Note: factor 2 is an approximate - should be the 97.5% quantile of a t-distribution with n-2 degree of freedom]

Assessing the accuracy of the coefficient estimates

Hypothesis testing $H_0 : \beta_1 = 0 \qquad H_a : \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

t-statistic follows a Student distribution with a degree of freedom of n-2, based on which an associated p-value is calculated.

- What if we have a large beta and a large p-value?
- What if we have a very small beta but also a very small p-value?

Assessing the accuracy of the coefficient estimates

For multiple linear regression

- Standard error, 95% confidence interval, t-statistic (and p-value) of the coefficients are calculated similarly (by softwares)
- Additional hypothesis testing: F-test for overall significance

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a : \text{ at least one } \beta_j \text{ is non-zero}$$

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)} \qquad \text{TSS} = \sum(y_i - \bar{y})^2$$

F-statistic follows an F-distribution, based on which an associated p-value can be calculated.
If this p-value is small (say, less than 5%), we can accept Ha and conclude at least 1 predictor is related to the response.
But how do we know which one? => Feature selection

Assessing the accuracy of the OLS model

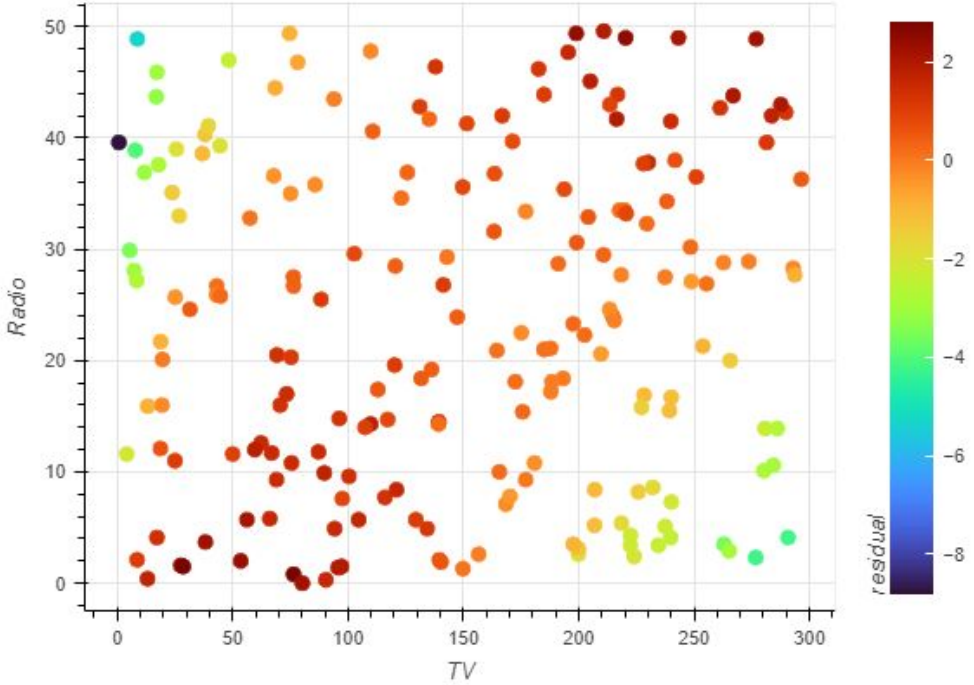
- R-squared: proportion of variance (of the response) explained

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- Residual standard error

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}$$

- Plotting the residual



Predictions with the OLS model

Potential sources of errors

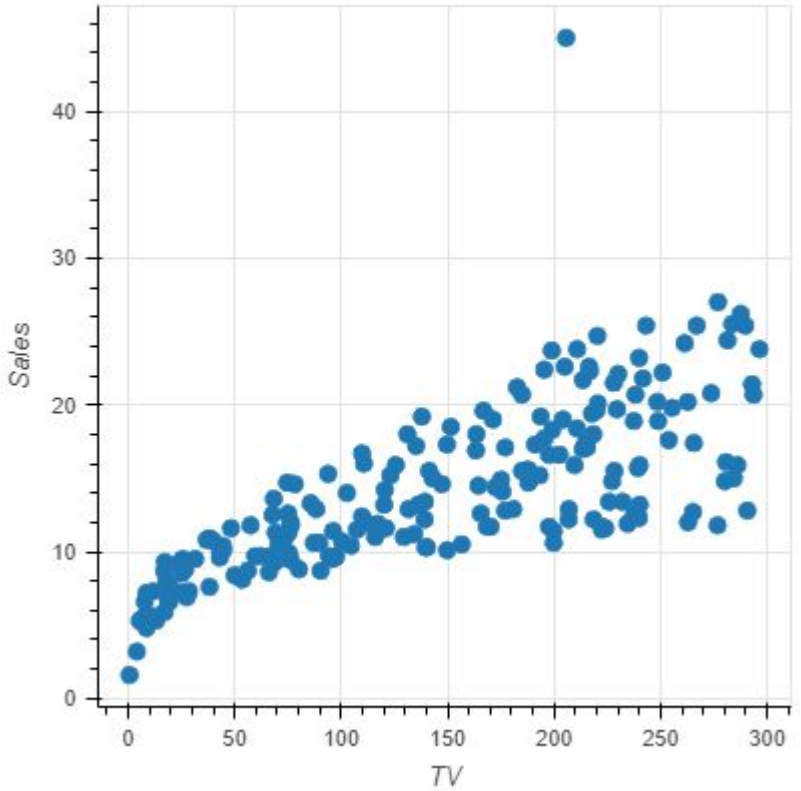
- Inaccuracy of the coefficient estimates (reducible errors) -> use *confidence interval* to determine how close the estimated response is to the true response.

- Model bias (the true relationship is not linear)

- Random error, e.g. noise (irreducible error) -> use *prediction interval* to how close the estimated response is to the observed response.

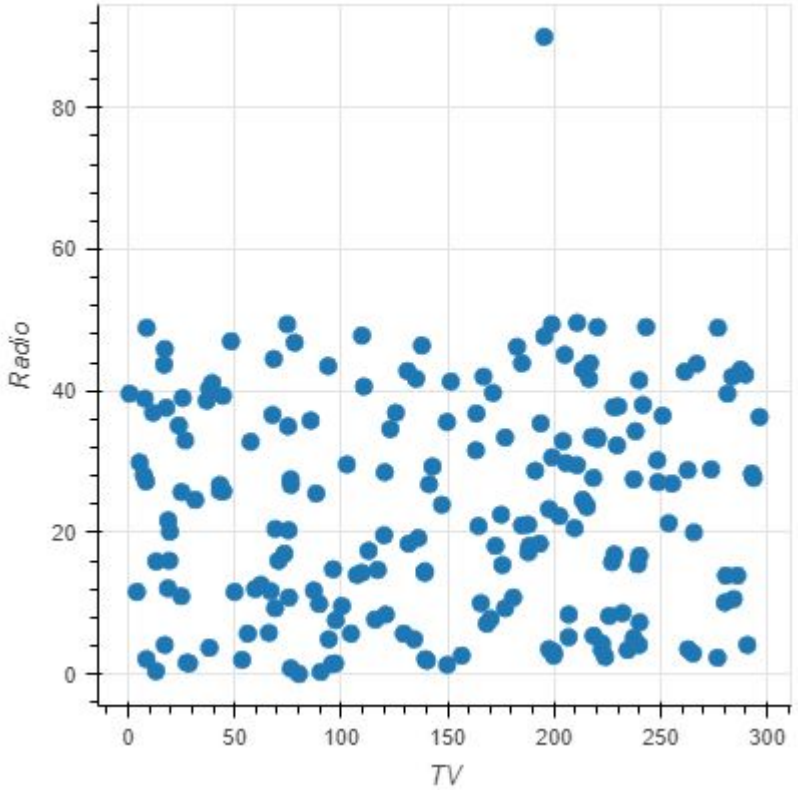
Potential problems - Outliers

An outlier is a data point for which *the response* is far from the value predicted by the fitted model. They can be identified by the residual plot.

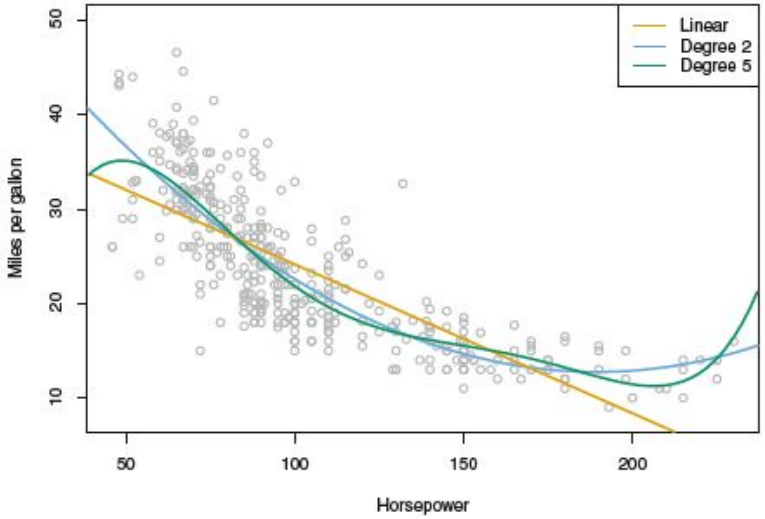


Potential problems - High leverage points

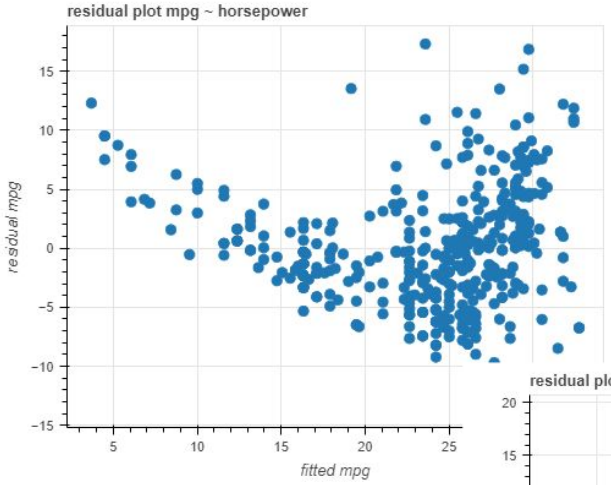
A high leverage point is a data point for which have *unusual predictors* (independent variables).

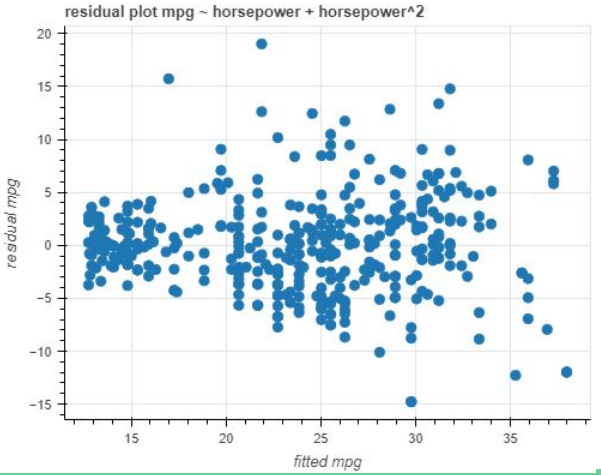


Potential problems - Non-linear relationship



$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$





Exercises

Use Google Colab https://colab.research.google.com/

- Write Python code to regress Sales on TV, Radio and the interaction term (Radio x TV)
- Plot of residual against the fitted value
- Compare R-squared and RSE from this new model against the model without the interaction term.
- Should we worry about collinearity?