Regression Analysis with Linear Models

Dr. Lan Du

Faculty of Information Technology, Monash University, Australia

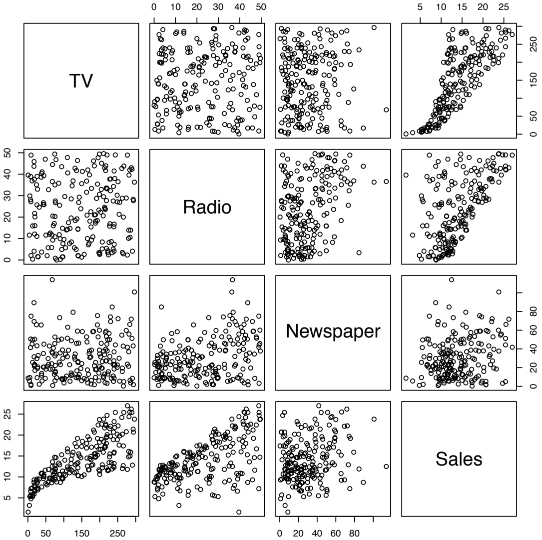
FIT5149 week 3

Outline

Outline

Linear Regression for the Advertising Data in ISL

Consider the advertising data used in "Introduction to Statistical Learning".

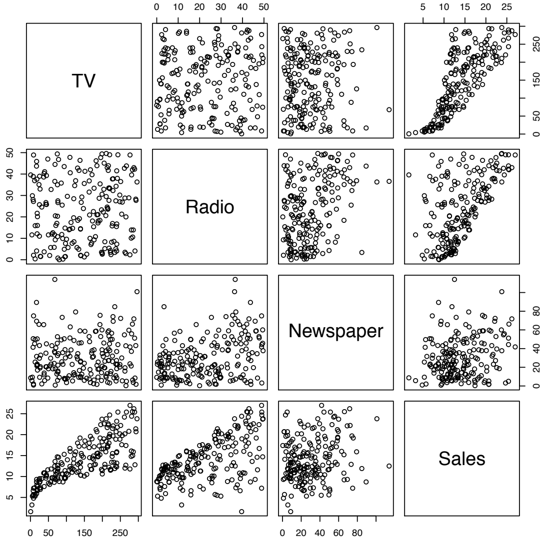


Questions we might ask:

- Is there a relationship between advertising budget and sales?

Linear Regression for the Advertising Data in ISL

Consider the advertising data used in "Introduction to Statistical Learning".

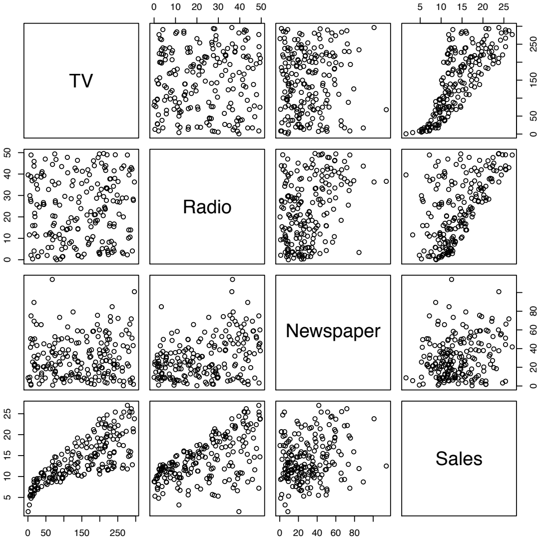


Questions we might ask:

- How strong is the relationship between advertising budget and sales?

Linear Regression for the Advertising Data in ISL

Consider the advertising data used in "Introduction to Statistical Learning".

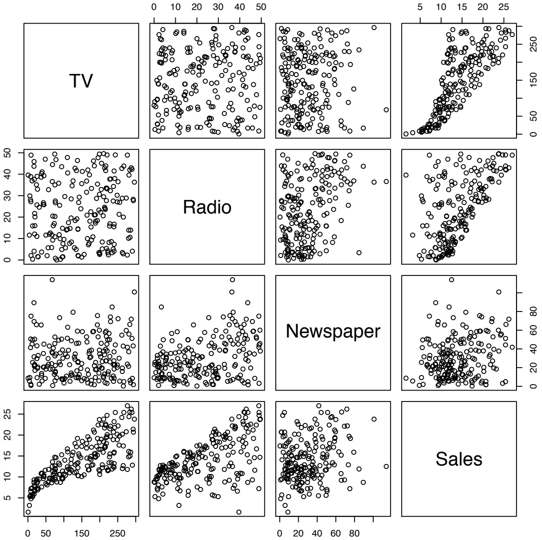


Questions we might ask:

- Which media contribute to sales?

Linear Regression for the Advertising Data in ISL

Consider the advertising data used in "Introduction to Statistical Learning".

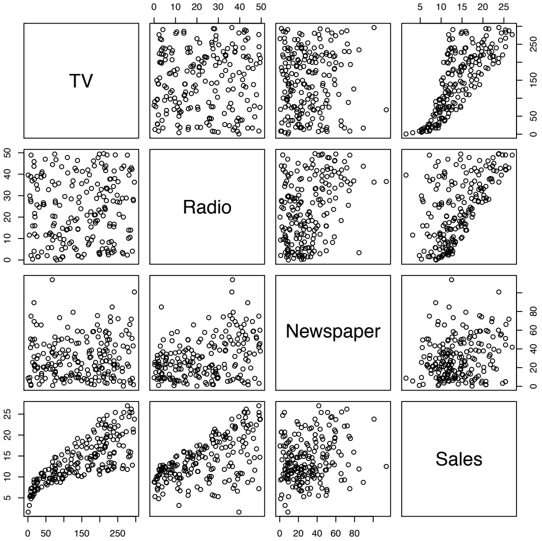


Questions we might ask:

- How accurately can we predict future sales?

Linear Regression for the Advertising Data in ISL

Consider the advertising data used in "Introduction to Statistical Learning".

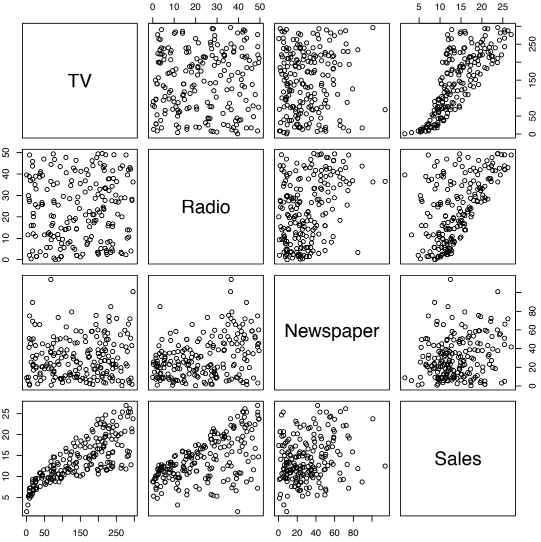


Questions we might ask:

- Is the relationship linear?

Linear Regression for the Advertising Data in ISL

Consider the advertising data used in "Introduction to Statistical Learning".



Questions we might ask:

- Is there synergy among the advertising media?

Simple Linear Regression

- Simple linear regression is a statistical method that allows us to predict a quantitative response Y on the basis of a single predicator Variable X. It assumes the relationship between Y and X can be model by a straight line:

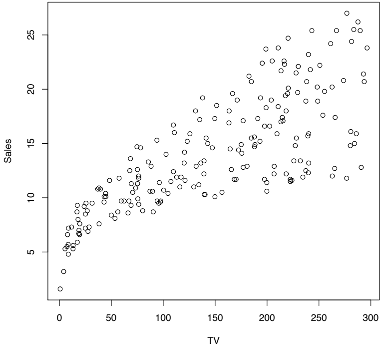
$$Y = \beta_0 + \beta_1 X + \epsilon$$

  where
  - β_0: the expected value of Y when $X = 0$.
  - β_1: the average change in Y for a 1-unit change in X.
  - ϵ: error term describes the random component of the linear relationship.
- Assumptions:
  - **Linearity**: The response variable Y has a linear relationship to the predictor variable X.
  - **Nearly normal residuals**: The errors must be independent and normally distributed.

$$\epsilon \sim \mathcal{N}(0, \sigma^2 \cdot I_{n \times n})$$

  - **Constant variability**: The Variance of the residuals is constant.

Example: Advertising data



$$\text{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 \times \text{TV}$$

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients,
  - Inference: describe the linear dependency between sales and budgets for TV advertisement.
  - Prediction: predict future sales given a budget plan for TV advertisement,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{y}$ indicates the prediction of Y on the basis of $X = x$.

The "Ordinary Least Squares" Regression.

- Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ be the prediction for Y based on the ith value of X. The ith residual (i.e., error) is defined as

$$e_i = y_i - \hat{y}_i$$

- We define the residual sum of squares (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2$$

or equivalent as

$$\begin{aligned}\text{RSS} &= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\end{aligned}$$

- The least square approaches chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimise the RSS.

How to Fit a Regression model with RSS in R

- The lm() function performs a least squares regression and creates a linear model object:

```
advData = read.csv("./Advertising.csv")
advData <- advData[c("TV", "Radio", "Newspaper", "Sales")]
attach(advData)
lmfit = lm(Sales~TV)
lmfit
```

```
Call:
lm(formula = Sales ~ TV)

Coefficients:
(Intercept)           TV
    7.03259      0.04754
```

where:

- Models for lm() are specified symbolically: *response* $\sim$ *predictor*
- The intercept $\hat{\beta}_0 = 7.026$ and the slope $\hat{\beta}_1 = 0.0475$

- The linear model object contains much more information that just the coefficients!

Interpret Simple Linear Regression Model

MONASH University

```
summary(lmfit)
```

```
Call:
lm(formula = Sales ~ TV)

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
```

Assessing the Accuracy of the Coefficient Estimates

```
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
```

- Coefficient - Std. Error: measures how precisely the model estimates the coefficient's unknown value.
  - $SE(\hat{\beta}_0) = 0.457843$: in the absence of any advertising, the average sales can vary by 457.843 units.
  - $SE(\hat{\beta}_1) = 0.002691$: for each \$1,000 increase in television advertising, the average increase in sales can vary by 2.691 units.

Assessing the Accuracy of the Coefficient Estimates

```
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
```

- Coefficient - Std. Error: measures how precisely the model estimates the coefficient's unknown value.
  - These standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form

    $$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

    That is, there is approximately a 95% chance that the interval

    $$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

    will contain the true value of β_1.

Assessing the Accuracy of the Coefficient Estimates

```
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
```

- Coefficient - Std. Error: measures how precisely the model estimates the coefficient's unknown value.
  - In the case of the advertising data
    - The 95% confidence interval for β_0 is $[6.130, 7.935]$
    - The 95% confidence interval for β_1 is $[0.042, 0.053]$
  - Use the confidence interval to assess the reliability of the estimate of the coefficient.
  - Standard errors can also be used to perform hypothesis tests on the coefficient.
    - H_0: There is no relationship between X and Y, i.e., $\beta_1 = 0$
    - H_a: There is some relationship between X and Y, i.e., $\beta_1 \neq 0$

Assessing the Accuracy of the Coefficient Estimates

```
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
```

- Coefficient - t statistics

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

which measures the number of standard deviations that $\hat{\beta}_1$ is away from 0.

  - Large t value indicates the null hypothesis could be rejected.
  - Small t value indicates rejecting the null hypothesis could cause a type-I error.

Question: How large is large?

Assessing the Accuracy of the Coefficient Estimates

```
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
```

- Coefficient - Pr(>|t|) (i.e., p-value): test for the predicative power of predictor variable, i.e., TV
  - Small p-value ($Pr(> |t|) < \alpha = 0.001$): reject the null hypothesis
    - Changes in the predictor's value are related to changes in the response variable.
  - Use the coefficient p-values to determine which terms to keep in the regression model.

Assessing the Accuracy of the Model

```
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
```

- Residual standard error (RSE): an estimate of the standard deviation of residuals, i.e., ϵ.

$$\text{RSE} = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

  - A measure of the quality of a linear regression fit, or a measure of the lack of fit of the model
  - The advertising data: $RSE = 3.259$.
    - Actual sales in each market deviate from the true regression line by approximately 3,259 units, on average.
    - The percentage error:

$$3,259/14,000 = 23\%$$

    where 14,000 is the mean value of sales.

Assessing the Accuracy of the Model

```
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
```

- The Coefficient of Determination (i.e., the R^2 statistic): measures the proportion of variability in Y that can be explained using X.

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

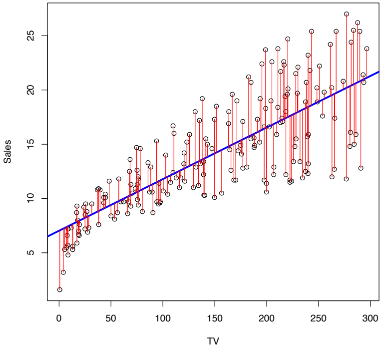
where the total sum of squares (TSS) is $\sum_{i=1}^{n}(y_i - \bar{y})^2$, and RSS is $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- $0 \leq R^2 \leq 1$: the larger R^2 is the better the model is fitting the actual data.
- The advertising data: $R^2 = 0.6119$.

Normality: are residuals normally distributed?

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124
```

Residuals are essentially the difference between the observed response values and the response values predicted by the model.



- Ideally, residuals should be normally distributed.

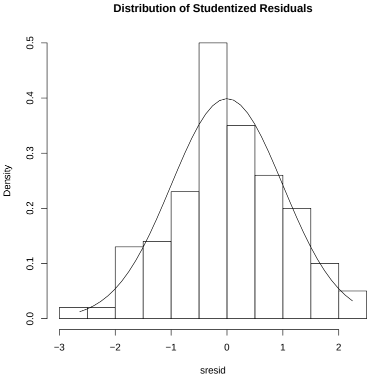
$$\mathbb{E}(\epsilon) = 0$$

- When assessing how well the model fit the data, you should look for a symmetrical distribution across these points on the mean value 0.

Normality: are residuals normally distributed?

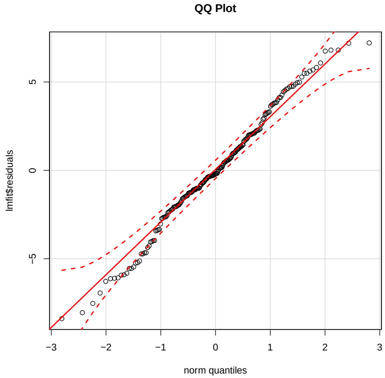
MONASH University

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124
```

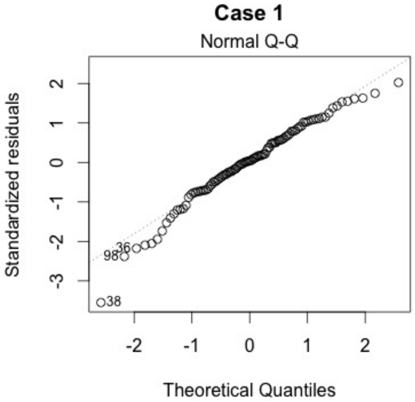


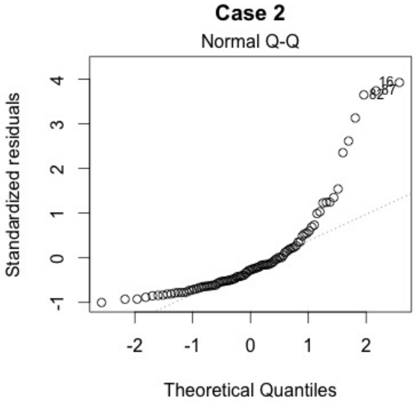
Normality: are residuals normally distributed?

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124
```



Normality: are residuals normally distributed?





Linearity: is the relationship between predictor and response variables linear?

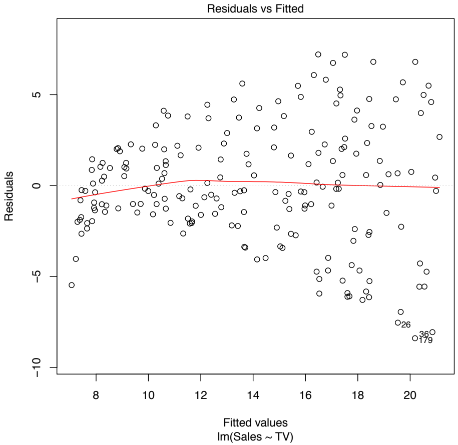
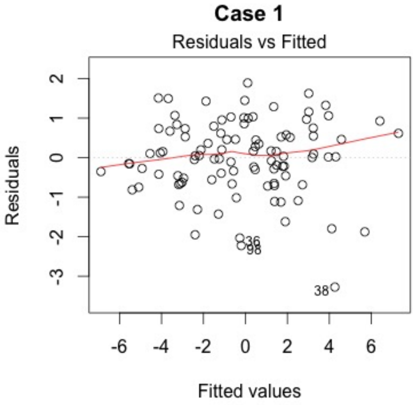
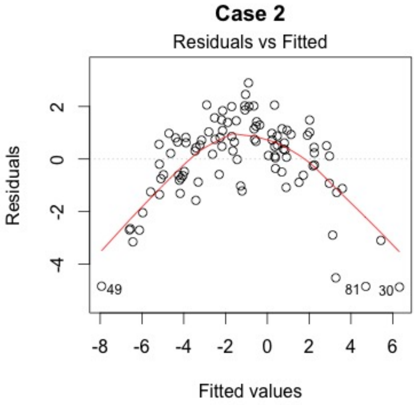


Figure: Plots of residuals versus predicted (or fitted) values for the Advertising

- If residuals are equally spread around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships.
- what is the difference in the plots between linear models trained on datasets
  - the relationship between predictors and the response variable is linear.
  - the relationship between predictors and the response variable is not linear.

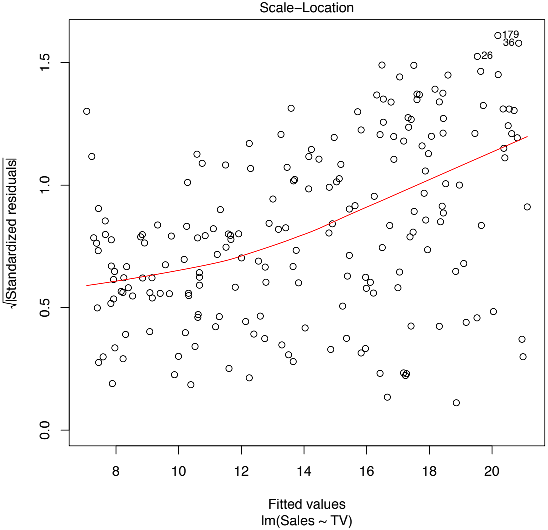
Linearity: is the relationship between predictor and response variables linear?





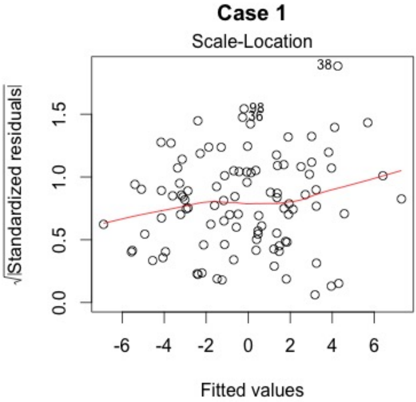
Constant variance (homoscedasticity)

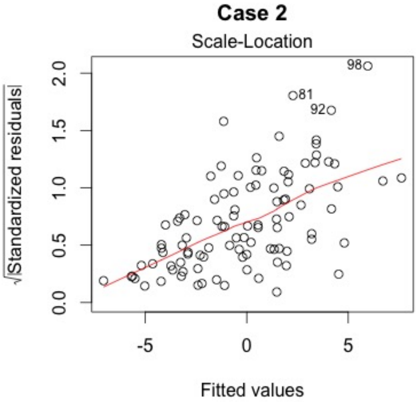
Are residuals spread equally along the ranges of predictors



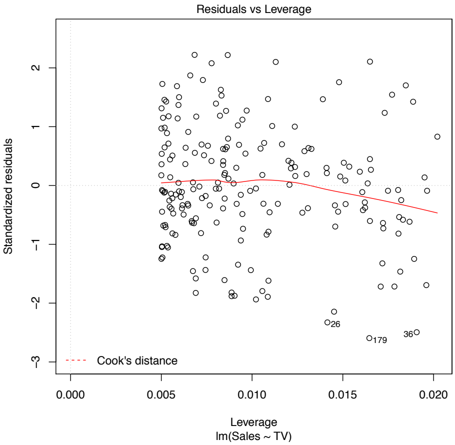
Constant variance (homoscedasticity)

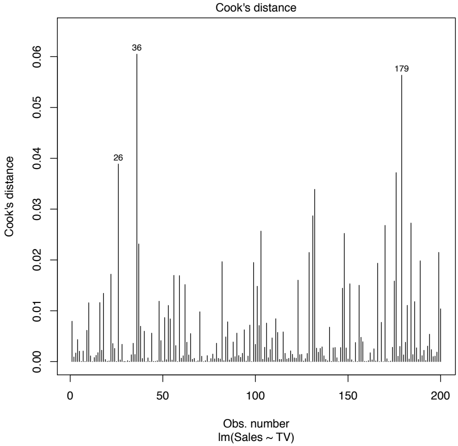
Are residuals spread equally along the ranges of predictors



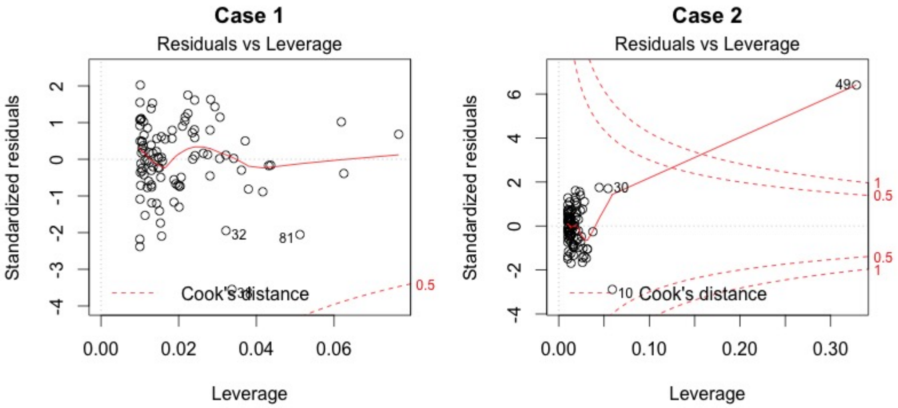


Residuals v.s. Leverage: what are the influential data sample in the fitting?





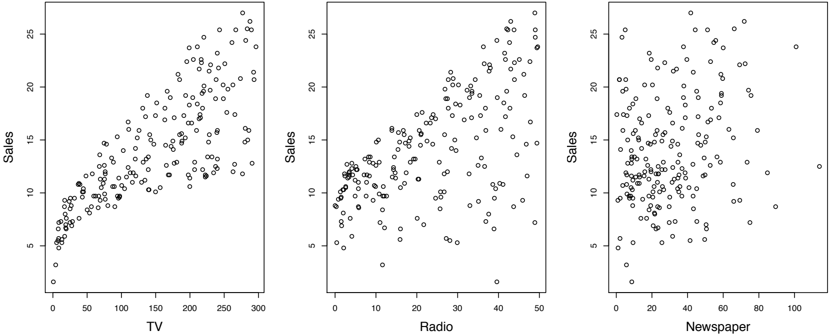
Residuals v.s. Leverage: what are the influential data sample in the fitting?



Watch out for outlying values at the upper right corner or at the lower right corner.

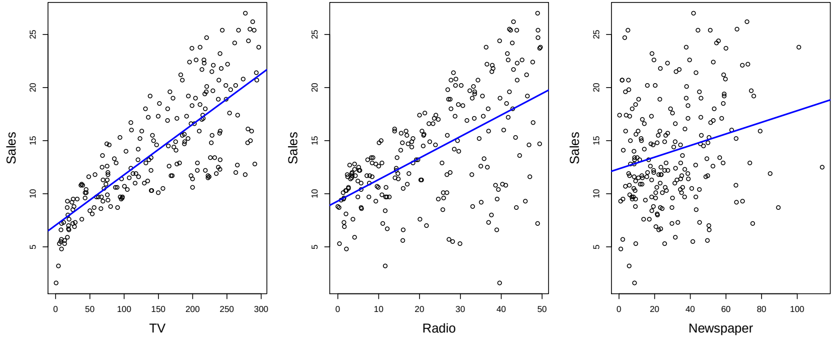
Outline

Example: the Advertising Data



- How we can extend our analysis of the advertising data in order to accommodate these two additional predictors?

Example: the Advertising Data



- Problems
  - Predict sales given the three advertising media budgets.
  - Ignore the correlation between the predictors, TV, Radio and Newspaper.

Multiple Linear Regression

- The multiple linear regression model:

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

  where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$
- β_j: the average effect on Y of a one unit increase in X_j, holding all other predictors fixed.
- In the advertising example, the model becomes

$$Sales = \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times Newspaper + \epsilon$$

Estimating the Regression Coefficients

- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make prediction using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

- We estimate $\beta_0, \beta_1, \ldots, \beta_p$ as the values that minimise the sum of squared residuals

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_p x_p\right)^2 \end{aligned}$$

This can be done using standard statistical software.

Results for Advertising Data

- Results:

```
mllg = lm(Sales~., data = advData)
summary(mllg)
```

```
Call:
lm(formula = Sales ~ ., data = advData)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Results for Advertising Data

Compare simple linear regression with multiple linear regression:

	Coefficients	Std. error	t value	p-value
Intercept	7.032594	0.457843	15.36	<2e-16
TV	0.047537	0.002691	17.67	<2e-16
Intercept	9.31164	0.56290	16.542	<2e-16
Radio	0.20250	0.02041	9.921	<2e-16
Intercept	12.35141	0.62142	19.88	< 2e-16
Newspaper	0.05469	0.01658	3.30	0.00115
Intercept	2.938889	0.311908	9.422	<2e-16
TV	0.045765	0.001395	32.809	<2e-16
Radio	0.188530	0.008611	21.893	<2e-16
Newspaper	-0.001037	0.005871	-0.177	0.86

The multiple linear regression suggests that there is no relationship between sales and newspaper while the simple linear regression implies the opposite.

Results for Advertising Data

Correlation matrix for TV, Radio, Newspaper, and sales.

```
cor(advData)
```

	TV	Radio	Newspaper	Sales
TV	1	0.0548086644658301	0.056647874965057	0.782224424861606
Radio	0.0548086644658301	1	0.354103750761175	0.576222574571055
Newspaper	0.056647874965057	0.354103750761175	1	0.228299026376165
Sales	0.782224424861606	0.576222574571055	0.228299026376165	1

- The correlation between radio and newspaper is 0.354.
  - A tendency to spend more on newspaper advertising in markets where more is spent on radio advertising.

Some Important Questions

- Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

F-Statistics

- Is there a relationship between the response and predictors?
  - Hypothesis testing
    - Null hypothesis: There is no relationship between Y and $X_1, X_2, \ldots, X_p$.

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

    - The alternative: There is at least one X_j related to Y.

$$H_a : \beta_j \neq 0, \exists j \in [1, p]$$

  - F-statistics: a good indicator of whether there is a relationship between our predictor and the response variables.

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

    - F-value close to 1: no relationship between Y and $X_1, X_2, \ldots, X_p$.
    - F-value greater than 1: a relationship between our predictor and the response variables.

F-Statistics

- Is there a relationship between the response and predictors?
  - For the F-value, how large is large?
    - n is large: small F-value provides strong evidence against H_0.
    - n is small: large F-value is needed.
  - Example: multiple linear regression on the advertising dataset

```
Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Model Fit

- How well does the model fit the data?

```
mllg = lm(Sales~., data = advData)
anova(mllg)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TV	1	3314.61816686865	3314.61816686865	1166.73075736576	1.80933660946536e-84
Radio	1	1545.61660306373	1545.61660306373	544.050125566422	1.88272164339468e-58
Newspaper	1	0.0887171654310898	0.0887171654310898	0.0312280451031693	0.859915050080576
Residuals	196	556.825262902188	2.84094521888871	NA	NA

Model Fit

- How well does the model fit the data?

```
lm1 = lm(Sales ~ TV)
lm2 = lm(Sales ~ TV + Radio)
lm3 = lm(Sales ~ TV + Radio + Newspaper)
anova(lm1,lm2, lm3, test="Chisq")
```

	Res.Df	RSS	Df	Sum of Sq	Pr(>Chi)
1	198	2102.53058313135	NA	NA	NA
2	197	556.913980067619	1	1545.61660306373	2.47937530553987e-120
3	196	556.825262902188	1	0.088717165431035	0.859732582779944

Confidence interval v.s. Prediction interval

MONASH University

- Given a set of predictor values, what response value should be predict, and how accurate is our prediction?
  - Prediction: given the estimated coefficients, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$,

$$\hat{Y} = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i X_i$$

```
lm = lm(Sales ~ TV + Radio, data = advData)
set.seed(0)
newTV <- runif(10, min(TV), max(TV))
newRadio <- runif(10, min(Radio), max(Radio))
nData = data.frame(TV = newTV, Radio = newRadio)
predict(lm, newdata=nData)
```

	TV	Radio
1	265.853362080292	3.0645990151912
2	79.210911691119	10.2163389150053
3	110.737037122599	8.75721492543817
4	170.092739543156	34.076333194226
5	269.257043501455	19.0515444234014
6	60.3373470077757	38.1841344319284
7	266.353829844948	24.6858824074268
8	280.040476926602	35.5938780099154
9	196.09790723836	49.198542303592
10	186.729022780899	18.8497448999435

1	15.6612982601506
2	8.46599326403736
3	9.63415841804624
4	17.1098156651422
5	18.8224865122765
6	12.8602208901317
7	19.7488735200483
8	22.4257437238907
9	21.1425653110146
10	15.0084950382048

Confidence interval v.s. Prediction interval

- Given a set of predictor values, what response value should be predict, and how accurate is our prediction?
  - To determine how close $\hat{Y}$ will be close to $f(X)$.
    - Confidence interval: use to quantify the uncertainty around the expected value of predictions (average of a group of predictions) — the uncertainty of predicting the average sales over a number of markets.

```
predict(lm, newdata=nData, interval = "confidence")
```

	fit	lwr	upr
1	15.6612982601506	15.1367929201415	16.1858036001596
2	8.46599326403736	8.10802220969713	8.82396431837759
3	9.63415841804624	9.29463411638258	9.9736827197099
4	17.1098156651422	16.8145727479215	17.405058582363
5	18.8224865122765	18.4051406295901	19.2398323949628
6	12.8602208901317	12.4435453179055	13.2768964623579
7	19.7488735200483	19.3467647872336	20.1509822528631
8	22.4257437238907	21.9584567202787	22.8930307275026
9	21.1425653110146	20.6566749977346	21.6284556242947
10	15.0084950382048	14.739150864574	15.2778392118356

Confidence interval v.s. Prediction interval

- Given a set of predictor values, what response value should be predict, and how accurate is our prediction?
  - To determine how close $\hat{Y}$ will be close to $f(X)$.
    - Prediction interval: use to quantify the uncertainty around a single prediction — e.g. the uncertainty of predicting sales given the budgets of TV and Radio adverting for a particular market.

```
predict(lm, newdata=nData, interval = "prediction")
```

	fit	lwr	upr
1	15.6612982601506	12.3042935926206	19.0183029276805
2	8.46599326403736	5.13094937199256	11.8010371560822
3	9.63415841804624	6.30104407294193	12.9672727631505
4	17.1098156651422	13.7809205231636	20.4387108071209
5	18.8224865122765	15.4805481418221	22.1644248827308
6	12.8602208901317	9.51836616279689	16.2020756174665
7	19.7488735200483	16.4088037724045	23.0889432676922
8	22.4257437238907	19.077202006984	25.7742854407974
9	21.1425653110146	17.7913768826121	24.4937537394172
10	15.0084950382048	11.681796859832	18.3351932165776

Outline

Linear Regression with Qualitative Predicators

MONASH University

- Some predictors are not quantitative but are qualitative, taking a discrete set of values.
- These are also called categorical predictors or factor variables.

	A	B	C	D	E	F	G	H	I	J	K	L
1		Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
2	1	14.891	3606	283	2	34	11	Male	No	Yes	Caucasian	333
3	2	106.025	6645	483	3	82	15	Female	Yes	Yes	Asian	903
4	3	104.593	7075	514	4	71	11	Male	No	No	Asian	580
5	4	148.924	9504	681	3	36	11	Female	No	No	Asian	964
6	5	55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331
7	6	80.18	8047	569	4	77	10	Male	No	No	Caucasian	1151
8	7	20.996	3388	259	2	37	12	Female	No	No	African Amer	203
9	8	71.408	7114	512	2	87	9	Male	No	No	Asian	872
10	9	15.125	3300	266	5	66	13	Female	No	No	Caucasian	279
11	10	71.061	6819	491	3	41	19	Female	Yes	Yes	African Amer	1350
12	11	63.095	8117	589	4	30	14	Male	No	Yes	Caucasian	1407
13	12	15.045	1311	138	3	64	16	Male	No	No	Caucasian	0
14	13	80.616	5308	394	1	57	7	Female	No	Yes	Asian	204
15	14	43.682	6922	511	1	49	9	Male	No	Yes	Caucasian	1081
16	15	19.144	3291	269	2	75	13	Female	No	No	African Amer	148
17	16	20.089	2525	200	3	57	15	Female	No	Yes	African Amer	0
18	17	53.598	3714	286	3	73	17	Female	No	Yes	African Amer	0
19	18	36.496	4378	339	3	69	15	Female	No	Yes	Asian	368
20	19	49.57	6384	448	1	28	9	Female	No	Yes	Asian	891

Figure: The credit card dataset that contain both quantitative variables (e.g., income, limit, rating, and age), and qualitative variables (e.g., gender, student, married, and ethnicity).

Linear Regression with Qualitative Predicators — continued

MONASH University

- Dummy coding — making many variables out of one
  - A categorical variable with k levels will be transformed into $k-1$ variables each with two levels.
  - For example, for the ethnicity variable we create two dummy variables. The first could be

$$x_{i,1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th persion is not Asian} \end{cases}$$

and the second could be

$$x_{i,2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th persion is not Caucasian} \end{cases}$$

Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American} \end{cases}$$

Linear Regression with Qualitative Predicators — continued

MONASH University

- Dummy coding — making many variables out of one
  - Example: the Credit dataset.

```
summary(lm(Balance ~ Ethnicity, data = credit_data))
```

```
Call:
lm(formula = Balance ~ Ethnicity, data = credit_data)

Residuals:
    Min      1Q  Median      3Q     Max
-531.00 -457.08  -63.25  339.25 1480.50

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          531.00      46.32  11.464   <2e-16 ***
EthnicityAsian       -18.69      65.02  -0.287    0.774
EthnicityCaucasian   -12.50      56.68  -0.221    0.826
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 460.9 on 397 degrees of freedom
Multiple R-squared:  0.0002188, Adjusted R-squared:  -0.004818
F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

Outline

Addictive and Linear assumptions

- Two of the most important assumptions on the relationship between predictors and response:

$$\widehat{Sales} = \beta_0 + \beta_1 \times TV + \beta_2 \times Radio$$

  - Additive — the effect of changes in X_j on Y is independent of X_i for $i \neq j$.
  - Linear — the change in Y due to one-unit change in X_j is constant, regardless of the value of X_j.
- Can we remove the additive assumption?

Interaction between variables

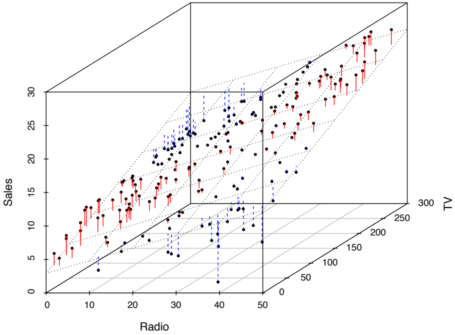


Figure: Over-estimate v.s. under-estimate without considering interaction between predictors

- Synergy effect (or interaction affect):
  - For example, given a fixed budget of $100, 000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio.
  - Spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.

Interaction between variables — continued

- Model with interaction terms takes the form

$$\begin{aligned} Sales &= \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times (TV \times Radio) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times Radio) \times TV + \beta_3 \times Radio + \epsilon \end{aligned}$$

  - β_3: the increase in the effectiveness of TV advertising for a one unit increase in radio advertising (or vice-versa)

Interaction between variables — continued

- results

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
Radio       2.886e-02  8.905e-03   3.241   0.0014 **
TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,    Adjusted R-squared:  0.9673
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

  - Strong evidence that $Ha : \beta_3 \neq= 0$: the true relationship is not additive

Interaction between variables — continued

MONASH University

- results

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
Radio       2.886e-02  8.905e-03   3.241   0.0014 **
TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,	Adjusted R-squared:  0.9673
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

- the R^2 and F-statistics:

	$Sales \sim TV + Radio$	$Sales \sim TV + Radio + TV * Radio$
R^2	0.8972	0.9678
	$(0.9678 - 0.8972)/(1 - 0.8972) \approx 69\%$	
$F - statistic$	859.6	1963

Interaction between variables — continued

- results

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
Radio       2.886e-02  8.905e-03   3.241   0.0014 **
TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,    Adjusted R-squared:  0.9673
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

  - Interpret coefficients:
    - An increase in TV advertising of \$1, 000 is associated with increased sales of

$$(\hat{\beta}_1 + \hat{\beta}_3 \times Radio) \times 1000 = 19 + 1.1 \times Radio$$

    - An increase in radio advertising of \$1, 000 will be associated with an increase in sales of

$$(\hat{\beta}_2 + \hat{\beta}_3 \times TV) \times 1000 = 29 + 1.1 \times TV$$

Interaction between variables — continued

- results

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
Radio       2.886e-02  8.905e-03   3.241   0.0014 **
TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,    Adjusted R-squared:  0.9673
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

  - The hierarchy principle: if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficient are not significant.

Summary

- What we have covered:
  - Simple linear regression with ordinary least squares
  - Various regression diagnostics
    - Assess the accuracy of the estimated coefficients
    - Assess the accuracy of the model
    - Residual analysis
  - Multiple linear regression
  - Categorial variables in regression
  - Extension of linear regression: interaction between variables
- What we haven't covered:
  - Outliers
  - High leverage points
  - Collinearity
  - Linear regression with K-Nearest Neighbors

See sections 3.3.3, 3.4 and 3.5 of "Introduction to Statistical learning"

Reference

- Reading materials:
  - "Linear Regression", Chapter 3 of "Introduction to Statistical Learning", 6th edition
  - "Linear Regression and ANOVA", Chapter 11 of "R Cookbook" by Paul Teetor, available online from Monash Library.
- Some figures in this presentation were taken from
  - "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani
  - `https://data.library.virginia.edu/diagnostic-plots/`
- Some of the slides are reproduced based on the slides from T. Hastie and R. Tibshirani