Line / Plane of best fit

Linear Regression

Linear in X (or some function of X)

ACTL3142 & ACTL5110 Statistical Machine Learning for Risk and Actuarial Applications

Disclaimer

Some of the figures in this presentation are 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

Overview

Line

- Simple Linear Regression
- Multiple Linear Regression
- Linear model Selection
- Potential problems with Linear Regression

Plane

Reading

James et al (2021), Chapter 3, Chapter 6.1

Linear Regression

- A classical and easily applicable approach for supervised learning
- Useful tool for predicting a quantitative response – Non categorical
- Many more advanced techniques can be seen as an extension of linear regression

Simple Linear Regression

Overview

$$Y = f(X) + \varepsilon$$

Output

Data

- Predict a quantitative response Y based on a single predictor variable X
- Approximately a linear relationship between X and Y

$$Y = \underline{\beta_0 + \beta_1 X} + \epsilon$$

$$f(X) = \beta_0 + \beta_1 X$$

- Use (training) data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
- Make predictions of Y_i (given $X = x_i$)

X is one-dimen here.

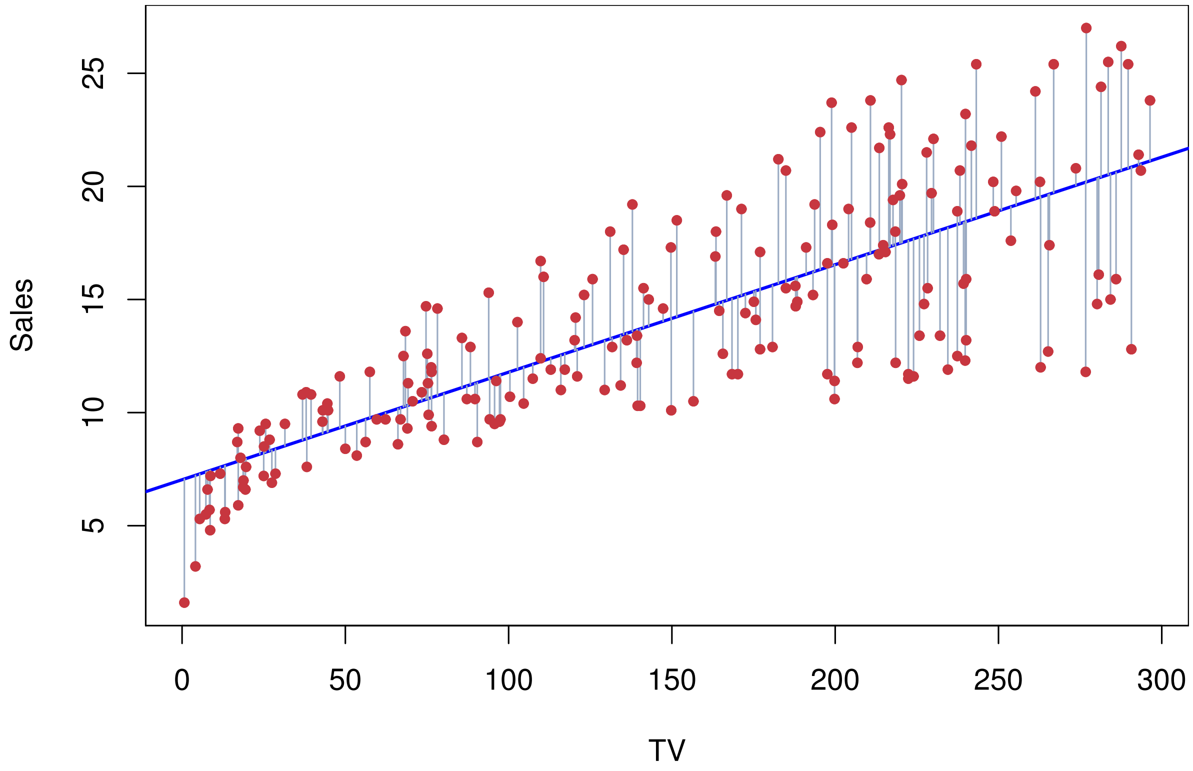
$$E[Y] = \beta_0 + \beta_1 E[X] + \cancel{E[\varepsilon]} \to 0 \longrightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Simple and easy to understand

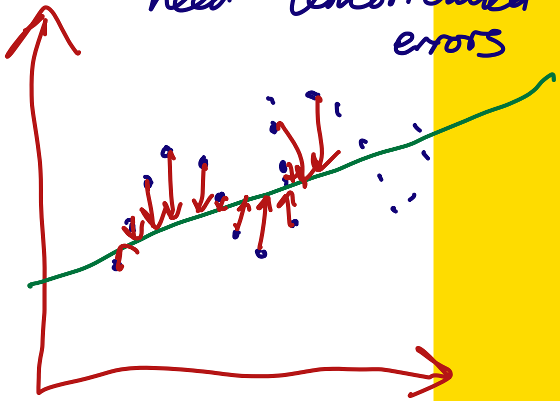
• X is assumed to be deterministic

Advertising Example

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}$$



Why we need uncorrelated errors



Assumptions of the Model

Specific vector of X.

- **Weak assumptions**

Error is 0 on average

Constant variance

$$\mathbb{E}(\epsilon_i|X=\underline{x})=0, \quad \mathbb{V}(\epsilon_i|X=\underline{x})=\sigma^2$$

$$\text{and} \quad Cov(\epsilon_i,\epsilon_j|X=\underline{x})=0$$

– Errors are uncorrelated

for $i=1,2,3,...,n$; for all $i \neq j$ and $\underline{x}=[x_1,\ldots,x_n]^\top$. In other words, errors have **zero mean**, **common variance** and are **uncorrelated**. Parameters estimation: Least Squares

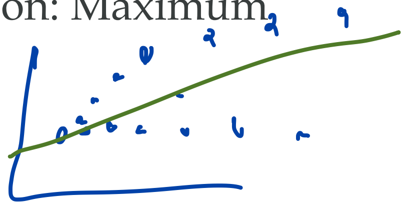
If met, the line Linear regression should be pretty good.

- **Strong assumptions**

It allows for hypothesis testing

$$\epsilon_i|X=\underline{x}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,\sigma^2)$$

for $i=1,2,3,...,n$. In other words, errors are **i.i.d. Normal** random variables with **zero mean** and **constant variance**. Parameters estimation: Maximum Likelihood or Least Squares



Least Squares Estimates (LSE)

"True" model is

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

- Most common approach to estimating $\hat{\beta}_0$ and $\hat{\beta}_1$
- Minimise the residual sum of squares (RSS)

$$\text{RSS} = \underbrace{\sum_{i=1}^{n} (y_i - \overset{f(x)}{\hat{y}_i})^2}_{} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- The least square coefficient estimates are (make sure you can derive these!)

Best estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2} = \frac{S_{xy}}{S_{xx}}$$

Slide 67, notations

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i$. See slide on S_{xy}, S_{xx} and sample (co-)variances. **Proof**: See Lab questions.

-Orange book.

LS Demo

Least Squares Estimates (LSE) - Properties

Under the **weak assumptions** we have **unbiased estimators**:

- $\mathbb{E}\left[\widehat{\beta}_0|X=\underline{x}\right]=\beta_0$ and $\mathbb{E}\left[\widehat{\beta}_1|X=\underline{x}\right]=\beta_1$.

$Var(\varepsilon_i|X=x_i)$

- An (unbiased) estimator of σ^2 is given by:

$$s^2=\frac{\sum_{i=1}^n\hat{\epsilon}_i^2}{n-2}=\frac{\sum_{i=1}^n\left(y_i-\left(\widehat{\beta}_0+\widehat{\beta}_1x_i\right)\right)^2}{n-2}=\frac{\text{RSS}}{n-2}=\text{RSE}^2$$

  where $\hat{\epsilon}_i=y_i-\widehat{y}_i=e_i$ are called the residuals and RSE the residual standard error.

Proof: See Lab questions.

Least Squares Estimates (LSE) - Uncertainty

Under the **weak assumptions** we have that the (co-)variance of the parameters is given by:

More confident in estimate with more data

$$
\begin{aligned}
\mathrm{Var}\left(\widehat{\beta}_0 | X = \underline{x}\right) &= \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right) = \underline{\sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)} \\
&= SE(\hat{\beta}_0)^2 \\
\mathrm{Var}\left(\widehat{\beta}_1 | X = \underline{x}\right) &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sigma^2}{S_{xx}} = SE(\hat{\beta}_1)^2 \\
\mathrm{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1 | X = \underline{x}\right) &= -\frac{\overline{x}\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = -\frac{\overline{x}\sigma^2}{S_{xx}}
\end{aligned}
$$

If $n \uparrow$, $\mathrm{Var}(\hat{\beta}_0) \downarrow$

$\mathrm{Var}(\hat{\beta}_1) \downarrow$

since $s^2 \downarrow$

Proof: See Lab questions.

$E[Y|X=3]$

$$\bar{x} = E[X] = \sum_i \frac{x_i}{n}$$

$x_i \in \underline{x}$

Maximum Likelihood Estimates (MLE)

- In the regression model there are three parameters to estimate: β_0, β_1, and σ^2.
- Under the **strong assumptions** (i.i.d Normal RV), the joint density of $Y_1, Y_2, \ldots, Y_n$ is the product of their marginals (independent by assumption) so that the likelihood is:

$$Y_i = \beta_0 + \beta_1 x_i + \underset{\downarrow}{\varepsilon_i} \quad N(0,\sigma^2)$$

$$\ell\left(\underline{y}; \beta_0, \beta_1, \sigma\right) = -\,n \log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 .$$

Proof: See Lab questions.

Under strong, you can estimate β_0, β_1, σ^2 by MLE.

- β_0 and β_1 estimates by MLE or LS are the same.

Maximum Likelihood Estimates (MLE)

Partial derivatives set to zero give the following MLEs:

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}},\\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x},\end{aligned}$$

and

$$\widehat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right)^2.$$

- Note that the parameters β_0 and β_1 have the same estimators as that produced from Least Squares.
- However, the MLE $\widehat{\sigma}^2$ is a biased estimator of σ^2.
- In practice, we use the unbiased variant s^2 (see slide).

• Since LS and MLE are the same, we more or less assume the strong assumptions hold when we do testing.

Assessing the Accuracy I

- How to assess the accuracy of the coefficient estimates? In particular, consider the following questions:
  - What are the confidence intervals for $\underbrace{\beta_0 \text{ and } \beta_1?}$
  - How to test the null hypothesis that there is no relationship between X and Y?
  - How to test if the influence of the exogeneous variable (X) on the endogenous variable (Y) is larger/smaller than some value?

Test if these are different from 0.

> **Note**
>
> For inference (e.g. confidence intervals, hypothesis tests), we need the strong assumptions!

Assessing the Accuracy II

- Checking MSE.

- How to assess the accuracy of the model?
- How to assess the accuracy of the predictions? In particular:
  - for the population regression line (i.e. mean response)?
  - for the actual value of the dependent variable (i.e. individual response)?

- R^2

- Mallow C_p
- Adj. R^2
- AIC
- BIC

Assessing the Accuracy of the Coefficient Estimates - Confidence Intervals

Using the **strong assumptions**, a $100\,(1-\alpha)\,\%$ confidence interval (CI) for β_1, and *resp.* for β_0, are given by:

- for β_1:

S ↓ as n ↑

$$\widehat{\beta}_1 \pm t_{1-\alpha/2,\underline{n-2}} \cdot \underbrace{\frac{s}{\sqrt{S_{xx}}}}_{\hat{SE}(\hat{\beta_1})}$$

As n increases, becomes narrower distribution

- for β_0:

$$\widehat{\beta}_0 \pm t_{1-\alpha/2,n-2} \cdot \underbrace{s\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}}}_{\hat{SE}(\hat{\beta}_0)}$$

as n increases, interval shrinks

See rationale slide.

Assessing the Accuracy of the Coefficient Estimates - Inference on the slope

- When we want to test whether the exogenous variable has an influence on the endogenous variable or if the influence is larger/smaller than some value.
- For testing the hypothesis

$$H_0 : \beta_1 = \widetilde{\beta}_1 \quad \text{vs} \quad H_1 : \beta_1 \neq \widetilde{\beta}_1$$

- Specific value for β_1

for some constant $\widetilde{\beta}_1$, we use the test statistic:

$$t(\widehat{\beta}_1) = \frac{\widehat{\beta}_1 - \widetilde{\beta}_1}{\hat{SE}(\hat{\beta}_1)} = \frac{\widehat{\beta}_1 - \widetilde{\beta}_1}{(s \, / \sqrt{S_{xx}} \,)}$$

Typically test $\beta_1 = 0$

- Is β_1 helpful?
- If $\hat{\beta}_1 = 0$, then $\hat{\beta}_0 = \bar{y}$.

which has a t_{n-2} distribution under the H_0 (see rationale slide).

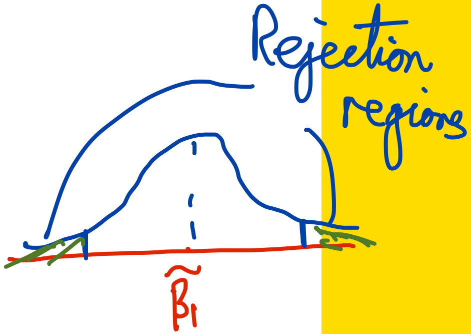
$$Y = f(X) + \varepsilon$$
$$= \beta_0 + \beta_1 X + \varepsilon$$

Assessing the Accuracy of the Coefficient Estimates - Inference on the slope

The decision rules under various alternative hypotheses are summarized below.

Decision Making Procedures for Testing $H_0 : \beta_1 = \tilde{\beta}_1$

Alternative H_1	**Reject H_0 in favor of H_1 if**
$\beta_1 \neq \tilde{\beta}_1$	$\left\lvert t\left(\widehat{\beta}_1\right)\right\rvert > t_{1-\alpha/2,n-2}$
$\beta_1 > \tilde{\beta}_1$	$t\left(\widehat{\beta}_1\right) > t_{1-\alpha,n-2}$
$\beta_1 < \tilde{\beta}_1$	$t\left(\widehat{\beta}_1\right) < -t_{1-\alpha,n-2}$



To test whether the regressor variable is significant or not, it is equivalent to testing whether the slope is zero or not. Thus, test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

Default test in regression outputs.

Assessing the Accuracy of the Coefficient Estimates - Inference on the intercept

Similarly, for testing the null hypothesis $H_0 : \beta_0 = \widetilde{\beta}_0$ for some constant $\widetilde{\beta}_0$, we use the test statistic:

$$t\left(\widehat{\beta}_0\right) = \frac{\widehat{\beta}_0 - \widetilde{\beta}_0}{\hat{SE}(\hat{\beta}_0)} = \frac{\widehat{\beta}_0 - \widetilde{\beta}_0}{\left(s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}\right)},$$

which has a t_{n-2} distribution under the H_0 (see rationale slide).

Assessing the Accuracy of the Coefficient Estimates - Advertising Example

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

$$\frac{0.0475}{0.0027} \rightarrow$$ Many std away from 0.

Reject at a p-value < 0.0001

Testing if

$\hat{\beta}_0 = 0$

$\hat{\beta}_1 = 0$

Assessing the Accuracy of the Model

If $\hat{\beta}_1 = 0$

$\hat{\beta}_0 = \bar{y}$

Partitioning the variability is used to assess how well the linear model explains the trend in data:

How much does data differ from average

$$\boxed{\underbrace{y_i - \overline{y}}_{\text{total deviation}}} = \underbrace{(y_i - \widehat{y}_i)}_{\text{unexplained deviation}} + \underbrace{(\widehat{y}_i - \overline{y})}_{\text{explained deviation}}.$$

Error explained by our model

Error we can't explain

We then obtain:

$$\underbrace{\sum_{i=1}^{n} (y_i - \overline{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}_{\text{SSM}},$$

TSS RSS ε_i^2

SSM = TSS − RSS

where:

- SST or TSS: **total sum of squares**;
- SSE or RSS: sum of squares error or **residual sum of squares**;
- SSM: **sum of squares model** (sometime called regression).

Proof: See Lab questions

Assessing the Accuracy of the Model

Interpret these sums of squares as follows:

- SST (or TSS) is the total variability in the absence of knowledge of the variable X;
- SSE (or RSS) is the total variability remaining after introducing the effect of X;
- SSM is the total variability "explained" because of knowledge of X.

This partitioning of the variability is used in ANOVA tables:

Source	Sum of squares	DoF	Mean square	F
Regression	$\mathrm{SSM} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	$\mathrm{DFM} = 1$	$\mathrm{MSM} = \frac{\mathrm{SSM}}{\mathrm{DFM}}$	$\frac{\mathrm{MSM}}{\mathrm{MSE}}$
Error	$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\mathrm{DFE} = n - 2$	$\mathrm{MSE} = \frac{\mathrm{SSE}}{\mathrm{DFE}}$	
Total	$\mathrm{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2$	$\mathrm{DFT} = n - 1$	$\mathrm{MST} = \frac{\mathrm{SST}}{\mathrm{DFT}}$	

- Also allows us to do F test → Testing if model is non-zero

- R^2

Assessing the Accuracy of the Model

MSE unit less

or hard to compare on diff datasets

Noting that:

$$\text{SSE} = \underbrace{S_{yy}}_{=\text{SST}} - \underbrace{\widehat{\beta}_1 S_{xy}}_{=\text{SSM}},$$

we can define the R^2 **statistic**, the square of the sample correlation, as:

$$R^2 = \left(\frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}\right)^2 = \widehat{\beta}_1 \frac{S_{xy}}{S_{yy}} = \frac{\widehat{\beta}_1 S_{xy}}{\text{SST}} = \frac{\overset{\text{Error explained}}{\text{SSM}}}{\underset{\text{Total variation}}{\text{SST}}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

- Tells you how good is your model fit

- R^2 is interpreted as the proportion of total variation in the y_i's explained by the variable x in a linear regression model.
- R^2 takes on a value between 0 and 1.
- R^2 is also called **coefficient of determination**.

$0 \leq R^2 \leq 1$

$R^2 = 1$

Proof: See Lab questions

Assessing the Accuracy of the Predictions - Mean Response

$Y = f(X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$

Suppose $x = x_0$ is a specified value of the *out of sample* regressor variable and we want to predict the corresponding Y value associated with it. The **mean** of Y is:

$$\begin{aligned}\mathbb{E}[Y \mid x_0] &= \mathbb{E}[\beta_0 + \beta_1 x \mid x = x_0] \\ &= \beta_0 + \beta_1 x_0.\end{aligned}$$

$\mathbb{E}[Y \mid x = x_0]$

Our (unbiased) estimator for this mean (also the fitted value of y_0) is:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Predict $\boxed{f(X)}$

The variance of this estimator is:

$$\mathrm{Var}(\hat{y}_0) = \underbrace{\left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right)}_{*} \sigma^2 \cancel{=} \mathrm{SE}(\hat{y}_0)^2 = s^2 *$$

Proof: See Lab questions.

Assessing the Accuracy of the Predictions - Mean Response

Using the **strong assumptions**, the $100\,(1-\alpha)\,\%$ confidence interval for $\beta_0 + \beta_1 x_0$ (mean of Y) is:

$$\underbrace{(\widehat{\beta}_0 + \widehat{\beta}_1 x_0)}_{\hat{y}_0} \boxed{\pm\, t_{1-\alpha/2,n-2} \times \underbrace{s\sqrt{\frac{1}{n} + \frac{(\overline{x}-x_0)^2}{S_{xx}}}}_{\hat{\mathrm{SE}}(\hat{y}_0)}},$$

Confidence interval for $f(x_0)$

as we have

$$\hat{y}_0 \sim \mathcal{N}(\beta_0 + \beta_1 x_0, \mathrm{SE}(\hat{y}_0)^2)$$

and

$$\frac{\hat{y}_0 - (\beta_0 + \beta_1 x_0)}{\hat{\mathrm{SE}}(\hat{y}_0)} \sim t(n-2).$$

Similar rationale to slide.

- Ignored variability from ε

$Y = f(x) + \varepsilon$

Assessing the Accuracy of the Predictions - Individual response

A **prediction interval** is a confidence interval for the **actual value** of a Y_i (not for its mean $\beta_0 + \beta_1 x_i$). We base our prediction of Y_i (given $X = x_i$) on:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

Here trying to predict $\underline{Y}$, not $f(x)$

The error in our prediction is:

$$Y_i - \widehat{y}_i = \beta_0 + \beta_1 x_i + \epsilon_i - \widehat{y}_i = \mathbb{E}[Y|X = x_i] - \widehat{y}_i + \epsilon_i.$$

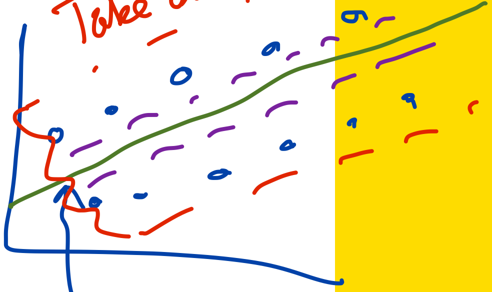
with

$$\mathbb{E}\left[Y_i - \widehat{y}_i | \underline{X} = \underline{x}, X = x_i\right] = 0, \text{ and}$$

$$\mathrm{Var}(Y_i - \widehat{y}_i | \underline{X} = \underline{x}, X = x_i) = \sigma^2 \Big(1 + \frac{1}{n} + \frac{(\overline{x} - x_i)^2}{S_{xx}}\Big).$$

Comes from ε

Take into acct var. from ε



Before we predicted average value

Proof: See Lab questions.

Assessing the Accuracy of the Predictions - Individual response

A $100(1-\alpha)\%$ **prediction interval** for Y_i, the value of Y at $X = x_i$, is given by:

$$\underbrace{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}_{\widehat{y}_i} \pm t_{1-\alpha/2,n-2} \cdot s \cdot \sqrt{\underline{1} + \frac{1}{n} + \frac{(\overline{x} - x_i)^2}{S_{xx}}},$$

Wider due to accounting for ε.

as

$$(Y_i - \widehat{y}_i | \underline{X} = \underline{x}, X = x_i) \sim \mathcal{N}\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(\overline{x} - x_i)^2}{S_{xx}}\right)\right), \text{ and}$$

$$\frac{Y_i - \widehat{y}_i}{s\sqrt{1 + \frac{1}{n} + \frac{(\overline{x} - x_i)^2}{S_{xx}}}} \sim t_{n-2}.$$

Multiple Linear Regression

Overview

- Extend the simple linear regression model to accommodate multiple predictors

$X_1, X_2, \dots, X_p$ orthogonal directions

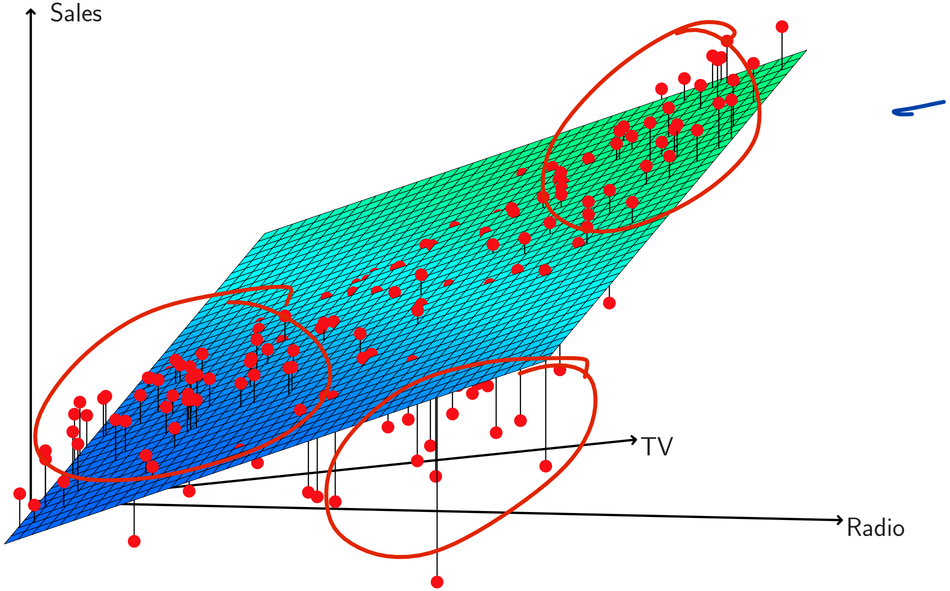
$$Y = \overset{\text{intercept}}{\beta_0} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- β_j: the average effect on Y of a one unit increase in X_j, holding all other predictors fixed

- Simple linear → line of best fit
- Multiple linear → plane of best fit

Advertising Example

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio}$$



- Not a perfect fit, discernable errors

X (uncorrelated errors)

Qualitative predictors

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

Yes/No

Suppose a predictor is qualitative (e.g. 2 different levels) - how would you model/code this in a regression? What if there are more than 2 levels?

$X =$ Yes code as 1

No code as 0

$X =$ Yes

No

Maybe

$$X = \begin{array}{cc} \text{No} & \text{Maybe} \\ \left(\begin{array}{c} 1 \\ 0 \\ 0 \end{array}\right. & \left.\begin{array}{c} 0 \\ 1 \\ 0 \end{array}\right) \end{array}$$

Linear Algebra and Matrix Approach

The model can be re-written as:

Y is a vector

$$Y = f(X) + \varepsilon$$

$$\underline{y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$$

$$f(x) = X\beta$$

Vector

X matrix

with

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$\underline{y} = [y_1, \ldots, y_n]^\top$$

$$\mathbf{X} = \begin{array}{c} \\ \\ \\ \\ \\ \end{array}\begin{array}{ccccc} & X_1 & X_2 & & X_p \\ \end{array}$$

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccccc} & X_1 & X_2 & & X_p \end{array} \\ \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1,p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{n,p} \end{bmatrix} \end{array}$$

$$\underline{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^\top$$

$$\underline{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_n]^\top$$

Note that the matrix $\mathbf{X}$ is of size $n \times p + 1$, the vectors $\underline{y}$, $\underline{\beta}$ and $\underline{\varepsilon}$ are column vectors.

Assumptions of the Model

$Y = XB + \varepsilon$

Weak Assumptions:

The error terms ε_i satisfy the following:

$$\begin{aligned} \mathbb{E}[\varepsilon_i|\mathbf{X}=\mathbf{x}] &= 0, && \text{for } i = 1, 2, \ldots, n; \\ \text{Var}(\varepsilon_i|\mathbf{X}=\mathbf{x}) &= \sigma^2, && \text{for } i = 1, 2, \ldots, n; \\ \text{Cov}(\varepsilon_i, \varepsilon_j|\mathbf{X}=\mathbf{x}) &= 0, && \text{for all } i \neq j. \end{aligned}$$

Same assumptions as before, but extend for multivariate

In words, the errors have **zero means, common variance**, and are **uncorrelated**. In matrix form, we have:

$$\mathbb{E}\left[\underline{\varepsilon}\right] = \underline{0}; \qquad \text{Cov}\left(\underline{\varepsilon}\right) = \sigma^2 \mathbf{I}_n,$$

where $\mathbf{I}_n$ is a matrix of size $n \times n$ with ones on the diagonal and zeros on the off-diagonal elements.

Strong Assumptions: $\varepsilon_i|\mathbf{X}=\mathbf{x} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$.

In words, errors are **i.i.d. normal** random variables with **zero mean** and **constant variance**.

Least Squares Estimates (LSE)

- Same least squares approach as in Simple Linear Regression
- Minimise the residuals sum of squared (RSS)

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_p x_{ip}\right)^2$$

$$= \underbrace{\left(\underline{y} - \mathbf{X}\underline{\beta}\right)^\top \left(\underline{y} - \mathbf{X}\underline{\beta}\right)}_{\text{Loss}} = \sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

$$- \frac{\partial \text{Loss}}{\partial \beta} = 0 \text{ at } (X^T X)^{-1} X^T y$$

- If $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ exists, it can be shown that the solution is given by:

$$\widehat{\underline{\beta}} = \underbrace{\left(\mathbf{X}^\top\mathbf{X}\right)}_{\text{will result in a matrix w. a column linearly dependent on the others}}^{-1}\mathbf{X}^\top\underline{y}.$$

- The corresponding vector of fitted (or predicted) values is

$$\widehat{\underline{y}} = \mathbf{X}\widehat{\underline{\beta}}.$$

Least Squares Estimates (LSE) - Properties

Under the **weak assumptions** we have **unbiased estimators**:

1. The least squares estimators are unbiased: $\mathbb{E}[\underline{\widehat{\beta}}] = \underline{\beta}$.
2. The variance-covariance matrix of the least squares estimators is: $\text{Var}(\underline{\widehat{\beta}}) = \sigma^2 \cdot \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$.
3. An unbiased estimator of σ^2 is:

$$s^2 = \frac{1}{n-p-1}\left(\underline{y} - \underline{\widehat{y}}\right)^\top \left(\underline{y} - \underline{\widehat{y}}\right) = \frac{\text{RSS}}{n-p-1},$$

   $p + 1$ is the total number of parameters estimated.

4. Under the **strong assumptions**, each $\widehat{\beta}_k$ is normally distributed. See details in see slide.

$$\mathbb{E}[\hat{B}] = \mathbb{E}[(X^TX)^{-1}X^Ty]$$

$$= (X^TX)^{-1}X^T\mathbb{E}[y]$$

$$=$$

$$= \beta$$

Test the Relationship Between the Response and Predictors

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

- Is our model better than no model.

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

- F-statistic $= \frac{(\text{TSS}-\text{RSS})/p}{\text{RSS}/(n-p-1)}$ - From ANOVA table
- Question: Given the individual p-values for each variable, why do we need to look at the overall F-statistics?

Test if $\hat{\beta}_j = 0$

t-value as: $\frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)}$

- By statistical chance if p is large at least one β will be non-zero.

- Accounts for # p and tests relationship wholistically

Analysis of variance (ANOVA)

The sums of squares are interpreted as follows:

- SST (or TSS) is the total variability in the absence of knowledge of the variables $X_1, \ldots, X_p$;
- SSE (or RSS) is the total variability remaining after introducing the effect of $X_1, \ldots, X_p$;
- SSM is the total variability “explained” because of knowledge of $X_1, \ldots, X_p$.

ANOVA

This partitioning of the variability is used in ANOVA tables:

Source	Sum of squares	DoF	Mean square	F	p-value
Regression	$\text{SSM} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$\text{DFM} = p$	$\text{MSM} = \frac{\text{SSM}}{\text{DFM}}$	$\frac{\text{MSM}}{\text{MSE}}$	$1 - F_{\text{DFM,DFE}}(F)$
Error	$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\text{DFE} = n - p - 1$	$\text{MSE} = \frac{\text{SSE}}{\text{DFE}}$		
Total	$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$\text{DFT} = n - 1$	$\text{MST} = \frac{\text{SST}}{\text{DFT}}$		

Tests if model is non-zero.

Model Fit and Predictions

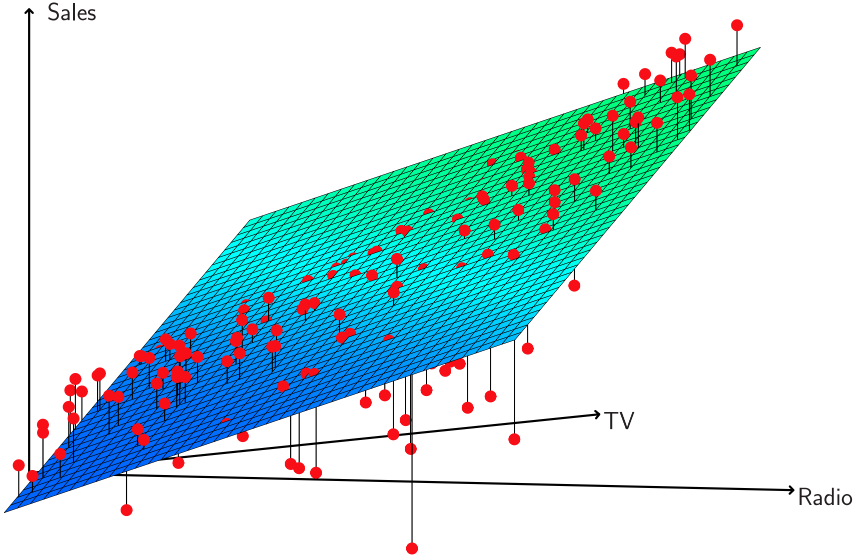
- Measure model fit (similar to the simple linear regression)
  - Residual standard error (RSE)
  - R^2
- Uncertainties associated with the prediction
  - $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ are estimates – Same as before in simple linear
  - linear model is an approximation
  - random error ϵ

$\hat{\beta}_j = 0$

t-stat $\dfrac{\hat{\beta}_j}{se(\hat{\beta}_j)}$

Advertising Example (continued)

Linear regression fit using TV and Radio:



What do you observe?

Other Considerations in the Regression Model

- Qualitative predictors
  - two or more levels, with no logical ordering
  - create binary (0/1) dummy variables
  - Need (#levels - 1) dummy variables to fully encode — R does this automatically
- Interaction terms ($X_i X_j$) (removing the additive assumption)
- Quadratic terms (X_i^2) (non-linear relationship)

$Y_i = \beta_0 + \beta_1 x_i$

$Y_i = \beta_0 + \beta_1 x_i^2$

week - 7.

$$x = \begin{pmatrix} T \\ T \\ F \\ T \\ F \\ M \\ F \end{pmatrix} = \begin{array}{cc} F & M \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ T & F \end{array}$$

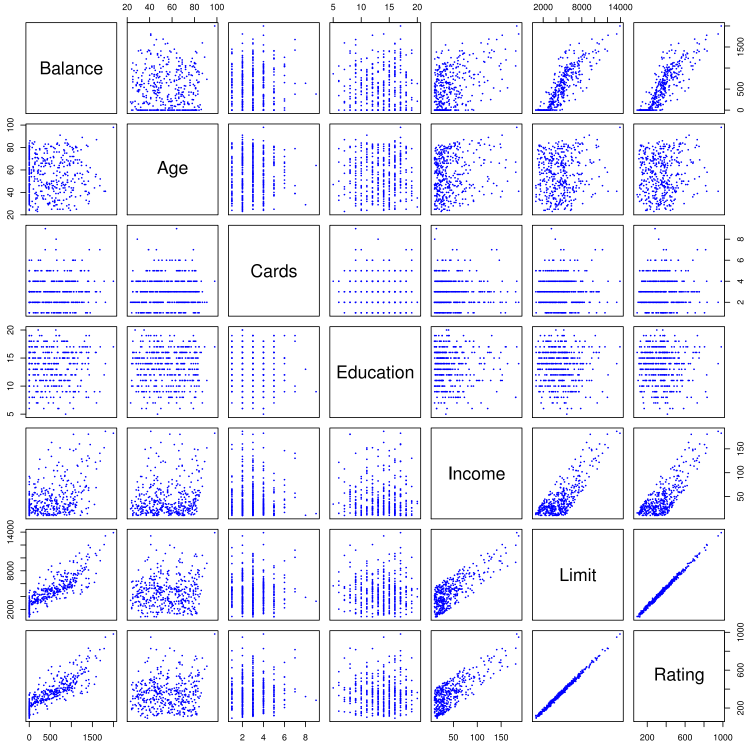
— Be careful in interpretations!

$$Y = \beta_0 + \underset{\overset{\uparrow}{F}}{\beta_1 X} + \underset{\overset{\uparrow}{M}}{\beta_2 X}$$

If $X = F$, then Y increases by β_1 on average compared to if $X = T$.

Linear model selection

The `credit` dataset



Qualitative covariates: `own`, `student`, `status`, `region`

Linear Model selection

$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p x_p.$

- Various approaches - we will focus on
  - Subset selection
  - Indirect methods
  - Shrinkage (also called Regularization) (Later in the course)
  - Dimension Reduction (Later in the course)

If X_i are quantitative

then: If X_i increases by 1, (others fixed)

Y_i increases by β_i on average

Subset selection

- The classic approach is subset selection
- Standard approaches include
  - Best subset
  - Forward stepwise
  - Backwards stepwise
  - Hybrid stepwise

$\left.\begin{array}{l}MSE\\R^2\end{array}\right\}$ always improve when p increases

Best subset selection

Consider a linear model with n observations and p potential predictors:

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

Algorithm:

- Consider the models with 0 predictors, and call this $\mathcal{M}_0$. This is the null model
- Consider all models with 1 predictor, pick the best fit, and call this $\mathcal{M}_1$
- $\ldots$
- Consider the model with p predictor, and call this $\mathcal{M}_p$. This is the full model
- Pick the best fit of $\underline{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p}$

Adj R^2, AIC, BIC, C_p

Best subset selection - behaviour

- Considers all possible models, given the predictors
- Optimal model $\mathcal{M}_k$ sets $p - k$ parameters to 0, the rest are found using the normal fitting technique
- Picks the best of all possible models, given selection criteria
- Very computationally expensive. Calculates:

$$\sum_{k=0}^{p} \binom{p}{k} = 2^p \text{ models}$$

— Very expensive if p is not small.

Stepwise Example: Forward stepwise selection

Algorithm:

- Start with the null model $\mathcal{M}_0$

, Improves your metric the most

- Consider the p models with 1 predictor, pick the best, and call this $\mathcal{M}_1$
- Extend $\mathcal{M}_1$ with one of the $p - 1$ remaining predictors. Pick the best, and call this $\mathcal{M}_2$
- $\ldots$
- End with the full model $\mathcal{M}_p$
- Pick the best fit of $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$

Stepwise subset selection - behaviour

- Considers a much smaller set of models, but the models are generally good fits
- Far less computationally expensive. Considers only:

$$\sum_{k=0}^{p-1}(p-k) = 1 + \frac{p(p+1)}{2} \text{ models}$$

- Grows quad instead of exponentially in p.

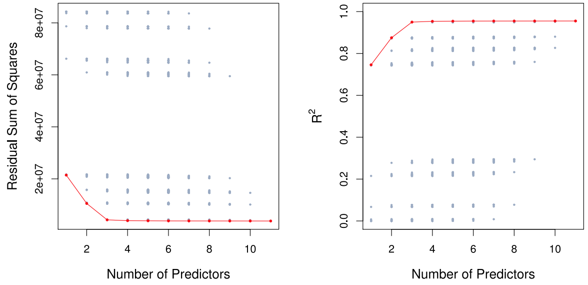
- Like best-subset, sets excluded predictor's parameters to 0
- Backward and forward selection give similar, but possibly different models
- Assumes each “best model” with n predictors is a proper subset of the one with size $n + 1$
  - In other words, it only looks one step ahead at a time
- Hybrid approaches exist, adding some variables, but also removing variables at each step

Example: Best subset and forward selection on *Credit* data

# Variables	Best subset	Forward stepwise
1	*rating*	*rating*
2	*rating, income*	*rating, income*
3	*rating, income, student*	*rating, income, student*
4	*cards, income, student, limit*	*rating, income, student, limit*

How to determine the “best” model

- Need a metric to compare different models
- R^2 can give misleading results as models with more parameters always have a higher R^2 on the training set:



RSS and R^2 for each possible model containing a subset of the ten predictors in the `Credit` data set.

- Want low test error:
  - Indirect: estimate test error by adjusting the training error metric due to bias from overfitting
  - Direct: e.g. cross-validation, validation set ← Week 5.

Indirect methods

1. C_p with d predictors: Mallow Cp

$$\frac{1}{n}(\text{RSS} + \underbrace{2d\hat{\sigma}^2}_{\text{penalty on \# predictors}})$$

- Unbiased estimate of test MSE if $\hat{\sigma}^2$ is an unbiased estimate of σ^2

2. Akaike information criteria (AIC) with d predictors:

$$\frac{1}{n}(\text{RSS} + \underbrace{2d\hat{\sigma}^2}) - \text{penalty}$$

- Proportional to C_p for least squares, so gives the same results

Indirect methods cont.

3. Bayesian information criteria (BIC) with d predictors

$$\frac{1}{n}(\text{RSS} + \log(n)\, d\hat{\sigma}^2)$$

- $\log(n) > 2$ for $n > 7$, so this is a much heavier penalty

4. Adjusted R^2 with d predictors

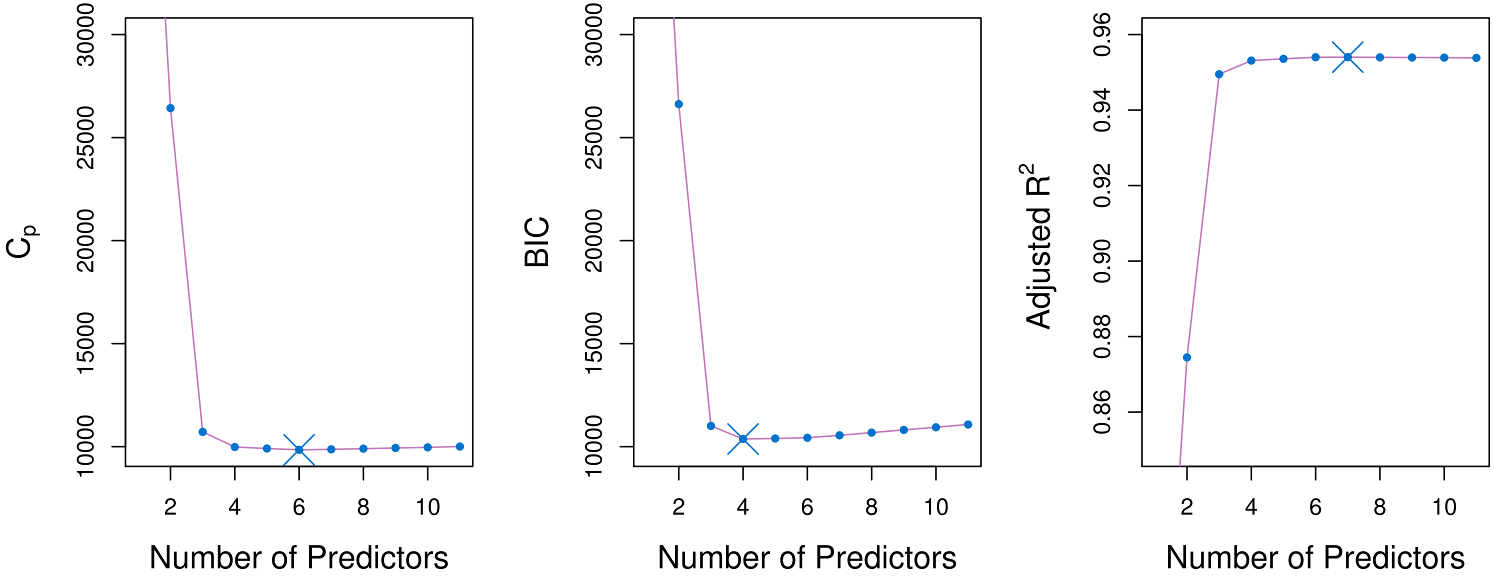
$$1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

R^2 with penalty

Adj. R^2 can be negative

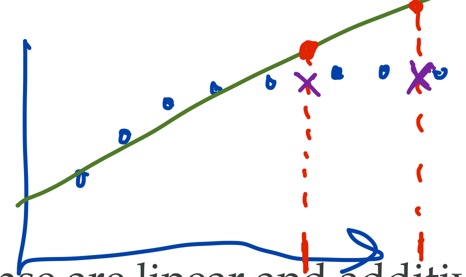
- Decreases in RSS from adding parameters are offset by the increase in $1/(n-d-1)$
- Popular and intuitive, but theoretical backing not as strong as the other measures

How to determine the "best" model - `Credit` dataset



Potential problems with Linear Regression

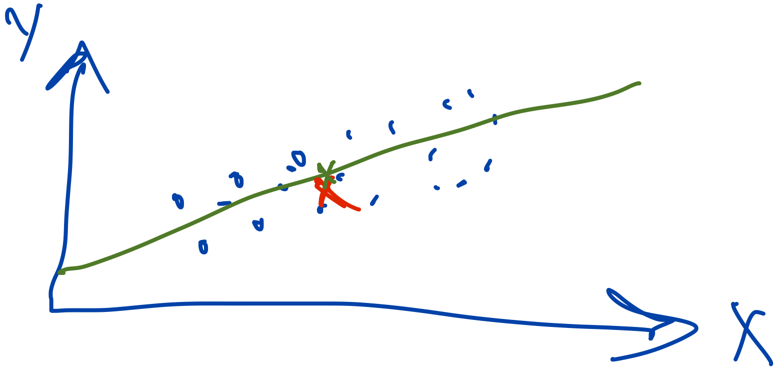
Potential Problems/Concerns



To apply linear regression properly:

- The relationship between the predictors and response are linear and additive (i.e. effects of the covariates must be additive);
- Homoskedastic (constant) variance;
- Errors must be independent of the explanatory variables with mean zero (weak assumptions);
- Errors must be Normally distributed, and hence, symmetric (only in case of testing, i.e., strong assumptions).

A "good" model is one with a low test error.

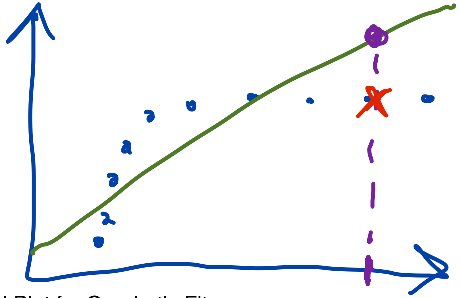


Potential Problems/Concerns

1. Non-linearity of the response-predictor relationships
2. Correlation of error terms
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points
6. Collinearity
7. Confounding effect (correlation does not imply causality!)

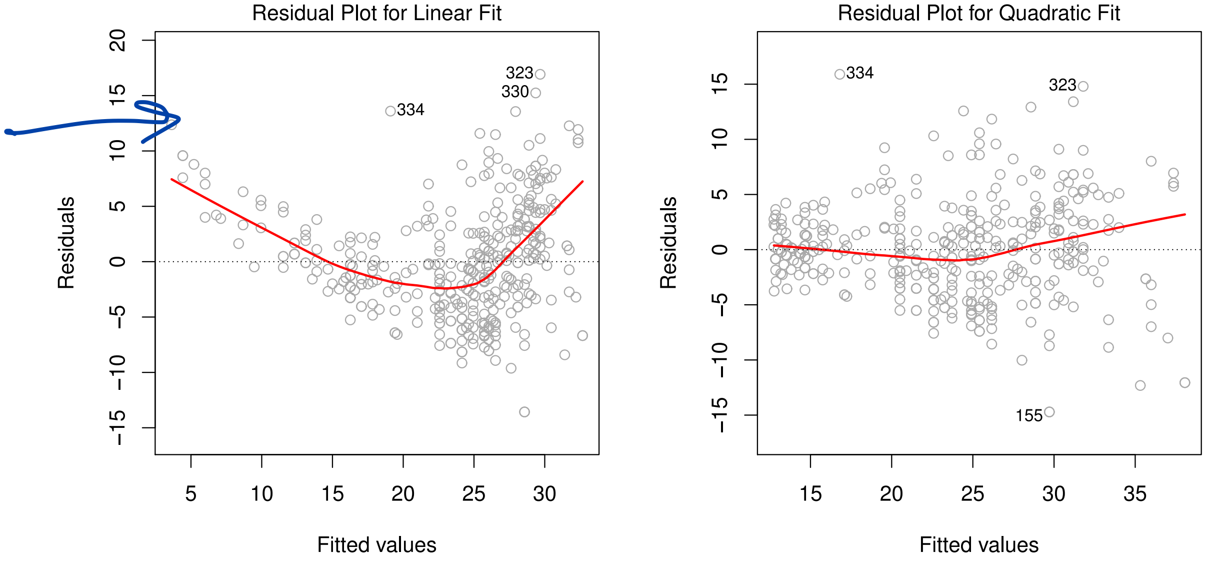
1. Non-linearities

Example: residuals vs fitted for `MPG` vs `Horsepower`:



$y_i - \hat{y}_i$

$Y = \beta_0 + \beta_1 x_1$



$Y = \beta_0 + \beta_1 x_1^2$

$Y = \beta_0 + \beta_1 \gamma(x_i)$

LHS is a linear model. RHS is a quadratic model.

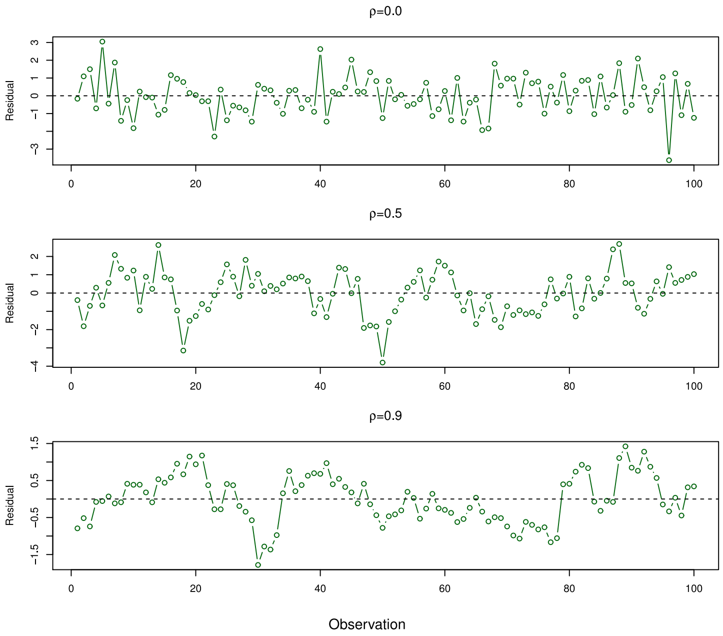
Quadratic model removes much of the pattern - we look at these in more detail later.

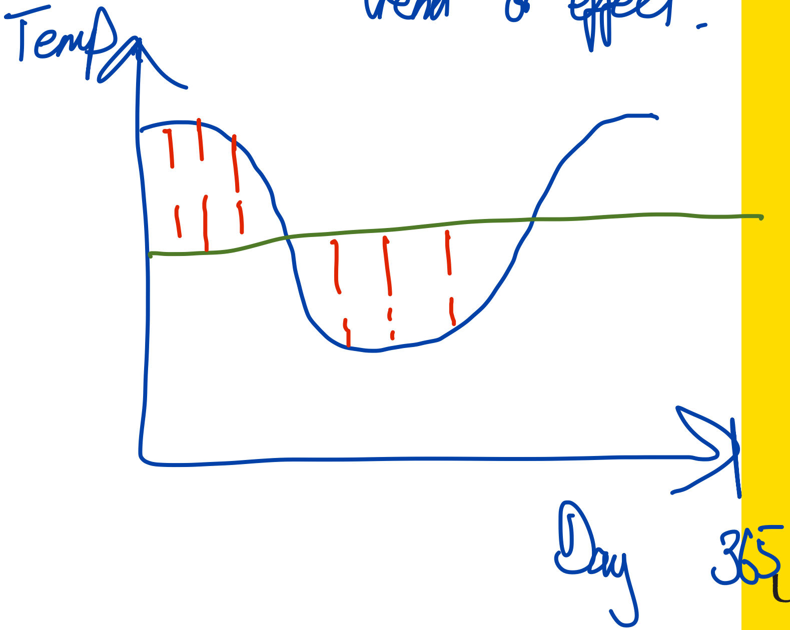
2. Correlations in the Error terms

- Time series have correlated errors. Regressions tend not to work on time series

- The assumption in the regression model is that the error terms are uncorrelated with each other.
- If they are not uncorrelated the standard errors will be incorrect.

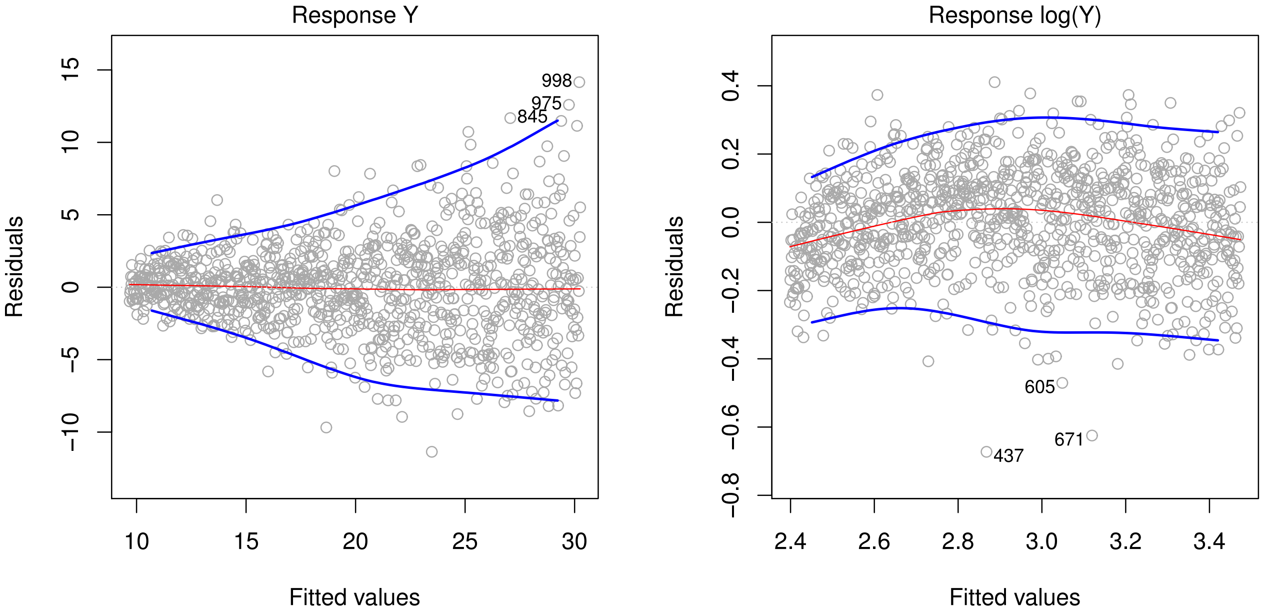
- Probably have not identified an underlying trend or effect.

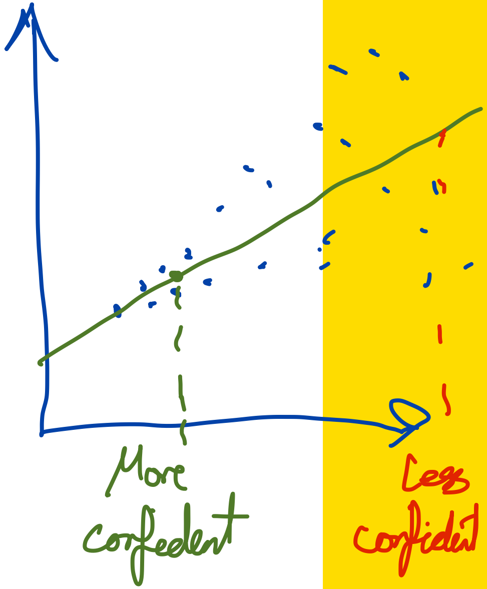




3. Non-constant error terms

The following are two regression outputs vs Y (LHS) and lnY (RHS)

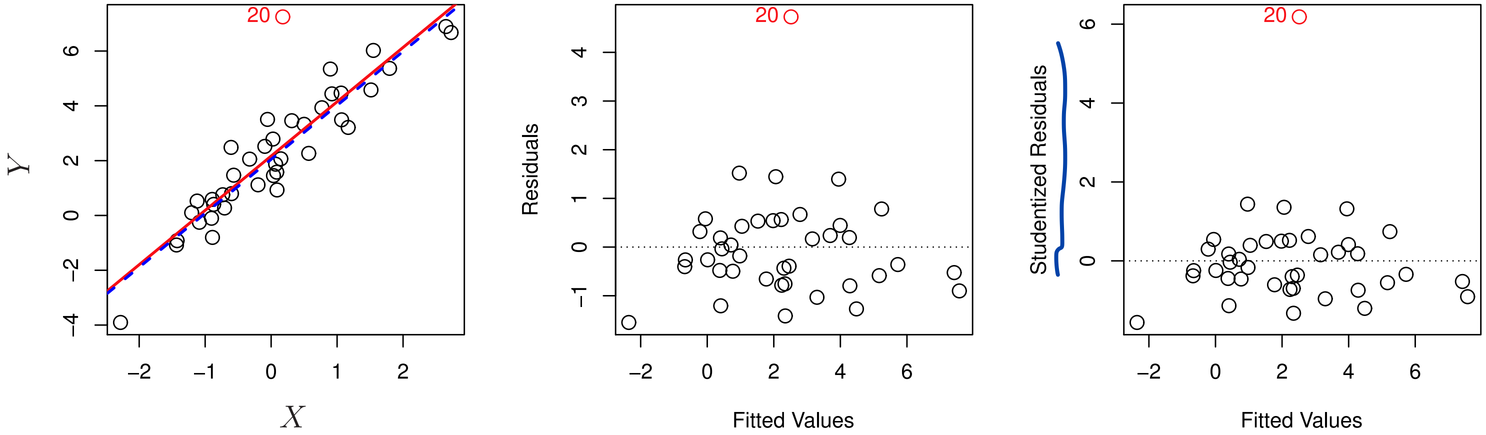


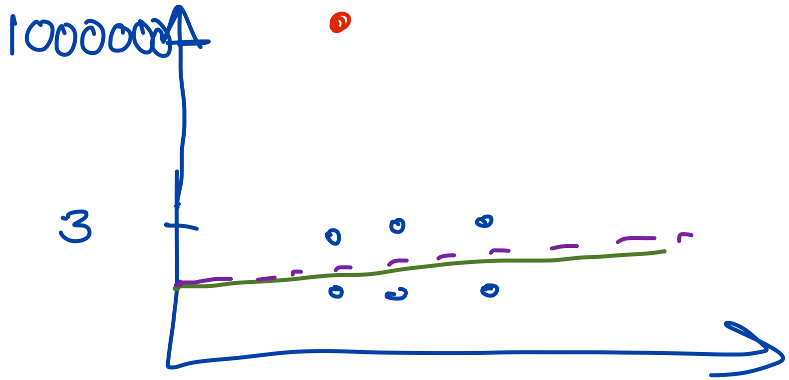


In this example log transformation removed much of the heteroscedasticity.

4. Ouliers

$$\frac{\text{Residuals}}{\text{S.e. residuals}}$$





$$\mu \quad (y-X^T\beta)^T(y-X^T\beta)$$

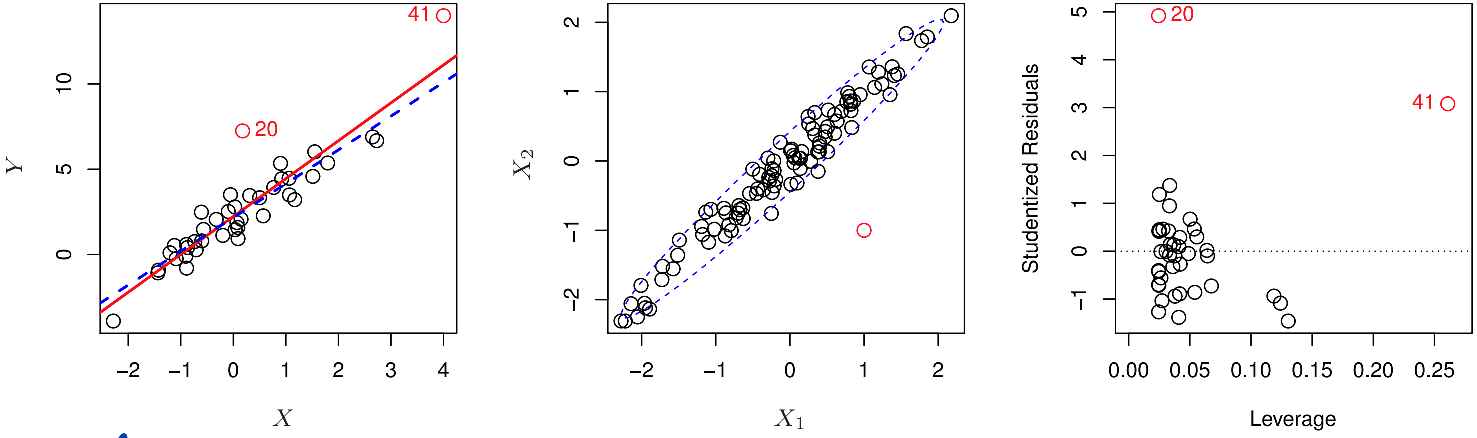
$$S \quad \sum(y_i \; \beta_0+\beta_1X_i)^2$$

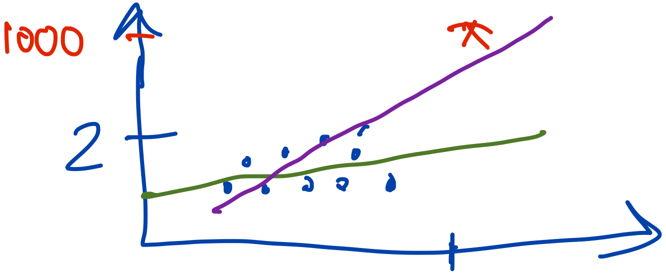
- Error in data collection
- Feature you need to indentify

5. High-leverage points

The following compares the fitted line with (RED) and without (BLUE) observation 41 fitted.

- Points that influence line/plane fit significantly





- Probably should model w. and without to see overall effect on AIC

High-leverage points

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

- Have unusual predictor values, causing the regression line to be dragged towards them
- A few points can significantly affect the estimated regression line
- Compute the leverage using the hat matrix:

$$\hat{y} = \underline{X\hat{\beta}}$$

$$\hat{y} = \underbrace{X(X^TX)^{-1}X^T}\,y$$

$$H = \underbrace{X(X^TX)^{-1}X^T}$$

The hat matrix puts a hat on y

- Note that

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij}Yj = h_{ii}Y_i + \sum_{j \neq i}^{n} h_{ij}Y_j$$

so each prediction is a linear function of all observations, and $h_{ii} = [H]_{ii}$ is the weight of observation i on its own prediction

- If $\underbrace{h_{ii} > 2(p+1)/n}$, the predictor can be considered as having a high leverage

h_{ii} is closer to 1, then it is high leverage.

6. Collinearity

$$X = \begin{array}{c} \text{int} \\ \left(\begin{array}{ccc} 1 & \cdot & \cdot \\ 1 & & \cdot \\ 1 & & \cdot \end{array}\right) \end{array}$$

- Two or more predictor values are closely related to each other
- Reduces the accuracy of the regression by increasing the set of plausible coefficient values
- In effect, the causes SE of the beta coefficients to grow. ← False conclusions on $\hat{\beta}_i = 0$ tests
- Correlation can indicate one-to-one (linear) collinearity

$X^T X$ is not full rank, means

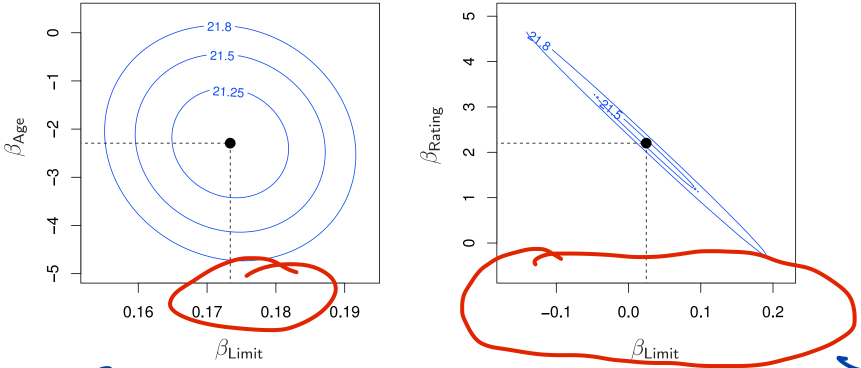
$(X^T X)^{-1}$ does not exist

$$\hat{y} = (X^T X)^{-1} X^T y$$

- If factors are not coded properly then $X^T X$ will not be full rank

Part one Weeks 1-4/5

Collinearity makes optimisation harder



β_{Limit} p-value small

p-value large.

- Contour plots of the values as a function of the predictors. `Credit` dataset used.
- Left: `balance` regressed onto `age` and `limit`. Predictors have low collinearity
- Right: `balance` regressed onto `rating` and `limit`. Predictors have high collinearity
- Black: coefficient estimate

Multicollinearity

$$X_j = \beta_0 + \beta_1 X_1 + \dots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \dots + \beta_p X_p.$$

- Use variance inflation factor

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

- $R^2_{X_j|X_{-j}}$ is the R^2 from X_j being regressed onto all other predictors
- Minimum 1, higher is worse (>5 or 10 is considered high)

Test for linear dependence

fit just predictor j on other predictors

Look at difference in s.e.

7. Confounding effects

- But what about confounding variables? Be careful, correlation does not imply causality![1]
- C is a **confounder** (confounding variable) of the relation between X and Y if:
  - C influences X and C influences Y,
  - but X does not influence Y (directly).

Just because X and Y appear related,
does not mean they are!

ice cream sales $\not\longleftrightarrow$ # shark attacks.

temperature /
# people at beach.

[1] Check this website on spurious correlations.

Confounding effects

- The predictor variable X would have an indirect influence on the dependent variable Y.
  - Example: Age ⇒ Experience ⇒ Probability of car accident. If experience can not be measured, age can be a proxy for experience.
- The predictor variable X would have no direct influence on dependent variable Y.
  - Example: Becoming older does not make you a better driver.
- Hence, a predictor variable works as a predictor, but action taken on the predictor itself will have no effect.

– Be careful with interpretation if you believe a confounding effect is present!

Confounding effects

How to correctly use/don't use confounding variables?

- If a confounding variable is observable: add the confounding variable.
- If a confounding variable is unobservable: be careful with interpretation!

So what's next

Generalisations of the Linear Model

In much of the rest of this course, we discuss methods that expand the scope of linear models and how they are fit:

(Binary)

Y is quantitative

- *Classification problems:* logistic regression ← Will it rain tomorrow?
- *Non-normality:* Generalised Linear Model
- *Non-linearity:* splines and generalized additive models; KNN, tree-based methods Week 7.
- *Regularised fitting:* Ridge regression and lasso
- *Non-parametric:* Tree-based methods, bagging, random forests and boosting, KNN (these also capture non-linearities) Week 9/10.

Appendices

Appendix: Sum of squares

Recall from ACTL2131/ACTL5101, we have the following sum of squares:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad \Longrightarrow \quad s_x^2 = \frac{S_{xx}}{n-1}$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 \qquad \Longrightarrow \quad s_y^2 = \frac{S_{yy}}{n-1}$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \quad \Longrightarrow \quad s_{xy} = \frac{S_{xy}}{n-1},$$

Here s_x^2, s_y^2 (and s_{xy}) denote sample (co-)variance.

Appendix: CI for β_1 and β_0

Rationale for β_1: Recall that $\widehat{\beta}_1$ is unbiased and $\text{Var}(\widehat{\beta}_1) = \sigma^2/S_{xx}$. However σ^2 is usually unknown, and estimated by s^2 so, under the **strong assumptions**, we have:

$$\frac{\widehat{\beta}_1 - \beta_1}{s/\sqrt{S_{xx}}} = \underbrace{\frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}_{\mathcal{N}(0,1)} \Bigg/ \underbrace{\sqrt{\frac{\frac{(n-2)\cdot s^2}{\sigma^2}}{n-2}}}_{\sqrt{\chi^2_{n-2}/(n-2)}} \sim t_{n-2}$$

as $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$ then $\frac{(n-2)\cdot s^2}{\sigma^2} = \frac{\sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 \cdot x_i)^2}{\sigma^2} \sim \chi^2_{n-2}$.

Note: Why do we lose two degrees of freedom? Because we estimated two parameters!

Similar rationale for β_0.

Appendix: Statistical Properties of the Least Squares Estimates

4. Under the strong assumptions of normality each component $\widehat{\beta}_k$ is normally distributed with mean and variance

$$\mathbb{E}[\widehat{\beta}_k] = \beta_k, \quad \mathbf{Var}(\widehat{\beta}_k) = \sigma^2 \cdot c_{kk},$$

and covariance between $\widehat{\beta}_k$ and $\widehat{\beta}_l$:

$$\mathrm{Cov}(\widehat{\beta}_k, \widehat{\beta}_l) = \sigma^2 \cdot c_{kl},$$

where c_{kk} is the $(k+1)^{\text{th}}$ diagonal entry of the matrix $\mathbf{C} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$.
The standard error of $\widehat{\beta}_k$ is estimated using $\mathrm{se}(\widehat{\beta}_k) = s\sqrt{c_{kk}}$.