Linear Regression

ACTL3142 & ACTL5110 Statistical Machine Learning for Risk Applications

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

Linear Regression

- A classical and easily applicable approach for supervised learning
- Useful tool for predicting a quantitative response
- Model is easy to interpret
- Many more advanced techniques can be seen as an extension of linear regression

Week 4/5.

Lecture Outline

- **Simple Linear Regression** – One predictor
- Multiple Linear Regression – multiple predictors
- Categorical predictors
- R Demo
- ANOVA
- Linear model selection – Ways to improve model
- Potential problems with Linear Regression
- So what's next
- Appendices

n = # data points

Overview

Suppose we have pairs of data $(\underline{y_1}, x_1), (y_2, x_2), ..., (\underline{y_n}, x_n)$ and we want to predict values of $\underline{y_i}$ based on $\underline{x_i}$?

- We could do a linear prediction: $y_i = mx_i + b$.
- We could do a quadratic prediction: $y_i = ax_i^2 + bx_i + c$.
- We could do a general non-linear function prediction: $y_i = f(x_i)$.

All of these methods are examples of models we can specify. Let's focus on the linear prediction. Some questions:

- How do we choose m and b? There are infinite possibilities?
- How do we know whether the line is a 'good' fit? And what do we mean by 'good'?

Is it useful?

$$Y = \underset{\substack{\uparrow \\ \text{model}}}{f(X)} + \varepsilon$$

Overview

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

Simple linear regression is a linear prediction.

- Predict a quantitative response $Y = (y_1, ..., y_n)^\top$ based on a single predictor variable $\underline{X = (x_1, ..., x_n)}^\top$
- Assume the 'true' relationship between X and Y is linear:

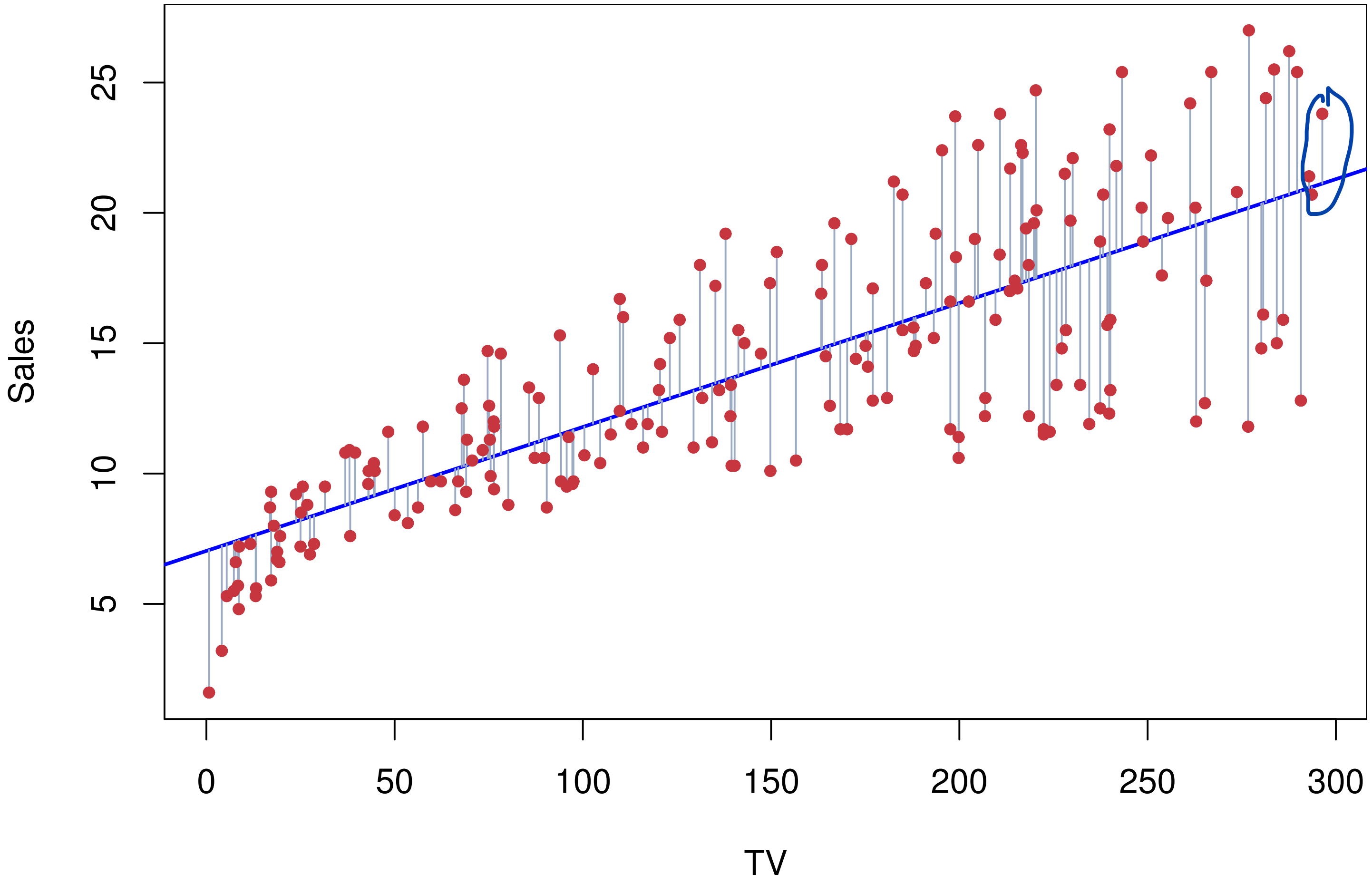
$$\underline{Y} = \underline{\beta_0} + \beta_1 X + \boxed{\epsilon},$$

$$b + m x_i$$

where $\epsilon = (\epsilon_1, ..., \epsilon_n)^\top$ is an error term with certain assumptions on it for identifiability reasons.

Advertising Example

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



Sum all these grey lines, it should be close to 0.

$Y = \beta_0 + \beta_1 X + \varepsilon$ → Weak, strong

Assumptions on the errors

- **Weak assumptions**

Constant

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

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

for $i = 1, 2, 3, ..., n$; for all $i \neq j$.

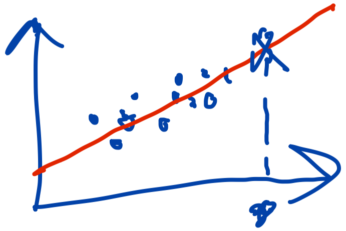
In other words, errors have **zero mean**, **common variance** and are conditionally **uncorrelated**. Parameters estimation: Least Squares

- **Strong assumptions**

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

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

Model estimation



- We have paired data $(y_1, x_1), ..., (y_n, x_n)$.
- We assume there is a 'true' relationship between the y_i and x_i described as

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

- And we assume ϵ satisfies either the weak or strong assumptions.
- How do we obtain estimates $\hat{\beta}_0$ and $\hat{\beta}_1$? If we have these estimates, we can make predictions on the mean:

$$\begin{aligned}\hat{y}_i &= \mathbb{E}[\underline{y_i}|X] = \mathbb{E}[\beta_0 + \beta_1 x_i + \epsilon_i|X] \\ &= \underline{\hat{\beta}_0 + \hat{\beta}_1 x_i}\end{aligned}$$

where we used the fact that $\underline{\mathbb{E}[\epsilon_i|X]} = 0$ and we estimate β_j by $\hat{\beta}_j$.

Weak + strong

Least Squares Estimates (LSE)

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x}_i)^2$$
$$= \widehat{Var}(x_i) \cdot (n-1)$$

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

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

- The least square coefficient estimates are

$$\left.\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2} = \frac{S_{xy}}{S_{xx}} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}\right\}$$

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.

LS Demo

Least Squares Estimates (LSE) - Properties

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

- $\mathbb{E}\left[\hat{\beta}_0|X\right] = \beta_0$ and $\mathbb{E}\left[\hat{\beta}_1|X\right] = \beta_1$.
- An (unbiased) estimator of σ^2 is given by:

$$s^2 = \frac{\sum_{i=1}^{n}\left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2}{n-2}$$

Proof: See Lab questions.

- What does this mean? Using LSE obtains on average the correct values of β_0 and β_1 if the assumptions are satisfied.
- How confident or certain are we in these estimates?

Least Squares Estimates (LSE) - Uncertainty

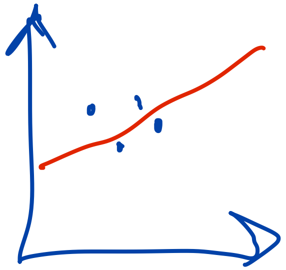
Under the **weak assumptions** we have that the (co-)variance of the parameters is given by:

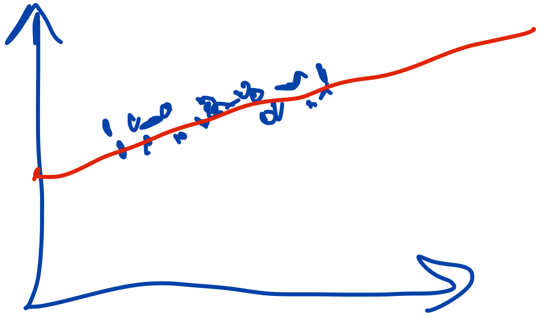
$$\begin{aligned}
\operatorname{Var}\left(\hat{\beta}_0|X\right) =& \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right) = \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right) \quad * \\
=& SE(\hat{\beta}_0)^2 \\
\operatorname{Var}\left(\hat{\beta}_1|X\right)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

$S_{xx} \uparrow \quad n \uparrow$

$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

Proof: See Lab questions. Verify yourself all three quantities goes to 0 as n gets larger.





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:

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

Proof: Since $Y = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, then $y_i \overset{\text{i.i.d.}}{\sim} N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$. The result follows.

Maximum Likelihood Estimates (MLE)

Partial derivatives set to zero give the following MLEs:

$$\left.\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2} = \frac{S_{xy}}{S_{xx}},\\ \hat{\beta}_0 &= \overline{y}-\hat{\beta}_1\overline{x},\end{aligned}\right\} \leftarrow \text{Same as LSE.}$$

and

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

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

$Y = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i$

Interpretation of parameters

How do we interpret a linear regression model such as $\hat{\beta}_0 = 1$ and $\hat{\beta} = -0.5$?

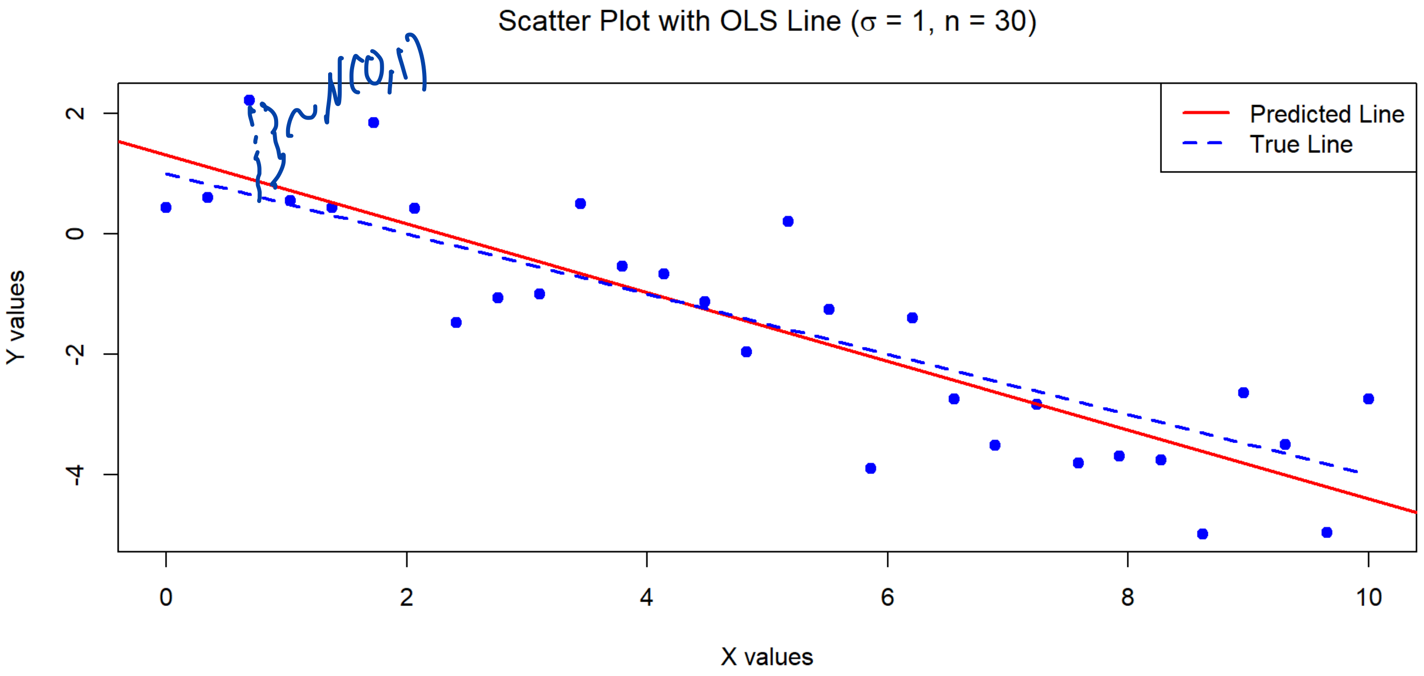
- The intercept parameter $\hat{\beta}_0$ is interpreted as the value we would predict if $x_i = 0$.
  - E.g., predict $\underline{y_i} = 1$ if $\underline{x_i} = 0$
- The slope parameter $\hat{\beta}_1$ as the expected change in the mean-response of y_i for a 1 unit increase in x_i.
  - E.g., we would expect y_i to decrease on average by -0.5 for every 1 unit increase in x_i.

Example 1

The below data was generated by $Y = 1 - 0.5 \times X + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 1)$ with $n = 30$.

$\hat{\beta}$

```
Estimates of Beta_0 and Beta_1:
 1.309629 -0.5713465

Standard error of the estimates:
 0.346858 0.05956626
```

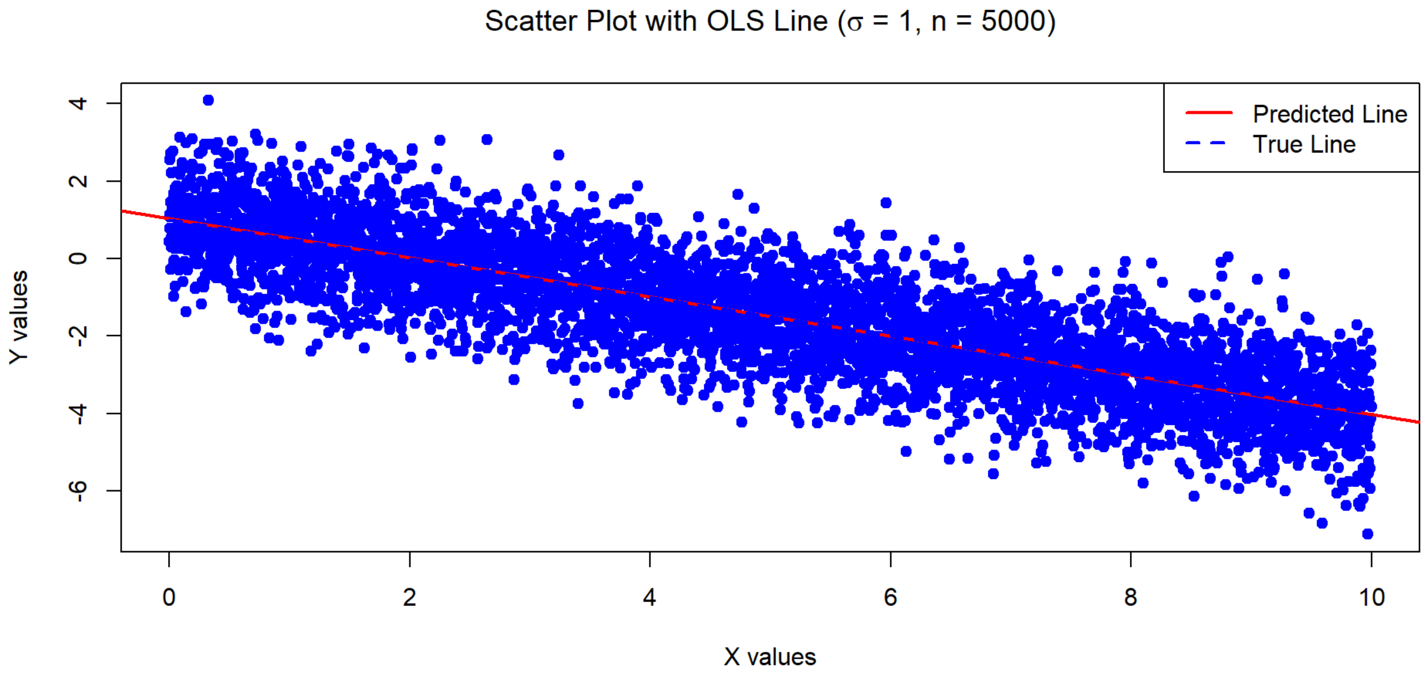


Example 2

The below data was generated by $Y = 1 - 0.5 \times X + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 1)$ with $n = 5000$.

```
Estimates of Beta_0 and Beta_1:
 1.028116 -0.5057372

Standard error of the estimates:
 0.02812541 0.00487122
```

This depends on the irreducible error a bit.

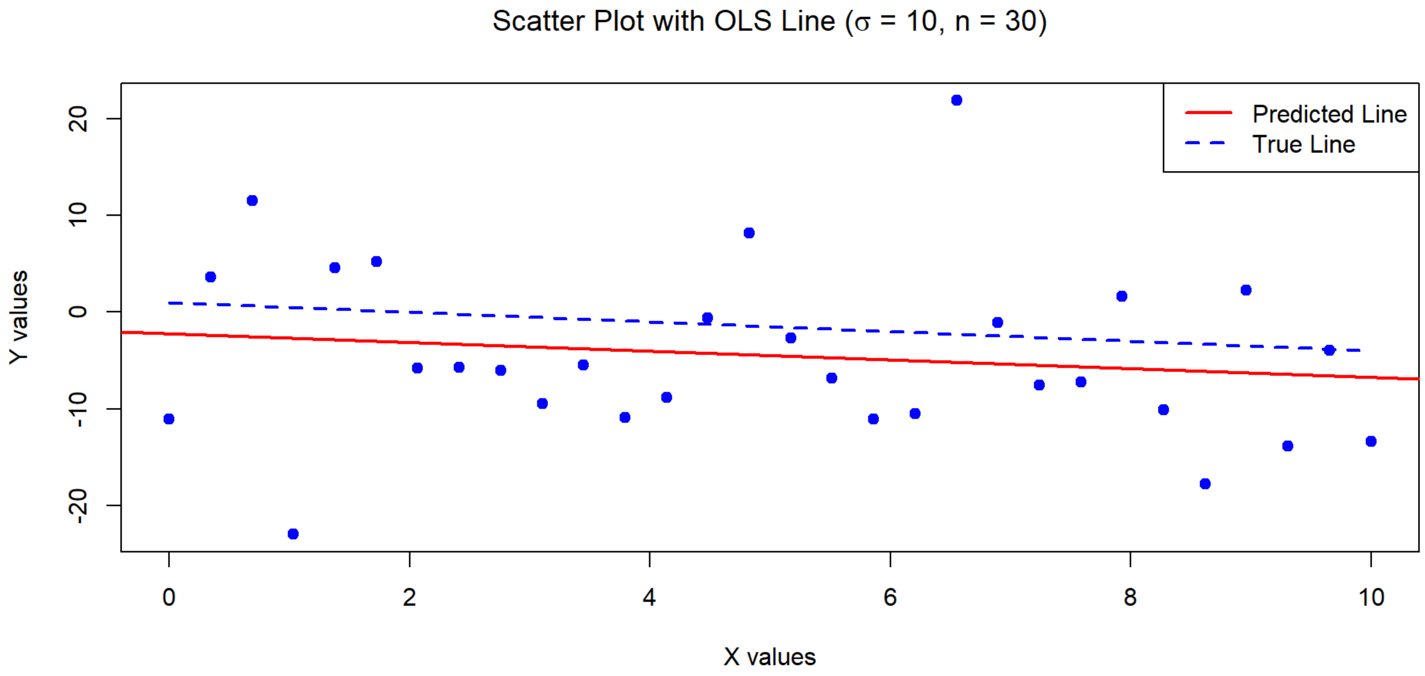


Example 3

The below data was generated by $Y = 1 - 0.5 \times X + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 100)$ with $n = 30$.

Variance

```
Estimates of Beta_0 and Beta_1:
 -2.19991 -0.4528679

Standard error of the estimates:
 3.272989 0.5620736
```



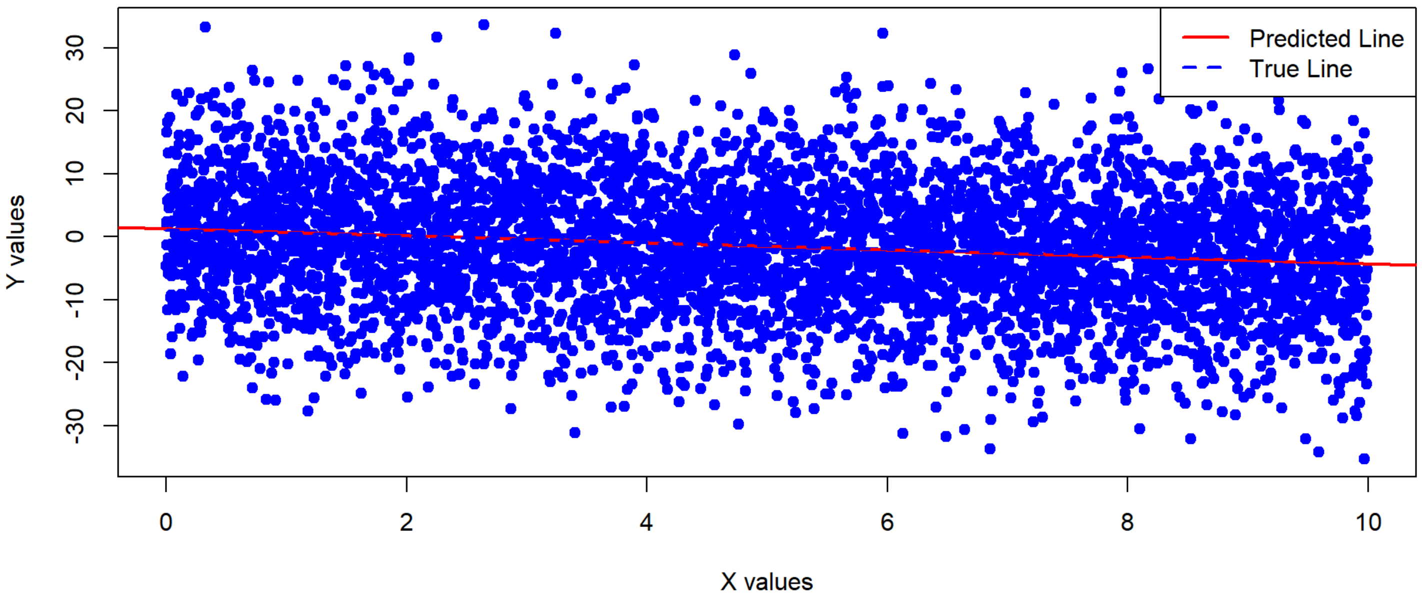
Example 4

The below data was generated by $Y = \underline{1} - \underline{0.5} \times X + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 100)$ with $n = 5000$.

```
Estimates of Beta_0 and Beta_1:
 1.281162 -0.5573716

Standard error of the estimates:
 0.2812541 0.0487122
```

More in 4 compared to 3.

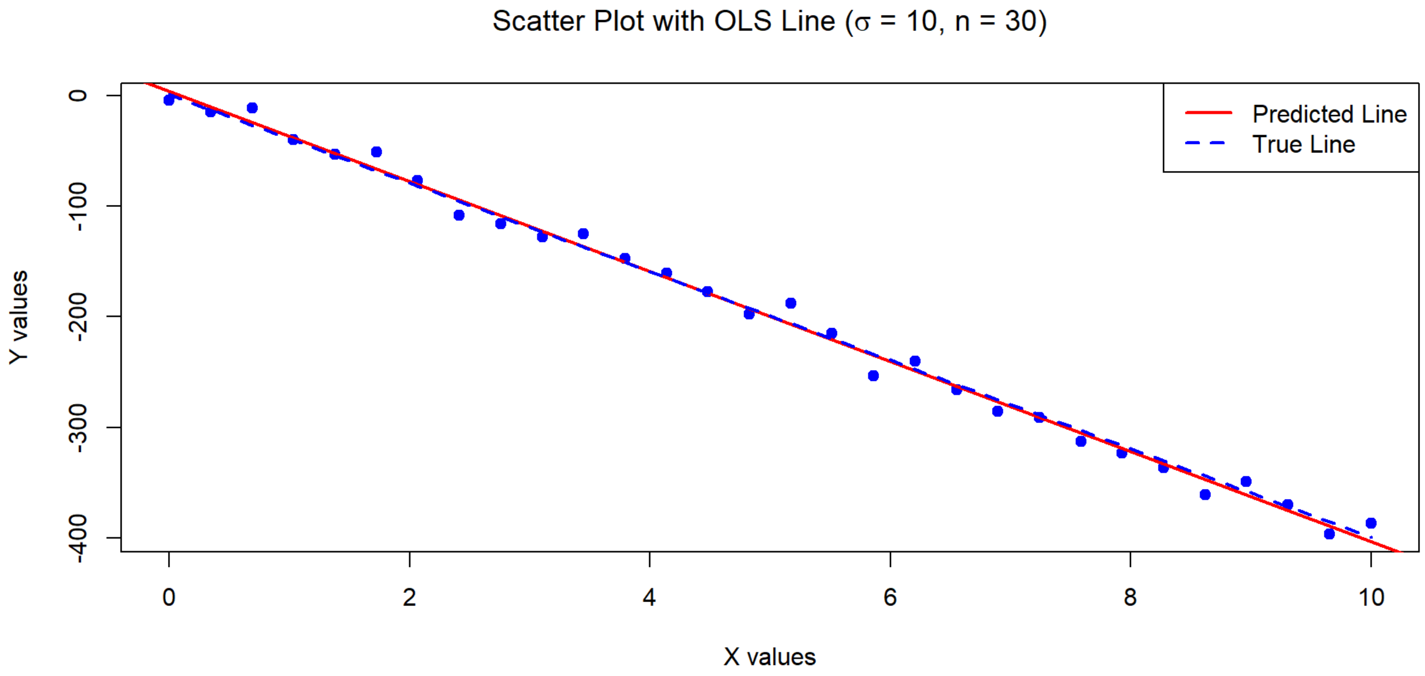
Scatter Plot with OLS Line (σ = 10, n = 5000)



Example 5

The below data was generated by $Y = 1 - 40 \times X + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 100)$ with $n = 30$.

```
Estimates of Beta_0 and Beta_1:
 4.096286 -40.71346

Standard error of the estimates:
 3.46858 0.5956626
```



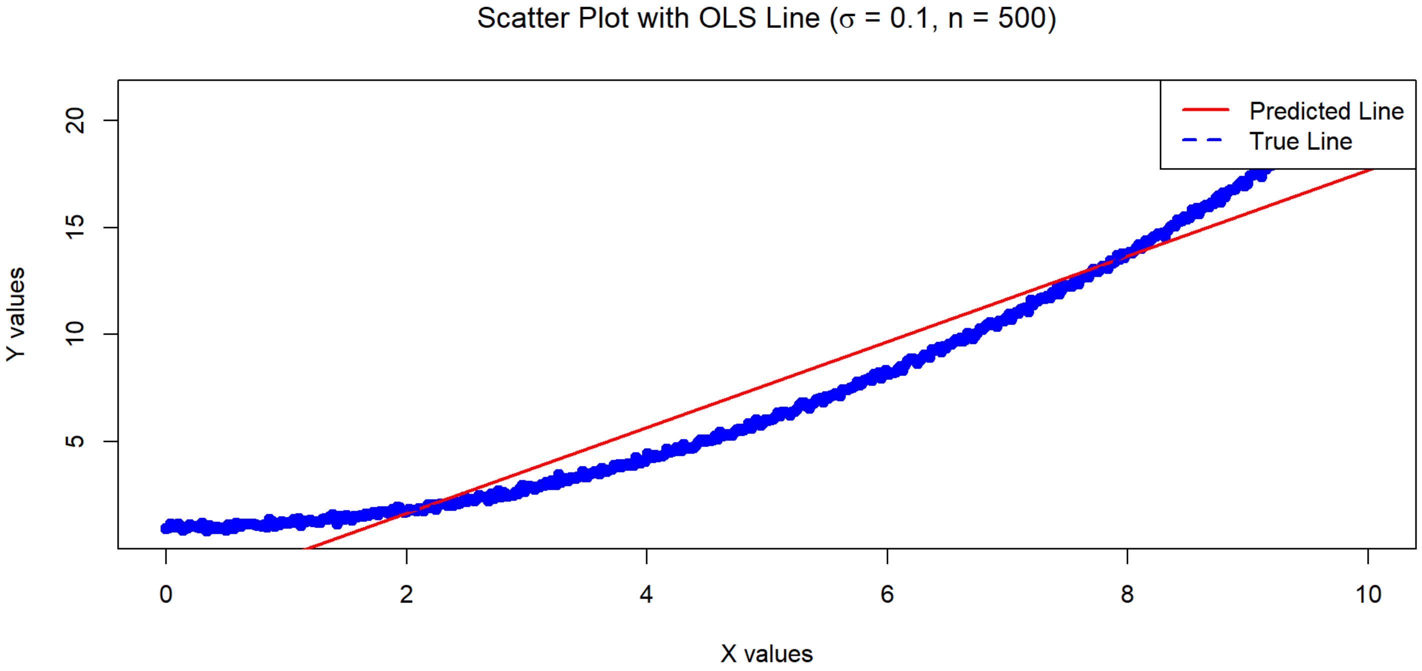
Assessing the models

- How do we know which model estimates are reasonable?
    - Estimates for examples 1, 2 and 4 seem very good (low bias and low standard error)
    - However we are less confident in example 3 (low bias but high standard error)
    - Pretty confident in example 5 despite a similar standard error to example 3.
    - Can we quantify this uncertainty in terms of confidence intervals / hypothesis testing?
- Consider the next example, it has low variance but it doesn't look 'right'.

Example 6

The below data was generated by $Y = 1 + 0.2 \times X^2 + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 0.01)$ with $n = 30$.

```
Estimates of Beta_0 and Beta_1:
 -2.32809 2.000979

Variances of the estimates:
 0.01808525 0.0005420144
```



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 β_0 and β_1?
  - How to test the null hypothesis that there is no relationship between $\underline{X}$ and $\underline{Y}$?
  - How to test if the influence of the exogenous variable (X) on the endogenous variable (Y) is larger/smaller than some value?

Note

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

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:

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

- for β_0:

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

See rationale slide.

S_{xx} ↑ as n ↑

the intervals shrink.

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

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

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

  $$t(\hat{\beta}_1) = \frac{\hat{\beta}_1 - \tilde{\beta}_1}{\hat{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \tilde{\beta}_1}{\left(s \,/\, \sqrt{S_{xx}}\,\right)} \quad \bigstar$$

  which has a t_{n-2} distribution under the H_0 (see rationale slide).
- The construction of the hypothesis test is the same for β_0.

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**
$\beta_1 \neq \tilde{\beta}_1$	$\left\lvert t\left(\hat{\beta}_1\right)\right\rvert > t_{1-\alpha/2,n-2}$
$\beta_1 > \tilde{\beta}_1$	$t\left(\hat{\beta}_1\right) > t_{1-\alpha,n-2}$
$\beta_1 < \tilde{\beta}_1$	$t\left(\hat{\beta}_1\right) < -t_{1-\alpha,n-2}$

- Typically only interested in testing $H_0 : \underline{\beta_1 = 0}$ vs. $H_1 : \beta_1 \neq 0$, as this informs us whether our β_1 is significantly different from 0.
    - I.e., including the slope parameter is worth it!
- Similar construction for β_0 test, and again typically only test against 0.

Example 1 - Hypothesis testing

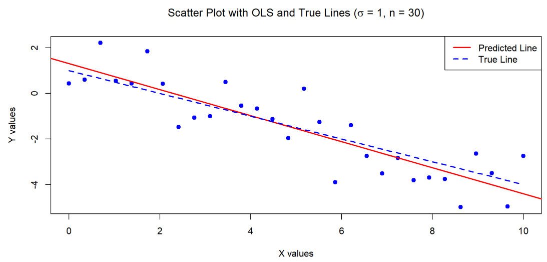
The below data was generated by $Y = 1 - 0.5 \times X + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 1)$ with $n = 30$.

Tested against $\hat{\beta}_j = 0$

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8580 -0.7026 -0.1236  0.5634  1.8463 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.30963    0.34686   3.776 0.000764 ***
X           -0.57135    0.05957  -9.592  2.4e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9738 on 28 degrees of freedom
Multiple R-squared:  0.7667,    Adjusted R-squared:  0.7583 
F-statistic:    92 on 1 and 28 DF,  p-value: 2.396e-10
```

Output from r.

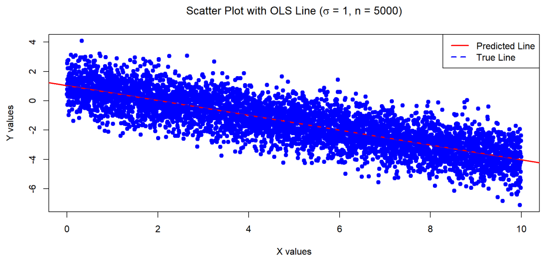
later



Example 2 - Hypothesis testing

The below data was generated by $Y = 1 - 0.5 \times X + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 1)$ with $n = 5000$.

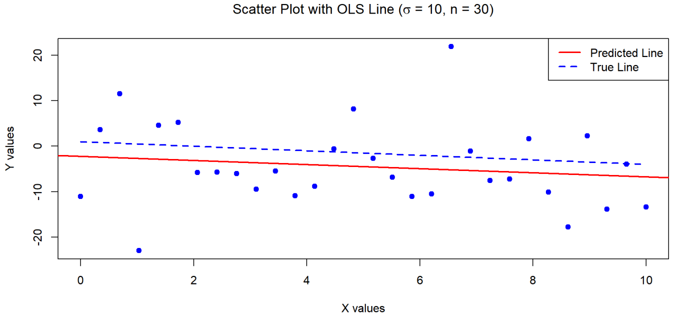
```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1179 -0.6551 -0.0087  0.6655  3.4684 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.028116   0.028125   36.55   <2e-16 ***
X           -0.505737   0.004871 -103.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9945 on 4998 degrees of freedom
Multiple R-squared:  0.6832,    Adjusted R-squared:  0.6831 
F-statistic: 1.078e+04 on 1 and 4998 DF,  p-value: < 2.2e-16
```



Example 3 - Hypothesis testing

The below data was generated by $Y = 1 - 0.5 \times X + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 100)$ with $n = 30$.

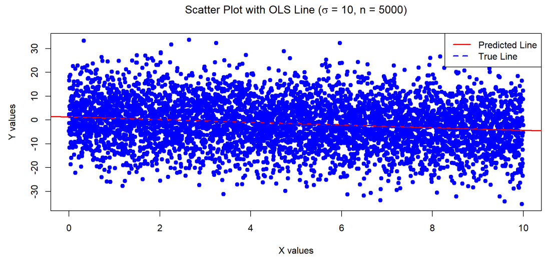
```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.306  -5.751  -2.109   5.522  27.049 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.1999     3.2730  -0.672    0.507
X            -0.4529     0.5621  -0.806    0.427

Residual standard error: 9.189 on 28 degrees of freedom
Multiple R-squared:  0.02266,	Adjusted R-squared:  -0.01225 
F-statistic: 0.6492 on 1 and 28 DF,  p-value: 0.4272
```



Example 4 - Hypothesis testing

The below data was generated by $Y = 1 - 0.5 \times X + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 100)$ with $n = 5000$.

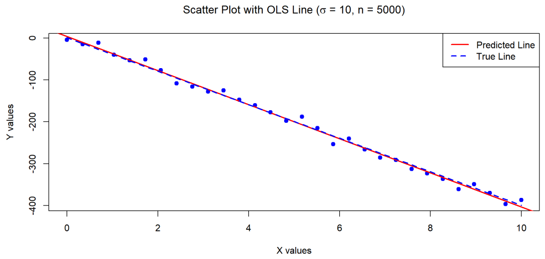
```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.179  -6.551  -0.087   6.655  34.684 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.28116    0.28125   4.555 5.36e-06 ***
X           -0.55737    0.04871 -11.442  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.945 on 4998 degrees of freedom
Multiple R-squared:  0.02553,	Adjusted R-squared:  0.02533 
F-statistic: 130.9 on 1 and 4998 DF,  p-value: < 2.2e-16
```



Example 5 - Hypothesis testing

The below data was generated by $Y = 1 - 40 \times X + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 100)$ with $n = 30$.

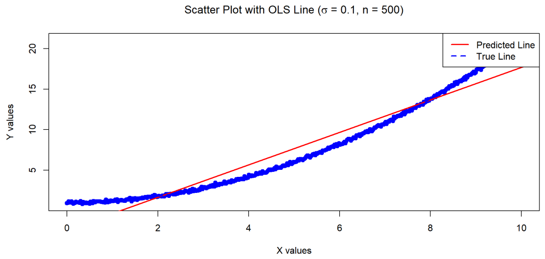
```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.580  -7.026  -1.236   5.634  18.463 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.0963     3.4686   1.181    0.248    
X           -40.7135     0.5957 -68.350   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.738 on 28 degrees of freedom
Multiple R-squared:  0.994, Adjusted R-squared:  0.9938 
F-statistic:  4672 on 1 and 28 DF,  p-value: < 2.2e-16
```



Example 6 - Hypothesis testing

The below data was generated by $Y = 1 + 0.2 \times X^2 + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 0.01)$ with $n = 30$.

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8282 -1.3467 -0.4217  1.1207  3.4041 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.32809    0.13448  -17.31   <2e-16 ***
X            2.00098    0.02328   85.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.506 on 498 degrees of freedom
Multiple R-squared:  0.9368,	Adjusted R-squared:  0.9367 
F-statistic:  7387 on 1 and 498 DF,  p-value: < 2.2e-16
```



Summary of hypothesis tests

Below is the summary of the hypothesis tests for whether β_j are statistically different from 0 for the six examples at the 5% level.

	1	**2**	**3**	**4**	**5**	**6**
β_0	Y	Y	N	Y	N	Y
β_1	Y	Y	N	Y	Y	Y

Does that mean the models that are significant at 5% for both β_0 and β_1 are equivalently 'good' models?

- No! Example 6 is significant but clearly the underlying relationship is not linear.

Assessing the accuracy of the model

We have the following so far:

- Data plotting with model predictions overlayed.
- Estimates of a linear model coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Standard errors and hypothesis tests on the coefficients.

But how do we assess whether a model is ‘good’ or ‘accurate’? Example 5 looks arguably the best while clearly example 6 is by far the worst.

Assessing the Accuracy of the Model

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

$$\underbrace{y_i - \overline{y}}_{\text{total deviation}} = \underbrace{(y_i - \hat{y}_i)}_{\substack{\text{unexplained deviation} \\ \| \\ \varepsilon_i}} + \underbrace{(\hat{y}_i - \overline{y})}_{\text{explained deviation}}.$$

We then obtain:

$$\text{Variance of } y. \longrightarrow \underbrace{\sum_{i=1}^{n} (y_i - \overline{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{\substack{\text{RSS} \\ = \sum \varepsilon_i^2}} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}_{\text{SSM}},$$

where:

- TSS: **total sum of squares**;
- 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:

- TSS is the total variability in the absence of knowledge of the variable X. It is the total square deviation away from its average;
- 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:

Explained

Unexplained

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

Have you explained enough variation to say your model is better than nothing

Assessing the Accuracy of the Model

Noting that:

$$\mathrm{RSS} = \boxed{\underbrace{S_{yy}}_{=\mathrm{TSS}} - \underbrace{\hat{\beta}_1 S_{xy}}_{=\mathrm{SSM}},}$$

we can define the R^2 **statistic** as:

$$R^2 = \left(\frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}\right)^2 = \hat{\beta}_1 \frac{S_{xy}}{S_{yy}} = \frac{\hat{\beta}_1 S_{xy}}{\mathrm{SST}} = \frac{\mathrm{SSM}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

sample

Square/ of correlation

- 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 is the square of the sample correlation between Y and X in simple linear regression.
  - Hence takes a value between 0 and 1.

Proof: See Lab questions

Summary of R^2 from the six examples

Below is a table of the R^2 for all of the six examples:

	1	**2**	**3**	**4**	**5**	**6**
R^2	0.76	0.68	0.02	0.03	0.99	0.89

- The R^2 for 1, 2, and 3, 4 are more or less equivalent.
  - As expected since we only changed n.
- Example 5 has the highested R^2 despite having an insignificant β_0.
- Example 6 has a higher R^2 than 1-4, despite it clearly not being linear.
- Example 6 does not satisfy either the weak or strong assumptions, the results cannot be trusted. (More on this later)
- **There is more to modelling than looking at numbers!**

Lecture Outline

- Simple Linear Regression
- **Multiple Linear Regression**
- Categorical predictors
- R Demo
- ANOVA
- Linear model selection
- Potential problems with Linear Regression
- So what's next
- Appendices

Overview

- Extend the simple linear regression model to accommodate multiple predictors

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

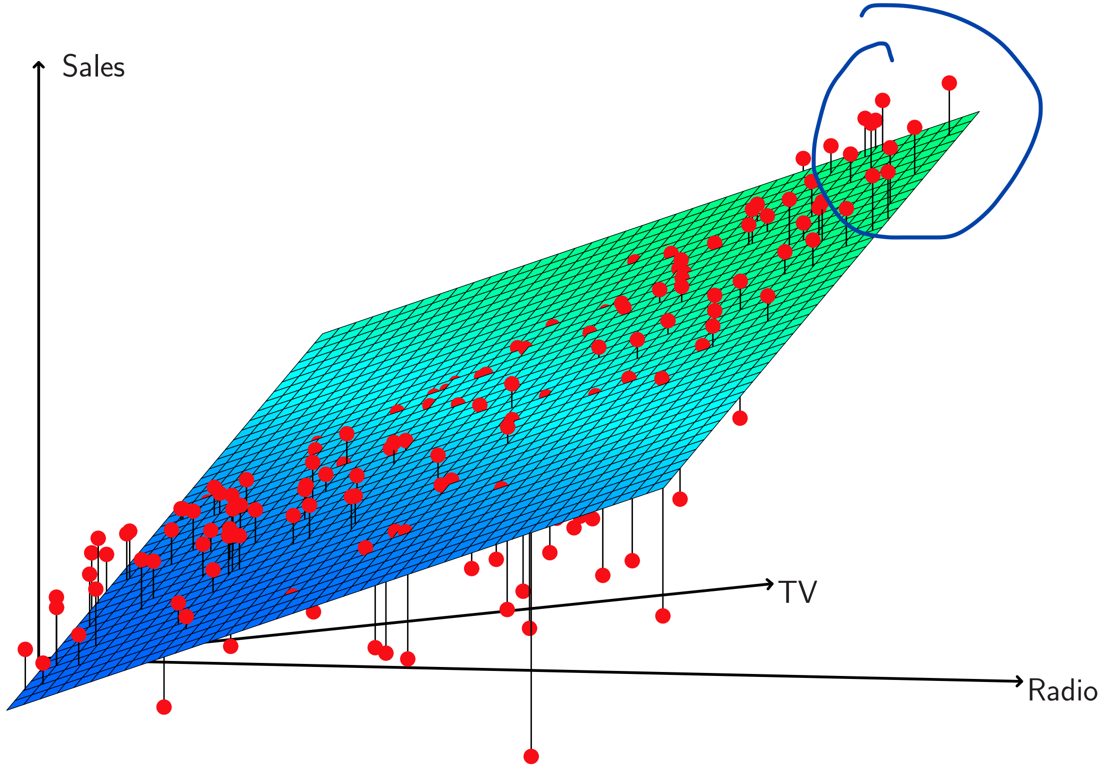
predictor variable number.

  - Recall $Y = (y_1, ..., y_n)^\top$ and we denote $X_j = (x_{1j}, x_{2j}, ..., x_{nj})^\top$.
  - Data is now paired as $(y_{11}, x_{11}, x_{12}, ..., x_{1p}), ..., (y_{n1}, x_{n1}, ..., x_{np})$.
- β_j: the average effect on y_{ij} of a one unit increase in x_{ij}, holding all $x_{ik}, k \neq j$ variables fixed.
- Instead of fitting a line, we are now fitting a (hyper-)plane
- Important note: If we denote x_i to be the i'th row of X, you should observe that the response Y is still linear with respect to the predictors since

$$y_i = \underline{\mathrm{x}_i}\beta + \epsilon_i$$

Advertising Example

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



Linear Algebra and Matrix Approach

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

$$Y = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

The model can be re-written as:

$$Y = X\beta + \epsilon$$

with $\beta = (\beta_0, \beta_1, ..., \beta_p)^\top$, Y and ϵ is defined the same as simple linear regression. The matrix X is given by

r $(n \times (p+1))$

- Try expanding first row.

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{(p+1,\ 1)}$$

Note that the matrix X is of size $(n, p + 1)$ and β is a $p + 1$ column vector.

- Verify all the dimensions make sense, expand it! Also verify simple linear regression can be recovered from this notation.
- Take careful note of the notation in different contexts. Here X is a matrix, while in simple linear regression it was a column vector. Depending on the context it should be obvious which is which.

Assumptions of the Model

Weak Assumptions:

The error terms ϵ_i satisfy the following:

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

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

$$\mathbb{E}\left[\epsilon\right] = \underline{0}; \qquad \mathrm{Cov}\left(\epsilon\right) = \sigma^2 I_n,$$

where I_n is the $n \times n$ identity matrix.

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

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)

$$\begin{aligned}\mathrm{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip}\right)^2 \\ &= (Y - X\beta)^\top (Y - X\beta) = \sum_{i=1}^{n}\hat{\epsilon}_i^2.\end{aligned}$$

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

$$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y.$$

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

$$\hat{Y} = X\hat{\beta}.$$

Least Squares Estimates (LSE) - Properties

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

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

$$s^2 = \frac{1}{n-p-1}\left(Y - \hat{Y}\right)^\top \left(Y - \hat{Y}\right) = \frac{\mathrm{RSS}}{n-p-1},$$

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

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

Lecture Outline

- Simple Linear Regression
- Multiple Linear Regression
- **Categorical predictors**
- R Demo
- ANOVA
- Linear model selection
- Potential problems with Linear Regression
- So what’s next
- Appendices

Qualitative predictors

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?

- Consider for example the problem of predicting salary for a potential job applicant:
  - A quantitative variable could be years of relevant work experience.
  - A two-category variable could be is the applicant currently an employee of this company? (T/F)
  - A multiple-category variable could be highest level of education? (HS diploma, Bachelors, Masters, PhD) How do we incorporate this qualitative data into our modelling?

Integer encoding

One solution - assign the values of the categories to a number.

- E.g., $(HS, B, M, P) = (1, 2, 3, 4)$.

Problem? The numbers you use specify a relationship between the categories. For example, we are saying a Bachelors degree is above a HS diploma (in particular, is worth 2x more). So $\beta_{edu}(B) = 2 \times \beta_{edu}(HS)$.

- $(HS, B, M, P) = (4, 7, 2, 3)$.

Now this gives an interpretation that a HS diploma is worth more than a PhD but less than a Bachelors?

- What if the categories are completely unrelated like colours (green, blue, red, yellow)?

One-hot encoding

Another solution is to use a technique called *one-hot encoding*. Create a set of binary variables that take 0 or 1 depending if the variable belongs to a certain category.

- **Use one-hot encoding when the categories have no ordinal relationship between them.**
- E.g., if if we have (red, green, green, blue) the dummy encoded matrix could be:

$$\begin{pmatrix} R \\ G \\ G \\ B \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where the first column represents red, second green and third blue.

Dummy encoding

Technically, we **cannot** use one-hot encoding in linear regression, but instead use a technique called *dummy encoding*.

We pick a base case, i.e. set the entry of the row of the matrix to be 0 if it's the base case.

Using the same example as before and we set 'Red' to be the base case we have:

$$\begin{pmatrix} R \\ G \\ G \\ B \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix},$$

where now the first column is green, second is blue. If both columns are 0, then it represents red (implicitly).

- Need this to prevent a singularity in $(X^\top X)$, since the first column of X are $1'$s (recall your definition of linear independence!)
- Bonus question: What if we remove the intercept column in our design matrix X? Do we still need a base case?

Lecture Outline

- Simple Linear Regression
- Multiple Linear Regression
- Categorical predictors
- **R Demo**
- ANOVA
- Linear model selection
- Potential problems with Linear Regression
- So what’s next
- Appendices

The matrix approach

TV	radio	sales
230.1	37.8	22.1
44.5	39.3	10.4
17.2	45.9	9.3
151.5	41.3	18.5
180.8	10.8	12.9
8.7	48.9	7.2
57.5	32.8	11.8
120.2	19.6	13.2
8.6	2.1	4.8
199.8	2.6	10.6
66.1	5.8	8.6

$$Y = X\beta + \epsilon$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

```r
library(tidyverse)
site <- url("https://www.statlearning.com/s/Advertising.csv")
df_adv <- read_csv(site, show_col_types = FALSE)
X <- model.matrix(~ TV + radio, data = df_adv);
y <- df_adv[, "sales"]
```

```r
head(X)
```

```
  (Intercept)    TV radio
1           1 230.1  37.8
2           1  44.5  39.3
3           1  17.2  45.9
4           1 151.5  41.3
5           1 180.8  10.8
6           1   8.7  48.9
```

```r
head(y)
```

```
# A tibble: 6 × 1
  sales
  <dbl>
1  22.1
2  10.4
3   9.3
4  18.5
5  12.9
6   7.2
```

Brief refresher

Fitting: Minimise the residuals sum of squares

$$\begin{aligned}\mathrm{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \ldots - \hat{\beta}_p x_{i,p}\right)^2 \\ &= (Y - X\beta)^\top (Y - X\beta)\end{aligned}$$

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

$$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y.$$

Predicting: The predicted values are given by

$$Y = X\hat{\beta}.$$

R's `lm` and `predict`

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

```r
model <- lm(sales ~ TV + radio, data = df_adv)
coef(model)
```

```
(Intercept)          TV       radio 
 2.92109991  0.04575482  0.18799423 
```

```r
X <- model.matrix(~ TV + radio, data = df_adv)
y <- df_adv$sales
beta <- solve(t(X) %*% X) %*% t(X) %*% y
beta
```

```
                  [,1]
(Intercept) 2.92109991
TV          0.04575482
radio       0.18799423
```

$$\hat{Y} = X\hat{\beta}.$$

```r
budgets <- data.frame(TV = c(100, 200, 300), radi
predict(model, newdata = budgets)
```

```
       1        2        3 
11.25647 17.71189 24.16731 
```

```r
X_new <- model.matrix(~ TV + radio, data = budget
X_new %*% beta
```

```
      [,1]
1 11.25647
2 17.71189
3 24.16731
```

Dummy encoding

Design matrices are normally an ‘Excel’-style table of covariates/predictors plus a column of ones.

If categorical variables are present, they are added as *dummy variables*:

```r
fake <- tibble(
  speed = c(100, 80, 60, 60, 120, 40),
  risk = c("Low", "Medium", "High",
           "Medium", "Low", "Low")
)
fake
```

```
# A tibble: 6 × 2
  speed risk
  <dbl> <chr>
1   100 Low
2    80 Medium
3    60 High
4    60 Medium
5   120 Low
6    40 Low
```

```r
model.matrix(~ speed + risk, data = fake)
```

```
  (Intercept) speed riskLow riskMedium
1           1   100       1          0
2           1    80       0          1
3           1    60       0          0
4           1    60       0          1
5           1   120       1          0
6           1    40       1          0
attr(,"assign")
[1] 0 1 2 2
attr(,"contrasts")
attr(,"contrasts")$risk
[1] "contr.treatment"
```

Dummy encoding & collinearity

Why do *dummy variables* drop the last level?

```
X_dummy = model.matrix(~ risk, data = fake)
as.data.frame(X_dummy)
```

```
  (Intercept) riskLow riskMedium
1           1       1          0
2           1       0          1
3           1       0          0
4           1       0          1
5           1       1          0
6           1       1          0
```

```
solve(t(X_dummy) %*% X_dummy)
```

```
            (Intercept)   riskLow riskMedium
(Intercept)           1 -1.000000       -1.0
riskLow              -1  1.333333        1.0
riskMedium           -1  1.000000        1.5
```

```
X_oh <- cbind(X_dummy, riskHigh = (fake$risk == "
as.data.frame(X_oh)
```

```
  (Intercept) riskLow riskMedium riskHigh
1           1       1          0        0
2           1       0          1        0
3           1       0          0        1
4           1       0          1        0
5           1       1          0        0
6           1       1          0        0
```

```
solve(t(X_oh) %*% X_oh)
```

```
Error in solve.default(t(X_oh) %*% X_oh): system is
computationally singular: reciprocal condition number =
6.93889e-18
```

Lecture Outline

- Simple Linear Regression
- Multiple Linear Regression
- Categorical predictors
- R Demo
- **ANOVA**
- Linear model selection
- Potential problems with Linear Regression
- So what's next
- Appendices

Test the Relationship Between the Response and Predictors

• Does our model explain a significant proportion of the variance in Y?

The below is a test to if the multiple linear regression model is significantly better than just predicting the mean $\bar{Y}$.

$Y = X\beta + \varepsilon$

• This test does not say $\beta_j \neq 0$

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

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

*

- F-statistic $= \frac{(\text{TSS}-\text{RSS})/p}{\text{RSS}/(n-p-1)} \sim F_{p,n-p-1}$
- Verify the F-test gives the same conclusion as the t-test on $\underline{\beta_1 \neq 0}$ for simple linear regression!
- Question: Given the individual p-values for each variable, why do we need to look at the overall F-statistics?
  - Because a model with all insignificant p-values may jointly still be able to explain a significant proportion of the variance.
  - Conversely, a model with significant predictors may still fail to explain a significant proportion of the variance.

Analysis of variance (ANOVA)

$TSS = \sum (Y_i - \bar{Y})^2 = (n-1)\cdot \widehat{Var}(Y)$

The sums of squares are interpreted as follows:

- TSS is the total variability in the absence of knowledge of the variables $X_1, \ldots, X_p$;
- RSS is the total variability remaining after introducing the effect of $X_1, \ldots, X_p$; $= \hat{\varepsilon}^T \hat{\varepsilon}$
- SSM is the total variability "explained" because of knowledge of $X_1, \ldots, X_p$.

ANOVA - Week 5 GLM

This partitioning of the variability is used in ANOVA tables:

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

$$TSS = RSS + SSM$$

$$SSM = TSS - RSS$$

Model Fit and Predictions

- Measure model fit (similar to the simple linear regression)
  - Residual standard error (RSE)
  - $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$
- Uncertainties associated with the prediction
  - $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ are estimates. Still have the t-tests to test individual significance.
  - linear model is an approximation
  - random error ϵ

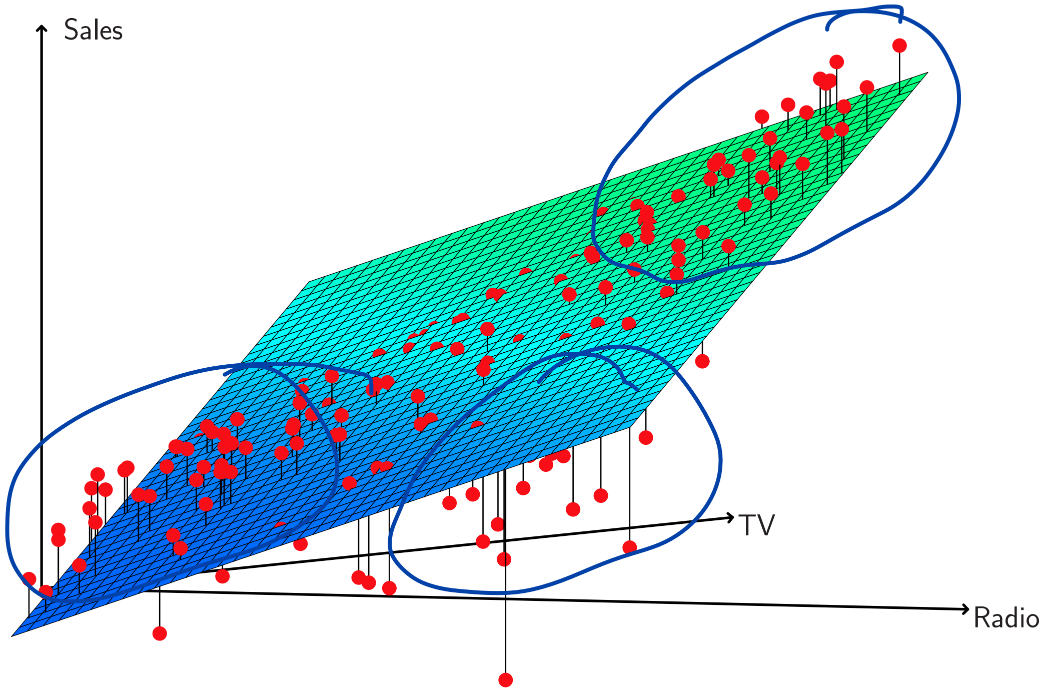
Constructed similarly as the SLR ~~to~~ tests

– True data may not be linear in X

$$Y = X\beta + \underline{\varepsilon}$$

Advertising Example (continued)

Linear regression fit using TV and Radio:



What do you observe?

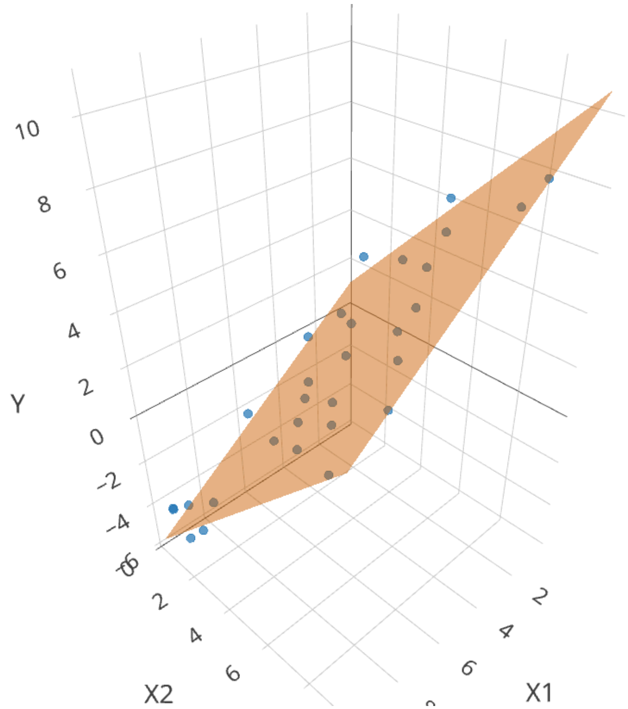
Other Considerations in the Regression Model

- Qualitative predictors
  - two or more levels, with no logical ordering - Dummy encoding
  - create binary (0/1) dummy variables
  - Need (#levels - 1) dummy variables to fully encode
- Interaction terms $(X_i X_j)$ (removing the additive assumption)
- Quadratic terms (X_i^2) (non-linear relationship) - Week 8

UNSW SYDNEY

Example 7 - Data plot

The below data was generated by $Y = 1 - 0.7 \times X_1 + X_2 + \epsilon$ where $X_1, X_2 \sim U[0, 10]$ and $\epsilon \sim N(0, 1)$ with $n = 30$.



Example 7 - Model summary

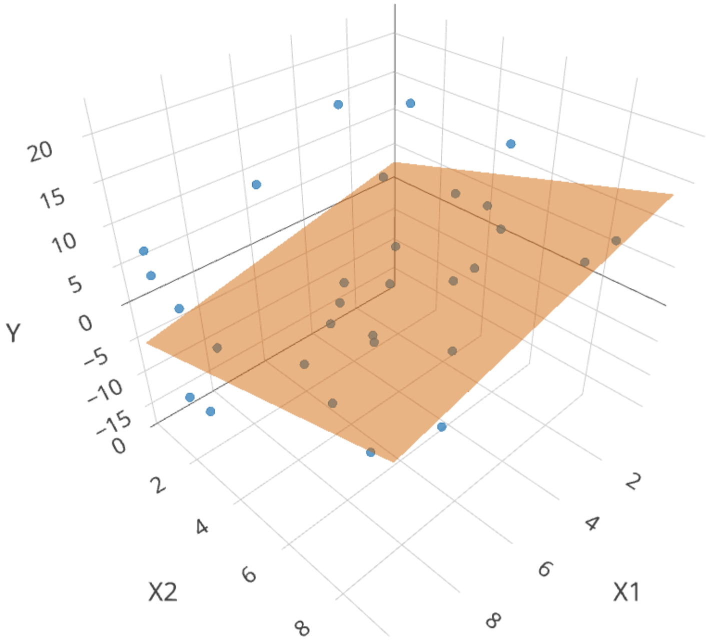
The below data was generated by $Y = 1 - 0.7 \times X_1 + X_2 + \epsilon$ where $X_1, X_2 \sim U[0, 10]$ and $\epsilon \sim N(0, 1)$ with $n = 30$.

```
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6923 -0.4883 -0.1590  0.5366  1.9996

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.22651    0.45843   2.675   0.0125 *
X1          -0.71826    0.05562 -12.913 4.56e-13 ***
X2           1.01285    0.05589  18.121  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8625 on 27 degrees of freedom
Multiple R-squared:  0.9555,    Adjusted R-squared:  0.9522
F-statistic: 290.1 on 2 and 27 DF,  p-value: < 2.2e-16
```

✓ Model explains significant proportion of the variance

Example 8 - Data plot

The below data was generated by $Y = 1 - 0.7 \times X_1 + X_2 + \epsilon$ where $X_1, X_2 \sim U[0, 10]$ and $\epsilon \sim N(0, 100)$ with $n = 30$.



Example 8 - Model summary

The below data was generated by $Y = 1 - 0.7 \times X_1 + X_2 + \epsilon$ where $X_1, X_2 \sim U[0, 10]$ and $\epsilon \sim N(0, 100)$ with $\underline{n = 30}$.

$H_0: \beta_1 = \beta_2 = 0$

$H_a: \beta_j \neq 0$

```
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max
-16.923  -4.883  -1.591   5.366  19.996

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2651     4.5843   0.712   0.4824
X1           -0.8826     0.5562  -1.587   0.1242
X2            1.1285     0.5589   2.019   0.0535 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.625 on 27 degrees of freedom
Multiple R-squared:  0.2231,	Adjusted R-squared:  0.1656
F-statistic: 3.877 on 2 and 27 DF,  p-value: 0.03309
```

None are significant at 5%

$SE[\varepsilon_i]$

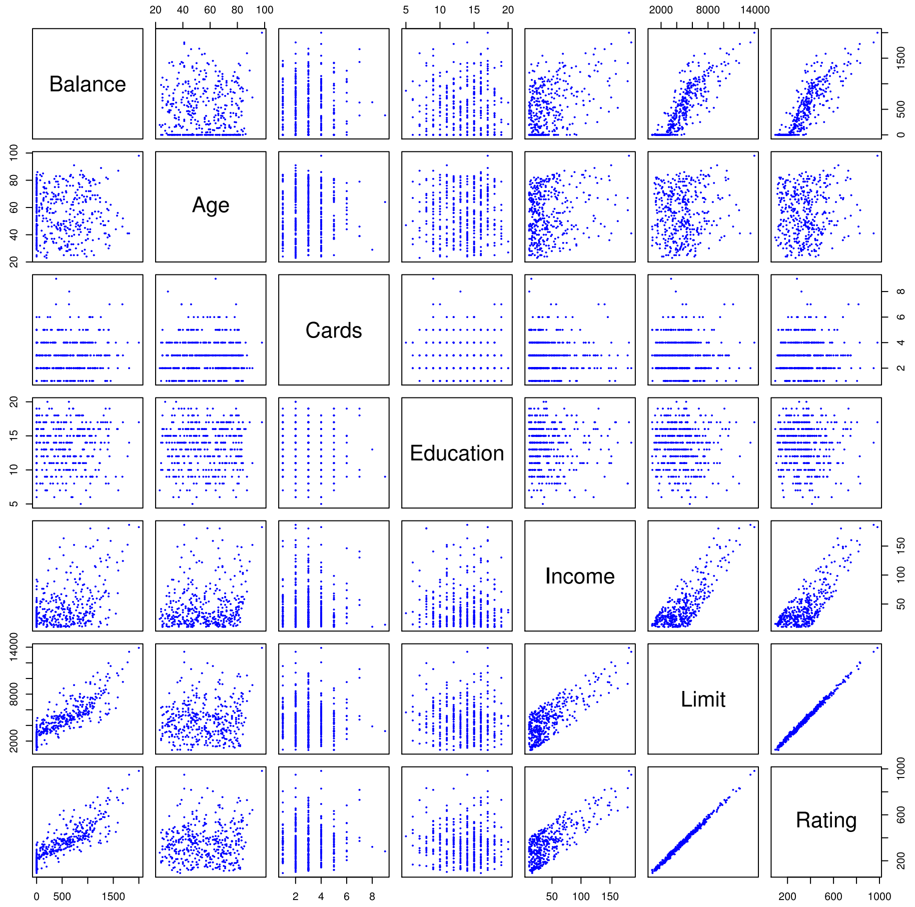
*

at 5%

Lecture Outline

- Simple Linear Regression
- Multiple Linear Regression
- Categorical predictors
- R Demo
- ANOVA
- **Linear model selection**
- Potential problems with Linear Regression
- So what's next
- Appendices

The credit dataset



Qualitative covariates: own, student, status, region

Linear Model selection

– How do we select the best model?

What do I mean here?

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

Subset selection

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

Best subset selection

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

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

Algorithm:

- Consider the models with $\underline{0}$ predictors, and call this $\underline{\mathcal{M}}_0$. This is the null model / $\beta_0 = \bar{Y}$
- Consider all models with 1 predictor, pick the best fit, and call this $\mathcal{M}_1$
- . . . $X_1, X_2, \ldots, X_p$ $Y = \beta_0 + \beta_j X_j$
- Consider the model with p predictor, and call this $\mathcal{M}_p$. This is the full model
- Pick the best fit of $\underline{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p}$

$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$

Take 'best' model from all of the p model

Best subset selection - behaviour

combinations

- 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:

Regression MSE

Classification Classification error.

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

10 predictors = 1024 models

100 predictors = 2^{100} huge

Stepwise Example: Forward stepwise selection

Algorithm:

$Y = \beta_0$

$Y = \beta_0 + \beta_j X_j$

Backwords

- Start with the null model $\mathcal{M}_0$
- 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 $\mathcal{M}_2$

  $Y = \beta_0 + \beta_j X_j + \beta_k X_k \qquad k \neq j$
- . . .
- End with the full model $\mathcal{M}_p$
- Pick the best fit of $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$

Backwords goes from $M_p \longrightarrow M_0$.

Stepwise subset selection - behaviour

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

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

- 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*

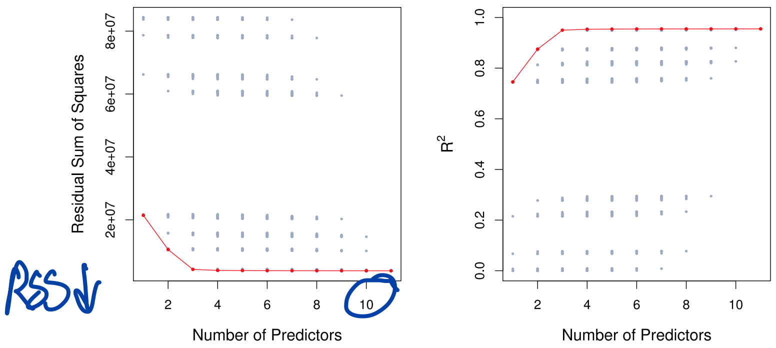
Best and stepwise do not neccessarily give the same 'best' model.

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:

R^2 is an increasing function of p (amount of predictor)

Going from $2 \rightarrow 3$ predictors cannot make R^2 worse.



RSS $\downarrow$

$R^2 \uparrow$

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 - To be covered later

Week 7.

Indirect methods

RSS + penalty

d = # predictors

1. C_p with d predictors:

$$C_p = \frac{1}{n}(\underline{\text{RSS}} + \boxed{2d\hat{\sigma}^2})$$

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

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

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

- Proportional to $\underline{C_p}$ for least squares, so gives the same results

All estimators of test error, through your training error.

Indirect methods cont.

3. Bayesian information criteria (BIC) with d predictors

$$\frac{1}{n}\left(\text{RSS} + \overset{2 \text{ for AIC}}{\underline{\log(n)}}\, d\hat{\sigma}^2\right)$$

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

Penalises extra parameters more harshly.

4. Adjusted R^2 with d predictors

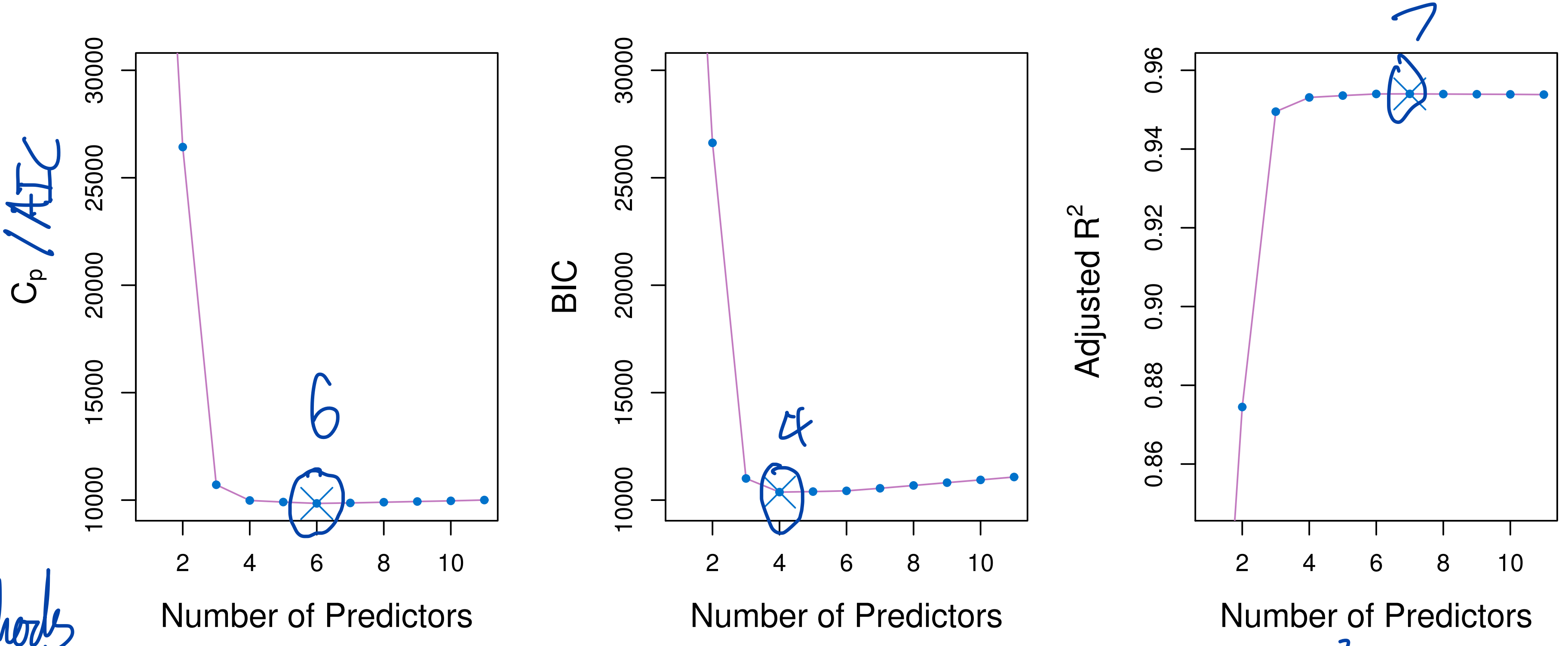
Can be negative

$$1 - \frac{\boxed{\text{RSS}/(n-d-1)}}{\text{TSS}/(n-1)} \quad \uparrow \text{ as } d \uparrow$$

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

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

How to determine the "best" model - Credit dataset



Direct methods might be better to use

Should have at least 4 predictors and no more than 8.

Lecture Outline

- Simple Linear Regression
- Multiple Linear Regression
- Categorical predictors
- R Demo
- ANOVA
- Linear model selection
- **Potential problems with Linear Regression**
- So what's next
- Appendices

Potential Problems/Concerns

$$Y = \beta_0 + \underbrace{\beta_1 X_1 + \cdots + \beta_p X_p}_{\text{linear in } X}$$

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; ← Weak assumptions
- 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). – MLE and LSE give same estimators.

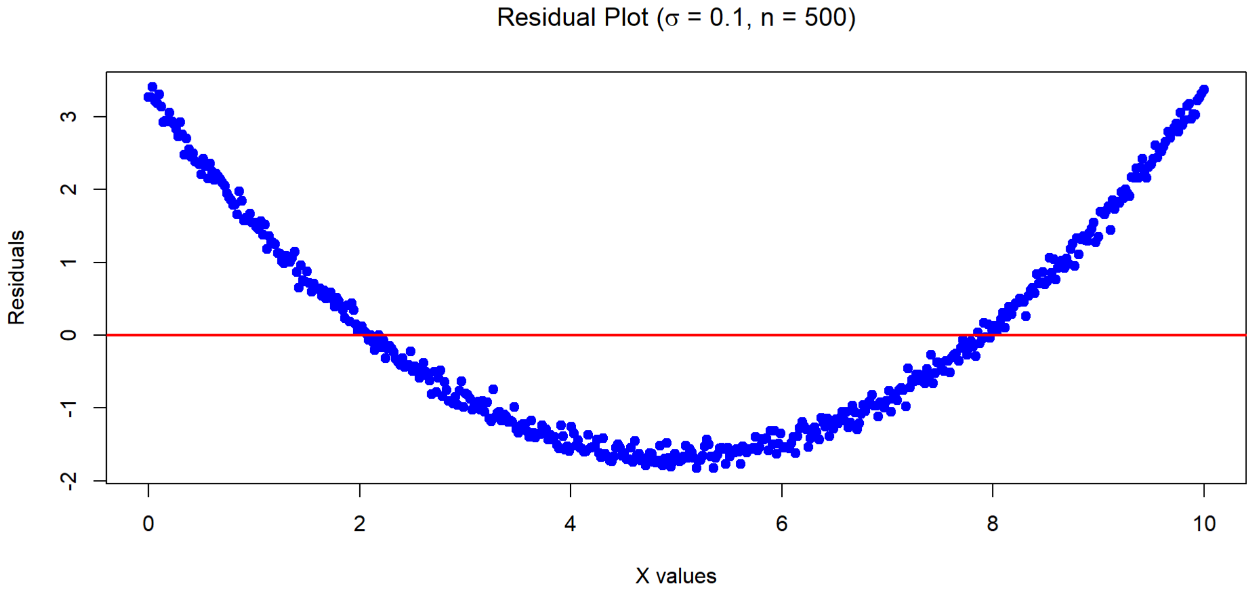
Should not use Lin regression on Example 6!

Recall Example 6 - The problems

Recall the below data was generated by $Y = 1 + 0.2 \times X^2 + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 0.01)$ with $n = 30$.

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

```
Mean of the residuals: -1.431303e-16
```



- Residuals do not have constant variance.
- Residuals indicate a linear model is not appropriate.

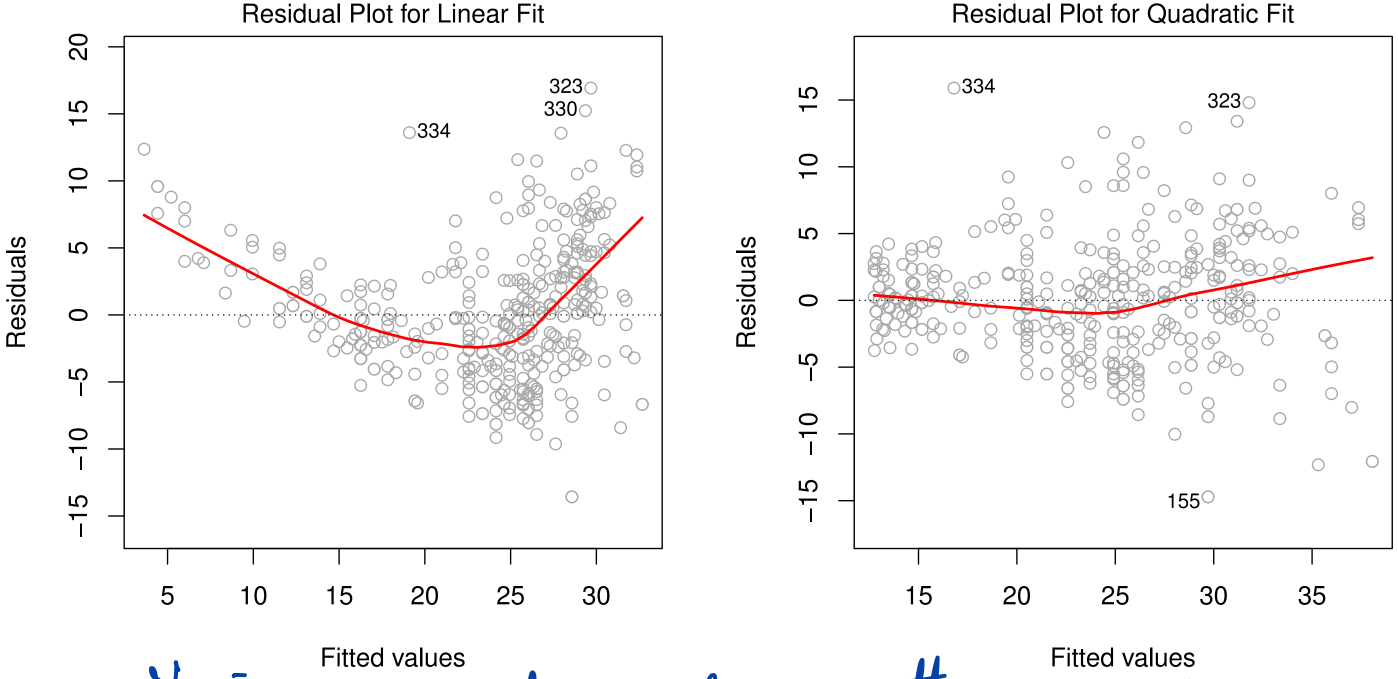
Makes sense since ✓ Y is quadratic in X.

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 - Think of example 6 as well.

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



- $Y = \beta_0 + \beta_1 X_1^2$

Variance is not constant, pattern present.

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

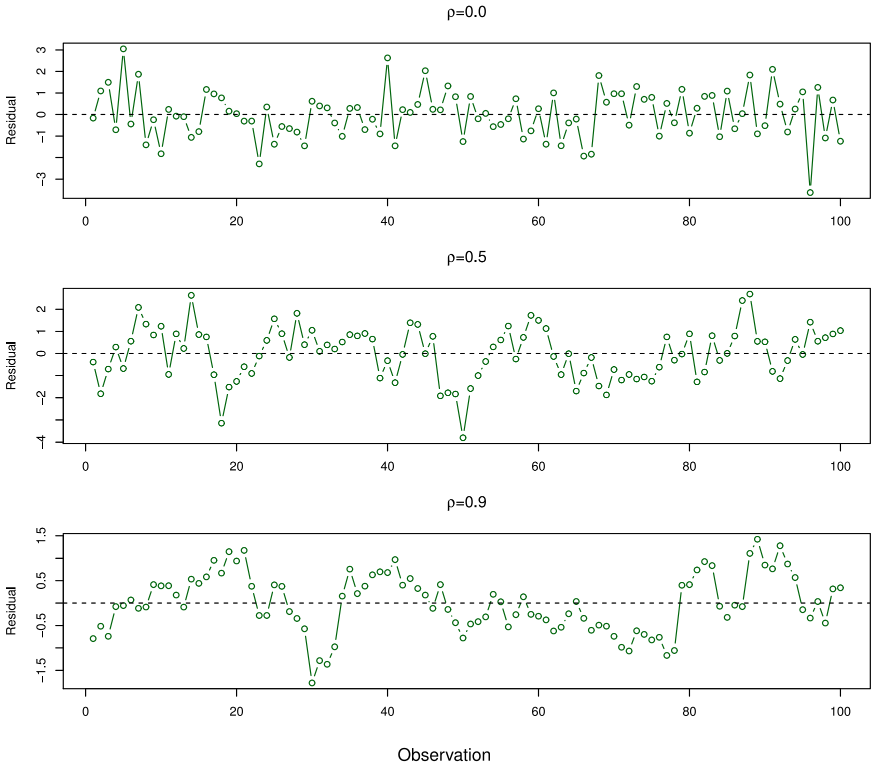
Weak assumptions

$Cov(\varepsilon_i, \varepsilon_j) = 0$

$i \neq j$

- 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.

Tests are ~~not~~ to be trusted



$\varepsilon_i = \varepsilon_{i-1} + V_i$

$V_i \sim N(0,1)$

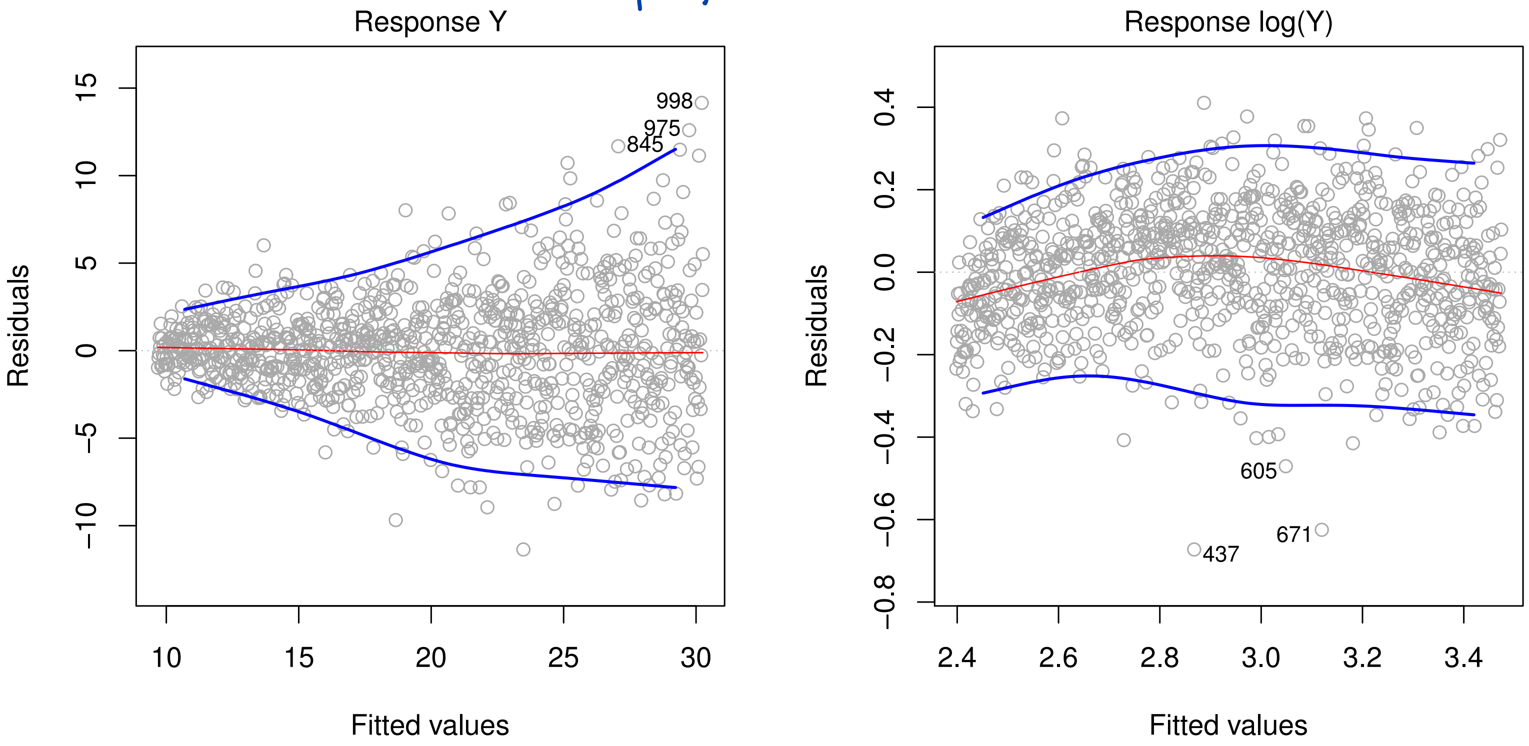
MA(1)

3. Non-constant error terms

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

Clear pattern. $Y = \beta_0 + \beta_1 X$

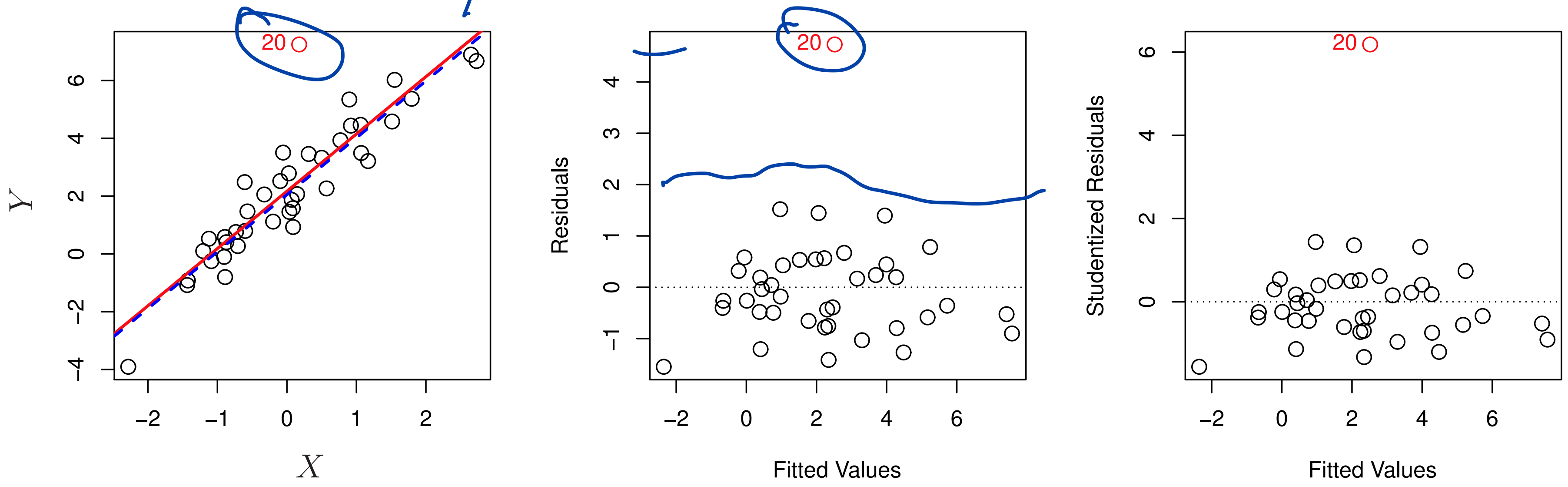
$\ln Y = \beta_0 + \beta_1 X$



In this example log transformation removed much of the heteroscedasticity.

4. Outliers

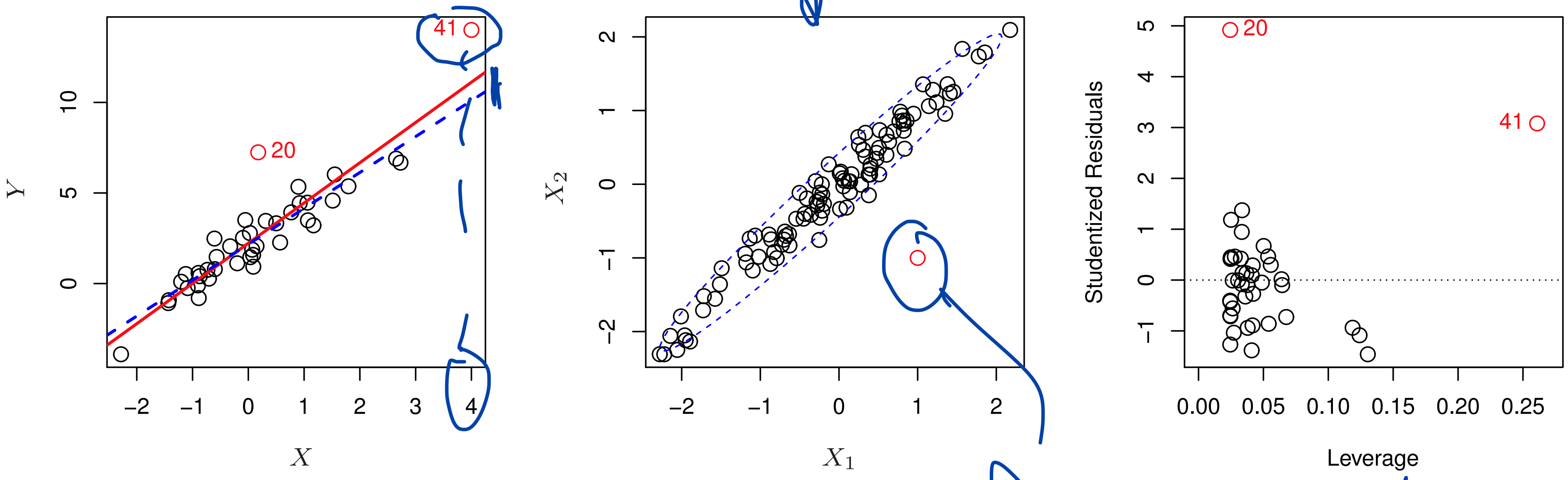
Red is line fitted with outlier, blue without.



5. High-leverage points

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

MLR problem



Fall outside usual range of predictors

High-leverage points

- 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:

$$H = X(X^\top X)^{-1}X^\top$$

$\hat{y} = HY$

- Note that

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij}y_j = \underline{\boxed{h_{ii}}y_i + \sum_{j\neq i}^{n} h_{ij}y_j}$$

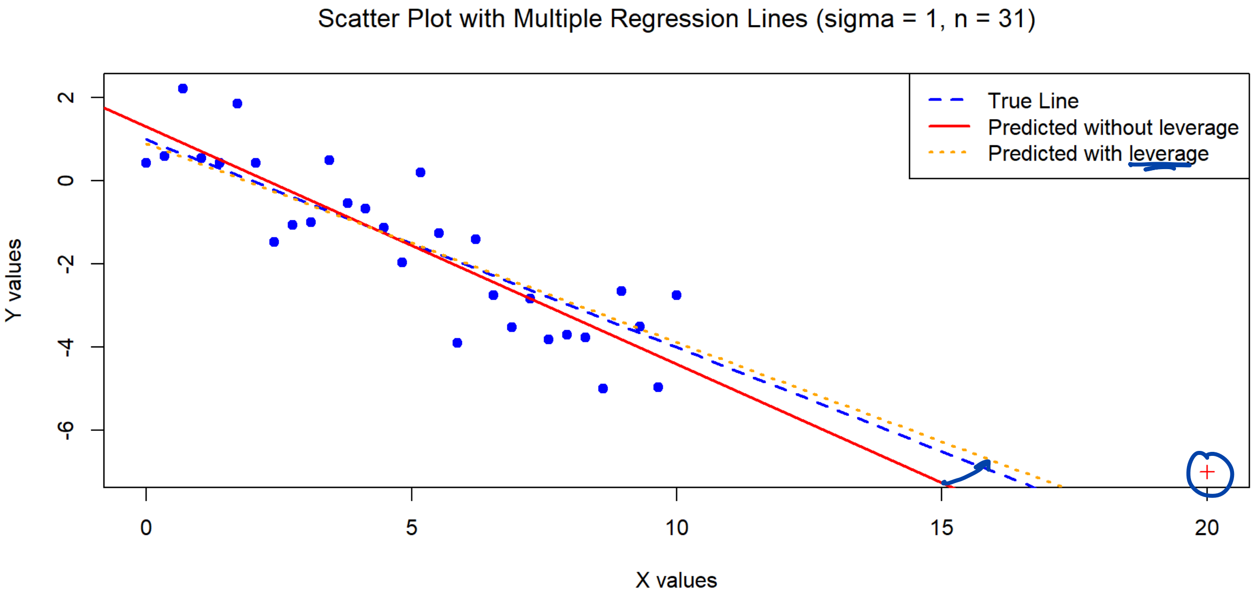
so each prediction is a linear function of all observations, and $h_{ii} = [H]_{ii}$ is the weight of observation i on its own prediction

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

High-leverage points (Example 1)

The below data was generated by $Y = 1 - 0.5 \times X + \epsilon$ where $X \sim U[0, 10]$ and $\epsilon \sim N(0, 1)$ with $n = 30$. We have added one high leverage point (made a red '+' on the scatterplot).

- This point $(y = -7, x = 20)$ has a leverage value of 0.47 >> 4/30, depsite it not being an outlier.



$X = 20$

$Y \approx 1 - 10$

$\approx -9 \quad -7$

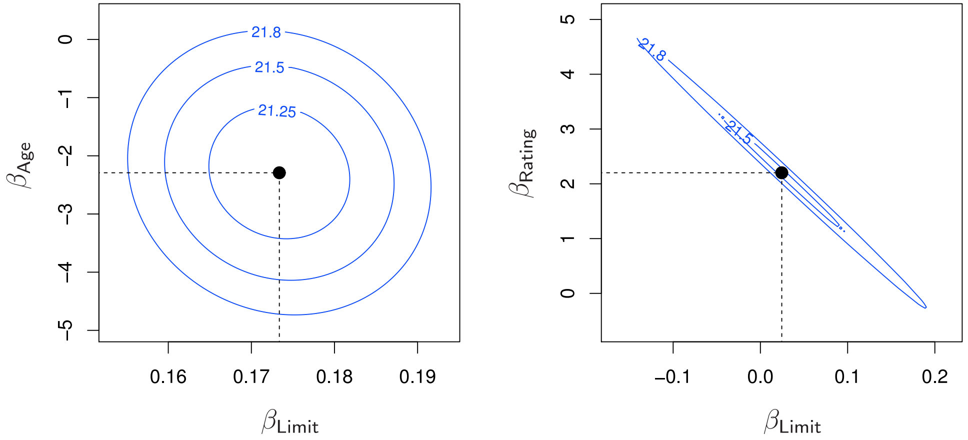
2 sd away

6. Collinearity

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

- Two or more predictor values are closely related to each other (linearly dependent)
- If a column is linearly dependent on another, the matrix $(X^\top X)$ is singular, hence non-invertible.
- 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.
- Correlation can indicate one-to-one (linear) collinearity

Collinearity makes optimisation harder



- 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

$$X_j = \gamma_0 + \gamma_1 X_1 + \dots + \gamma_{j-1} X_{j-1} + \gamma_{j+1} X_{j+1} + \dots + \gamma_p X_p.$$

- Use variance inflation factor

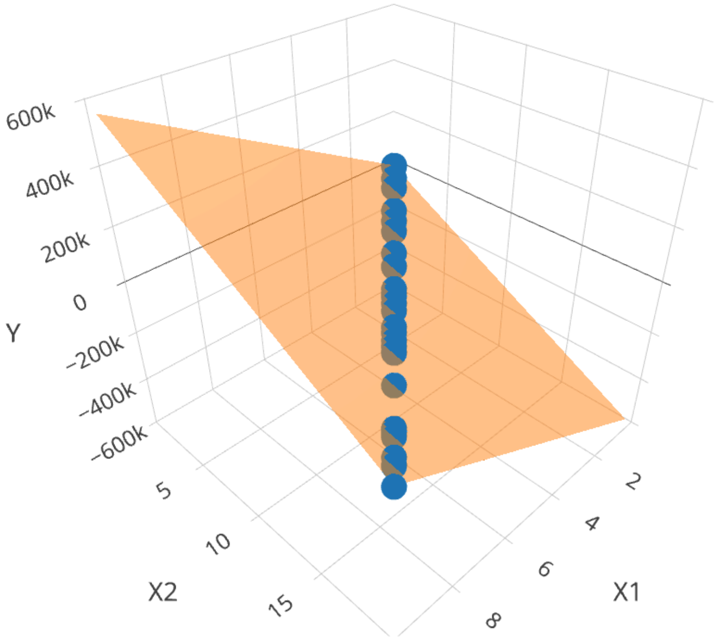
$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{\underline{X_j}|\boxed{X_{-j}}}}$$

$R^2_{X_j|X_{-j}}$ should be near 0 if X_j is not well explained by X_{-j} (Week linear relationship)

- $R^2_{X_j|X_{-j}}$ is the R^2 from $\underline{X_j}$ being regressed onto all other predictors
- Minimum 1, higher is worse ($>$ 5 or 10 is considered high)
- Recall R^2 measures the strength of the linear relationship between the response variable (X_j) against the explanatory variables (X_{-j}).

Multicollinearity example - Plot

The below data was generated by $Y = 1 - 0.7 \times X_1 + X_2 + \epsilon$ where $X_1 \sim U[0, 10]$, $X_2 = 2X_1$ and $\epsilon \sim N(0, 1)$ with $n = 30$.



Multicollinearity example - Summary and VIF

The below data was generated by $Y = 1 - 0.7 \times X_1 + X_2 + \epsilon$ where $X_1 \sim U[0, 10]$, $X_2 = 2X_1 + \varepsilon$, where $\varepsilon \sim N(0, 10^{-8})$ is a small change (to make this work) and $\epsilon \sim N(0, 1)$ with $n = 30$.

```
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.32126 -0.46578  0.02207  0.54006  1.89817 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)  5.192e-01  3.600e-01   1.442   0.1607  
X1           5.958e+04  3.268e+04   1.823   0.0793 .
X2          -2.979e+04  1.634e+04  -1.823   0.0793 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8538 on 27 degrees of freedom
Multiple R-squared:  0.9614,	Adjusted R-squared:  0.9585 
F-statistic: 335.9 on 2 and 27 DF,  p-value: < 2.2e-16

VIF for X1: 360619740351

VIF for X2: 360619740351
```

- High SE on the coefficient estimates making them unreliable.

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).

[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 $\Rightarrow$ Experience $\Rightarrow$ Aptitude for mathematics. 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: Being old doesn't necessarily mean you are good at maths!
- Hence, a predictor variable works as a predictor, but action taken on the predictor itself will have no effect.

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!

If you could use math test scores instead of age, it would be better!

- Can you measure experience in math another way?

Lecture Outline

- Simple Linear Regression
- Multiple Linear Regression
- Categorical predictors
- R Demo
- ANOVA
- Linear model selection
- Potential problems with Linear Regression
- **So what’s next**
- Appendices

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:

- *Classification problems:* logistic regression
- *Non-normality:* Generalised Linear Model - Week 4/5
- *Non-linearity:* splines and generalized additive models; KNN, tree-based methods
- *Regularised fitting:* Ridge regression and lasso
- *Non-parametric:* Tree-based methods, bagging, random forests and boosting, KNN (these also capture non-linearities)

Lecture Outline

- Simple Linear Regression
- Multiple Linear Regression
- Categorical predictors
- R Demo
- ANOVA
- Linear model selection
- Potential problems with Linear Regression
- So what's next
- **Appendices**

Appendix: Sum of squares

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

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

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

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

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

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

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

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 $\hat{\beta}_k$ is normally distributed with mean and variance

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

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

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

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

Simple linear regression: Assessing the Accuracy of the Predictions - Mean Response

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:

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

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

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

The variance of this estimator is:

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

Proof: See Lab questions.

Simple linear regression: 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:

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

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.

Simple linear regression: 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:

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

The error in our prediction is:

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

with

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

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

Proof: See Lab questions.

Simple linear regression: 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:

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

as

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

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