

Statistical modeling in logistics

MMS075

Lecture 4 – Linear regression (cont.):
interaction terms, assumptions & problems,
summary of advertising example

Acknowledgement: 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

Outline – combined lecture & exercise class

- Multiple linear regression (cont.)
 - Reviewing variable selection
 - Interaction terms
 - Practical implications of advertising example analysis
- Potential problems with linear regression
 - Non-linear relationships
 - Correlation of error terms
 - Non-constant variance of error terms (heteroscedasticity)
 - Outliers
 - High leverage points
 - Collinearity
- Feedback

Recommended resources

Reading in [ISL](#): Ch 3 intro, Sections 3.3.2-3.3.3 and 3.4, 7.1 for theory, end of 3.6.2, 3.6.4, 6.5.1, 6.5.2 and 7.8.1 for R codes

The videos from the [Statistical Learning](#) course are available at [this link](#). Relevant videos for the new material today:

- [Interactions and Nonlinearity](#) (14:16)
- [Lab: Linear Regression](#) (22:10)

For understanding the details of variable selection and model quality measures:

- [Linear Model Selection and Best Subset Selection](#) (13:44)
- [Forward Stepwise Selection](#) (12:26)
- [Backward Stepwise Selection](#) (5:26)
- [Estimating Test Error Using Mallow's Cp, AIC, BIC, Adjusted R-squared](#) (14:06)
- [Lab: Best Subset Selection](#) (10:36)
- [Lab: Forward Stepwise Selection and Model Selection Using Validation Set](#) (10:32)

Multiple linear regression continued

Reviewing variable selection

Interaction terms – Combined effect of predictors

Practical implications of advertising example analysis

Recall: best subset selection

- Measures for the quality of the model (you get them from software):

Mallow's C_p

AIC (Akaike's information criterion)

BIC (Bayesian information criterion)

Adjusted R^2

Models with small values of these measures are preferred (i.e. small values of Mallow's C_p , AIC or BIC indicate a higher quality model)

Models with a large value of adjusted R^2 are preferred, i.e. large values of adjusted R^2 indicate a higher quality model

- Best subset selection: check all subsets of predictors & choose model with best quality according to one of these measures
- Illustrated in R for Carseats when modelling Sales (next slides)

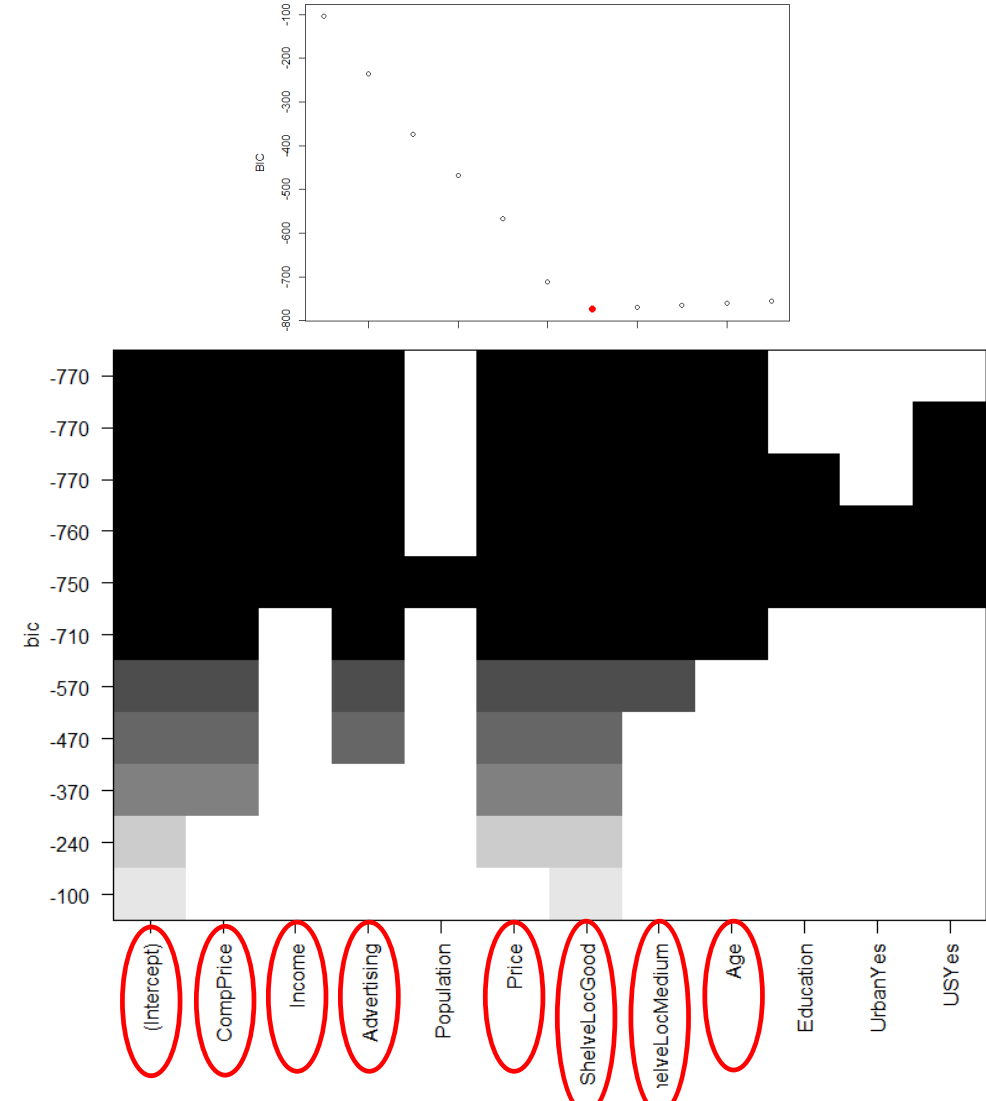
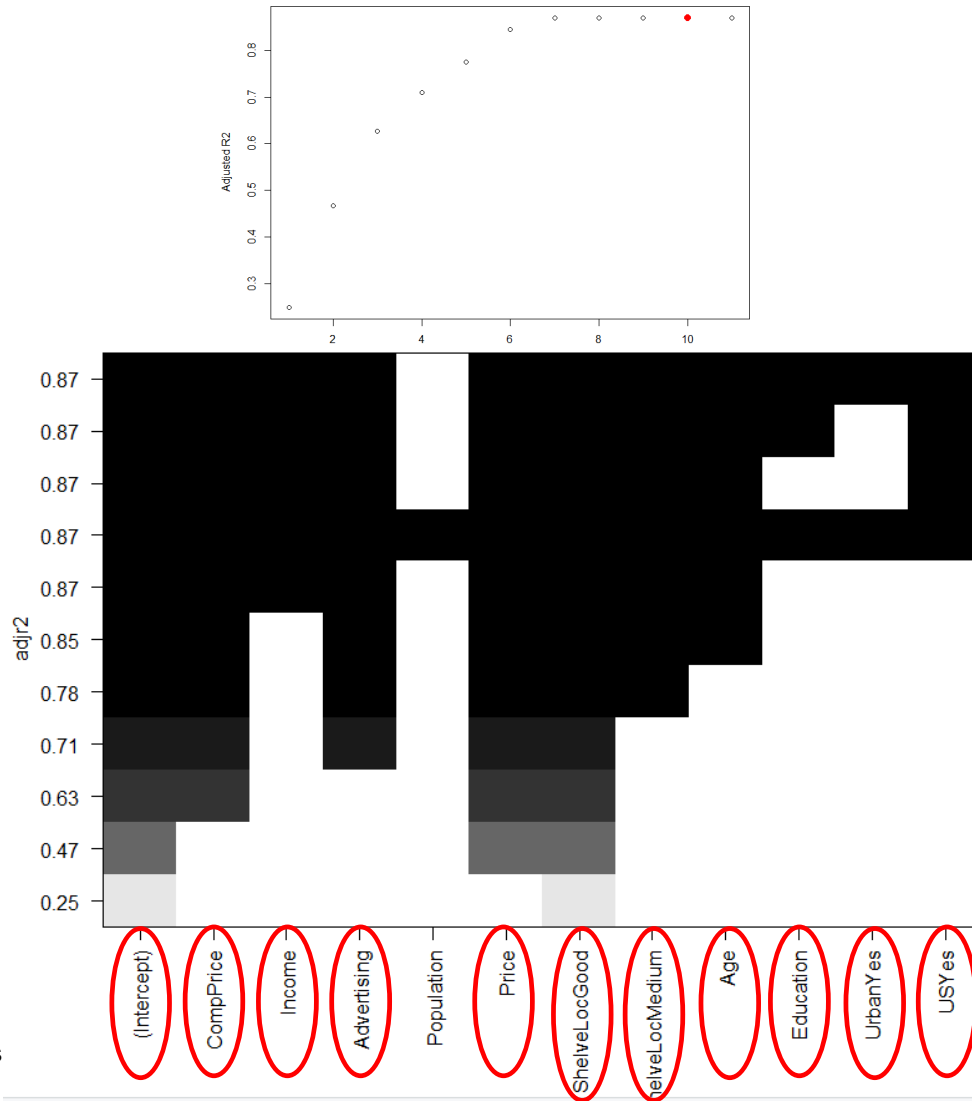
R output for best subset selection

- Output specifies which variables to include for the best model of each size

		CompPrice	Income	Advertising	Population	Price	ShelveLocGood	ShelveLocMedium	Age	Education	UrbanYes	USYes
1	(1)	" "	" "	" "	" "	" "	"*"	" "	" "	" "	" "	" "
2	(1)	" "	" "	" "	" "	"*"	"*"	" "	" "	" "	" "	" "
3	(1)	"*"	" "	" "	" "	"*"	"*"	" "	" "	" "	" "	" "
4	(1)	"*"	" "	"*"	" "	"*"	"*"	" "	" "	" "	" "	" "
5	(1)	"*"	" "	" "	" "	"*"	"*"	"*"	" "	" "	" "	" "
6	(1)	"*"	" "	"*"	" "	"*"	"*"	"*"	"*"	" "	" "	" "
7	(1)	"*"	"*"	"*"	" "	"*"	"*"	"*"	"*"	" "	" "	" "
8	(1)	"*"	"*"	"*"	" "	"*"	"*"	"*"	"*"	" "	" "	"*"
9	(1)	"*"	"*"	"*"	" "	"*"	"*"	"*"	"*"	"*"	" "	"*"
10	(1)	"*"	"*"	"*"	" "	"*"	"*"	"*"	"*"	"*"	"*"	"*"
11	(1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"

- For example, the best 4-variable model for predicting Sales includes:
 - CompPrice
 - Advertising
 - Price
 - The dummy variable ShelveLocGood representing level Good for ShelveLoc

The overall best models with adj R² and BIC



Coefficients in the best model

- These can be obtained by the function "coef", see for example the command below:
`coef(BestSSModel,which.min(BestSSsummary$bic))`

- The output can be used to specify the final model from the best subset selection algorithm using BIC:

$$\widehat{\text{Sales}} = 5.475 + 0.0926 \times \text{CompPrice} + 0.0158 \times \text{Income} + 0.116 \times \text{Advertising} - 0.095 \times \text{Price} + 4.836 \times \text{ShelveLocGood} + 1.952 \times \text{ShelveLocMedium} - 0.046 \times \text{Age}.$$

- Note: the best model according to adjusted R^2 would be a 10-predictor model → it can matter which quality measure we use

Multiple linear regression continued

Reviewing variable selection & Polynomial regression, transformations

Interaction terms – Combined effect of predictors

Practical implications of advertising example analysis

Additive assumption

What happens if we increase the value of predictor 1 by one unit?

- The predicted response before and after the increase:

$$\hat{y}_{\text{before}} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

$$\hat{y}_{\text{after}} = \hat{\beta}_0 + \hat{\beta}_1 (x_1 + 1) + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

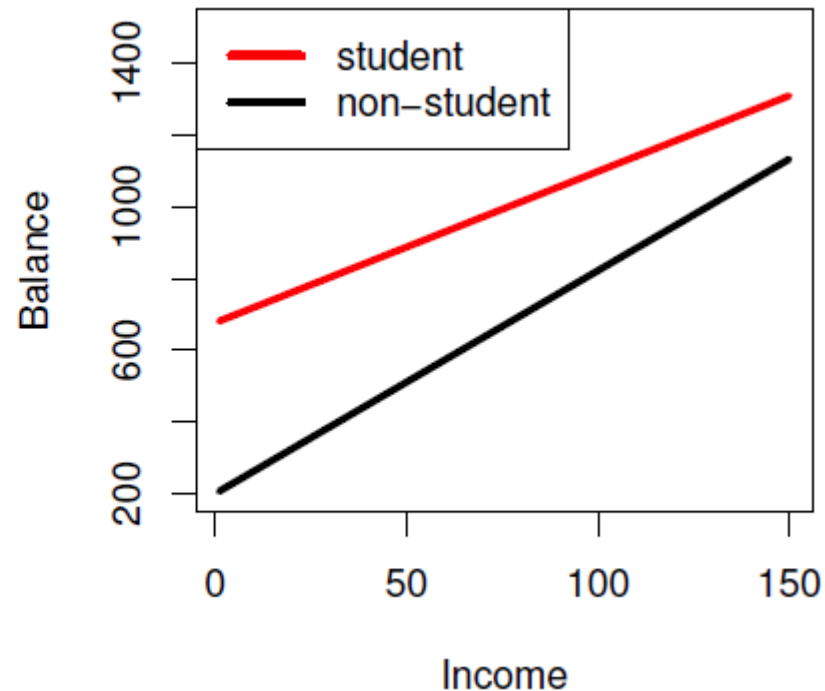
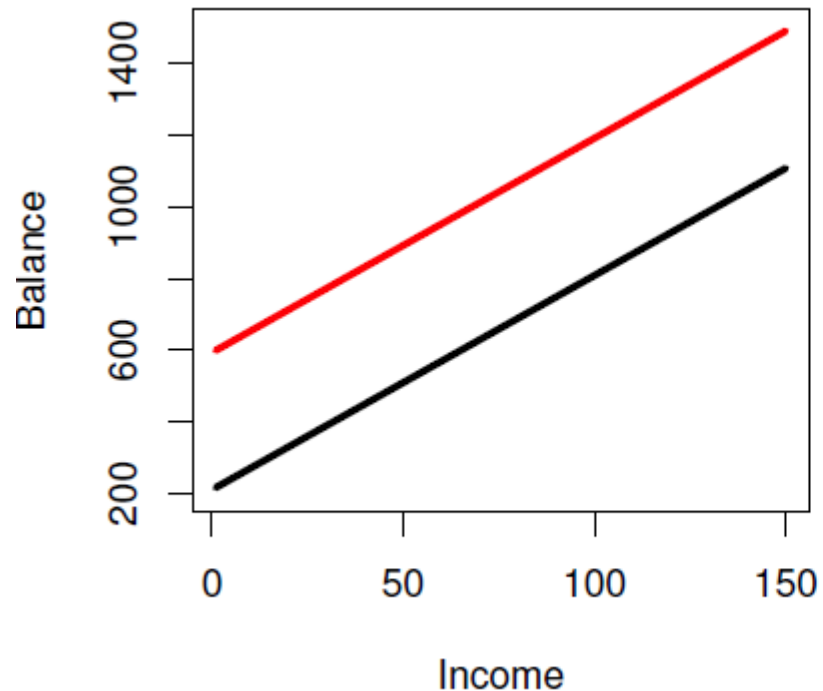
- The difference, $\hat{y}_{\text{after}} - \hat{y}_{\text{before}} = \hat{\beta}_1$, is the average effect of increasing predictor 1 by one unit on the response.
- This does not depend on the value of any other variable!

This is not always realistic

- In real-world situations, changing one parameter often has different effect on the response value in different groups
- Increasing tax of luxury items may affect monthly spendings of rich people differently compared to other people
- Effect of increased number of trainings on muscle mass may depend on gender
- Example (ISL): how does balance (average credit card debt) depend on income (in thousands of dollars)? This may depend on student status (next slide)

Effect of increased income on balance

- On the left: model fit without considering potential interaction between student status and income on their effect on balance. On the right: model fit with interaction term (defined on next slide)



In figure to the right, slope for students is lower than slope for non-students

This suggests smaller effect of income increase on credit card balance for students compared to non-students

Interaction effect / synergy effect

- There is interaction (or synergy) between predictors X_i and X_j if the value of X_j influences the effect of X_i on the response Y

- Include this in the linear regression model by adding a new predictor defined as the product of X_i and X_j :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \beta_{p+1} X_i X_j + \varepsilon$$

Interaction term

- This indeed introduces a dependence: by basic algebra, the coefficient of X_i is $\hat{\beta}_i + x_j \hat{\beta}_{p+1}$, i.e. it depends on the value of X_j

Advertising example

- Can spending money on radio advertisement increase effectiveness of TV advertisements?

- Check R outputs to find out:

```
call:
lm(formula = sales ~ TV + radio)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7977 -0.8752  0.2422  1.1708  2.8328

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92110    0.29449   9.919  <2e-16 ***
TV           0.04575    0.00139  32.909  <2e-16 ***
radio        0.18799    0.00804  23.382  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
call:
lm(formula = sales ~ TV + radio + TV:radio)

Residuals:
    Min       1Q   Median       3Q      Max
-6.3366 -0.4028  0.1831  0.5948  1.5246

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.75022020  0.24787137  27.233  <2e-16 ***
TV           0.01910107  0.00150415  12.699  <2e-16 ***
radio        0.02886034  0.00890527   3.241  0.0014 **
TV:radio      0.00108649  0.00005242  20.727  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

The model has improved,
but how to interpret it?

Interpretation of interaction

- The prediction model with the coefficient estimates:

$$\widehat{\text{sales}} = 6.7502 + 0.0191 \times \text{TV} + 0.0289 \times \text{radio} + 0.0011 \times \text{TV} \times \text{radio}$$


- Effect of \$1000 increase of TV advertisements on sold units:

$$1000 \times (0.0191 + 0.0011 \times \text{radio}) = 19.1 + 1.1 \times \text{radio}$$

- Effect of \$1000 increase of radio advertisements on sold units:

$$1000 \times (0.0289 + 0.0011 \times \text{TV}) = 28.9 + 1.1 \times \text{TV}$$

Effect indeed depends on
value of the other variable



Hierarchical principle

- It can happen that the interaction term is significant, but one or both main effects (whose product defines the interaction term) are not significant
- Even in this case (i.e. even in case of high p-values), include both main terms in the model!
- This is important for interpretation of results

Multiple linear regression continued

Reviewing variable selection & Polynomial regression, transformations

Interaction terms – Combined effect of predictors

Practical implications of advertising example analysis

Questions about advertising example

- The advertising example has been used in various ways throughout the discussion of linear regression
- What could the management learn from these results? What are the practical implications?
- What are the most important questions and how to answer them? See next slides

Relationship between advertising & sales?

- Fit multiple regression model with all variables & check model significance

```
Call:
lm(formula = sales ~ TV + radio + newspaper)

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

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

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

F-test gives strong evidence against the hypothesis of all coefficients being 0
→ yes, there is a relationship between advertising budget and sales.

How strong is the relationship?

- R^2 value and RSE indicate strength of the relationship

```
call:
lm(formula = sales ~ TV + radio + newspaper)

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

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

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

R^2 value indicates that about 90% of the variability in sales can be explained by the predictors.

RSE gives a lack of fit of explained by the predictors and has a value of 1686. This could be compared to the average of the response values.

Which ad types contribute to sales?

- Check variable significance (i.e. t-test p-values) in multiple linear regression model.

```
call:
lm(formula = sales ~ TV + radio + newspaper)

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

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

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

P-values of TV and radio indicate significant contribution for these ad types. However, in the presence of the other two variables, newspaper is not significant

What is the effect size?

- How large is the effect of each medium on sales?
- Estimates are given in previous R output. As for confidence intervals, see R output below:

	2.5 %	97.5 %
(Intercept)	2.32376228	3.55401646
TV	0.04301371	0.04851558
radio	0.17154745	0.20551259
newspaper	-0.01261595	0.01054097

CI for newspaper contains 0, showing again that in the presence of TV and radio budget, the newspaper budget is not a significant predictor for sales

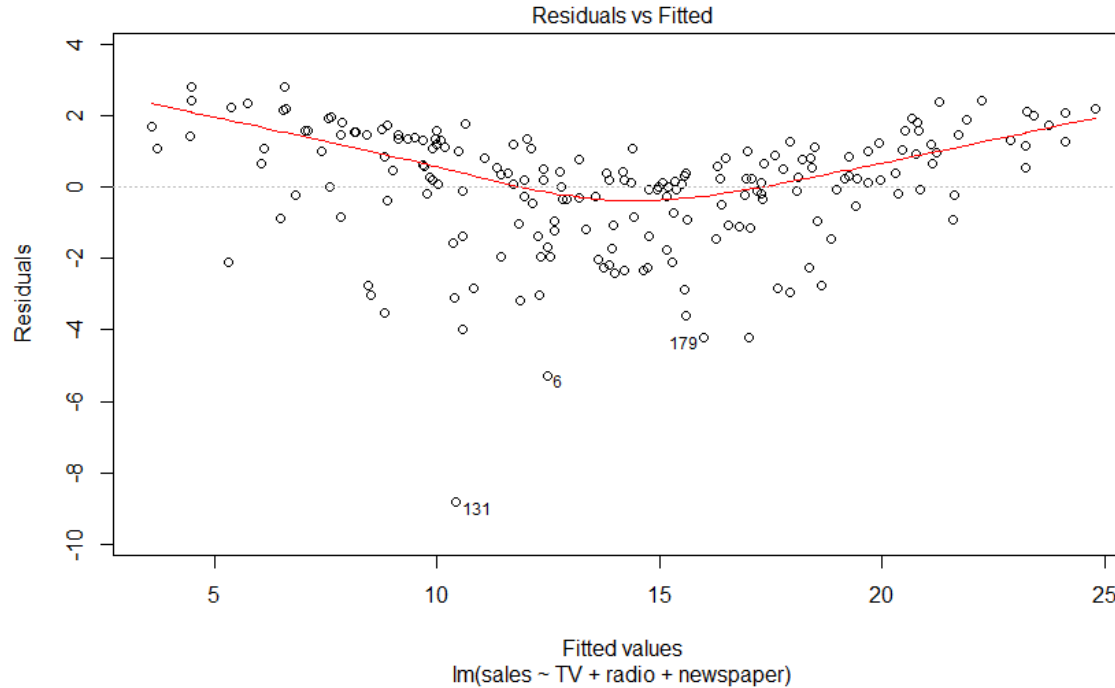
- Conclusion: \$1000 extra on TV advertisements is expected to result in 43-49 extra units sold. \$1000 extra on radio ads is expected to result in 172-206 extra units sold. \$1000 extra on newspaper ads may result in **-13 to 11** extra units sold.

How accurately can we predict sales?

- We predict average sales by the confidence interval
- Specific sales are predicted by the prediction interval (which is wider due to potential individual deviation from average)
- Both are available from R using the “predict” function:
`predict(...,interval="predict")`
`predict(...,interval="confidence")`

Is the relationship linear?

- Residual plot (see later) suggests non-linear effect



- Transformation of sales could be considered to handle this

Is there synergy among advertising media?

- Yes, recently discussed results have indicated synergy between TV and radio budget

The p-value of the interaction term is very small
 R^2 has increased from including interaction

```
Call:
lm(formula = sales ~ TV + radio)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7977 -0.8752  0.2422  1.1708  2.8328

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92110    0.29449   9.919  <2e-16 ***
TV           0.04575    0.00139  32.909  <2e-16 ***
radio       0.18799    0.00804  23.382  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = sales ~ TV + radio + TV:radio)

Residuals:
    Min       1Q   Median       3Q      Max
-6.3366 -0.4028  0.1831  0.5948  1.5246

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.75022020  0.24787137  27.233  <2e-16 ***
TV           0.01910107  0.00150415  12.699  <2e-16 ***
radio       0.02886034  0.00890527   3.241  0.0014 **
TV:radio     0.00108649  0.00005242  20.727  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

Potential problems with linear regression

Non-linear relationships – use transformations

Correlation of error terms – e.g. time series data; experimental design is important

Non-constant variance of error terms (heteroscedasticity)

Outliers – unusual response values

High leverage points – observations with large effect


Collinearity – linear dependence between predictors

Assumptions for linear regression

Recall model equation for linear regression:

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

Random error
with mean 0
and variance σ^2



Using this equation for modelling implicitly assumes:

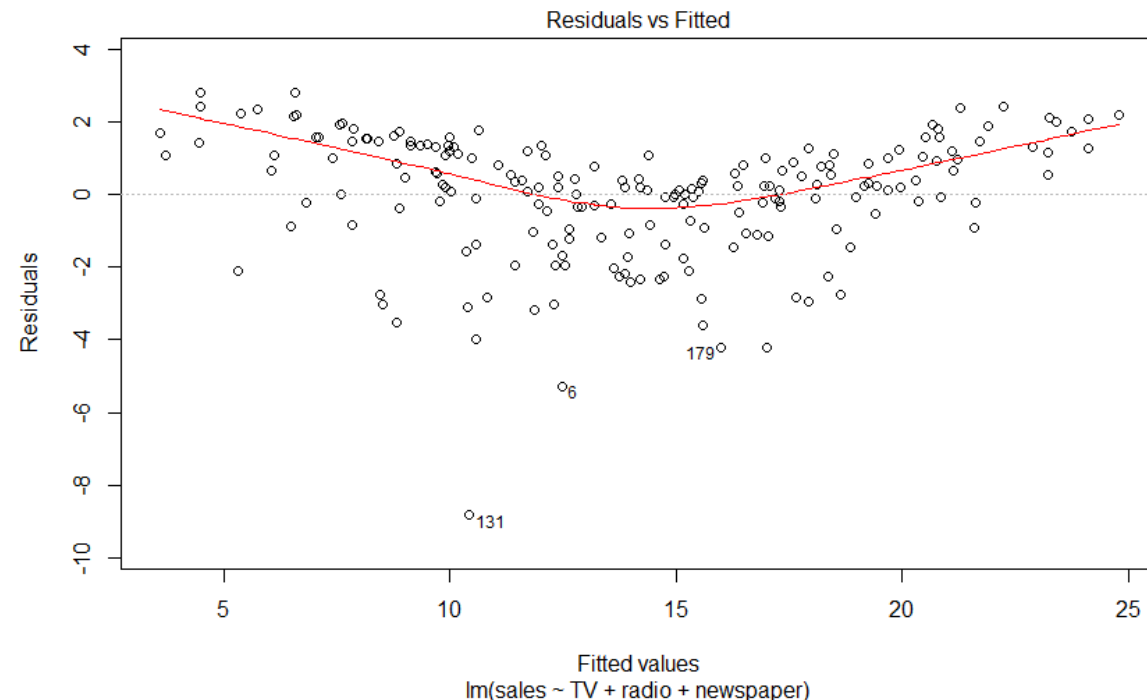
- Linear dependence of response on predictors
- Independence & same variance of random error terms

What can go wrong with these assumptions and otherwise? We will discuss most common issues

Essential tool: the residual plot

- Several problems can be identified by appropriate plots
- For each data point i , we have an observed response y_i , a predicted (/fitted) value \hat{y}_i and their difference is the residual:
$$e_i = y_i - \hat{y}_i$$

- **Residual plot:**
 - predicted values on x-axis
 - residuals on y-axis
- Pattern indicates problem



Potential problems with linear regression

Non-linear relationships – use transformations

Correlation of error terms – e.g. time series data; experimental design is important

Non-constant variance of error terms (heteroscedasticity)

Outliers – unusual response values

High leverage points – observations with large effect

Collinearity – linear dependence between predictors

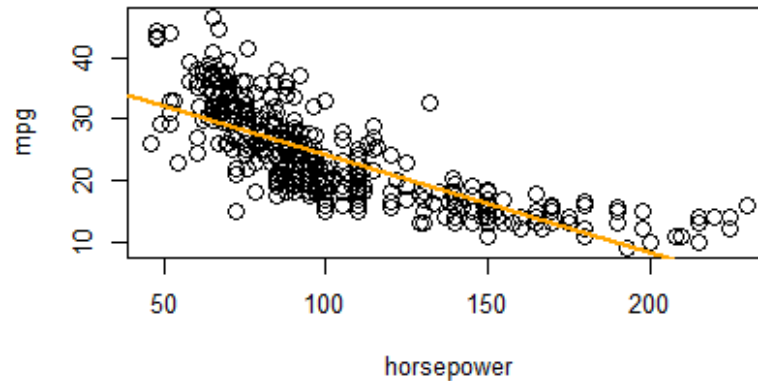
Consequences of non-linear relationship

- Linear dependence of response on predictors is the most basic assumption for linear regression
- True relationship is almost never linear. If it is close enough → the linear model can still be useful
- If true relationship is far from linear → none of coefficient estimates, predictions or standard errors can be trusted

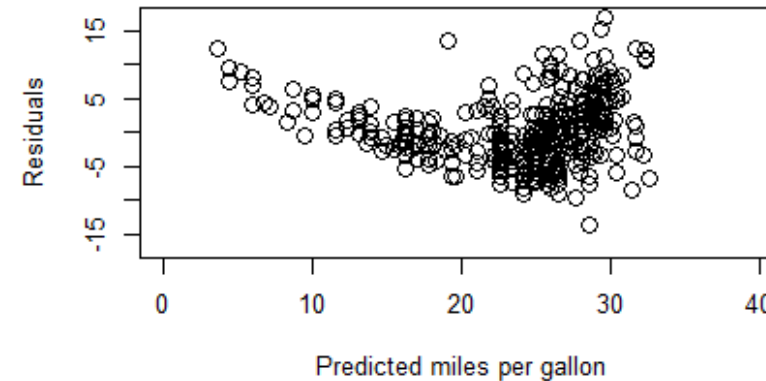
How to spot non-linear dependence?

- Check scatter plots or residual plots; for mpg vs horsepower:

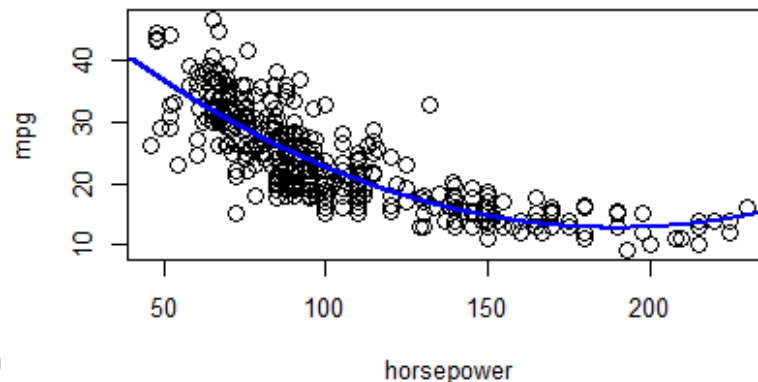
Linear model, scatter plot



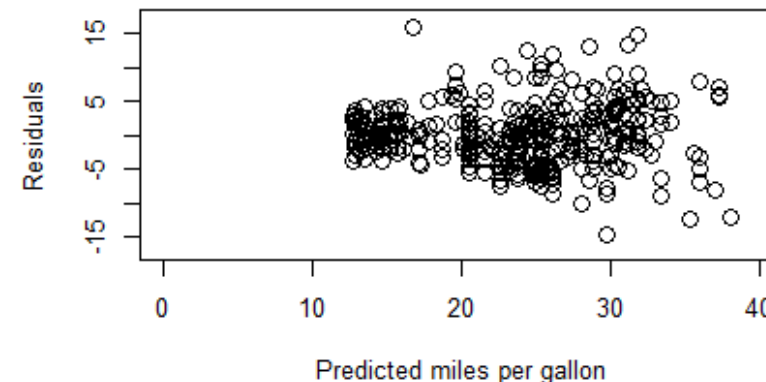
Linear model, residual plot



Quadratic model, scatter plot



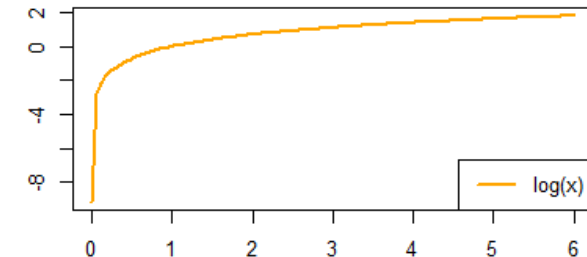
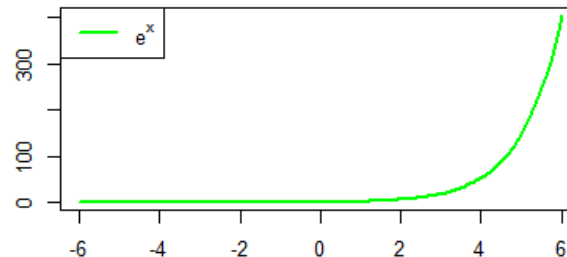
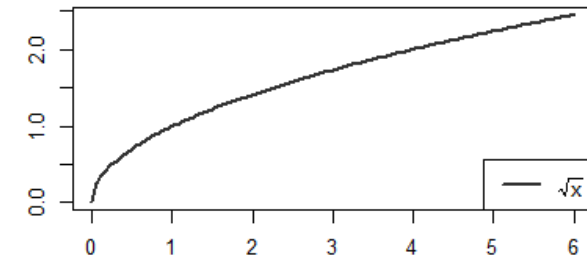
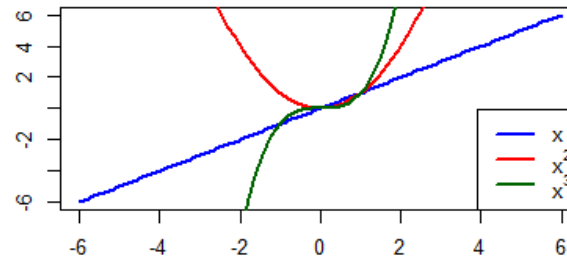
Quadratic model, residual plot



How to address a non-linear relationship?

- Simple solution: include non-linear transformations of the predictors in the linear model, e.g. X^2 , X^3 , $\log X$, \sqrt{X} depending on suspected type of relationship

Does the plot look similar to any of these functions?
If it does, try that transformation!



Potential problems with linear regression

Non-linear relationships – use transformations

Correlation of error terms – e.g. time series data; experimental design is important

Non-constant variance of error terms (heteroscedasticity)

Outliers – unusual response values

High leverage points – observations with large effect

Collinearity – linear dependence between predictors

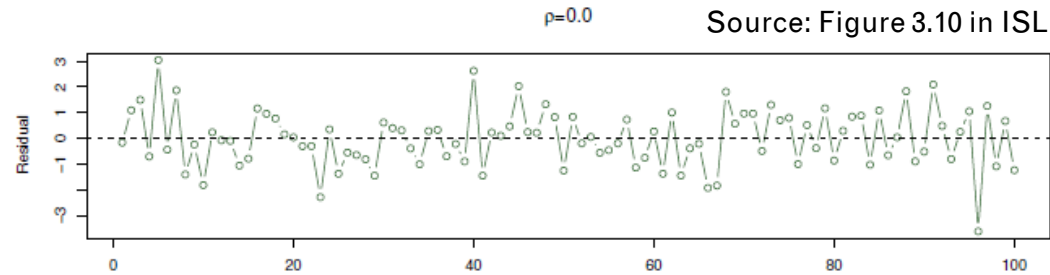
When are error terms correlated?

- Correlated error terms frequently occur in **time series data** (data points are subsequent measurements in time). E.g. temperature values at adjacent time points may have similar values (→ positively correlated errors)
- **Tracking in residuals** (i.e. adjacent residual values having similar values) may be a sign of correlated error terms
- Similar background factors for study subjects can also cause correlation
- Next slide shows plots of residuals with different levels of correlation on left side (Figure 3.10 in ISL), and on the right side a similar plot for the multiple linear regression model predicting sales based on TV and radio advertisements as predictors.

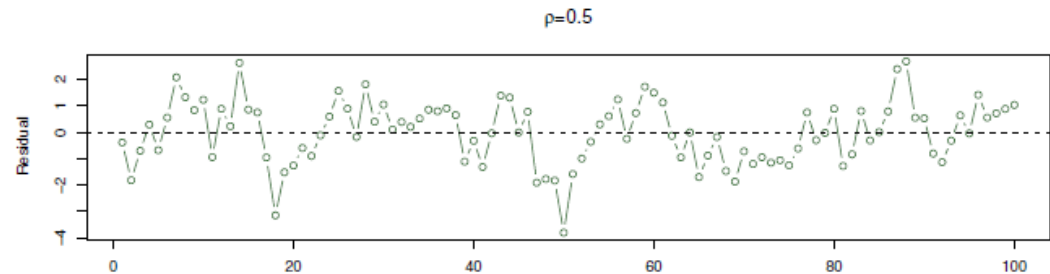
Is there tracking in residuals?

Correlation
level

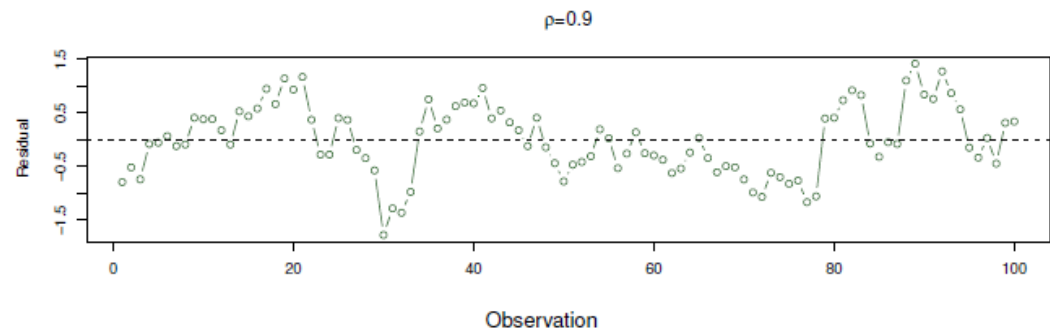
None



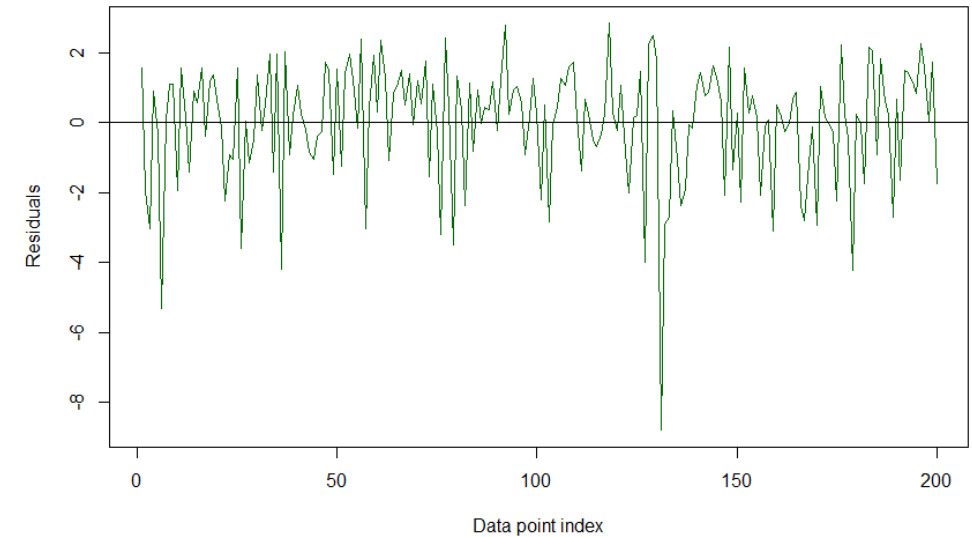
Moderate



Strong

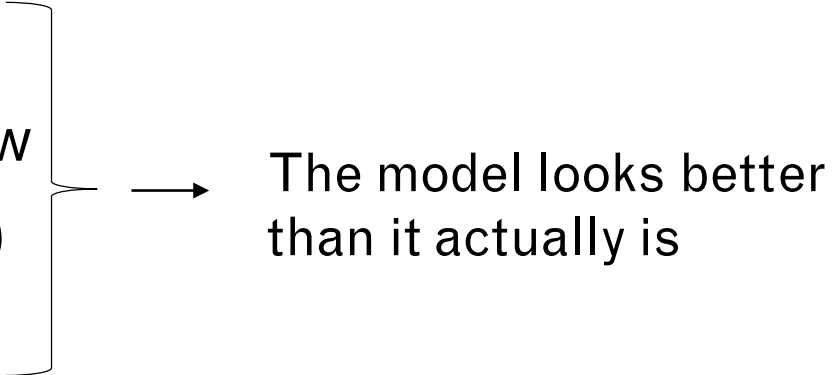


Plot of residuals for Advertising model with $\text{lm}(\text{sales} \sim \text{TV} + \text{radio})$



- Which one is this most similar to?
- Is ordering by index meaningful?

Consequences of correlated error terms

- Error terms are underestimated
 - Confidence & prediction intervals are too narrow (and don't provide the promised confidence level)
 - p-values are lower than they should be
- 
- The model looks better than it actually is
- Methods have been developed* to address correlated error terms
 - Try to avoid/mitigate correlated error terms by good experimental design

* For the details, see e.g. Brown, T. A. (2015). Confirmatory factor analysis for applied research. Guilford Publications

Potential problems with linear regression

Non-linear relationships – use transformations

Correlation of error terms – e.g. time series data; experimental design is important

Non-constant variance of error terms (heteroscedasticity)

Outliers – unusual response values

High leverage points – observations with large effect

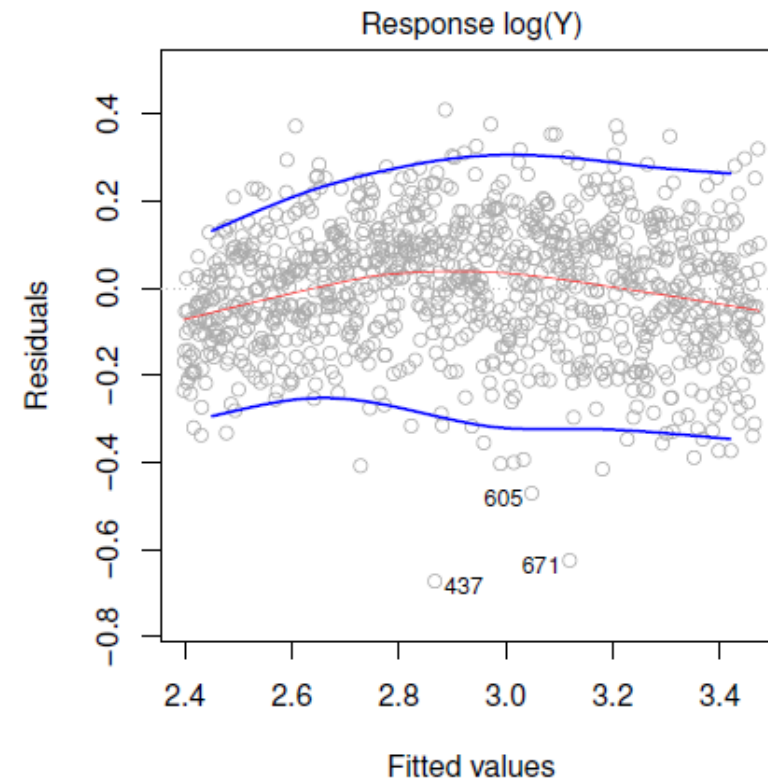
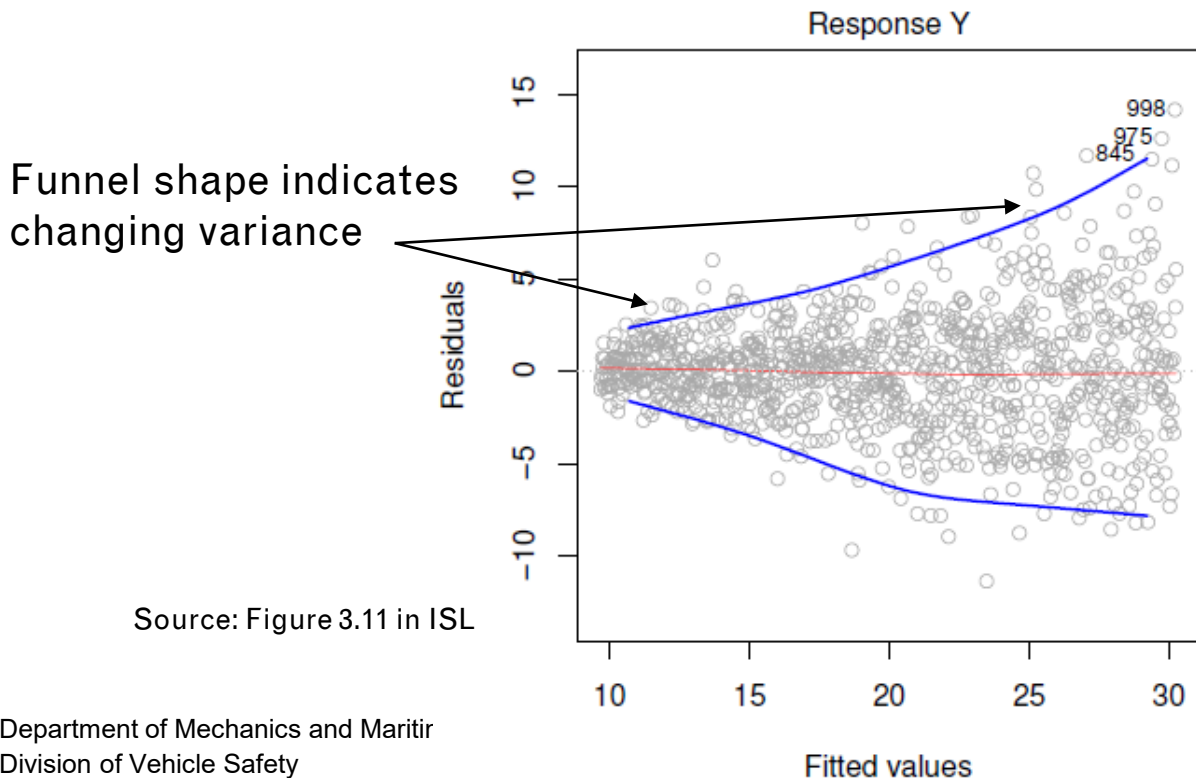
Collinearity – linear dependence between predictors

Why is constant variance of error important?

- RSE was used as an estimate of the standard error σ of ε and was included in the definition of standard errors & confidence intervals of coefficients, etc.
- If we cannot talk about using the same variance σ^2 throughout, we may have problems with all corresponding measures and concepts:
 - Standard errors
 - Confidence intervals
 - Hypothesis testing
- Non-constant variance of error term is called **heteroscedasticity**

How to spot & handle heteroscedasticity?

- Check residual plots for patterns (e.g. funnel shape, bow tie shape).
- Typical situation: variance of error terms increases with the value of the response \rightarrow transforming the response by the log function may help:



No evidence of changing variance in residual plot with transformed response

Potential problems with linear regression

Non-linear relationships – use transformations

Correlation of error terms – e.g. time series data; experimental design is important

Non-constant variance of error terms (heteroscedasticity)

Outliers – unusual response values

High leverage points – observations with large effect

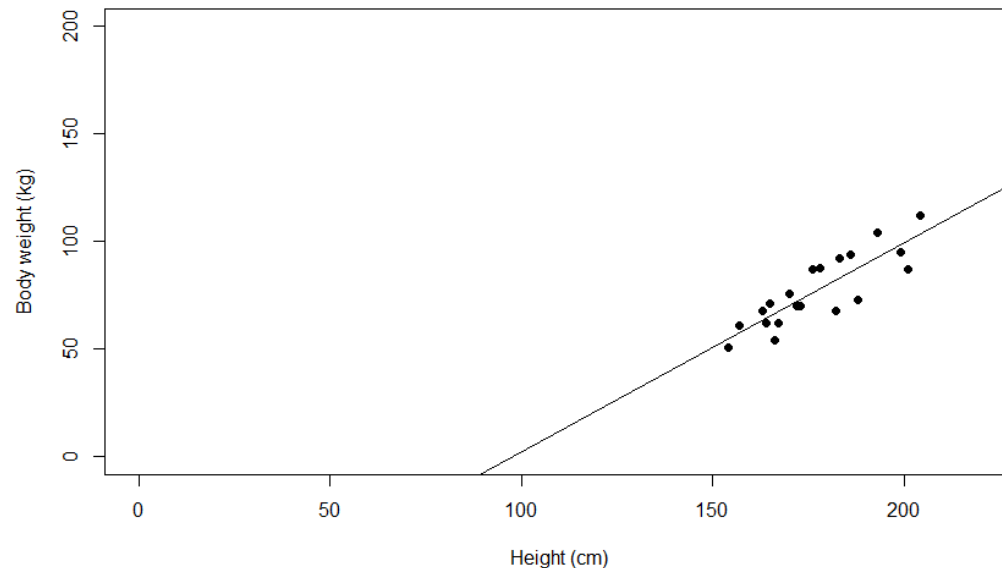
Collinearity – linear dependence between predictors

Outliers: points with unusual response

- Those points are called **outliers** whose response value y_i is far from the predicted value \hat{y}_i
- Possible reasons for presence of outliers include:
 - Error in data collection/coding (→ remove/fix observation value)
 - Missing predictor in the model (→ check for potential predictors)
 - Large random error
- If predictor values of an outlier are in usual predictor ranges
→ effects on coefficient not so large, but error increases a lot.
- If predictor values of an outlier are outside the usual ranges
→ effects on both coefficients and errors may be very large.

Example: body weight vs height

- Class with 20 students, see body weight vs height plot below
- Regression line and R output are shown below



```
Call:
lm(formula = Bodyweight ~ Height)

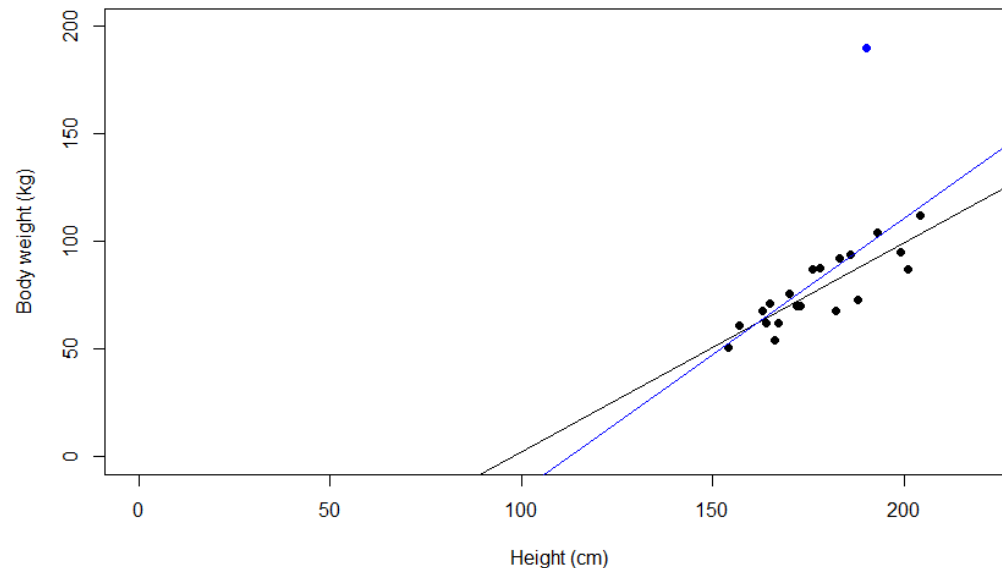
Residuals:
    Min       1Q   Median       3Q      Max
-14.9108  -4.2229   0.4685   8.1552  11.2213

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -95.1240    25.3518  -3.752   0.00146 **
Height        0.9736     0.1427   6.821 0.00000219 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.093 on 18 degrees of freedom
Multiple R-squared:  0.7211,    Adjusted R-squared:  0.7056 
F-statistic: 46.53 on 1 and 18 DF,  p-value: 0.000002192
```

Effect of large weight value

- What changes if a student of 190cm and 190kg joins the class?
- Regression line in blue and new R output are shown below



```
call:
lm(formula = Bodyweight3 ~ Height3)

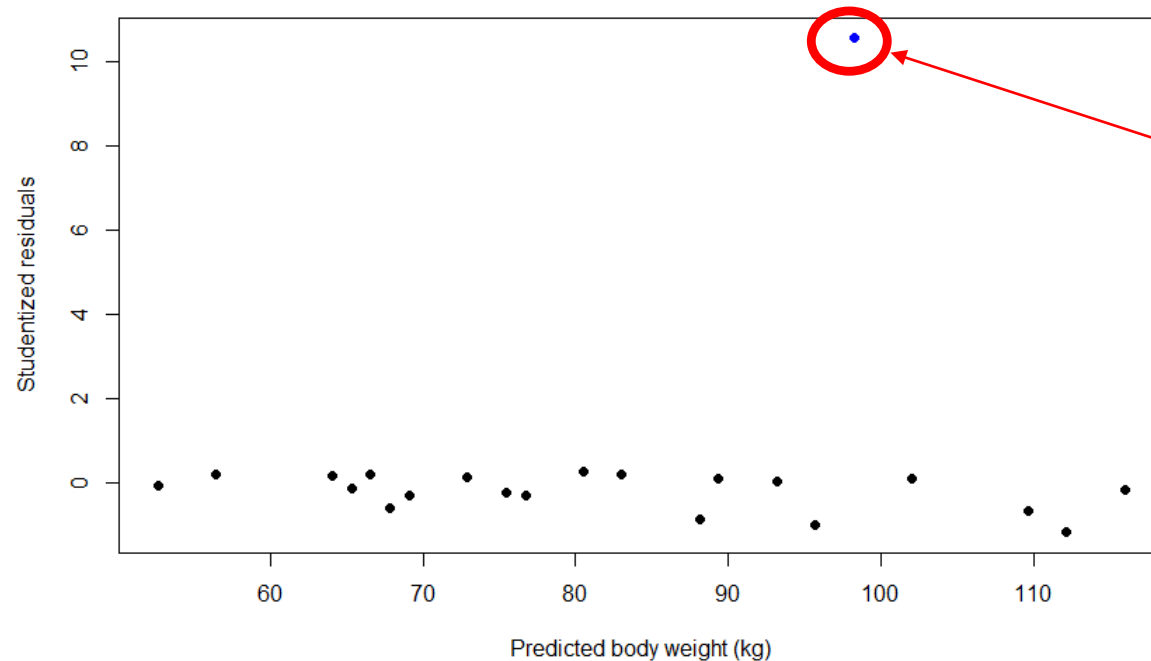
Residuals:
    Min       1Q   Median       3Q      Max
-25.167  -7.111  -1.649   3.954  91.763

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -142.370    65.052  -2.189  0.04132 *
Height3       1.266     0.365   3.470  0.00257 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.71 on 19 degrees of freedom
Multiple R-squared:  0.3879,    Adjusted R-squared:  0.3556
F-statistic: 12.04 on 1 and 19 DF,  p-value: 0.002566
```

How to spot outliers?

- **Studentized residual plot** may give indications. Studentized residuals < -3 or > 3 indicate potential outliers → point of new student in plot below may be an outlier



In the body weight vs height example, the studentized residual of the newly joined student is above 10 → very strong evidence of being an outlier in this model

Potential problems with linear regression

Non-linear relationships – use transformations

Correlation of error terms – e.g. time series data; experimental design is important

Non-constant variance of error terms (heteroscedasticity)

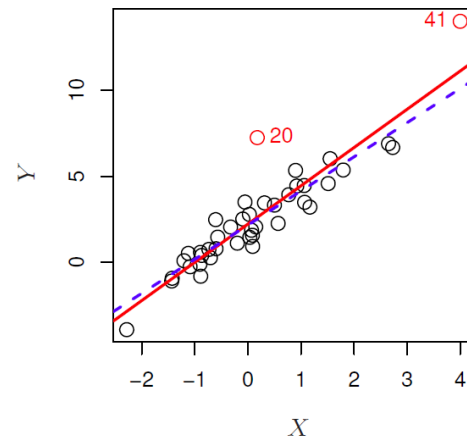
Outliers – unusual response values

High leverage points – observations with large effect

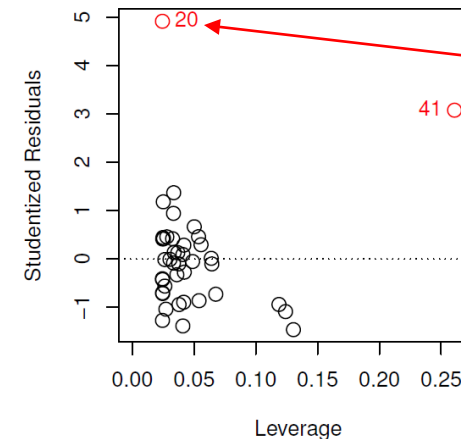
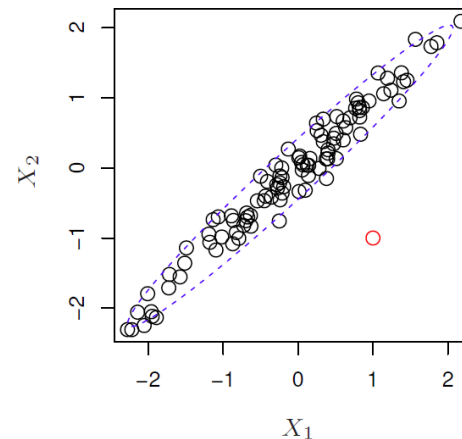
Collinearity – linear dependence between predictors

High leverage points

- Those points are called **high leverage points** whose predictor values are outside the usual predictor ranges or have an unusual combination of usual predictor values (see middle figure below).



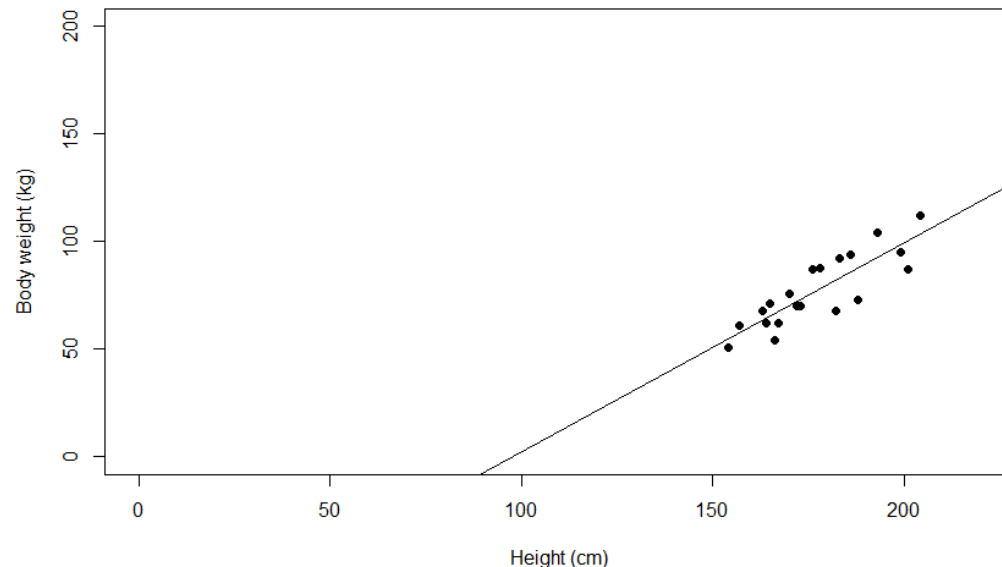
Source: Figure 3.13 in ISL



- The response values at high leverage points have large impact on regression line → An outlier at a high leverage point is a particularly dangerous combination

Example: body weight vs height

- Same class with 20 students as before, plot & output below
- A single extra observation with miscoded height can change everything, see next slide



```
Call:
lm(formula = Bodyweight ~ Height)

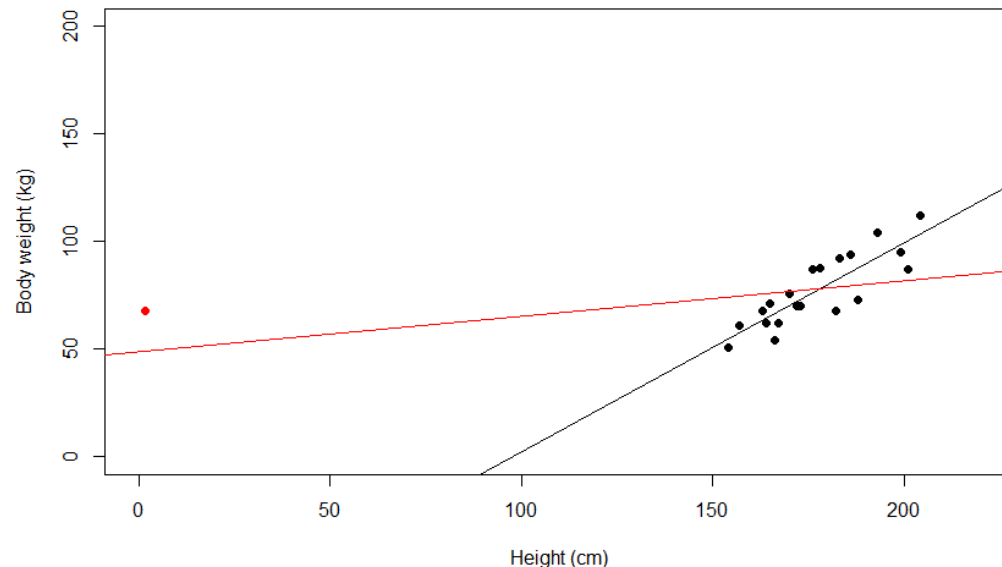
Residuals:
    Min       1Q   Median       3Q      Max
-14.9108  -4.2229   0.4685   8.1552  11.2213

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -95.1240    25.3518  -3.752   0.00146 **
Height        0.9736     0.1427   6.821 0.00000219 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.093 on 18 degrees of freedom
Multiple R-squared:  0.7211,    Adjusted R-squared:  0.7056
F-statistic: 46.53 on 1 and 18 DF,  p-value: 0.000002192
```

Body weight vs height, coding error

- A new student of 171 cm & 68 kg joins the class. However, her height is miscoded as 1.71, because of using m instead of cm.
- The effect of this error on the regression model is dramatic:



```
call:
lm(formula = Bodyweight2 ~ Height2)

Residuals:
    Min       1Q   Median       3Q      Max
-23.386 -11.002  -5.199  12.833  29.371

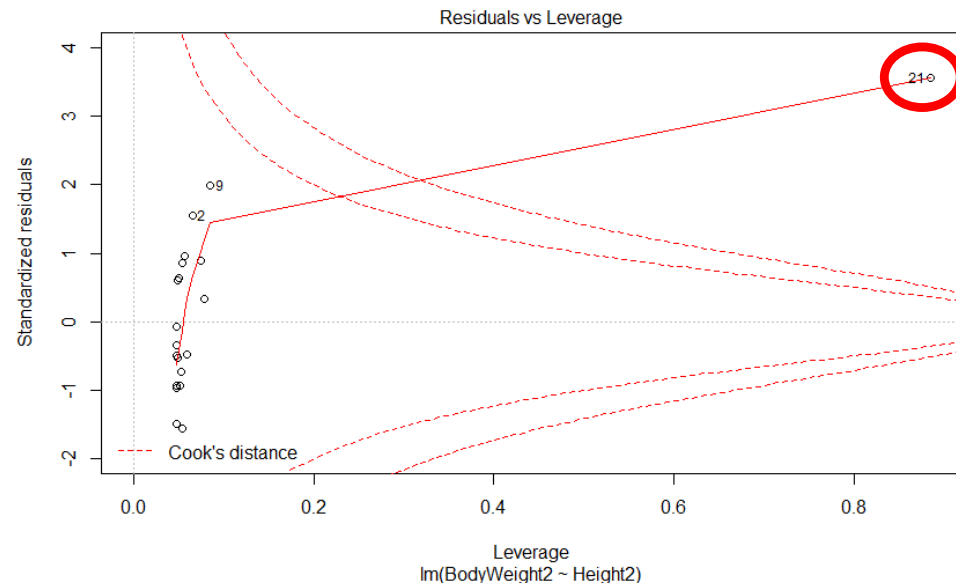
Coefficients:
(Intercept) 48.99688
Height2      0.16486

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.41 on 19 degrees of freedom
Multiple R-squared:  0.1673,    Adjusted R-squared:  0.1234
F-statistic: 3.817 on 1 and 19 DF,  p-value: 0.06563
```


How to spot high leverage points?

- Get leverage statistics from software. This is always between $1/n$ and 1. A value much larger than $(p+1)/n$ for a given point indicates high leverage.
- R has built-in residuals vs leverage graphs when plotting models:



Observation 21 (i.e. the miscoded one) has very high leverage statistics

Potential problems with linear regression

Non-linear relationships – use transformations

Correlation of error terms – e.g. time series data; experimental design is important

Non-constant variance of error terms (heteroscedasticity)

Outliers – unusual response values

High leverage points – observations with large effect

Collinearity – linear dependence between predictors

Collinearity

- **(Multi-)Collinearity** occurs when a predictor has a strong linear relationship with one or more other predictors
- Coefficients are then unreliable – which predictor to attribute the effect to? It results in large standard errors for coefficients
- It can be diagnosed by a **variable inflation factor (VIF)** of 5 or more for any variable. For variable j , this depends on the R^2 value of the linear model predicting X_j by all other predictors:

$$\text{VIF} = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

Collinearity – credit card example in ISL

- Assume: we want to predict credit card balance by a model that includes age and credit card limit, see R output below

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.734e+02  4.383e+01  -3.957 9.01e-05 ***
Limit        1.734e-01  5.026e-03  34.496 < 2e-16 ***
Age         -2.291e+00  6.725e-01  -3.407 0.000723 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 230.5 on 397 degrees of freedom
Multiple R-squared:  0.7498,    Adjusted R-squared:  0.7486
F-statistic:  595 on 2 and 397 DF,  p-value: < 2.2e-16

```


- We now remember that credit score, representing creditworthiness, may be relevant to include in the model:

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -259.51752    55.88219  -4.644 4.66e-06 ***
Limit        0.01901     0.06296   0.302 0.762830
Age         -2.34575     0.66861  -3.508 0.000503 ***
Score        2.31046     0.93953   2.459 0.014352 *

```

Limit is suddenly non-significant
– what happened?



Checking multicollinearity by VIF

- To understand why limit is not significant in the presence of credit score, we try linearly predicting each predictor from the other two:

```
Call:
lm(formula = Limit ~ Age + Score)

Residuals:
    Min       1Q   Median       3Q      Max
-411.14 -127.83   14.16  128.33  468.22

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -529.2872    35.7560  -14.803  <2e-16 ***
Age          -0.2644     0.5328   -0.496    0.62
Score         14.8747     0.0594  250.419  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 182.6 on 397 degrees of freedom
Multiple R-squared:  0.9938,    Adjusted R-squared:  0.9937
F-statistic: 3.168e+04 on 2 and 397 DF,  p-value: < 2.2e-16
```

$$VIF_{\text{Limit}} = \frac{1}{1-0.9938} = 161$$

```
Call:
lm(formula = Age ~ Limit + Score)

Residuals:
    Min       1Q   Median       3Q      Max
-32.276 -13.845   -0.536   14.733   34.773

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.311923    3.349568   15.020  <2e-16 ***
Limit        -0.002345    0.004725   -0.496    0.620
Score         0.046376    0.070486    0.658    0.511
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.2 on 397 degrees of freedom
Multiple R-squared:  0.01126,    Adjusted R-squared:  0.006275
F-statistic: 2.26 on 2 and 397 DF,  p-value: 0.1057
```

$$VIF_{\text{Age}} = \frac{1}{1-0.01126} = 1$$

```
Call:
lm(formula = Score ~ Limit + Age)

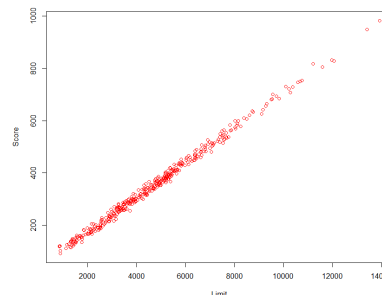
Residuals:
    Min       1Q   Median       3Q      Max
-31.817  -8.617  -1.304    8.607   30.429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.727e+01    2.327e+00   16.019  <2e-16 ***
Limit         6.681e-02    2.668e-04  250.419  <2e-16 ***
Age          2.349e-02    3.570e-02    0.658    0.511
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.24 on 397 degrees of freedom
Multiple R-squared:  0.9938,    Adjusted R-squared:  0.9937
F-statistic: 3.169e+04 on 2 and 397 DF,  p-value: < 2.2e-16
```

$$VIF_{\text{Score}} = \frac{1}{1-0.9938} = 161$$

- The very high VIF values if Limit and Score suggest collinearity. The reason for this is their linear dependence. see the Score vs Limit plot below:



Feedback

Student representatives

- The names of student representatives for this course & contact information will be published on the [course homepage](#)
- Please inform the student representatives about your impression of the course so far, e.g.:
 - Overall opinion
 - Potential issues
 - Improvement suggestions
- This is important even if you give feedback via www.menti.com

Feedback quiz - optional

Feedback is essential to me so that I can improve the lectures during the course. All comments about today's class, assignment 1 or the course in general are welcome!

If you are willing to give feedback, please follow these steps:

1. Go to www.menti.com
2. Enter the code 30 81 24
3. Answer the questions or enter other comments related to the course