

Statistical modeling in logistics

MMS075

Lecture 3: Multiple linear regression –
qualitative variables, variable selection,
prediction intervals, non-linear relationships

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

Assignment 1 information

- Assignment 1 has been published in Canvas, deadline: Monday, Feb 10, 23:59. Make sure to submit a solution before the deadline, even if it is imperfect.
- You can use any software you want to get the solution, but I can help best with R if needed (and software like Excel may not necessarily be able to address later assignments)
- The course is given in English → submissions in English are encouraged, but submissions in Swedish will also be accepted

Exam information

- The exam will not contain very long computations. A formula sheet will be provided, so you don't need to learn the formulas by heart, but you need to understand the concepts.
- At the last lecture (on March 3), we will review what to focus on
- This is a course in statistical modelling, not R → the exam will not contain questions about specific commands in R.
- The course is given in English → at the exam, answers in English are encouraged, but answers in Swedish are also accepted.

Outline – combined lecture & exercise class

- Multiple linear regression (cont):
 - Reviewing hypothesis testing of relationship with response
 - Qualitative predictors
 - Variable selection
 - Uncertainty in prediction
 - Non-linear relationships
- Feedback

Recommended resources

Reading in [ISL](#):

- Sections 3.2.2-3.3.2 for the theory
- Sections 2.3.4-2.3.5 and 3.6.4-3.6.6 for R codes
- Enrollment in the online course [Statistical Learning](#) is unfortunately not available anymore. The videos from the course are available at [this link](#).
The most relevant ones for today are:
 - [Model Selection and Qualitative Predictors](#) (14:51)
 - [Interactions and Nonlinearity](#) (14:16)

Multiple linear regression continued

Reviewing hypothesis testing of relationship with response

Qualitative predictors – Predictors taking categories as values

Variable selection – Best subset, forward, backward, mixed methods

Uncertainty in prediction – Confidence and prediction intervals

Non-linear relationships – Polynomial regression

p-value and corresponding decisions

- For a specific null hypothesis and test statistic, the corresponding p-value is the probability of having **at least as extreme values of the test statistic as the one observed, assuming that the null hypothesis is true.**
- If the p-value is very small*, that's strong evidence against the null hypothesis → decision: **reject the null hypothesis**
- If the p-value is not very small, that's not enough evidence against the null hypothesis → decision: **fail to reject the null hypothesis**

*Compared to a significance level α which is typically chosen as 0.1, 0.05 (this is most common) or 0.01

Model significance vs variable significance

- A linear regression **model is significant** if all variables together as a group have a relationship with response; use F-test:

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

$$H_a : \text{at least one } \beta_j \text{ is not zero}$$

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

- In a linear regression model, **a specific predictor X_j is significant** if in the presence of all other predictors, it has a relationship with the response; use t-test to test this:

$$H_0 : \beta_j = 0$$

$$H_a : \beta_j \neq 0$$

$$t = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$$

Exercise 1 example (simple linear regression)

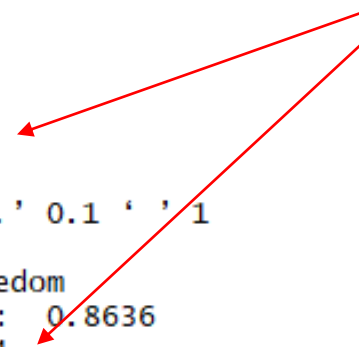
- Output from R: p-value of $0.0465 < 0.05 \rightarrow$ decision: reject H_0 , conclude: there is a relationship between number of employees on the project and completion time
- **Important: p-value < 0.05 if and only if 95% confidence interval for the slope does not contain 0**

```
Call:
lm(formula = CompTime ~ Employees)

Residuals:
    1     2     3     4 
-10 -10  10  10 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.0000    13.2288   7.559  0.0171 *
Employees    -1.0000     0.2236  -4.472  0.0465 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.14 on 2 degrees of freedom
Multiple R-squared:  0.9091,    Adjusted R-squared:  0.8636 
F-statistic: 20 on 1 and 2 DF,  p-value: 0.04654
```



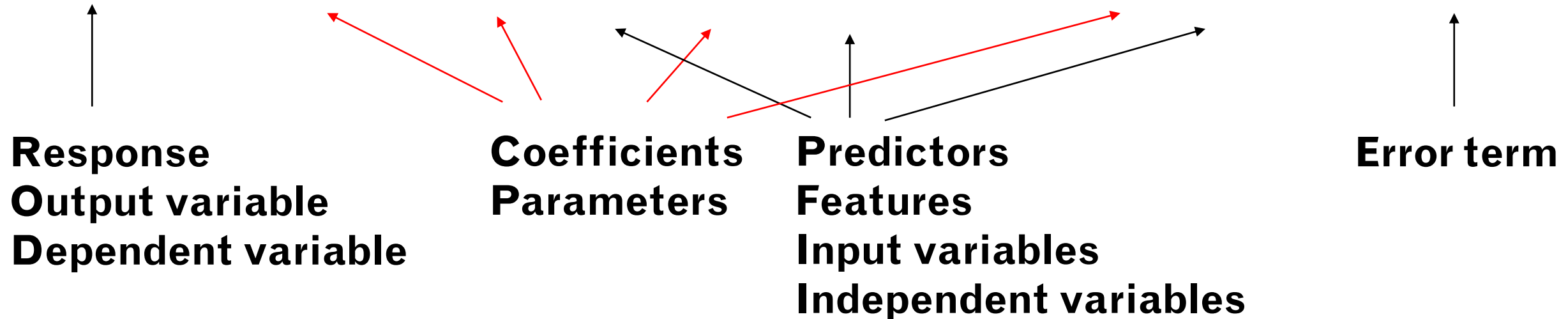
Note: the p-values of the t-test and the F-test are exactly the same!

**For simple linear regression:
model significance =
significance of the (single)
predictor**

Multiple linear regression: model definition

- Outcome linearly depends on $p \geq 2$ predictors:

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



- All the above terms are used regularly in different contexts

Here: variable significance \neq model sign.

- See computer output for the advertising example:

```
call:
lm(formula = AdData$sales ~ AdData$TV + AdData$radio + AdData$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 ***
AdData\$TV	0.045765	0.001395	32.809	<2e-16 ***
AdData\$radio	0.188530	0.008611	21.893	<2e-16 ***
AdData\$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

Not all p-values for t-test are the same as the p-value for the F-test!

For multiple linear regression:

- **Model significance is NOT the same as significance of predictors**
- **Model significance is NOT the same as having individual significance of ≥ 1 predictor**

Multiple linear regression continued

Reviewing hypothesis testing of relationship with response

Qualitative predictors – Predictors taking categories as values

Variable selection – Best subset, forward, backward, mixed methods

Uncertainty in prediction – Confidence and prediction intervals

Non-linear relationships – Polynomial regression

Using qualitative predictors

- In all examples so far, both the predictors and the response were numerical
- Consider now the situation when some predictors are categorical (i.e. take categories as possible values, not numbers)
- Linear models can easily include categorical **predictors**. (For categorical response, other models will be used, see later lectures)
- Review definition of quantitative-qualitative data in SJO915 slides:

- Quantitative (or numerical) data :

Example:

- The weights of new born children: 3050g, 4120g, 2990g, 3650g, ...
- The ages of respondents: 23, 58, 19, 44,...

Quantitative data can further be described by distinguishing between **discrete** and **continuous** types!

- Qualitative (or *categorical* or *attribute*) data :

consists of names or labels (representing categories)

Example:

- The genders (*male/female*) of professional athletes
- Survey responses: *yes, no, undecided*
- Course grades: *A, B, C, D, or F*

Binary predictors in regression models

- To include binary (two-level) predictors taking only two possible values, e.g. Yes/No – define a new variable, called **dummy variable**:

$$X_{\text{dummy}} = \begin{cases} 1 & \text{if the value of predictor is 'Yes'} \\ 0 & \text{if the value of predictor is 'No'} \end{cases}$$

- The coefficient of X_{dummy} gives the effect of changing the value of the predictor from 'No' to 'Yes' (see examples on next slides)
- One could also define 1 for 'No' and 0 for 'Yes' – the coefficient would then show the effect of changing the variable from 'Yes' to 'No'

Example: cost of transporting goods

- Model how the cost of transporting goods from Sweden depends on distance by linear regression (which may be an oversimplification). This gives the following estimated costs:

$$\widehat{\text{cost}} = \hat{\beta}_0 + \hat{\beta}_{\text{dist}} \times \text{distance}$$

- The cost may also depend on whether or not the destination is in an EU country → define a dummy variable:

$$X_{\text{EU}} = \begin{cases} 1 & \text{if the destination is in the EU} \\ 0 & \text{if the destination is outside the EU} \end{cases}$$

- The updated model is then:

$$\widehat{\text{cost}} = \hat{\beta}_0 + \hat{\beta}_{\text{dist}} \times \text{distance} + \hat{\beta}_{\text{EU}} \times X_{\text{EU}}$$

How to interpret this model?

- Recall the updated model:

$$\widehat{\text{cost}} = \hat{\beta}_0 + \hat{\beta}_{\text{dist}} \times \text{distance} + \hat{\beta}_{\text{EU}} \times X_{\text{EU}}$$

- The estimated cost for a destination within EU at distance = 2000 km:

$$\widehat{\text{cost}} = \hat{\beta}_0 + \hat{\beta}_{\text{dist}} \times 2000 + \hat{\beta}_{\text{EU}}, \text{ because here } X_{\text{EU}} = 1$$

- The estimated cost for a destination outside the EU at distance = 2000 km:

$$\widehat{\text{cost}} = \hat{\beta}_0 + \hat{\beta}_{\text{dist}} \times 2000, \text{ because here } X_{\text{EU}} = 0$$

- The difference between the cost estimates is exactly $\hat{\beta}_{\text{EU}}$

Qualitative predictors with >2 levels

- Consider a variable with more than 2 values:
 - E.g. transport destination: EU, US, China, Other → 4 levels here
 - Ethnicity (ISL example): African American, Asian, Caucasian → 3 levels
 - Transport mode: Rail, Water, Air, Road → 4 levels
 - IKEA sofa names: Ektorp, Nockeby, Kivik, Klippan, Karlstad, ... → >10 levels
- These categories are not ordered → it would be **WRONG** to use a single dummy variable with values 0, 1, 2, ...

$$X_{\text{dummy}} = \begin{cases} 3 & \text{if the value of predictor is Other} \\ 2 & \text{if the value of predictor is China} \\ 1 & \text{if the value of predictor is US} \\ 0 & \text{if the value of predictor is EU} \end{cases}$$

Define several binary dummy variables!

- For qualitative predictors with L levels, $L-1$ dummy variables need to be defined
- Why $L-1$? Because they indicate the difference compared to a baseline level which won't need a dummy variable
- The baseline level is freely chosen. All direct comparisons will be made against this level → choose it in the most relevant way

Transporting goods with 4 destinations

- Example: for destination, we can choose EU as baseline and define 3 dummy variables (with the L=4 levels on slide 19):

$$X_{\text{US}} = \begin{cases} 1 & \text{if destination is in the United States} \\ 0 & \text{if destination is not in the United States} \end{cases}$$

$$X_{\text{China}} = \begin{cases} 1 & \text{if destination is in China} \\ 0 & \text{if destination is not in China} \end{cases}$$

$$X_{\text{Other}} = \begin{cases} 1 & \text{if destination is in the Other category} \\ 0 & \text{if destination is not in the Other category} \end{cases}$$

Interpretation of coefficients

- The model including the three dummy variables:

$$\widehat{\text{cost}} = \hat{\beta}_0 + \hat{\beta}_{\text{dist}} \times \text{distance} + \hat{\beta}_{\text{US}} \times X_{\text{US}} + \hat{\beta}_{\text{China}} \times X_{\text{China}} + \hat{\beta}_{\text{Other}} \times X_{\text{Other}}$$

- The estimated cost for a destination at distance=7500 km in the US (above) and in China (below):

$$\widehat{\text{cost}} = \hat{\beta}_0 + \hat{\beta}_{\text{dist}} \times 7500 + \hat{\beta}_{\text{US}}$$

$$\widehat{\text{cost}} = \hat{\beta}_0 + \hat{\beta}_{\text{dist}} \times 7500 + \hat{\beta}_{\text{China}}$$

Comparisons beyond baseline

- EU was baseline in the example – can the model make a comparison between US as destination compared to China as destination?

Cost of transporting to US = Cost of transporting to EU + $\hat{\beta}_{US}$

Cost of transporting to China = Cost of transporting to EU + $\hat{\beta}_{China}$

→ Cost of transporting to China = Cost of transporting to US - $\hat{\beta}_{US}$ + $\hat{\beta}_{China}$

(Note: there is no distance that would make the first two equations meaningful, which indicates problems with this model)

Multiple linear regression continued

Reviewing hypothesis testing of relationship with response

Qualitative predictors – Predictors taking categories as values

Variable selection – Best subset, forward, backward, mixed methods

Uncertainty in prediction – Confidence and prediction intervals

Non-linear relationships – Polynomial regression

Which variables to include in a model?

Typical situation:

- We want to understand or predict a certain variable Y (outcome)
- We have observations available about Y together with values of many other, both relevant and irrelevant parameters

Which ones of the other parameters can/should be used as predictors in a model that explains/predicts Y ?

Some classical approaches are discussed in the next slides

Why not include all variables?

- The model with all variables gives smallest RSS and is the linear model that provides best fit with existing observations (largest R^2) – why should we need anything else?
- Possible problem: the resulting model fits existing observations too well and may not perform well on new data (**overfitting**)
- Overfitting will be discussed extensively in later lectures

Example: advertising data

- Recall the output for the model including all three predictors:

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

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

```
Coefficients:
(Intercept)      Estimate Std. Error t value Pr(>|t|)
AdData$TV         0.045765  0.001395  32.809  <2e-16 ***
AdData$radio      0.188530  0.008611  21.893  <2e-16 ***
AdData$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
```

t-test fails to provide evidence of a relationship between sales and newspaper in the presence of TV and radio
→ the newspaper term should probably not be included in the model as it may just cause overfitting

- Is it convincing that the model should be exactly like this:

$$\text{sales} \approx 2.939 + 0.046 \times \text{TV} + 0.189 \times \text{radio} - 0.001 \times \text{newspaper} ?$$

Best subset selection

- How to measure which model **explains the response best compared to its size/complexity?**

- 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

Problem: too many subsets!

- Number of subsets grows exponentially in the number of variables:

p	1	2	3	4	5	10	20	30	40
# subsets	2	4	8	16	32	1024	1048576	1073741824	1099511627776

- Best subset selection may be difficult/impossible to do if there are many variables
- Instead: consider a method based on p-values of t-tests of variable significance! (next slides)

Backward selection

- One sign of possible overfitting in the variable with all models was the presence of non-significant variables → if there are non-significant* ones, drop the one with highest p-value!
- Investigate the resulting model with $p-1$ variables.
 - If all variables are significant → STOP and use this model
 - If there are non-significant ones → drop the one with highest p-value
- Keep dropping the variable with the highest p-value until you get to a model where each variable is significant

*Greater than or equal to a significance level which is typically chosen as 0.1 or 0.05

Backward selection on advertising data

Step 1: output of model with all three predictors:

```

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

```

newspaper term is non-significant
→ drop it from the model

Step 2: output of model with TV and radio as predictors (i.e. after having dropped newspaper):

```

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

```

Each variable is very significant
→ STOP and use this model

Conclusion: use $\widehat{\text{sales}} = 2.921 + 0.046 \times \text{TV} + 0.188 \times \text{radio}$

Forward selection

Step 0: Instead of dropping the worst variables after starting with many, one can also start from 0 variables (only intercept – this is called the **null model**) and keep adding the best variables:

Step 1: Check R^2 values resulting from adding the remaining variables to the current model one-by-one

Step 2:

If the variable whose addition provides the highest R^2 value has low p-value → add it to the model and go back to step 1

If this variable has a high p-value → do not add it to the model and STOP

Forward selection on advertising data

Step 0 – R output from null model:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.0225     0.3689   38.01  <2e-16 ***
---

```

Step 1 – R outputs from the three possible models after null:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.032594    0.457843   15.36  <2e-16 ***
TV           0.047537    0.002691   17.67  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared: 0.6119, Adjusted R-squared: 0.6099

```

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.31164    0.56290   16.542  <2e-16 ***
radio        0.20250    0.02041    9.921  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.275 on 198 degrees of freedom
Multiple R-squared: 0.332, Adjusted R-squared: 0.3287

```

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.35141    0.62142   19.88  < 2e-16 ***
newspaper    0.05469    0.01658    3.30  0.00115 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.092 on 198 degrees of freedom
Multiple R-squared: 0.05212, Adjusted R-squared: 0.04733

```

This is largest; step 2: the p-value of TV is very low → we add it to the model and re-do step 1 (see next slide)

Forward selection for ad data (cont.)

Step 1 again – consider additions to the model that already contains TV by looking at their R outputs:

```

              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

```

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.774948    0.525338  10.993  < 2e-16 ***
TV            0.046901    0.002581  18.173  < 2e-16 ***
newspaper    0.044219    0.010174   4.346  2.22e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.121 on 197 degrees of freedom
Multiple R-squared:  0.6458,    Adjusted R-squared:  0.6422

```

This is largest; step 2: the p-value of radio is very low → we add it to the model and re-do step 1

Step 1 again – consider additions to TV+radio model. Only one is possible:

```

              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

```

This is largest (and it is the only option left anyway).

Step 2: the p-value of newspaper is very high
→ we DO NOT add it to the model and STOP; we use the previous model (without adding newspaper) as final

Conclusion: use $\widehat{\text{sales}} = 2.921 + 0.046 \times \text{TV} + 0.188 \times \text{radio}$

Potential non-significant variables

- Forward selection ensures that each newly added variable will be significant (only those with low p-value are added)
 - However, variables added in previous steps might become non-significant as a result from the addition of a new variable!
- Final model from forward selection can possibly contain some non-significant variables
- This is addressed by mixed selection, which is forward selection with some backward steps included (see next slide)

Mixed selection

Step 0: Start with forward selection (from null model, and in each step, consider the variable whose addition gives best fit)

Step 1: After each variable addition, check the p-value for each variable in the model & remove variables with p-value above a threshold

Step 2: Continue as in forward selection until a new variable is added

→ In final model, all variables have low p-value & all variables outside the model would have large p-value if added to the model

Multiple linear regression continued

Reviewing hypothesis testing of relationship with response
Qualitative predictors – Predictors taking categories as values
Variable selection – Best subset, forward, backward, mixed methods
Uncertainty in prediction – Confidence and prediction intervals
Non-linear relationships – Polynomial regression

Prediction point estimate

- Prediction as a point estimate (best guess) follows from the model equation. For completion time vs employee number example:

$$\widehat{\text{Time}} = 100 - 1 \times \text{Employees}$$

- For example, for 70 employees, this equation gives a point estimate of 30 days for completion time
- What is the margin of error? Can we provide an interval that is likely to contain the true value?

Predict average or specific outcome?

What is an interval that is likely to provide the true average completion time (that we would observe as an average over many years) if Employees = 70?

This is called a **confidence interval**; R output gives [-6;66]

What is an interval that is likely to provide an interval for the completion time for a specific occasion (e.g. next year, and not the long-time average)?

This is called a **prediction interval**; R output gives [-41;101]

Confidence interval vs prediction interval

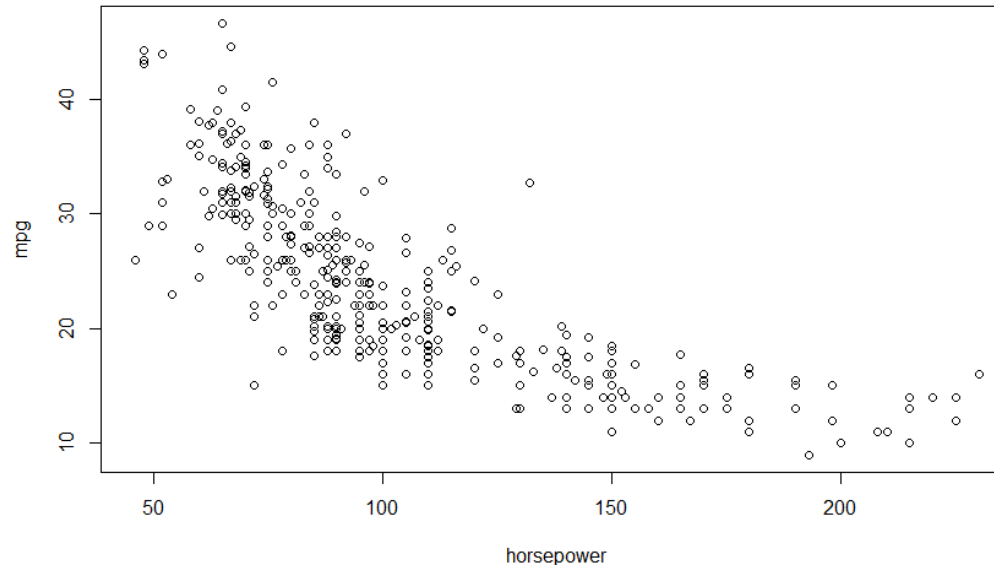
- Both intervals are centered at the prediction point estimate
- Prediction interval is always wider – while the long-term average may be contained in the confidence interval, there can be deviations from this average on specific occasions
- In the completion time vs employee number example, the intervals are very wide because the regression model was based on very few data points

Multiple linear regression continued

Reviewing hypothesis testing of relationship with response
Qualitative predictors – Predictors taking categories as values
Variable selection – Best subset, forward, backward, mixed methods
Uncertainty in prediction – Confidence and prediction intervals
Non-linear relationships – Polynomial regression

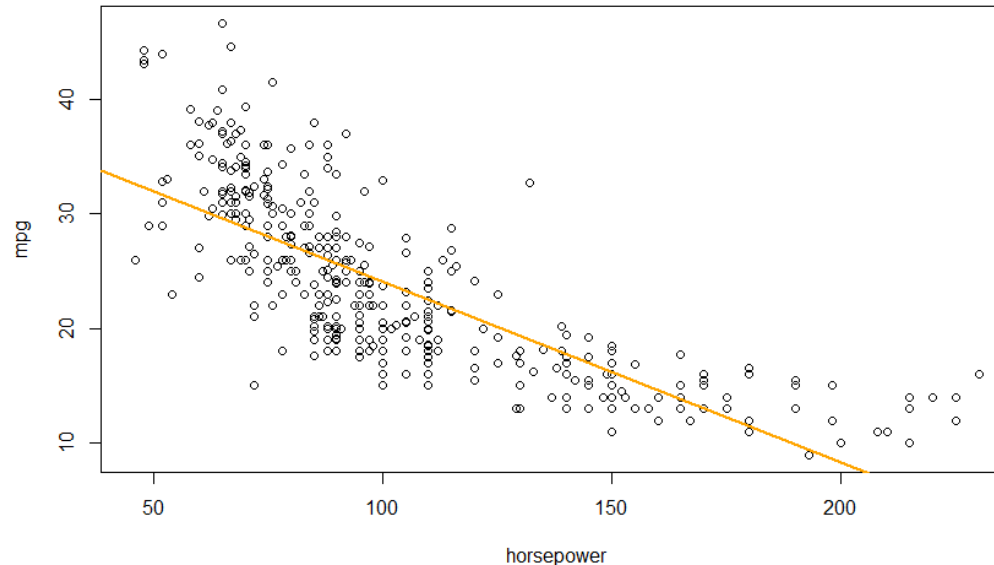
Possible issue: non-linear dependence

- Often: the dependence of Y on the predictors is non-linear
- This might be visible on scatter plots; example from ISL: miles per gallon (mpg) is plotted for cars with different horsepower



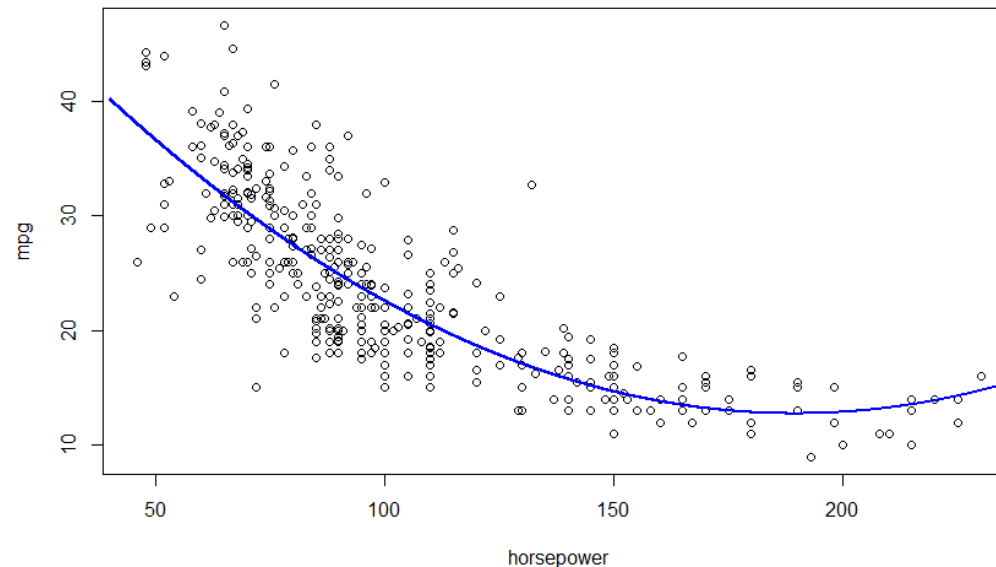
Linear regression does not work well

- Simple linear regression finds the best **line** that fits Y values and does not give good predictions if Y values are not close to any line
- Can we change the model somehow to address such cases?



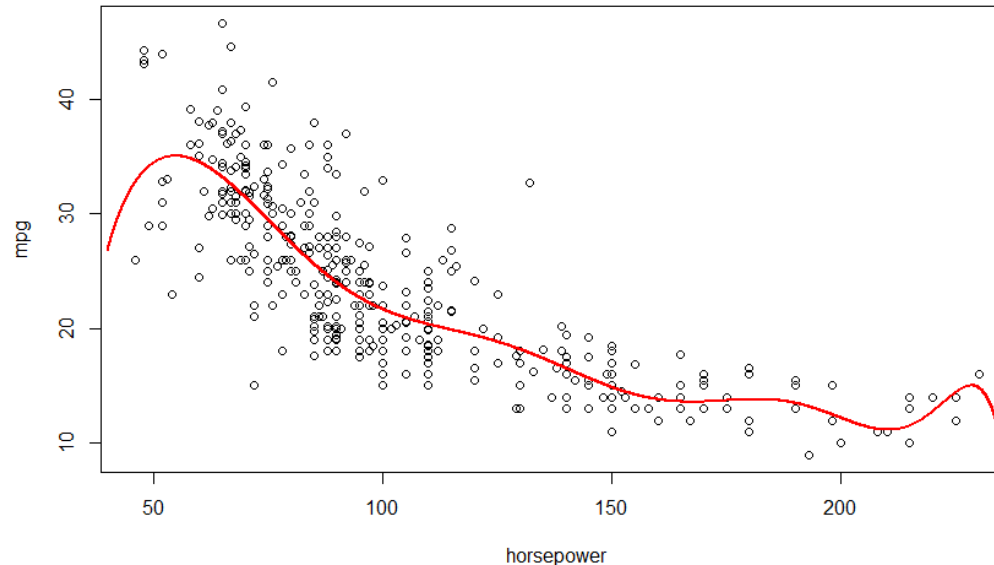
Try higher power terms as predictors!

- If a quadratic/cubic dependence is suspected → simply include a higher degree term of the predictor in the linear model!
- Model including quadratic term gives a better fit:



Choosing complexity

- Which powers of predictors should be included?
- Including too high powers may lead to overfitting – see example with polynomial of degree 10 below



Comparing model fit

Linear

```
call:
lm(formula = mpg ~ horsepower)

Residuals:
    Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66  <2e-16 ***
horsepower  -0.157845   0.006446  -24.49  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

Quadratic

```
call:
lm(formula = mpg ~ I(horsepower^2) + horsepower)

Residuals:
    Min       1Q   Median       3Q      Max
-14.7135  -2.5943  -0.0859   2.2868  15.8961

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.9000997   1.8004268   31.60  <2e-16 ***
I(horsepower^2) 0.0012305   0.0001221   10.08  <2e-16 ***
horsepower  -0.4661896   0.0311246  -14.98  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.374 on 389 degrees of freedom
Multiple R-squared:  0.6876,    Adjusted R-squared:  0.686
F-statistic: 428 on 2 and 389 DF,  p-value: < 2.2e-16
```

Cubic

```
call:
lm(formula = mpg ~ I(horsepower^3) + I(horsepower^2) + horsepower)

Residuals:
    Min       1Q   Median       3Q      Max
-14.7039  -2.4491  -0.1519   2.2035  15.8159

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.068e+01  4.563e+00  13.298  < 2e-16 ***
I(horsepower^3) -2.147e-06  2.378e-06  -0.903   0.3673
I(horsepower^2) 2.079e-03  9.479e-04   2.193   0.0289 *
horsepower  -5.689e-01  1.179e-01  -4.824  2.03e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.375 on 388 degrees of freedom
Multiple R-squared:  0.6882,    Adjusted R-squared:  0.6858
F-statistic: 285.5 on 3 and 388 DF,  p-value: < 2.2e-16
```

- Quadratic model gives clear improvement on linear
- No significant improvement by cubic term

Feedback

Feedback quiz

Feedback is essential to me so that I can improve the lectures during the course. Please take your chance to optimize your learning experience!

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

1. Go to www.menti.com
2. Enter the code 14 71 24
3. Rate your experience today in slide 1
4. Wait until I change slide
5. Answer to the questions in slide 2