Multiple (non) linear regression

Jiří Kléma

Department of Computer Science,
Czech Technical University in Prague

Lecture based on **ISLR book** and its accompanying slides

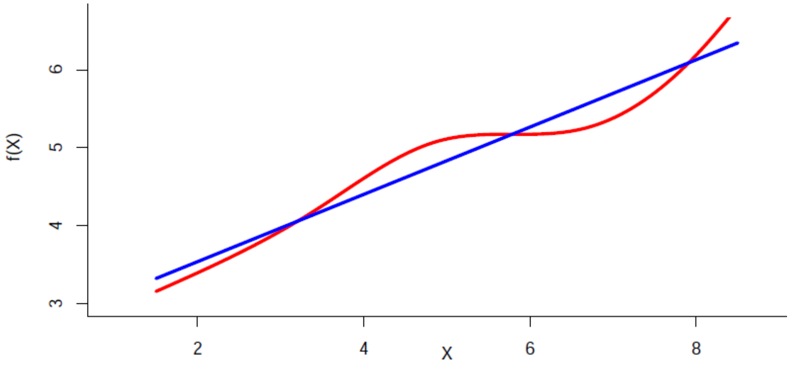
http://cw.felk.cvut.cz/wiki/courses/b4m36san/start

Agenda

- Linear regression
  - a simple model with a single predictor,
  - parameters, interpretation, hypotheses testing,
  - generalization towards multiple linear regression,
  - special issues: qualitative predictors, outliers, collinearity,
- linear model selection and regularization
  - subset selection,
  - regularization = shrinkage, lasso, ridge regression,
  - choosing the optimal model, estimating test error,
- moving beyond linearity
  - polynomial regression,
  - step functions, splines,
  - local regression,
  - generalized additive models.

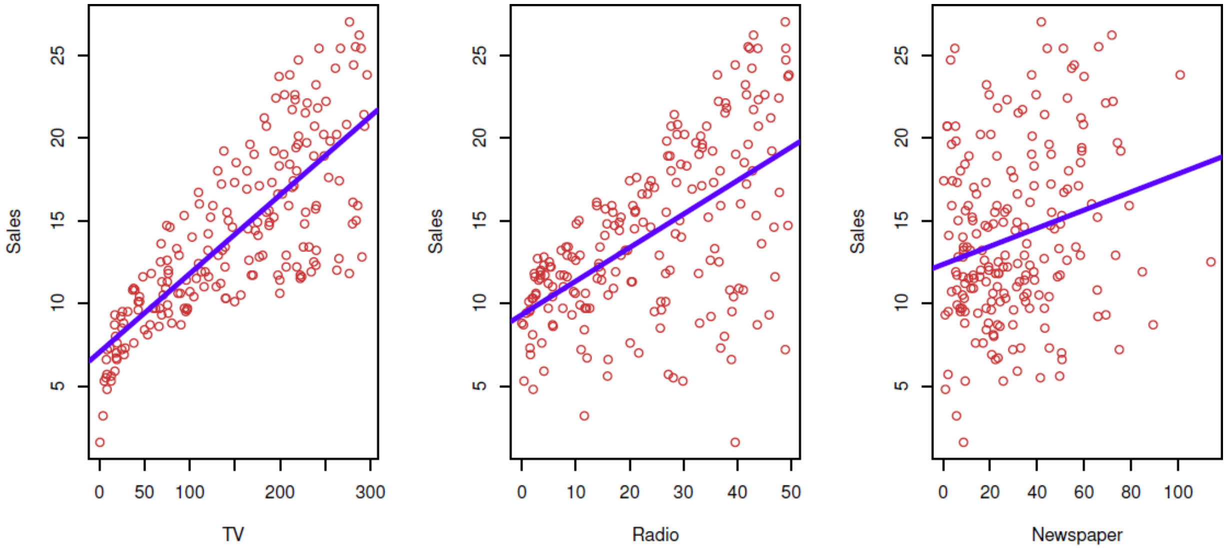
Linear regression

- Assumption of linearity
  - often simplifying assumption only
    - ∗ true regression functions are rarely linear,
    - ∗ still, the linear model extremely useful, both conceptually and practically,
  - the simplification increases ability to learn
    - ∗ helps to cope with the well-known curse of dimensionality,
  - the simplification brings interpretability
    - ∗ a reasonable number of parameters with clear meaning,
  - good performance preserved in case of moderate violation
    - ∗ linear models can be extended otherwise.



Linear regression for the advertising data

- Consider the advertising data shown below, questions we might ask:
  - is there a relationship between advertising budget and sales?
  - how strong is the relationship between advertising budget and sales?
  - which media contribute to sales?
  - how accurately can we predict future sales?
  - is the relationship linear?
  - is there synergy among the advertising media?



Simple linear regression using a single predictor X

- We assume a model

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

  - where β_0 and β_1 are two unknown constants (coefficients, parameters),
  - they represent the **intercept** and **slope**,
  - ϵ is the error term, normally distributed with 0 mean, the same variance at every X, independent,

- we predict the future values of independent variable using

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

  - where the hat symbol denotes an estimated value,
  - $\hat{y}$ indicates a prediction of Y on the basis of $X = x$.

Estimation of the parameters by least squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X,
- then $e_i = y_i - \hat{y}_i$ represents the ith residual,
- we define the **residual sum of squares** (RSS) as

$$RSS = e_1^2 + e_1^2 + \cdots + e_m^2 = \sum_{i=1}^{m} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- the least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS

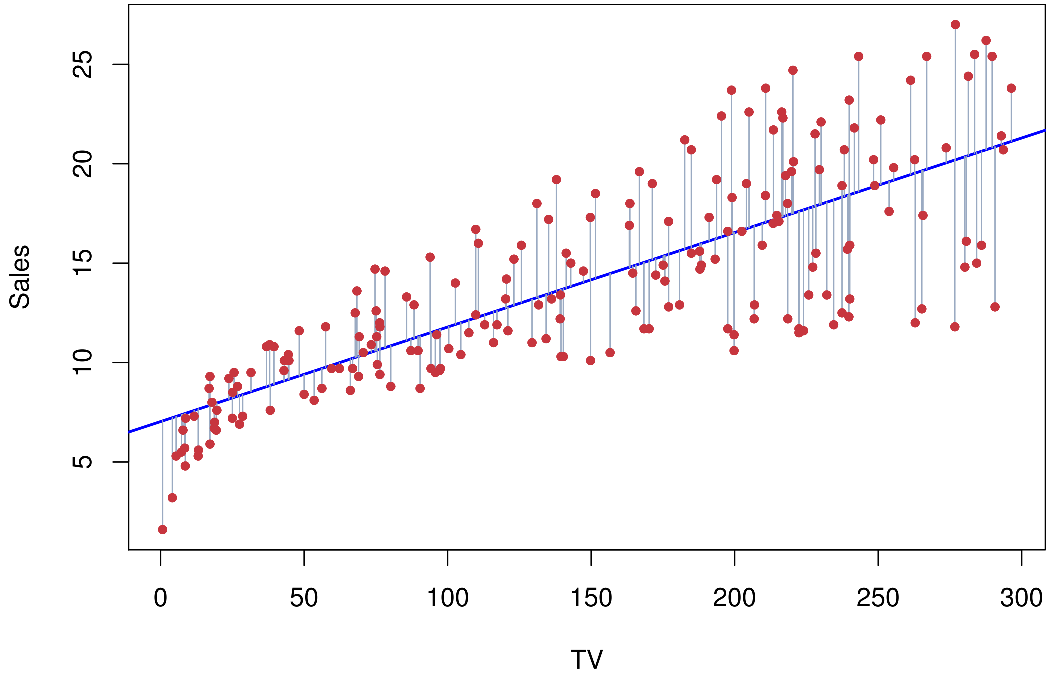
$$\frac{\partial RSS}{\partial \hat{\beta}_1} = 0 \rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m} (x_i - \bar{x})^2}$$

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = 0 \rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- where $\bar{x}$ and $\bar{y}$ are the sample means for X and Y.

Example: advertising data

- The least squares fit for the regression of $sales$ onto TV
  - a linear fit captures the essence of the relationship,
  - although it is somewhat deficient in the left of the plot,
  - $\hat{\beta}_0$: no TV advertising, around 7 (thousand) units sold,
  - $\hat{\beta}_1$: \$1,000 more spent on TV associated with selling ∼48 additional units.

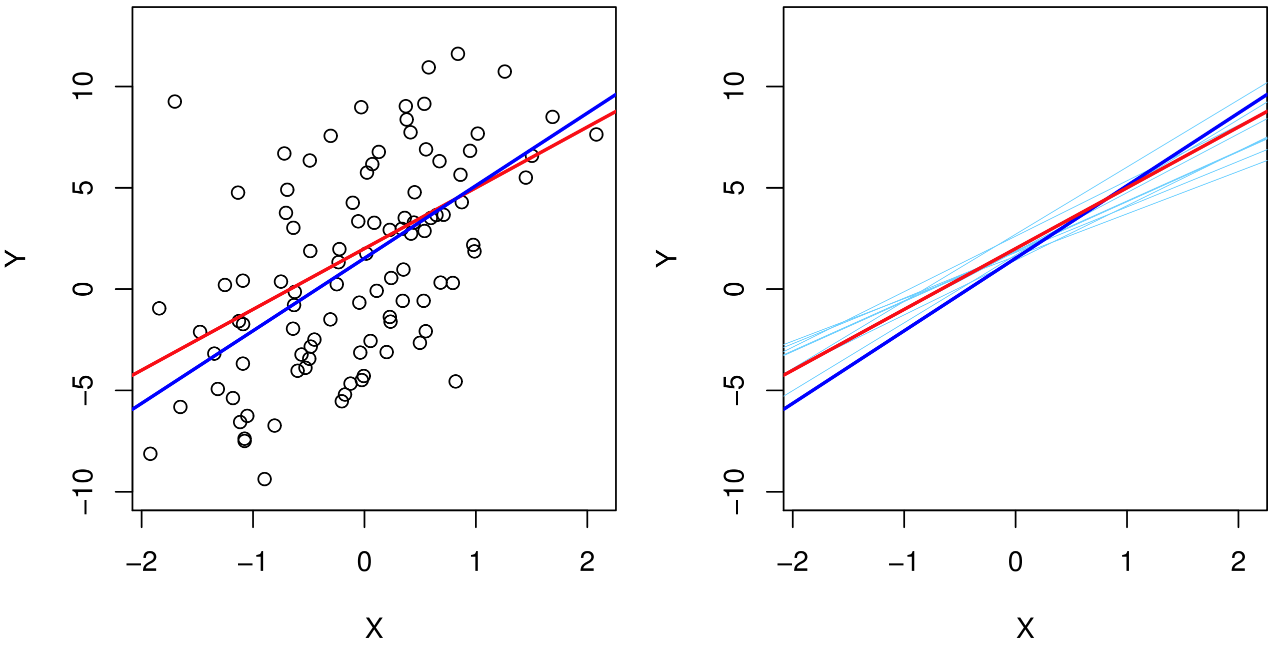


Assessing the accuracy of the coefficient estimates

- Standard error of an estimator reflects how it varies under repeated sampling

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{m}(x_i - \bar{x})^2} \quad SE(\hat{\beta}_0)^2 = \sigma^2\left(\frac{1}{m} + \frac{\bar{x}^2}{(x_i - \bar{x})^2}\right)$$

  - where $\sigma^2 = Var(\epsilon)$,
  - **residual standard error** RSE is σ estimate, $RSE = \sqrt{RSS/(m-2)}$.



Coefficient confidence intervals

- Confidence interval (CI)
  - $100(1-\alpha)\%$ confidence interval is a range of values that encompasses the true (unknown) population parameter in $100(1-\alpha)\%$ repeated sampling trials like this,
  - there is approximately a 95% chance that the interval below will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample)

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

  - a more precise (and general) CI estimate is based on $1-\alpha/2$ quantile of a t-distribution with (m-2) degrees of freedom

$$[\hat{\beta}_1 - t_{1-\alpha/2,m-2}SE(\hat{\beta}_1), \hat{\beta}_1 + t_{1-\alpha/2,m-2}SE(\hat{\beta}_1)]$$

  - for the advertising data, the 95% CI for β_1 is [0.042, 0.053].

Hypothesis testing

- The most common hypothesis test
  - H_0: there is no relationship between X and Y,
  - H_A: there is some relationship between X and Y,
- mathematically this corresponds to testing
  - $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$,
- the test stems from the standard error and t-statistic given by

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

  - the statistic will have a t-distribution with m-2 degrees of freedom,
  - the corresponding p-value is the probability of observing any value equal to $|t|$ or larger,
  - both under the H_0 assumption, i.e., assuming $\beta_1 = 0$.

Assessing the overall accuracy of the model

- **R-squared** gives the fraction of variance explained by the model

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

  - where $TSS = \sum_{i=1}^{m} (y_i - \bar{y})^2$ stands for the total sum of squares,
  - and $RSS = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$ stands for the residual sum of squares,
- while RSE mentioned earlier gives an absolute measure of its lack of fit,
- it can be shown that in this simple linear regression setting $R^2 = r^2$,
  - where r is the correlation between X and Y.

$$r = \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{m}(y_i - \bar{y})^2}}$$

Advertising data results: X=TV, Y=Sales

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

Quantity	Value
RSE	3.26
R^2	0.612
F-statistic	312.1

Multiple linear regression

- Here, a following model is assumed

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

- interpretation of the regression coefficients
  - β_j represents the **average** effect on Y of a one unit increase in X_j, **holding all other predictors fixed**,
  - it perfectly fits the balanced design where the predictors are uncorrelated,
  - correlations among predictors cause problems
    - the variance of all coefficients tends to increase, sometimes dramatically,
    - interpretations become hazardous, when X_j changes, everything else changes,
  - example: $Y =$ number of tackles by a football player in a season; W and H are his weight and height; fitted regression model is $\hat{Y} = b_0 + .50W - .10H$; how do we interpret $\hat{\beta}_2 < 0$?
  - claims of causality should be avoided for observational data.

Estimation of the parameters for multiple regression

- No principal changes from the simple model,
- the prediction formula is

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

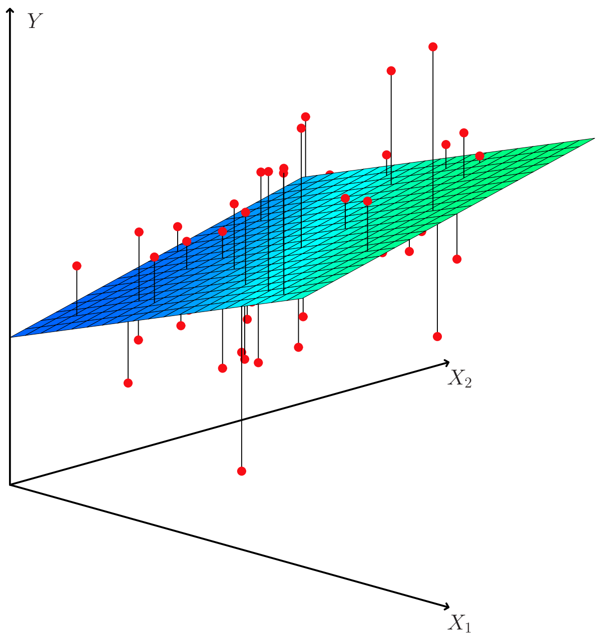
- the parameter estimates obtained by RSS minimization

$$RSS = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{m} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2$$

- **ordinary least squares** estimation
  - the most simple approach,
- **generalized least squares**
  - allow efficient β estimation when heteroscedascity or correlations are present.

Some important questions

1. Is at least one of the p predictors useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of them useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?



Is at least one predictor useful?

- Formally: $H_0: \ \beta_1 = \beta_2 = \cdots = \beta_p = 0$ vs H_A: at least one $\beta_j \neq 0$,
- this test is performed by computing the F-statistic

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

- in fact, we compare fit of the full (RSS) and intercept only model (TSS)
  - technically, we compute the ratio between explained and unexplained variance of the full model,
- provided H_0 is true
  - $E((TSS - RSS)/p) = E(RSS/(m-p-1)) = \sigma^2$ and F is close to 1,
  - the test is adjusted to the number of predictors p and the sample size m,
- F-statistic is compared with quantiles of $F(p, m-p-1)$ distribution.

Advertising data results: the full model

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Correlations	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Quantity	Value
RSE	1.69
R^2	0.897
F-statistic	570

Deciding on the important variables

- Selection cannot be directly based on the observed predictor p-values
  - namely for large p, risk of false discoveries due to multiple comparisons,
- build and compare (a lot of) alternative models
  - use a criterion that balances the training error and model size,
  - to exemplify, Mallow's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R^2, and cross-validation (CV),
- the most direct search methods
  - **all subsets** regression is not feasible, $\mathcal{O}(2^p)$,
  - **forward stepwise selection** starts with the null model and gradually adds the variable that results in the lowest RSS, $\mathcal{O}(p^2)$,
  - **backward stepwise selection** starts with the full model, gradually removes the variable with the largest p-value in the last model, i.e., the least significant one, $\mathcal{O}(p^2)$, cannot be used if $p > m$,
- a more systematic overview will be done in model regularization section.

Categorical predictors

- So far, categorical and continuous independent variables treated separately
  - often we need to study them concurrently, employ them in regression,
- for binary predictors we create a new 0/1 variable X_i with the resulting model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} x_i = 1 : \beta_0 + \beta_1 + \epsilon_i \\ x_i = 0 : \beta_0 + \epsilon_i \end{cases}$$

  - interpretation: β_0 is the average outcome in the zero group, $\beta_0 + \beta_1$ is the average outcome in the positive group, β_1 is the average difference in outcomes between groups,
- however, a -1/1 dummy variable could be introduced too

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} x_i = 1 : \beta_0 + \beta_1 + \epsilon_i \\ x_i = -1 : \beta_0 - \beta_1 + \epsilon_i \end{cases}$$

  - the predictions will be identical, β values and interpretation change.

Categorical predictors

- for predictors with l levels we typically create $l-1$ dummy variables,
- e.g., $ethnicity$ $\in$\{Asian, Caucasian, African American\} could be captured by

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

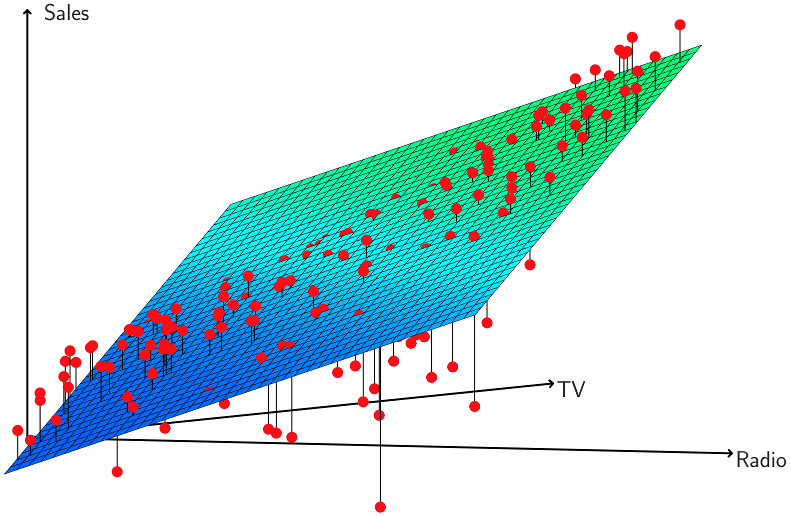
- the level with no dummy variable (African American) is known as the baseline,
- the dummy variables appear in the regression equation

$$y_i = \beta_0 + \beta_1 x_{ii} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American} \end{cases}$$

  - interpretation: β_0 is the average outcome in the baseline group, β_1 is the average outcome increase in the first group, β_2 is the average outcome increase in the second group (both wrt baseline).

Interaction in the advertising data?

- Regular linear model is additive
  - e.g., the average effect on sales of a one-unit increase in TV is always β_{TV}, regardless of the amount spent on radio,
- however, there is an interaction between TV and radio spending
  - when advertising is split between the two media, the additive linear model tends to underestimate sales → there is synergy between the predictors.



Modeling interactions

- Add an interaction term into the model, in the case of advertising problem

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{TV} \times \text{radio}) + \epsilon$$

- in this model, the effect of TV changes with the value of radio (and vice versa)

$$\text{sales} = \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \epsilon$$

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + (\beta_2 + \beta_3 \times \text{TV}) \times \text{radio} + \epsilon$$

- results of this model confirm the role of the interaction

	Coefficient	Std. error	t-statistic	p-value
Intercept	6,7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Modeling interactions – interpretation and hints

- The p-value for the interaction term TV×radio is extremely low
  - indicating that there is strong evidence for $H_a : \beta_3 \neq 0$,
- R^2 for the interaction model is 96.8%, compared to
  - compared to only 89.7% for the model using TV and radio without an interaction term,
  - $\frac{96.8-89.7}{100-89.7} = 69\%$ of the variability in sales that remains after fitting the additive model has been explained by the interaction term,
- the coefficient estimates in the table suggest that
  - increase in TV advertising of $1,000 is associated with increased sales of $(\beta_1 + \beta_3 \times \text{radio}) \times 1000 = 19 + 1.1 \times \text{radio}$ units,
- If we include an interaction in a model
  - we should also include the main effects (TV and radio in our case), even if the p-values associated with their coefficients are not significant.

How regression got its name . . .

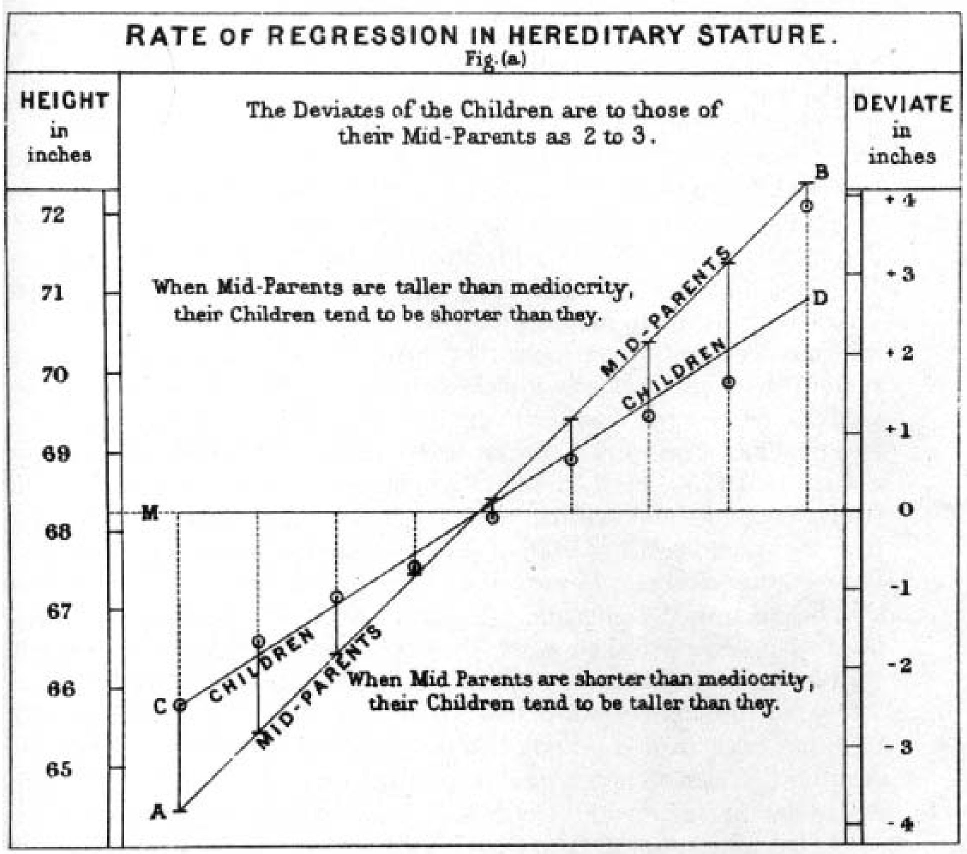

RATE OF REGRESSION IN HEREDITARY STATURE.
Fig. (a.)
HEIGHT in inches
The Deviates of the Children are to those of their Mid-Parents as 2 to 3.
DEVIATE in inches
72
71
70
69
68
67
66
65
+4
+3
+2
+1
0
-1
-2
-3
-4
B
D
M
C
A
MID-PARENTS
CHILDREN
When Mid-Parents are taller than mediocrity, their Children tend to be shorter than they.
When Mid Parents are shorter than mediocrity, their Children tend to be taller than they.

Figure 8.8. Galton's graphical illustration of regression; the circles give the average heights for groups of children whose midparental heights can be read from the line AB. The difference between the line CD (drawn by eye to approximate the circles) and AB represents regression toward mediocrity. (From Galton, 1886a.)

Choosing the optimal model

- Three classes of methods
  - subset selection – already mentioned, repetitive application of least squares on various reduced sets of predictors,
  - dimension reduction – ordinary least squares regression in a L-dimensional subspace, see the lectures on dimension reduction,
  - **shrinkage** – we fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates, this shrinkage (also known as regularization) has the effect of reducing variance and can also perform variable selection,
- RSS and R^2 are not suitable for selecting the best model
  - at least, among a collection of models with different numbers of predictors,
  - they are related to the training error,
  - the model containing all of the predictors will always have the smallest RSS and the largest R^2.

Adjusted R-squared

- Unlike the R^2 statistic, it pays a price for unnecessary predictors,
- for a least squares model with p variables

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(m-p-1)}{TSS/(m-1)}$$

  - a maximization criterion (unlike C_p , AIC, and BIC),
  - a heuristic criterion, vaguely motivated in statistical theory,
  - irrelevant variables bring a small decrease in RSS, this decrease is outweighed by decrease in $m - p - 1$ (neither TSS nor m changes with p),
  - a comparative measure, different meaning than R^2 (a measure of fit), can be e.g. negative.

Shrinkage Methods

- **Ridge regression**
  - recall that the least squares fitting procedure estimates $\beta_0, \ldots, \beta_p$ with the values that minimize

$$RSS = \sum_{i=1}^{m} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2$$

  - in contrast, the ridge regression coefficient estimates are the values that minimize

$$\sum_{i=1}^{m} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2 = RSS + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2$$

  - where $\lambda \geq 0$ is a tuning parameter, to be determined separately, typically, CV is used,
  - $\lambda \sum_j \hat{\beta}_j^2$ is a **shrinkage penalty** with the effect of shrinking the β_j estimates towards zero.

Shrinkage Methods

- **Ridge regression**
  - the standard least squares coefficient estimates are scale equivariant
    - multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$.
    - regardless of how the jth predictor is scaled, $X_j\hat{\beta}_j$ will remain the same,
  - in contrast, the ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant
    - ridge regression should be applied after standardizing the predictors

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

Ridge regression improvement over least squares

- **Bias-variance trade-off**
  - suppose we have fit a model $\hat{f}(x)$ to some training data,
  - let (x_0, y_0) be a test observation drawn from the same population,
  - if the true model is $Y = f(X) + \epsilon$ (with $f(x) = E(Y|X = x)$) then

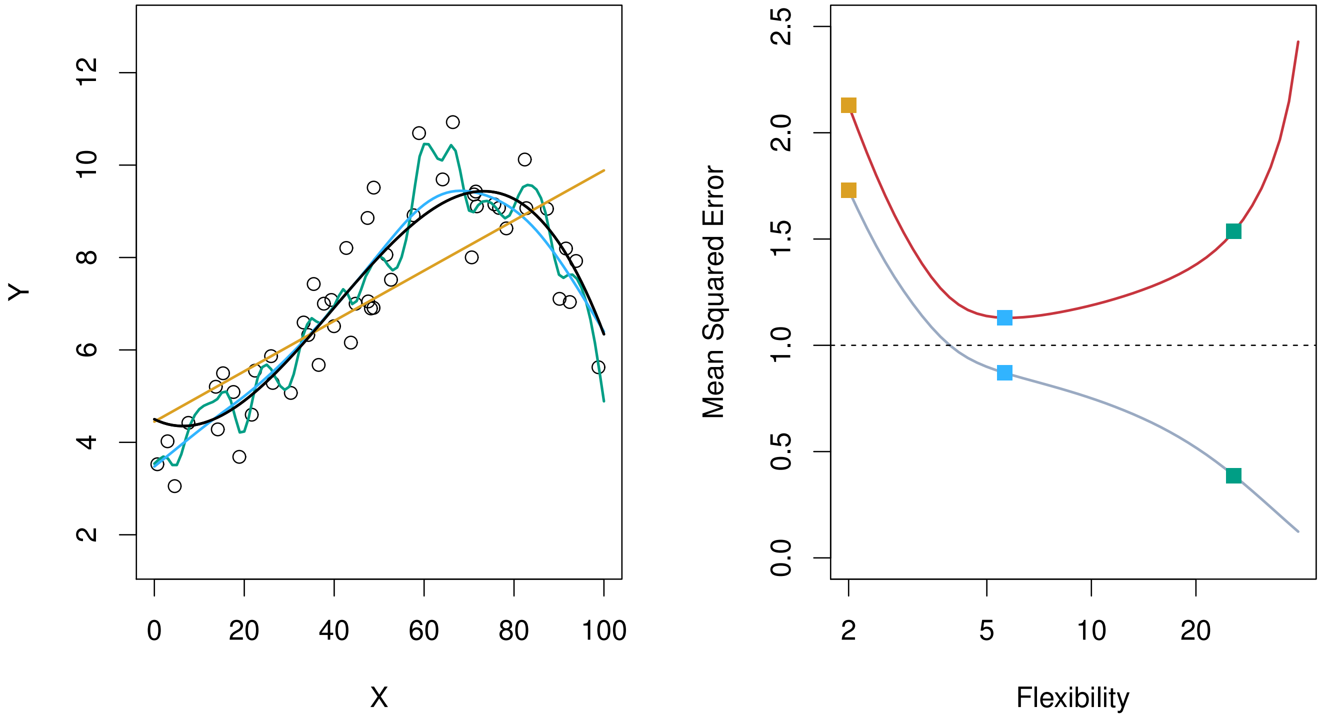
$$E\big(y_0 - \hat{f}(x_0)\big)^2 = Var(\hat{f}(x_0)) + \big[Bias(\hat{f}(x_0))\big]^2 + Var(\epsilon)$$

$$Bias(\hat{f}(x_0)) = E(\hat{f}(x_0)) - f(x_0)$$

  - the error can be decomposed into model **variance**, **bias** and irreducible error,
  - typically as the flexibility of $\hat{f}$ increases, its variance increases, and its bias decreases,
  - choosing the flexibility based on average test error amounts to a bias-variance trade-off.

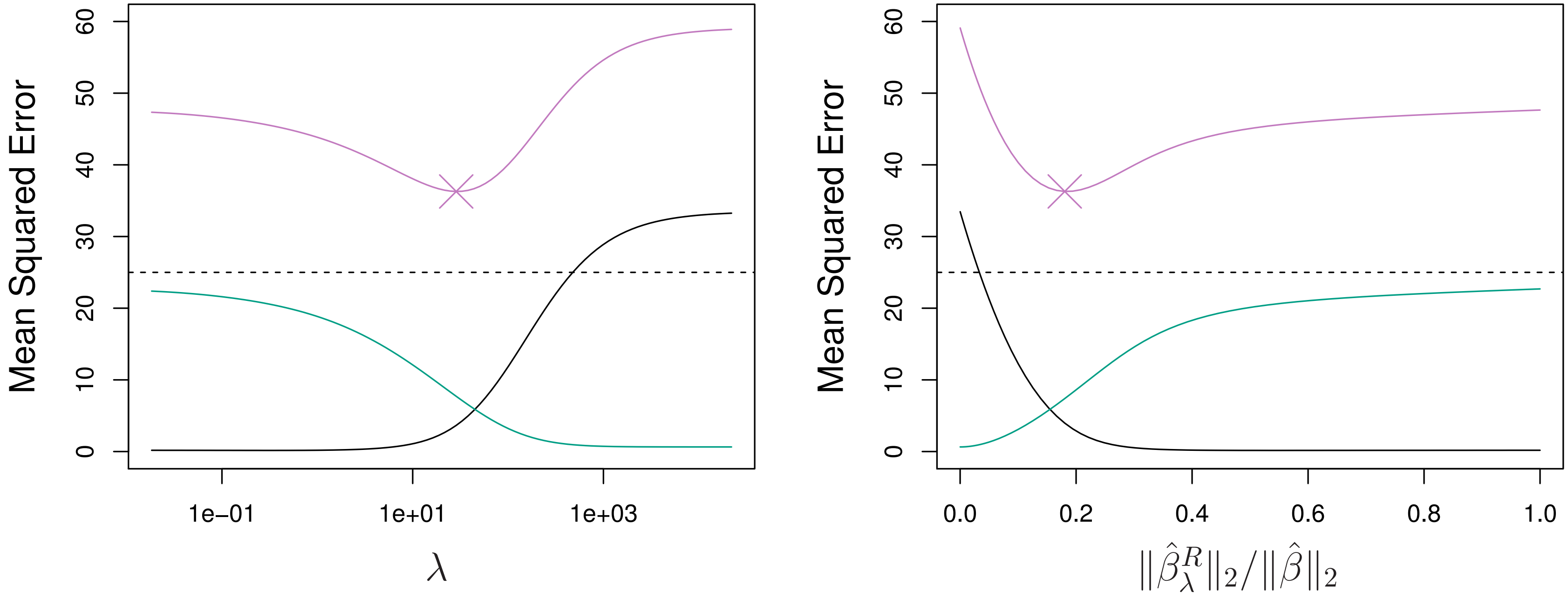
Ridge regression improvement over least squares

- **Relationship between model flexibility and accuracy**
  - and the bias of the train error towards more flexible/overfit models.



Black curve is truth. Orange, blue and green curves/squares correspond to fits/models of different flexibility. Red curve on right is generalization error, grey curve is the training error.

Ridge regression improvement over least squares



Simulated data with $n = 50$ observations, $p = 45$ predictors, all having nonzero coefficients. Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and the ration between the norms of ridge and ordinary regression coefficients. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.

Shrinkage Methods

- **The lasso**
  - ridge regression will include all p predictors in the final model
    - $*$ disadvantage, does not help in feature selection,
  - lasso overcomes this problem by minimizing

$$\sum_{i=1}^{m}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\hat{\beta}_j| = RSS + \lambda\sum_{j=1}^{p}|\hat{\beta}_j|$$

    - $*$ in statistical parlance, the lasso uses an ℓ_1 penalty instead of an ℓ_2 penalty applied in ridge regression,
  - the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when λ is sufficiently large,
  - the lasso yields **sparse models** and performs **variable selection**.

The variable selection property of the lasso

- Why is it that the lasso, unlike ridge regression, results in coefficient estimates that are exactly equal to zero?
- One can show that the lasso and ridge regression coefficient estimates solve the problems minimize

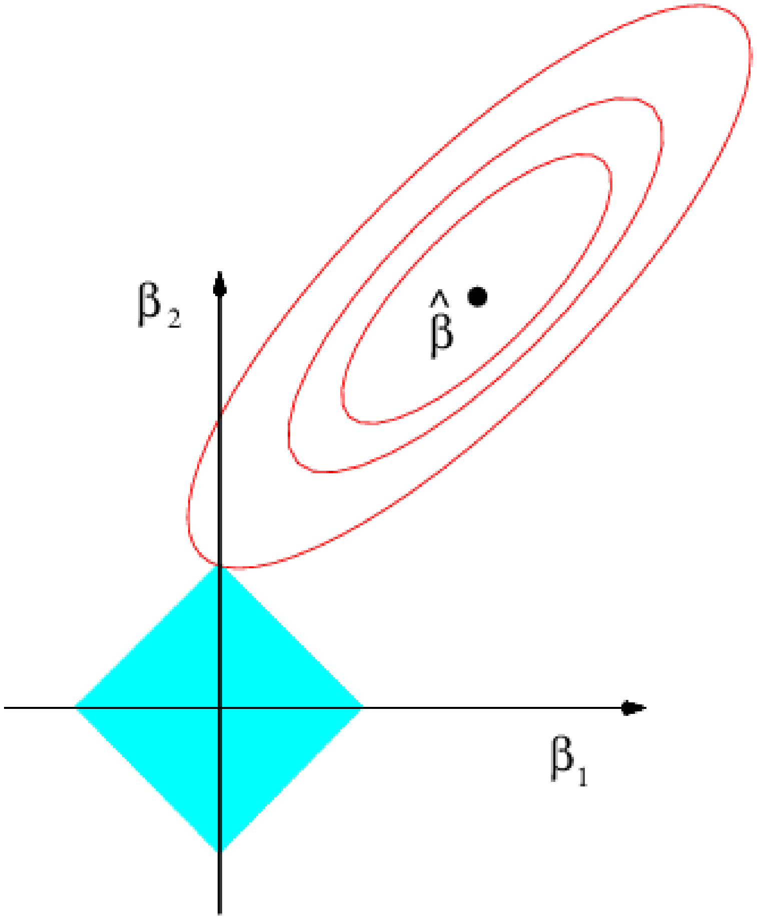
$$\min_{\hat\beta} \sum_{i=1}^{m} (y_i - \hat\beta_0 - \sum_{j=1}^{p} \hat\beta_j x_{ij})^2 \text{ subject to } \sum_{j=1}^{p} |\hat\beta_j| \le s$$

and

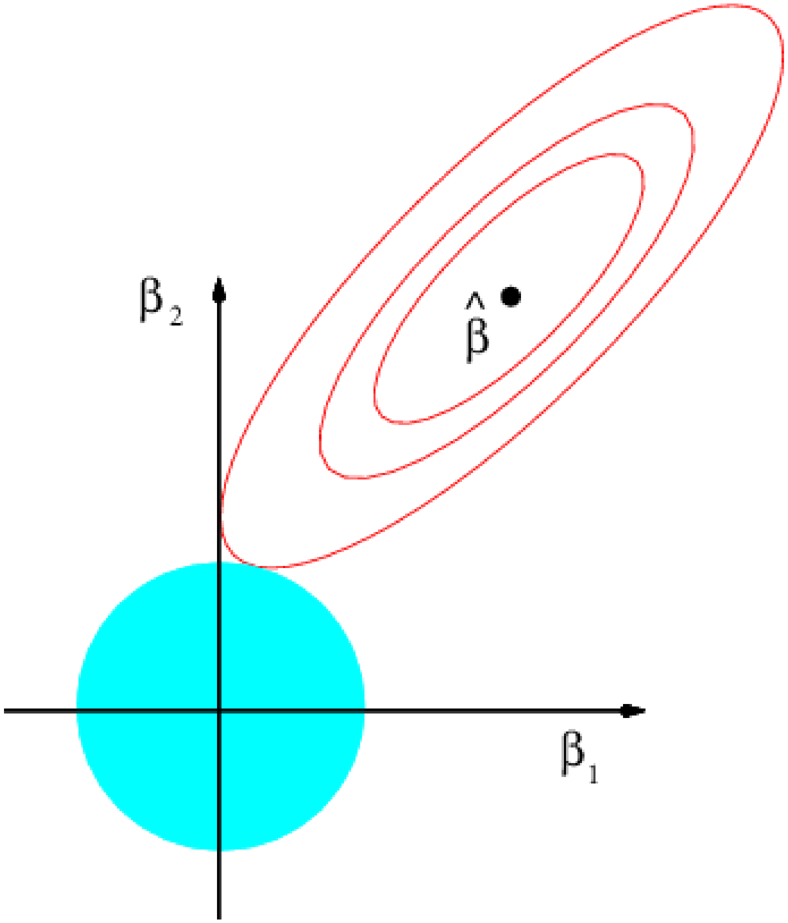
$$\min_{\hat\beta} \sum_{i=1}^{m} (y_i - \hat\beta_0 - \sum_{j=1}^{p} \hat\beta_j x_{ij})^2 \text{ subject to } \sum_{j=1}^{p} \hat\beta_j^2 \le s$$

- in other words, for every value of λ there is some s such that the alternative definitions lead to the same regression coefficient estimates.

Lasso and ridge regression: geometric interpretation



Lasso



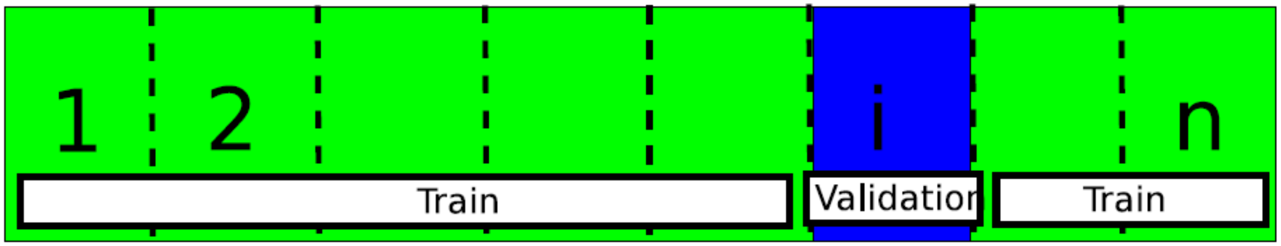
Ridge regression .

Selecting the value of tuning parameter

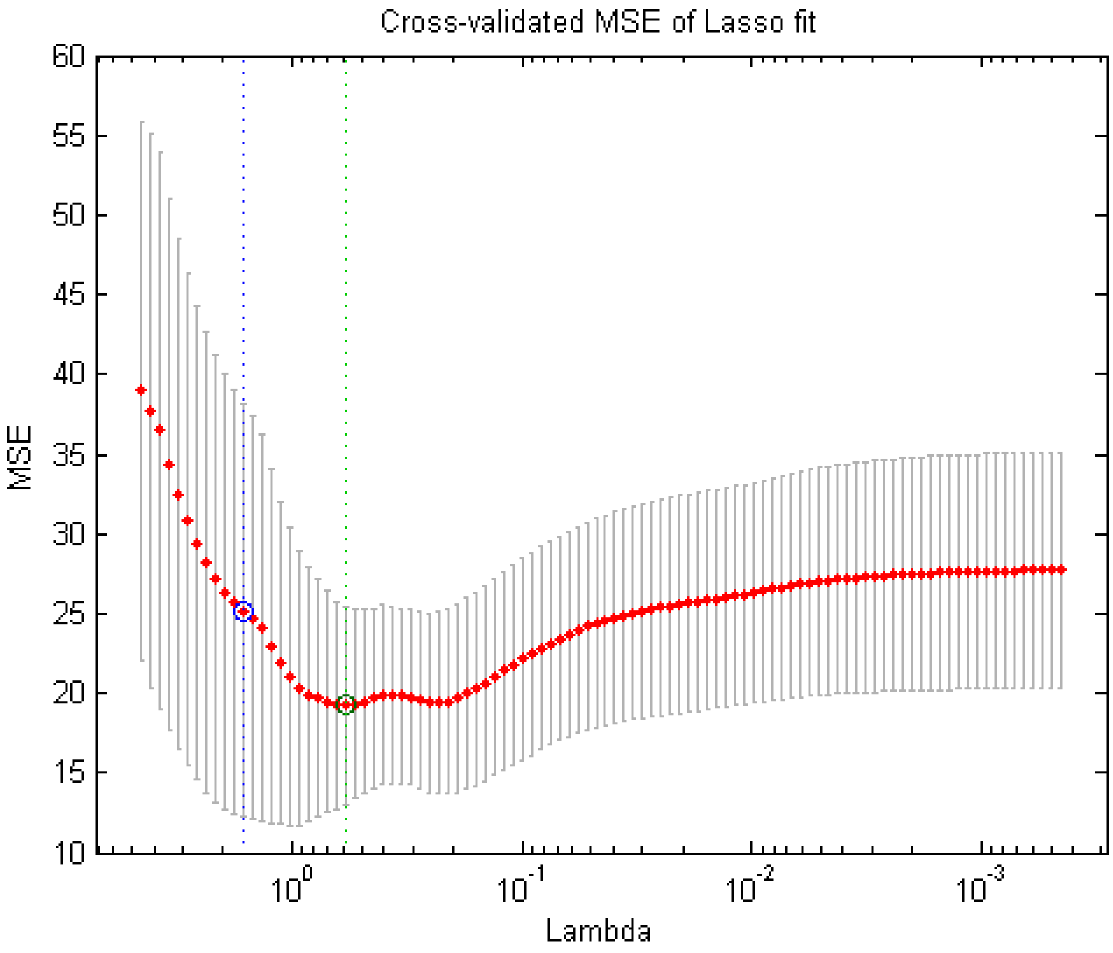
- which λ value (or equivalently, the value of the constraint s) is the best?
- employ **cross-validation**
  - use fresh/unseen data to estimate the expected generalization error,
  - calculate mean squared error $MSE_{Te} = \frac{RSS}{m} = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$,
- in our case, proceed as follows
  - choose a grid of λ values,
  - compute the cross-validation error rate for each value of λ,
  - select either λ_{min} for which the cross-validation error is smallest,
  - or select λ_{1se}, the largest value of λ such that error is within 1 standard error of the minimum,
- Finally, the model is re-fit
  - using all of the available observations,
  - and the selected value of the tuning parameter.

Cross-validation

- the training error can dramatically underestimate the test error
  - it is a positively biased estimate of the future generalization error,
- **hold-out** makes the most easy approach to model testing
  - split the available data between train and validation set (70:30 ratio),
  - sufficient for large data sets,
- k-fold cross-validation
  - randomly split observations into k folds of (approximately) equal size,
  - in k gradual runs perform training on $k-1$ folds and testing on the remaining fold (always different),
  - eventually, k estimates of the test error are averaged.



Comparison between λ_{min} and λ_{1se}



λ_{min} corresponds to the green line, always larger than λ_{1se} which is in blue.

Summary

- Multiple linear regression
  - a simple model with the strong assumption of linearity,
  - helps to understand concurrent effects on a target continuous variable,
- model selection and regularization may improve prediction and understanding
  - neither ridge regression nor the lasso will universally dominate the other,
  - lasso performs better when the response is a function of only a relatively small number of predictors, however ...
  - ... the number of predictors related to the response is never known a priori,
  - cross-validation can help to decide which approach is better on a particular data set and select a value of the tuning parameter.

The main references

:: Resources (slides, scripts, tasks) and reading

- G. James, D. Witten, T. Hastie and R. Tibshirani: **An Introduction to Statistical Learning with Applications in R.** Springer, 2014.
- K. Markham: **In-depth Introduction to Machine Learning in 15 hours of Expert Videos**. Available at R-bloggers.