Linear regression

Paweł Kłeczek

WEAIIB, Katedra Automatyki i Robotyki, ISS

2017/18

Linear regression

- a very simple approach for supervised learning
- a useful tool for predicting a quantitative response
- a good jumping-off point for newer approaches (many fancy statistical learning approaches can be interpreted as generalizations/extensions of linear regression)

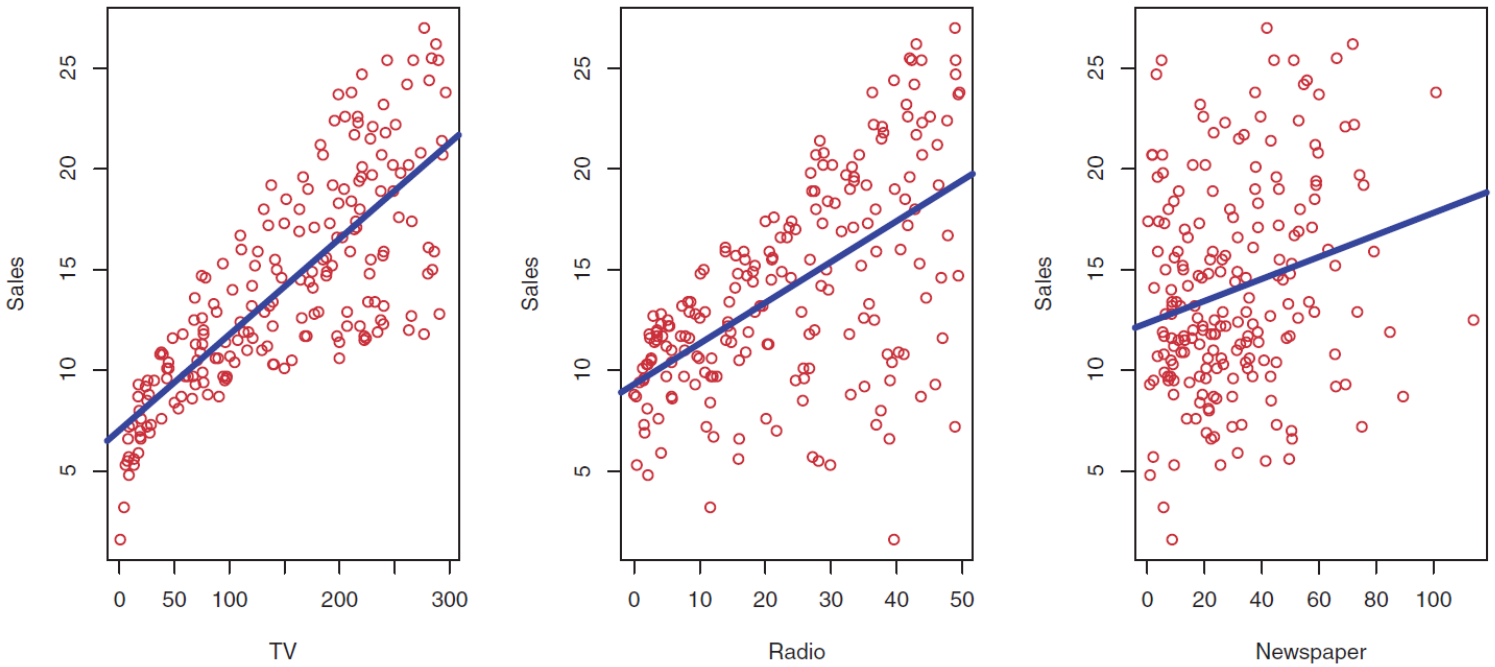
Supervised Learning:

have access to the “right answer” for each example in the data

Regression Problem:

predict real-valued output

A simple regression task



The `Advertising` data set ($n = 200$). In each plot the blue line represents a simple model that can be used to predict sales (the simple least squares fit of sales to that variable).

Task: Suggest, on the basis of this data, a marketing plan for next year that will result in high product sales.

What information would be useful in order to provide such a recommendation?

A simple regression task

Some important questions:

- 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 estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Linear regression can be used to answer each of these questions!

Simple linear regression (SLR)

Assumption: a linear relationship between X and Y:

$$Y \approx \beta_0 + \beta_1 X$$

Example:

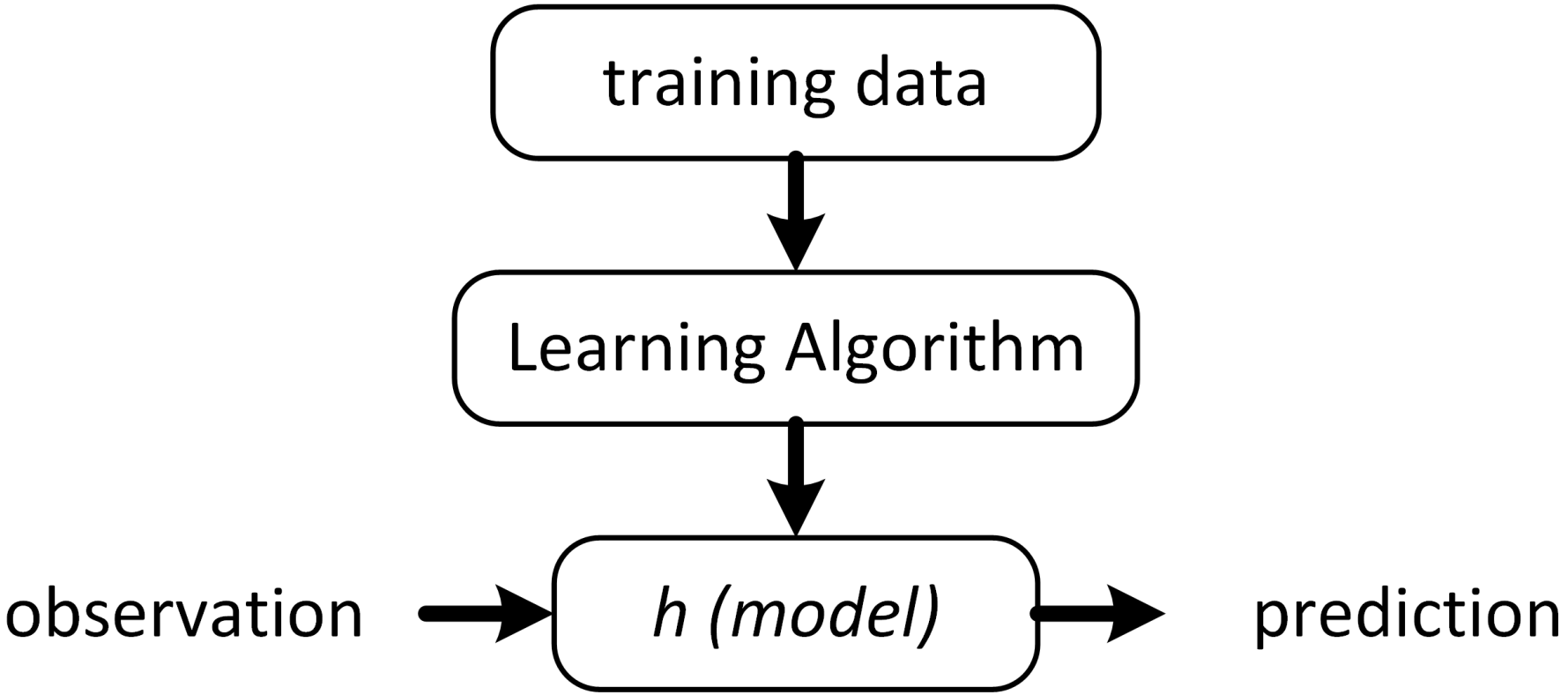
X – TV advertising, Y – sales

Regress sales onto TV by fitting the model:

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

β_0, β_1 – unknown constants that represent the intercept and slope terms in the linear model (aka model coefficients, model parameters)

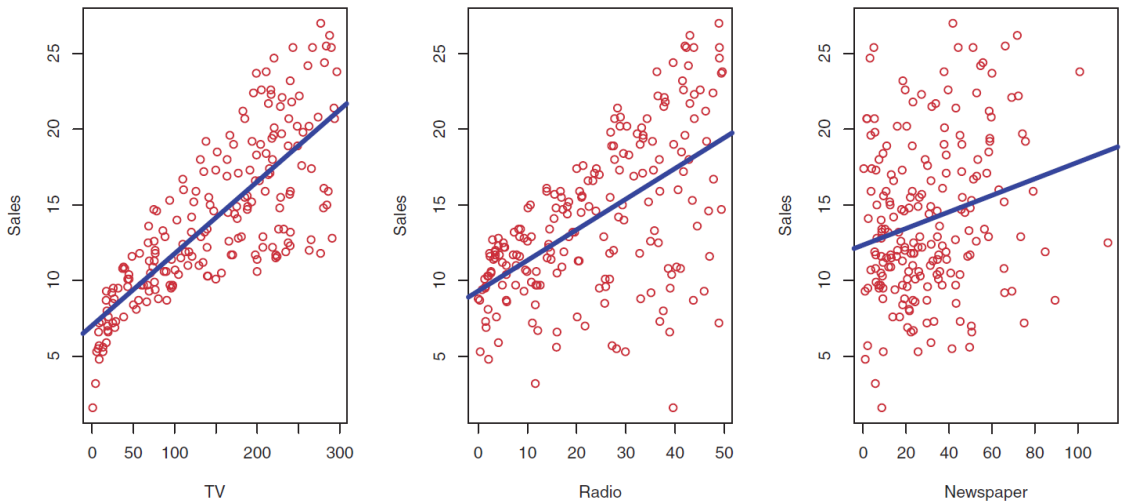
SLR workflow



training data $\Rightarrow$ estimate $\hat{\beta}_0$, $\hat{\beta}_1$ (model) $\Rightarrow$ predictions ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1$ x)

Estimating the coefficients

In practice, β_0 and β_1 are unknown – we must use data to estimate them: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ – n observation pairs (eg. TV advertising budget and product sales)



The `Advertising` data set.

Goal: obtain coefficient estimates $\hat{\beta}_0$, $\hat{\beta}_1$ such that the linear model fits the available data well (= find an intercept $\hat{\beta}_0$ and a slope $\hat{\beta}_1$ such that the resulting line is as close as possible to the n data points)

Residual sum of squares (RSS)

Most common way of measuring closeness:
minimizing the least squares criterion

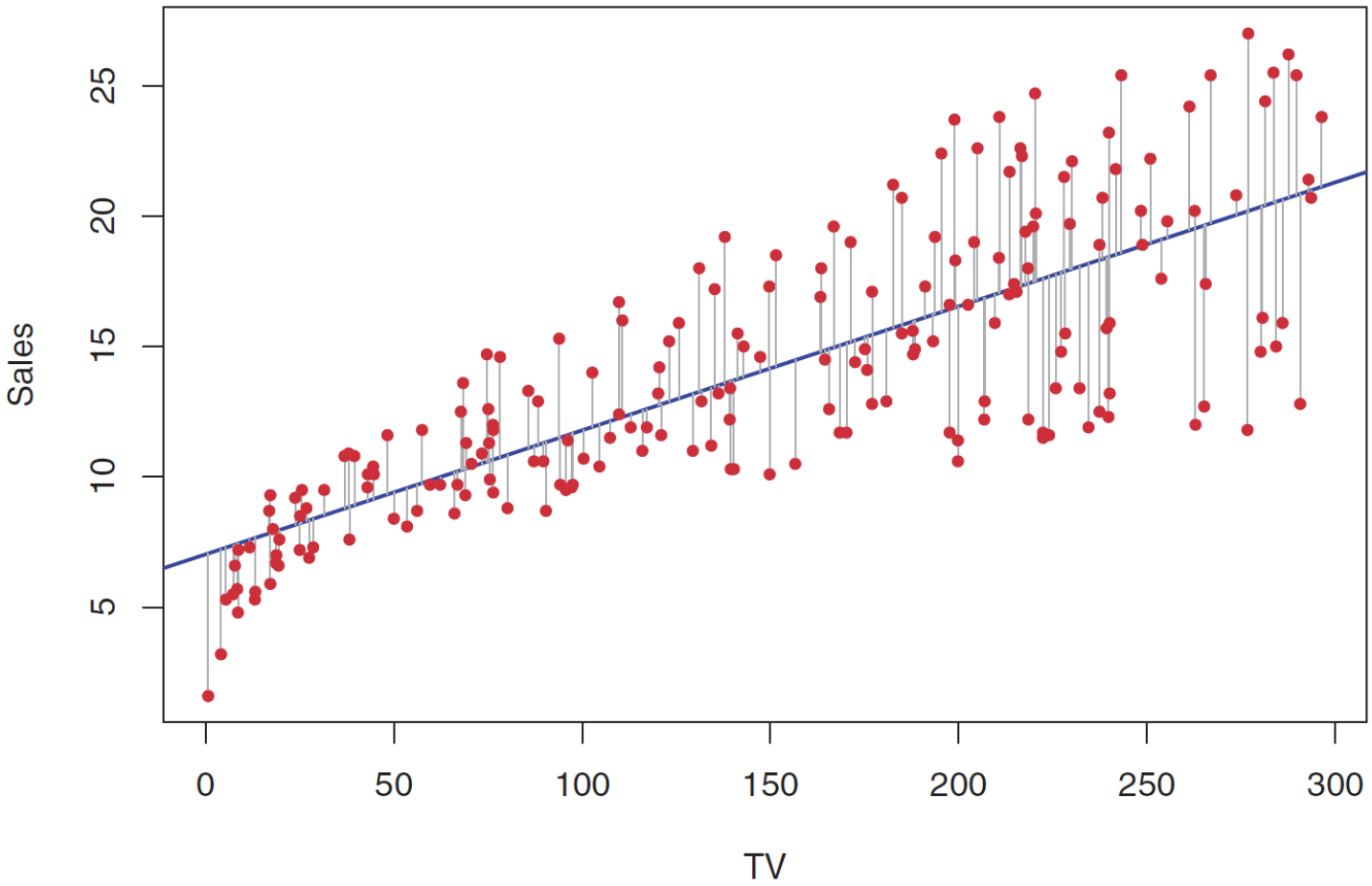
Residual sum of squares (RSS):

$$\text{RSS} = e_1^2 + e_2^2 + \ldots + e_n^2$$

(ith residual: $e_i = y_i - \hat{y}_i$)

Least squares approach: choose $\hat{\beta}_0$, $\hat{\beta}_1$ to minimize the RSS

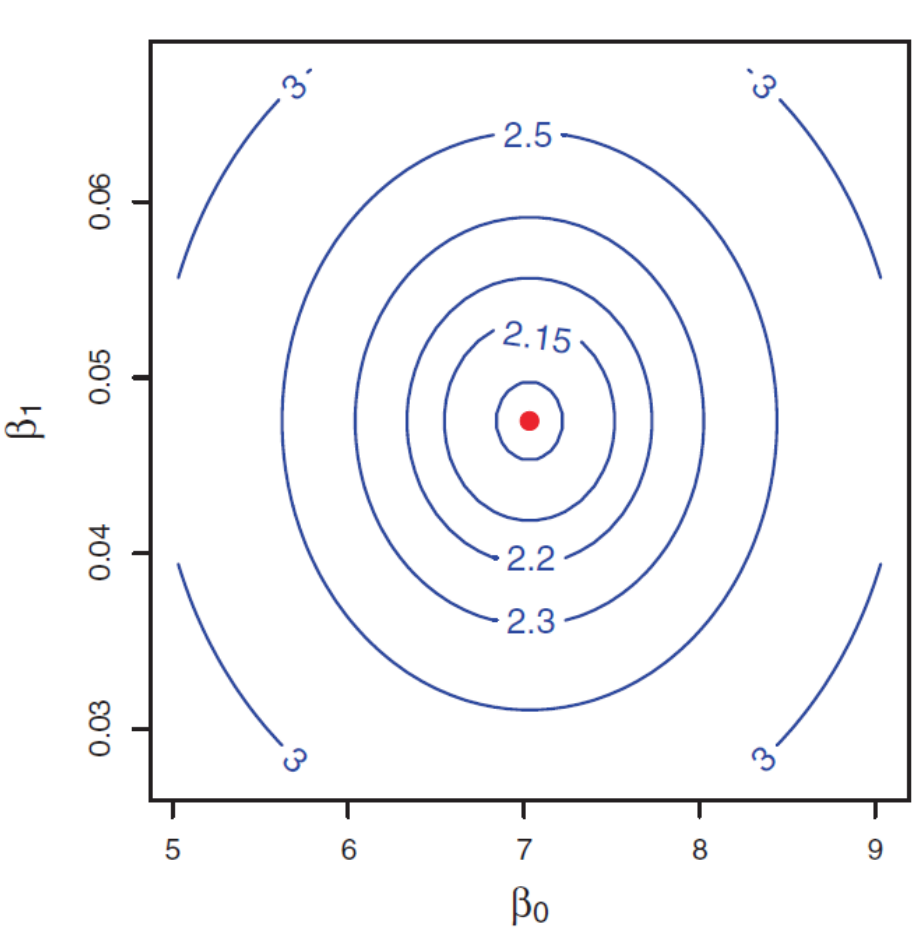
RSS example

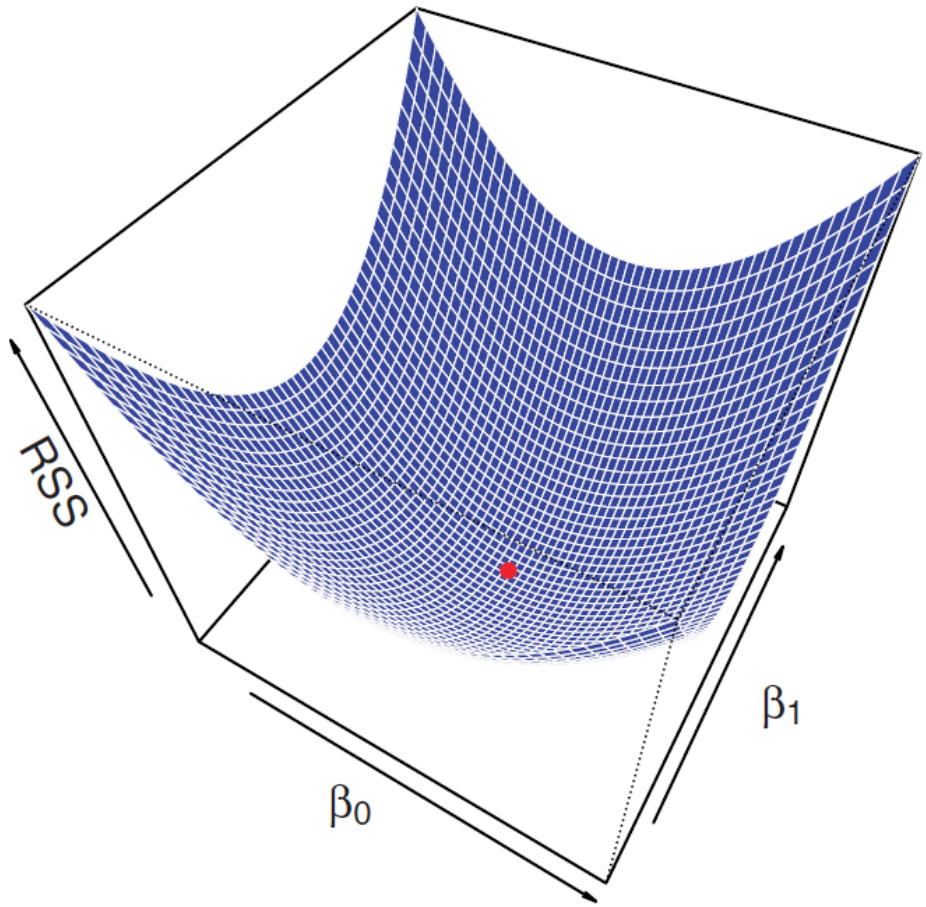


$\hat{\beta}_0 = 7.03$, $\hat{\beta}_1 = 0.0475$

Interpretation: “an additional \$1,000 spent on TV advertising is associated with selling approximately 47.5 additional units of the product”

RSS example





The red dot represents the pair of least squares estimates $(\hat{\beta}_0, \hat{\beta}_1)$

Assessing the accuracy of the coefficient estimates

Assumption: the true relationship between X and Y takes the form

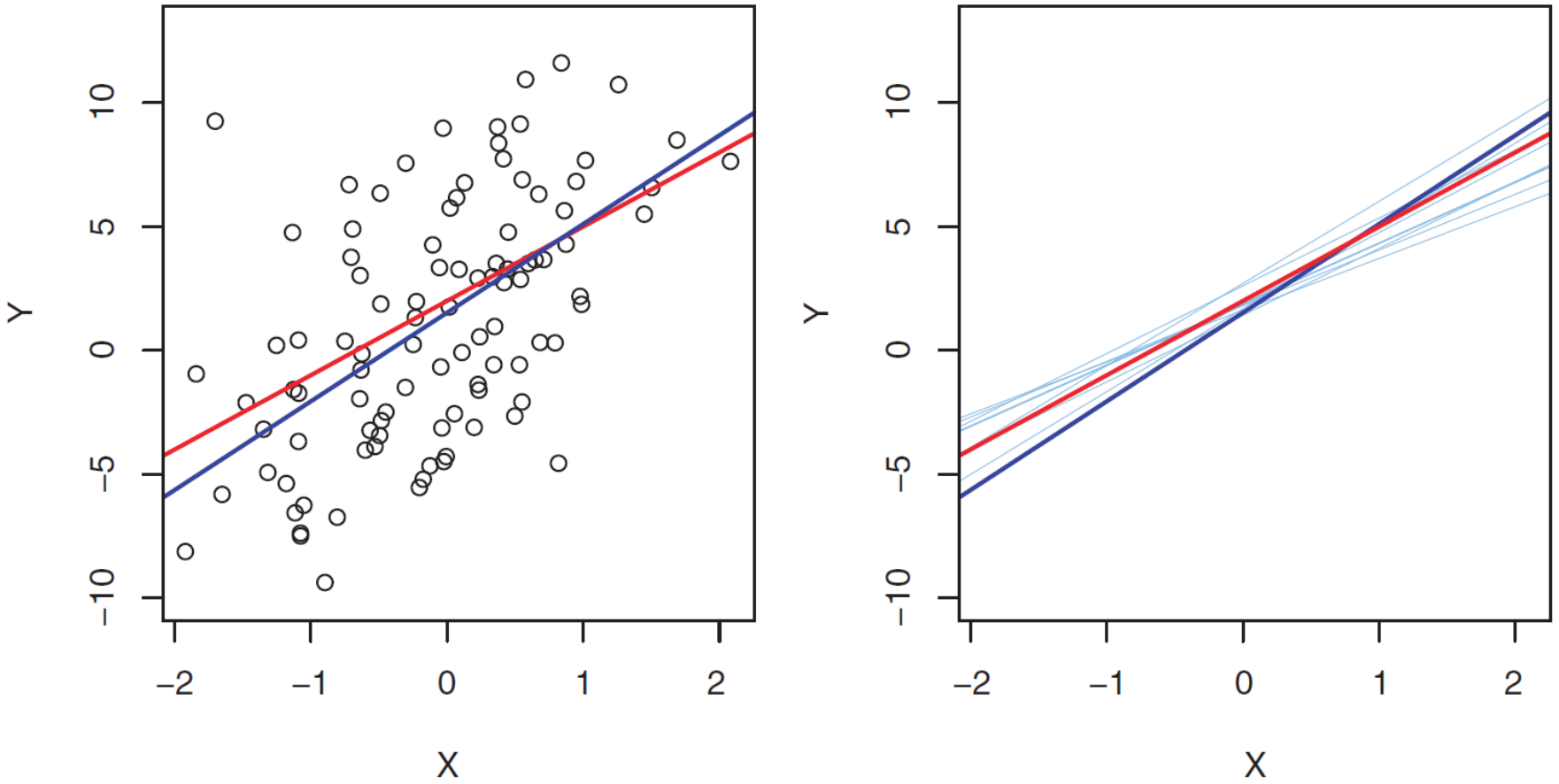
$$Y = f(X) + \epsilon$$

f – some unknown function,
ϵ – a mean-zero random error term (i.e. the true relation is not linear; typical assumption: the error term is independent of X)

Population regression line (best linear approximation to the true relationship between X and Y):

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

Assessing the accuracy of the coefficient estimates



The red line – the population regression line. The blue line – the least squares line.

Bias

Estimate μ using $\hat{\mu}$ (sample mean) $\Rightarrow$ unbiased (on average, we expect $\hat{\mu} = \mu$)

On the basis of one particular set of observations $\hat{\mu}$ might over-/underestimate μ, but if we could average a huge number of estimates of μ obtained from a huge number of sets of observations, then this average would exactly equal μ.

An unbiased estimator does not systematically over- or under-estimate the true parameter.

The property of unbiasedness holds for the least squares coefficient estimates as well!

Residual Standard Error (RSE)

How accurate is the sample mean $\hat{\mu}$ as an estimate of μ?

Answer – standard error of $\hat{\mu}$, $\mathrm{SE}(\hat{\mu})$:

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

σ – the standard deviation of each of the realizations y_i of Y

Roughly speaking: the average amount that the estimate $\hat{\mu}$ differs from the actual value of μ (note: this deviation shrinks with n!)

In general, σ^2 is not known, but can be estimated from the data.
Residual standard error (the estimate of σ):

$$\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$$

Confidence intervals (CI)

A 95% CI – a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

95% CIs for linear regression (approximation):

- for β_1: $\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$
- for β_0: $\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$

Hypothesis tests on the coefficients

The most common hypothesis test – testing the *null hypothesis*:

- H_0 : There is no relationship between X and Y ($H_0 : \beta_1 = 0$)
- H_a : There is some relationship between X and Y ($H_0 : \beta_1 \neq 0$)

Is $\hat{\beta}_1$ sufficiently far from zero that we can be confident that β_1 is non-zero?

How far is far enough?

Hypothesis tests on the coefficients

Depends on the accuracy of $\hat{\beta}_1 \ldots$ (i.e. depends on $\mathrm{SE}(\hat{\beta}_1)$):

- $\mathrm{SE}(\hat{\beta}_1)$ is small? – even relatively small values of $\hat{\beta}_1$ may provide strong evidence that $\beta_1 \neq 0$ (i.e. that there is a relationship between X and Y)
- $\mathrm{SE}(\hat{\beta}_1)$ is large? – $\hat{\beta}_1$ must be large in absolute value in order for us to reject the null hypothesis

In practice – compute a t-statistic:

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

(measures the number of standard deviations that $\hat{\beta}_1$ is away from 0)

If there really is no relationship between X and Y , then we expect that t will have a t-distribution with $n - 2$ degrees of freedom.

Hypothesis tests on the coefficients

p-value: the probability of observing any value $y : y \geq |t|$ (assuming $\beta_1 = 0$)
Interpretation: a small p-value – "is unlikely to observe such a substantial association between the predictor and the response due to chance" $\approx$ there is an association

Reject the null hypothesis if the p-value is small enough (typically: 5 or 1%)

For the `Advertising` data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget.

	Coeff.	Std. err.	t-stat.	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Assessing the accuracy of the model

How to quantify the extent to which the model fits the data?

- residual standard error (RSE) – the average amount that the response will deviate from the true regression line
- R^2 statistic – the proportion of variance explained

Residual Standard Error (RSE)

An estimate of the standard deviation of ϵ (the average amount that the response will deviate from the true regression line).

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}}$$

RSE is small if the predictions obtained using the model are very close to the true outcome values.

Problem: RSE provides an absolute measure of lack of fit of the model to the data (measured in the units of Y). **What constitutes a good RSE?**

R^2 statistic

R^2 statistic – the proportion of variance explained:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} \in [0, 1]$$

($\text{TSS} = \sum(y_i - \bar{y})^2$ – total sum of squares; measures the total variance in the response Y)

- TSS – the amount of variability inherent in the response before the regression is performed
- RSS – the amount of variability left unexplained after performing the regression
- $R^2 \approx 1$ – a large proportion of the variability in the response has been explained by the regression
- $R^2 \approx 0$ – the regression didn't explain much of the variability in the response (our linear model is wrong? inherent error σ^2 is high? maybe both?)

Multiple linear regression (MLR)

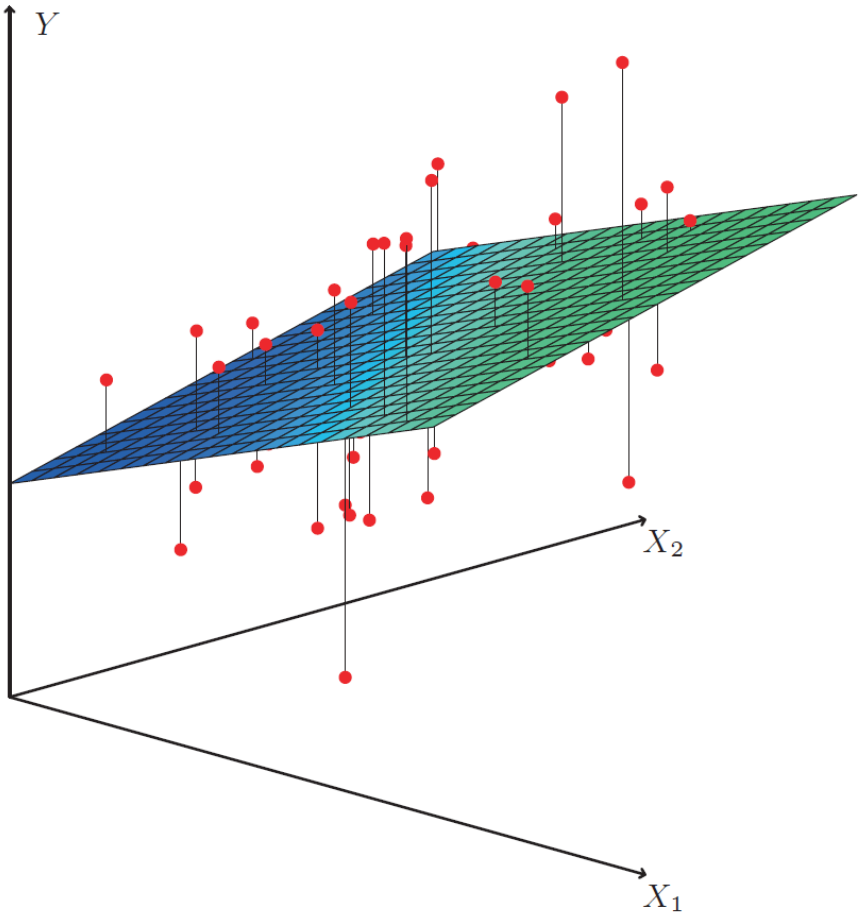
How can we extend our analysis of the advertising data in order to accommodate additional predictors (eg. the radio and newspapers)?

Extend the simple linear regression model (for p predictors):

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

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

Estimating the regression coefficients



In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.

Estimating the regression coefficients

For the `Advertising` data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.

	Coeff.	Std. err.	t-stat.	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

Correlation matrix for the `Advertising` data.

	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

Some important questions (when performing MLR)

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

Is there a relationship between the response and predictors?

Is there a relationship between the response and predictors?

- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$
- H_a : at least one $\beta_j \neq 0$

This hypothesis test is performed by computing the F-statistic:

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

- no relationship between the response and predictors? – $F \approx 1$
- H_a is true? – $F > 1$

Is there a relationship between the response and predictors?

More information about the least squares model for the regression of number of units sold on TV, newspaper, and radio advertising budgets in the `Advertising` data.

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

How large the F-statistic needs to be before we can reject H_0 in favor of H_a?

Is there a relationship between the response and predictors?

Depends on the values of n and p:

- n is large? – even if F-statistic that is just a little larger than 1 might still provide evidence against H_0
- n is small? – need a larger F-statistic to reject H_0

p-value – the partial effect of adding that variable to the model (information about whether each individual predictor is related to the response, after adjusting for the other predictors)

The approach of using an F-statistic to test for any association between the predictors and the response works when p is relatively small, and certainly small compared to n.

Deciding on important variables (variable selection)

Consider models built from different subsets of variables, evaluate using various statistics:

- Mallow's Cp
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- adjusted R^2
- ...

Evaluate all possible models? Not feasible for large p (# models = 2^p!)

Three classical approaches for variable selection:

- forward selection
- backward selection
- mixed selection

Model fit

R^2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the response.

Rationale: Adding another variable to the least squares equations must allow us to fit the training data (though not necessarily the testing data) more accurately $\Rightarrow$ R^2 statistic (which is also computed on the training data) must increase

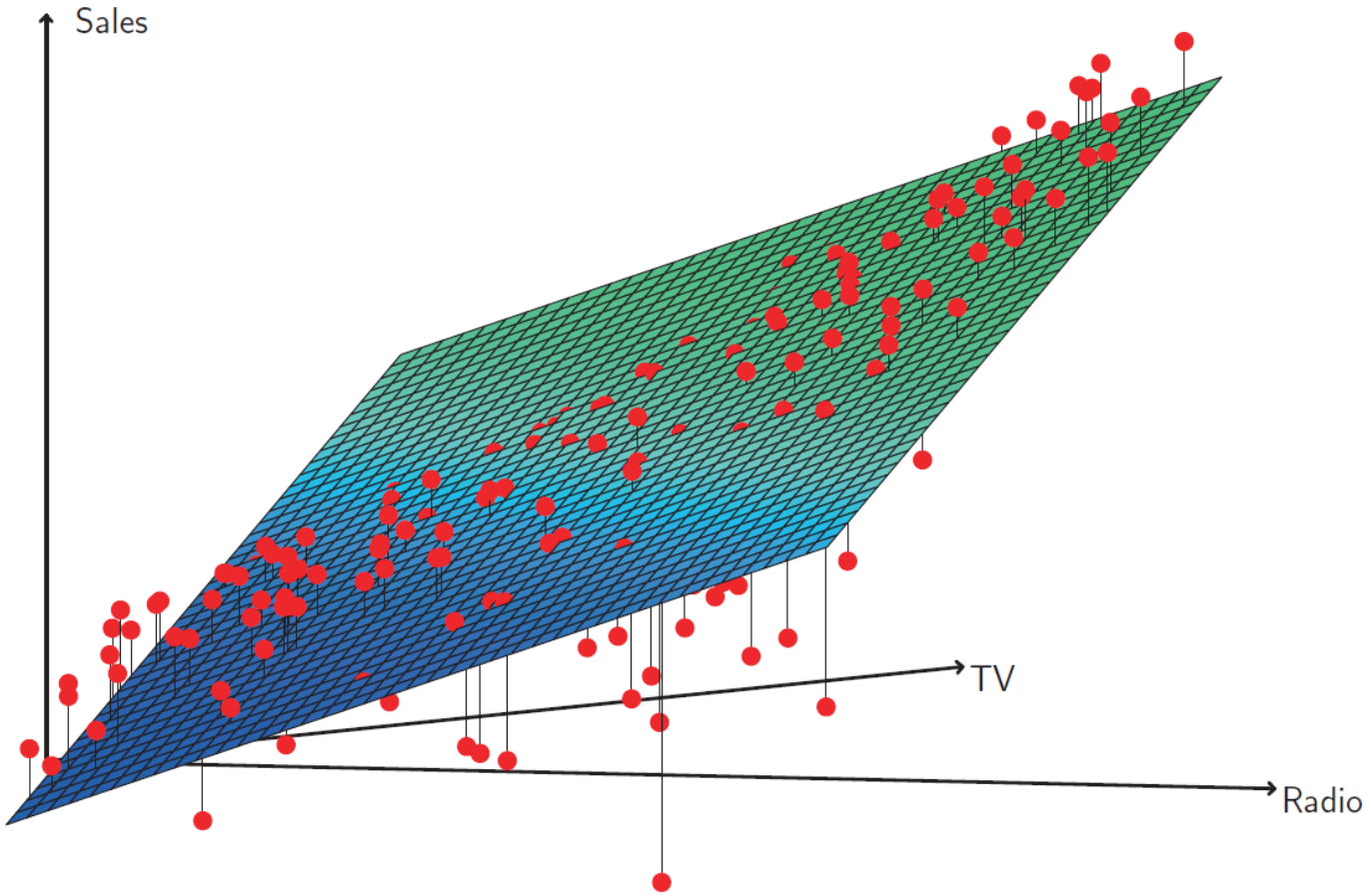
How RSE can increase when a new predictor is added to the model given that RSS must decrease?! In general, RSE is defined as:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}}$$

(simplifies for a simple linear regression)

Models with more variables can have higher RSE if the decrease in RSS is small relative to the increase in p.

Model fit



For the `Advertising` data, a linear regression fit to sales using TV and radio as predictors. From the pattern of the residuals, we can see that there is a pronounced non-linear relationship in the data.

Predictions

Even if we knew $f(X)$ (i.e. the true values for all β_i) the response value cannot be predicted perfectly due to the random error ϵ in the model (i.e. the irreducible error).

How much will Y vary from $\hat{Y}$?

Answer: Use prediction intervals!

Prediction intervals

Prediction intervals (PIs) are always wider than confidence intervals, because they incorporate both:

- the error in the estimate for $f(X)$ (the reducible error)
- the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error).

Use a CI to quantify the uncertainty surrounding the *average* y over a large set of observations (interpretation: 95% of intervals of this form will contain the true value of $f(X)$).

Use a PI to quantify the prediction uncertainty surrounding y for a *particular* observation (interpretation: 95% of intervals of this form will contain the true value of Y for this observation).

Note: both intervals are centered at the same value, but PI is substantially wider than CI

Qualitative predictors (with 2 levels)

A qualitative predictor (a *factor*) only has two *levels* (possible values)? – Create an indicator (a *dummy variable*) that takes on two possible numerical values, eg. for gender:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

and use this variable as a predictor in the regression equation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 x_i + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male} \end{cases}$$

Interpretation:

- β_0 – the average credit card balance among males
- $\beta_0 + \beta_1$ – the average credit card balance among females
- β_1 – the average difference in credit card balance between females and males

Qualitative predictors (with 2 levels)

The decision to code females as 1 and males as 0 in is arbitrary, and has no effect on the regression fit, but does alter the interpretation of the coefficients. Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable using a -1/+1 coding scheme:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male} \end{cases}$$

This results in the model

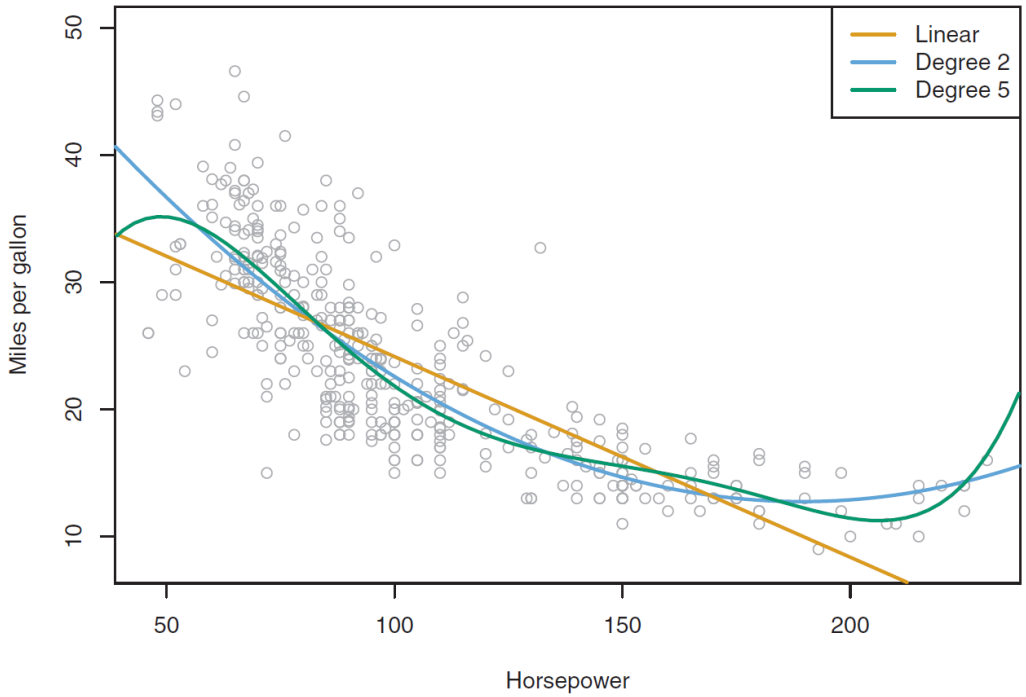
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 x_i + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 - \beta_1 x_i + \epsilon_i & \text{if } i\text{th person is male} \end{cases}$$

Interpretation:

- β_0 – the overall average credit card balance (ignoring the gender effect)
- β_1 – the amount that females are above the average and males are below the average

Non-linear relationships

Try polynomial regression!



The `Auto` data set and different regression fits.

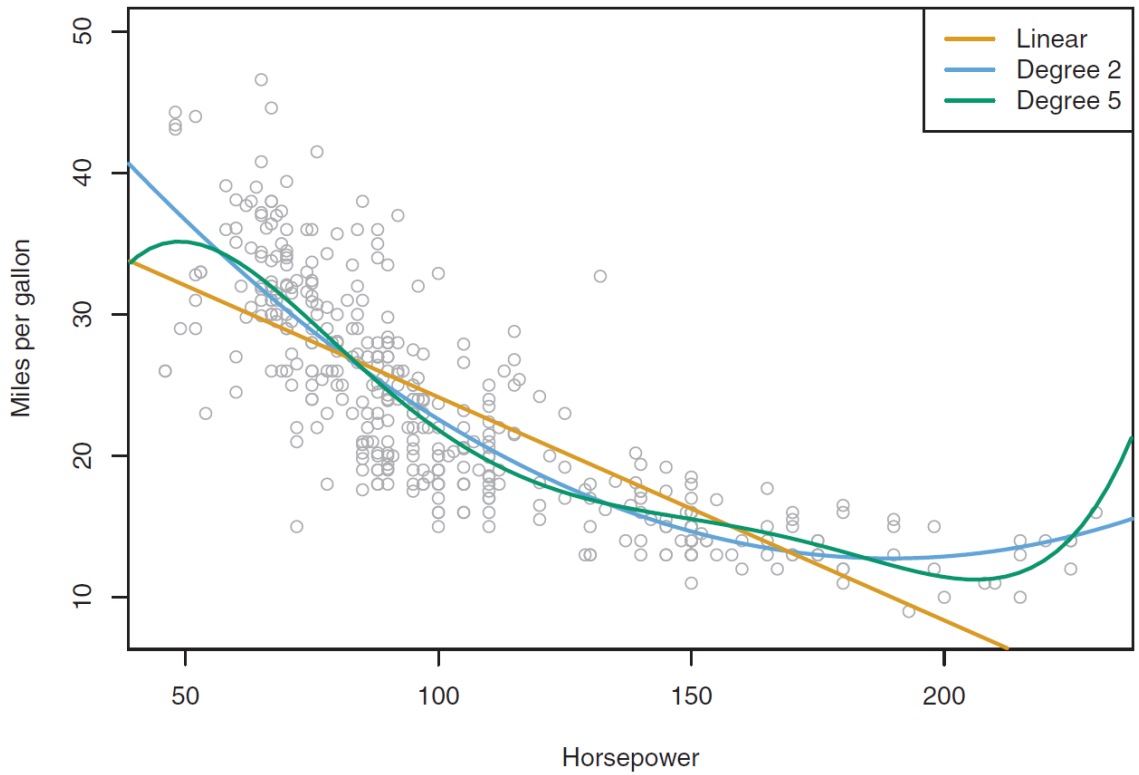
$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \epsilon$ (may provide a better fit?)

Involves predicting mpg using a non-linear function of horsepower, but **it is still a linear model**! – It is simply a multiple linear regression model with $X_1 = \text{horsepower}$ and $X_2 = \text{horsepower}^2$.

Overfitting

R^2 of the linear fit: 0.606
R^2 of the quadratic fit: 0.688 (and p-value for the quadratic term is highly significant)

Why not include horsepower^3, horsepower^4, or even horsepower^5 then?



Overfitting...

Potential problems

Most common problems when fitting linear regression model to a dataset:

- Non-linearity of the response-predictor relationships
- Correlation of error terms
- Non-constant variance of error terms
- Outliers
- High-leverage points
- Collinearity

In practice, identifying and overcoming these problems is as much an art as a science. . .

References

- *An Introduction to Statistical Learning (with Applications in R)*. G. James, D. Witten, T. Hastie, R. Tibshirani (2013)