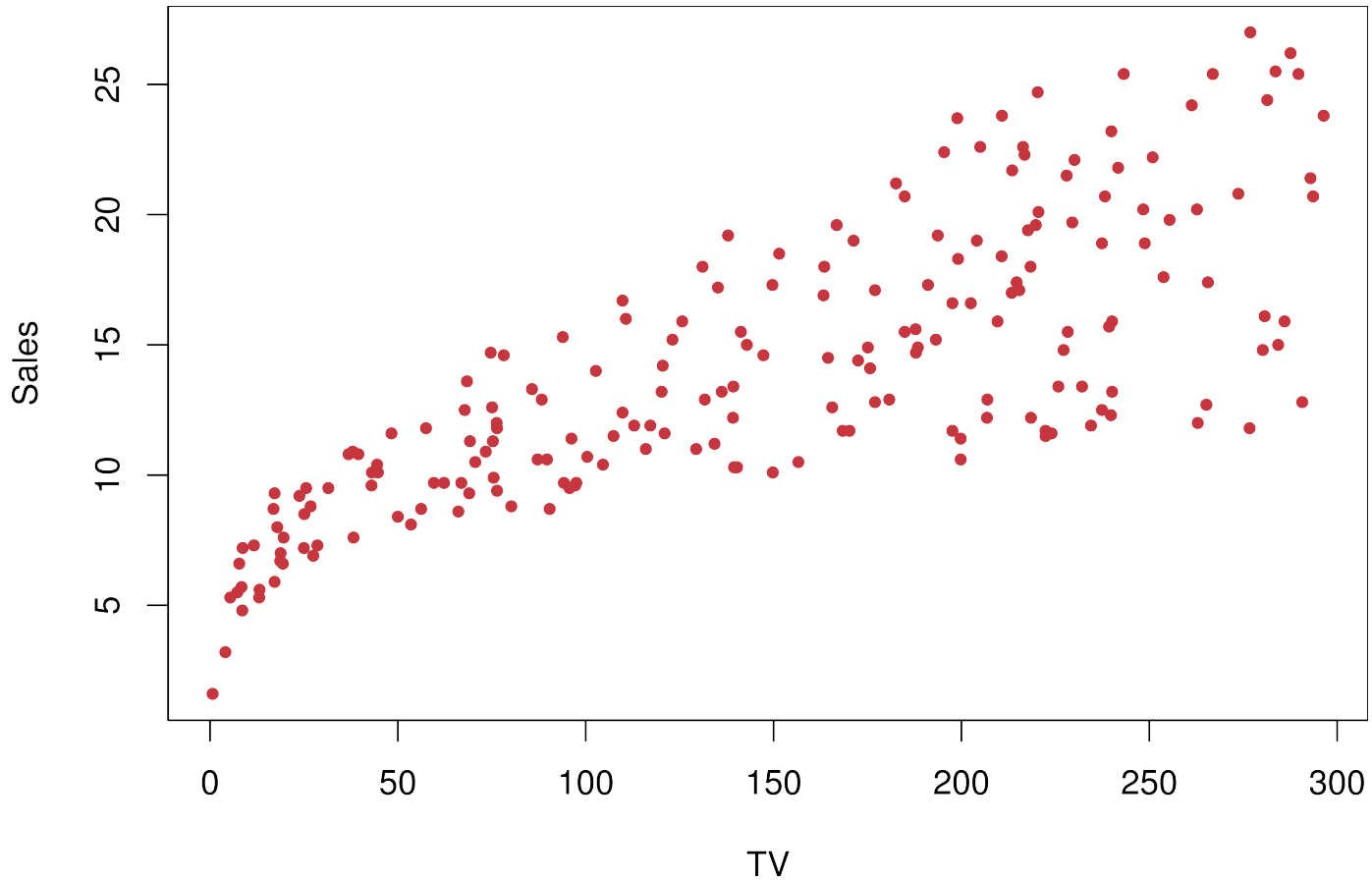
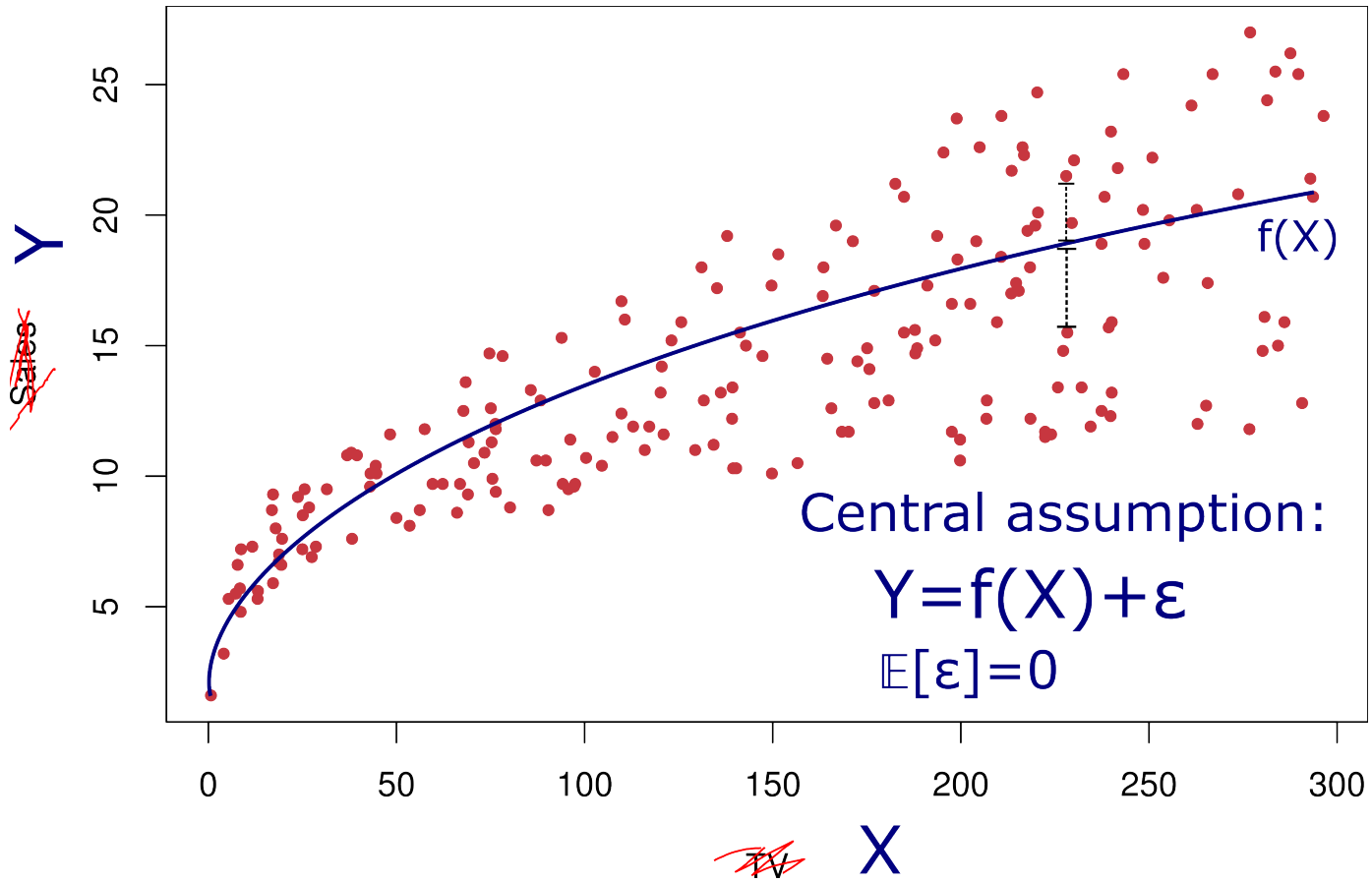
A general learning problem - regression settting



Red dots are past observations for my company.
Question: How much will I sell if I spend $50 on TV advertising? $200? $400?

A general learning problem - regression settting



Strategy: Fit a function $f(X)$ to the data. Input the X to get an estimate output $Y(X)$.

Question: Which $f(X)$ should we choose?

A general learning problem - regression settting

Assume a functional relationship between X and Y as

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

Here the noise ε has mean zero, and is (generally) uncorrelated between observations and with constant variance.

Alternative formulation in terms of the conditional mean of Y given features X:

$$\mathbb{E}(Y \mid X) = f(X)$$

A natural prediction $\hat{Y}$ for a new observation with input X_0 is

$$\hat{Y} = \mathbb{E}(Y \mid X_0) = f(X_0)$$

(We see later in the course a scientific justification)

Linear regression, one feature

In linear regression we take $f(X)$ to be linear:

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Here β_0 is the *intercept* (sometimes called bias) and β_1 is the *slope*. Both are *coefficients* (intercept often excluded).

The noise terms ε are assumed to be

- independent
- Gaussian with mean 0 and
- *constant variance* σ^2 (unknown).

May sound ridiculous but this just might be the most important ML model!

Linear regression - general formulation

More generally with p features and an intercept term the model is

$$Y = X\beta + \varepsilon$$

that is,

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{i1} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The matrix X is called the *design matrix*.

Linear regression - general formulation

More generally with p features and an intercept term the model is

$$Y = X\beta + \varepsilon$$

that is,

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{i1} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The matrix X is called the *design matrix*.

With partner: If $n = p = 1$, what do we get when multiplying this out?

Categorical features

Some variables are not continuous. E.g. `Has kids`, or `Is a student`.

Introduce one *dummy variable* for each level A of feature X:

$$\mathbb{1}_{\{X=A\}} = \begin{cases} 1, & X = A \\ 0, & X \neq A \end{cases}$$

This is referred to also as *one-hot encoding* in ML.

Categorical features

Many ways of parameterising a model with a group-specific mean:

With

$$Y = \beta_0 \mathbb{1}_{\{X=A\}} + \beta_1 \mathbb{1}_{\{X=B\}} + \varepsilon,$$

coefficients are the group means.

With

$$Y = \beta_0 + \beta_1 \mathbb{1}_{\{X=B\}} + \varepsilon,$$

β_0 is the mean in group A, and β_1 the *difference* in group means between groups B and A.

Complex functional relationships

Perhaps we realize that our output, Y is not simply linear in the input variables x_1, x_2 we need a nonlinear term $x_1 x_2$. So...

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

With partner: Can you rename something to make this look exactly like Linear Regression with 3 input variables?

Complex functional relationships

Perhaps we realize that our output, Y is not simply linear in the input variables x_1, x_2 we need a nonlinear term $x_1 x_2$. Clever trick: Use

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

The trick: Introduce transformation of features as new features, e.g. $x_1 x_2 = x_3$ (and less crucially $\beta_{12} = \beta_3$) to obtain,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

The relationship is complex if plotted against the original features, but is linear in the new features.

Interaction terms (final thing)

We call a term like $x_1 x_2$ an interaction term.
The model with an interaction term has a nice interpretation.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Consider what happens between y and x_1 for different fixed values of $x_2 = \alpha$:

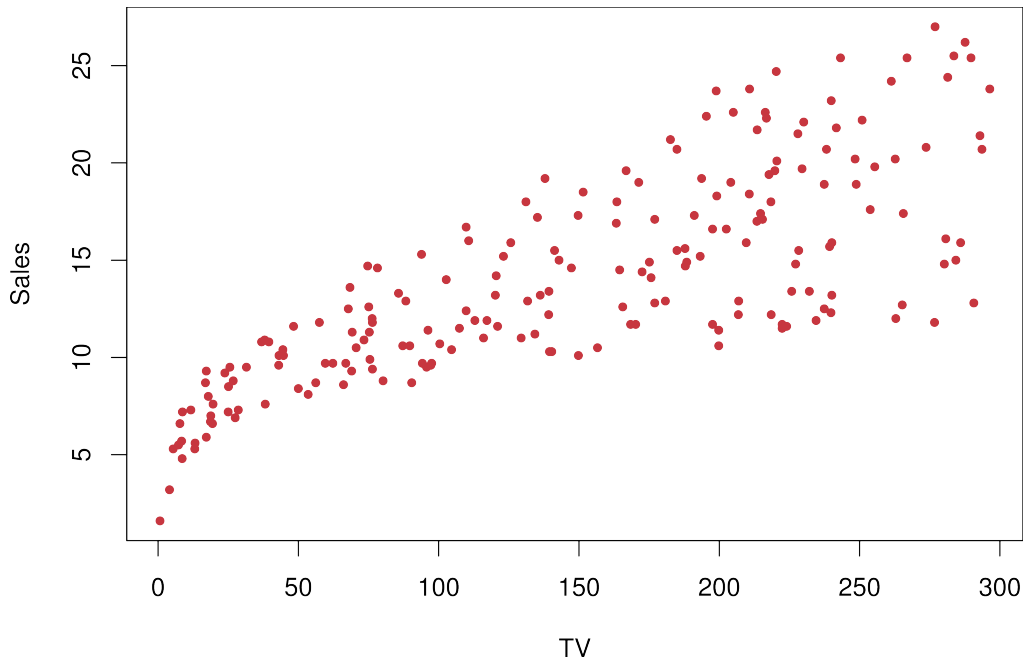
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 \alpha + \beta_3 x_1 \alpha$$

This reduces to

$$Y = (\beta_0 + \beta_2 \alpha) + (\beta_1 + \beta_3 \alpha) x_1$$

A first approach to model building

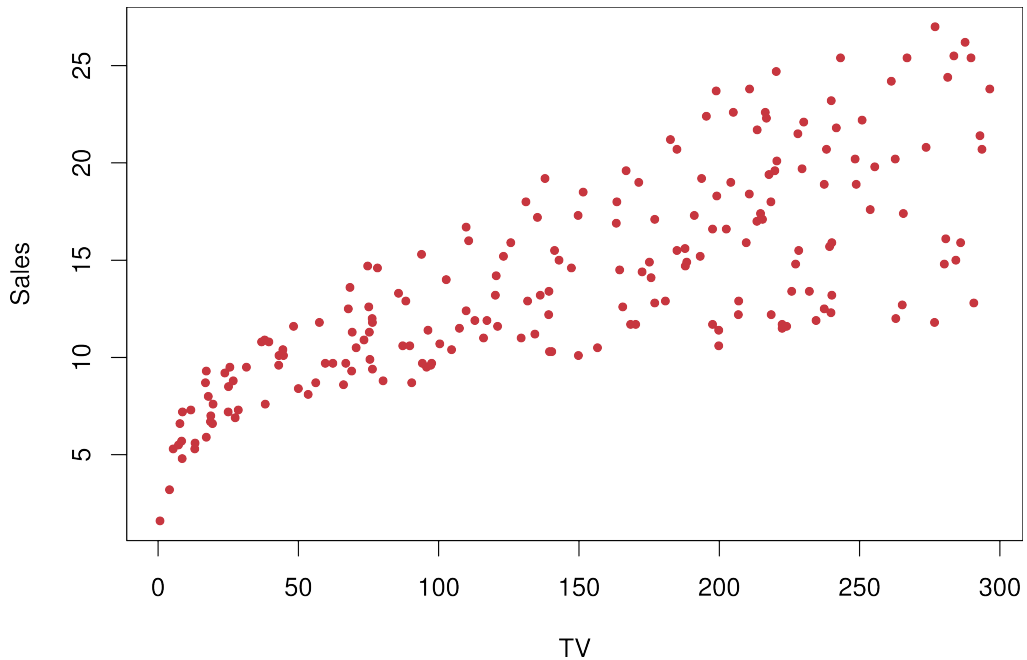
Make some scatter plots against single features.



Curved? Transform the features, add quadratic terms etc.

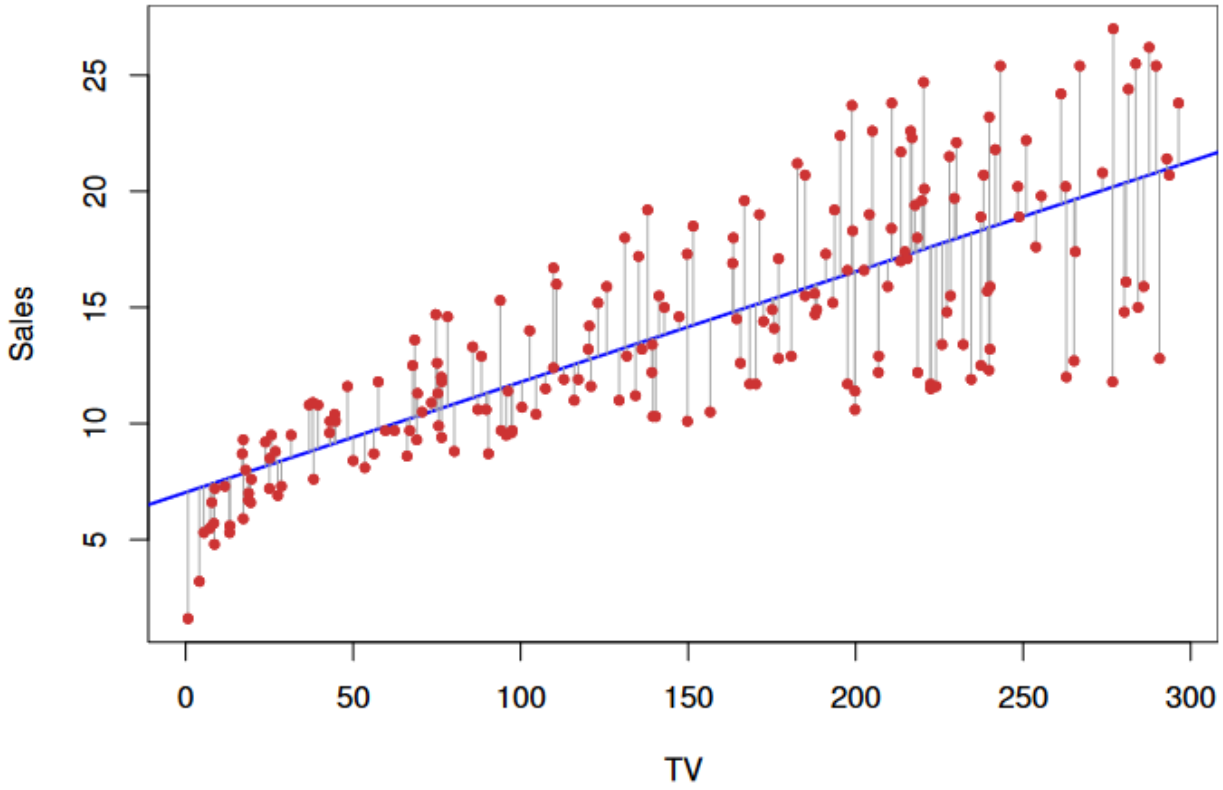
A first approach to model building

Make some scatter plots against single features. (**Always!**)



Curved? Transform the features, add quadratic terms etc.

Estimating coefficients β



(Fig from ISLwR – clearly bad model fit here!)

Ordinary least squares regression finds β that minimises the sum of squared errors - the residual sum of squares:

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

Estimating parameters β and σ^2

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

More generally, we prefer to fit our model to maximize the *model likelihood*: Given the datapoints and the model we assume (a linear model), which model parameters would be the most likely to give us the data?

When we assume **Gaussian** noise, minimizing RSS and using Maximum Likelihood will give the same result.

Estimating parameters β and σ^2

Maximum Likelihood starts with defining a likelihood function

$$p_Y(y_i \mid X) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2}\frac{(y_i - x_i^T\beta)^2}{\sigma^2}}.$$

For a choice of β and σ, the likelihood of the model given all data points would be,

$$\prod_{i=1}^{n} p_Y(y_i \mid X) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2}\frac{(y_i - x_i^T\beta)^2}{\sigma^2}}.$$

We want to find the β and σ that maximize this.. Or rather.. Minimize the negative logarithm of this:

$$\ell(\beta, \sigma^2) = -\frac{1}{2}\left(n \log \sigma^2 + \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T\beta)^2 \right)$$

Strategy: First solve for β, as this optimum is the same for all values of σ^2. Then plug in $\hat{\beta}$ and maximise the profile likelihood for σ^2.

Minimising RSS to get β: Normal equations

$$\text{RSS} = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = (Y - X\beta)^T (Y - X\beta)$$

Differentiate and set equal to zero for each $\beta_r, r = 0, \ldots, p$:

$$\frac{\partial \text{RSS}}{\partial \beta_r} = 2 \sum_{i=1}^{n} x_{ir} (y_i - x_i^T \beta)$$

Collecting all $p + 1$ equations this becomes

$$X^T (Y - X\beta) = 0$$

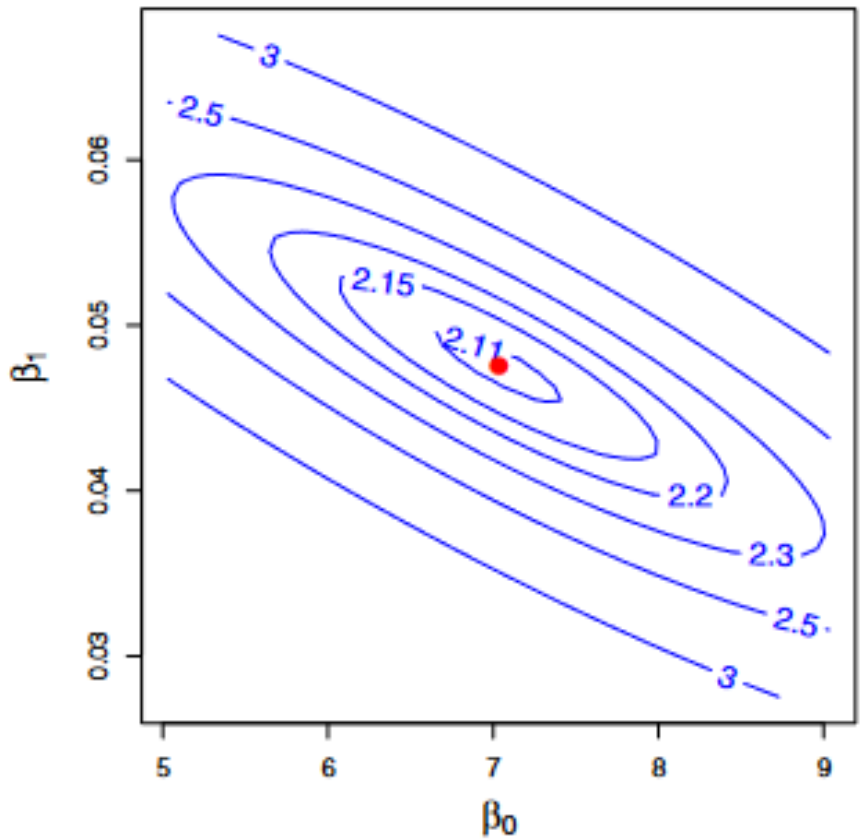
The solution has a nice closed form:

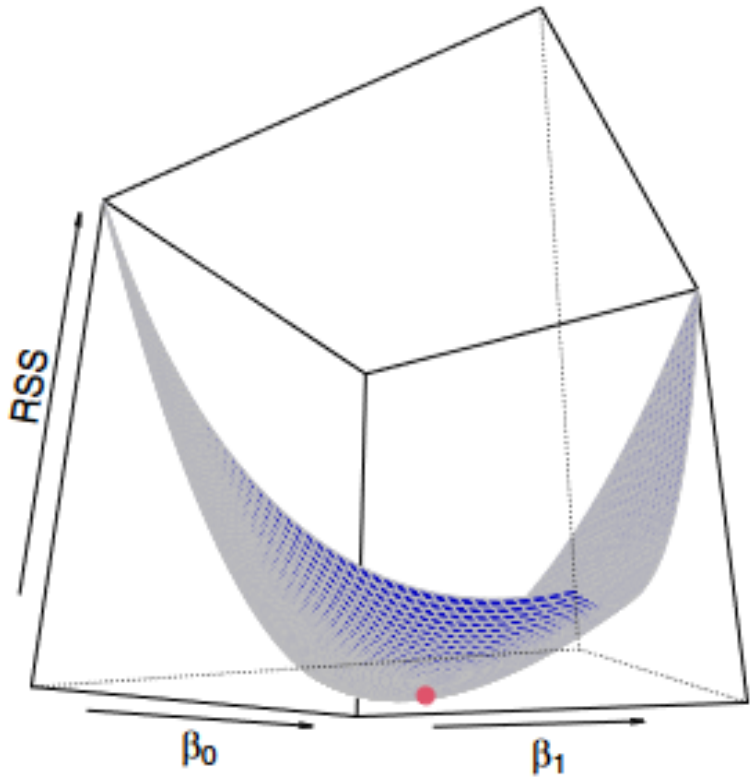
$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Note that we cannot have more parameters than observations (need $p < n$).

The minimum clearly exists

For the Sales-TV data, the RSS as a function of β looks like this:





Estimated residual variance

Fixing β at $\hat{\beta}$ and minimising the log-likelihood in σ^2 gives

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2$$

Usually we normalise instead by $n-p$ for unbiasedness, so the residual variance estimate is taken to be

$$RSS/(n-p).$$

(If n is large and p is small, you will not see the difference)

Likelihoods. Jeez what's the point?

When we choose a model, we need to fit it.

A great way to do this is with a *maximum likelihood estimator* (MLE),

To do this, we define a likelihood function for our model and find the parameters that, under the assumed model, would make the observed data be the most likely.

(This is computationally heavy at first, but Python have great implementations that we can use.)

Making predictions from a linear model

Given a set of features $x_i = (x_{i1}, \ldots, , x_{ip})$, we may now use the model to predict Y_i.

A natural prediction is the mean of Y_i, which is simply the value of the regression line at x_i (more generally, the value of the *linear predictor* at x_i). Replace $\hat{\beta}$ with the expression we found for it using Maximum Likelihood to get:

$$\hat{Y} = X\hat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{H} Y.$$

The matrix H is called *hat matrix* because it "puts a hat on Y" when it transforms Y into $\hat{Y}$. Its diagonal elements h_{ii} are called *leverage* and are important in model checking.

Confidence intervals and Prediction intervals

Interpreting coefficients using confidence intervals. (For Lin Reg, see Eq. (3.8) in ISL)

	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

TABLE 3.1. *For the* `Advertising` *data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of* \$1,000 *in the TV advertising budget is associated with an increase in sales by around 50 units. (Recall that the* `sales` *variable is in thousands of units, and the* `TV` *variable is in thousands of dollars.)*

In the case of the advertising data, the 95 % confidence interval for β_0 is $[6.130, 7.935]$ and the 95 % confidence interval for β_1 is $[0.042, 0.053]$. Therefore, we can conclude that in the absence of any advertising, sales will, on average, fall somewhere between 6,130 and 7,935 units. Furthermore, for each \$1,000 increase in television advertising, there will be an average increase in sales of between 42 and 53 units.

More complex intervals, e.g. for $\hat{f}(x)$, can be made using that $\hat{\beta}$ is approximately (multivariate) Gaussian.

Prediction interval for $\hat{Y}$ (individual values) is always wider than the confidence interval for $\hat{f}(x)$ (an expectation).

Building and scrutinizing models

Building and scrutinizing models

Which features should be in the model?

Check by doing scatter plots (Always! Remember?)

How should features enter the model?

- transformations
- interactions between features

Fit the model.

Does the model fit well?

- Check whether model assumptions are met
- Find out *in what way* the assumptions are unsuitable.

More generally... *scrutinize* the model.

Scrutinize: Test for whether a specific coefficient is zero

A test for $\beta_j = 0$ can be tested by comparing

$$\frac{\beta_j - 0}{\hat{\text{SE}}(\beta_j)}$$

to a t-distribution with $n - 2$ degrees of freedom.

	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

TABLE 3.1. *For the* `Advertising` *data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of* \$1,000 *in the TV advertising budget is associated with an increase in sales by around 50 units. (Recall that the* `sales` *variable is in thousands of units, and the* `TV` *variable is in thousands of dollars.)*

Model output always gives the estimated SE for coefficients, $\hat{\text{SE}}(\beta_j)$.
Alternatively... Use F-statistic ($F \gg 1$ if there is a relationship between response and predictor).

Test for whether several coefficients are zero

Example:

$$M_1 : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon$$
$$M_0 : Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Model M_0 is a special case of M_1 where $\beta_2 = 0$ and $\beta_3 = 0$.

The F-test statistic measures how much the RSS changes when we use the simpler model instead of the more complex one.

$$F = \frac{(\mathrm{RSS}_{M_0} - \mathrm{RSS}_{M_1})/q}{\mathrm{RSS}_{M_1}/(n-p-1)} = \left(\frac{\mathrm{RSS}_{M_0}}{\mathrm{RSS}_{M_1}} - 1\right) / \frac{q}{(n-p-1)}$$

Where q is the number of parameters dropped in reducing M_1 to M_0. Higher F if more reduction in RSS for M_1 or/and q is small compared to $n-p-1$.

Test for whether *all* coefficients are zero

A special case of the F-test indicates whether the set of explanatory variables included is useful at all in explaining the response.

$$M_1 : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon$$
$$M_0 : Y = \beta_0 + \varepsilon$$

Hypothesis testing

Two nested models can be compared by F-tests.

Any set of models can be compared by an information criterion, e.g. AIC or BIC.

Strategies for model selection:

- Forward selection (start by including no/few variables)
- Backward (start by including all/many variables)
- Alternating between forwards and backward selection

Don't test too much, as then you run into issues with multiple testing.

Use model inspection to see if you have a well-fitting model.

Residuals

Inspecting residuals is useful for checking whether there are hints that the model is wrong and, if so, in which ways.

The (raw) *residuals* are the estimated errors

$$e_i = y_i - \hat{y}_i = y_i - x_i^T \beta$$

We can also look at *standardized* residuals that have been scaled by their standard error - they are all approximately standard normal.

$$\frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}$$

Residuals against fitted values or explanatory variables

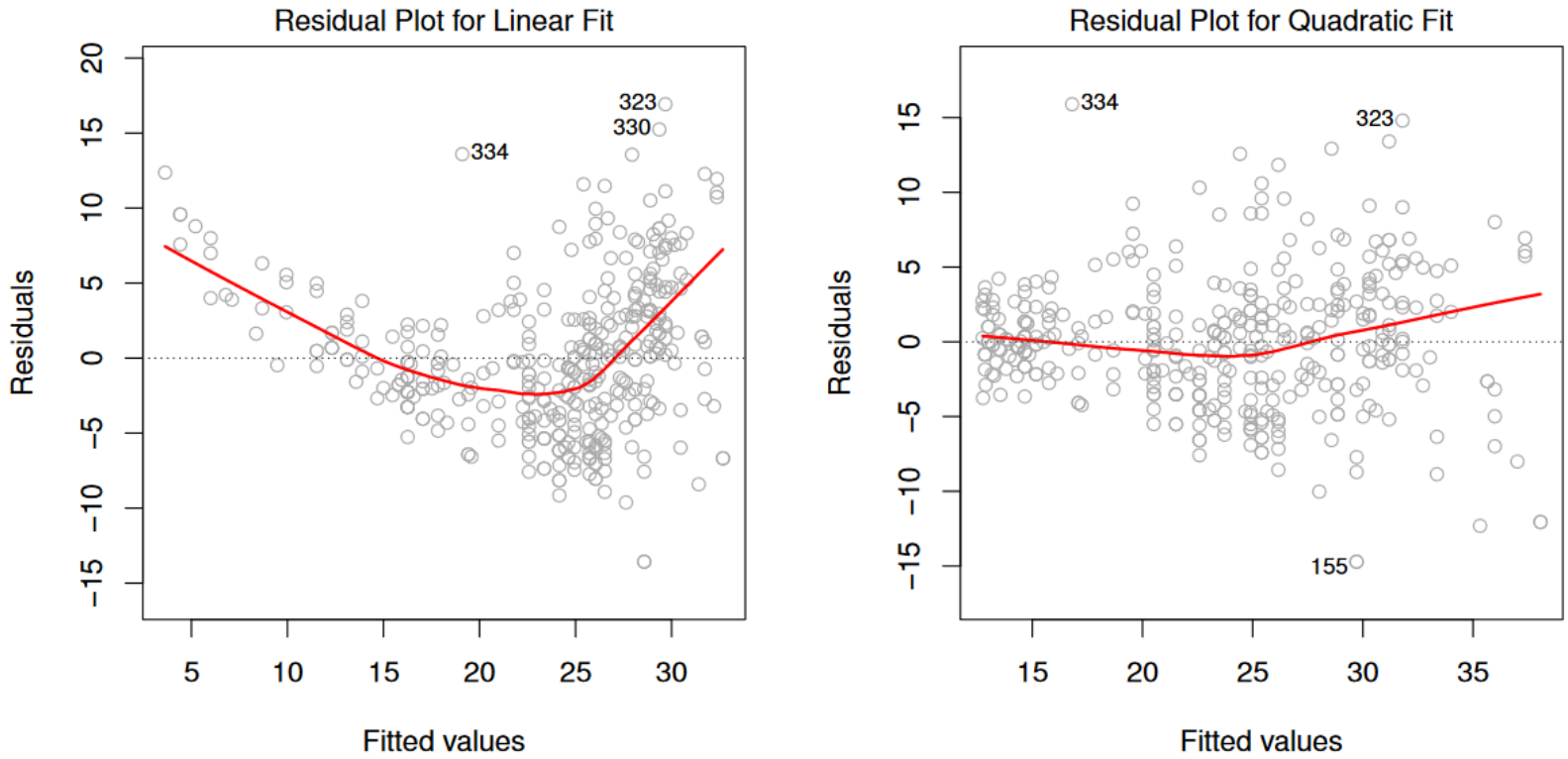
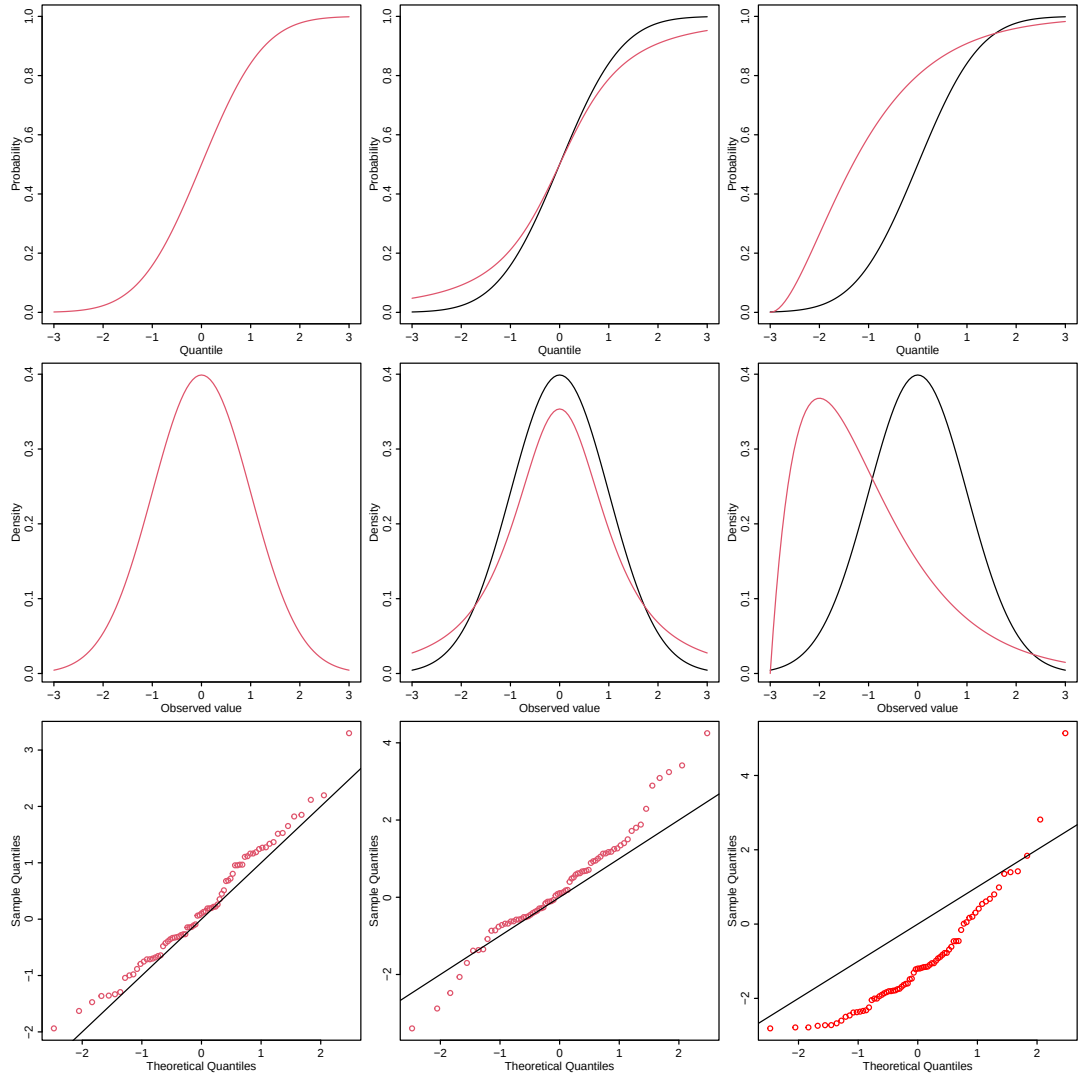


FIGURE 3.9. *Plots of residuals versus predicted (or fitted) values for the* `Auto`

- Is there any sign of curvature looking at the mean? (The mean of residuals should be constant zero)
- Is there any sign of unequal variance?

Quantile-quantile plots of standardized residuals



Check whether residuals are indeed approximately standard normal - if so, the plot resembles a straight line.

Influential points

Be aware when an observation has a combination of high residual and high leverage (potential influence on the regression fit).