Linear Regression

Irina Mohorianu
CSCI

09/02/2021

Section 1

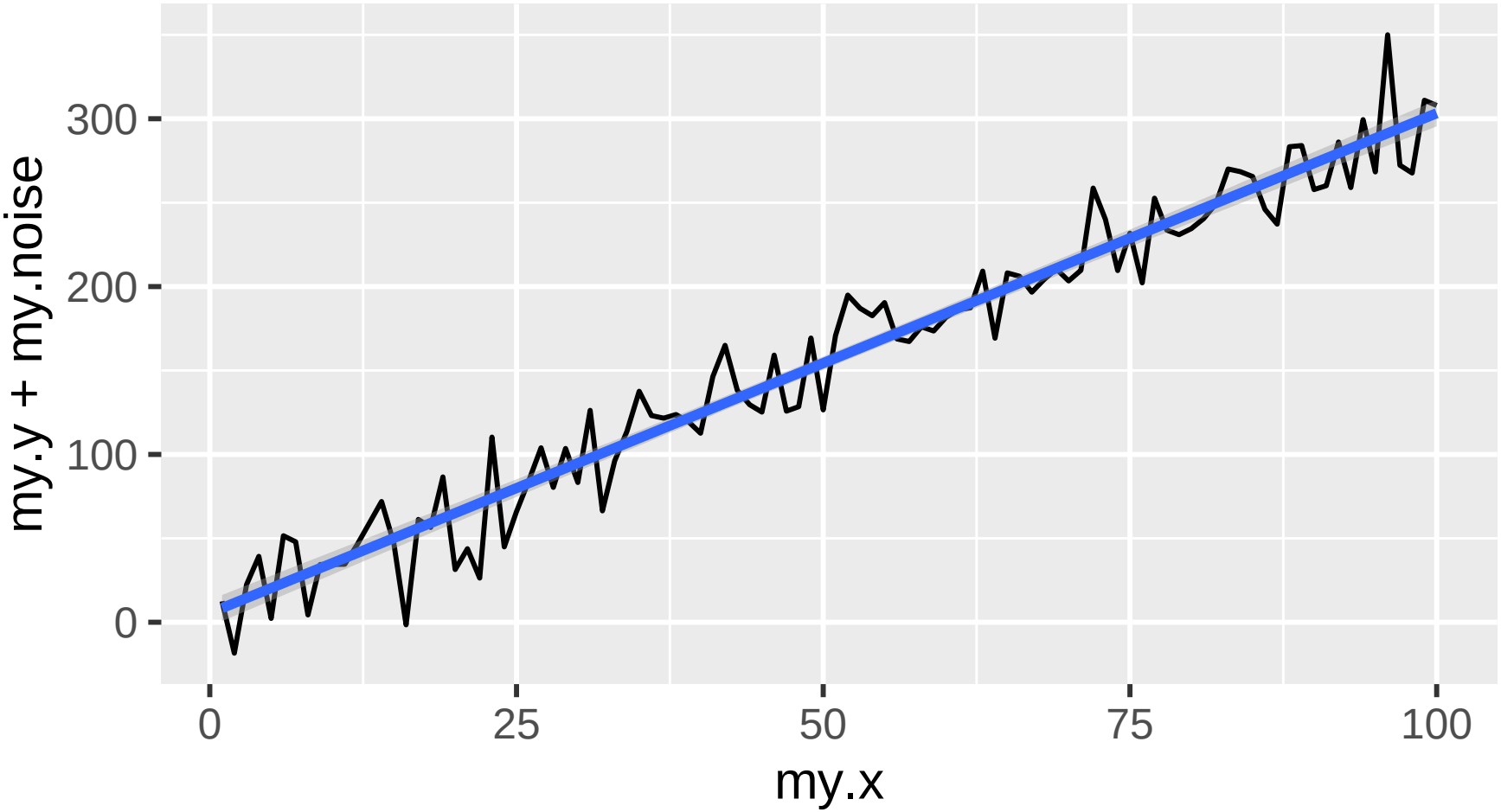
Linear Regression. Definitions.

Linear Regression. Definition

linear regression = a very simple approach for supervised learning.
We assume a linear dependence of Y on $X_1, X_2, ..., X_p$.

Linear Regression. Definition

linear regression = a very simple approach for supervised learning.
We assume a linear dependence of Y on $X_1, X_2, ..., X_p$.



Linear Regression. Definition

We assume the model

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

where β_0 and β_1 are two unknown constants that represent the *intercept* and the *slope*; β_0 and β_1 are also known as *coefficients* or *parameters* and ϵ is the error term.

Linear Regression. Definition

We assume the model

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

where β_0 and β_1 are two unknown constants that represent the *intercept* and the *slope*; β_0 and β_1 are also known as *coefficients* or *parameters* and ϵ is the error term.

Given the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict the output, $\hat{y}$ using

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

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

Advertising data

Advertising data:

```
head(my.advertise)
```

```
##      TV Radio Newspaper Sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 151.5  41.3      58.5  18.5
## 5 180.8  10.8      58.4  12.9
## 6   8.7  48.9      75.0   7.2
```

```
summary(my.advertise)
```

```
##        TV              Radio          Newspaper          Sales      
##  Min.   :  0.70   Min.   : 0.000   Min.   :  0.30   Min.   : 1.60  
##  1st Qu.: 74.38   1st Qu.: 9.975   1st Qu.: 12.75   1st Qu.:10.38  
##  Median :149.75   Median :22.900   Median : 25.75   Median :12.90  
##  Mean   :147.04   Mean   :23.264   Mean   : 30.55   Mean   :14.02  
##  3rd Qu.:218.82   3rd Qu.:36.525   3rd Qu.: 45.10   3rd Qu.:17.40  
##  Max.   :296.40   Max.   :49.600   Max.   :114.00   Max.   :27.00
```

Why is Linear Regression useful?

In this lecture, we review some of the key ideas underlying the linear regression model, as well as the least squares approach that is most commonly used to fit this model

1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media contribute to sales?

Why is Linear Regression useful?

In this lecture, we review some of the key ideas underlying the linear regression model, as well as the least squares approach that is most commonly used to fit this model

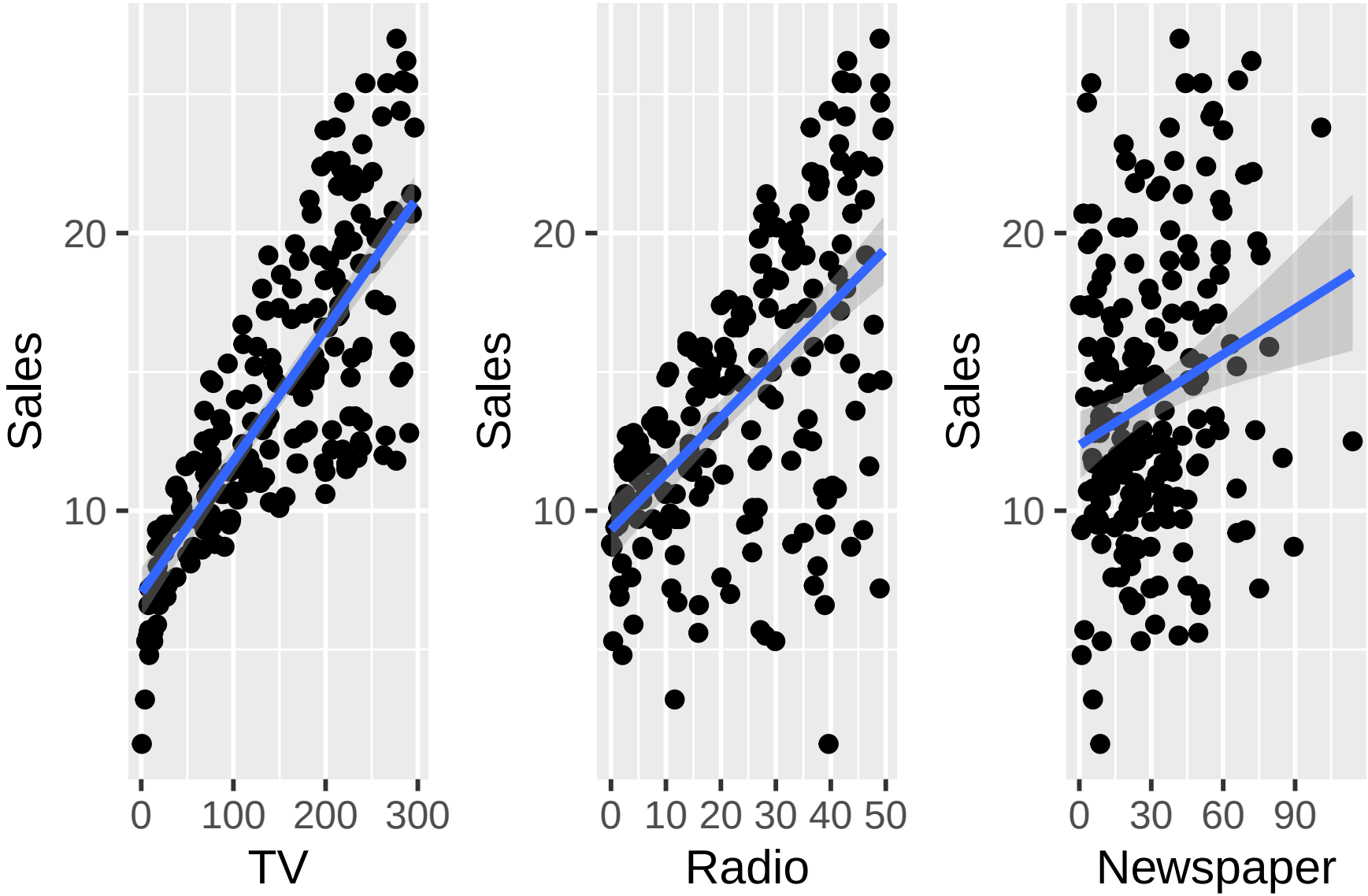
1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media contribute to sales?
4. How accurately can we estimate the effect of each medium on sales?
5. How accurately can we predict future sales?
6. Is the relationship linear?

Why is Linear Regression useful?

In this lecture, we review some of the key ideas underlying the linear regression model, as well as the least squares approach that is most commonly used to fit this model

1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media contribute to sales?
4. How accurately can we estimate the effect of each medium on sales?
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there synergy among the advertising media?

Advertising data 2



Advertising data 3

More formally:

$$sales \approx \beta_0 + \beta_1 \times TV + \epsilon$$

$$sales \approx \beta_0 + \beta_1 \times Radio + \epsilon$$

$$sales \approx \beta_0 + \beta_1 \times Newspaper + \epsilon$$

Advertising data 3

More formally:

$$sales \approx \beta_0 + \beta_1 \times TV + \epsilon$$

$$sales \approx \beta_0 + \beta_1 \times Radio + \epsilon$$

$$sales \approx \beta_0 + \beta_1 \times Newspaper + \epsilon$$

or with all predictors into one model:

$$sales \approx \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times Newspaper + \epsilon$$

Section 2

Coefficient estimates. Accuracies.

Estimation of coefficients using 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 $\epsilon_i = y_i - \hat{y}_i$; ϵ_i is the ith residual.

Estimation of coefficients using 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 $\epsilon_i = y_i - \hat{y}_i$; ϵ_i is the ith residual.

We define the *residual sum of squares (RSS)* as

$$RSS = \epsilon_1^2 + \epsilon_2^2 + ... + \epsilon_n^2$$

which is equivalent to

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + ... + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Estimation of coefficients using Least Squares

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimise the RSS. The minimising values are:

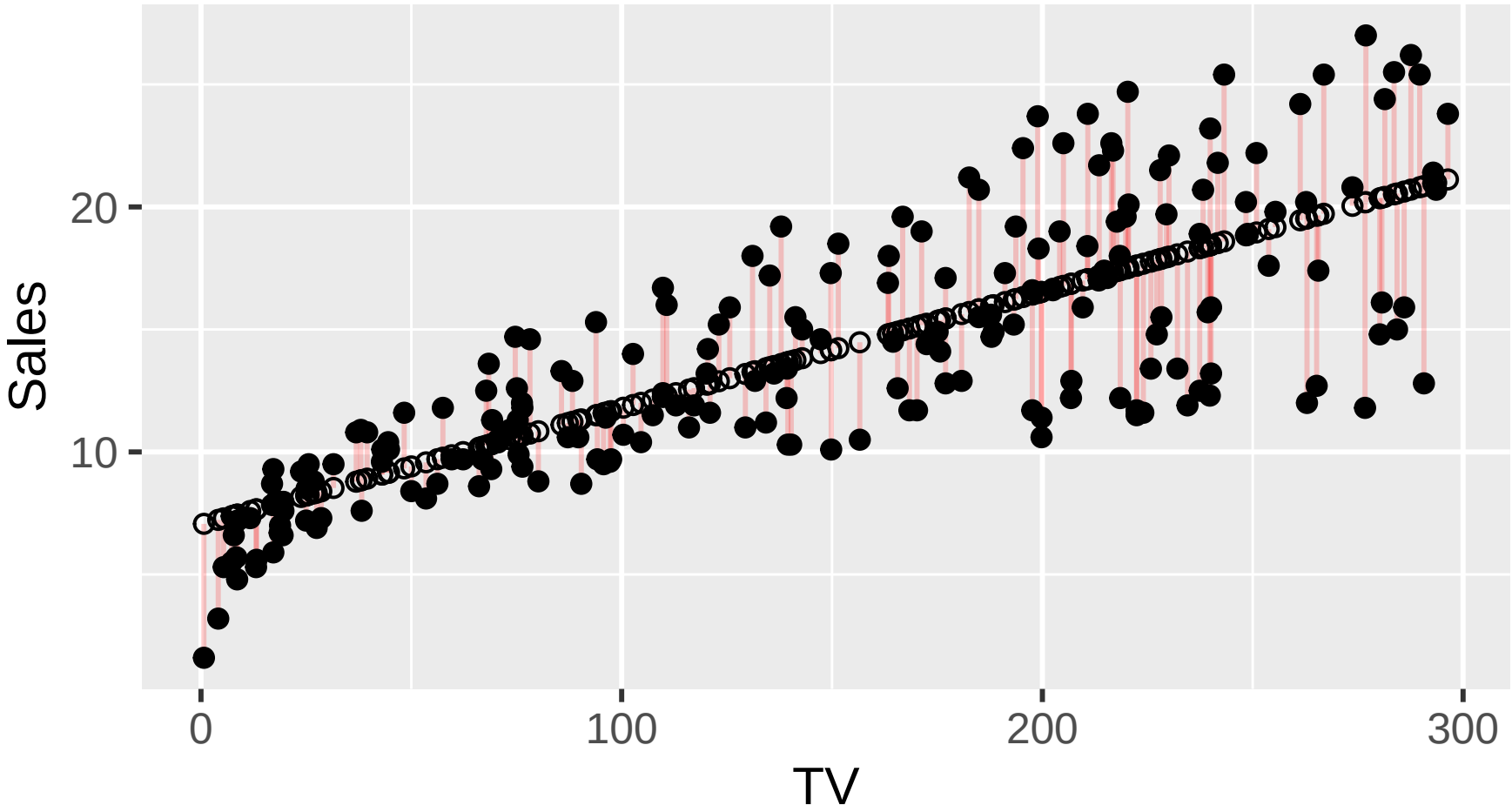
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_1 \bar{x}$$

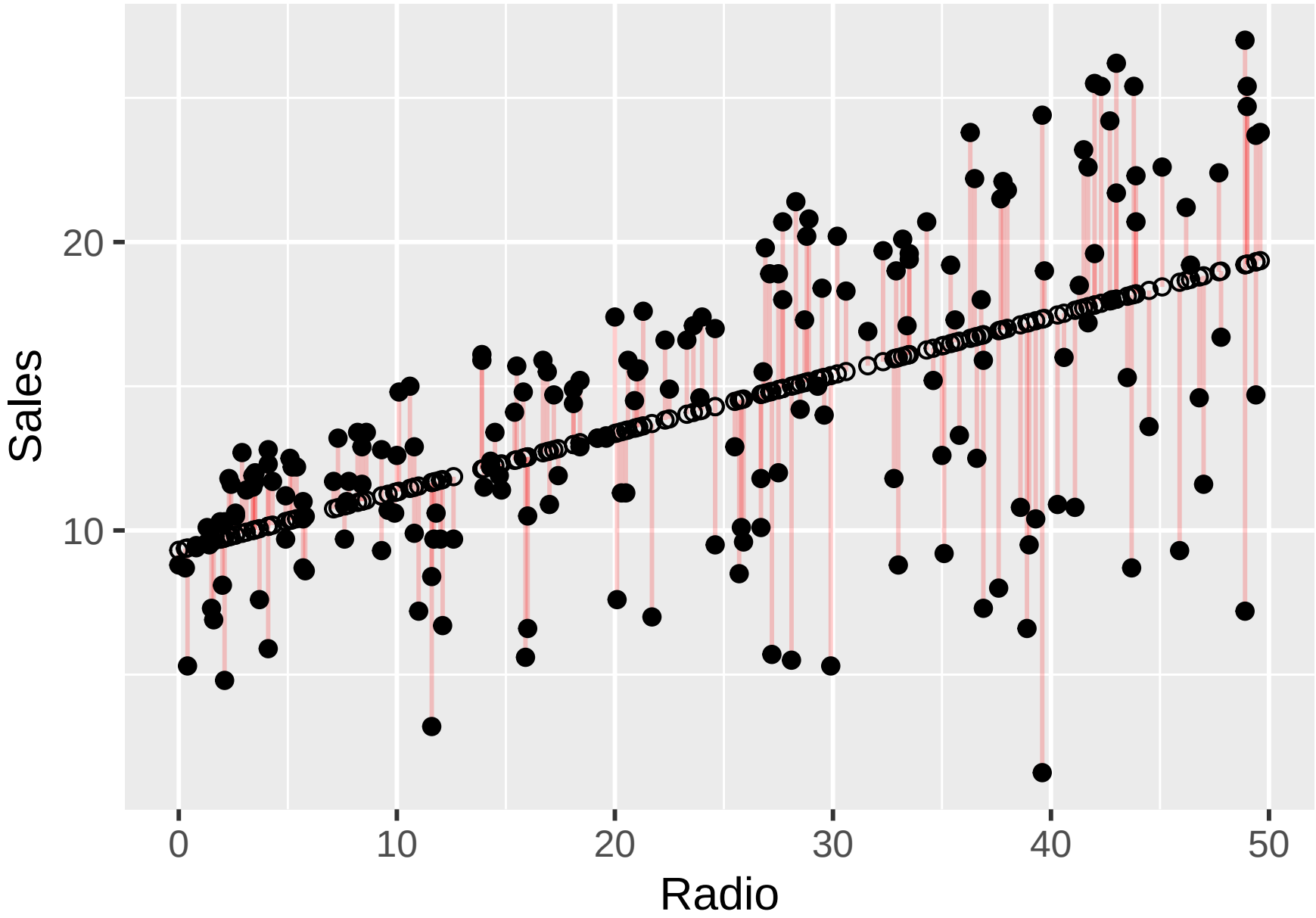
where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Estimation of coefficients using Least Squares

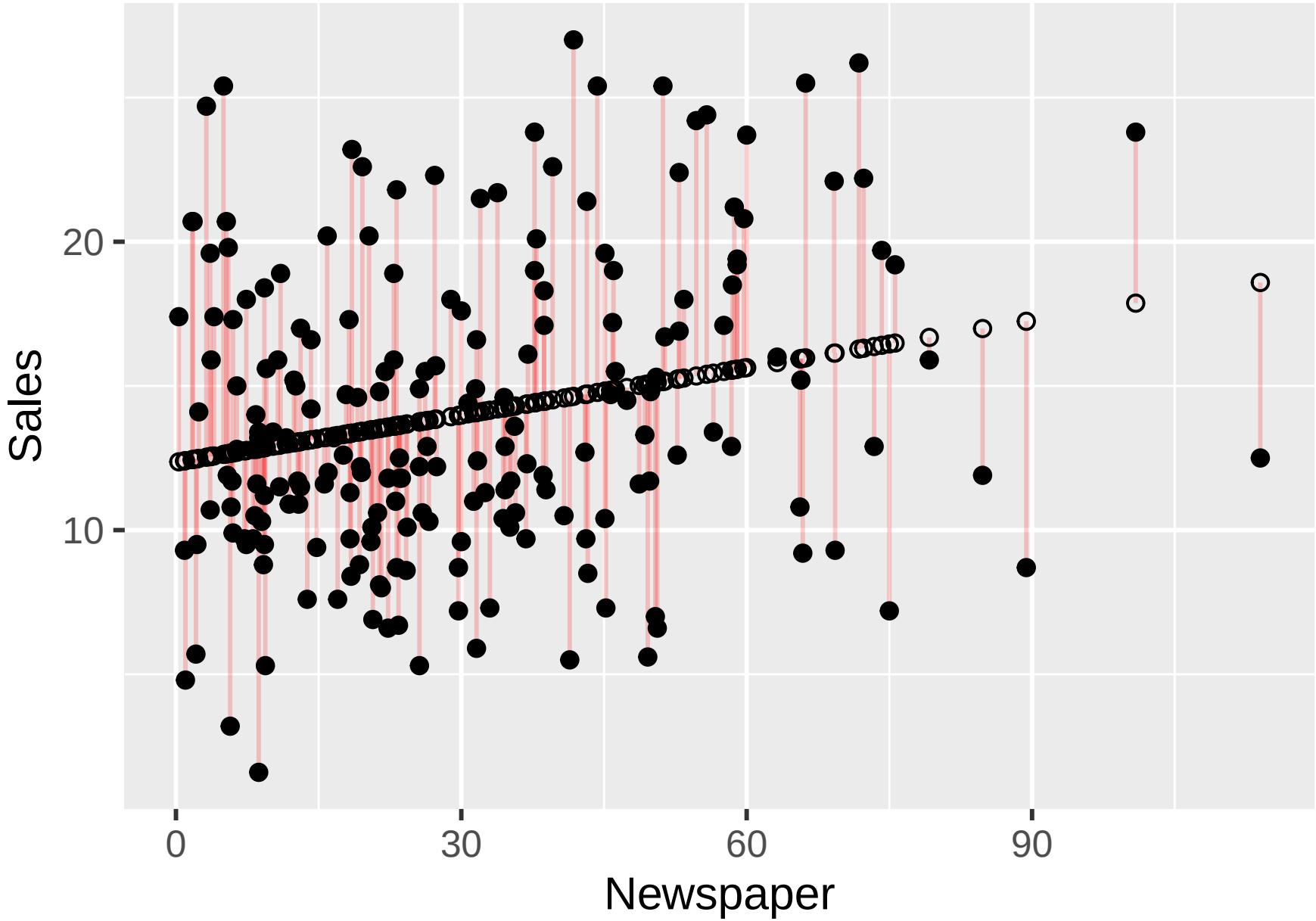


The least squares fit for the regression of Sales onto TV. For this predictor, the linear fit captures the essence of the relationship; however the errors are large for large values.

Estimation of coefficients using Least Squares



Estimation of coefficients using Least Squares



Section 3

Assessing errors.

Assessing the Accuracy of the Coefficient Estimates.

The standard error of an estimator reflects how it varies under repeated subsampling.

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and

$$SE(\hat{\beta}_0)^2 = \sigma^2\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\}$$

where $\sigma^2 = Var(\epsilon)$.

Assessing the Accuracy of the Coefficient Estimates.

The standard error of an estimator reflects how it varies under repeated subsampling.

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and

$$SE(\hat{\beta}_0)^2 = \sigma^2\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\}$$

where $\sigma^2 = Var(\epsilon)$.

The standard errors can be used to compute confidence intervals. A 95% CI is defined as the range of values such that with 95% probability the range will contain the true unknown value of the parameter.

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1).$$

Hypothesis testing

Standard errors can be used to perform *hypothesis testing* on the coefficients. The most common hypothesis test involves testing the *null hypothesis* of:

H_0 There is no relationship between X and Y

H_1 There is some relationship between X and Y

Hypothesis testing

Standard errors can be used to perform *hypothesis testing* on the coefficients. The most common hypothesis test involves testing the *null hypothesis* of:

H_0 There is no relationship between X and Y

H_1 There is some relationship between X and Y

Or, more formally:

H_0 $\beta_1 = 0$

H_1 $\beta_1 \neq 0$

since if $\beta_1 = 0$ the model becomes $Y = \beta_0 + \epsilon$ and X is not associated with Y.

Hypothesis testing

To test the H_0 we compute a t-statistic:

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

This will be a t-distribution with $n - 2$ degrees of freedom.

Using R, we can compute the probability of observing any value $\geq |t|$. We call this probability **p-value**.

Hypothesis testing on advertising data 1

```
fit <- lm(Sales ~ TV, data = my.advertise)
summary(fit)
```

```
##
## Call:
## lm(formula = Sales ~ TV, data = my.advertise)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             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
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Hypothesis testing on advertising data 2

```
fit <- lm(Sales ~ Radio, data = my.advertise)
summary(fit)
```

```
##
## Call:
## lm(formula = Sales ~ Radio, data = my.advertise)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7305  -2.1324   0.7707   2.7775   8.1810
##
## Coefficients:
##             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
## F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

Hypothesis testing on advertising data 3

```
fit <- lm(Sales ~ Newspaper, data = my.advertise)
summary(fit)
```

```
##
## Call:
## lm(formula = Sales ~ Newspaper, data = my.advertise)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2272  -3.3873  -0.8392   3.5059  12.7751
##
## Coefficients:
##             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
## F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

Assessing overall accuracy of the model

Residual standard error:

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

where the *residual sum of squares* is $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_1)^2$

Assessing overall accuracy of the model

Residual standard error:

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

where the *residual sum of squares* is $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_1)^2$

R^2 fraction of variance explained is

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

where $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the total sum of squares.

Assessing overall accuracy of the model 2

An R^2 statistic that is close to 1 indicates that a large proportion of variability is explained by the regression. Values close to 0 occur when the linear model is wrong or the error σ^2 is high.

Assessing overall accuracy of the model 2

An R^2 statistic that is close to 1 indicates that a large proportion of variability is explained by the regression. Values close to 0 occur when the linear model is wrong or the error σ^2 is high.

The link between correlation:

$$Cor(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

and R^2 is $R^2 = Cor(X, Y)^2$.

Section 4

Multiple linear regression. Variable importance.

Advertising data. Multiple linear regression

More formally:

$$sales \approx \beta_0 + \beta_1 \times TV + \epsilon$$

$$sales \approx \beta_0 + \beta_1 \times Radio + \epsilon$$

$$sales \approx \beta_0 + \beta_1 \times Newspaper + \epsilon$$

or with all predictors into one model:

$$sales \approx \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times Newspaper + \epsilon$$

Advertising data. Multiple linear regression

More formally:

$$sales \approx \beta_0 + \beta_1 \times TV + \epsilon$$

$$sales \approx \beta_0 + \beta_1 \times Radio + \epsilon$$

$$sales \approx \beta_0 + \beta_1 \times Newspaper + \epsilon$$

or with all predictors into one model:

$$sales \approx \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times Newspaper + \epsilon$$

Where β_j is the average effect on Y corresponding to X_j, *holding all other predictors fixed.*

Interpretation of the regression coefficients

- Uncorrelated predictors (ideal scenario)
  - each coefficient can be estimated and tested separately
  - independent interpretations *"a unit change in X_j is associated with a β_i change in Y"* are correct

Interpretation of the regression coefficients

- Uncorrelated predictors (ideal scenario)
  - each coefficient can be estimated and tested separately
  - independent interpretations *"a unit change in X_j is associated with a β_i change in Y"* are correct
- Correlated predictors
  - the variance of all coefficients tends to increase
  - interpretations are difficult, when X_j changes, other predictors change as well, leading to a composite effect on Y

Interpretation of the regression coefficients

- Uncorrelated predictors (ideal scenario)
  - each coefficient can be estimated and tested separately
  - independent interpretations *"a unit change in X_j is associated with a β_i change in Y"* are correct
- Correlated predictors
  - the variance of all coefficients tends to increase
  - interpretations are difficult, when X_j changes, other predictors change as well, leading to a composite effect on Y
- **Claims of causality should be avoided**

Interpretation of the model

Is there a relationship between the response and the predictors?

Interpretation of the model

Is there a relationship between the response and the predictors?

For the simple linear regression we checked $\beta_1 = 0$.

Interpretation of the model

Is there a relationship between the response and the predictors?

For the simple linear regression we checked $\beta_1 = 0$.

For a multiple regression setting with p predictors we have:

H_0 $\beta_1 = \beta_2 = ... = \beta_p = 0$

H_1 at least one β_j is non zero.

Estimation of coefficients. Multiple Regression

Given the estimates $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ the predicted values are:

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

Estimation of coefficients. Multiple Regression

Given the estimates $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ the predicted values are:

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

The estimated $\beta_0, \beta_1, ..., \beta_p$ minimise the RSS (sum of squared residuals):

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y_1})^2$$

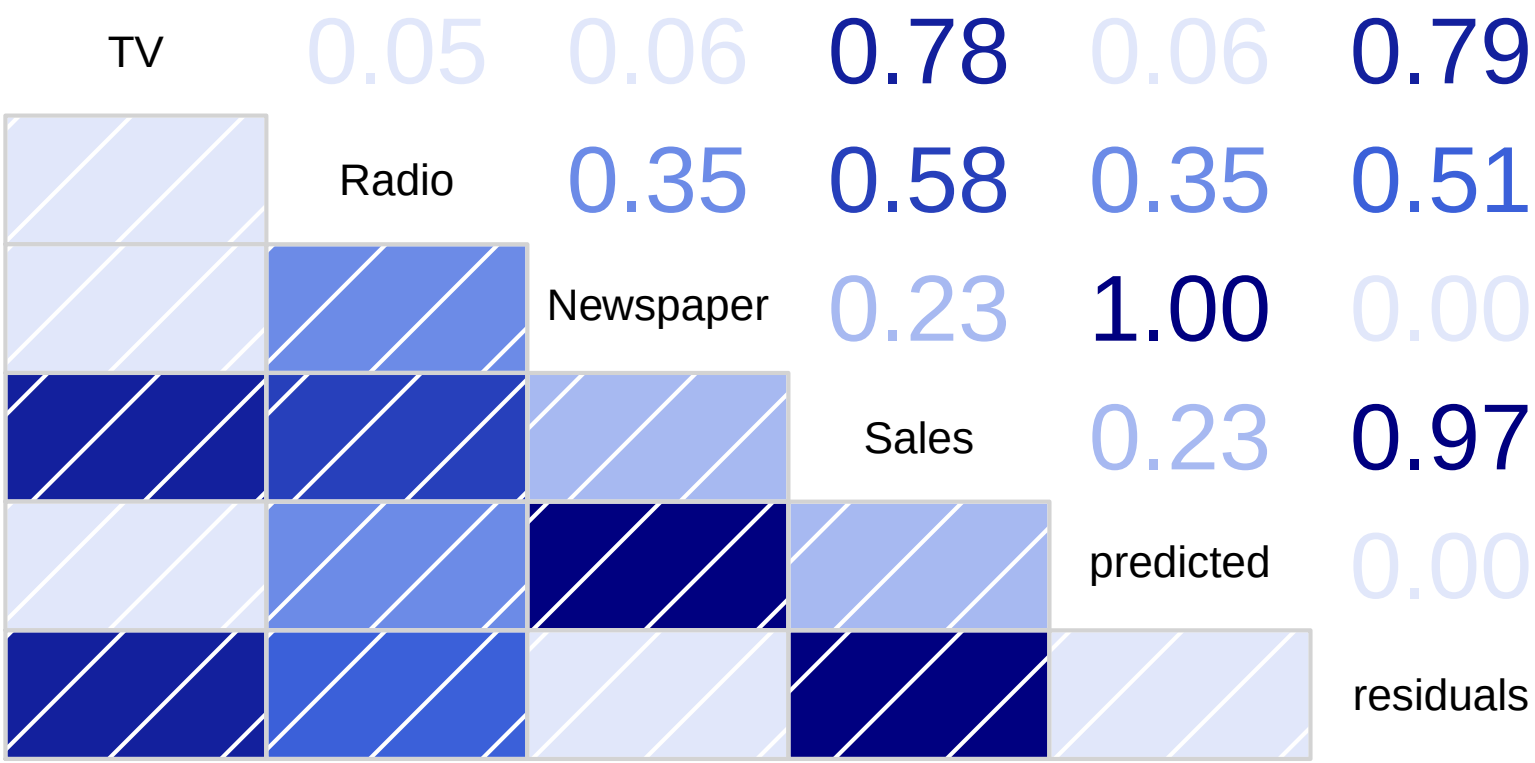
Multiple regression on the advertising data

```
all.predictors <- lm(Sales ~ TV + Radio + Newspaper, data = my.advertise)
summary(all.predictors)
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper, data = my.advertise)
##
## 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
```

Multiple regression on the advertising data

```
library(corrgram)
cor.matrix <- cor(my.advertise)
corrgram::corrgram(cor.matrix, order=FALSE,
         upper.panel=panel.cor)
```



Multiple regression on the advertising data. Questions

1. Is at least one of the predictors $X_1, X_2, ..., X_p$ useful in predicting the response?

Multiple regression on the advertising data. Questions

1. Is at least one of the predictors $X_1, X_2, ..., X_p$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?

Multiple regression on the advertising data. Questions

1. Is at least one of the predictors $X_1, X_2, ..., X_p$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?

Multiple regression on the advertising data. Questions

1. Is at least one of the predictors $X_1, X_2, ..., X_p$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors 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?

Section 5

Identification of discriminative variables.

Important variables

- The brute force approach is called "all subsets" or best subsets regression i.e. we compute the least squares fit for all possible subsets and then choose between them based on some criteria that balances training error with model size.

Important variables

- The brute force approach is called "all subsets" or best subsets regression i.e. we compute the least squares fit for all possible subsets and then choose between them based on some criteria that balances training error with model size.
- However we often can't examine all possible models, since they are 2^p of them; for example when p $=$ 40 there are over a billion models! Instead we need an automated approach that searches through a subset of them.

Forward selection

- Begin with the *null model* - a model that contains an intercept but no predictors

Forward selection

- Begin with the *null model* - a model that contains an intercept but no predictors
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.

Forward selection

- Begin with the *null model* - a model that contains an intercept but no predictors
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.

Forward selection

- Begin with the *null model* - a model that contains an intercept but no predictors
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold

Forward selection on the advertising data

```
model.T <- lm(Sales ~ TV, data = my.advertise)
model.R <- lm(Sales ~ Radio, data = my.advertise)
model.N <- lm(Sales ~ Newspaper, data = my.advertise)
```

```
## [1] "model.T"
## [1] "Residual standard error: 3.259 on 198 degrees of freedom"
## [1] "Multiple R-squared: 0.612, Adjusted R-squared: 0.61"
## [1] "F-statistic: 312.145 on 1 and 198 DF, p-value: < 2.2e-16"

## [1] "model.R"
## [1] "Residual standard error: 4.275 on 198 degrees of freedom"
## [1] "Multiple R-squared: 0.332, Adjusted R-squared: 0.329"
## [1] "F-statistic: 98.422 on 1 and 198 DF, p-value: < 2.2e-16"

## [1] "model.N"
## [1] "Residual standard error: 5.092 on 198 degrees of freedom"
## [1] "Multiple R-squared: 0.052, Adjusted R-squared: 0.047"
## [1] "F-statistic: 10.887 on 1 and 198 DF, p-value: 0.001148"
```

Forward selection on the advertising data

```
model.TR <- lm(Sales ~ TV + Radio, data = my.advertise)
model.TN <- lm(Sales ~ TV + Newspaper, data = my.advertise)
```

```
## [1] "model.TR"
## [1] "Residual standard error: 1.681 on 197 degrees of freedom"
## [1] "Multiple R-squared: 0.897, Adjusted R-squared: 0.896"
## [1] "F-statistic: 859.618 on 1 and 197 DF, p-value: < 2.2e-16"

## [1] "model.TN"
## [1] "Residual standard error: 3.121 on 197 degrees of freedom"
## [1] "Multiple R-squared: 0.646, Adjusted R-squared: 0.642"
## [1] "F-statistic: 179.619 on 1 and 197 DF, p-value: < 2.2e-16"
```

We select the model.TR.

Forward selection on the advertising data

```
model.TRN <- lm(Sales ~ TV + Radio + Newspaper, data = my.advertise)
```

```
## [1] "model.TRN"
## [1] "Residual standard error: 1.686 on 196 degrees of freedom"
## [1] "Multiple R-squared: 0.897, Adjusted R-squared: 0.896"
## [1] "F-statistic: 570.271 on 1 and 196 DF, p-value: < 2.2e-16"
```

model.TRN has larger error than model.TR, and comparable R^2 Our best model is model.TR

Backward selection

- Start with all variables in the model; fit the model
- Remove the variable with the largest p-value i.e. the variable that is the least statistically significant

Backward selection

- Start with all variables in the model; fit the model
- Remove the variable with the largest p-value i.e. the variable that is the least statistically significant
- The new $p - 1$-variable model is fit, and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached e.g. when all remaining variables have a significant p-value above some threshold

Backward selection on the advertising data

```
model.TRN <- lm(Sales ~ TV + Radio + Newspaper, data = my.advertise)
summary(model.TRN)
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper, data = my.advertise)
##
## 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
```

The Newspaper prediction has the largest p-value i.e. it will be excluded.

Backward selection on the advertising data

```
model.TR <- lm(Sales ~ TV + Radio, data = my.advertise)
summary(model.TR)
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio, data = my.advertise)
##
## 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
```

All predictors are significant; therefore this is the final model.

Mixed selection

- this is a combination of forward and backward selection
- start with no variables in the model; add the variable that provides the best fit

Mixed selection

- this is a combination of forward and backward selection
- start with no variables in the model; add the variable that provides the best fit
- as more predictors are added, the p-values of the predictors already included will start to fluctuate; exclude the ones that rise above an *a priori* defined threshold
- continue performing the forward and backward steps until all variables in the model have sufficiently low p-values

Mixed selection

- this is a combination of forward and backward selection
- start with no variables in the model; add the variable that provides the best fit
- as more predictors are added, the p-values of the predictors already included will start to fluctuate; exclude the ones that rise above an *a priori* defined threshold
- continue performing the forward and backward steps until all variables in the model have sufficiently low p-values
- **Backward selection cannot be performed if $p > n$**
- **Forward selection can always be used; however it is a greedy approach (remedied by the mixed selection)**

Section 6

Summary

Overview of the main points for Linear Regression

- coefficient estimation for single parameter regression
- assessment of discriminative power for single and multiple regression
- identification of "optimal" model using incremental selection of predictors

Quotes

"Essentially all models are wrong, but some are useful".

— George Box

"The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively."

— Fred Mosteller and John Tukey, paraphrasing George Box