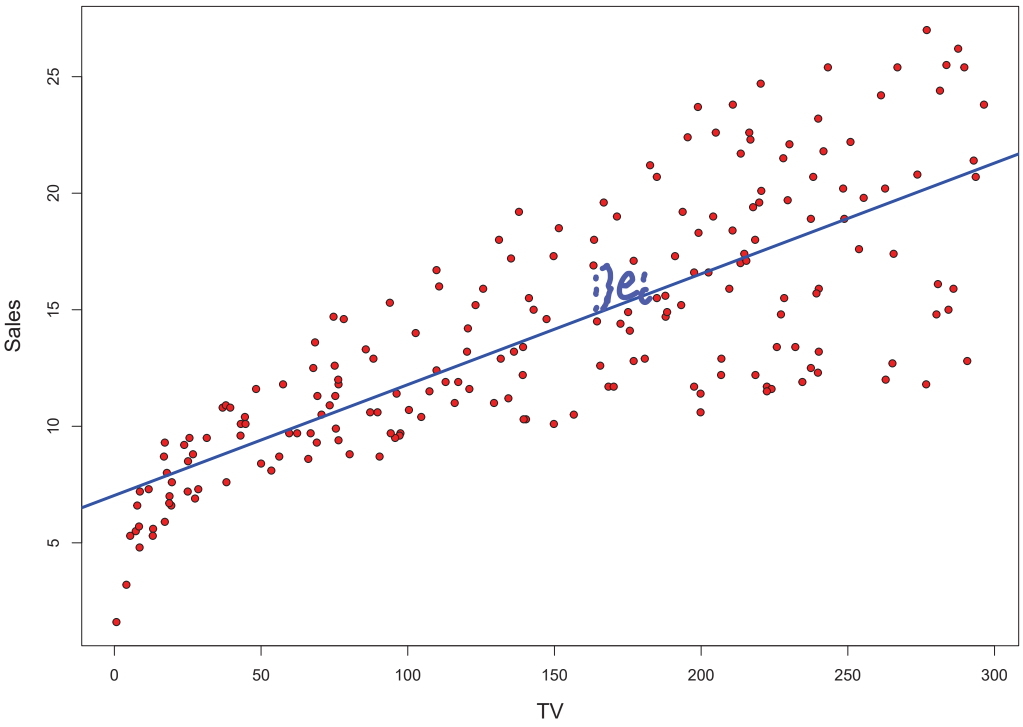


Least squares fit for the regression of sales onto TV. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

```
> names(lm_fit)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

> lm_fit$coefficients
(Intercept)          TV
 7.03259355  0.04753664
```

$\hat{\beta}_0 \rightarrow$ $\leftarrow \hat{\beta}_1$

```
> lm_fit$residuals
          1           2           3           4           5
 4.12922549  1.25202595  1.44977624  4.26560543 -2.72721814
          6           7           8           9          10
-0.24616232  2.03404963  0.45350227 -2.64140866 -5.93041431
         11          12          13          14          15
-1.57476548  0.16128975  1.03603441 -1.96741599  2.26517814
```

. . .

Assessing the Accuracy of the Coefficient Estimates

The standard error of an estimator tells us the average amount that the estimate differs from the actual value. The standard errors associated with β_0 and β_1 are

$$\mathrm{SE}(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \qquad \mathrm{SE}(\hat{\beta}_1) = \sigma\sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

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

In general, σ is not known, but can be estimated from the data. This estimate is known as the residual standard error, and is given by the formula

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

Confidence Intervals

These standard errors can be used to compute confidence intervals. A $(1-\alpha)100\%$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}\mathrm{SE}(\hat{\beta}_1)$$

where $t_{n-2,\alpha/2}$ is the $100(1-\alpha/2)$th percentile (quantile) of the t distribution with $n-2$ degrees of freedom.

For the advertising data,

```
> confint(lm_fit, level = 0.95)
```

```
                 2.5 %     97.5 %
(Intercept) 6.12971927 7.93546783
TV          0.04223072 0.05284256
```

the 95% confidence interval for β_1 is [0.042, 0.053].

Prediction

The estimated regression line is

$$\hat{\mu}_{Y|X}(x) = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$\hat{y}$ is the predicted value of Y for specific values of X $(X = x)$.

Example: What is the predicted sales for a market whose TV advertising is 100?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 7.0326 + 0.0475x$$

$$\hat{y} = 7.0326 + 0.0475 \times 100 = 11.7826 \text{ units}$$

This value represents two types of prediction:

1) The average sales of all markets whose TV advertising is 100. ($\hat{\mu}_{Y|X}(100)$)
2) The sales of an individual market whose TV advertising is 100. ($\hat{y}$)

A $(1-\alpha)100\%$ confidence interval for a predicted response is

average response

$$\hat{y} \pm t_{n-2,\alpha/2}\text{RSE}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

$\hat{\mu}_{Y|X}(x)$

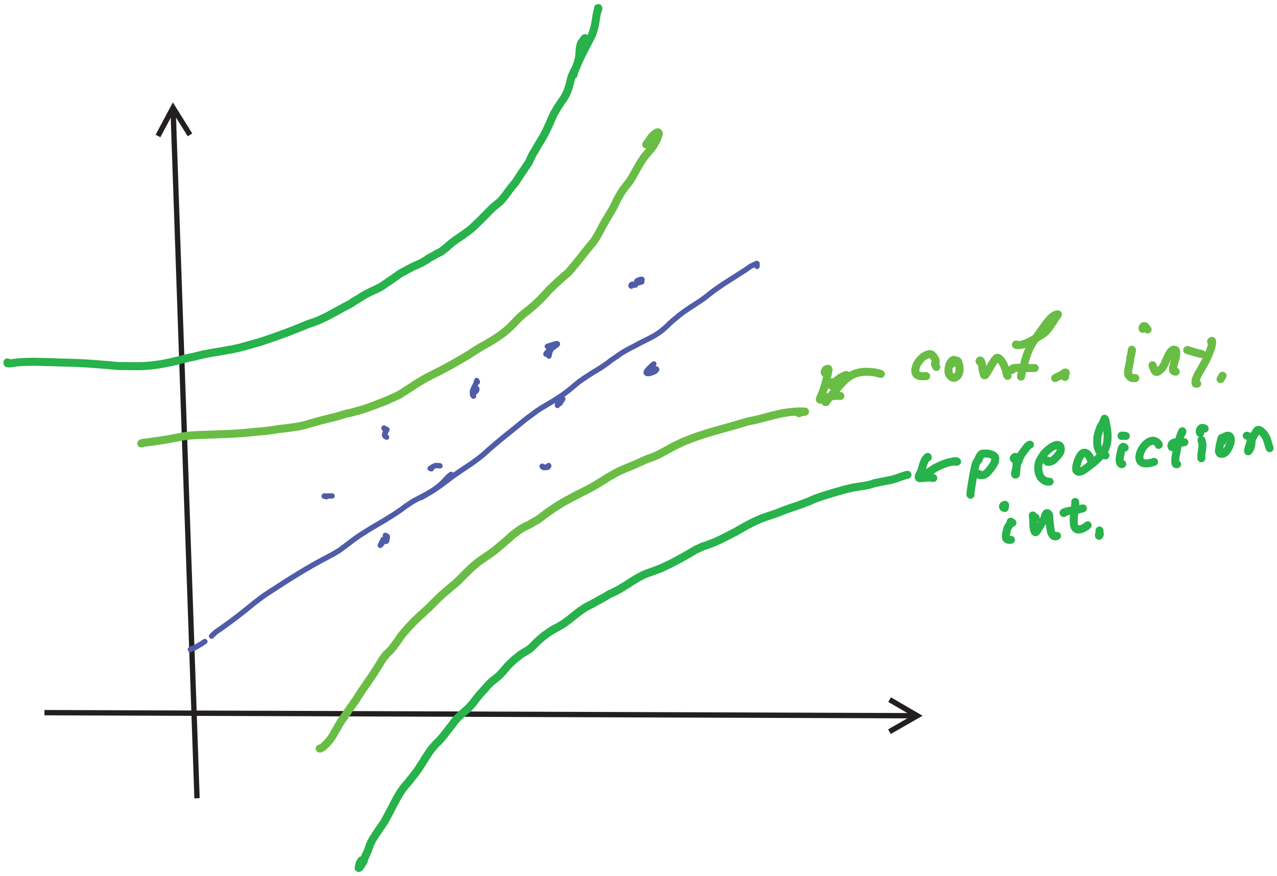
x = value we are predicting for

The square root is smallest when x is near $\bar{x}$.

A $(1-\alpha)100\%$ prediction interval for a predicted response is

an individual response

$$\hat{y} \pm t_{n-2,\alpha/2}\text{RSE}\underbrace{\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}}_{\text{gives a larger error.}}$$



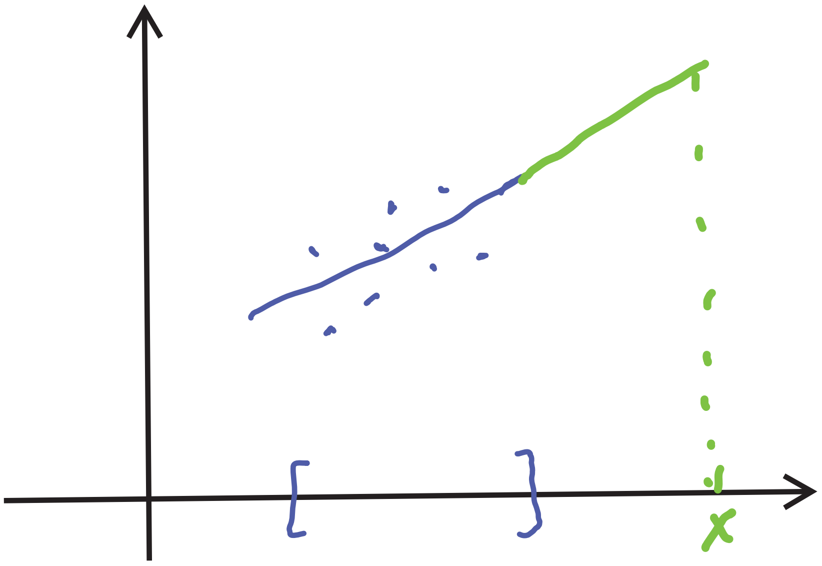
The predict() function can be used to produce confidence intervals and prediction intervals for the prediction of sales for a given value of TV advertising budget.

```
> predict(lm_fit,data.frame(TV=(c(20,100,250))), interval="confidence",
level = 0.95)
        fit       lwr       upr
1  7.983326  7.170396  8.796257
2 11.786258 11.267820 12.304695
3 18.916754 18.206189 19.627319
```

```
> predict(lm_fit,data.frame(TV=(c(20,100,250))), interval="prediction",
level = 0.95)
        fit       lwr      upr
1  7.983326  1.505984 14.46067
2 11.786258  5.339251 18.23326
3 18.916754 12.451461 25.38205
```

For instance, the 95% confidence interval associated with a TV value of 100 is (11.27, 12.30), and the 95% prediction interval is (5.34, 18.23). As expected, the confidence and prediction intervals are centered around the same point (a predicted value of 11.79 for sales when TV equals 100), but the latter are substantially wider.

Caution: Don't do prediction for values of x that are far outside of the range of the data. Don't extrapolate!



Hypothesis Testing

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

$$H_0: \text{There is no relationship between } X \text{ and } Y$$

versus the *alternative hypothesis*

$$H_a: \text{There is some relationship between } X \text{ and } Y$$

- Mathematically, this corresponds to testing

  $$H_0: \ \beta_1 = 0$$

  versus

  $$H_a: \ \beta_1 \neq 0$$

  since if $\beta_1 = 0$ then the model ($Y = \beta_0 + \beta_1 X + \epsilon$), reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

- To test the null hypothesis, we compute a t-*statistic*, given by

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

- This will have a t-distribution with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$.

- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the p-value. Typical p-value cutoffs for rejecting the null hypothesis are 5% or 1%.

For the Advertising data,

```
> summary(lm_fit)

Call:
lm(formula = sales ~ TV, data = Adv)

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
```

	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

Both p-values are very small. so we can conclude that $\beta_0 \neq 0$ and $\beta_1 \neq 0$.

Assessing the Overall Accuracy of the Model

- The quality of a linear regression fit is typically assessed using two related quantities: the *residual standard error* (RSE) and the R^2 statistic.
- The RSE is an estimate of the standard deviation of ϵ. It is computed

using the formula

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

where RSS was defined before by

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

- The RSE provides an absolute measure of lack of fit of the model $Y = \beta_0 + \beta_1 X + \epsilon$ to the data. But since it is measured in the units of Y, it is not always clear what constitutes a good RSE.
- The R^2 statistic provides an alternative measure of fit. It takes the form of a proportion—the proportion of variance explained—and so it always takes on a value between 0 and 1, and is independent of the scale of Y.
- R^2 can be calculated by

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

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

An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression. A number near 0 indicates that the regression did not explain much of the variability in the response; this might occure because the linear model is wrong, or the inherent error σ^2 is high, or both.