3.1 Simple Linear Regression

Prof. Lauren Perry

Simple Linear Regression

Model:

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

where X consists of a single predictor variable.

- The *intercept*, β_0, and the *slope*, β_1, make up the models *parameters* or *coefficients*.

When we use the estimated model to make predictions, we write

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

- Conceptually, this is a 2D extension of using a sample mean $\bar{x}$ to estimate a population mean μ.

Estimating the Coefficients

- We can think of our data as n points of the form (x_i, y_i).
- Our goal is to estimate β_0 and β_1 so that the model fits the data well.
    - That is, so that
    $$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$
    for each $i \in \{1, \ldots, n\}$.
    - Idea: the line is as close as possible to all n data points.

Least Squares

The *least squares criterion* focuses on "closeness" as a measure of how close each response value y is to the predicted value $\hat{y}$:

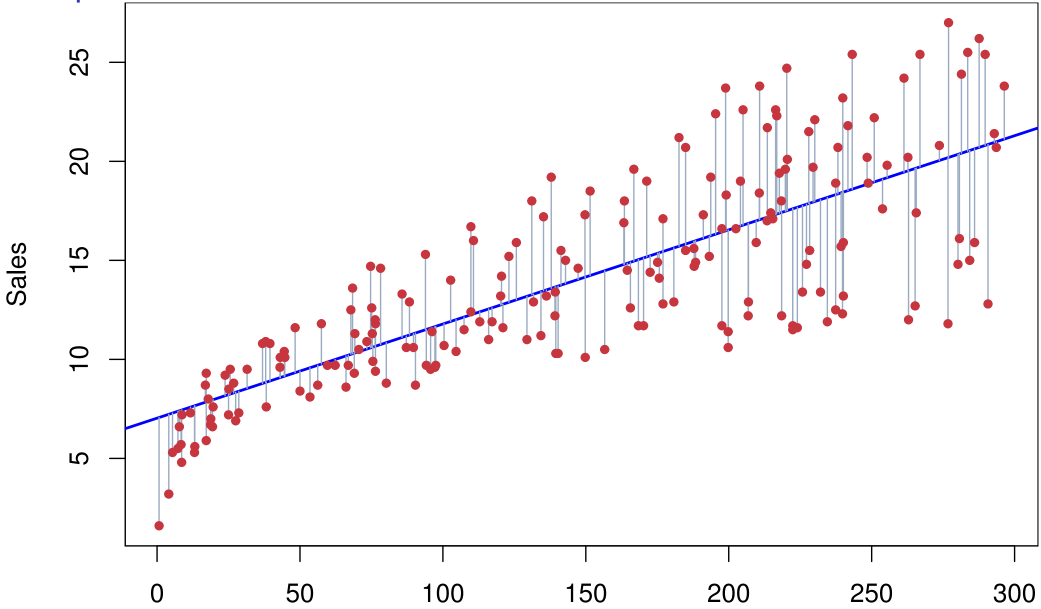
$$e_i = y_i - \hat{y}_i$$

where e_i is the ith *residual*.

Then the *residual sum of squares* is

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

Least Squares



Least Squares

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

$$
\begin{aligned}
\text{RSS} &= e_1^2 + e_2^2 + \cdots + e_n^2 \\
&= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 \\
&= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2
\end{aligned}
$$

which we minimize by taking the derivatives

$$
\frac{\delta \text{RSS}}{\delta \hat{\beta}_0} \quad \text{and} \quad \frac{\delta \text{RSS}}{\delta \hat{\beta}_1}
$$

Least Squares

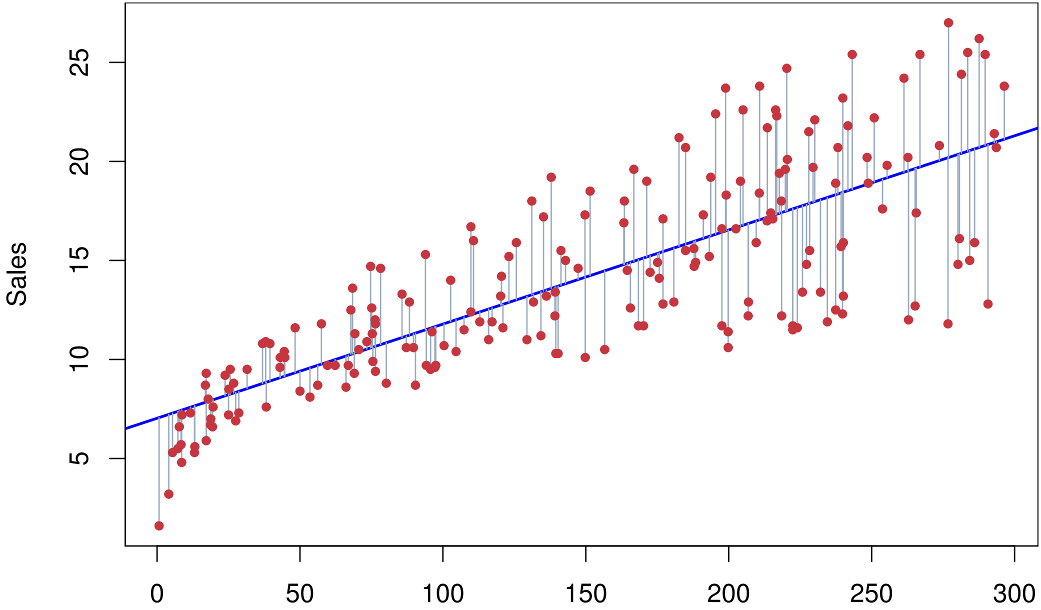
This minimization problem yields

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

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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



Assessing Accuracy of Coefficient Estimates

When we assume f is linear, we say

$$Y = f(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon$$

- where β_0 is the intercept term.
  - This is the expected value of Y when $X = 0$.
- and β_1 is the slope.
  - This is the average increase in Y for a one-unit increase in X.

Assessing Accuracy of Coefficient Estimates

The model

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

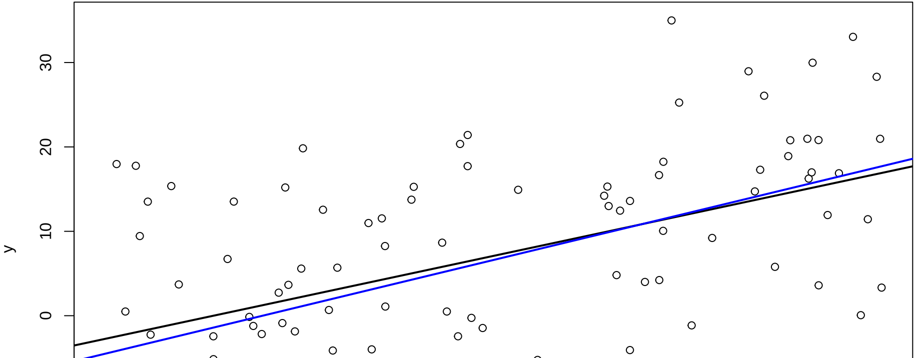
defines the (unknown) *population regression line*, the best linear approximation to the true relationship between X and Y.

The estimated line

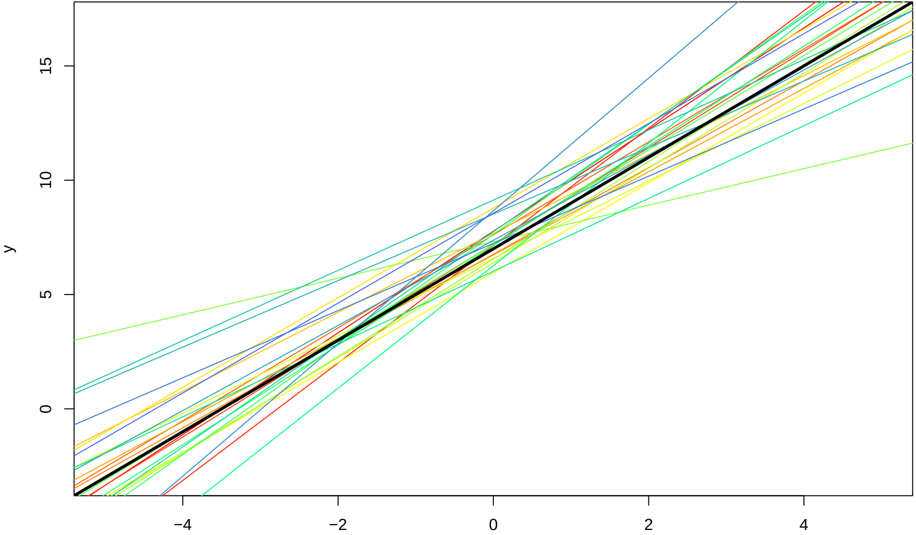
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

is the *least squares regression line*.

```r
f.x <- function(x){2*x + 7 + rnorm(length(x),0,10)}
x <- runif(100, -5, 5)
y <- f.x(x)
plot(x,y)
abline(7, 2, col='black', lwd=2)
abline(lm(y~x), col='blue', lwd=2)
```



Example: Generating Many Samples



Example: Generating Many Samples

```r
rand.lines <- function(){
  x <- runif(100, -5, 5)
  y <- 2*x + 7 + rnorm(length(x),0,10)
  lm(y ~ x)$coefficients
}
coefs <- replicate(25, rand.lines())

colfunc <- colorRampPalette(c("red","yellow","springgreen","royalblue"))
colrs <- colfunc(25)

plot(-5:5, 2*(-5:5)+7, type='l', lwd=2, xlab='x', ylab='y')
for(i in 1:25) abline(coefs[,i], col=colrs[i])
```

Assessing Accuracy of Coefficient Estimates

Least squares estimates are *unbiased*. Idea:

- Take a large number of samples and calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ for each.
- If we were to find the mean of all the estimates of $\hat{\beta}_0$, it would be β_0.
- ...and if we were to find the mean of all the estimates of $\hat{\beta}_1$, it would be β_1.
- We can see this visualized in the previous plot.

Assessing Accuracy of Coefficient Estimates

As in using $\bar{x}$ to estimate μ, a regression line from a single sample may or may not be a good estimate.

- How variable is it?
  - When we use $\bar{x}$ to estimate μ, the variability is

$$\mathrm{Var}(\bar{x}) = \mathrm{SE}(\bar{x})^2 = \frac{\sigma^2}{n}$$

  - SE tells us roughly how far a typical estimate differs from μ.

Assessing Accuracy of Coefficient Estimates

So what about the regression line?

For $\hat{\beta}_0$,

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

and for $\hat{\beta}_1$,

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

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

- Assumption: the errors ϵ_i are uncorrelated and have common variance.

Estimating σ

In general, σ is unknown, but can be estimated from the data:

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

- This is also called the *residual standard error*.

Confidence Intervals for β_0 and β_1

A general confidence interval looks like

$$\text{point estimate} \pm (\text{critical value}) \times (\text{standard error})$$

For β_i,

$$\hat{\beta}_i \pm t_{df,\alpha/2} \times \text{SE}(\hat{\beta}_i)$$

- We use the t-distribution under the assumption that the errors are approximately Gaussian (normal).

Hypothesis Tests for β_0 and β_1

The most common hypothesis test in this setting involves

- (*Null hypothesis*) H_0: There is no relationship between X and Y.
- (*Alternative hypothesis*) H_A: There is some relationship between X and Y.

Hypothesis Tests for β_0 and β_1

Mathematically, this is just

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0$$

Because, if $\beta_1 = 0$, then the model is just $Y = \beta_0 + \epsilon$, which does not depend on X.

- Note: in the model $Y = \beta_0 + \epsilon$, we find $\hat{\beta}_0 = \bar{y}$.

Hypothesis Tests for β_0 and β_1

Two ways to test these hypotheses:

1. Use the confidence interval approach (check if 0 is in the interval for $\hat{\beta}_1$).
2. Compute a *test statistic*

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

   which measures how many standard deviations $\hat{\beta}_1$ is from 0.
   - From here, we typically calculate the *p-value*, or the probability of observing a value as extreme as $\hat{\beta}_1$ if in fact $\beta_1 = 0$.

Hypothesis Tests for β_0 and β_1

In practice, we never do this by hand.

```
mod1 <- lm(Loblolly$age ~ Loblolly$height)
summary(mod1)
```

```
##
## Call:
## lm(formula = Loblolly$age ~ Loblolly$height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5528 -0.7378  0.1421  0.6925  2.8966
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.757380   0.229203   3.304  0.00141 **
## Loblolly$height 0.378274   0.005979  63.272  < 2e-16 ***
## ---
```

Assessing Model Accuracy

Having concluded that β_1 is nonzero, we want to examine the extent to which the model fits the data.

Linear regression model quality assessed using two measures:

1. Residual standard error 2 R^2

Residual standard error

Recall: $\text{RSE} = \hat{\sigma}$.

- This is a measure of how far - on average - linear regression line estimates deviate from the truth.
  - A "good" RSE will depend on problem context (e.g., units).
- RSE is considered a *lack of fit* measure.
  - If predictions are very close to true outcomes, RSE will be small (and vice versa).

R^2 Statistic

RSE is measured in units of Y, so it may be unclear what a "good" RSE is.

The R^2 statistic

- is the proportion of variance explained by the model.
- always takes values between 0 and 1.

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

where $\text{TSS} = \sum(y_i - \bar{y})^2$

Sum of Squares

- TSS is the *total sum of squares*, the total variance in Y.
- RSS is the *residual sum of squares*, the variability leftover after the regression is performed.
- Another measure, ESS, is the *explained sum of squares* and is the variability in Y that is explained by the regression model:

$$\text{TSS} = \text{RSS} + \text{ESS}$$

Thus, $R^2 = \frac{\text{ESS}}{\text{TSS}}$ is the proportion of variability in Y that can be explained by the linear regression model.

R^2 Statistic

"Good" R^2 values are those closer to 1.

. . . How close to 1?

It depends!

Correlation

We can also measure the (linear) *correlation* between two variables.

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

In the linear regression context, the square of the correlation is the R^2 we just saw.

Overall model fit

```
mod1 <- lm(Loblolly$age ~ Loblolly$height)
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: Loblolly$age
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## Loblolly$height  1   5076  5076.0  4003.3 < 2.2e-16 ***
## Residuals       82    104     1.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cor(Loblolly$height, Loblolly$age)
```

```
## [1] 0.9899132
```