Linear Regression Analysis

- The regression line
- The regression effect
- Properties of the residuals
- Sampling distribution of Least Squares Estimates
Correlation Coefficient

Review: correlation coefficient

\[ \rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y} \]

\[ s_x^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 \]

\[ s_y^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2 \]

\[ s_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) \]

Correlation measures the strength of the linear association between \( x \) and \( y \)

\[-1 \leq \rho \leq 1\]

- \(|\rho|\) close to 1 \(\rightarrow\) strong linear association
- \(|\rho|\) close to 0 \(\rightarrow\) no linear association

Warning: only linear association
The Regression Line

- Goes through the point of averages
- Slope: increase of 1 SD in $x$ leads to an increase in $\rho$ SD in $\hat{y}$.
- Intercept: predicted response when $x = 0$ (not very important)

Danger: Do not extrapolate
> n <- 15
# Fix the x’s
> x <- runif(n,0,10)

# Fix sigma
> sigma <- 3

# Generate errors
> error <- rnorm(n,mean = 0, sd = sigma)

# Construct observations
> y <- -1 + 2*x + error
# Plot Data
> plot(x,y)
Figure 1: Plot of x versus y
# Fit Least Squares

`> fm <- lm(y ~ x)`

`> fm`

**Call:**
`lm(formula = y ~ x)`

**Coefficients:**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.9387</td>
</tr>
<tr>
<td>x</td>
<td>2.0347</td>
</tr>
</tbody>
</table>
# Sampling distribution of intercept and slope
# Construct 1000 datasets from the same model

\[
\begin{align*}
\text{beta} & \leftarrow \text{NULL} \\
B & \leftarrow 1000 \\
\text{for} \ (i \ in \ 1:B) \ & \text{do} \\
& \quad \{ \\
& \quad \quad \text{error} \leftarrow \text{rnorm}(n, \text{mean} = 0, \text{sd} = \text{sigma}) \\
& \quad \quad y \leftarrow -1 + 2*x + \text{error} \\
& \quad \quad \text{fm} \leftarrow \text{lm}(y \sim x) \\
& \quad \quad \text{beta} \leftarrow \text{rbind(beta, fm$coefficients)} \\
& \quad \} \\
\text{beta0} & \leftarrow \text{beta[,1]} \\
\text{beta1} & \leftarrow \text{beta[,2]} 
\end{align*}
\]
Figure 2: Histogram of the estimates $\hat{\beta}_0$
Figure 3: Histogram of the estimates $\hat{\beta}_1$
# Calculate Statistics

> mean(beta0)
[1] -0.8526184
> mean(beta1)
[1] 1.980757
> sd(beta0)
[1] 1.810449
> sd(beta1)
[1] 0.3243954

# Compare with true values

> sx <- sqrt(sum((x - mean(x))^2))
> sigma*sqrt(1/n + mean(x)^2/sx^2)
[1] 1.765777
> sigma/sx
[1] 0.3197968