Simple Linear Regression
(ISLR 3.1)

Yingbo Li

Southern Methodist University

STAT 4399

Outline

1. Parameter Estimation
2. Hypothesis Testing
3. Assessing the Accuracy for the Regression Model

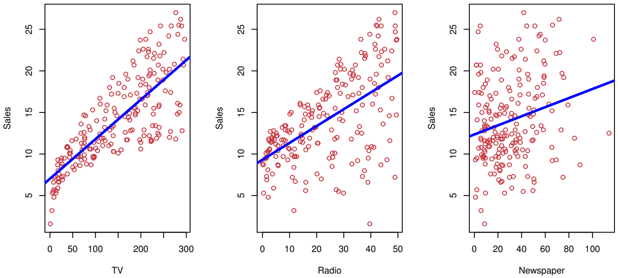
Simple linear regression

- One (continuous) response Y, vs. one predictor X.
- There is approximately a linear relationship between X and Y:

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

where both *regression coefficients* β_0 and β_1 are unknown.

Recall the `Advertising` data



Today, we study the linear relationship between `TV` ad budget and `sales`.

Regression line

- Training data: $n = 200$

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

- We assume a simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \text{for } i = 1, 2, \ldots, n.$$

  - β_0: intercept (unknown parameter)
  - β_1: slope (unknown parameter)
  - ϵ_i: mean zero error, usually assumed to be i.i.d. $\mathrm{N}(0, \sigma^2)$.

- Regression line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

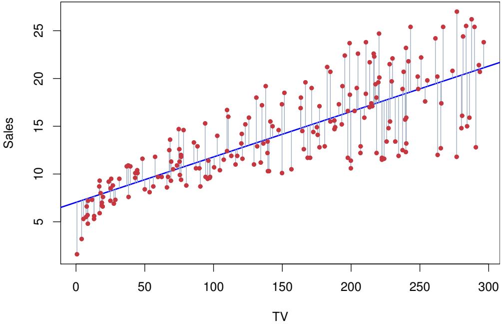
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates based on the training data.
  - $\hat{Y}$ is a prediction of the response on the basis of a certain value of X.

Residual sum of squares

The regression line should be the closest to all $n = 200$ data points.

For each $i = 1, 2, \ldots, n$, its

- fitted value: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- actual value: Y_i
- residual: $\hat{\epsilon}_i = Y_i - \hat{Y}_i$



How to measure the closeness? *Residual Sum of Squares (RSS)*

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Ordinary least square (OLS) estimators

OLS estimators are chosen to minimize the RSS.

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

OLS estimators

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2},$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

where $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$ and $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.

Note, the value of $\hat{\beta}_0$ ensures that the regression line passes the center of the training data $(\bar{X}, \bar{Y})$.

```
> lm1 = lm(Sales ~ TV, data = Advertising);
> summary(lm1);

Call:
lm(formula = Sales ~ TV, data = Advertising)

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
```

Interpretations

$$\texttt{Sales} \approx 7.0326 + 0.0475 \times \texttt{TV}$$

- Slope: for each unit increase in X, we would expect Y to increase/decrease by $\hat{\beta}_1$ unit on average.
  - An additional \$1000 spent on TV advertising is associated with selling additional 47.5 units of product on average.
- Intercept: markets with zero X are expected to have $\hat{\beta}_0$ in Y on average.

Assessing the accuracy of $\hat{\beta}_0, \hat{\beta}_1$

If we use another training dataset, our estimates $\hat{\beta}_0, \hat{\beta}_1$ will be different. Then how accurate (close to the truth) our estimates using the current dataset $\hat{\beta}_0 = 7.0326$, $\hat{\beta}_1 = 0.0475$ are?

Standard Errors (SE) of the estimators: $SE(\hat{\beta}_j)^2 = V(\hat{\beta}_j)$, for $j = 0, 1$:

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right].$$

If the variance $\sigma^2 = V(\epsilon)$ is unknown, then estimate it by

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

Here $\hat{\sigma}$ is called the *residual standard error*.

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a point estimate (e.g. sample mean $\bar{X}$) to estimate a parameter (e.g. population mean μ) is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably will not hit the exact population parameter. If we report a range of plausible values – a confidence interval – we have a good shot at capturing the parameter.

Confidence interval

When n is large (say > 30), a $100(1-\alpha)\%$ *confidence interval* (CI) on β_j:

$$\hat{\beta} \pm z_{\alpha/2} \times SE(\hat{\beta}),$$

Here $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of a standard normal distribution. For 95% CI, $\alpha = 0.05$, and

```
> qnorm(0.975, mean = 0, sd = 1)
[1] 1.959964
```

Recall that:

```
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 ***
```

95% CI for β_1 is

$$\begin{aligned}\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1) &= 0.0475 \pm 1.96 \times 0.0027 \\ &= [0.0422, 0.0528]\end{aligned}$$

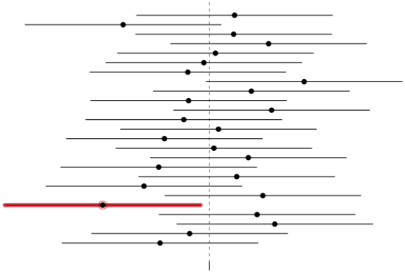
Interpretation of CI

We are $100(1-\alpha)\%$ confident that *population parameter* is between *[l, u]*.

What do we mean by "we are 95% confident ..."?

- Suppose we took many samples and built a confidence interval from each sample using the equation $\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)$.
- Then about 95% of those intervals would contain the true slope (β_1).

- The figure on the left shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true slope β_1, and one does not.



We are 95% confident that the slope β_1 is between $[0.0422,\ 0.0528]$.

Hypothesis testing

Is there a significant relationship between X and Y?

- This is equivalent to ask: is the slope β_1 zero or not?
- The 95% CI for β_1 is $[0.0422, 0.0528]$; since it does not contain zero, β_1 is not very likely to be zero.
- Another way is to do hypothesis testing:

$$H_0 : \beta_1 = 0 \longleftrightarrow H_1 : \beta_1 \neq 0$$

t-test:

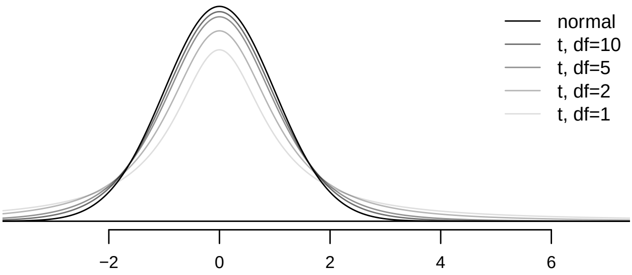
- Under the null hypothesis H_0, the test statistic

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

follows a Student-t distribution with degrees of freedom $df = n - 2$.

The t distribution

- Always centered at zero, like the standard normal distribution.
- This distribution also has a bell shape, but its tails are *thicker* than the normal model's.
- Has a single parameter: *degrees of freedom* (df).



What happens to shape of the t distribution as df increases?

t-test

Recall that:

```
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
```

The p-value is $< 2 \times 10^{-16}$. So we *reject the null hypothesis*.

The data provide convincing evidence that there exists a significant linear relationship between TV ad budget and sales.

How to find the p-value for a one-sided t-test?

$$H_0 : \beta_1 = 0 \longleftrightarrow H_1 : \beta_1 > 0$$

Sums of squares

- Total sum of squares: total variability of Y

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- Sum of squares of regression: variability in Y explained by the model

$$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

- Residual sum of squares: variability in Y left unexplained

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

There relationship:

$$TSS = SS_{reg} + RSS$$

R^2: the fraction of variance explained

$$R^2 = \frac{SS_{reg}}{TSS}$$

```
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
```

61.19% of the variability in sales can be explained by TV ad budget using the simple linear model.

For simple linear regression,

$$R^2 = r^2, \quad r = Cor(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

```
> cor(Advertising$TV, Advertising$Sales)^2;
[1] 0.6118751
```