Chapter 2: Simple Linear Regression (Cont'd)

- 2.3 Hypothesis Testing
- 2.4 Confidence Intervals for Parameters
- 2.5 Prediction
- More Detail on Degrees of Freedom; ANOVA
- 2.6 Coefficient of Determination
- 2.12 Maximum Likelihood Estimation; Gauss-Markov Theorem
- 2.11 Regression Through the Origin
- 2.10 Issues Which May Arise - Hazards of Regression

2.3 Hypothesis Testing

- β_1 summarizes the relationship between x and y.

- How do we know this relationship is real?

- Is there a way to test if there is no relation between x and y? That is to see if $\beta_1 = 0$.

2.3 Hypothesis Testing

- In order to test the relationship, the assumption that ε is $N(0, \sigma^2)$ is necessary.

- With this assumption, we recognize that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are normally distributed.

- The parameters (mean and variance) in the distributions mentioned above can be derived, and estimated from the data.

- The hypothesis testing we have learned about can be applied, based on the information above.

Distributions of $\widehat{\beta}_1$ and $\widehat{\beta}_0$ (assuming normal responses)

- y_i is $N(\beta_0 + \beta_1 x_i, \sigma^2)$, and
  $\widehat{\beta}_1 = \sum_{i=1}^n a_i y_i \qquad (a_i = \frac{x_i - \bar{x}}{S_{xx}})$
  $\Rightarrow \widehat{\beta}_1$ is $N(\beta_1, \frac{\sigma^2}{S_{xx}})$.
  Also, $\frac{SSE}{\sigma^2}$ is χ^2_{n-2} (independent of $\widehat{\beta}_1$) so

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}} \sim t_{n-2}$$

  $\widehat{\beta}_0 = \sum_{i=1}^n c_i y_i \qquad (c_i = \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}})$
  $\Rightarrow \widehat{\beta}_0$ is $N(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right))$ and

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$$

Inference for β_1:

$$H_0 : \beta_1 = \beta_{10} \quad \text{vs.} \quad H_1 : \beta_1 \neq \beta_{10}$$

Under H_0,

$$t_0 = \frac{\widehat{\beta}_1 - \beta_{10}}{\sqrt{MSE/S_{xx}}} \quad \text{or} \quad \frac{\widehat{\beta}_1 - \beta_{10}}{\text{SE}(\beta_1)}$$

has a t-distribution on $n-2$ degrees of freedom.

$$\text{p-value } = P(|t_{n-2}| > t_0)$$

e.g. testing significance of regression for `cars` data:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

Recall the regression output for this example (which has $n = 50$ observations):

```
cars.lm <- lm(dist ~ speed, data = cars)
summary(cars.lm)$coefficients
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -17.58      6.758   -2.60 1.23e-02
## speed           3.93      0.416    9.46 1.49e-12
```

$\widehat{\beta}_1 = 3.932$ and $\text{SE}(\beta_1) = 0.416$.

2.3 Hypothesis Testing e.g. cont'd

Therefore,

$$t_0 = \frac{3.932 - 0}{0.416} = 9.464$$

p-value $= P(|t_{48}| > 9.464) = 2(1 - P(t_{48} \leq 9.464)) = 1.49 \times 10^{-12}$

R command: `2*(1-pt(9.464, 48))`

2.4 $(1-\alpha)$ Confidence Intervals

- slope:

$$\widehat{\beta}_1 \pm t_{n-2,\alpha/2} s.e.$$

or

$$\widehat{\beta}_1 \pm t_{n-2,\alpha/2} \sqrt{\text{MSE}/S_{xx}}$$

`cars` e.g. (95% confidence interval)

$$3.932 \pm t_{48,.025}(0.416)$$

or

$$3.932 \pm 2.011(0.416) = 3.932 \pm 0.835$$

R command: `qt(.975, 48)`

- intercept:

$$\widehat{\beta}_0 \pm t_{n-2,\alpha/2} s.e.$$

e.g. $$-17.579 \pm 2.011(6.758) = -17.579 \pm 13.589$$