Chapter 2: Simple Linear Regression (Cont'd)

- 2.4 Confidence Intervals
- 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

2.4 Confidence Interval for Mean Response at Given x

- $\mu_0 = E[y|x_0] = \beta_0 + \beta_1 x_0$
  where x_0 is a possible value that the predictor could take.

- $\widehat{\mu}_0 = \widehat{E[y|x_0]} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ is a point estimate of the mean response at that value.

- To find a $1 - \alpha$ confidence interval for $\mu_0 = E[y|x_0]$, we need the variance of $\widehat{\mu}_0 = \widehat{E[y|x_0]}$:

$$\text{Var}(\widehat{\mu}_0) = \text{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0)$$

$$\widehat{\beta}_0 = \sum_{i=1}^{n} c_i y_i$$

$$\widehat{\beta}_1 x_0 = \sum_{i=1}^{n} b_i x_0 y_i.$$

2.4 Confidence Interval for Mean Response - Variance

- Therefore, $\widehat{\mu}_0 = \sum_{i=1}^n (c_i + b_i x_0) y_i$ so

$$\text{Var}(\widehat{\mu}_0) = \sum_{i=1}^n (c_i + b_i x_0)^2 \text{Var}(y_i)$$

$$= \sigma^2 \sum_{i=1}^n (c_i + b_i x_0)^2$$

$$= \sigma^2 \sum_{i=1}^n \left(\frac{1}{n} + \frac{(x_0 - \bar{x})(x_i - \bar{x})}{S_{xx}} \right)^2$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right).$$

- Estimator: $\widehat{\text{Var}(\widehat{\mu}_0)} = \text{MSE} \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$.

2.4 Confidence Interval for Mean Response - Formula and Example

- The confidence interval is then

$$\widehat{\mu}_0 \pm t_{n-2,\alpha/2}\sqrt{\text{MSE}\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

- e.g. Compute a 95% confidence interval for the expected stopping distance when the speed is 15 mph.

$$x_0 = 15, \widehat{\mu}_0 = -17.58 + 3.93(15) = 41.41$$

$$n = 50, \bar{x} = 15.4, S_{xx} = 1370, \text{MSE} = 4.93$$

2.4 Confidence Interval for Mean Response - Example (cont'd)

- Interval:

$$41.41 \pm t_{48,.025}\sqrt{4.93\left(\frac{1}{50}+\frac{(15-15.4)^2}{1370}\right)}$$

$$= 41.41 \pm 2.01\sqrt{0.1} = 41.41 \pm 0.63$$

$$= (40.77, 42.04)$$

- **Interpretation:** under the conditions of the road and car that were present when the data were collected, we can be approximately 95% confident that the mean stopping distance is between 40.77 feet and 42.04 feet when the speed of the car is 15 mph.

2.4 Confidence Interval for Mean Response - R Code

- R code:

```
predict(cars.lm, newdata=data.frame(speed = 15),
        interval="confidence")

##    fit lwr upr
## 1   41   37   46
```

- Exercise: Write your own R function to compute this interval.

2.5: Predicted Response at Given x

- If a new observation were to be taken at x_0, we would predict it to be $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon$ where ε is unobservable independent noise.

- y_0 is a random variable with mean $\mu_0 = \beta_0 + \beta_1 x_0$ which we estimated earlier with $\widehat{\mu}_0$.

- Our goal: find an interval of the form $\widehat{\mu}_0 \pm c$ which contains the true value of y_0 with probability $1 - \alpha$.

- That is, find c for which

$$P(\widehat{\mu}_0 - c \leq y_0 \leq \widehat{\mu}_0 + c) = 1 - \alpha.$$

2.5: Predicted Response at Given x

- Basic algebraic manipulations lead to

$$P(-c \leq \widehat{\mu}_0 - \mu_0 + \varepsilon \leq c) = 1 - \alpha. \tag{1}$$

- The central term above is normally distributed with mean 0 and variance: $\text{Var}(\widehat{\mu}_0) + \sigma^2 = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$.

- Estimator: $\widehat{\text{Var}(y_0 - \widehat{y}_0)} = \text{MSE} \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$. Thus,

$$\frac{\widehat{\mu}_0 - \mu_0 + \varepsilon}{\sqrt{\text{MSE}(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}}$$

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

2.5: Predicted Response at Given x

- Dividing all terms in Equation (1) by $\sqrt{\text{MSE}(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}})}$ gives

$$P\left(\frac{-c}{\sqrt{\text{MSE}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}} \leq T \leq \frac{c}{\sqrt{\text{MSE}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}}\right) = 1-\alpha$$

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

- The prediction interval is then

$$\widehat{\mu}_0 \pm t_{n-2,\alpha/2}\sqrt{\text{MSE}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

2.5: Predicted Response at Given x - Example

- `Cars` e.g.: Compute a 95% prediction interval for the stopping distance for a single new observation where the speed is 15 mph.

- As before,

$$x_0 = 15, \widehat{\mu}_0 = 41.41$$

$$n = 50, \bar{x} = 15.4, S_{xx} = 1370, \text{MSE} = 4.93.$$

There is a 95% probability that the stopping distance for a randomly selected car travelling at 15 mph will be in the interval:

$$41.41 \pm t_{48,.025}\sqrt{4.93\left(1 + \frac{1}{50} + \frac{(15-15.4)^2}{1370}\right)}$$

$$= 41.41 \pm 2.01\sqrt{5.03} = 41.41 \pm 4.51 = (36.9, 45.92)\text{ft}.$$

2.5: Predicted Response at Given x - R Code

- R code:

```
predict(cars.lm, newdata=data.frame(speed = 15),
        interval="prediction")

##    fit lwr upr
## 1   41   10   73
```

- Exercise: Write your own R function to compute this interval.

Degrees of Freedom

- A random sample of size n coming from a normal population with mean μ and variance σ^2 has n degrees of freedom: $Y_1, \ldots, Y_n$.
- Each linearly independent restriction reduces the number of degrees of freedom by 1.
- $\sum_{i=1}^{n} \frac{(Y_i-\mu)^2}{\sigma^2}$ has a $\chi^2_{(n)}$ distribution.
- $\sum_{i=1}^{n} \frac{(Y_i-\bar{Y})^2}{\sigma^2}$ has a $\chi^2_{(n-1)}$ distribution. (Calculating $\bar{Y}$ imposes one linear restriction.)
- $\sum_{i=1}^{n} \frac{(Y_i-\widehat{Y}_i)^2}{\sigma^2}$ has a $\chi^2_{(n-2)}$ distribution. (Calculating $\widehat{\beta}_0$ and $\widehat{\beta}_1$ imposes two linearly independent restrictions.)
- Given the x_i's, there are 2 degrees of freedom in the quantity $\widehat{\beta}_0 + \widehat{\beta}_1 x_i$.
- Given x_i and $\bar{Y}$, there is one degree of freedom: $\widehat{y}_i - \bar{Y} = \widehat{\beta}_1(x_i - \bar{x})$.

Unbiased estimator for σ^2

$$\frac{1}{n - \text{\# parameters estimated}} \sum_{i=1}^{n} e_i^2$$

$$= \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$$

* n observations $\Rightarrow$ n degrees of freedom

* 2 degrees of freedom are required to estimate the parameters

* the residuals retain $n - 2$ degrees of freedom

Analysis of Variance: Breaking down (analyzing) Variation

- The variation in the data (responses) is summarized by

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \text{TSS}$$

(Total sum of squares)

- 2 sources of variation in the responses:

1. variation due to the straight line relationship with the predictor
2. deviation from the line (noise)

- $\underbrace{y_i - \bar{y}}_{\text{deviation from data center}} = \underbrace{y_i - \widehat{y}_i}_{\text{residual}} + \underbrace{\widehat{y}_i - \bar{y}}_{\text{difference: line and center}}$

Analysis of Variance (cont'd)

From the previous slide, we have

$$y_i - \bar{y} = e_i + \hat{y}_i - \bar{y}.$$

Squaring both sides and summing over all i from 1 to n gives

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum (e_i + \hat{y}_i - \bar{y})^2 = \sum e_i^2 + \sum (\hat{y}_i - \bar{y})^2$$

since

$$\sum e_i \hat{y}_i = 0 \quad \text{and } \bar{y} \sum e_i = 0.$$

Therefore,

$$S_{yy} = SSE + \sum (\hat{y}_i - \bar{y})^2 = SSE + SS_R.$$

The last term is the regression sum of squares.

Relation between SS_R and β_1

- We saw earlier that

$$SSE = S_{yy} - \widehat{\beta}_1 S_{xy}$$

- Therefore,

$$SS_R = \widehat{\beta}_1 S_{xy} = \widehat{\beta}_1{}^2 S_{xx}$$

Note that, for a given set of x's SS_R depends only on $\widehat{\beta}_1$.

$$MS_R = SS_R/d.f. = SS_R/1$$

(1 degree of freedom for slope parameter)

Expected Sums of Squares

- $$E[SS_R] = E[S_{xx}\hat{\beta}_1{}^2]$$

$$= S_{xx}\left(\text{Var}(\hat{\beta}_1) + (E[\hat{\beta}_1])^2\right)$$

$$= S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right)$$

$$= \sigma^2 + \beta_1^2 S_{xx}$$

$$= E[MS_R]$$

  Therefore, if $\beta_1 = 0$, then SS_R is an unbiased estimator for σ^2.
- Development of $E[\text{MSE}]$:

$$E[S_{yy}] = E[\sum(y_i - \bar{y})^2]$$

$$= E[\sum y_i^2 - n\bar{y}^2] = \sum E[y_i^2] - nE[\bar{y}^2]$$

Development of E[MSE] (cont'd)

Consider the 2 terms on RHS, separately:

1. term:

$$E[y_i^2] = \text{Var}(y_i) + (E[y_i])^2$$

$$= \sigma^2 + (\beta_0 + \beta_1 x_i)^2$$

$$\sum E[y_i^2] = n\sigma^2 + n\beta_0^2 + 2n\beta_0\beta_1\bar{x} + \sum \beta_1^2 x_i^2$$

2. term:

$$E[\bar{y}^2] = \text{Var}(\bar{y}) + (E[\bar{y}])^2$$

$$= \sigma^2/n + (\beta_0 + \beta_1\bar{x})^2$$

$$nE[\bar{y}^2] = \sigma^2 + n\beta_0^2 + 2n\beta_0\beta_1\bar{x} + n\beta_1^2\bar{x}^2$$

Development of E[MSE] (cont'd)

$\Rightarrow$

$$E[S_{yy}] = (n-1)\sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2$$

$$= E[SS_T]$$

$\circ$

$$E[SSE] = E[S_{yy}] - E[SS_R]$$

$$= (n-1)\sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2 - (\sigma^2 + \beta_1^2 S_{xx})$$

$$= (n-2)\sigma^2$$

$$E[\text{MSE}] = E[SSE/(n-2)] = \sigma^2$$

Another approach to testing $H_0 : \beta_1 = 0$: ANOVA

- Under the null hypothesis, both MSE and MS_R estimate σ^2.
- Under the alternative, only MSE estimates σ^2.

$$E[MS_R] = \sigma^2 + \beta_1^2 S_{xx} > \sigma^2$$

- A reasonable test is

$$F_0 = \frac{MS_R}{\text{MSE}} \sim F_{1,n-2}$$

- Large $F_0 \Rightarrow$ evidence against H_0.
- Note $t_\nu^2 = F_{1,\nu}$ so this is really the same test as

$$t_0^2 = \left[\frac{\widehat{\beta}_1}{\sqrt{\text{MSE}/S_xx}}\right]^2 = \frac{\widehat{\beta}_1{}^2 S_{xx}}{\text{MSE}} = \frac{MS_R}{MSE}$$

The ANOVA table

Source	df	SS	MS	F
Reg.	1	$SS_R = \widehat{\beta}_1{}^2 S_{xx}$	$MS_R = \widehat{\beta}_1{}^2 S_{xx}$	$\frac{MS_R}{\text{MSE}}$
Error	$n-2$	$SSE = S_{yy} - \widehat{\beta}_1{}^2 S_{xx}$	$SSE/(n-2)$	
Total	$n-1$	S_{yy}		

`cars` example:

```
anova(cars.lm)   # R code

## Analysis of Variance Table
##
## Response: dist
##           Df Sum Sq Mean Sq F value  Pr(>F)
## speed      1  21185   21185    89.6 1.5e-12 ***
## Residuals 48  11354     237
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: there is strong evidence that the slope is nonzero.
(recall that the t-statistic for testing $\beta_1 = 0$ had been 9.46= $\sqrt{89.6}$)

- Exercise: Write an R function to compute these ANOVA quantities.

2.6 R^2 - Coefficient of Determination

- $R^2 =$ is the fraction of the response variability explained by the regression:

$$R^2 = \frac{SS_R}{S_{yy}}$$

- $0 \leq R^2 \leq 1$. Values near 1 imply that most of the variability is explained by the regression.

- `cars` data:

$$SS_R = 21185 \text{ and } S_{yy} = 3.25 \times 10^4$$

so

$$R^2 = \frac{21185}{32539} = .651$$

2.6 R^2 - Coefficient of Determination

R output: (Look for "Multiple R-squared" below.)

```
summary(cars.lm)

##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29.07  -9.53  -2.27    9.21  43.20
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.579      6.758   -2.60    0.012 *
## speed          3.932      0.416    9.46  1.5e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15 on 48 degrees of freedom
## Multiple R-squared:  0.651,Adjusted R-squared:  0.644
## F-statistic: 89.6 on 1 and 48 DF,  p-value: 1.49e-12
```

2.6 R^2 - Coefficient of Determination - Properties

- Another interpretation:

$$E[R^2] \doteq \frac{E[SS_R]}{E[S_{yy}]} = \frac{\beta_1^2 S_{xx} + \sigma^2}{(n-1)\sigma^2 + \beta_1^2 S_{xx}}$$

$$= \frac{\beta_1^2 \frac{S_{xx}}{n-1} + \frac{\sigma^2}{n-1}}{\sigma^2 + \beta_1^2 \frac{S_{xx}}{n-1}} \doteq \frac{\beta_1^2 \frac{S_{xx}}{(n-1)}}{\sigma^2 + \beta_1^2 \frac{S_{xx}}{(n-1)}}$$

for large n. (Note: this differs from the textbook.)

Properties of R^2

- Thus, R^2 increases as
  1. S_{xx} increases (x's become more spread out)
  2. σ^2 decreases

- Cautions
  1. R^2 does not measure the magnitude of the regression slope.
  2. R^2 does not measure the appropriateness of the linear model.
  3. A large value of R^2 does not imply that the regression model will be an accurate predictor.

2.12 Maximum Likelihood Estimation

- Normal assumption is required:

$$f(y_i|x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(y_i-\beta_0-\beta_1 x_i)^2}$$

$$\text{Likelihood: } L(\beta_0, \beta_1, \sigma) = \prod f(y_i|x_i)$$

$$\propto \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}$$

- Maximize with respect to β_0, β_1, and σ^2.

- (β_0, β_1): Equivalent to minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

(i.e. Least-squares) σ^2: SSE/n, biased.

Gauss-Markov Property

- Linearity Property:

$$S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x})$$

  so $\widehat{\beta}_1$ is linear in the y_i's.

- Gauss-Markov Theorem: Among linear unbiased estimators for β_1 and β_0, $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are best (i.e. they have smallest variance).

- Exercise: Find the expected value and variance of $\frac{y_1 - \bar{y}}{x_1 - \bar{x}}$. Is it unbiased for β_1? Is there a linear unbiased estimator with smaller variance?

2.11 - Regression through the Origin

- intercept = 0

$$y_i = \beta_1 x_i + \varepsilon$$

- Max. Likelihood and L-S $\Rightarrow$

$$\text{minimize} \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2 \rightsquigarrow \widehat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$e_i = y_i - \widehat{\beta}_1 x_i$$

$$SSE = \sum e_i^2$$

- Unbiased Variance Estimator:

$$\widehat{\sigma}^2 = \text{MSE} = \frac{SSE}{n-1}$$

2.11 - Regression through the Origin

- Properties of $\widehat{\beta}_1$:

$$E[\widehat{\beta}_1] = \beta_1$$

$$\text{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum x_i^2}$$

- $1 - \alpha$ C.I. for β_1:

$$\widehat{\beta}_1 \pm t_{n-1,\alpha/2}\sqrt{\frac{\text{MSE}}{\sum x_i^2}}$$

2.11 - Regression through the Origin

- $1-\alpha$ C.I. for $E[y|x_0]$:

$$\widehat{\mu}_0 \pm t_{n-1,\alpha/2}\sqrt{\frac{\text{MSE}x_0^2}{\sum x_i^2}} \quad \text{since} \quad \text{Var}(\widehat{\beta}_1 x_0) = \frac{\sigma^2}{\sum x_i^2}x_0^2$$

- $1-\alpha$ P.I. for y, given x_0:

$$\widehat{\mu}_0 \pm t_{n-1,\alpha/2}\sqrt{\text{MSE}\left(1+\frac{x_0^2}{\sum x_i^2}\right)}$$

2.11 - Regression through the Origin

R code:

```
cars.lm <- lm(dist ~ speed  - 1, data = cars)
summary(cars.lm)

##
## Call:
## lm(formula = dist ~ speed - 1, data = cars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -26.18 -12.64  -5.46   4.59  50.18
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## speed    2.909      0.141    20.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16 on 49 degrees of freedom
## Multiple R-squared:  0.896,Adjusted R-squared:  0.894
## F-statistic:  423 on 1 and 49 DF,  p-value: <2e-16
```

2.11 - Regression through the Origin

- Predicting the stopping distance for a car travelling 15 mph:

```
predict(cars.lm, newdata=data.frame(speed = 15),
            interval="prediction")
##    fit lwr upr
## 1   44   11   77
```

2.10 Hazards of Regression

- Extrapolation: predicting y values outside the range of observed x values. There is no guarantee that a future response would behave in the same linear manner outside the observed range.
  e.g. Consider an experiment with a spring. The spring is stretched to several different lengths `x` (in cm) and the restoring force `F` (in Newtons) is measured:

```
x    F
3    5.1
4    6.2
5    7.9
6    9.5
> spring.lm <- lm(F~x - 1,data=spring)
```

2.10 Hazards of Regression - Extrapolation Example (cont'd)

```
> summary(spring.lm)

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x   1.5884     0.0232    68.6  6.8e-06
```

The fitted model relating `F` to `x` is

$$\widehat{F} = 1.58x$$

Can we predict the restoring force for the spring, if it has been extended to a length of 15 cm?

2.10 Hazards of Regression

- High leverage observations: x values at the extremes of the range have more influence on the slope of the regression than observations near the middle of the range.

- Outliers can distort the regression line. Outliers may be incorrectly recorded OR may be an indication that the linear relation or constant variance assumption is incorrect.

2.10 Hazards of Regression

- A regression relationship does not mean that there is a cause-and-effect relationship.

  e.g. The following data give the number of lawyers and number of homicides in a given year for a number of towns:

no. lawyers	no. homicides
1	0
2	0
7	2
10	5
12	6
14	6
15	7
18	8

  Note that the number of homicides increases with the number of lawyers. Does this mean that in order to reduce the number of homicides, one should reduce the number of lawyers?

2.10 Hazards of Regression

- Beware of nonsense relationships.

  e.g. It is possible to show that the area of some lakes in Manitoba is related to elevation.

  Do you think there is a real reason for this? Or is the apparent relation just a result of chance?