36-401 Class Notes and Exercise for 9/12/00

Howard Seltman

Sampling Distributions, continued

$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 X_i,\ \sigma^2) \quad i \perp\!\!\!\perp j$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \sim \mathrm{MVN}_n \left(\begin{bmatrix} \beta_0 + X_1\beta_1 \\ \beta_0 + X_2\beta_1 \\ \vdots \\ \beta_0 + X_n\beta_1 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \right) \equiv \mathrm{MVN}_n \left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n \right)$$

—aside—

$$b_1 = \frac{\sum \left[(X_i - \bar{X})(Y_i - \bar{Y}) \right]}{\sum (X_i - \bar{X})^2} = \frac{1}{\sum (X_i - \bar{X})^2} \left\{ \sum \left[(X_i - \bar{X}) Y_i \right] - \sum \left[(X_i - \bar{X}) \bar{Y} \right] \right\}$$

$$= \frac{\sum \left[(X_i - \bar{X}) Y_i \right]}{\sum (X_i - \bar{X})^2}$$

$$\mathrm{Var}\left(\frac{(X_1 - \bar{X})}{\sum (X_i - \bar{X})^2} Y_1 \right) = \sigma^2 \frac{(X_1 - \bar{X})^2}{\left[\sum (X_i - \bar{X})^2 \right]^2}$$

$$\Rightarrow \mathrm{Var}(b_1) = \sigma^2 \frac{\sum (X_i - \bar{X})^2}{\left[\sum (X_i - \bar{X})^2 \right]^2} = \frac{\sigma^2}{\mathrm{TSS}_X}$$

—end aside—

$$b_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{TSS_X} \right) \Rightarrow \frac{b_1 - \beta_1}{\sigma / \sqrt{TSS_X}} \sim \mathrm{N}(0, 1)$$

$$\frac{b_1 - \beta_1}{s_e / \sqrt{TSS_X}} \sim \mathrm{t}_{n-2}$$

What We Get When Assumptions Are Met

- Estimates of coefficients (b_0 and b_1) are unbiased
- Estimate of residual variance (σ_e^2) is unbiased
- Above estimates are minimum variance (most efficient)
- Standard errors of the coefficients are unbiased
- t and F tests are valid

Testing and Confidence Intervals for Simple Regression

- t-test and meaning of p-value

$$\frac{b-\beta_h}{\mathrm{estSE}_b} \sim t_{n-K} \qquad \text{p-value} = 2(1-\mathrm{pt}(\mathrm{abs}(t),\ n-K))$$

- SS, MS, df

```
> rslt_lm(myy~myx)
> anova(rslt)
 Analysis of Variance Table

Response: myy

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
myx        1  11.29121 11.29121 99.85358 2.149778e-05
Residuals  7   0.79154  0.11308
```

- "Omnibus" F-test
  F-statistic: 99.85358 on 1 and 7 degrees of freedom, the p-value is 2.149778e-05

- Confidence intervals for coefficients
  Need, e.g., $t_{(0.05,\mathrm{df})} = \mathrm{qt}(1-0.05/2, \mathrm{df})$
  CI is $b_1 \pm t_{(0.05,\mathrm{df})}\mathrm{estSE}_{b_1}$

- Prediction interval for $Y \mid X = X_i$
  CI is $\mu_i \pm t_{(0.05,\mathrm{df})}\mathrm{estSE}_{Y_i^*}$

- Confidence interval for $E(Y \mid X = X_i)$
  for CI drop "1" from formula for $\mathrm{estSE}_{Y_i^*}$

- Joint confidence regions

```
  Correlation of Coefficients:
       (Intercept)
  SQFT -0.9586
```

Regression Through the Origin

```
spill_read.table("spill.dat")
summary(lm(OilLost ~ cnt - 1, spill))
```

- Misleading R^2 and F

Breakout Session for 9/12/00

Exercise 1

The prevalence of respiratory symptoms was recorded for 9 groups of subjects exposed to differing levels of dust in their work environment. Dust exposure was measured as particles/ft^3/year scaled by 10^6. The direct outcome variable is "relative risk", the ratio of symptom prevalence at a given exposure level to symptom prevalence in the absence of workplace dust.

```
> dust_data.frame(exposure=c(75,100,150,350,600,900,1300,1650,2250),
              RR=c(1.10,1.05,0.97,1.90,1.83,2.45,3.70,3.52,4.16))
> summary( lm(RR ~ exposure, dust) )
Coefficients:  Value Std. Error t value Pr(>|t|)
(Intercept)   1.0360 0.1688     6.1358  0.0005
   exposure   0.0015 0.0002     9.9927  0.0000

Residual standard error: 0.3363 on 7 degrees of freedom
Multiple R-Squared: 0.9345
F-statistic: 99.85 on 1 and 7 degrees of freedom, the p-value is 2.15e-05
```

Is the intercept different from 0? What is your guess at the "correct" value of the intercept just from reading the experimental protocol? Why? What is the t-value to test if the intercept is different from that value? Do you think this represents a "significant" t-value?

Exercise 2

```
> summary( lm(RR-1 ~ exposure, dust) )
Residuals:
     Min    1Q   Median      3Q    Max
 -0.3406 -0.14 -0.05667 0.02818 0.6623

Coefficients:  Value Std. Error t value Pr(>|t|)
(Intercept)   0.0360 0.1688     0.2132  0.8373
   exposure   0.0015 0.0002     9.9927  0.0000

Residual standard error: 0.3363 on 7 degrees of freedom
Multiple R-Squared: 0.9345
F-statistic: 99.85 on 1 and 7 degrees of freedom, the p-value is 2.15e-05
```

(Note that -1 on the left side of the tilde in the regression formula just subtracts one from the Y variable.) Is the intercept significantly different from zero in the second regression run? What is a 95% CI (confidence interval) for b_0?

Exercise 3

```
> summary( lm(RR-1 ~ exposure-1, dust) )
Call: lm(formula = RR - 1 ~ exposure - 1, data = dust)
Residuals:
     Min      1Q   Median      3Q    Max
 -0.3598 -0.1086 -0.06121 0.04207 0.6663

Coefficients:
           Value Std. Error t value Pr(>|t|)
exposure  0.0016  0.0001    16.2955  0.0000

Residual standard error: 0.3156 on 8 degrees of freedom
Multiple R-Squared: 0.9708
F-statistic: 265.5 on 1 and 8 degrees of freedom, the p-value is 2.024e-07
```

Is the third regression better than the second? What do we need to do to see if the model is adequate? How do you interpret the model substantively, i.e. how could you explain this to a statistically naive researcher in Environmental Studies?