

STAT 252

Using more than one explanatory variable in regression

Why more than one explanatory variable?

- Reduces Bias
 - ▣ We can “control for the effects of” one x-variable when attempting to assess the relationship between the other x-variable and the y variable.
 - Ex: If homes with more bedrooms tend to also have more bathrooms, we can estimate the effect of bedrooms on the price, **net** the effect of bathrooms.
- Reduce SSE
 - ▣ With the residuals smaller (by and large) the SSE and thus the SE for slope of the x-variable in question is reduced, making it more likely that we can spot a relationship, if there really is one.
 - Note: s is smaller when SSE is smaller and s is one of the two terms in every standard error equation in the regression context
 - The lower s is, the smaller these standard errors will be \leftrightarrow the more precision we have

Multiple Regression

□ Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$

- ▣ Where the errors (ε 's) are:
 - Normal
 - Mean zero – no matter what
 - Common SD (σ) – Note: spread around population relationship the same no matter what
 - Independent of each other
- ▣ Note: we cannot ever really know β 's ε 's or σ , but we can estimate them.
- ▣ Note: this is often called a “first order” model because
 - There are no quadratic, cubic or more complicated terms
 - A 2nd order model would allow for quadratic terms
 - There are no products of explanatory variables
 - A 2nd order model could include an explanatory variable that is the product of two other variables

Multiple Regression (2 Explanatory variables)

□ Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

□ We estimate the β 's ε 's and σ via “least squares”

▣ For any choice of estimated β 's we calculate

$$SSE = \sum (y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2))^2$$

and estimate β 's by choosing $\hat{\beta}'s$ that minimize SSE.

▣ Note: this is the best fitting plane to our data

□ More generally, we estimate β 's by minimizing SSE

▣ And if SSE is lowered, we are essentially able to explain more about the variability in our response variable

Choosing Explanatory Variables

- More details later, but for now ...
 - Use scatterplots and common sense to help figure out which x's should be included in the regression model
 - The more explanatory variables
 - the more complex the interpretation
 - the more likely you'll be including variables that really shouldn't be there
 - The fewer explanatory variables
 - the clearer the interpretation
 - The more likely you'll neglect a variable that should be there

Real Estate Example (Minitab)

The regression equation is
 $\text{price}(K) = 52.2 + 0.0841 \text{ sq ft} - 35.8 \text{ bedrooms} + 37.8 \text{ bathrooms}$

Predictor	Coef	SE Coef
Constant	52.18	25.06
sq ft	0.08414	0.01937
bedrooms	-35.76	13.90
bathrooms	37.77	13.32

S = 24.2560 R-Sq = 71.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	30713	10238	17.40	0.000
Residual Error	21	12355	588		
Total	24	43069			

- The regression equation
 - Coefficients
 - Interpretation
 - Fitted values/prediction
- Residuals, SSE, MSE, s and R²

Inference for β 's

- Estimates: $\hat{\beta}_i$ is used to estimate β_i
- Standard Errors (SE's) $S_{\hat{\beta}_i}$
 - Complex formula (take Stat 324)
 - Depends on
 - n – the larger the sample size, the smaller the SE
 - s – the smaller the errors, the smaller the SE
 - The joint distribution of the x variables (the shape of the scatterplots relating the x's to each other)
 - More on this later

Testing β 's

- Is there a relationship (having controlled for the effect of other explanatory variables)?
- Hypotheses
 - Null: $\beta_i=0$ (no relationship)
 - Alternative: $\beta_i \neq 0$ (there is a relationship)
- Test statistic:
$$t = \frac{\hat{\beta}_i - 0}{S_{\hat{\beta}_i}}$$
 - df = n-(k+1)
- P-values and conclusions as before
 - Note: can have one-tailed tests

CIs for β 's

- We are $100 \times (1 - \alpha)\%$ sure that β_i is in

$$\hat{\beta}_i \pm t_{\alpha/2} s_{\hat{\beta}_i}$$

- Note: $df = n - (k + 1)$
- Example: bathrooms
- Interpretation (be sure to include the conditioning!)
 - We are 95% sure that each additional bathroom changes the price of a home by ...
 - We are 95% sure that each additional bathroom changes the price of a home by ..., if the size of the home and the number of bedrooms is unchanged.

R^2 and Adjusted R^2

- The multiple coefficient of determination, R^2
 - Describes the proportion of variability of y that can be considered explained by the x variables

$$R^2 = 1 - \frac{SSE}{SSTotal} = \frac{SSTotal - SSE}{SSTotal} = \frac{SSReg}{SSTotal}$$

- Adjusted R^2
 - Takes into account the number of explanatory variables

$$R_{adj}^2 = 1 - \left(\frac{n-1}{n-(k+1)} \right) \frac{SSE}{SSTotal} = 1 - \left(\frac{n-1}{n-(k+1)} \right) (1 - R^2)$$

- A fairer way to compare across models of different sizes

Inference for Individual Slopes (cont)

- In Minitab
 - Stat \rightarrow Regression \rightarrow Regression
- T-tests for $\beta = 0$ versus $\beta \neq 0$ are provided
- CIs to estimate the slope
- Examples:
 - If we control for the rainfall levels, is there a relationship between year and corn yield?
 - If we control for the size of a pumpkin (circumference and height) what is the relationship between color (1=fair, 2=good, 3=excellent) and price?

Model Utility Test

- Strange definition of **model utility**: "Are any of the explanatory variables useful predictors of the response?"
- Suppose we want to see if **any** of the k explanatory variables is related to the response variable
 - We could do k separate t-tests
 - But that would increase the chance of at least one Type I error (thinking there is a relationship if, in fact, there is none)
 - Combined procedure via the ANOVA table part of the Regression output

F-test (in regression)

- Null: $\beta_1 = \beta_2 = \dots = \beta_k = 0$ (no relationship between any explanatory variable and the response)
- Alternative: at least one $\beta \neq 0$ (there at least one explanatory variable that is related to the response)
- Test Statistic
 - ▣ From Minitab
 - ▣ By Hand $\rightarrow F = \frac{MS_{Reg}}{MSE} = \frac{R^2 / k}{(1 - R^2) / (n - (k + 1))}$
- P-value from Software (next slide)
- Conclusion as always
- Example: $R^2 \rightarrow F \rightarrow P\text{value}$

F-test (in regression)

- P-value from Software
 - ▣ Graph \rightarrow Probability Distribution Plot
 - View Probability
 - ▣ Distribution F
 - ▣ "numerator df"=k
 - ▣ "denominator df"=n-(k+1)
 - ▣ Shaded Area
 - X-value
 - Right Tail
 - ▣ P-value is Upper-tail area
 - The chance, assuming the null hypothesis is true, that the F-ratio would be at least as large as the one we did see ... just by luck alone.