STA 302 / 1001

Note Title 11/10/2011

My new mantra : Answers must be

unambiguous , accurate – for any marks

deep – for a good mark

Back to Breakdown example - unequal variance

Untransformed data, no weights

fitted equation: $\hat{y} = 1886.2 - 54.0\,x$

($\hat{y}$: time; x: voltage)

Untransformed data, with weights $\propto \frac{1}{\text{var at each } x_i}$

fitted equation: $\hat{y} = 119.7 - 3.13\,x$

($\hat{y}$: time; x: voltage)

Equation is completely — because large values that occurred

at values of x at which there is large variance are down weighted

Looked at standardized residuals for weighted LS.
↳ use these always for weighted LS rather than raw residuals.

Student. vs. $\hat{y}$ – increasing, possibly curvature, approx. constant variance.
r_i vs $\hat{y}$

Normal quantile plot – distribution of standardized residuals is right-skewed.

Weighted LS fixed non-constant variance problem but not other problems

Need a transformation to fix other problems

If assumptions of normality and independence hold, the same general procedures for tests and CIs (t & F) apply as in unweighted least squares

eg $t_{obs} = \dfrac{b_{1w}}{s_w / \sqrt{\sum w_i (x_i - \bar{x})^2}}$

where $WSXX$ = "weighted SXX" $= \sum w_i (x_i - \bar{x})^2$

and $s_w^2 = \dfrac{\sum w_i (y_i - b_{0w} - b_{1w} x_i)^2}{n-2}$

Under $H_0: \beta_1 = 0$, t_{obs} is an observation from t_{n-2} distribution

Multiple Regression

- more than one explanatory variable

Why do this? - multiple x's arise naturally

- Want to control for some x's to consider effect on y of other x's over and above control variables
- Want to fit a polynomial
- Want to compare regression line for 2 or more groups

Multiple Regression

p = number of explanatory variables in the model

Model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i,$

$i = 1, \ldots, n$

Need to estimate $p+2$ parameters ($p+1$ β's + σ^2)

Need $n > p+2$ observations (minimum)

In matrix notation, $Y = X\beta + e$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

"Design matrix"

$$X = \underbrace{\begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}}_{p+1}{}_{n \times (p+1)}$$

Gauss-Markov Assumptions: $E(\underset{\sim}{\varepsilon} \mid X) = \underset{\sim}{0}$

$\text{Var}(\underset{\sim}{\varepsilon} \mid X) = \sigma^2 I$

For inference need $\underset{\sim}{\varepsilon}$ have a multivariate normal distribution

Least squares estimates: $\underset{\sim}{b} = (X'X)^{-1} X' Y$

$$\underset{\sim}{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}$$

Residuals $\hat{e} = Y - \hat{Y} = Y - Xb = Y - X(X'X)^{-1}X'Y$

$= (I - H) Y$

Estimate of σ^2: $S^2 = MSE = \frac{\sum \hat{e}_i^2}{\text{df error}} = \frac{\hat{e}'\hat{e}}{\text{df error}}$

df error $= n - p - 1$

$RSS = \hat{e}'\hat{e} = Y'(I-H)Y$

rank $(I-H)$
$=$ rank (I) $-$ rank (H)
$= n - (p+1)$

assuming columns of X are linearly independent

Need $E(RSS) = (n-p-1)\sigma^2$ to show S^2 is unbiased

estimator for σ^2

Proof: Same as SLR proof

except $\text{trace}(I - H) = \text{trace}(I) - \text{trace}(H)$

$= n - \text{trace}\left(\underbrace{X'X (X'X)^{-1}}_{I_{p+1}} \right)$

$= n - (p+1)$ //

$H = X(X'X)^{-1}X'$

$\text{trace}(H) = \text{trace}\left(X'X (X'X)^{-1} \right)$

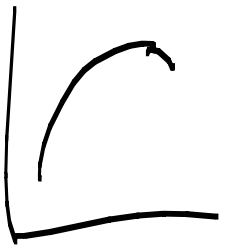
Rainfall Example

Y: corn yield (bushels / acre) in 6 US states, 1890 - 1927

Explanatory variable: rainfall (inches)

Straight line not adequate

Non-monotonic relationship, curvature



Fit quadratic model:

$$Y = \beta_0 + \beta_1 \text{rain} + \beta_2 \text{rain}^2 + e$$

- multiple regression: 2 predictors: rain, rain^2
- still linear regression (linear in β's)

$p = 2$

X matrix has 3 columns

$n = 38$

df for error $= 38 - 3 = 35$

df for SST $= n - 1 = 37$

df for SSReg $= 2 = p$

Fitted equation: $\hat{\text{yield}} = -5.015 + 6.004\,\text{rain} - 0.229\,\text{rain}^2$

$b_0 \qquad b_1 \qquad b_2$

In general, in multiple regression interpret coefficient of a predictor variable (β_j or its est. b_j) as change in mean value of y associated with a one unit change in the predictor variable (X_j) with all other variables held constant.

For rainfall example, this is impossible.

Can say for this example

If rain increases from 8" to 9", estimated change in mean of yield is $6.004 - .229 (81 - 64)$
~ 2.1 bu/acre.

If rain increases from 14" to 15", estimated change in mean of yield is

$$6.004 - .229\left(15^2 - 14^2\right)$$

$$= -.637 \text{ bu/acre}$$

Is coefficient of quadratic term = 0?

Test: $H_0: \beta_2 = 0$ vs $H_a: \beta_2 \neq 0$

Test statistic: $t_{obs} = \dfrac{b_2}{s.e.(\hat{\beta}_2)}$

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

3rd diagonal entry of $\widehat{Var}(\hat{\beta})$ is $\widehat{Var}(\hat{\beta}_2)$

Substitute s^2 (=MSE) for σ^2 and take square root to get $s.e.(\hat{\beta}_2)$

$$t_{obs} = \frac{-.229}{.0886} = -2.6$$

.0886 → from SAS

Under H_0, this is an observation from a t distribution with 35 df
$= n-p-1$

$p = .0140$

Conclude that we have some evidence that coefficient of rain^2 is not 0 so we should keep it in the model; even with linear term in model.

Note: Test of $H_0: \beta_j = 0$ gives an indication of whether or not the j^{th} predictor variable statistically significantly contributes to the estimation/prediction of Y over and above the other predictor variables.

That is, the assumes the other variables

are in the model

R^2 in Multiple Regression

$$R^2 = \frac{SS_{Reg}}{SST}$$

For rainfall example, quadratic regression $R^2 = \frac{209.}{704.} = .297$

called the "coefficient of multiple determination"

It's not the square of a correlation anymore.

It's guaranteed to increase if you add more predictor variables.

Why?