Notes

$s_x^2 = \frac{1}{n-1}\sum(x_i-\bar{x})^2 \rightarrow s_x = \sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}$

$\text{Correlation}_{xy} = r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}$ → sample covariance

$\sum e_i^2 = \sum(\hat{y}_i - y_i)^2 = SSE$ (sum, squared error) has $(n-2)$ df.

$s^2 = \sum(y_i-\bar{y})^2/(n-1)$ = unbiased estimator of variance of single population

s_{res}^2 = error term variance = $MSE = \frac{SSE}{n-2} = \sum e_i^2/n-2$. Note $E[MSE] = \sigma_{res}^2$

↳ $s_{res} = \sqrt{MSE}$

$t_{b_n} = b_n/s(b_n)$

$s^2(b_1) = MSE/\sum(x_i-\bar{x})^2$

$s^2(b_0) = MSE[1/n + \bar{x}^2/\sum(x_i-\bar{x})^2]$

$SSR = \sum(\hat{Y}_i-\bar{Y})^2$, w/ df = 1 → $MSR = SSR/1$

$SSTO = \sum(Y_i-\bar{Y})^2$, w/ df = $n-1$ → $SSTO = SSR + SSE$

$SSE = \sum(Y_i-\hat{Y}_i)^2$, w/ df = $n-2$

$F^* = MSR/MSE$

$R^2 = SSR/SSTO = 1 - SSE/SSTO$

Notes 2

Regression

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{systematic}} + \underbrace{\varepsilon_i}_{\text{unsystematic}}$$

$$E[Y_i] = \beta_0 + \beta_1 X_i$$

$$\sigma^2[Y_i] = \sigma^2[\beta_0 + \beta_1 X_i + \varepsilon_i] = \sigma^2$$

Notes on ε_i:

1) $E[\varepsilon_i | X_i] = 0$
2) $\sigma^2[\varepsilon_i | X_i] = \sigma^2$ → Homoskedastic
3) $\sigma[\varepsilon_i, \varepsilon_j | X_i] = 0 \ \forall \ \varepsilon_i$
4) $\sigma[X_i | \varepsilon_i] = 0$
5) Assume $E[\varepsilon] \neq 0$ → include any error in intercept

know This

OLS

Choose b_1, b_2 s.t. minimize $\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$

$$Q = \sum [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i) \longrightarrow \text{Set equal to zero \& solve for normal equations.}$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum X_i (Y_i - \beta_0 - \beta_1 X_i)$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Normal Equations:

$$\sum Y_i = n b_0 + b_1 \sum X_i \rightarrow \frac{1}{n} \sum Y_i = b_0 + \frac{b_1}{n} \sum X_i \rightarrow \bar{Y} = b_0 + b_1 \bar{X}$$

Then, $\frac{\partial Q}{\partial \beta_1} = -\sum X_i (Y_i - b_0 - b_1 X_i) = 0$

$$\Rightarrow \sum X_i (Y_i - (\bar{Y} - b_1 \bar{X}) - b_1 X_i) = 0$$

$$\Rightarrow \sum X_i Y_i - \sum X_i (\bar{Y}) + \sum X_i b_1 \bar{X} - \sum b_1 X_i^2 = 0$$

$$\Rightarrow \sum X_i Y_i - \bar{Y} \sum X_i + b_1 \bar{X} \sum X_i - b_1 \sum X_i^2 = 0$$

$$\Rightarrow \sum X_i Y_i - \bar{Y} n \bar{X} + b_1 n \bar{X}^2 - b_1 \sum X_i^2 = 0$$

$$\Rightarrow \sum X_i Y_i - n \bar{Y} \bar{X} = b_1 \sum X_i^2 - b_1 n \bar{X}^2$$

$$\Rightarrow \sum X_i Y_i - n \bar{Y} \bar{X} = b_1 [\sum X_i^2 - n \bar{X}^2]$$

$$\Rightarrow \sum (X_i Y_i) - \sum \bar{Y} \bar{X} = b_1 [\sum X_i^2 - n \bar{X}^2]$$

$$\Rightarrow \sum (X_i - \bar{X})(Y_i - \bar{Y}) = b_1 [\sum (X_i - \bar{X})^2]$$

$$\Rightarrow b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}}$$

Note: $E[b_0] = \beta_0$, $E[b_1] = \beta_1$

** Note: $\sum X_i (Y_i - b_0 - b_1 X_i) = 0$

Notes 3

Gauss-Markov

OLS estimators are BLUEs

Best
Linear
Unbiased → $E[b_x] = \beta_x$
Estimators

} No better linear estimators

Conditions:

1) Population process must be linear → $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
2) Data is from random sample
3) No perfect multicollinearity
4) Zero conditional Mean → $E(\varepsilon_i) = 0$ regardless of x_i
5) Homoskedastic
6) No serial correlation

- Value of response variable = response
- $E[Y_i]$ = mean response
  - Mean of prob. dist. of $[Y|X]$
    - $\hat{Y}$ = point estimate of $[\text{mean response}|X) = [E[Y_i]|X]$
- Fitted value $(\hat{Y}_i)$ ∀ observed values (Y_i)

- residual = $e_i = Y_i - \hat{Y}_i$
  - $\Sigma e_i = 0 = \bar{e}$
    - Σe_i^2 is minimized
  - $S_{xe} = \frac{\Sigma(x_i - \bar{x})(e_i - \bar{e})}{(n-1)}$ ← covariance of x & e

$$= \frac{\Sigma x_i e_i - \bar{x}\Sigma e_i}{(n-1)} = \frac{\Sigma x_i e_i}{n-1} = \frac{\Sigma x_i (Y_i - b_0 - b_1 x_i)}{n-1} = 0$$

Note: residual ≠ error!
error is $\varepsilon_i = Y_i - E[Y_i]$ w/ $E[Y_i]$ is unknown & can't be calculated

- Point $(\bar{x}, \bar{y})$ is always on regression line.

- $\Sigma Y_i = \Sigma \hat{Y}_i$
  - $\Sigma Y_i = \Sigma \hat{Y}_i + \Sigma e_i$ (→ 0)
    - $\therefore \bar{Y} = \bar{\hat{Y}}$

- $S_{\hat{Y}e} = \frac{\Sigma(\hat{Y}_i - \bar{\hat{Y}})(e_i - \bar{e})}{n-1} = 0$

Notes 4

Estimation of error term variance

Recall $s^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$ is estimate of σ^2 of single population

↳ LD called "mean square"

* Then, note that, for the regression model, the variance associated w/ each observation Y_i is the same as that associated w/ each error term ε_i.

$\therefore SSE = \sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2$ → (Residual sum of squares)

w/ $df = n-2$, two lost to estimating β_0 & β_1 for $\hat{Y}_i$

Then the mean square error =

$$s^2 = \frac{SSE}{n-2} = MSE$$

Note: MSE is an unbiased estimator of σ^2: $E[MSE] = \sigma^2$

$\therefore s = \sqrt{MSE}$ is estimator of standard deviation σ.

Normal Error Regression Model

Up until now, the error terms were not necessarily normally distributed.

However, to establish interval estimates & perform tests (t/F-test)

we must assume:

ε_i are iid $\sim N(0, \sigma^2)$

↳ Recall, can center @ 0 by adding value to intercept term.

Notes 5

Hypothesis Tests

Note: $\beta_1 = 0$ does not imply no association between X & Y,
$\beta_1 = 0$ implies all probability distributions @ any level of X are identical w/ same mean & variance.
↳ Imagine horizontal regression line.
↳ W/ $\beta_1 \neq 0$, prob. dist. still normal w/ same variance, but mean depends on X.

Recall: $b_1 = \sum(x_i-\bar{x})(Y_i-\bar{Y}) / \sum(x_i-\bar{x})^2$, $E(b_1)=\beta_1$, $\sigma^2(b_1) = \sigma^2 / \sum(x_i-\bar{x})^2$

Then, $s^2(b_1) = MSE / \sum(x_i-\bar{x})^2$

Then we can express b_1 as a linear combination of Y_i's

$b_1 = \sum k_i^{\beta_1} Y_i$ w/ $k_i^{\beta_1} = \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2}$

(Note: b/c k_i are a func of x_i, & x_i are constant, b_1 is lin. comb. of Y_i)

Then, recall that $b_1 = Cov(x,y) / Var(x)$

→ $b_1 = \sum(Y_i-\bar{Y})(x_i-\bar{x}) / \sum(x_i-\bar{x})^2$

↳ $\sum(x_i-\bar{x})(Y_i-\bar{Y}) = \sum Y_i(x_i-\bar{x}) - \bar{Y}\sum(x_i-\bar{x})$
→ $= \sum Y_i(x_i-\bar{x}) - \bar{Y}[n\bar{x}-n\bar{x}]^0$
→ $= \sum Y_i(x_i-\bar{x})$

$\therefore b_1 = \frac{\sum Y_i(x_i-\bar{x})}{\sum(x_i-\bar{x})^2} = \sum k_i^{\beta_1} Y_i$ //

* Properties of $k_i^{\beta_1}$'s (numerator of $k_i^{\beta_1}$)

Know these ↳

1) $\sum k_i^{\beta_1} = 0$ → b/c $\sum(x_i-\bar{x}) = \sum x_i - \sum\bar{x} = n\bar{x} - n\bar{x} = 0$

2) $\sum x_i k_i^{\beta_1} = 1$ → $\sum x_i k_i^{\beta_1} = \sum x_i(x_i-\bar{x}) / \sum(x_i-\bar{x})^2 = \frac{\sum x_i^2 - \sum x_i\bar{x}}{\sum x_i^2 - 2\sum x_i\bar{x} + \sum\bar{x}^2}$
→ $= (\sum x_i^2 - n\bar{x}^2) / (\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2) = (\sum x_i^2 - n\bar{x}^2) / (\sum x_i^2 - n\bar{x}^2) = 1$ //

Notes 6

* Properties of $k_i^{b_1}$ cont.

know this

3) $\sum (k_i^{b_1})^2 = 1/\sum (x_i - \bar{x})^2$

↳ $\sum k_i^2 = \left[\frac{\sum (x_i-\bar{x})}{\sum (x_i-\bar{x})^2}\right]^2 = \frac{\sum (x_i-\bar{x})^2}{\left[\sum (x_i-\bar{x})\right]^2} = \frac{1}{\sum (x_i-\bar{x})^2}$ //

∴ b/c b_1 is a linear combination of Y_i, and Y_i is normally distributed, b_1 is a normally distributed random variable.

• b_1 is an unbiased point estimator of β_1:

$E[b_1] = E[\sum k_i^{b_1} Y_i]$

$= \sum k_i^{b_1} E[Y_i] = \sum k_i^{b_1} E[\beta_0 + \beta_1 x_i] = \beta_0 \sum k_i^{b_1} + \beta_1 \sum k_i^{b_1} x_i$ ($\sum k_i^{b_1} \to 0$, $\sum k_i^{b_1} x_i \to 1$)

∴ $E[b_1] = \beta_1$ //

• $\sigma^2(b_1) = \sigma^2/\sum (x_i-\bar{x})^2$:

$\sigma^2(b_1) = \sigma^2(\sum k_i Y_i) = \sum k_i^2 \sigma^2(Y_i) = \sum k_i^2 \sigma^2 = \sigma^2 \sum k_i^2$

$= \sigma^2/\sum (x_i-\bar{x})^2$ //

Then, b/c $b_1 \sim N(\,)$, $(b_1-\beta_1)/s(b_1) \sim t_{(n-1)}$

◦ Confidence Interval for $\beta_1 = b_1 \pm t^*_{(1-\alpha/2,\, n-2)}\, s(b_1)$

Note: $t^* = b_1/s(b_1)$ — w/ $s(b_1) = \sqrt{MSE/\sum (x_i-\bar{x})^2} = \sqrt{s^2(b_1)}$

• Confidence Interval for β_0 :

$s^2(b_0) = MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i-\bar{x})^2}\right]$

Notes 7

Properties of $k_i^{\beta_0}$

$b_0 = \sum k_i^{\beta_0} Y_i$ w/ $k_i^{\beta_0} = \frac{1}{n} - \frac{\bar{X}(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$

b_0 is a linear combination of Y_i's

$$b_0 = \sum k_i^{\beta_0} Y_i = \frac{\sum Y_i}{n} - \frac{\bar{X}\sum Y_i(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} = \bar{Y} - \bar{X}\underbrace{\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}}_{b_1}$$

$b_0 = \bar{Y} - \bar{X} b_1$

1) $\sum k_i^{\beta_0} = 1$:

$$\sum k_i^{\beta_0} = \sum \left[\frac{1}{n} - \frac{\bar{X}(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} \right] = \sum \frac{1}{n} - \frac{\bar{X}\sum (X_i - \bar{X})}{\sum (X_i - \bar{X})^2} = 1 - \frac{\bar{X}(n\bar{X} - n\bar{X})}{\sum (X_i - \bar{X})^2}, \quad n\bar{X} - n\bar{X} = 0$$

$\therefore \sum k_i^{\beta_0} = 1$ //

2) $\sum k_i^{\beta_0} X_i = 0$:

$$\sum k_i^{\beta_0} X_i = \sum \left[\frac{X_i}{n} - \frac{\bar{X} X_i (X_i - \bar{X})}{\sum (X_i - \bar{X})^2} \right] = \sum \frac{X_i}{n} - \frac{\sum \bar{X} X_i (X_i - \bar{X})}{\sum (X_i - \bar{X})^2} = \bar{X} - \bar{X} \frac{\sum X_i (X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$

$$\rightarrow = \bar{X} - \bar{X} \frac{\sum X_i (X_i - \bar{X})}{\sum (X_i - \bar{X})(X_i - \bar{X})} = \bar{X} - \bar{X} \frac{\sum X_i (X_i - \bar{X})}{\sum X_i (X_i - \bar{X}) - \bar{X}\sum (X_i - \bar{X})}$$

$$\rightarrow = \bar{X} - \bar{X} \frac{\sum (X_i - \bar{X})}{\sum (X_i - \bar{X}) - \bar{X}(n\bar{X} - n\bar{X})}, \quad \sum (X_i - \bar{X}) = 1, \quad n\bar{X} - n\bar{X} = 0$$

$= \bar{X} - \bar{X} = 0$ //

B/c b_0 is a linear combination of Y_i & Y_i is independent, normally distributed, then b_0 is a normally distributed random variable.

b_0 is unbiased:

$$E[b_0] = E\left[\sum k_i^{\beta_0} Y_i\right] = \sum k_i^{\beta_0} E[Y_i] = \sum k_i^{\beta_0} E[\beta_0 + \beta_1 X_i]$$

$$= \beta_0 \underbrace{\sum k_i}_{1} + \beta_1 \underbrace{\sum k_i X_i}_{0} = \beta_0 \text{ //}$$

$\sigma^2(b_0)$:

$$\sigma^2(b_0) = \sigma^2\left(\sum k_i^{\beta_0} Y_i\right) = \sum k_i^2 \sigma^2(Y_i) = \sum k_i^2 \sigma^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right]$$

Notes 8

Then, $(b_0 - \beta_0)/s(b_0) \sim t_{(n-2)}$

$CI = b_0 \pm t_{(1-\alpha/2, n-2)} * s(b_0)$ w/ $s(b_0) = MSE\left[\frac{1}{n} + \frac{\bar{x}}{\sum(x_i-\bar{x})^2}\right]$

Interval Estimate

$E[Y_h]$ = mean response when $X = X_h$ → Y_h is for one level of X

↳ $\hat{Y}_h = b_0 + b_1 X_h$

↳ $E[\hat{Y}_h] = E[Y_h]$

↳ $\sigma^2(\hat{Y}_h) = \sigma^2\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(x_i-\bar{x})^2}\right]$

↳ $s^2(\hat{Y}_h) = MSE\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(x_i-\bar{x})^2}\right]$ *

- Y_h are normally distributed ∵

$E[\hat{Y}_h] = E[b_0 + b_1 X_h] = E(b_0) + X_h E[b_1] = \beta_0 + \beta_1 X_h$

Notes on Y_h ∵

- confidence intervals and errors in inference of mean response are calculated by repeated samples @ each level of X, b/c x_i is known.
- Variance of Y_h is smallest when $X_h = \bar{X}$

- Confidence limits apply when a single mean response is being estimated. (Not sensitive to normality of errors)

Notes

Intervals for a mean response

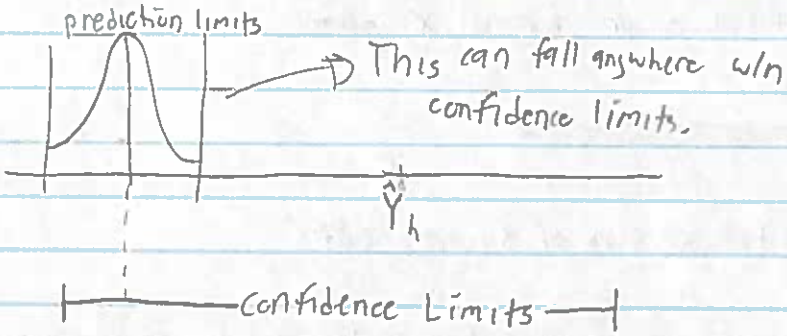
New Observation of Y w/ corresponding level X: $Y_{h(new)}$

Note: $\hat{Y}_{h(new)}$ is an estimate of $Y_{h(new)}$, the prediction of a single outcome drawn from distribution of Y. Will be wider b/c predicting a single value.

$\hat{Y}_h$ is an estimate of $E\{Y_h\}$ - estimate of mean of distribution of Y, given a level of X.

Note: Mean of distribution of Y estimated by $\hat{Y}_h$

But, Variance of distribution of Y estimated by MSE



Therefore, prediction limits for $Y_{h(new)}$ must account for both:

1) Variation in location of distribution
2) Variation w/ the probability distribution

Then: $\hat{Y}_h \pm t_{(1-\alpha;n-2)} * s\{pred\}$ is C.I.

$$\sigma^2(pred) = \sigma^2(Y_{h(new)} - \hat{Y}_h) = \sigma^2(Y_{h(new)}) + \sigma^2(\hat{Y}_h) = \sigma^2 + \sigma^2(\hat{Y}_h)$$

w/ $\hat{Y}_h = \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$

$$\therefore s^2(pred) = MSE + s^2(\hat{Y}_h) = MSE\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

This means prediction intervals widen as X_h deviates from $\bar{X}$.

Note: Prediction ≠ confidence limits:
Confidence Interval - prediction on a parameter
Prediction Interval - Statement on a value of a random variable

In R:
1) Build lm() model
2) predict(lm_model, df(Variable = x), interval = 'prediction')
→ Gives fit, lwr, upr for all X given.

~~Confidence Bands for Regression~~

ANOVA - Partition sum of squares i.d.f.

Recall: Variance of a single population estimated by sample variance:
$(s^2 = \Sigma(Y_i - \bar{Y})^2/n-1)$
w/ numerator = sum of squares
[Note: denom = (n-1) b/c lose 1 df in estimating μ_y w/ $\bar{Y}$.]

Then, Total Sum of Squares:
$$SSTO = \Sigma(Y_i - \bar{Y})^2$$

Error/Residual Sum of Squares:
$$SSE = \Sigma(Y_i - \hat{Y}_i)^2 = \Sigma e_i$$

Regression Sum of Squares:
$$SSR = \Sigma(\hat{Y}_i - \bar{Y})^2$$
◇ Measure of what part of variation in Y_i is associated w/ regression line.

estimate $\bar{Y}$ → $n-1$ df; estimate $\bar{Y}_i$, start w/ $\hat{Y}_i$ → 1 df; estimate β_0, β_1 → $n-2$ df

Then, $SSTO = SSR + SSE$

Note: SSR can be viewed as how "smart" the regression model is compared to naïevely picking the mean.
SSE can be viewed as how much "smarter" the model could be.

Sum of squares divided by df is called mean squares

$MSE = SSE/n-2$
$MSR = SSR/1$

ANOVA TABLE

	SS	df	MS
Reg	$\Sigma(\hat{Y}_i - \bar{Y})^2$	1	$\Sigma(\hat{Y}_i - \bar{Y})^2/1$
Error	$\Sigma(Y_i - \hat{Y}_i)^2$	$n-2$	$\Sigma(Y_i - \hat{Y})^2/n-2$
Tot	$\Sigma(Y_i - \bar{Y})^2$	$n-1$	

Recall: $E(MSE) = \sigma^2$ → Mean of sampling distribution of MSE
$E(MSR) = \sigma^2 + \beta_1^2 \Sigma(x_i - \bar{x})^2$

When $\beta_1 = 0$, then MSE & MSR have same sampling distribution location

If $\beta_1 \neq 0$, then sampling distribution of $MSR > MSE$
b/c $\beta_1^2 \Sigma(x_i - \bar{x})^2 > 0$

Therefore, MSE v. MSR comparison a good test for $\beta_1 = 0$:

If $MSR \approx MSE$, $\beta_1 \approx 0$
If $MSR >> MSE$, $\beta_1 \neq 0$

↳ $F^* = MSR/MSE$

Note: This is an upper-tail test

↳ Large F^* support $H_a: \beta_1 \neq 0$
↳ $F^* \approx 1$ support $H_0: \beta_1 = 0$
↳ $F^* > F_{(1-\alpha, n-2)}$ then do not reject.

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

↳ R^2 is a measure of goodness-of-fit of regression line.
R^2 is the proportionate reduction of total variance associated w/ the use of predictor variable X.

* R^2 is the fraction of sample variation of Y explained by the variation in X.

$$R^2 = r_{xy}^2 = r_{y\hat{y}}^2 \quad \text{w/ } r = \text{correlation}$$

Notes: $R^2 = 1$ if perfect fit.
$R^2 = 0$ if $B_1 = 0$.

Remember, high R^2 does not imply good predictions can be made, nor does it imply the regression line is a good fit.
$R^2 = 0$ does not imply X & Y not related.

Note: $t^2 = F$

Running the model in R

lm_mod = lm(data$y ~ data$X)
Summary(lm_mod) will give residual quarters, coefficients w/ p-values, standard error, R^2, F-stat.
We can plot the data & regression line: plot(x,y)
abline(lm_mod)
summary(lm_mod)$fstatistic gives F-Stat. (Big is good).

anova(lm_mod) gives ANOVA table w/ F-value for all explanatory variables

Summary(lm_mod)$coefficients to get coefficients.
residuals(lm_mod)
fitted(lm_mod)

Non-Consistency of Error Variance (Heteroskedasticity)

Plot residuals against X. → Plot (x, residuals(lm_mod))
Look for random distribution → residualPlot(lm_mod)
(Note: same as plotting residuals against $\hat{Y}$)

Raise flags if variance not consistent.
→ i.e. megaphone shape.

Test using Brown-Forsythe: → For homoskedasticity

① Divide data into 2 groups, one w/ high X-values, one w/ low.
② Test for $\sigma_1^2 = \sigma_2^2$ w/ σ_i^2 = variance of residuals : H_0
(2.5) Tests residuals around the median

Note: This is a variation of levene, which centers around mean.
∴ In R: leveneTest(lm_mod, center = median, group = as.factor(grouping))
→ Look @ p-value

Breusch-Pagan Test: → For homoskedasticity

Assumes error terms are independent & normally distributed &
error term ε_i denoted by σ_i^2 is related to X s.t.:

$$\log_e(\sigma_i^2) = \gamma_0 + \gamma_1 X$$

and σ_i^2 either increases or decreases w/ level of X, depending on sign of γ_1.

Then, H_0: $\gamma_1 = 0$ → homoskedastic.
H_a: $\gamma_1 \neq 0$

In R:
→ Studentized BP
bptest(lm_mod)

ncvtest(lm_mod) → Not studentized

Outliers

Recall: Standardize a random variable by:

$$\left(X - \mu_X / \sigma_X\right)$$

Then, standardize a residual by:

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = e_i / \sqrt{MSE}$$

Studentize a residual by:

$$e_i^* = e_i / \sqrt{MSE_{(-i)}}$$

Normality of Error Terms

Shapiro-Wilk Test for Normality:

H_0: Normally Distributed data

H_a: Non-Normally distributed

In R: shapiro.test(lm_mod$residuals)

↳ extracts studentized residuals

Can check visually w/ QQ-Plots: qqplot(lm_mod)

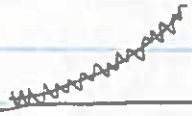
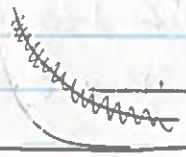
↳ Plots theoretical v. empirical distribution.

↳ If they closely match, assume empirical ≈ theoretical.

Transformations

Can transform y/x if SLR model not a good fit.

Common examples:

Regression pattern	Transformation
	$x' = \sqrt{x}$
	$x' = x^2$ or $x' = \exp(x)$
	$x' = 1/x$ or $x' = \exp(-x)$

Can also introduce constant if x is near zero:

i.e. $x' = 1/x \rightarrow x' = 1/(x+k)$ w/ k=constant

* Notes If heteroskedastic, usually transform dependent variable.

LD

Residual pattern	Transformation
	$y' = \sqrt{y}$
	$y' = \log_{10} y$
	$y' = 1/y$

Box-Cox

Family of Power transformations:

$$Y' = Y^{\lambda}$$

Box-Cox is MLE of λ
↳ Numerical search for best $\hat{\lambda}$

① Calculate new Y_i:

$$W_i = \begin{cases} k_1(Y_i^{\lambda} - 1) & \text{if } \lambda \neq 0 \\ k_2(\log_e Y_i) & \text{if } \lambda = 0 \end{cases}$$

$$k_1 = 1/\lambda k_2^{(\lambda-1)}$$

$$k_2 = \left(\prod Y_i\right)^{1/n} = \text{geometric mean.}$$

② Regress W_i on X to get SSE_{λ}

③ Minimize $SSE_{\hat{\lambda}}$

In R: boxcox(lm-toluca)

Multiple Linear Regression Intro

Prediction

Functional Form: Predict Y using $\hat{Y} = \hat{f}(x)$ → in which case, only care about accurate predictions

Inference

Want to understand how Y changes in response to X's

↳ In which case, $\hat{f}$ not a black-box

Estimating f

Parametric Approach:

Two-step approach:

① Making assumption(s) about functional form of f.

② Use training data to fit the model (estimate β_k's)

↳ Reduces problem to estimating set of parameters.

↳ Note: This is an easier method, but can result in over-simplifying

- Non-parametric models, i.e. SVM, Random Forest
  - ↳ Not restricted to assumptions, but often require large datasets
    - ↳ Not interpretable

Supervised: Data is "labeled" w/ response variable Y.

Unsupervised: No response variable Y. i.e. clustering

Recall: $MSE = \sum (Y_i - \hat{f}(x))^2 / n$

Want to minimize test MSE, which is model trained on different data than it is tested on.

↳ Note: Model with lowest training MSE does not always have lowest testing MSE

↳ Often due to overfitting

Two-Predictor MLR

$$E[Y_i] = b_0 + b_1 X_{1i} + b_2 X_{2i}$$

Then, b_1 = change in mean response $E[Y_i]$ per unit increase in X_1, holding all else constant (ceteris paribus)

Note that X_1 & X_2 have additive effects, meaning they do not interact.

β_1, β_2 are referred to as partial regression coefficients

(p-1) Predictor MLR → (p = parameters)

w/ ε_i iid $\sim N(0, \sigma^2)$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{p-1} X_{i} + \varepsilon_i$$

This is called a first-order regression model w/ p-1 predictor variables

Also is: $Y_i = \sum_{k=0}^{p-1} \beta_k X_{ik} + \varepsilon_i$ w/ $X_{i0} = 1$

This forms a hyperplane.

OR, in Linear Algebra terms:

$$\underset{n\times 1}{\hat{Y}} = HY \quad \text{w/} \quad \underset{n\times n}{H} = X(X'X)^{-1}X'$$

$$X'Xb = X'Y' \rightarrow \hat{Y} = Xb$$

MLR ANOVA

	SS	df	MS
Regression	SSR	$p-1$	$\frac{SSR}{p-1}$
Error	SSE	$n-p$	$\frac{SSE}{n-p}$
Total	SSTO	$n-1$	

Notes: You can't unexplain error.
R^2 only increases w/ more predictors

$F^* = MSR/MSE$, if $\leq F(1-\alpha, p-1, n-p)$ fail to reject H_0.

$R^2 = SSR/SSTO = 1 - SSE/SSTO$

$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO}$

↳ Decreases w/ more variables.

$\frac{b_k - \beta_k}{s(b_k)} \sim t(n-p)$ w/ $CI = b_k \pm t_{(1-\alpha/2, n-p)}^* s(b_k)$

$\hat{Y}_h \pm t_{(1-\alpha/2, n-p)}^* s(\hat{Y}_h)$ w/ $s^2(\hat{Y}_h) = X_h' s^2(b) X_h$ (probably not needed)

Remember - can not extrapolate outside the bands of your training data. The predictor variables jointly define the region.

Diagnostics

In R: cor(data) will give matrix of correlations, want high values in column/row for response variable

residualPlots(lm_mod) will also give test statistics for all predictor variables,

↳ When testing residuals against predictors, done by adding predictor squared to the model & testing for significance.

Notes 20

Extra Sum of Squares

[or marg. increase in SSR]

Extra Sum of Squares measures the marginal reduction in the error sum of squares when one or several predictor variables are added.

- If model is $Y \sim X_1$, SSR is SSR for X
- If model is $Y \sim X_1 + X_2 + X_3 + \cdots + X_{p-1}$, SSR is $SSR(X_1, X_2 \ldots X_{p-1})$

Note: $SSR(X_1 \ldots X_{p-1}) \neq SSR(X_1) + SSR(X_2) + \cdots + SSR(X_{p-1})$

Rather: $SSR(X_1, X_2 \ldots X_{p-1}) = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) + \vdots + SSR(X_{p-1}|X_1 \ldots X_{p-2})$

anova(mod) in R

TYPE I MLR ANOVA TABLE → Sequentially added Variables

source	SS	df	MS
Regression	$SSR(X_1, X_2, X_3)$	3	$MSR(X_1, X_2, X_3)$
X_1	$SSR(X_1)$	1	$MSR(X_1)$
$X_2 \mid X_1$	$SSR(X_2 \mid X_1)$	1	$MSR(X_2 \mid X_1)$
$X_3 \mid X_1, X_2$	$SSR(X_3 \mid X_1, X_2)$	1	$MSR(X_3 \mid X_1, X_2)$
Error	$SSE(X_1, X_2, X_3)$	$n-4$	$MSE(X_1, X_2, X_3)$
Total	SSTO	$n-1$	

Notes: SSTO Always constant.

$SSR(X_1, X_2) = SSR(X_2, X_1)$

Notes 21

TYPE II MLR ANOVA TABLE → Anova(mod) in R

Regression	$SSR(X_1, X_2, X_3)$	3
$X_1 \mid X_2, X_3$	$SSR(X_1 \mid X_2, X_3)$	1
$X_2 \mid X_1, X_3$	" "	1
$X_3 \mid X_1, X_2$	" "	1
Error	$SSE(X_1, X_2, X_3)$	$n-4 = n-p$
Total	SSTO	$n-1$

Recall: Test significance of single $\beta_k = 0$ by t-test

Test significance $\forall \beta_k = 0$ w/ F-test

Alternative methods:

Partial F-test → $H_0: \beta_k = 0$

$$F^* = \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_F}$$

Note: overall F-test if testing $\beta_k = 0 \ \forall k$

Pf: $F^* = \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_F}$ w/ $>$ predictors

$$= \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{(n-3) - (n-4)} \div \frac{SSE(X_1, X_2, X_3)}{n-4}$$ (MSE)

w/ $SSE(X_1, X_2) - SSE(X_1, X_2, X_3) = SSR(X_3 \mid X_1, X_2)$

$\therefore F^* = MSR(X_3 \mid X_1, X_2) / MSE(X_3, X_2, X_1) = (t^*)^2$

Notes 22

Coefficient of Partial Determination

Measures the marginal contribution of one X-variable when all others in model already.

$$R^2_{Y_{new|old}} = \frac{SSR(new|old)}{SSE(old)}$$

i.e. adding X_1 if X_2 exists:

$$R^2_{Y1|2} = \frac{SSR(X_1|X_2)}{SSE(X_2)}$$

↳ It is the proportionate reduction of variation in Y remaining after <old> is included in the model that is gained by also including <new>.

Correlation, Uncorrelation, & Multicollinearity

Uncorrelated

$SSR(X_2|X_1) = SSR(X_2)$ iff X_1 & X_2 are perfectly uncorrelated.

Then R^2 values are simply additive.

Similarly, b_1 does not change w/ introduction of X_2.

Perfect correlation

X is no longer full rank,

In R:

lm(Y~X1+X2) w/ cor(X1, X2)=1

results in NA for coefficient estimate of X_2.

Note: Multicollinearity does not generally inhibit our ability to obtain a good fit.

↳ However, SE's for regression coefficients may grow exponentially, resulting in losing statistical significance of predictors.

(See MLR Slide 90/91 for ex.)

Notes 23

Multicollinearity Cont.

When predictor variables are correlated, the marginal contribution of any one predictor variable can vary significantly.

Note: There are times when adding a predictor that is highly correlated to an existing predictor can increase extra sum of squares term:

$SSR(x_2|x_1) > SSR(x_2)$

Then x_1 is a Supressor variable

Polynomial Regression Models

Used when: 1) True response is polynomial

2) True response unknown, but approximated to be polynomical.

Note: Should center x_i (express as deviation from $\bar{x}$) to avoid multicollinearity between x_i & x_i^2

Note: Be wary of predictors to a power greater than 3, this can be tough to interperate.

Hierarchical Rule - If you include x_i^n as a predictor, you should include x_i^k $\forall$ $0 \leq k \leq n$ (keep all lower orders too)

LD viewed as providing more "base" info on the shape.

LD holds for interaction terms too.

Interpretation: No longer same as single-order models.

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_1^2$$

$$\frac{\partial \hat{Y}}{\partial x_1} = b_1 + 2b_2 x_1$$

Note: A polynomial regression function to the $(n-1)$ order can perfectly fit the data

Notes 24

Ramsey Reset

Tests if quadratic term is suitable to add to model.

Done w/ t-test:

1) Regress Y on all predictors to get Y_i
2) Regress Y on all predictors and $\hat{Y}_i^n$ w/ $n \in \mathbb{R}$
3) Test $H_0: \beta_{Y_i^n} = 0$

Note: Can be run on multiple orders at once & can test all w/ partial F-test.

In R:

```
library(lmtest)
resettest(mod, power = 2, type = "fitted")
```

Interaction Terms

Effects are no longer additive

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2$$

Then one unit change in X_1 results in an $b_1 + b_3 X_2$ change in mean response

That means effects of changes in one predictor depend on the level of the other.

Two types of interactions:

1) Reinforcement - Positive coefficient on interaction term

2) Interference - Negative coefficient on interaction term.

Note: once again, centering the predictors can help avoid multicollinearity.

In R:

```
lm(Y ~ X1 + X2 + X1:X2, data = df)
```

Note X:Y: * gives all possible interactions between.

Notes 25

Qualitative Predictors (Indicative / Dummy)

Transform factored variables into (c-1) variables w/ c = # factors. Each new variable is binary (usually) depending on if the criteria is met for a given observation

Qualitative / Quantitative Interactions

$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2$ w/ $X_2 = 0/1$.

Then, if $X_2 = 0$, simply SLR w/ X_1

If $X_2 = 1$, slope & intercept shift.

$\hat{Y} = (b_0 + b_2) + (b_1 + b_3) X_1$

Note: If this shifted regression model intersects w/ the old one, called a disordinal interaction. Else, an ordinal interaction

Model Selection

- Kutner suggests 6-10 observations for every feature in model
- Too few predictors can lead to:
  - Biased coefficient estimates
  - Biased mean responses
  - Biased prediction intervals
  - Biased MSE

Notes 26

Model Selection Cont.

$R^2_p = 1 - \frac{SSE_p}{SSTO}$ w/ R^2_p monotonically non-decreasing

↳ Looks for diminishing returns to scale.

$R^2_{adj,p} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SSTO}$

Mallow's C_p

Concerned w/ MSE of the n fitted values for every subset of the model.

MSE involves $\hat{Y}_i - M_i$ w/ M_i = true mean response @ X_i level.

[Bias for $\hat{Y}_i$] = $E\{\hat{Y}_i\} - M_i$

where $E\{\hat{Y}_i\}$ is the expected value of i^{th} fitted value.

If $E\{Y_i\}$ differs from M_i, there is bias.

Then:

If $C_p >> p$, substantial bias

$C_p < p$, no bias

$C_p = p$ if all predictors used

↳ Parameters

Note: Need a solid (well selected) pool of predictors for effective use of Mallow's C_p.

This is because MSE must be an unbiased estimator of σ^2

Notes 27

AIC & SBC (BIC)

Akaike's Information Criteria (AIC) &

$$AIC_p = n\ln(SSE_p) - n\ln n + 2p$$

(1) (3) (3)

1) Decreases w/ p
2) Constant term
3) Penalize large p
4) Penalize large p & n

Schwartz's Bayesian Criterion &

$$SBC_p = n\ln(SSE_p) + n\ln(n) + \ln(n)p$$

(1) (2) (4)

This means SBC penalizes for high n.
Both penalize for high p.
Smaller values are good!

$Press_p$

Prediction Sum of Squares (PRESS) is a measure of how well the model fits.
Calculated same as SSE, but each $\hat{Y}_i$ obtained by regressing Y on predictors without observation i, and using that model for prediction.

LO Leave-one-out SSE

Small PRESS is good.

Best Subsets Method

SSE_p, R^2_p, $R^2_{a,p}$, C_p, AIC_p, SBC_p, $PRESS_p$

Run all → and see which subset performs best across the board.

This only identifies a "good" model(s) & further testing is still needed

Notes 28

Model Validation

Mean Squared Prediction Error (MSPR)

$$MSPR = \frac{\sum_{i=1}^{n^*}(Y_i - \hat{Y}_{i(1)})^2}{n^*}$$ ← Run only on test data

LO MSPR = MSE of testing

Want MSPR ≃ MSE

Added Variable Plots

Regress $Y \sim X_1 + \dots X_{n-1}$

and

$X_n \sim X_1 + \dots X_{n-1}$

Note: Always go through $(0,0)$

LOBLC regression line always passes through $(\bar{x}, \bar{y})$

& $\bar{e}_i = 0$

Plot residuals against each other.

Visible relationship means X_n probably valuable.

Leverage Value – Outlier Recognition for X

h_{ii} = Diagonal Values of H (hat matrix) = Leverage of X_i

large h_{ii} signifies outliers. Considered large if:

$h_{ii} > 2\bar{h} = \frac{2p}{n}$

$h_{ii} > .5$ is high

$.2 < h_{ii} < .5$ is moderate

$h_{ii} = 1$, Then $\hat{Y}_i = Y_i$

Notes 29

Outlier Recognition - Influential Observations

$$DFFITS = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$ = Difference between fitted values w/ & w/o X_i.

High DFFITS signify X_i is an outlier.

Cook's Distance

Similar to DFFITS but measures how much better the fit is for all points.

$$D_i = \frac{\sum (Y_j - \hat{Y}_{j(i)})^2}{pMSE}$$

↳ Big Value = Big Influence

* Y outliers *

Studentized Residuals

$r_i = e_i / s(e_i)$

↳ have constant variance, can compare to $\bar{r}$

Deleted Residual

$d_i = Y_i - \hat{Y}_{i(i)}$

$t_i = d_i / s(d_i)$ = Studentized deleted residual.

↳ Big means outlier

DFBETAS

Algebraically the same as Cook's

Multicollinearity Recognition

VIF - Formalized, widely accepted measure.

$$VIF = \frac{1}{1 - R_k^2}$$ w/ $R_k^2 = R^2$ when $X_k \sim X_1 + \dots + X_{p-2}$

VIF = 1 means X_k has no relationship to other predictors

VIF > 1 means correlation exists between X_k & one or more predictors

If $R_k^2 = 1$, VIF is unbounded.

$$\overline{VIF} = \frac{\sum VIF_p}{p-1}$$ → $\overline{VIF} >> 1$ is cause for concern

↳ Notes VIF > 10, signifies severe multicollinearity.

Notes 30

Autocorrelation

$X_{i,t}$ depends on $X_{i(t-1)}$ = Autocorrelation

Results in underestimated MSE variance and std. errors

Durbin-Watson test for Autocorrelation

$$D = \frac{\sum (e_t - e_{t-1})^2}{\sum e_i^2} \quad \text{w/ } e = \text{residual}$$

H_0: $\rho = 0$

$D > d_U$ fail to reject H_0

$D < d_L$ reject H_0

$d_L < D < d_U$?? Inconclusive

Quiz 1
MSAN 601

September 01, 2016

Question 1 (2 pts)

In the SLR model, the probability distribution of Y (i.e., Y_i) has the same mean and variance for all levels of X (i.e., X_i). True or False? Explain.

Answer

False. The variance remains the same, however the mean depends on the level of X.

Question 2 (2 pts)

The number of points above the fitted regression line is always equal to the number of points below it. True or False? Explain.

Answer

False. For the sum of residuals to be zero, there need not be equal number of points on either side of the regression line, but there has to be at least one.

Question 3 (1 pt)

In SLR, what does β_1 measure?

Answer

In SLR, β_1 is the slope of the regression line. It measures the average change in Y that the model predicts for a unit change in X.

Question 4 (4 pts)

In the context of an SLR model, prove the following:

1. $E[Y_i] = \beta_0 + \beta_1 X_i$

   $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
   $\therefore E[Y_i] = E[\beta_0 + \beta_1 X_i + \epsilon_i]$
   $\therefore E[Y_i] = E[\beta_0] + E[\beta_1 X_i] + E[\epsilon_i]$
   $\therefore E[Y_i] = E[\beta_0] + E[\beta_1 X_i] + E[\epsilon_i]$
   $\therefore E[Y_i] = \beta_0 + \beta_1 X_i + 0$
   $\because \beta_0, \beta_1, X_i$ are constants, $E[constant] = constant$ and $E[\epsilon_i] = 0$ (SLR model assumption)
   $\therefore E[Y_i] = \beta_0 + \beta_1 X_i$

2. $V(Y_i) = constant \ \forall \ i$

   $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
   $\therefore V[Y_i] = V[\beta_0 + \beta_1 X_i + \epsilon_i]$
   $\therefore V[Y_i] = V[\beta_0] + V[\beta_1 X_i] + V[\epsilon_i]$

$\therefore V[Y_i] = 0 + 0 + V[\epsilon_i]]$
$\because \beta_0, \beta_1, X_i$ are constants, $V[constant] = 0$ and $V[\epsilon_i] = \sigma^2 = constant$ (SLR model assumption)
$\therefore V[Y_i] = constant$

Question 5 (2 pts)

For the SLR model, $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, how many random variables are there. Explain.

Answer

For the SLR model, there are two random variables Y_i and ϵ_i. β_0, β_1 *and* X_i have fixed values whereas the values ϵ_i and Y_i take depend on their distributions. $\epsilon_i \sim N(0, \sigma^2)$ whereas the for the distribution of Y, mean depends on the level of X and variance is equal to that of ϵ_i.

Question 6 (4 pts)

Write out the normal error regression model and its assumptions (in English and math).

Answer

Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Assumptions:

1. $E[\epsilon_i | X_i] = 0$, the expected value of the error terms is zero (exogeneity).
2. $\sigma^2[\epsilon_i | X_i] = \sigma^2$ ($constant$), the variance of the error terms is constant (homoscedasticity).
3. $\sigma[\epsilon_i, \epsilon_j | X_i] = 0 \forall i \neq j$, error terms are independent and identically distributed and follow the normal distribution ($N \sim (0, \sigma^2)$).
4. $\sigma[X_i, \epsilon_i] = 0$, there is no correlation between the error terms and predictors.

Question 7 (3 pts)

Write in English and mathematically (using correct symbols) how a residual is computed. Given the regression line $\hat{Y} = 3 + 4.5X$, compute the residual for $(2, 11)$.

Answer

A residual is the difference between the actual and fitted value of the response variable.
$e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X$
$e_i = 11 - (3 + 4.5(2)) = -1$

Question 8 (1)

What does a negative value of β_1 indicate about the relation between X and Y?

Answer

It indicates that X and Y are inversely proportional or there is a negative correlation between the two.

Question 9 (6 pts)

Prove that $\bar{X}$ and $\bar{Y}$ always lie on the regression line.

Answer: SLR slides 32 - 36.

Quiz 2
MSAN 601

Question 1 (6 pts)

Write out the hypothesis test which tests for the statistical significance of β_1 for an SLR model. Be sure to include the null and alternate hypothesis, the critical value including degrees of freedom (two-tailed test) for $\alpha = 0.05$ and an interpretation of both possible results.

Answer

$H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$

$t^* = \frac{b_1}{s\{b_1\}}$, degrees of freedom = n-2

Interpretation of t^*:
If $|t^*| > t_{0.975;n-2}$, reject H_0. Thus, with 95% confidence we can say that β_1 is statistically significantly different from zero.
If $|t^*| \leq t_{0.975;n-2}$, do not reject H_0. Thus, with a 95% confidence, we fail to reject the null hypothesis and conclude that β_1 is not statistically significantly different from zero.

Question 2 (6 pts)

There is an SLR model with n = 17, $b_1 = 3.48$, $s\{b_1\} = 0.54$. Test the null hypothesis that β_1 is statistically significantly greater than 0, assuming $\alpha = 0.05$. Include null and alternate hypothesis, show all relevant calculations, and explicitly state your conclusion.

Answer

$H_0 : \beta_1 \leq 0$, $H_a : \beta_1 > 0$

$t^* = \frac{3.48}{0.54} = 6.44$

$t_\alpha = 1.753$ One-sided, $n - 2 = 15, \alpha = 0.05$

$t^* > t_\alpha$, thus, we reject the null hypothesis.

Thus, with a 95% confidence, we reject the null hypothesis and conclude that β_1 is statistically significantly greater than zero.

Question 3 (6 pts)

Fill in the blank values in the ANOVA table below

	Sum of Squares	df	Mean Square	F
Regression	**1494.465**	**1**	1494.465	**536.90**
Residual	**1664.54**	598	**2.78**	
Total	3159.009	**599**		

Question 4 (4 pts)

How can you use MSE and MSR to test whether $\beta_1 = 0$? Be complete in your answer including all relevant mathematical details.

Answer

$E[MSE] = \sigma^2$ and $E[MSR] = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$

We can use the F-test to test whether $\beta_1 = 0$.

$H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$

$F^* = \frac{MSR}{MSE}$, this is an upper-tail test as $MSE \geq MSR$. If $\beta_1 = 0$, F^* will be close to 1 and we fail to reject the null hypothesis. At higher values of F^*, however, we reject the null.

If $F^* \leq F(1 - \alpha; 1; n - 2)$, do not reject H_0 else reject H_0 with a confidence level of $1 - \alpha$.

Question 5 (6)

Prove: $SSTO = SSE + SSR$

Answer

See slide 142 in SLR slides.

Question 6 (3)

What is coefficient of determination? What are its limitations?

Coefficient of determination is a measure of the goodness of fit of the regression line. It is equal to the fraction of the sample variation in Y that can be explained by the variation in X.

Answer

Limitations:

1. High R^2 does not imply that good predictions can be made.
2. High R^2 does not imply that the estimated regression line is a good fit.
3. $R^2 \sim 0$ does not imply that X and Y are not related.

Question 7 (3)

How is R^2 calculated? How is its value related to Pearson's coefficient of correlation? What is the range of values it can take?

Answer

$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$
R^2 = square of Pearson's coefficient of correlation
$0 \leq R^2 \leq 1$

Question 8 (4)

What do you understand by a prediction interval? How does it differ from a confidence interval? (no math required)

Answer

A prediction interval is an estimate of an actual value of Y for given X. A confidence interval is an interval for the mean values of the Y for given X. Thus, a prediction interval is always wider than a confidence interval.

Question 9 (2)

The prediction intervals become wider as we predict Y for X values away from $\bar{X}$. True or false? Explain.

Answer

True. $s^2\{pred\} = MSE\Big[1 + \frac{1}{n} + \frac{(X_h - \bar{X})}{\sum(X_i - \bar{X})^2}\Big]$, from the formula, as X_h goes further away from $\bar{X}$, the prediction interval widens. This happens because as the difference between X_h and $\bar{X}$ increases, we can make less precise predictions.

Quiz 3
MSAN 601

September 14, 2016

Question 1 (2 pts)

If a histogram of residuals is generated, which assumption would one be attempting to verify? What would one expect as the center / mean value of the residuals of the histogram?

Answer

Normality of the error terms. We would expect the mean to be zero.

Question 2 (8 pts)

1. For what purpose would one use a Brown-Forsythe test?
2. What is it good at detecting (be specific).
3. Name a weakness of the Brown-Forsythe test?
4. How does the Brown-Forsythe test differ from the Breush-Pagan test?

Answer

1. To verify the homoscedasticity of the error terms.
2. Megaphone or inverted megaphone-shaped heteroskedasticity.
3. Grouping of residuals is subjective.
4. BP test assumes that the error terms are independent and normally distributed. BF test uses a two-sample t-test to verify if the two groups have equal median whereas BP uses chi-square test if variance of an error term is a function of X.

Question 3 (2 pts)

When generating a residual plot, give **two** different examples of what can be on the x and y axes. E.g., on the x-axis we plot ... and on the y-axis we plot ....

Answer

1. residuals (y) on fitted values ($\hat{y}$)
2. residuals (y) on predictor (x)

Question 4 (3 pts)

What is the difference between a standardized and studentized residual? Which plots—standardized or studentized residual plot—would you prefer to employ to detect outliers? Why?

Answer

A standardized residual is calculated as $\frac{e_i}{\sqrt{MSE}}$ while a studentized residual is calculated as $\frac{e_i}{\sqrt{MSE_{(-i)}}}$. Thus, a studentized residual measures the residual of the data point when its influence has been removed.

If the point is an outlier, MSE will decrease considerably making the residual more prominent. Thus, I would use studentized residual method to detect outliers.

Question 5 (2 pts)

Draw a sketch of what a QQ-plot would look like if the data were left-skewed.

Answer

See SLR slides.

Question 6 (3 pts)

How do outliers affect the regression line? What are some of the cases when you can ignore them?

Answer

Outliers can alter the regression line significantly by affecting its slope, for example. Outliers can be ignored when they are a result of an error in recording, reporting, a miscalculation or equipment malfunctioning.

Question 7 (4 pts)

1. What is the purpose of transforming a variable?
2. When trying to correct for heteroskedasticity **only** in an SLR model, which variable do you transform? What issues can you run into in making such a transformation, and how would you correct for them?

Answer

1. Transformation helps when SLR model doesn't fit the given data or the residuals violate an assumption.
2. Transform the response variable first. This may change the linear relation into a curvilinear one. Transforming the predictor will help resolve this.

Question 8 (4 pts)

Given the model

$$\log(y) = 2.19 + 0.17 \log(x)$$

if x is decreases by 1%, y would be expected to change by how much?

Answer

$y = e^{2.19+0.17log(x)}$
$y = e^{2.19}x^{0.17}$
x is decreased by 1%, $\therefore y_1 = e^{2.19}(0.99)x^{0.17}$

Thus, $y_1 = 0.99829y$, i.e., y decreases by $(1 - 0.99^{0.17}) \sim 0.17\%$, therefore $\sim 2\%$

Question 9 (2 pts)

Given a compelling argument both for and against using linear regression as a statistical learning tool.

Answer

Linear regression models are very simple to construct and interpret. However, they are restrictive in their assumptions, and may not be the most accurate when predicting in comparison to other machine learning techniques.

Question 10 (3 pts)

Given an MLR model with $n = 492$ and eight predictor variables, and $SSE = 4,000$, compute MSE.

Asnwer

MSE $= \frac{4000}{492-9} = 8.28$

Quiz 4
MSAN 601

September 21, 2016

Question 1 (3 pts)

What does 'the effect of predictor variables on the mean response is additive' mean in the context of first-order regression model?

Answer

In a first-order regression model, the effect of a predictor variable on the mean response does not depend on the level of other predictors in the model. In other words, the predictors do not interact and thus, are said to have an additive effect.

Question 2 (4 pts)

The GLRM in the matrix form can be written as $Y = X\beta + \epsilon$. Write the dimensions of Y, X, β *and* ϵ matrices. (Assume n data points and $p-1$ predictors.)

Answer

Y: n × 1
X: n × p
β: p × 1
ϵ: n × 1

Question 3 (3 pts)

How does adjusted R^2 help in determining if a predictor should be included in the model?

${R_a}^2 = 1 - \frac{n-1}{n-p}\frac{SSE}{SSTO}$

When a predictor is added to an existing model, SSE decreases and p increases. Thus, ${R_a}^2$ can either increase or decrease. A good predictor will help decrease the SSE enough to account for increased value of p. As a result, the second term of the above equation will decrease and thus, ${R_a}^2$ will increase. These predictors should be included in the model.

Question 4 (6 pts)

For an MLR model $Y = 0.5 + 0.85X_1 - 0.32X_2 + 0.33X_3$, interpret how a unit increase in each of the parameters $X_1,\ X_2,\ X_3$ affect the response Y.

Answer

Holding all other predictors at a constant level,

- For a unit increase in X_1, we expect the response Y to increase by 0.85.
- For a unit increase in X_2, we expect the response Y to decrease by 0.32.
- For a unit increase in X_3, we expect the response Y to increase by 0.33.

Question 5 (3 pts)

Explain the statement 'You cannot unexplain error'.

Answer

The statement is used in the context of linear regression. As more variables are added to the model, SSE decreases and R^2 increases. Additional variables always help to explain more error and not less. Thus, we cannot unexplain error.

Question 6 (6 pts)

For an MLR, $SSTO = 500$, $SSE(X_1, X_2) = 110$, $SSR(X_1|X_2) = 5$
Find: $SSR(X_2),\ SSR(X_1, X_2),\ SSE(X_2)$

Answer

multiple answers here due to typo in question

Question 7 (2 pts)

What is coefficient of partial determination?

Answer

A coefficient of partial determination is used to measure the marginal contribution of a predictor when other predictors are already included in the model.

Question 8 (2 pts)

What happens to the MLR model when perfectly correlated variables are included as predictors?

Answer

If perfectly correlated variables are included as predictors, the design matrix X isn't full rank and thus cannot be inverted.

Question 9 (4 pts)

What is multicollinearity? How does it affect an MLR model? How can it be detected?

Answer

Multicollinearity exists when two or more predictors in an MLR are highly correlated.
It can lead to an increase in the standard error of regression coefficients affecting their statistical significance. It also affects their interpretation.

Observing the relation between the predictors using a scatter plot is the simplest method to detect multicollinearity. One can also observe the marginal contribution of additional predictor in reducing SSE. If the predictors are correlated, the contribution will be very small.

Quiz 5
MSAN 601

September 30, 2016

Question 1 (2 pts)

How is centering of the predictor variable useful in case of a second order polynomial model with one predictor variable?

Answer

Centering helps reduce multicollinearity.

Question 2 (2 pts)

While using the Hierarchical Approach to fit polynomial models, what rule must be followed about the polynomial terms of a predictor?

Answer

If a higher order polynomial term of a predictor is used in the model, then all the lower order terms of that predictor must also be included in the model.

Question 3 (6 pts)

Explain what an interaction term is. Give a simple numerical example and explain it.

Answer

An interaction term is used in a regression model when there exists a relation between the predictors such that the influence of two predictors on the response is not additive.

A model with two predictors and an interaction term takes a form of

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

The change in the mean response with a unit increase in X_1 when X_2 is held constant is $\beta_1 + \beta_3 X_2$. The increase in X_1 depends on the level of X_2.

Question 4 (6 pts)

Discuss two types of interaction effects. Write an example equation for each.

Answer

1. Reinforcement: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
2. Interference: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 - \beta_3 X_1 X_2$

Question 5 (4 pts)

Consider the predictors `age` (numeric) and `gender` (M and F), and the response `height` (numeric). While constructing a model to predict height using age and gender, how many regressors will a model that is linear in the predictors have? How will the values F and M of gender be represented in the model? Write out the *theoretical* regression equation.

Answer

The model will have 2 predictors, one for age and one for gender.

The gender predictor can be coded as 0 for F and 1 for M.

The model equation is: $Y = \beta_0 + \beta_1\ age + \beta_2\ gender$, where gender = 0 when F and 1 when M

Question 6 (6 pts)

What is the purpose of the Ramsay RESET test? What are the advantages and disadvantages of using said test? Write out the null and alternative hypotheses.

Answer

Ramsey RESET is a top-down approach to determine if the predictors raised to their second or higher powers and the interaction terms are significant in the model. The test is very generalized and can only be used as a guidance. Further steps will be needed to determine the predictors to be included in the model. However, it helps save time and efforts while modeling.

H_0: $\alpha_1 = ... = \alpha_k = 0$, H_a: *not all* $\alpha_k = 0$

where $\alpha_1, ..., \alpha_k$

Question 7 (6 pts)

Consider the following fitted regression equation:

$$\widehat{height} = 3 + 0.33age + 0.266age^2$$

Interpret a positive, one-unit change in the value of the predictor `age`.

Answer

Taking the derivative, $\frac{\partial\ \widehat{height}}{\partial\ age} = 0.33 + (2)0.266\ age$

$\therefore \frac{\partial\ \widehat{height}}{\partial\ age} = 0.33 + 0.532\ age$

Thus, in going from the age of 0 to 1 units, height increases by 0.33, in going from the age of 1 to 2 units, height increases by $0.33 + 0.532\ (1) = 0.862$, in going from the age of 2 to 3 units, height increases by $0.33 + 0.532\ (2) = 1.394$ and so on.