**Least Squares Estimators:** , . Unbiased and have minimum variance. is estimated by , the fitted line. Error is i.i.d and norm dist; fitted line always goes thru . Mean of error , error has constant variance , error is uncorr w/ each other , and independent to if is independent. , a random variable, is the sum of a deterministic term and random term .

**Residuals:** . Observable while error is not. Sum is zero, residual/error sum of squares is minimized, 2 df, , , , . Should have constant variance (just like the error term).

**Estimated Variance:** , residual/error mean square in regression model is an unbiased estimator of pop variance . Est of pop standard deviation is just , standard error of regression. Smaller suggests that that the model is explaining a more significant portion of the variability in the data.

**Confidence Interval for and :** . For CI , just replace with . But ; .

**2-sided t-test:** Determine linear association or different from 0 (), , . Test statistic is . If conclude . If conclude .

**1-sided t-test:** Positive association, for example, , . If conclude . If conclude .

**Interval Est., CI, and PI of :** Point estimator . . Can use in place of . Confidence Interval is given as . Prediction Interval is given as where .

**Conf. Band:** Describes likely region of unknown real line. . .

**Sums of Squares:** . has 1 df for simple LinReg.

|  |  |  |  |
| --- | --- | --- | --- |
| **Source of Var** | **SS** | **df** | **MS** |
| Regression |  |  |  |
| Error |  |  |  |
| Total |  |  |  |

**F-test:** Determine linear association (), , . So the reduced model is . Test stat is . If conclude . If conclude . in simple LinReg , more specifically, the stat for the slope coefficient. This is because both stats are essentially testing the same hypothesis in simple linear regression: whether the slope differs from zero.

**General Linear Test:** Identical , . where in stands for “full model”

If conclude . If conclude . For test, and .

**Coef of Det:** . Degree of *linear* assoc; high good model useful pred; low no relationship.

With high you could still have very wide prediction intervals and alt. models might fit better; all low tells you is

that there is little evidence for a linear relationship, data could still behave in other ways.

**Diagnostic plots:** scatter, vs or , histogram or QQ plot of . Help spot outliers, non-normality, skew, nonlinearity.

**Lack of Fit:** Assume ’s are indep, iden dist, and dist of have same variance . ,

. Test stat where , .

. Full model is , reduced model is .

**Remedies:** Nonconstant var: trans or use weighted least sq; Nonnormal errors: trans , Gen linear; Nonlinear rel: try nonlinear reg, trans or add more predictors. **Box-Cox:**  no trans; square root; nat log.

**Bonferroni CIs:** For , and where . Simultaneous CIs for , .

**Working-Hotelling:** . . This is a confidence interval for simultaneous estimates mean response based on the Confidence Band for the entire regression line.

**General Bonferroni Procedure:** where . Reasonable when , the number of simultaneous CIs, is small; for ’s.

**Simultaneous Prediction Intervals:** SCHEFFE: where . BONFERRONI: where .

**Matrix knowledge:** Matrices are always written *rows x columns. Rank* is # of lin indep columns or rows; cannot exceed rows, columns. (**ABC**)T = **C**T**B**T**A**T, (**ABC**)-1 = **C**-1**B**-1**A**-1, (**A**T)-1 = (**A**-1)T.

**Matrix Reg:** For entries for , **Y** is , **X** is w/ the first column being all s, is w/ row and row , and is . Also **Y X**, , and **I** (). So the normal error regression model is **Y = X +** .

**Least Sq Est:** is an unbiased estimator of with **X**T**X**-1**X**T**Y**, in the top row and in the bottom. Hence **b**  and **b X**T**X**-1 so is multivariate norm **X**T**X**-1.

**Residuals:**  **YI HY**, , **I H**, , is an unbiased est of **I H**.

**Mean Response:** Let **X**Th be . Then  **X**Thand est. variance of is **X**Th**X**T**X**-1**X**h.

**New Observation Prediction:** pred**X**Th**X**T**X**-1**X**h estimates pred**X**Th**X**T**X**-1**X**h.

**PI vs CI vs CB:** CI: Shows the likely range of values associated with a parameter. Used to quantify the uncertainty around a parameter estimate from a statistical model or a sample statistic. Interpretation: If we were to draw many samples from a population and calculate the confidence interval for each sample, we'd expect about 95% (for a 95% confidence level) of those intervals to contain the true population parameter.

PI: Predicts what range an unknown observation will fall. Used to quantify the uncertainty around individual predicted values from a statistical model, accounting not only for the uncertainty in the model parameters but also for the inherent variability in the response. Interpretation: Gives a range in which a new observation will fall, with a certain level of confidence, given specific predictor values. This interval is typically wider than a confidence interval for the mean response because it accounts for the added variability of individual observations.

CB: Provides a range around a regression line (or curve) in which we're confident the true regression line (or curve) will lie. Interpretation: Represents uncertainty about the line (or curve) across the range of the data. The band tends to be narrower near the mean of the predictor variable and wider as you move away from it.