STAT 704 --- Chapter 2: Inference in Regression

Inference about the slope β_1:

• It can be shown that the sampling distribution of b_1 is

<u>Proof</u>:

• So

but σ^2 is unknown, so we estimate it with

Then

Hence, a $(1 - \alpha)100\%$ CI for β_1 is:

Note that testing H_0: $\beta_1 = 0$ is often important in SLR.
• Under the SLR model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$**, if $\beta_1 = 0$, then**

• In that case, *X* is

To test H_0: $\beta_1 = 0$ at significance level α, we use the test statistic:

Rejection rule and P-value depend on the alternative hypothesis:

• What if we want to test a nonzero value of β_1, e.g., H_0: $\beta_1 = 3$?

• Typically we find these CIs and t* and P-values using SAS or R.

<u>Example</u> (Toluca refrigeration company):
***X* = Lot Size (to produce a certain part)**
***Y* = Work Hours (needed to produce a certain part)**

Interval Estimation of $E(Y_h)$

• We often wish to estimate the mean *Y*-value at a particular *X*-value, say X_h.

• We know a point estimate for this mean $E(Y_h)$ is simply

• This estimate has variability depending on which sample we obtain. (Why?)

• To account for the variability, we develop a CI for $E(Y_h)$.

<u>Note</u>: $\hat{Y}_h$ **is a**

so $\hat{Y}_h$ **has a**

• So estimating σ^2 with MSE and using earlier principles, a $(1 - \alpha)100\%$ CI for $E(Y_h)$ is:

• Note this CI is narrowest when and gets wider

Prediction Interval for *Y*-value of a New Observation

• Suppose we have a new data point with $X = X_h$.
• We wish to predict the *Y*-value for this observation.
• Point prediction is

• What about a prediction interval?
• There are <u>two</u> sources of sampling variability for this <u>predicted</u> *Y*:
(1)

(2)

• Our CI for $E(Y_h)$ only involved the <u>first</u> source.
• Our Prediction Interval for $Y_{h(new)}$ will be _________

• Variance of the prediction error is:

Estimating σ^2 with MSE, our $(1 - \alpha)100\%$ PI for $Y_{h(new)}$ is:

Example (Toluca data):

• With a 90% CI, estimate the mean number of work hours for lots of size 65 units.

• With a 90% PI, predict the number of work hours for a new lot having size 65 units.

Note: Working and Hotelling developed $100(1 - \alpha)\%$ confidence bands for the entire regression line.
(see Sec. 2.6 for details)

Picture:

<u>Analysis of Variance Approach to Regression</u>

• Our regression line is a way to use the predictor (*X*) to explain how the response (*Y*) varies.

• This can be represented mathematically by <u>partitioning</u> the <u>total sum of squares</u> (SSTO).

SSTO = $\sum (Y_i - \bar{Y})^2$ **is a measure of the total (sample) variation in the *Y* variable.**

• Note SSTO =

Picture:

• When we account for X,

we would use

SSE = $\sum (Y_i - \hat{Y}_i)^2$ **is a measure of how much *Y* varies <u>around the regression line</u>.**

SSR =

SSR measures how much of the variability in *Y* is explained by the regression line (by *Y*'s linear relationship with *X*).

• Thus SSE measures

Degrees of freedom:

• To directly compare "explained variation" to "unexplained variation," we must divide by the proper d.f. to obtain the corresponding <u>mean square</u>:

If MSR >> MSE, then the regression line explains a lot of the variation in *Y*, and we say the regression line fits the data well.

Summary: ANOVA Table

• Note the expected Mean Squares: MSR is expected to be large than MSE if and only if

• So testing whether the SLR model explains a significant amount of the variation in *Y* is equivalent to testing

• Consider the ratio MSR / MSE. If H_0 is true, we expect this to be near

• If H_0 is true, this ratio has

Leads us to

Test statistic

RR:

• Note that $F^* = (t^*)^2$ and that this F-test (in SLR) is equivalent to the t-test of H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$.

Example:

General Linear Test

• Note if H_0: $\beta_1 = 0$ holds, our "reduced model" is

• It can be shown that the least-squares estimate of β_0 here is

• Thus SSE for the reduced model is

• Note that the SSE(R) can never be less than the SSE for the full model, SSE(F).
• Including a predictor can never cause the model to explain less variation in *Y*.

→
• If SSE(R) is only a little more than SSE(F), then the predictor is

• We can generally test this with an F-test:

• This principle of comparing SSE(R) and SSE(F) based on "reduced" and "full" models will be used often in more advanced regression models.

R^2 and r

• The coefficient of determination
is the <u>proportion</u> of total sample variation in *Y* that is explained by its linear relationship with *X*.

• The closer R^2 is to 1, the

Correlation coefficient r =

• Note

Values of r near 0 →

Values of r near 1 →

Values of r near –1 →

Cautions about R^2 and r:
• R^2 could be high, but predictions may not be precise.
• R^2 could be high, but the linear regression model may not be the best fit

• R^2 and r could be near 0, but X and Y could still be related

• R^2 can be inflated when sample X values are widely spaced

Example (Toluca data):

Correlation Models

• In regression models:

• If we simply have two continuous variables X and Y without natural response/predictor roles, a correlation model may be appropriate.
• Convenience store example:

• If appropriate, we could assume X and Y have a bivariate normal distribution.
• Five parameters:
• Investigation of the linear association between X and Y is done through inferences on ρ_{XY}.
• r is a point estimate of ρ_{XY}.
• Testing H_0: $\rho_{XY} = 0$ is equivalent to

• A CI for ρ_{XY} requires Fisher's z-transformation:

For large samples, a $(1 - \alpha)100\%$ CI for

• Then use Table B.8 in book to back-transform endpoints to get CI for ρ_{XY}.

Example:

Cautions about Regression

• When predicting future values, the conditions affecting *Y* and *X* should remain similar for the prediction to be trustworthy.

• Beware of extrapolation (predicting *Y* for values of *X* outside the range of *X* in the data set). The relationship observed between *Y* and *X* may not hold for such *X* values.

• Concluding that *Y* and *X* are linearly related (that $\beta_1 \neq 0$) does not imply a causal relationship between *X* and *Y*.

• Beware of making multiple predictions or inferences simultaneously – generally the Type I error rate is affected.

• The least-squares estimates are not unbiased if *X* is measured with error.
• This is when the *X* values we observe in our data are not the true predictor values for those observations.
• In this case, the estimated coefficients are biased toward zero.
• Advanced techniques are needed to deal with this issue.