Contents

Statistical Inference for Simple Linear Regression (Ch.2)

The Normal Assumption

The error term: ε_i's are i.i.d. (independent and identically distributed) $N(0, \sigma^2)$. Hence,

$$Y_i \text{ } are \text{ } i.i.d. N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Inferences concerning regression coefficients

- **Sampling distribution of b_1 (i.e., $\widehat{\boldsymbol{\beta}}_1$)**
  Under the Normal assumption, the OLS (same as MLE) estimator b_1 has distribution
  $$\frac{(b_1 - \beta_1)}{se(b_1\ )} \sim t(n-2)$$
  where: β_1 *is the true value of the slope.*
  $se(b_1)$, or written as $s(b_1)$, is the standard error of b_1.
  $se(b_1) = \sqrt{\frac{s^2}{\sum(X_i - \bar{X})^2}}$
  s^2 = MSE = $\sum(y_i - \widehat{y}_i)^2/(n-2)$.

- **Confidence Interval for β_1** is constructed as:
  $$b_1 \pm t_{(1-\alpha/2,n-2)} se(b_1),$$
  where: b_1 is the estimate (OLS and MLE are the same in this case) of β_1;
  $t_{(1-\alpha/2,n-2)}$ is the critical value for d.f.= n-2 at (1- α)100% confidence level;
  $se(b_1) = \sqrt{\frac{s^2}{\sum(X_i - \bar{X})^2}}$ is the standard error of b_1;

- **Test of significance for β1** is constructed as:
  $H_0: \beta_1 = \beta_{10}$ vs $H_a: \beta_1 \neq \beta_{10}$
  $t_{obs} = \frac{(b_1 - \beta_{10})}{se(b_1)}$
  $p - value = 2 * P(\ t > |t_{obs}|)$, where $t \sim t(n-2)$
  Reject H_0 if p-value < α.
  Or, use critical value, reject H_0 if $|\ t_{obs}| \geq t_{crit}, t_{crit} = t(1 - \alpha/2, n-2)$

where: β_{10} is the "hypothesized" value for the slope (often, $\beta_{10} = 0$).
If H_a is one-sided, adjust the p-value computation is one-sided as well.

- If Ha: beta > 0, then p-value = P($t_{(df = n-2)} > t_{obx}$)
- If Ha: beta < 0, then p-value = P($t_{(df = n-2)} < t_{obx}$)

➢ **CI and test for β_0** are similar, with standard error:

$$se(b_0) = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right)}$$

➢ Theories behind the above results (FYI):

- *Y_i's* are independent, Normal random variables. *b_0* and *b_1* are linear combinations of the *Y_i's*. Properties of Normal distribution assure that:

$$b_1 \sim N(\beta_1, \sigma^2(b_1)),\ where\ \sigma^2(b_1) = \sigma^2 / \sum(X_i - \bar{X})^2$$

$$b_0 \sim N(\beta_0, \sigma^2(b_0)),\ where\ \sigma^2(b_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right)$$

- In practice, σ^2 (the population variance of the errors) is unknown. It is estimated by s^2 = MSE. Hence, changing the Normal distribution into t-distribution.

➢ If the distribution of ε_i is not normal but is relatively symmetric, then the CIs and significance tests are reasonable approximations.

Confidence interval and Prediction interval

➢ A distinction between the Confidence Interval (CI) and the Prediction Interval (PI) is made by the difference between mean response ($E(Y_h) = \mu_h$) and a single response ($Y_{h(new)}$) at x_h.

➢ Same point estimation:

$$\hat{\mu}_h(ie, \hat{Y}_{mean}) = \hat{Y}_{h(new)} = b_0 + b_1 X_h$$

➢ Different standard errors:

For the mean response: $se(\hat{Y}_{mean}) = \sqrt{s^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)}$

For single response (why?):

$$se(\hat{Y}_{h(new)}) = \sqrt{s^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} + 1\right)} = \sqrt{\left[se(\hat{Y}_{mean})\right]^2 + MSE}$$

➢ The CI (for mean response) and PI (for single response) are:

(point est.) ± (critical value)(std.err of the point est.)

where the critical value is $t_{(1-\alpha/2, n-2)}$.

- The **Working-Hotelling Confidence Band** gives combined "confidence intervals" at *all* X_h:

$$\hat{\mu}_h \pm W \times se(\hat{Y}_{mean}), \qquad where\ W = \sqrt{2F_{(1-\alpha;\,2,n-2)}}$$

$F_{(1-\alpha;\,2,n-2)}$ can be found using an F-distribution calculator.
We will discuss more on the topic of multiple CI's and tests in later chapters.

The ANOVA table

- Partitioning **Sum of Squares:**
  We will use the Sum of Squares as a measurement of the variation.
  Total variation (measured by Sum of Squares) in Y: $SSTO = \sum \left(Y_i - \bar{Y}\right)^2$
  Variation in Y that can be explained by X: $SSR = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$
  Variation due to randomness (unexplained variation): $SSE = \sum \left(Y_i - \hat{Y}_i\right)^2$
  Note that: (proof?) *SSTO = SSR + SSE*

- Partitioning **Degrees of Freedom:**

$$df_{Total}\ (n\text{-}1) = df_{Reg}\ (1) + df_{Error}\ (n\text{-}2)$$

- The ANOVA table.

Source	Sum of Squares	df	Mean Squares	F	P-value
Regression(Model)	SSR	1	MSR=SSR/df_R	F=MSR/MSE	$P(F_{(dfR,\ dfE)} > F)$
Error(Residual)	SSE	n-2	MSE=SSE/df_E		
Total	SSTO	n-1			

Q. What is that p-value in the ANOVA table for?

Coefficient of determination

- R^2 = SSR/SSTO is the proportion of the variation in Y (measured by the sum of squares) that can be determined/explained by the linear regression model using X.

- In SLR, $R^2 = (r)^2$, where r is the correlation coefficient between Y and X.

- R^2 is between (0, 1).

- For models with the same number of predictors, a larger value of R^2 is preferred. However, note that R^2 increases when more predictors are added to the model. (Adjusted-R^2 will be introduced in multiple linear regression.)

- R^2 tells the strength of association assuming the current model is appropriate. In SLR (straight-line), if the association between the variables is curved, R^2 may not be meaningful.

Some considerations:

- The normality assumption.

- Interpreting the regression parameters and R^2.

- The Power of a test.
  - What is Power?
  - Things that can affect the power:
    - The difference between the "truth" and "hypothesis".
    - The variability of the data (i.e., the standard error of the statistic).
    - The sample size.
    - The significance level.
  - More details on this topic in the text, p.51 and Table B.5 (p. 669)

- Extrapolation.

- Correlation vs causation.

Regression Diagnostics (Ch.3)

What to diagnose?

- The assumptions about the model.

Diagnostic tools

- The residuals: $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$
- Most diagnostic graphs and tests are based on the residuals. Sometimes, the residuals may be further "processed" (more details later).

The Lack-of-fit test

- Test for Linearity (and how the model fits the data set in general).
- Only applies to the data sets where there are 2 or more observations ($Y_{i,j}$) at a given X_j value.
- Partitioning the $SSE = \sum_{i,j}\left(Y_{i,j} - \hat{Y}_{i,j}\right)^2$

  Sum of Squares of Lack of Fit: $SS_{LF} = \sum_{i,j}\left(\bar{Y}_j - \hat{Y}_{i,j}\right)^2$

  Sum of Squares of Pure Error: $SS_{PE} = \sum_{i,j}\left(Y_{i,j} - \bar{Y}_j\right)^2$

  $SSE = SS_{LF} + SS_{PE}$

- Partitioning the Degrees of Freedom:

  $df_E\ (n-2) = df_{LF}\ (c-2) + df_{PE}\ (n-c)$

  where n is the sample size, and c is the number of different X values in the data set.

- Expanded ANOVA Table.

Source	Sum of Squares	df	Mean Squares	F	P-value
Regression	SSR	1	$MSR=SSR/df_R$	F=MSR/MSE	$P(F_{(dfR,\ dfE)} > F)$
Error	SSE	n-2	$MSE=SSE/df_E$		
Lack of Fit	SS_{LF}	c-2	$MS_{LF}=SS_{LF}/df_{LF}$	$F_{LF} = MS_{LF}/MS_{PE}$	$P(F_{(dfLF,\ dfPE)} > F_{LF})$
Pure Error	SS_{PE}	n-c	$MS_{PE}=SS_{PE}/df_{PE}$		
Total	SSTO	n-1			

Q. What are the hypotheses associated with the p-values?

Please refer to handouts *Diagnostics.pdf* and *Regression_Diagnostics.doc* on Blackboard for more details on regression diagnosis.

Simultaneous Intervals (Ch.4)

Why correct for significance level?

Let CI_1 and CI_2 be the confidence intervals for two parameters: β_1 and β_2, respectively. We will use "c" if the interval contains the true parameter value, "nc" if the interval does not contain the true parameter.

P(both "c") < P(CI_1 "c") = $1 - \alpha$,

P(both "c") = P(CI_1 "nc") + P(CI_2 "nc") - P(CI_1 "nc" or CI_2 "nc")

$\geq (1 - \alpha) + (1 - \alpha) - 1$

$\geq 1 - 2\alpha$

The Bonferroni correction for joint CI

- We want the probability that "*both intervals are correct*" to be $\geq 1 - \alpha$. That is, we have an "error budget" of α that can be spent on 2 CI's. So we split the "budget" equally and each CI will have $\alpha^* = \alpha/2$.

- For example, for 95% simultaneous (joint) CI for β_0 and β_1, each CI will have confidence level 97.5% using Bonferroni correction.

- In general, if we consider g CI's simultaneously at $(1 - \alpha)$ family level, we will use:
$$1 - \alpha^* = 1 - (\alpha/g)$$
for each CI, i.e., $b_k \pm t_{(1-\alpha^*/2, n-2)} s(b_k)$

- Bonferroni correction is a conservative approach.

Simultaneous/Joint confidence Interval for mean response ($E(Y_h) = \mu_h$)

- Use Bonferroni correction to adjust the critical t-value:
$$\hat{\mu}_h \pm B \times se(\hat{Y}_{mean}), \qquad where\ B = t_{(1-\alpha/(2g);\, n-2)}$$

- Recall Working-Hotelling:
$$\hat{\mu}_h \pm W \times se(\hat{Y}_{mean}), \qquad where\ W = \sqrt{2F_{(1-\alpha;\, 2,\ n-2)}}$$

- Which one is better? Choose the smaller value of B and W (why?).

Simultaneous/Joint Prediction Interval for individual response (*$Y_{h(new)}$*)

- Use Bonferroni correction to adjust the critical t-value:
$$\hat{Y}_{h(new)} \pm B \times se\left(\hat{Y}_{h(new)}\right),\ where\ B = t_{(1-\alpha/(2g);\, n-2)}.$$

- Scheffé: consider g PI's simultaneously:
$$\hat{Y}_{h(new)} \pm S \times se\left(\hat{Y}_{h(new)}\right),\ where\ S = \sqrt{gF_{(1-\alpha;\ g,\ n-2)}}.$$

- Which one is better? Choose the smaller value of *B* and *S* (why?).

Simultaneous tests

- Adjust the significance level for each test, using Bonferroni correction.

Miscellaneous topics (Ch.4, cont.)

Regression through origin

- $Y_i = \beta_1 X_i + e_i$
- Estimation equations need to be changed.
- Residuals may not sum to 0.
- ANOVA table needs adjustments (SSTO may be less than SSE).

Measurement Error

- For Y, this is usually not a problem. Just adds to the random error.
- For X, we can get biased estimators of our regression parameters.
- See text p.165 for more details.

Inverse Prediction

- Sometimes it is called *calibration*.
- Given Y_h, predict the corresponding value of X, by solving the fitted linear equation.
- Approximate CI can be obtained.
- See text p.169 for more details.

Choice of X levels

- Many questions: How many levels? How far between 2 levels? Range of X? How many observations at each level?
- Answer: It depends!
- See text p.171 for more details.