INFERENCE FOR MULTIPLE LINEAR REGRESSION

Recall Terminology:

p predictors $x_1, x_2, \dots, x_p$

(Some might be indicator variables for categorical variables.)

k-1 *non-constant* terms $u_1, u_2, \dots, u_{k-1}$

Each u_j is a function of $x_1, x_2, \dots, x_p$:

$$u_j = u_j(x_1, x_2, \dots, x_p)$$

For convenience, we often set $u_0 = 1$

(constant function/term)

$$\underline{u} = \begin{bmatrix} u_0 \\ u_1 \\ \cdot \\ \cdot \\ u_{k-1} \end{bmatrix} = \begin{bmatrix} 1 \\ u_1 \\ \cdot \\ \cdot \\ u_{k-1} \end{bmatrix}$$

Assumptions so far:

1. $E(Y|\underline{x})$ (or $E(Y|\underline{u})$) $= \eta_0 + \eta_1 u_1 + \dots + \eta_{k-1} u_{k-1}$

$$= \underline{\eta}^T \underline{u}$$

(Linear Mean Function)

2. $Var(Y|\underline{x})$ (or $Var(Y|\underline{u})$) $= \sigma^2$

(Constant Variance)

Additional Terminology:

(Similar to simple linear regression)

- $\hat{y}_i = \hat{\underline{\eta}}^T \underline{u}_i$ (*i^{th} fitted value* or *i^{th} fit*)
- $\hat{e}_i = y_i - \hat{y}_i$ (*i^{th} residual*)
- $RSS = RSS(\hat{\underline{\eta}}) = \sum (y_i - \hat{y}_i)^2 = \sum (\hat{e}_i)^2$

(*residual sum of squares)*

Results from Assumptions (1) and (2):

(Similar to simple linear regression)

- $\hat{\eta}_j$ is an unbiased estimator of η_j.
- $\hat{\sigma}^2 = \frac{1}{n-k} RSS$ is an unbiased estimator of σ^2.

Note: In simple regression, k = 2.

Example: Haystacks

Additional Assumptions Needed for Inference:

(3) Y|$\underline{x}$ is normally distributed
(Recall that this will be the case if $\underline{X}$,Y are multivariate normal.)

(4) The y_i's are independent observations from the Y|$\underline{x}_i$'s.

Consequences of Assumptions (1) - (4) for Inference for Coefficients:

- Y|$\underline{x}$ ~ N($\underline{\eta}^T\underline{u}$, σ^2)
- $\hat{\sigma}^2$ is a multiple of a χ^2 random variable with

  n-k degrees of freedom -- so we say

  $\hat{\sigma}^2$ and RSS have df = n-k.
- There is a formula for s.e.($\hat{\eta}_j$).

  (We'll use software to calculate it.)
- $\frac{\hat{\eta}_j - \eta_j}{s.e.(\hat{\eta}_j)}$ ~ t(n-k) for each j.

Note:

- The consequences listed above are also valid replacing (3) by the weaker assumption that $Y|\underline{x}_i$ is normally distributed for i = 1, 2, … , p.

- If the $Y|\underline{x}_i$'s are not normal, but are not too ill-behaved and n is large enough, the consequences above are still approximately true, thanks to the CLT.

- *Example*: Haystacks

Caution: Multiple Testing

Recall: If you set an α level for hypothesis tests, then a p-value less than α tells you that (at least) one of the following holds:

i) The model does not fit

ii) The null hypothesis is false.

iii) The sample at hand is one of the less than α percent of samples for which you would falsely reject the null hypothesis.

*If you are doing **two** hypothesis tests with the same data:*

- There is **no** guarantee that the "bad" samples (for which you falsely reject the null) are the same for both tests.

- *In general, the probability of falsely rejecting one of the two null hypotheses is **greater** than* α.

When doing two hypothesis tests with the same data, you typically need an *overall* significance level α:

- That is, you want to be able to say that, if the model fits and *both* null hypotheses are true, then the probability of falsely rejecting *at least* one of the two null hypotheses using your decision rule is α.

- To do this, you typically need *lower* significance levels for each test individually.

- One way to be sure of having an overall significance level α when doing k hypothesis tests with the same data is the *Bonferroni* method: Require significance level α/k for each test individually.

- There are various other methods that allow individual significance levels higher than α/k, but they only apply in specific situations.)

For this reason, in model-building in regression, p-values for hypothesis tests are often interpreted as just loose guides for what might or might not be reasonable.

A similar situation holds for confidence intervals:

- To be able to say, "We have produced these two intervals by a procedure which, for 95% of all suitable samples, produces a first interval containing η_0 and a second interval containing η_1" (i.e., if you want an *overall* confidence level 95%), the two individual confidence intervals need to have *individual* confidence level *greater* than 95%.

- Bonferroni will also work here: requiring individual confidence levels 97.5% will suffice to give overall confidence level 95% for two confidence intervals.

- In regression, we can also use *confidence regions*; see Section 10.8 for more details.

Inference for Means: Recall from simple regression:

$$\text{Var}\,(\hat{E}(Y|x)) = \text{Var}(\hat{E}(Y|x)|\, x_1, \ldots, x_n)$$

$$= \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}\right).$$

So

$$\text{s.e}\,(\hat{E}(Y|x) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$$

$$= \hat{\sigma} \times (\text{a function of } x \text{ and the } x_i\text{'s but not the } y_i\text{'s})$$

An analogous computation (best done by matrices -- see Section 7.9) in the multiple regression model gives

$$\text{Var}\,(\hat{E}(Y|\underline{x})) = \text{Var}(\hat{E}(Y|\underline{x})|\, \underline{x}_1, \ldots, \underline{x}_n)$$

$$= h\sigma^2,$$

where $h = h(\underline{u})$ ($= h(\underline{x})$ by abuse of notation) is a function of $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_n$, called the *leverage*.

In simple regression,

$$h(x) = \frac{1}{n} + \frac{(x - \bar{x})^2}{SXX}$$

Note that $(x - \bar{x})^2$ (hence also h(x)) is a (non-linear) measure of the distance from x to $\bar{x}$. Similarly, in multiple regression, h($\underline{x}$) is a type of measure of the distance from $\underline{u}$ to the *centroid*

$$\underline{\bar{u}} = \begin{bmatrix} 1 \\ \bar{u}_1 \\ \cdot \\ \cdot \\ \cdot \\ \bar{u}_{k-1} \end{bmatrix},$$

(i.e., it is a monotone function of $\sum (u_j - \bar{u}_j)^2$.) In particular:

- The further $\underline{u}$ is from $\underline{\bar{u}}$, the larger Var ($\hat{E}(Y|x)$) is, so the less precisely we can estimate E(Y|x) or y.
- For example, an x-outlier could give a large h, and hence make inference less precise.

Example: 1 predictor

Define:

$$\text{s.e. } (\hat{E}(Y|\underline{x})) = \hat{\sigma}\sqrt{h(u)}$$

Summarize:

- The larger the leverage, the larger s.e. ($\hat{E}(Y|x)$) is, so the less precisely we can estimate E(Y|x).
- The leverage depends just on the $\underline{x}_i$'s, not on the y_i's.

Similarly to simple regression:

- The sampling distribution of $\hat{E}(Y|\underline{x})$ is normal
- $\dfrac{\hat{E}(Y \mid \underline{x}) - E(Y \mid \underline{x})}{\text{s.e.}(\hat{E}(Y \mid \underline{x})} \sim t(n\text{-}k).$
- Thus we can do hypothesis tests and find confidence intervals for the conditional mean response E(Y|$\underline{x}$)

Again,

- The consequences listed above are also valid replacing (3) by the weaker assumption that $Y|\underline{x}_i$ is normally distributed for i = 1, 2, … , p.

- If the $Y|\underline{x}_i$'s are not normal, but are not too ill-behaved and n is large enough, the consequences above are still approximately true, thanks to the CLT.

Prediction: Results are similar to simple regression:

- Prediction error = $Y|\underline{x} - \hat{E}(Y|\underline{x})$

- $\text{Var}(Y|\underline{x} - \hat{E}(Y|\underline{x})) = \sigma^2(1 + h(\underline{u}))$

- Define s.e. $(Y_{pred}|\underline{x}) = \hat{\sigma}\sqrt{1+h}$

- $\dfrac{Y \mid \underline{x} - \hat{E}(Y \mid \underline{x})}{se(y_{pred} \mid \underline{x})} \sim t(n-k)$, so we can form prediction intervals.

Caution: As with simple regression, for prediction, we need the assumption that $E(Y|\underline{x})$ is normal (or very close to normal, with approximate results).

Example: Haystacks