INFERENCE FOR MULTIPLE LINEAR REGRESSION

Terminology: Similar to terminology for simple linear regression

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

Results similar to those in 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.
- $\hat{\sigma}^2$ is a multiple of a χ^2 distribution with n-k degrees of freedom -- so we say $\hat{\sigma}^2$ and RSS have df = n-k.

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|x_i's.

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

- Y|$\underline{x}$ ~ N($\underline{\eta}^T \underline{u}$, σ^2)
- There is a formula for s.e.($\hat{\eta}_j$). (We'll use software to calculate it.)
- $\dfrac{\hat{\eta}_j - \eta_j}{s.e.(\hat{\eta}_j)}$ ~ t(n-k) for each j.

Example: Haystacks

Inference for Means:

In simple regression, we saw

$$\text{Var}(\hat{\text{E}}(Y|x)) = \text{Var}(\hat{\text{E}}(Y|x)|\ x_1, \ldots, x_n) = \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}\right).$$

So

$$\text{s.e}(\hat{\text{E}}(Y|x) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}} = \hat{\sigma} \text{ times a function of the } x_i\text{'s (but not the } y_i\text{'s)}$$

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

$$\text{Var}(\hat{E}(Y|\underline{x})) = \text{Var}(\hat{E}(Y|\underline{x})|\ \underline{x}_1, \dots, \underline{x}_n) = h\sigma^2,$$

where h = h($\underline{u}$) (= h($\underline{x}$) by abuse of notation) is a function of $\underline{u}_1, \underline{u}_2, \dots, \underline{u}_n$, called the *leverage*. (The name will be explained later.)

In simple regression,

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

Note that $(x-\bar{x})^2$ (hence h(x)) is a 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*

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

(that is, 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. (Thus an outlier could give a large h, and hence make inference less precise.)

Example: 1 predictor

Define:

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

Summarizing:

- 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:

$$\frac{\hat{E}(Y|\underline{x}) - E(Y|\underline{x})}{\text{s.e.}(\hat{E}(Y|\underline{x})} \sim t(n-k).$$

Thus we can do hypothesis tests and find confidence intervals for the conditional mean response E(Y|$\underline{x}$)

Prediction: Results are similar to simple regression:

- Prediction error = $Y|\underline{x} - \hat{E}(Y|\underline{x})$
- $\mathrm{Var}(Y|\underline{x} - \hat{E}(Y|\underline{x})) = \sigma^2(1 + h(\underline{u})) = \sigma^2 + \mathrm{Var}(E(Y|\underline{x}))$
- Define s.e. $(Y_{pred}|\underline{x}) = \hat{\sigma}\sqrt{1+h}$
- $\dfrac{Y \mid \underline{x} - \hat{E}(Y \mid \underline{x})}{se(y_{pred} \mid \underline{x})}$ ~ t(n-k),so we can form prediction intervals.

Example: Haystacks