Lecture 12 (CR.3)

Basic picture of regression:

\[ y = \alpha + \beta x \]

\[ (x_3, y_3) \]

\[ \hat{y}_3 = \alpha + \beta x_3 \]

\[ (x_3, \hat{y}_3) \]

\[ \text{error} = [y_3 - \alpha - \beta x_3] \]

\[ \hat{\beta} = \frac{x\bar{y} - \bar{x}\bar{y}}{x^2 - \bar{x}^2} \]

\[ \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \]

\[ z = \bar{y} - \hat{\beta} \bar{x} \]

Think of \( \alpha = \text{Arterial Blood Pressure (ABP)} \) measured noninvasively

\( \alpha = \text{Intracranial Pressure (ICP)} \) measured invasively.

There is a different (more useful?) way of looking at this via variance.

This way, we will arrive at quantities called \( R^2 \) and \( s_e \), which together assess how good the fit is.

- Suppose we measure Table Length, \( y \).
- Repeat, and histogram:

\[ s_y = \sqrt{\frac{s_{yy}}{n-1}} \sim 0 \]

- One may report:

True length = 150±10 cm

- Now, suppose you are unhappy with the large \( s_y \).
- You may wonder, could some of that variability be due to something else that is varying every time you make a measurement of \( y \).
- \( \alpha = \text{temperature} \) ? \( \text{humidity} \)?
- If so, then by measuring \( y \) and \( \alpha \), we may be able to reduce the \( \pm \) of our report, by specifying \( y \) at a given \( \alpha \).
Analysis of variance (ANOVA):

How much of the variation in $y$ is due to the linear relationship between $y$ and $x$?

Variance of $y = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$

$SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

$SS_{explained} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

$SS_{unexplained} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Total variation in $y$.

Variation in $y$ explained by (or due to) $x$.

Variation in $y$ unexplained by $x$.

$SS_{total} = SS_{explained} + SS_{unexplained}$

$\text{SST} = \frac{SS_{explained}}{n-1} + \frac{SS_{unexplained}}{n-2}$

Variability is reduced from $\pm (10)^2$ to something smaller, say $\pm (3)^2$.

Therefore, $\frac{SS_{explained}}{SS_{total} \times 100}$, called $R^2$, measures how good the fit is.

Percent variation in $y$, explained by $x$.

$(\text{Bad Model/fit}) 0 < R^2 < 1$ (Good Model/fit)

The other piece, $SS_{unexplained} = SSE$, is a sum-squared, and so can be "Averaged" to provide a measure of typical error. Specifically, $\sqrt{\frac{SSE}{n-2}} = \hat{\sigma}$ = Std. dev. of errors ~ typical error.

Typo in book, p. 121

Move on $R^2$, $\hat{\sigma}$ later.
Picture for the above decomposition: (ANOVA)

When there is a linear relationship between $x$ and $y$,
then some portion of the variation in $y$ can be attributed
to (or explained by) $x$. That portion is $SS_{\text{exp.}}$, and the (unexplained) rest is $SS_{\text{unexp}} = SSE$. So the variability in $y$, $SST$, is reduced to $SSE$.
Example (from previous lecture):

\[ \text{SST} = \sum_{i}(y_i - \overline{y})^2 = \ldots = 6251.2 \]

\[ \text{SSE} = \sum_{i}(y_i - \hat{y}_i)^2 = \text{last column in table in prev. lecture.} \]

\[ = (-1.5)^2 + (5.1)^2 + (11.5)^2 + (-30.3)^2 + (65.1)^2 = 1307 \]

\[ R^2 = \frac{\text{Coeff. of det.}}{\text{SST - SSE}} = \frac{6251.2 - 1307}{6251.2} = 0.79. \]

Conclusion: 79\% of the variability (or variation) in \( y \) (weight) is due to (can be explained by)
the linear relation with \( x \) (height).

\[ R^2 \]

Note: \( R^2 \) is not a square of anything; at least not generally. Later, you'll see why it's written
symbol as \( R^2 \), or even as \( r^2 \) (e.g., in books).

\[ \text{It is coefficient of determination} \]

The other piece of the decomposition:

\[ s_e = \sqrt{\frac{1307}{5-2}} = 20.9 \text{ pounds} \]

Conclusion: The typical deviation of the data about the fit
(i.e. error) is about 21 pounds.
Come up with another example of $x$ and $y$ (like ABP and ICP), where regression can help in predicting $y$ from $x$ in a situation where without regression the "cost" of measuring $y$ directly is extremely high (like ICP).

Suppose all we have are data on a single variable $y$: $y_i$, $i = 1, 2, 3, \ldots, n$. Show that the predictor that minimizes SSE is the sample mean $\bar{y}$. Hint: let $Y$ denote the prediction, and the minimize SSE.

Consider the following decomposition:

$$
\sum_i (y_i - \bar{y})^2 = \sum_i [(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)]^2
$$

$$
= \sum_i (y_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 + 2 \sum_i (y_i - \bar{y})(y_i - \hat{y}_i)
$$

Show that the last term is zero if $\hat{y}_i = \hat{\beta} x_i$.

Hint: use $\hat{y} = \bar{y} - \hat{\beta} \bar{x}$ to simplify the expression. First, only when close to the end, use $\hat{\beta} = \frac{\bar{x}y - \bar{x}\bar{y}}{\bar{x}^2 - \frac{1}{n}}$. 


For the data shown in problem 3.22

a) Compute the OLS fit

b) Compute the total variation, SST.

c) Decompose it into explained and unexplained.

d) Compute $R^2$, and interpret (in English).

e) Compute the std. dev. of errors, and interpret (in English).

All by hand. You may use computer to compute sums, means, std. deviations, but not a function that does regression or analysis of variance.