# 18. Simple Linear Regression

## 18.2.1. In previous set of notes

## 18.2.2. The Method of Least Squares (continued)

The estimate of $\sigma_{y|x}$ is obtained by

$$s_{y|x} = \frac{1}{(n-q-1)}\sum_{i=1}^{n} e_i^2 = \frac{1}{(n-q-1)}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \text{ , where}$$

q is the number of explanatory variables (q=1 for simple linear regression). We'll call $s_{y|x}$ the standard deviation from regression (other names: regression standard error, residual standard error, root mean square error). Whatever it's called, it measures the variability of y about the regression line at any x.

Now we have estimates of parameters in our linear model. Before we do inference, we illustrate some SLR "hand" calculations using the meadowfoam data.

### Number of flowers per plant ($y_i$)

```
62.3 77.4 55.3 54.2 49.6 61.9 39.4 45.7 31.3 44.9
36.8 41.9 77.8 75.6 69.1 78.0 57.0 71.1 62.9 52.2
60.3 45.6 52.6 44.4
```

### Light intensity ($x_i$)

```
 150  150  300  300  450  450  600  600  750  750
 900  900  150  150  300  300  450  450  600  600
 750  750  900  900
```

Some summary statistics

```
      ***  Summary Statistics for data in:  Case0901 ***

           flowers intensity
    Mean: 56.13750  525.0000
 Total N: 24.00000   24.0000
Std Dev.: 13.73339  261.6835
```

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=$$

$$\hat{\beta}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=$$

$$\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}=$$

Thus, our estimated simple linear regression line is

$$\hat{y}=\hat{\alpha}+\hat{\beta}x=$$

Interpretation?

We could use this to calculate the $\hat{y}_i$ for every i and then calculate $s_{y|x}$ using the equation above, but we'll leave that to S-Plus. S-Plus says

$$s_{y|x} = 8.94$$

### 18.2.3. Inference for Regression Coefficents

If $y \sim N(\mu_{y|x}, \sigma^2_{y|x})$, then it is not too difficult to show that our regression parameter estimators are also normally distributed:

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2_{y|x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \text{ and}$$

$$\hat{\alpha} \sim N\left(\alpha, \sigma^2_{y|x}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\right)$$

Thus, we can perform hypothesis tests and construct confidence intervals for each regression parameter estimate. We continue to use the meadowfoam data.

$H_0$: $\beta=\beta_0$
$H_A$:

IF we knew $\sigma^2_{y|x}$, then we could use the z tables:

$$z = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})}$$

where $se(\hat{\beta})$ is the square root of the variance of $\hat{\beta}$ given above. (NOTE: Some hat's may be missing; will add during lecture.) But, we must estimate $se(\hat{\beta})$ by plugging in $s_{y|x}$ for $\sigma_{y|x}$. When we do we get

$$t = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})}$$

which follows a Student's t distribution with (n-q-1) df.

$\hat{\beta} =$

$$se(\hat{\beta}) = \frac{s_{y|x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} =$$

t=

## p-value

$P(t_{(n-q-1)} \geq t) =$

## Conclusion / Interpretation?

For most applications of SLR (not all), we are only interested in making inference about the slope parameter, although we could do testing for the intercept by the same procedure as above.

## (2-sided) (1-$\alpha$)*100% Confidence Interval for the slope, $\beta$:

$\hat{\beta} \pm t_{(n-q-1),\frac{\alpha}{2}} se(\hat{\beta}) =$

## Interpretation?

S-Plus

Statistics>Regression>Linear…

```
      *** Linear Model ***

Call: lm(formula = flowers ~ intensity, data = Case0901, na.action = na.exclude)
Residuals:
    Min     1Q  Median    3Q   Max
 -15.73 -7.805 0.01857 6.186 13.27

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  77.3850   4.1612    18.5969   0.0000
  intensity  -0.0405   0.0071    -5.6816   0.0000

Residual standard error: 8.94 on 22 degrees of freedom
Multiple R-Squared: 0.5947
F-statistic: 32.28 on 1 and 22 degrees of freedom, the p-value is 0.0000103
```

### 18.2.4. Inference for Predicted Values

We may also want to "estimate" the mean of y at some x or "predict" a new y at some x (note: POB calls each "prediction"). In each case, we may also want to get some idea of the precision of your "prediction", so we'll construct confidence intervals (for the mean) and prediction intervals (for new y).

Predict the mean y at some x value:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x =$$

$$se(\hat{y}) = s_{y|x}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} =$$

$$\hat{y} \pm t_{(n-q-1),\frac{\alpha}{2}} se(\hat{y}) =$$

## Predict a new y value at some x value:

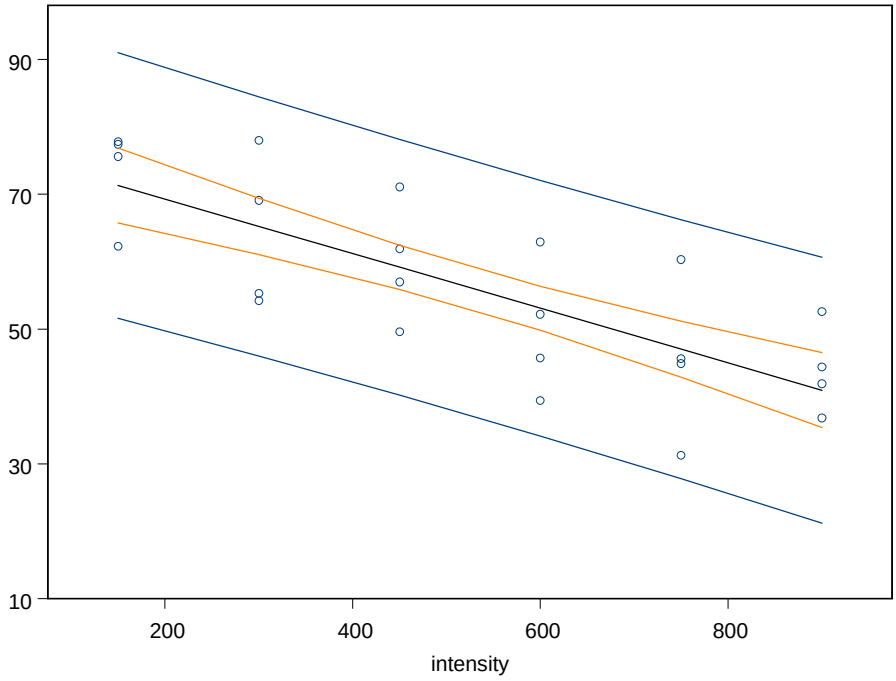
$$\tilde{y} = \hat{\alpha} + \hat{\beta}x =$$

$$se(\tilde{y}) = s_{y|x}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{s_{y|x}^{\ 2} + se(\hat{y})} =$$

$$\tilde{y} \pm t_{(n-q-1),\frac{\alpha}{2}} se(\tilde{y}) =$$

We could calculate confidence intervals and prediction intervals at many x values and “connect the dots” to get confidence bands (for the mean) and prediction bands (for new y values).

S-Plus confidence bands and prediction bands for the meadowfoam data"




## 18.3. Evaluation of the Model

When we fit a regression model to data, we would like to get a measure of how "good" the model is and to examine if our assumptions appear to be met or not. Checking our model is sometimes referred to as regression diagnostics. This model checking is usually part of an iterative modeling framework:

1) explore the data with graphs, etc.
2) formulate a model
3) check the model (repeat 1 and 2 if necessary)
4) make inference for parameters or predictions
5) present results

### 18.3.1. The Coefficient of Determination

We saw the coefficient of determination in the last chapter (notes). It's the same in this chapter, except we have different notation ($R^2$):

$$R^2 = r^2$$

That is, it's the square of the Pearson (sample) correlation coefficient. Same interpretation as given before: $R^2$ is proportion of the total variability in the observed y values explained by the linear regression of y on x (the linear association between y and x).

Perhaps a more intuitive formula is

$$R^2 = \frac{s_y^2 - s_{y|x}^2}{s_y^2}$$

Calculating the (sample) variance among the $y_i$ values in the meadowfoam example, we get $s_y^2$=188.60592. We already know $s_{y|x}^2$=8.94$^2$. Thus,

$R^2$=

(Which is on the S-Plus printout above.)

### 18.3.2. Residual Plots

Recall the residuals

$$e_i = y_i - \hat{y}_i$$

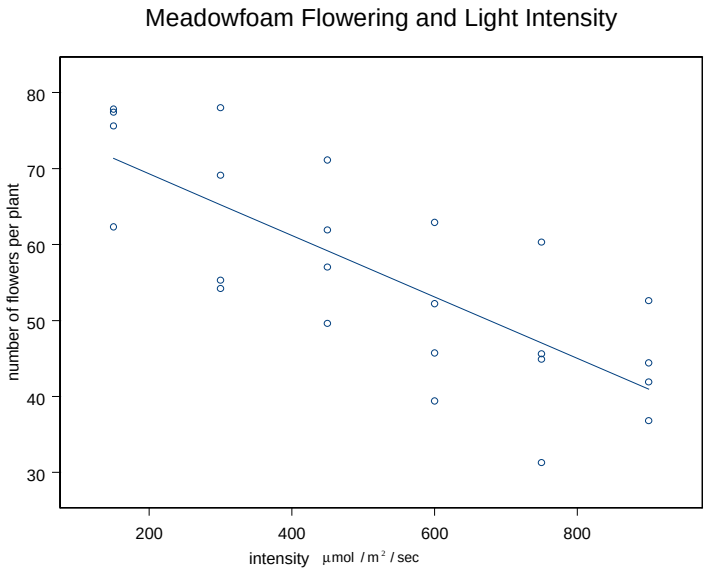
These residuals can be thought of as the sample versions of the errors $\varepsilon_i$, which should have mean zero and constant variance according to our model (this is equivalent to what we said earlier: y is assumed to have mean

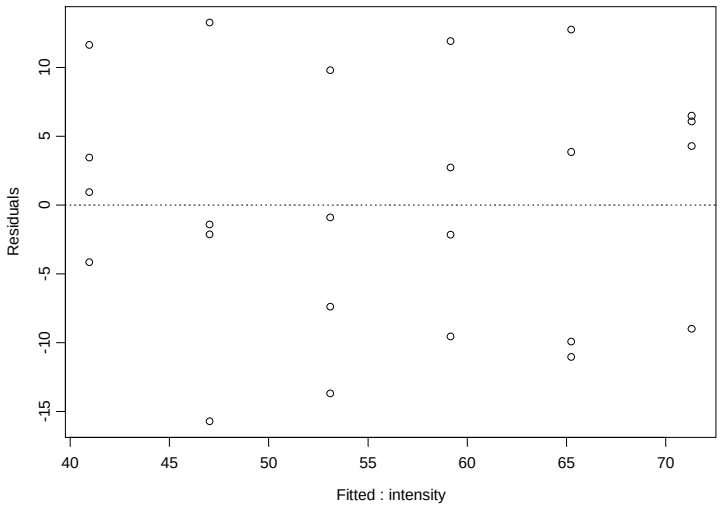
$$\mu_{y|x} = \alpha + \beta x$$

and constant variability about the mean as indicated by

$$\sigma_{y|x}.$$

This is equivalent to the $\varepsilon$ having zero mean for all x values and (the same) constant variability (standard deviation). Also, the assumption of normality on y at any x is equivalent to an assumption of normality on $\varepsilon$ at any x. Thus, if our model is "good," we would expect to see the residuals, $e_i$ "evenly scattered" about zero with no systematic departures from zero or from the constant variance assumption. A often used graphical tool for examining the residuals is the plot of the residuals verses the predicted values:




Things to look for

- outliers—residual values that are “far” from zero
- constant “spread” of residuals about zero—a common departure from this is to see a “fan-shaped” pattern in the residual plot
- symmetry about zero—a rough check on the normality assumption
- trends

### 18.3.3. Transformations

One approach to handling departures from model assumptions is to transform the data. The power transformations on both x and y are common:

$$x^p \text{ and/or } y^p$$

where p is usually in (…,-3,-2,-1,-1/2, 0, ½, 1, 2, 3,…), where $x^0$ or $y^0$ is taken to mean ln(x) or ln(y), respectively.

Questions on Lab 12?