The Analysis of Variance for Simple Linear Regression

- the total variation in an observed response about its mean can be written as a sum of two parts - its deviation from the fitted value plus the deviation of the fitted value from the mean response

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

- squaring both sides gives the total sum of squares on the left, and two terms on the right (the third vanishes)

- this is the analysis of variance decomposition for simple linear regression

$$SST = SSE + SSR$$

- as always, the total is

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

- the residual sum of squares is

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}))^2 \\ &= SS_{YY} - 2\hat{\beta}_1 SS_{XY} + \hat{\beta}_1^2 SS_{XX} \\ &= SS_{YY} - \hat{\beta}_1^2 SS_{XX} \\ &= SS_{YY} - \hat{\beta}_1 SS_{XY} \\ &= SS_{YY} - \frac{SS_{XY}^2}{SS_{XX}} \end{aligned}$$

- the regression sum of squares is

$$\begin{aligned} SSR &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(\hat{\beta}_1(x_i - \bar{x}))^2 \end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^{n} \hat{\beta}_1^2 (x_i - \bar{x})^2 \\ &= \hat{\beta}_1^2 SS_{XX} = \hat{\beta}_1 SS_{XY} = \frac{SS_{XY}^2}{SS_{XX}} \end{aligned}$$

- in completing the square above, the third term is

$$\begin{aligned} & 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\ &= 2\sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{\beta}_1(x_i - \bar{x}) \\ &= 2\hat{\beta}_1 \sum_{i=1}^{n} \hat{e}_i (x_i - \bar{x}) = 2\hat{\beta}_1 SS_{\hat{e}X} \\ &= 0 \end{aligned}$$

  using the result that the residuals are uncorrelated with the predictors

- the degrees of freedom are $n - 1$, $n - 2$ and 1 corresponding to SST, SSE and SSR

- the results can be summarized in tabular form

Source	DF	SS	MS
Regression	1	SSR	MSR = SSR/1
Residual	$n-2$	SSE	MSE = SSE/(n-2)
Total	$n-1$	SST	

Example: For the Ozone data

- $SST = SS_{YY} = 1014.75$
- $SSR = \frac{SS_{xy}^2}{SS_{xx}} = (-2.7225)^2/.009275 = 799.1381$
- $SSE = SST - SSR = 1014.75 - 799.1381 = 215.62$
- degrees of freedom: total = 4-1=3, regression = 1, error = 2

- goodness of fit of the regression line is measured by the **coefficient of determination**

$$R^2 = \frac{SSR}{SST}$$

- this is the proportion of variation in y explained by the regression on x
- R^2 is always between 0, indicating nothing is explained, and 1, indicating all points must lie on a straight line
- for simple linear regression R^2 is just the square of the (Pearson) correlation coefficient

$$\begin{aligned} R^2 &= \frac{SSR}{SST} = \frac{SS_{XY}^2/SS_{XX}}{SS_{YY}} \\ &= \frac{SS_{XY}^2}{SS_{XX}SS_{YY}} \\ &= r^2 \end{aligned}$$

- this gives another interpretation of the correlation coefficient - its square is the coefficient of determination, the proportion of variation explained by the regression

- note that with R^2 and SST, one can calculate

$$SSR = R^2 SST$$

and

$$SSE = (1 - R^2) SST$$

Example: Ozone data

- we saw $r = -.8874$, so $R^2 = .78875$ of the variation in y is explained by the regression

- with $SST = 1014.75$, we can get

$$\begin{aligned} SSR &= R^2 SST = .78875(1014.75) \\ &= 800.384 \end{aligned}$$

and

$$\begin{aligned} SSE &= (1 - R^2)SST \\ &= (1 - .78875)1014.75 = 214.3659 \end{aligned}$$

- these answers differ slightly from above due to round-off error

A statistical model for simple linear regression

- we assume that an observed response value y_i is related to its predictor x_i according to the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- where β_0 and β_1 are the intercept and slope
- ϵ_i is an additive random deviation or 'error', assumed to have zero mean and constant variance σ^2
- any two deviations ϵ_i and ϵ_j are assumed to be independent

- the mean of y_i is

$$\mu_{x_i} = \beta_0 + \beta_1 x_i$$

  which is linear in x_i

- the variance is assumed to be the same for each case, and this justifies giving each case the same weight when minimizing SSE

- under these assumptions, the least squares estimators

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}}$$

  and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

  have good statistical properties

- among all linear unbiased estimators, they have minimum variance

- an unbiased estimator has a sampling distribution with mean equal to the parameter being estimated

- the variance of the deviations σ^2 is estimated using the average squared residual,

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{SSE}{n-2} = MSE$$

  where division is by $n-2$ here because two β's have been estimated

- to make inferences about the model parameters we also need to assume that the deviations ϵ_i are normally distributed

Statistical inferences for regression

Standard errors for regression coefficients

- regression coefficient values, $\hat{\beta}_0$ and $\hat{\beta}_1$, are point estimates of the true intercept and slope, β_0 and β_1 respectively.

- using our assumptions about the deviations, and the rules for mean and variance, the sampling distribution of the slope estimator can be derived to be

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{SS_{xx}})$$

- this means that if we had a large number of data sets and calculated the slope estimate each time, their histogram would look normal, be centered around the true slope and have variance as given above

- the standard deviation of $\hat{\beta}_1$ is $\sqrt{\frac{\sigma^2}{SS_{xx}}}$

- the value of σ^2 is unknown, so the estimator MSE is used in its place to produce the standard error of the estimate $\hat{\beta}_1$, as

$$SE_{\hat{\beta}_1} = \frac{\sqrt{MSE}}{\sqrt{SS_{xx}}} = \frac{s}{\sqrt{SS_{xx}}}$$

- the standard error for the intercept estimator $\hat{\beta}_0$ is

$$SE_{\hat{\beta}_0} = \sqrt{MSE(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}})}$$

Example: Ozone data

- standard errors for the regression coefficients are estimated below.
- $SS_{xx} = .009275$ and $MSE = 107.80$
- $SE_{\hat{\beta}_1} = \sqrt{MSE/SS_{xx}} = \sqrt{107.80/.009275} = 107.81$

- $SE_{\hat{\beta}_0} = \sqrt{MSE(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}})} = \sqrt{107.80((1/4) + (.0399/.009275))} = 10.77$

Tests for regression coefficients

- the most common and useful test is whether or not the relationship between the response and predictor is significant
- $H_0 : \beta_1 = 0$, *there is no linear relationship*
- $H_a : \beta_1 \neq 0$, *there is a linear relationship*
- the alternative is usually two sided
- the test statistic is

$$T = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$$

  and this is compared to the t_{n-2} distribution

- on occasion, we specify a value $\beta_{1,0}$ other than 0 in the null hypothesis
- then the test statistic becomes

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE_{\hat{\beta}_1}}$$

- one can also test hypotheses about the intercept
- $H_0 : \beta_0 = \beta_{0,0}$,
- $H_a : \beta_0 \neq \beta_{0,0}$
- often we are interested in whether the intercept is zero
- the test statistic is

$$T = \frac{\hat{\beta}_0 - \beta_{0,0}}{SE_{\hat{\beta}_0}}$$

and this is compared to the t_{n-2} distribution

Example: Ozone data

- we saw $\hat{\beta}_1 = -293.531$ and $SE_{\hat{\beta}_1} = 107.81$
- the test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ gives

$$T = \frac{-293.531}{107.81} = -2.7227$$

- comparing to the $t_{4-2=2}$ distribution gives $P = .11$ exactly, or $.10 < P < .20$ using the tables
- in spite of the high correlation calculated earlier, the relationship between ozone and yield is not significant using $\alpha = .10$ or smaller

Example: Tree data.

- earlier we obtained $\hat{\beta}_1 = 11.036$, $n = 20$, $r = .976$, $s_y = 91.7$ and $s_x = 8.1$ for the straight line fit

- we can determine that

$$SS_{XX} = 19s_x^2 = 19(8.1)^2 = 1246.59$$

and

$$SST = SS_{YY} = 19(91.7)^2 = 159,768.9$$

- from this we can calculate

$$\begin{aligned} SSE &= (1 - R^2)SST \\ &= (1 - .976^2)159768.9 = 7576.88 \end{aligned}$$

and

$$MSE = \frac{SSE}{n-2} = \frac{7576.88}{18} = 420.9378$$

- the standard error of the slope estimate is

$$\begin{aligned} SE_{\hat{\beta}_1} &= \sqrt{\frac{MSE}{SS_{XX}}} \\ &= \sqrt{\frac{420.9378}{1246.59}} = .5811 \end{aligned}$$

- the test statistic for an association between diameter and usable volume is

$$T = \frac{11.036}{.5811} = 18.99$$

and there are $20 - 2 = 18$ degrees of freedom

- the P value is less than .01, using the tables, so we conclude that the linear association between usable volume and diameter at chest height is statistically significant

- if you compare with the computer output shown earlier, you will see that the values calculated by hand are slightly different, due to round-off error

```
MTB > regress c2 1 c1;
SUBC> residuals c3.
The regression equation is
volume = - 191 + 11.0 diameter

Predictor        Coef        Stdev    t-ratio        p
```

Constant	-191.12	16.98	-11.25	0.000
diameter	11.0413	0.5752	19.19	0.000

s = 20.33   R-sq = 95.3%   R-sq(adj) = 95.1%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	152259	152259	368.43	0.000
Error	18	7439	413		
Total	19	159698			

Confidence intervals for regression coefficients

- confidence intervals are constructed using the standard errors as follows

$$\hat{\beta}_i \pm t_{\alpha/2,n-2} SE_{\hat{\beta}_i}$$

  for $i = 0$ or $i = 1$

- the degrees of freedom for the t distribution are the same as the degrees of freedom associated with MSE

Example: Ozone data

- 95% confidence intervals for β_1 and β_0 are computed as follows

- $t_{\alpha/2,n-2} = t_{.025,2} = 4.303$
- for the slope, β_1:
  $-293.531 \pm 4.303(107.81)$

$$(-757.4, \quad 170.3)$$

- note that this interval contains zero, which confirms that the slope is not significantly different from zero
- for the intercept, β_0:
  $253.434 \pm 4.303(10.77)$

$$(207.1, \quad 299.8)$$

Estimating the mean of Y at $x = x^*$

- the estimated mean of Y when $x = x^*$ is

$$\hat{\mu}_{x^*} = \hat{\beta}_0 + \hat{\beta}_1 x^* = \bar{y} + \hat{\beta}_1(x^* - \bar{x})$$

- because both $\hat{\beta}_0$ and $\hat{\beta}_1$ have normal sampling distributions, μ_{x^*} does as well

- the mean of this distribution is the true mean

$$\mu_{x^*} = \beta_0 + \beta_1 x^*$$

because both $\hat{\beta}_0$ and $\hat{\beta}_1$ have means equal to their population values

- the variance of $\hat{\mu}_{x^*}$ is

$$\sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}} \right)$$

which is the sum of the variances of $\bar{y}$ and $\hat{\beta}_1(x^* - \bar{x})$

- in short

$$\hat{\mu}_{x^*} \sim N\left(\beta_0 + \beta_1 x^*, \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}} \right) \right)$$

- the standard error of $\hat{\mu}_{x^*}$ is

$$SE_{\hat{\mu}_{x^*}} = \sqrt{MSE \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}} \right)}$$

- a confidence interval for the mean $\mu_{x^*} = \beta_0 + \beta_1 x^*$ when $x = x^*$ is given by

$$\hat{\mu}_{x^*} \pm t_{\alpha/2,n-2} SE_{\hat{\mu}_{x^*}}$$

Example: Ozone data

- a 95% confidence interval for the mean yield at $x = 0.10$ is obtained as follows
- when $x^* = 0.10$, the estimated mean is

$$\hat{\mu}_{.1} = 253.434 - 293.531(0.1) = 224.08$$

- the standard error of this estimate is

$$SE_{\hat{\mu}_{.1}} = \sqrt{107.8\left(\frac{1}{4} + \frac{(0.1 - .0875)^2}{.009275}\right)} = 5.36$$

- the table value is $t_{\alpha/2,n-2} = t_{.025,2} = 4.303$
- the half width of the interval, or margin of error, is

$$t_{\alpha/2,n-2} SE_{\hat{\mu}_{.1}} = 4.303(5.36) = 23.08$$

- so the interval is 224.08 ± 23.08 or

$$(201, 247.16)$$

Predicting a new response value at $x = x^*$

- in making a **prediction interval** for a future observation on y when $x = x^*$, we need to incorporate two sources of variation
- the first is the variation in the estimate $\hat{\mu}_{x^*}$ about the actual mean μ_{x^*}
- the second is the variation of the new response y about its mean
- the error of prediction is

$$\begin{aligned} y - (\hat{\beta}_0 + \hat{\beta}_1 x^*) &= (y - (\beta_0 + \beta_1 x^*)) - \\ &\quad (\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)) \end{aligned}$$

- the first term in brackets on the right hand side of this expression is ϵ^*, which has a $N(0, \sigma^2)$ distribution.

- the second term is the deviation of $\hat{\mu}_{x^*}$ from the actual mean μ_{x^*} which we have seen is

$$N\left(0, \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}\right)\right)$$

- as y represents a future observation, the distributions of the two terms are independent, and it follows that the distribution of the prediction error $y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$ is

$$N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}\right)\right)$$

- the standard error of the prediction error is estimated by

$$\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}\right)}$$

- and the prediction interval for y is given by

$$\hat{\beta}_0+\hat{\beta}_1 x^* \pm t_{\alpha/2,n-2}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{SS_{xx}}\right)}$$

Ozone example: A 95% prediction interval

for y when $x = 0.10$ is calculated.

- when $x^* = 0.10$, the prediction is

$$\hat{\mu}_{x^*} = 253.434 - 293.531(0.1) = 224.08$$

- the standard error of prediction is

$$\begin{aligned} SE_{y^*} &= \sqrt{107.8\left(1+\frac{1}{4}+\frac{(0.1-.0875)^2}{.009275}\right)} \\ &= 11.69 \end{aligned}$$

- the margin of error is $t_{\alpha/2,n-2} SE_{y^*} = 4.303(11.69) = 50.29$
- so the prediction interval is

$$224.08 \pm 50.29$$

- or $(173.79, 274.37)$

Tree example: Minitab can be used to find confidence intervals for the mean at x^* and for prediction intervals for a new value at x^*.

- the output below was obtained using $Stat > Regression > Options$, where a diameter of 30 in. was used

```
MTB > Name c3 "CLIM1" c4 "CLIM2" c5 "PLIM1" c6 "PLIM2"
MTB > Regress c2 1 c1;
SUBC>   Constant;
SUBC>   Predict 30;
SUBC>     CLimits 'CLIM1'-'CLIM2';
SUBC>     PLimits 'PLIM1'-'PLIM2';
SUBC>   Brief 2.

Regression Analysis: C2 versus C1

The regression equation is
C2 = - 191 + 11.0 C1


Predictor      Coef  SE Coef       T      P
Constant   -191.12    16.98  -11.25  0.000
C1         11.0413   0.5752   19.19  0.000
```

```
S = 20.3290   R-Sq = 95.3%   R-Sq(adj) = 95.1%


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  152259  152259  368.43  0.000
Residual Error  18    7439     413
Total           19  159698


Predicted Values for New Observations

New
Obs     Fit  SE Fit        95% CI              95% PI
  1  140.11    4.63  (130.38, 149.85)  (96.31, 183.92)


Values of Predictors for New Observations

New
Obs    C1
  1  30.0
```

- for this dataset we previously saw that $n = 20$, $SS_{XX} = 1246.59$ and $MSE = 420.9378$

- the mean diameter is $\bar{x} = 28.45$, so the standard error for estimating the mean at $X = 30$ is

$$
\begin{aligned}
SE_{\hat{\mu}_{x*}} &= \sqrt{420.9378 * \left(\frac{1}{20} + \frac{(30 - 28.45)^2}{1246.59}\right)} \\
&= 4.6753
\end{aligned}
$$

- this is close to the $SE\ Fit$ given in the output