Stat 328 Regression Summary

Simple Linear Regression Model

The basic (normal) "simple linear regression" model says that a response/output variable y depends on an explanatory/input/system variable x in a "noisy but linear" way. That is, one supposes that there is a linear relationship between x and mean y,

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

and that (for fixed x) there is around that mean a distribution of y that is normal. Further, the model assumption is that the standard deviation of the response distribution is constant in x. In symbols it is standard to write

$$y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is normal with mean 0 and standard deviation σ. This describes one y. Where several observations y_i with corresponding values x_i are under consideration, the assumption is that the y_i (the ϵ_i) are independent. (The ϵ_i are conceptually equivalent to unrelated random draws from the same fixed normal continuous distribution.) The model statement in its full glory is then

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for $i = 1, 2, ..., n$
ϵ_i for $i = 1, 2, ..., n$ are independent normal $(0, \sigma^2)$ random variables

The model statement above is a perfectly theoretical matter. One can begin with it, and for specific choices of β_0, β_1 and σ find probabilities for y at given values of x. In applications, the real mode of operation is instead to take n data pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ and use them to make *inferences about the parameters* β_0, β_1 and σ and to make *predictions based on the estimates* (based on the empirically fitted model).

Descriptive Analysis of Approximately Linear (x, y) Data

After plotting (x, y) data to determine that the "linear in x mean of y" model makes some sense, it is reasonable to try to quantify "how linear" the data look and to find a line of "best fit" to the scatterplot of the data. The **sample correlation between y and x**

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

is a measure of strength of linear relationship between x and y.

Calculus can be invoked to find a slope β_1 and intercept β_0 minimizing the sum of squared vertical distances from data points to a fitted line, $\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$. These "**least squares**" values are

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and

$$b_0 = \overline{y} - b_1\overline{x}$$

It is further common to refer to the value of y on the "least squares line" corresponding to x_i as a **fitted or predicted value**

$$\widehat{y}_i = b_0 + b_1 x_i$$

One might take the difference between what is observed (y_i) and what is "predicted" or "explained" ($\widehat{y}_i$) as a kind of leftover part or "**residual**" corresponding to a data value

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

The sum $\sum_{i=1}^{n}(y_i - \overline{y})^2$ is most of the sample variance of the n values y_i. It is a measure of raw variation in the response variable. People often call it the "**total sum of squares**" and write

$$SStot = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

The sum of squared residuals $\sum_{i=1}^{n}e_i^2$ is a measure of variation in response remaining unaccounted for after fitting a line to the data. People often call it the "**error sum of squares**" and write

$$SSE = \sum_{i=1}^{n}e_i^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

One is guaranteed that $SSTot \geq SSE$. So the difference $SSTot - SSE$ is a non-negative measure of variation accounted for in fitting a line to (x, y) data. People often call it the "**regression sum of squares**" and write

$$SSR = SStot - SSE$$

The **coefficient of determination** expresses SSR as a fraction of $SSTot$ and is

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

which is interpreted as "the fraction of raw variation in y accounted for in the model fitting process."

Parameter Estimates for SLR

The descriptive statistics for (x, y) data can be used to provide "single number estimates" of the (typically unknown) parameters of the simple linear regression model. That is, the slope of the least squares line can serve as an estimate of β_1,

$$\widehat{\beta}_1 = b_1$$

and the intercept of the least squares line can serve as an estimate of β_0,

$$\widehat{\beta}_0 = b_0$$

The variance of y for a given x can be estimated by a kind of average of squared residuals

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-2}$$

Of course, the square root of this "regression sample variance" is $s = \sqrt{s^2}$ and serves as a single number estimate of σ.

Interval-Based Inference Methods for SLR

The normal simple linear regression model provides inference formulas for model parameters. The normal simple linear regression model provides inference formulas for model parameters. **Confidence limits for σ** are

$$s\sqrt{\frac{n-2}{\chi^2_{\text{upper}}}} \quad \text{and} \quad s\sqrt{\frac{n-2}{\chi^2_{\text{lower}}}}$$

where χ^2_{upper} and χ^2_{lower} are upper and lower percentage points of the χ^2 distribution with $\nu = n - 2$ degrees of freedom. And **confidence limits for β_1** (the slope of the line relating mean y to x ... the rate of change of average y with respect to x) are

$$b_1 \pm t\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

where t is a quantile of the t distribution with $\nu = n - 2$ degrees of freedom. The text calls the ratio $s/\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}$ the standard error of b_1 and uses the symbols SE_{b_1} for it. In these terms, the confidence limits for β_1 are

$$b_1 \pm tSE_{b_1}$$

Confidence limits for $\mu_{y|x} = \beta_0 + \beta_1 x$ (the mean value of y at a given value x) are

$$(b_0 + b_1 x) \pm ts\sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

One can abbreviate $(b_0 + b_1 x)$ as $\widehat{y}$, and the text calls $s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$ the standard error of the mean and uses the symbol $SE_{\widehat{\mu}}$ for it. (JMP unfortunately call$s$ this the Std Err Pred y.) In these terms, the confidence limits for $\mu_{y|x}$ are

$$\widehat{y} \pm tSE_{\widehat{\mu}}$$

(Note that by choosing $x = 0$, this formula provides confidence limits for β_0, though this parameter is rarely of independent practical interest.)

Prediction limits for an additional observation y at a given value x are

$$(b_0 + b_1 x) \pm ts\sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

The text (unfortunately) calls $s\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$ the standard error for "estimating" an individual y and uses the symbol $SE_{\widehat{y}}$ for it. (JMP equally unfortunately calls this "Std Err Individ y.") It is on occasion useful to notice that

$$SE_{\widehat{y}} = \sqrt{s^2 + SE_{\widehat{\mu}}^2}$$

Using the $s_{\widehat{y}}$ notation, prediction limits for an additional y at x are

$$\widehat{y} \pm tSE_{\widehat{y}}$$

Hypothesis Tests and SLR

The normal simple linear regression model supports hypothesis testing. $\mathbf{H_0: \beta_1 = \#}$ can be tested using the test statistic

$$T = \frac{b_1 - \#}{\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}} = \frac{b_1 - \#}{SE_{b_1}}$$

and a t_{n-2} reference distribution. $\mathbf{H_0: \mu_{y|x} = \#}$ can be tested using the test statistic

$$T = \frac{(b_0 + b_1 x) - \#}{s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}} = \frac{\widehat{y} - \#}{SE_{\widehat{\mu}}}$$

and a t_{n-2} reference distribution.

ANOVA and SLR

The breaking down of $SSTot$ into SSR and SSE can be thought of as a kind of "**an**alysis **o**f **va**riance" in y. That enterprise is often summarized in a special kind of table. The general form is as below.

ANOVA Table (for SLR)

Source	SS	df	MS	F
Regression	SSR	1	$MSR = SSR/1$	$F = MSR/MSE$
Error	SSE	$n-2$	$MSE = SSE/(n-2)$	
Total	$SStot$	$n-1$		

In this table the ratios of sums of squares to degrees of freedom are called "mean squares." The mean square for error is, in fact, the estimate of σ^2 (i.e. $MSE = s^2$).

As it turns out, the ratio in the "F" column can be used as a test statistic for the hypothesis $\mathbf{H_0: \beta_1 = 0}$. The reference distribution appropriate is the $F_{1,n-2}$ distribution. As it turns out, the value $F = MSR/MSE$ is the square of the t statistic for testing this hypothesis, and the F test produces exactly the same p-values as a two-sided t test.

Standardized Residuals and SLR

The theoretical variances of the residuals turn out to depend upon their corresponding x values. As a means of putting these residuals all on the same footing, it is common to "standardize" them by dividing by an estimated standard deviation for each. This

produces **standardized residuals**

$$e_i^* = \frac{e_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}}$$

These (if the normal simple linear regression model is a good one) "ought" to look as if they are approximately normal with mean 0 and standard deviation 1. Various kinds of plotting with these standardized residuals (or with the raw residuals) are used as means of "model checking" or "model diagnostics."

Multiple Linear Regression Model

The basic (normal) "multiple linear regression" model says that a response/output variable y depends on explanatory/input/system variables $x_1, x_2, \ldots, x_k$ in a "noisy but linear" way. That is, one supposes that there is a linear relationship between $x_1, x_2, \ldots, x_k$ and mean y,

$$\mu_{y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

and that (for fixed $x_1, x_2, \ldots, x_k$) there is around that mean a distribution of y that is normal. Further, the standard assumption is that the standard deviation of the response distribution is constant in $x_1, x_2, \ldots x_k$. In symbols it is standard to write

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

where ϵ is normal with mean 0 and standard deviation σ. This describes one y. Where several observations y_i with corresponding values $x_{1i}, x_{2i}, \ldots, x_{ki}$ are under consideration, the assumption is that the y_i (the ϵ_i) are independent. (The ϵ_i are conceptually equivalent to unrelated random draws from the same fixed normal continuous distribution.) The model statement in its full glory is then

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i \quad \text{for } i = 1, 2, ..., n$$
$$\epsilon_i \text{ for } i = 1, 2, ..., n \text{ are independent normal } (0, \sigma^2) \text{ random variables}$$

The model statement above is a perfectly theoretical matter. One can begin with it, and for specific choices of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ and σ find probabilities for y at given values of $x_1, x_2, \ldots, x_k$. In applications, the real mode of operation is instead to take n data vectors $(x_{11}, x_{21}, \ldots, x_{k1}, y_1), (x_{12}, x_{22}, \ldots, x_{k2}, y_2), ..., (x_{1n}, x_{2n}, \ldots, x_{kn}, y_n)$ and use them to make *inferences about the parameters* $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ and σ and to make *predictions based on the estimates* (based on the empirically fitted model).

Descriptive Analysis of Approximately Linear $(x_1, x_2, \ldots, x_k, y)$ Data

Calculus can be invoked to find coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ minimizing the sum of squared vertical distances from data points in $(k+1)$ dimensional space to a fitted surface, $\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}))^2$. These "**least squares**" values DO NOT have simple formulas (unless one is willing to use matrix notation). And in particular, one can NOT simply somehow use the formulas from simple linear regression in this more complicated context. We will call these minimizing coefficients $b_0, b_1, b_2, \ldots, b_k$ and need to rely upon JMP to produce them for us.

It is further common to refer to the value of y on the "least squares surface" corresponding to $x_{1i}, x_{2i}, \ldots, x_{ki}$ as a **fitted or predicted value**

$$\widehat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$$

Exactly as in SLR, one takes the difference between what is observed (y_i) and what is "predicted" or "explained" ($\widehat{y}_i$) as a kind of leftover part or "**residual**" corresponding to a data value

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

The **total sum of squares**, $SSTot = \sum_{i=1}^{n}(y_i - \overline{y})^2$, is (still) most of the sample variance of the n values y_i and measures raw variation in the response variable. Just as in SLR, the sum of squared residuals $\sum_{i=1}^{n} e_i^2$ is a measure of variation in response remaining unaccounted for after fitting the equation to the data. As in SLR, people call it the **error sum of squares** and write

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

(The formula looks exactly like the one for SLR. It is simply the case that now $\widehat{y}_i$ is computed using all k inputs, not just a single x.) One is still guaranteed that $SSTot \geq SSE$. So the difference $SSTot - SSE$ is a non-negative measure of variation accounted for in fitting the linear equation to the data. As in SLR, people call it the **regression sum of squares** and write

$$SSR = SStot - SSE$$

The **coefficient of (multiple) determination** expresses SSR as a fraction of $SSTot$ and is

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

which is interpreted as "the fraction of raw variation in y accounted for in the model fitting process." This quantity can also be interpreted in terms of a correlation, as it turns out to be the square of the sample linear correlation between the observations y_i and the fitted or predicted values $\widehat{y}_i$.

Parameter Estimates for MLR

The descriptive statistics for $(x_1, x_2, \ldots, x_k, y)$ data can be used to provide "single number estimates" of the (typically unknown) parameters of the multiple linear regression model. That is, the least squares coefficients $b_0, b_1, b_2, \ldots, b_k$ serve as estimates of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. The first of these is a kind of high-dimensional "intercept" and in the case where the predictors are not functionally related, the others serve as rates of change of average y with respect to a single x, *provided the other x's are held fixed*.

The variance of y for a fixed values $x_1, x_2, \ldots, x_k$ can be estimated by a kind of average of squared residuals

$$s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-k-1}$$

The square root of this "regression sample variance" is $s = \sqrt{s^2}$ and serves as a single number estimate of σ.

Interval-Based Inference Methods for MLR

The normal multiple linear regression model provides inference formulas for model parameters. **Confidence limits for σ** are

$$s\sqrt{\frac{n-k-1}{\chi^2_{\text{upper}}}} \quad \text{and} \quad s\sqrt{\frac{n-k-1}{\chi^2_{\text{lower}}}}$$

where χ^2_{upper} and χ^2_{lower} are upper and lower percentage points of the χ^2 distribution with $\nu = n-k-1$ degrees of freedom. **Confidence limits for β_j** (the rate of change of average y with respect to x_j) are

$$b_j \pm tSE_{b_j}$$

where t is a quantile of the t distribution with $\nu = n-k-1$ degrees of freedom and SE_{b_j} is a standard error of b_j. There is no simple formula for SE_{b_j}, and in particular, one can NOT simply somehow use the formula from simple linear regression in this more complicated context. It IS the case that this standard error is a multiple of s, but we will have to rely upon JMP to provide it for us.

Confidence limits for $\mu_{y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ (the mean value of y at a particular choice of the $x_1, x_2, \ldots, x_k$) are

$$\widehat{y} \pm tSE_{\widehat{\mu}}$$

where $SE_{\widehat{\mu}}$ depends upon which values of $x_1, x_2, \ldots, x_k$ are under consideration. There is no simple formula for $s_{\widehat{\mu}}$ and in particular one can NOT simply somehow use the formula from simple linear regression in this more complicated context. It IS the case that this standard error is a multiple of s, but we will have to rely upon JMP to provide it for us.

Prediction limits for an additional observation y at a given vector $(x_1, x_2, \ldots, x_k)$ are

$$\widehat{y} \pm tSE_{\widehat{y}}$$

where it remains true (as in SLR) that $SE_{\widehat{y}} = \sqrt{s^2 + SE_{\widehat{\mu}}^2}$ but otherwise there are no simple formulas for it, and in particular one can NOT simply somehow use the formula from simple linear regression in this more complicated context.

Hypothesis Tests and MLR

The normal multiple linear regression model supports hypothesis testing. **$H_0: \beta_j = \#$** can be tested using the test statistic

$$T = \frac{b_j - \#}{SE_{b_j}}$$

and a t_{n-k-1} reference distribution. **$H_0: \mu_{y|x_1,x_2,\ldots,x_k} = \#$** can be tested using the test statistic

$$T = \frac{\widehat{y} - \#}{SE_{\widehat{\mu}}}$$

and a t_{n-k-1} reference distribution.

ANOVA and MLR

As in SLR, the breaking down of $SSTot$ into SSR and SSE can be thought of as a kind of "**analysis of variance**" in y, and summarized in a special kind of table. The general form for MLR is as below.

ANOVA Table (for MLR Overall F Test)

Source	SS	df	MS	F
Regression	SSR	k	$MSR = SSR/k$	$F = MSR/MSE$
Error	SSE	$n-k-1$	$MSE = SSE/(n-k-1)$	
Total	$SStot$	$n-1$		

(Note that as in SLR, the mean square for error is, in fact, the estimate of σ^2 (i.e. $MSE = s^2$).)

As it turns out, the ratio in the "F" column can be used as a test statistic for the hypothesis $\mathbf{H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0}$. The reference distribution appropriate is the $F_{k,n-k-1}$ distribution.

"Partial F Tests" in MLR

It is possible to use ANOVA ideas to invent F tests for investigating whether some whole group of β's (short of the entire set) are all 0. For example one might want to test the hypothesis

$$\text{H}_0: \beta_{l+1} = \beta_{l+2} = \cdots = \beta_k = 0$$

(This is the hypothesis that only the first l of the k input variables x_i have any impact on the mean system response ... the hypothesis that the first l of the x's are adequate to predict y ... the hypothesis that after accounting for the first l of the x's, the others do not contribute "significantly" to one's ability to explain or predict y.)

If we call the model for y in terms of all k of the predictors the **"full model"** and the model for y involving only x_1 through x_l the **"reduced model"** then an F test of the above hypothesis can be made using the statistic

$$F = \frac{(SSR_{\text{Full}} - SSR_{\text{Reduced}})/(k-l)}{MSE_{\text{Full}}} = \left(\frac{n-k-1}{k-l}\right)\left(\frac{R^2_{\text{Full}} - R^2_{\text{Reduced}}}{1 - R^2_{\text{Full}}}\right)$$

and an $F_{k-l,n-k-1}$ reference distribution. $SSR_{\text{Full}} \geq SSR_{\text{Reduced}}$ so the numerator here is non-negative.

Finding a p-value for this kind of test is a means of judging whether R^2 for the full model is "significantly"/detectably larger than R^2 for the reduced model. (Caution here, statistical significant is not the same as practical importance. With a big enough data set, essentially *any* increase in R^2 will produce a small p-value.) It is reasonably common to expand the basic MLR ANOVA table to organize calculations for this test statistic. This is

(Expanded) ANOVA Table (for MLR)

Source	SS	df	MS	F
Regression	SSR_{Full}	k	$MSR_{\text{Full}} = SSR/k$	$F = MSR_{\text{Full}}/MSE_{\text{Full}}$
$x_1, ..., x_l$	SSR_{Red}	l		
$x_{l+1}, ..., x_k \vert x_l, ..., x_k$	$SSR_{\text{Full}} - SSR_{\text{Red}}$	$k-l$	$(SSR_{\text{Full}} - SSR_{\text{Red}})/(k-l)$	$\frac{(SSR_{\text{Full}} - SSR_{\text{Red}})/(k-l)}{MSE_{\text{Full}}}$
Error	SSE_{Full}	$n-k-1$	$MSE_{\text{Full}} = SSE_{\text{Full}}/(n-k-1)$	
Total	$SStot$	$n-1$		

Standardized Residuals in MLR

As in SLR, people sometimes wish to standardize residuals before using them to do model checking/diagnostics. While it is not possible to give a simple formula for the "standard error of e_i" without using matrix notation, most MLR programs will compute these values. The standardized residual for data point i is then (as in SLR)

$$e_i^* = \frac{e_i}{\text{standard error of } e_i}$$

If the normal multiple linear regression model is a good one these "ought" to look as if they are approximately normal with mean 0 and standard deviation 1.

Intervals and Tests for Linear Combinations of β's in MLR

It is sometimes important to do inference for a linear combination of MLR model coefficients

$$L = c_0\beta_0 + c_1\beta_1 + c_2\beta_2 + \cdots + c_k\beta_k$$

(where $c_0, c_1, \ldots, c_k$ are known constants). Note, for example, that $\mu_{y|x_1,x_2,\ldots,x_k}$ is of this form for $c_0 = 1, c_1 = x_1, c_2 = x_2, \ldots,$ and $c_k = x_k$. Note too that a difference in mean responses at two sets of predictors, say $(x_1, x_2, \ldots, x_k)$ and $(x_1', x_2', \ldots, x_k')$ is of this form for $c_0 = 0, c_1 = x_1 - x_1', c_2 = x_2 - x_2', \ldots,$ and $c_k = x_k - x_k'$.

An obvious estimate of L is

$$\widehat{L} = c_0b_0 + c_1b_1 + c_2b_2 + \cdots + c_kb_k$$

Confidence limits for L are

$$\widehat{L} \pm t(\text{standard error of } \widehat{L})$$

There is no simple formula for "standard error of $\widehat{L}$". This standard error is a multiple of s, but we will have to rely upon JMP to provide it for us. (Computation of $\widehat{L}$ and its standard error is under the "Custom Test" option in JMP.)

H_0: L = # can be tested using the test statistic

$$T = \frac{\widehat{L} - \#}{\text{standard error of } \widehat{L}}$$

and a t_{n-k-1} reference distribution. Or, if one thinks about it for a while, it is possible to find a reduced model that corresponds to the restriction that the null hypothesis places on the MLR model and to use a "$l = k - 1$" partial F test (with 1 and $n - k - 1$ degrees of freedom) equivalent to the t test for this purpose.