STAT 500

Simple Linear Regression: ANOVA Table and R^2

Regression Analysis: ANOVA

- Write the deviation from the overall sample mean as $Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$ where $\hat{Y}_i = b_o + b_1 X_i$

- Partition the corrected total sums of squares

$$
\begin{aligned}
SS_{corrected\ total} &= \sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\
&= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 \\
&\quad +2 \sum_i (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\
&= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 \\
&= SS_{residuals} + SS_{model}
\end{aligned}
$$

Regression Analysis: ANOVA

- Cross product term is

$$
\begin{aligned}
2\sum_i (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) &= 2\sum_i e_i(b_o + b_1 x_i - \bar{Y}) \\
&= 2(b_o - \bar{Y})\sum_i e_i + 2b_1 \sum_i e_i x_i \\
&= 0 \quad \text{because } \sum_i e_i = \sum_i e_i x_i = 0
\end{aligned}
$$

Note that

$$
\begin{aligned}
SS_{model} &= \sum_i (\hat{Y}_i - \bar{Y})^2 = \sum_i (b_o + b_1 x_i - \bar{Y})^2 \\
&= b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2
\end{aligned}
$$

Regression Analysis: ANOVA

- $SS_{model} = SS_{total}$ - SS_{error}

$$= \Sigma_i (\hat{Y}_i - \bar{Y})^2$$
$$= \Sigma_i (b_o + b_1 x_i - \bar{Y})^2$$
$$= b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- SS_{model} is also denoted by $SS_{regression}$

- SS_{error} is also denoted by $SS_{residuals}$ or SSE

Regression Analysis: ANOVA

SS_{error} has $n-2$ degrees of freedom because

- Two parameters must be estimated to calculate $\hat{Y}_i$
- The residuals satisfy two constraints

$$\sum e_i = 0 \qquad \text{and} \qquad \sum e_i x_i = 0$$

ANOVA Table

Source	df	Sums of Squares
Model	1	$SS_{\text{model}} = \Sigma_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
Error	$n-2$	$SS_{\text{error}} = \Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
Total	$n-1$	$SS_{\text{Total}} = \Sigma_{i=1}^{n}(Y_i - \bar{Y})^2$

ANOVA Table: Example

Source	df	SS	MS	F	p-value
Model	1	0.22573	0.22573	2961.55	$< \mathbf{0.0001}$
Error	15	0.00114	0.00007622		
Total	16	0.22688			

Mean Squares

- MS_{error}
  - $\hat{\sigma}^2 = MS_{\text{error}} = SS_{\text{error}}/(n-2)$
  - $\hat{\sigma}^2$ is an unbiased estimate of σ^2

$$E(MS_{\text{error}}) = \sigma^2$$

- MS_{model}
  - $E(MS_{\text{model}}) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$
  - When β_1=0, $E(MS_{\text{model}}) = \sigma^2$. Otherwise, $E(MS_{\text{model}}) > \sigma^2$.

F-test for Significance of Model

- $H_0 : \beta_1 = 0 \rightarrow Y_i = \beta_0 + \epsilon_i$
- $H_a : \beta_1 \neq 0 \rightarrow Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- Test Statistic:

$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$

- Reject H_0 if

$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}} > F_{1,n-2,1-\alpha}$$

F-test: Example

- $H_0 : \beta_1 = 0$
- $H_a : \beta_1 \neq 0$
- $F = 2961.55$ with p-value < 0.0001.
- Reject H_0 and conclude there is a significant linear relationship between boiling point of water and log of barometric pressure.

Coefficient of Determination (R^2)

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{Total}}}$$

- Fraction of variation in the response variable that can be explained by the linear regression model with the explanatory variable x.

- Expressed as percentage: $0\% \le R^2 \le 100\%$

- Large values of R^2 indicate better model fit.

R^2: Example

$$\boldsymbol{R^2 = \frac{SS_{\textsf{model}}}{SS_{\textsf{Total}}} = \frac{0.22573}{0.22688} = 0.9950}$$

99.50% of the variation in log(barometric pressure) can be explained by the linear regression model with boiling point of water.

STAT 500

Simple Linear Regression: Inference for Parameters and Prediction Intervals

Inference for Model Parameters

- Population Slope - $\boldsymbol{\beta_1}$
- Population Intercept - $\boldsymbol{\beta_0}$
- Conditional Mean - $\boldsymbol{\mu_{Y|x}}$

Inference for the Slope (β_1)

- Discuss inference for β_1 in detail (then summarize the rest)

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- b_1 is a linear combination of normal random variables (the Y_i's) so b_1 is normally distributed with

$$E(b_1) = \beta_1 \qquad \text{Var}(b_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- $b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$

Inference for the Slope (β_1)

Examine

$$\text{Var}(b_1) = \frac{\sigma^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}$$

A more precise estimate of β_1 can be obtained by :

- Spreading out the X values
- Getting a larger sample i.e. more (X, Y) pairs
- Making the error variance smaller

Inference for the Slope (β_1)

- Use MS_{error} to estimate σ^2

$$(\text{Note that } MS_{error} \sim \frac{\sigma^2\chi^2_{n-2}}{n-2} \;)$$

- Standard error of b_1 is $S_{b_1} = \sqrt{MS_{error} / \sum_{i=1}^{n}(x_i - \bar{x})^2}$

- $(b_1 - \beta_1)/S_{b_1}$ has a t-distribution with $n-2$ d.f.

Hypothesis Test for β_1

- Null and Alternative Hypotheses

$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 \neq 0$$

- Test Statistic

$$T = \frac{b_1 - 0}{S_{b_1}}$$

- Reject H_0 if $|T| > t_{n-2,1-\alpha/2}$
- Note that $T^2 = F$, this t-test for β_1 is the same as the F-test for significance of model from ANOVA Table.
- One-sided alternative hypothesis is possible for the t-test: $H_a : \beta_1 > 0$ or $H_a : \beta_1 < 0$

CI for β_1

- $100(1-\alpha)\%$ confidence interval for β_1:

$$b_1 \pm t_{n-2,1-\alpha/2} S_{b_1}$$

Forbes Data

Weisberg, Sanford, *Applied Linear Regression*, Wiley, 1980.

- James D. Forbes collected data in the mountains of Scotland
- n=17 locations (at different altitudes)
- Objective: Predict barometric pressure (in inches of mercury) from boiling point of water (X) in ^{o}F.
- Use Y=log(barometric pressure)
- Motivation: Fragile barometers of the 1840's were difficult to transport

Analysis of the Forbes Data

- Test $H_o : \beta_1 = 0 \ (Y_i = \beta_0 + \epsilon_i)$
  versus $H_a : \beta_1 \neq 0 \ (Y_i = \beta_0 + \beta_1 x_i + \epsilon_i)$

- Evaluate

$$t = \frac{b_1 - 0}{S_{b_1}} = \frac{.020623 - 0}{0.000379} = 54.42$$

- The least squares estimate of the slope is 54 standard errors away from zero (p-value $<<$ .0001).
  It is extremely unlikely that an estimate that far from zero could occur simply because of random errors when β_1 is actually zero.
  Consequently, reject the null hypothesis and conclude that the slope is positive.

Analysis of the Forbes Data

- A 95% confidence interval for the slope indicates that the slope is "very well" estimated from these data

$$b_1 \pm t_{15,.975} S_{b_1}$$

$$\Rightarrow \quad \mathbf{0.020623 \pm (2.131)(0.00037895)}$$

$$\Rightarrow \quad (0.0198,\ 0.0214)$$

Inference for the Intercept (β_0)

- $b_o = \bar{Y} - b_1\bar{x} \sim N(\beta_o, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}))$

- b_o has standard error $S_{b_o} = \sqrt{MS_{error}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$

- Reject $H_o : \beta_o = 0$ if $|t| = \left|\frac{b_o - 0}{S_{b_o}}\right| > t_{n-2,1-\alpha/2}$

- $100(1-\alpha)\%$ confidence interval for β_o is

$$b_0 \pm t_{n-2,1-\alpha/2}\ S_{b_0}$$

Inference for the Intercept (β_0)

- Rarely considered
- Values of x must be near 0 for meaningful interpretations
- Would be most likely to use confidence interval

Analysis of the Forbes Data

- Test $H_o: \beta_0 = 0$ $(Y_i = \beta_1 x_i + \epsilon_i)$
  versus $H_a: \beta_0 \neq 0$ $(Y_i = \beta_0 + \beta_1 x_i + \epsilon_i)$

- Evaluate $t = \dfrac{b_0 - 0}{S_{b_0}} = \dfrac{-0.971 - 0}{0.0769} = -12.6$

- The least squares estimate of the intercept is 12.6 standard errors away from zero (p-value << .0001). Reject the null hypothesis and conclude that the intercept is negative. (No practical motivation)

- A 95% confidence interval for the intercept is

  $$b_0 \pm t_{15,.975} S_{b_0} \Rightarrow -0.971 \pm (2.131)(0.0769) \Rightarrow (-1.135, -0.807)$$

Inference for Conditional Means

Inference for $\mu_{Y|x} = E(Y|X = x) = \beta_o + \beta_1 x$

- Estimate is $\hat{\mu}_{Y|x} = b_o + b_1 x$
- $\hat{\mu}_{Y|x}$ is a linear function of two normally distributed random variables (b_0 and b_1, not independent)
- $\hat{\mu}_{Y|x}$ is $N\left(\beta_o + \beta_1 x, \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}\right)\right)$
- Note: value of x does not need to be present in sample.

Inference for Conditional Means

- standard error is

$$S_{\hat{\mu}_{Y|x}} = S_{b_o+b_1x} = \sqrt{MS_{error}\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}$$

- $100(1-\alpha)\%$ confidence interval for $\beta_o + \beta_1 x$ is

$$(b_o + b_1 x) \pm t_{n-2,1-\alpha/2}\ S_{\hat{\mu}_{Y|x}}$$

Confidence Region for a Line Segment

Use the Scheffe' procedure to get simultaneous confidence intervals for every x in an entire line segment:

$$(b_o + b_1 x) \pm \sqrt{2F_{2,n-2,1-\alpha}}\; S_{b_o+b_1 x}$$

for $a \leq x \leq b$

Prediction

Predict the value for Y at given x:

$$Y_{new} = \beta_o + \beta_1 x + \epsilon$$

- Estimate is still $\hat{Y} = b_o + b_1 x$
- Standard error is

$$S_{pred} = \sqrt{MS_{error}\left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

- $100(1 - \alpha)\%$ prediction interval:

$$(b_o + b_1 x) \pm t_{n-2,1-\alpha/2}\ S_{pred}$$

Comparison

- Confidence Interval for Condition Mean $\mu_{Y|x}$
  - Inference for a point on the population regression line given value of x
  - Source of inference is estimating regression line
- Prediction Interval for Y
  - Inference for a point in the scatterplot of all population values given value of x.
  - Sources of inference are estimating regression line AND predicting Y given the regression line.

Analysis of the Forbes Data

- Construct a 95% confidence interval for the mean of possible log-pressure measurements when the boiling point of water is x=209 ^oF

- Estimated mean is

$$\hat{\mu}_{Y|x} = b_0 + b_1 x = -0.9710 + (.0206)(209) = 3.339$$

- Evaluate the standard error of this estimate

$$S_{\hat{\mu}_{Y|x}} = \sqrt{.0000762 \left(\frac{1}{17} + \frac{(209 - 202.953)^2}{530.78} \right)} = 0.00312$$

- A 95% confidence interval is

$$\hat{\mu}_{Y|x} \pm t_{15,.975} S_{\hat{\mu}_{Y|x}} \Rightarrow 3.339 \pm (2.131)(0.00312) \Rightarrow (3.333, 3.346)$$

Analysis of the Forbes Data

- Apply the exponential function to the end points to get an *approximate* confidence interval for the mean pressure

  (28.02, 28.39) inches of Hg

- This could be computed with either the REG procedure or the GLM procedure in SAS by adding an additional line to the data file with X=209 and a missing value for Y

Analysis of the Forbes Data

Scheffe procedure for constructing a 95% confidence region for a segment of the true regression line

Evaluate $(\boldsymbol{b_0 + b_1 x}) \pm \sqrt{\boldsymbol{2F_{(2,n-2),1-\alpha}}} \boldsymbol{S_{b_0+b_1x}}$

$$\Rightarrow (\boldsymbol{b_0 + b_1 x}) \pm \sqrt{\boldsymbol{2F_{(2,15),0.95}}} \boldsymbol{S_{b_0+b_1x}}$$

$$\Rightarrow (\boldsymbol{b_0 + b_1 x}) \pm (\mathbf{2.713}) \sqrt{\mathbf{.0000762} \left(\frac{\mathbf{1}}{\mathbf{17}} + \frac{(\boldsymbol{x} - \mathbf{202.953})^{\mathbf{2}}}{\mathbf{530.78}} \right)}$$

Analysis of the Forbes Data

- Construct a 95% prediction interval for a log-pressure value when the boiling point of water is x=209 ^{o}F

- Prediction is the estimated mean

  $$\hat{Y} = b_0 + b_1 x + error = -0.9710 + (.0206)(209) + 0 = 3.339$$

- Evaluate the standard error of the prediction (include the variation of the associated random error, estimated as $MS_{error} = .0000762$)

  $$S_{pred} = \sqrt{.0000762\left(1 + \frac{1}{17} + \frac{(209 - 202.953)^2}{530.78}\right)} = 0.00927$$

Analysis of the Forbes Data

- A 95% prediction interval is

$$\hat{y} \pm t_{15,.975} S_{pred} \quad \Rightarrow \quad \mathbf{3.339 \pm (2.131)(0.00927)}$$

$$\Rightarrow \quad \mathbf{(3.319, 3.359)}$$

- Apply the exponential function to the end points to get an *approximate* prediction interval for barometric pressure:
  (27.63, 28.76) inches of Hg

- This could be computed with either the REG or GLM procedure in SAS by adding an additional line to the data file with X=209 and a missing value for Y