Equations for Univariate Linear Regression

Basic computations:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{X})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{Y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

SUMMARY OUTPUT

Regression Statistics

Multiple R	$r = \sqrt{r^2}$
R Square	$r^2 = \frac{SSR}{SST} = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$
Adjusted R Square	*Used in Multiple Regression*
Standard Error	$s = \sqrt{\frac{SSE}{n-2}}$
Observations	n

ANOVA

	df	***SS***	***MS***	***F***	***Significance F***
Regression	1	$SSR = \hat{\beta}_1^2 S_{xx}$	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$	From F table with *1* and *n-2* d.f.
Residual	$n-2$	$SSE = SST - SSR$	$MSE = \frac{SSE}{n-2}$		
Total	$n-1$	$SST = S_{yy}$			

	Coefficients	***Standard Error***	***t Stat***	***P-value***
Intercept	$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$	$SE(\hat{\beta}_0) = s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n S_{xx}}}$	$t = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)}$	From *t* distn. With *n-2 d.f.*
Slope	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$	$SE(\hat{\beta}_1) = \frac{s}{\sqrt{S_{xx}}}$	$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$	From *t* distn. With *n-2 d.f*

A 100(1-α)% **confidence interval** for μ^*, the true mean at a value x^* is given by

$$\hat{\mu}^* - t_{n-2,\alpha/2} s \sqrt{\frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{S_{xx}}} \le \mu^* \le \hat{\mu}^* + t_{n-2,\alpha/2} s \sqrt{\frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{S_{xx}}}$$

where $\hat{\mu}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$

A 100(1-α)% **prediction interval** for Y^*, a single observation taken at a value x^* is given by

$$\hat{Y}^* - t_{n-2,\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{S_{xx}}} \le Y^* \le \hat{Y}^* + t_{n-2,\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{S_{xx}}}$$

where $\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$

Example:

Football Example (data to be posted on BLACKBOARD)

Basic computations:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{44313.2}{117} = 378.745 \qquad \bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{3150.6}{117} = 26.928$$

$$S_{xx} = \sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 17180049.34 - \frac{1}{117}44313.2^2 = 396633.15$$

$$S_{yy} = \sum_{i=1}^{n}\left(y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 90093.8 - \frac{1}{117}3150.6^2 = 5253.8$$

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n}\left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right) \\ &= \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \\ &= 1232567.56 - \frac{1}{117}44313.2 \times 3150.6 \\ &= 39292.62 \end{aligned}$$

Regression coefficients:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{39292.62}{396633.15} = 0.099065$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 26.928 - 0.099065 \times 378.745 = -10.5923$$

***R^2* computation:**

$$r^2 = \frac{SSR}{SST} = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{39292.62^2}{396633.15 \times 5253.8} = 0.7409$$

$$r = \sqrt{r^2} = \sqrt{0.740899721} = 0.860755$$

ANOVA computations:

$$SST = S_{yy} = 5253.8$$

$$SSR = \hat{\beta}_1^2 S_{xx} = \left(0.0990654\right)^2 \times 396633.15 = 3892.539$$

$$SSE = SST - SSR = 5253.8 - 3892.539 = 1361.261$$

$$MSR = \frac{SSR}{1} = \frac{3892.539}{1} = 3892.539$$

$$MSE = \frac{SSE}{n-2} = \frac{1361.261}{115} = 11.83705$$

$$F = \frac{MSR}{MSE} = \frac{3892.539}{11.83705} = 328.84$$

Significance F = FDIST(328.84,1,115)=1.617E-35

Standard error of regression:

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1361.261}{115}} = 3.4405$$

Significance of regression coefficients:

$$SE\left(\hat{\beta}_0\right) = s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}} = 3.4405\sqrt{\frac{17180049.34}{117 \times 396633.15}} = 2.09337$$

$$t = \frac{\hat{\beta}_0}{SE\left(\hat{\beta}_0\right)} = \frac{-10.59235}{2.09337} = -5.0599 \quad \text{for } \beta_0$$

P-value for β_0 = TDIST(-5.0599,115,2)=1.6054E-06

$$SE\left(\hat{\beta}_1\right) = \frac{s}{\sqrt{S_{xx}}} = \frac{3.4405}{\sqrt{396633.15}} = 0.0054629$$

$$t = \frac{\hat{\beta}_1}{SE\left(\hat{\beta}_1\right)} = \frac{0.0990654}{0.0054629} = 18.1341 \quad \textit{for } \beta_1$$

P-value for β_1 = TDIST(18.1341,115,2)=1.6175E-35

95% CI for β_0 =

$$\hat{\beta}_0 - t_{n-2,0.025} SE\left(\hat{\beta}_0\right) \le \beta_0 \le \hat{\beta}_0 + t_{n-2,0.025} SE\left(\hat{\beta}_0\right)$$

$$1.1563636 - 2.306 \times 0.074028 \le \beta_0 \le 1.1563636 + 2.306 \times 0.074028$$

$$0.98565 \le \beta_0 \le 1.32707$$

95% CI for β_1 =

$$\hat{\beta}_1 - t_{n-2,0.025} SE\left(\hat{\beta}_1\right) \le \beta_1 \le \hat{\beta}_1 + t_{n-2,0.025} SE\left(\hat{\beta}_1\right)$$

$$-10.5923 - 1.981 \times 0.0054629 \le \beta_0 \le -10.5923 + 1.981 \times 0.0054629$$

$$-14.739 \le \beta_0 \le -6.446$$

Example: 95% <u>Confidence</u> interval for μ* at POINTS=320 (NU's value)

$$\hat{\mu}^* = \hat{\beta}_0 + \hat{\beta}_1 x^* = -10.5923 + 0.0990654 \times 320 = 21.109$$

$$\hat{\mu}^* - t_{n-2,\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{S_{xx}}} \le \mu^* \le \hat{\mu}^* + t_{n-2,\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{S_{xx}}}$$

$$21.109 - 1.981 \times 3.4405\sqrt{\frac{1}{117} + \frac{\left(320 - 378.745\right)^2}{396633.15}} \le \mu^* \le 21.109 - 1.981 \times 3.4405\sqrt{\frac{1}{117} + \frac{\left(320 - 378.745\right)^2}{396633.15}}$$

$$20.21356 \le \mu^* \le 22.00359$$

Example: 95% <u>Prediction</u> interval for Y* at POINTS=320

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^* = -10.5923 + 0.0990654 \times 320 = 21.109$$

$$\hat{Y}^* - t_{n-2,\alpha/2}\, s\sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{S_{xx}}} \le Y^* \le \hat{Y}^* + t_{n-2,\alpha/2}\, s\sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{S_{xx}}}$$

$$21.109 - 1.981 \times 3.4405\sqrt{1 + \frac{1}{117} + \frac{\left(320 - 378.745\right)^2}{396633.15}} \le Y^* \le 21.109 - 1.981 \times 3.4405\sqrt{1 + \frac{1}{117} + \frac{\left(320 - 378.745\right)^2}{396633.15}}$$

$$14.235 \le Y^* \le 27.982$$

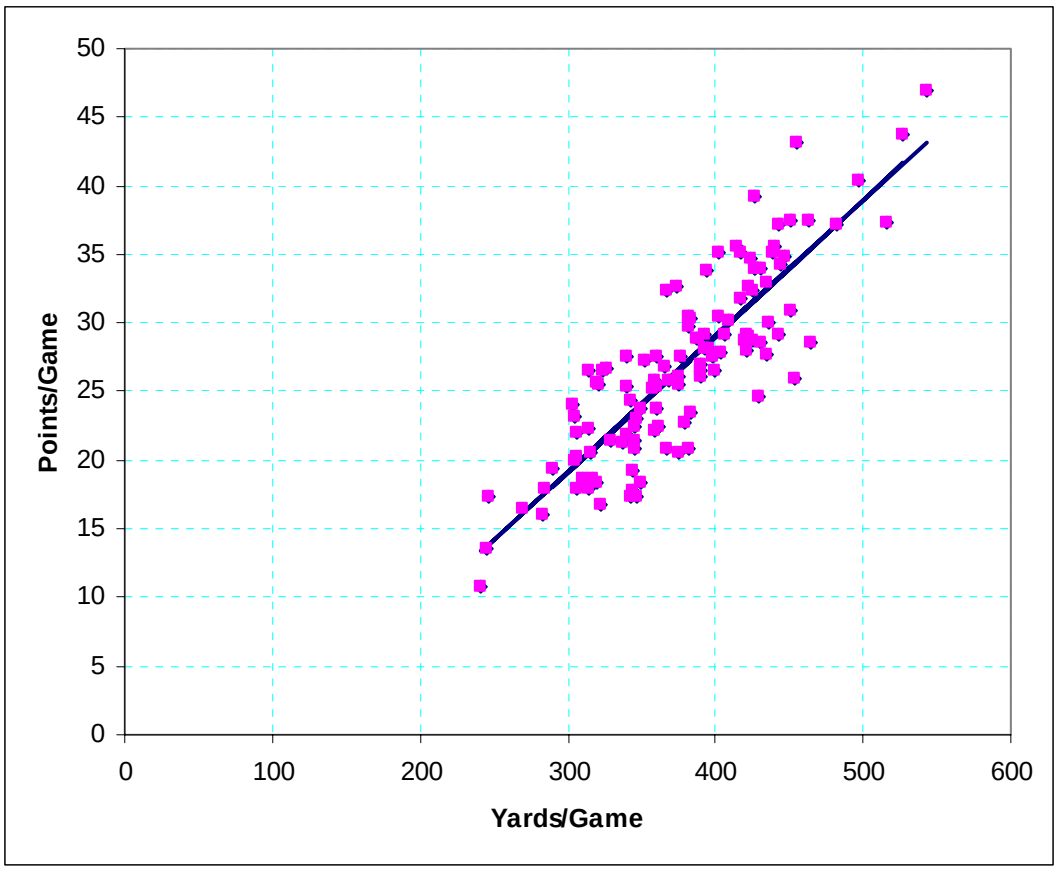
Here is the EXCEL OUTPUT

Regression Statistics	
Multiple R	0.8608
R Square	0.7409
Adjusted R Square	0.7386
Standard Error	3.4405
Observations	117

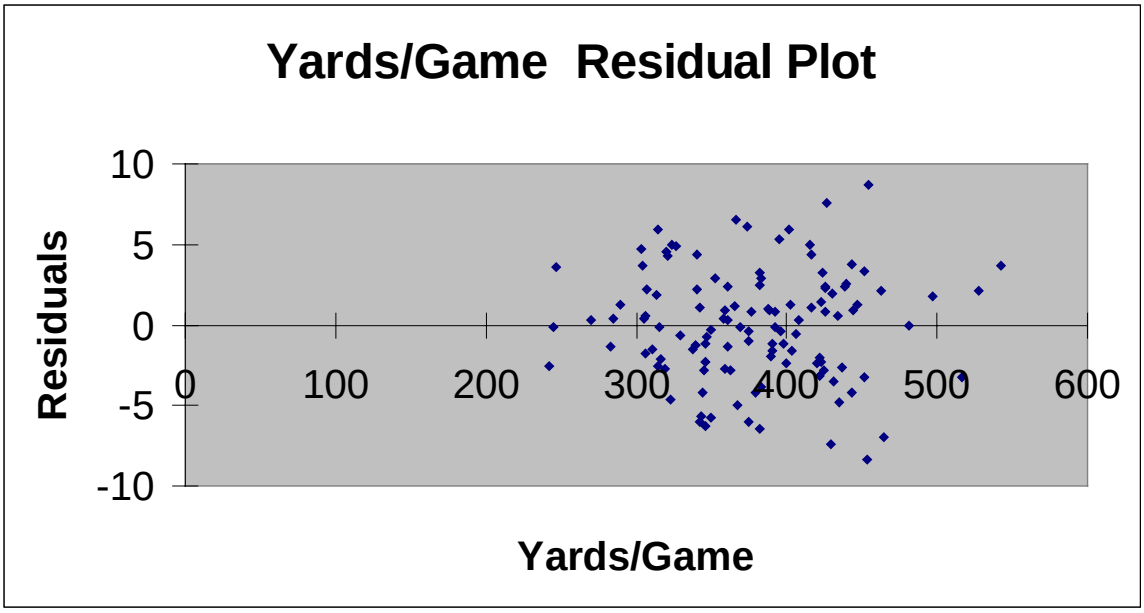
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3,892.54	3,892.54	328.84	1.61748E-35
Residual	115	1,361.26	11.84		
Total	116	5,253.80			

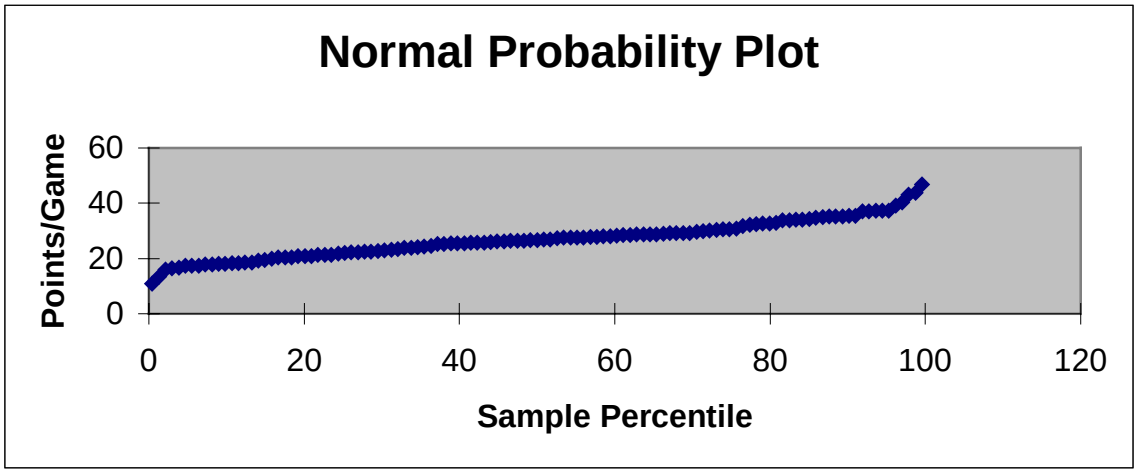
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-10.59234834	2.0934	-5.0599	1.6054E-06	-14.73891018	-6.445786513
Yards/Game	0.09906540	0.0055	18.1341	1.6175E-35	0.08824436	0.109886434



There is clearly a good fit between the data and the estimated line.



The data exhibit a **slight** degree of dependence of the errors on the yards per game. This is known as **heteroskedasticity**. The impact of this is that the estimators of β_0 and β_1 are still unbiased estimators but they are not the minimum variance estimators in the class of linear unbiased estimators. Dealing with heteroskedasticity is beyond the scope of this course and is a topic dealt with in an intermediate econometrics class. See (Kmenta, J., 1971, Elements of Econometrics, MacMillan Press, New York, section 8.1).



The errors seem to be approximately Normally distributed from this plot. There is some evidence that the errors come from a distribution with a light tail compared to the Normal distribution. (see Tamhane and Dunlop Figure 4.10(d)).