# 36-401 Class Notes - September 13, 2001

## Chapter 2: Bivariate Regression Analysis

- Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

  What do $\beta_0, \beta_1$ represent?

  What are the assumptions on $\epsilon_i,\ i = 1, 2, \ldots, n$?

  Estimate/Prediction: $\hat{Y}_i = b_0 + b_1 X_i$
  $\hat{Y}^* = b_0 + b_1 X^*$

  What is the difference between $\beta_0, \beta_1$ and $b_0, b_1$?

  Residual error: $e_i = Y_i - \hat{Y}_i$

  What does the residual measure?

- Estimation method: "Least Squares"

$$
\begin{aligned}
\text{RSS} &= \sum_{i=1}^{n} e_i^2 \\
&= \sum \left(Y_i - \hat{Y}_i\right)^2 \\
&= \sum \left(Y_i - b_0 - b_1 X_i\right)^2
\end{aligned}
$$

Minimization of RSS gives

$$\begin{aligned} b_1 &= \frac{\sum\left[(X_i - \bar{X})(Y_i - \bar{Y})\right]}{\sum(X_i - \bar{X})^2} = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x} \\ b_0 &= \bar{Y} - b_1\bar{X} \end{aligned}$$

What do the following variables represent/define? How are they determined?

- $s_x$? $s_y$?
- $s_{xy}$?
- $r$?

Estimate of $\sigma^2$: $s_e^2 = \frac{\text{RSS}}{n-K}$, where $n$ = sample size and $K$ = number of parameters estimated

What is K for a linear regression problem?

$s_e$ is "residual standard error"

What does the residual standard error tell us?

- Coefficient of Determination: $\text{R}^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{s^2_{\hat{Y}}}{s^2_Y}$

  Note: RSS+ESS=TSS if $\hat{Y}$ is based on LS estimates of $b_0$ and $b_1$.

  What does $R^2$ provide?

$0 \leq R^2 \leq 1$.
What is the interpretation when $R^2 = 0$? when $R^2 = 1$?

- Sample Covariance of X and Y

$$s_{xy} = \frac{\sum\left[(X_i - \bar{X})(Y_i - \bar{Y})\right]}{n-1}$$

Sample Correlation of X and Y

$$r = \frac{s_{xy}}{\sqrt{s_x^2}\sqrt{s_y^2}} \text{ or } \frac{s_{xy}}{s_x s_y}$$

$-1 \leq r \leq 1$, where $0 \leq\mid r \mid\leq 1$ tells the strength of the linear trend and $\text{sgn}(r)$ tells if the trend is positive or negative

What interpretations can we make from $r$? What if $r = 1$? if $r = -1$?

What is the relationship between $R^2$ and $r^2$ for linear regression models?

- Sampling Distributions and Standard Errors

What assumptions need to be made regarding the errors to result in normal sampling distributions for $b_0, b_1$? What are their associated parameters?

Estimated standard error of $b_1$:

$$\text{SE}_{b_1} = \frac{s_e}{\sqrt{\text{TSS}_x}}$$

Estimated standard error of $b_0$:

$$\text{SE}_{b_0} = s_e\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\text{TSS}_x}}$$

Estimated standard error of prediction of $Y^* = b_0 + b_1 X^*$:

$$\text{SE}_{Y^*} = s_e\sqrt{\frac{1}{n} + \frac{(X^* - \bar{X})^2}{\text{TSS}_x} + 1}$$

(drop +1 for estimated standard error of mean of Y for a given X)

How do we interpret the standard errors?

# Breakout Session - September 13, 2001

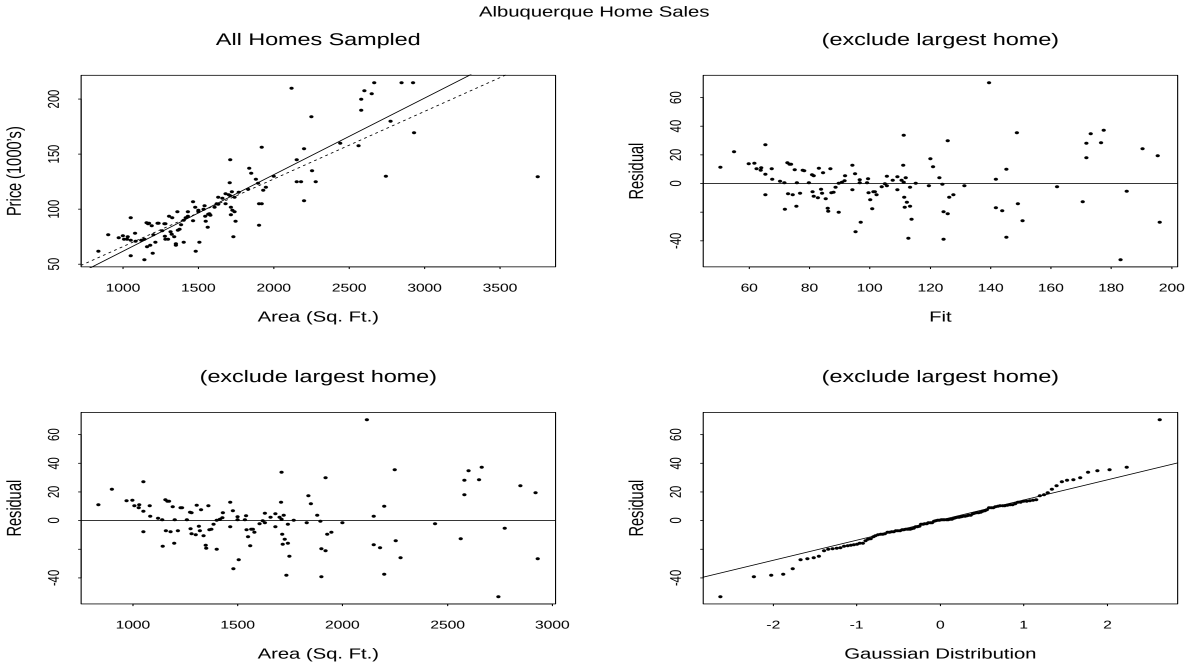
The data for this exercise come from a file maintained by the multiple listing real estate agency in Albuquerque, NM. The file contains prices of home resales along with descriptive data on the home, including square feet, age, number of certain features, custom built or not, and corner location or not. Realtors can use this kind of data to find an expected selling price for a home with given characteristics. For now we are just looking at how well the size of the house (in square feet of usable floor space) can be used to predict the price of the house.

```
> homedat_read.table("homedat.dat",T)
> dim(homedat)      # [1] 117    8
> names(homedat) # "PRICE" "SQFT"  "AGE"   "FEATS" "NE"    "CUST"  "COR"   "TAX"
> homedat$PRICE_homedat$PRICE/10  # go from hundreds to thousands
> par(mfrow=c(2,2), oma=c(0,0,2,0))
> plot(homedat$SQFT, homedat$PRICE, xlab="Area (Sq. Ft.)", ylab="Price")
> abline(lm(PRICE~SQFT, data=homedat), lty=2)  # lty=2 makes dashed line
> homelm_lm(PRICE~SQFT, homedat, subset=SQFT<3500)
> abline(homelm)
> mtext("Albuquerque Home Sales", outer=2, cex=1.3)
> summary(homelm)
# Call: lm(formula = PRICE ~ SQFT, data = homedat, subset = SQFT < 3500)
# Residuals:
#     Min     1Q Median    3Q   Max
#  -53.13 -8.979 0.4996 9.936 70.55
#
# Coefficients:
#               Value Std. Error  t value Pr(>|t|)
# (Intercept)  -7.6208   5.7177    -1.3329   0.1852
#        SQFT   0.0695   0.0034    20.7433   0.0000
#
# Residual standard error: 17.53 on 114 degrees of freedom
# Multiple R-Squared: 0.7906
> mean(homedat$SQFT[homedat$SQFT<3500])   #  [1] 1635.784
> var(homedat$SQFT[homedat$SQFT<3500])    #  [1] 238134.1

> plot(homelm$fit, homelm$res, xlab="Fit", ylab="Residual",
>      main="(exclude largest home)")
> abline(h=0)

> plot(homedat$SQFT[homedat$SQFT<3500], homelm$res, xlab="Area (Sq. Ft.)",
>      ylab="Residual", main="(exclude largest home)")
> abline(h=0)

> qqn(homelm$res, main="(exclude largest home)", ylab="Residual")
> ps("AlbHome")
```




Based on the S-plus code and the first four plots, what model is being fit?

What is included in the error term in this example?

What assumptions appear to be violated?

Using the simple linear model, what is the expected home price for a 1500 square foot home resale in Albuquerque? Write out the standard errors for a) $\beta_0$, b) $\beta_1$, c) the expected price for a home with 1500 square feet, and d) the price of a random individual home with 1500 square feet. Use numbers, not symbols!

Verify that the results match the printout. Interpret the standard errors.