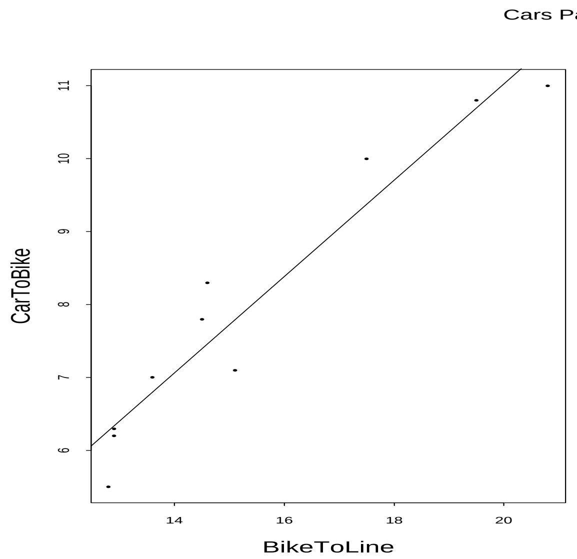
36-401 Class Notes and Exercise for 9/7/00

Howard Seltman

Simple Linear Regression




```
> bikelm_lm(cyc$Car~cyc$Center)  or
> bikelm_lm(Car~Center, data=cyc)
> summary(bikelm)
Call: lm(formula = cyc$Car ~ cyc$Center)
Residuals:
     Min      1Q  Median     3Q    Max
 -0.7699 -0.4485 0.03493 0.3561 0.8415

Coefficients:
                 Value  Std. Error  t value  Pr(>|t|)
(Intercept)   -2.1825  1.0567    -2.0654  0.0727
 cyc$Center  0.6603  0.0675    9.7858  0.0000

Residual standard error: 0.5821 on 8 df
Multiple R-Squared: 0.9229
F-statistic: 95.76 on 1 and 8 df
the p-value is 9.975e-06
```

- Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2) \quad \epsilon_i \perp\!\!\!\perp \epsilon_j$
  Design or covariate space
  Estimate/Prediction: $\hat{Y}_i = b_0 + b_1 X_i \quad \hat{Y}^* = b_0 + b_1 X^*$
  Residual error: $e_i = Y_i - \hat{Y}_i$

- Estimation method: "Least Squares"

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

  Meaning of "SS"?
  Minimization of RSS gives

$$b_1 = \frac{\sum \left[(X_i - \bar{X})(Y_i - \bar{Y})\right]}{\sum (X_i - \bar{X})^2} \qquad b_0 = \bar{Y} - b_1 \bar{X}$$

  Estimate of σ^2: $s_e^2 = \frac{\mathrm{RSS}}{n-K}$ $\quad s_e$ is "residual standard error"

- Coefficient of Determination: $\mathrm{R}^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$
  Note: RSS+ESS=TSS if $\hat{Y}$ is based on LS estimates of b_0 and b_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

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

- Sampling Distributions and Standard Errors
  Assumptions on ϵ: normal, independent, same σ^2

  Estimated standard error of b_1:

$$\mathrm{SE}_{b_1} = s_e\sqrt{\frac{1}{\mathrm{TSS}_x}}$$

  Estimated standard error of b_0:

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

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

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

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

Breakout Session for 9/7/00

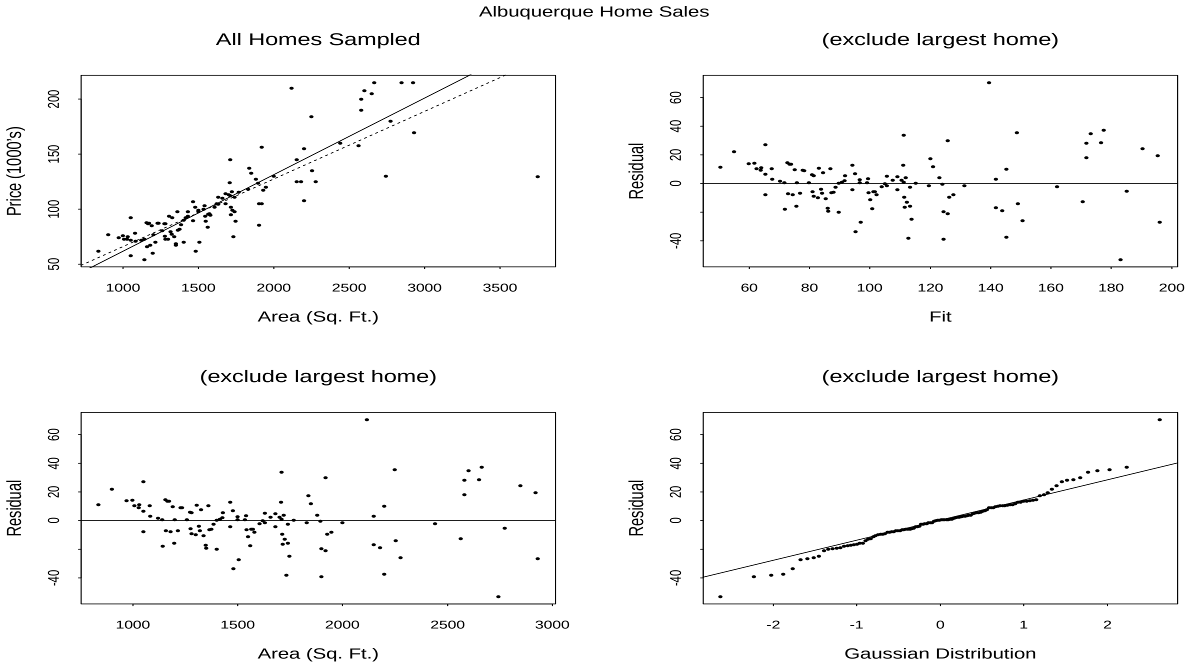
The data for this exercise come from a file maintained by the multiple listing real estate agency in Albuquerque, NM. The file contains prices of home resales along with descriptive data on the home, including square feet, age, number of certain features, custom built or not, and corner location or not. 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 are appropriate null and alternative hypotheses?

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) β_0, b) β_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.