Linear Regression

The term "regression" describes a class of models for studying the relationship between a response variable Y and covariates (also called explanatory variables or regressors) $X^{(1)}, \ldots, X^{(p)}$.

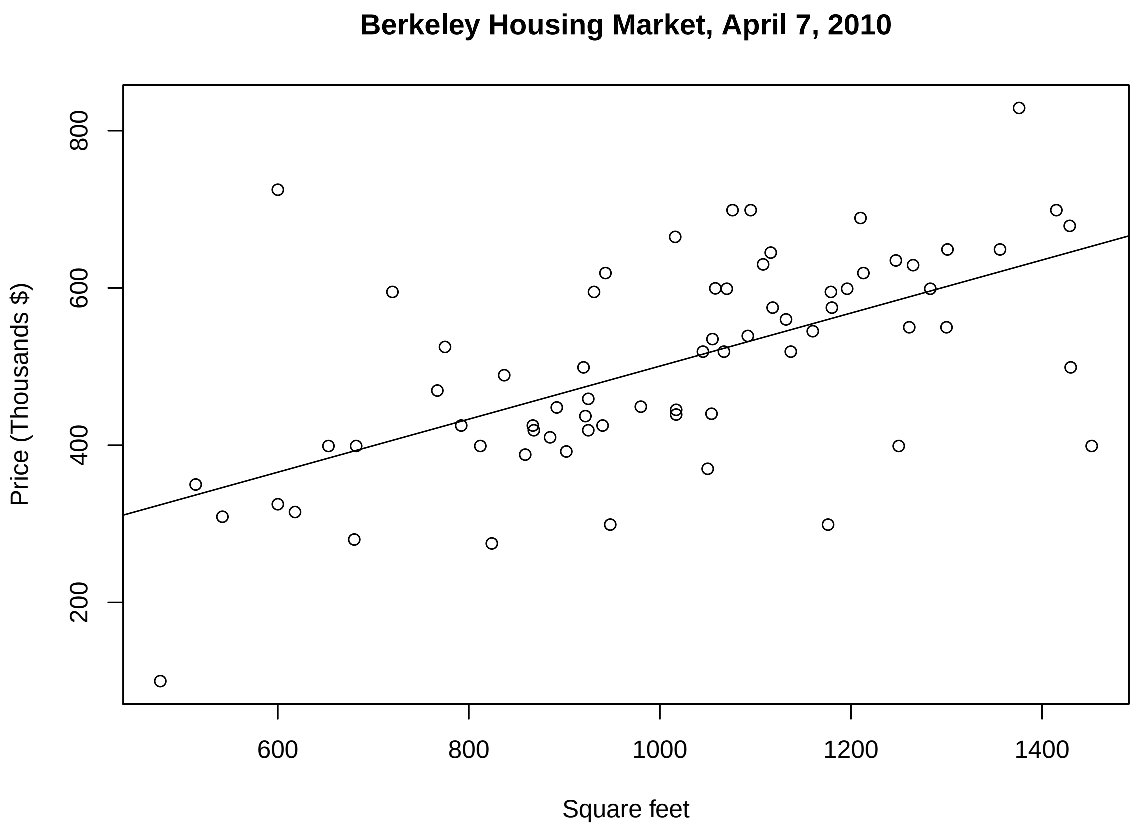
The assumption of linearity is less restrictive than it might seem, since the $X's$ can consist of nonlinear transformations of other variables of interest.

We'll start with the simple linear regression model, which means $p = 1$ and

$$\begin{aligned} E[Y|X = x] &= \beta_0 + \beta_1 x \\ V[Y|X = x] &= \sigma^2 \end{aligned}$$

We're not (yet) assuming anything else about $p(Y|X)$.

We observe pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, and based on this we estimate β_0, β_1, and σ^2. For example, here is some data from www.zillow.com:



The model for an individual observation is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $E[\epsilon_i] = 0$ and $V[\epsilon_i] = \sigma^2$.

The fitted regression line is $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, and the fitted values are $\hat{Y}_i = \hat{r}(X_i)$. The residuals are

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

A classical way of estimating β_0 and β_1 is by minimizing the residual sum of squares

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

The least squares estimates are

$$\begin{aligned}
\hat\beta_1 &= \frac{\sum_{i=1}^n (X_i - \bar X)(Y_i - \bar Y)}{\sum_{i=1}^n (X_i - \bar X)^2} \\
\hat\beta_0 &= \bar Y - \hat\beta_1 \bar X
\end{aligned}$$

Once we have $\hat\beta_0$ and $\hat\beta_1$, we may form an unbiased estimator of σ^2 via

$$\hat\sigma^2 = \frac{1}{n-2}\sum_{i=1}^n \hat\epsilon_i^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat\beta_0 - \hat\beta_1 X_i)^2$$

In the housing example, $\hat\beta_0 = 163.3$ and $\hat\beta_1 = 0.337$. We may interpret $\hat\beta_1$ to mean that for every additional square foot, the average price increases by $337.

Now add the assumption that $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$. Equivalently, $Y_i | X_i \sim N(\mu_i, \sigma^2)$, where $\mu_i = \beta_0 + \beta_1 X_i$.

Conditioning on X (treating X as random is known as "errors in variables" and is beyond the scope of this course), we have a likelihood

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \right\}$$

The MLEs for β_0 and β_1 are the same as the least squares estimates. The MLE for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Some basic properties of β_0 and β_1:

1. They are unbiased: $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$.

2. They are consistent: $\hat{\beta}_0 \xrightarrow{P} \beta_0$ and $\hat{\beta}_1 \xrightarrow{P} \beta_1$.

3. They are asymptotically normal:

$$\frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} \xrightarrow{D} N(0,1) \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \xrightarrow{D} N(0,1)$$

The variances are

$$
\begin{aligned}
V[\hat{\beta}_0] &= \frac{\sigma^2}{n}\frac{\sum_{i=1}^n X_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\\
V[\hat{\beta}_1] &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\\
Cov[\hat{\beta}_0, \hat{\beta}_1] &= -\frac{\sigma^2 \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}
\end{aligned}
$$

To estimate standard errors, we plug in $\hat{\sigma}^2$ (either unbiased or MLE) for σ^2. This allows us to construct confidence intervals and carry out tests.

Usually the test we're interested in is for $H_0 : \beta_1 = 0$. For this we can construct a Wald test using $W = \hat{\beta}_1 / \widehat{se}(\hat{\beta}_1)$.

In R, much of this calculation can be carried out using the `lm` function.

```
> linmod <- lm(price~sqft, data = berkhousing)
> summary(linmod)

Call:
lm(formula = price ~ sqft, data = berkhousing)

Residuals:
     Min       1Q   Median       3Q      Max
-260.983  -51.817    3.214   46.845  359.347

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 163.22699   57.16475   2.855  0.00572 **
sqft          0.33738    0.05517   6.115  5.6e-08 ***

... (more stuff)
```

Plotting the points and adding the fitted line:

```
plot(berkhousing$sqft, berkhousing$price,
     xlab = "Square feet", ylab = "Price (Thousands $)")
abline(linmod) # Add the fitted line to the plot
```

Here is some code to compute the p-value for the Wald test for $\beta_1 = 0$. In this case it is very small, as was the p-value for the t-test that `lm` computed.

```
> beta1 <- linmod$coefficients[2]
> se.beta1 <- summary(linmod)$coefficients[2,2]
> W <- beta1/se.beta1
> 2*pnorm(-abs(W))
        sqft
9.670514e-10
```

`linmod` and `summary(linmod)` are lists, but they print in special ways. To see what's inside the list, use `names(linmod)` and `names(summary(linmod))`.

What if we want to predict Y from X? We need to be careful what we mean by this: are we talking about

- the fit $\hat{r}(x_*) = \widehat{E}[Y|X = x_*]$? This is the mean of a distribution.

- a new observation Y_* for $X = x_*$? This is a sample from a distribution.

Our confidence intervals are different, depending on which one we want.

```
> predict(linmod, newdata = data.frame(sqft = 1000), interval = "confidence")
       fit      lwr      upr
1 500.6042 474.5419 526.6664
> predict(linmod, newdata = data.frame(sqft = 1000), interval = "prediction")
       fit     lwr      upr
1 500.6042 282.698 718.5103
```

The coefficient of variation, usually just called R^2, is the ratio of "explained" sum of squares to total sums of squares:

$$
\begin{aligned}
R^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= 1 - \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}
\end{aligned}
$$

R^2 ranges from zero (no variance explained) to one (all variance explained – a perfect fit).