Linear Regression

Guy Lebanon

Linear Regression Model and Least Squares Estimation

Linear regression is probably the most popular model for predicting a RV $Y \in \mathbb{R}$ based on multiple RVs $X_1, \ldots, X_d \in \mathbb{R}$. It predicts a numeric variable using a linear combination of variables $\sum \theta_i X_i$ where the combination coefficients θ_i are determined by minimizing the sum of squared prediction error on the training set. We use below the convention that the first variable is always one i.e., $X_1 \equiv 1$ in order to facilitate having a constant term in the linear combination and being able to write it in matrix form

$$\hat{Y} = \theta_1 + \sum_{i=2}^{d} \theta_i X_i = \sum_{i=1}^{d} \theta_i X_i = \theta^\top X.$$

A more mathematical view of linear regression is that it is a probabilistic model for Y given X that assumes

$$Y = \theta^\top X + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \qquad \text{or equivalently} \tag{1}$$

$$Y|X \sim N(\theta^\top X, \sigma^2). \tag{2}$$

In other words, linear regression assumes that given X, Y is normally distributed with mean that is linearly increasing in X and constant variance. In particular, no assumption is made on the distribution $p(X)$.

The variable Y and the variables X_i usually represent numeric quantities e.g., weight, age, salary, ozone measurement, temprature, etc. However, in some cases X_i may represent a binary categorical variable $X_i \in \{0, 1\}$ e.g., gender, sickness etc. More general categorical variables that take values in a finite unordered set $\{1, \ldots, c\}$ e.g., color or race are converted to $c - 1$ binary variables that are turned on with value 1 if the variable matches the corresponding value and 0 otherwise. The choice of the last c value is indicated by all $c - 1$ variables being zero.

It is important to note that $X_1, \ldots, X_d$ may be non-linear functions of some original data $X'_1, \ldots, X'_r$ e.g., $X_1 = X'_1, X_2 = (X'_2)^2, X_3 = X'_1 X'_2$, $X_4 = \exp(X_2) \log X_1$, etc. The resulting linear regression model $\theta^\top X$ is linear in X but non-linear in the original data X'. This gives linear regression substantial flexibility for modeling cases where Y and X' are non-linearly related. All that is needed is a nonlinear trasnformation $X' \mapsto X$ for which a linear relationship exists between Y and X.

The training data in regression analysis is usually multiple iid samples $(X^{(i)}, Y^{(i)}) \overset{\text{iid}}{\sim} p(X, Y) = p(X)p(Y|X), i = 1, \ldots, n$ where $p(|X)$ is defined by (1) and $p(X)$ is an arbitrary marginal. In the case of observational data $p(X)$ corresponds to nature and in the case of experimental data $p(X)$ corresponds to the experimental design. We represent the conditional relationship between the training data $(X^{(i)}, Y^{(i)}), i = 1, \ldots, n$ in matrix form $\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\epsilon}$ where $\mathbf{Y} = (Y^{(1)}, \ldots, Y^{(n)}) \in \mathbb{R}^{n \times 1}$, $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a matrix whose rows are $X^{(i)}$ $i = 1, \ldots, n$, and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$ ($\boldsymbol{\epsilon}$ is the vector of noise values $\epsilon_i = Y^{(i)} - \theta^\top X^{(i)} \sim N(0, \sigma^2)$, $i = 1, \ldots, n$ corresponding to the training data and is therefore a multivariate normal vector). The matrix $\mathbf{X}$ and the vectors $\mathbf{Y}, \boldsymbol{\epsilon}$ correspond to the n training set instances and are denoted in bold face to avoid confusion with the random variables X, Y, ϵ. Thus

$$\mathbf{Y}|\mathbf{X} \sim N(\mathbf{X}\theta, \sigma^2 I).$$

The standard way of obtaining the parameter vector θ is by minimizing the sum of square deviations (also known as residual sum of squares or RSS) of the observations from the model predictions

$$\hat{\theta} = \arg\min_{\theta} \text{RSS}(\theta) \quad \text{where} \quad \text{RSS}(\theta) = \|\mathbf{Y} - \mathbf{X}\theta\|^2 = \sum_{i=1}^{n} (Y^{(i)} - \theta^\top X^{(i)})^2$$

which is equivalent to the maximum conditional likelihood estimator $\hat{\theta} = \arg\max_\theta \sum_i \log p(Y^{(i)}|X^{(i)})$. The minimization above is that of a quadratic function and has a closed form expression, derived below. Differentiating the RSS criteria with respect to θ and setting it to 0 gives the set of normal equations

$$\nabla \text{RSS}(\theta) = 0 \quad \Leftrightarrow \quad \sum_i (Y^{(i)} - \theta^\top X^{(i)}) X_j^{(i)} = 0 \quad \forall j \quad \text{or} \quad \mathbf{X}^\top \mathbf{X}\theta = \mathbf{X}^\top \mathbf{Y}$$

$$\Rightarrow \quad \hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

The predictions on the training data $\mathbf{X}$ made by the model $\hat{\theta}$ is

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\theta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \mathbf{H}\mathbf{Y} \quad \text{where} \quad \mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top. \tag{3}$$

In the special case when the columns $u_1, \ldots, u_n$ of $\mathbf{X}$ are orthogonal the components of the least squares projection $\hat{\theta}$ become the standard orthogonal basis projections $\hat{\theta}_j = \langle u_j, \mathbf{Y}\rangle / \|u_j\|^2$.

We conclude with introducing the coefficient of determination also known as R^2 which together with $\text{RSS}(\hat{\theta})$ measures the model quality. It is defined as the square of the sample correlation coefficient between the training values $Y^{(i)}$ and the fitted values $\hat{Y}^{(i)} = \hat{\theta}^\top X^{(i)}$

$$R^2 = (\text{Sample-Correlation}(Y, \hat{Y}))^2 = \frac{\left(\sum_{i=1}^n (Y^{(i)} - \bar{Y})(\hat{Y}^{(i)} - \bar{\hat{Y}})\right)^2}{\sum_{i=1}^n (Y^{(i)} - \bar{Y})^2 \sum_{i=1}^n (\hat{Y}^{(i)} - \bar{\hat{Y}})^2}$$

It ranges between 0 and 1 where 1 indicates perfect linear fit with no noise $\sigma \to 0$.

Linear Regression in R

R has a very convenient linear regression function called `lm`. Its first argument is a linear model formula and its second argument is the data frame on which the training takes place. It returns an object which can be queried by looking at its variables or using functions that operate it like `predict` which predicts new Y values using the estimated linear model.

The formula argument desribes the variables X_i. The variable to the left of the tilde is the response and the variables to the right of the tilde are the explanatory X_i variables. Both can be transformations of variables of the data frame and the constant term is included by default. For example `log(x1) log(x2)` models $log x_1 = \theta_1 + \theta_2 \log x_2 + \epsilon$. Multiple explanatory variables are denoted with plus signs. A product operator corresponds all possible variable products, potentially with the constant term. R detects automatically categorical variables and expands them to one or more binary variables as described in the previous section. More details on formulas may be found in the examples below or in the R documention.

Below, we model diamond price as a linear function of diamond carat (plus a constant term). Note the formula in the `lm` function and the access to the least squares parameters using the `coef` function applied to the object returned from the linear model.

```
> diamSmall=diamonds[sample(1:dim(diamonds)[1],500),]; # reduce dataset size
> M1=lm(price~carat,diamSmall); # price=theta_1+theta_2 * carat + epsilon
> theta=coef(M1); theta # least squares parameter estimates \hat\theta
```

```
(Intercept)       carat
  -2174.092    7658.815
```

```
> # scatter plot of price vs carat overlayed with estimated regression line
> print(ggplot(diamSmall,aes(carat,price))+geom_point()+
+      geom_abline(intercept=theta[1],slope=theta[2],size=2,color=I("red")))
> predict(M1,data.frame(carat=c(3,4,5))) # model prediction for new X values
```

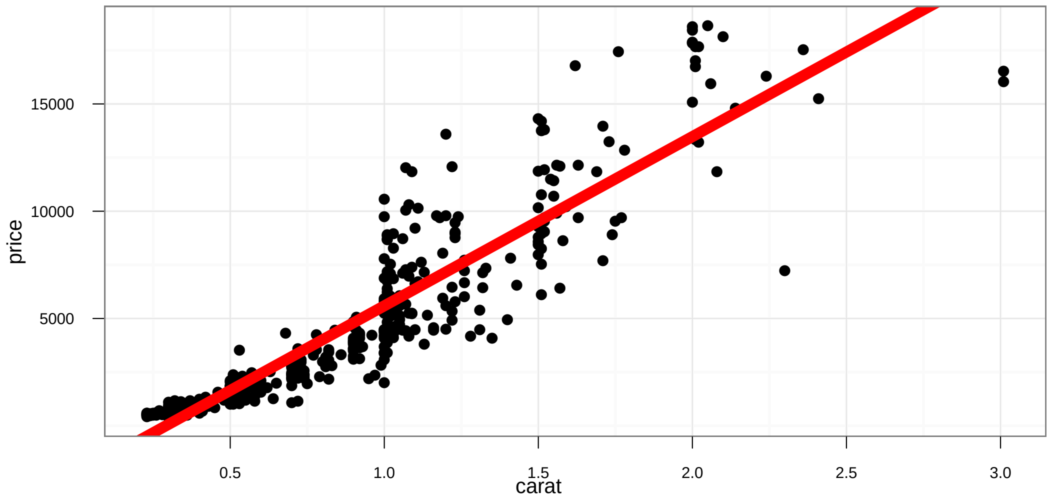
```
       1        2        3
20802.35 28461.17 36119.98
```

```
> summary(M1)$r.squared # R∧2 coefficient of determination
```

```
[1] 0.8272807
```

```
> mean(residuals(M1)∧2) # RSS (sum of squared residuals) normalized by n
```

```
[1] 3324694
```



As we see from the scatter plot above, the regression line captures the general trend between the two variables. However, upon closer examination we see that the the trend has some non-linear trend in it. To account for it we add variables that are non-linear transformations of the original variable and regress linearly on these multiple parameters.

```
> # price=theta_1+theta_2*carat+theta_3*carat∧2+theta_3*carat∧3+epsilon
> M2=lm(price∼carat+I(carat∧2)+I(carat∧3),diamSmall);
> theta=coef(M2); theta # least squares parameter estimates \hat\theta
```

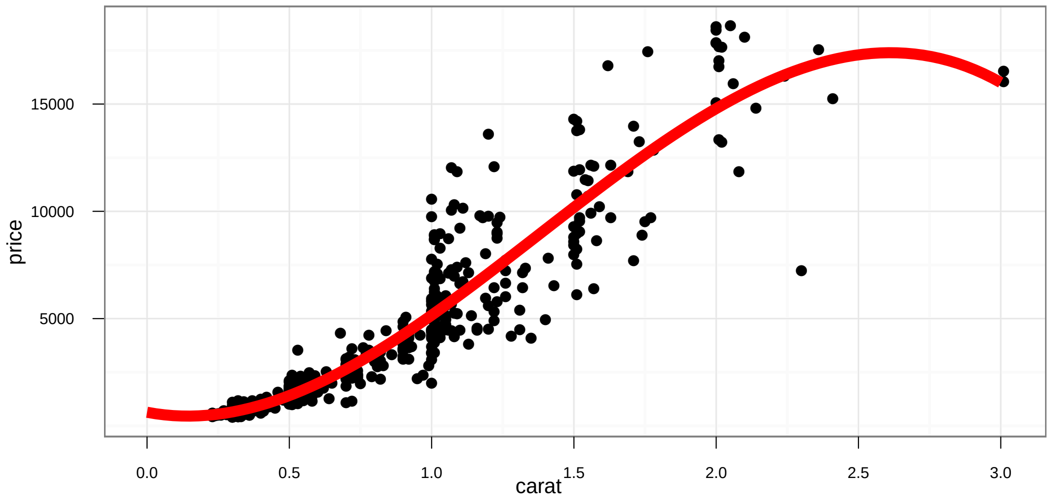
```
(Intercept)        carat   I(carat∧2)   I(carat∧3)
   633.7231   -2538.7808    9316.7619   -2254.2733
```

```
> cX=seq(0,3,length=100); # carat grid for plotting predictions
> cY=theta[1]+theta[2]*cX+theta[3]*cX∧2+theta[4]*cX∧3;
> print(ggplot(diamSmall,aes(carat,price))+geom_point()+
+       geom_line(aes(x=cX,y=cY),size=2,color=I("red")))
> summary(M2)$r.squared # R∧2 coefficient of determination
```

```
[1] 0.8695391
```

```
> mean(residuals(M2)∧2) # RSS (sum of squared residuals) normalized by n
```

```
[1] 2279123
```



This model seems to be a better fit based on the scatter plot overlayed with the regression model. Note that the regression model is a linear surface in the space $(1, X', X'^2, X'^3)$ and is a non-linear curve in the space of the original carat variable X'. The better fit can also be confirmed by comparing the R^2 scores and the RSS values.

As another example consider the code below which adds diamond color as an explanatory variable to carat. Note that R detects it as a categorical variable and creates multiple binary variables in the linear combination as described in the first section above.

```
> M3=lm(price~carat+color,diamSmall); # adding a categorical variable: color
> theta=coef(M3); theta # note coefficients corresponding to multiple binary variables
```

```
(Intercept)       carat      colorE      colorF      colorG      colorH
 -1996.3660   8233.5031   -190.5149   -521.6163   -321.3877   -684.9762
     colorI      colorJ
 -1323.6239  -2681.5287
```

```
> summary(M3)$r.squared # R^2 coefficient of determination
```

```
[1] 0.8666104
```

```
> mean(residuals(M3)^2) # RSS (sum of squared residuals) normalized by n
```

```
[1] 2330287
```

As we see both the R^2 and the RSS improve considerably indicating that adding the diamond color the the linear model (using multiple binary variables) provides a better fit to the data.

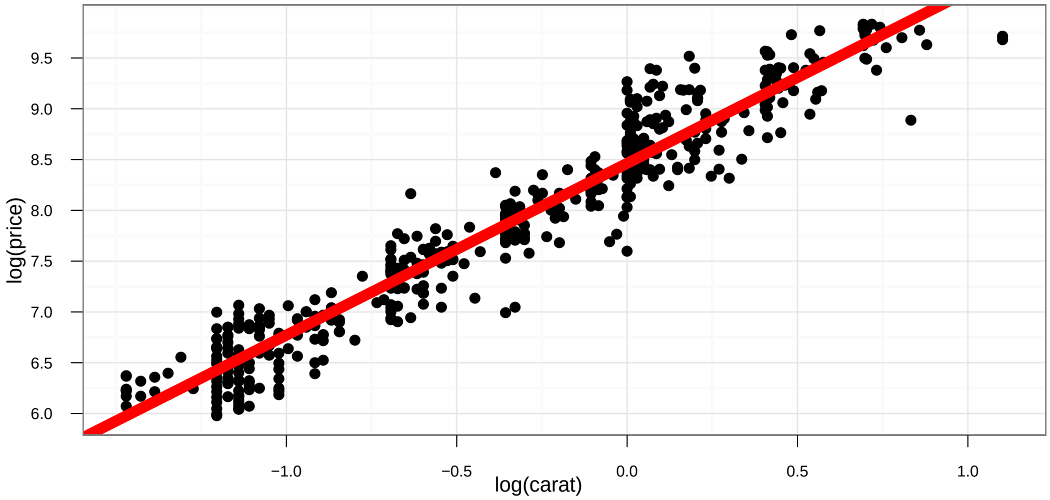
As the following figure shows perhaps the best fit is obtained from transforming both the carat and price variables with a log function.

```
> M4=lm(log(price)~log(carat),diamSmall); # adding a categorical variable: color
> theta=coef(M4); theta
```

```
(Intercept)  log(carat)
   8.462654    1.691301
```

```
> print(ggplot(diamSmall,aes(log(carat),log(price)))+geom_point()+
+       geom_abline(intercept=theta[1],slope=theta[2],size=2,color=I("red")))
> summary(M4)$r.squared # R^2 coefficient of determination
```

```
[1] 0.9291413
```



R allows for considerable flexibility in the specification of the formula in `lm`. In the examples above, we saw formulas where additive terms are added with plus signs and implicitly including the constant term. To remove the constant term add +0 to the formula. To encode interaction (by product) of two terms use the : operator. To add all possible products between two groups of variables use the * operator (including products with the constant term). To add higher powers use the ˆ operator. Use the as-is function `I()` to escape these symbols and interpret them literally. Variables can be dropped using the - symbol. We give below some examples.

formula	model	formula	model
$y \sim x$	$y = \theta_1 + \theta_2 x$	$y \sim x + 0$	$y = \theta_1 x$
$y \sim x + z$	$y = \theta_1 + \theta_2 x + \theta_3 z$	$y \sim x : z$	$y = \theta_1 + \theta_2 xz$
$y \sim x * z$	$y = \theta_1 + \theta_2 x + \theta_3 z + \theta_4 xz$	$y \sim x + z + x : z$	$y = \theta_1 + \theta_2 x + \theta_3 z + \theta_4 xz$
$y \sim (x + z + w)\hat{}2$	$y = \theta_1 + \theta_2 x + \theta_3 z + \theta_4 w + \theta_5 xz + \theta_6 xw + \theta_7 zw$	$y \sim I(x + z)$	$y = \theta_1 + \theta_2 (x + z)$
$y \sim (x + z + w)\hat{}2 - zw$	$y = \theta_1 + \theta_2 x + \theta_3 z + \theta_4 w + \theta_5 xz + \theta_6 xw$	$\log(y) \sim \log(x)$	$\log(y) = \theta_1 + \theta_2 \log(x)$

Deriving the Distribution of $\hat\theta$

Since the MLE $\hat\theta$ has a closed form we can derive useful expressions for the distribution of $\hat\theta$ conditioned on the training X values in the matrix $\mathbf{X}$ (note that no assumption is made on $p(X)$)

$$\begin{aligned}
&\mathsf{E}(\hat\theta|\mathbf{X}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \mathsf{E}(\mathbf{Y}|\mathbf{X}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\theta = \theta \\
&\mathsf{Var}(\hat\theta|\mathbf{X}) = \mathsf{Var}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}|\mathbf{X}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \mathsf{Var}(\mathbf{Y}|\mathbf{X})((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)^\top = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2 I\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}
\end{aligned}$$

where we use the fact that $(A^\top)^{-1} = (A^{-1})^\top$. Since the MSE is the sum of the bias squared (which is zero) and the variance we have that the $d \times d$ MSE matrix $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$ decreases to the zero matrix as $n \to \infty$ (the entries of $\mathbf{X}^\top\mathbf{X}$ get bigger as n increases and therefore the entries of $(\mathbf{X}^\top\mathbf{X})^{-1}$ decrease). We also see that the MSE is linearly related to the inherent noise level σ^2 of ϵ. Conditioned on $\mathbf{X}$ the vector $\mathbf{Y}$ is multivariate normal. Since $\hat\theta$ is a linear function of $\mathbf{Y}$ it is also multivariate Gaussian (since a linear transform of a multivariate normal RV is a multivariate normal RV). We therefore have

$$\hat\theta|\mathbf{X} \sim N(\theta, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}). \tag{4}$$

Estimating σ^2

In some cases we are interested not only in estimating θ but also in estimating σ^2. We define the vector of residuals i.e., differences between training $Y^{(i)}$ values and predictions $\hat Y^{(i)} = \hat\theta^\top X^{(i)}$

$$\mathbf{e} \overset{\text{def}}{=} \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}.$$

We note that $\mathsf{E}(\mathbf{e}|\mathbf{X}) = \mathbf{X}\theta - \mathbf{X}\,\mathsf{E}(\hat\theta|\mathbf{X}) = 0$ and that $\mathsf{Var}(\mathbf{e}|\mathbf{X}) = (\mathbf{I}-\mathbf{H})\,\mathsf{Var}(\mathbf{Y})(\mathbf{I}-\mathbf{H})^\top = (\mathbf{I}-\mathbf{H})\sigma^2\mathbf{I}(\mathbf{I}-\mathbf{H})^\top = \sigma^2(\mathbf{I}-\mathbf{H}^\top-\mathbf{H}+\mathbf{H}\mathbf{H}^\top) = \sigma^2(\mathbf{I}-\mathbf{H})$ since $\mathbf{H}$ is symmetric (verify!) and $\mathbf{H}\mathbf{H}^\top = \mathbf{H}\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{H}$. Since $\mathbf{e}$ is a linear function of $\mathbf{Y}$ which conditioned on $\mathbf{X}$ is Gaussian $\mathbf{e}|\mathbf{X}$ is also a Gaussian and we have

$$\mathbf{e}|\mathbf{X} \sim N(0, \sigma^2(\mathbf{I}-\mathbf{H})).$$

The residual sum of squares (RSS) defined above may be simplified to

$$\text{RSS}(\hat\theta) = \mathbf{e}^\top\mathbf{e} = \mathbf{Y}^\top(\mathbf{I}-\mathbf{H})^\top(\mathbf{I}-\mathbf{H})\mathbf{Y} = \mathbf{Y}^\top(\mathbf{I}-\mathbf{H})\mathbf{Y} = \mathbf{Y}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}\hat\theta = \mathbf{Y}^\top\mathbf{Y} - \hat\theta^\top\mathbf{X}^\top\mathbf{Y}$$

where we used again the fact that $\mathbf{H}$ is symmetric and $(\mathbf{I}-\mathbf{H})^\top(\mathbf{I}-\mathbf{H}) = (\mathbf{I}-\mathbf{H})$ (proven above).

Proposition 1. *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a matrix of rank d. Then $S^2 \overset{\text{def}}{=} RSS(\hat\theta)/(n-d)$ is an unbiased estimator of σ^2 which is independent of $\hat\theta$. Furthermore, $RSS(\hat\theta)/((n-d)\sigma^2) = S^2/\sigma^2 \sim \chi^2_{n-d}$.*

Proof. Using the lemma below, the fact that $\mathbf{H}$ is a projection matrix, the trace of a projection matrix is its rank, and the fact that $\mathsf{E}(\mathbf{Y})$ is already projected by $\mathbf{H}$ (see [1] for more details on these statements as well as a more detaled proof) to obtain

$$\begin{aligned}
&\mathsf{E}(\text{RSS}(\hat\theta)) = \mathsf{E}(\mathbf{Y}^\top(\mathbf{I}-\mathbf{H})\mathbf{Y}) = \sigma^2\text{tr}(\mathbf{I}-\mathbf{H}) + \mathsf{E}(\mathbf{Y})^\top(\mathbf{I}-\mathbf{H})\,\mathsf{E}(\mathbf{Y}) = \sigma^2\text{tr}(\mathbf{I}-\mathbf{H}) = \sigma^2(n-d). \\
&\mathsf{Cov}(\hat\theta, \mathbf{Y}-\mathbf{X}\hat\theta) = \mathsf{Cov}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}, (\mathbf{I}-\mathbf{H})\mathbf{Y}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathsf{Var}(\mathbf{Y})(\mathbf{I}-\mathbf{H}) = 0
\end{aligned}$$

and therefore $\hat{\theta}$ is independent of $\mathbf{e}$ (for two normal RVs zero covariance implies independence) and also of $\text{RSS}(\hat{\theta}) = \mathbf{e}^\top\mathbf{e}$. Finally,

$$\text{RSS}(\hat{\theta}) = \mathbf{Y}^\top(\mathbf{I}-\mathbf{H})\mathbf{Y} = (\mathbf{Y}-\mathbf{X}\theta)^\top(\mathbf{I}-\mathbf{H})(\mathbf{Y}-\mathbf{X}\theta) = \mathbf{e}^\top(\mathbf{I}-\mathbf{H})\mathbf{e}$$

which is a quadratic form with a rank $n-p$ matrix and therefore correspond to χ^2_{n-d} distribution. □

Lemma 1. *Let X be a random (column) vector with mean μ and variance Σ. Then $\mathsf{E}(X^\top AX) = tr(A\Sigma) + \mu^\top A\mu$.*

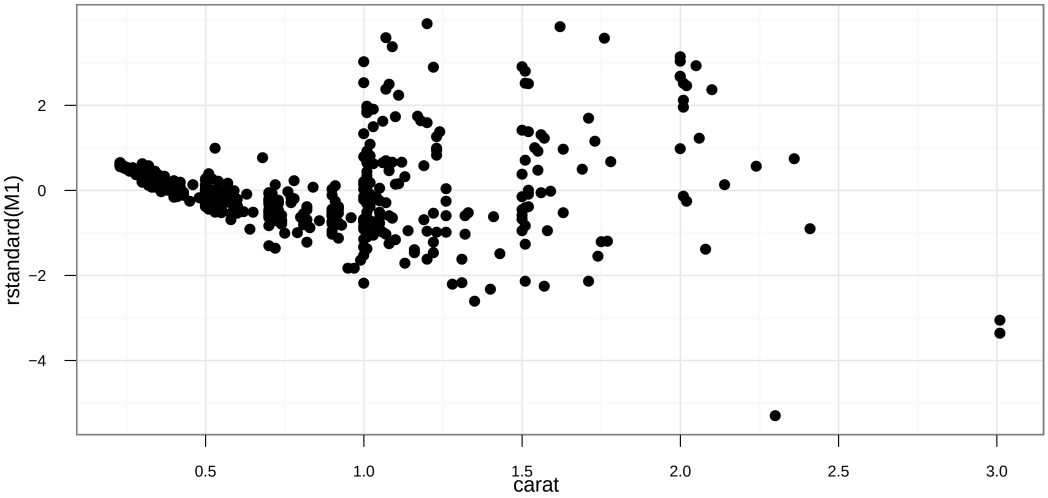
Proof. Using the fact that $\text{tr}(AB) = \text{tr}(BA)$

$$\mathsf{E}(X^\top AX) = \mathsf{E}(\text{tr}(X^\top AX)) = \mathsf{E}(\text{tr}(AXX^\top) = \text{tr}(\mathsf{E}(AXX^\top)) = \text{tr}(A\,\mathsf{E}(XX^\top)) = \text{tr}(A\Sigma) + \text{tr}(A\mu\mu^\top) = \text{tr}(A\Sigma) + \mu^\top A\mu$$

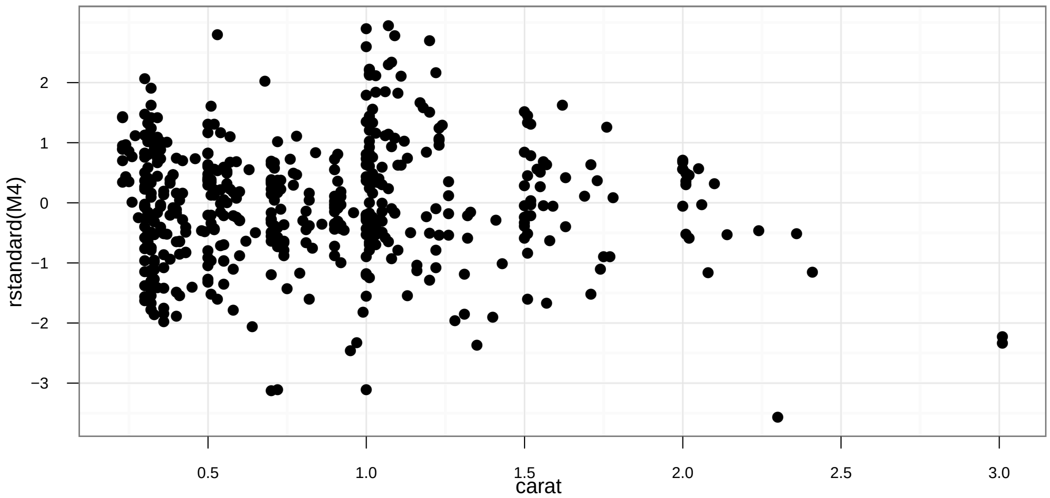
□

Since $\mathbf{e}|\mathbf{X} \sim N(0,\sigma^2(\mathbf{I}-\mathbf{H}))$ we have that the standartized residuals have the distribution $(\mathbf{I}-\mathbf{H})^{-1}\mathbf{e}|\mathbf{X} \sim N(0,\sigma^2)$. It is useful to examine graphically whether the standartized residuals indeed follow a normal distribution as the linear regression model suggests. If they do, this can lead to a graphical estimation of σ^2. If they don't we may conclude that the linear regression model is incorrect (although potentially useful nevertheless). As the following figures show, the distribution of the standartized residual conditioned on X is clearly non-normal for the M1 model above. For different X values we get a different standatized residual distribution which in some cases is not even centered at 0 indicating a systematic underprediction and overprediction trends. In the case of the M4 model the distribution of the standartized residuals conditioned on X is rather similar to the theoretical $N(0,\sigma^2)$ regardless of the value of X (except perhaps for high carats where the assumption breaks down). We therefore conclude that the linear regression assumption is much more accurate for model M4 that it is for model M1. This fact is in agreement with the scatter plot and regression line figures shown in the previous section.

```
> print(qplot(carat,rstandard(M1),data=diamSmall))
```



```
> print(qplot(carat,rstandard(M4),data=diamSmall))
```



Confidence Intervals and Hypothesis Tests

Using the above analysis we can use pivotal quantities to obtain large sample (z-value) or small sample (t-value) confidence intervals or hypothesis tests. Specifically, since $\hat{\theta}|\mathbf{X} \sim N(\theta, \sigma^2(\mathbf{X}\mathbf{X}^\top)^{-1})$, the t-statistic

$$\frac{\theta_j - \hat{\theta}_j}{\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}\mathrm{RSS}(\hat{\theta})/(n-d)}} \sim T_{n-d} \tag{5}$$

can be used to obtain confidence intervals or hypothesis tests concerning θ_j. For example,

$$P\left(\hat{\theta}_j - t_{\alpha/2}\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}\mathrm{RSS}(\hat{\theta})/(n-d)} \;\le\; \theta_j \;\le\; \hat{\theta}_j + t_{\alpha/2}\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}\mathrm{RSS}(\hat{\theta})/(n-d)}\right) \ge 1-\alpha$$

where $t_{\alpha/2}$ is the $\alpha/2$ quantile of the t distribution with $n-d$ degrees of freedom. It is important to realize that confidence intervals or hypothesis tests such as the ones based on (5) should be interpreted with respect to fixed $\mathbf{X}$.

Gauss Markov Theorem

The following theorem provides a strong motivation for using the least squares estimator $\hat{\theta}$ as opposed to a different one.

Theorem 1 (Gauss-Markov)**.** *$\hat{\theta}$ is BLUE (best linear unbiased estimator) i.e. among all unbiased estimators that are linear functions of $Y^{(1)}, \ldots, Y^{(n)}$ it has the smallest variance*[1] *(conditioned on $\mathbf{X}$).*

Proof. We already know that $\hat{\theta}$ is linear and unbiased. Let $\tilde{\theta} = \mathbf{AY}$ be a any other unbiased linear estimator of θ. Then to prove the theorem we need to show that $\mathsf{Var}(\tilde{\theta}) - \mathsf{Var}(\hat{\theta})$ is positive semidefinite. Since

$$\tilde{\theta} = \mathbf{AY} = (\mathbf{D} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)(\mathbf{X}\theta + \boldsymbol{\epsilon}) = (\mathbf{DX} + \mathbf{I})\theta + (\mathbf{D} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\boldsymbol{\epsilon}$$

(for some matrix $\mathbf{D}$) for $\tilde{\theta}$ to be unbiased we must have $\mathbf{DX} = 0$. We then have

$$\begin{aligned}\mathsf{Var}(\tilde{\theta}) &= \mathsf{E}(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^\top = \mathsf{E}(\mathbf{D} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top(\mathbf{D}^\top + \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}) \\ &= \sigma^2(\mathbf{D}\mathbf{D}^\top + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{DX}(\mathbf{X}^\top\mathbf{X})^{-1} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}^\top) \\ &= \sigma^2\mathbf{D}\mathbf{D}^\top + \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} \quad \Rightarrow \quad \mathsf{Var}(\tilde{\theta}) - \mathsf{Var}(\hat{\theta}) \quad \text{is positive semi-definite.}\end{aligned}$$

□

Nevertheless, there has been a recent surge of interest in biased estiamtors for linear regression that obtain lower MSE than the least squares estimators, in particular in high dimensional cases. Note that this does not contradict the above result as it states optimality among unbiased estimators only.

[1] A variance matrix A is better than a variance matrix B here if $A \ge B$ which is defined as $A - B$ is positive semidefinite.

References

[1] G. A. Seber and A. J. Lee. *Linear Regression Analysis.* Wiley Interscience, 2003.